POLITECNICO DI MILANO

POLITECNICO
MILANO 1863

WHAT MAKES CITIES ATTRACTIVE FOR PEOPLE?

Investigating Public Policies and Online Discussions Through a Machine Learning Model using Text Mining & Sentiment Analysis

Master Thesis

Yasser Elsedawy

2020

WHAT MAKES CITIES ATTRACTIVE FOR PEOPLE?

Investigating Public Policies and Online Discussions Through a Machine Learning Model using Text Mining & Sentiment Analysis

Master Thesis

Presented to:
Politecnico Di Milano
Department of Architecture, Built environment and Construction Engineering (ABC)
Via Giuseppe Ponzio, 31, 20133 Milano MI

Supervisor:
Roberta Capello
Co-Supervisor:
Andrea Caragliu

By:
Yasser Emam Ragab Mohamed Elsedawy

2020
Milano

Table of Contents

List of Tables

List of Figures

Acknowledgement

I would like to show my gratitude to my supervisor Professor Roberta Capello and my Co-Supervisor Professor Andrea Caragliu for the continuous support, patience, guidance and knowledge during my Master study and research. Moreover, I would like to show my appreciation to Politecnico di Milano for the support and funding to carry out this degree. Lastly, I wish to acknowledge the help, support and great love of my parents: Emam and Amina, my girlfriend: Anna, and my brother: Karim as they are the reason for always moving forward and this work would not have been possible without their contribution.

Abstract

Traditionally, the attractiveness of a city is measured by quantitative indicators. Nonetheless, the role of qualitative interpretations of online discussions has been nominal in academic studies, a shortcoming this dissertation addresses. Therefore, the aim of this work is to identify whether the analysis of online discussions can strengthen the understanding of cities' perception for inhabitants, tourists, investors, and high skilled immigrants. We investigate to what extent and in what ways the analysis of online discussions can identify issues and deepen the understanding of people's behaviors. We do so to support decision-making for policymakers.

Quantitative and qualitative analysis is carried out using the Python programming language. Firstly, the quantitative analysis is performed using regression modeling on a dataset of 42 European cities. The findings of the analysis demonstrate that cities' attractiveness can be measured based on the average net salary, the average rental price in the city center, and the number of e-commerce active population. In addition to quantitative analysis, the qualitative analysis is performed using text-mining and sentiment analysis, in which 148 million words are collected, processed, analyzed, and clustered into positive, negative, and neutral sentiments. The findings of the analysis of online discussions endorse the possibility to identify people's concerns, needs, and preferences, some of which are climate change, gender identity, Brexit, gun control, and immigration. The significance of this study is that it recommends the use of text-mining and sentiment analysis to support decision-making by enabling policymakers to extract relevant, valid, and comprehensive information. Furthermore, it can monitor people's behavior, and opinions during the implementation of public policies.

Chapter 1 **Background**

1.1 Introduction

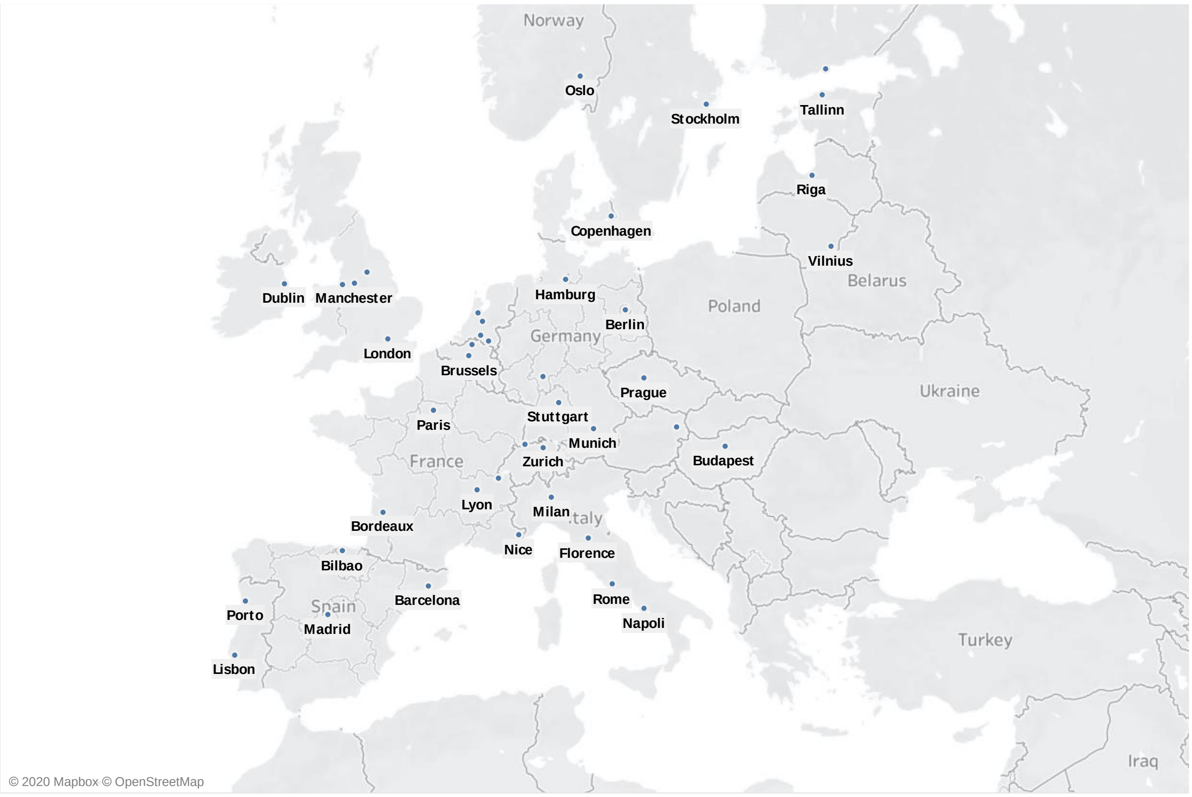
The satisfaction of cities' stakeholders is among the top priorities for decision-makers. Cities are being challenged to compete with each other to attract both people and investments, whilst cities are expanding physically, they currently incorporate around half of the world population. According to the **(United Nations, 2018**)**,** by 2050, more than two-thirds of the world population will be living in cities. These urban areas generate over 80% of global GDP **(World Bank, 2020),** making them influential in driving wealth and development. Nevertheless, this speed and magnitude of urbanization bring challenges, which have to be encountered with accelerated demand for affordable housing, connected transport systems, and jobs, among other needs. Almost half a billion urban residents are vulnerable to storms and sea-level rise as they live in coastal areas endangering $4.7 trillion in assets exposed to coastal floods. Building safe, smart, resilient, and sustainable cities require enormous efforts, and that constitute effective policies. Governments and municipalities have a crucial responsibility for taking actions to shape the future of cities' development and to create opportunities for all **(Global Business Outlook, n.d.).** Additionally, cities' role against climate change is essential as they consume 2/3 of the world's energy. As cities develop, their exposure to climate risks increases. Sustainable Development Goal (SDG 11) acknowledges the important role of cities in the global policy agenda and sets the charge for them to be both sustainable and inclusive in moving the world towards a resilient future. For all the benefits they confer, cities are as well a major cause of pollution, heat, and waste. Presently, cities account for around 70% of the planet's energy-related carbon dioxide emissions **(INTERNATIONAL ENERGY AGENCY, 2016).** Furthermore, increases in population and motorized vehicles, with manufacturing and fossil fuel burning, have given rise to low air quality in many urban areas. The majority of city residents breathe air that exceeds the World Health Organization's guidelines for safe exposure to fine particulate pollution (PM2.5), one of the most dangerous urban threats to human health.

Moreover, cities experience the urban heat island effect, a rise in local temperature as a result of high-density living conditions, and the clearing of natural land cover. This heat, met by rising temperatures from climate change, can lead to a lot of health complications. The Urban Environment and Social Inclusion Index (UESI), provides an unprecedented level of detail into the state of the environment and social equity in cities. Using high-resolution and large-scale data, the UESI reveals how residents living in the same city often experience urban environments in vastly different and unequal ways. In total, 90 out of 162 cities are disproportionately troubling lower-income populations with air pollution, urban heat, and lack of accessible transport **(Data-Driven Lab, 2018)**. City residents with lower income are as well less likely to have the means to adapt to these challenges, which can conserve and worsen inequality. Due to that, cities are failing on SDG 11 to provide inclusive and sustainable urban growth.

To reduce pollution and increase accessibility for residents, integrated transit systems, and sustainable buildings can be introduced, whilst the addition of green spaces can help to reduce urban heat. With the right policies and programs, cities have a strong likelihood of achieving SDG 11. Thus, 2020 is a critical year for climate action. It is the year the Paris Agreement goes into effect and nations begin implementing, as well as increasing, their climate plans. According to the Global Climate Action from Cities, Regions, and Businesses 2019 report, more than 6,000 subnational actors and 1,500 businesses in ten high-emitting countries around the world have committed to emission reduction targets that could lead to an additional (1.4-2.2) Carbon footprint in 2030, approximately 4% of today's global emissions **(NewClimate Institute, 2019).** This number is on top of what national governments have already pledged to the Paris Agreement. Cities have the challenge of creating greater sustainability and inclusion. Cities with urban innovation, great employment opportunities, and lower infrastructure costs, are estimated to have savings of $17 trillion by 2050 **(World Resources Institute, 2018)**. What comes after will determine how close cities can stay within their warming targets and turn away the worst effects of climate change.

Besides implementing the right policies and programs, it's essential for cities to understand people's opinions and sentiments regarding the current obstacles that cities have to face. Those sentiments could be extracted from online discussions on social media and online platforms. In an attempt to support the retrieval of crucial information that may prevent current policies from achieving the expected outcomes and reasonably facilitate the evaluation of future policies. Therefore, in this dissertation, we attempt to create a link between decision-makers and cities' stakeholders. We do so to eliminate the existing dialogue barriers. The main hypothesis of this work is that cities, which have access to the information and insights from online discussions, and leverage on it, will report greater attractiveness than cities that do not. Whilst businesses in the private sector have been implementing data science tools for many years, most organizations in the public sector have not leveraged yet on that knowledge. With this work, we attempt to contribute by covering a wide range of case studies. The work is carried out using data science capabilities in order to benefit from data generated from cities. Therefore, originate a real impact on the built environment and people's lives.

The work consists of two main parts. The first part is addressing quantitative analysis and it is carried out in chapters two and three. In chapter two, data science functionalities and Python programming language are introduced. Within Python, different libraries are applied, such as Numpy/Scipy for numerical operations, Pandas for data manipulation and analysis, MatPlotLib, and Seaborn for plotting, Scikit-Learn for Machine learning modeling, VADER for Sentiment analysis, and NLTK for natural language processing. Furthermore, in chapter three, we start to investigate a sample of 42 cities within the European Union. Such selection is based on similar regulations, policies, and easy travel between them. The selected cities are represented in the following map: Paris, London, Milan, Madrid, Munich, Berlin, Barcelona, Stockholm, Hamburg, Frankfurt, Rome, Amsterdam, Stuttgart, Brussel, Vienna, Dublin, Naples, Budapest, Lisbon, Oslo, Lyon, Copenhagen, Leeds, Rotterdam, Zurich, Helsinki, Manchester, Prague, Antwerp, Porto, Utrecht, Bordeaux, Bilbao, Eindhoven, Basel, Liverpool, Geneva, Nice, Florence, Riga, Vilnius, Tallinn.



Map based on Longitude (generated) and Latitude (generated). The marks are labeled by City. Details are shown for City.

Figure 1. 1 *A Geographical map showing the selected cities locations*

Moreover, chapter three highlights and analyzes 16 main indicators within each city. The data for those indicators were collected across a number of different platforms, such as Eurostat, Istat, Statista, World Bank, World Population Review, Government platforms, Teleport, OECD, and Numbeo.

The indicators are as following: foreign_visitors, local_visitors, employment, internet_access, income_households, e_commerce_active, gdp, population, foreign_population, rental_price_center, avg_net_salary, air_quality, urban_greenery, life_expectancy, startups_number, taxi_cost_center.

Each of the previously mentioned indicators was chosen based on a specific ground, in order to investigate their significance concerning cities' attractiveness, those grounds are explained below:

- **Urbanization and the built environment**: as it is the infrastructure of cities, that includes transportation, roadways, buildings, and land-use. Built environment design and development can help mitigate climate change, support adaptation, and improve the environment and public health **(National Environment Health Association, n.d.)**, it can also lead people to flee out of their home city to relocate to another, according to a report by **(C40 Cities Climate Leadership Group, n.d.)**, climate action has a range of wider benefits for the health and prosperity of cities and their citizens, from green jobs and growth, to active, happier lives and cleaner air and water. This has an immediate, tangible impact on people's lives and the built environment. Moreover, cities are more productive than in rural areas. They provide efficient infrastructure, services, communications, and skilled labor forces **(URBAN Intergroup at the European Parliament, 2011)**. They can achieve the economies of scale, agglomeration, and urbanization. Cities are the driving forces of national economic development. Therefore, the air quality index, urban greenery index, and life expectancy are selected to represent urbanization and the built environment.
- **Economic growth** means an increase in GDP and an increase in the value of national output, income, and expenditure. Essentially the benefit of economic growth is higher living standards **(Economics Help, 2019**), higher real incomes, and the ability to devote more resources to areas like health care and education. Thus, Economic growth is one of the best estimators of reasons to relocate to another city. In general, people leave their homes looking for better opportunities and quality of life. However, when cities grow to certain levels, they start to produce negative impacts such as overcrowding, congestion, and pollution. The role of cities is to maximize positive externalities and to minimize the negative impacts. Therefore, GPD and Employment are selected to represent the economic growth of cities.
- **Real Estate, housing and neighborhood** conditions for students have a huge impact on their progress in life. This insight has led practitioners and policymakers to emphasize the role of place as they design new interventions. Evidence also demonstrates that as students progress in their education, housing conditions continue to affect their choices **(U.S.**

Department of Housing and Urban Development, n.d.). Housing needs affect students' success and progress. When students cannot cover their living expenses, they will often try to look to move or relocate to a more convenient and affordable place, or to a more affordable city. According to **(World Economic Forum, 2019)**, cities are growing at an unprecedented rate, presenting an incredible opportunity for the development of local economies. However, their residents are in need of good, affordable housing, and this remains a challenge around the world. Enabling environment for affordable housing can be developed with the right infrastructure, investment, and macroeconomic policies targeted towards social and financial inclusion. The challenge of affordability requires not just short-term fixes but also long-term strategies. Solutions will need to address both the supply side and the demand side of the housing market and involve public-sector, private-sector, and non-profit stakeholders. Therefore, the rental price of a medium size apartment, the average net salary, the cost of a taxi from the main airport or train station to the city center, and the level of income for households are selected.

- **Digitalization** can improve the efficiency of cities by minimizing the waste of time and resources. It can also increase productivity and drive economic activities. Therefore, internet access and e-commerce active population were selected.
- **Ease of doing business as** it means the regulatory environment is more favorable to the starting and operation of a local firm. Therefore, the number of startups in each city is selected.
- **Demographic characteristics:** population and foreign population, are selected.
- **The attractiveness of cities:** the number of foreign visitors and local visitors per year were chosen to measure the inflows demand in each city.

In chapter three we evaluate the relationship between the selected indicators. The data is tested for correlation and a causality effect. A significant relationship between cities' inflows and between the indicators is investigated. Regression analysis is implemented, both linear simple regression and multiple linear regression. Moreover, a prediction model is created based on the data and evaluated with different methodologies, in order to attempt to predict the number of people attracted by each city.

The second part of the work is addressing qualitative analysis and is carried out in chapter four. In this part, text mining and sentiment analysis are performed. The purpose is to investigate how the analysis of online discussions, using sentiment analysis, could contribute to the understanding of cities' attractiveness. The motive behind the use of sentiment analysis is due to the fact that it provides people's opinions about a service, or a policy, which helps policymakers enhance that service or policy's quality. Sentiment analysis creates a link between decision-makers and cities' stakeholders, hence, enables business decisions.

Since all stakeholders such as inhabitants, tourists, immigrants, or business owners, interact online together, in addition, it is currently not possible to extract online data about an individual stakeholder type. Thus, in this work, we try to identify the topics discussed by analyzing the context of the comments obtained. Text mining is applied to 10 cities chosen from the dataset of 42 cities. Reddit, an American social news aggregation, web content rating, and discussion website is used. To collect the data, we search on Reddit for comments mentioning selected keywords in the past 2 years. The selected keywords are related to the cost of living in a certain city, the quality of life in that city, investing opportunities, and studying opportunities in that city. For instance, the keywords are as follows for the city of London (Invest in London, Study in London, Quality of Life in London, Cost of Living in London). The text collected is taking into account only English speakers.

VADER (Valence Aware Dictionary and sEntiment Reasoner) is used to apply sentiment analysis on the collected text. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER is fully open sourced under the MIT License **(Hutto & Gilbert, 2014).**

After applying sentiment analysis, the text for each city is interpreted and transformed into insights. In addition, an investigation is carried out in order to identify the relationship between the two parts of the work, the quantitative and the qualitative analysis. We do so to examine in what ways qualitative analysis can support in generating more insights into the reasons why some cities are more attractive than others.

Additionally, word clustering is applied. Word Clustering is a technique for partitioning sets of words into subsets of semantically similar words. It automatically analyzes text data to determine cluster words for a set of text. Lastly, the results are demonstrated as clusters of words in both sentiments, positive and negative. In the final chapter of the work, chapter five, conclusions, and recommendations are established.

1.2 Related work

Compared to the increased research on organizational attractiveness of firms in the private sector in the past 10 years, a scarcer number of academic studies on the attractiveness of a city has been introduced. Moreover, there have been nominal studies addressing online discussions about the attractiveness of a city.

In order to obtain the number of academic studies carried out in the last 10 years, Data Science capabilities are applied to Google scholar. The data based on city branding and company branding is collected and compared. The results in Figures 1.1 and 1.2 demonstrate that the maximum number of studies conducted on city branding is inferior to the minimum number of studies carried out on company branding. The data shows a wide gap between the two topics. In addition, it is evident that the gap trend is decreasing between them, as more studies have been moving towards the concept of city branding, less has been moving towards company branding.

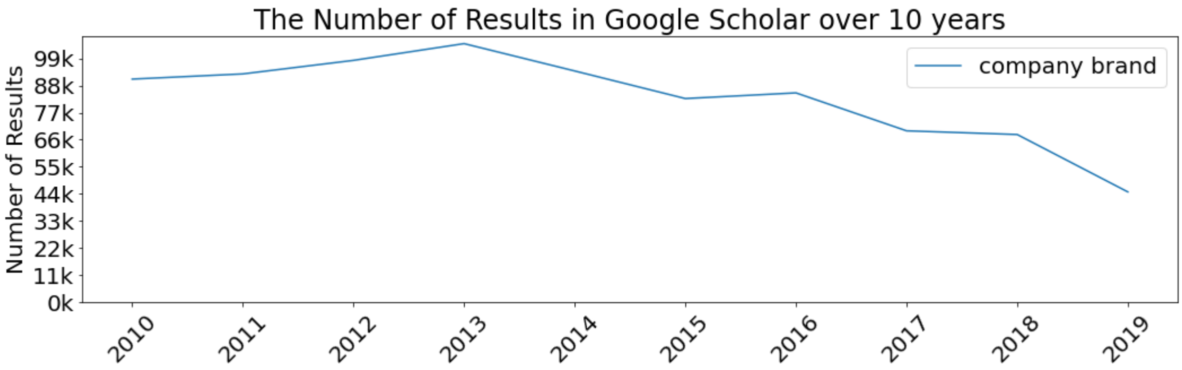


Figure 1. 2 *Google Scholar results for company branding research in the last 10 years*

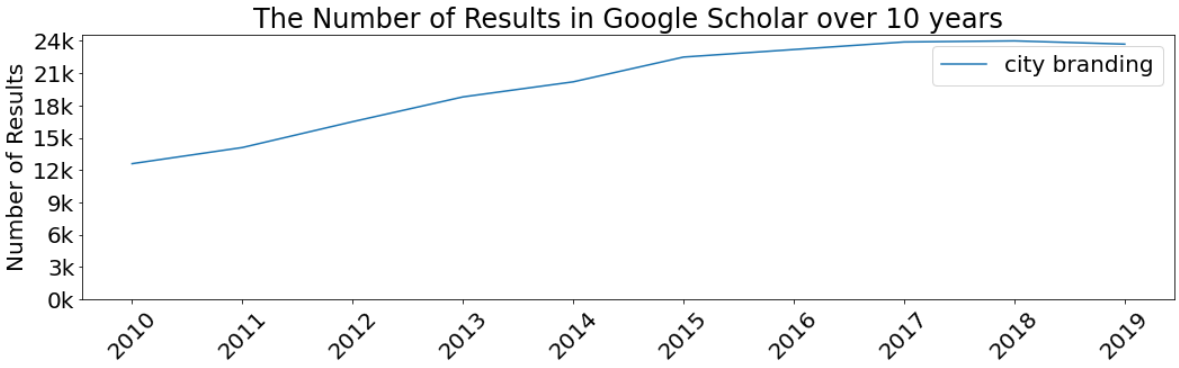


Figure 1. 3 *Google Scholar results for city branding research in the last 10 years*

In the following figure, different keywords are used to investigate the number of studies carried out on the concept of cities' attractiveness. The results are showing that more research was made using the keyword city branding than both urban attractiveness and cities' attractiveness.

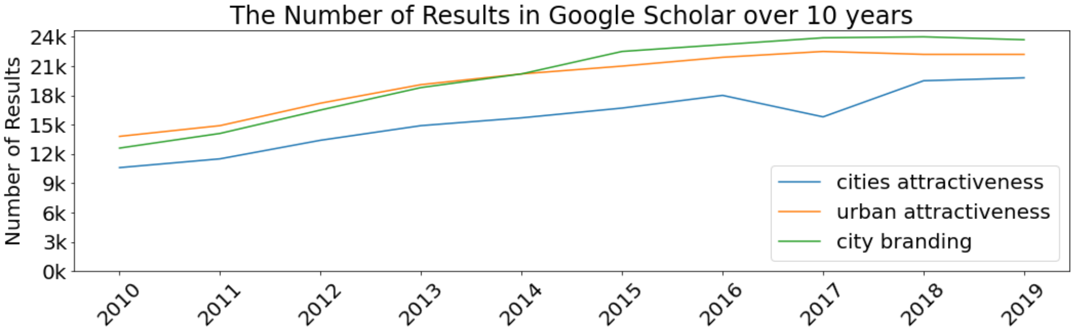


Figure 1. 4 *Google Scholar results for city branding, cities attractiveness and urban attractiveness research in the last 10 years*

In figure 1.5, sentiment analysis keywords are added to further investigate the studies. The outcome demonstrates that the "cities' attractiveness" keyword, when associated with the "sentiment analysis" keyword, results in a higher number of studies than of "urban attractiveness" and "city branding" keywords. Overall, it is evident that there is a minimum number of studies carried out on the notion of sentiment analysis regarding cities' attractiveness in comparison to other research areas.

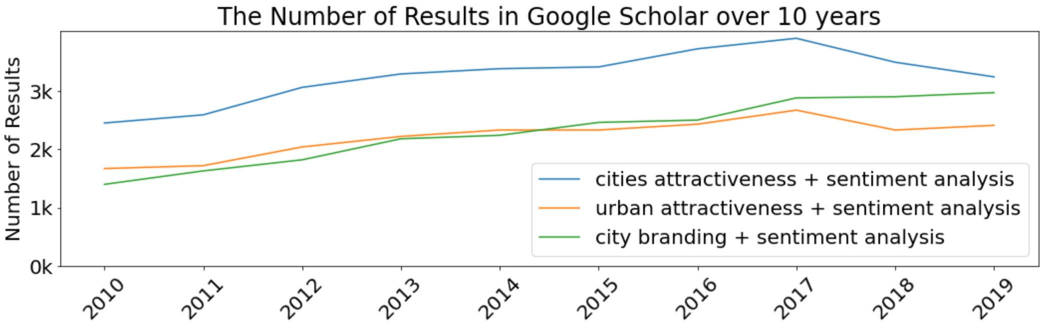


Figure 1. 5 *Google Scholar results for city branding, cities attractiveness and urban attractiveness with sentiment analysis keyword research in the last 10 years*

In the following figures, 420 research papers are collected from the period of 2010 to 2020 (see appendix), those papers are collected based on keywords corresponding to city attractiveness or city branding. The main goal is to extract the most frequent words used in those papers over the past 10 years. In figure 1.4, the most frequent words used in titles of academic studies are visualized. For instance, urban, city, cities, smart development, spatial, and sustainable. We do so in order to understand the main focus of those studies.

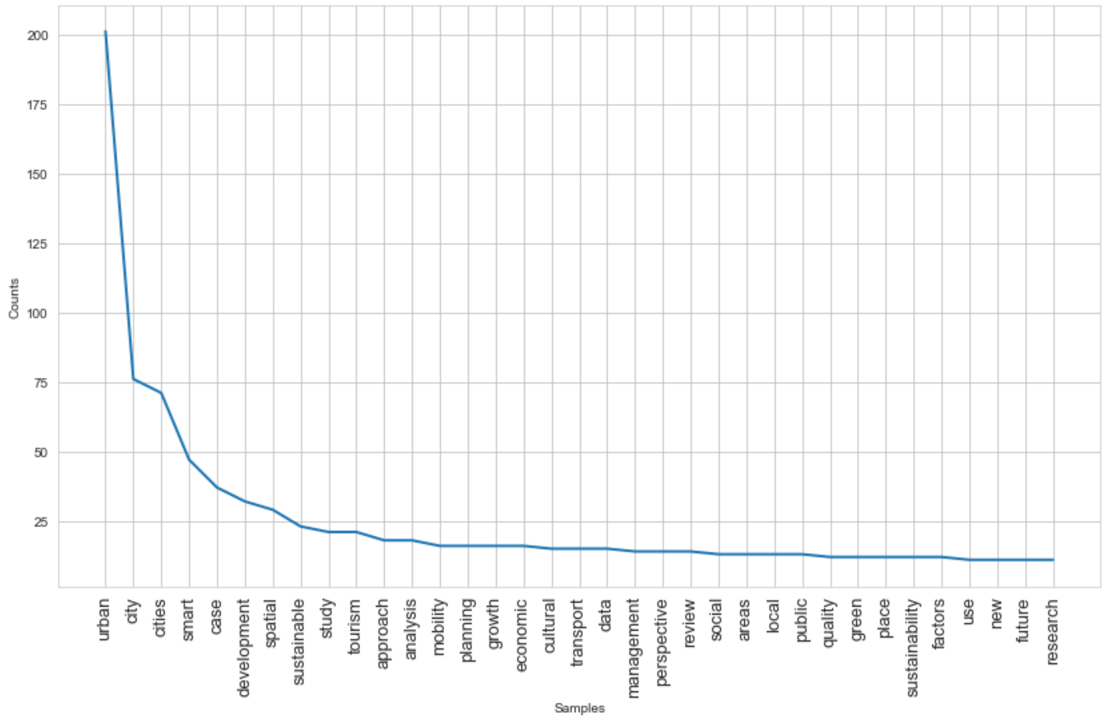


***Figure 1. 6** Most frequent words mentioned in research titles related to Cities attractiveness over the last 10 years*

In the following figure, the same words are visualized on a word cloud, the bigger the size of a word the higher the frequency of it repeating in the collected studies.

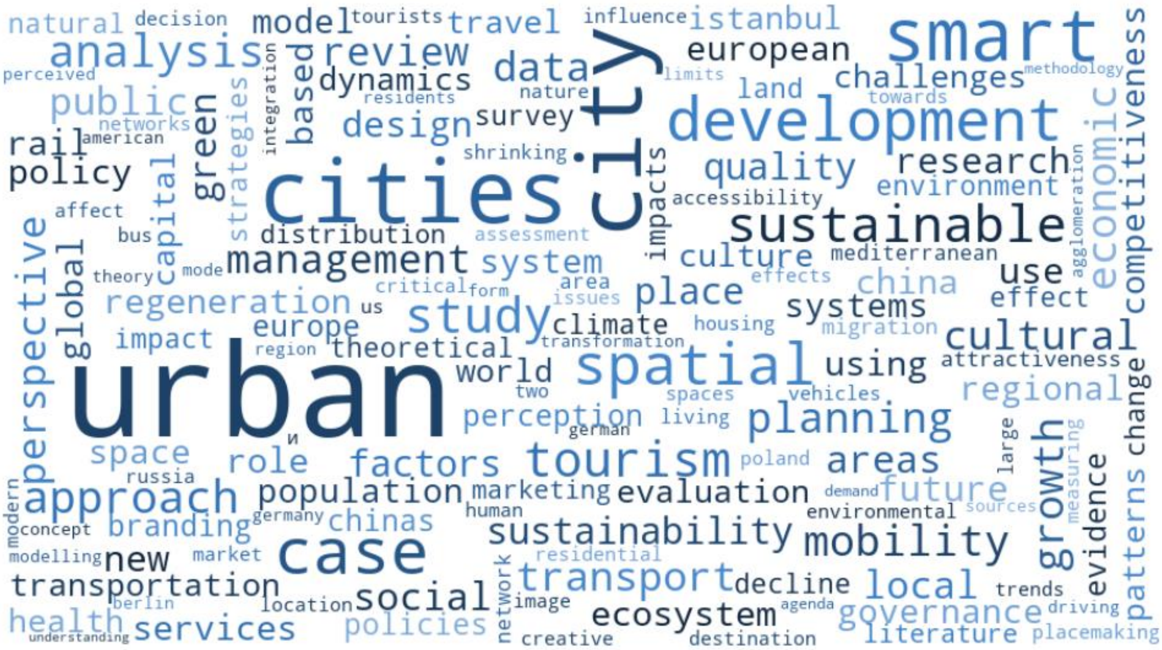


Figure 1. 7 *A word cloud of the most frequent words mentioned in research titles related to Cities attractiveness over the last 10 years*

In the following two figures, the same process is carried out in order to analyze the paper's content. The most frequently used words are as follows: cities, attractiveness, urban, city, development, economic, population, transport, quality, social, planning, growth, cultural. This sheds the light on the concepts that cities' attractiveness was linked with in the past work.

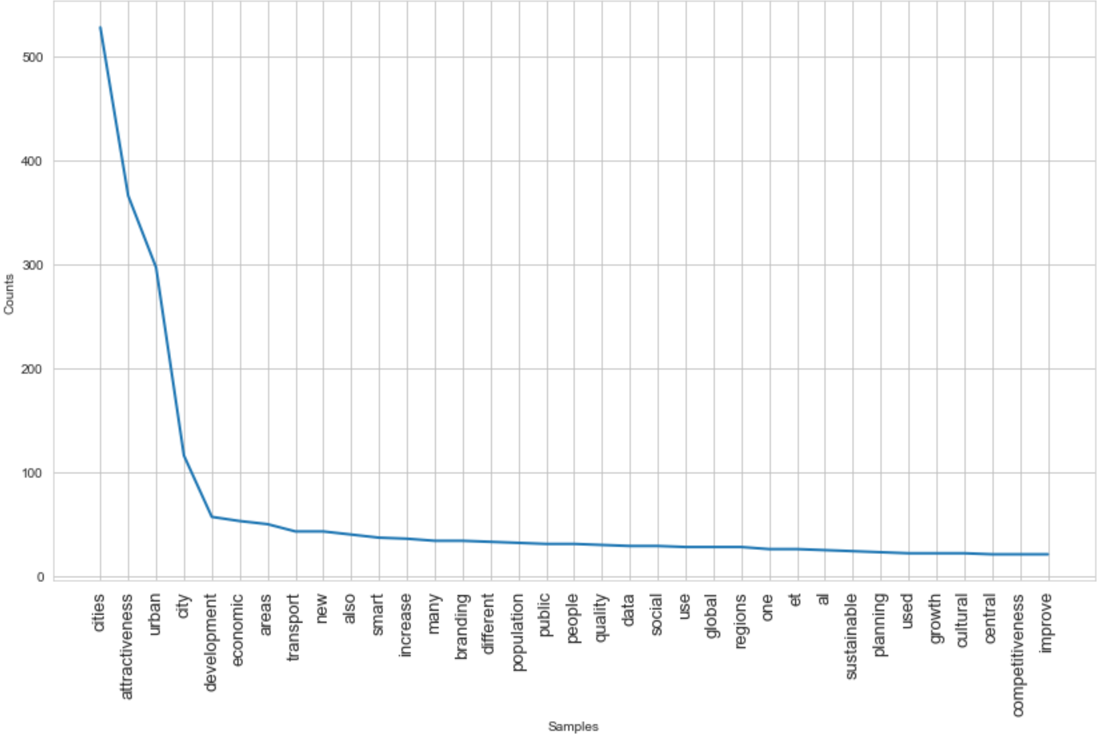


Figure 1. 8 *Most frequent words mentioned in research contents related to Cities attractiveness over the last 10 years*

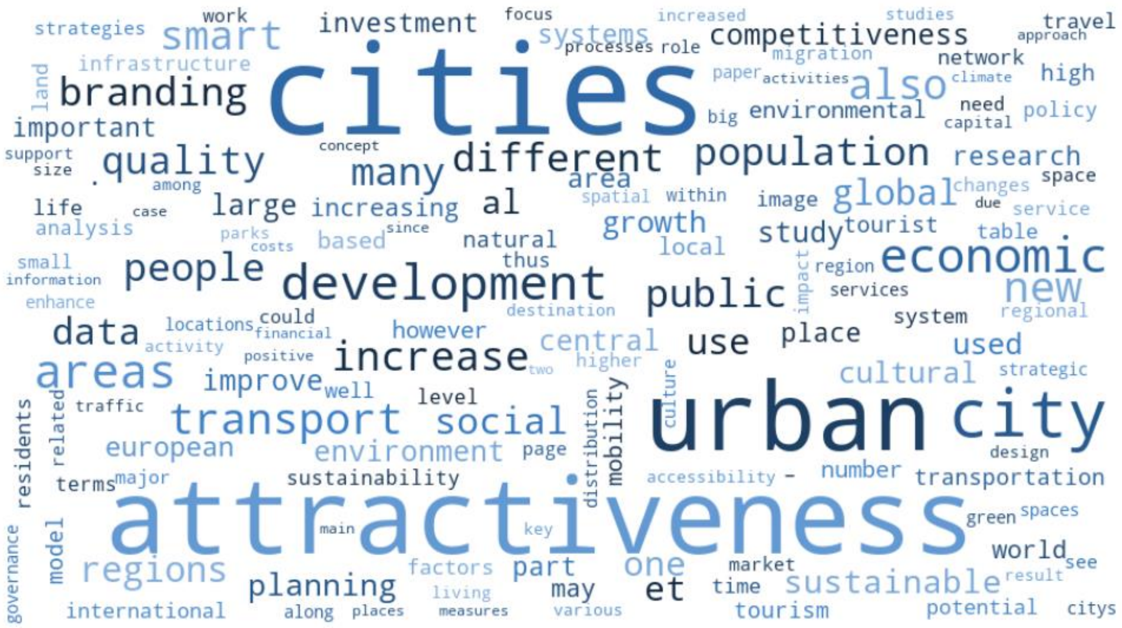


Figure 1. 9 *A word cloud of the most frequent words mentioned in research contents related to Cities attractiveness over the last 10 years*

Lastly, the most frequently used publishers are visualized, such as Elsevier, Springer, Taylor & Francis Group, Research Gate, and Emerald.

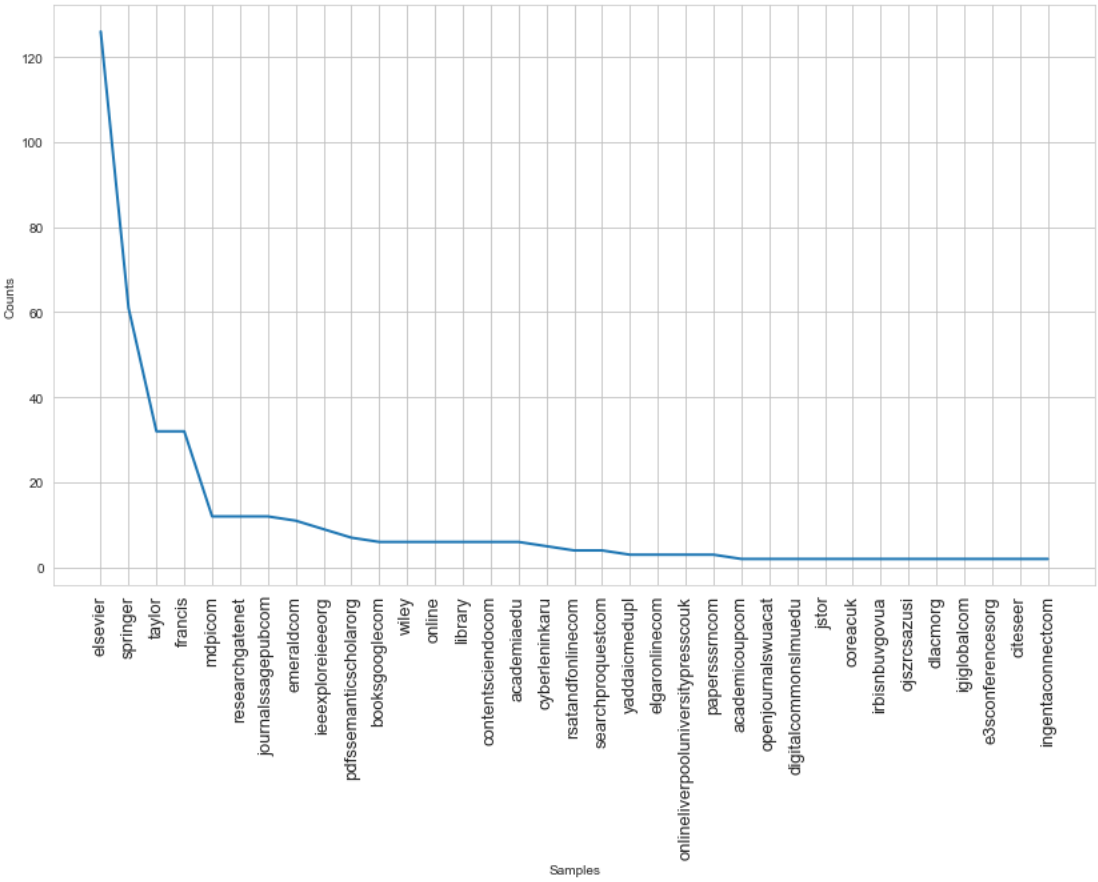


***Figure 1. 10** Most frequent publishers for research related to Cities attractiveness over the last 10 years*

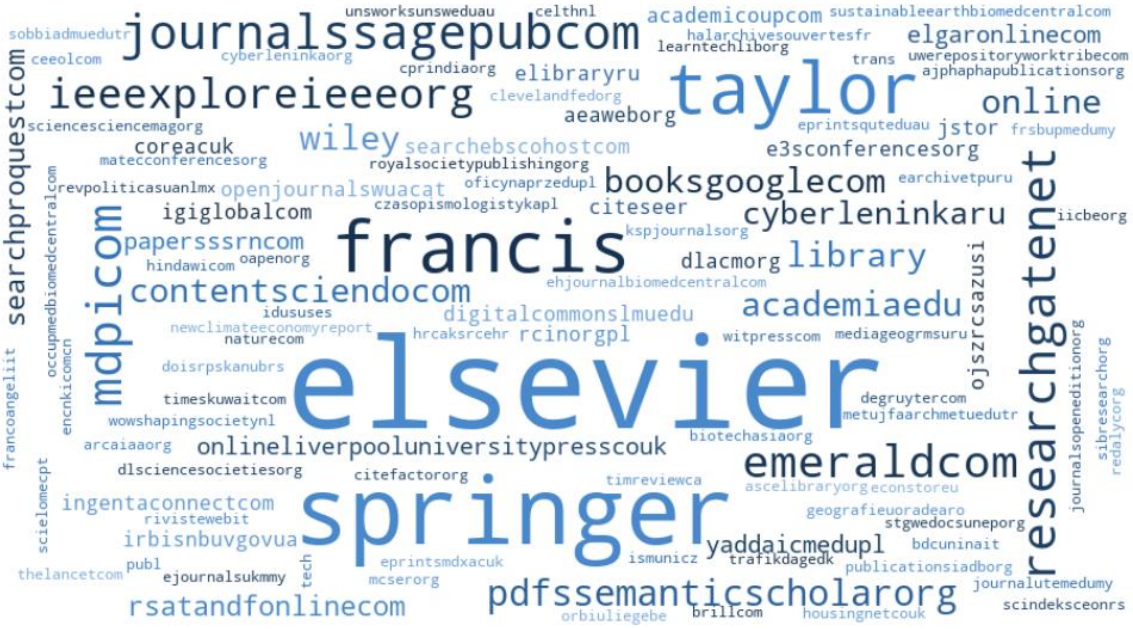


***Figure 1. 11** A word cloud for the most frequent publishers for research related to Cities attractiveness over the last 10 years*

To conclude, if the attractiveness of a firm can improve its financial performance in a competitive market, in this dissertation we investigate whether higher city attractiveness could influence the success of its strategy in international competition in promoting international investment and immigration. We try to examine the gap between qualitative data (people's perceptions in online discussions) and quantitative data (cities' indicators). Since, the success of policies does not rely only on governments or policymakers but also relies on people's behaviors, as they are the main stakeholder whose decisions and actions are affected by their individual perceptions **(Pyhälä, et al., 2016)**.

Chapter 2 **Data Science**

Data science can be thought of as the basis for empirical research, it turns data into insights or even actions, for data is to be used to inform the hypotheses and provide observations. In many cases, this data is used either by businesses or by scientists to inform their understanding of a phenomenon. Because there is often a large amount of data which we can mine for insights, we often call this big data. Insight is a term we use to refer to the data product of data science. It is extracted from a diverse amount of data through a combination of exploratory data analysis and modelling. The questions we ask are sometimes quite specific, but sometimes it takes looking at the data and patterns in it to come up with a specific question. Another important point to recognize is that data science is not a static, one-time analysis. It involves a process where the models we generate lead to insights and those insights are then improved by gathering further empirical evidence, or simply, data. For instance, a book retailer like amazon.com can constantly improve the model of a customer's book preferences using the customer demography, his or her previous purchases and prior book reviews by the customer. Their models also likely take into account the similarity of customers to detect common interests. The book retailer can also use this information to predict which customers are likely to like a new book and take action to market the book to those customers. This is where we see insights being turned into action **(The University of California, San Diego, 2020).**

Using data science and analysis of the past and current information, data science generates actions. This is not just an analysis of the past, but rather a generation of actionable information for the future. This is what we call a prediction and it is very similar to a weather forecast. When we decide what to wear based on the forecast for the day, we are taking action based on insight delivered to you. Business leaders and decision-makers take action based on the evidence provided by their data science teams. Because companies take action based on these insights, data science teams need to be experts in their practice to ensure those insights are well-reasoned. Data science has been around for a very long time. Scientists have always used data to gain insight based on observations. This data growth combined with the advances in storage, networking, and computing at scale has brought us to a new era of data science. Many dynamic data-driven applications in this new era build upon data-driven predictions to support decisions, just like the Amazon book prediction example. It is nearly impossible to find an industry, scientific discipline, or engineering venture today that is not impacted by data science. One need only look at the major trends in smart cities, precision medicine, energy management, and smart manufacturing to see how it is shaping the economy today, and all these fields are looking for experts in a combination of advanced data analytics, the traditional modelling, and simulations.

Data can include anything from user preferences and purchasing history on websites to scientific data from remote sensors and instruments and personal health data from variable devices and social media data related to customer satisfaction, political trends, health epidemics, law enforcements and terrorist activities, as well as medical data from drug trials, treatment options, and patient population. Every minute, 204 million emails are sent, 200,000 photos are uploaded, and 1.8 million likes are generated on Facebook. On YouTube, 2.78 million videos are viewed, and 72 hours of video are uploaded **(Mohiuddin & Al-Sakib Khan, 2018)**. A data-oriented business currently collects data in the order of terabytes, but petabytes are becoming more common to the daily lives. The bottom line is that all of these sources point to an exponential growth in data volume and storage. While many of us are excited by the opportunities offered by big data, this rapid growth also comes with a number of management and analysis challenges, least of which is information overload. The challenges aren’t just to manage the data but to try to see how everything is connected. Finding the connections between the kinds of data sets has the potential to lead to interesting discoveries. Such an endeavor requires proper use of data management, data-driven methods, scalable tools for dynamic coordination and scalable execution, and a skilled interdisciplinary workforce.

This is where we come in the picture. By putting the time into skills and programming in Python, statistics, machine learning, and big data, we will be ready to take on the topic of this dissertation by exploring the trends and the online discussion on what makes cities attractive for relocation. Data science team often comes together to analyze situations or answer questions in business or science which no single person could solve on their own. There are lots of moving parts to the solution, but in the end, all these parts should come together to provide actionable insight based on data science. Being able to use evidence-based insight into the decisions is more important now than ever.

2.1 Data Science Process

Generally speaking, data science starts with a team of people with an overarching broad question and of course some data to explore. We can say that we start with data and questions and we build a process around how we come up with data-driven insight. A process is a conceptual entity in the beginning and defines the core set of steps to solve the question. There are many ways to look at the process. One way of looking at it is as two distinct activities, mainly data engineering and data analysis. A more detailed way of looking at the process reveals five distinct steps or activities of the data science process, namely acquire, prepare, analyze, report and act. We can simply say data science happens at the boundary of all these steps. Ideally, this process should support experimental work and dynamic scalability on big data and cloud platforms. This five-step process can be used in alternative ways in real-life big data applications if we add the dependencies of different tools to each other. The influence of big data pushes for alternative scalability approaches at each step of the process. Another way to look at this process is seeing all these steps with reporting needs in different forms or drawing all these activities as an iteration process including build, explore and scale for big data as steps. Scalable data analysis needs alternative data management techniques, systems, analytical tools and methods as well as nodes of scalability based on dynamic data and computing load, change in physical infrastructure and streaming data specific emergencies arising from special events. This scalable process should be programmable through the utilization of reusable and reproducible programming interfaces to analytical tools, visualization environments and user reporting environments. This final dimension, the programming of data science steps in Python.

We should focus on formulating a well-stated data science problem. Without this, we won't have a clear goal in mind or knowledge when we have solved the problem. An example question is how can sales figures in call center logs become viable to evaluate any product? Or in a manufacturing process, how can data from multiple sensors on a piece of equipment be used to detect equipment failure? How cities can we understand their customers and market better to achieve effective targeted marketing? Next, we need to assess the situation with respect to the problem or opportunity we have defined. This is a step where we need to exercise caution, analyzing risks, costs, benefits, contingencies, regulations, resources and requirements of the situation. What are the requirements of the problem? What are the assumptions and constraints? What resources are available to us? This is terms of both personnel and capital such as computer systems, equipment etc. What are the main costs associated with this project? What are the potential benefits? What risks are there in pursuing a project? What are the contingencies to potential risks? Answers to these questions will help us get a better overview of the situation and a better understanding of what the question involves and how we will guide the programming to solve the question with all these in mind. Then, we need to define the goals and objectives. Defining success criteria is also very important. What do we hope to achieve by the end of this dissertation? Having clear goals and success criteria will help us to assess the dissertation throughout its lifecycle. Once we know the problem we want to address and understand the constraints and goals, then we can formulate the plan to come up with the answer that is the solution to the business problem or the analytics we are trying to achieve. As a summary, defining the questions we're looking to find answers for is a huge factor contributing to the success of any data science project. By following the explained set of steps, we can formulate better questions to solve using analytical skills and link them to scientific and business value. Let's now look into the basic steps in the Data Science process;

- The first step is data acquisition, acquiring the dataset and importing it to the data science. A lot of data exists in conventional, relational databases, like structured data coming from organizations. The tool of choice access data from databases is SQL, which is supported by all relational database management systems. Additionally, most database systems come with a figural application environment that allows us to query and explore the data sets in the database. Data can also exist in files such as text files and Excel spreadsheets. Scripting languages are generally used to get data from files. A scripting language, like Python, is a high-level programming language that can be general-purpose or specialized for specific functions. Other common scripting languages with support for processing files are JavaScript, PHP, Perl, R, Octa and MATLAB to name a few. An increasingly popular way to get data is from websites. Webpages are written using a set of senders approved by a world wide web consortium, or W3C. This includes a variety of formats and services. One common format is XML or Extensible Markup Language or JSON, which both use markup symbols and tabs to describe the contents on a webpage. Many websites also host webs services which provide programmatic access to their data. There are several types of web services, the most popular one being REST since it's easy to use. REST stands for Representational State Transfer, and it is an approach for implementing webs services with performance scalability and maintainability in mind. WebSocket services are also becoming more popular since they allow Realtime notifications from the websites. NoSQL storage systems are increasingly used to manage a variety of data types. These data stores are databases that do not represent data in a table format with columns or rows, as with conventional relational databases. Examples of these data stores include Cassandra, MongoDB and HBase. NoSQL data stores provide APIs to allow users to access the data. These APIs can be used directly or in an application that needs to access the data, like a Python script. Additionally, most NoSQL systems provide data access via a web interface, such as REST.

- The second step is data preparation, data exploration and visualization and performing data cleaning, understanding the nature of the data (quality and format) and doing preliminary analysis. Exploring data is a big part of the two-step data preparation activity. We want to do some preliminary investigation in order to gain a better understanding of the specific characteristics of the data. We will be looking for things like correlations, general trends, outliers. And without this step, we will not be able to use the data effectively. Correlation figures can be used to explore the dependencies between different variables in the data. General trends show us a simple figure of how the data is progressing over time. And outliers show us the data points that are distant from other data points. Plotting outliers will help us double-check for errors in the data due to measurement. In some cases, outliers that are not errors might make us find a rare event. Additionally, summary statistics provide numerical values to describe the data. Some basic summary statistics that we should compute for the data set are mean, median, mode, range, and standard deviation. Mean and median are measures of the location of specific values. The mode is the value that occurs most frequently in the data set. And the range and standard deviation are measures of spread in the data. Looking at these measures will give us an idea of the nature of the data. They can tell us if there's something wrong with the data. For example, if the range of the values for age in the data includes negative numbers or a number much greater than a hundred, there's something suspicious in the data that needs to be examined. Visualization techniques also provide a quick and effective representation. A heat map, for instance, can quickly give us an idea where the hot spots are. Many different types of figures can be used. Histograms show the distribution of the data and can show skewness or unusual dispersion. Boxplots are another type of plot for showing data distribution. Line figures are useful for seeing how values in the data change over time. Spikes in the data are also easy to spot. Scatter plots can show a correlation between two variables

  Pre-process data stage includes first is to clean the data to address data quality issues.

Examples for quality issues: like a customer with two different addresses recorded at two different sales locations but these recordings don't agree, missing customer age in Geographical studies, an invalid step code, for example, a six-digit zip code, outliers like a sensor failure that cause values to be much higher or lower than expected for a period of time. Some approaches we can take to address these data quality issues, we can remove the data records with missing values, or we can merge duplicate records. This would record a way to determine how to resolve conflicting values. Perhaps it makes sense to retain the nearer value whenever there's a conflict. For invalid values, the best estimate for a reasonable value can be used as a replacement. For example, a missing age value for an employee can be filled in based on a reasonable estimate on the employee's length of employment. Outliers can also be removed if they are not important to the task. In order to address all these data quality issues effectively, knowledge about the application, such as how the data was collected, the user population, the intended users of the application etc. are important. This domain knowledge is essential to making informed decisions on how to handle incomplete or incorrect data. we also need to be careful about the changes we make to avoid coming to incorrect conclusions and be sure to keep records of the changes you make

The second is to transform the data to make it suitable for analysis. This step is known by many names, data manipulation, data pre-processing, data wrangling and data munging. Some operations for this data munging, wrangling, pre-processing includes scaling, transformation, feature selection, dimensionality reduction and data manipulation. Scaling involves changing the range of values to be between a specified range such as from zero to one. This is done to avoid having certain features with large values from dominating the results. For example, in analyzing data with height and weight the magnitude of the weight values is much greater than the magnitude of the height values. So, scaling all values to be between zero and one will equalize contributions from both height and weight features. Various transformations can be performed on the data to reduce noise and variability. One such transformation is called log transformation. Log transformation generally results in data with less variability which may help with the analysis. Other filtering techniques can also be used to remove variability in the data. Of course, this comes at the cost of less detailed data, so these factors must be weighed for the specific application. Feature selection can involve removing redundant or irrelevant features, combining features and creating new features. During the exploring data step, we may have discovered that two features are very correlated. one of these features can be removed without negatively affecting the analysis results. For example, the purchase price of a product and the amount of sales tax paid are likely to be very correlated. Eliminating the sales tax amount then will be beneficial. Removing redundant or irrelevant features will make the subsequent analysis simpler. we may want to combine features or create new ones. For example, adding applicants' education level as a feature to a loan approval application would make sense. There are also algorithms to automatically determine the most relevant features based on various mathematical properties. Dimensionality reduction is useful when the dataset has a large number of dimensions. It involves finding a smaller subset of dimensions that captures most of the variation in the data. This reduces the dimensions of the data while eliminating irrelevant features and makes analysis simpler. A technique commonly used for dimensionality reduction is called principal component analysis. Raw data often has to be manipulated to be in the correct format for the analysis. For example, from samples recording daily changes in stock prices, we may want to capture the price changes for a particular market segment, for example, real estate or healthcare. This would require determining which stocks belong to which market segment, grouping them together, and perhaps computing the mean, range and standard deviation for each group.

- The third step includes doing data analysis, we analyze the prepared dataset using statistical analysis and machine learning. We start by selecting analytical techniques and building models and analyzing results. This step can take a couple of iterations (repetition) on its own or might require a data scientist to go back to steps one and two to get more data or package data in a different way. Data analysis involves building a model from the data which is called input data. The input data is used by the analysis technique to build a model. What the model generates is the output data. The main categories of analysis techniques are classification, regression, clustering, association analysis and figure analysis. In classification, the goal is to predict the category of the input data. An example of this is predicting the weather as being sunny, rainy, windy or cloudy. Another example is to classify a tumour as either benign or malignant. In this case, the classification is referred to as binary classifications as they are only two categories. When the model has to predict a numeric value instead of a category, then the task becomes a regression problem. An example of regression is to predict the price of a stock over time. The stock price is a numeric value, not a category, so this is a regression task instead of a classification task. Other examples of regression are estimating the weekly sales of a new product and predicting the score on a test. Next is clustering. In clustering, the goal is to organize similar items into groups. An example is grouping a company's customer base into distinct segments for more effective targeted marketing like seniors, adults and teenagers. Other examples include determining different groups of weather patterns like rainy, cold or snowy

The goal in association analysis is to come up with a set of rules to capture associations between items or events. The rules are used to determine when items or events occur together. A common application of association analysis is known as market basket analysis which is used to understand customer purchasing behaviour. For instance, a supermarket chain used association analysis to discover a connection between two seemingly unrelated products. They discovered that many customers who go to the supermarket late on Sunday night to buy diapers also tend to buy beer. This information was then used to place beer and diapers close together on Sundays and they saw a jump in sales of both items. This is the famous diaper-beer connection. Then we want to use figure analytics to analyze the data. This kind of data comes about when we have a lot of entities and connections between those entities like social networks. Some examples where figure analytics can be useful are exploring the spread of a disease or epidemic by analyzing hospitals and doctors' records, identification of security threats by monitoring social media, email and text data, and optimization of mobile communication network traffic to ensure data quality and reduce dropped calls

Modelling starts with selecting one of these techniques we listed as the appropriate analysis technique depending on the type of problem we have. Then, we construct the model using the data that we have prepared. To validate the model, we apply it to new data samples. This is to evaluate how well the model does on data that was used to construct it. The common practice is to divide the prepared data into a set of data for constructing the model and reserving some of the data for evaluating the model after it has been constructed. We can also use new data prepared the same way as the data that was used to construct the model. Evaluating the model depends on the type of analysis technique we used. For classification and regression, we will have the correct output for each sample in the input data. Comparing the correct output and the output predicted by the model provides a way to evaluate the model. For clustering, the groups resulting from clustering should be examined to see if they make sense for the application. For example, do the customer segments reflect the customer base? Are they helpful for use in your targeted marketing campaigns

For association analysis and figure analysis, some investigation will be needed to see if the results are correct. For example, network traffic delays need to be investigated to see if what the model predicts is actually happening and whether the sources of the delays are where they are predicted to be in the real system. After we have evaluated the model to get a sense of its performance on the data, we will be able to determine the next steps. Some questions to consider are, should the analysis be performed with more data in order to get better model performance? Would using different data help? For example, in the clustering results, is it difficult to distinguish customers from distinct regions? Would adding zip codes to the input data will help generate final grain customer segments? Do the analysis results suggest a more detailed look at some aspect of the problem? For example, predicting sunny weather gives very good results but rainy weather predictions are just not that accurate. This means we should take a closer look at the examples for rainy weather. Perhaps there is some irregularity in those samples or perhaps there are some missing data that needs to be included in order to completely capture rainy weather. The ideal situation would be that the model performs very well with respect to success criteria that were determined when we defined the problem at the beginning.

- This fourth step includes presentation and reporting of insights, Findings from these analyses are typically turned into reports and presented to the stakeholders who need to take actions. The first thing is determining what part of the analysis is most important to offer the biggest value to the scientific community or the company in the industry to a particular audience. In deciding what to present, we should ask ourselves these questions. What are the main results? What value do these results provide based on the specific domain that we are working on, and the application the question led me to? How can the model add to this application? In other words, how do the results compare to the success criteria determined at the beginning for that application's specific purpose? We need to include the answers to these questions in the report or presentation. So, we should make those questions and answers the main topics and be sure to have the data or visualizations to back them up and we should keep in mind that the analysis may show results that are counter to what we were hoping to find or results that are inconclusive or puzzling. We need to show these results as well. We should remember that the point of reporting the findings is to determine what the next steps should be. All findings must be presented so that informed decisions can be made. If we think about it, the biggest danger is to make it seem like the results tell a clear story when they actually don’t. The techniques that we discuss and explore in data can be used here as well. Scatter plots, line figures, heat maps and other types of figures are effective ways to present the results visually. We should also have tables with details from the analysis as a backup. There are many visualization tools that are available. Some of the most popular open-source ones are listed here, including R, a software package for data analysis, which has powerful visualization capabilities. Python, that we will use in this dissertation, is a general-purpose programming language, or a scripting language, that allows us to use a number of packages to support data analysis and figures. D3 is a JavaScript library for producing interactive web-based visualizations and data-driven comments. The leaflet is a lightweight, mobile-friendly JavaScript library to create interactive maps. Lastly, Tableau and Google Charts allow us to create visualizations in the profiles so we can share them or put them on a site or a blog, and they provide cross-platform compatibility to mobile devices. Timeline is a JavaScript library that allows us to create timelines over these results. We'll be using tools that connect well with Python's Jupyter notebooks.

- The last step includes acting on the data and turning these insights into data-driven actions. Based on what we have found, it is likely we have actions you could take to improve the city. Now, we need to figure out how to implement the actions. What is necessary to add this action to the process, or application? How should it be automated, if it can be? The stakeholders need to be identified and become involved in this change. Just as with any process improvement changes, we need to monitor and measure the impact of the action on the process or application. Be sure to think about what data we should collect during and after the change to properly evaluate its impact. Evaluating Results from the implemented action will determine the next steps. Is there additional analysis that needs to be performed, in order to yield even better results? What data should be revisited? Are there additional, or further ideas that should be explored? For example, let's not forget what big data enables us. We could be taking real-time action based on rapidly streaming information. In business, we need to define what part of the business needs real-time action to be able to implement the operation or the interaction with the customers. With public service, we need to know what action to take when certain events are observed in the data. Once we define these real-time actions, we need to make sure that there are automated systems, or processes, in the organization, or scientific research group, that perform such actions and provide failure recovery in case of problems. As a summary, big data and data science are only useful if the insights can be turned into actions, and the action should be carefully defined and evaluated.

2.2 Python for Data Science

Data science happens at the intersection of computer science, mathematics, and business or scientific expertise. Even at this level, all of these require deeper knowledge and skills in areas like domain expertise, data engineering, statistics, and computing. And even deeper analysis of these skills based on data science job listings would lead you to skills like machine learning, statistical modelling, relational algebra, business passion, problem-solving, and data visualization. That's a lot of skills to have for a single person. Given such a wide range of skills across multiple definitions of data scientists seem impossible. Data scientists are teams of people who act like one. This is why we say data science is a team sport, referring to the breadth of information and skills it takes to make it happen. However, there are still common traits to the data scientists. For example, data scientists are passionate about the story and meaning behind data, they understand the problem they are trying to solve, and aim to find the right analytical methods to solve this problem. And they all have an interest in engineering solutions to solve problems.

Python is a clear leader in many data science categories. Although learning any of the programming languages, including R Java, C, Scala, and Julia is a good idea. There are specific reasons why we pick Python. Instead of explaining why Python is a good language for data science, let's focus on why data scientists love Python. Python is an open language with a vibrant community. Thanks to the efforts of this community, it offers an ever-growing set of data management, analytical processing, and visualization libraries, some of which we will review in this dissertation. Such libraries make Python applicable to every step of the data science process. Lastly, but very importantly, the Jupyter Notebooks make Python-based analysis more producible and repeatable, as well as provides built-in training and communication support to help with team communication. We will start by learning about Jupyter Notebooks, followed by NumPy and Pandas to ingest and analyze data efficiently. We will add the visualization libraries, including Matplotlib, and continue with applying machine learning libraries in Scikit-Learn to create models.

The learning goal is to be able to articulate the benefits of Python as a programming language. We'll eventually be using Jupyter, which was originally the combination of three languages, Julia, Python and R. These days, Jupyter supports over 40 programming languages. But w. For now, we're starting in Python so that you have a strong programming basis to build upon going forward. Python, even though it's only been around since 1991, is already one of the world's most popular programming languages. Python is powerful, we will be able to accomplish in just a few lines of Python what might have taken three or five times as many lines in another language, like java or C++. Python is surprisingly fast for interpretive language. Generally, interpretive languages can be slower because they'd run on top of an interpreter, rather than being compiled directly from the machine on which they're running. Python plays well with other languages as well. Python, because it's interpreted, can run everywhere. Moreover, because Python is open source and free, you can install it anywhere. Python has a number of features, like dynamic typing and automatic memory management, which make it both easy to learn Python and also easy to read Python code.

Chapter 3 Machine Learning

3.1 Introduction

In this chapter, the importance of Machine Learning. Machine learning is a field of study that focuses on computer systems that can learn from data. That is, these systems, often called models, can learn to perform a specific task by analyzing lots of examples for a particular problem. For instance, a Machine Learning model can learn to recognize an image of a cat by being shown lots and lots of images of cats. This notion of learning from data means that a Machine Learning model can learn a specific task on its own. The Machine Learning algorithm is programmed to learn from the data that there's nothing in the algorithm or program which directly aims to learn the given task. In other words, Machine Learning models are not given the step by step instructions on how to recognize the image of a cat. Instead, the model learns on its own what features are important in determining that a picture contains a cat from the data that it has analyzed.

Because the model learns to perform this task from data, it's good to note that the amount and quality of data available for building the model are important factors in how well the model learns from the task. Because Machine Learning models can learn from data, they can be used to discover hidden patterns and trends in the data as well. These trends and patterns lead to valuable insights into the data; thus, the use of Machine Learning allows for data-driven functions to be made for a particular problem.

To summarize, the field of Machine Learning focuses on the study and construction of computer systems that can learn from data without being explicitly programmed. Machine Learning algorithms and techniques are used to build models to discover hidden patterns and trends in the data, allowing for data-driven decisions to be made. Machine Learning has been used in many different applications, many of which we encounter in the everyday life, perhaps without even realizing it. One application of Machine Learning is credit card fraud detection. Every time we use the credit card, the current purchase is analyzed against your history of credit card transactions to determine if the current purchase is a legitimate transaction or a potentially fraudulent one. If the purchase is very different from the past purchases, such as for a big-ticket item in a category that you never show any interest in, or at the point of sale location is from another country, then it will be flagged as a suspicious activity. In that case, the transaction may be denied, or we might get a phone call from the credit card company to confirm that the purchase was indeed made by us. This is a very common use of Machine Learning that's encountered in everyday life.

Another example application of Machine Learning encountered in daily life is a handwritten digit recognition. When we deposit a handwritten check into an ATM, a Machine Learning process is used to read the numbers written on the check to determine the amount of the deposit. Handwritten digits are trickier to decipher than typed digits, due to the many variations in people's handwriting. A Machine Learning system can sift through different variations to find similar patterns to distinguish a one from a nine, for instance. Recommendations on websites is another example application of Machine Learning that most people have experienced firsthand. After we buy an item on a website, we will often get a list of related items. Often, this will be displayed as customers who bought this item also bought these items, or, you may also like. These related items have been associated with the item we purchased by a Machine Learning model and are now being shown to us since we may also be interested in them. This is a common application of Machine Learning used often in sales and marketing. These were just a few examples among many.

Data science has its roots and statistics, machine learning, artificial intelligence, and computer science among other fields. Machine Learning is the part of this field that encompasses the algorithms and techniques used to learn from data. There are a number of other terms we hear in conversations about Machine Learning. The term data mining became popular around the time that the use of databases became commonplace. Data mining was used to refer to activities related to finding patterns in databases and data warehouses. There are some practical data management aspects to data mining related to accessing data from databases, but the process of finding patterns in data is similar to Machine Learning and often uses similar algorithms and techniques. Predictive analytics is another term, and it refers to analyzing data in order to predict future outcomes. This term is usually used in the business context to describe activities such as sales forecasting or predicting the purchase of the behavior of a customer. But again, the techniques used to make these predictions are the same techniques for Machine Learning. Data science is a new term that's used to describe the process of processing and analyzing data to extract meaning. Machine Learning techniques can also be used here. Because the term data science became popular at the same time that big data began appearing, data science usually refers to extracting meaning from big data.

There are different categories of machine learning techniques for different types of problems. The main categories are classification, regression, cluster analysis, and association analysis.

A. In classification, the goal is to predict the category of the input data. An example of this is predicting the weather as being sunny, rainy, windy, or cloudy. The input data, in this case, would be sensor data specifying the temperature, relative humidity, atmospheric pressure, wind speed, wind direction, etc. The target or what we are trying to predict would be the different weather categories, like sunny, windy, rainy, and cloudy. The categories to be predicted are called classifications. Another example is to classify a tumor as benign or malignant. In this case, the classification is referred to as binary classification since there are only two categories, but we can have many categories as well.

B. When the model has to predict a numeric value instead of a category, then the task becomes a regression problem. An example of regression is to predict the price of a stock. The stock price is a numeric value, not a category, so this is a regression task. It is not a classification task. If we were to predict whether the stock price will rise or fall, then that would be a classification problem but if we are predicting the actual price of the stock, then that's a regression problem. That is the main difference between classification and regression. To summarize, in classification we are predicting a category, and in regression, we're predicting a numeric value. Another regression application in addition to the stock example is for prediction of the amount of rain for a region. Predicting if it will rain the next day. That is predicting between one of the two categories, rain or no rain. So, that's a classification problem. But if we are predicting the amount of rain, which is a numeric value, that would be a regression problem.

C. In cluster analysis, the goal is to organize similar items in the dataset into groups. A very common application of cluster analysis is referred to as customer segmentation. This means that we're separating the customer base into different groups or segments based on customer types. For instance, it would be beneficial to segment the customers into seniors, adults, and teenagers. These groups have likely different likes and dislikes and have different purchasing behaviors. When companies segment customers into different groups like this, they may be able to provide targeted marketing ads for each group's particular interests. Note that cluster analysis is also referred to as clustering in different contexts.

D. Association analysis: The goal here is to come up with a set of rules to capture associations between items or events. The rules are used to determine when items or events occur together. A common application of association analysis is known as market basket analysis, which is used to understand customer purchasing behavior. For example, association analysis can reveal that banking customers who have checking or deposit accounts also tend to be interested in other investment vehicles such as money market accounts. This information can be used for cross-selling. Other common applications can recommend similar items based on purchasing or browsing the history of customers. Finding items that are often purchased together and offered based on these related items at the same time to drive sales of both items would be very beneficial. Identification of web pages that are often accessed together can also provide us with a good basis for association analysis.

For the techniques discussed, there are two ways of conducting the learning itself. These categories are referred to as supervised and unsupervised learning.

i. In supervised approaches, the target, which is what the model is predicting, is provided. This is referred to as having labeled data because the target is labeled for every sample that we have in the dataset. Referring back to the example of predicting the weather category of sunny, windy, rainy, or cloudy, every sample in the dataset is labeled as being one of these four categories. So, the data is labeled, and predicting the weather category is a supervised task. In general, classification and regression are supervised approaches.

ii. In unsupervised approaches, on the other hand, the target that the model is predicting is unknown or unavailable. This means that we have unlabeled data, so we can't train using these labels. Going back to the cluster analysis example of segment customers into different groups, the samples in the data are not labeled with the correct group. Instead, the segmentation is performed using a clustering technique to group items based on characteristics of what they have in common. Thus, the data is unlabeled and the task of grouping customers into different segments is an unsupervised one. In general, cluster analysis and association analysis are unsupervised approaches.

3.2 Regression Analysis

In this section, regression analysis is discussed. When the machine learning model has to predict a numeric value instead of a category, then the task becomes a regression problem. An example of regression is to predict the price of a stock. The stock price is numeric and not a category, so this is a regression task instead of a classification task. Here are some examples where regression can be used. We can use it for forecasting the high temperature for the next day or estimating the average housing price for a particular region or determining the demand for a new product, a new book based on similar existing products or like in the case estimating or predicting the demand for cities visitors per year based on the previously mentioned variables. The model has to predict this target value for each sample. Since the target label is provided for each sample, as a numeric value here, the regression task is a supervised one.

The importance of regression analysis for businesses is that it helps determine which factors matter most, or which it can ignore, and how those factors interact with each other. The regression analysis is used for forecasting and finding the causal relationship between variables. Understanding the importance of regression analysis, the advantages of linear regression, as well as the benefits of regression analysis and the regression method of forecasting can help any business gain a far greater understanding of the variables that can impact its success into the future. Smart companies use regression analysis to make decisions about all kinds of business issues. How to impact sales or employee retention or recruiting the best people. Most companies use it to explain a phenomenon they want to understand (e.g. why did customer service calls drop last month?); predict things about the future (e.g. what will sales look like over the next six months?); or to decide what to do (e.g. should we go with this promotion or a different one?). The importance of regression analysis is that it is all about data: data means numbers and figures that actually define the business. The advantages of regression analysis are that it can allow you to essentially crunch the numbers to help you make better decisions for any business currently and into the future.

How is the regression analysis used in forecasting? The regression method of forecasting involves examining the relationship between two different variables, known as the dependent and independent variables. As in the case, we want to forecast future yearly visitors for the city and we have noticed that the number of visitors rises and fall, depending on factors that can be related to economy, environment or cost and quality of life, and thus, the number yearly visitors would be the dependent variable (y), because they "depend" on those factors, which they can be called the independent variables (x1,x2,…,xn). We would need to investigate how closely those variables are related.

In general, building a regression model involves two phases, a training phase in which the model is built, and a testing phase in which the model is applied to new data that the model hasn't seen before. The model is built using training data and evaluated on test data. The goal is building a regression model that performs well on training data as well as generalized to new data. The two datasets are used as follows, the training dataset is used to train the model, which is to adjust the parameters of the model to learn the input to output mapping. The test data set is used to evaluate the performance of the model on new data or the leftover data. Training a linear regression model means adjusting the parameters to fit the regression line to the samples. The regression line can be determined using what's referred to as the least-squares' method or OLS. The least-squares method finds the regression line that makes the sum of the residuals as small as possible. In other words, we want to find the line that minimizes the sum of the squared errors of prediction. The goal of linear regression then is to find the best-fitting straight line through the samples using the least-squares method. Once the regression model is built, we can use it to make predictions. In linear regression, if there's only one input variable, then the task is referred to as simple linear regression. In cases with more than one input variable, then, it's referred to as multiple linear regression. To summarize, the linear regression captures the linear relationships between a numerical output and the input variables. The least-squares method can be used to build a linear regression model by finding the best-fitting line through the samples.

In this table, the difference between Correlation and Regression analysis is explained.

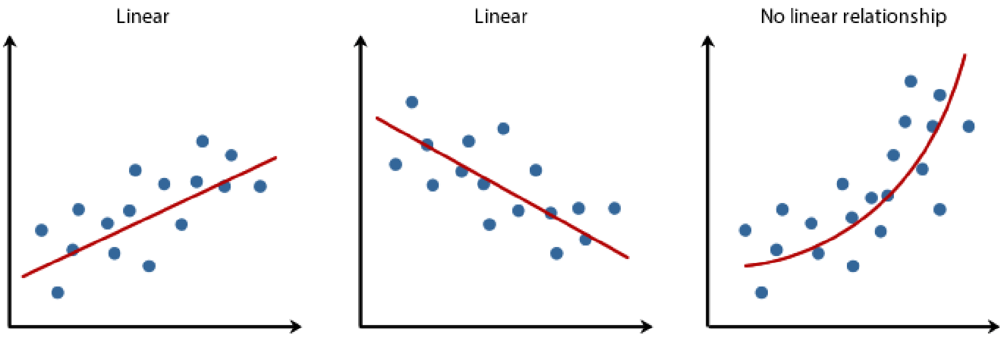
Table 3. 1 *Correlation and Regression comparison*

Correlation	Regression
Correlation does not imply causation	Regression imply causation
Measures the relationship between variables	Measures how one variable affect one another
The variables move together	The variables don't have connection only cause and effect
The relationship between x and y is the same between y and x	The relationship between x and y is not the same as y and x
Figure: Single point	Figure: regression Line, is the best fitting line through the data points

A few assumptions is made when linear regression is used to model the relationship between a response and a predictor. These assumptions are essential conditions that should be met before we draw inferences regarding the model estimates or before we use a model to make predictions. The necessary OLS assumptions, which are used to derive the OLS estimators in linear regression models, are discussed below **(Learn By Doing, Inc., n.d.).**

OLS Assumption 1: The linear regression model is “linear in parameters.” To verify if the relationship between two variables is linear by testing the x1 against Y in a scatter plot if the result is a straight line, so a linear regression is suitable. If not, then a fix is needed. Fixes can be to run a nonlinear regression, exponential transformation or Log transformation

OLS Assumption 2: There is a random sampling of observations. This assumption of OLS regression says that: The sample taken for the linear regression model must be drawn randomly from the population. The number of observations taken in the sample for making the linear regression model should be greater than the number of parameters to be estimated. This makes sense mathematically too. If a number of parameters to be estimated (unknowns) are more than the number of observations, then estimation is not possible. If a number of parameters to be estimated (unknowns) equal the number of observations, then OLS is not required. We can simply use algebra. It should not be the case that dependent variables impact independent variables. This is because, in regression models, the causal relationship is studied and there is not a correlation between the two variables. For example, if you run the regression with inflation as your dependent variable and unemployment as the independent variable, the OLS estimators are likely to be incorrect because, with inflation and unemployment, we expect correlation rather than a causal relationship. The error terms are random. This makes the dependent variable random.



Copyright 2014. Laerd Statistics.

Figure 3. 1 *A representation of the difference between linear and nonlinear relationship*

OLS Assumption 3: The conditional mean should be zero. The expected value of the mean of the error terms of OLS regression should be zero given the values of independent variables. In other words, the distribution of error terms has zero mean and doesn't depend on the independent variables. Thus, there must be no relationship between the X and the error term.

OLS Assumption 4: There is no multicollinearity (or perfect collinearity). In a simple linear regression model, there is only one independent variable and hence, by default, this assumption will hold true. However, in the case of multiple linear regression models, there is more than one independent variable. The OLS assumption of no multicollinearity says that there should be no linear relationship between the independent variables. The reasoning is that if a variable A can be represented using variable B with a correlation of 90%-100%, there is no point using both. In such a situation, it is better to drop one of the independent variables from the linear regression model. If the relationship (correlation) between independent variables is strong (but not exactly perfect), it can still cause problems in OLS estimators. Hence, this OLS assumption says that you should select independent variables that are not correlated with each other. Therefore, we will not include all the selected variables in the model since there are some that are strongly correlated with each other.

OLS Assumption 5: Spherical errors: There are homoscedasticity and no autocorrelation. According to this OLS assumption, the error terms in the regression should all have the same variance, then the linear regression model has heteroscedastic errors and likely to give incorrect estimates. This OLS assumption of no autocorrelation says that the error terms of different observations should not be correlated with each other. For example, when we have time-series data (e.g. yearly data of unemployment), then the regression is likely to suffer from autocorrelation because unemployment next year will certainly be dependent on unemployment this year. Hence, error terms in different observations will surely be correlated with each other. In simple terms, this OLS assumption means that the error terms should be IID (Independent and Identically Distributed).

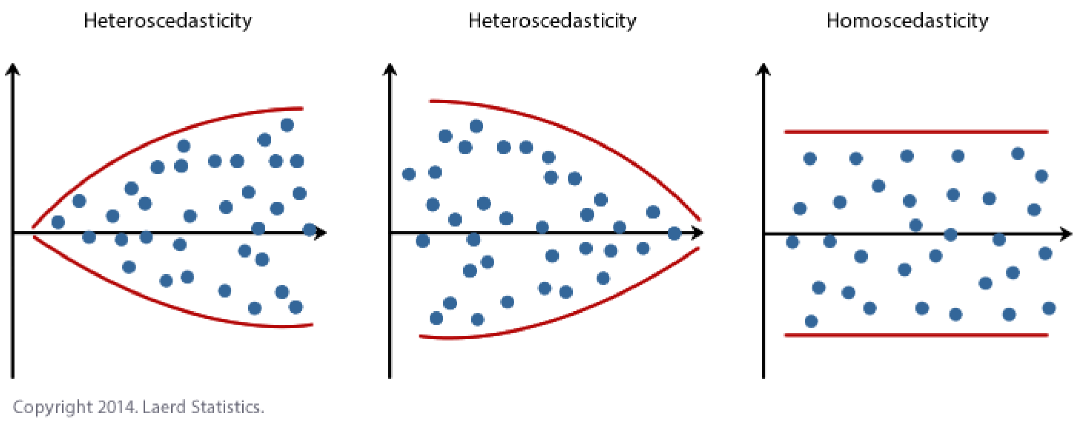


Figure 3. 2 *A representation of the difference between heteroskedasticity and homoscedasticity*

The above diagram shows the difference between Homoscedasticity and Heteroscedasticity. The variance of errors is constant in case of homoscedasticity while it's not the case if errors are heteroscedastic.

OLS Assumption 6: Error terms should be normally distributed. This assumption states that the errors are normally distributed, conditional upon the independent variables. This OLS assumption is not required for the validity of the OLS method; however, it becomes important when one needs to define some additional finite-sample properties. Note that only the error terms need to be normally distributed. The dependent variable Y need not be normally distributed

.

3.2.1 Correlation and Data Exploring

Correlation is one of the most common and most useful statistics. A correlation is a single number that describes the degree of relationship between two variables. The correlation, denoted by r, measures the amount of linear association between two variables, r is always between -1 and 1 inclusive. The R-squared value, denoted by R2, is the square of the correlation. It measures the proportion of variation in the dependent variable that can be attributed to the independent variable. The R-squared value R2 is always between 0 and 1 inclusive.

For instance, a perfect positive linear association. Correlation r = 1; R-squared = 1.00. While a large positive linear association. Correlation r = 0.9; R=squared = 0.81. In a small positive linear association. The points are far a bit from the trend line. Correlation r = 0.45; R-squared = 0.2025. And when there is no association. Correlation r = 0.0; R-squared = 0.0. For a correlation to be considered meaningful, it depends on the discipline. Those are some rough guidelines **(The UCWbL @ DePaul University, n.d.)**.

Table 3. 2 *Correlation guidelines*

Discipline	r meaningful if	R_2 meaningful if
Physics	r < -0.95 or 0.95 < r	$0.9 < R^2$
Chemistry	r < -0.9 or 0.9 < r	$0.8 < R^2$
Biology	r < -0.7 or 0.7 < r	$0.5 < R^2$
Social Sciences	r < -0.55 or 0.55 < r	$0.3 < R^2$

To make it easy for table representation, each of the variables isassigned to an Alphabetical letter as follows in the next table:

Table 3. 3 *Variables with their corresponding Alphabetical letter*

foreign_visitors	A
local_visitors	B
employment	C
internet_access	D
income_households	E
e_commerce_active	F
gdp	G
population	H
foreign_population	I
rental_price_center	J
avg_net_salary	K
air_quality	L
urban_greenery	M
life_expectancy	N
startups_number	O
taxi_cost_center	P

At first, we test for correlation, in the next table the correlation matrix is represented. The number of foreign visitors in cities (A) is not correlated with 3 of the indicators (rental_price_center, life_expectancy, taxi_cost_center), negatively correlated with 3 indicators (avg_net_salary, air_quality, urban_greenery), and strongly correlated with 9 of the indicators. We can also notice the 98% correlation between Employment and population, and 96% between employment and e_commerce_active.

Table 3. 4 *Variables correlation matrix*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	1.00	0.61	0.77	0.78	0.75	0.76	0.81	0.78	0.78	0.26	-0.16	-0.46	-0.28	0.19	0.61	0.30
B	0.61	1.00	0.62	0.60	0.83	0.62	0.74	0.60	0.43	0.17	-0.05	-0.16	0.00	0.25	0.24	0.35
C	0.77	0.62	1.00	0.99	0.77	0.96	0.76	0.98	0.74	0.23	-0.04	-0.47	-0.21	0.25	0.61	0.35
D	0.78	0.60	0.99	1.00	0.75	0.95	0.75	1.00	0.75	0.20	-0.07	-0.53	-0.28	0.28	0.61	0.33
E	0.75	0.83	0.77	0.75	1.00	0.79	0.93	0.74	0.68	0.43	0.13	-0.31	-0.11	0.30	0.57	0.52
F	0.76	0.62	0.96	0.95	0.79	1.00	0.78	0.93	0.81	0.33	0.05	-0.36	-0.09	0.24	0.72	0.35
G	0.79	0.64	0.66	0.65	0.80	0.69	1.00	0.64	0.70	0.29	-0.07	-0.35	-0.13	0.12	0.63	0.35
H	0.78	0.60	0.98	1.00	0.74	0.93	0.74	1.00	0.74	0.17	-0.11	-0.56	-0.32	0.29	0.58	0.32
I	0.78	0.43	0.74	0.75	0.68	0.81	0.80	0.74	1.00	0.35	0.02	-0.43	-0.12	0.18	0.85	0.21
J	0.26	0.17	0.23	0.20	0.43	0.33	0.37	0.17	0.35	1.00	0.83	0.13	0.19	0.52	0.42	0.58
K	-0.16	-0.05	-0.04	-0.07	0.13	0.05	-0.0	-0.11	0.02	0.83	1.00	0.35	0.30	0.44	0.08	0.46
L	-0.46	-0.16	-0.47	-0.53	-0.31	-0.36	-0.38	-0.56	-0.43	0.13	0.35	1.00	0.71	-0.16	-0.26	-0.09
M	-0.28	0.00	-0.21	-0.28	-0.11	-0.09	-0.14	-0.32	-0.12	0.19	0.30	0.71	1.00	-0.15	-0.01	0.07
N	0.19	0.25	0.25	0.28	0.30	0.24	0.32	0.29	0.18	0.52	0.44	-0.16	-0.15	1.00	0.06	0.40
O	0.61	0.24	0.61	0.61	0.57	0.72	0.70	0.58	0.85	0.42	0.08	-0.26	-0.01	0.06	1.00	0.22
P	0.30	0.35	0.35	0.33	0.52	0.35	0.41	0.32	0.21	0.58	0.46	-0.09	0.07	0.40	0.22	1.00

Moreover, in the next table, information about the data and variables are presented, the count column represents the sample number, the mean column represents the average values for each column, the std column represents the standard deviation which is the how much the numbers of a variable differ from the mean value for that variable, while the min and max column represents the minimum and maximum values in each variable, lastly the 25%, 50%, 75%, It describes the distribution of the data. 50% should be a value that describes the median of the data. 25% and 75% is the border of the upper/lower quarter of the data. We can notice that the standard deviation figures are often very high which indicates high variations in the data.

Table 3. 5 *Variables descriptive analysis*

	count	mean	std	min	25%	50%	75%	max
A	42	4454600.83	4280449.97	439281	1614110	2766374.5	5685981.5	19244789
B	42	5092873.02	5458086.6	560772	1520904.25	2942463	6865928	26417043
C	42	1965833.01	1467440.18	591511.7	958133.031	1370037.18	2441824.06	6655851.7
D	42	2569903.42	1961376.66	664437.8	1235788.52	1744511.95	3170524.8	8802610.88
E	42	97682.663	94335.2925	13451	41812.2675	71944.2	112514.97	413501.36
F	42	1850147.18	1621396.79	247216.2	811243.475	1295270.23	2251871.63	8084030.4
G	42	155375.7	155175.3	26307.20	69312.54	96908.67	189896.0	733874.75
H	42	2768260.1	2068799.82	761702	1305999.5	1869089.5	3633380.75	8982256
I	42	716654.717	994942.633	30000	207744.735	455671.963	807155.243	4748613.41
J	42	1216.42095	474.698291	510.4	968	1056	1386	2464
K	42	34509.1095	14139.7251	15191.44	25489.86	33264.44	38052.96	79200
L	42	0.62261905	0.20604356	0.22	0.505	0.655	0.795	0.99
M	42	0.70714286	0.18606694	0.1	0.65	0.735	0.8525	1
N	42	81.1807143	2.15056362	73.63	81.045	81.32	82.43	83.36
O	42	694.333333	1584.85176	30	151.25	255.5	556	9805
P	42	37.0966667	20.0303807	7	20.63	33	47.4675	90.26

By plotting any two correlated variables on a scatter plot we can investigate if there is a high data variation. Thus, according to the figure, high variations in the data exists. Fixing this issue can be done by using log transformation. The log transformation can be used to make highly skewed distributions less skewed and to make all of the features comparable. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics

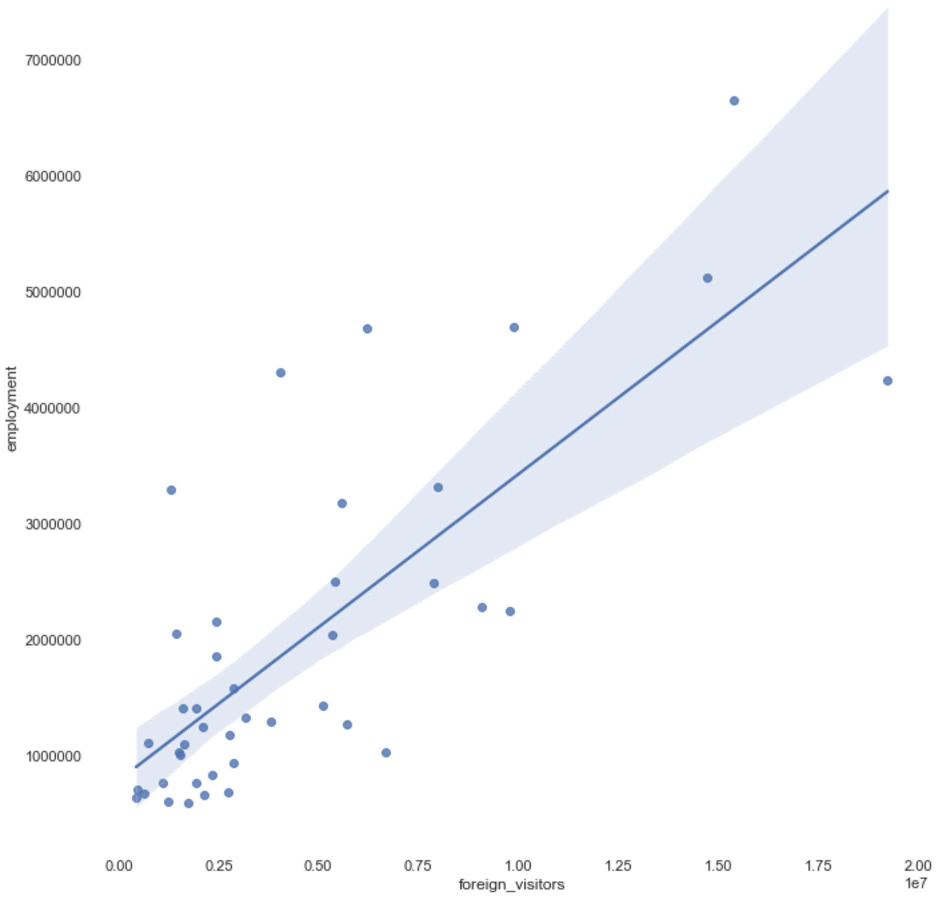


Figure 3. 3 *Scatter plot between foreign visitors and employment*

The next matrix represents the correlation after the log transformation, we have 5 strongly correlated variables, 2 somewhat strong, and the rest of the variables are the same as before.

Table 3. 6 *Variables correlation matrix after log transformation*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	1.00	0.58	0.74	0.74	0.57	0.73	0.71	0.75	0.58	0.16	-0.22	-0.44	-0.16	0.13	0.39	0.34
B	0.58	1.00	0.68	0.68	0.78	0.66	0.76	0.68	0.47	0.25	-0.01	-0.22	0.01	0.39	0.30	0.40
C	0.74	0.68	1.00	0.99	0.76	0.94	0.81	0.99	0.75	0.26	0.03	-0.47	-0.20	0.32	0.40	0.43
D	0.74	0.68	0.99	1.00	0.76	0.93	0.82	1.00	0.77	0.24	0.01	-0.53	-0.27	0.36	0.37	0.43
E	0.57	0.78	0.76	0.76	1.00	0.80	0.86	0.74	0.67	0.60	0.36	-0.34	-0.08	0.57	0.40	0.59
F	0.73	0.66	0.94	0.93	0.80	1.00	0.83	0.91	0.81	0.41	0.21	-0.34	-0.04	0.37	0.48	0.47

G	0.75	0.66	0.75	0.74	0.73	0.78	1.00	0.74	0.61	0.31	0.01	-0.40	-0.10	0.26	0.44	0.46
H	0.75	0.68	0.99	1.00	0.74	0.91	0.81	1.00	0.75	0.20	-0.05	-0.56	-0.31	0.34	0.33	0.40
I	0.58	0.47	0.75	0.77	0.67	0.81	0.72	0.75	1.00	0.38	0.21	-0.45	-0.12	0.27	0.50	0.35
J	0.16	0.25	0.26	0.24	0.60	0.41	0.37	0.20	0.38	1.00	0.84	0.10	0.22	0.64	0.49	0.65
K	-0.22	-0.01	0.03	0.01	0.36	0.21	0.08	-0.05	0.21	0.84	1.00	0.31	0.32	0.52	0.28	0.54
L	-0.44	-0.22	-0.47	-0.53	-0.34	-0.34	-0.43	-0.56	-0.45	0.10	0.31	1.00	0.71	-0.18	-0.06	-0.13
M	-0.16	0.01	-0.20	-0.27	-0.08	-0.04	-0.12	-0.31	-0.12	0.22	0.32	0.71	1.00	-0.15	0.14	0.09
N	0.13	0.39	0.32	0.36	0.57	0.37	0.33	0.34	0.27	0.64	0.52	-0.18	-0.15	1.00	0.03	0.50
O	0.39	0.30	0.40	0.37	0.40	0.48	0.48	0.33	0.50	0.49	0.28	-0.06	0.14	0.03	1.00	0.32
P	0.34	0.40	0.43	0.43	0.59	0.47	0.50	0.40	0.35	0.65	0.54	-0.13	0.09	0.50	0.32	1.00

The following figure is showing a heatmap representation of the correlation matrix.

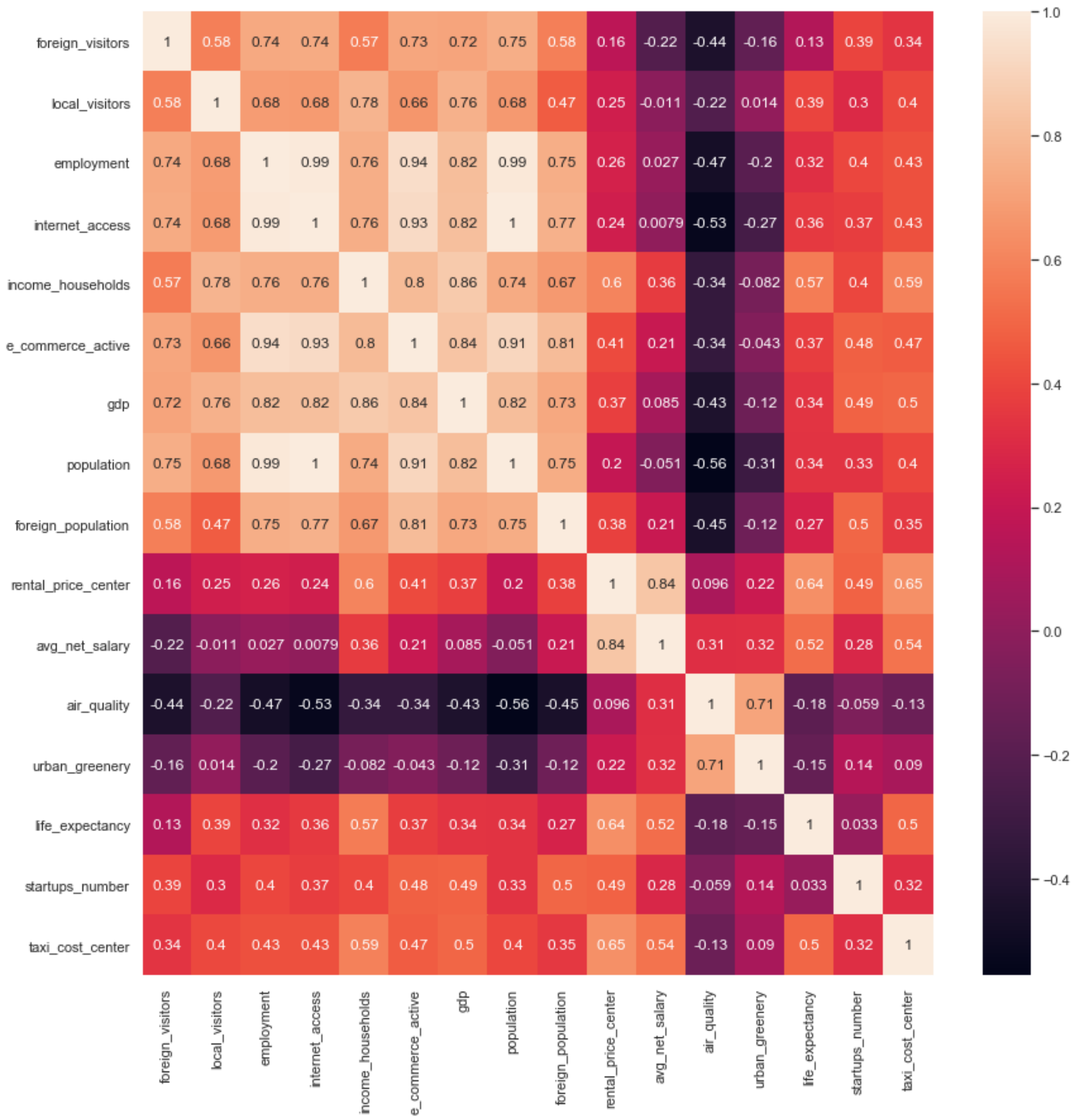


Figure 3. 4 *Heatmap representing the correlation between all the variables*

In the following table, data and variables after the log transformation is presented. Accordingly, data variation became less and for instance to find out how skewed the data are, note that for the 'foreign_visitors' the mean is higher than the median, which means the data is right-skewed in this case.

Table 3. 7 *Variables descriptive analysis after log transformation*

	count	mean	std	min	25%	50%	75%	max
A	42	6.47522457	0.40105783	5.64274242	6.20790994	6.44190716	6.7547752	7.28431315
B	42	6.50605045	0.42571075	5.74878632	6.18208869	6.46869027	6.83652303	7.4218842
C	42	6.19249633	0.29265337	5.77196334	5.98119608	6.13655018	6.38739434	6.82320364
D	42	6.30248886	0.30379738	5.82245433	6.09179694	6.24167391	6.50102411	6.9446115

E	42	4.83579134	0.36736675	4.12875457	4.62116452	4.85699579	5.05119532	5.61647694
F	42	6.13971235	0.32922458	5.39307693	5.90816472	6.1123393	6.35244837	6.90762794
G	42	5.035676	0.362133	4.420075	4.840691	4.986357	5.277395	5.865622
H	42	6.33631973	0.30296342	5.8817851	6.11594272	6.27158167	6.55983632	6.95338543
I	42	5.60591415	0.46600779	4.47712125	5.31752594	5.65864901	5.90682928	6.67656681
J	42	3.05468786	0.16419439	2.70791067	2.98587536	3.02366392	3.14159547	3.3916407
K	42	4.50697719	0.16309878	4.18159894	4.40601724	4.52197968	4.58038824	4.89872518
L	42	-0.2343764	0.16898833	-0.6575773	-0.2967717	-0.1837714	-0.0996589	-0.0043648
M	42	-0.1738221	0.16880334	-1	-0.1870866	-0.1338031	-0.0693566	0
N	42	1.90929814	0.01185023	1.8670548	1.90872622	1.91019704	1.9160853	1.92095771
O	42	2.45411736	0.52810743	1.47712125	2.17965137	2.40725032	2.74493724	3.9914476
P	42	1.50292967	0.25380452	0.84509804	1.31391459	1.51831445	1.67639439	1.95549533

For the coming analysis, strongly correlated variables only are used when running the simple linear regression model.

The next figure is showing a boxplot for each of the selected variables. A box plot is a figure that gives you a good indication of how the values in the data are spread out. Boxplots are a standardized way of displaying the distribution of data based on a five-number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”) as explained in the figure after. Median (50th Percentile) is the middle value of the dataset, first quartile (Q1/25th Percentile) is the middle number between the smallest number (not the “minimum”) and the median of the dataset, third quartile (Q3/75th Percentile) is the middle value between the median and the highest value (not the “maximum”) of the dataset, interquartile range (IQR) is the 25th to the 75th percentile and outliers (shown as green circles), in the figure we can see that there is no outliers and that and that most of the variables are normally distributed.

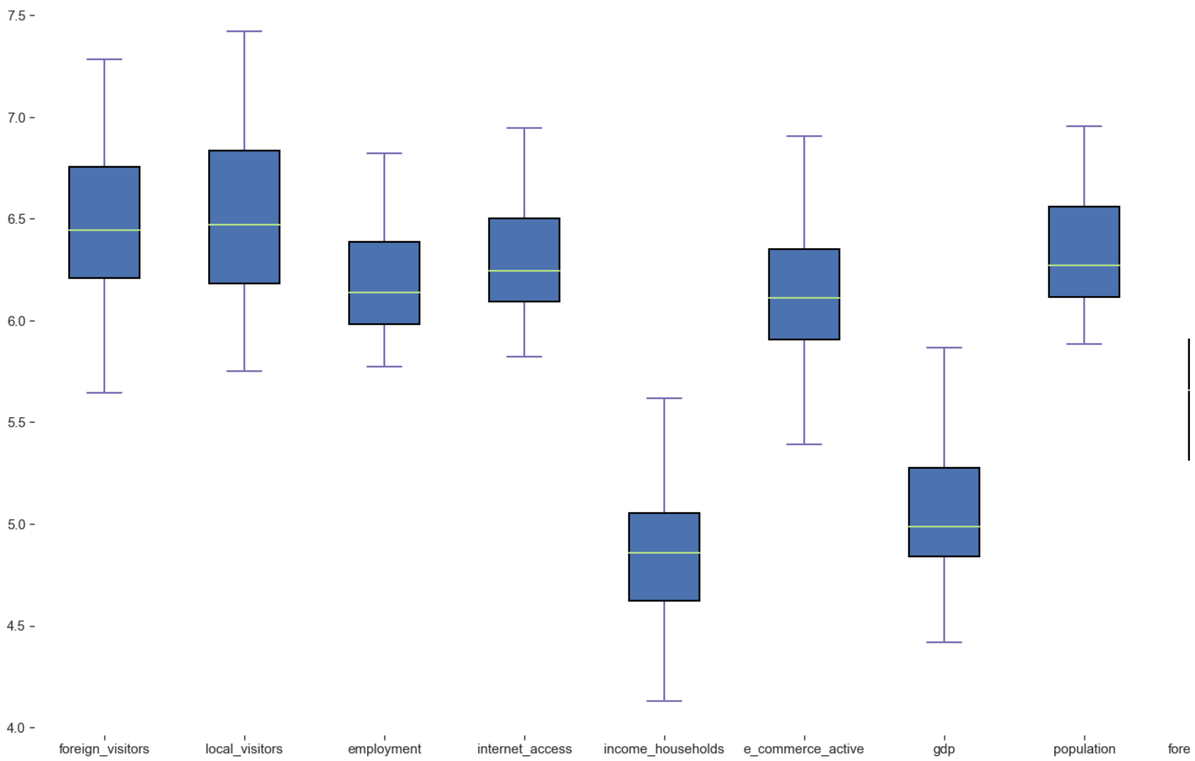


Figure 3. 5 *A box plot representing the selected variables*

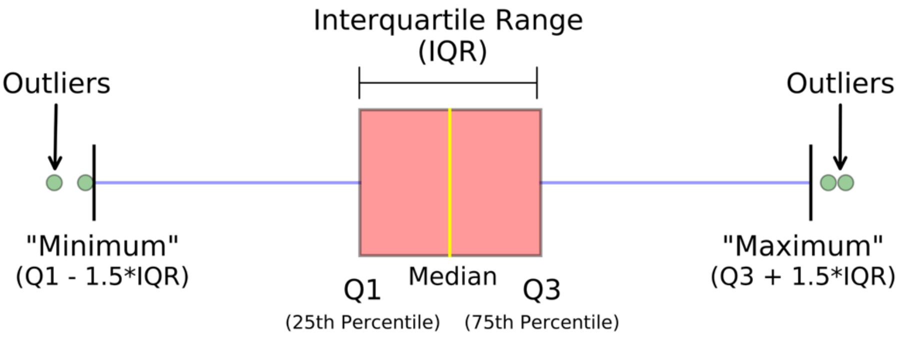


Figure 3. 6 *An explanation of the box plot results, source: towards data science*

Since the dependent variable ‘y’ in the regression model is ‘foreign_visitors’, in the next following figures we will be comparing that variable to the other variables on a scatter plot for data visualization and interpretation.

In the next figure, we can see the relationship between ‘foreign_visitors’ and ‘population’, there is a positive correlation, what does a “positive relationship” mean in this context? It means that, in general, higher scores on one variable tend to be paired with higher scores on the other and that lower scores on one variable tend to be paired with lower scores on the other, as in a lot of cases when population increases in a city, the number of foreign visitors also increases.

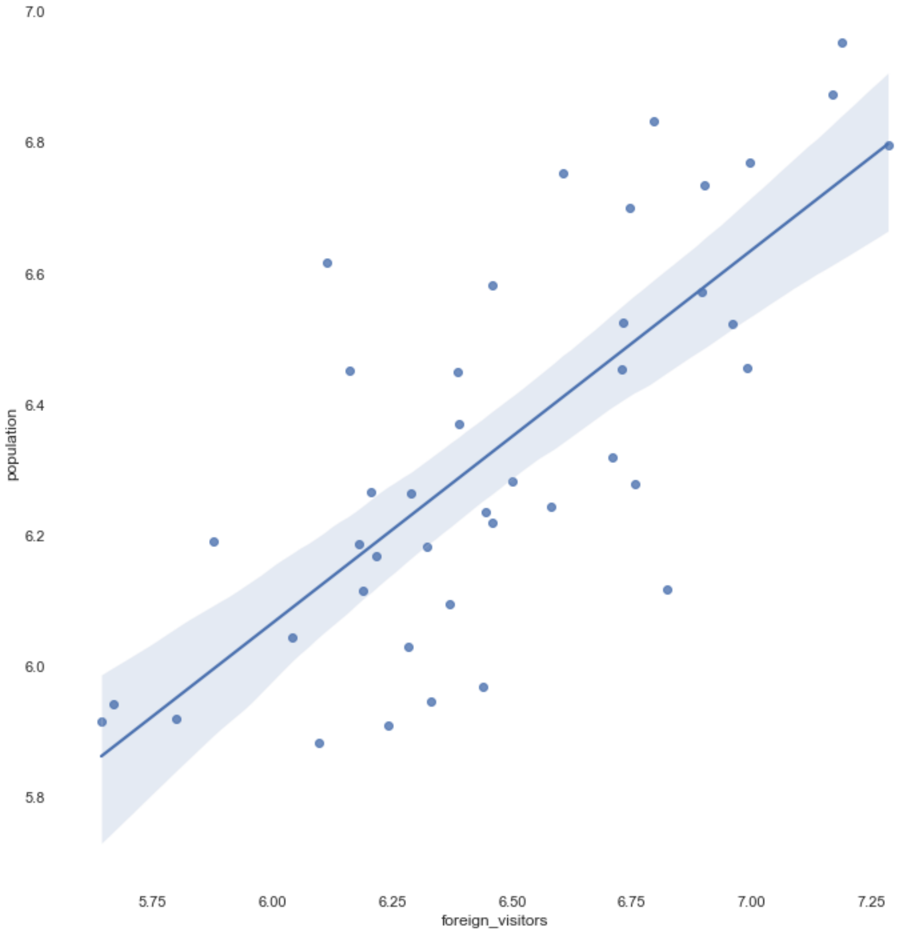


Figure 3. 7 *A scatter plot between foreign visitors and population*

In the next figure, the relationship between ***'foreign_visitors'*** and ***'employment'*** is demonstrated, there is also a positive correlation when the number of employment populations increases in a city, the number of foreign visitors as well increase.

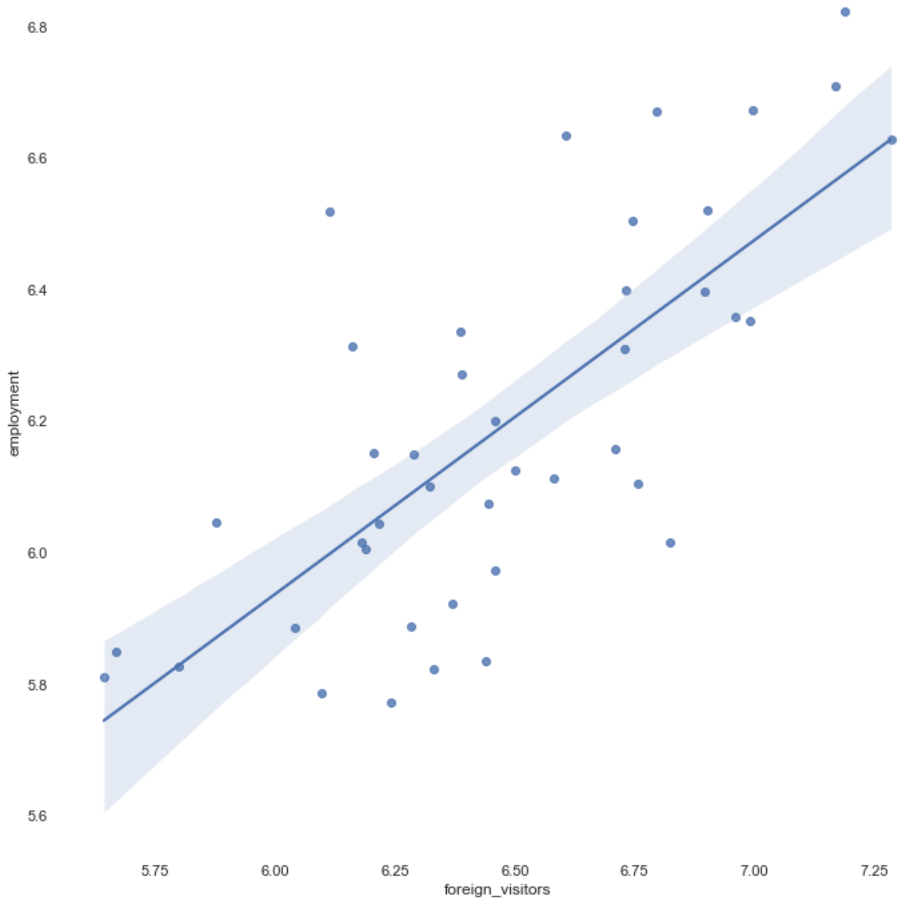


Figure 3. 8 *A scatter plot between foreign visitors and employment after log transformation*

In the following figure, we divided the number of foreign visitors and employment by population to represent the relationship between ***'foreign_visitors_per_capita'*** and ***'employment_per_capita'***, the result is showing that there is absolutely no relationship between the two variables.

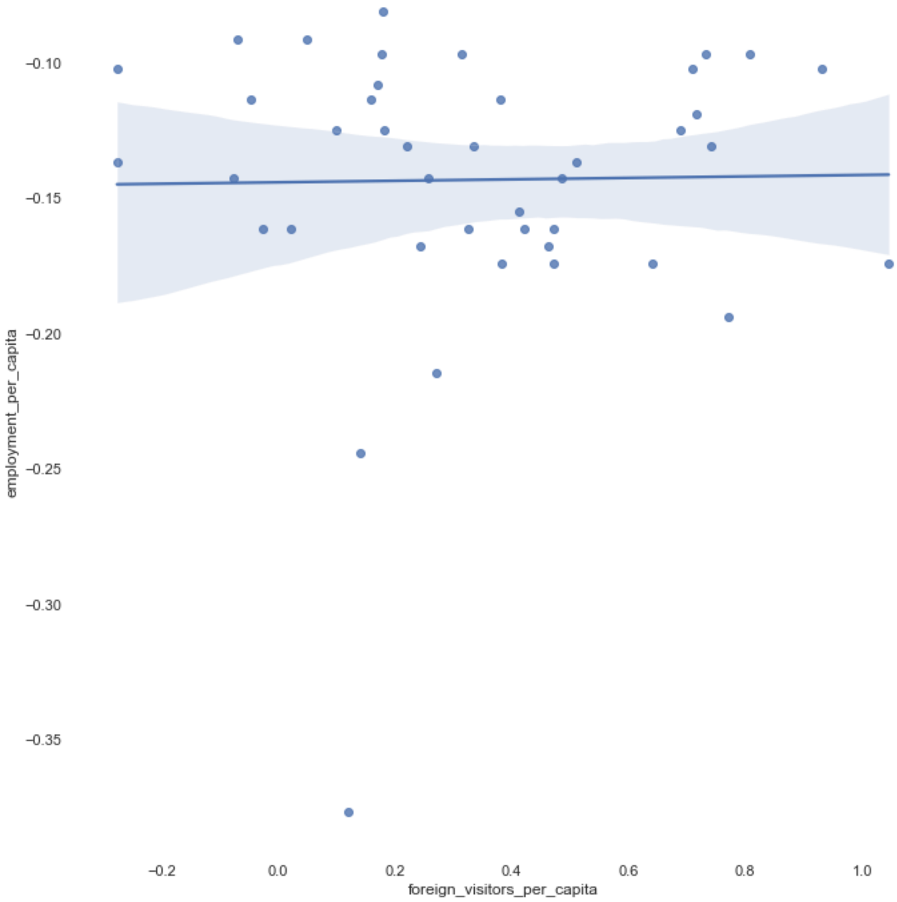


Figure 3. 9 *A scatter plot between foreign visitors and employment after log transformation*

In the upcoming figure, we show the relationship between ‘foreign_visitors’ and 'local_visitors', some cities' local visitors are higher than their foreign visitors and some have the same. However, in 58% of the cases the relationship is positive.

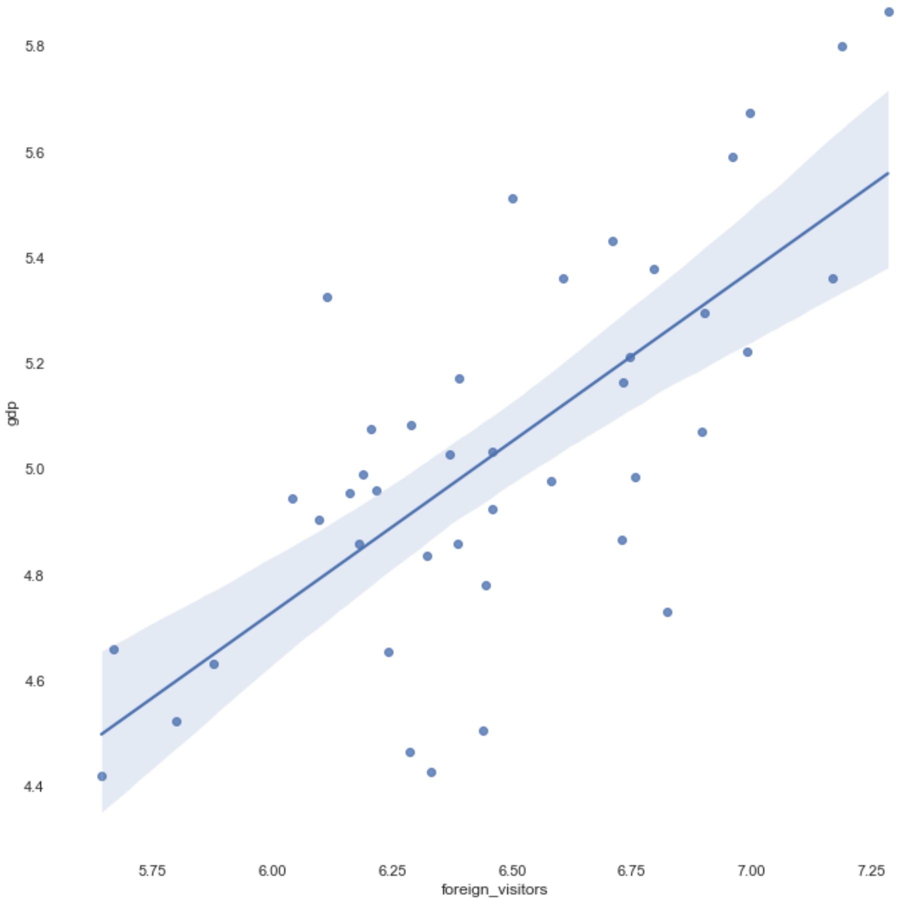


Figure 3. 10 *A scatter plot between foreign visitors and GDP*

The linear regression equation has the form y= a + bx, where 'y' is the dependent variable, what we are trying to investigate and predict, 'x' is the independent variable, all the variables that will help us predict 'y', while 'b' is the slope of the line and 'a' is the y-intercept. We will run a simple linear regression model on all the variables against 'foreign_visitors', we will be doing this over Python while using the OLS method. At the beginning we will be using all the sample to evaluate the variables and to choose the best variable in order to build the prediction mode, however, when the time comes to build the model we will be using a training sample of 60% of the data sample which equals 25 observations, keeping 40% for the testing and predicting model, which is 17 observations. In the next table, we can find the results of the linear regression. Where the independent variables will be run individually against each of the dependent variables, resulting in 8 simple linear regression models, using a confidence level of 95%.

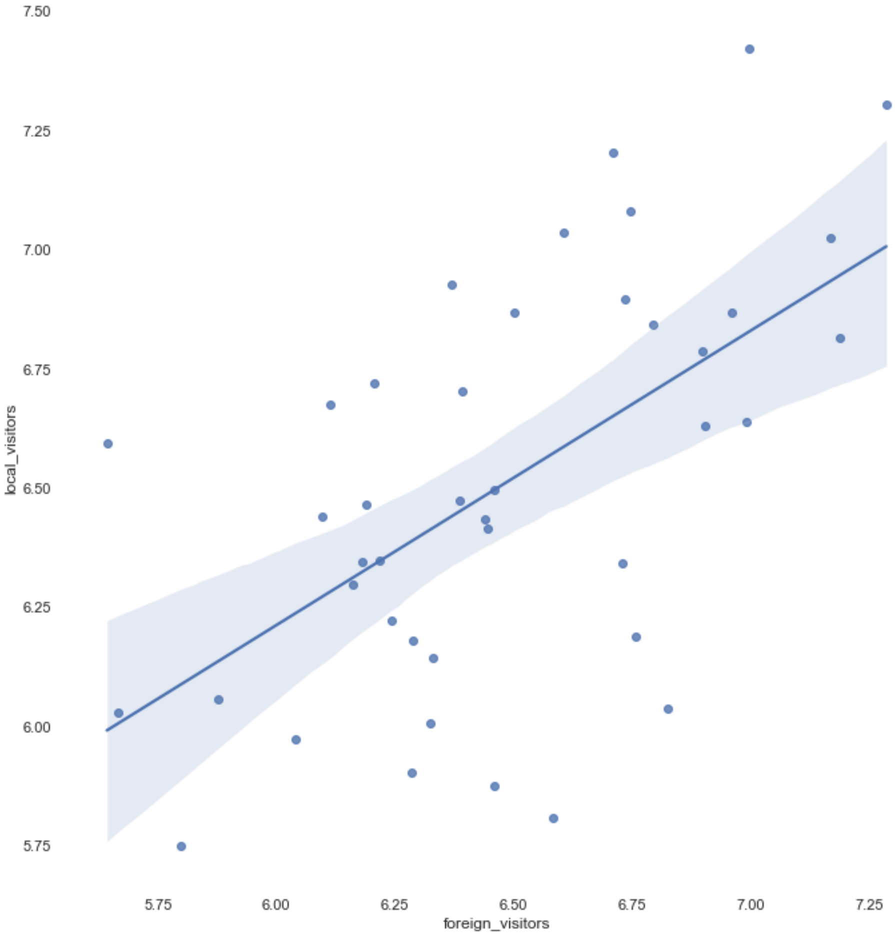


Figure 3. 11 *A scatter plot between foreign visitors and local visitors*

In the next figure, we can see the relationship between ***'foreign_visitors'*** and ***'foreign_population'***, there is a positive correlation, however not a very strong one, in some cases we can see clusters with low foreign populations and high foreign visitors and in others, we can see the opposite.

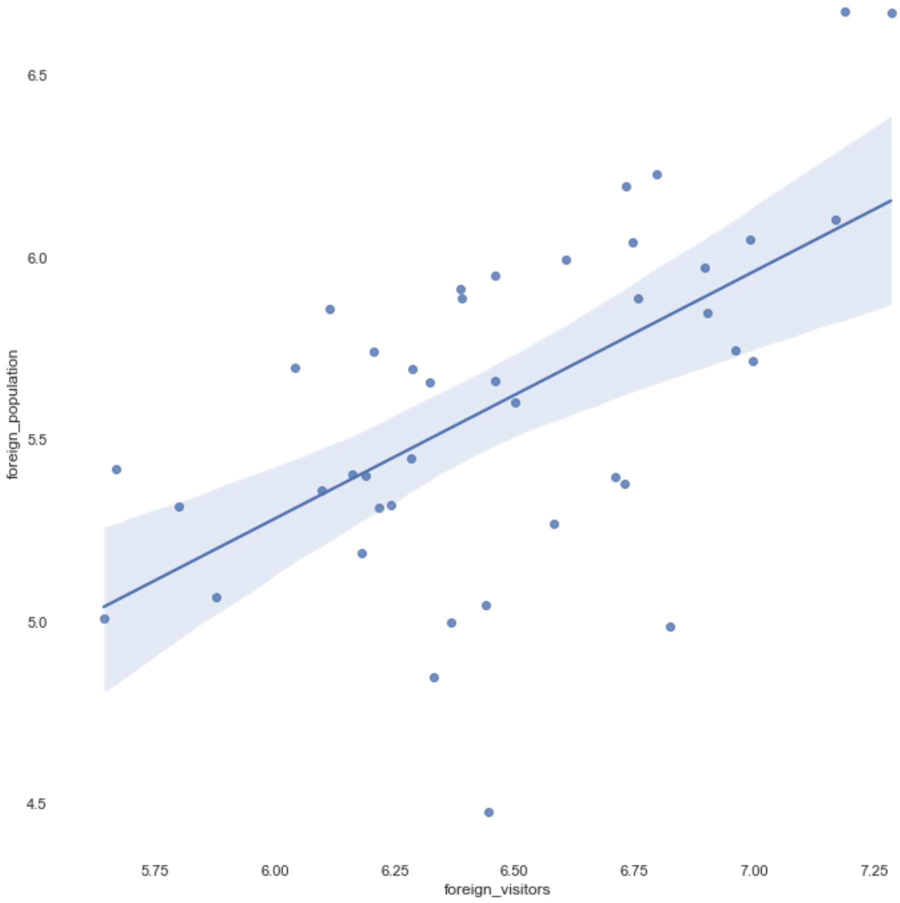


***Figure 3. 12** A scatter plot between foreign visitors and foreign population*

This figure represents the relationship between ‘foreign_visitors’ and 'income_households', again we can notice a positive relationship 57% with 43% of the cases that go against the trend.

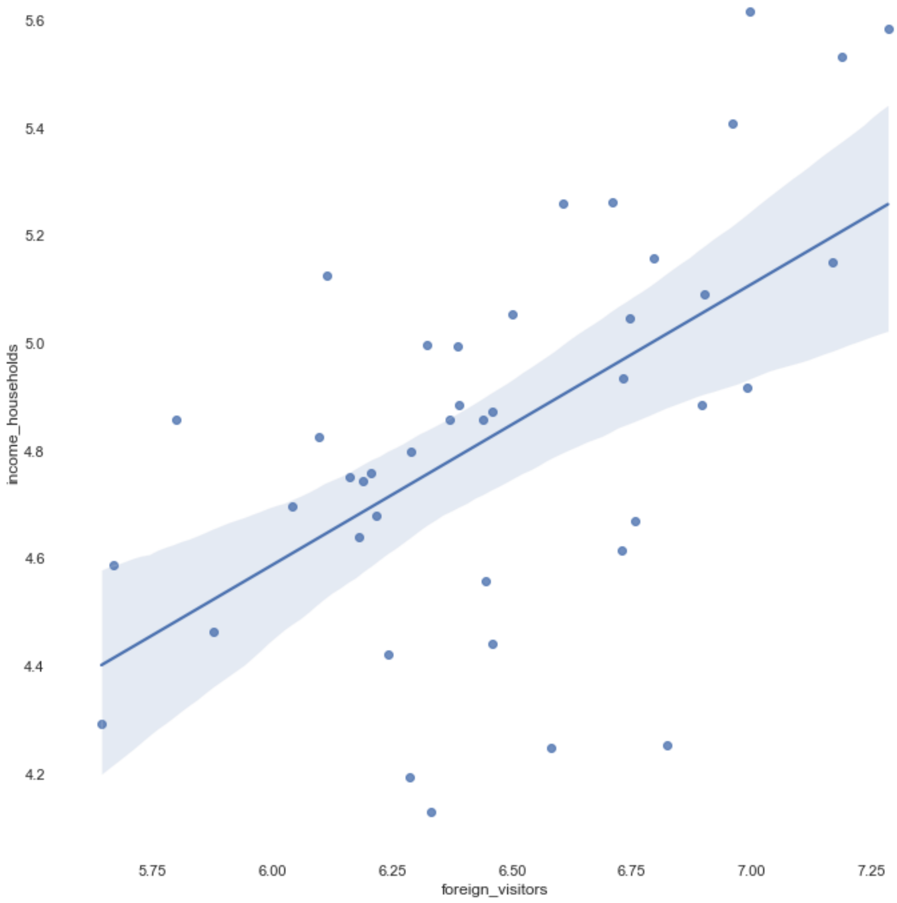


Figure 3. 13 *A scatter plot between foreign visitors and income households*

In the last two scatters we have the relationship between ‘foreign_visitors’ and 'internet_access' and 'e_commerce_active' with a correlation of 74% and 73% respectively.

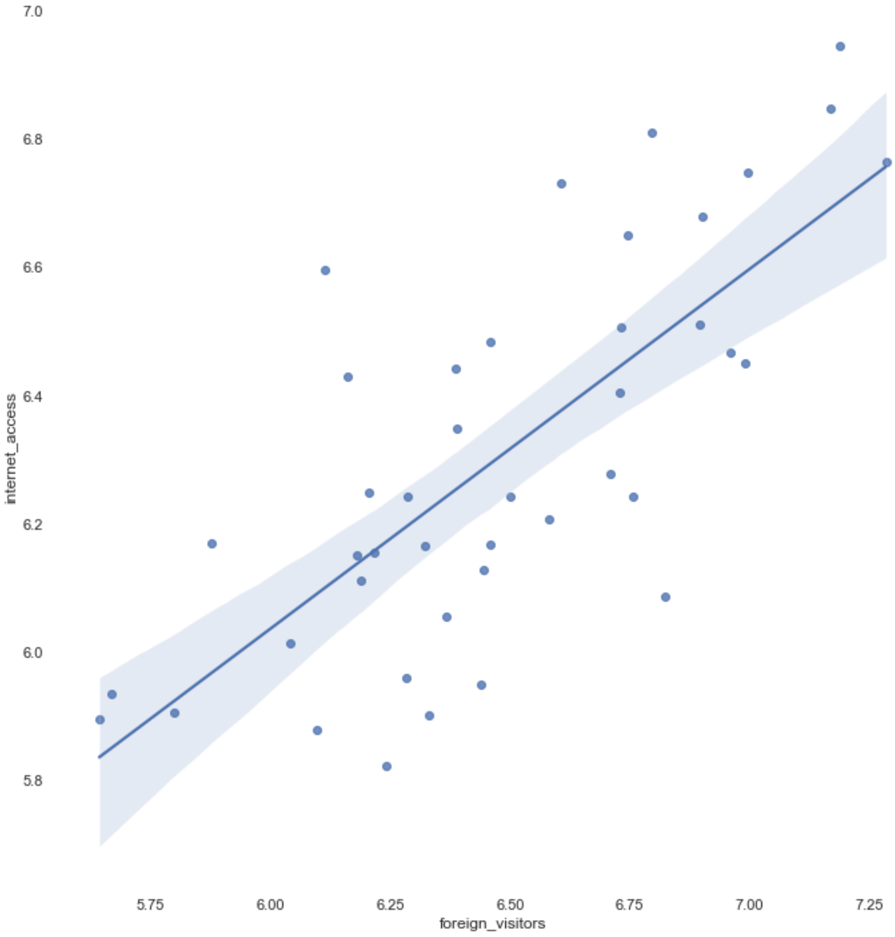


Figure 3. 14 *A scatter plot between foreign visitors and internet access*

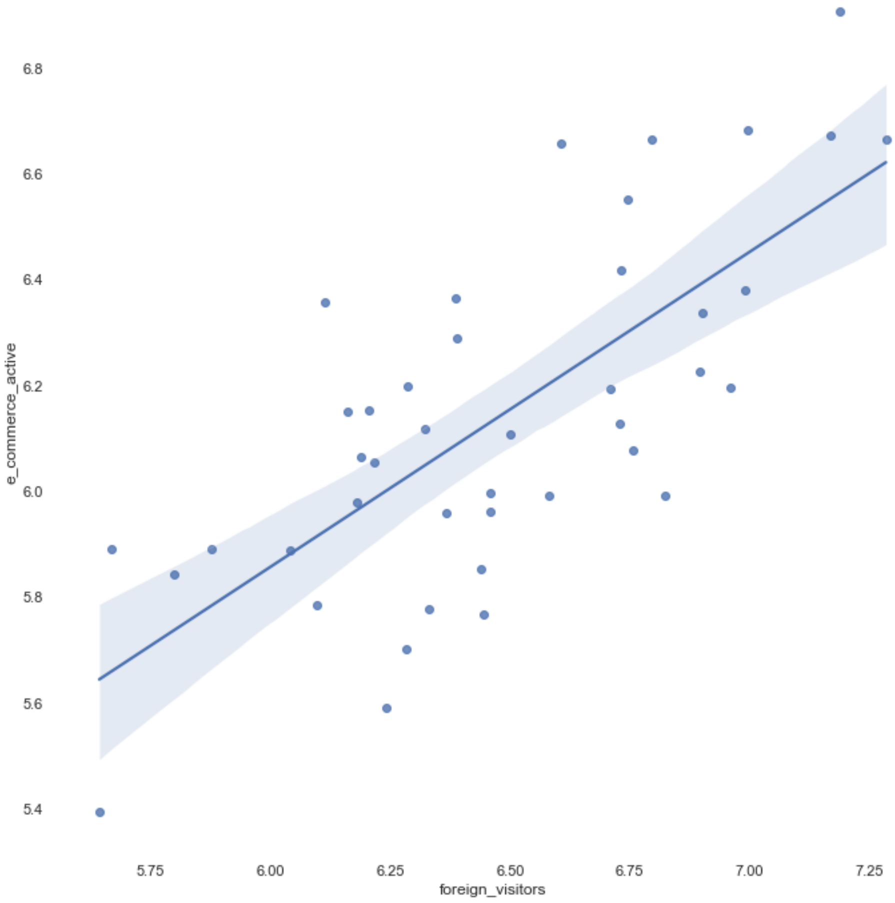
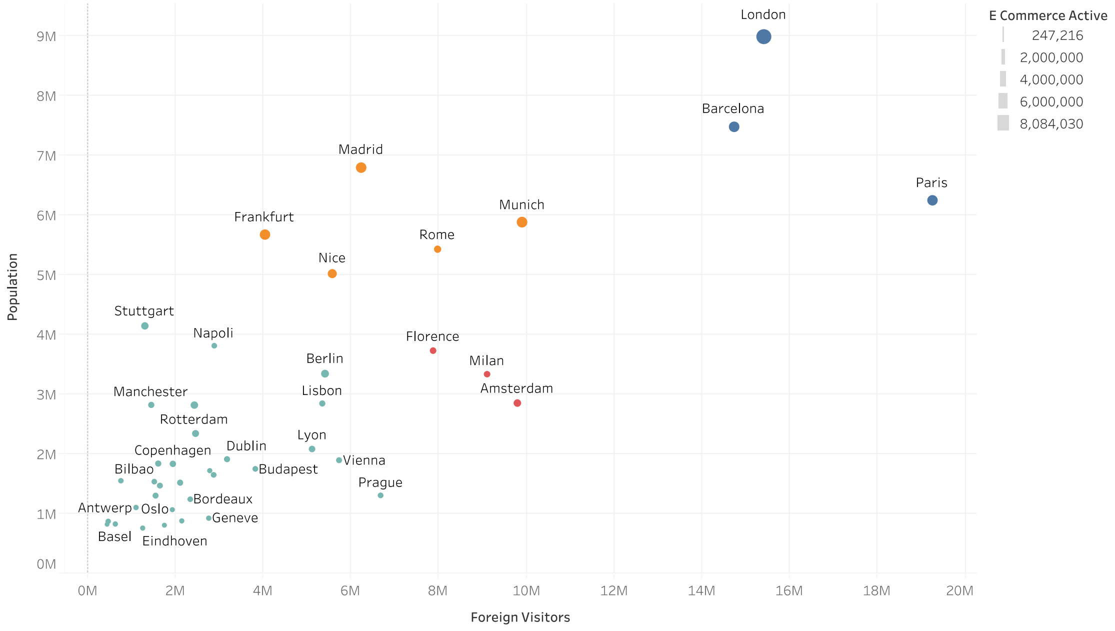


Figure 3. 15 *A scatter plot between foreign visitors and e-commerce active*

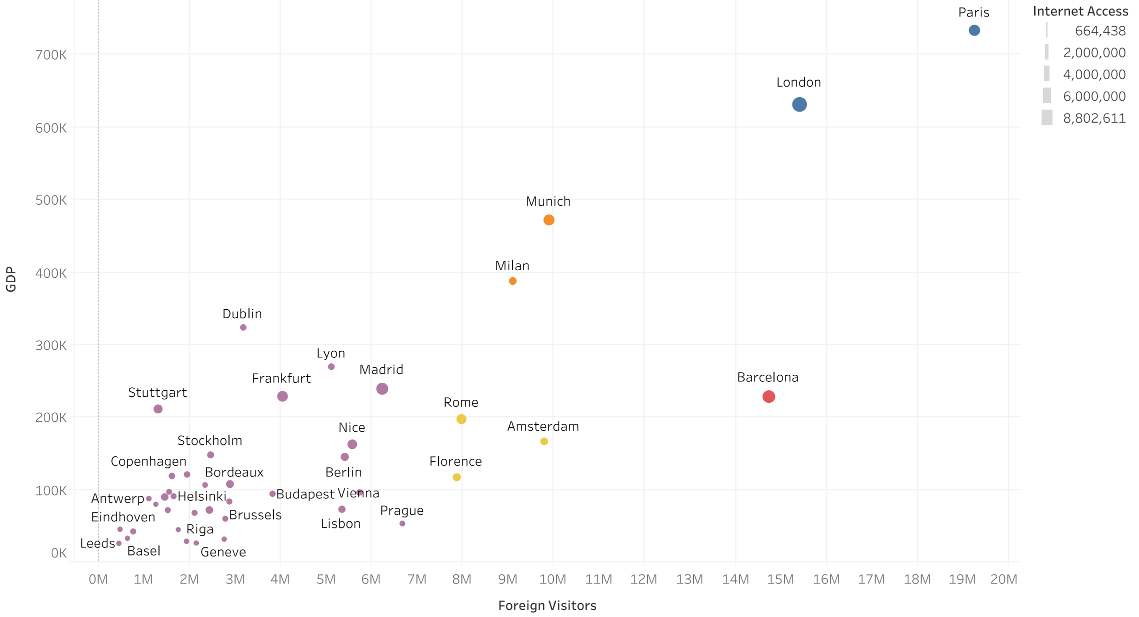
In the next visualization, the relationship between 'foreign_visitors' and 'population' filtered with the 'e-commerce_active' is shown. The trend of higher GDP leads to higher e-commerce active and higher numbers of foreign visitors is noticed.



The trend of sum of Population for Foreign Visitors. Color shows details about City & Foreign Visitors (group). Size shows sum of E Commerce Active. The marks are labeled by City. The data is filtered on Action (City), which keeps 42 members.

Figure 3. 16 *A clustered scatter plot between foreign visitors and population filtered by e-commerce active*

In the next figure, the relationship between 'foreign_visitors' and 'GDP' filtered with the 'internet access' is shown, as before, the trend of higher GDP leads to higher internet access and higher numbers of foreign visitors.



The trend of sum of GDP for Foreign Visitors. Color shows details about City & Foreign Visitors 1 (group). Size shows sum of Internet Access. The marks are labeled by City. The data is filtered on Action (City), which keeps 42 members.

Figure 3. 17 *A clustered scatter plot between foreign visitors and GDB filtered by internet access*

In the next visualization, we have the relationship between 'foreign_visitors' and 'local_visitors'. A higher number of foreign visitors does not necessarily means a higher number of local visitors.

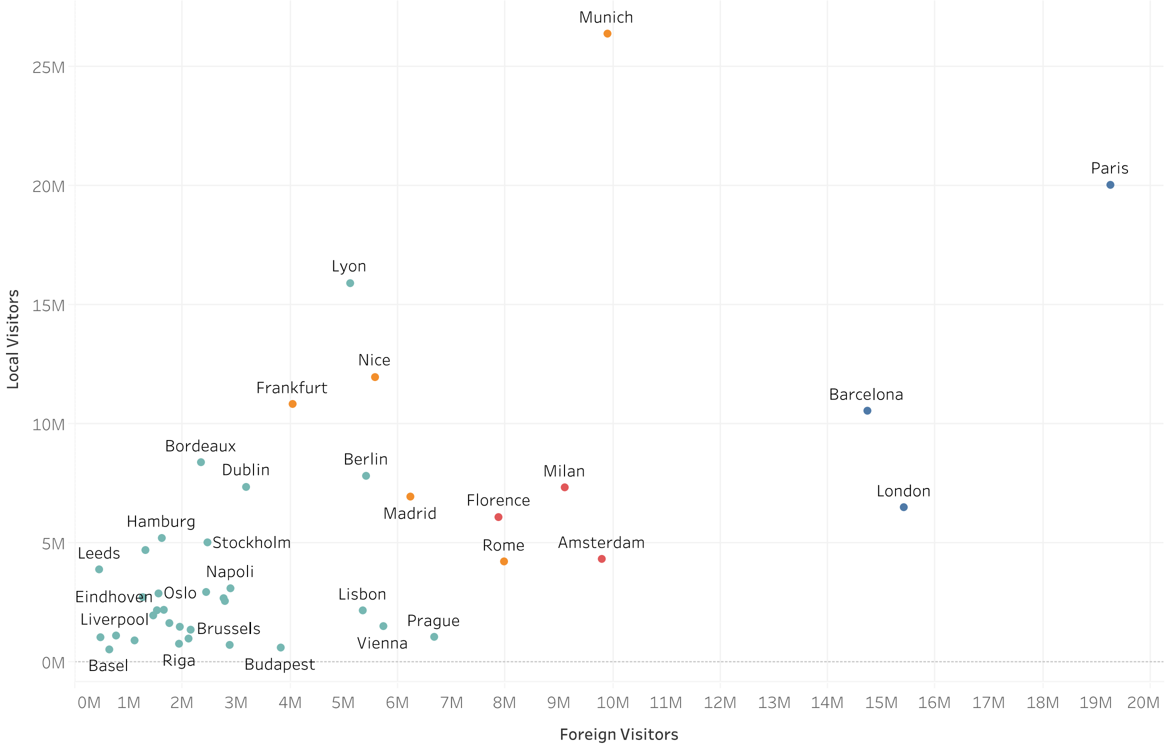


Figure 3. 18 *A clustered scatter plot between foreign visitors and local visitors*

Generally, we emphasized that a city can be attractive due to its economic situation, when GDP and employment were high the number of foreign visitors was also high, moreover, population of cities as well had a strong correlation, that might lead to the fact that the more people live in a certain city, the more people knowing about it.

3.2.2 Simple Linear Regression

The linear regression equation has the form y= a + bx, where 'y' is the dependent variable, what we are trying to investigate and predict, 'x' is the independent variable, all the variables that will help us predict 'y', while 'b' is the slope of the line and 'a' is the y-intercept. A simple linear regression model is applied on all the variables against 'foreign_visitors', using Python and the OLS method. At the beginning all the sample is used to evaluate the variables and to choose the best variable in order to build the prediction mode, however, when the time comes to build the model a training sample is used of 60% of the data sample which equals 25 observations, keeping 40% for the testing and predicting model, which is 17 observations. In the next table, we can find the results of the linear regression. Where the independent variables will be run individually against each of the dependent variables, resulting in 8 simple linear regression models, using a confidence level of 95%.

Table 3. 8 *Simple linear regression results*

Independent Variables	Coefficient	Significance (p-value)	Standard Error	F-stats
local_visitors	0.5484	0.0000590	0.121	20.5
Employment	1.0109	0.0000000362	0.146	47.74
internet_access	0.9781	0.0000000267	0.140	48.68
income_households	0.6214	0.0000863	0.142	19.17
e_commerce_active	0.8837	0.000000100	0.133	44.43
Gdp	0.7925	0.000000139	0.122	41.97
Population	0.9989	0.00000000873	0.137	52.88
foreign_population	0.5031	0.0000925	0.110	20.76

To interpret the results, explaining what the data referring to is explained below:

· The Standard Error (std err), is an indication of the reliability of the mean. A small std err is an indication that the sample mean is a more accurate reflection of the actual population mean. Larger sample size will normally result in a smaller std err.

· The significance helps determine whether the relationships that we observe in the sample also exist in the larger population. This is represented by the p-value, the value for each independent variable tests the null hypothesis that the variable has no correlation with the dependent variable. It is standard practice to use the p-values to decide whether to include variables in the final model. For the results above, we would consider working with all the variables that their p-value is less than 0.05 (95% confidence level), which means all of them are significant enough to be in the model. Thus, keeping values that their p-value is bigger than 0.05 significance level can reduce the model's precision.

· The F-statistic value in regression is the result of a test where the null hypothesis is that all of the regression coefficients are equal to zero. In other words, the f-test decides whether the added variables improve the model or not, usually, a higher f-statistic is better.

· The regression coefficients are estimates of the actual population parameters. To obtain unbiased coefficient estimates that have the minimum variance, and to be able to trust the p-values, the model must satisfy the six classical assumptions of OLS linear regression. The sign of each coefficient indicates the direction of the relationship between a predictor variable and the response variable. A positive sign indicates that as the predictor variable increases, the response variable also increases. A negative sign indicates that as the predictor variable increases, the response variable decreases.

From the table above we have concluded that for a simple linear regression the best variable to fit the data is 'population' as it has the lowest p-value and the highest f-stats while having a good standard error. Hence, the prediction equation would be represented as follows: Y = 0.1461 + 0.9989 * X, It also can be: foreign visitors = 0.1461 + 0.9989 *population, telling us that the foreign visitors each year is predicted to increase 0.7162 times when the population goes up by one.

3.2.2.1 Measuring Model Fitness

In the following table we will be testing some methodologies to assess and check the fitness of the model.

Table 3. 9 *Assessment methodologies for simple linear regression*

Dep. Variable:	foreign_visitors	R-squared:	0.569
Indep. Variable:	population	Adj. R-squared:	0.559
Model:	OLS	F-statistic:	52.88
Method:	Least Squares	Prob (F-statistic):	0
Df Residuals:	40	Df Model:	1
No. Observations:	42	Log-Likelihood:	-3.0240
Covariance Type:	nonrobust	AIC:	10.05
Omnibus:	0.723	BIC:	13.52
Prob(Omnibus):	0.696	Durbin-Watson:	2.391
Skew:	-0.286	Jarque-Bera (JB):	0.739
Kurtosis:	2.693	Prob(JB):	0.691
t-stats	7.272	Cond. No.	138
MAE	0.22	RMSE	0.27
MAPE	3.52		

- R-squared is a measure of fitness for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between the model and the dependent variable on a convenient (0 – 1) scale. An R-squared equal zero means that a model does not explain any of the variations in the response variable around its mean, while an R-squared equals one represents a model that explains all of the variations in the response variable around its mean. Usually, the larger the R2, the better the regression model fits our observations. In practice, we will never see a regression model with an R2 of 100%. In general, studies that try to explain human behavior like ours generally have R2 values of less than 50%. Therefore, our model is a good fit.

· Adj. R-squared: since R-squared tends to reward us for including too many independent variables in a regression model, and it doesn't provide any incentive to stop adding more. Adjusted R-squared uses different approaches to help us fight that impulse to add too many. The protection that adjusted R-squared provides is critical because too many terms in a model can produce results that you can't trust. These statistics help us include the correct number of independent variables in our regression model. We use adjusted R-squared to compare the goodness-of-fit for regression models that contain different numbers of independent variables. For instance, if we are comparing a model with five independent variables to a model with one variable and the five variable model has a higher R-squared. Is the model with five variables actually a better model, or does it just have more variables? To determine this, we need to compare the adjusted R-squared values. The adjusted R-squared adjusts for the number of variables in the model. Importantly, its value increases only when the new variable improves the model fit but starts to decrease when a new variable doesn't improve the model fit by a sufficient amount. Hence, we won't be interpreting the Adj. R-squared value now, we will in the next part when we implement a multiple linear regression.

· Prob(F-Statistic): This tells the overall significance of the regression. This is to assess the significance level of all the variables together unlike the t-statistic that measures it for individual variables. Usually, when the value is closer to zero, it implies that overall the regressions are strong and meaningful.

· AIC/BIC: It stands for Akaike's Information Criteria and is used for model selection. It penalizes the error mode in case a new variable is added to the regression equation. It is calculated as the number of parameters minus the likelihood of the overall model. A lower AIC implies a better model. Whereas, BIC stands for Bayesian information criteria and is a variant of AIC where penalties are made more severe. In general, a lower BIC means that a model is considered to be more likely to be the true model. Our AIC and BIC are not that low which means that the model can get better.

Log-likelihood values cannot be used alone as an index of fit because they are a function of sample size but can be used to compare the fit of different coefficients. Because we want to maximize the log-likelihood, the higher value is better. For example, a log-likelihood value of -3 is better than -7. We will discuss this more in the multiple linear regression model.

· Covariance type: In statistics, robust regression is a form of regression analysis designed to overcome some limitations of traditional parametric and non-parametric methods. OLS has favorable properties if its underlying assumptions are true, but can give misleading results if those assumptions are not true; thus, OLS is said to be not robust to violations of its assumptions.

· t-statistic(t) is the coefficient divided by its standard error and is used in hypothesis testing via Student's t-test. The regression software compares the t statistic on our variable with values in the Student's t distribution to determine the P-value, which is the number that we really need to be looking at.

· Df Residuals: is the sample size minus the number of parameters being estimated

· Omnibus tests whether the explained variance in a set of data is significantly greater than the unexplained variance, overall. One example is the F-test in the analysis of variance. There can be legitimate significant effects within a model even if the omnibus test is not significant. We usually hope to see a relatively small number.

· Prob(Omnibus) is one of the assumptions of OLS is that the errors are normally distributed. The omnibus test is performed in order to check this. Here, the null hypothesis is that the errors are normally distributed. Prob(Omnibus) is supposed to be close to 1 in order for it to satisfy the OLS assumption. In this case, Prob(Omnibus) is 0.696, which is very good.

· Durbin-Watson: Another assumption of OLS is of homoscedasticity. This implies that the variance of errors is constant. A value between 1 to 2 is preferred. Here, it is ~2 implying that the regression results are acceptable from the interpretation side of this metric.

· Jarque-Bera (JB)/Prob(JB) should be in line with the Omnibus test. It is also performed for the distribution analysis of the regression errors. It is supposed to agree with the results of the Omnibus test.

· The condition number (Cond. No.) measures the sensitivity of a function's output to its input. When two predictor variables are highly correlated, which is called multicollinearity, the coefficients or factors of those predictor variables can fluctuate erratically for small changes in the data or the model. Ideally, similar models should be similar. Multicollinearity can produce inaccurate results. We hope to see a relatively small number.

· Skew is a measure of data symmetry. We want to see something close to zero, indicating the residual distribution is normal. This value also drives the Omnibus. This result has a small value, and therefore good, skew.

- Kurtosis is a measure of "peakiness", or curvature of the data. Higher peaks lead to greater Kurtosis. Greater Kurtosis can be interpreted as a tighter clustering of residuals around zero, implying a better model with few outliers.

For the following 3 measures, they will be based on a 40% test sample and 60% train sample to measure the accuracy of the prediction.

- Mean Absolute Error (MAE): measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight, MAE does not indicate underperformance or overperformance of the model. Each residual contributes proportionally to the total amount of error, meaning that larger errors will contribute linearly to the overall error. A small MAE suggests the model is great at prediction, while a large MAE suggests that the model may have trouble in certain areas. An MAE of 0 means that the model is a perfect predictor of the outputs (but this will almost never happen). Our average error, MAE is 0.22, which means that our model is working good towards predicting but it can get better
- The mean absolute percentage error (MAPE) is a statistical measure of how accurate a forecast system is. It measures this accuracy as a percentage and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values. MAPE has managerial appeal and is a measure commonly used in forecasting. The smaller the MAPE the better the forecast. Our MAPE= 3.52, which is not optimal.
- Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.
  Our error rate, RMSE = 0.27, usually any error rate bigger than 0.5 is considered as a bad prediction model, which means that our model is somewhat a good fit.

Overall the model is significant, it has a good R2 value, p-value, and Prob(F-Statistic) values are close to zero, the model also has a relatively small standard error. In the next figure, we are presenting the relationship between the true data we collected, and the prediction data resulted from the model, we can notice a strong correlation equals r = 0.834

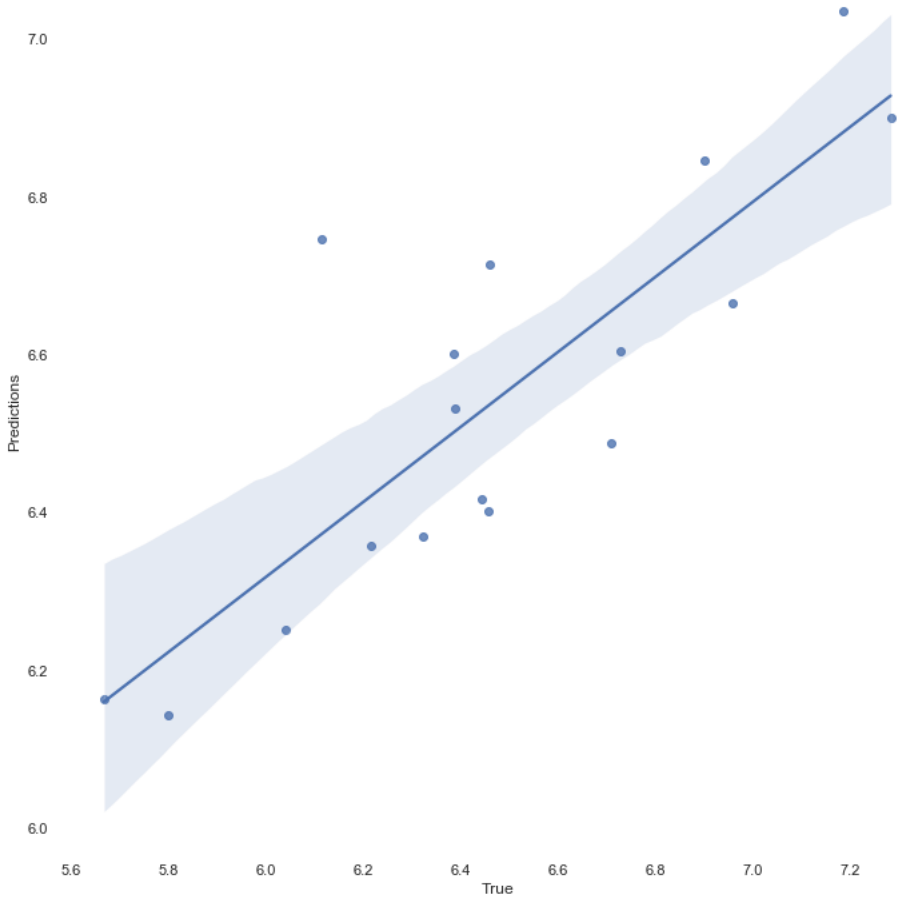


Figure 3. 19 *A scatter plot between the true values (test) and the predictions values (trained), for the simple linear regression*

In the next part of the work we will try to fit a multiple regression model to get better result and to try and get a better prediction model.

3.2.3 Multiple Linear Regression

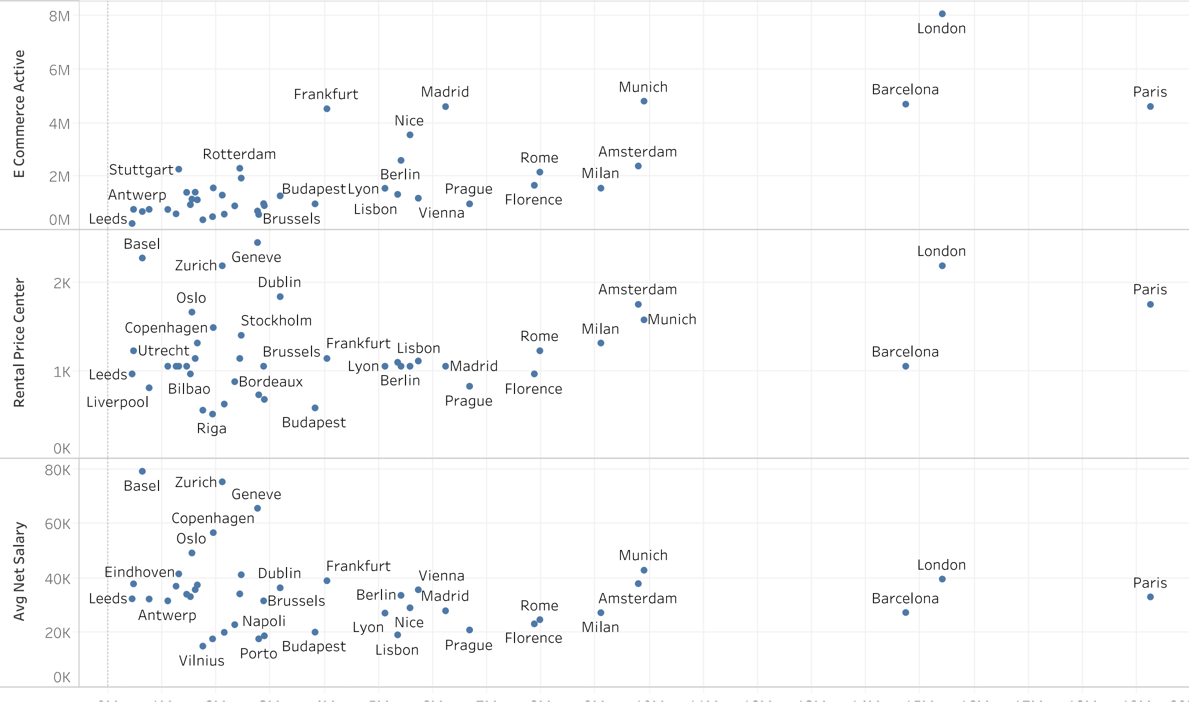
Cities attractiveness to people cannot depend only on one variable, thus, we need more variables in our regression model to try to reach a better result, people choose to visit a city for all kind of reasons, those reasons cannot be put all in our regression, and so we will be only interested in how our selected variables can perform and predict people’s visiting a certain city. It is important to note that the multiple regression is not about the best fitting line anymore, but about the best fitting model. Therefore, as we have discussed before, Adjusted R-squared and p-value are supposed to help us fight the incentives of adding too many variables in one model. The adjusted R-squared adjusts for the number of variables in the model. Importantly, its value increases only when the new variable improves the model fit but starts to decrease when a new variable doesn't improve the model fit by a sufficient amount.

The first step to start with is adding all the independent variables in the model together against the dependent variable, *'foreign_visitors'*, and then we will start eliminating the variables with p-value > 0.05 one after the other until we reach a point when all the variables are significant. After doing that multiple times, the results below have been achieved.

***Table 3. 10** Multiple linear regression results*

Independent Variables	Coefficient	Significance (p-value)	Standard Error
e_commerce_active	0.7723	0.0000000156	0.107
rental_price_center	1.6675	0.000323	0.391
avg_net_salary	-2.2824	0.000000625	0.367

In this visualization, we tried to represent the relationship between the three independent variables with the dependent variable together, we can notice a similar trend in all of the figures.



The trends of sum of E Commerce Active, sum of Rental Price Center and sum of Avg Net Salary for Foreign Visitors. The marks are labeled by City.

***Figure 3. 20** Three scatter plots between foreign visitors and e-commerce active, Rental price center and avergae net salary*

The adjusted R2 value, in that case, goes from 0.559 to 0.759, a lot higher than the adjusted R2 value for the Simple linear regression, thus, our model is performing better. The new variables together are creating a more fitted model. The reason why the population won't perform better with other variables is that if used with other correlated variables will be a violation of the no multicollinearity OLS assumption which states that when two predictor variables are highly correlated, the coefficients or factors of those predictor variables can produce inaccurate results.

Hence, our dependent variable y= 'foreign_visitors' and our independent variables x1= 'e_commerce_active ', x2=' rental_price_center' and x3=' avg_net_salary' Now, we can use the coefficients to build our prediction equation in our new model would be as following:
Y = 6.9265 + 0.7723* X1 + 1.6675 * X2 – 2.2824*X3
It also can be:
foreign_visitors' = 6.9265 + 0. 7723* e_commerce_active +1. 6675 * rental_price_center- 2.2824*avg_net_salary
Telling us that foreign visitors each year is predicted to increase 0.7723 times when the number of e-commerce active population goes up by one and to increase 1.6675 times when the average rental price goes up by one and to decrease 2.2824 times when the average net salary in a city goes up by one, that is not necessary to happen directly, it might be due to that average net salary is an indication of a certain quality of life and that people get attracted more often to visit or live in a less expensive city and have a medium life quality than living in an expansive city with higher qualities. It would be interesting in the future to test if a city can achieve to have a high life standard with a low and affordable cost of living.

3.2.3.1 Measuring Model Fitness

In the following table we will be testing the same methodologies used in the simple linear regression model to assess and check the fitness of this model.

Table 3. 11 *Assessment Methodologies for multiple linear regression*

Dep. Variable:	foreign_visitors	R-squared:	0.777
Model:	OLS	Adj. R-squared:	0.759
Method:	Least Squares	F-statistic:	44.09
No. Observations:	42	Prob (F-statistic):	0
t1	7.2	Log-Likelihood:	10.781
t2	4.2	AIC:	-13.56
T3	-6.226	BIC:	-6.611
Covariance Type:	nonrobust	Df Residuals:	38
Omnibus:	0.207	Durbin-Watson:	2.331

Prob(Omnibus):	0.902	Jarque-Bera (JB):	0.001
Skew:	0.013	Prob(JB):	0.999
Kurtosis:	3.002	Cond. No.	308
MAE	0.17	RMSE	0.21
MAPE	2.79		

To interpret the results as before, we will start by exploring the most important indicators and comparing them with the linear regression model.

· R-squared as we said before R2 measures the strength of the relationship between the model and the dependent variable on a convenient (0 – 1) scale. In our simple linear regression model, we had an R2 equals 0.559, while our new model provides a value of 0.777, we can see a huge improvement in our multiple linear regression model. Therefore, we have a strong positive change. However, let’s put in mind that R2 always rewards for including more variables, thus, it’s important to check the adjusted R2 value for increase or decrease.

· Adj. R-squared, as we mentioned before, provides protection against too many variables. The adjusted R2 value increases only when the new variable improves the model fit. The results here are high as it can get and equals 0.759. Thus, we have a strong fit and a positive change.

· The F-statistic decides whether our added variables to the simple linear regression improves the model or not, in this case, we haven’t used the same variable as in the first model. Prob(F-statistic): This tells the overall significance of the regression. This is to assess the significance level of all the variables together unlike the t-statistic that measures it for individual variables. Usually, when the value is closer to zero, it implies that overall the regressions are strong and meaningful.

· AIC/BIC: A lower AIC implies a better model, and a lower BIC means that a model is considered to be more likely to be the true model. Our AIC and BIC were not that low in our simple linear regression model, however here the values are -13.56 and -6.611 respectively, which means that we have a better model.

· Log-likelihood as we said before cannot be used alone but can be used to compare the fit of different coefficients. Because we want to maximize the log-likelihood, the higher value is better. For instance, In the first model, we had a log-likelihood value of -3.0502 and in our new model we have a log-likelihood value of 10.781, which again indicates a better-fitted model

· Prob(Omnibus) is supposed to be close to 1 in order for it to satisfy the OLS assumption. In this case, Prob(Omnibus) increased to a value higher very close to 1 which is also a positive change.

· Durbin-Watson: Another assumption of OLS is of homoscedasticity. This implies that the variance of errors is constant. A value between 1 to 2 is preferred. Here, our value is outside the range, however, it went down in comparison to before, implying that the regression results are acceptable from the interpretation side of this metric.

· Kurtosis: a measure of "peakiness", or curvature of the data. Higher peaks lead to greater Kurtosis. Greater Kurtosis can be interpreted as a tighter clustering of residuals around zero, implying a better model with few outliers.

· As before we will try to investigate the accuracy of our model with the mean absolute error, mean absolute percentage error, and Root Mean Square Error. The results are MAE= 0.17, MAPE=2.79, and RMSE= 0.21 which means that our model has better accuracy than before and a good predicting capability.

In conclusion, our model achieved better results in all of the most indicators that measure if the model is significant, it has a better R2 value, a better-adjusted R2, A p-value, and Prob(F-Statistic) equals zero, the model also has a small standard error. The next figure is showing the relationship between the test/true data and the predictions, a strong positive correlation is found equals r = 0.914, which corresponds with the previous investigations.

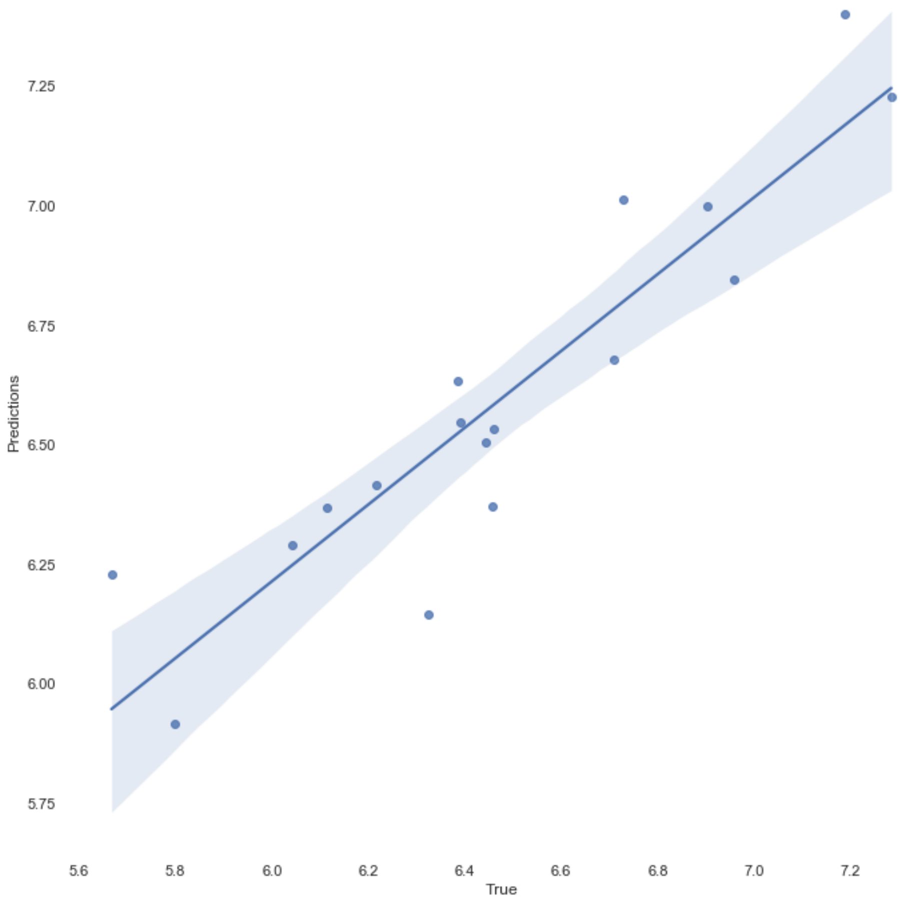


Figure 3. 21 *A scatter plot between the true values (test) and the predictions values (trained), for the multiple linear regression*

Chapter 4 **Natural Language Processing (Text Mining)**

4.1 Introduction

In this chapter, we will focus on Natural Language Processing using a popular Python package called NLTK. Natural Language Processing, or shortly NLP, is a data science term used to refer to the interaction of computers, and natural language humans use. This is not an easy task because human language is ambiguous. As humans, we are good at understanding the context of something said to us and link that to the understanding of the concepts around it. Doing this algorithmically, however, is not easy. This is what the field of NLP tries to improve and develop algorithms and data techniques that make it possible in an effective and fast fashion. Chances are, we have used an NLP application, if we used an online summary of news or books, looked for keywords generated for most popular twitter topics, or used a virtual assistant on the phone. NLP techniques are applied in speech recognition engines, like Siri, Google Now, or Alexa. These engines are designed to learn what and how a human talks over time, and constantly improve their accuracy. Similarly, automatic translators, like Google Translate or Facebook automatic translation of posts use an LP, using some recent very effective neural network-based techniques that take not only words and phrases into account, but also the context, by looking at the word surrounding the text they are translating. Chatbots that can answer questions via Facebook Messenger are another example of NLP. They use NLP engines to process the questions, often simply categorize them, and match them to existing answers to the question. NLTK is the most popular Python package for NLP. It is an open-source library that provides modules for importing, cleaning, preprocessing text data, in human language, and then apply computational linguistics algorithms, or machine learning algorithms, like sentiment analysis, to these datasets. NL techniques depend on large amounts of text or other linguistics data.

The first step in NLP is generally to split the text into words. This process might appear simple, but it can be monotonous to handle all corner cases. Corner cases include inconsistent use of punctuation, or contractions, or shortened versions of words. They can also be hyphenated words that include characters like the New York-based example. To understand how we can build machine learning features from words, we need to apply something called Bag-of-words. The Bag-of-words model is a very simple representation of a body of text as a loose set of words. It flattens any text into an unordered collection of words. Although it disregards the sentence structure associated with the words, this simple technique is very useful to identify a topic or sentiment in text, like if a product review has a negative or positive sentiment, or what a body of text talks about. It's also best practice to filter out stopwords and maybe even the punctuations from the bag-of-words before further analysis. Some of the stopwords can be as following: ['i', 'me', 'my', 'myself', 'we', 'the', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a'] which occur a lot but don't have a big significance in identifying the context of the text being processed.

We will have a lot of words to analyze, so how can we find out the frequency of each word that is how many times each word appears in this corpus? We will use the counter object from the collection package in Python for this purpose. We can provide it with a list of words. And it returns an object wire, which we can find out the reputation of each word. Once we have the frequency, in a counter, for each word, we will see how we can plot the distribution of words, using matplotlib. The visualization is particularly useful if we are comparing two or more data sets. Flatter distribution indicates a large vocabulary, while a peak distribution indicates a restricted vocabulary. Often due to a focused topic or a specialized language. We will also create histograms of the words for visualization of the frequencies.

After we are done with the previous steps, we can begin to test and implement the sentiment analysis but first, what is Sentiment Analysis? Sentiment analysis or opinion mining is the automated process of identifying and extracting the subjective information that underlies a text. This can be either an opinion, a judgment, or a feeling about a particular topic or subject. The most common type of sentiment analysis is called 'polarity detection' and consists of classifying a statement as 'positive', 'negative', or 'neutral'. For example, let's take this sentence: "I don't like traveling to berlin: it's really huge and I get lost a lot". A sentiment analysis model would automatically tag this as Negative.

Sentiment analysis is a sub-field of Natural Language Processing (NLP), and it has been getting a lot of attention in recent years due to its many exciting applications in a variety of fields, ranging from business to political studies. Thanks to sentiment analysis, entities can understand the reputation of their brand, and cities can have the possibility to know what people are saying about them. By analyzing social media posts, product reviews, customer feedback, or responses, they can be aware of how their people feel about their service. They can also track specific topics and get relevant insights on how people are talking about those topics. Sentiment analysis is particularly useful for social media monitoring because it goes beyond metrics that focus on the number of likes or retweets and provides a qualitative point of view. We will be performing Sentiment Analysis on data from Reddit using machine learning which can help us understand how people are talking about the sample cities. Reddit is an American social news aggregation, with more than 330 million active users, and more than 130,000 active communities. Reddit can allow businesses to reach a broad audience and connect with customers. This is one of the reasons why social listening, monitoring conversation, and feedback in social media has become a crucial process in social media marketing. Monitoring social media discussion, either from Reddit, Facebook or Twitter, allows entities to understand their audience, keep on top of what's being said about their business, brand, and their competitors, and discover new trends. In this part of the work, we'll take a closer look at sentiment analysis and how we can analyze and understand emotions on Reddit.

Nearly 80% of the world's digital data is unstructured, and data obtained from social media sources is no exception to that. Since the information is not organized in any predefined way, it's difficult to sort and analyze. The micro-blogging content coming from Reddit, Twitter, or Facebook poses serious challenges, not only because of the amount of data involved but also because of the kind of language used in them to express sentiments, i.e., short forms, memes, and emoticons. Sifting through huge volumes of this text data is difficult as well as time-consuming. Though it may seem easy on paper, Sentiment Analysis is actually a tricky subject. There are various reasons for that: Understanding emotions through the text is not always easy. Sometimes even humans can get misled, so expecting a 100% accuracy from a computer is like asking for the Moon. A text may contain multiple sentiments all at once. For instance, “The intent behind the movie was great, but it could have been better”. The above sentence consists of two polarities, i.e., Positive as well as Negative. So how do we conclude whether the review was Positive or Negative? Computers aren't too comfortable in comprehending Figurative Speech. Figurative language uses words in a way that deviates from their conventionally accepted definitions in order to convey a more complicated meaning or heightened effect. Let us understand it better with an example. “The best I can say about Barcelona is that it was interesting.” Here, the word ’interesting’ does not necessarily convey positive sentiment and can be confusing for algorithms. Heavy use of emoticons and slangs with sentiment values in social media texts like that of Reddit, Twitter, or Facebook also makes text analysis difficult. For example a “ :)” denotes a smiley and generally refers to positive sentiment while “:(” denotes a negative sentiment on the other hand. Also, acronyms like “LOL“, ”OMG” and commonly used slangs like “Nah”, “meh”, ”giggly” etc. are also strong indicators of some sort of sentiment in a sentence **(Pandey, 2018)**. These are a few of the problems encountered not only with sentiment analysis but with NLP as a whole. In fact, these are some of the Open-ended problems of the Natural Language Processing field. Fortunately, thanks to the developments in Machine Learning and NLP, it is now possible to create models that learn from examples and can be used to process and organize text data. Reddit sentiment analysis systems allow us to sort large sets of text and detect the polarity of each statement automatically. And the best part, it's fast and simple, saving valuable hours and allowing us to focus on tasks where we can make a bigger impact. These are some of the main advantages of Reddit sentiment analysis:

· Scalability: let’s say we need to analyze hundreds of comments discussing a specific topic. While we could do that manually, it would take hours and hours of manual processing and would end up being inconsistent and impossible to scale. By performing Reddit sentiment analysis, we can automate this task and obtain cost-effective results in a very short time.

· Consistent Criteria: analyzing sentiment in a text is a subjective task. When done manually, the same comment may be perceived differently by two members of the same team, and the results will probably be biased. By training a machine learning model to perform sentiment analysis on Reddit, we can set the parameters to analyze all the data and obtain more consistent and accurate results.

To give a real-life example of Sentiment analysis, the 2016 US Presidential Elections is very important, and for many reasons. Apart from the political aspect, the major use of analytics during the entire canvassing period garnered a lot of attention. During the elections, millions of Twitter data points, belonging to both Clinton and Trump, were analyzed and classified with a sentiment of either positive, neutral, or negative. Some of the interesting outcomes that emerged from the analysis were:

The tweets that mentioned ‘@realDonaldTrump’ were greater than those mentioning ‘@HillaryClinton’, indicating the majority were tweeting about Trump. For both candidates, negative tweets outnumbered the positive ones. The Positive to Negative Tweet ratio was better for Trump than for Clinton. This is the power that sentiment analysis brings to the table and it was quite evident in the U.S elections.

To use Sentiment Analysis with Reddit Data, we can identify four main steps in this process: Data gathering, Data preparation, The creation of the sentiment analysis model, and Visualization of the results.

1. Data Gathering: If we want to perform Reddit sentiment analysis, the first step is to gather the data. This is the data that we will use for: training the machine learning model and running the actual sentiment analysis on the data.
2. Data Preparing: Preprocessing and Cleaning text, once we have captured the comments, we need for the sentiment analysis, it’s time to prepare the data. As we mentioned earlier, social media data is unstructured. That means it’s raw, noisy, and needs to be cleaned before we can start working on the sentiment analysis model. This is an important step because the quality of the data will lead to more reliable results. Preprocessing a Reddit dataset involves a series of tasks

like removing all types of irrelevant information like emojis, special characters, and extra blank spaces. It can also involve making format improvements, delete duplicate comments, or tweets that are shorter than three characters.

3. Creating a Reddit Sentiment Analysis Model: At this point, we probably have a few ideas about types of content we would like to analyze, so it may be the right time to use machine learning for text analysis.

4. Data Visualization of the Results: Data visualization tools are a powerful resource to take the sentiment analysis results to the next level. Raw results, in other words, long lists of text, will never get as much attention as a beautifully designed chart or figure that show insight in a visual and engaging way.

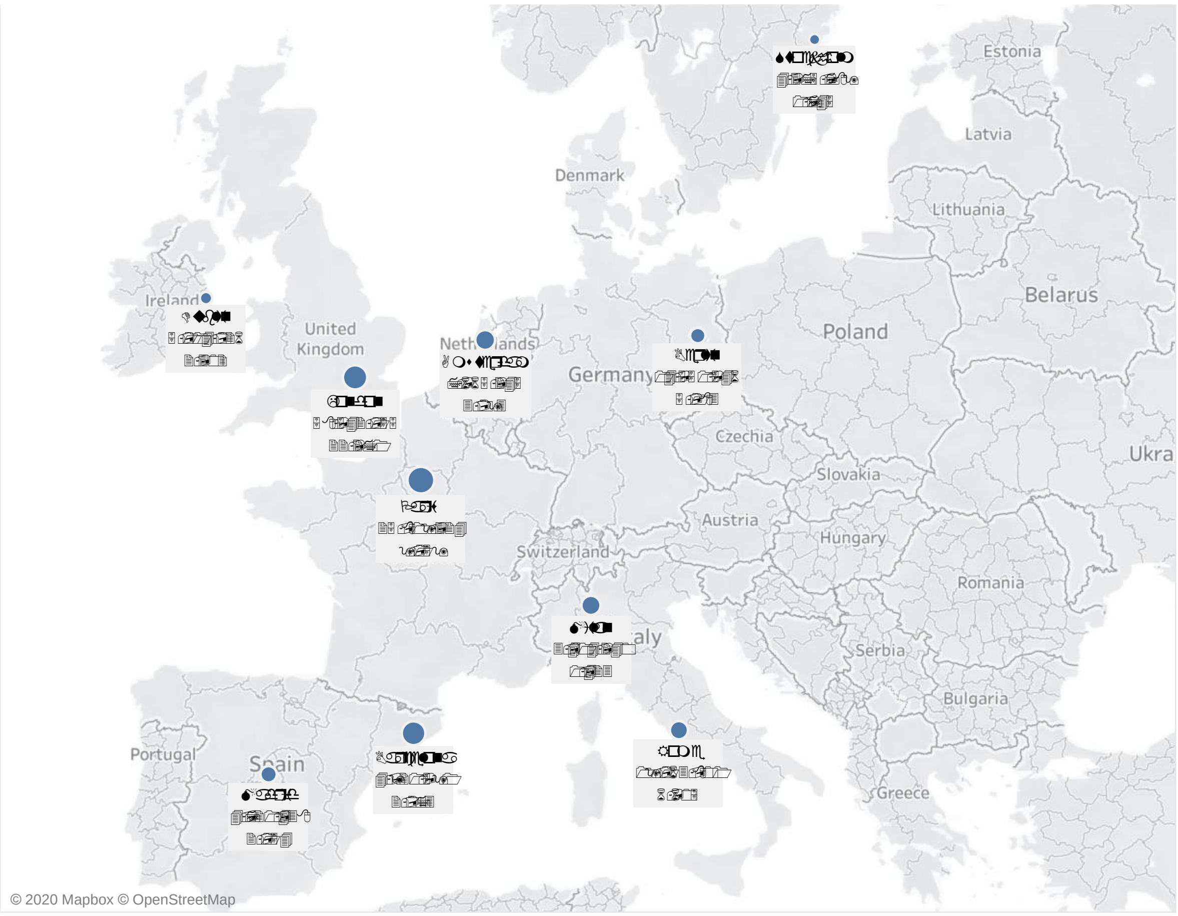
For the Sentiment Analysis, VADER (Valence Aware Dictionary and sEntiment Reasoner) is used **(Hutto & Gilbert, 2014)**. VADER uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative. VADER has been found to be quite successful when dealing with social media texts, NY Times editorials, movie reviews, and product reviews. This is because VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is. It is fully open-sourced under the MIT License. VADER has a lot of advantages over traditional methods of Sentiment Analysis, including It works exceedingly well on social media type text, yet readily generalizes to multiple domains. It doesn’t require any training data but is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon. It is fast enough to be used online with streaming data, and it does not severely suffer from a speed-performance tradeoff.

After Vader sentiment analysis is applied and two different text data frames are obtained, word clustering is applied. Word Clustering is a technique for partitioning sets of words into subsets of semantically similar words and is increasingly becoming a major technique used in a number of NLP tasks ranging from word sense or structural disambiguation to information retrieval and filtering. It automatically analyzes text data to determine cluster words for a set of text. This is known as ‘unsupervised’ machine learning because it doesn’t require a predefined list of tags or training data that’s been previously classified by humans. Since word clustering doesn’t require training, it’s a quick and easy way to start analyzing the data. However, we can’t guarantee we will receive accurate results, which is why we are including this method after applying the sentiment analysis. By the end of the analysis, we should have results that are demonstrating the clusters of words in both sentiments, positive and negative.

In the analysis, text mining is used on 10 cities chosen from the dataset of 42 cities, we search on Reddit for comments that mentioned those cities in the last 2 years and we investigate what is being discussed regarding chosen keywords, the chosen keywords are linked with the cost of living in a certain city, the quality of life in that city, investing opportunities, and studying opportunities in that city. To be specific the keywords are as follows for the city of London (Invest in London, Study in London, Quality of Life in London, Cost of Living in London).

4.2 Reddit Sentiment Analysis

The following figure is showing the number of comments and the number of words collected for each city and filtered with the number of foreign visitors, the bigger the circle the higher the foreign visitors.



[illegible]

Figure 4. 1 A geographical representation for the number of words and comments collected for each city

In the next part of the work, we explore each city individually.

4.2.1 Amsterdam

[illegible]

Figure 4. 2 *Geographical representation for the number of the words and comments collected for the city of Amsterdam*

For Amsterdam, we managed to collect 3554 comment, with a total of 10,527,539 words. Out of those 3554 comments, 3093 are unique values, and 461 duplicate which was removed from the analysis, keeping 7,665,245 words to process and analyze.

4.2.1.1 Text Processing and Visualization

Considering that we cannot go directly from raw text to fitting a machine learning model. In this step we will perform the following procedures on the collected text:

- Remove all punctuation: For instance; ('!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~')
- Remove all stop words: Stop words are those words that do not contribute to the deeper meaning of the phrase.
- Normalize words: It is common to convert all words to one case.
- Returns a list of the cleaned text

After applying the previous methods, we now have 638,586 clean, simple and easy to interpret text. In the following table, the difference between the data before processing it and the data after is shown, we are also exploring the frequency of words per each comment. The processed text has a total average of around 200 comments, while not processed texts has around 2500 words as average. The longest processed comment has 1116 words.

Table 4. 1 *Amsterdam text descriptive analysis*

Before Processing	After Processing
count 3093	count 3093
mean 2478.255739	mean 206.461688
std 2514.500647	std 202.778318
min 20	min 1
25% 689	25% 63
50% 1575	50% 130
75% 3313	75% 277
max 10572	max 1116

The previous data is represented in the following figures, notice that most of the words per comment before processing were between (0-1500) and after they went down to the range of (0-100).

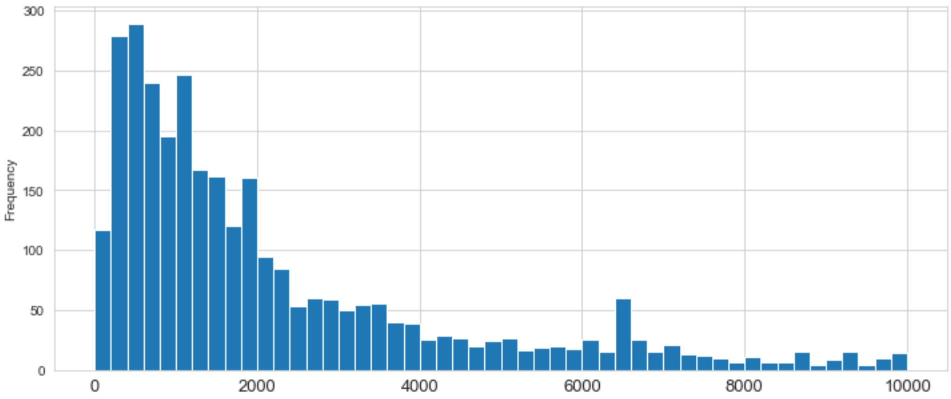


Figure 4. 3 *Amsterdam words per comments before processing*

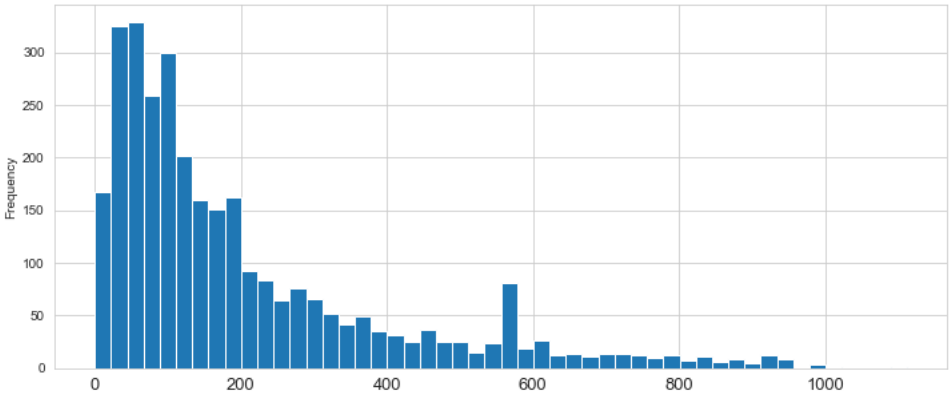


Figure 4. 4 *Amsterdam words per comments after processing*

In the following figure we can visualize the top 35 most frequent words acquired from the text.

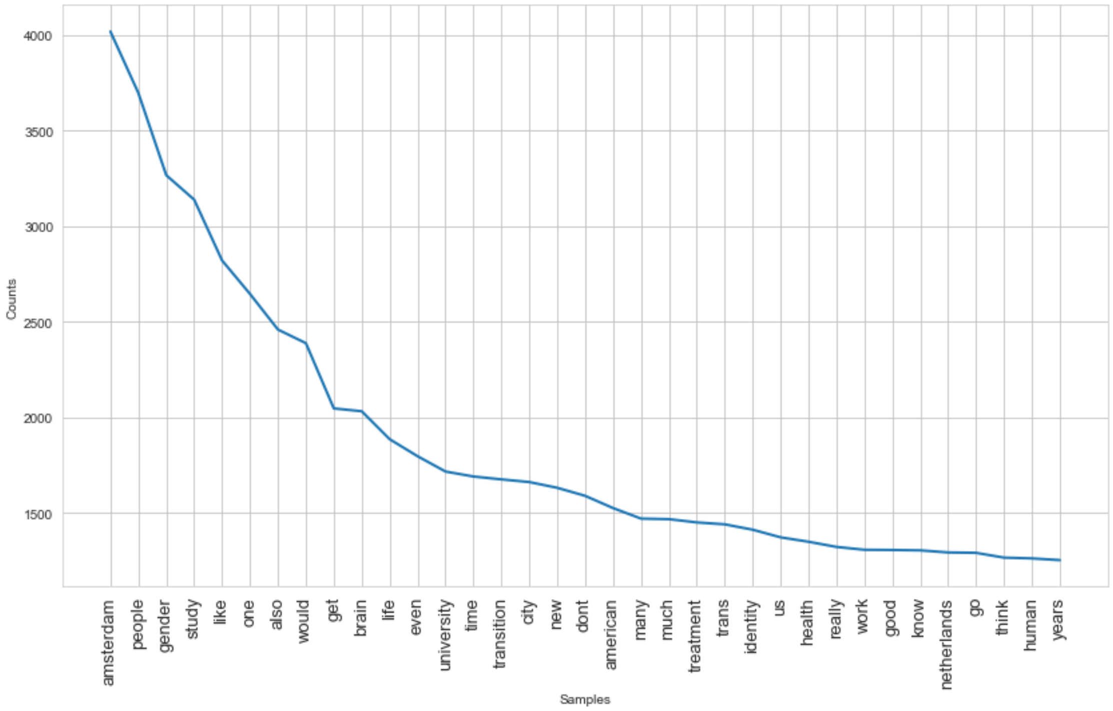


Figure 4. 5 *Amsterdam 35 most frequent words for all sentiments*

We can sort the word counts and plot their values on Logarithmic axes to check the shape of the distribution. Log scales show relative values instead of absolute ones. This visualization in the next figure shows a flatter distribution which indicates a large vocabulary while a peaked distribution shows a restricted vocabulary often due to a focused topic or specialized language.

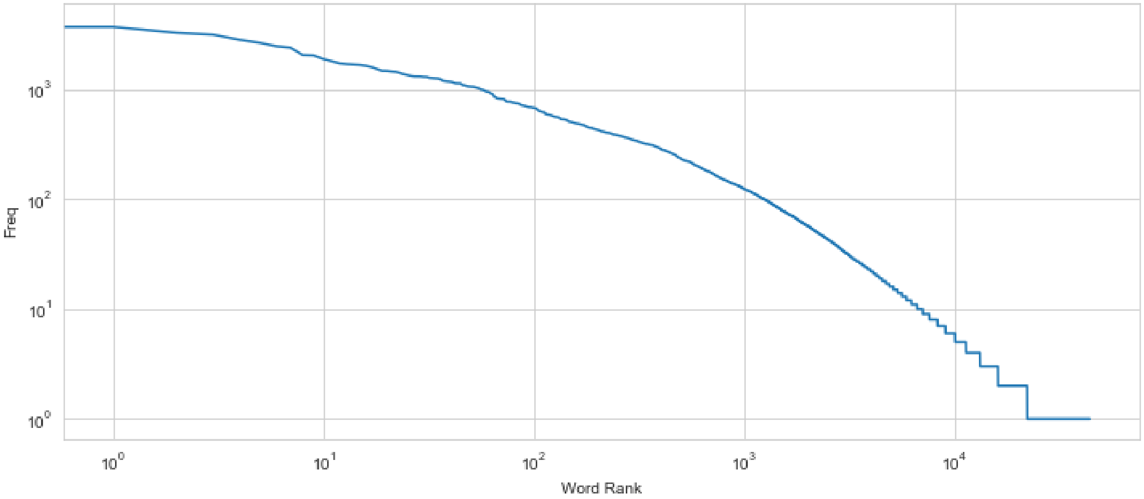


Figure 4. 6 *Amsterdam word counts distribution on a logarithmic scale*

While on the following representation, a word cloud is done on the most 150 frequent words, in a word cloud the size of each word indicates its frequency or importance. Thus, the more often a specific word appears in the text, the bigger and bolder it appears in the word cloud. We can notice that the word

‘Amsterdam’ is on the top of the list which makes sense counting on the keywords we used, then we can notice that people usually discuss universities, studying, gender and Dutch language topics.

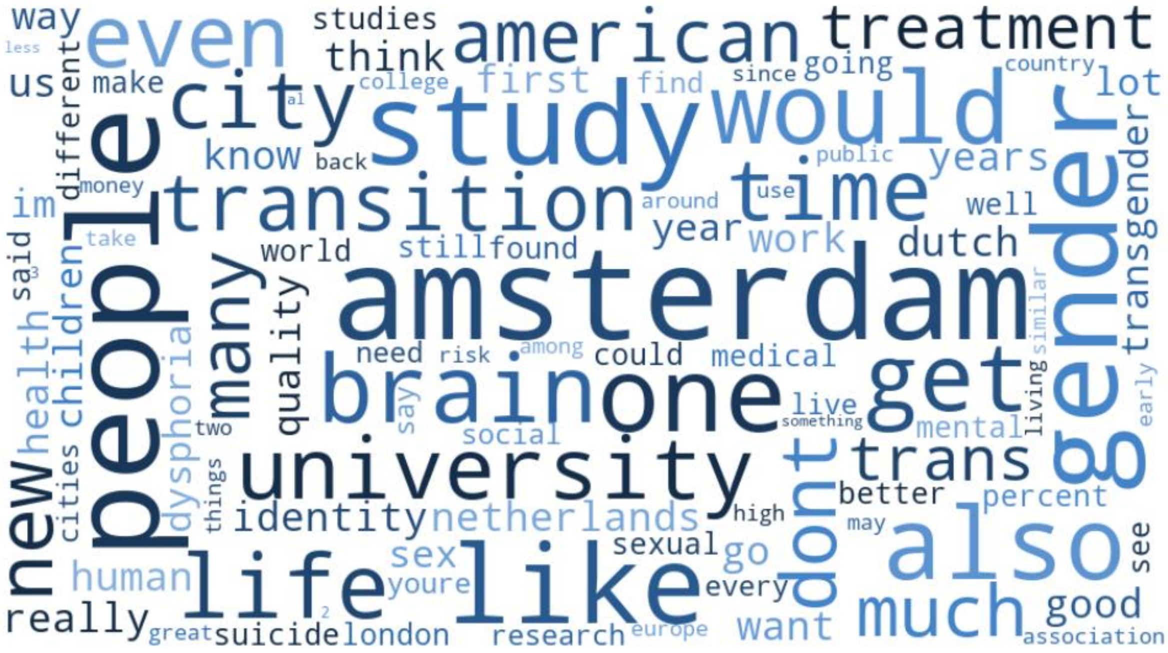


Figure 4. 7 *Amsterdam word cloud on all sentiments*

4.2.1.2 Sentiment Analysis & Words Clustering

In the last part of the work we carried out a text analysis to obtain some insights and to understand the text since we are dealing with unsupervised learning problems, however, we still don’t have a clear idea what the text can indicate, to be specific we don’t have a context. Thus, it’s finally time to use Sentiment Analysis on the text to label the text and classify them into 3 categories (Positive, Neutral, Negative) using VADER (Valence Aware Dictionary and sEntiment Reasoner). Here the VADER algorithms try to discover natural structure in data. The algorithm looks for similar patterns and structures in the data points and groups them into clusters. The classification of the data is done based on the clusters formed.

VADER produces four sentiment metrics from these word ratings, which you can see below. The first three, positive, neutral and negative, represent the proportion of the text that falls into those categories. As you can see, a sentence can be rated as 45% positive, 55% neutral and 0% negative. The final metric, the compound score, it is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1 (most extreme positive). In this case, the example sentence has a rating of 0.69, which is strongly positive.

Table 4. 2 *VADER sentiment example*

Sentiment metric	Value
Positive	0.45
Neutral	0.55
Negative	0.00
Compound	0.69

It’s important to note that Vader analysis primarily depends on the things that we removed from the text before, like Punctuation: for instance, the use of an exclamation mark(!), increases the magnitude of the intensity without modifying the semantic orientation. For example, “The food here is good!” is more intense than “The food here is good.” and an increase in the number of (!), increases the magnitude accordingly. Conjunctions: Use of conjunctions like “but” signals a shift in sentiment polarity, with the sentiment of the text following the conjunction being dominant. “The food here is great, but the service is horrible” has mixed sentiment, with the latter half dictating the overall rating. Therefore, we will be applying the sentiment analysis on a semi-processed version of the text, we will keep all the punctuation, smiley faces, exclamation marks and we will only remove things that do not contribute to the analysis and only cause noise such as _((<br/>)/(<a).*(>). *(</a>)//(&amp/(&gt/ http\S+"/*)_, after removing the previous noise we will be working with 6,870,776 words in this step. However, later on further analysis we will be using the fully cleaned version of the text to look for insights, only this time we will have the text labeled positive, neutral and negative for further analysis.

According to the documentation of Vader sentiment analysis for researches and academic work it’s best to assume that a compound result of anything >=0.05 will be considered as positive sentiment, and anything that falls below <=-0.05 will be considered as negative sentiment, while anything falls between (-0.05<score<0.05) will be considered as neutral sentiment.

In the next table we have the results of the analysis. 67.3% of the comments collected are positive, 29.1% are negative and 3.6% are neutral as represented in the figure.

Table 4. 3 *Amsterdam sentiment results*

Negative (-1)	Positive (1)	Neutral (0)	Total
899	2084	110	3093

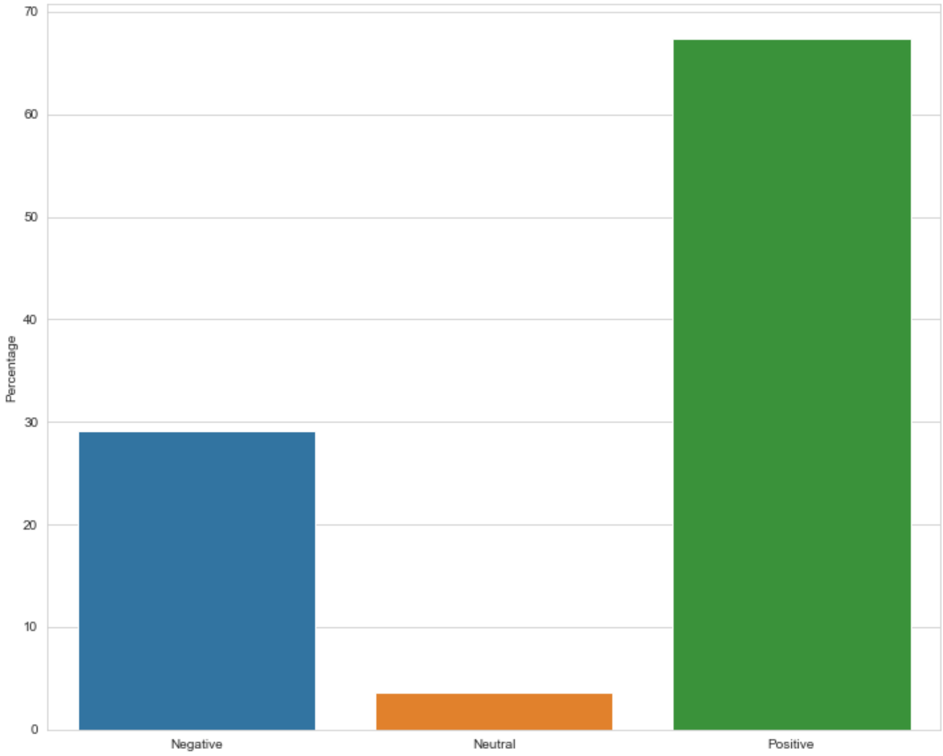


Figure 4. 8 *Amsterdam Sentiment results in percentage(%)*

In the last part, we had the chance to explore a figureical representation for word frequency without labeling the data to positive and negative. In the next visualization we will be exploring the same figures, however, this time it will be with showing the word frequency for both positive and negative sentiments.

As before, In the following figures we can visualize the top 35 positive and most frequent words and the most 150 frequent words. We can notice a change in the words comparing to the one from before. The sentiment analysis tells us that people talked in a positive way about topics like *study, life, university, city, work and the Dutch language* when talked about Amsterdam.

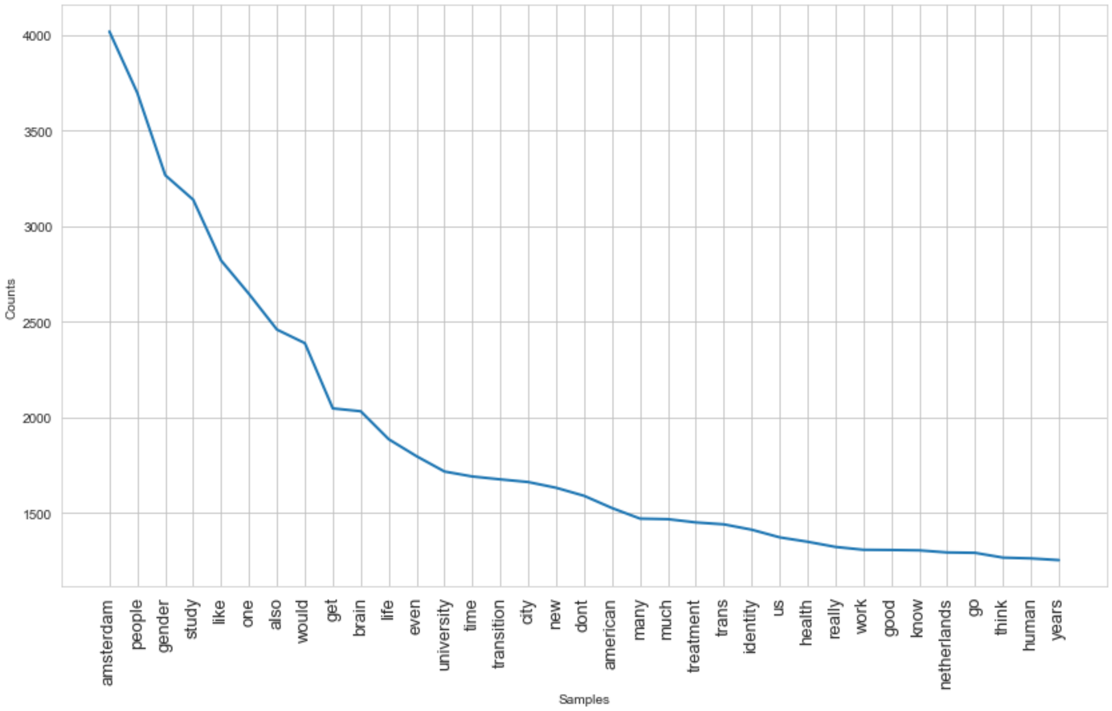


Figure 4. 9 *Amsterdam 35 most frequent words for positive sentiment*

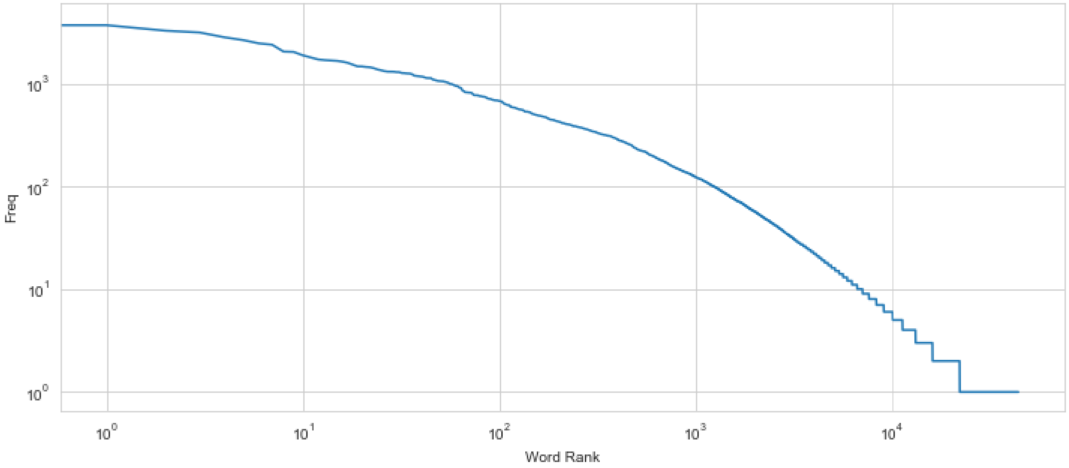


Figure 4. 10 *Amsterdam word counts distribution for positive sentiment on a logarithmic scale*

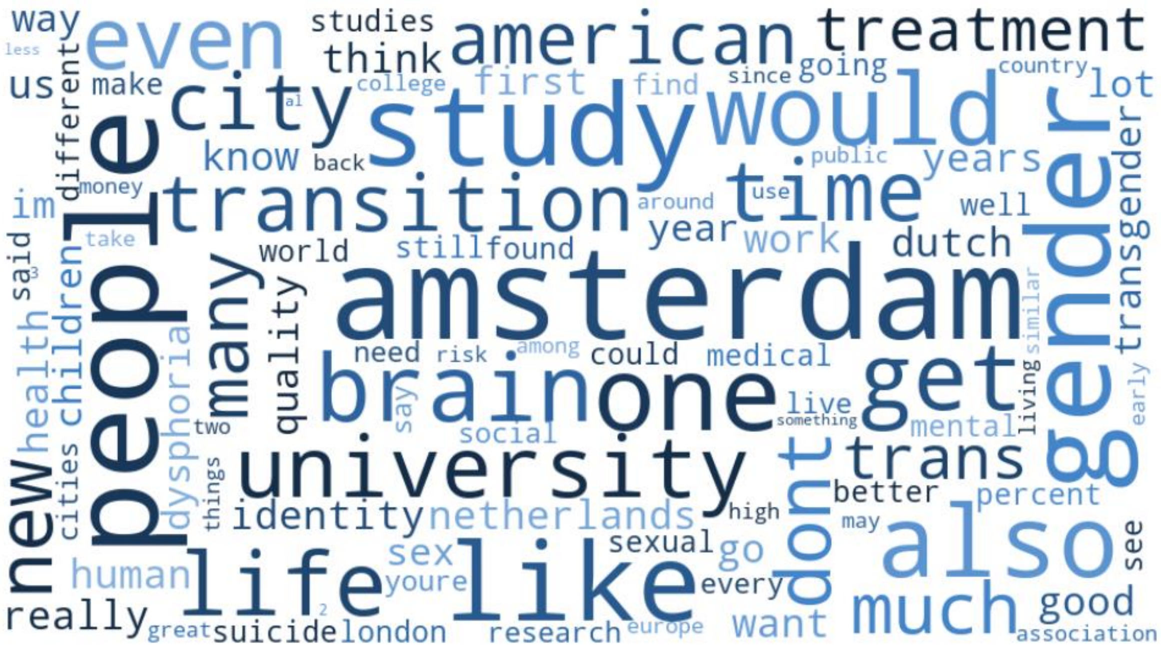


Figure 4. 11 *Amsterdam word cloud on positive sentiments*

For the next figure we are presenting the same, however, it's for the negative words. The sentiment analysis tells us that people talked in a negative way about topics like *transgender, mental health and medical treatments* when talking about Amsterdam. This might be relating to the fact that in 2013, the Dutch Parliament approved a bill that would allow transgender people to legally change their gender on their birth certificates and other official documents without undergoing sterilization and sex reassignment surgery, and the law took effect in 2014 **(Wikipedia, n.d.).**

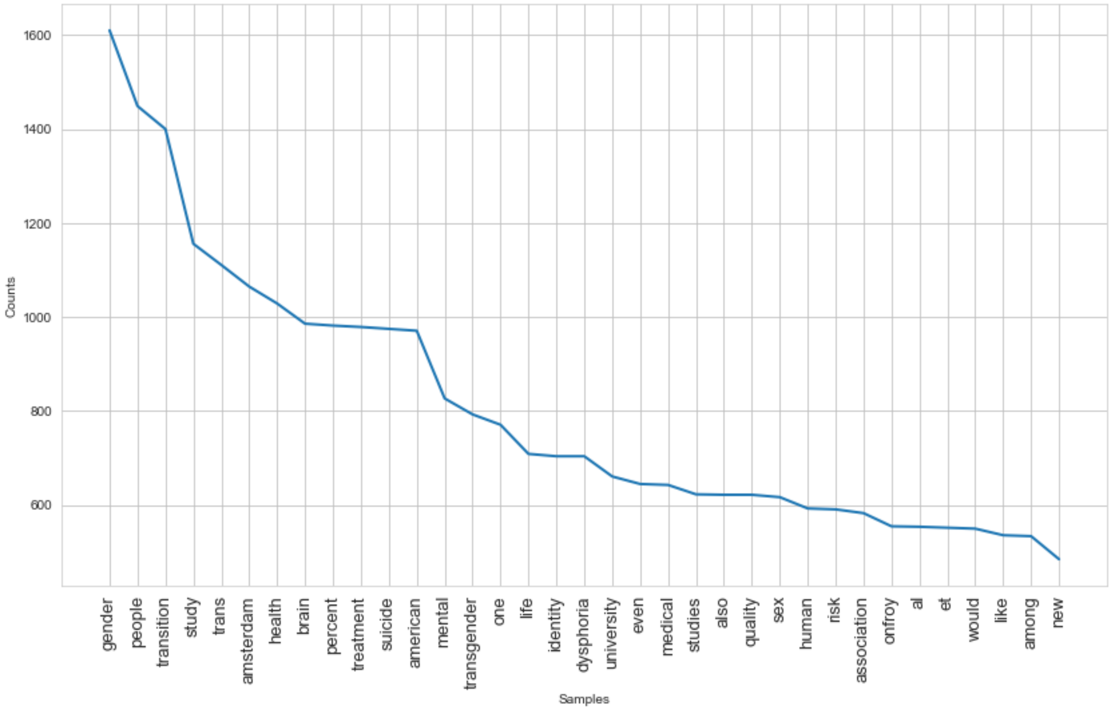


Figure 4. 12 *Amsterdam 35 most frequent words for negative sentiment*

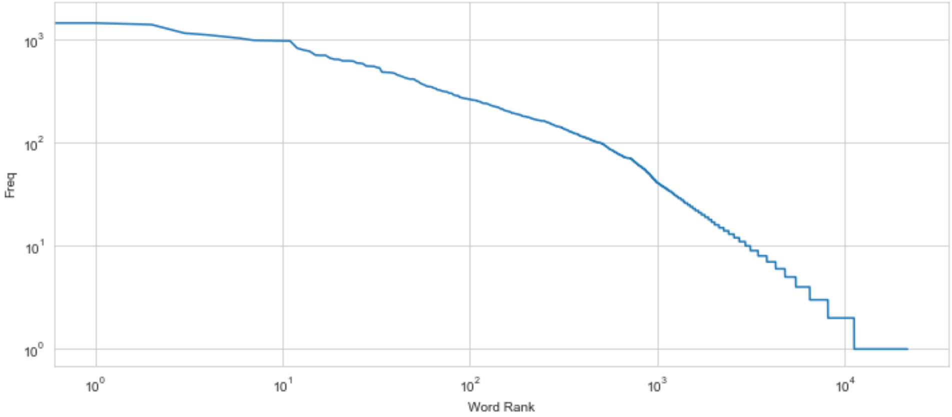


Figure 4. 13 *Amsterdam word counts distribution for negative sentiment on a logarithmic scale*

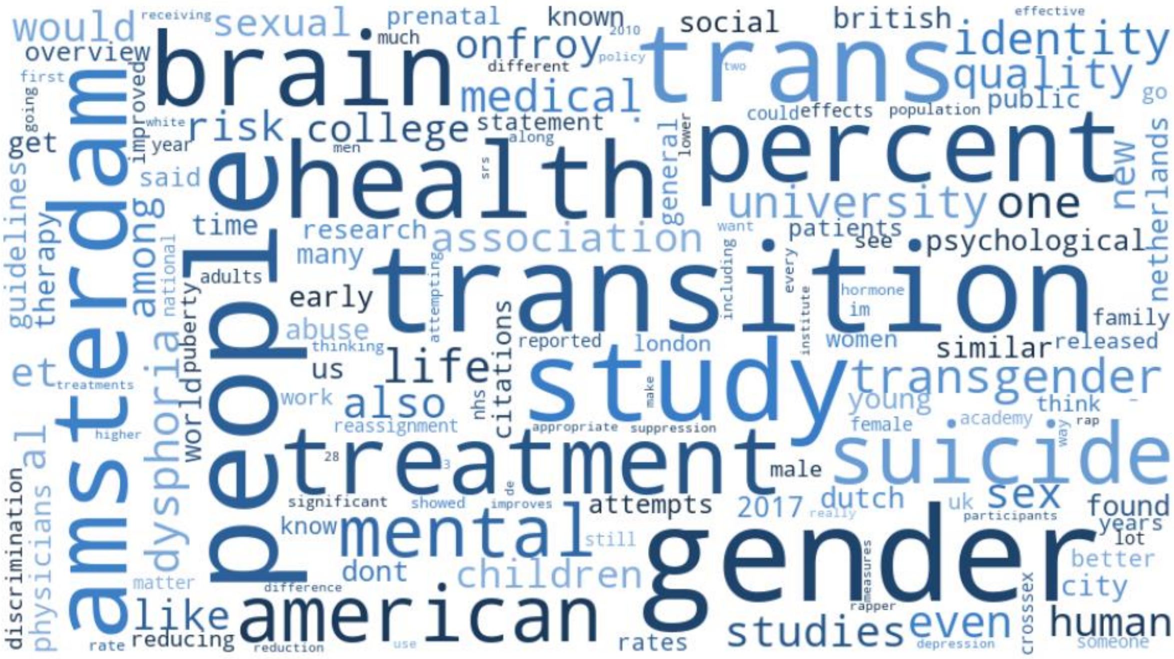


Figure 4. 14 *Amsterdam word cloud on negative sentiments*

Moreover, we will apply word clustering technique to the text. First, we will be investigating the group of words that is discussing a specific topic in the positive text, and then we will be doing the same for the negative text. This step is very important to get a closer understanding on what people are discussing online through a bag of words and to discover meaningful implicit subjects across all text. After we will be exploring some of the results obtained by showing examples of the sentiments.

In the next section we come understand that people discussed the *gender identity* topic in both a positive and a negative way, we also know that people discussed positively the *quality and city life* in Amsterdam, Utrecht city was mentioned a lot of times when talking about Amsterdam. However, people are concerned with topics include *prostitution, drugs and mental health*.

- Positive Clusters & Discussions

```
[[' gender', ' children', ' dysphoria', ' transgender', ' puberty', ' boys', ' girls', ' age', ' identity', ' brain'], ['
amsterdam', ' study', ' like', ' people', ' time', ' invest', ' im', ' really', ' years', ' think'], [' life', ' people', ' city', '
like', ' quality', ' living', ' amsterdam', ' cities', ' good', ' dont'], [' dutch', ' university', ' study', ' student', '
students', ' amsterdam', ' english', ' netherlands', ' utrecht', ' wiki'], [' brain', ' gender', ' identity', ' human', '
sexual', ' prenatal', ' overview', ' treatment', ' american', ' transition']]
```

- ['I have been living 6 years in NL (Amsterdam, to be precise) and I have been living now for 5 months in Germany (Berlin). I can only answer for the personal/life part. \\- Language: I don\'t know why everyone says that "Berlin is an international city, you can get by with English". My guess is that they probably have never been to the Netherlands. You can have a social life without speaking German, that is true (Berlin is a capital, plenty of expats). However, forget the level of the Dutch English skills among the general population and government institutions: I took a super-intensive German course to arrive at a reasonable A2 level in 3 months, and it was life-saving. Go to the mechanics because the car breaks, talk to the phone company because Internet had problems, go to the town-hall for papers, talk to the agency to rent the apartment, buy some pieces of furniture and get them delivered. Not a single English-word was spoken. All activities that I did also in the NL and I have never asked myself "will that person speak English", because it was basically a given fact. \\- cost of living: Berlin is cheaper than Amsterdam food-wise and better if you compare the overall quality of the ingredients you can get (you know, vegetables with some taste, instead of the tasteless tomatoes you get at AH in NL :P). It is just a bit cheaper rent-wise. Of course the amount you pay for rent depends on a lot of things, but for similar situations (e.g.: center of the city) expect the prices to be similar. Regarding entertainment, again, a bit cheaper but not that much. \\- I cannot really tell you the difference between Germans and the Dutch on a personal level. I work in research so most of my interactions have always been inside the larger expat community, so English was always the main language spoken plus it\'s all people that are used to travel and meet new people all the time, so social interactions come easy. I got the impression that the stick-to-the-rules part is true but on the other hand it\'s not like the Dutch are these crazy people setting things on fire on a daily basis :D (disclaimer: I am Italian, so I definitely see the difference Italy/Germany or Italy/NL when it comes to general behavior. I don\'t see that much of a difference between Germany and NL). In summary: I think that the main differences you can find on the daily life are mainly two: 1) the level of native-language required to get by is way higher in Germany than NL 2) if you like small, well organized and clean cities, the NL is better. If you like big cities, where you don\'t mind that it takes 1 hour to go anywhere using public transport but on the other hand you have plenty of things to do, than Berlin is better (and probably Frankfurt is similar, but I cannot really speka for that). Hope it helps!']

- ['It's a very livable city if you stay out of the centre, tourist-brimmed district. The clean air and the infrastructure are miles ahead of what Malta will be in the next 10 years, even if the transport ministry in Malta gets a knock on the head one day and decides to at least finally start building segregated public transport and bike lanes, which is very doable considering that narrow roads also exist here in Amsterdam. There still are some shortcomings and little inconveniences in my opinion, but in general the quality of life for what you pay expenses compared to an expat going to Malta is much, much better.']
- [Amsterdam is highly underrated. For expats you you can look at some really cool places for €1.3-1.5k /mo. Before bills. Expensive in comparison to a lot of places, but Adam is also a HFT/propshop hub where you can make some serious money. Tbh, crank your leetcode and study, get a killer offer and rent will be the last thing on your mind. Best of luck :)]

- Negative Clusters & Discussions

[[' parliamentarianism', ' muslim', ' europe', ' obsolete', ' history', ' lefts', ' class', ' trade', ' jews', ' party'], [' people', ' study', ' amsterdam', ' like', ' dont', ' im', ' prostitution', ' think', ' work', ' time'], [' gender', ' brain', ' identity', ' human', ' transgender', ' sexual', ' overview', ' prenatal', ' university', ' male'], [' onfroy', ' 2017', ' released', ' known', ' rap', ' rapper', ' bail', ' lastfm', ' album', ' robbery'], [' transition', ' percent', ' trans', ' suicide', ' health', ' treatment', ' mental', ' gender', ' dysphoria', ' brain']]

- These thousands of professionals are also working under an intensely politicised atmosphere where any deviation from transgender orthodoxy results in accusations of bigotry due to the pervasive influence of transgender activists. This is a summary of one of the studie. In this Dutch study they identified 127 children who were referred to the Gender Identity clinic in Amsterdam when they were under the age of 12. They then looked to see if these children were still gender dysphoric by the time they reached adolescence at age 15. 47 (37%) of these children had persisted. However 80 (64%) of children had either desisted (52) or were no longer traceable(28)
- [Earlier investigations into homosexual violence against homosexuals in Amsterdam showed that the suspects were as indigenous as they were of Moroccan descent, 36 percent each. Because native Dutch people in that age group make up 39 percent of the total and Moroccans 16 percent, this study showed that there was Moroccan overrepresentation. perpetrator for assault: Dutch: 37,5% Foreign: 50% Mixed: 12.5%]

4.2.1.3 Conclusion

The results of VADER analysis are not only remarkable but also very encouraging. The outcomes highlight the tremendous benefits that can be attained by the use of VADER in cases of micro-blogging sites wherein the text data is a complex mix of a variety of text.

If we took a closer look at how Amsterdam stands against other cities, we will see that it comes in 4th place for the number of foreign people visiting, 7th for the percentage of positive comments online, 8th place for GDP and 6th for e-commerce active, 8th for internet access and population. Having almost 67% from the collected sample showing a positive sentiment about Amsterdam is indeed a good indicator, showing little discomfort for the remaining negative discussions and what they could represent in the future will be very challenging for policy makers as it seems that people are generally happy about the cost and quality of life in Amsterdam, however, online heated online discussions are focused on the polices that is undertaken by the city like the legalization of cannabis, and how it can has its effects on the city's livability and people's mental health, as well as, legalization of prostitution, and gender identity.

On February 2020, Amsterdam's mayor was considering banning tourists from buying cannabis in the city's coffee shops, to solve the problem of the overcrowding in the city's red-light district. It's been said that a third of foreign tourists would be less likely to visit Amsterdam again if they couldn't buy cannabis in coffee shops **(Euronews, 2020).**

4.2.2 Barcelona

***Figure 4. 15** Geographical representation for the number of the words and comments collected for the city of Barcelona*

For Barcelona city, we managed to collect 2246 comment, with a total of 5,594,895 words. Out of those 3554 comments, 2073 are unique values, with a total of 4,921,591 words to process and analyze.

4.2.2.1 Text Processing and Visualization

After carrying out text processing as before, we now have 453,057 clean text. In table 4.4, the processed text has a total average of around 200 words per comment, while not processed text has around 2300 words per comment as average. The longest processed comment has 1201 words.

***Table 4. 4** Barcelona text descriptive analysis*

Before Processing	After Processing
count 2073.000000	count 2073.000000
mean 2374.139411	mean 218.551375
std 2539.111869	std 237.805565
min 26.000000	min 3.000000
25% 592.000000	25% 55.000000
50% 1300.000000	50% 119.000000
75% 3118.000000	75% 284.000000

max 10391.000000	max 1201.000000

We can see the previous data represented in the following figures

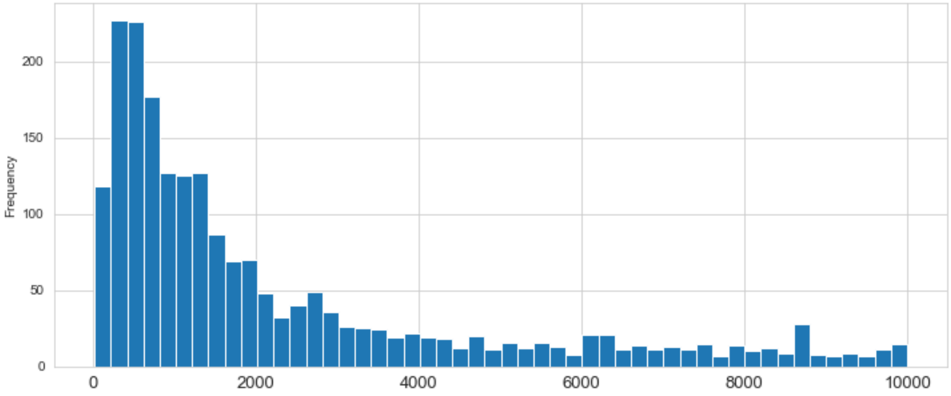


Figure 4. 16 *Barcelona words per comments before processing*

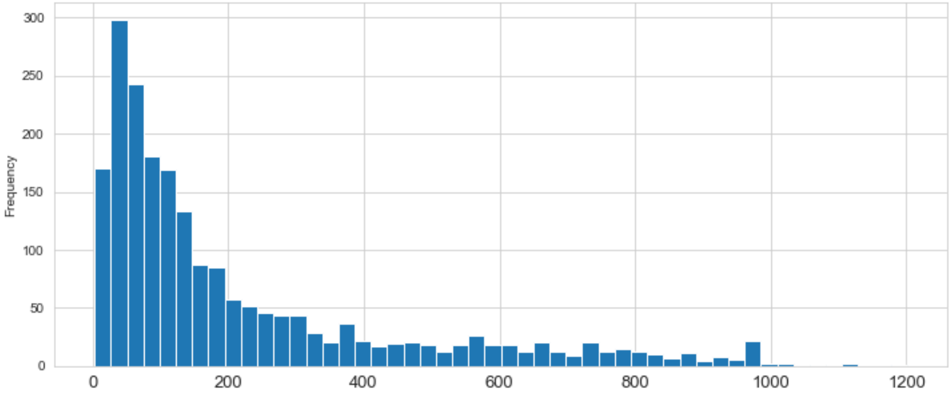


Figure 4. 17 *Barcelona words per comments after processing*

In the following figure we can visualize the top 35 most frequent words acquired from the text. While on the figure after, a word cloud is done on the most 150 frequent words. We can notice that people discussed conversations about Barcelona team, the players, studying, and life.

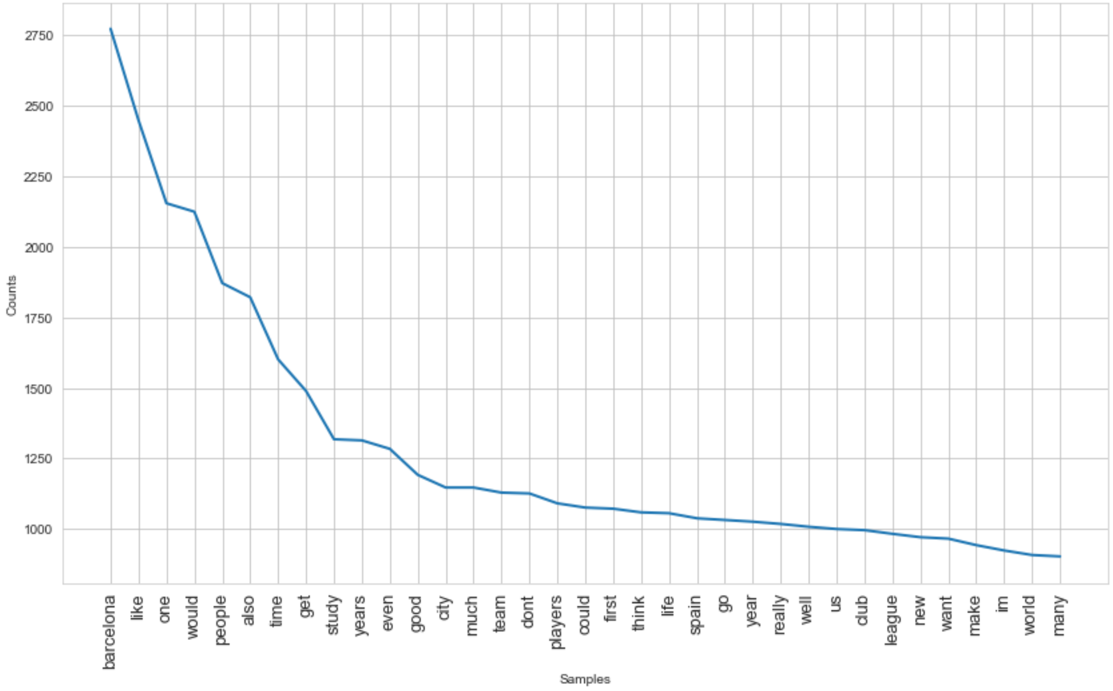


Figure 4. 18 *Barcelona 35 most frequent words for all sentiments*

We can sort the word counts and plot their values on Logarithmic axes as usual in the next figure.

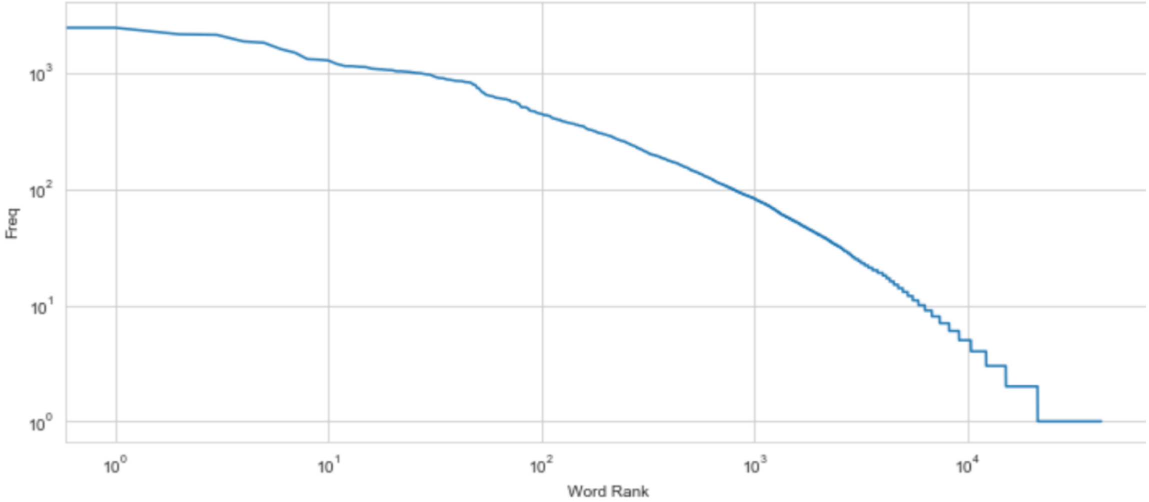


Figure 4. 19 *Barcelona word counts distribution on a logarithmic scale*

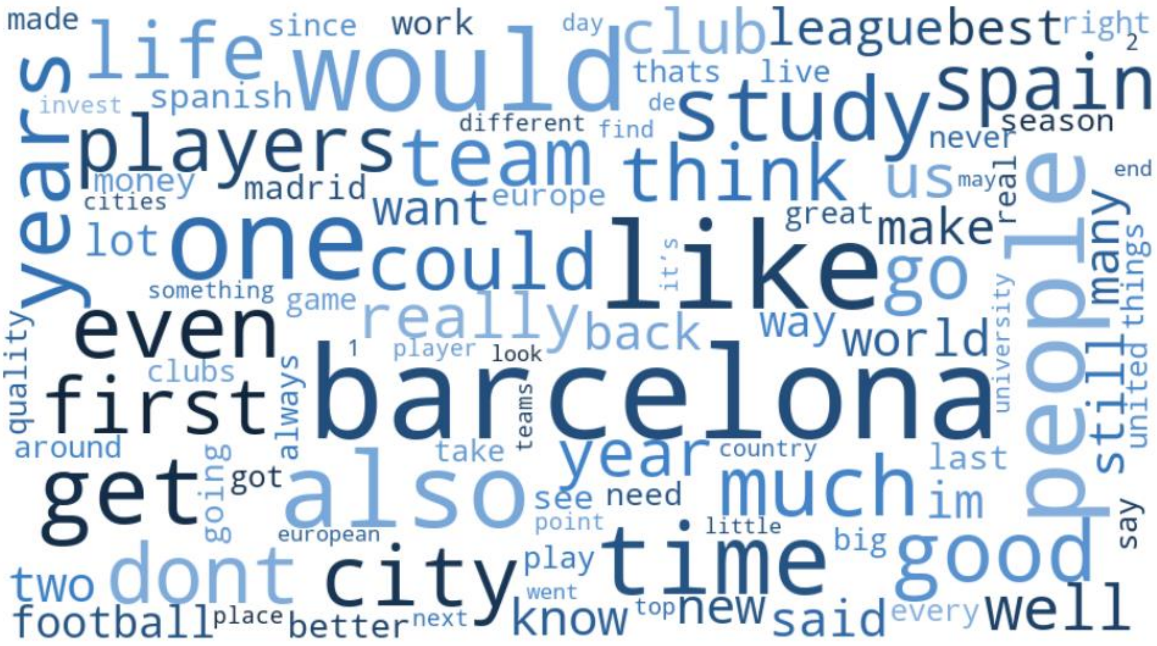


Figure 4. 20 Barcelona word cloud on all sentiments

4.2.2.2 Sentiment Analysis & Word Clustering

Moreover, we will use sentiment Analysis on the text to label it. As before we will work on the semi-processed text with a total 4,778,991 words. In the next table we have the results of the analysis. 77.9% of the comments collected are positive, 19.84% are negative and 2.26% are neutral as represented in the figure.

Table 4. 5 *Barcelona sentiment results*

Negative (-1)	Positive (1)	Neutral (0)	Total
411	1615	47	2073

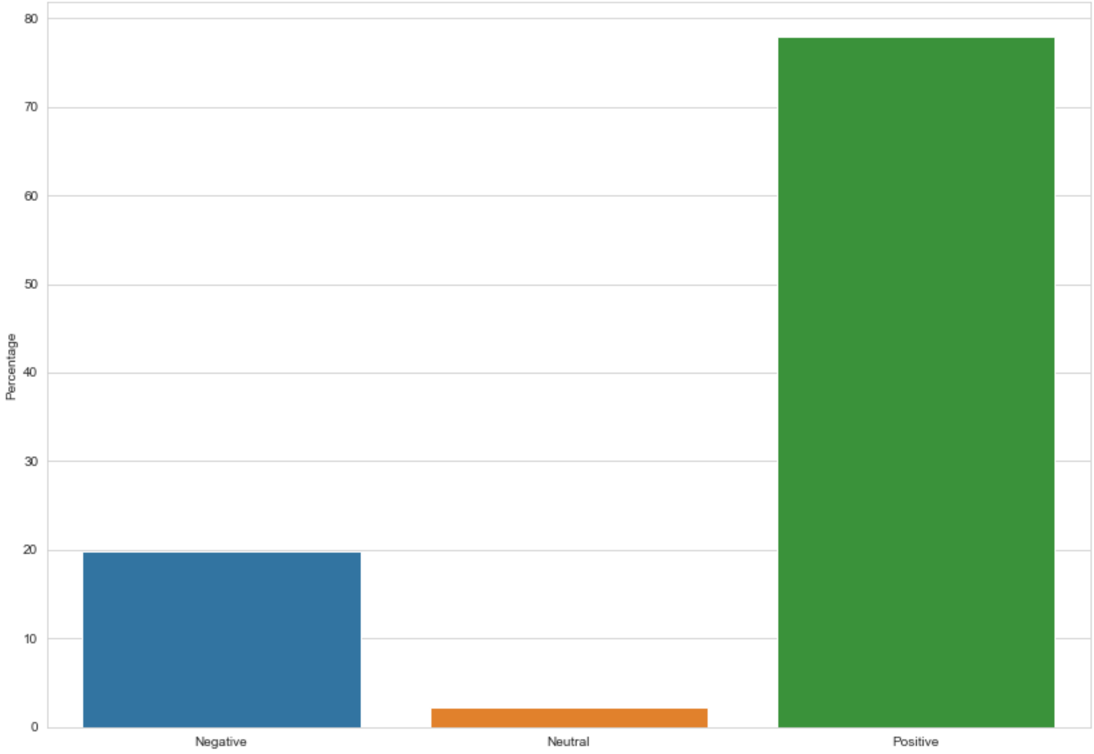


Figure 4. 21 *Barcelona sentiment results in percentage (%)*

In the following figure we can visualize the top 35 positive and most frequent words acquired from the analysis. While on the figure after, a word cloud is done as before. We can notice a slight change in the words comparing to the one from before as the percentage of the positive sentiment are over 70%. The sentiment analysis tells us that people talked in a positive way about topics like *Barcelona team, player, study, life, and football.*

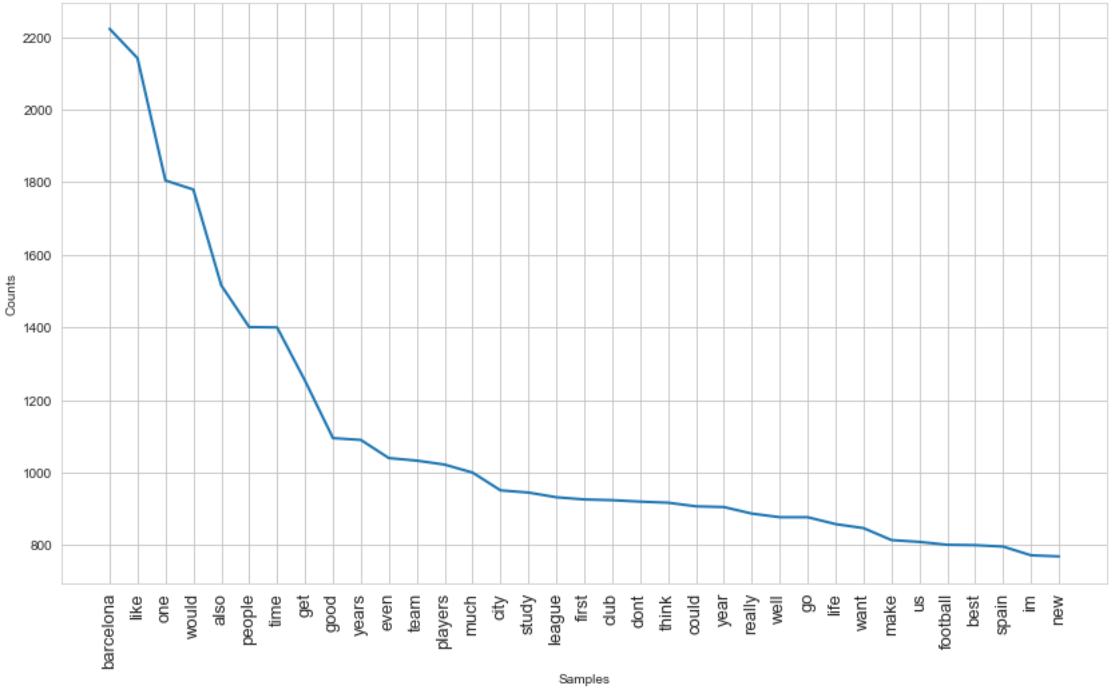


Figure 4. 22 *Barcelona 35 most frequent words for positive sentiment*

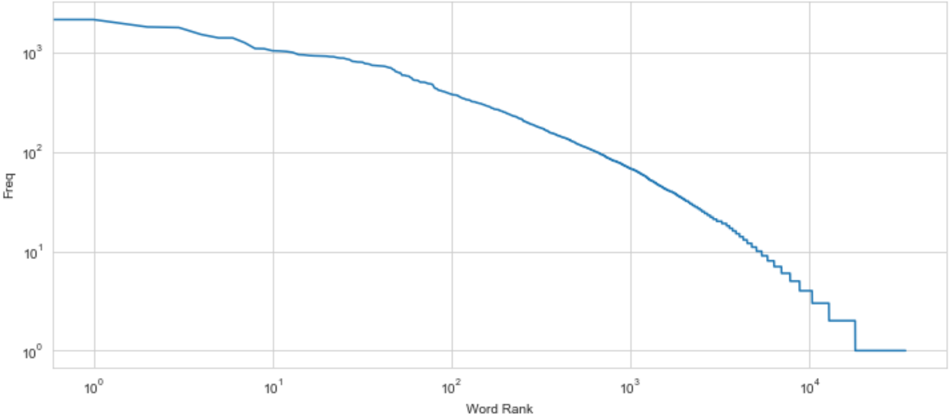


Figure 4. 23 *Barcelona word counts distribution for positive sentiment on a logarithmic scale*

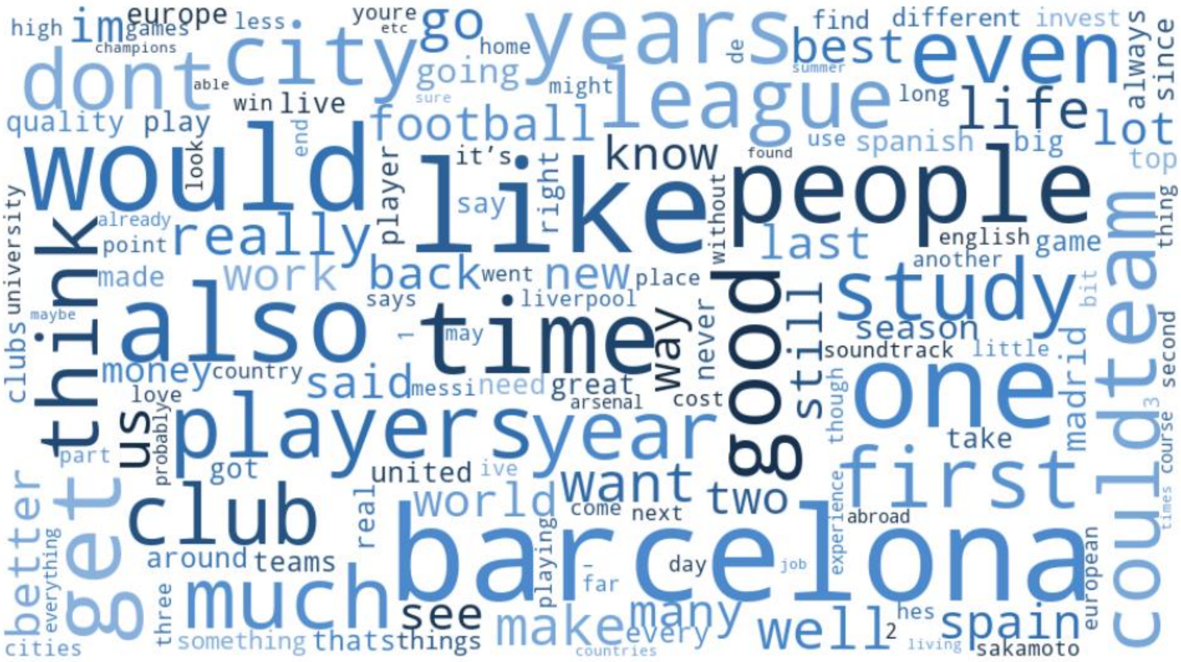


Figure 4. 24 *Barcelona word cloud on positive sentiments*

The analysis demonstrate that people spoke in a negative way about topics related to *the country, Spain, the government, and the Catalan people*, those in fact are political topics relating to the ongoing issues between the Catalonia region and the government of Spain as Catalonia's drive for independence pushed Spain into its biggest political crisis for 40 years. The region had its autonomy suspended for almost seven months by Madrid after a failed bid to break away in 2017. In October 2019, Spain's Supreme Court sentenced nine Catalan politicians and activists to jail terms of between nine and 13 years for that independence bid **(BBC, 2019)**.

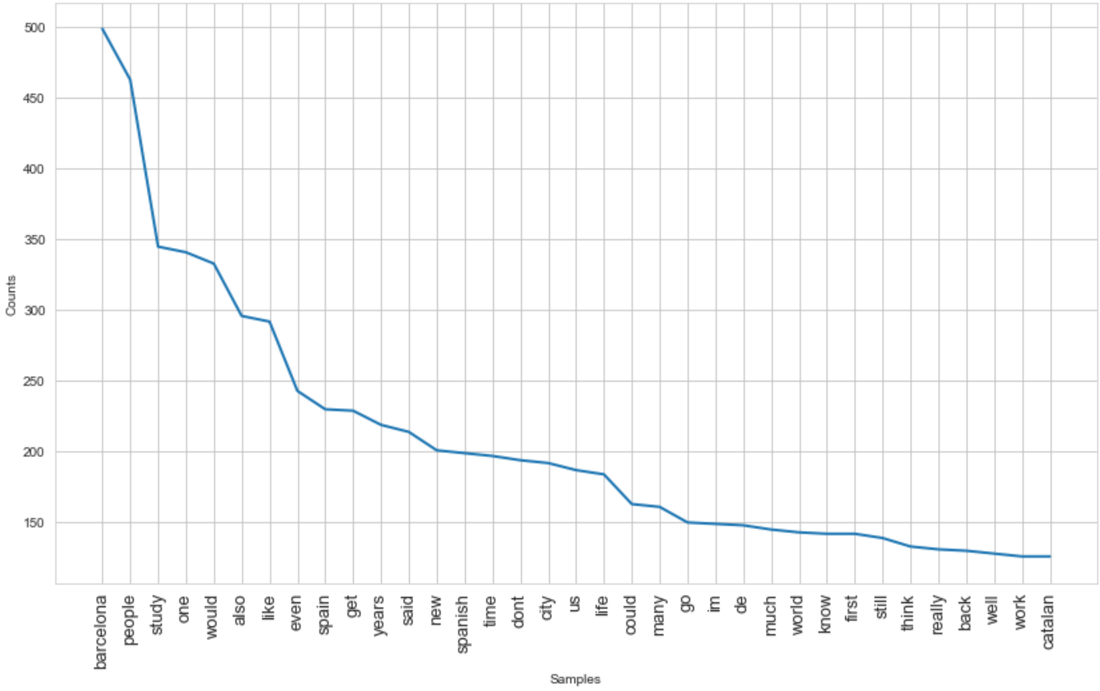


Figure 4. 25 *Barcelona 35 most frequent words for negative sentiment*

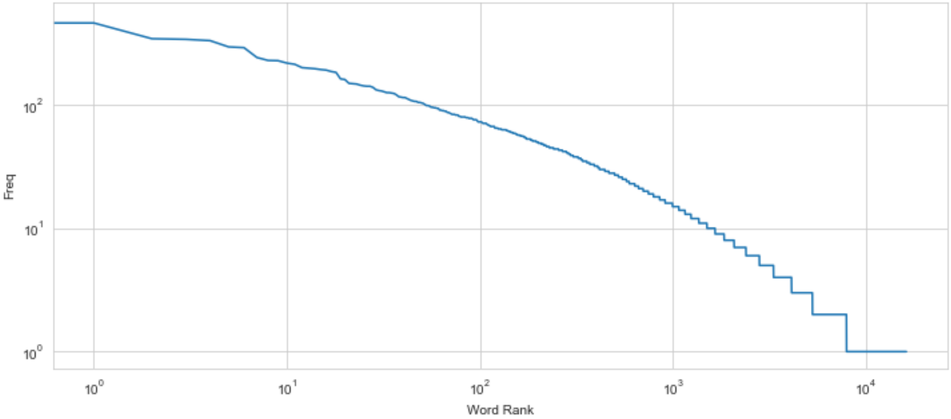


Figure 4. 26 *Barcelona word counts distribution for negative sentiment on a logarithmic scale*

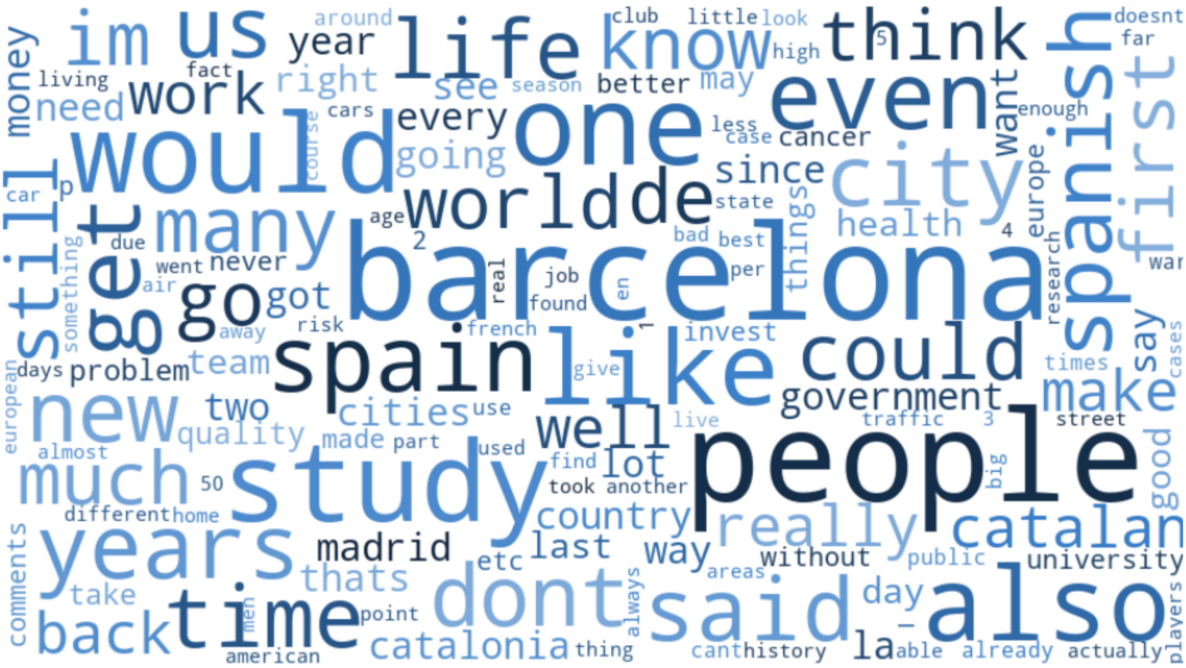


Figure 4. 27 Barcelona word cloud on negative sentiments

In the next part, word clustering results is carried out.

- Positive Clusters & Discussions

[[' life', ' quality', ' spain', ' living', ' people', ' barcelona', ' live', ' city', ' cities', ' cost'], [' study', ' barcelona', ' like', ' people', ' im', ' time', ' city', ' invest', ' europe', ' spain'], [' sakamoto', ' soundtrack', ' album', ' film', ' 1999', ' 1992', ' sylvian', ' 2002', ' 2005', ' piano'], [' abroad', ' study', ' spanish', ' barcelona', ' english', ' school', ' students', ' spain', ' people', ' like'], [' players', ' club', ' team', ' league', ' football', ' clubs', ' like', ' teams', ' season', ' united']]

- [‘Software developers can get 30k a year with 5 years of experience but could also get 60k in the Netherlands, 80k in Switzerland or 100k+ in the Valley.   30k outside Madrid / Barcelona buys you roughly the same quality of life as 60k in the Netherlands or 80k in Switzerland... and a lot more than 100k+ in the Valley (no reasonable amount of money will buy you quality of life in that hellhole). most of those jobs are in Madrid or Barcelona where 30k won't get you far unless you live with your parents. You haven't looked much. There are thousands of readily open positions in the smaller hubs like Málaga, Valencia, Bilbao, Sevilla or Coruña.   It's not supply and demand because the emigration of developers is massive here and its really hard to get senior developers here. That's exactly what supply and demand looks like. Pay them more and they'll come, most expats that I talk to (yes, this is anecdotal evidence) would love to come back.’]

- ['I worked in Barcelona for two years and it was the best two years of my life! However, the pay is rubbish. Enough to live a nice quality of life, eat and drink out a lot, rent a decent apartment but never enough to save. If you're ill, take leave, or even students cancel classes you can lose money. I also went back to the uk every summer to do a summer school, which I didn't mind because I'm from there and it was nice to be near friends and family, you get free board so make decent cash, and as I was earning under the tax threshold had a nice lump sum to move back to Spain with in September. Saying that, I don't know what the implications of two months with no work would mean for USA citizens.']
- ['On my last trip before deciding to move, I was out late on a Sunday night in Barcelona and walked past a group of little old ladies sitting at a table on a terrace drinking beer and talking and laughing, and thought that is something I'll never see in the US. When I got back home, seeing that picture made something snap, and I decided to sell everything and make a move. I had already been offered to sell my part of my company so it was a quick decision. 17 years later I am glad I did it. Quality of life is better. Cost of living is lower. The weather is a lot better than what we had in Seattle. Free healthcare is available, though I pay for private healthcare. Sure there are downsides like everywhere. We have loudmouth idiots here too (they call them brother-in-laws as a general term) and the bureaucracy and poor organization can be very frustrating at times. Still it was a good decision.']

- Negative Clusters & Discussions

[[' city', ' club', ' barcelona', ' life', ' dont', ' people', ' invest', ' money', ' im', ' cancer'], [' barcelona', ' like', ' study', ' life', ' game', ' abroad', ' team', ' time', ' got', ' quality'], [' spain', ' country', ' high', ' people', ' barcelona', ' study', ' europe', ' government', ' madrid', ' im'], [' spanish', ' piano', ' catalan', ' albéniz', ' study', ' speak', ' catalonia', ' guitar', ' works', ' music'], [' people', ' 50', ' terrorists', ' years', ' terrorism', ' trump', ' study', ' blood', ' took', ' caught']]

- ["Europe has a lot of positive things, but not as a whole, only in variations as your cross from region to region. Poor and struggling people in Barcelona make a lot of US cities feel great. Roma settlements in Slovakia are no less imporverished and no better quality of life than poor black towns in rural Alabama--both get government assistance, both have a lot of obstacles. In Europe you can still have a shitty life, but it's more than you're unlikely to have much social mobility, upward or downward. You aren't as exposed to as many random, large opportunities as in the US, but you also aren't exposed to as much chance of total ruin. You may not accumulate as much debt (but trying buying a house in major European metro areas, yikes), yet you also probably won't accumulate as much savings to your economic peers in the US. America has a lot more extremes, and both positive and negative chances of reaching those, Europe more stable but limited comfort."]

- [‘Sorry, you don't have a clue of what is going on in Catalunya. Like, you have NO idea. Please get your facts checked, and I don't know, perhaps get a dose of reality and start throwing whatever media you use for being informed to the trash can. So are you talking about Spain bombing Catalunya? Well, you should know both the Italians and the Germans helped fascist Spain bomb Barcelona (and many other Catalan towns). The current Spanish regime is a direct descendant of 1936-1975 fascist Spain, only adapted to maintain appearances, but still run by the hardcore fascists, especially in the judiciary and military/police branches. You should know a famous hero of Spanish nationalism, which btw is a banal nationalism of the worst kinds, meaning Spanish nationalism does not even seem to officially exist, which makes its values, including supremacist views invisible to many people, said this: "You have to bomb Barcelona at least once every 50 years". Of course, Spanish heroes are mainly military Generals with long histories of spilling their own people's blood for the sake of their Empire. Just like the invasion of America and the bloodshed that ensued killing millions in the name of religion and the kingdom of Spain, or the romantic views even today you get of Torquemada, who "defended Christianity" (hence Spain) with the Inquisition against heretics (ie. anyone, heretic or not, not in a good standing with those in power).’]
- ["well there are many things that we would achieve: Economically speaking we would have control over the airports and maritime ports and barcelona could develop better, for example the barcelona port currently brings 19% of the benefits to the portuary(?) system of spain and only gets 0.7% back, also the airport of el prat needs expanding and the central government is doing everything it can to not allow it. There's also collaborations with other countries like the 5G corridor between catalonia and occitania (it was stopped by the foreign affairs minister). There's also the cultural and linguistic problems. Currently as a catalan even if i go to a catalan university i am not guaranteed to be able to take all my classes in catalan because there's always someone that demands X class to be in spanish (even if it was stated before that the class would be in catalan) and the teachers mostly always accept. If we were independent we would be able to study fully in the language. Other things like not being accused of terrorism or stuff like that for speaking catalan would be nice too (it's happened too many times). Those are mostly the reasons i want independence, other stuff would be nice to have but those depend on catalan politicians, for example Puigdemont spoke some time ago about modernizing things like the use of the electronic vote that Estonia already uses, etc.. There's also the aran valley, they have their own language too (occitan), culture, institutions.. i believe the catalan government promised the aranese goverment to be able to vote wether they want to remain in catalonia or remain in spain after independence.. These problems would easily be solved with more autonomy but we've reached a point where a spanish politician showing some sort of wish to dialog with independentists means they are suiciding electorally speaking hahah. The mass media of spain really likes to show it's hate towards catalans (or independentists, people don't really differentiate the two in spain). Things like lying, stuff like that, we've seen it recently with the 9 independentists arrested accused of terrorism a few days ago, it just feels like we're occupied by spain rather than a part of spain. Stuff like this also happened with the basque independentist movement, police tortured basque

> activists, there was the whole ETA thing, at least with the catalan independence movement there's no terrorist group.. Spain doesn't really show any sign of change towards becoming a better state, sadly."]

4.2.2.3 Conclusion

Barcelona city online discussions sentiments are highly concentrated on four topic, cost of life, quality of life, the football team, and the Catalonia independence issue. Most of the discussions agree on the very low cost of life that Barcelona has to offer, addition to the advantages that the Spanish culture has, however, a lot of concerns are focused on the low wages as well. More heated discussions are focused on the independence issues and what the people of Catalonia are facing and their position against the government of Spain.

According to **(BBC, 2019)** , Catalonia's drive for independence pushed Spain into its biggest political crisis for 40 years. In October 2019, Spain's Supreme Court sentenced nine Catalan politicians and activists to jail. It's important to mention that Barcelona stands 3rd for the numbers of yearly foreign visitors against other cities, 1st for the percentage of positive comments online, it comes in 6th place for their GDP, 2nd place for e-commerce active, population and the number of people that has internet access, in general, a 77% positive sentiment is a very good indicator of the public opinion about Barcelona.

4.2.3 Berlin

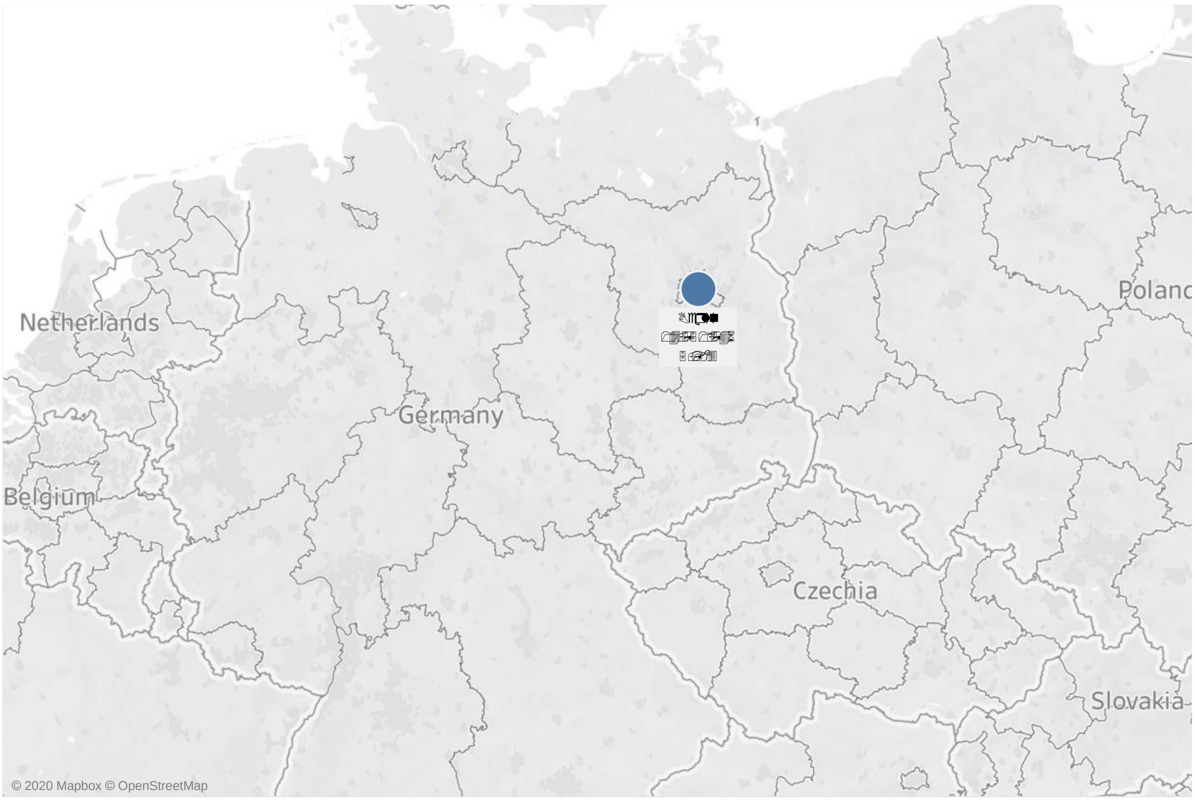


Figure 4. 28 *Geographical representation for the number of the words and comments collected for the city of Berlin*

We managed to collect 5737 comment, with a total of 17,236,639 words, 5183 are unique values, with a total of 14,551,546 words to process and analyze.

4.2.3.1 Text Processing and Visualization

After processing the text, we now have 1,296,454 words to analyze.

In the following table, we can see the difference between the data before processing it and the data after, we are also exploring the frequency of words per each comment. The processed text has a total average of around 250 words per comment, while not processed texts has around 2600 words as average. The longest processed comment has 3432 words.

Table 4. 6 *Berlin text descriptive analysis*

Before Processing	After Processing
count 5183.000000	count 5183.000000
mean 2626.680108	mean 248.133706
std 2552.941645	std 242.476741
min 32.000000	min 3.000000

25% 710.000000	25% 67.000000
50% 1606.000000	50% 150.000000
75% 3944.500000	75% 371.000000
max 39772.000000	max 3432.000000

We can see the previous data represented in the following figures

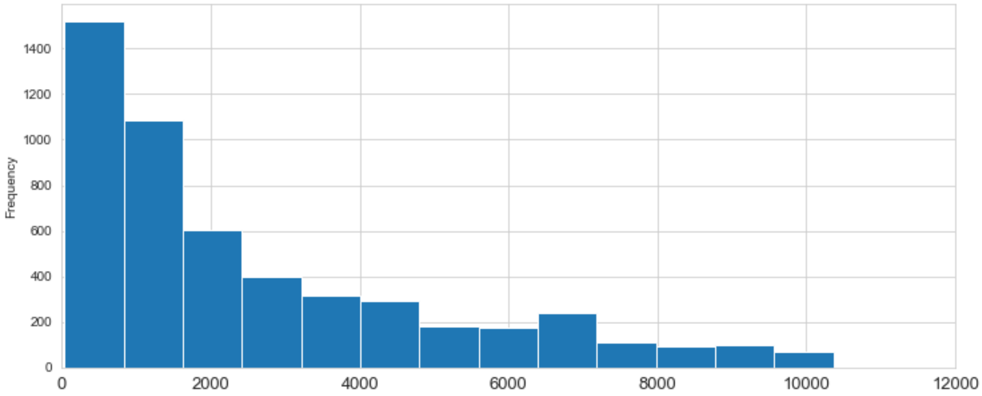


Figure 4. 29 *Berlin words per comments before processing*

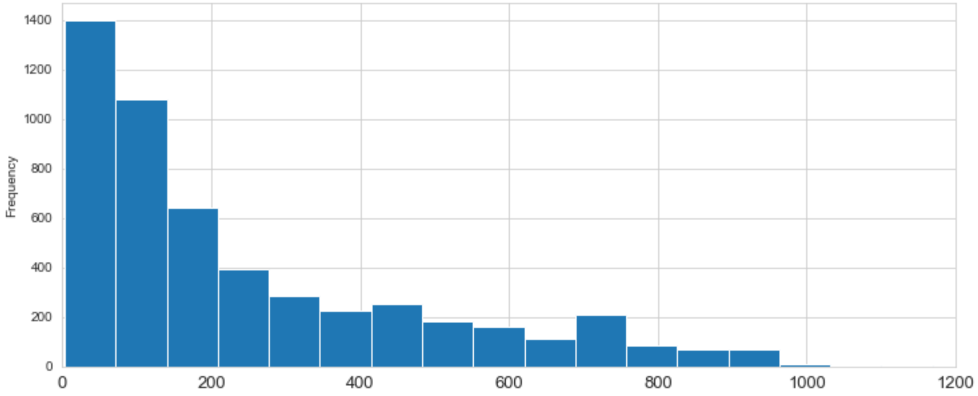


Figure 4. 30 *Berlin words per comments after processing*

In the following figure we will visualize the most frequent words and the word cloud. We can notice the topics that people usually talk about are world war, studying, Muslims and the German language.

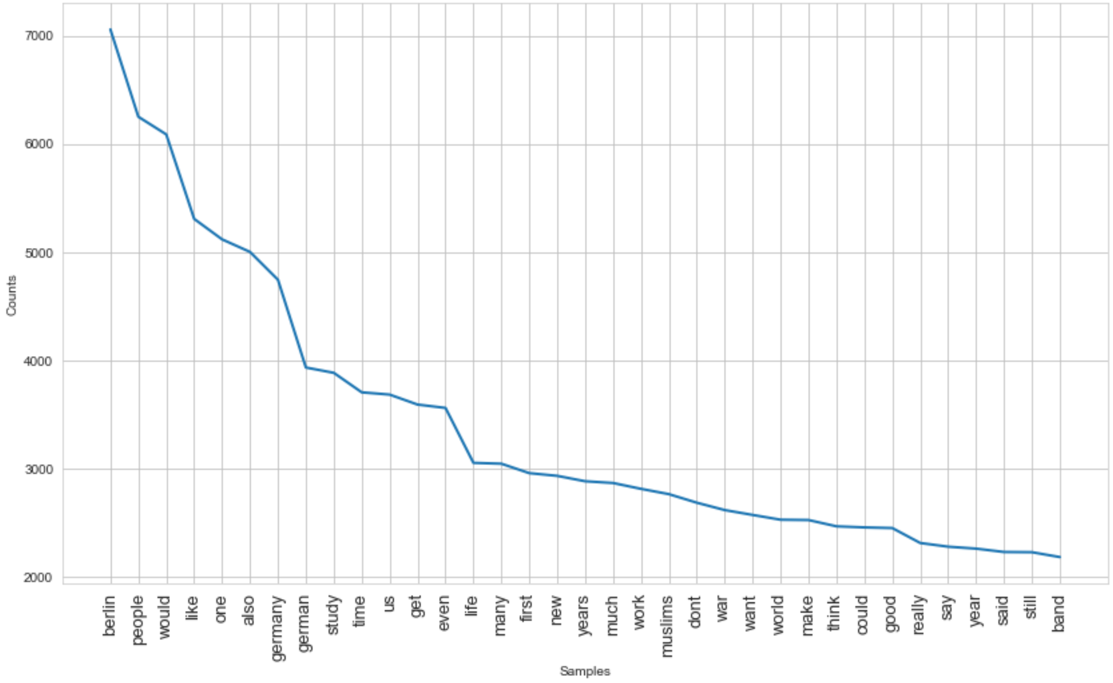


Figure 4. 31 *Berin 35 most frequent words for all sentiments*

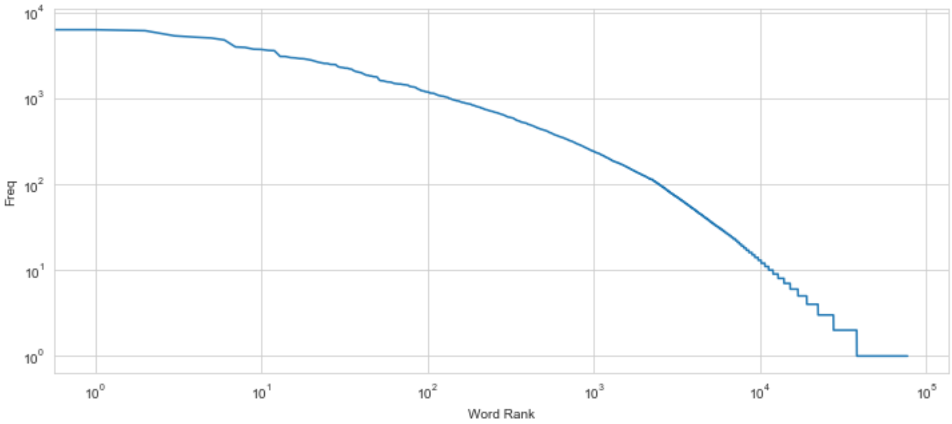


Figure 4. 32 *Berlin word counts distribution on a logarithmic scale*

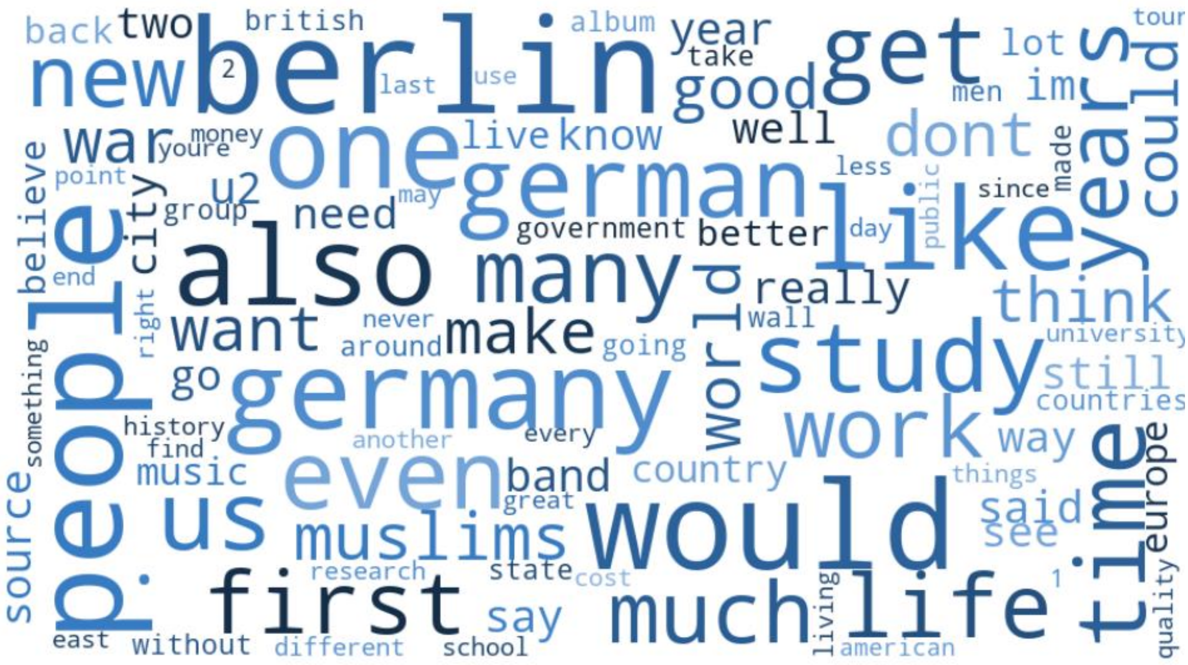


Figure 4. 33 Berlin word cloud on all sentiments

- Sentiment Analysis & Word Clustering

it’s time to use Sentiment Analysis on the text to label it. In the next table we have the results of the analysis. 65.88% of the comments collected are positive, 31.67% are negative and 2.45% are neutral as represented in the figure.

Table 4. 7 Berlin sentiment results

Negative (-1)	Positive (1)	Neutral (0)	Total
1641	3415	127	5183

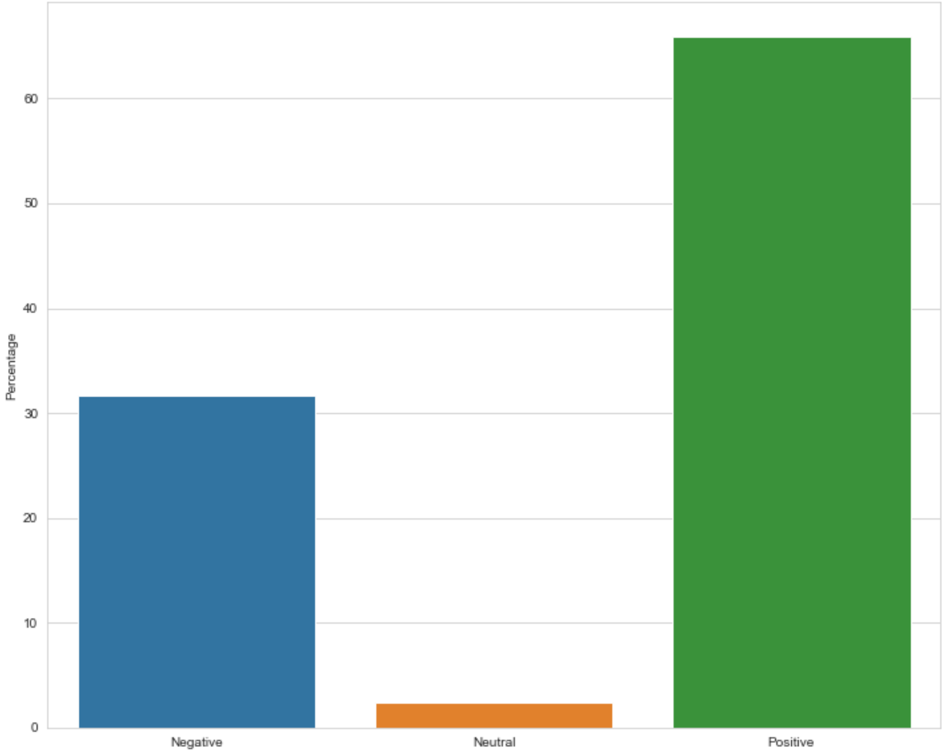


Figure 4. 34 *Berlin sentiment results in percentage (%)*

In the following figure we can visualize the top 35 positive and most frequent words and the word cloud for the most 150 frequent words. We can notice a slight change in the words comparing to the one from before as the percentage of the positive sentiment is high. The sentiment analysis tells us that people talked in a positive way about topics like *work, German language, music, study team, player, study and life.*

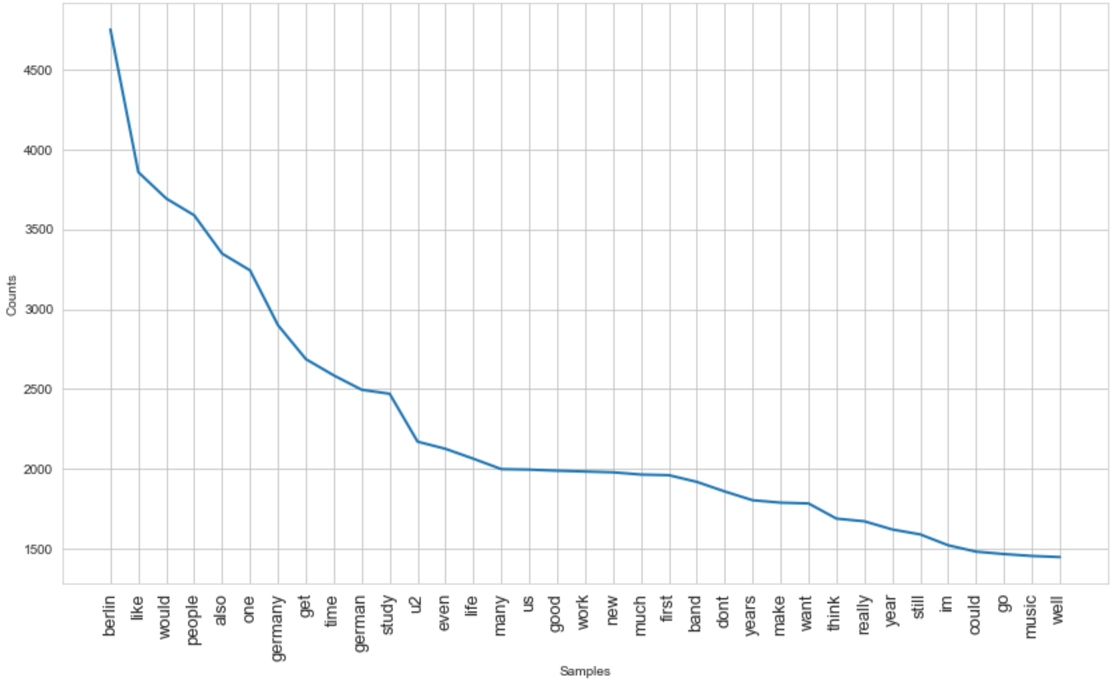


Figure 4. 35 *Berlin 35 most frequent words for positive sentiment*

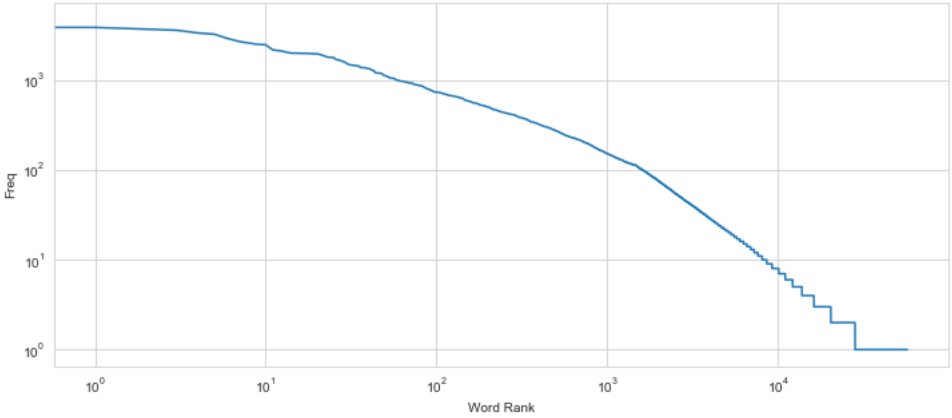


Figure 4. 36 *Berlin word counts distribution for positive sentiment on a logarithmic scale*

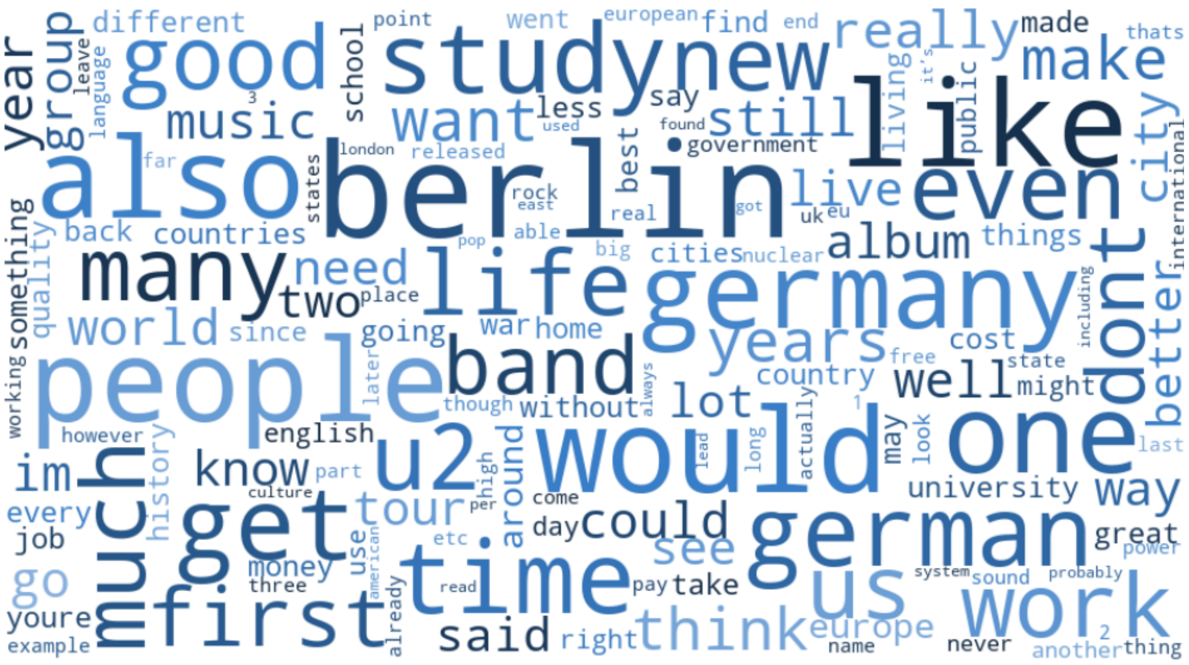


Figure 4. 37 *Berlin word cloud on positive sentiments*

However, for the negative sentiment, people spoke mainly about, *war, Muslims, suicide and attacks*. Those topics can be related to the fact that Germany has had some recent terrorist attacks, the last one when a far-right terrorist has killed at least nine people in a city in western Germany **(BBC, 2020).**

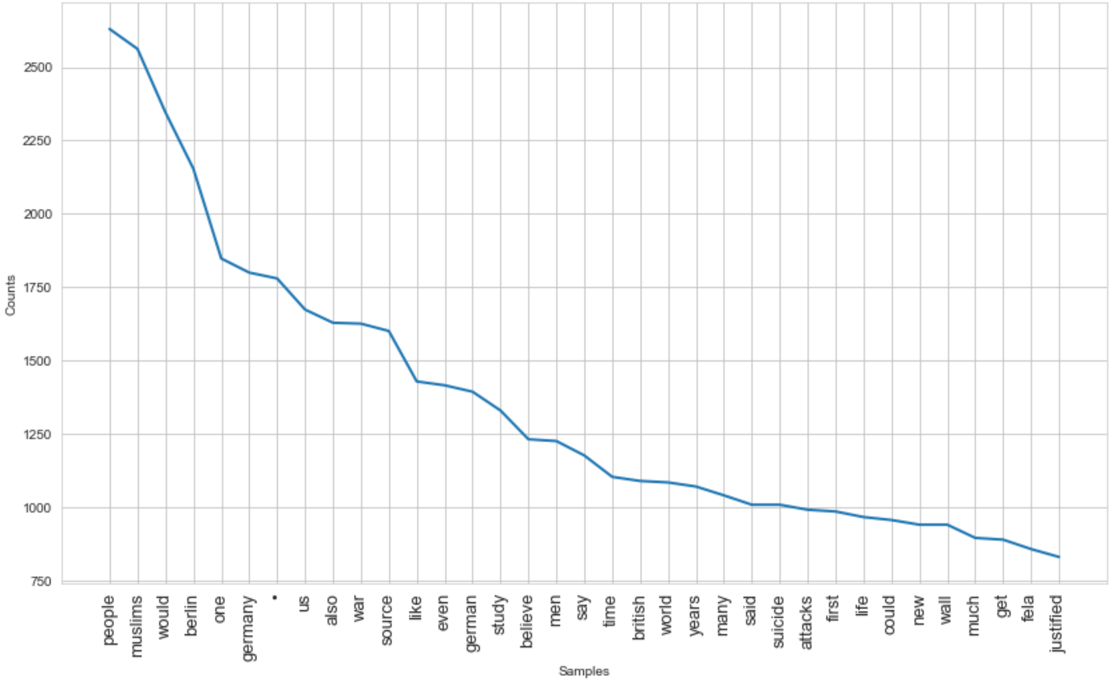


Figure 4. 38 *Berlin 35 most frequent words for negative sentiment*

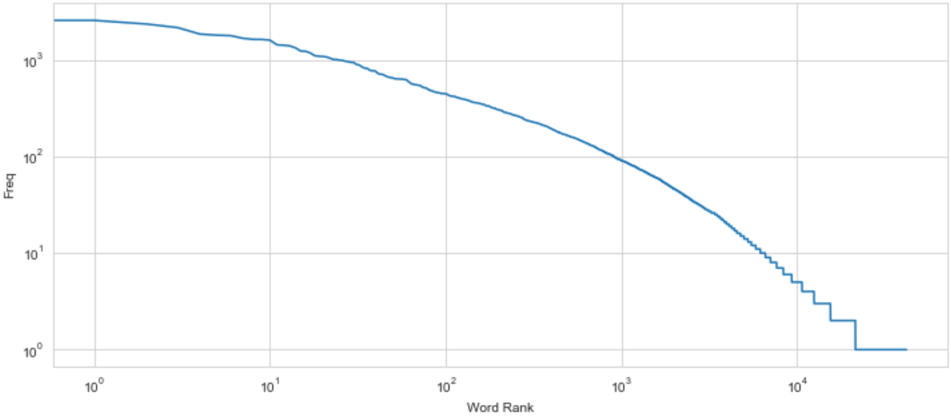


Figure 4. 39 *Berlin word counts distribution for negative sentiment on a logarithmic scale*

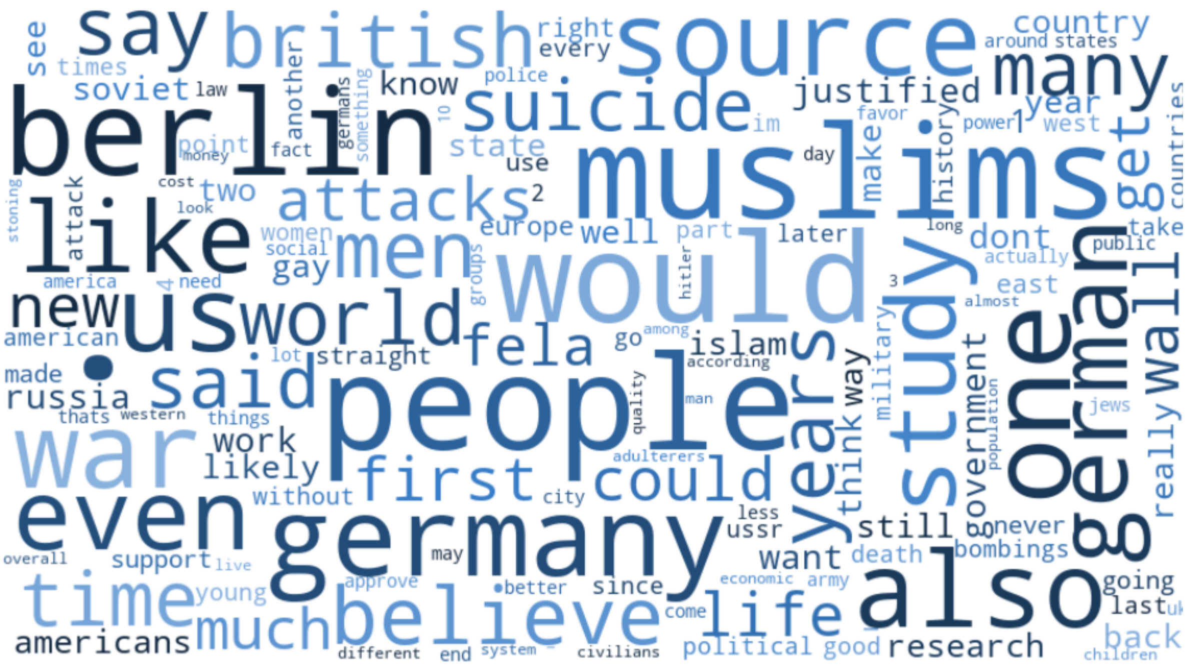


Figure 4. 40 *Berlin word cloud on negative sentiments*

In the next part, word results is carried out.

- Positive Clusters & Discussions

```
[[' nuclear', ' germany', ' people', ' german', ' new', ' government', ' wall', ' study', ' said', ' berlin'], [' u2', '
band', ' album', ' tour', ' bono', ' music', ' rock', ' pop', ' group', ' sound'], [' city', ' life', ' quality', ' cities', '
berlin', ' living', ' people', ' like', ' live', ' better'], [' berlin', ' study', ' like', ' people', ' im', ' dont', ' time', '
really', ' want', ' invest'], [' german', ' germany', ' visa', ' english', ' language', ' job', ' work', ' university', '
study', ' degree']]
```

- ['All right. If it\'s 47000 Euro , then you have 2395,29 Euro monthly, post-tax and insurance. So if you say you will have less than 1000 Euro monthly, that means you think about paying 1400 Euro in rent? Yes, that may happen if you need to live in the city center. It\'s the same in most big cities, no matter where. As capitals go, Berlin is still among the cheaper ones. There aren\'t really "rural" areas close to Berlin, but rents certainly get cheaper as you move further from the center. And Germany (this may be another feature of "life quality" that not every country has) has good public transport, at least around major cities, which enables people to commute into the city using said public transport. In general, due to the high demand, finding housing in Berlin is hard. People often need months to find something. And if you think about paying 1400 Euro rent from a 2400 Euro income... plenty of landlords wouldn\'t rent a flat that expensive to someone with that income, if they have other applicants who make more or who have two earners in a family.']
- ["lol you really that 5% is the reason for the difference? :P that's not what I meant. I meant, a middle class person with a 45k euros (brutto) income can live comfortably, have a car, go on holidays, comprehensive healthcare coverage even in the most expensive german cities (Berlin-Hamburg-Frankfurt-Munich). Tax rate in total for that amount of money for a single person (not married) is about 30-35% (so deduct that and you get the netto income). That's what I meant by healthy middle class and policies that make life good for anyone. Finding affordable housing is a huge effort, public transportation can be pretty expensive depending on circumstances, you know the usual problems, but overall quality life is more affordable."]
- ["Why don't you look at Germany? Most universities are free (200/term but you get a free travelcard for the whole year which makes it basically free). The websites are often in German, but nearly all is taught in English. In Berlin there is UdK, there#s also other famous schools Dussledoerf, the Staedel(for Masters) in frankfurt. Otherwise look at other Northern European countries where it is free to study, and give stipends to poor students. (exclude England from this list)"]

- Negative Clusters & Discussions

[[' study', ' protesters', ' activists', ' live', ' parents', ' berlin', ' unemployed', ' 92', ' leftwing', ' finds'], [' muslims', ' justified', ' believe', ' suicide', ' attacks', ' islam', ' british', ' bombings', ' approve', ' say'], [' war', ' german', ' germany', ' hitler', ' soviet', ' people', ' world', ' germans', ' berlin', ' history'], [' fela', ' russia', ' felas', ' music', ' band', ' putins', ' nigeria', ' kuti', ' ransomekuti', ' nigerian'], [' people', ' wall', ' source', ' berlin', ' dont', ' germany', ' study', ' like', ' work', ' life']]

- [“A survey by Prof. Ruud Koopmans at Humboldt University in Berlin revealed that over 45% of German Muslims and 70% of Dutch Muslims consider the religious rules of Islam to be more important than the secular laws of the country where they are living. [Of muslims 18-29 years old, polled in 2007,](26% of Muslims in the US Think suicide bombings can be justified. That's nearly 1/3. Here are other countries as reference, though I suggest you take a look through the study I linked to, it's enlightening. 35% of Muslims in Great Britain Think suicide bombings can be justified 42% of Muslims in France Think suicide bombings can be justified 22% of Muslims in Germany Think suicide bombings can be justified 29% of Muslims in Spain Think suicide bombings can be justified "]

- ["There are some examples of firms that pay US level wages in Europe and the UK. Palantir in London is one such firm. Swiss pay is generally similar to American pay and taxes might be lower, depending on your situation. I.e. can a New Grad expect a similar pay/benefit situation as a Software Engineer for Google in Berlin as in San Francisco? For that specific example, Berlin new grads get less than SF/SV new grads at Google. Whether or not they have lower quality of life is debatable. There's someone here who voluntarily moved from the US to Germany at Google so hopefully they'll see this thread and elaborate."]
- ["My observation is that PHP pays less than some other languages, perhaps 20% less. There are lots of experienced migrants fighting for jobs after finishing graduate diplomas etc. That is a common route in my experience - do a cheap graduate diploma - get the post study work visa. Find a job then use the job to get points for residency. The salary range is probably $45k to $75k. With the right skills you could get more. Many jobs will be looking for full-stack devs who have good javascript skills too. Do you have a work visa ? Probably tough trying to find a job without a valid visa. Wages are falling due to market saturation. Also note that housing here is very very expensive. Why don't you look for a job somewhere else in the EU ? I heard there were lots of startups in Berlin for example. "]

4.2.3.2 Conclusion

Berlin comes 8th as a popular foreign destination, 9th for the percentage of positive comments, 10th on GDP, 5th for e-commerce active, 6th internet access and population. Berlin achieved 65% positive answers form the total collected, some topics about terrorism are real issues online, as Germany has some 4 million Muslims, while Berlin has around 350,000, this is due to the labor migration in the 1960s and several waves of political refugees since the 1970s as a part of the German easy policies of inviting foreign workers to the country **(Pew Forum on Religion and Public Life, 2017).**

According to a study by **(Leipzig University, 2018),** 56% of Germans sometimes thought the many Muslims made them feel like strangers in their own country, up from 43% in 2014. In 2018, 44% thought immigration by Muslims should be banned, up from 37% in 2014. During the last couple of years, a lot of Muslims and Islamic institutions were attacked through racism or other kinds of discrimination, this is a crucial topic for policy makers in Berlin to focus on before the situation can get out of control, online discussions can gave us a perspective on how things really are, branding and marketing can turn the wheel around and promote a culture of tolerance nonviolence and peace and draft the political, social, and economic participation of Muslim communities living in the city.

4.2.4 Dublin

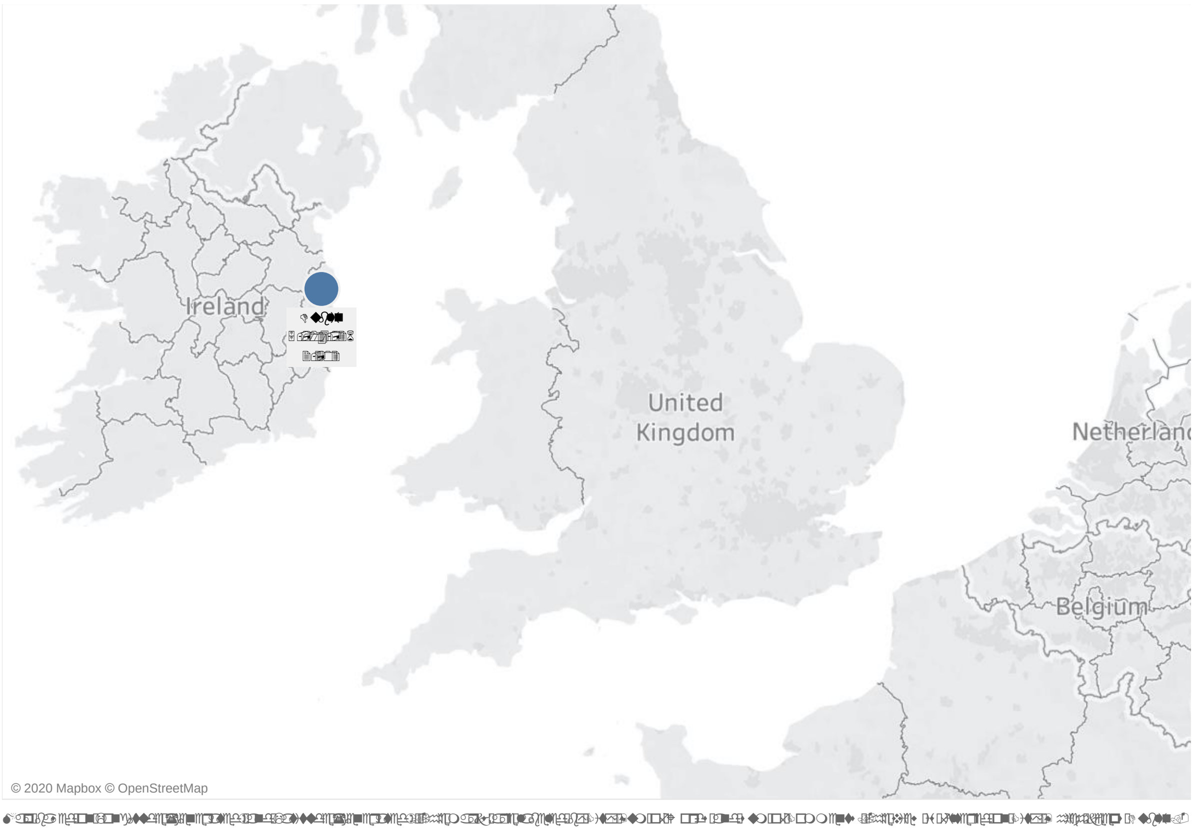


Figure 4. 41 *Geographical representation for the number of the words and comments collected for the city of Dublin*

We managed to collect 2537 comment, with a total of 6,039,560 words, 2302 are unique values, with a total of 5,114,026 words to process and analyze.

4.2.4.1 Text Processing and Visualization

After processing the text, we now have 476,785 words to analyze. In the following table, we can see the difference between the data before processing it and the data after, we are also exploring the frequency of words per each comment. The processed text has a total average of around 200 words per comment, while not processed texts has around 2150 words as average. The longest processed comment has 1094 words.

Table 4. 8 *Dublin text descriptive analysis*

Before Processing	After Processing
count 2302.000000	count 2302.000000
mean 2155.703736	mean 206.106864
std 2341.154289	std 227.611260
min 43.000000	min 4.000000
25% 541.000000	25% 51.000000
50% 1128.500000	50% 107.000000
75% 2857.500000	75% 265.000000
max 9954.000000	max 1094.000000

We can see the previous data represented in the following figures

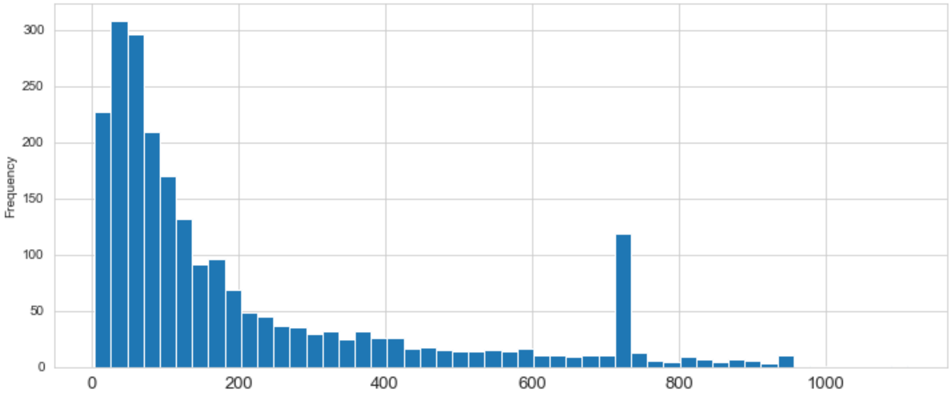


Figure 4. 42 *Dublin words per comments before processing*

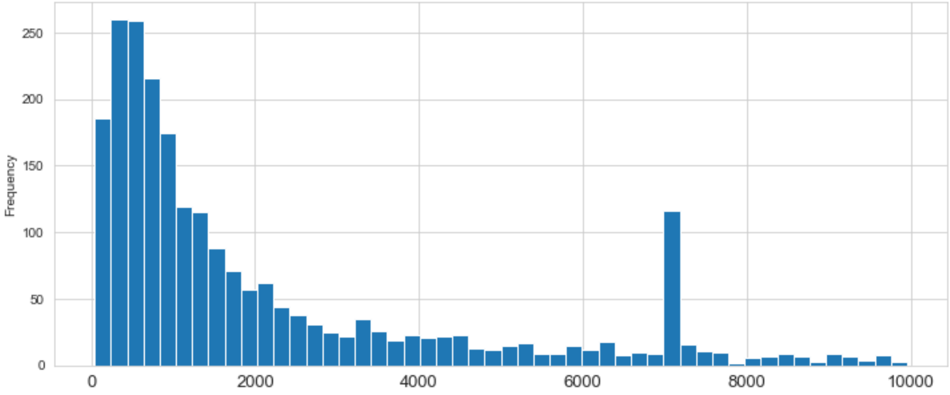


Figure 4. 43 *Dublin words per comments after processing*

In the following figure, we can notice that people’s most frequent words are music, Irish, Ireland, study. Then, we can sort the word counts and plot their values on Logarithmic axes to check the shape of the distribution.

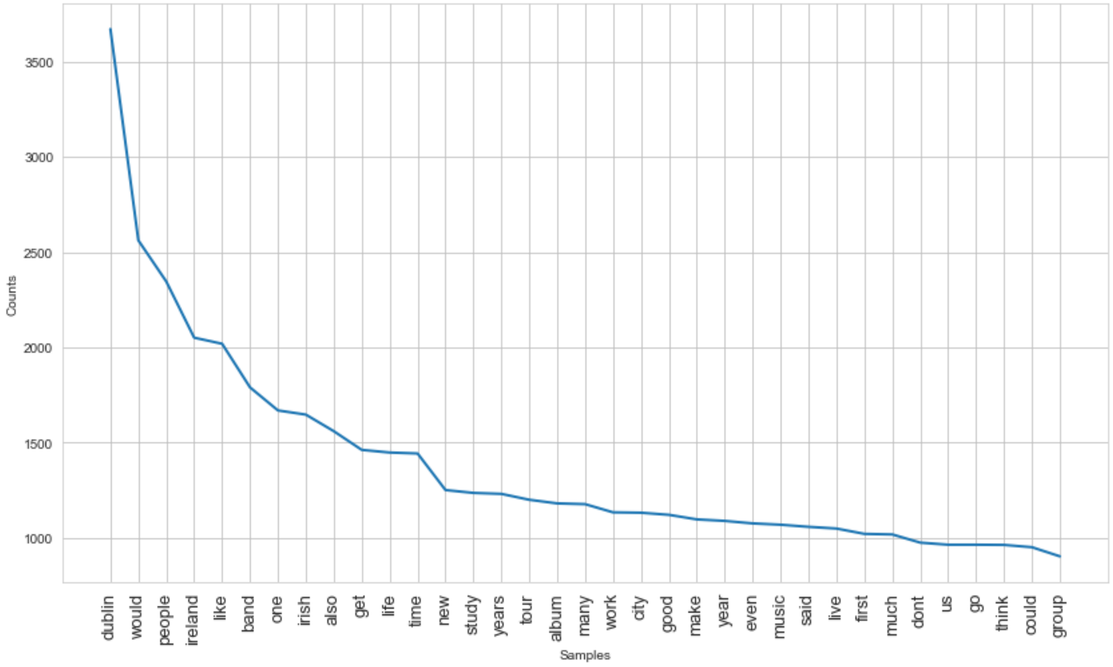


Figure 4. 44 *Dublin 35 most frequent words for all sentiments*

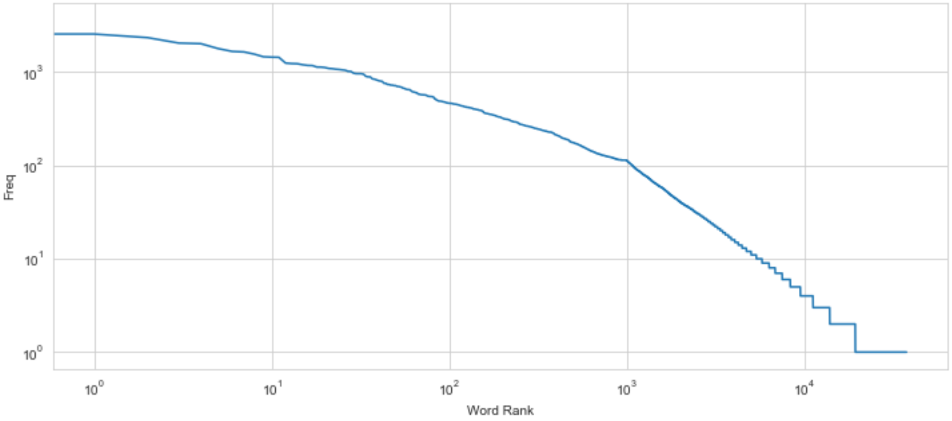


Figure 4. 45 *Dublin word counts distribution on a logarithmic scale*

Then we can visualize with a word cloud.

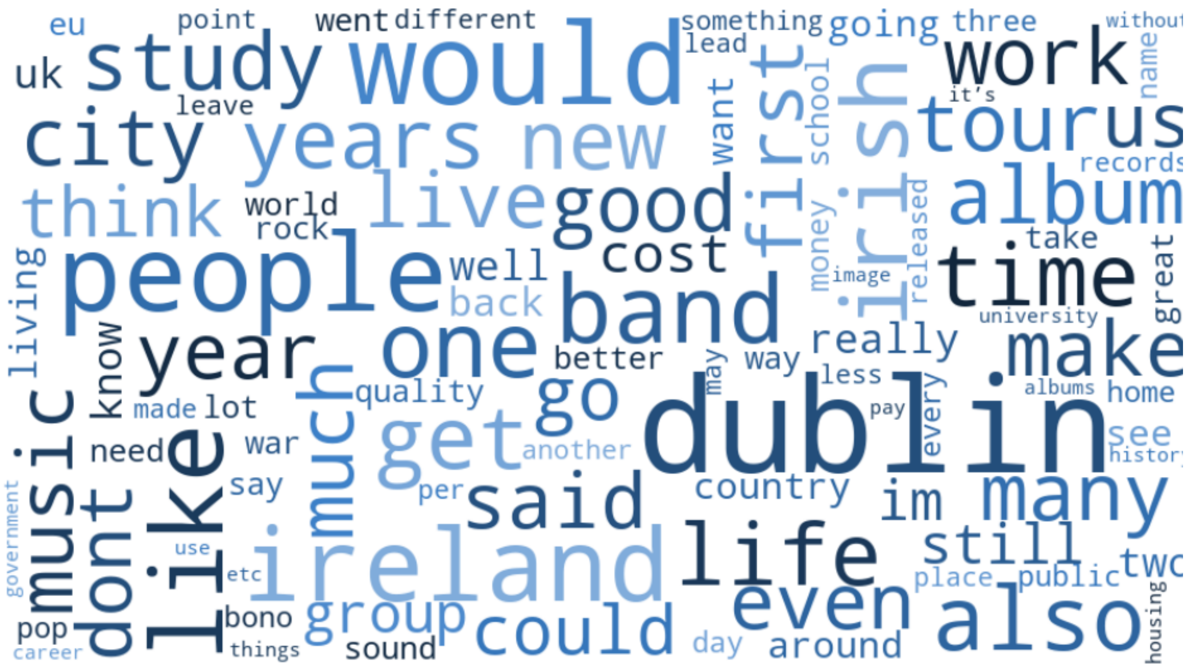


Figure 4. 46 *Dublin word cloud on all sentiments*

4.2.4.2 Sentiment Analysis & Words Clustering

In the next table we have the results of the analysis. 72.55% of the comments collected are positive, 24.63% are negative and 2.82% are neutral as represented in the figure.

Table 4. 9 *Dublin sentiment results*

Negative (-1)	Positive (1)	Neutral (0)	Total
567	1670	65	2302

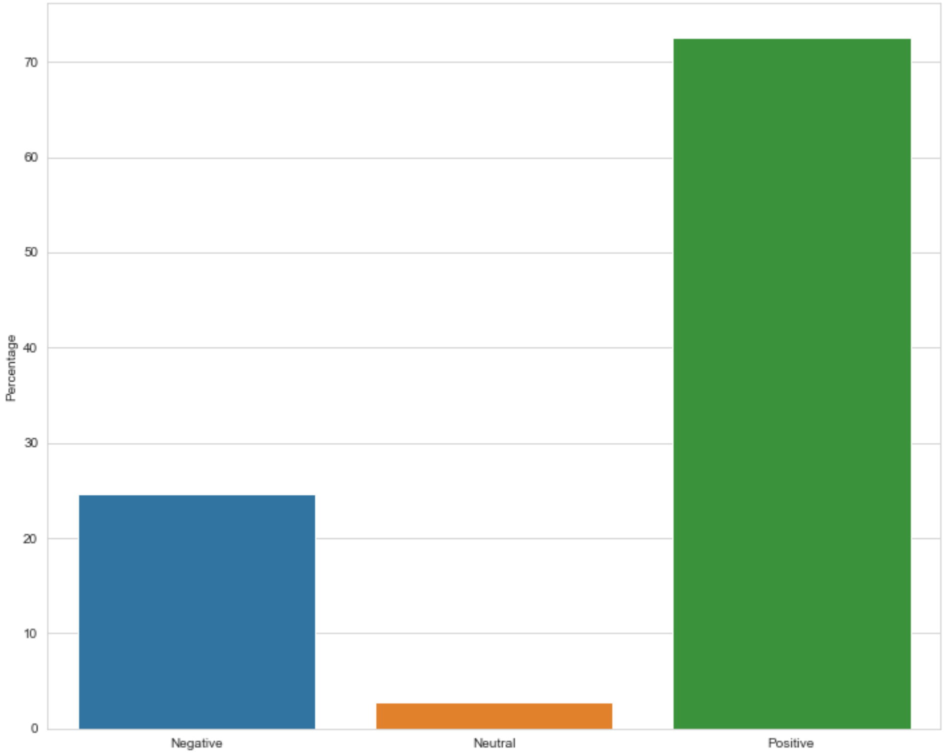


Figure 4. 47 *Dublin sentiment results in percentage (%)*

In the following figure we can visualize the most frequent words and the word cloud for the positive sentiment. We can notice that the words are almost identical as before: *life, work, study,* as well as, *band, music, album*, are repeated, which indicate topics related to Irish music culture.

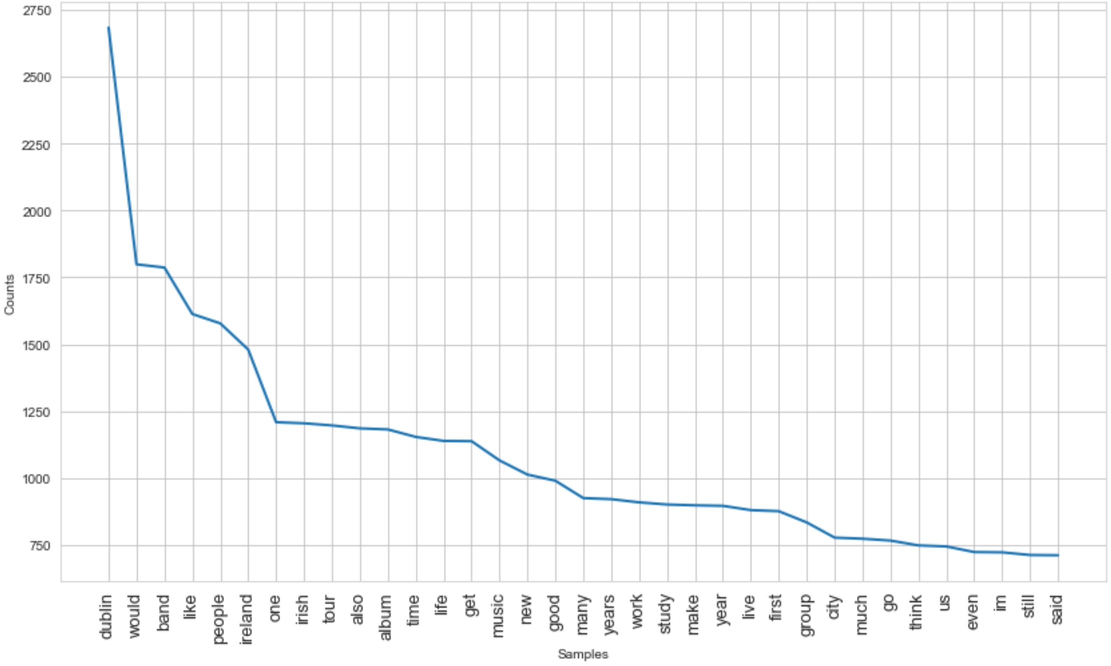


Figure 4. 48 *Dublin 35 most frequent words for positive sentiment*

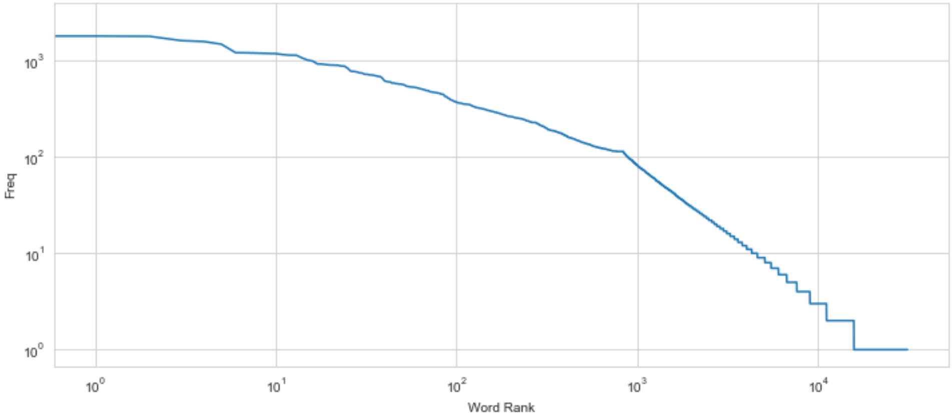


Figure 4. 49 *Dublin word counts distribution for positive sentiment on a logarithmic scale*

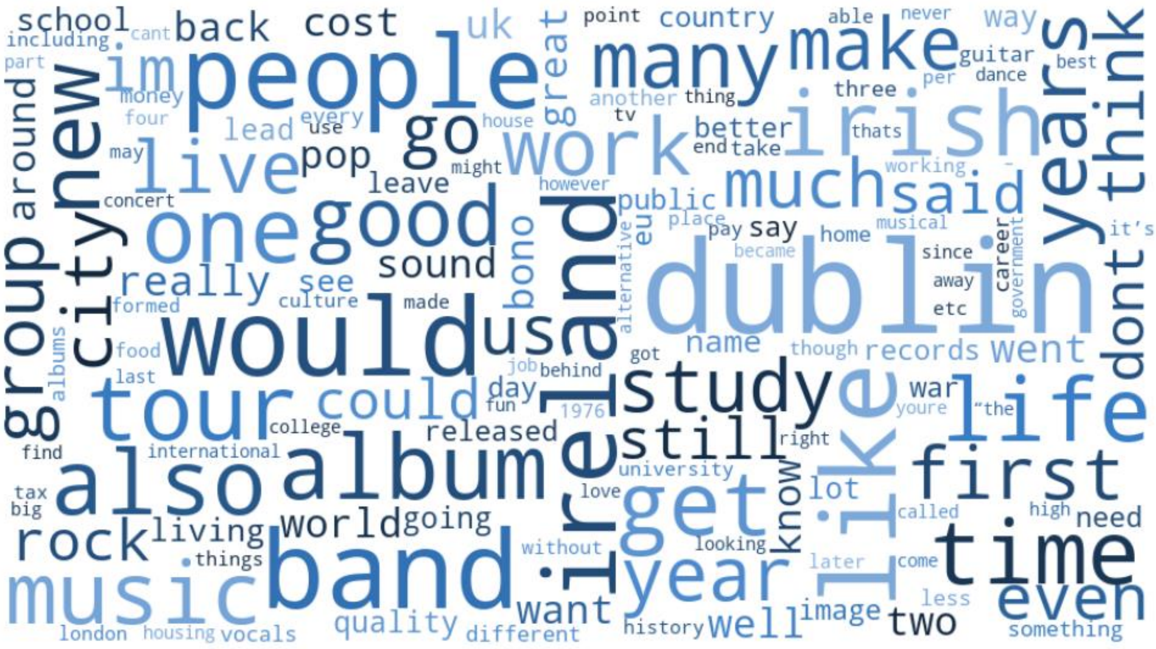


Figure 4. 50 *Dublin word cloud on positive sentiments*

For the next figure we are presenting the same for the negative words. The sentiment analysis tells us that people talked in a negative way about almost the same topics related to the *city, work and life,* we will get more insight on those topics later.

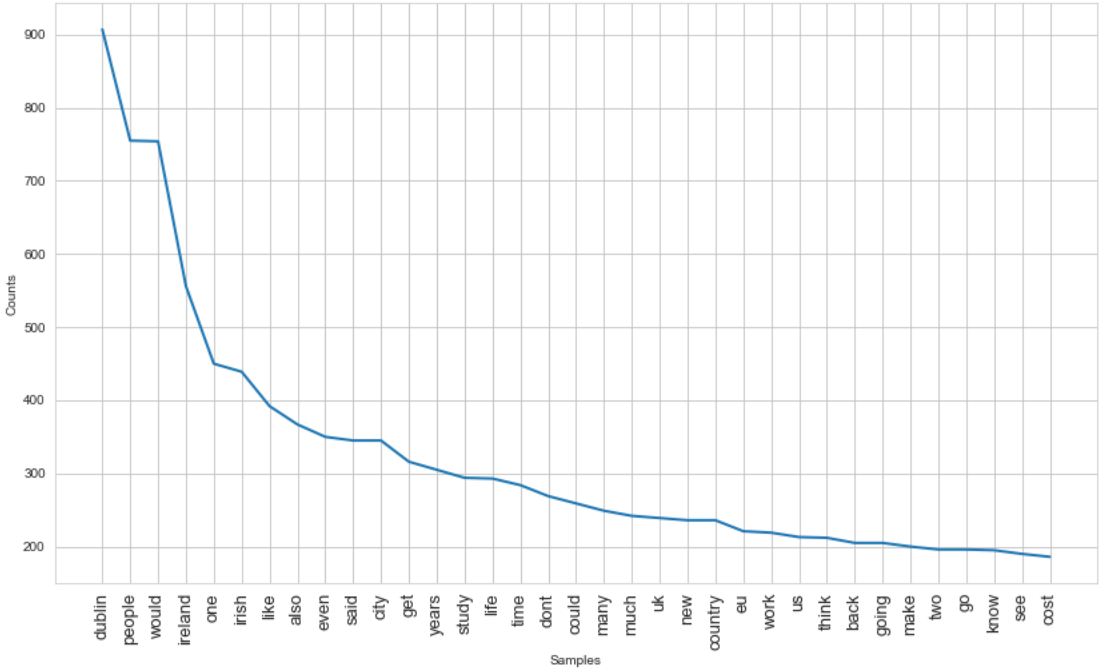


Figure 4. 51 *Dublin 35 most frequent words for negative sentiment*

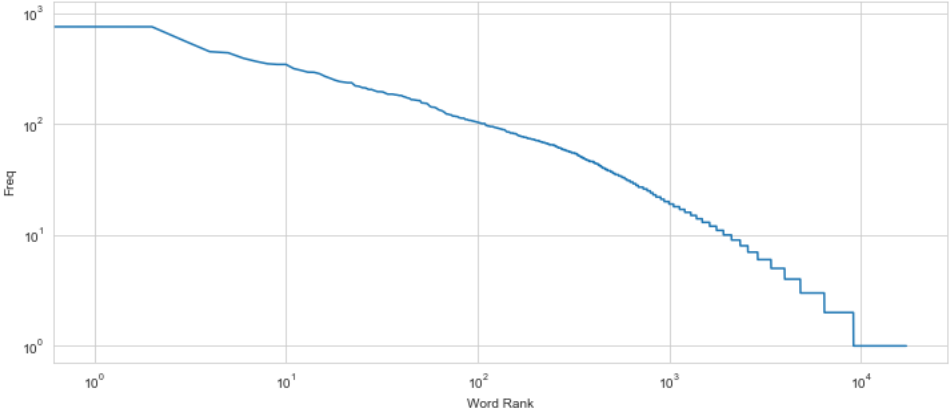


Figure 4. 52 *Dublin word counts distribution for negative sentiment on a logarithmic scale*

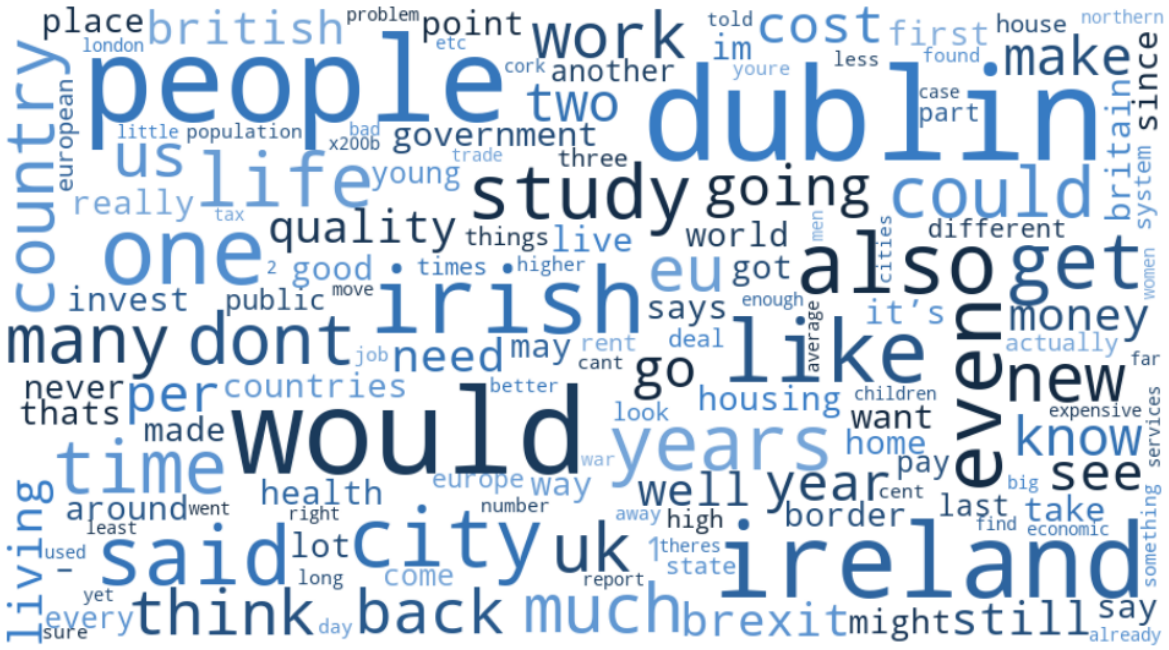


Figure 4. 53 *Dublin word cloud on negative sentiments*

In the next part, word clustering results is carried out.

- Positive Clusters & Discussions

[[[' quality', ' dublin', ' life', ' live', ' city', ' living', ' people', ' better', ' ireland', ' rent'], [' people', ' study', ' dont', ' know', ' want', ' university', ' like', ' english', ' dublin', ' im'], [' band', ' album', ' tour', ' music', ' bono', ' rock', ' group', ' pop', ' sound', ' albums'], [' dublin', ' study', ' time', ' invest', ' like', ' people', ' ireland', ' irish', ' city', ' abroad'], [' eu', ' ireland', ' tax', ' uk', ' brexit', ' ni', ' irish', ' northern', ' government', ' companies']]

- ["For comparison I'm a STEM PhD student in Edinburgh (from Dublin) and I'm living pretty comfortable on a stipend of ~£15000, living in a 1-bed flat that's a short cycling distance to the city center and where I work. Still have enough money left over to go for pints, go to shows, flights home, and put away a bit as savings. I was living even more comfortably in a flat share right in the middle of the city up until recently but I got sick of my flatmates and decided to pay the extra bit to have some space to myself as I finish up. All my friends back in Dublin who are doing PhDs are either living with their parents or getting support from their parents. I have no idea how anyone would survive in Dublin off the stipend alone, which is a symptom that something has gone very wrong. Also, it's very strange to see people in this thread say that PhD students aren't workers. I went back to academia after working in industry and it's much more similar to working than studying. I get that it is both, but it's more like working while studying part-time than vice versa. Like, it's research, you can't study what doesn't exist yet. You have to make it yourself and study what does exist when you can."]

- ["i have the same feeling, lived/worked in dublin for a summer and while it's true there are issues that at the moment make it unrealistic for me to want to move there i definitely would love (in terms of quality of life) settling down in ireland or living there for a longer period. i've led a pretty peripatetic life so far, and that's where i've felt the most welcome by people and regulations alike."]
- ["i emigrated to ireland from the south of france 14 months ago. it's close enough to france that i visit my parents twice a year.   the weather is shit, the food is horrible, public transportation is barren. everything else is great. i am happier here, i have a steady job and opportunities that in france wouldn't be there for me, i feel safe in dublin even if it has issues. not always the case in my hometown.   people here love gambling, hate building houses, there are feral 12 y/o groups in the streets, half of a tesco is dedicated to crisps and their various flavours (with popular hits like buffalo, shrimp flavour and all time classic cheese and onion). their favourite brand of crisp/anthropomorphic potato supreme leader is tayto. they will die for him. it's so big they have an amusement park. a potato trying to get you to eat potatoes sounds weird to me but i nod and go along. i don't like crisps anyway. don't let tayto know.   the other staple of the irish cuisine is the chicken filet (t not silent, it doesn't pretend to be french, a redeeming quality). it's a chicken sandwich. yeah that's it.   irish people i know, namely the gf from up north and her friends are more religious than us. but in the sense of belonging to a community through upbringing, rituals, indoctrination, tribalism, a shared proclivity towards the flute or the drums, 18th century uniforms and sometimes, not often, but sometimes yes indeed a wholly misplaced belief in a superior being.   at least in france everybody more or less agrees that god is dead, life is cruel and senseless, nothing could be worse and everything could be better. here people are happy and believe in the future. somewhat. "]

- Negative Clusters & Discussions

```
 [[' irish', ' ireland', ' eu', ' brexit', ' uk', ' people', ' british', ' britain', ' dublin', ' government'], [' malley', '
dublin', ' study', ' parents', ' people', ' rents', ' years', ' ireland', ' children', ' start'], [' quality', ' life', ' dublin', '
expensive', ' live', ' salary', ' people', ' rent', ' living', ' city'], [' city', ' people', ' dublin', ' bus', ' invest', ' dont', '
like', ' airbnb', ' country', ' cork'], [' im', ' study', ' work', ' dublin', ' students', ' time', ' going', ' year', ' dont', '
people']]
```

- ["think of the cost of living. not just extortionate rent, but insurance, parking, fuel, food, crazy heating and electricity bills, healthcare, it all adds up to the point that even if you would earn quite a bit more than wherever you're living, it's just not worth it. i know people earning 2k euros a month in ireland who live much much worse than people earning minimum wage in eastern europe. we don't worry about freezing in winter, being abused by landlords in a myriad of ways, owning a car and many other things. it's even more about how much you spend than how much you earn that defines the quality of life and i wouldn't move to dublin for anything less than 5k a month ever."]

- ['need to move all the poor and those on long-term social welfare somewhere far out of dublin. leave dublin for those in full time employment/study. ']
- ['ireland does have a lot of tech jobs. its all in dublin though and its really expensive to live there with a decent quality of life. maybe work a couple of years to beef up your experience and salary, and then go there. london and dublin are very hard to afford when you are young, broke and starting out.']

4.2.4.3 Conclusion

Dublin seems to have a lot of positives over the negatives, and it seems that it's all connected, the discussion are concentrated about cost and quality of life in the city, most agree that the quality of life is high, however to live in the city, you would need high salary due to the high cost of life. Dublin in fact, a more expensive place to live than some of the priciest places on the planet, according to a new survey **(The Journal, 2019)**. Dublin stands on 4th place for GDP, which explains that situation, however it also comes late on 9th place for foreign visitors and 5th for the percentage of positive comments. The online discussions seem to be aligned with the data.

4.2.5 London

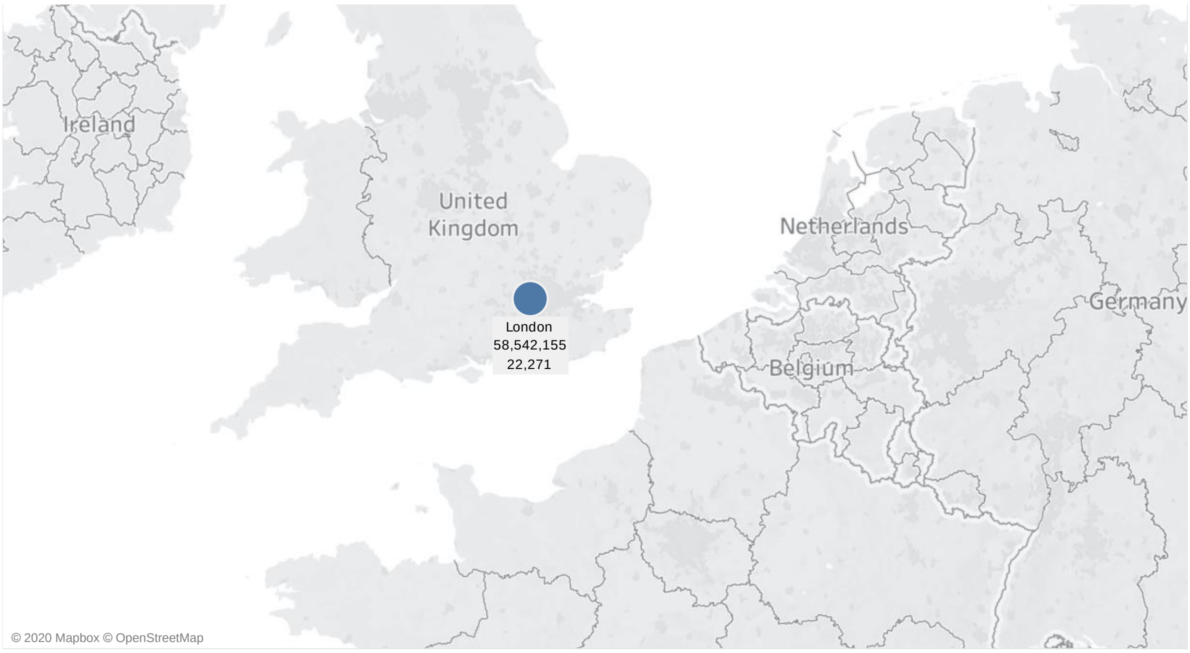


Figure 4. 54 *Geographical representation for the number of the words and comments collected for the city of London*

We managed to collect 24693 comment, by far the highest number for now, with a total of 70,532,807 words, 22271 are unique values, with a total of 58,542,155 words to process and analyze.

4.2.5.1 Text Processing and Visualization

After processing the text, we now have 5,258,601 words to analyze.

In the following table, we can see the difference between the data before processing it and the data after, we are also exploring the frequency of words per each comment. The processed text has a total average of around 235 words per comment, while not processed texts has around 2480 words as average. The longest processed comment has 21002 words, making it the highest so far.

***Table 4. 10** London text descriptive analysis*

Before Processing	After Processing
count 22271.000000	count 22271.000000
mean 2486.101343	mean 235.053612
std 2997.851495	std 281.203007
min 1.000000	min 0.000000
25% 661.000000	25% 63.000000
50% 1417.000000	50% 133.000000
75% 3467.000000	75% 326.000000
max 235376.000000	max 21002.000000

We can see the previous data represented in the following figures

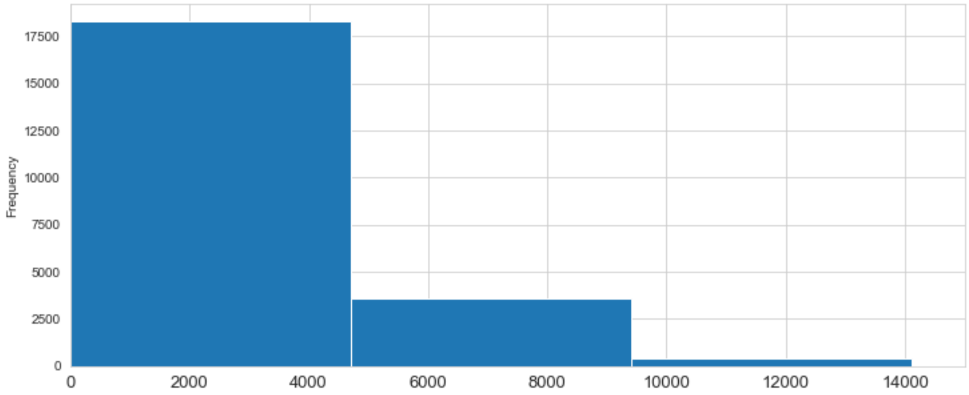


***Figure 4. 55** London words per comments before processing*

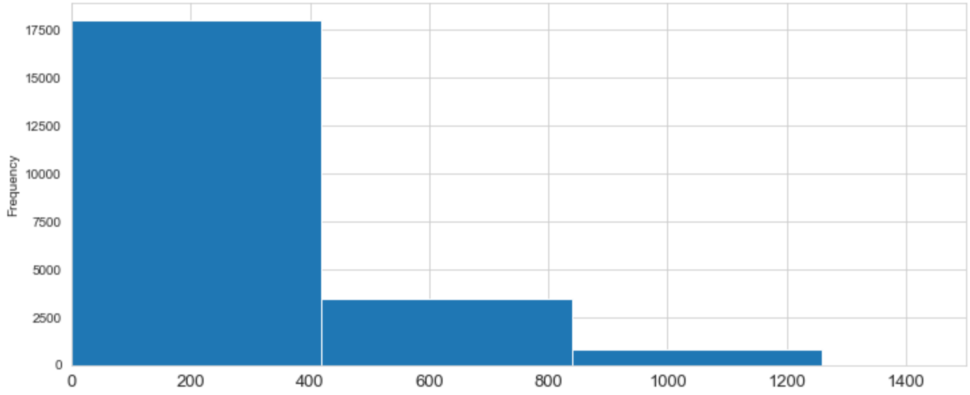


Figure 4. 56 *London words per comments after processing*

In the following figure we can visualize the top 35 most frequent words and the most 150 frequent words, in a word cloud. We can notice that people usually discuss *study, life, UK, work, US* topics.

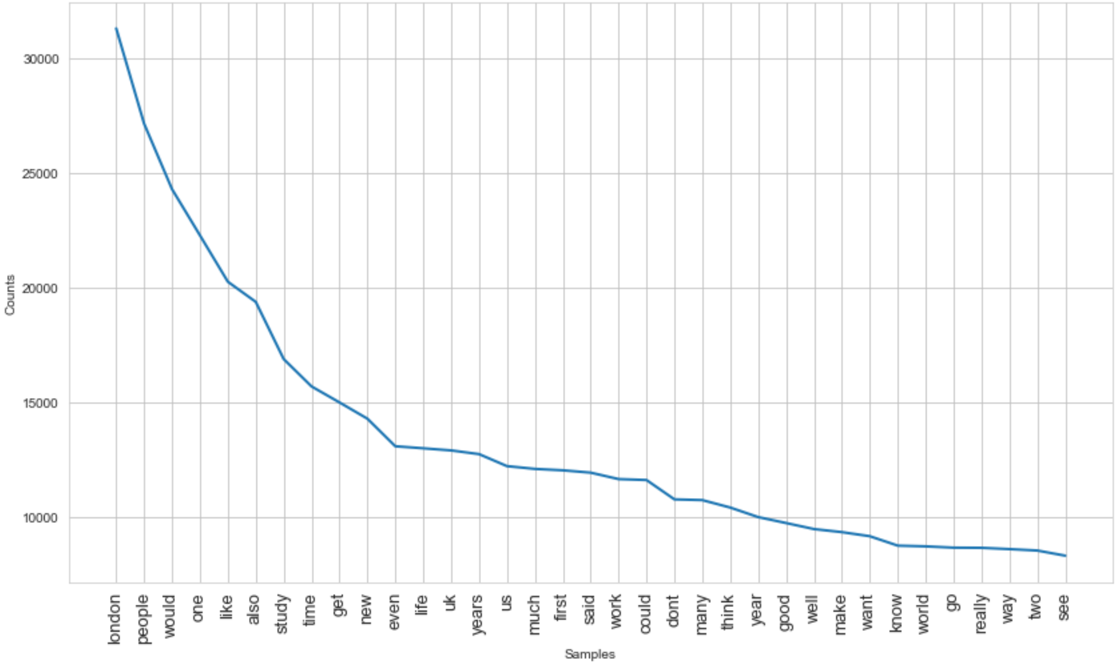


Figure 4. 57 *London 35 most frequent words for all sentiments*

We can sort the word counts and plot their values on Logarithmic axes to check the shape of the distribution.

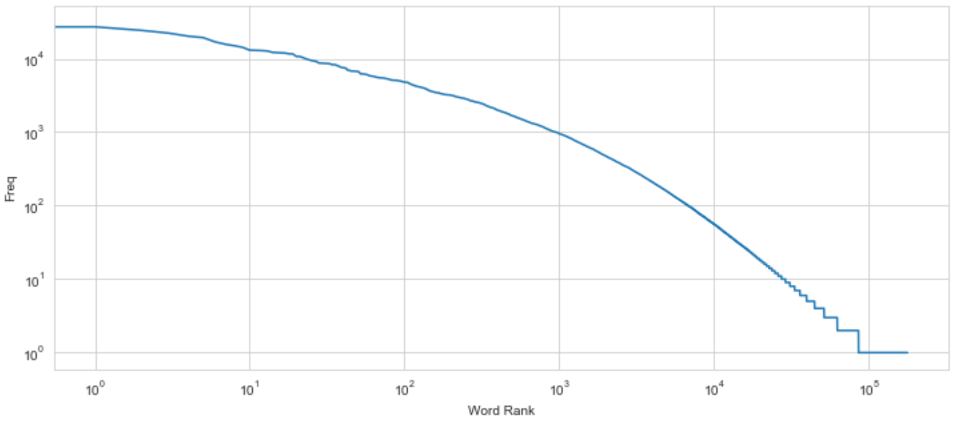


***Figure 4. 58** London word counts distribution on a logarithmic scale*

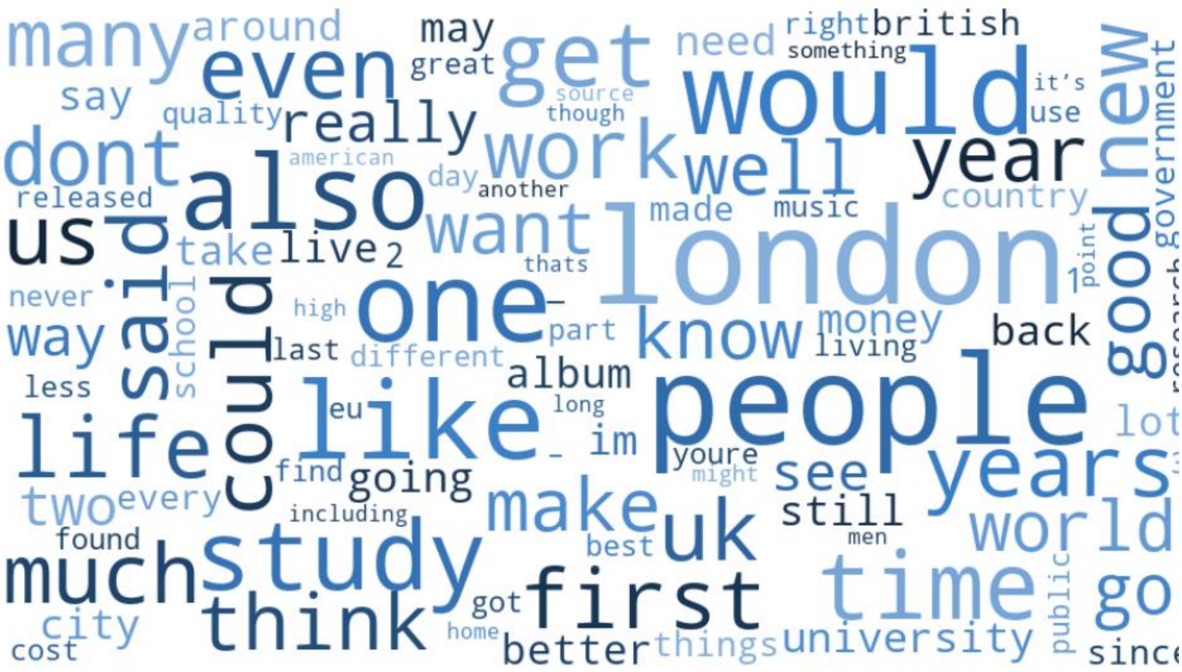


***Figure 4. 59** London word cloud on all sentiments*

- Sentiment Analysis & Words Clustering

In the next table we have the results of the analysis. 67.42% of the comments collected are positive, 30.15% are negative and 2.43% are neutral as represented in the figure.

***Table 4. 11** London sentiment results*

Negative (-1)	Positive (1)	Neutral (0)	Total
6713	15017	541	22271

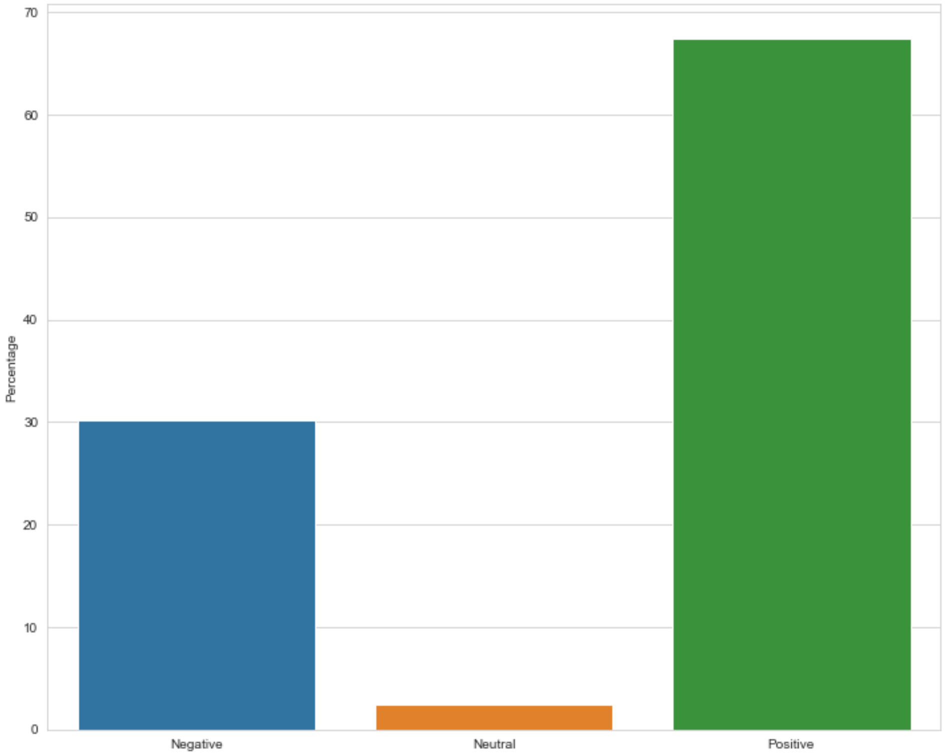


Figure 4. 60 *London sentiment results in percentage (%)*

In the following figure we can visualize the top 35 positive and most frequent words and the most 150 frequent words. We can notice that the words are almost identical due to the 67% of positive sentiment.

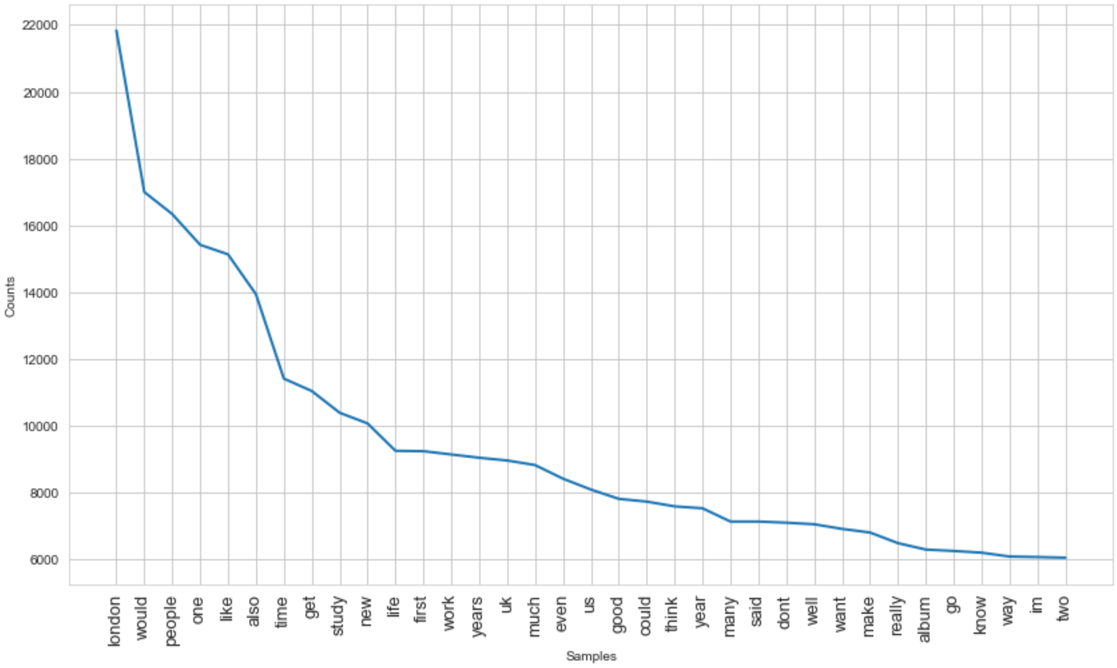


Figure 4. 61 *London 35 most frequent words for positive sentiment*

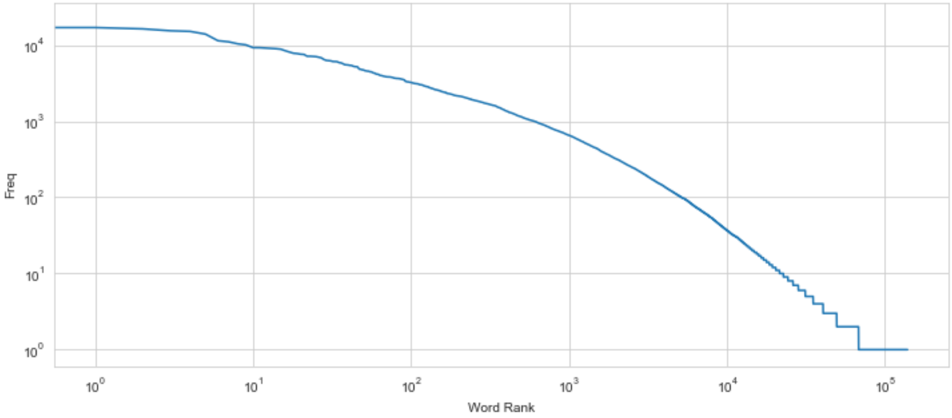


Figure 4. 62 *London word counts distribution for positive sentiment on a logarithmic scale*

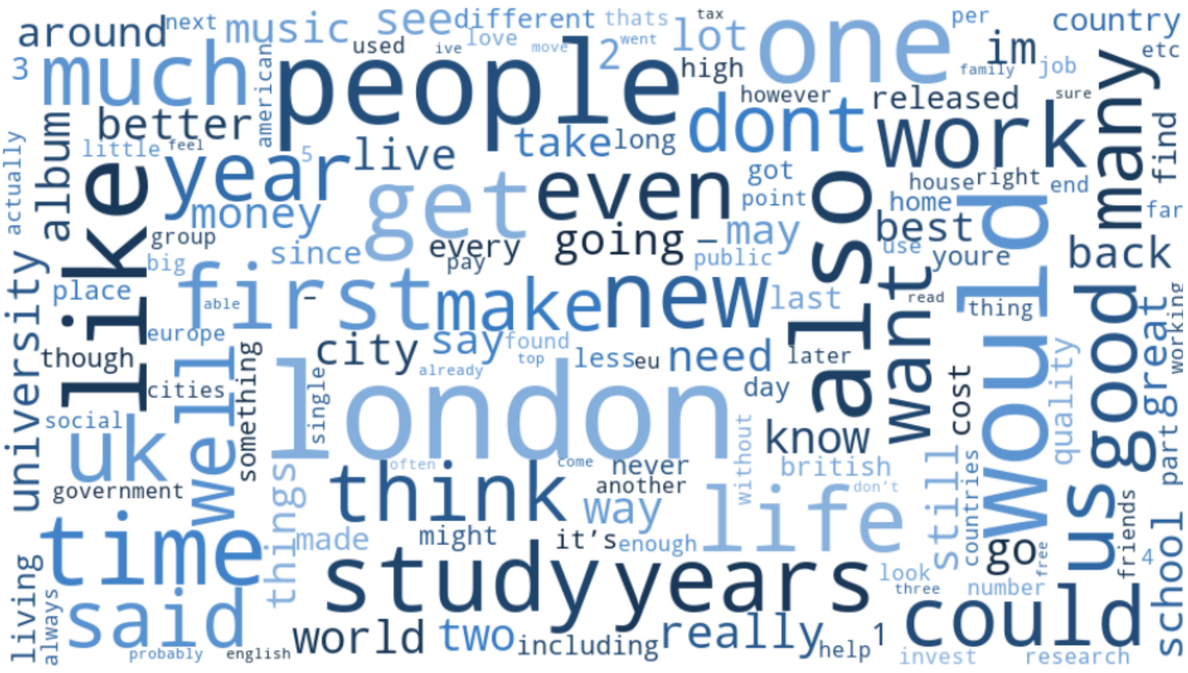


Figure 4. 63 *London word cloud on positive sentiments*

However, for the negative words. The sentiment analysis tells us that people talked in a negative way about topics similar to before such as study, life, UK and government.

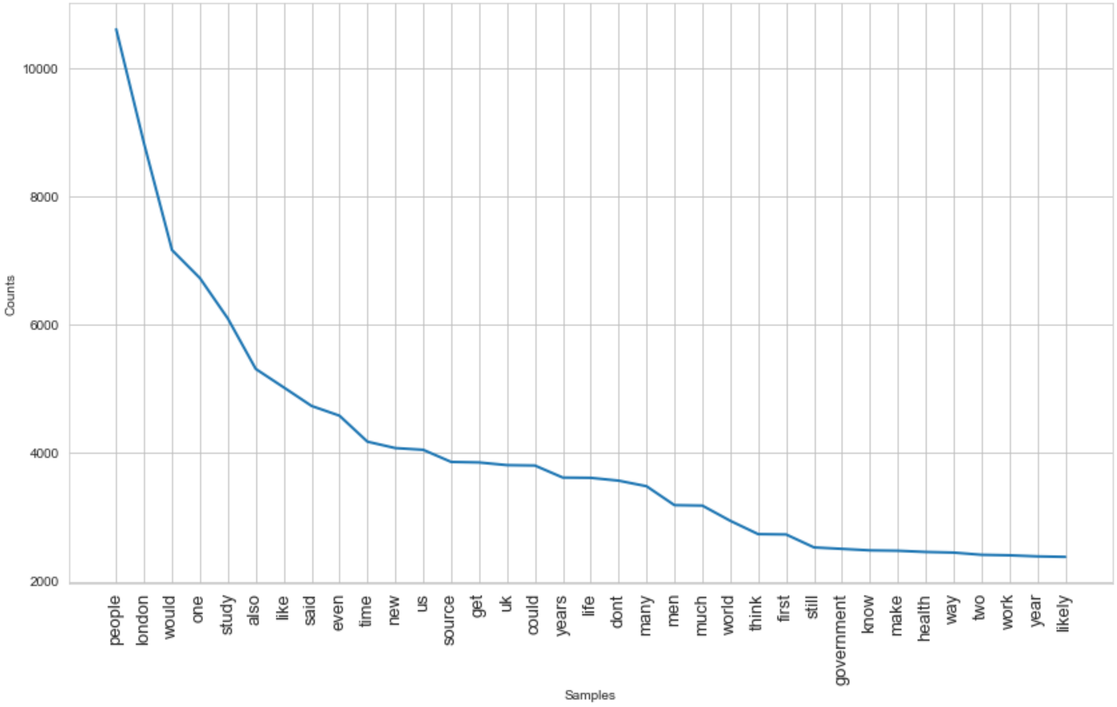


Figure 4. 64 *London 35 most frequent words for negative sentiment*

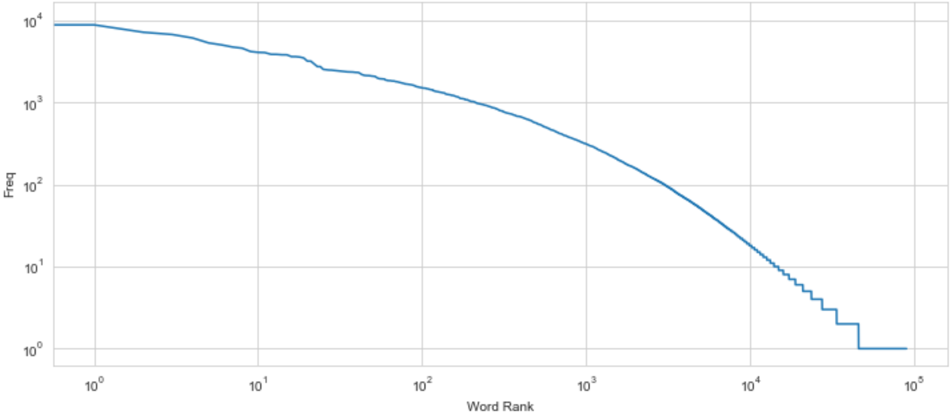


Figure 4. 65 *London word counts distribution for negative sentiment on a logarithmic scale*

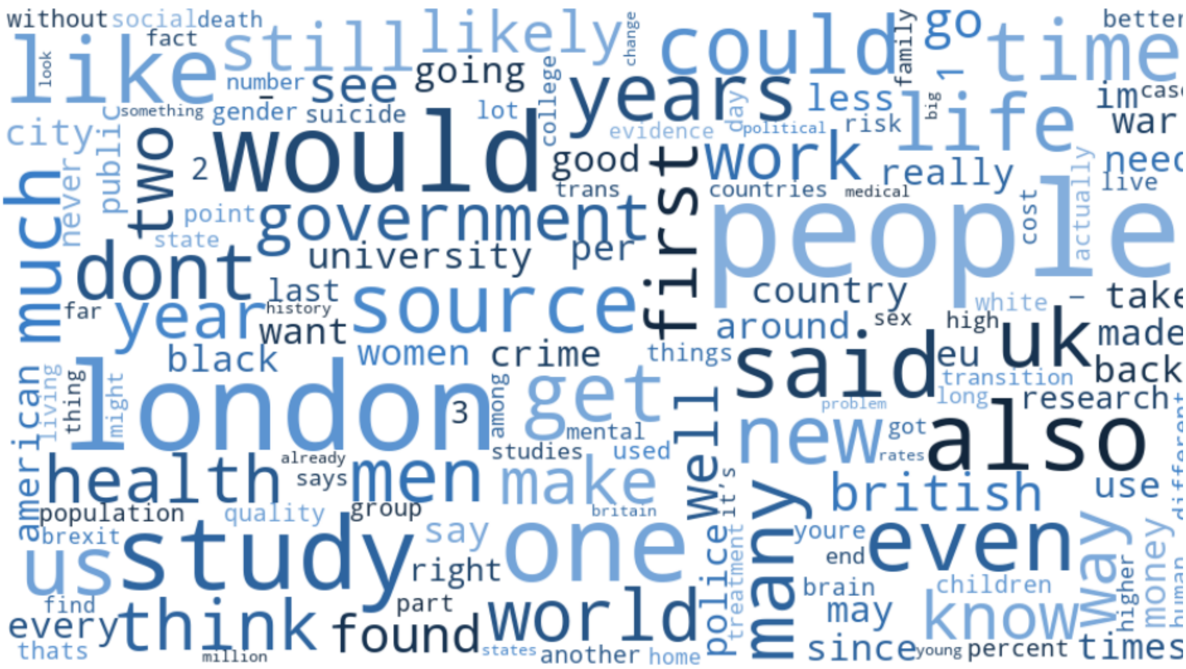


***Figure 4. 66** London word cloud on negative sentiments*

In the next part, word clustering results is carried out.

- Positive Clusters & Discussions

```
[[' study', ' london', ' people', ' university', ' like', ' new', ' brain', ' invest', ' said', ' research'], [' im', ' like', ' study', ' time', ' london', ' people', ' really', ' dont', ' think', ' want'], [' life', ' quality', ' london', ' living', ' live', ' city', ' people', ' cost', ' better', ' like'], [' eu', ' uk', ' people', ' government', ' brexit', ' tax', ' invest', ' money', ' investment', ' labour'], [' album', ' released', ' björk', ' zevon', ' drake', ' simone', ' music', ' band', ' songs', ' hit']]
```

- ["many of my friends come from portugal came to study in england, especially london and they absolutely loved it. besides, my boyfriend is portuguese, he can't see himself coming back to portugal one day.   they publicly said they often watch pl in their childhood because of the language they were taught early at the school"]
- ['if you look out the original press release (scotland is actually the best region after london: london is the best performing region, with nearly seven in 10 (69%) students coming to the capital to study wanting to stay and work there after graduating – more than twice the number of any other region. scotland was the next best performing region with 32% of students planning to stay post-graduation. in england, the north west ranked highest with 28%.   the east and south east were the worst performing, with 12% and 14% planning to stay, respectively – although this is less surprising given

london's proximity (both regions had a disproportionate number of students planning to move to the capital).   more alarming were the figures for the east and west midlands, where only 17% plan to stay.']

- ['if you want to practice law in the uk, namely in london, you should study law in england. you can either study law at the undergraduate or at the graduate level, called a graduate degree in law. the gdl is a one year "conversion" which you undertake after a non-law undergraduate degree. i would advise that if you decide to undertake this route, you study in an english-language degree. in any event, you will be more employable in england if you study in the uk for as long as possible. you should also be aware of the difference between a solicitor and a barrister if you hope to practice in england.']

- Negative Clusters & Discussions

```
[[' people', ' life', ' london', ' uk', ' dont', ' eu', ' quality', ' money', ' like', ' im'], [' gun', ' guns', ' crime', ' police',
' people', ' firearms', ' violent', ' homicide', ' firearm', ' violence'], [' transition', ' gender', ' brain', ' trans', '
treatment', ' suicide', ' percent', ' identity', ' dysphoria', ' mental'], [' study', ' london', ' people', ' said', ' time',
' like', ' women', ' university', ' years', ' new'], [' source', ' band', ' fela', ' men', ' blacks', ' black', ' gay', '
album', ' likely', ' straight']]
```

- ["things that will certainly happen if brexit goes through pound will drop massively when the final exit goes through tariffs and trade obstacles make goods and services more expensive in the uk meaning lower quality of life and purchasing power for the citizens   lots of companies that didn't move yet, dependent on eu trade will just move overseas to either ireland or another eu country altogether resulting in simultaneous loss of tax income by the government, job losses for uk citizens and a shrinkage of gdp   london will lose its place in the top 3 trade hubs to new york, frankfurt and paris brain drain of experts and educated citizens towards the eu   nhs will be put on a massive strain as the costs of medical infrastructure, repairs and medicine will all rise due to most of these being imported from the eu and not developed domestically.   reduced trading and partnerships within the commonwealth as the benefit of trading with the uk is diminished due to their industries being hampered by this move. the commonwealth will have to divert a part of the trading they are doing with the uk right now to other eu members due to their single-market benefit.   general economic decline as uk companies need to look for new suppliers and logistics as the usual eu based ones will likely not be profitable anymore without a single market.   things that are   likely   to happen if brexit goes through   irish unification   rebirth of ira activities   gibraltar trying to join spain as their entire economy is based on being a part of the eu   argentine most likely using the fallout to try and scoop up the falkland islands possibly sparking some skirmishes or perhaps even a real war.   scottish referendum and possibly the secession from the uk altogether   uk islands all over the world would try to find a legal way to claim they are part of scotland and not wales/england's uk, meaning most islands stop being part of the uk as well.   national unrest especially within london as most of these developments will result in resentment from a majority of people that now reject brexit and feel like they are unheard. which could lead to new forms of terrorist groups in addition to the classic ira."]

- ["research has found no evidence of an average causal impact of immigration on crime. one study found no evidence of an average causal impact of immigration on crime in england and wales,. no causal impact and no immigrant differences in the likelihood of being arrested were found for london, which saw large immigration changes and if you read the whole article through and through, you will notice the same pattern over and over. whenever immigrants have a higher crime rate, it's usually because of drugs. adding to that immigrants get harsher sentences than citizens and thus are usually overrepresented in prison. and sure, i'm not sugar coating it, there are some countries that do worse but even there it's usually because immigrants are male and in the age group that in local populations also has the highest crime rate. if you correct for that and socioeconomic status they are getting even in those countries pretty close to native population. adding to that, people who have witnessed war and conflict are more likely to be criminal (ptsd, same as veterans).   it's not the europeans fault either, it's not? are we in bizarro world where colonialism didn't happen and europe is continuing to sell weapons to war torn countries?"]
- [" because public transport doesn't work outside london. people can't get to work unless they have a car.   the government spends money on public transport where lots of people work. if more worked in other areas and less in london more money would be spent in those locations   companies can want lots of things. you can't operate if you have no staff. most people in london moved there for a job, if the job was elsewhere people would have moved there instead, people follow jobs   people don't want to live where the internet is non functional and public transport doesn't work. this message is currently being taken by carrier pigeon to my friend in london to post on reddit because there is no internet outside the m25   given you have based this upon precisely nothing,   basic principles of supply and demand   i'm not sure why i'm responding to this but.. higher taxes with a lower cost of life leading to a higher take home vs living on london is still attractive.   at the moment people aren't moving with the same taxes with a lower cost of life but you think they would for higher taxes with a lower cost of life. i've no idea why you responded either   fix public transport. and tax regionalism will have no impact on this whatsoever"]

4.2.5.2 Conclusion

London is by far the highest number of words and comments collected, London also comes first for population, internet access and e-commerce active, while comes 2nd for GDP and the number of foreign visitors per year, however, London comes in 6th place for the percentage of positive comments online. There are various topics found, from music, to culture to politics and economics and even football, what’s important here to highlight is the heated topics over Brexit and quality of life in London, some discussions are concerned with what will happen after Brexit with companies relocating their HQ to other cities, concerns are also focused on whether the quality of life will be lower than it is now. In fact, Continued uncertainty over Britain's future trading relationship with the EU could cost the UK economy £4.4bn annually by the end of this year and £15bn by the end of the decade, a new report has warned **(The Independent, 2020)**.

Some comments also mentioned the London public transport as it's either overcrowded or over expensive for some and the only way around it is to own a car. Its's important to know that London's monthly travel cost is considered the 'most expensive in world' **(BBC, 2017)**. To tackle these problems policy makers are required to develop a strategy to ensure the public of the upcoming uncertainty of their city's future.

4.2.6 Madrid

Figure 4. 67 *Geographical representation for the number of the words and comments collected for the city of Madrid*

We managed to collect 2285 comment, by far the highest number for now, with a total of 5,200,200 words, 2154 are unique values, with a total of 4,721,438 words to process and analyze.

4.2.6.1 Text Processing and Visualization

After processing the text, we now have 436,066 words to analyze. In the following table, we can see the difference between the data before processing it and the data after, we are also exploring the frequency of words per each comment. The processed text has a total average of around 200 words per comment, while not processed texts has around 2140 words as average. The longest processed comment has 1117 words.

Table 4. 12 *Madrid text descriptive analysis*

Before Processing	After Processing
count 2154	count 2154
mean 2140.982823	mean 201.855617
std 2323.068638	std 219.181474
min 30	min 3
25% 554.5	25% 52
50% 1218.5	50% 114
75% 2788.2	75% 263
max 10000	max 1117

We can see the previous data represented in the following figures

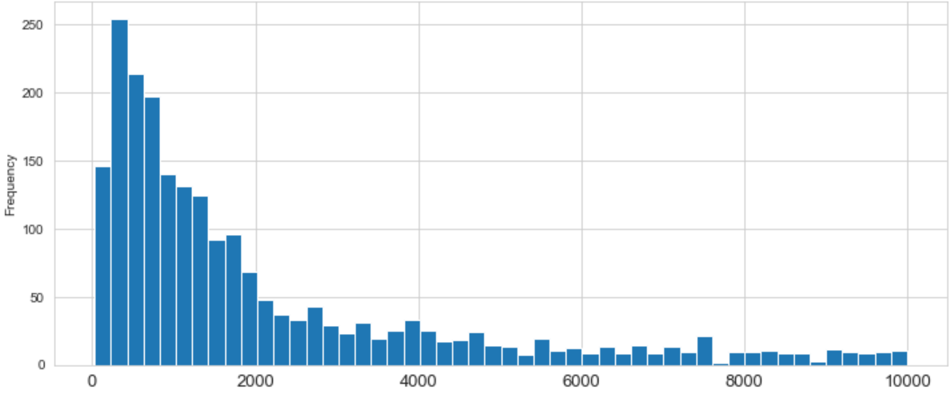


Figure 4. 68 *Madrid words per comments before processing*

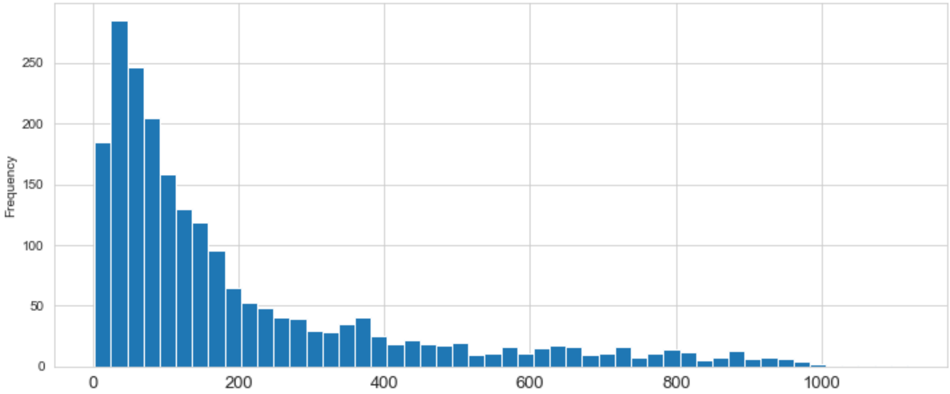


Figure 4. 69 *Madrid words per comments after processing*

In the following figure we can visualize the top 35 most frequent words and the most 150 frequent words, we can notice that people usually discuss the *Madrid club, team, players, study and city* topics.

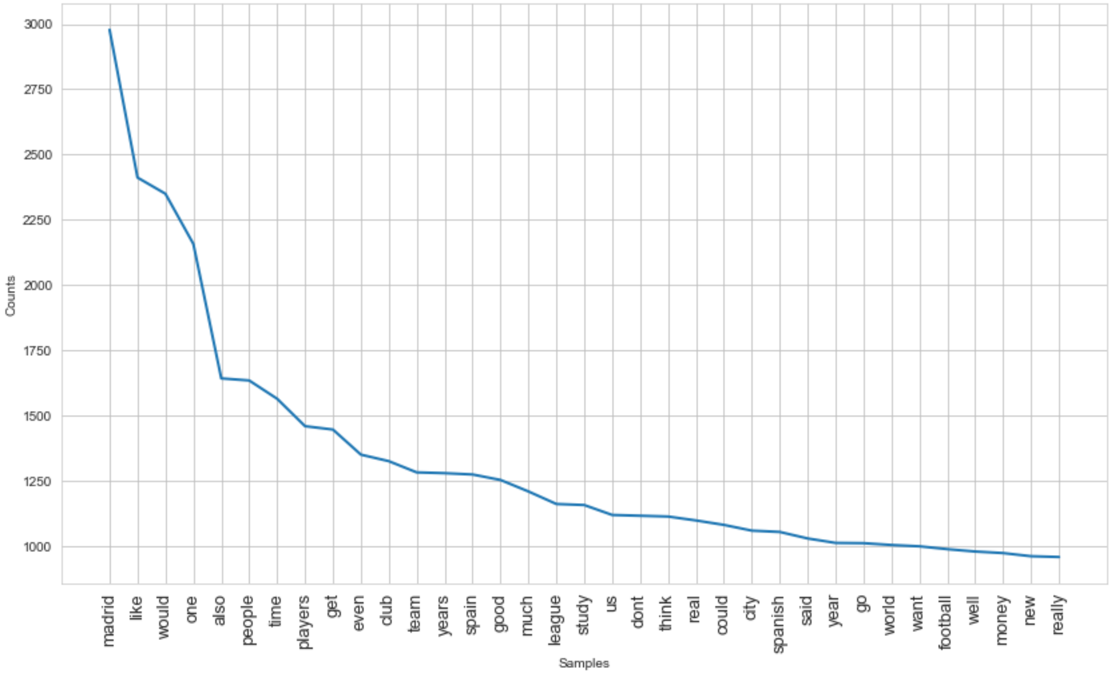


Figure 4. 70 *Madrid 35 most frequent words for all sentiments*

We can sort the word counts and plot their values on Logarithmic axes to check the shape of the distribution.

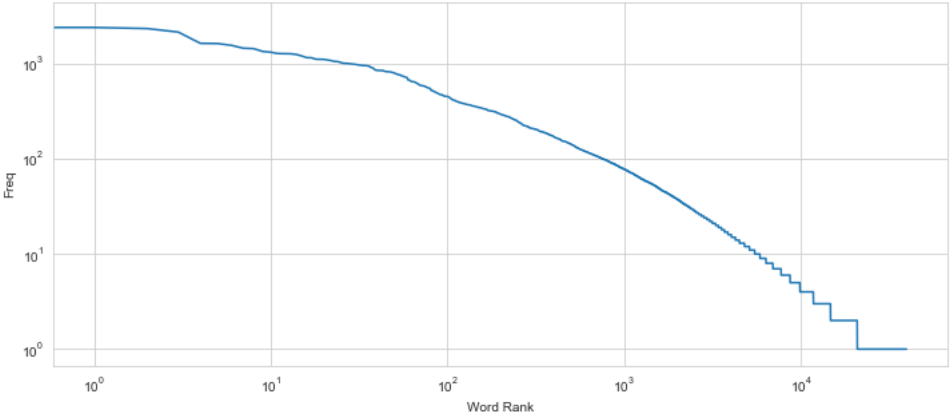


Figure 4. 71 *Madrid word counts distribution on a logarithmic scale*

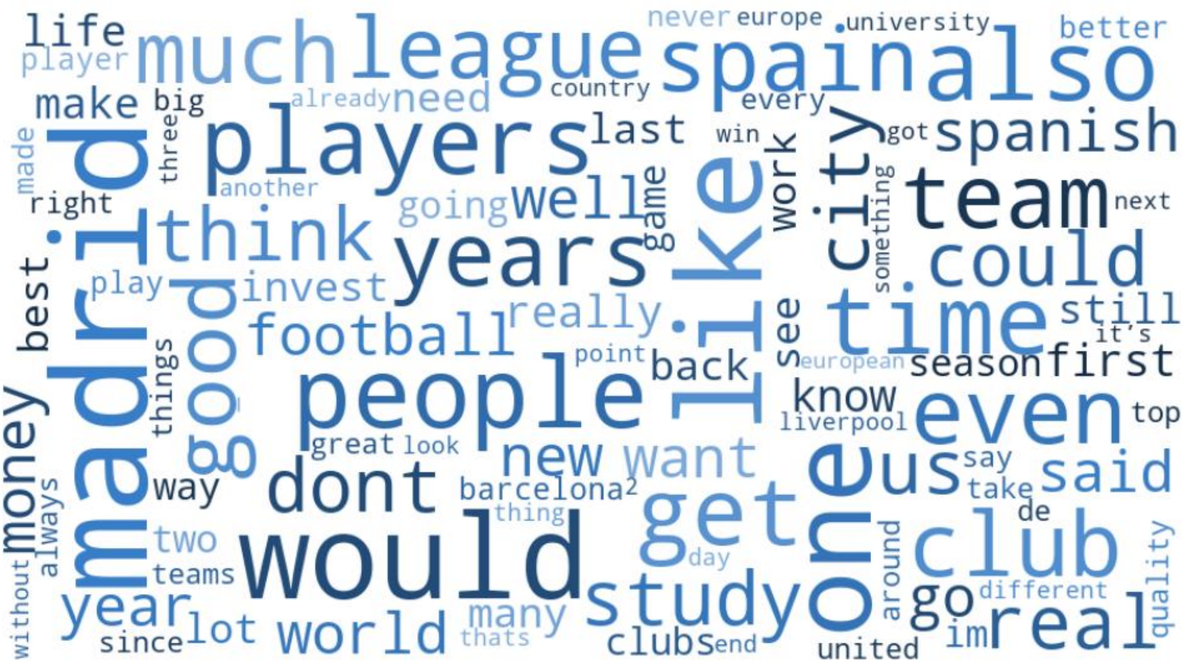


Figure 4. 72 *Madrid word cloud on all sentiments*

4.2.6.2 Sentiment Analysis & Words Clustering

In the next table, 77.4% of the comments collected are positive, 19.86% are negative and 2.74% are neutral as represented in the figure.

Table 4. 13 *Madrid sentiment results*

Negative (-1)	Positive (1)	Neutral (0)	Total
428	1667	59	2154

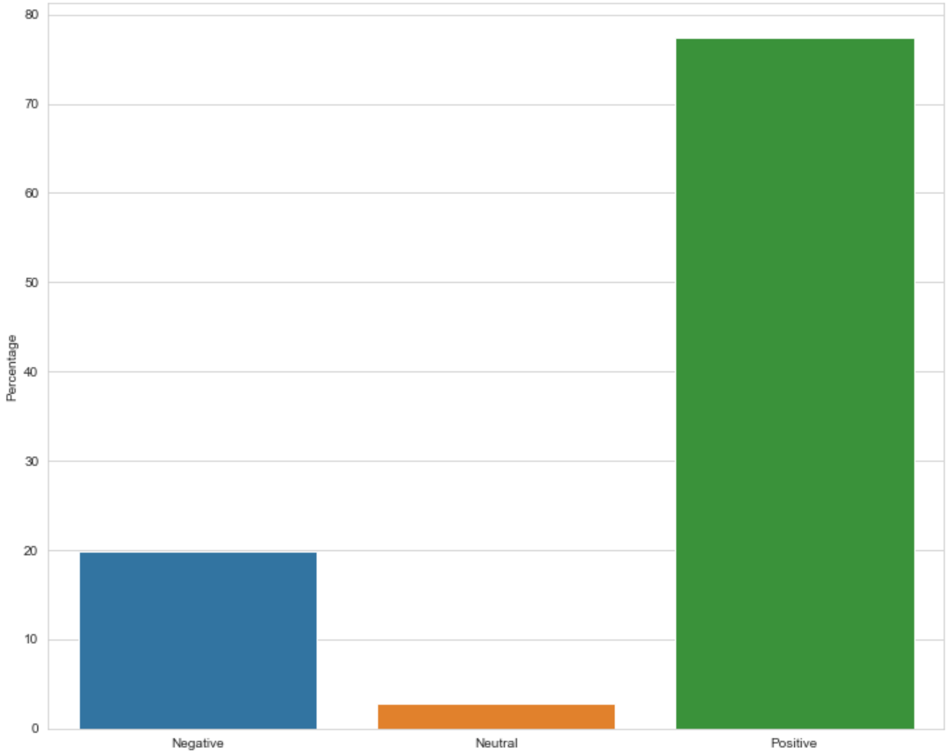


***Figure 4. 73** Madrid sentiment results in percentage (%)*

In the following figure, we can notice a slight change in the words comparing to the one from before as the percentage of the positive sentiment are 77%.

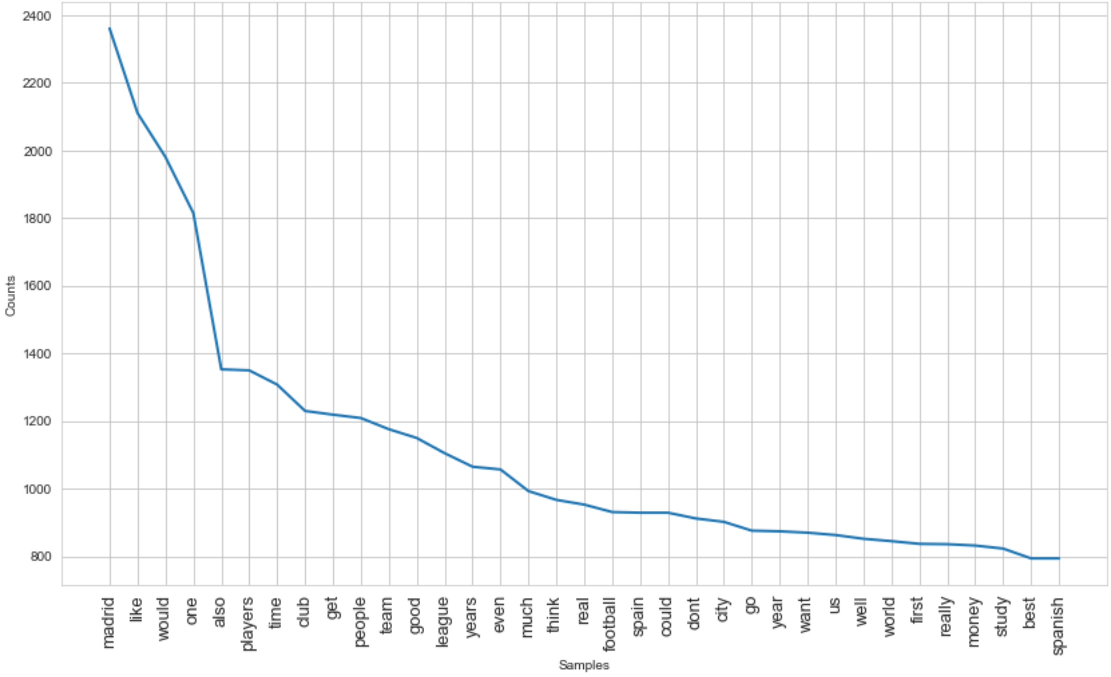


***Figure 4. 74** Madrid 35 most frequent words for positive sentiment*

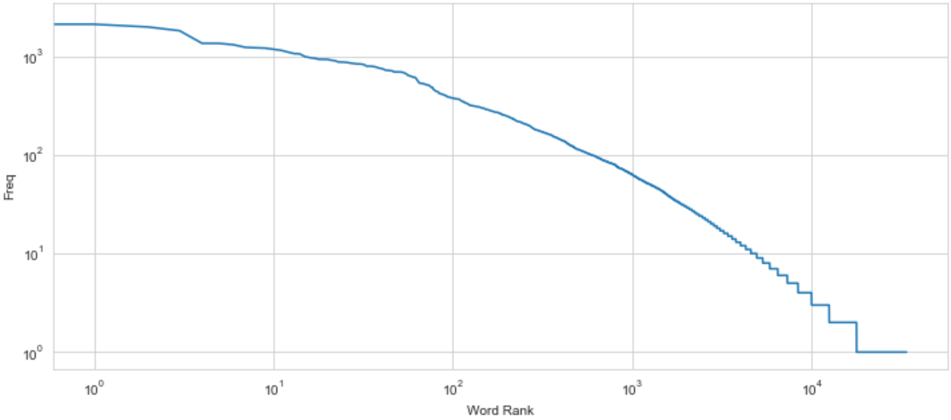


Figure 4. 75 *Madrid word counts distribution for positive sentiment on a logarithmic scale*

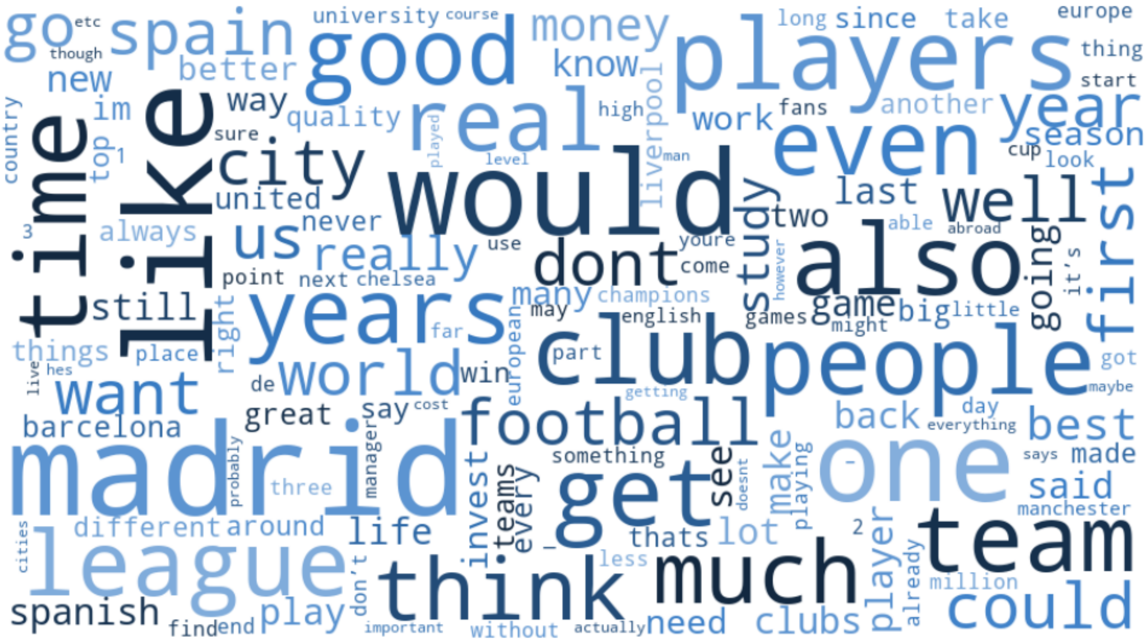


Figure 4. 76 *Madrid word cloud on positive sentiments*

For the next figure we are presenting the negative words. The sentiment analysis tells us that people talked in a negative way about topics related to the study, football team, invest, and the government.

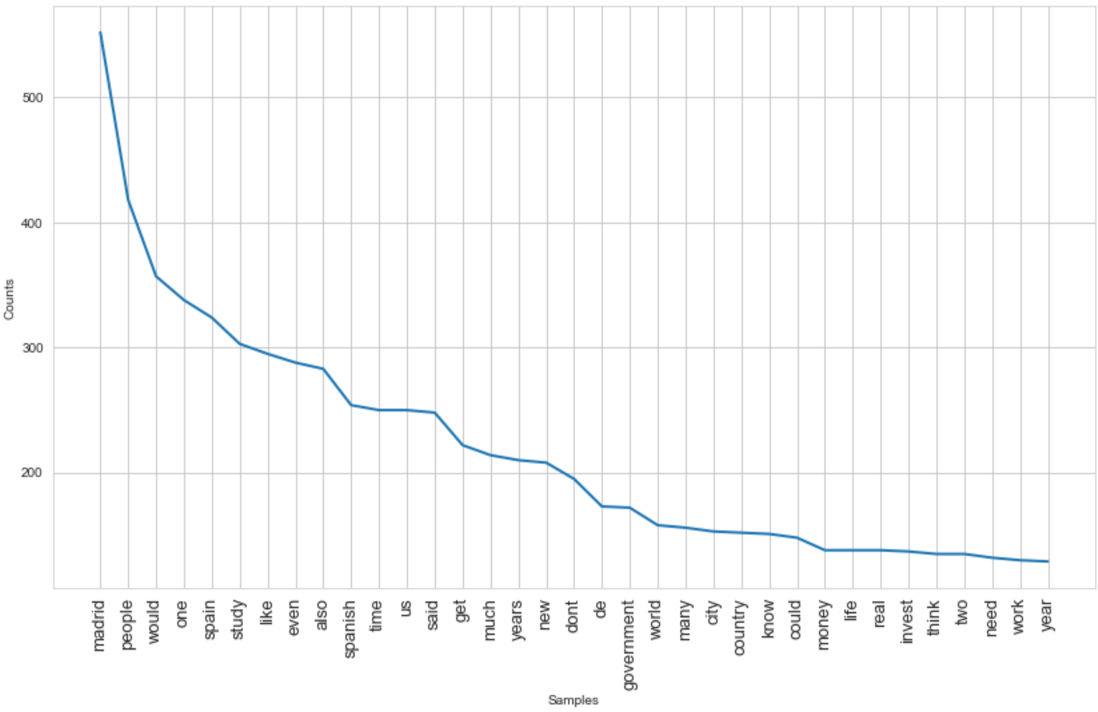


Figure 4. 77 *Madrid 35 most frequent words for negative sentiment*

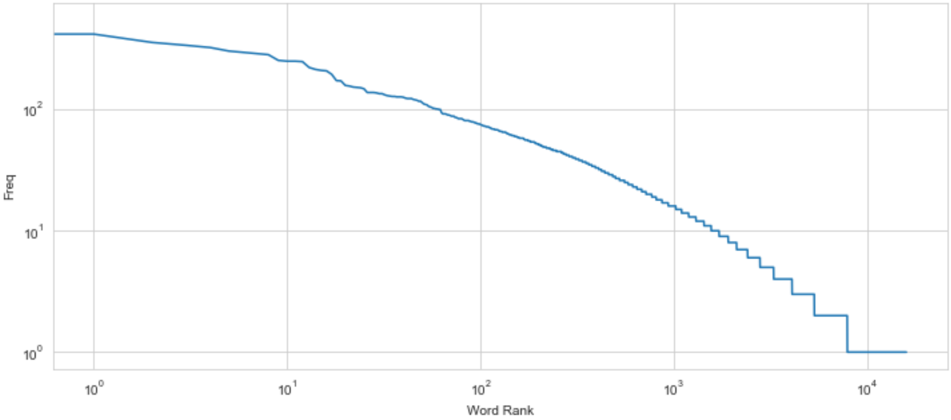


Figure 4. 78 *Madrid word counts distribution for negative sentiment on a logarithmic scale*

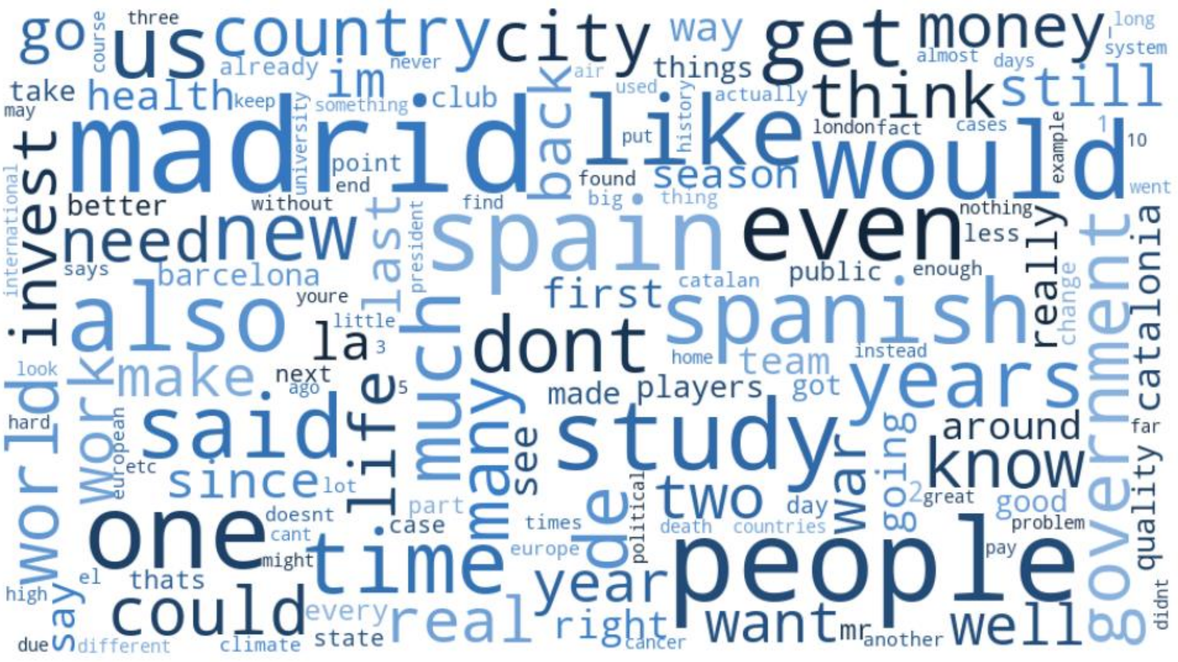


Figure 4. 79 *Madrid word cloud on negative sentiments*

In the next part, word clustering results is carried out.

- Positive Clusters & Discussions

```
 [[' madrid', ' invest', ' real', ' study', ' like', ' new', ' climate', ' time', ' people', ' world'], [' club', ' league', '
clubs', ' football', ' team', ' liverpool', ' teams', ' money', ' players', ' united'], [' abroad', ' study', ' program', '
school', ' spanish', ' madrid', ' university', ' college', ' students', ' spain'], [' spain', ' people', ' life', ' spanish', '
like', ' madrid', ' cities', ' live', ' city', ' barcelona'], [' players', ' player', ' team', ' invest', ' like', ' play', ' hes', '
think', ' real', ' madrid']]
```

- ["Hey I lived in Madrid for a while too! People are right in saying nothing will compare. I love Chicago and have many friends there (I'm currently in Minneapolis), nothing about it is like any European city so I'm not sure if it could really replace Madrid for you. I really miss the quality of life in Madrid (minus the smog), such a great city."]
- ["Life in Madrid was great. I lived in the city center, and was never really struggling for money. Madrid has some really cool areas to drink and eat in, and the food and weather are great. There is a large community of expat teachers, and it is easy to make friends with the locals . A lot of my really good friends now are actually people I taught during my time in Spain. However, I have heard that salaries have been dropping over the last couple of years, and that the cost of living in Madrid centre is rising quickly. "]

- ["I've spent a fair bit of time in Madrid as I just finished up a year abroad in Spain. I didn't study in Madrid but was generally satisfied with my experience on the metro when I visited. The link to Barajas is very convenient too. Thanks for the explanation on your experience as a commuter. Are you Spanish btw? I'm looking to return to Spain to teach English, not sure how the real job market is but I adore the country."]
- ['That was just what I was about to say. Definetly the best public university for social sciences. Also, it has very new and well kept facilities (as it was founded in 1989) and it has the youngest teachers among all public universities. I live in Madrid -where the university is- and study in the UC3M, and I have to say -although this may be a relatively biased opinion- that the Carlos III University is considered to be the best public university here in the capital, especially in degrees like economics, law, international studies and other degrees along those lines.']

- Negative Clusters & Discussions

```
[[' quality', ' life', ' madrid', ' dont', ' people', ' like', ' said', ' think', ' season', ' bad'], [' spain', ' spanish', '
people', ' study', ' government', ' catalan', ' catalonia', ' time', ' madrid', ' war'], [' people', ' berlin', ' high', '
study', ' cancer', ' light', ' world', ' city', ' population', ' madrid'], [' invest', ' players', ' money', ' madrid', '
real', ' club', ' season', ' team', ' dont', ' year'], [' cannabis', ' samples', ' study', ' rome', ' contaminated', '
madrid', ' sold', ' dangerous', ' health', ' human']]
```

- ['Cannabis resin sold on the streets of Madrid is contaminated with dangerous levels of faecal matter, a study says. Traces of E.coli bacteria and the Aspergillus fungus were found by analysts who examined 90 samples bought in and around the Spanish capital. The samples of hashish were wrapped up in plastic "acorns" were the worst offenders, reportedly because of the way they are smuggled into the country. Some 40% of these also had the aroma of faeces, the study\'s lead author said. Buying, selling and importing cannabis is against the law in Spain, as is using it in public - although it is technically legal to grow it for personal use, provided it is not publicly visible, and to consume it in private. How was the study carried out? José Manuel Moreno Pérez, a pharmacologist from the Universidad Complutense in Madrid, collected hashish samples (also referred to as hash or resin) directly from street dealers, both in the city and the surrounding suburbs. The aim was to determine whether the drugs sold were suitable for human consumption. His research team then separated the contaminated samples by shape, with some of them resembling "acorns" and others "ingots", to see if one shape had more contaminates than the other. Wasn\'t even weed but hashish.']
- ["I'm white, but I'm also Venezuelan. I was little when I moved to Spain and have lived here since (my family is from here). You just can't imagine how differently people treat me when they learn I was

born in Venezuela. I'm writing this not to pretend I get the same discrimination other people do, but it may help if there are some instances when people might suddenly discover where your characters are from from or if they only talk with someone through the phone and then meet, etc. I'm pretty pale, and have no accent, so people only learn that I'm Venezuelan when I tell them or when they see my papers. For example, some years ago I was trying to find more about some uni courses in X University. The process goes and there comes the moment: I fill in my info and write I was born in Caracas. Well, the dude attending me changed mood radically. He spent ten minutes trying to convince me that I'm not Spanish (?) He was so, so mad. I was supposed to get an email from him to continue the process. ']

- 'I\'m sorry if I sound rude, but you sound to me a bit naive about Spain and Madrid to make such a big investment there for just 4 to 5 years. What do you do for a living? It is hard to find a job in Madrid, and harder to get paid about 1000 €/ month. It is almost impossible to get a job and study at the same time. The educative system is not friendly to workers, and the work environment is not friendly to students either. You go there with no job, as a student and want to ask for a mortgage? They won\'t give it to you. If you have the whole money for a house you can buy it and pay 100% for it, but no bank will lend you money. About pet friendly rents... I have 4 cats. Most places dont want pets but some big housing companies have a no pet policy in their contract but won\'t have a problem if you bring pets as long as they don\'t cause any problem. I was with banco Santander, the whole building belong to them and it was "no pets", most people had pets and there was never a problem. I wish you the best in Madrid. A very nice place to live in. But please go pay a visit before and start the job hunting while in your country. You will see its not as nice ad you imagine it.'

4.2.6.3 Conclusion

Madrid online discussion results is mainly filled with topics related to football due to being home to the greatest football club in history, Real Madrid, and for that reason, more in depth analysis is needed for cities that has their name attached to popular events or clubs. Madrid had a very good percentage of 77% of positive sentiment coming 2nd after Barcelona and it had very good comments on their level of quality of life and culture. Some of the negative sentiments' were based on discrimination against non-Spanish nationals and the difficulty of finding jobs as a student. Madrid however comes 7th for the number of foreign visitors, 3rd for population and internet access, 4th for e-commerce active, 5th of GDP.

4.2.7 Milan

Figure 4. 80 *Geographical representation for the number of the words and comments collected for the city of Milan*

We managed to collect 1495 comment, by far the highest number for now, with a total of 3,717,959 words, 1423 are unique values, with a total of 3,414,240 words to process and analyze.

4.2.7.1 Text Processing and Visualization

After processing the text, we now have 313,400 words to analyze. In the following table, we can see the difference between the data before processing it and the data after, we are also exploring the frequency of words per each comment. The processed text has a total average of around 220 words per comment, while not processed texts has around 2300 words as average. The longest processed comment has 1343 words.

Table 4. 14 *Milan text descriptive analysis*

Before Processing	After Processing
count 1423.000000	count 1423.000000
mean 2331.779339	mean 219.435699
std 2399.341321	std 224.517736
min 33.000000	min 3.000000
25% 617.000000	25% 59.000000
50% 1358.000000	50% 127.000000
75% 3160.500000	75% 301.500000
max 9992.000000	max 1343.000000

We can see the previous data represented in the following figures

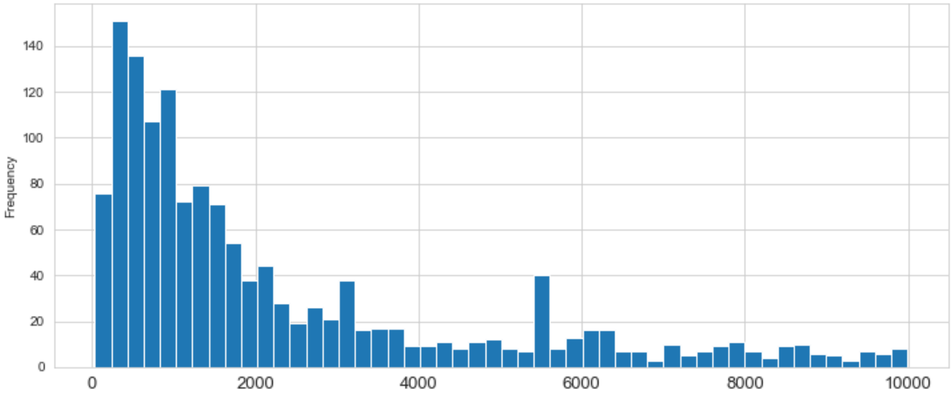


Figure 4. 81 *Milan words per comments before processing*

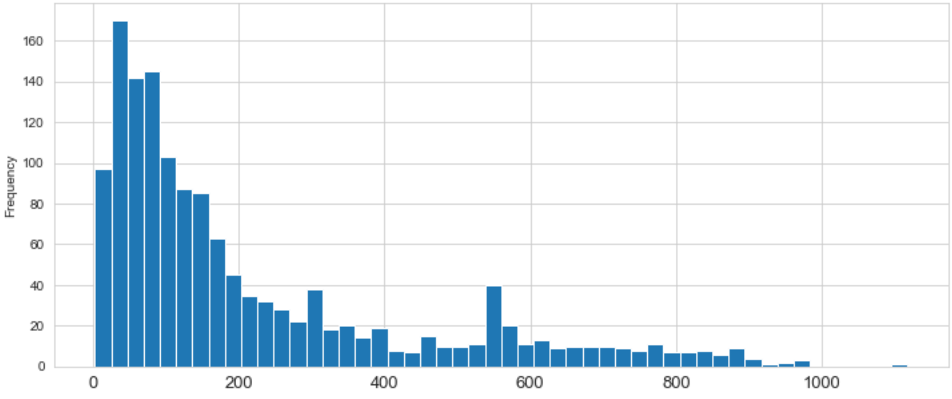


Figure 4. 82 *Milan words per comments before processing*

In the following figure we can visualize the top 35 most frequent words and the most 150 frequent words, we can notice that people usually discuss *Italy, study, Milan clubs, city, and life* topics.

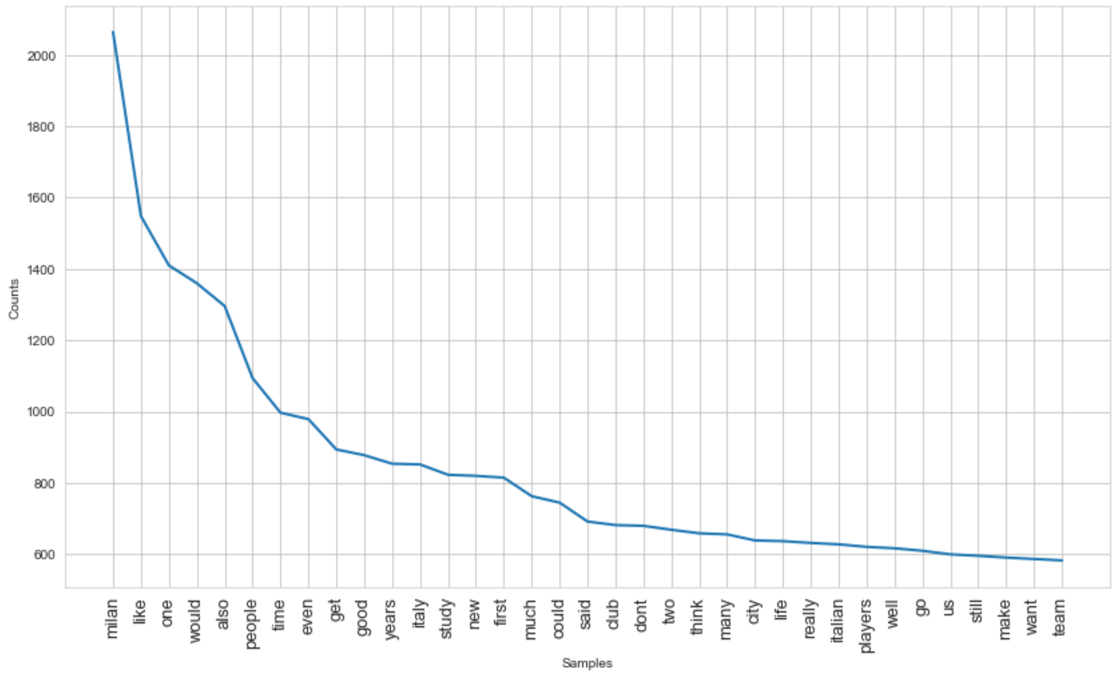


Figure 4. 83 *Milan 35 most frequent words for all sentiments*

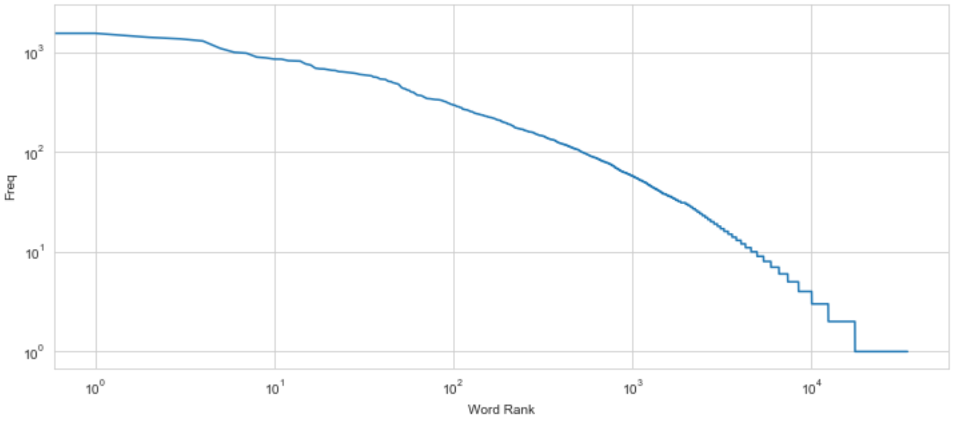


Figure 4. 84 *Milan word counts distribution on a logarithmic scale*

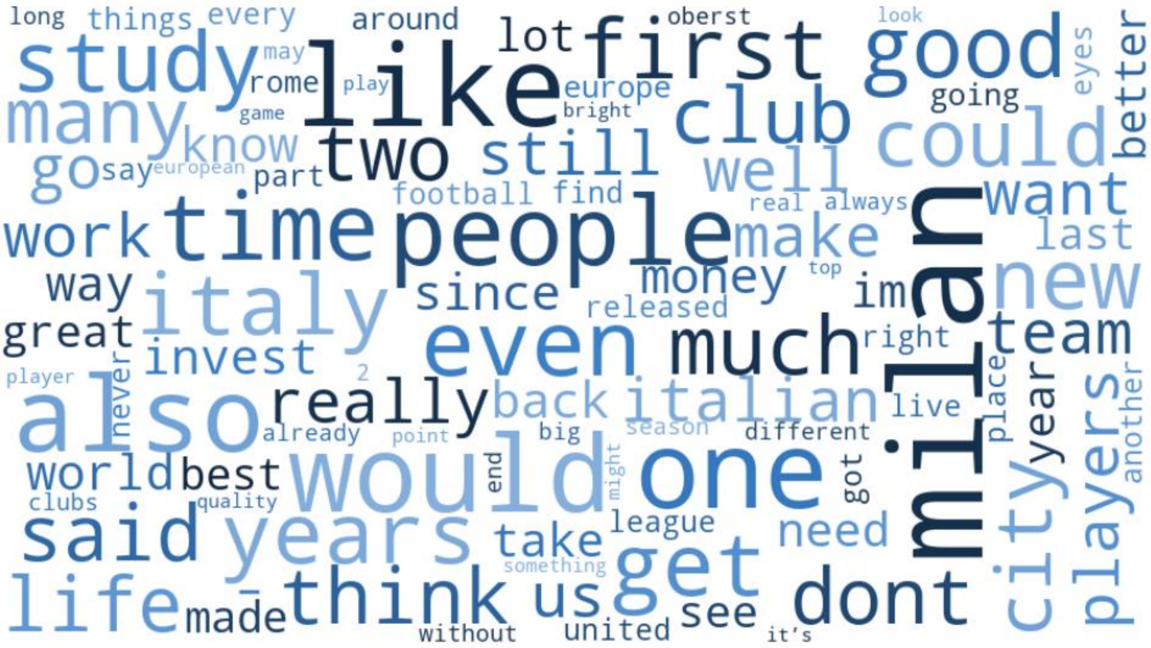


Figure 4. 85 *Milan word cloud on all sentiments*

4.2.7.2 Sentiment Analysis & Words Clustering

In the next table we have, 74.7% of the comments collected are positive, 22.9% are negative and 2.4% are neutral as represented in the figure.

Table 4. 15 *Milan sentiment results*

Negative (-1)	Positive (1)	Neutral (0)	Total
326	1063	34	1423

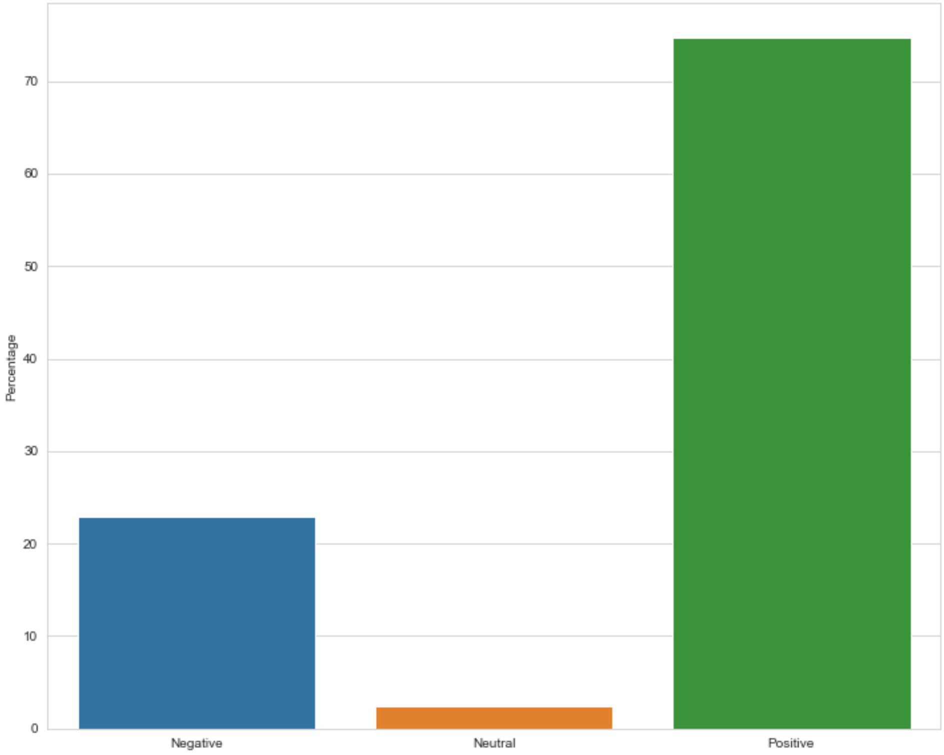


Figure 4. 86 *Milan sentiment results in percentage (%)*

In the following figure, the positive sentiment analysis looks very similar to the one from before.

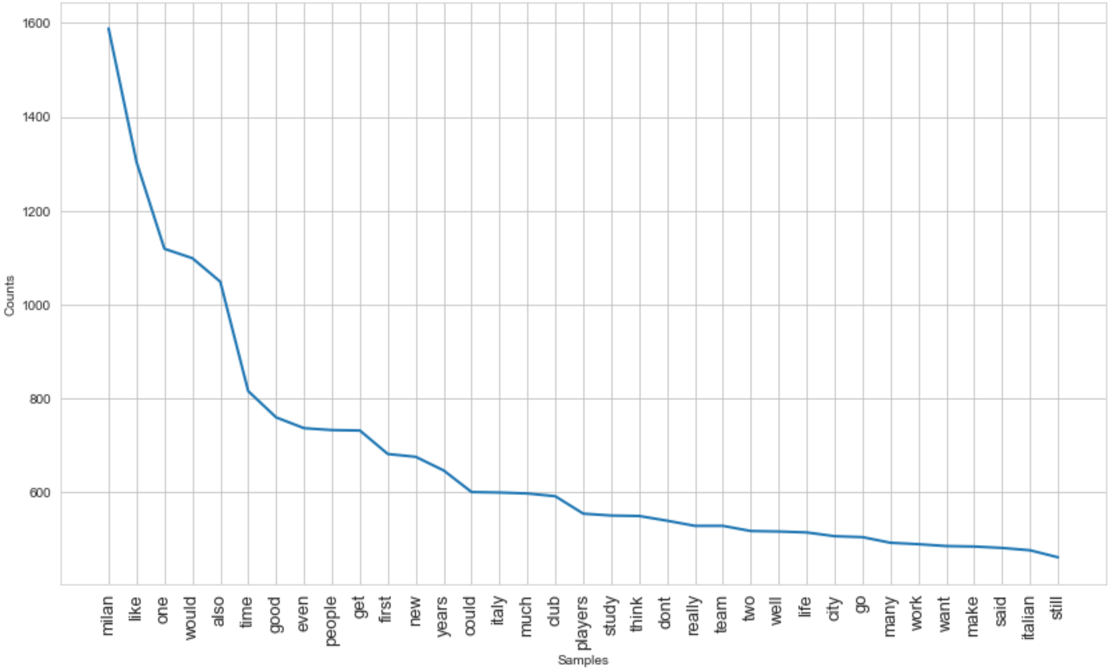


Figure 4. 87 *Milan 35 most frequent words for positive sentiment*

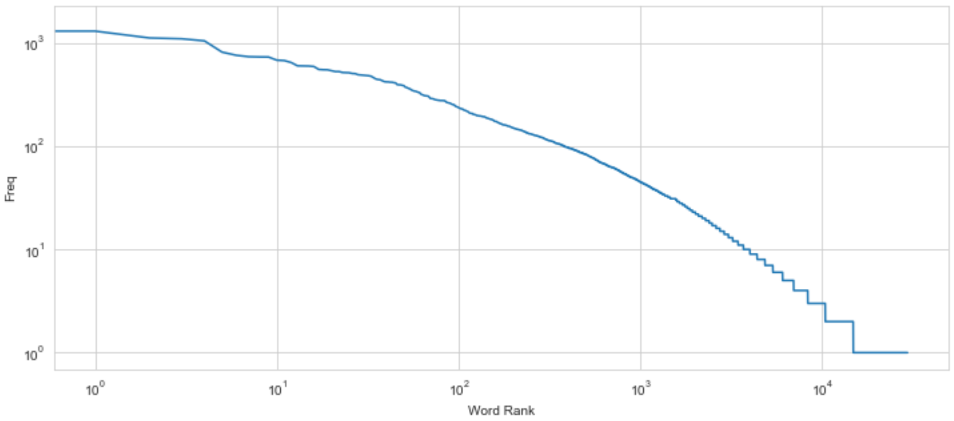


Figure 4. 88 *Milan word counts distribution for positive sentiment on a logarithmic scale*

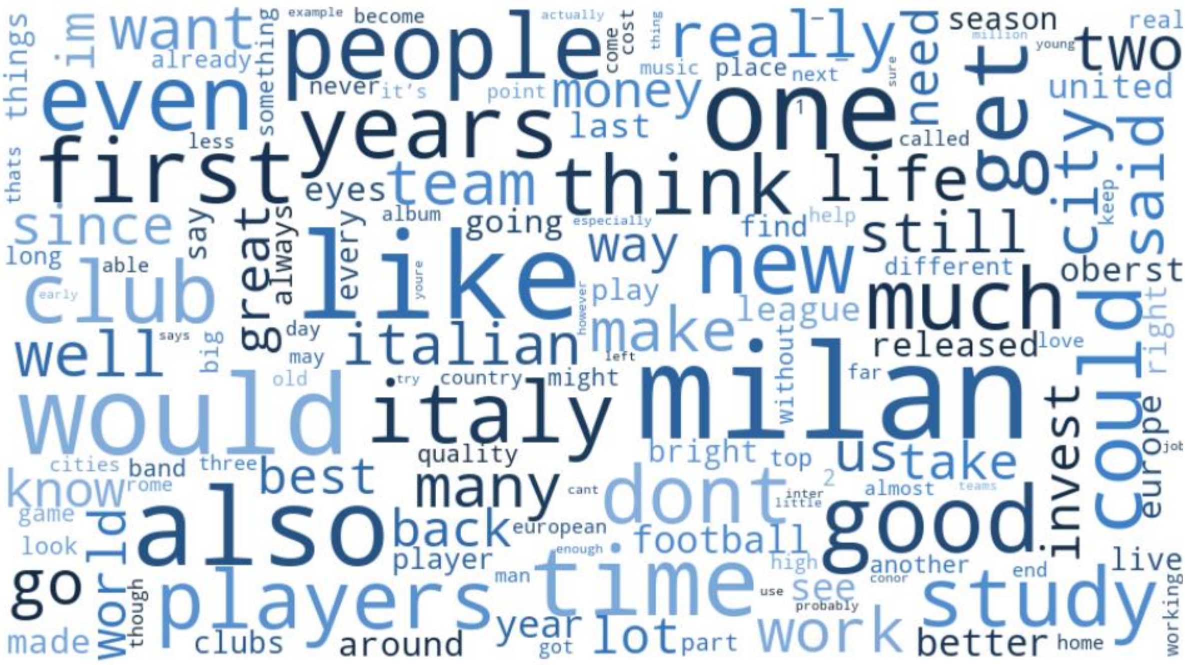


Figure 4. 89 *Milan word cloud on positive sentiments*

For the next figure we are presenting the negative sentiment. The analysis tells us that people talked in a negative way about topics related to the Coronavirus, Italy, study, which seems very related to the main and positive sentiment as well, that’s why it’s very important to get a deeper look at the words clusters and examples.

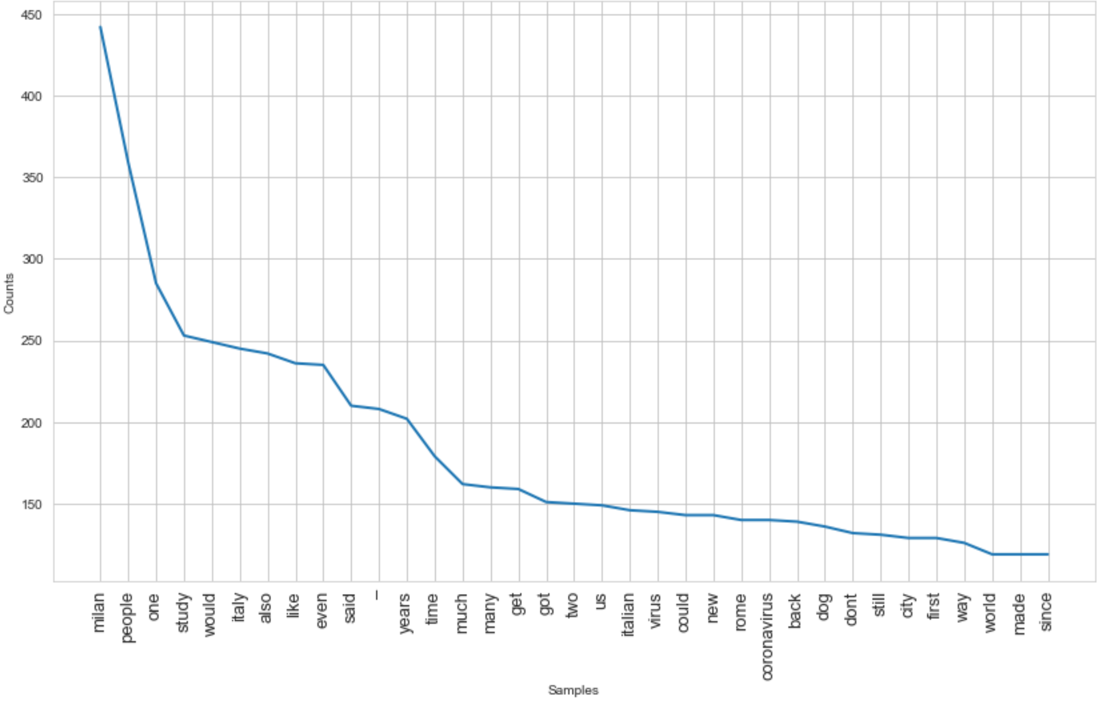


Figure 4. 90 *Milan 35 most frequent words for negative sentiment*

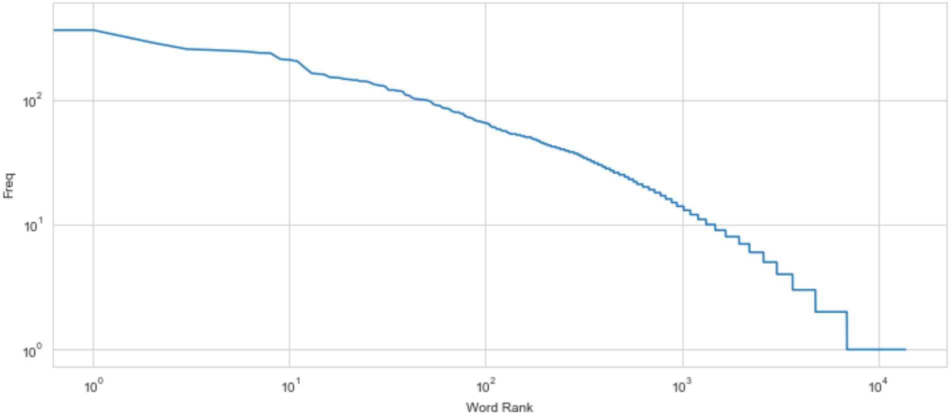


Figure 4. 91 *Milan word counts distribution for negative sentiment on a logarithmic scale*

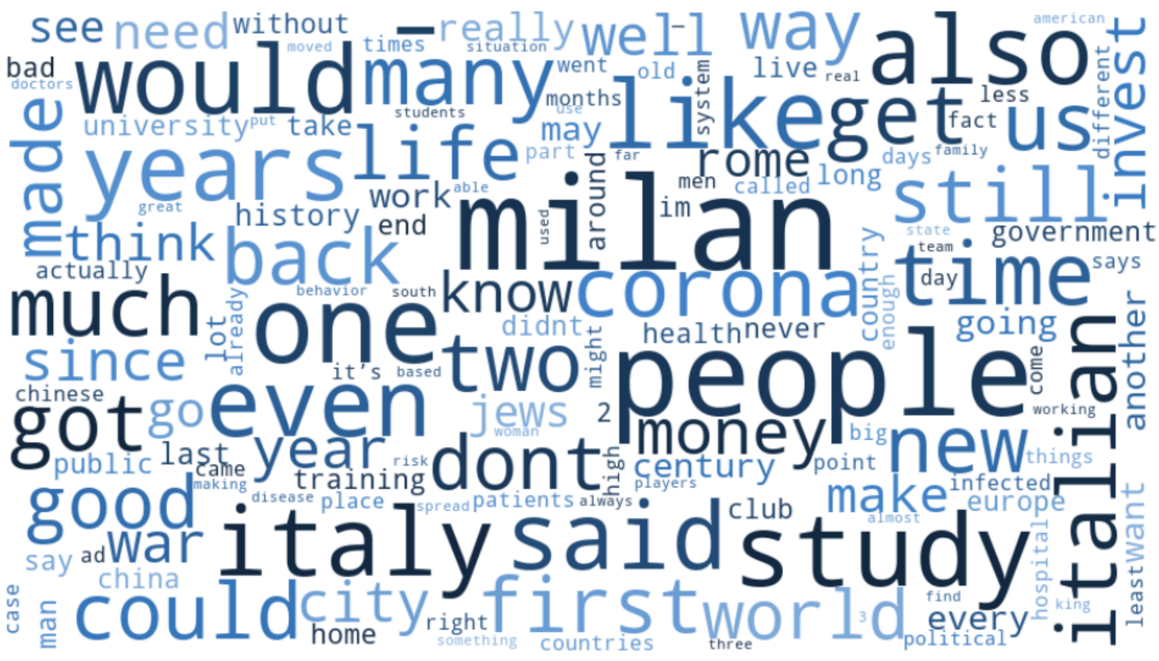


Figure 4. 92 Milan word cloud on negative sentiments

In the next part, word clustering results is carried out.

- Positive Clusters & Discussions

```
[[' invest', ' milan', ' stadium', ' new', ' need', ' ac', ' elliott', ' world', ' fans', ' like'], [' milan', ' like', ' study', '
italian', ' people', ' time', ' new', ' italy', ' good', ' life'], [' club', ' players', ' team', ' invest', ' clubs', ' money', '
football', ' league', ' teams', ' milan'], [' oberst', ' released', ' bright', ' eyes', ' band', ' conor', ' album', '
albums', ' interview', ' omaha'], [' study', ' italy', ' milan', ' like', ' city', ' people', ' im', ' dont', ' life', ' really']]
```

- ['At the end of the Erasmus I went to Milan, Bergamo, Venice, Verona, Lago di Garda and Genova, and it was so funny to say “Ciao, buona sera”, “Grazie mille”, “Arrivederci” and “Questa pizza è un successo” all the time hahaha. Back in Mexico I started liking Italy even more, I listen to L’Officina della Camomilla on a regular basis and love watching Italian cinema. And I’m still learning Italian, I really want to become fluent. My love for Italy is so big that I really see myself living there someday, so I’m gonna study and work hard to achieve it. Viva l’Italia e gli italiani! □□□']
- ["What matters to every man at the end is quality of life. I imagine Food, Sun and living with Italians with their culture and hospitality is an important factor to consider for many players, especially the south americans, but the gap between SerieA and PL is still too wide, most players would still decide to play in a top PL team rather than in Italy, regardless of the bad food and rain. Let's hope Milan in the near future can help shorten that gap."]

- ["Milanese person here. Even though my English is not that good I still think I can have a basic conversation with another person that can speak on the same (basic) level. Do most Italians there speak English? I'm talking about caffes (?) , restaurants , street vendors etc. Yes, most Italians speak an at least acceptable level of English, especially younger generations. Retail workers and the like usually speak it very good. How much (roughly) would a place for 2 people staying 3 nights cost? We're not looking for some fancy hotel , just a decent place that we can sleep over the night. A lot. I used to give a room on AirBnB for rent at it was priced 35 euros per night. It was almost always reserved and this despite me having occasional bad reviews. I suggest you either seek a hostel or rent an AirBnB a little outside of the city, such as in Monza, from where you can reach Milan in about 30 mins by train. How much Euro would be needed for 3 day stay over there? We'll be spending mostly on food , drinks , some souvenirs , maybe visit a cathedral or something along the lines. I'd say 150 euros to have a high margin. Be sure to apply the common moneysaving tips of visiting every town: seek for restaurants where locals eat. I might as well give you restaurant suggestions if you tell me what you like. Do you have any place that you want to recommend to us to visit? We will probably search on Google about things to do in Milano , but if you have anything to share to us would be great! The interior of the Cathedral of Milan, a walk on the roof of the Cathedral. The Pinacoteca di Brera if you are into classic art and the Museo del Novecento if you are into modern art. The Navigli and the old town (known as Cinque Vie and Carrobbio on Google Maps) are pretty neat places for a walk, as well as City Life if you have a kink for brutalist architecture. The Sempione park is worth a walk, as well as a trip on the Branca Tower, it costs just 5 euros and the sight from there is amazing. For nightlife and bars I suggest the aforementioned Navigli. "]

- Negative Clusters & Discussions

[[' italy', ' italian', ' italians', ' people', ' south', ' english', ' study', ' southern', ' milan', ' speak'], [' study', ' city', ' work', ' milan', ' people', ' live', ' easy', ' cities', ' living', ' training'], [' money', ' club', ' ffp', ' clubs', ' invest', ' milan', ' uefa', ' psg', ' debt', ' elliott'], [' players', ' invest', ' got', ' team', ' moved', ' milan', ' cl', ' dont', ' married', ' teams'], [' people', ' study', ' corona', ' im', ' milan', ' said', ' years', ' like', ' dont', ' italy']]

- ["Hi! 4 year in a Bocconi residence. Residual places are very few, usually 2 3 for each residence, so it's not so easy to obtain one. The price is good for Milan, rent is in the 6800 8000€ range per year, so a 560 680 per month, all single rooms. Rent in Milan starts from 550€ without electricity, gas etc., so as you can see the price is not so different, except that in the residence the rent is flat, no bills etc. I really like the residence life, I think it's great for an undergrad because you meet a lot of people from Bocconi, your learn how to study, where are the places and the slang for calling them, little tips for professors, university life, how to hang out etc.."]

- ["You'll need to know italian at a decent level in order to have a job in Italy, finding one knowing english alone is hard. It may still be feasible in Milan, which is the most international city there is in the country.   Of the cities you've mentioned, I'd stay away from Rome"]
- ['The evidence shows us that the first Italian feminists were, for the most part, enthusiastic precursors of fascism. There is no doubt that from the first moment there were women in the new movement of the former socialist leader Benito Mussolini, but little is said that they had, for the most part, feminists. Many of them were those who participated in the activities of the squadrons, revolutionary clash groups of workers\' origin. Some fascists such as the young heroine squad Inés Donati joined the Fasci of Combattimento, groups that are characterized by their violent clashes against communist militiamen. Then groups were formed only for women (the Fasci Femminili). The first one was created in Monza (Milan) and others quickly followed. "]
- ['Politecnico in Milan is of course your best choice. I study electronics and comunications engineering in "Politecnico Di Bari" in southern italy. As Osspn said it lacks the funding other universities, especially in the northern part of italy have, but it provides a well rounded preparation in all the subject you need, the structure is good and it has some well equipped labs. But as I said you should try to get into Politecnico of Milan, it\'s the best for computer science, electronics, engineering. Travel as much as you can, because both in northern part and southern part there are beautiful cities. Good Luck and Forza milan from Barletta.’]

Conclusion

Milan as Madrid before, the name is strongly attached with the Milan football clubs, however, we were able to distinguish some topics concerning the quality and cost of life in the city, some people praised the quality of life that Milan has to offer however, some had negative sentiment against the high cost of life in the city. Other parts of the comments were discussing people’s feedback of a trip or a travel they made to Italy and the Italian culture, as Italy is one of the hottest touristic destinations in the world. Some of the discussions also was focused on whether people who don’t speak the Italian language would be able to find jobs in Milan and the answers were less positive than negative, that topic is crucial to policy makers as it would limit the possibility of attracting skilled workers to the city of Milan.

As a matter of fact, a new study by **(Confindustria, 2019),** said that thousands of technical roles in Italy will go unfilled in the coming years because of a lack of skilled workers applying for jobs. According to the study, from 2019-2021 there will be some 193,000 empty job vacancies in the food, technology, mechanical, textile, chemical and wood-furniture sectors due to the lack of young Italians taking the required training courses, this along with the lack of attracting skilled foreign workers can be a hot topic for policy makers in the upcoming period.

Milan came 3rd for the percentage of positive comments and 5th for the number of yearly foreign visitors, having in mind that Milan has one of the lowest populations and e-commerce active percentage among the cities.

4.2.8 Paris

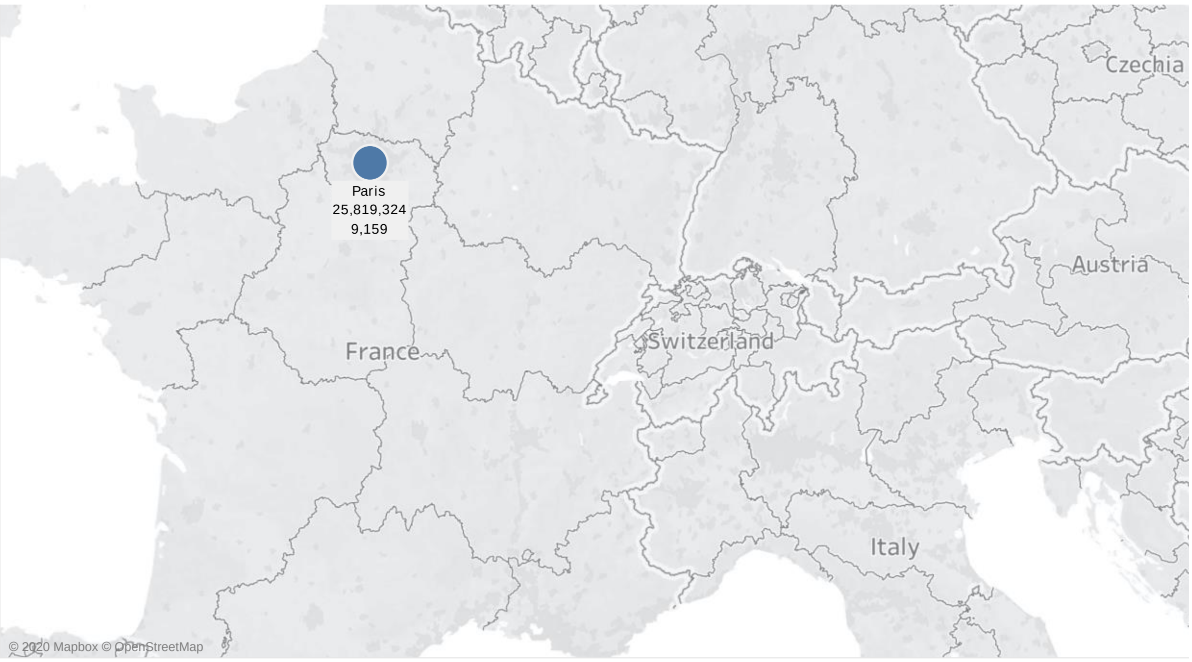


Figure 4. 93 *Geographical representation for the number of the words and comments collected for the city of Paris*

We managed to collect 10001 comment, with a total of 29,436,544 words, 9159 are unique values, with a total of 25,819,324 words to process and analyze.

4.2.8.1 Text Processing and Visualization

After processing the text, we now have 2,320,753 words to analyze. In the following table, as before we can see the difference between the data before processing it and the data after. The processed text has a total average of around 250 words per comment, while not processed texts has around 2600 words as average. The longest processed comment has 3432 words.

Table 4. 16 *Paris text descriptive analysis*

Before Processing	After Processing
count 9159.000000	count 9159.000000

mean 2677.385522	mean 252.142374
std 2523.101915	std 238.631919
min 1.000000	min 0.000000
25% 828.000000	25% 77.500000
50% 1719.000000	50% 162.000000
75% 3852.000000	75% 360.000000
max 39772.000000	max 3432.000000

We can see the previous data represented in the following figures

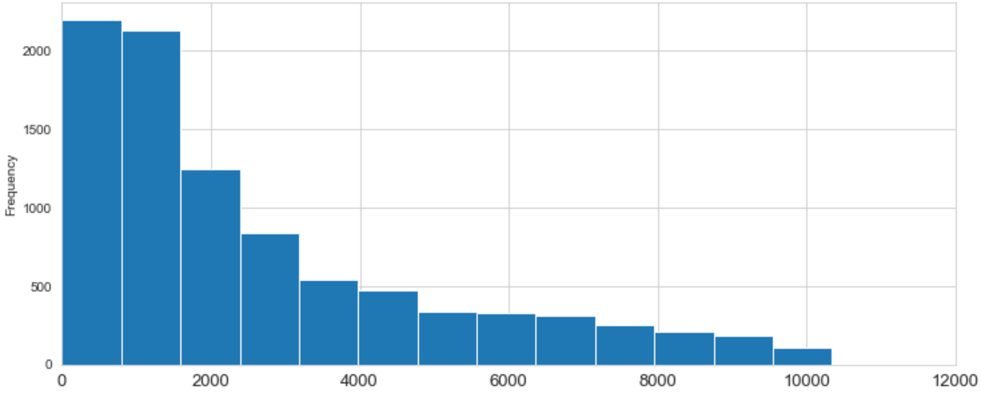


Figure 4. 94 *Paris words per comments before processing*

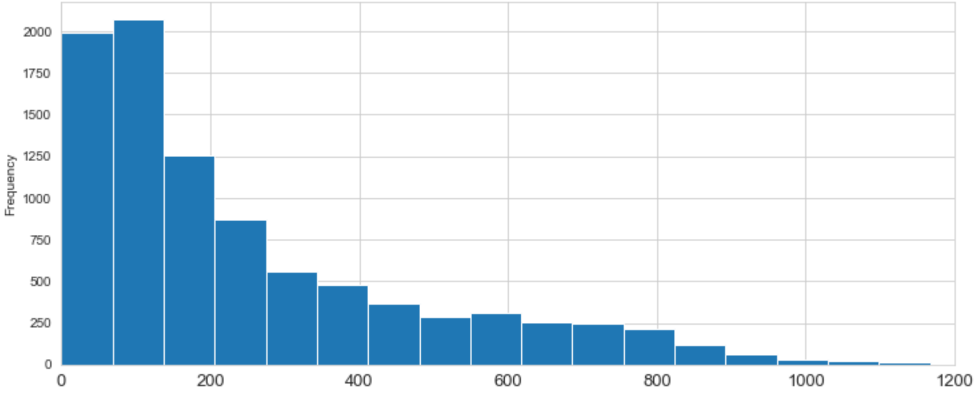


Figure 4. 95 *Paris words per comments after processing*

In the following figure we can visualize the top 35 most frequent words and the most 150 frequent words, we can notice that people usually discuss topics related to the keywords: climate change, world, study, work, and French.

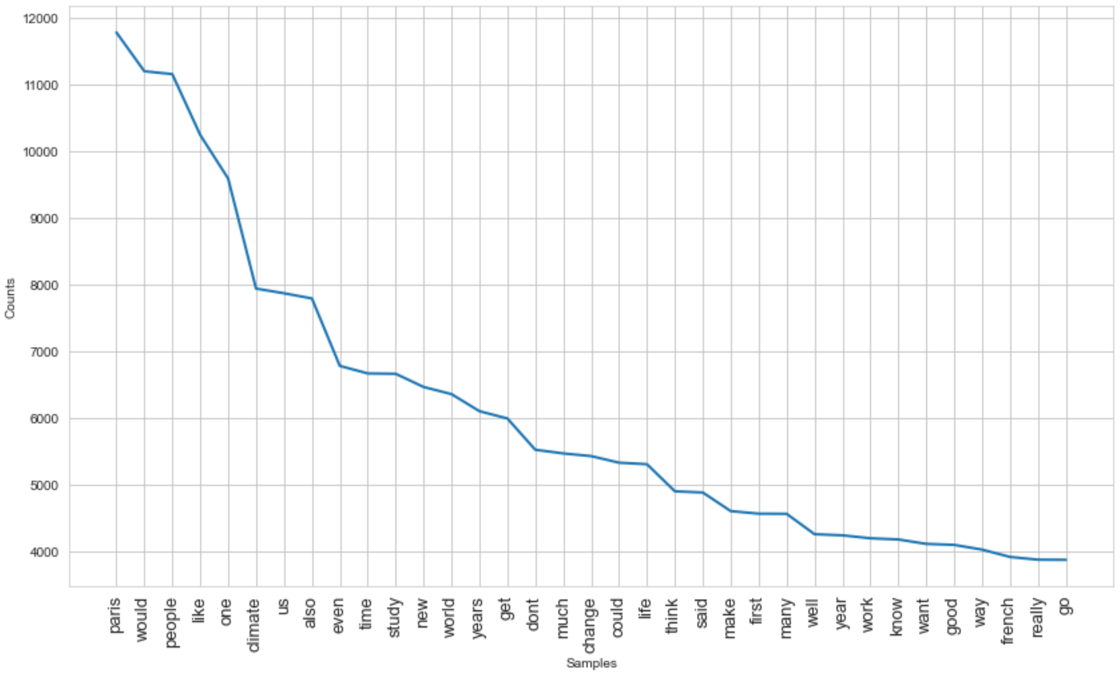


Figure 4. 96 *Paris 35 most frequent words for all sentiments*

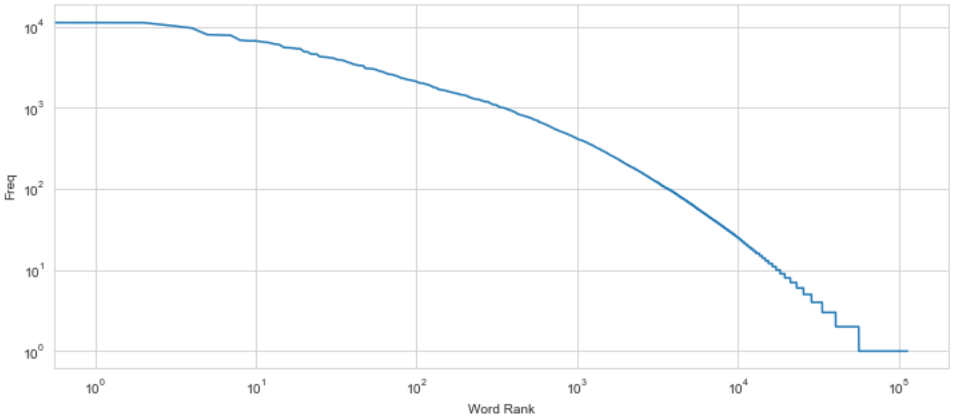


Figure 4. 97 *Paris word counts distribution on a logarithmic scale*

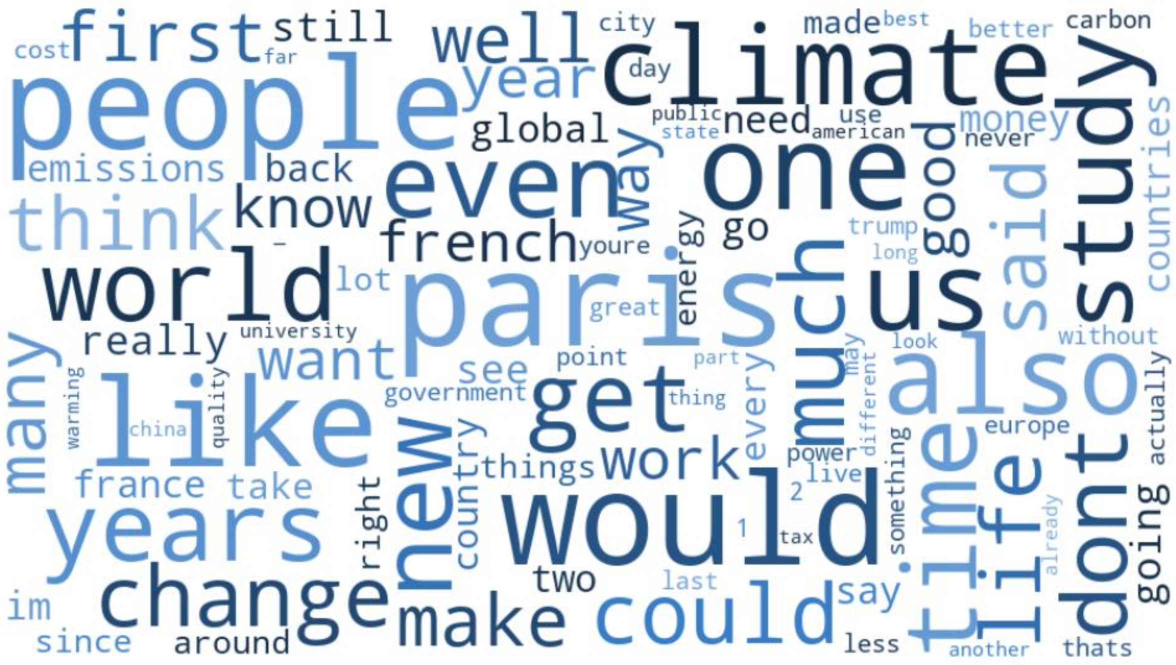


Figure 4. 98 *Paris word cloud on all sentiments*

4.2.8.2 Sentiment Analysis & Words Clustering

In the next table, we have the results of the analysis. 69.3% of the comments collected are positive, 28.9% are negative and 1.8% are neutral as represented in the figure.

Table 4. 17 *Paris sentiment results*

Negative (-1)	Positive (1)	Neutral (0)	Total
2640	6348	171	9159

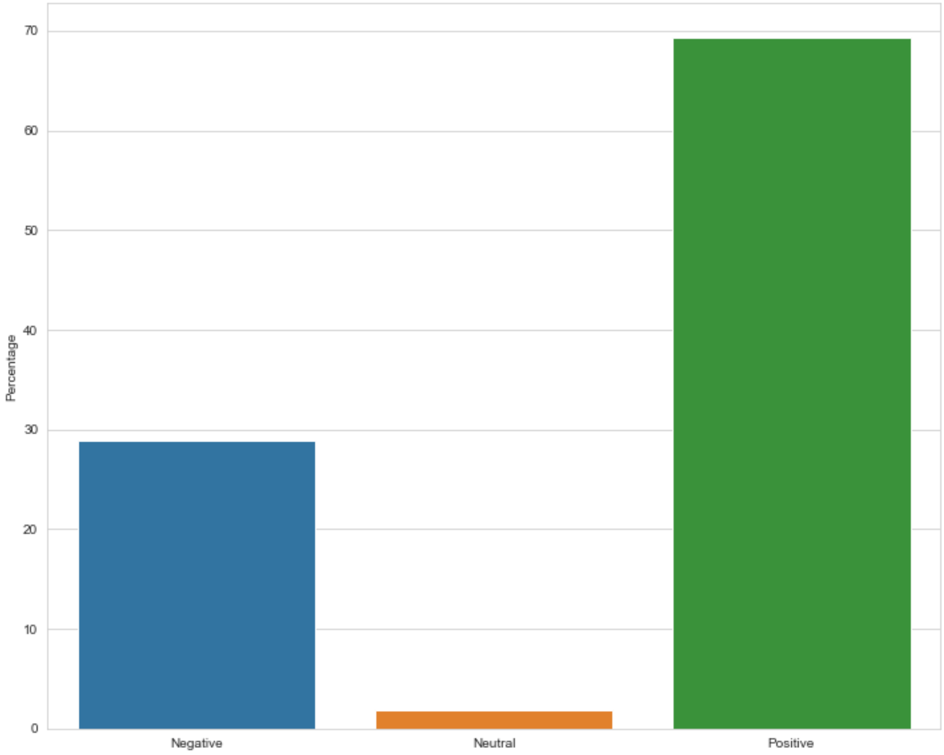


Figure 4. 99 *Paris sentiment results in percentage (%)*

In the following figure we can notice a slight change in the words for positive sentiment comparing to the one from before.

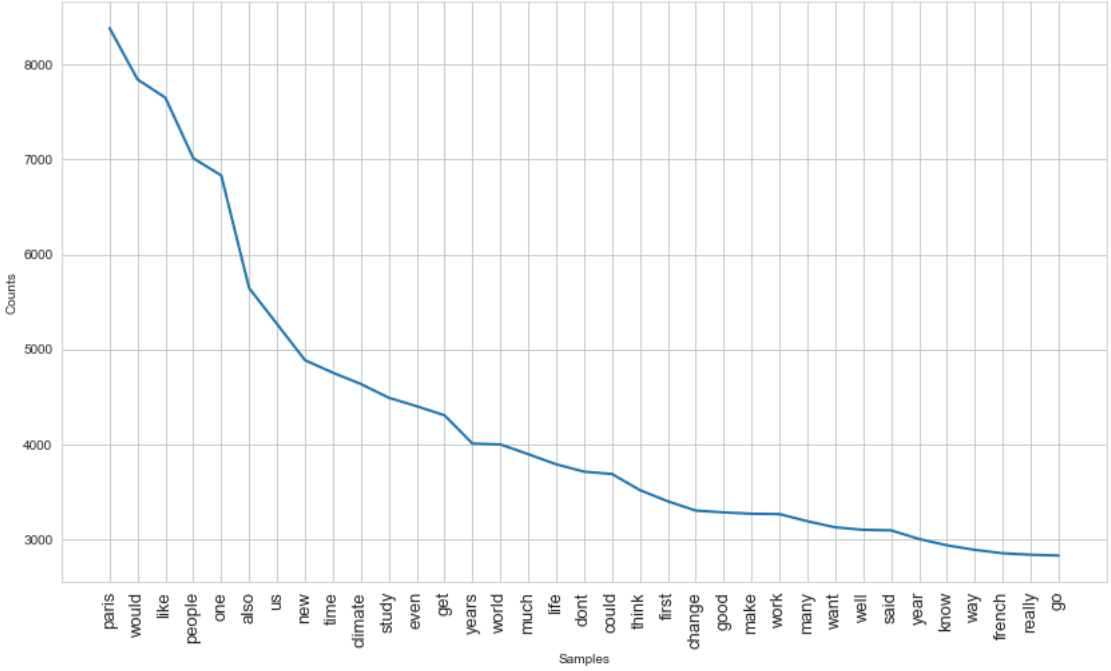


Figure 4. 100 *Paris 35 most frequent words for positive sentiment*

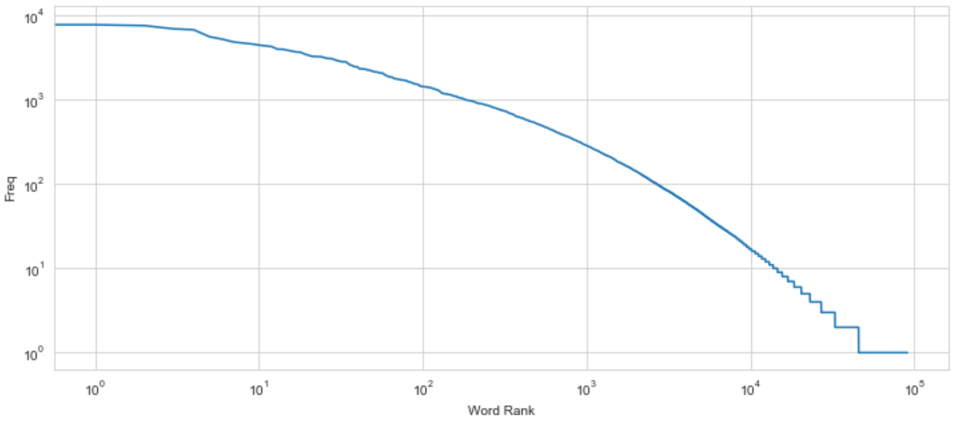


Figure 4. 101 *Paris word counts distribution for positive sentiment on a logarithmic scale*

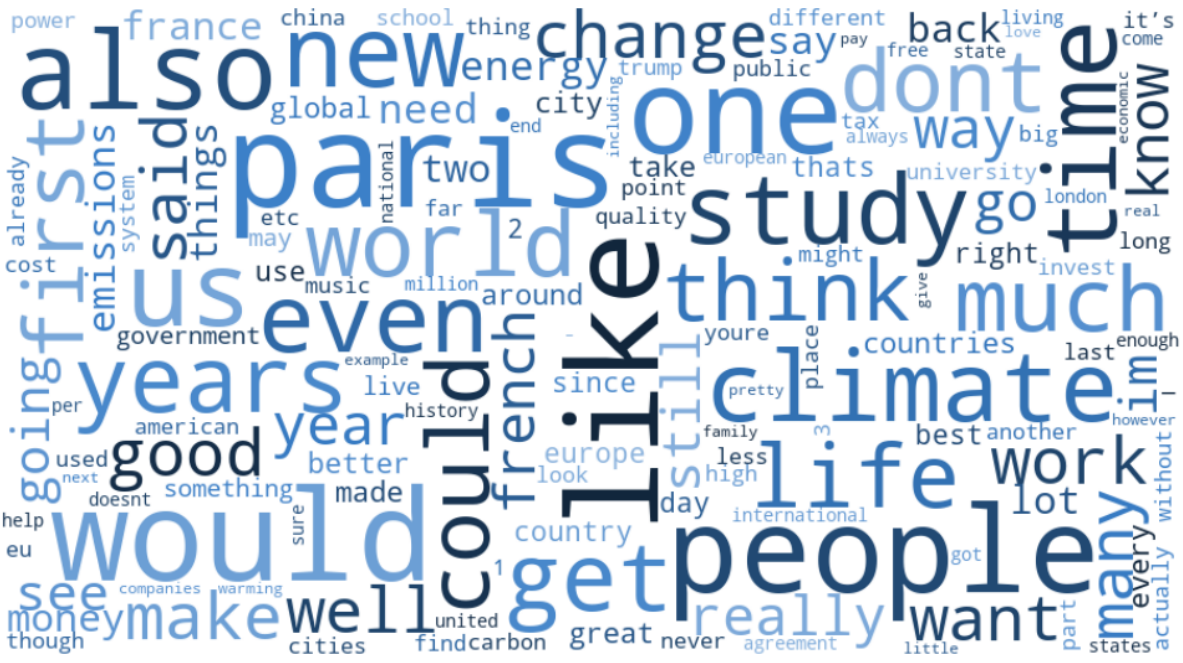


Figure 4. 102 *Paris word cloud on positive sentiments*

For the next figure, the sentiment analysis tells us that people talked in a negative way about the same topics from before however with less frequency, adding to the list keywords such as: Trump, global, countries, and war, this will be elaborated more in the coming section.

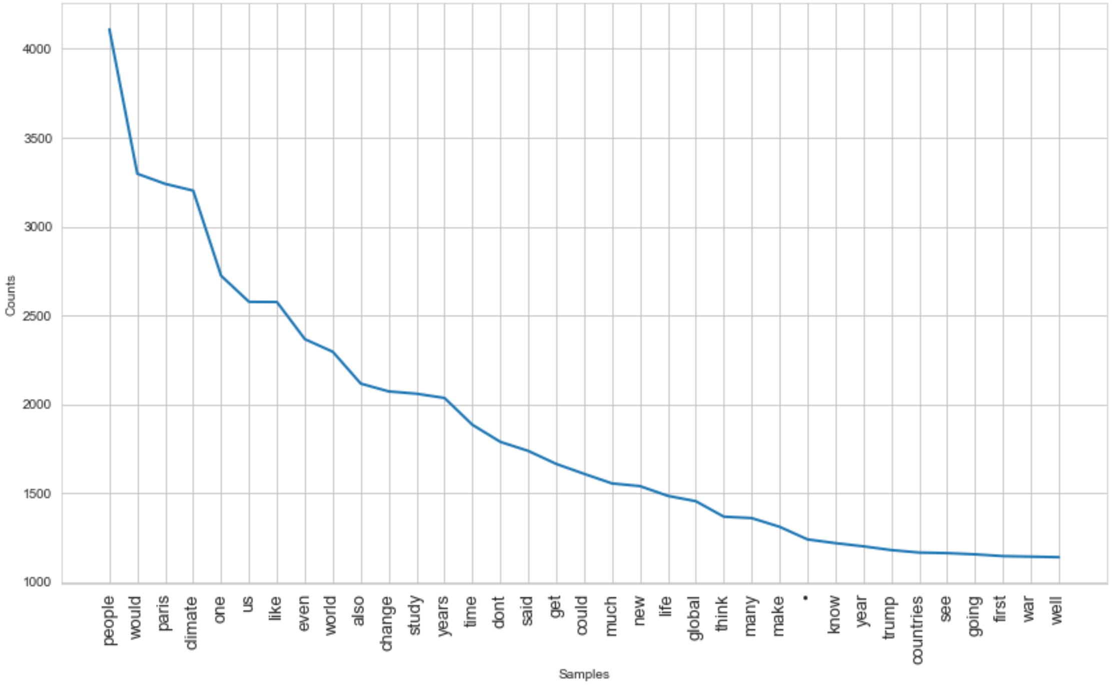


***Figure 4. 103** Paris 35 most frequent words for negative sentiment*

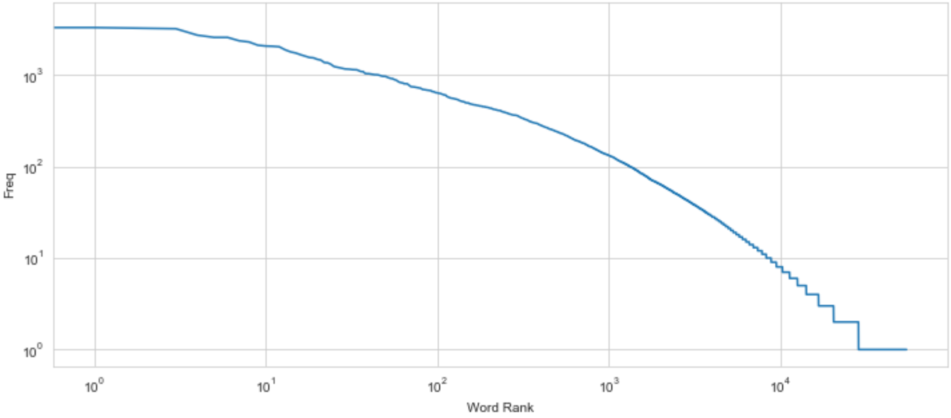


***Figure 4. 104** Paris word counts distribution for negative sentiment on a logarithmic scale*

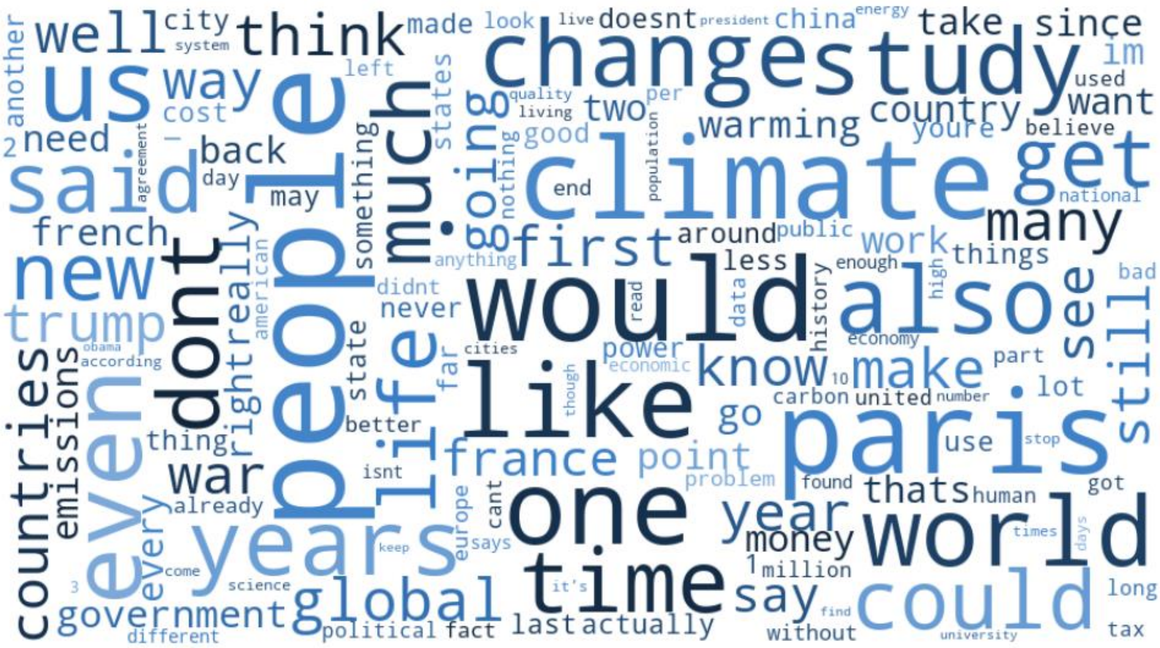


Figure 4. 105 Paris word cloud on negative sentiments

In the next part, word clustering results is carried out.

- Positive Clusters & Discussions

[[' study', ' paris', ' like', ' time', ' new', ' people', ' life', ' years', ' invest', ' music'], [' climate', ' change', ' warming', ' global', ' emissions', ' agreement', ' carbon', ' world', ' scientists', ' degrees'], [' energy', ' emissions', ' climate', ' carbon', ' tax', ' trump', ' china', ' coal', ' countries', ' agreement'], [' people', ' like', ' life', ' dont', ' city', ' live', ' cities', ' think', ' want', ' paris'], [' french', ' language', ' abroad', ' english', ' france', ' study', ' school', ' paris', ' speak', ' program']]

- [“The investors want governments to speed up the shift from fossil fuels to renewable energy, back rules requiring companies to report climate-related information in their financial statements and put a meaningful price on carbon emissions.   In a separate investor-backed study released Wednesday, the London-based Transition Pathway Initiative said none of the world #039 s top 50 oil and gas companies are in line with the Paris goal of capping global warming at 2C. By contrast, the researchers found that about 20% of the 59 electric utility companies examine are aligned with that target.   quot We, as a major institutional investor, are concerned that transition risk - the large and growing gap between government targets and company ambitions - is a major source of investment risk."]
- [“It sounds like you thought about this financially, but not from an academic aspect. Master's are intense because you only have 2 years to research and write a paper."]

- ["If you find a residence near the school/university in Paris, that would dramatically improve your life. - If you can afford the cost of living (and the tuiton fees, if applicable), go for ESSEC. It's the best on the list in terms of education, reputation, pretty much everything."]
- ['The Paris agreement is certainly a step in the right direction, but it is only a step," said Monier. "It puts us on the right path to keep warming under 3° C, but even under the same level of commitment of the Paris Agreement after 2030, the study indicates a 95 percent probability that the world will warm by more than 2° C by 2100. This is from the article you mentioned. Ya it is not enough but it definitely has an effect.']
- 'Paris is definitely the place with the most offers, followed by Lyon. But there are tech jobs in every medium city in France to name a few Nice, Rennes, Nantes, Orléans, Tours, Bordeaux, Montpelier, Lille I'm from Orléans and studied in Tours, they are pretty medium cities, but they have some opportunities and a very good quality of life. If you're looking for a big city, Paris is great, you have big airports, good public transportation but it's expensive, hard to get a big appartement in the center for a reasonable price. Lyon is a really cool city too, it's souther than Paris, really beautiful too and good airport and transport system too. Rent is not as bad as Paris. If you're looking for a place that's easiest to live in, it depends on the climate you're looking for. For the visa thing, I have no idea how difficult it is, never came across an American who works here. I don't know how much it costs either, but if you have a good profile/experiences, you will find a company to take you in for sure.'

- Negative Clusters & Discussions

```
[[' jackson', ' michael', ' boys', ' study', ' girls', ' nude', ' broke', ' paris', ' book', ' went'], [' abroad', ' gun', '
study', ' marx', ' mass', ' shootings', ' paris', ' guns', ' people', ' like'], [' people', ' trump', ' dont', ' like', ' tax', '
money', ' think', ' climate', ' want', ' world'], [' paris', ' people', ' study', ' french', ' france', ' life', ' like', ' city', '
war', ' time'], [' climate', ' change', ' warming', ' global', ' world', ' emissions', ' years', ' carbon', ' co2', '
earth']]
```

- [' The US did this after pulling out of the disastrous Paris agreement. Thank you President Trump. [It was entities that remained committed to regulations put in place before the Trump Presidency that drove the drop in emissions.](Trump's policies have been trying to reverse the very policies that has caused this drop. Overall, U.S. greenhouse gas emissions fell about 2 percent in 2019, according to preliminary estimates by Rhodium Group, an economic analysis firm. The previous year, strong economic growth and other factors had pushed emissions up roughly 3 percent. The 2019 drop was driven by a nearly 10 percent fall in emissions from the power sector, the biggest decline in decades, according to Rhodium. And the story there is all about coal. Coal generation in the U.S. fell by 18 percent last year, the largest annual decline on record, according to Rhodium. Another study, published in December, found a smaller but still dramatic drop for coal generation last year.']

- ["Yes! Paris is pretty dangerous! If a woman walks alone, I'd say 50% of the time some random guy will come to you to block your way, then ask you to go with them or ask for money. Even in broad daylight. Do they really expect that a woman will say yes? And why do they have so much free time to harass people? And what's up with letting those lowlife people immigrate easily to harass women and steal, while other people I know with STEM degrees, who wants to legit study/work there have a hard time getting visa. Don't understand the recent trend to deny reality."]
- [" Isn't the US leading in reducing emissions already? Why pay out big money for shit like the Paris Climate Accord? Because the US has massive wealth in large part because of historical and ongoing emissions that are still far higher than the per-capita emissions of the countries money from the Paris Climate Accord would help support. People in the US - and every other developed nation - are literally murdering people so that they can have a higher quality of life. This isn't hyperbole. People are dying, we know why, and plenty of people who are aware aren't interested in changing anything. "]

4.2.8.3 Conclusion

Paris come 1st among all the cities for the number of foreign visitors per year, it also comes 1st as of their GDP, 4th in population, percentage of positive sentiment online, and internet access and 3rd on e-commerce active. Most of Paris' heated discussions online are focused on the gun spread issues and on the climate change and global warming topics, along with some results concerning the livability in the city. It's important to know that France has strict gun control laws, but illegal weapons are widespread, and these have been used in most of the deadly terror attacks in the country in recent years. **(The Local France, 2017)** Its also important to know that in 2016, The Paris Agreement was signed, the agreement was within the United Nations Framework Convention on Climate Change (UNFCCC), dealing with greenhouse-gas-emissions mitigation, adaptation, and finance. Under the Paris Agreement, each country must determine, plan, and regularly report on the contribution that it undertakes to mitigate global warming. In 2017, some studies have said that none of the major industrialized nations were implementing the policies they had imagined and have not met their pledged emission reduction targets, and even if they had, the sum of all member pledges (as of 2016) would not keep global temperature rise "well below 2 °C". Although the United States has declared that it will withdraw from the Paris Agreement, this cannot be taken into effect until the day after the 2020 presidential election. And for that reason, the internet is full of both positive and negative sentiments about the Paris agreement, some supported the United States position of withdrawing from the agreement as according to the comments, USA is already reducing its emissions massively, and some others considered it a move in the wrong direction as USA is the second largest producer of carbon footprint in the world after China **(Wikipedia, 2020).**

If we compare what people have talked about online and compared it to the real market situation, we can find that the total climate emissions in USA rose in 2018 but fell slightly, by 2%, last year **(ecoRI News, 2020)** . However, the decline in 2019 was due almost entirely to the continuing decrease in coal consumption. The decline was largely due to market forces, rather than policies. Some coal plants were closed in favor of cheaper natural gas and renewable energy. The reality is that the United States has achieved no significant reductions in greenhouse gases over the past three years, which will make meeting treaty targets more challenging **(Inside Climate News, 2020).**

Other discussions about Paris included the degree of safety for women and others were about the cost of life, commuting or studying in the city.

4.2.9 Rome

Figure 4. 106 *Geographical representation for the number of the words and comments collected for the city of Rome*

We managed to collect 7281 comment, with a total of 23,129,850 words, 6605 are unique values, with a total of 19,163,801 words to process and analyze.

4.2.9.1 Text Processing and Visualization

After processing the text, we now have 1701501 words to analyze. In the following table, the processed text has a total average of around 250 words per comment, while not processed texts has around 2800 words as average. The longest processed comment has 6969 words.

Table 4. 18 *Rome text descriptive analysis*

Before Processing	After Processing
count 6605.00000	count 6605.000000
mean 2819.14595	mean 256.505526
std 2701.72518	std 243.075577
min 2.00000	min 0.000000
25% 881.00000	25% 82.000000
50% 1914.00000	50% 175.000000
75% 4100.00000	75% 370.000000
max 80211.000000	max 6969.000000

We can see the previous data represented in the following figures

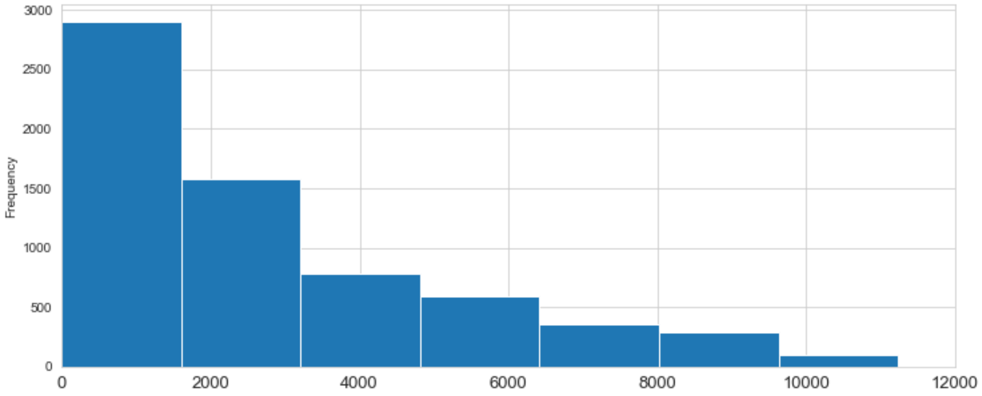


Figure 4. 107 *Rome words per comments before processing*

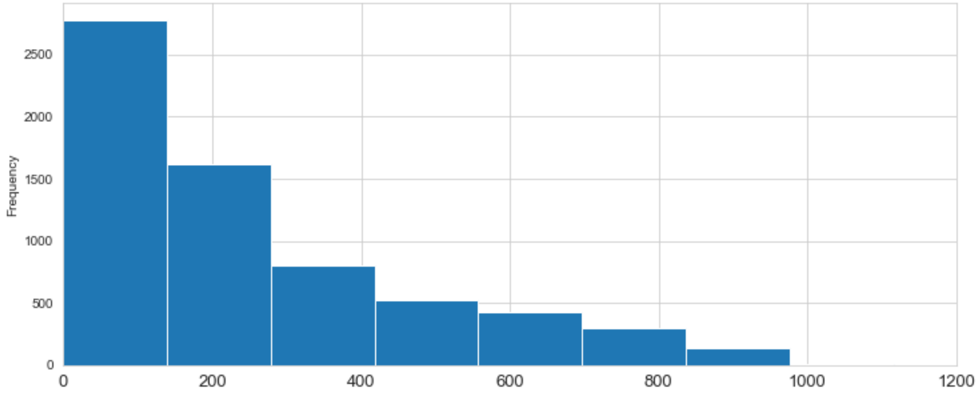


Figure 4. 108 *Rome words per comments after processing*

In the following figure we can visualize the top 35 most frequent words acquired and the most 150 frequent words, we can notice that people usually discuss roman, time, study, history and church.

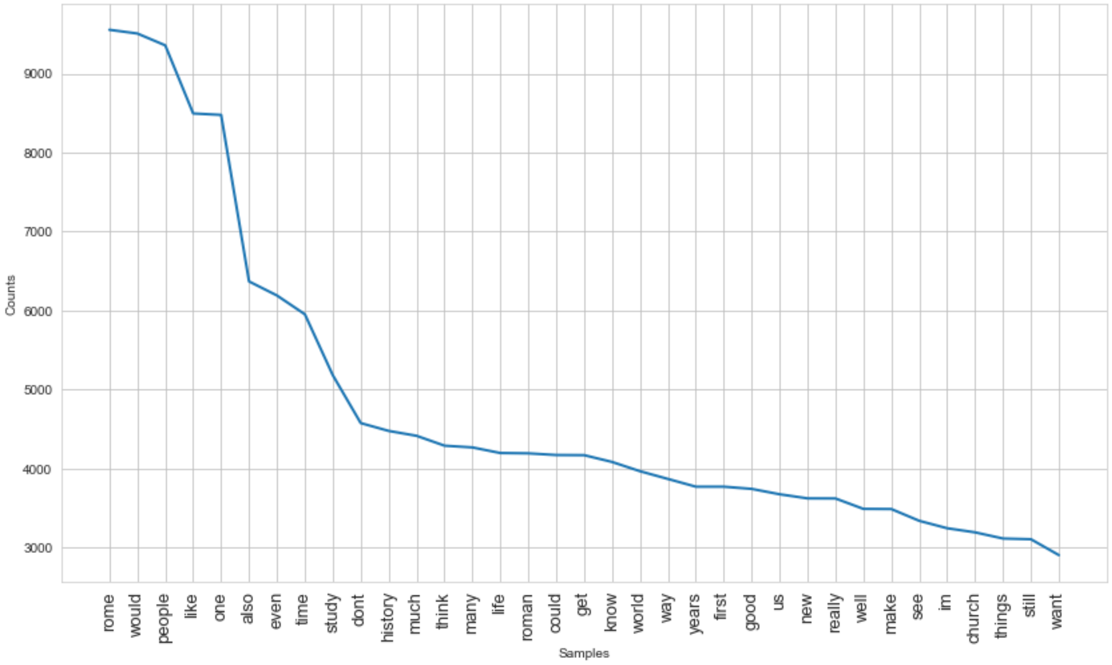


Figure 4. 109 *Rome 35 most frequent words for all sentiments*

We can sort the word counts and plot their values on Logarithmic axes to check the shape of the distribution.

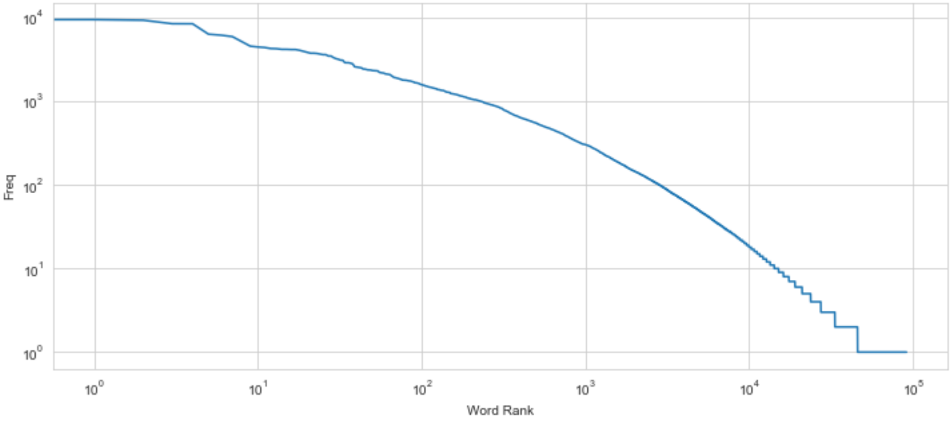


Figure 4. 110 *Rome word counts distribution on a logarithmic scale*

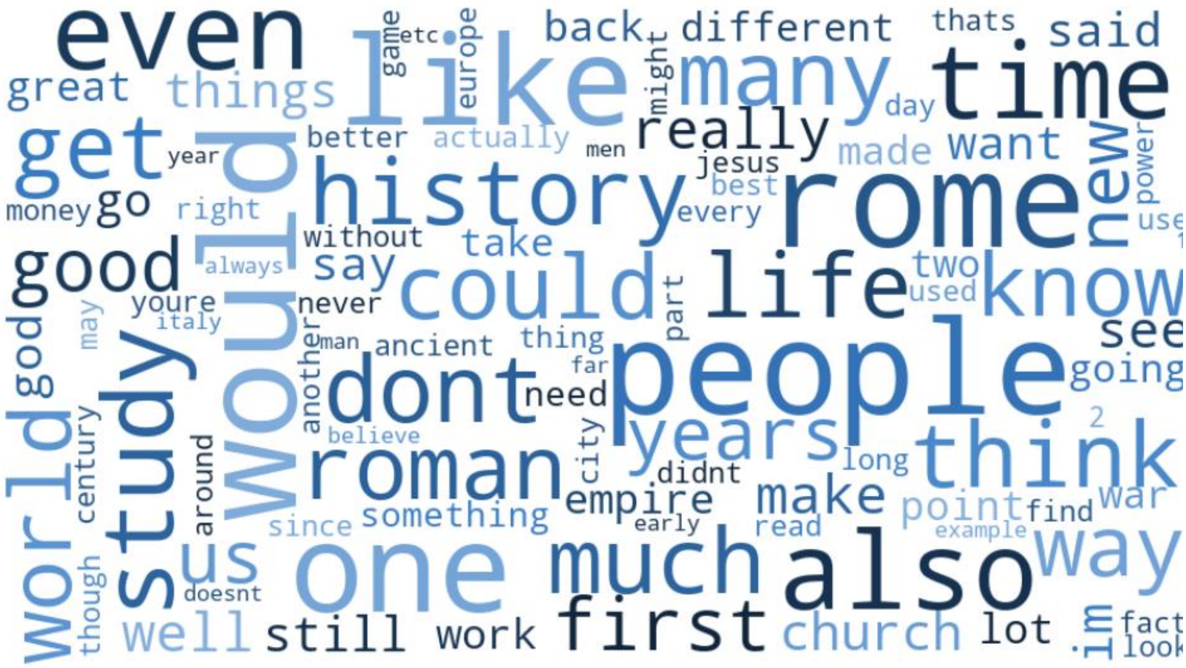


Figure 4. 111 *Rome word cloud on all sentiments*

4.2.9.2 Sentiment Analysis & Words Clustering

it’s time to use Sentiment Analysis on the text to label it. In the next table we have the results of the analysis. 66.23% of the comments collected are positive, 32.24% are negative and 1.53% are neutral as represented in the figure.

Table 4. 19 *Rome sentiment results*

Negative (-1)	Positive (1)	Neutral (0)	Total

2130	4374	101	6605

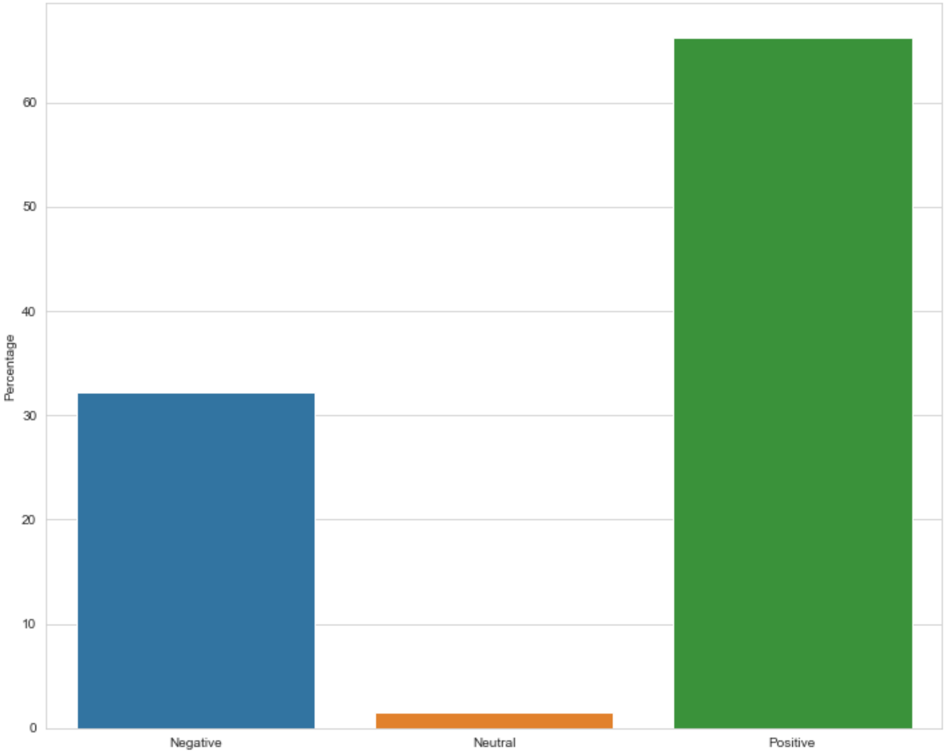


Figure 4. 112 *Rome sentiment results in percentage (%)*

In the following figure, we can notice a slight change in the positive words comparing to the one from before.

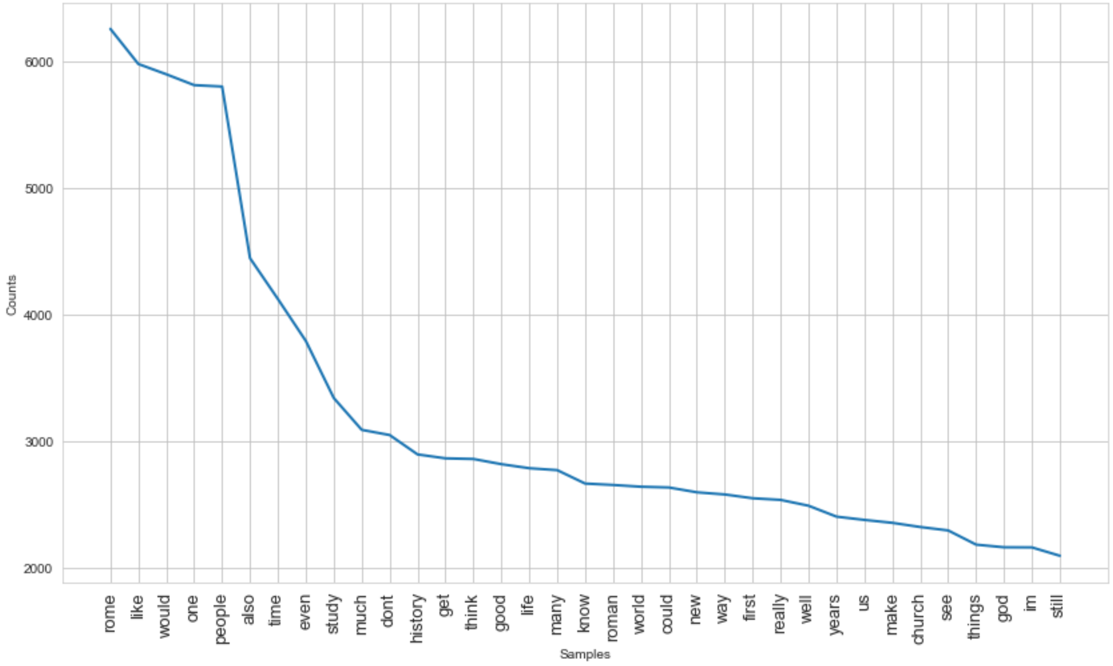


Figure 4. 113 *Rome 35 most frequent words for positive sentiment*

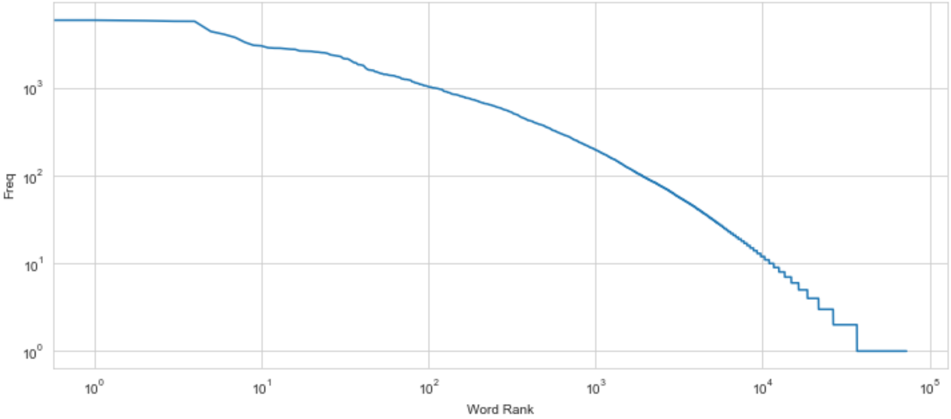


Figure 4. 114 *Rome word counts distribution for positive sentiment on a logarithmic scale*

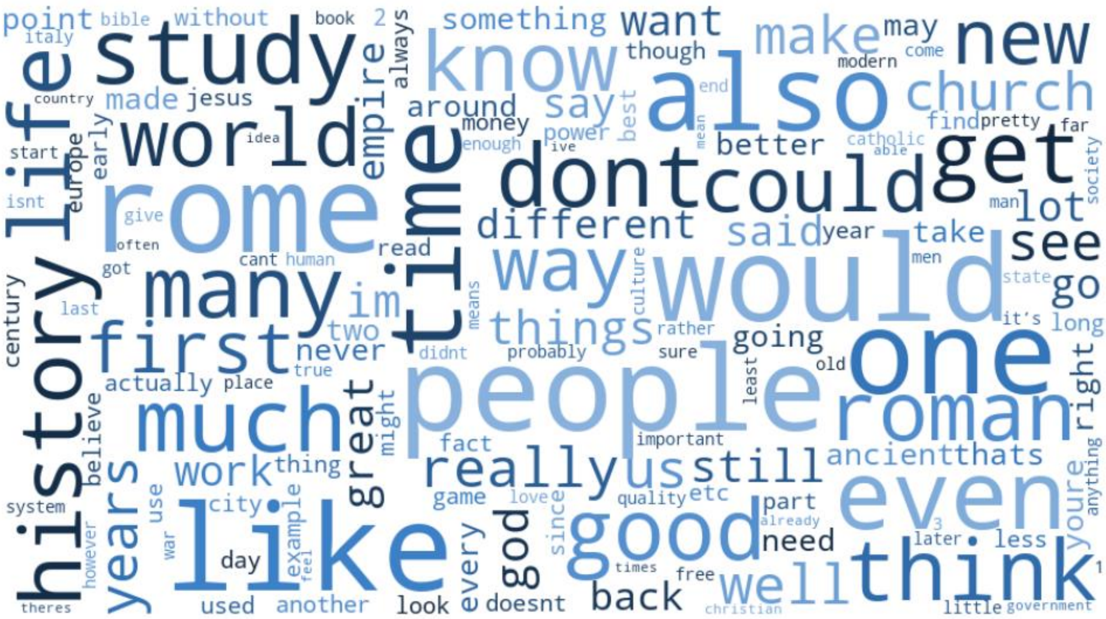


Figure 4. 115 *Rome word cloud on positive sentiments*

In the next figure, the analysis tells us that people talked in a negative way about similar topics to the positive sentiment, the words are in general related to the Roman empire and the Roman times.

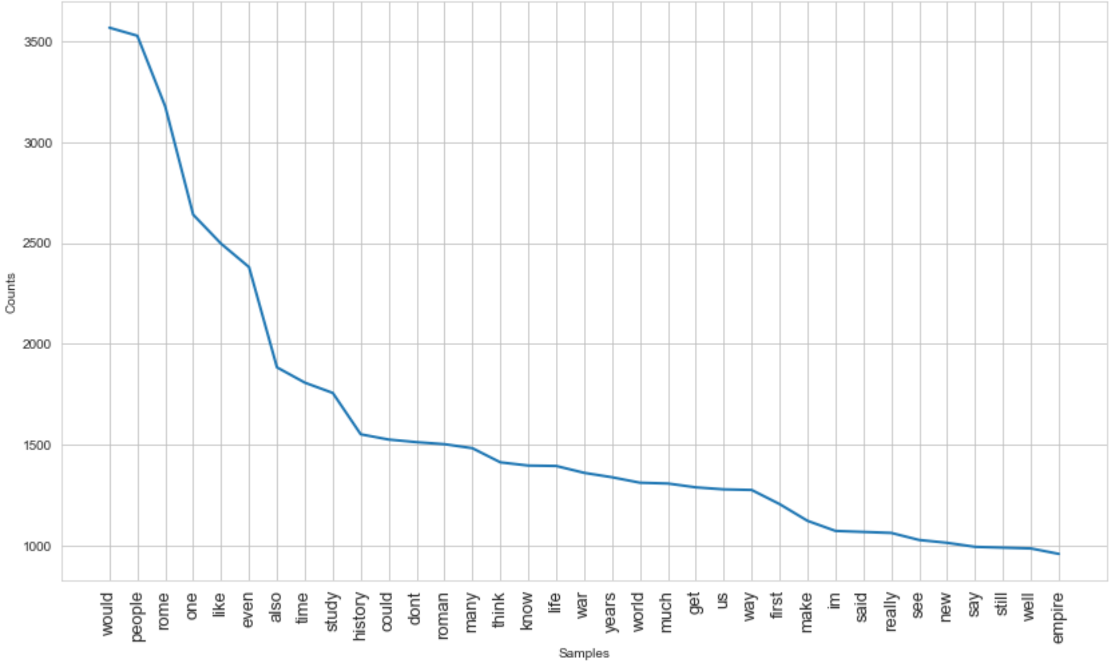


Figure 4. 116 *Rome 35 most frequent words for negative sentiment*

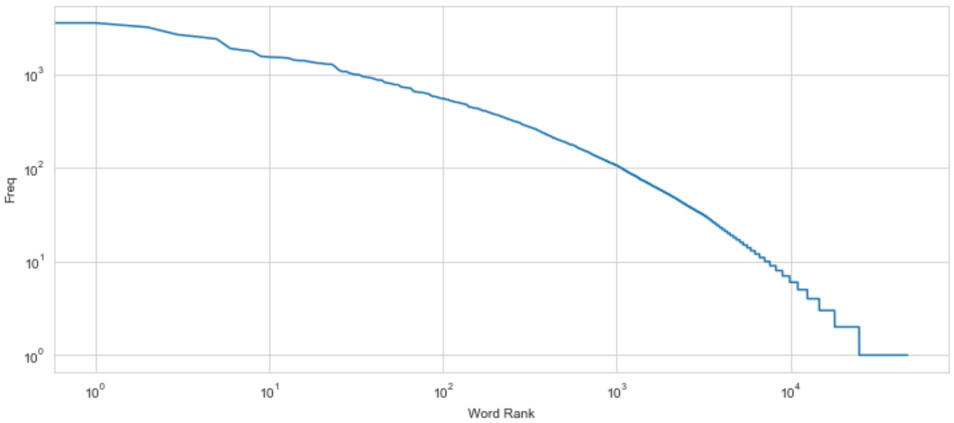


Figure 4. 117 *Rome word counts distribution for negative sentiment on a logarithmic scale*

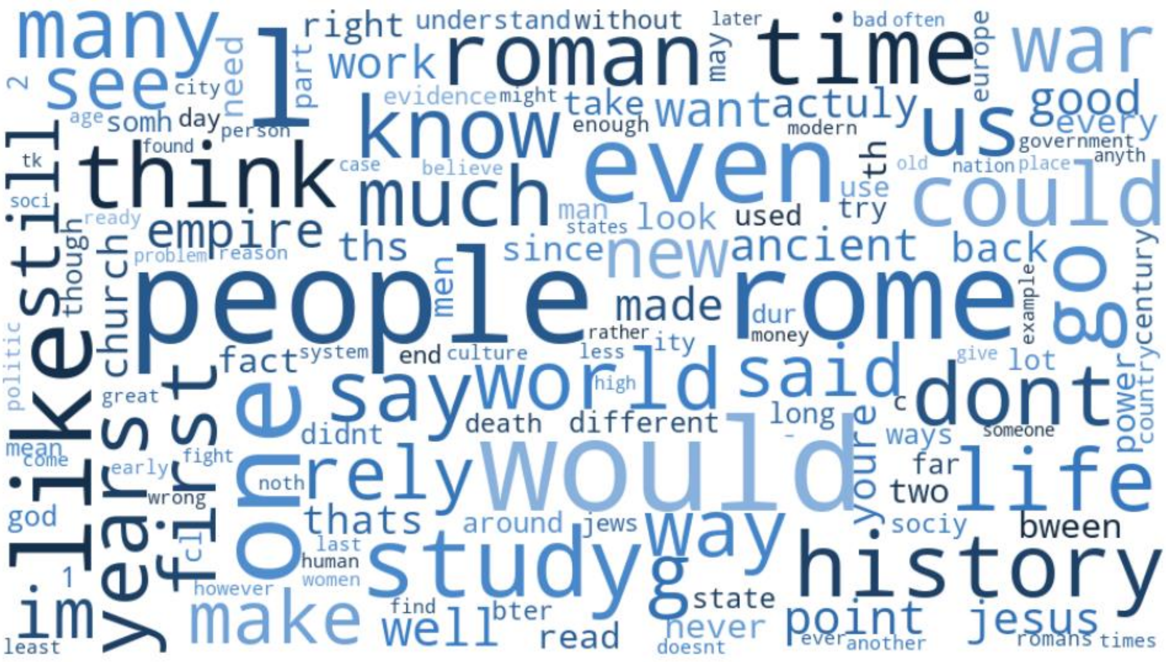


Figure 4. 118 *Rome word cloud on negative sentiments*

In the next part, word clustering results is carried out.

- Positive Clusters & Discussions

[[' [[' church', ' god', ' jesus', ' catholic', ' bible', ' christ', ' faith', ' orthodox', ' christianity', ' christian'], [' abroad', ' italy', ' study', ' rome', ' italian', ' city', ' like', ' school', ' trip', ' people'], [' people', ' like', ' dont', ' think', ' time', ' im', ' know', ' world', ' make', ' life'], [' rome', ' history', ' roman', ' study', ' like', ' ancient', ' empire', ' time', ' life', ' world'], [' game', ' dlc', ' games', ' campaign', ' factions', ' play', ' warhammer', ' war', ' total', ' units']]

- ["Amsterdam, Paris, Barcelona, Rome too crowded, touristy, and expensive (in relation to local salaries and wages). Copenhagen, too expensive and difficult af to even get your foot into anyway, they're the opposite of Sweden in terms of openness to foreigners there. Stockholm, darker and colder than Berlin, expensive and very difficult to find an apartment but probably more realistic than Copenhagen. They have a pretty strong music scene as well. I personally rotate between Lisbon, Vienna, Amsterdam, Stockholm, and Prague in basically that order."]

- [‘I will say, and I think this will be of the more controversial opinions in this forum, Technocracy does not necessarily demand strict adherence to empiricism -- Specifically with regards to religion. You can be religious and scientifically minded, and generally speaking, when it comes to quality of life, satisfaction, and stress management, religion has many beneficial qualities. A scientific society should make use of a religion of the state, like ancient Rome.']
- [“The more we possess the faster we can develop and progress and improve quality of life. Historically, the high points of civilization are always when the population is maximal. When populations decline, civilization collapse is underway. We saw it in ancient Rome, China and thousands of other empires. It raises the question is it possible to manage a soft landing with a declining population? History says no. The status quo of ever increasing populations create a host of new problems, but with enough human resources (capital, ingenuity, labor) I believe they can be solved. For example, getting off this planet is a good start."]
- ['Study the histories of Christian women, including female saints! Christianity consists of not only Biblical history, but post-Biblical history as well. Not only do you learn more about Christian women this way, but for me learning about how we got from the last chapter of Revelations to today makes me feel closer to God and stronger in my faith. Here’s a fun fact: women are likely the only reason Christianity wasn’t entirely wiped out during a plague that went through Rome. While most people were leaving the city and leaving their sick and dying loved ones behind, Christian women were the ones who stayed behind to care for them. While they were nursing sick people back to health, the sick people asked them why they would risk their lives for them, and the women told them about Jesus. This meant that, despite a major population drop and weakening of other religions, the number of followers of Christianity skyrocketed. Some scholars believe that this is the most important event in the post-Biblical history of Christianity.']
- [“We do have more evidence of the economic infrastructure of Rome being used to actually enrich its provinces to at least some capacity. Britain was staunchly mercantilist and capitalist in a way Rome was not and it's not easy to compare the two, hence why I said black and white. Rome was just as capable of, and easily did, destructive behavior. But I was mostly trying to push back against the psuedohistorical narrative of colonialism being largely good for places, and Rome was not colonial in the way that we know later empires such as Britain to be. Rome for the most part allowed provinces to do their own thing so long as they paid fealty and taxes to Rome and supplied them with other things there were no massive (or at least, as massive) economic disruptions to native populations. I mean I'd love to see sources on the increased quality of life claims but they don't really make much sense in the

context of numerous other sources that document with statistics how wealth was leaving places on huge scales."]

- Negative Clusters & Discussions

```
[[' roman', ' ity', ' empire', ' rome', ' war', ' romans', ' itian', ' study', ' army', ' history'], [' people', ' rome', '
like', ' history', ' study', ' dont', ' think', ' life', ' war', ' time'], [' nuclear', ' youre', ' post', ' physics', ' consider', '
comment', ' laws', ' uranium235', ' furious', ' defy'], [' transgender', ' trans', ' gender', ' 4500', ' ago', ' roles', '
thirdgender', ' priests', ' texts', ' document'], [' jesus', ' church', ' god', ' jews', ' christianity', ' jewish', ' bible', '
christian', ' catholic', ' christ']]
```

- [“Look back at ancient Rome and the gladiators. Even with the knowledge that stepping into the arena could mean death, people wanted to be gladiators.   There will always be people willing to risk their health as an entertainer because of the tremendous financial rewards. "]
- [" The existence of Christians in Rome 100 years after the Death of Jesus is evidence of Jesus existing? Absolutely, in tandem with mentions from non-Christian sources mere decades after his existence, the evidence of an oral tradition, and the culmination of the Testaments.   You're seriously holding that up? I am, along with the overwhelming majority of other people that study this period of history for a living. You know, the experts in this field. But if angry atheist blogs and youtube vids gain serious academic standing we might be in danger.   The truth of the matter is that for the majority of the people from antiquity that we learn about, Jesus has some of the least amount of evidence supporting his existence. And how much primary source work have you done to determine this? Because again, any read through most primary sources is going to be pulling up name after name after name of people that came decades to centuries beforehand. You have fundamentally flawed understanding of the historiofigurey of the ancient world.   "]
- ['It does not. They\'re called that because of the dramatic drop in quality of life and especially record keeping after the fall of Rome. The people that called it the dark ages were themselves Catholics. The "ignorance" that they saw in that era was besides religion.']

Conclusion

Rome came 6th as the number of foreign visitors, 8th for the percentage of positive comments, 7th for GDP and e-commerce active, 5th for population and internet access. It was extremely hard to find any topics related to Rome’s cost and quality of life as most of the online discussions collected were about Rome’s history and the Roman Empire, some are history based and others are religious based, and that’s exactly the point, Rome is considered to be an open air history museum, people usually travel or visit or relocate to

Rome due to the unprecedent culture the city has to offer, Rome is always the place for the curious, wanderers and explorers.

4.2.10 Stockholm

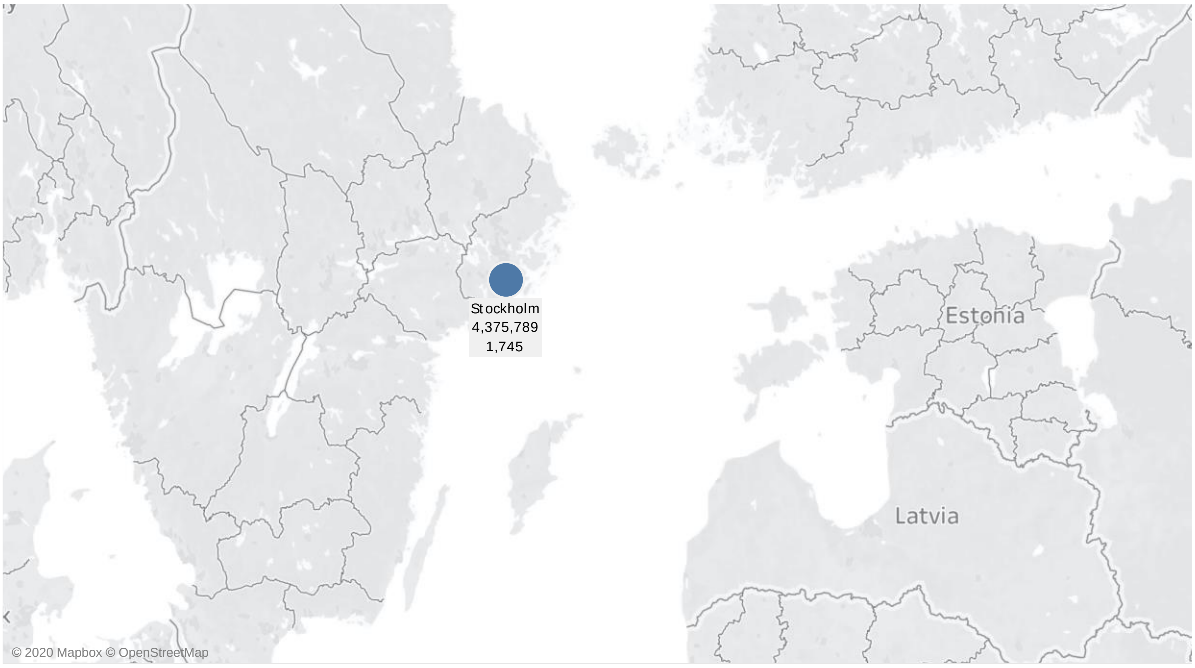


Figure 4. 119 *Geographical representation for the number of the words and comments collected for the city of Stockholm*

We managed to collect 1859 comment, with a total of 4,787,336 words, 1754 are unique values, with a total of 4,375,789 words to process and analyze.

4.2.10.1 Text Processing and Visualization

After processing the text, we now have 391,677 words to analyze. In the following table, the processed word has a total average of 220 comments, while not processed texts has around 2400 words as average. The longest processed comment has 1280 words.

Table 4. 20 *Stockholm text descriptive analysis*

Before Processing	After Processing
count 1745.000000	count 1745.000000
mean 2415.222923	mean 223.446991
std 2275.331630	std 211.785273
min 30.000000	min 3.000000

25% 770.000000	25% 73.000000
50% 1598.000000	50% 146.000000
75% 3333.000000	75% 306.000000
max 10968.000000	max 1280.000000

We can see the words per comments length represented in the following figures

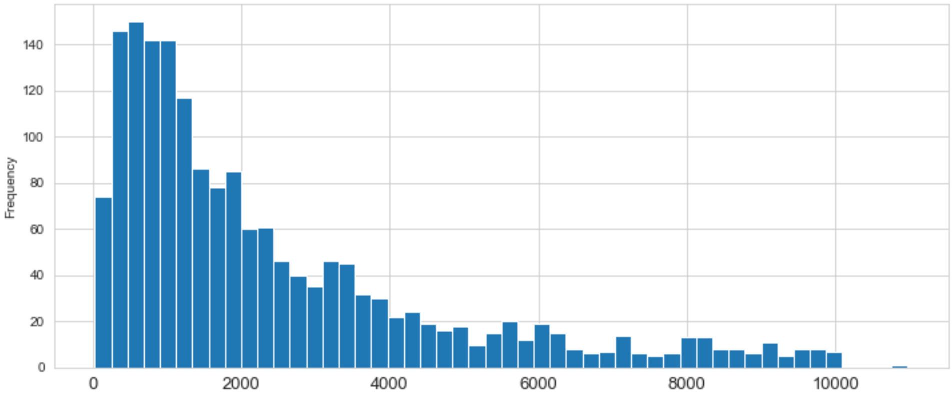


Figure 4. 120 *Stockholm words per comments before processing*

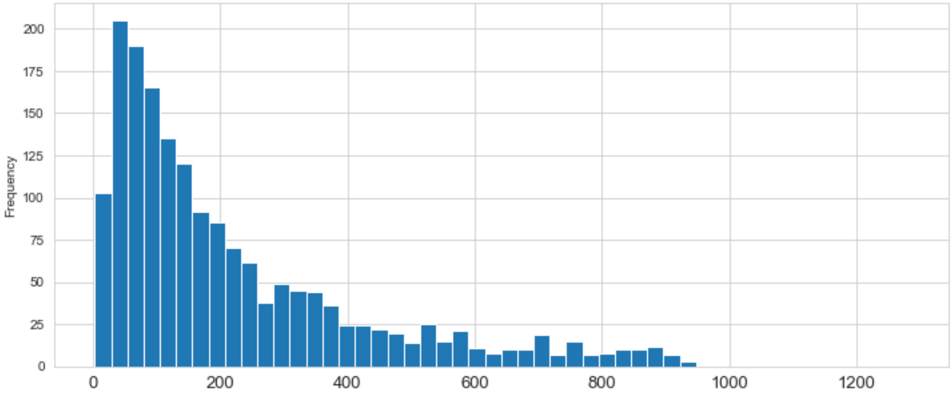


Figure 4. 121 *Stockholm words per comments after processing*

In the following figure, we can notice that people usually discuss words like study, like, university, Swedish and work.

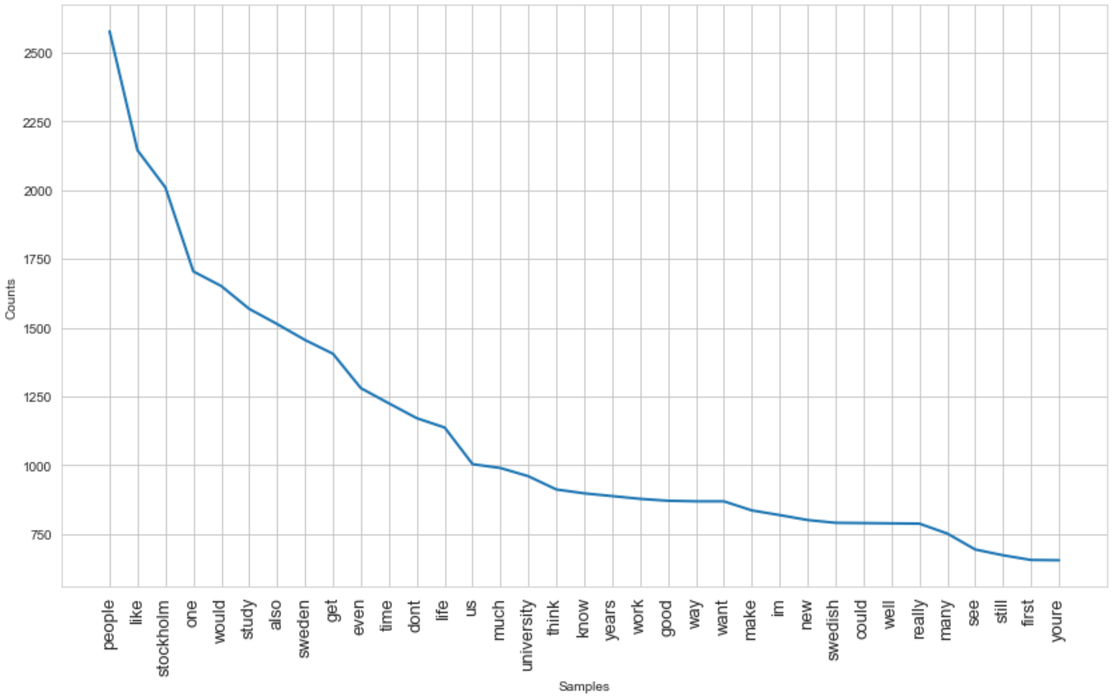


Figure 4. 122 *Stockholm 35 most frequent words for all sentiments*

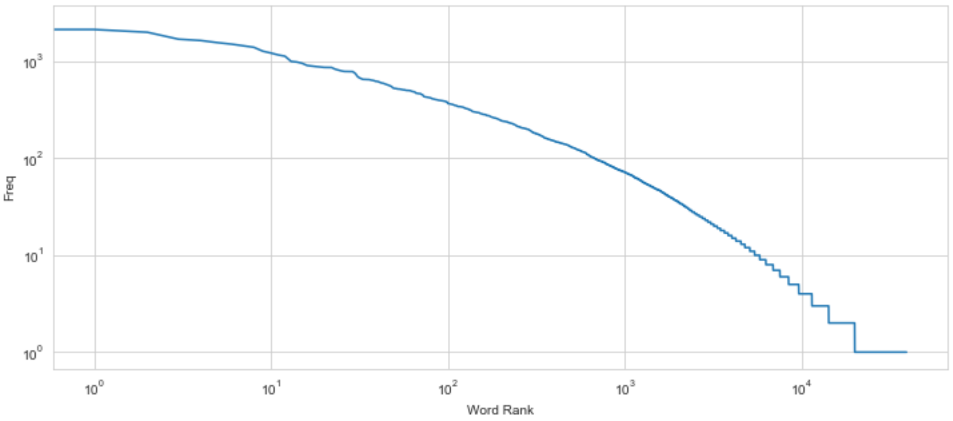


Figure 4. 123 *Stockholm word counts distribution on a logarithmic scale*

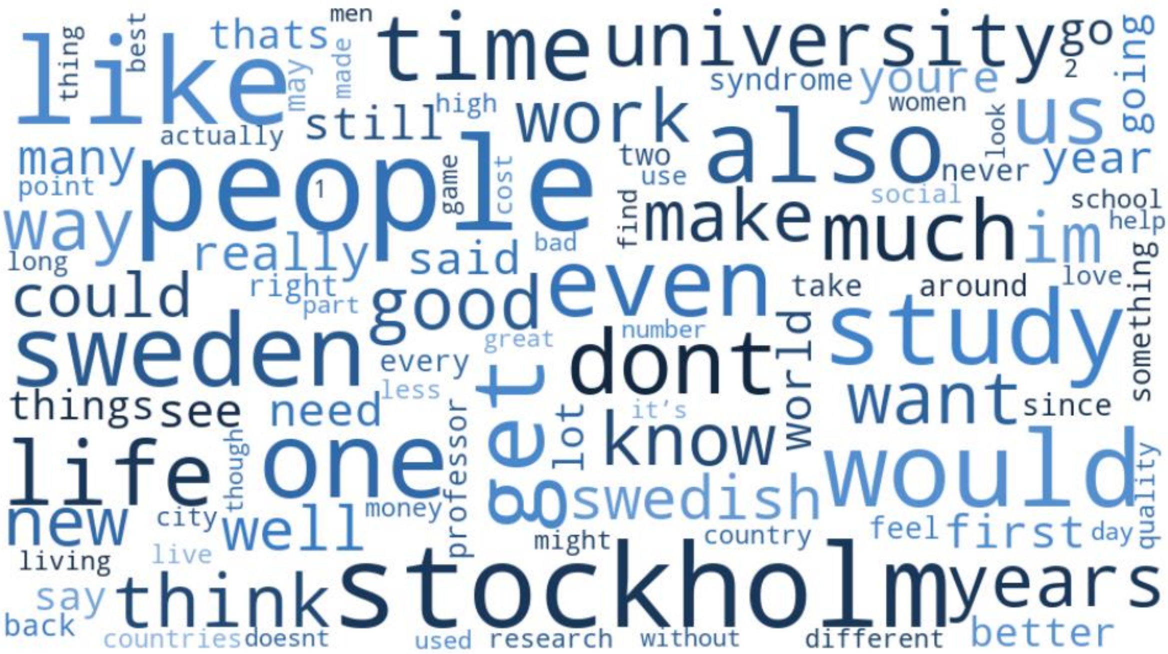


Figure 4. 124 *Stockholm word cloud on all sentiments*

4.2.10.2 Sentiment Analysis & Words Clustering

it’s time to use Sentiment Analysis on the text to label it. In the next table we have the results of the analysis. 62.55% of the comments collected are positive, 35.45% are negative and 2% are neutral as represented in the figure.

Table 4. 21 *Stockholm sentiment results*

Negative (-1)	Positive (1)	Neutral (0)	Total
618	1091	36	1745

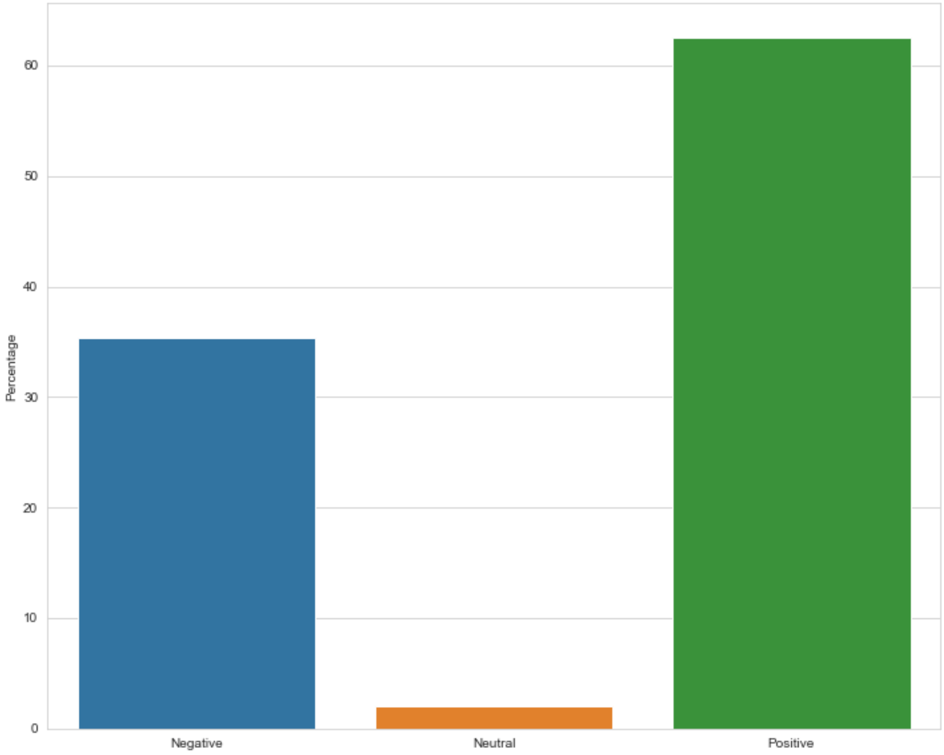


***Figure 4. 125** Stockholm sentiment results in percentage (%)*

In the following figure, the analysis tells us that people talked in a positive way about similar topics as before.

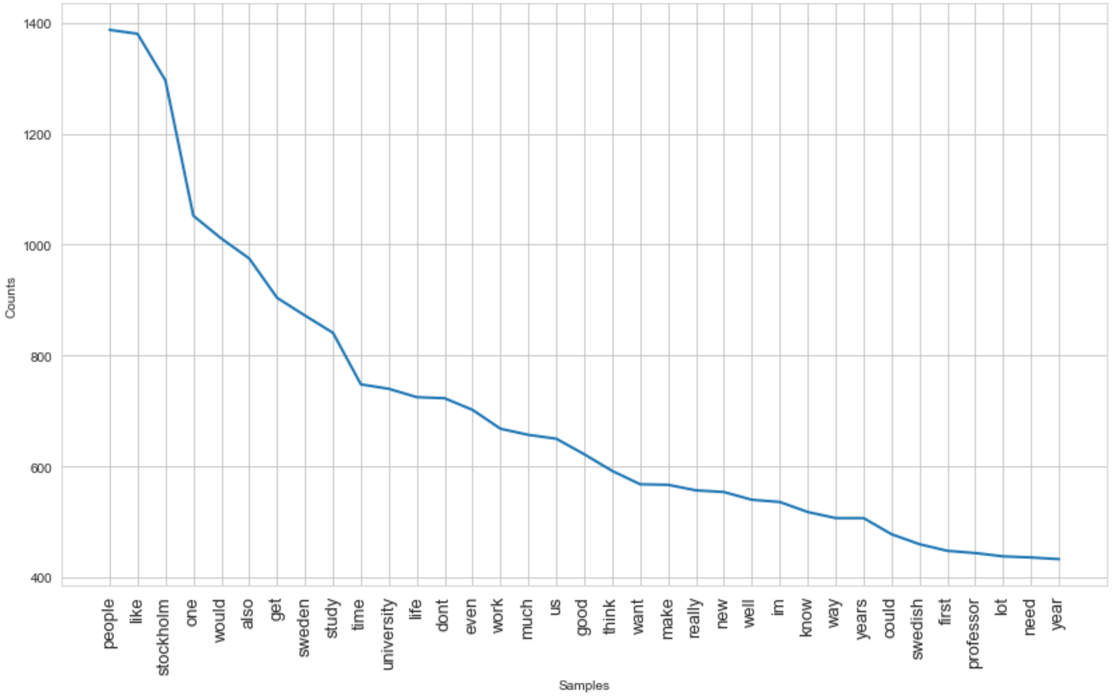


***Figure 4. 126** Stockholm 35 most frequent words for positive sentiment*

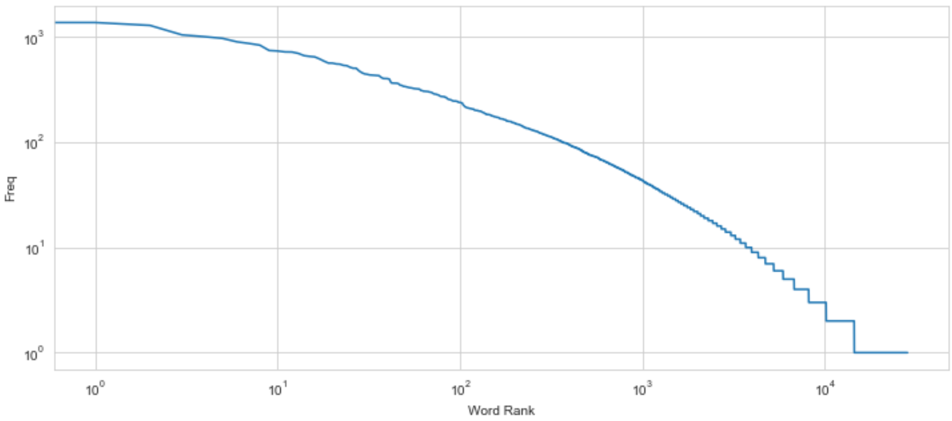


Figure 4. 127 *Stockholm word counts distribution for positive sentiment on a logarithmic scale*

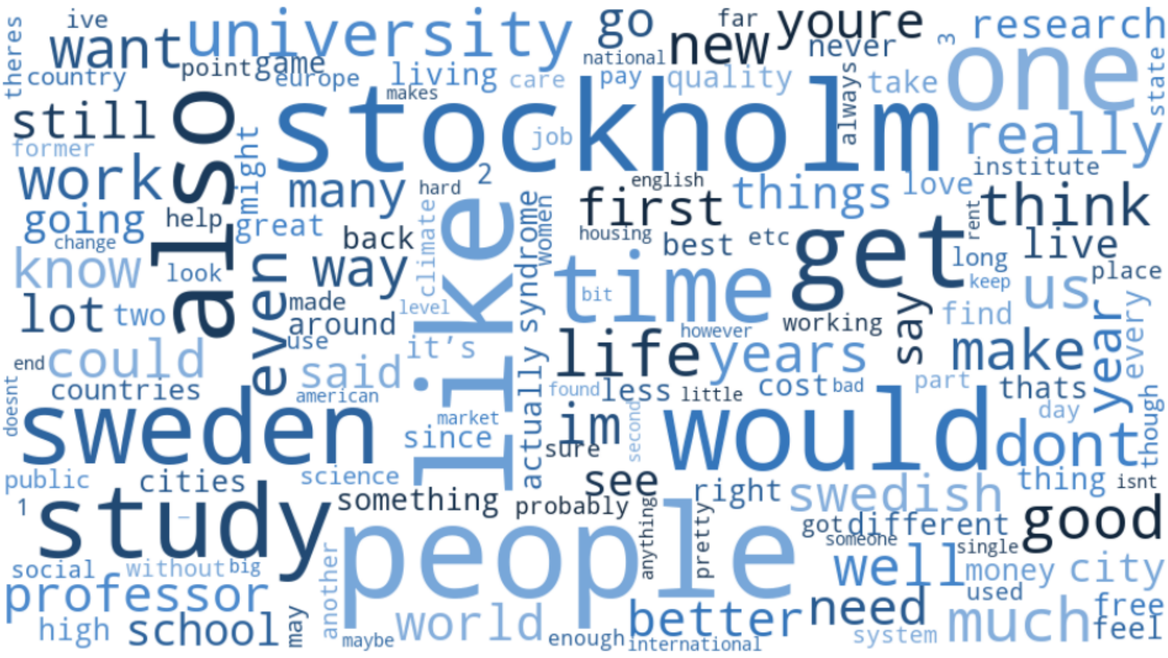


Figure 4. 128 *Stockholm word cloud on positive sentiments*

For the negative sentiment, the analysis tells us that the only difference from before is when people attach Stockholm to the Stockholm Syndrome.

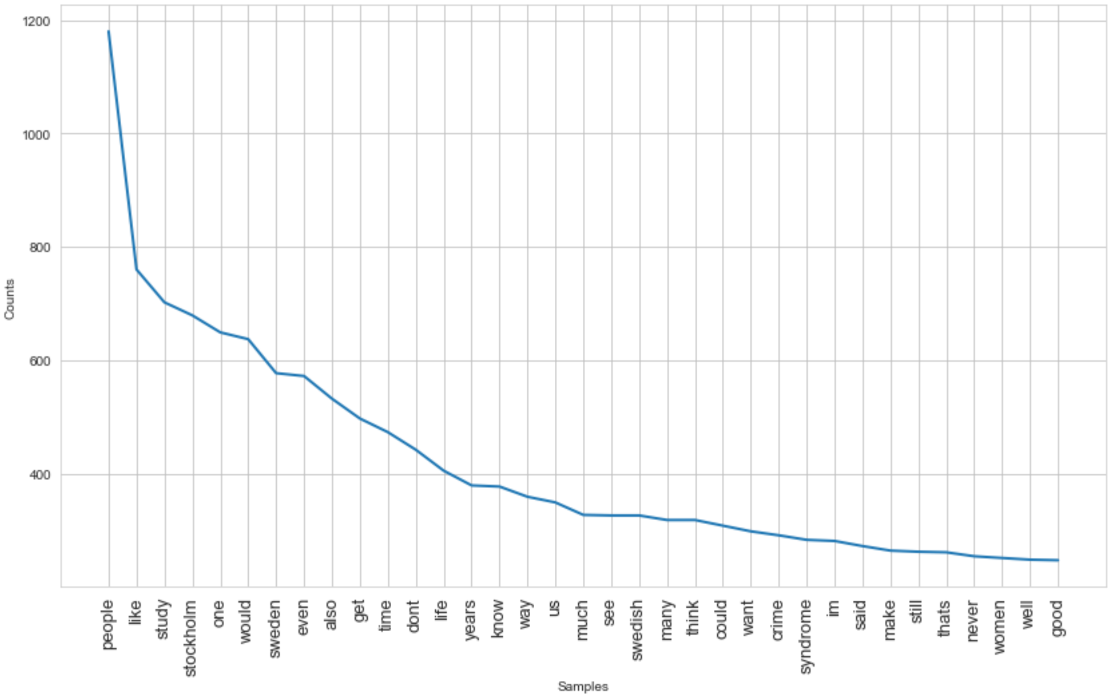


***Figure 4. 129** Stockholm 35 most frequent words for negative sentiment*

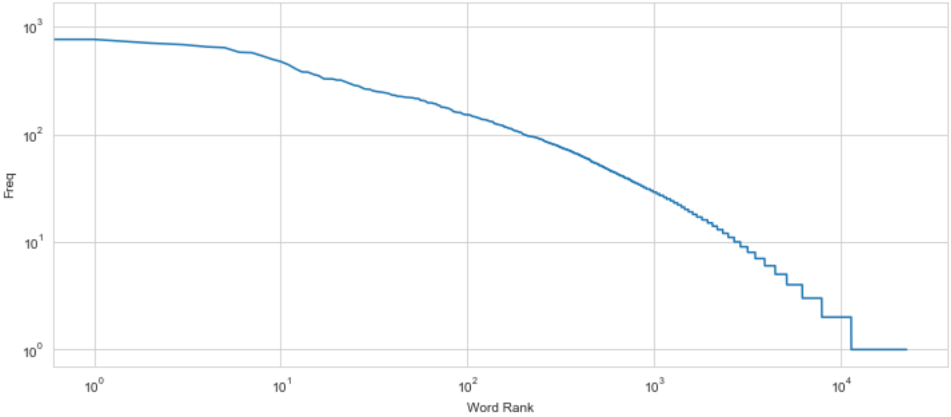


***Figure 4. 130** Stockholm word counts distribution for negative sentiment on a logarithmic scale*

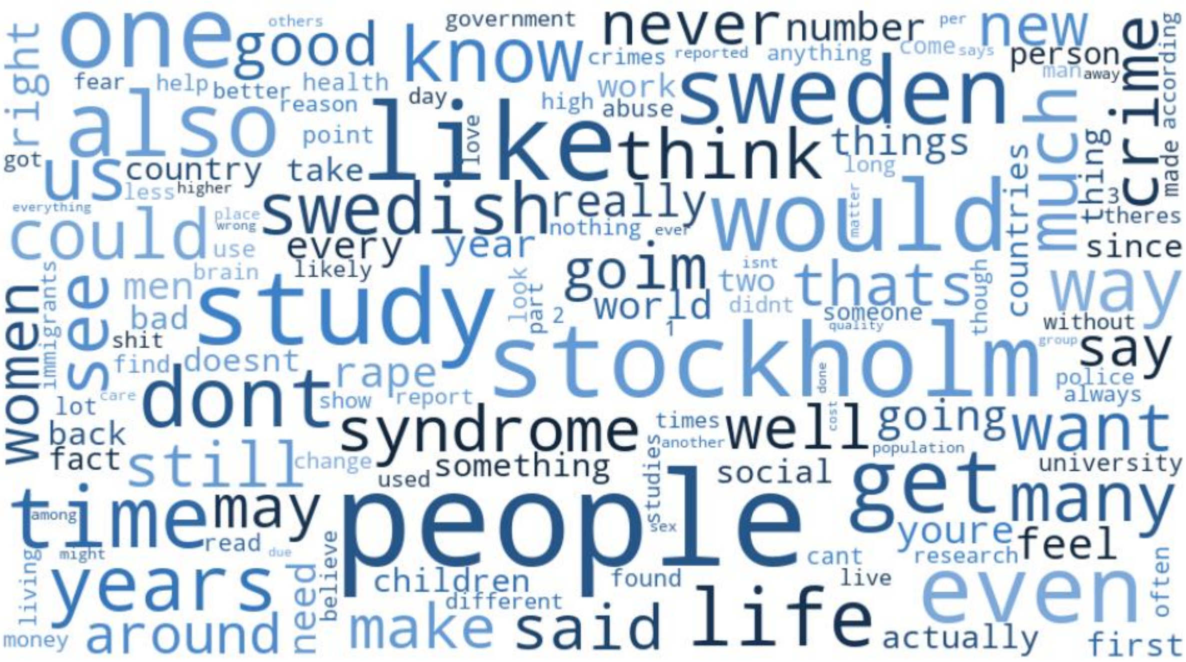


Figure 4. 131 *Stockholm word cloud on negative sentiments*

In the next part, word clustering results is carried out.

- Positive Clusters & Discussions

[[' game', ' like', ' people', ' im', ' time', ' dont', ' life', ' youre', ' really', ' love'], [' study', ' people', ' stockholm', ' like', ' syndrome', ' women', ' good', ' dont', ' time', ' think'], [' sweden', ' swedish', ' university', ' school', ' english', ' study', ' student', ' uppsala', ' stockholm', ' want'], [' living', ' life', ' rent', ' people', ' sweden', ' cost', ' quality', ' housing', ' like', ' city'], [' professor', ' university', ' emeritus', ' science', ' retired', ' atmospheric', ' research', ' institute', ' climate', ' sciences']]

- ["Sweden and especially Stockholm is a bit more pricy than other destinations, Stockholm specifically has a shortage of housing so prices are high. I've heard that the party life is pretty good but its not as wild as one of the student cities. (E.g.. Örebro or Lund) A pint can range pretty wildly, best price will be 30 sek (3 euro) but often closer to 4.5 euro. If cost is a huge concern I could understand going somewhere cheaper. That said, Sweden is absolutely gorgeous. Stockholm has a lot of cool cultural destinations and great atmosphere. So even if you can't study here I hope you get the chance to visit! (Ps. It's gonna get cold, so pack the good jackets)"]
- ['Its a very reasonable concern, unfortunately its true that you will likely not have much luck in Stockholm, the bigger the city is the less help people are willing to give. You really should look up something called "folkhögskola", it\'s a special type of school we have in Sweden that is great at giving you a chance to work on your life along with studies. Many of them offer housing on site at a low cost

with the only requirement that you do your part in school. Not being great at Swedish might make it harder to find one that suits you so you might want to look for someone who can help you with it, maybe someone at your SFI classes? A teacher or a guidance councelor? Best of luck bro, you\'ll find a way.']

- ['Having a lot of friends in Sweden, I find it is idealized in Canada compared to reality. While still having a very good quality of life, crime has been spiking, waiting lists to get an appartment in Stockholm span decades, and the sun setting down in mid-afternoon gets pretty depressing.']

- Negative Clusters & Discussions

```
[[' test', ' dont', ' pass', ' study', ' apply', ' stockholm', ' lessons', ' syndrome', ' driving', ' email'], [' sweden', '
crime', ' crimes', ' rape', ' swedish', ' immigrants', ' backgrounds', ' suspected', ' foreign', ' people'], [' study', '
sweden', ' people', ' swedish', ' stockholm', ' suicide', ' university', ' health', ' years', ' sex'], [' conservative', '
authoritarianism', ' conservatives', ' trump', ' fear', ' gop', ' brain', ' winning', ' cultural', ' social'], [' people', '
like', ' life', ' dont', ' syndrome', ' time', ' think', ' stockholm', ' want', ' know']]
```

- [" Some facts about Sweden and the migration crisis. The official figures show a population of 8 million in 1969 and a projected population of 10 million by 2017, with (on current growth rates) the population reaching 11 million by 2024. This requires Sweden at normal levels of population increase to be building 71,000 new residences a year to meet the needs of the country by 2020, or 426,000 new residences in total by that date. Although there is a presumption that the Swedish people, like their political elites, were always in favour of such migration, the facts suggest otherwise. In 1993 the newspaper Expressen broke one of the great taboos of Swedish politics and published a rare opinion poll on the country's actual views. Under the headline 'Throw them out' the paper revealed that 63 per cent of Swedish people wanted immigrants to go back to their home countries. An accompanying article by the paper's editor-in-chief, Erik Månsson, noted that, 'The Swedish people have a firm opinion on immigration and refugee policies. Those in power have the opposite opinion. It does not add up. It is an opinion bomb about to go off. That is why we are writing about this, starting today. Telling it just like it is. In black and white. Before the bomb goes off.'
- ["25% on most things in sweden. sweden is not a good place to live in if you want to save up money or make big money. its a good place to live if you're okay with being part of a middle class that has a pretty good quality of life. however if you want to live in the cities (stockholm etc) its pretty shit either way because the rent is insane compared to what you get and where you get it. ive moved away from sweden multiple times because ive grown tired of the absurdly high taxes and the high cost of living. ill probably move away from sweden again in the coming year because it just does not feel worthwhile to work and live here when i can make almost twice as much by moving elsewhere."]

4.2.10.3 Conclusion

Stockholm online discussion was rich of multiple topics, it's important to know that Stockholm's positive sentiment percentage comes in last place in comparison to the other cities, it also comes last for the number of people visit the city, it has one of the smallest populations and therefore lower GDP, lower internet access and e-commerce active population than other cities. However, most of Stockholm online discussions is saying that the city has a very good and high standard quality of life, however, much like Dublin, there were a lot of complains about the cost of life and whether or not a person who lives there can actually be able to save money, there were also some negative sentiments about the availability of residential Real Estate, as per the comments, Sweden if open for migration and the country has a high growth rate in population in comparison to new residences to meet that demand. It's also important to notice that a lot of the results obtained are connected to the Stockholm syndrome, which is much like Milan and Madrid connected with Football clubs.

It's important to know that Sweden during the "migrant crisis" of 2015/16, the country took in over 150,000 asylum seekers and received the most asylum applications per capita of any European country with over 750,000 asylum applications have been filed since 2000.

Chapter 5 Results & Future Work

5.1 Results

This dissertation aimed to shed light on the role of online discussion analysis and on its possible contribution to the understanding of cities' perception. Based on quantitative and qualitative analysis, it has been concluded that cities' indicators are not sufficient in order to comprehensively evaluate cities' attractiveness, due to a discrepancy measured and identified between those indicators and online discussions' analysis. Therefore, the analysis of online discussions affirms its capability to obtain a deeper understanding of cities' attractiveness when linked to cities' indicators, as stated in the initial hypothesis. Ultimately, extracting pivotal information could enable policymakers in identifying problems and making data-driven decisions.

For further elaboration, in the following chart, the number of comments collected per each city is presented, in comparison to 'GDP' and 'foreign_visitors' indicators. It is evident that London and Paris have the highest number of comments collected, as well as, ranking in the first two places for 'GDP' and 'foreign visitors' indicators. Barcelona, in contrast, has a similar number of foreign visitors per year as London, with a significantly lower number of comments collected in comparison to London. This does not indicate that Barcelona is less popular than London, instead, it may suggest that the politically strong position of London on the world map as a leading economic and business center makes it more likely for people to speak about it.

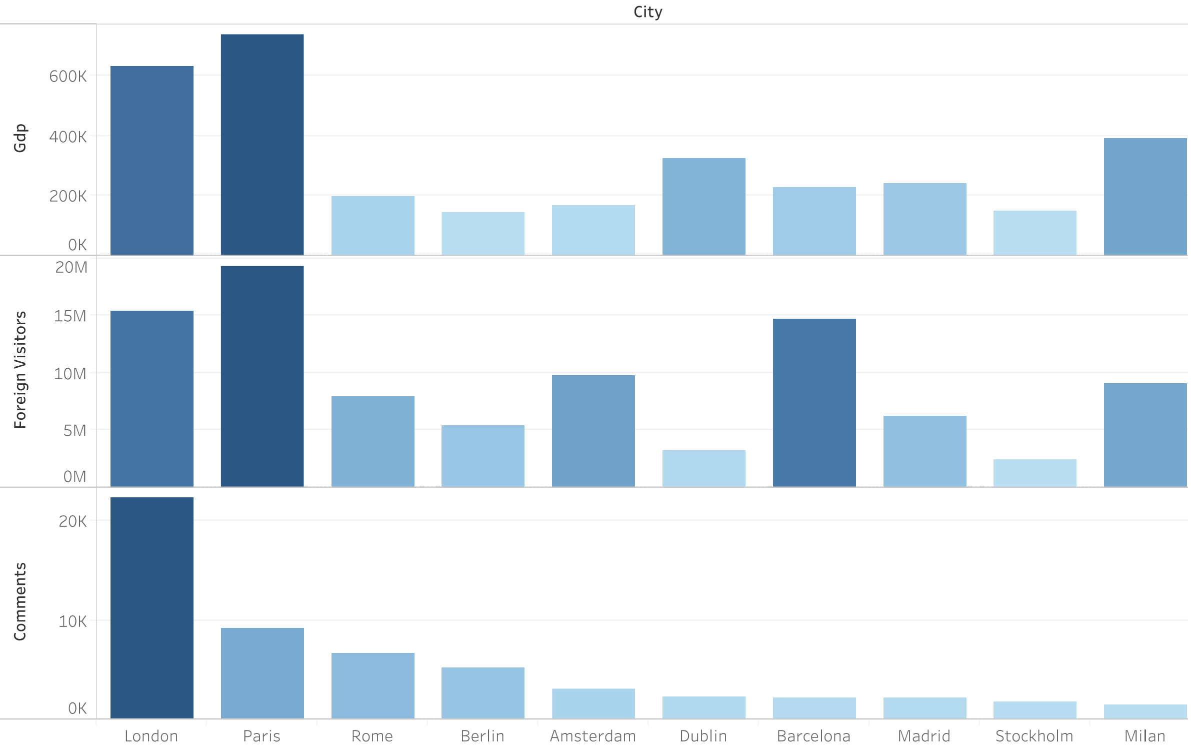


Figure 5. 1 *A comparison between the number of comments collected, foreign visitors, and GDP for each city*

In the following figure, the ratio between the positive and negative comments is compared to 'GDP' and 'foreign visitors' indicators. The results show that Barcelona, Madrid, and Milan are the first 3 places to represent the highest positive sentiment ratio, which indicates that the number of comments, overall, is not necessarily a positive indicator.

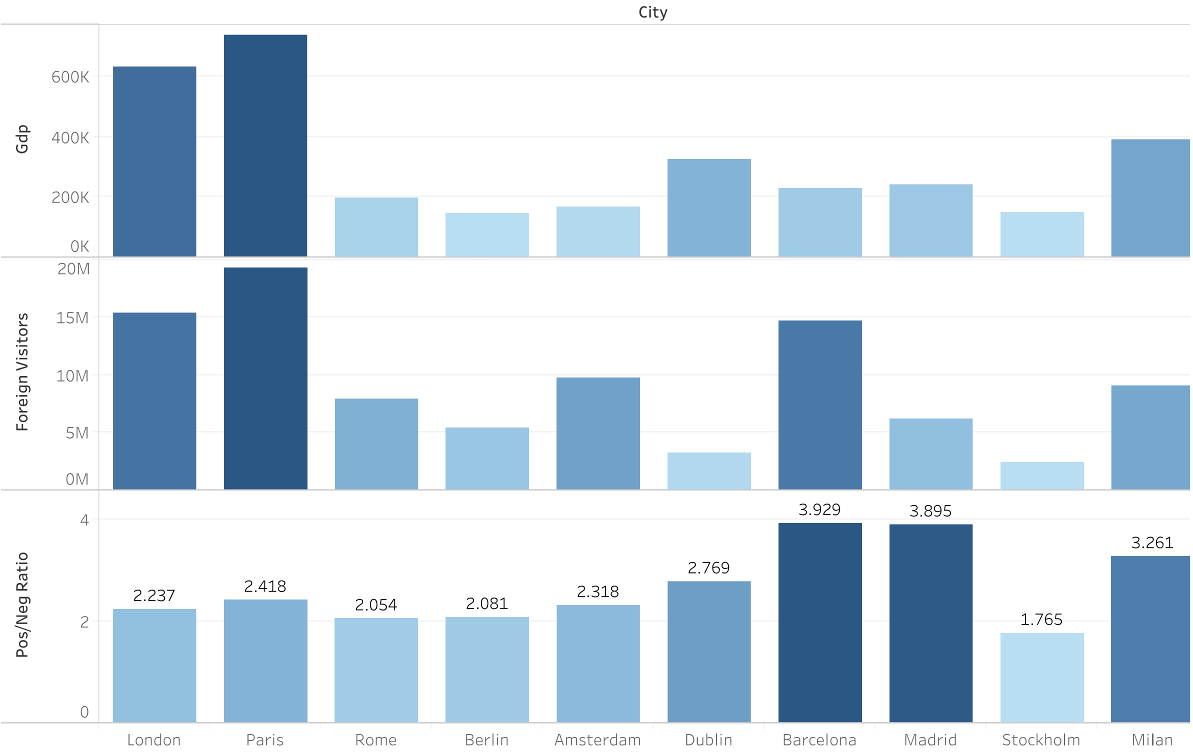


Figure 5. 2 *A comparison between the ratio between the positve and negative comments number, foreign visitors, and GDP for each city.*

In the following figure, the number of comments collected is compared to the ratio between the positive and negative comments for further elaboration and visualization. A higher number of comments, predominantly, does not imply a higher number of positive comments.

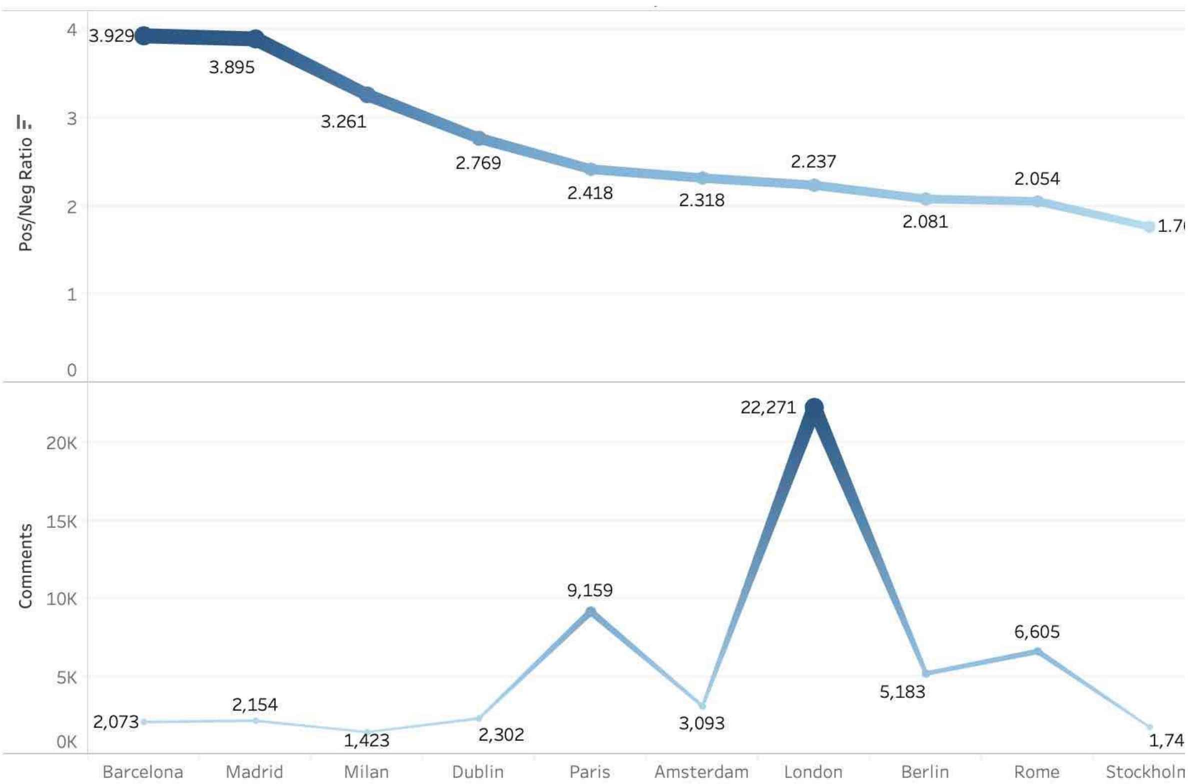


Figure 5. 3 *A comparison between the number of comments collected and the ratio between the positve and negative comments number for each city.*

In the following two figures, the main findings from the text mining and sentiment analysis are presented. The 10 clusters of the most frequent words for both, the positive and the negative sentiment for each of 10 selected cities are demonstrated. For instance, the results of Amsterdam indicate that people are generally satisfied with the cost and quality of life, and the policy of legalization of gender identity. In contrast, concerns are directed towards the legalization of cannabis and its consequence on people's mental health and the city's livability, as well as, legalization of prostitution. Moreover, Barcelona achieved the highest positive sentiment ratio, most of the discussions imply the low cost of living, in contrast, a lot of concerns are focused on the low wages and the Catalonia independence issues. For Berlin, an increased number of concerns for terrorism and labor migration is shown. As for Dublin, the discussion is focused on the high quality of life in the city, whilst, concerns about the cost of living are stated, some of the comments are indicating that living near the city of Dublin is a more convenient option.

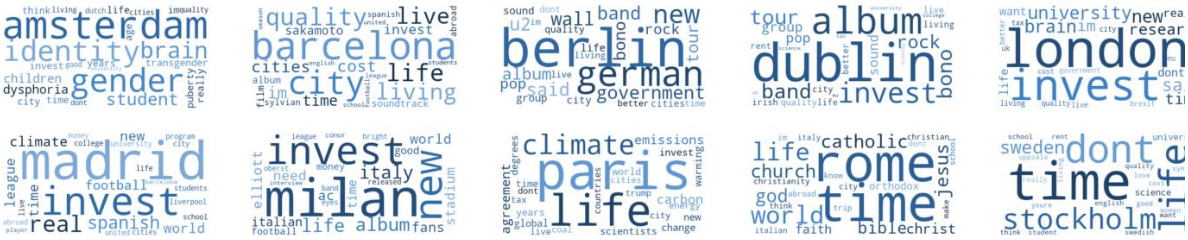


Figure 5. 4 *10 clusters of the most frequent words for the positive sentiment*

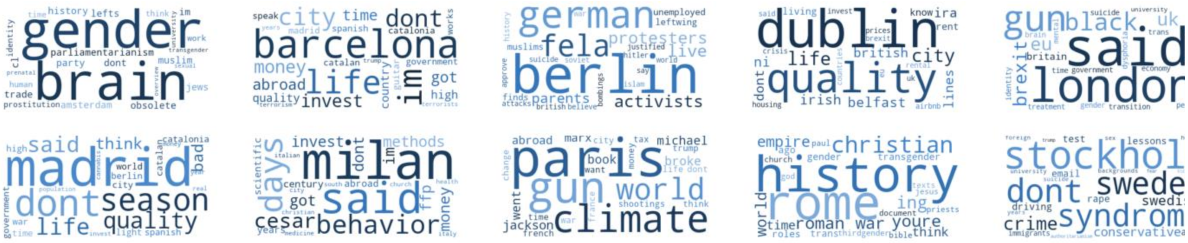


Figure 5. 5 *10 clusters of the most frequent words for the negative sentiment*

As previously mentioned, London has the highest number of comments collected, the discussion is mainly related to the Post-Brexit situation and its effects on the quality of life in the city. Madrid was one of the best results of positive sentiments, the results were mainly related to football and the high quality of life. Similarly, Milan's discussion is related to football and the good quality of life, however, concerns were focused on the cost of living and the strain of English speakers relocating to Milan for work due to the language barrier. For Paris, discussions are focused on the gun spread issues in the country, climate change, and global warming topics, along with other discussions about the degree of safety for women and the cost of life, commuting or studying in the famous city. For Rome, most of the topics obtained were related to either the history of the city, the Roman Empire, or religion, indicating that Rome's brand perception will always be attached to its culture. Lastly, Stockholm online discussion is considering the high standard and the quality of life, while concerns are directed towards the cost of life, asylum seekers, migration and its effects on the availability of residential Real Estate.

The significance of these findings is that they confirm the hypothesis of this dissertation. Cities that have evident and clear insights from online discussions are able to better understand their stakeholders' concerns, needs, requests, and preferences. Furthermore, cities that are able to leverage this information and insights are ultimately able to support their decision-making process based on a significant amount of text data.

5.2 Future Work

Entities and public bodies might have a large number of opinions and feedback posted on social media; however, it is still impossible to analyze it manually without any error or bias. Therefore, for future work, the use of different sentiment analysis techniques on the same sample of text and comparing the results together is recommended. As well as, the use of different online platforms such as Facebook, Twitter, and Quora, to not only understand people's opinions online but to understand if those opinions change by changing the type of platform analyzed. In addition, in order to be able to undertake econometric analysis, analyzing the remaining 32 cities using text mining and sentiment analysis is as well required.

Anxiety, excitement, hope, and skepticism are some of the aspects that expected to be measured in the future using sentiment analysis. There is general speculation that sentiment analysis needs to move beyond a one-dimensional scale that can only measure positive and negative sentiment. Thus, entities and public administrations should become more aware of the applications of sentiment analysis within their area of specialties for the days to come.

Bibliography

BBC, 2017. [Online]
Available at: bbc.com/news/uk-england-london-39806865

BBC, 2019. [Online]
Available at: https://www.bbc.com/news/world-europe-41584864

BBC, 2020. [Online]
Available at: https://www.bbc.com/news/world-europe-51567971

C40 Cities Climate Leadership Group, n.d. *BENEFITS OF CLIMATE ACTION,* s.l.: C40 Cities Climate Leadership Group.

Confindustria, 2019. [Online]
Available at: https://www.thelocal.it/20190218/these-are-the-thousands-of-job-vacancies-that-italy-cant-fill

Data-Driven Lab, 2018. *METRICS FOR SUSTAINABLE AND INCLUSIVE CITIES.* [Online]
Available at: https://datadrivenlab.org/wp-content/uploads/2018/12/2018_UESI_Full_Report.pdf

Economics Help, 2019. *Benefits of economic growth.* [Online]
Available at: https://www.economicshelp.org/macroeconomics/economic-growth/benefits-growth/

ecoRI News, 2020. *No Progress Made to Reduce U.S. Greenhouse-Gas Emissions.* [Online]
Available at: https://www.ecori.org/climate-change/2020/1/22/no-progress-made-to-reduce-us-greenhouse-gas-emissions

Euronews, 2020. [Online]

Available at: https://www.euronews.com/2020/02/15/amsterdam-mayor-considers-banning-tourists-from-buying-cannabis

Global Business Outlook, n.d. [Online]

Available at: https://www.globalbusinessoutlook.com/habitat-iii-that-once-every-20-years-global-urban-event/

Hutto, C. & Gilbert, E., 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.

Inside Climate News, 2020. *U.S. Emissions Dropped in 2019: Here's Why in 6 Charts.* [Online]

Available at: https://insideclimatenews.org/news/07012020/infographic-united-states-emissions-2019-climate-change-greenhouse-gas-coal-transportation

INTERNATIONAL ENERGY AGENCY, 2016. *Energy Technology Perspectives.* [Online]

Available at: https://datadrivenlab.org/wp-content/uploads/2018/12/2018_UESI_Full_Report.pdf

Learn By Doing, Inc., n.d. *Key Assumptions of OLS: Econometrics Review.* [Online]

Available at: https://www.albert.io/blog/key-assumptions-of-ols-econometrics-review/

Leipzig University, 2018. *Germans are becoming more and more intolerant.* [Online]

Available at: https://www.sueddeutsche.de/politik/auslaenderfeindlichkeit-muslime-studie-rechtsextremismus-1.4199261

Mohiuddin, A. & Al-Sakib Khan, . P., 2018. *Data Analytics: Concepts, Techniques, and Applications.* s.l.:s.n.

National Environment Health Association, n.d. *Built Environment and Climate Change.* [Online]

Available at: neha.org

NewClimate Institute, 2019. *GLOBAL CLIMATE ACTION FROM CITIES, REGIONS AND BUSINESSES.* [Online]

Available at: https://newclimate.org/wp-content/uploads/2019/09/Report-Global-Climate-Action-from-Cities-Regions-and-Businesses_2019.pdf

Pandey, P., 2018. *Simplifying Sentiment Analysis using VADER in Python (on Social Media Text).* [Online]

Available at: https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f

Pew Forum on Religion and Public Life, 2017. *The Growth of Germany's Muslim Population.* [Online]

Available at: https://www.pewforum.org/essay/the-growth-of-germanys-muslim-population/

Pyhälä, A., Fernández, L. Á. & Lehvävirta, H., 2016. Global environmental change: local perceptions, understandings, and explanations. *PubMed Central.*

The Independent, 2020. *Brexit uncertainty could cost UK economy £15bn by end of decade, report finds.* [Online]

Available at: https://www.independent.co.uk/news/uk/politics/brexit-economy-boris-johnson-trade-deal-eu-cost-rand-a9304356.html

The Journal, 2019. *Dublin is more expensive to live in than both London and Vienna.* [Online]

Available at: https://www.thejournal.ie/dublin-cost-of-living-index-4926640-Dec2019/

The Local France, 2017. *Five things to know about guns in France.* [Online]

Available at: https://www.thelocal.fr/20171004/five-things-to-know-about-guns-in-france

The UCWbL @ DePaul University, n.d. *Linear Correlation.* [Online]

Available at: https://condor.depaul.edu/sjost/it223/documents/correlation.htm

The University of California, San Diego, 2020. *Python for Data Science.* [Online]

Available at: https://www.edx.org/course/python-for-data-science-2

U.S. Department of Housing and Urban Development, n.d. *PD&R Edge.* [Online]

Available at:

https://www.huduser.gov/portal/pdredge/pdr_edge_featd_article_030915.html

United Nations, 2018. *Revision of World Urbanization Prospects.* [Online]

Available at: https://www.un.org/development/desa/publications/2018-revision-of-world-urbanization-prospects.html

URBAN Intergroup at the European Parliament, 2011. *The Economic Role of Cities,* s.l.: s.n.

Wikipedia, 2020. [Online]

Available at: https://en.wikipedia.org/wiki/Paris_Agreement

Wikipedia, n.d. *LGBT rights in the Netherlands.* [Online]

Available at: https://en.wikipedia.org/wiki/LGBT_rights_in_the_Netherlands

World Bank, 2020. *World Bank.* [Online]

Available at: https://www.worldbank.org/en/topic/urbandevelopment/overview#2

[Accessed 2020].

World Economic Forum, 2019. *Making Affordable Housing a Reality in Cities,* s.l.: s.n.

World Resources Institute, 2018. [Online]

Available at: https://www.wri.org/blog/2018/11/cities-can-save-17-trillion-preventing-urban-sprawl

Appendix

	Title	Year	Author(s)
1	Sustainable Urban Mobility Planning (SUMP) at subregional area level with the use of transportation model	2017	G Karoń, G Krawczyk, A Urbanek…
2	Spatial quality of natural elements and safety perception in urban parks	2015	A Hami, F Emami
3	Modelling uncertainties in long-term predictions of urban growth: a coupled cellular automata and agent-based approach	2015	AM El Saeid Mustafa, I Saadi, M Cools…
4	Ecosystem service implementation and governance challenges in urban green space planning—The case of Berlin, Germany	2015	N Kabisch
5	Spatial heterogeneity in the costs of the economic crisis in Europe: are cities sources of regional resilience?	2015	R Capello, A Caragliu, U Fratesi
6	Resilient communities. Social infrastructures for sustainable growth of urban areas. A case study	2017	A Boeri, D Longo, V Gianfrate…
7	Evaluating cultural routes for a network of competitive cities in the mediterranean sea: the eastern monasticism in western mediterranean area	2015	F Calabrò, D Campolo, G Cassalia…
8	Urban transport justice	2016	S Gössling
9	Analysing urban green space accessibility and quality: A GIS-based model as spatial decision support for urban ecosystem services in Brussels	2017	P Stessens, AZ Khan, M Huysmans, F Canters
10	Smart cities and smart mobilities	2017	P Geoffron
11	Urban peripheries as growth and conflict spaces: The development of new towns in China	2015	X Yang, J Day, SS Han

12	Towards an integrated evaluation approach for cultural urban landscape conservation/regeneration	2018	F Nocca, LF Girard
13	Spatial dynamics of population in Kolkata urban agglomeration	2015	V Yadav, RB Bhagat
14	Smart Governance: urban regeneration and integration policies in Europe. Turin and Malmö case studies	2015	C Testoni, A Boeri
15	Agglomeration economies and smart cities	2015	AMB Barufi, K Kourtit
16	Does the visibility of greenery increase perceived safety in urban areas? Evidence from the place pulse 1.0 dataset	2015	X Li, C Zhang, W Li
17	Urban diversity: Disentangling the cultural from the economic case	2015	B Vormann
18	Characterization of allergen emission sources in urban areas	2016	P Cariñanos, C Adinolfi…
19	The concept of smart city in the theory and practice of urban development management	2016	D Sikora-Fernandez, D Stawasz
20	Sustainability and competitiveness in Australian cities	2015	R Hu
21	Perspective of decreasing of road traffic pollution in the cities	2018	A Galkin, O Lobashov, S Capayova…
22	Transport disadvantage, car dependence and urban form	2016	G Mattioli, M Colleoni
23	Discovering urban activity patterns in cell phone data	2015	P Widhalm, Y Yang, M Ulm, S Athavale, MC González
24	Urban agriculture between pioneer use and urban land grabbing: The case of “Prinzessinnengarten” Berlin	2015	M Clausen
25	Community attachment and resident attitude toward old masonry walls and associated trees in urban Hong Kong	2015	AY Lo, CY Jim
26	Supporting sustainable transportation	2016	M Renner
27	Preference heterogeneity in mode choice based on a nationwide survey with a focus on urban rail	2016	Z Zheng, S Washington, P Hyland, K Sloan…
28	Uncovering urban human mobility from large scale taxi GPS data	2015	J Tang, F Liu, Y Wang, H Wang
29	Urban Nature: Perception and Acceptance of Alternative Green Space Management and the Change of Awareness after Provision of Environmental Information. A …	2017	PA Unterweger, N Schrode, O Betz
30	Imperatives for greening cities: A historical perspective	2017	Y Feng, PY Tan
31	Residents' preferred policy actions for shrinking cities	2016	MH Guimarães, LC Nunes, AP Barreira…
32	The “renaturation” of urban rivers: The case of the St Charles River in Quebec	2015	A Brun

33	Copro-necrophagous beetles (Coleoptera: Scarabaeinae) in urban areas: A global review	2016	L Ramírez-Restrepo, G Halffter
34	[HTML][HTML] Autonomous driving and urban land use	2016	D Heinrichs
35	What drives planning in a shrinking city? Tales from two German and two American cases	2017	K Pallagst, R Fleschurz, S Said
36	Urban transformation in Istanbul and Budapest: Neoliberal governmentality in the EU's semi-periphery and its limits	2015	E Akçalı, U Korkut
37	Perception and preference of trees: A psychological contribution to tree species selection in urban areas	2016	T Gerstenberg, M Hofmann
38	The economic development–vibrant center connection: Tracking high-growth firms in the DC region	2016	E Malizia, Y Motoyama
39	Spatial and welfare effects of automated driving: will cities grow, decline or both?	2019	G Gelauff, I Ossokina, C Teulings
40	Options and strategies for balanced development for liveable cities: an epilogue	2016	VP Singh, B Maheshwari, B Thoradeniya
41	A methodology to evaluate accessibility to bus stops as a contribution to improve sustainability in urban mobility	2019	MV Corazza, N Favaretto
42	Visioning the bus system of the future: stakeholders' perspective	2015	A Musso, MV Corazza
43	Participatory governance in smart cities: the urbanAPI case study	2017	K Soomro, Z Khan, D Ludlow
44	Local Fiscal Capability and Liberalization of Urban Hukou	2016	L Zhang, M Li
45	Identifying the spatial effects and driving factors of urban PM2. 5 pollution in China	2017	Z Cheng, L Li, J Liu
46	Exploring knowledge management perspectives in smart city research: A review and future research agenda	2019	J Israilidis, K Odusanya, MU Mazhar
47	13. Greener cities, a wild card for ticks?	2016	F Gassner, KM Hansford…
48	Facilitating the selection of city logistics measures through a concrete measures package: A generic approach	2016	K Papoutsis, E Nathanail
49	Travel Demand Management (TDM) case study for social behavioral change towards sustainable urban transportation in Istanbul	2017	İ Batur, M Koç
50	Understanding and managing visitor pressure in urban tourism	2017	K Koens, A Postma
51	Global cities–cities changed social-cultural space	2017	OY Matveeva
52	Economically incentivising smart urban regeneration. Case study of Port Louis, Mauritius	2018	Z Allam, P Newman

53	Rules of the roost: characteristics of nocturnal communal roosts of rainbow lorikeets (Trichoglossus haematodus, Psittacidae) in an urban environment	2015	AK Jaggard, N Smith, FR Torpy, U Munro
54	The choice of means of transport and daily movements in urban environment	2016	L Zajickova, V Vozenilek…
55	Demography decline, population aging, and modern financial approaches to urban policy	2018	G Carbonaro, E Leanza, P McCann…
56	Dynamics of development of the largest cities- Evidence from Poland	2019	C Kowalczyka, J Kil, K Kurowska
57	Do urban tourism hotspots affect Berlin housing rents?	2017	P Schäfer, J Hirsch
58	Spatial inequalities in big Indian cities	2015	P Sidhwani
59	Metropolisation and the evolution of systems of cities in the Czech Republic, Hungary and Poland since 1950	2015	N Zdanowska
60	Repositioning cities through star architecture: how does it work?	2018	N Alaily-Mattar, J Dreher, A Thierstein
61	Rescaling and refocusing smart cities research: From mega cities to smart villages	2018	A Visvizi, MD Lytras
62	The dynamic evolution and moving tracks of the center of gravity for the spatial pattern of China's urban development	2018	C Fang, D Yu, H Mao, C Bao, J Huang
63	The Limits to Artist-Led Regeneration: Creative Brownfields in the Cities of High Culture	2016	L Andres, O Golubchikov
64	Urban shrinkage in Germany: An entangled web of conditions, debates and policies	2017	A Nelle, K Großmann, D Haase, S Kabisch, D Rink…
65	An autonomous taxi service for sustainable urban transportation	2017	D Nicolaides, D Cebon, J Miles
66	Valuing co-benefits to make low-carbon investments in cities bankable: the case of waste and transportation projects	2017	K Rashidi, M Stadelmann, A Patt
67	City boosterism and place-making with light rail transit: A critical review of light rail impacts on city image and quality	2017	F Ferbrache, RD Knowles
68	The impact of urban growth patterns on urban vitality in newly built-up areas based on an association rules analysis using Geographical 'big data'	2018	Q He, W He, Y Song, J Wu, C Yin, Y Mou
69	Theoretical approach to the study of quality of life in rural and urban settlements	2016	R Goran, B JELISAVKA
70	Commute mode diversity and public health: A multivariate analysis of 148 US cities	2018	C Frederick, W Riggs…
71	Is social polarization related to urban density? Evidence from the Italian housing market	2018	V Antoniucci, G Marella
72	Health and climate related ecosystem services provided by street trees in the urban environment	2016	JA Salmond, M Tadaki…

73	A dynamic analysis of tourism determinants in Sicily	2015	D Provenzano
74	A study of the effect of a high-speed rail station on spatial variations in housing price based on the hedonic model	2015	B Geng, H Bao, Y Liang
75	A 4-dimensional model and combined methodological approach to inclusive urban planning and design for ALL	2019	N Rebernik, BG Marušić, A Bahillo, E Osaba
76	How infrastructure can promote cycling in cities: Lessons from Seville	2015	R Marqués, V Hernández-Herrador…
77	" New" direction of urban development from a Central European perspective	2016	Z Bujdoso, T Kovács, S CSABA, Z Brambauer
78	Strategic foresight for (coastal) urban tourism market complexity: The case of Bournemouth	2016	S Carlisle, A Johansen, M Kunc
79	Strategies to improve sustainability in urban landscape, literature review	2016	M Rakhshandehroo, MJM Yusof…
80	Toolbox: Measurable performance information based tools for co-creation of resilient, ecosystem-based urban plans with urban designers, decision-makers …	2016	FHM Van De Ven, RPH Snep, S Koole…
81	ASSESSMENT OF SHORT-TERM VACATION CONDITIONS IN URBAN AGGLOMERATIONS'OF KAZAKHSTAN	2015	A ABILOV, V YASKEVICH, M JUNUSSOVA…
82	Digital urbanisms: Exploring the spectacular, ordinary and contested facets of the media city	2015	J Vuolteenaho, K Leurs, J Sumiala
83	Financing the Urban Transition: Policymakers' Summary. Coalition for Urban Transitions	2017	G Floater, D Dowling, D Chan…
84	Tales of transforming cities: Transformative climate governance capacities in New York City, US and Rotterdam, Netherlands	2019	K Hölscher, N Frantzeskaki, T McPhearson…
85	Towards sustainable urban transport in Singapore: Policy instruments and mobility trends	2019	M Diao
86	Smart city governance: A local emergent perspective	2016	A Meijer
87	Tourism, transformation and urban ethnic communities: The case of Matonge, Brussels	2015	A Diekmann, I Cloquet
88	Opportunities and benefits of green balconies and terraces in urban conditions	2017	E Mladenović, M Lakićević, L Pavlović…
89	The issues of infill development in cities of the Tyumen region	2018	D Izvin, V Lez'Er, A Kopytova
90	Cities Network Along the Silk Road	2017	P Ni, M Kamiya, R Ding
91	[HTML][HTML] Willingness to commute among future physicians: a multicenter cross-sectional survey of German medical students	2018	J Quart, T Deutsch, S Carmienke…

92	Contribution for an urban geomorphoheritage assessment method: proposal from three geomorphosites in Rome (Italy)	2017	A Pica, GM Luberti, F Vergari, P Fredi…
93	[HTML][HTML] From shacks to skyscrapers: multiple spatial rationalities and urban transformation in Accra, Ghana	2016	L Fält
94	Urban regeneration in historic downtown areas: an ex-ante evaluation of traffic impacts in Athens, Greece	2015	K Kepaptsoglou, MG Karlaftis, I Gkotsis…
95	Towards the integration of urban sound planning in urban development processes: the study of four test sites within the SONORUS project	2015	S Alves, L Estévez-Mauriz, F Aletta…
96	Perception of density by pedestrians on urban paths: An experiment in virtual reality	2018	D Fisher-Gewirtzman
97	Smart Mobility in Italian Metropolitan Cities: A comparative analysis through indicators and actions	2018	R Battarra, C Gargiulo, MR Tremiterra…
98	Measurement of intellectual capital of Lithuanian cities by a composite index	2015	R Krušinskas, J Bruneckienė
99	City-brand building–From city marketing to city branding	2017	B Melović, S Mitrović, A Djokaj
100	City Branding in China's Northeastern Region: How Do Cities Reposition Themselves When Facing Industrial Decline and Ecological Modernization?	2018	M Han, M De Jong, Z Cui, L Xu, H Lu, B Sun
101	The power of communities in smart urban development	2016	K Borsekova, A Vanova, K Vitalisova
102	The act of (future) cycling: testing urban designs and conducting research in virtual reality	2015	G de Leeuw, J de Kruijf
103	Measuring urban competitiveness: ranking European large urban zones	2017	L Sáez, I Periáñez, I Heras-Saizarbitoria
104	The influence of transport infrastructure development on sustainable living environment in Lithuania	2016	A Griškevičiūtė-Gečienė, D Griškevičienė
105	The role of attitudes, transport priorities, and car use habit for travel mode use and intentions to use public transportation in an urban Norwegian public	2015	Ö Şimşekoğlu, T Nordfjærn, T Rundmo
106	Economics of small wind turbines in urban settings: An empirical investigation for Germany	2015	B Grieser, Y Sunak, R Madlener
107	Adaptation of ANFIS model to assess thermal comfort of an urban square in moderate and dry climate	2016	S Kariminia, S Motamedi, S Shamshirband…
108	Efficiency evaluation of urban transport using the DEA method	2018	S Hajduk
109	Emerging threats in urban ecosystems: a horizon scanning exercise	2015	MC Stanley, JR Beggs, IE Bassett…
110	Transport issues and sustainable mobility in smart cities	2015	E Venezia, S Vergura
111	Engaging with global urban governance	2018	M Acuto

112	Urban resilience: A conceptual framework	2019	PJG Ribeiro, L Gonçalves
113	Think regionally, act locally?: Gardening, cycling, and the horizon of urban spatial politics	2017	JG Stehlin, AR Tarr
114	Recreational ecosystem services in European cities: Sociocultural and Geographical contexts matter for park use	2018	LK Fischer, J Honold, A Botzat, D Brinkmeyer…
115	Space Syntax: A method to measure urban space related to social, economic and cognitive factors	2018	A Van Nes, C Yamu
116	Competitive dynamics between criminals and law enforcement explains the super-linear scaling of crime in cities	2015	S Banerjee, P Van Hentenryck, M Cebrian
117	Simulating urban growth driven by transportation networks: A case study of the Istanbul third bridge	2015	IE Ayazli, F Kilic, S Lauf, H Demir, B Kleinschmit
118	Dark cities? Developing a methodology for researching dark tourism in European cities	2016	R Powell, J Kennell
119	Classification of smart city research-a descriptive literature review and future research agenda	2019	P Gupta, S Chauhan, MP Jaiswal
120	Economic evaluation of urban heritage: An inclusive approach under a sustainability perspective	2015	L Dalmas, V Geronimi, JF Noël, JTK Sang
121	Urban resilience at eye level: Spatial analysis of empirically defined experiential landscapes	2019	K Samuelsson, J Colding, S Barthel
122	New spatial mobility patterns in large Spanish cities: From the economic boom to the great recession	2018	J Bayona-i-Carrasco, F Gil-Alonso…
123	Factors influencing the use of urban greenways: A case study of Aydın, Turkey	2016	A Akpinar
124	Revitalization of urban public spaces: An overview	2015	M Ramlee, D Omar, RM Yunus, Z Samadi
125	Managing crowds: The possibilities and limitations of crowd information during urban mass events	2015	LB Zomer, W Daamen, S Meijer…
126	Children Living with 'Sustainable' Urban Architectures	2015	J Horton, S Hadfield-Hill…
127	Spatiotemporal effects of main impact factors on residential land price in major cities of China	2017	S Yang, S Hu, W Li, C Zhang, JA Torres
128	Integrating ecosystem services into urban park planning & design	2016	DC Ibes
129	Perceived urban design qualities and affective experiences of walking	2016	M Johansson, C Sternudd, M Kärrholm
130	Measuring the sustainability of urban water services	2015	RC Marques, NF da Cruz, J Pires
131	Beyond neoliberal imposition: state–local cooperation and the blending of social and economic objectives in French urban development corporations	2016	G Pinson, C Morel Journel

132	Culture in sustainable urban development: Practices and policies for spaces of possibility and institutional innovations	2018	S Kagan, A Hauerwaas, V Holz, P Wedler
133	Characteristics of urban agglomerations in different continents: history, patterns, dynamics, drivers and trends	2018	W Loibl, G Etminan…
134	Crowdsourcing functions of the living city from Twitter and Foursquare data	2016	X Zhou, L Zhang
135	Identifying design criteria for urban system 'last-mile'solutions–a multi-stakeholder perspective	2016	TS Harrington, J Singh Srai, M Kumar…
136	Smart cities in a smart world	2015	B Murgante, G Borruso
137	Relevant factors in sustainable urban development of urban planning methodology and implementation of concepts for sustainable planning (Planning …	2016	E Sofeska
138	"Barcelona in common": A new urban regime for the 21st-century tourist city?	2018	AP Russo, A Scarnato
139	Planning and design elements for transit oriented Developments/Smart cities: Examples of cultural borrowings	2016	J Black, K Tara, P Pakzad
140	An explanation of urban sprawl phenomenon in Shiraz Metropolitan Area (SMA)	2018	B Bagheri, SN Tousi
141	Defining and refining the research agenda for Australian cities	2017	R Freestone, B Randolph, A Wheeler
142	Investigation of visitors' motivation, satisfaction and cognition on urban forest parks in Taiwan	2016	YC Wang, JC Lin, WY Liu, CC Lin…
143	Urban design, public space and the dynamics of creative milieux: a photofigureic approach to Bairro Alto (Lisbon), Gràcia (Barcelona) and Vila Madalena (São Paulo)	2015	P Costa, R Lopes
144	Parents' education, school-age children and household location in American cities	2015	W Sander, W Testa
145	The governance of smart cities: A systematic literature review	2018	RWS Ruhlandt
146	Everyday wild: Urban natural areas, health, and well-being	2019	AE Cheesbrough, T Garvin, CIJ Nykiforuk
147	[HTML][HTML] Autonomous vehicles and the future of urban tourism	2019	SA Cohen, D Hopkins
148	Urban and rural population growth in a spatial panel of municipalities	2017	D Firmino Costa da Silva, JP Elhorst…
149	Sustainable mobility in the low carbon city: Digging up the highway in Odense, Denmark	2017	P Fenton
150	Smart Cities Evaluation–A Survey of Performance and Sustainability Indicators	2018	D Petrova-Antonova, S Ilieva

151	Distribution of tourists within urban heritage destinations: a hot spot/cold spot analysis of TripAdvisor data as support for destination management	2020	E van der Zee, D Bertocchi…
152	Urban form breeds neighborhood vibrancy: A case study using a GPS-based activity survey in suburban Beijing	2018	J Wu, N Ta, Y Song, J Lin, Y Chai
153	City Blueprints: baseline assessments of water management and climate change in 45 cities	2016	CJ Van Leeuwen, SHA Koop, RMA Sjerps
154	Data-driven participation: Algorithms, cities, citizens, and corporate control	2016	M Tenney, R Sieber
155	The analyses of socio-economic development tendencies of the capital cities in the modern Russia	2016	IS Glebova, SN Kotenkova…
156	Commercial farming within the urban built environment–Taking stock of an evolving field in northern countries	2018	K Benis, P Ferrão
157	Addressing big data challenges in smart cities: a systematic literature review	2016	S Chauhan, N Agarwal, AK Kar
158	Urban geoheritage complexity: Evidence of a unique natural resource from Shiraz city in Iran	2018	T Habibi, AA Ponedelnik, NN Yashalova, DA Ruban
159	Urban design in favor of human thermal comfort for hot arid climate using advanced simulation methods	2017	A Barakat, H Ayad, Z El-Sayed
160	Sino-western tourists' place attachment to a traditional Chinese urban destination: A tale from Hangzhou, China	2016	Z Xu
161	Financing low-carbon, climate-resilient cities	2018	S Colenbrander, M Lindfield, J Lufkin…
162	Exploring local consequences of two land-use alternatives for the supply of urban ecosystem services in Stockholm year 2050	2016	JH Kain, N Larondelle, D Haase, A Kaczorowska
163	Affordable housing and urban regeneration in Portugal: a troubled tryst	2015	R Branco, S Alves
164	Theoretical aspects for use technologies formation and implementation of urban development land monitoring	2015	OV Pyrkova
165	The recreational attractiveness of Dąbie District and the character of its landscape	2017	E Sochacka-Sutkowska, A Pilarczyk
166	CRITICAL RECONSIDERATION OF SOCIAL AND CULTURAL ELEMENTS THAT CONSTITUTE THE CREATIVE ECOSYSTEM OF A CITY: EXAMPLES FROM …	2017	M Uršič
167	The future of smart cities: Open issues and research challenges	2018	H Schaffers
168	Facetten der Reurbanisierung	2017	R Hamm, A Jäger, K Keggenhoff

169	The Jessica Initiative: An Instrument for Urban Sustainable Development. Examples of Urban Regeneration in Silesia (Poland) and Central Moravia (Czech Republic) …	2015	K Tarnawska, J Rosiek
170	Tourism and mobility. Best practices and conditions to improve urban livability	2015	RA La Rocca
171	The effect of packaging attributes on consumer buying decision behavior in major commercial cities in Ethiopia	2017	GA Imiru
172	Benchmarking urban competitiveness in Europe to attract investment	2015	L Sáez, I Periáñez
173	Investigating territorial specialization in tourism sector by ecosystem services approach	2019	F Scorza, B Murgante, G Las Casas, Y Fortino…
174	Creating Inclusive cities: A review of indicators for measuring sustainability for urban infrastructure in India	2016	S Bhattacharya, SA Patro…
175	Urban house prices: A tale of 48 cities	2015	KA Kholodilin, D Ulbricht
176	Quality of life, multimodality, and the demise of the autocentric metropolis: A multivariate analysis of 148 mid-size US cities	2019	CA Talmage, C Frederick
177	From pedestrianisation to commercial gentrification: The case of Kadıköy in Istanbul	2017	D Özdemir, İ Selçuk
178	Cities Network Along the Silk Road	2017	P Ni, M Kamiya, R Ding
179	Understanding the urban spatial structure of Sub-Saharan African cities using the case of urban development patterns of a Ghanaian city-region	2019	FSK Agyemang, E Silva, M Poku-Boansi
180	Smart city near to 4.0—an adoption of industry 4.0 conceptual model	2017	M Postránecký, M Svítek
181	Evaluating the capability of walkability audit tools for assessing sidewalks	2018	M Aghaabbasi, M Moeinaddini, MZ Shah…
182	Algorithms to assess music cities: Case study—Melbourne as a music capital	2017	AJ Baker
183	Size distribution of cities: A kinetic explanation	2019	S Gualandi, G Toscani
184	Towards universal health coverage via social health insurance in China: systemic fragmentation, reform imperatives, and policy alternatives	2017	AJ He, S Wu
185	Modeling territorial attractiveness in the metropolis	2019	EGC Nuño, RFL Pacheco
186	Would environmental pollution affect home prices? An empirical study based on China's key cities	2017	Y Hao, S Zheng
187	[PDF][PDF] Projective cities: Organizing large cultural development initiatives	2015	N Wåhlin
188	[PDF][PDF] How Much Does Technology Affect the Management of Cities in Latin American and the Caribbean	2018	M De Halleux, A Estache…

189	City Branding, Sustainable Urban Development and the Rentier State. How do Qatar, Abu Dhabi and Dubai present Themselves in the Age of Post Oil and Global …	2019	M De Jong, T Hoppe, N Noori
190	The impacts of high-speed rail extensions on accessibility and spatial equity changes in South Korea from 2004 to 2018	2015	H Kim, S Sultana
191	From SMART Cities to SMART City-Regions: Reflections and Proposals	2017	I Greco, A Cresta
192	Knowledge-based urban development of cross-border twin cities	2016	T Makkonen, A Weidenfeld
193	Climate change and environmental degradation and the drivers of migration in the context of shrinking cities: A case study of Khuzestan province, Iran	2019	AR Khavarian-Garmsir, A Pourahmad…
194	Combining cognitive mapping and MCDA for improving quality of life in urban areas	2018	PAM Faria, FAF Ferreira, MS Jalali, P Bento…
195	Urban sustainable competitiveness: a comparative analysis of 500 cities around the world	2017	P Ni, Y Wang
196	Neighbourhood change and spatial polarization: The roles of increasing inequality and divergent urban development	2018	T Modai-Snir, M van Ham
197	The limits of growth: A case study of three mega-projects in Istanbul	2017	E Dogan, A Stupar
198	Simulating block-level urban expansion for national wide cities	2017	Y Long, K Wu
199	Preferences for cultural urban ecosystem services: Comparing attitudes, perception, and use	2015	C Bertram, K Rehdanz
200	Human capital spillovers in Dutch cities: consumption or productivity?	2017	VA Venhorst
201	Think Cities: the accelerator for sustainable planning	2020	G Carfantan, F Daniel, L D'orazio…
202	A current inventory of vacant urban land in America	2016	GD Newman, AOM Bowman, R Jung Lee…
203	The smart city ecosystem as an innovation model: lessons from Montreal	2016	MR Khomsi
204	Assessing the capacity to govern flood risk in cities and the role of contextual factors	2018	S Koop, F Monteiro Gomes, L Schoot, C Dieperink…
205	The concept of a walkable city as an alternative form of urban mobility	2017	K Turoń, P Czech, M Juzek
206	The significance of digital data systems for smart city policy	2017	K Kourtit, P Nijkamp, J Steenbruggen
207	Migration mobility of population and otkhodnichestvo in modern Russia	2015	TG Nefedova
208	A summated rating scale for measuring city image	2015	S Gilboa, ED Jaffe, D Vianelli, A Pastore, R Herstein

209	Spatial patterns, driving forces, and urbanization effects of China's internal migration: County-level analysis based on the 2000 and 2010 censuses	2015	T Liu, Y Qi, G Cao, H Liu
210	Beyond town and gown: Universities, territoriality and the mobilization of new urban structures in Canada	2015	JPD Addie, R Keil, K Olds
211	Engaging urban nature: improving the understanding of public perceptions of the role of biodiversity in cities	2019	V Campbell-Arvai
212	Public Transport in the Era of ITS: The Role of Public Transport in Sustainable Cities and Regions	2016	X Roselló, A Langeland, F Viti
213	[PDF][PDF] CONTRASTING GLOBAL IMAGERY TO LOCAL REALITIES IN THE POSTCOLONIAL WATERFRONTS OF MALAYSIA'S CAPITAL CITIES.	2016	Q Stevens, M Kozlowski, N Ujang
214	What drives planning in a shrinking city? Tales from two German and two American cases	2017	K Pallagst, R Fleschurz, S Said
215	Relationship between types of urban forest and PM2. 5 capture at three growth stages of leaves	2015	T Nguyen, X Yu, Z Zhang, M Liu, X Liu
216	Human diffusion and city influence	2015	M Lenormand, B Gonçalves…
217	Does bus accessibility affect property prices?	2019	L Yang, J Zhou, OF Shyr
218	Smart and Fermented Cities: An Approach to Placemaking in Urban Informatics	2019	G Freeman, J Bardzell, S Bardzell, SY Liu…
219	City data fusion: Sensor data fusion in the internet of things	2016	M Wang, C Perera, PP Jayaraman, M Zhang…
220	Smart city concept: What it is and what it should be	2016	I Zubizarreta, A Seravalli…
221	Effective destination advertising: Matching effect between advertising language and destination type	2015	J Byun, SCS Jang
222	Spatial distribution of knowledge-intensive business services in a small post-communist economy	2017	J Ženka, J Novotný, O Slach, I Ivan
223	Population and spatial distribution of urbanisation in Peninsular Malaysia 1957-2000	2017	T Masron, U Yaakob, NM Ayob…
224	Capital city dynamics: Linking regional innovation systems, locational policies and policy regimes	2016	H Mayer, F Sager, D Kaufmann, M Warland
225	Events and placemaking	2017	M De Brito, G Richards
226	The Role of City and Host University Images on Students' Satisfaction with the Assigned Destination	2017	R Roostika
227	Subways and urban growth: Evidence from earth	2018	M Gonzalez-Navarro, MA Turner
228	Assessment tools for urban sustainability policies in Spanish Mediterranean tourist areas	2017	P Martí, A Nolasco-Cirugeda, L Serrano-Estrada
229	Vehicular social networks: Enabling smart mobility	2017	Z Ning, F Xia, N Ullah, X Kong…

230	Calculation of high-rise construction limitations for non-resident housing fund in megacities	2018	O Iliashenko, S Krasnov…
231	Mine Sited after Mine Activity: The Brownfields Methodology and Kuzbass Coal Mining Case	2019	M Cehlár, J Janočko, Z Šimková, T Pavlik, M Tyulenev…
232	A concept of forecasting origin-destination air passenger demand between global city pairs using future socio-economic scenarios	2015	I Terekhov, V Gollnick
233	City boosterism and place-making with light rail transit: A critical review of light rail impacts on city image and quality	2017	F Ferbrache, RD Knowles
234	Lost in complexity, found in dispersion:'Peripheral'development and deregulated urban growth in Rome	2015	L Salvati
235	An evolutionary theory of urban systems	2018	D Pumain
236	Uneven growth: tackling city decline	2016	A Pike, D MacKinnon, M Coombes…
237	Modelling of sustainable development of megacities under limited resources	2019	S Sergeev, T Kirillova, I Krasyuk
238	Domestic tourist market in the population estimates: a sociological analysis	2016	EV Frolova, OV Rogach, EE Kabanova…
239	EVALUATING THE LOCATION OF REGIONAL RETURN CENTERS IN REVERSE LOGISTICS THROUGH INTEGRATION OF GIS, AHP AND INTEGER …	2015	AZ Acar, İ Önden, K Kara
240	Public utility companies in liberalized markets–The impact of management models on local and regional sustainability	2017	S Mejia-Dugand, O Hjelm, L Baas
241	Static vs. dynamic agglomeration economies: Spatial context and structural evolution behind urban growth	2017	R Camagni, R Capello, A Caragliu
242	Visioning the bus system of the future: stakeholders' perspective	2015	A Musso, MV Corazza
243	The smart citizen factor in trustworthy smart city crowdsensing	2016	M Pouryazdan, B Kantarci
244	Comparing China's urban systems in high-speed railway and airline networks	2018	H Yang, F Dobruszkes, J Wang, M Dijst…
245	Regeneration strategies in shrinking urban neighbourhoods—Dimensions of interventions in theory and practice	2015	W Schenkel
246	Consumer happiness derived from inherent preferences versus learned preferences	2016	Y Tu, CK Hsee
247	Boosting city image for creation of a certain city brand	2019	A Shirvani-Dastgerdi, G De-Luca
248	The influence of capital system categories on human development index in Brazil	2015	AC Fachinelli, CP Giacomello, F Larentis
249	Dissemination of electric vehicles in urban areas: Major factors for success	2016	A Ajanovic, R Haas

250	A complex network perspective for characterizing urban travel demand patterns: figure theoretical analysis of large-scale origin–destination demand networks	2017	M Saberi, HS Mahmassani, D Brockmann, A Hosseini
251	What Liberal World Order?	2017	M Leonard
252	City-integrated renewable energy for urban sustainability	2016	DM Kammen, DA Sunter
253	Fifty shades of green	2017	MJ Nieuwenhuijsen, H Khreis, M Triguero-Mas…
254	Multicriteria evaluation of sustainable energy solutions for Colosseum	2017	O Loikkanen, R Lahdelma, P Salminen
255	City logistics-a strategic element of sustainable urban development	2016	S Kauf
256	Urban Development in Poland, from the Socialist City to the Post-Socialist and Neoliberal City.[in:] V	2016	G Węcławowicz
257	Urban design: an important future force for health and wellbeing	2016	S Kleinert, R Horton
258	Emerging technologies and cultural tourism: Opportunities for a cultural urban tourism research agenda	2017	C Garau
259	Identification of key energy efficiency drivers through global city benchmarking: A data driven approach	2017	X Wang, Z Li, H Meng, J Wu
260	An exploration of smart city approaches by international ICT firms	2019	D van den Buuse, A Kolk
261	Automated vehicles and how they may affect urban form: A review of recent scenario studies	2019	D Stead, B Vaddadi
262	Vicious advice: Analyzing the impact of TripAdvisor on the quality of restaurants as part of the cultural heritage of Venice	2017	A Ganzaroli, I De Noni, P van Baalen
263	Urban tourism (s): is there a case for a paradigm shift?	2015	C Pasquinelli
264	Regional quality of living in Europe	2015	P Lagas, F van Dongen, F van Rijn, H Visser
265	Numerical study of the impact of vegetation coverings on sound levels and time decays in a canyon street model	2015	G Guillaume, B Gauvreau, P L'Hermite
266	Quantifying potential benefits of horizontal cooperation in urban transportation under uncertainty: a simheuristic approach	2016	CL Quintero-Araujo, A Gruler, AA Juan
267	The green branding of Hong Kong: visitors' and residents' perceptions	2016	CS Chan, LM Marafa
268	Disruptive innovation in the era of global cyber-society: With focus on smart city efforts	2017	A Rucinski, R Garbos, J Jeffords…
269	The 'actually existing smart city'	2015	T Shelton, M Zook, A Wiig

270	A hypothesis of the dimensional organization of the city construct. A starting point for city brand positioning	2015	HG Larsen
271	Travel costs and urban specialization patterns: Evidence from China's high speed railway system	2017	Y Lin
272	[HTML][HTML] Living near to attractive nature? A well-being indicator for ranking Dutch, Danish, and German functional urban areas	2017	MN Daams, P Veneri
273	[PDF][PDF] Asymmetries of the North Caucasus federal district subjects' social ecological economic development under macroeconomic tendencies	2015	AH Dikinov
274	Characterizing Geographical preferences of international tourists and the local influential factors in China using geo-tagged photos on social media	2016	S Su, C Wan, Y Hu, Z Cai
275	Modern city positioning case study as a tool of territorial marketing: Magadan, Russia	2015	NA Romanova, TA Brachun, EA Dmitrieva
276	Assessing the sources of competitiveness of the US states	2017	M Akpinar, Ö Can, M Mermercioglu
277	Re-examining historical energy transitions and urban systems in Europe	2016	M Chabrol
278	Interregional mobility of talent in Spain: The role of job opportunities and qualities of places during the recent economic crisis	2018	S Sánchez-Moral, A Arellano…
279	Bike-sharing as an element of integrated Urban transport system	2017	P Czech, K Turoń, R Urbańczyk
280	Renewing a historical legacy: Tourism, leisure shopping and urban branding in Paris	2015	C Rabbiosi
281	A computer vision system to localize and classify wastes on the streets	2017	MS Rad, A von Kaenel, A Droux, F Tieche…
282	Tourism and high speed rail in Spain: Does the AVE increase local visitors?	2017	D Albalate, J Campos, JL Jiménez
283	Functional effectiveness and modern mechanisms for national urban systems globalization: The case of Russia	2019	A Arkhipov, D Ushakov
284	Can a small city be considered a smart city?	2017	IM Lopes, P Oliveira
285	Smart city indicators: A systematic literature review	2016	F Purnomo, H Prabowo
286	Toward improved public health outcomes from urban nature	2015	DF Shanahan, BB Lin, R Bush…
287	[HTML][HTML] The green soul of the concrete jungle: the urban century, the urban psychological penalty, and the role of nature	2018	RI McDonald, T Beatley…
288	The economics of density: Evidence from the Berlin Wall	2015	GM Ahlfeldt, SJ Redding, DM Sturm, N Wolf
289	[HTML][HTML] Interregional youth migration in Russia: a comprehensive analysis of Geographical statistical data	2016	I Kashnitsky, N Mkrtchyan, O Leshukov

290	Cagliari and smart urban mobility: Analysis and comparison	2016	C Garau, F Masala, F Pinna
291	Is urban spatial development on the right track? Comparing strategies and trends in the European Union	2019	C Cortinovis, D Haase, B Zanon, D Geneletti
292	Deep learning the city: Quantifying urban perception at a global scale	2016	A Dubey, N Naik, D Parikh, R Raskar…
293	Challenges of the new urban world	2015	K Kourtit, P Nijkamp, MD Partridge
294	Reading vulnerabilities through urban planning history: An earthquake-prone city, Adapazarı case from Turkey	2016	E Orhan
295	US regional population growth 2000–2010: Natural amenities or urban agglomeration?	2017	DS Rickman, H Wang
296	[PDF][PDF] Sensory Attractiveness of Local Smoked Bacons	2018	B Kusz, J Kilar, D Kusz
297	Place quality and urban competitiveness symbiosis? A position paper	2016	N Esmaeilpoorarabi, T Yigitcanlar…
298	Smart city governance: A local emergent perspective	2016	A Meijer
299	Artificial Intelligence and Urbanization: The Rise of the Elysium City	2017	JM Munoz, A Naqvi
300	[PDF][PDF] Project management in the sphere of tourism (using the example of Taganrog)	2018	EA Vetrova, EE Kabanova…
301	Integrating physical and social sensing to enable smart city mobility services	2016	RG Qiu, L Qiu, Y Badr
302	Recycling the city new perspective on the real-estate market and construction industry	2015	E Micelli, A Mangialardo
303	Refugees, asylum seekers, and policy in OECD countries	2016	TJ Hatton
304	Smart metropolitan regional development: economic and spatial design strategies	2019	TMV Kumar
305	Urban attraction policies for international academic talent: Munich and Vienna in comparison	2017	C Reiner, S Meyer, S Sardadvar
306	Possibilities and limits of brand repositioning for a second-ranked city: The case of Brisbane, Australia's “New World City”, 1979–2013	2016	A Insch, B Bowden
307	Framing urban habitats: The small and medium towns in the peripheries	2015	T de Noronha, E Vaz
308	[HTML][HTML] L'image de marque des villes wallonnes	2016	L Scatton, S Schmitz
309	What have we learned about the causes of recent gentrification?	2016	J Hwang, J Lin
310	[HTML][HTML] All about the 'wow factor'? The relationships between aesthetics, restorative effect and perceived biodiversity in designed urban planting	2017	H Hoyle, J Hitchmough, A Jorgensen
311	Key factors for defining an efficient urban transport interchange: Users' perceptions	2016	S Hernandez, A Monzon

312	A Critical Interpretation of the" Quality of Place". Between Attractiveness and Post-rurality in Chianti	2019	M Battaglia, C Certomà, M Frey
313	Automated vehicles and the city of tomorrow: A backcasting approach	2019	E González-González, S Nogués, D Stead
314	The sociocultural sources of urban buzz	2016	D Arribas-Bel, K Kourtit…
315	Home from Home? locational choices of international "creative class" workers	2015	J Brown
316	Classifying urban residential areas based on their exposure to crime: A constructivist approach	2018	SCR Marques, FAF Ferreira…
317	Reframing place promotion, place marketing, and place branding-moving beyond conceptual confusion	2018	M Boisen, K Terlouw, P Groote, O Couwenberg
318	Evaluation of the strategic plans' impact on the competitive attractiveness and sustainability of regional development	2018	BM Grinchel
319	[PDF][PDF] Pervasive NFC-based solution for the analysis of tourism data in an environment of smart cities	2015	EAC Agredo, LC Martínez-Acosta, A Chantre…
320	[PDF][PDF] Shopping centres as the subject of Polish Geographical research	2016	A Rochmińska
321	[PDF][PDF] Mardin'in Doğal ve Kültürel Çekiciliklerinin Destinasyon Pazarlaması Kapsamında İncelenmesi	2017	O Atsız, İ Kızılırmak
322	Valuing green infrastructure in an urban environment under pressure—The Johannesburg case	2013	A Schäffler, M Swilling
323	Methodologies for local development in smart society	2012	L Batagan
324	Real wage inequality	2013	E Moretti
325	A memetic algorithm with a large neighborhood crossover operator for the generalized traveling salesman problem	2010	B Bontoux, C Artigues, D Feillet
326	The Postcolonial Dimension	2013	V Watson
327	[PDF][PDF] Identifying Spatial Structure of Urban Functional Centers Using Travel Survey Data: A Case Study of Singapore.	2013	C Zhong, X Huang, SM Arisona, G Schmitt
328	Economic geofigurey: A review of the theoretical and empirical literature	2013	SJ Redding
329	Determining factors of a city's tourism attractiveness	2011	A Calvo de Mora Schmidt, JM Berbel-Pineda…
330	Planning, competitiveness and sprawl in the Mediterranean city: The case of Athens	2010	I Chorianopoulos, T Pagonis, S Koukoulas, S Drymoniti
331	Implementing electric vehicles in urban distribution: A discrete event simulation	2013	P Lebeau, C Macharis, J Van Mierlo…
332	[PDF][PDF] The planning and governance of Asia's mega-urban regions	2011	AA Laquian

333	Tattooing and body piercing-what motivates you to do it?	2010	B Antoszewski, A Sitek, M Fijałkowska…
334	[PDF][PDF] Scenarios for active learning in smart territories.	2013	C Giovannella, A Iosue, A Tancredi, F Cicola, A Camusi…
335	Urban history and cultural resources in urban regeneration: a case of creative waterfront renewal	2013	M Sepe
336	Mega events and urban conflicts in Valencia, Spain: Contesting the new urban modernity	2011	L del Romero Renau, C Trudelle
337	A gold medal for the market: the 1984 Los Angeles Olympics, the Reagan era, and the politics of neoliberalism	2012	R Gruneau, R Neubauer
338	The FDI location decision: Distance and the effects of spatial dependence	2014	F Blanc-Brude, G Cookson, J Piesse…
339	Pats: A framework of pattern-aware trajectory search	2010	LY Wei, WC Peng, BC Chen…
340	Mystery and thriller tourism: Novel solutions for European cities	2013	W Strielkowski
341	[PDF][PDF] Creativity, Culture & the City: A question of interconnection	2011	C Landry
342	Smart Learning Eco-Systems:“fashion” or “beef”?	2014	C Giovannella
343	UNESCO biosphere reserves in an urbanized world	2012	AC de la Vega-Leinert, MA Nolasco…
344	'Peak car'—themes and issues	2013	P Goodwin, K Van Dender
345	Skin color, physical appearance, and perceived discriminatory treatment	2011	J Hersch
346	The dishonest relationship between city marketing and culture: Reflections on the theory and the case of Budapest.	2011	M Kavaratzis
347	[PDF][PDF] GIS Application in Urban Green space Per Capita Evaluation	2012	H Laghai, H Bahmanpour
348	Residential satisfaction of elderly in the city centre: The case of revitalizing neighbourhoods in Prague	2012	J Temelová, N Dvořáková
349	Simulating urban networks through multiscalar space-time dynamics: Europe and the united states, 17th-20th centuries	2010	A Bretagnolle, D Pumain
350	Lille 2004 and the role of culture in the regeneration of Lille métropole	2011	D Paris, T Baert
351	[HTML][HTML] Urban growth and decline: the role of population density at the city core	2011	K Fee, D Hartley
352	[PDF][PDF] Culture, creativity, cultural economy: A review	2014	J O'Connor, M Gibson
353	The quest to become a world city: Implications for access to water	2014	M Nastar
354	Creativity, culture tourism and place-making: Istanbul and London film industries	2010	MD Alvarez, B Durmaz, S Platt…

355	Neighborhoods and social interaction	2014	SC Brown, J Lombard
356	Delta urbanism: planning and design in urbanized deltas–comparing the Dutch delta with the Mississippi River delta	2013	H Meyer, S Nijhuis
357	A ballpark and neighborhood change: Economic integration, a recession, and the altered demography of San Diego's Ballpark District after eight years	2012	MB Cantor, MS Rosentraub
358	System solutions in urban water management: The Lodz (Poland) perspective	2011	I Wagner, M Zalewski
359	A combined fuzzy MCDM approach for selecting shopping center site: An example from Istanbul, Turkey	2010	S Önüt, T Efendigil, SS Kara
360	[HTML][HTML] Urban health challenges in Europe	2013	RJ Lawrence
361	European capital of culture designation as an initiator of urban transformation in the post-socialist countries	2014	T Lähdesmäki
362	Place marketing, place branding and foreign direct investments: Defining their relationship in the frame of local economic development process	2010	T Metaxas
363	Current trends in Smart City initiatives: Some stylised facts	2014	P Neirotti, A De Marco, AC Cagliano, G Mangano…
364	Jobs-housing balance in an era of population decentralization: An analytical framework and a case study	2011	BPY Loo, ASY Chow
365	City marketing and place branding: A critical review of practice and academic research.	2012	N Muñiz Martinez
366	Relaxing Hukou: Increased labor mobility and China's economic geofigurey	2012	M Bosker, S Brakman, H Garretsen…
367	Maintain, demolish, re-purpose: Policy design for vacant land management using decision models	2014	MP Johnson, J Hollander, A Hallulli
368	Introduction: Religion takes place: Producing urban locality	2013	M Burchardt, I Becci
369	Russian periphery is dying in movement: a cohort assessment of Russian internal youth migration based on Census data	2014	I Kashnitsky, N Mkrtchyan
370	Impacts of high speed rail on railroad network accessibility in China	2014	SL Shaw, Z Fang, S Lu, R Tao
371	[PDF][PDF] Media and the city: Making sense of place	2011	M Georgiou
372	Constraint-based approaches for balancing bike sharing systems	2013	L Di Gaspero, A Rendl, T Urli
373	A sustainable perspective on urban freight transport: Factors affecting local authorities in the planning procedures	2010	M Lindholm
374	Governing global city regions in China and the West	2010	RK Vogel, HV Savitch, J Xu, AGO Yeh, W Wu…

375	Implementing sustainable urban travel policies in China	2012	PAN Haixiao
376	[HTML][HTML] Modelling the potential effect of shared bicycles on public transport travel times in Greater Helsinki: An open data approach	2013	S Jäppinen, T Toivonen, M Salonen
377	The influence of the Olympic Games on Beijing consumers' perceptions of their city tourism development	2010	V Ratten, R Tsiotsou, I Kapareliotis…
378	Cultural triangle and beyond: a spatial analysis of cultural industries in Istanbul	2011	ZM Enlil, Y Evren, I Dincer
379	Comparing smart and digital city: initiatives and strategies in Amsterdam and Genoa. Are they digital and/or smart?	2014	RP Dameri
380	A tale of a musical city: Fostering self-brand connection among residents of Austin, Texas	2012	E Kemp, CY Childers, KH Williams
381	Competitiveness of Urban Business Tourism: An Innovative Study of Evaluation System and Approach [J]	2010	LI Xi
382	How does the household structure shape the urban economy?	2010	S Tscharaktschiew, G Hirte
383	The History of Architecture and Art and How It Is Seen by Tourists	2013	S Kaczmarek
384	An integrative theoretical model for improving resident-city identification	2014	S Zenker, S Petersen
385	Bulgarian urban settlements in the early 21st century	2010	M Ilieva, I Iliev
386	[PDF][PDF] Environmental challenges of urbanization: A case study for open green space management	2013	TPZ Mpofu
387	Assessing the what is beautiful is good stereotype and the influence of moderately attractive and less attractive advertising models on self-perception, ad attitudes, and …	2014	I Vermeir, D Van de Sompel
388	[PDF][PDF] Inertia: Spurious loyalty or action loyalty?	2011	LW Wu
389	The new urban world: Challenges and policy	2014	K Kourtit, P Nijkamp, N Reid
390	Understanding spatial differentiation in urban decline levels	2014	JJ Hoekveld
391	Cellular automata in urban spatial modelling	2012	S Iltanen
392	[PDF][PDF] Quality of life assessment based on spatial and temporal analysis of the vegetation area derived from satellite images	2011	MI Vlad, D Brătăşanu
393	Liveability of tall residential buildings	2011	B Yuen
394	[HTML][HTML] Small-size urban settlements: Proposed approach for managing urban future in developing countries of increasing technological capabilities, the case of …	2014	AA Abou-Korin
395	Building city brands through sport events: Theoretical and empirical perspectives	2012	H Westerbeek, M Linley

396	[PDF][PDF] Distribution of Urban Green Spaces - an Indicator of Topophobia - Topophilia of Urban Residential Neighborhoods. Case Study of 5th District of Bucharest …	2011	LA Cucu, CM Ciocănea, DA Onose
397	The sustainability of an entrepreneurial city?	2014	K Davidson, B Gleeson
398	Interrelations between travel mode choice and trip distance: trends in Germany 1976–2002	2010	J Scheiner
399	UV radiation as an attractor for insects	2012	A Barghini, BA Souza de Medeiros
400	Public construction project delivery process in Singapore, Beijing, Hong Kong and Sydney	2013	Y Ke, FYY Ling, Y Ning
401	[PDF][PDF] Slow tourists: A comparative research based on Cittaslow principles	2011	HR Yurtseven, O Kaya
402	Place-making processes in unconventional cultural practices. The case of Turin's contemporary art festival Paratissima	2014	FS Rota, C Salone
403	[PDF][PDF] Urban Challenges in South-East Asia	2011	YK Sheng
404	Metropolitan city growth and management in post-liberalized India	2012	A Shaw
405	[PDF][PDF] What makes an attractive employer: significant factors from employee perspective?	2011	SS Pingle, HK Sodhi
406	Measuring the economic, social and environmental performance of European island regions: emerging issues for European and regional policy	2013	I Spilanis, T Kizos, M Vaitis…
407	A combined site proximity and recreation index approach to value natural amenities: An example from a natural resource management region of Murray-Darling Basin	2012	S Tapsuwan, DH MacDonald, D King…
408	Economic viability	2010	C Jones, C Leishman, C MacDonald, A Orr…
409	Changes in employee commuting: a comparative analysis of employee commuting to major Slovenian employment centers from 2000 to 2009	2011	D Bole
410	Innovations for intergenerational neighborhoods	2012	I Ammann, M Heckenroth
411	[PDF][PDF] Brief survey of soft computing techniques used for optimization of TSP	2013	P Panwar, S Gupta
412	The shift to competitiveness and a new phase of sprawl in the Mediterranean city: Enterprises guiding growth in Messoghia–Athens	2014	I Chorianopoulos, G Tsilimigkas, S Koukoulas…
413	Socioeconomic effect on perception of urban green spaces in Guangzhou, China	2013	CY Jim, X Shan
414	The role of public libraries in culture-led urban regeneration	2013	D Skot-Hansen, CH Rasmussen, H Jochumsen
415	[PDF][PDF] The use of ITS for improving bus priority at traffic signals	2010	C Morellato, M Sdun

416	[PDF][PDF] Managers and entrepreneurs in creative and knowledge-intensive industries: what determines their location? Toulouse, Helsinki, Budapest, Riga and Sofia	2010	E Dainov, A Sauka
417	Urban rail systems investments: an analysis of the impacts on property values and residents' location	2011	F Pagliara, E Papa
418	How local authority decision makers address freight transport in the urban area	2012	M Lindholm
419	The migration dynamics of the “creative class”: evidence from a study of artists in Stockholm, Sweden	2013	T Borén, C Young
420	Globalization and the production of city image in Guangzhou's metro station advertisements	2011	H Zhu, J Qian, Y Gao