POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione

Corso di Laurea Magistrale in Ingegneria Informatica

A business plan methodology modified to take advantage of Data Science

Supervisor: Prof. Chiara Francalanci

Assistant Supervisor: Ing. Paolo Ravanelli

Master of Science Thesis of:

Alessandro Tribi

Student ID 804188

Academic Year 2014-2015

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

In the last recent years, Data Science has been widely employed by companies and governments to turn data into insight for making better decisions. In particular, concerning the public sector, this phenomenon has acquired increasing importance since 2007, when the eight principles of Open Government Data have been defined. These data are produced or commissioned by public bodies and they are made freely accessible and usable by everyone.

The aim of this work is to propose a revised methodology for writing a business plan, in order to take advantage of Data Science. This methodology is then used to write a business plan about the proposal for a road safety action. The action is intended for Rome, where, after having performed an analysis on data about road accidents, a serious issue regarding young drivers has been revealed. Rome has an extremely high number of young car drivers involved in personal injury accidents. The value is even more evident if compared to London.

A more detailed study on the data has allowed to identify the risky driving as the cause of the problem. For this reason, a case study showing a solution to a similar problem has been searched and it has been identified in Staffordshire, a county of the United Kingdom. In this region, after the introduction of an enhanced driving course, which is called Pass Plus Extra and it aims to increase the road safety awareness, a strong fall in the number of young drivers involved in serious crashes has been achieved. After having evaluated the effectiveness of the course by means of Data Science techniques, the economic and social advantages that the introduction of a similar enhanced driving course would have caused in Rome in the last four years have been estimated and the results are extremely positive. The innovative methodology presented in this work is supposed to be suitable not only in this case, but also in all the other contexts in which complete and updated data are available.

SOMMARIO

Negli ultimi anni, la scienza dei dati (Data Science) è stata largamente impiegata da imprese e governi per trasformare i dati in informazione utile, in modo da poter prendere le decisioni migliori. In particolare, per quanto riguarda il settore pubblico, questo fenomeno ha acquisito sempre maggior importanza a partire dal 2007, quando sono stati definiti gli otto principi dei dati governativi aperti (Open Government Data). Questi dati sono prodotti o commissionati da enti pubblici e vengono resi liberamente accessibili e utilizzabili da tutti.

Lo scopo di questa tesi è proporre una metodologia di redazione di studio di fattibilità modificata, per trarre vantaggio dalla Data Science. Questa metodologia viene poi utilizzata per redigere uno studio di fattibilità riguardante la proposta di attuazione di un intervento per la sicurezza stradale. L'intervento riguarda la città di Roma, dove, in seguito all'analisi dei dati riguardanti gli incidenti stradali, è stato rivelato un aspetto preoccupante relativo ai giovani conducenti. Roma, infatti, presenta un numero estremamente alto di giovani conducenti di automobili coinvolti in incidenti stradali con danni a persone. Il dato è ancor più evidente se paragonato a quello di Londra.

Uno studio più approfondito sui dati ha permesso di identificare come causa del problema la guida pericolosa dei giovani conducenti di Roma. Per questo motivo, si è cercato un caso di studio che presentasse una soluzione ad un problema simile ed è stato individuato nello Staffordshire, una contea del Regno Unito. In questa regione, in seguito all'introduzione di un corso di guida avanzata, chiamato Pass Plus Extra e improntato a migliorare la consapevolezza dei rischi legati alla guida, è stato registrato un netto calo del numero di giovani conducenti coinvolti in gravi incidenti. Dopo aver valutato l'efficacia del corso tramite tecniche di Data Science, sono stati stimati i vantaggi economici e sociali che l'introduzione di un simile corso di guida avanzata avrebbe portato a Roma negli ultimi quattro anni e i risultati sono estremamente positivi.

Si ritiene che la metodologia innovativa presentata in questa tesi non sia applicabile solo al caso in esame, ma anche in tutti gli altri contesti in cui si abbiano a disposizione dati completi e aggiornati.

1. INTRODUCTION

The concept of Open Data has acquired great importance in the last few years. In some countries, like the United Kingdom and the United States of America, they are regularly analysed and investigated to detect possible issues or identify new opportunities. In these countries, not only private companies, but also governments and local authorities are taking advantage of the great potentiality of open data. The same can't be argued about Italy, where the open data management still presents some evident lacks and it could be highly improved.

One of the few examples of an updated Italian open data portal containing interesting information is represented by the official portal of Rome. This work starts from data about road accidents taken from this open data portal and data about the same topic taken from the national open data portal of the United Kingdom. Then, a comparison between Rome and London is performed and a serious issue for Rome is identified, which is the extremely high number of young drivers involved in personal injury accidents. Finally, inside a completely revised business plan structure, conceived to take advantage of data science, a possible remedy is found, the quick fix action is evaluated and the economic and social impact that it would have had, if it had been applied in Rome four years ago, is precisely calculated.

The most innovative aspect of this work is represented by the proposal of an innovative methodology for the realisation of a business plan, based on the huge potentiality of data science. The business plan is related to the possible introduction of a road safety action and a fundamental role is held by open data and data science techniques, as they are used to get extremely precise results that aren't achievable by traditional approaches in this area. In fact, all the actions

taken up to now in the road safety field in Italy have been based on and evaluated by general statistics. One of the main shortcomings of this type of studies is that they can't describe in detail the features of an issue and the consequences of a public intervention.

The structure of the work is the following.
In chapter 2, as a result of a literature review, the state of the art of open data, open government, data analytics and works on open data about road accidents is provided.
In chapter 3 all the operation needed to retrieve and analyse open data are described. Then, a comparison between Rome and London is performed and a detailed qualitative analysis of the issue about young road casualties in Rome is produced.
In chapter 4 a modified version of the classic business plan structure is shown. This structure, revised to take advantage of data science, is then used to write a business plan about a possible road safety action in Rome.
Finally, chapter 5 concerns the conclusions of this work.

2. STATE OF THE ART

In this chapter the main topics covered in this work are presented. Their state of the art, resulted from a literature review, is described. The following themes are treated: open data, open government and open government data (section 2.1), data analytics (section 2.2) and previous works on open data about road accidents (section 2.3).

2.1 Open Data, Open Government and Open Government Data

This section introduces the concepts of *Open Data*, *Open Government* and *Open Government Data*, by mentioning some relevant historical facts and giving some useful definitions and information.

2.1.1 Open Data History and Definition

Robert King Merton, an American sociologist, was the first one who theorized that knowledge has to be shared for the common good. In his essay "*The Normative Structure of Science*", published in 1942, he described the modern science as a community in which the scientists' behaviour should be controlled by four norms: universalism, "communism" (later renamed as communalism), disinterestedness and organized scepticism. According to the second norm, in particular, the scientific knowledge is identified as a common property and it's argued that

the results of scientific researches should be freely accessible to everyone, in order to let the science and the knowledge grow. [15]

The birth of the new information technologies, among which Internet has definitely had a crucial impact, led to more and more considerations about knowledge sharing and therefore data sharing.
In 2004 the Open Knowledge Foundation, a global non-profit network, was founded. Their mission is well illustrated by their own words:

"We envision a world where:

- *knowledge creates power for the many, not the few.*
- *data frees us to make informed choices about how we live, what we buy and who gets our vote.*
- *information and insights are accessible – and apparent – to everyone."* [12]

They released the first version of the Open Definition in 2005 and it is maintained today by an Advisory Council. The definition identifies the characteristics that data must have to be recognized as *open*:

"To summarize the most important:

- *Availability and Access: the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.*
- *Re-use and Redistribution: the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets.*
- *Universal Participation: everyone must be able to use, re-use and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed."* [13]

Nowadays there are several fields in which open data are used (for instance government, science, economics, etc.). If open data are produced or commissioned by public bodies, they are called *Open Government Data.*
In December 2007, thirty open government advocates defined the 8 principles of Open Government Data. Public government data shall be considered open if they are:

- Complete – *"All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations."*
- Primary – *"Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms."*
- Timely – *"Data is made available as quickly as necessary to preserve the value of the data."*
- Accessible – *"Data is available to the widest range of users for the widest range of purposes."*
- Machine processable – *"Data is reasonably structured to allow automated processing."*
- Non-discriminatory – *"Data is available to anyone, with no requirement of registration."*
- Non-proprietary – *"Data is available in a format over which no entity has exclusive control."*
- License-free – *"Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed."* [16]

2.1.2 Open Government

In order to explain the concept of *Open Government*, it is worth starting from the "*Memorandum on Transparency and Open Government"*, signed by the President of U.S.A. Barack Obama on January 21, 2009. It contains the three main

principles upon which all Open Government initiatives taken by other countries since then are based.
The principles are the following:

"Government should be transparent. Transparency promotes accountability and provides information for citizens about what their Government is doing. [...]

Government should be participatory. Public engagement enhances the Government's effectiveness and improves the quality of its decisions. Knowledge is widely dispersed in society, and public officials benefit from having access to that dispersed knowledge. […] Executive departments and agencies should also solicit public input on how we can increase and improve opportunities for public participation in Government. […]

Government should be collaborative. […] Executive departments and agencies should use innovative tools, methods, and systems to cooperate among themselves, across all levels of Government, and with nonprofit organizations, businesses, and individuals in the private sector. Executive departments and agencies should solicit public feedback to assess and improve their level of collaboration and to identify new opportunities for cooperation." [11]

Thus, the concept of *Open Government* is closely related to the concept of *Open Government Data* mentioned above.

United Kingdom and Italy, like other several countries, have been inspired by this memorandum. The former published the report "*Putting the frontline first: smarter government*" in December 2009, in which it was stated that one of the key action of the new plan was to "*radically open up data and public information to promote transparent and effective government and social innovation"* and it was announced the release of over a thousand public datasets made free for reuse. [7] In January 2010 the official Open Data portal of the UK Government (data.gov.uk) launched publicly and now it contains thousands of datasets about

different themes, among which environment, health, transport, education and society.
The latter joined the Open Government Partnership in September 2011, an international organization that looks for strong commitments from each member country to promote transparency and empower citizens by taking advantage of new technologies. The partnership currently includes 69 participating countries.
In April 2012 Italy summarized its programs and initiatives in an Action Plan and subsequently the Italian Open Data portal launched (dati.gov.it). After more than three years, the portal still has a big problem: most of the data are not updated, they belong to datasets placed into the portal and then forgotten, so they can't be considered interesting and they seem to go against the Open Government principles. [8]

Not only national Open Data portals were born in these years, but also local authorities launched their own ones. The portal of the city of Rome (dati.comune.roma.it) has been used in this work and so it deserves to be cited. Launched on October 3, 2012, now it contains plenty of data about 11 different areas, among which tourism, road accidents, environment, society and public administration. The data are continuously updated, for this reason they are a very good resource for people who want to get useful information from them.

2.1.3 The benefits of Open Government Data

Open Government Data can be searched and manipulated using standard tools, each citizen with basic knowledge of Information Technology can exploit them and, potentially, create value from them. The main beneficiaries of the value created from these data are: government, citizens and wider economy.

Government – Open Government Data are often provided with the intention of increasing the overall efficiency and effectiveness of government operations.

They improve transparency and accountability, they produce innovative and personalized public services and they enhance the interaction processes between citizens and government. [17]
The release of government data online and their reuse can lead to a considerable decrease of the number of questions daily received by public authorities. This produces a reduction in work-load and costs and makes easier for public employees to answer to the remaining questions, because necessary information is also faster to find. A good example of what has just been stated is the case of the Netherlands, where the Ministry of Education has published education-related data for reuse and this has significantly improved the efficiency of the relative public services. [13]

Citizens – The publication of Open Government Data is obviously considered an important and innovative service for citizens: it increases the public participation and it gives more responsibilities to the public authority, as everyone can verify how it is working. [17] In Finland and in the United Kingdom, just to give an idea, there are two projects, respectively called "tax tree" and "where does my money go", that let the people know how the government spends tax money. Moreover, people can make better decisions in their own life, there are plenty of examples of new mobile apps and new services built upon Open Government Data. For example, they can help in finding the best place to live (like *mapumental.com* in the UK and *mapnificent.net* in Germany), the nearest place where it is possible to walk the dog or even the nearest public toilet (like *findtoilet.dk* in Denmark). [13]

Wider economy – According to the Open definition, open data can be used for commercial purposes. It's a proven fact that, if it's allowed the reuse of some data at very low or zero cost, developers and private enterprises can take advantage of this information and create more and more products to be marketed. As a consequence, national economy is strengthened and the government can receive revenue in the form of taxes. [17]

2.2 Data Analytics

In this section *Data Analytics* and the three categories in which it can be divided are presented. Data Analytics covers a huge number of methods and techniques. Two of them, the Clustering technique and the Holt-Winters method, have been used in this study and therefore a brief explanation of them is provided too.

2.2.1 Data Analytics Definition

Data Analytics is the science of analysing raw data in order to transform these data into insight and make better business decisions.
Thanks to the very useful information and knowledge it gives to the managers, its popularity has extremely increased in the last years and now a new concept has been introduced in the business world: *analytics-as-a-service*. This term refers to the utilization of web-delivered technologies for performing Data Analytics, in order to take advantage of *as-a-service* characteristics, like pay-per-use and high scalability.
Data Analytics is divided into three main categories: Descriptive Analytics, Predictive Analytics and Prescriptive Analytics.

2.2.2 Descriptive Analytics, Clustering and K-medoids

"*What happened?*"

Descriptive Analytics focuses on the past. It takes as input data regarding what happened until now, it summarizes them and it tries to discover interesting information that can be useful for describing a problem or determining opportunities, it can illustrate a scenario in which it's possible to take advantage of that particular situation.

Statistics (sums, averages, etc.), reports about historical facts and transactions and every other technique that aggregates raw data and make them interpretable by humans belong to this category. [3]

One of the most used technique is *clustering*. The main task of *clustering* (or *cluster analysis*) is to find similarities between data and divide them into different groups, so that data objects inside the same cluster are similar to each other and dissimilar to objects belonging to the other groups.
There are a lot of different algorithms able to perform this action and there isn't an algorithm that can always be defined as the best one, each algorithm could be the most appropriate to be used for every specific case. It depends on the type of the data (they can be numeric, ordinal, binary, categorical), the data characteristics (for instance high/low number of outliers or missing values) and the algorithm characteristics that are needed (it can be more/less scalable or easy/hard to be interpreted).

Moreover, clustering methods can be divided into 4 categories, depending on how they work. The categories are:

Hierarchical	Initially, groups are composed by an element. Then, at each step, the most similar groups are merged together, until a single group is obtained (or, on the contrary, the algorithm starts from a cluster containing all data objects and, at each step, it splits up one group until each cluster is composed by a single item).
Partitioning	Given k clusters, the algorithm assigns each data object to a cluster, trying to obtain a result in which items of the same group have high similarity.

Density-based	A local similarity criterion is used, as density-connected points are grouped together. Clusters of arbitrary shape can be discovered.
Model-based	Data are thought to be generated by a mathematical model. Some probability distribution is used to assign each item to the clusters.

Table 1 – CLUSTERING METHODS CATEGORIES

After an algorithm has finished its execution, usually, an evaluation of its results is performed. Based on intra-class and inter-class similarity, it is possible to choose the number of clusters that better represents the data.

A clustering problem that is worth examining in depth, as it has been analysed in this study, is called *k-medoid*. It refers to the class of the partitioning methods and its goal is to minimize the distance between the elements of each cluster and the item that represents the most central point (medoid) of that cluster.
The most used and famous algorithm for finding a solution (a local minimum) to the k-medoid problem is the *PAM* algorithm (*Partitioning Around Medoids*). It can work with the *dissimilarity matrix*, that is a matrix containing all the pairwise distances between the items, or directly with the data, by calculating the distance matrix it needs at first. It is composed by two phases: the *build* phase and the *swap* phase.
During the build phase, the observation whose sum of dissimilarities is minimum is chosen as the first medoid. Then, the other k-1 initial medoids are selected iteratively by minimizing the distance of the other data objects to their nearest medoid.
At the end of the build phase, k initial medoids are obtained and the swap phase begins. The algorithm swaps each selected medoid with a non-selected object and, if the sum of the dissimilarities between all objects and their respective

medoid decreases, then a new configuration has been found. The process continues until no further optimization is possible.
The final k medoids are the objects that best represent their own cluster and all the other observations are assigned to the cluster whose medoid is more similar to them.

2.2.3 Predictive Analytics and Holt-Winters Method

"*What will happen?*"

Here the focus is on the future. As the word *predictive* suggests, Predictive Analytics uses associations between data and trends to determine how a phenomenon is going to proceed in the future. [3]
Data mining, text mining and machine learning techniques belongs to this category, but also statistical time series forecasting is very popular. In general, each model and each technique that takes data about the past to predict future data is considered as part of Predictive Analytics.

The method used in this study is called *Holt-Winters* method, it performs time series forecasting and it is used when there is seasonality in the data. There are two versions of the method: the additive method and the multiplicative one. The former is preferred when the variations are roughly constant in the series, the latter is better when the variations change proportionally to the level of the series and therefore it's worth considering the percentage of the variations.
Both of them are comprised of 4 equations, the first one represents the forecast equation and the other ones are needed to calculate the estimated level [l_t], trend [b_t] and seasonal component [s_t] at time t. The equations of the additive model, which is the model that has been used here, are:

$$\hat{Y}_{t+h} = l_t + hb_t + s_{t-p+1+(h-1) \bmod p}$$
$$l_t = \alpha(Y_t - s_{t-p}) + (1-\alpha)(l_{t-1} + b_{t-1})$$

$$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1}$$
$$s_t = \gamma(Y_t - l_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-p}$$

where h = 1, 2, … represents the number of steps ahead into the future (starting from the last observation we have) and p stands for the period length of the seasonality (for instance, monthly data has p = 12). α, β and γ are called *smoothing parameters*. They are needed to calculate the weighted average inside their respective equations and they can be specified by the forecaster or obtained from the observed data. In this case the values that minimize the sum of the squared errors (and therefore maximize the forecast accuracy) are chosen.

In the first equation, the h-step ahead forecast is obtained by adding the level at time t, the trend component and the seasonal index taken from the last year of the observations. The level is calculated as a weighted average between the observation without its seasonality and the non-seasonal forecast for time t. The weighted average inside the trend equations makes use of the difference between the last two levels and the previous value of trend. Lastly, the seasonal index is calculated as a weighted average between the current seasonal index and the index that refers to the same season in the previous year.

2.2.4 Prescriptive Analytics

"*What should I do?*"

The goal of Prescriptive Analytics is to find the best actions to improve the current situation. Typically, a Data Analytics expert shows to a manager some possible scenarios, better than the current one, in which the company could be placed in the future and the actions needed to reach that situation. This is obtained after the execution of some mathematical algorithm on the provided data, based on constraints and requirements given as input.

Prescriptive Analytics, together with managers' experience, can get the best course of action for each particular situation and let the company reach its objectives. [3]

2.3 Previous works on Open Data about Road Accidents

This part describes how open data about road accidents have been analysed by previous works and researches so far. The focus is on accidents occurred in Rome and in the United Kingdom, the two cases covered by this work.

2.3.1 Rome Open Data

So far, the only works based on Rome's open data about road accidents are websites and applications that provide a map visualization of the crashes. This is probably due to the fact that open data have appeared in Italy very recently and their enormous potential has not been fully exploited yet.
An example of what has just been stated could be "*Roma Crash Map*", a visualization tool, developed some years ago, able to show the density of car accidents in the different Rome municipalities, also by grouping them according to their characteristics, such as light conditions, weather conditions and periods of time.

2.3.2 United Kingdom Open Data

There are several websites containing map visualization tools that work on data taken from the UK open data portal. *http://www.cyclestreets.net/collisions/* and *http://www.crashmap.co.uk* are just two examples of this category.
Both show a map in which the user can detect the exact position of all road accidents from 2008 to 2014 and they also provide some information about each single crash. Three different types of markers, distinguished by their colour (ranging from yellow to red), differentiate between accidents with slight, serious and fatal consequences.
The former website focuses on accidents involving cyclists. The latter lets the user filter crashes by the location, the year in which they took place, their severity and the types of casualties.

However, it's worth paying close attention to the researches conducted on crashes open data. They are often commissioned by public authorities and their goal is to find some interesting aspects that can help in improving road safety.
One of the most recent one was published in March 2015 and commissioned by the Welsh Government. Its purpose was to evaluate the effects of the Pass Plus Cymru, a course designed to improve young drivers' skills. In addition to consultations with people from the road safety staff and a careful literature review to analyse the impact of other similar courses, the researchers took in consideration several statistics based on data from the UK open data portal. [14]

3. PROBLEM ANALYSIS

This chapter focuses on the analysis of the data about drivers involved in personal injury accidents in Rome and in London. In particular, the first section explains all the operations performed to get the data that have been used and make them ready to be examined by a data analytics tool. In the second section a comparison between Rome and London is shown by means of analyses on specific attributes and a problem relating to young drivers in Rome comes to light. Finally, the third section illustrates the implementation of a clustering analysis on Rome's data and the obtained results. These are used to better describe the characteristics of the issue about young drivers involved in car accidents in Rome.
All data analyses have been performed by using the R software and its libraries.

3.1 Data Retrieval and Preliminary Operations

This work is based on data about road accidents taken from two open data portals. The first one is the official open data portal of the United Kingdom Government and road safety data can be found at *https://data.gov.uk/dataset/road-accidents-safety-data*. The second one is the open data portal of the city of Rome and the webpage from which it's possible to download data about road accidents is *http://dati.comune.roma.it/cms/it/incidenti_stradali.page*.
Moreover, data about daily rainfall in Rome and data about the population of Rome and the population of London are used. Daily rainfall data are taken from the open data portal of the Lazio region and precisely from *http://dati.lazio.it/catalog/dataset/serie-storica-agrometeo*. Data regarding the population

of Rome can be found at *http://dati.comune.roma.it/cms/it/popolazione_societa.page* and the estimates of the population of London are taken from the Office for National Statistics (ONS) website, *http://www.ons.gov.uk*.
In the following paragraphs the main characteristics of the data found in these datasets are presented. Then, a brief explanation of the operations needed to make these data ready to be investigated by data analytics tools is provided.

3.1.1 UK Road Safety Data

The data are provided in CSV format and they belong to STATS19 database, which is a collection of personal injury road accidents that took place in Great Britain from 1979. The data about accidents from 2005 and 2014 are more precise about the age of the drivers, as they include the exact age. For this reason, they have been used in this work. Previous data, instead, only provide information about five-year and ten-year age bands.
There are three different datasets: Accidents, Vehicles and Casualties. Only the first two of them have been taken into account, because they are useful for the purposes of this study.
Moreover, a document that acts as a data guide is provided. It contains tables specifying the meaning of each value associated to the variables.
In order to better understand the differences between the values of a certain variable, the STATS20 manual has also been used. It explains in detail all the information a Police Officer has to gather when a road accident resulted in a personal injury is reported. The manual can be found at the following link: *https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/230596/stats20-2011.pdf*.

3.1.2 Road Accidents Data of Rome

The data taken from the portal of Rome are of two different formats: JSON and XML. JSON data are the most recent and updated ones, but they lack information about accidents occurred on June 30 and December 31 for each year, so XML data regarding these two days need to be employed too.
Four different datasets are provided: Accidents, Vehicles, Persons and Pedestrians. The first three of them have been used in this work and the time interval considered is between January 2008 and June 2015, data regarding periods of time after this point are not available yet.
Each single dataset covers a period of six months inside this time interval.

3.1.3 Other Data

The data about the population of Rome and London can be directly downloaded from their respective websites and they are contained in CSV files.
In order to get the daily rainfall data from the open data portal of the Lazio region, instead, it is necessary to write and run a script able to access to the remote resources and export JSON data. These data, then, are saved to a JSON file. There is one weather dataset per year and each dataset is identified by an id, which is needed to access the resource. The script, written in Python language, is the following.

```
import json
import urllib.request

id = '3144dd7a-0828-4404-9152-8a0569647ca2'
url = ('http://dati.lazio.it/catalog/api/action/datastore_search?' +
        'resource_id=%s' +
        '&q=ROMA' +
        '&limit=6000') % (id)
data = urllib.request.urlopen(url)
json_string = data.read().decode('utf-8')
weather_string = json.loads(json_string)
weather_records = weather_string['result']['records']
rainfall_records = [obj for obj in weather_records
    if(obj['Grandezza'] == 'PREC_TOTG' and
    (obj['Stazione'] == 'ROMA Lanciani' or obj['Stazione'] == 'ROMA Ponte Nona'))]
with open('weatherRoma2015.json', 'w') as outFile:
    json.dump(rainfall_records, outFile)
```

Figure 1 – PYTHON SCRIPT TO ACCESS WEATHER DATA

It selects data regarding the total daily rainfall ("*PREC_TOTG*") registered by two weather stations of Rome, "*ROMA Lanciani*" and "*ROMA Ponte Nona*". The former station is taken as the primary source of data, because it is the nearest to the centre of Rome, but it lacks information about ten days for the entire period from January 2008 to June 2015, hence data from the latter station are taken into account for these specific days. The additional info about how many weather observations are needed (up to 6,000) has to be included because, otherwise, only a part of them are sent back in the response.

3.1.4 Preliminary Operations on Data

Files containing UK road safety data, data about daily rainfall in Rome and data about the population of Rome and London can directly be read by the R tool, as there aren't any obstacles to the creation of data frames for these data.
On the contrary, as it has been stated before, the data about road accidents in Rome present some problems.

First of all, two different formats (JSON and XML) have to be considered, so, in order to obtain the final data frame to be used, observations taken from files in both formats are required.

In addition, some fields of the JSON data contain strings taken "as is" from user input and not all these strings can be considered valid JSON strings. There are several cases, in fact, in which special characters, such as the double quotation mark (") and the backslash (\), are not preceded by the escape character, which is the backslash itself. Each string having this issue cannot be accepted as valid inside a JSON structure. In order to solve this situation, an other script has been written in Python language and it has been run for every dataset having the issue.

The following script handles the dataset about road accidents occurred in Rome in the first six months of 2014.

```
import re

pattern = '"DaSpecificare":"(.+)", "NaturaIncidente":"'

with open('json_incidenti_01.01.2014_30.6.2014.json', encoding='ISO-8859-1') as inFile:
    with open('accidents2014_1.json', 'w') as outFile:
        for line in inFile:
            result = re.search(pattern, line)
            if result:
                startPos = result.start()
                endPos = result.end()
                substring = result.group(1)
                substring = substring.replace('\\','\\\\')
                substring = substring.replace('"','\\"')
                line = (line[:startPos] +
                        '"DaSpecificare":"' +
                        substring +
                        '", "NaturaIncidente":"' +
                        line[endPos:])
            outFile.write(line)
```

Figure 2 – PYTHON SCRIPT TO HANDLE JSON FORMATTING ERRORS

For each line of the JSON file, a pattern is searched. The pattern makes use of a regular expression: the sequence of symbols *(.+)* indicates a string with length greater than zero placed between the substring on the left and the substring on the right. Basically the script looks for the string that represents the value of

the attribute *DaSpecificare*, which is followed by the attribute *NaturaIncidente* inside the JSON file.
When such a string is found, it is passed to the variable *substring* (*result.group(1)* refers to the searched string) and, if it contains backslashes or double quotation marks, an escape character is placed on their left. Every time a backslash character is written inside a Python string, it has to be expressed by two backslashes, because otherwise its role of escaping character is taken into account. After this operation has been completed, the line is reconstructed and it can be written on the output file.

Once these operations have been completed, all data can be imported into R (R libraries "*jsonlite*" and "*XML*" have been used) and stored into different data frames. Then, data frames containing observations of the same type but regarding different time periods need to be aggregated into single data collections. Finally, the data frames Accidents and Vehicles about UK data are joined to obtain the final single data frame containing UK observations. The same is done with the the data frames regarding accidents, vehicles, people and daily rainfall about Rome.

As a last remark, it's worth specifying that not all the observations about Rome's accidents have been confirmed: a few of them present the logical value *False* for the attribute *CONFERMATO* and, thus, they have not been taken into account in this work. They typically represent unrevised copies of existing observations or partial data with several missing values that have been inserted into the dataset by mistake.

3.2 Comparison between Rome and London

Among the data about accidents in the United Kingdom, the observations regarding the crashes in London have been considered, because London is the city with the most similar features compared to Rome and, then, it can be used to make a valid comparison and find the specific characteristics of the accidents occurred in Rome.
London is identified by 34 different values of the attribute *Local_Authority_.Highway*. inside the UK dataset: one corresponds to the Heathrow airport and the code is "*EHEATHROW*", the remaining 33 values correspond to the 33 districts (or boroughs) into which London is divided and all the codes related to them begin with "*E09*".

3.2.1 Car Drivers' Age Differences

As a first comparison between Rome and London, the car drivers' age differences have been investigated. The *Age_of_Driver* is present as attribute in the UK data, while data about accidents occurred in Rome only provides the year of birth. Therefore, what can be done is to calculate the difference between the year in which the accident took place and the year of birth and the result represents a good estimate of the driver's age (actually it could be the exact age or it could be wrong by one year, by considering the driver as one year older). From now on this value will be called *Estimated_Age* for convenience and it will be considered as a further variable of Rome's data.
In addition, the value "*Car*" inside the *Vehicle_Type* attribute of the UK data, refers both to cars and light quadricycles and there is no way to distinguish between them. Among the "*Instructions for the Completion of Road Accident Reports*" it is specified that the term "Car" includes each type of car "*and similar four-wheel drive vehicles*". [4] For this reason, in order to make UK data as much

as possible comparable to Rome's data, the vehicle types "*Autovettura privata*" (car) and "*Quadriciclo leggero*" (light quadricycle) have been taken into account for Rome's observations. Actually, data regarding light quadricycles seem to be little relevant for the comparison, as only the 1.1% of the car drivers analysed, in the case of Rome, refers to light quadricycles.
In order to select only data about drivers, the value "*Conducente*" of the attribute *TipoPersona* has been specified.

The time period between January 2008 and December 2014 has been considered, as data from both data frames (Rome and United Kingdom) are available for these years. The outcome of this comparison is shown in the graph below.

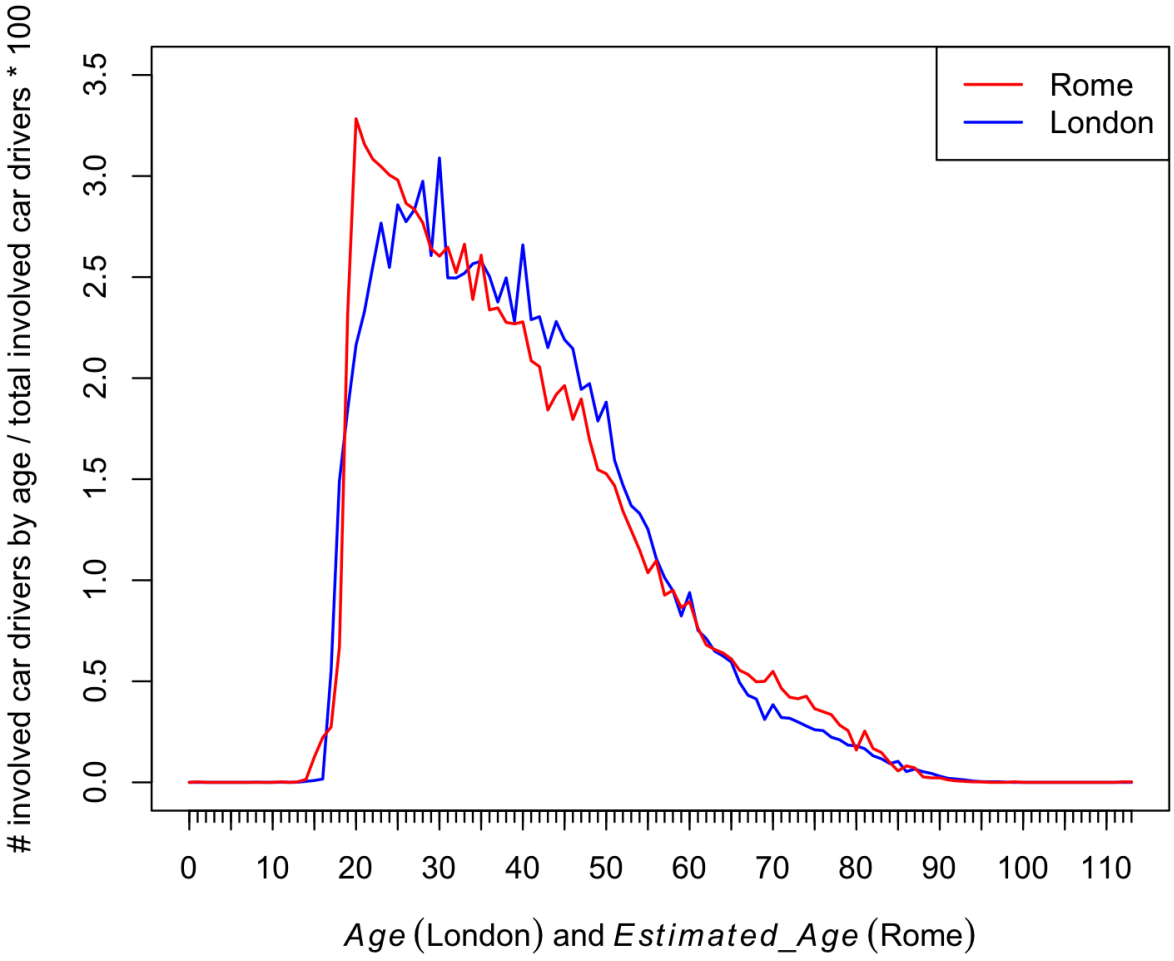

Figure 3 – NUMBER OF INVOLVED CAR DRIVERS BY AGE [Years: 2008-2014]

The horizontal axis indicates the car drivers' age (as it has been stated before, in the case of Rome the difference between the year of the accident and the year of birth is considered). The vertical axis represents the percentage of car drivers with a specific age involved in personal injury accidents compared to the total number of car drivers involved. Observations having missing values relating to attributes *Age_of_Driver* and *Estimated_Age* have not been taken into account.

The first thing that stands out is that young drivers seem to be more involved in personal injury accidents in Rome. In fact, the red line reveals a peak for the percentage referring to the 20 year-old drivers, which is not present in the blue line. Moreover, if the first five years in which it is possible to drive a car are considered, the following results are achieved:

$$Rome\ (Estimated\_Age\ 18-22): 12.52\%$$
$$London\ (Age\ 17-21): \ 8.38\%$$

The difference is even more evident given that, due to the *Estimated_Age* attribute, which can be the real age or the age overestimated by one year, some of the 22-year-old car drivers are classified as having *Estimated_Age* of 23 and, therefore, they are not included in the percentage.

3.2.2 Trends for Young Drivers by Year and Vehicle Type

In order to get more information about this phenomenon, the trend of the number of young drivers involved in accidents by year has been analysed. In addition, the data about the population have been employed for having a more precise idea of the problem. As a result, a graph showing how many young drivers and riders have been involved in personal injury accidents compared to the population of the two cities has been created.

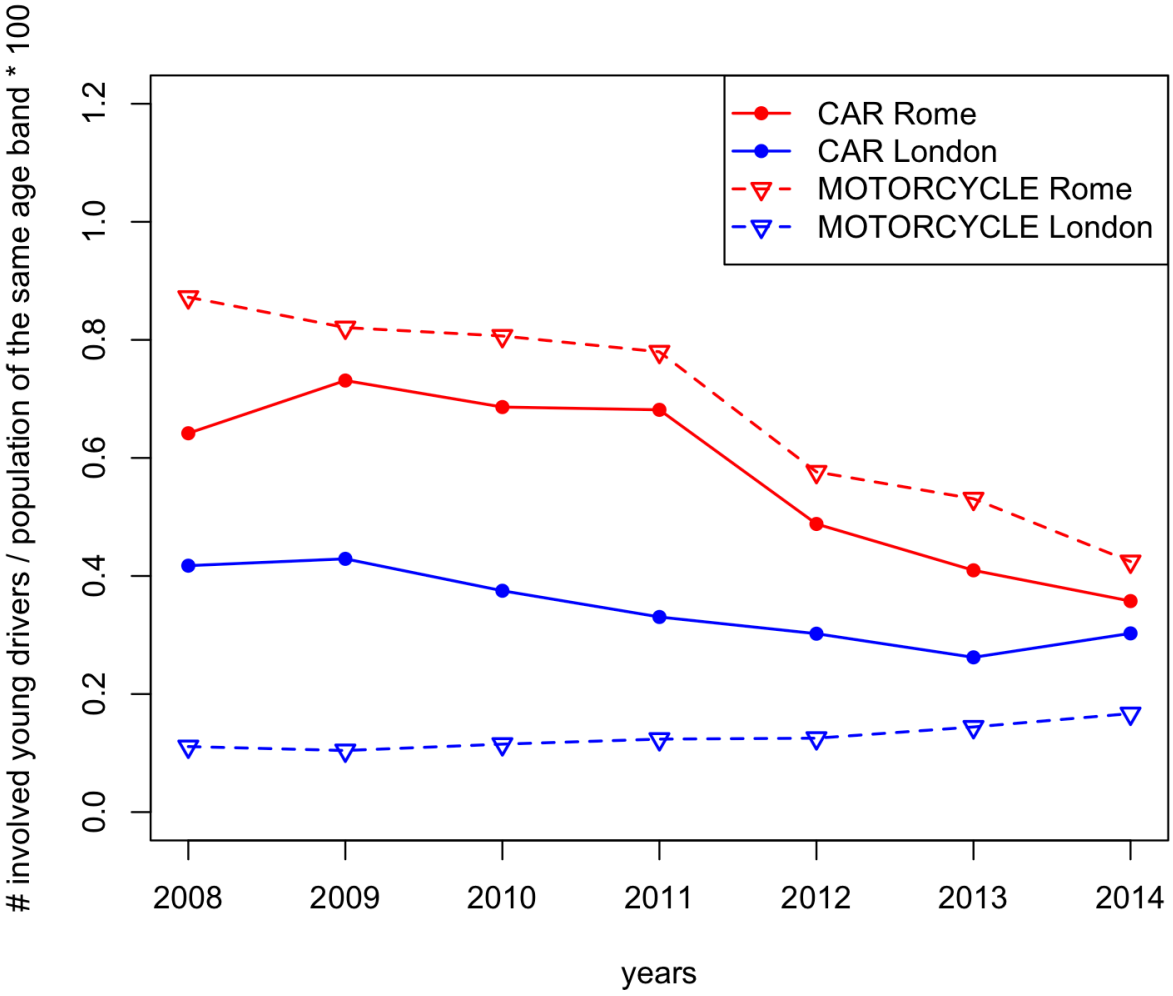

Figure 4 – NUMBER OF INVOLVED YOUNG DRIVERS BY VEHICLE AND YEAR
[Age: 18-22 Rome, 17-21 London]

The graph above clearly indicates a cultural difference between Rome and London referring to motor vehicles: in Rome young people seem to prefer motorcycles, as they are more involved in road accidents than cars; in London young car drivers are much more involved in accidents than riders. Therefore, the situation of Rome is alarming: the 2014 value about car drivers, despite being close to the London's one, is really worrisome, because, if the value about motorcycle riders is added, the result indicates that young people in Rome have a much higher inclination to accidents than young people in London.

As a last remark about the graph, there is a clear decrease in the number of young drivers and riders involved in accidents in Rome in 2012, but this is mainly due to the fall in the number of registrations for driving license courses (-19% in 2011 compared to the previous year and the same trend in 2012), the fall in the number of cars sold and the rise in gasoline price. [1] [18] For this reason, this decrease can't be seen as a completely positive fact.

3.2.3 Trends for Young Car Drivers by Hour and Day of Week

Concluding the comparison between Rome and London, it's interesting to analyse the different hours in which accidents have occurred during the week.
The information about the specific day of week is not present in Rome's data, but, having the value of the date of each accident, the variable *Day_of_Week* has been added to the data frame and its values have been easily calculated.
What is obtained is quite surprising and it's shown in the following graphs.
In both of them, the blue line represents the mean of the values about Monday, Tuesday, Wednesday and Thursday, this choice has been made because they have similar trends through the daily hours. Moreover, only the value of hours has been taken into account, discarding the value of minutes.

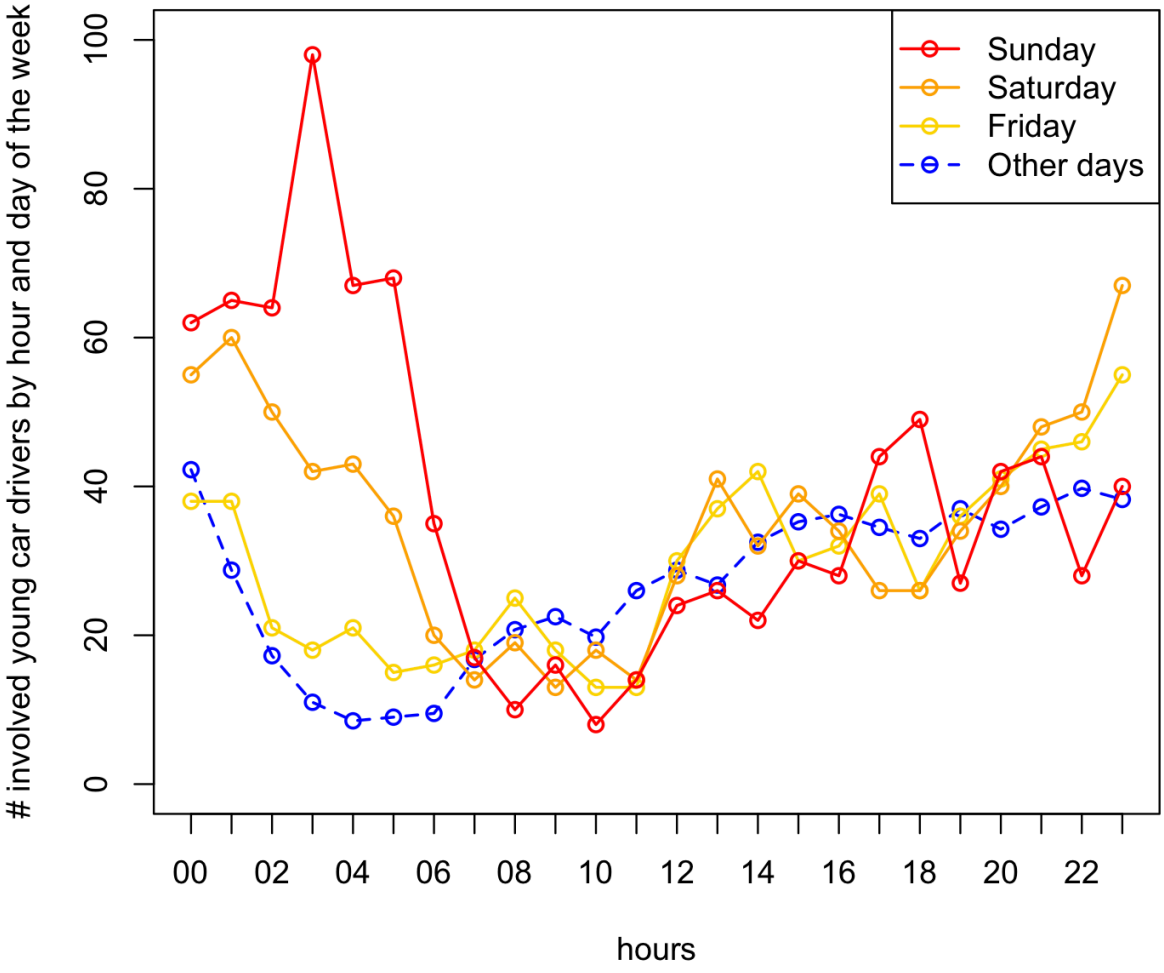

Figure 5 – NUMBER OF INVOLVED YOUNG CAR DRIVERS BY HOUR AND DAY OF THE WEEK [Rome, Years: 2008-2014] [Estimated_Age: 18-22]

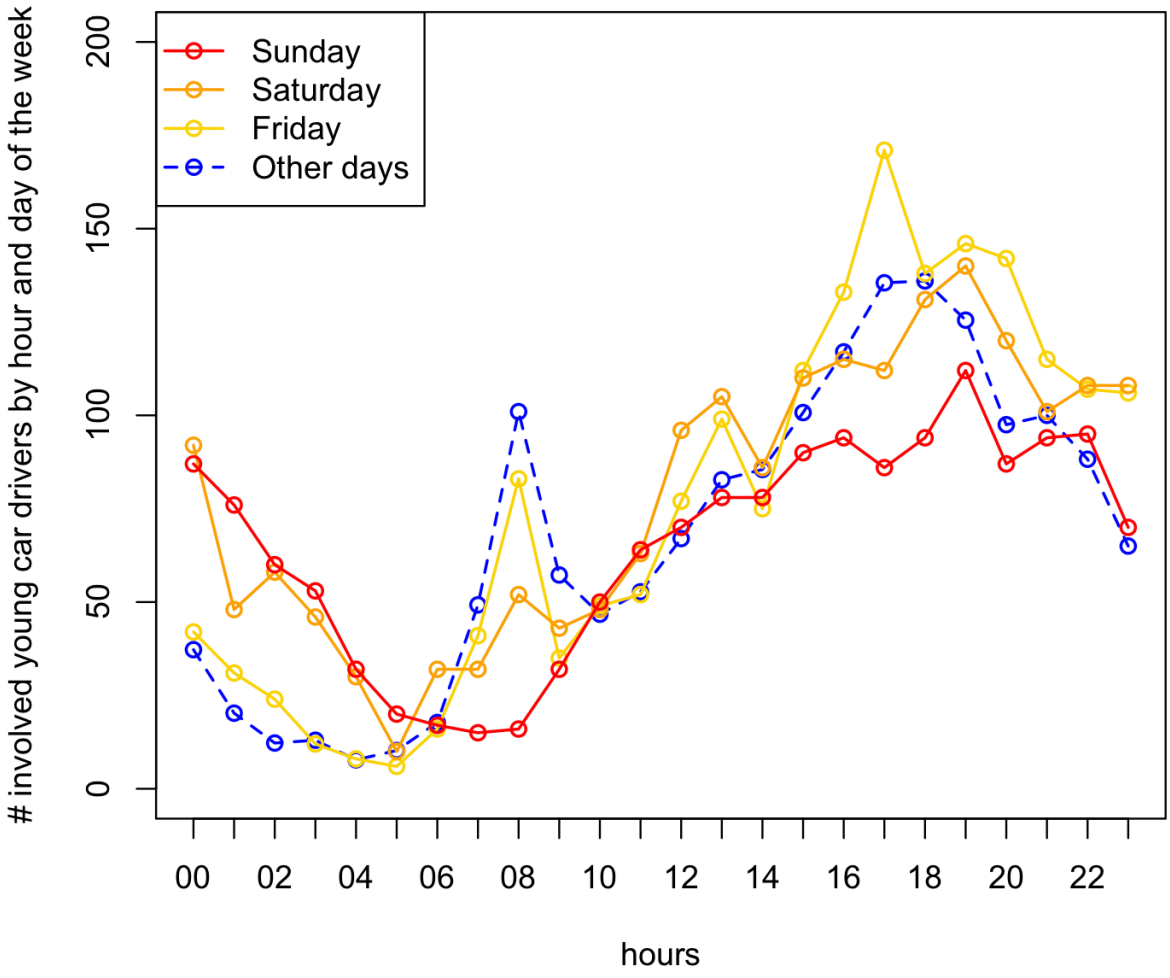

Figure 6 – NUMBER OF INVOLVED YOUNG CAR DRIVERS BY HOUR AND DAY OF THE WEEK [London, Years: 2008-2014] [Age: 17-21]

The main differences between Rome and London are two. First of all, in London there is a peak in the business days from 7:00am to 9:59am, which is not present in Rome. Maybe this could mean that in Rome the car is less used for going to school or to work. The second big difference regards the highest peak: in Rome it is placed on Sunday between 3:00am and 3:59am, while in London on Friday around 5:00pm. It is clear that young people in Rome drive much more unsafely at night during the weekend, especially when it's time to go home.

3.3 Clustering on Rome's Data

In order to investigate more in detail the characteristics of the accidents in which young car drivers are involved in Rome, a clustering analysis has been performed on these data (the time period is still from January 2008 to December 2014). R software and its library *cluster* have been used and, in particular, a partitioning algorithm called *PAM* (*Partitioning Around Medoids*) has been employed to assign each observation to a specific cluster. This algorithm tries to minimize the dissimilarity between elements of the same group and, given k as the number of clusters in which we want to divide the observations, it finds k data objects that represent the most central observations of each group (the distance between them and the other elements of the group is minimum). These k elements are called *medoids*. After the algorithm has terminated, the features that characterize each cluster can be identified and, as a consequence, an overview of the characteristics of the personal injury accidents involving young car drivers is obtained.

3.3.1 Data Pre-Processing

First of all, not all the attributes of Rome's data about car drivers involved in accidents are useful for this particular study. For this reason, a dimensionality reduction has been performed to eliminate irrelevant fields. The variables of interest are: *DataOraIncidente* (date and time of the accident), *Localizzazione2*, *ParticolaritaStrade*, *FondoStradale* (from now on called *Road_Surface_Conditions*), *CondizioneAtmosferica* (*Weather_Conditions*), *Illuminazione* (*Light_Conditions*), *AnnoNascita* (year of birth) and *Sesso* (*Sex_of_Driver*). As it has been stated before, using the date and the time of the accident and the year of birth, three new attributes have been calculated: *Estimated_Age*, *Day_of_Week* and

Hour. In addition, *Localizzazione2* and *ParticolaritaStrade* contains all the information needed to characterize the junction details, hence the new variable *Junction_Type* has been created starting from them.

Secondly, only the observations having *Estimated_Age* between 18 and 22 have been taken into account, as they represent the first five years in which it is possible to drive a car in Italy and, moreover, this is the car drivers' age band that seems to be most in difficulty compared to London. Regarding the vehicle type, instead, the values "*Autovettura privata*" (car) and "*Quadriciclo leggero*" (light quadricycle, much less relevant than car in this case, as it is present only in a minimum part of the observations) have been chosen, as they are comprised in the more general term "*Car*" used in the UK data and they have been employed in the comparison with London. In order to select only the data objects about drivers, then, the value "*Conducente*" has to be specified for the variable *TipoPersona* (type of person).

Furthermore, a discretization has been applied to values about total daily rainfall, which have been divided into five classes in the following way, obtaining the new variable *Daily_Rainfall_Class*:

Total Daily Rainfall (mm)	***Daily_Rainfall_Class***
0	0
0.1 – 9.9	1
10.0 – 24.9	2
25.0 – 49.9	3
> 49.9	4

Table 2 – DAILY RAINFALL CLASSES

Each daily rainfall class has been then assigned to observations indicating a wet road surface, while class 0 has been set for all the other observations. In fact, the additional information about daily rainfall has been used to differentiate between situations in which maybe there was a great amount of water on the surface and situations in which the road surface was probably little more than damp.

Finally, regarding the other attributes, in order to maintain a classification criterion that allows to compare these data with the UK data, some values have been aggregated and all the performed transformations are shown in the tables below.

Weather_Conditions

Before aggregation	After aggregation
"*Sereno*" (sunny) "*Nuvoloso*" (cloudy)	"*Fine*"
"*Pioggia in atto*" (raining) "*Grandine in atto*" (hail)	"*Rain*"
"*Nebbia*" (fog) "*Foschia*" (mist)	"*Fog*"
"*Nevicata in atto*" (snowing)	"*Snow*"
"*Vento forte*" (high wind) "*Sole radente*" (slanting sunlight)	"*Other*"

Table 3 – AGGREGATIONS ON WEATHER CONDITIONS VALUES ABOUT ROME'S DATA

Road_Surface_Conditions

Before aggregation	After aggregation
"*Asciutto*" (dry)	"*Dry*"
"*Bagnato (brina)*" (wet - hoarfrost) "*Bagnato (pioggia)*" (wet - rain) "*Bagnato (umidità in atto)*" (damp)	"*Wet*"
"*Ghiacciato*" (ice)	"*Ice*"
"*Con neve*" (snow)	"*Snow*"
"*Sdrucciolevole*" (slippery) "*Viscido da liquidi oleosi*" (oil) "*Con grandine*" (hail)	"*Other*"

Table 4 - AGGREGATIONS ON ROAD SURFACE CONDITIONS VALUES ABOUT ROME'S DATA

Light_Conditions

Before aggregation	After aggregation
"*Ore Diurne*" (daylight)	*2*
"*Sufficiente*" (sufficient)	*1*
"*Insufficiente*" (insufficient) "*Inesistente*" (no lighting)	*0*

Table 5 – AGGREGATIONS ON LIGHT CONDITIONS VALUES ABOUT ROME'S DATA

ParticolaritaStrade [PS] ***Localizzazione2 [L]***	***Junction_Type***
Before aggregation	After aggregation
[PS] "*Rotatoria*" (roundabout)	"*Roundabout*"
[PS] "*Intersezione semaforizzata*" [L] "*all'intersezione semaforizzata con*" (junction with automatic traffic signal)	"*Junction with Traffic Signal*"
[PS] "*Intersezione regolata dal vigile*" (junction with authorised person)	"*Junction with Authorised Person*"
[PS] "*Incrocio*" (junction) [PS] "*Intersezione non regolata/non segnalata*" (uncontrolled/unsignalled junction) [PS] "*Intersezione stradale segnalata*" (signalled junction)	"*Other Junction*"
[PS] other 15 values indicating places different from a junction	"*No Junction*"

Table 6 – AGGREGATIONS ON JUNCTION TYPE VALUES ABOUT ROME'S DATA

A clarification is needed about the last table, because getting exhaustive information about junction type is not so easy. "*Junction with Traffic Signal*" refers both to the value "*Intersezione semaforizzata*" of *ParticolaritaStrade* and to the value "*all'intersezione semaforizzata con*" of *Localizzazione2*. In this last case, simultaneously, *ParticolaritaStrade* always contains values indicating a junction. These values, for all the other instances of *Localizzazione2*, correspond to different junction types.

At the end of the data pre-processing operations, 5,083 observations about young drivers are organized in a table having nine variables. Three of them are categorical and they are *Weather_Conditions* (possible values: "*Fine*", "*Rain*", "*Fog*", "*Snow*", "*Other*"), *Road_Surface_Conditions* ("*Dry*", "*Wet*", "*Ice*", "*Snow*", "*Other*") and *Junction_Type* ("*Roundabout*", "*Junction with Traffic Signal*", "*Junction with Authorised Person*", "*Other Junction*", "*No Junction*"). Five are ordinal and they are *Hour* (from "*00*" to "*23*"), *Day_of_Week* (from *1* [Monday] to *7* [Sunday]), *Daily_Rainfall_Class* (from *0* to *4*), *Light_Conditions* (from *0* to *2*) and *Estimated_Age* (from *18* to *22*). One is binary and its value depends on the gender of the driver.

3.3.2 Choice of the Number of Clusters

In R, the library *cluster* has a function, called *daisy*, that is used to compute the dissimilarity matrix between data objects. The table obtained at the end of the pre-processing phase has been given as input to this function and, as a result, a numeric value of the distance between each couple of observations has been obtained. In order to perform this calculation, the *Gower's coefficient* has been used as the metric. According to this metric, the dissimilarity between two objects is achieved as the weighted mean of the contributions of each attribute. If a variable is categorical or binary, the distance between two values is 0 in the case of equal values and 1 otherwise. If a variable is ordinal, instead, the distance between two values is the difference between these values, divided by the total range. The function can also work with missing values and, if a variable contains a missing value, that variable is not considered in the distance computation.

As is established practice, once the distance matrix is computed, the optimal number of clusters is found by trying to divide the data objects into k groups, for k comprised in a specific range, and choosing the value of k for which the similarity between objects of the same cluster and the dissimilarity between objects

of different clusters is maximum. This operation has been performed making use of the PAM algorithm of the library *cluster*, with values of k ranging from 2 to 50. What has been obtained is shown in the following graph.

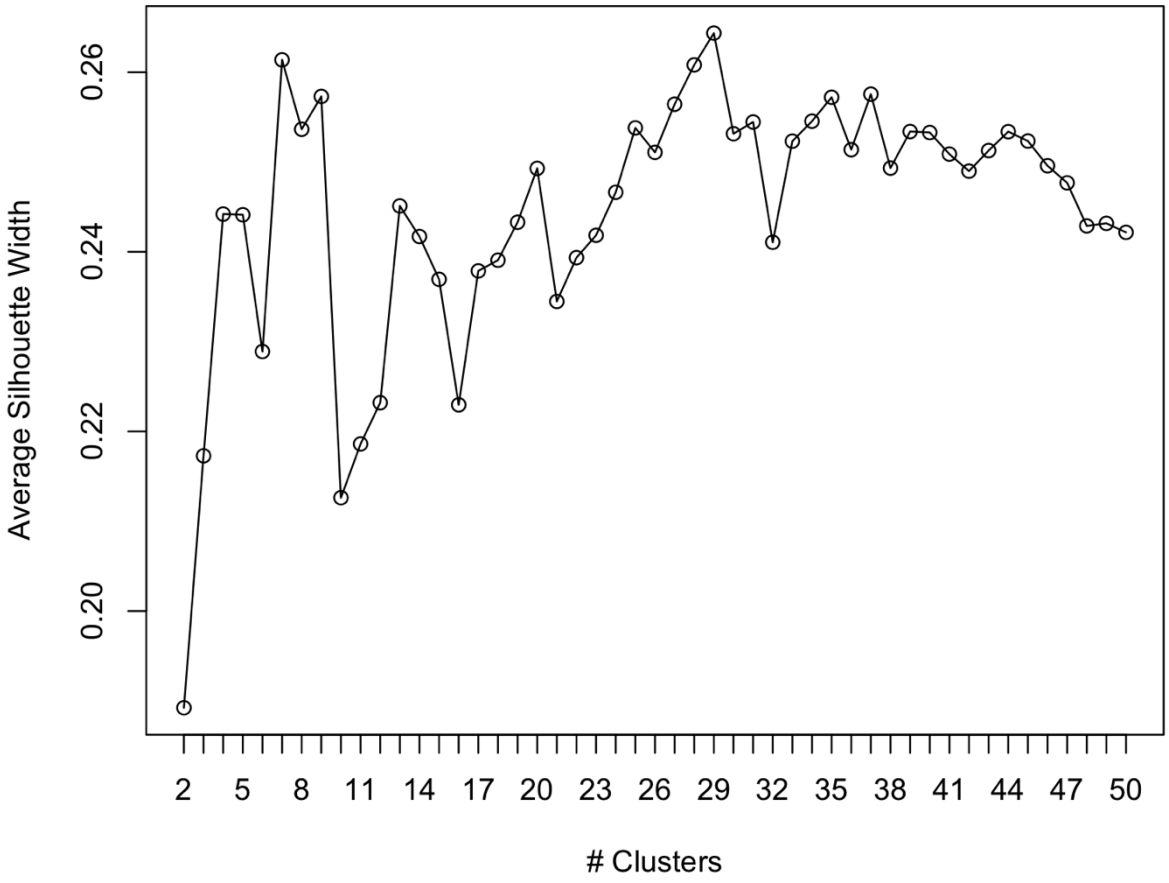

Figure 7 – AVERAGE SILHOUETTE WIDTH OF CLUSTERS IN ROME'S DATA

The average silhouette width is a measure of clustering validity, its value can range from -1 to 1 and a high value of average silhouette width means that there is a good similarity between objects belonging to the same cluster and a substantial dissimilarity between objects of different clusters.
The highest value of average silhouette width is reached for k = 29, hence the characteristics of these 29 clusters have been investigated.

3.3.3 Characteristics of the Clusters

In the following table the features that characterize each cluster are illustrated. For each variable, the bold font indicates a characteristic that is common to all the observations of the cluster (or at most some missing values are present for that variable). On the contrary, if a feature is present in the vast majority of the elements of a group, but not in all of them, the italic font is used. A slash means that the variable is not relevant for the characterization of that cluster.

	# of data objects	gender	age	weather conditions	road surface conditions	daily rainfall class	hour	light conditions	day of week	junction type
1.	273	**M**	/	*Fine*	**Dry or Other**	**0**	**15-23**	**0-1**	/	**Other or Roundabout**
2.	162	**F**	*19-22*	*Fine*	*Dry*	**0**	**17-23**	**0-1**	/	*Other or Roundabout*
3.	171	**M**	/	*Fine*	*Dry*	*0*	*04-20*	*2*	*Mon-Wed*	**Other or Roundabout**
4.	243	**M**	*18-20*	**Fine or Fog**	**Dry or Other**	**0**	**05-20**	**2**	**Mon-Fri**	*No Junction*
5.	205	**M**	*21-22*	**Fine**	*Dry*	**0**	*00-06 or 17-23*	**0-1**	**Mon-Fri**	**No Junction**
6.	81	**M**	/	**Rain**	**Wet**	*1-4*	/	/	/	*Other or Traffic Signal*

7.	80	/	/	*Fine*	**Wet**	**0-2**	*00-06 or 16-23*	*0-1*	/	*Traffic Signal*
8.	112	**M**	/	**Rain**	*Wet*	*1-4*	*04-20*	*2*	/	*No Junction*
9.	93	/	/	**Fine**	**Wet**	**0-3**	/	/	/	*Other*
10.	157	*M*	/	*Rain*	**Wet**	*1-4*	*00-06 or 16-23*	*0-1*	/	*No Junction*
11.	237	**F**	*18-20*	*Fine*	*Dry*	**0**	*05-21*	*2*	/	**No Junction**
12.	138	*M*	/	**Fine or Fog**	**Wet**	**0-3**	*00-07 or 18-23*	*0-1*	/	*No Junction*
13.	110	**F**	/	*Fine*	**Wet**	**0-3**	/	/	/	*No Junction*
14.	113	**F**	/	*Fine*	*Dry*	**0**	*05-21*	*2*	/	**Traffic Signal or Roundabout**
15.	195	**M**	**18-20**	*Fine*	*Dry or Other*	**0**	**00-06 or 17-23**	*0-1*	*Mon-Fri*	*No Junction*
16.	222	**M**	/	*Fine*	*Dry*	**0**	**05-21**	**2**	*Sat-Sun*	**No Junction or Roundabout**
17.	200	**M**	/	**Fine**	*Dry*	**0**	**00-06**	**0-1**	/	**Other or Roundabout**
18.	321	**F**	/	**Fine**	*Dry or Other*	*0*	*00-06 or 16-23*	*0-1*	/	*No Junction*
19.	258	/	**21-22**	**Fine**	*Dry*	**0**	*05-21*	*2*	*Mon-Fri*	*No Junction*
20.	101	**M**	/	**Fine**	**Wet**	/	*04-21*	*2*	/	*No Junction*

21.	120	**F**	/	**Rain or Fog**	**Wet or Other**	*1-4*	/	/	/	*No Junction*
22.	268	**M**	/	*Fine*	**Dry**	**0**	*00-07 or 16-23*	*0-1*	/	*Traffic Signal*
23.	251	**M**	/	**Fine or Fog**	*Dry*	**0**	*05-21*	**2**	*Thu-Sun*	*Other*
24.	254	**M**	/	*Fine*	*Dry*	**0**	**00-07 or 17-23**	*0-1*	**Sat-Sun**	*No Junction*
25.	76	**F**	/	*Rain*	**Wet**	*1-4*	/	/	/	*Other or Traffic Signal*
26.	320	**F**	/	**Fine**	*Dry*	*0*	*04-20*	**2**	/	*Other*
27.	95	**F**	/	**Fine**	*Dry*	*0*	*00-06 or 18-23*	*0-1*	/	*Traffic Signal*
28.	111	**F**	/	**Fine**	*Dry*	*0*	*00-05*	*1*	/	*Other*
29.	116	**M**	/	**Fine**	**Dry**	**0**	*05-21*	**2**	/	*Traffic Signal*

Table 7 – CLUSTERS IN ROME'S DATA

These clusters represent the most recurring types of car accidents in which young drivers have been involved in Rome. Their characteristics make them mutually exclusive. Observations having missing values are included in the clusters and, if the missing values are relating to variables not relevant for the characterization of the group, these data objects are even useful to determine the features of the cluster in which they are placed.

As a first consideration, the night hours in the weekend, which is the time period previously identified as the most dangerous, are a specific feature of the cluster 24, which, in addition, comprises only male drivers and mostly refers to accidents occurred in a place different from a junction. This could mean that young male drivers are more likely to take risks than females during these hours (high speed, drugs, alcohol) or, maybe, that young females don't drive a lot at night in the weekend.
An other interesting aspect is that, except some rare cases, the age doesn't represent a discrimination criterion, therefore, it can be argued that there isn't a big variation between the behaviours of the drivers in the age band of 18-22 years old.
The information about the total daily rainfall, instead, helps to distinguish between accidents with various danger levels. In fact, wet road surface conditions with daily rainfall class 0, which means that the weather station did not register any rainfall, are present only in clusters having fine weather conditions and, therefore, a lower level of danger.
Regarding the most dangerous situation about weather and road surface conditions, male drivers are once again the most involved category, as they characterize three of the five considered clusters and, by summing up the number of observations inside these clusters, male drivers are about three-fifths of the total number.

Fine weather conditions with dry road surface conditions are the most common combination of values about these variables, but this is probably due to the fact that, in Rome, the number of days in which it doesn't rain in a year is much greater than the number of rainy days.
Finally, some values aren't considered in the clusters, as they aren't so frequent in the dataset. For instance, the value "*Ice*" about the road surface conditions is very rare and it can't be used to characterize any group.

As a final remark, it's worth specifying that the number of observations indicated in the second column of the table above is not the actual number of observations having those characteristics. In fact, inside each cluster, there could be outliers (elements very different from the others), data objects wrongly assigned to that cluster and elements with a few different features. For this reason, after having identified the characteristics of each group, the elements not having entirely the features of any cluster have been counted and they are 178. Being 5083 the total number of observations, they only represent a small portion of data (3.50%) and some of them (11) also contain missing values.

4. METHODOLOGY

The aim of this work is to propose a business plan methodology modified to take advantage of the huge potentiality of open data and data science techniques. The four elements that a business plan can never lack are summarised in the following graph. Other parts, such as the description of the company and the development team, are not useful to deal with in this case.

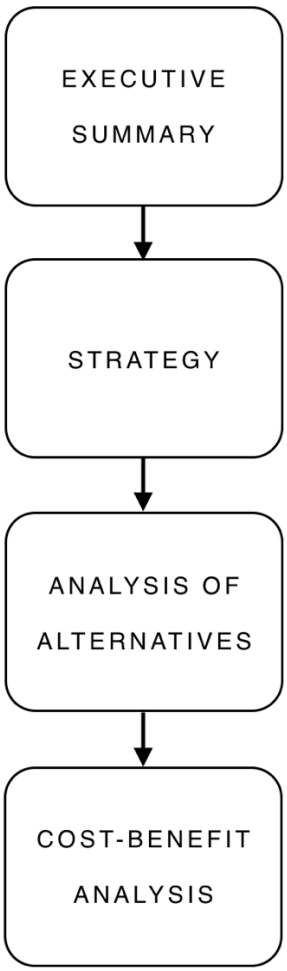

Figure 8 – TRADITIONAL BUSINESS PLAN STRUCTURE

The executive summary contains a brief recap of the current situation, an explanation of the remarkable outcomes that will be obtained if that action will be applied and the description of the importance of such a work, highlighting its key strengths as compared to other possible interventions. This is the first part of a business plan, but it is usually composed at last, as it summarizes the most relevant information included in the other sections of the business plan.
The strategy is about the decisions taken to realise the entire work and it contains the detailed procedures that will be followed to get the results. Moreover, here, the strategy also refers to the market description, the quantification of the market potential and the selection of case studies useful to support what is proposed.
The analysis of alternatives focuses on the comparison between this type of work and other types of actions that could be performed in that sector of interest.
Finally, the cost-benefit analysis shows the economic advantages that could be attributed to the work.

This structure has been completely revised giving a determining role to data science. In the modified business plan structure a bottom-up approach is followed, because all the sections are obtained starting from data. Data science leads the realisation of the entire work and what is achieved is a more precise and persuasive result.
The following graph describes the elements of which the innovative business plan structure is composed. The executive summary is still present as the first part of the business plan and the last part that is advisable to write, as it tries to convince the reader of the strength of the work, briefly explaining what is obtained in the other components.

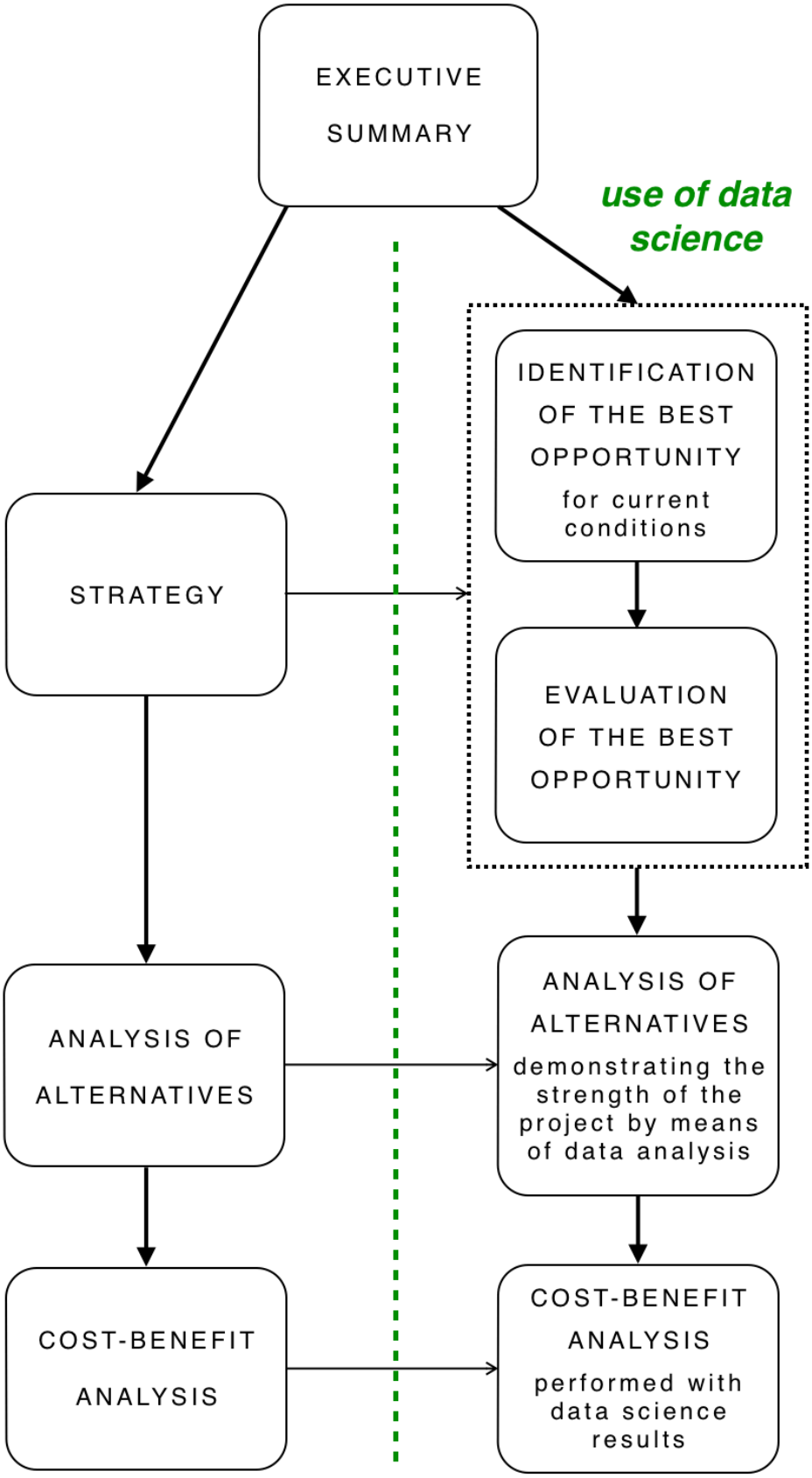

Figure 9 – MODIFIED BUSINESS PLAN STRUCTURE

The strategy has been divided into two phases: the identification of the best opportunity and its evaluation.
The first phase starts from a deep data analysis that identifies the features of the examined phenomenon or particular characteristics of the population on which the study is focused. At this point, some specific aspects stand out, because they are identified by particularly interesting trends of a variable or they are not expected to be present or they correspond to a total lack of a characteristic. These aspects coincide with issues or opportunities that have to be exploited. Once having identified the most relevant issue or the best opportunity, still using data science techniques, a precise characterization of it can be achieved.
In the second phase, similar actions that were taken before or similar case studies are searched. Then, analysing the data about that case, the effects of that action are investigated. Here, thanks to some elements such as a big change of the trend relative to a characteristic, the effectiveness of that action is proven. This positive evaluation allows to have a precise idea of the potential consequences that could be obtained in the examined case.

The analysis of alternatives focuses on the effectiveness of the proposed action, compared to all the other interventions that could be taken to reach the objective. It is based on the same data analysed in the "Identification of the best opportunity" phase and it aims at demonstrating that, according to the features revealed in the data, all the alternatives have to be rejected, because data clearly indicates this action as the most appropriate one.

The cost-benefit analysis is performed by means of the results obtained in the evaluation phase. What is contained in the modified cost-benefit analysis is by far more precise, as analyses achieved by means of data science techniques arrive at a level of accuracy that traditional cost-benefit analyses, typically based only on general statistics and experience, can't reach.

Following this innovative business plan structure based on data science, a business plan concerning the introduction of a road safety action in Rome, aimed at strongly reducing the number of young car drivers involved in personal injury accidents, has been realised and its four main components are presented in the following sections.

4.1 Executive Summary

Open Data mean innovation.
They can lead to social and economic benefits and they can be used to generate income, also by governments. In this work, the enormous potentiality of open data is exploited and the results clearly show a great social and economic opportunity for Rome.

The area of interest is defined by the actions taken to decrease the number of road casualties and, in particular, the focus is on the reduction of the number of young car drivers involved in personal injury accidents. These type of crashes causes a huge amount of social costs and it represents a serious issue especially in Rome. This work precisely evaluates the effects of an enhanced driving course, called Pass Plus Extra, that was introduced in Staffordshire, a county of the United Kingdom, in April 2006. Then, an analysis of the impact that a similar course, aimed to increase the safety awareness and the ability of young drivers, would have had in Rome in the last four years is provided. The results are astonishing: benefits would have been reached by the local authority, the young drivers, the driving instructors and the insurance companies. In particular, more than 7 million euros would have been saved about social costs, while just about € 265,000 would have been the total amount paid by the local authority to support the scheme.

In Italy road safety interventions have always been decided and evaluated by making use of generic statistics. There's no trace of deep open data analyses about this topic on the web. On the contrary, in the United Kingdom previous studies have taken advantage of open data concerning road crashes, but no open data analysis has demonstrated the effectiveness of Pass Plus Extra course up to now. This work, thus, proposes a completely innovative approach for facing the issue of young road casualties by means of data science techniques and it identifies a big opportunity of which Rome can take advantage.

4.2 Strategy

After having performed a deep analysis on data about accidents occurred in Rome (shown in Chapter 3 in detail), the attention has been focused on the serious issue about young car drivers involved in personal injury accidents. Then, having identified the road safety awareness as the most relevant aspect on which rely for solving or at least mitigating the problem, a research on previous similar actions taken by other governments have been performed and an interesting case has been found. This refers to Staffordshire, a county of the United Kingdom, where a strong decrease in the number of crashes involving young car drivers has coincided with the introduction of an enhanced driving course called Pass Plus Extra.

For this reason, a precise evaluation of this course has been realised and an estimate of the impact that the introduction of a similar course would have had in Rome in the last four years have been calculated.

The following sections contains a detailed description of what have been done in these phases.

4.2.1 Identification of the Best Opportunity

The Chapter 3 of this work contains a detailed description of the data analysis that has been performed on Rome's data about road accidents. To summarize the most relevant aspects, it has been found that Rome, compared to London, presents an extremely high number of young car drivers involved in serious crashes. In particular, an abnormal peak of accidents is reached at night during the weekend and, after having performed a clustering analysis, it has been demonstrated that this type of crashes is mainly related to male drivers and the place of the accident is mainly far from a junction. Moreover, it seems to be mostly related to fine weather conditions and dry road surface conditions, but this is probably due to the fact that, in a year, the days without precipitation are the large majority in Rome.

The obtained results clearly indicate that the most relevant factor that can be considered the cause of these accidents is the risky driving. Young drivers are not conscious of the consequences of drunk driving, excessive speed and fatigued driving. Thus, the most appropriate action to take in this case is about road safety awareness.

4.2.2 Evaluation of the Best Opportunity: the Case Study about Pass Plus Extra

Considering the United Kingdom, it has been found that, during the last few years, great effort has been put into reducing the number of young road casualties. Among the road safety actions that have been taken, one in particular is worth being mentioned, as it is an optional course introduced by the Government's Driving Standards Agency (DSA) in 1995 and it is still available at the moment. The course is called "*Pass Plus*" and it consists of 6 modules, regarding the following aspects of driving:

- in town
- in all weathers
- on rural roads
- at night
- on dual carriageways
- on motorways

All these modules should be covered as practical lessons, even though sometimes it is not allowed due to local conditions and, in these cases, they are covered as theory only. The course takes at least six hours and there isn't any test to pass at the end, but, when the the driver has reached the required standard, a certificate is awarded. Some insurance companies also offer a cheaper car insurance, which is usually very expensive for young drivers, to people having the Pass Plus certificate. [5] The cost of the course varies from £ 120 to £ 180 depending on where the driver lives, the instructor and how long the training takes. Moreover, in order to encourage more young drivers to take the Pass Plus course, several local authorities provide subsidies that can reduce the cost.
Among them, there are three very interesting cases, in which, in order to obtain the subsidy, a workshop aimed to improve road safety awareness has to be attended. These cases are represented by the regions of Wales, Staffordshire and Cumbria and the enhanced version of Pass Plus are respectively called "*Pass Plus Cymru*", "*Pass Plus Extra*" and "*Pass Plus +*".

Pass Plus Cymru was launched in Wales in 2006. The course comprises two parts: a practical part, which consists of the six practical modules of the original Pass Plus, and a workshop discussion, which takes two and a half hours, about driver's attitude, alcohol and drugs. Thanks to a Welsh Government subsidy, the young driver has to pay just £ 20 for the complete course and, since 2006, the proportion of young drivers participating in Pass Plus Cymru has increased constantly. [14]
In Staffordshire a similar scheme has been provided since April 2006 and it is called "*Pass Plus Extra*". A two-hour workshop, during which some video clips

are shown, represents the "extra" part compared to the original Pass Plus course. The client has to pay only £ 60 and the rest (£ 70) is funded by the local authority. Staffordshire has put great effort into publicizing the course, they have made use of local press, television, radio and fliers and they have received much attention from other local authorities and road safety units in the United Kingdom. [19]

In Cumbria, Pass Plus + was available from 2007 to March 2016 and the workshop was provided by the Cumbria Fire and Rescue Service. [2] [10]

The trends of the number of involved young car drivers, compared to the trends of the number of involved older drivers, have been investigated with reference to these three regions. What has been obtained is shown in the graph below.

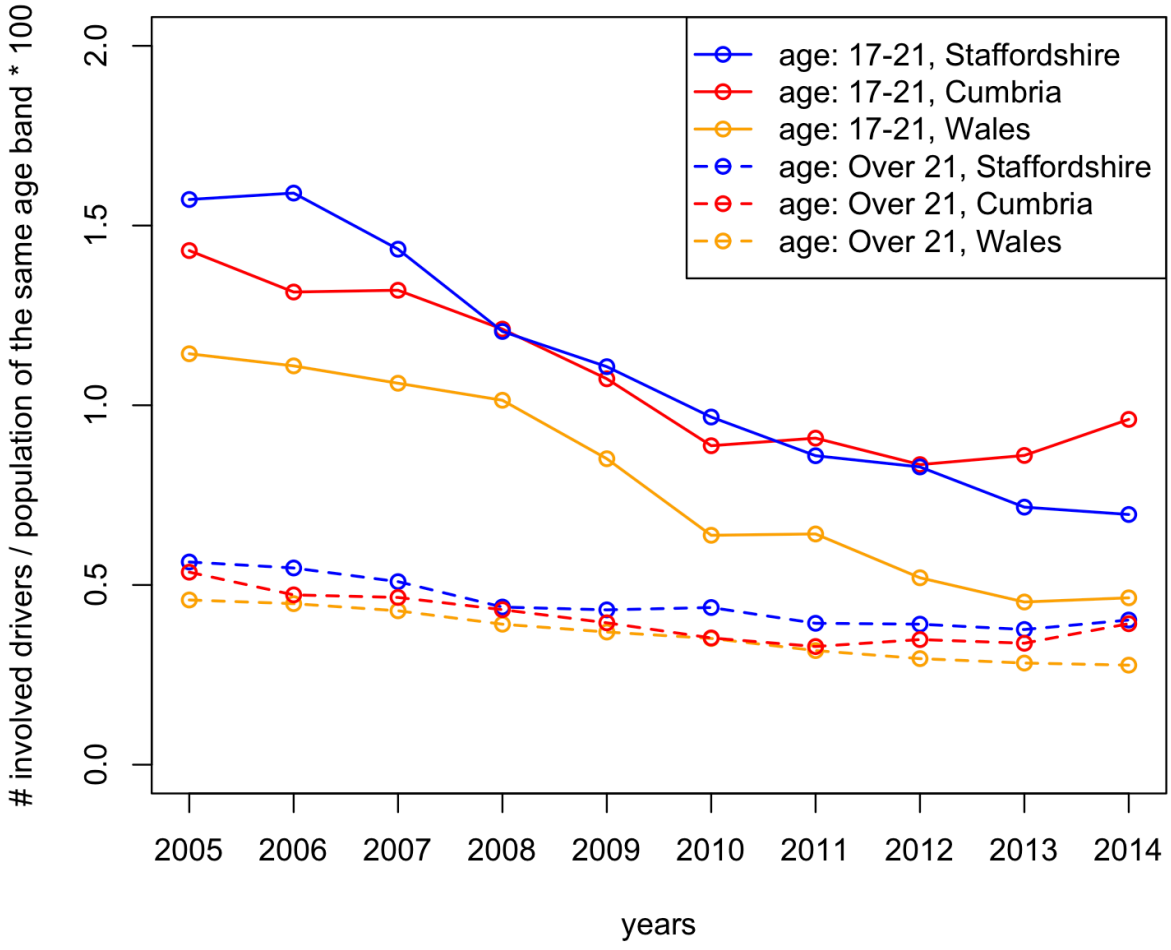

Figure 10 – NUMBER OF INVOLVED CAR DRIVERS BY AGE BAND AND REGION

The graph clearly shows that, in the years following the introduction of the enhanced course, a strong reduction of the number of the involved young car drivers has been achieved. In two cases this number has even reached a value very close to the number of involved older car drivers and, among these cases, Staffordshire has been chosen to be further analysed, because the positive effects seem to be obtained already a year after the introduction of Pass Plus Extra.

A Pass Plus Extra evaluation was commissioned by the Staffordshire County Council in 2010 and a report was produced following that study. The report contains information about the applicants for Pass Plus Extra for the time period between April 2006 and March 2010, therefore, data about young drivers involved in personal injury accidents during that period have been analysed in this work.

As a first general result, it has been found that, actually, after the introduction of the enhanced course, the number of young drivers involved has decreased more in Staffordshire than in all other counties of England. Stoke-on-Trent, the district of Telford and Wrekin and the area covered by the Derbyshire County Council haven't been taken into account for the comparison, as they joined the scheme for a part of the considered period.

The following table shows the trends of the number of the involved drivers having age between 17 and 21 years, highlighting how it has changed since April 2006, when Pass Plus Extra was launched.

	STAFFORDSHIRE		**REST OF ENGLAND**	
	# Involved Young Drivers	*Comparison with 2005/06 value*	*# Involved Young Drivers*	*Comparison with 2005/06 value*
2005/06	790		29,632	
2006/07	793	+0%	29,504	-0%
2007/08	714	-10%	27,887	-6%
2008/09	586	-26%	24,963	-16%
2009/10	594	-25%	23,612	-20%

Table 8 – COMPARISON STAFFORDSHIRE/REST OF ENGLAND ON THE NUMBER OF INVOLVED YOUNG DRIVERS

Considering the year between April 2005 and March 2006 as the reference period, it seems to be no difference regarding the first year after the introduction of Pass Plus Extra, as in both cases there isn't any substantial variation. On the contrary, starting from April 2007, Staffordshire has registered a greater reduction of young car drivers involved in personal injury accidents (-4%). The trend is confirmed also in the two following years, when, compared with the rest of England, the reduction is even more consistent (-10% and -5% in 2008/09 and 2009/10 respectively).

Moreover, the value relating to the first year of the course can be explained by the fact that a low number of drivers completed the course process in that year and this is shown in the following table, which contains data taken from the Pass Plus Extra evaluation report. To be more precise, these data refer to all territories covered by Pass Plus Extra in those years (Staffordshire, Stoke-on-Trent, Derbyshire and Telford), but, actually, 77% of the applicants have been registered as resident in Staffordshire. [19]

	# Clients Completing the Process
2006/07	177
2007/08	360
2008/09	367
2009/10	464

Table 9 – NUMBER OF CLIENTS COMPLETING THE PASS PLUS EXTRA PROCESS BY YEAR

These first results seem to suggest that Pass Plus Extra has actually had a role in reducing the number of involved young drivers.
Even though it isn't possible to prove the impact of Pass Plus Extra for sure, because it isn't specified if a driver has participated to the enhanced course in the collision data, a further analysis on these data is still feasible. In the following paragraphs a clustering analysis is performed on Staffordshire's data regarding the time period between April 2005 and April 2006 (one year before Pass Plus Extra), in order to find the clusters on which Pass Plus Extra has probably had a greater effect in the following years.

4.2.3 Evaluation of the Best Opportunity: Staffordshire's Data Pre-Processing

As previously done with Rome's data, the first operation to perform is data pre-processing. The data to take into account must have *Age_of_Driver* between 17 and 21 and refer to car accidents (value "9" of the variable *Vehicle_Type*) occurred in Staffordshire (value "E10000028" of the attribute *Local_Authority_.Highway.*) between April 2005 and April 2006. The attributes selected are:

Day_of_Week, *Time*, *Junction_Detail*, *Junction_Control*, *Light_Conditions*, *Weather_Conditions*, *Road_Surface_Conditions*, *Sex_of_Driver* and *Age_of_Driver*.
790 observations are obtained, but 5 of them contains the value "7" for the attribute *Light_Conditions*, that means "Darkness: street lighting unknown" and it can't be properly inserted into the range of values between "Daylight" and "Darkness: no street lighting". For this reason, they have been removed.

In addition, in order to make these data comparable with Rome's data, some aggregations and transformations have been performed on single values of the attributes and all these operations are described in the tables below.

Weather_Conditions

Before aggregation	After aggregation
"*1*" (Fine without high winds)	"*Fine*"
"*2*" (Raining without high winds) "*5*" (Raining with high winds)	"*Rain*"
"*7*" (Fog or Mist)	"*Fog*"
"*3*" (Snowing without high winds) "*6*" (Snowing with high winds)	"*Snow*"
"*4*" (Fine with high winds) "*8*" (Other)	"*Other*"

Table 10 – AGGREGATIONS ON WEATHER CONDITIONS VALUES ABOUT STAFFORDSHIRE'S DATA

Road_Surface_Conditions

Before aggregation	After aggregation
"*1*" (Dry)	"*Dry*"
"*2*" (Wet/Damp) "*5*" (Flood)	"*Wet*"
"*4*" (Frost/Ice)	"*Ice*"
"*3*" (Snow)	"*Snow*"

Table 11 – AGGREGATIONS ON ROAD SURFACE CONDITIONS VALUES ABOUT STAFFORDSHIRE'S DATA

Light_Conditions

Before aggregation	After aggregation
"*1*" (Daylight)	*2*
"*4*" (Darkness: street lights present and lit)	*1*
"*2*" (Darkness: street lights present but unlit) "*3*" (Darkness: no street lighting)	*0*

Table 12 – AGGREGATIONS ON LIGHT CONDITIONS VALUES ABOUT STAFFORDSHIRE'S DATA

Junction_Detail [JD] ***Junction_Control [JC]***	***Junction_Type***
Before aggregation	After aggregation
[JD] "*1*" (Roundabout) [JD] "*2*" (Mini roundabout)	"*Roundabout*"
[JD] "0" (Not at or within 20 metres of junction) [JD] "8" (Using private drive or entrance)	"*No Junction*"
[JC] "*1*" (Authorised person)	"*Junction with Authorised Person*"
[JC] "*2*" (Automatic traffic signal)	"*Junction with Traffic Signal*"
[JC] "*3*" (Stop sign) [JC] "*4*" (Give way or uncontrolled)	"*Other Junction*"

Table 13 – AGGREGATIONS ON JUNCTION TYPE VALUES ABOUT STAFFORDSHIRE'S DATA

Values such as "Oil or diesel" or "Mud", which are present in Rome's data, in the UK dataset are placed in an other variable in combination with different values of the *Road_Surface_Conditions* attribute. For this reason, they aren't present in this case.
Regarding the junction type, the values 1, 2, 3 and 4 of *Junction_Control* have been considered only in the case in which *Junction_Detail* contains values not referring to roundabouts or places far from a junction.

Finally, for the attribute *Time* only the value of hours has been taken into account and, concerning the attribute *Day_of_Week*, the original classification, which considered Sunday as the first day and Saturday as the last day of the week, has been transformed to make Sunday much closer to Saturday (Sunday

now corresponds to day 7 and Saturday to day 6), in the sense that accidents occurred in the weekend have similar characteristics, especially for the hour in which they take place.

After the phase of data pre-processing, 785 observations have been obtained and they contain eight variables. Four of them are ordinal and they are: *Day_of_Week* (from *1* [Monday] to *7* [Sunday]), *Hour* (from "*00*" to "*23*"), *Light_Conditions* (from *0* to *2*) and *Age_of_Driver* (from *17* to *21*). Three are categorical and they are: *Weather_Conditions* (possible values: "*Fine*", "*Rain*", "*Fog*", "*Snow*", "*Other*"), *Road_Surface_Conditions* ("*Dry*", "*Wet*", "*Ice*", "*Snow*") and *Junction_Type* ("*Roundabout*", "*Junction with Traffic Signal*", "*Junction with Authorised Person*", "*Other Junction*", "*No Junction*") and one, *Sex_of_Driver*, is logical.

4.2.4 Evaluation of the Best Opportunity: Clusters in Staffordshire's Data

Proceeding with the cluster analysis, the best number of cluster has to be chosen. Again, like before with Rome's data, the *daisy* function of the R library *cluster* has been used, the dissimilarity matrix between the observations has been computed making use of the *Gower's coefficient* and, for k between 2 and 50, data have been divided into k clusters applying the PAM algorithm.
The average silhouette widths of the obtained clusters are shown in following graph.

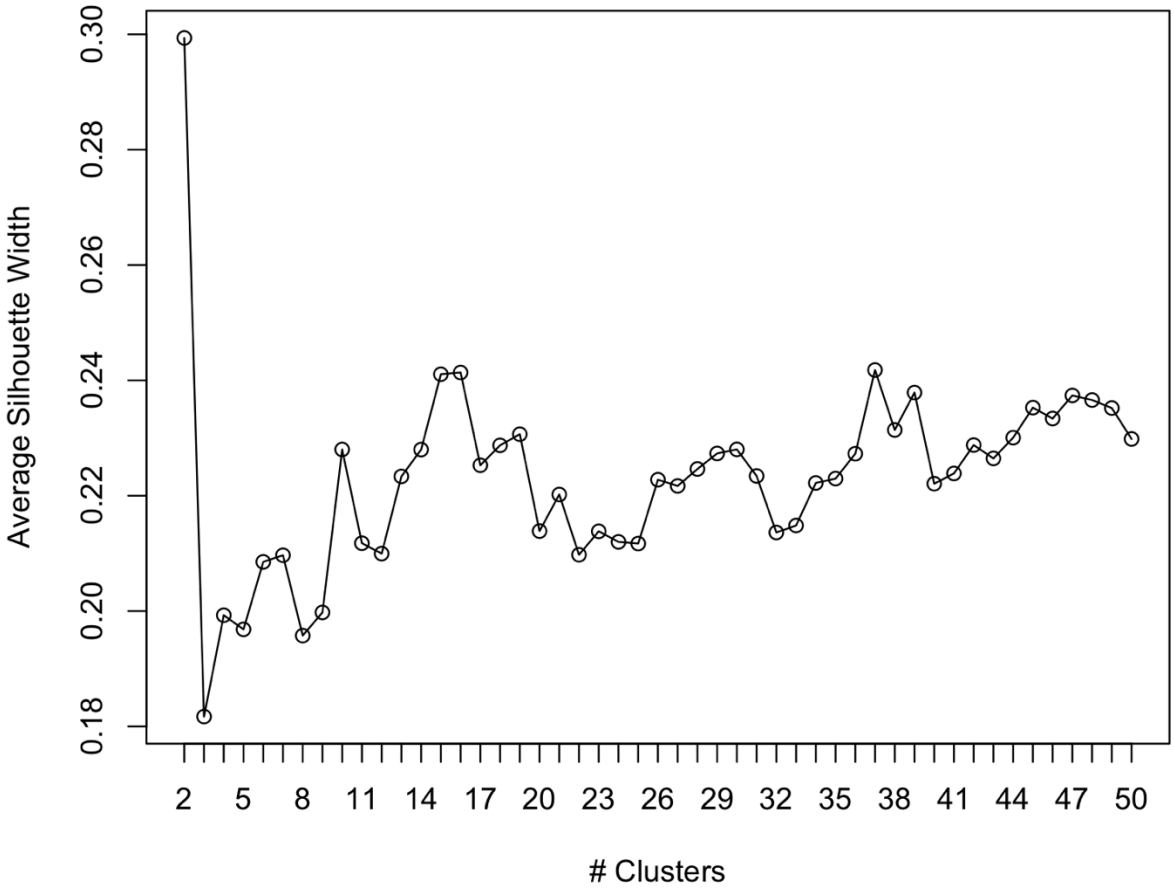

Figure 11 – AVERAGE SILHOUETTE WIDTH OF CLUSTERS IN STAFFORDSHIRE'S DATA

The highest value is reached in the case of two clusters, but this can't be accepted as the best value, as two clusters don't give enough information. Looking at other peaks, among 16 and 37, which have very similar values, the first one has been chosen as the best number of clusters, because it is sufficient to identify the main characteristics of the different groups. A k value equal to 37, instead, would be too much high and the resulting groups would contain too few observations to determine valid features.

After having determined the best number of clusters, their features have been analysed. The table below describes the characteristics of the 16 clusters. The bold font represents a feature present in all the observations of the group (or at

most there are some missing values for that variable) and the italic font indicates a feature that is in common for the vast majority of the elements of a cluster, but it isn't present in all of them. The slash is used if the variable is not relevant to characterize that cluster.

	# of data objects	gender	age	weather conditions	road surface conditions	hour	light conditions	day of week	junction type
1.	34	*M*	/	**Fine**	**Dry**	**00-05 or 16-23**	**0-1**	/	**No Junction**
2.	54	*F*	/	/	*Wet*	/	*1-2*	/	*No Junction*
3.	53	**M**	/	/	**Wet**	*06-20*	*2*	/	*Other*
4.	60	**F**	/	*Fine*	*Dry*	/	/	*Fri-Sun*	/
5.	112	**M**	/	**Fine**	*Dry*	*04-21*	*2*	/	**No Junction or Traffic Signal**
6.	67	**M**	/	*Fine*	**Dry**	**04-21**	*2*	/	*Other*
7.	38	**M**	/	*Rain or Other*	**Wet**	*07-20*	*2*	/	**No Junction or Traffic Signal or Roundabout**
8.	77	*F*	/	*Fine*	*Dry or Frost*	/	**1-2**	*Mon-Thu*	/

9.	52	**M**	/	*Fine*	**Wet or Frost**	**09-23**	/	/	*No Junction*
10.	51	*M*	/	*Fine*	*Dry*	/	**1-2**	/	*Roundabout*
11.	48	*M*	/	*Fine*	**Dry**	**00-02 or 16-23**	*0-1*	/	**Other or Traffic Signal**
12.	22	**F**	/	/	*Wet or Frost*	*16-23*	*0*	/	*No Junction*
13.	26	**F**	/	/	**Wet**	/	/	/	*Other or Roundabout*
14.	37	*M*	/	*Fine or Fog*	*Wet or Frost*	*00-08*	/	/	*No Junction*
15.	33	*M*	/	/	*Wet*	*16-23*	**0-1**	/	**Other or Traffic Signal or Roundabout**
16.	21	**M**	/	*Rain or Other*	**Wet or Frost**	*16-23*	*0-1*	/	**No Junction**

Table 14 – CLUSTERS IN STAFFORDSHIRE'S DATA [Apr '05 / Mar '06]

The age of the driver isn't useful to describe any cluster, hence, it seems to be no substantial difference among the behaviours of young drivers in the age band of 17-21 years old.

As it has been found analysing Rome's data, in Staffordshire young male drivers represent the most involved category too. The male gender is present in 11 clus-

ters out of 16 and, by summing up the number of observations about male drivers, the result indicates that young male drivers are about twice as much as the number of young female drivers.

Looking at the features that probably describe an accident caused by a risky behaviour, such as "No Junction" together with night-time hours, the clusters 1, 12, 14 and 16 have to be taken into consideration and, again, what can be argued is that young male drivers are surely the most involved category in these groups.

The following paragraph illustrates the effects of the introduction of Pass Plus Extra on these clusters, highlighting the groups that have mostly changed their trend compared to the trend of the same type of accidents regarding the rest of England.

4.2.5 Evaluation of the Best Opportunity: Effects of Pass Plus Extra on Single Clusters

Five clusters in particular seem to have been affected by the introduction of Pass Plus Extra. The five tables below describe their trend in comparison with the trend of observations having the same features but regarding the rest of England. As previously specified, the rest of England doesn't include Stoke-on-Trent, the district of Telford and Wrekin and the area covered by the Derbyshire County Council, as, for a part of the considered years, they joined Pass Plus Extra scheme.

	STAFFORDSHIRE		**REST OF ENGLAND**	
	# Involved Young Car Drivers	*Comparison with 2005/06 value*	*# Involved Young Car Drivers*	*Comparison with 2005/06 value*
2005/06	106		3,807	
2006/07	104	-2%	3,884	+2%
2007/08	82	-23%	3,588	-6%
2008/09	61	-42%	3,039	-10%
2009/10	57	-46%	2,893	-14%

Table 14 – COMPARISON STAFFORDSHIRE/REST OF ENGLAND ON THE TREND OF THE CLUSTER 5

	STAFFORDSHIRE		**REST OF ENGLAND**	
	# Involved Young Car Drivers	*Comparison with 2005/06 value*	*# Involved Young Car Drivers*	*Comparison with 2005/06 value*
2005/06	39		948	
2006/07	25	-36%	976	+3%
2007/08	29	-26%	929	-2%
2008/09	24	-38%	783	-17%
2009/10	26	-33%	843	-11%

Table 15 – COMPARISON STAFFORDSHIRE/REST OF ENGLAND ON THE TREND OF THE CLUSTER 10

	STAFFORDSHIRE		REST OF ENGLAND	
	# Involved Young Car Drivers	*Comparison with 2005/06 value*	*# Involved Young Car Drivers*	*Comparison with 2005/06 value*
2005/06	15		276	
2006/07	13	-13%	324	+17%
2007/08	5	-67%	348	+26%
2008/09	17	+13%	263	-5%
2009/10	6	-60%	254	-8%

Table 16 – COMPARISON STAFFORDSHIRE/REST OF ENGLAND ON THE TREND OF THE CLUSTER 12

	STAFFORDSHIRE		REST OF ENGLAND	
	# Involved Young Car Drivers	*Comparison with 2005/06 value*	*# Involved Young Car Drivers*	*Comparison with 2005/06 value*
2005/06	26		523	
2006/07	17	-35%	513	-2%
2007/08	14	-46%	467	-11%
2008/09	11	-58%	449	-14%
2009/10	16	-38%	375	-28%

Table 17 - COMPARISON STAFFORDSHIRE/REST OF ENGLAND ON THE TREND OF THE CLUSTER 14

	STAFFORDSHIRE		**REST OF ENGLAND**	
	# Involved Young Car Drivers	*Comparison with 2005/06 value*	*# Involved Young Car Drivers*	*Comparison with 2005/06 value*
2005/06	42		1,314	
2006/07	32	-24%	1,450	+10%
2007/08	28	-33%	1,200	-9%
2008/09	38	-10%	1,034	-21%
2009/10	22	-48%	955	-27%

Table 18 – COMPARISON STAFFORDSHIRE/REST OF ENGLAND ON THE TREND OF THE CLUSTER 15

Even if in two cases (clusters 12 and 15) there is an unexpectedly high value relating to 2008/09, these are the five clusters that most have a different trend compared to the rest of England. The trend of the clusters 5 and 15 is particularly interesting, as they almost decrease by half their number of observations in the fourth year after the introduction of the enhanced course.
Four of the five most affected clusters regard young male drivers and, in addition, in the Pass Plus Extra evaluation report it is specified that "*The number of male clients applying for the scheme has been consistently higher than female applicants*". [19] This seems to be a further evidence of the fact that the enhanced course has really had a crucial role in the decrease of accidents involving young car drivers.

Two of the four groups previously identified as the most related ones to a risky behaviour are among the five most affected clusters. The other two, respectively the cluster 1 and the cluster 16, presents a trend very similar to the rest of England, as the number of observations is slightly decreased in both cases.

For further information, an example of cluster that haven't clearly modified its trend compared to the rest of England is shown. It is represented by the cluster 13, which is characterized by young female drivers involved in car accidents in a junction without traffic signal or in a roundabout, with wet road surface conditions.

	STAFFORDSHIRE		**REST OF ENGLAND**	
	# Involved Young Car Drivers	*Comparison with 2005/06 value*	*# Involved Young Car Drivers*	*Comparison with 2005/06 value*
2005/06	36		1,440	
2006/07	35	-3%	1,564	+9%
2007/08	35	-3%	1,447	+0%
2008/09	36	-0%	1,350	-6%
2009/10	33	-8%	1,311	-9%

Table 19 – COMPARISON STAFFORDSHIRE/REST OF ENGLAND ON THE TREND OF THE CLUSTER 13

4.3 Analysis of Alternatives

The analysis on Rome's data about road accidents and the comparison with London has highlighted a huge difference concerning the number of young car drivers involved in personal injury accidents (Figure 3). This is the issue on which it is worth focusing and, thinking about the cause of the problem and the possible remedies, several hypotheses could be made. The following table summarizes

them and explains the reasons why all of them, except the last one, has to be rejected.

Hypothesis on the main cause	*Why it has to be rejected / accepted*
Poor quality driving school / Poor driving skills	REJECTED because the highest peak about young car drivers involved in serious crashes is reached at night during the weekend (Figure 5), therefore in a moment of the day in which the traffic intensity is at its minimum and there aren't many difficult decisions to take while driving.
Bad designed junction points	REJECTED because the value "No Junction", indicating a place far from a junction, is present in about a half of the clusters obtained from Rome's data about young car drivers (Table 7). Moreover, this value characterizes the cluster #24, which is exactly the group that refers to young car drivers involved in personal injury accidents at night during the weekend.
Adverse weather conditions / Bad road surface conditions	REJECTED because the majority of the observations and the majority of the clusters obtained from Rome's data (Table 7) are characterized by fine weather conditions and dry road surface conditions. The cluster #24 (crashes at night during the weekend) exhibits these characteristics too.
Poor light conditions	REJECTED because there's no cluster clearly characterized by the value 0 (insufficient or no lighting) of the attribute about light conditions (Table

	7). All the clusters indicating accidents occurred at night exhibit both values 0 and 1 (sufficient) about light conditions.
Risky driving	ACCEPTED because the clustering analysis has explicitly identified a cluster that is characterized by hours concerning the night-time during the weekend (cluster #24 in Table 7), which is the period considered as the most dangerous for young car drivers in Rome (Figure 5). This cluster exhibits also the value "No Junction", indicating that these crashes occurred in a place far from a junction and, thus, most probably due to risky driving.

Table 20 - HYPOTHESES ON THE MAIN CAUSE OF THE ISSUE ABOUT YOUNG CAR DRIVERS INVOLVED IN SERIOUS CRASHES IN ROME

Having identified the risky driving as the main cause of the issue, actions for increasing road safety awareness have to be considered and, among the case studies that have been found, Pass Plus Extra course, introduced in Staffordshire in April 2006, stands out. Thus, it is possible to evaluate its effects and have an idea of the potential social and economic benefits that a similar action would lead to in Rome.

Similar results have never been achieved by previous works, not based on data science techniques. For instance, talking about previous studies performed on young road casualties relating to car accidents occurred in Rome and, more in general, in Italy, not more than generic statistics can be found.

Regarding young road casualties in United Kingdom, instead, there are several examples of studies and evaluation reports about local and national interventions taken to strongly reduce this issue. In particular, Pass Plus course has been subject to a great number of studies, but the common characteristic of all these works is that they haven't been able to reach a definitive judgement of it, due to a lack of complete information.

Two works, especially, is worth being considered.
The first one is an evaluation report about Pass Plus Extra. The report is based on the results obtained by a questionnaire sent to 1000 drivers who have completed the course. Only 28% of these drivers have answered the questionnaire and this makes the information retrieved interesting but not completely reliable. [19] The great lack of works based on surveys and questionnaire, in fact, is that it's very difficult to obtain a high number of answers and, moreover, it isn't proven that all the received answers are actually true.
The second work to take into account concerns the Welsh version of Pass Plus, the so called Pass Plus Cymru. Its main feature is that it is based on a complete literature review about all the best actions taken in this sector in the United Kingdom. In this work, also due to the lack of information, inside open data regarding car accidents, about enhanced courses completed by drivers involved in crashes, it's specified that only some assumptions can be made on the general impact of the course. [14]

The main innovation introduced with this work, thus, is the very detailed analysis on open data about road accidents, considering all their interesting features, with the purpose of obtaining results never reached before. These outcomes are much more precise and compelling than any other statistical analysis and they demonstrate the great power of open data analytics.

4.4 Cost-Benefit Analysis

As it has been described by data analysis, the serious issue about young road casualties that is affecting Rome is very similar to the problem that was present in Staffordshire until April 2006 and that has been faced by local authorities introducing Pass Plus Extra. The enhanced driving course has contributed to strongly reduce the number of accidents involving young drivers and, therefore, it's interesting to estimate the consequences that a similar action would have caused in Rome in the last four years if it had been applied in 2012.

For this reason, assuming to obtain the same effects of Pass Plus Extra in Rome, an estimate of the number of young drivers that would have been saved from personal injury accidents is provided at first. Then, a cost-benefit analysis is performed for determining, given the same costs, the same benefits and the same rate of young drivers participating to the course, how much Rome would have taken advantage of a similar action.

4.4.1 Possible Effects of Pass Plus Extra in Rome

Regarding the car accidents occurred in Rome in 2015, the available open data cover crashes until June 2015, hence a forecast about the number of young drivers involved in serious car accidents in the last six months of that year is needed. In order to perform this operation, the data from 2012 to June 2015 and the library *forecast* of R software have been used and, in particular, the Holt-Winters method has been chosen, as it is especially suitable when data present seasonality.
The following graph shows the number of young car drivers involved in serious crashes in Rome by month. The blue line represents the performed forecast, the grey area indicates the prediction interval with a confidence level of 85%. This

means that the actual values about the last six months of 2015 should lie within that interval with probability 0.85.

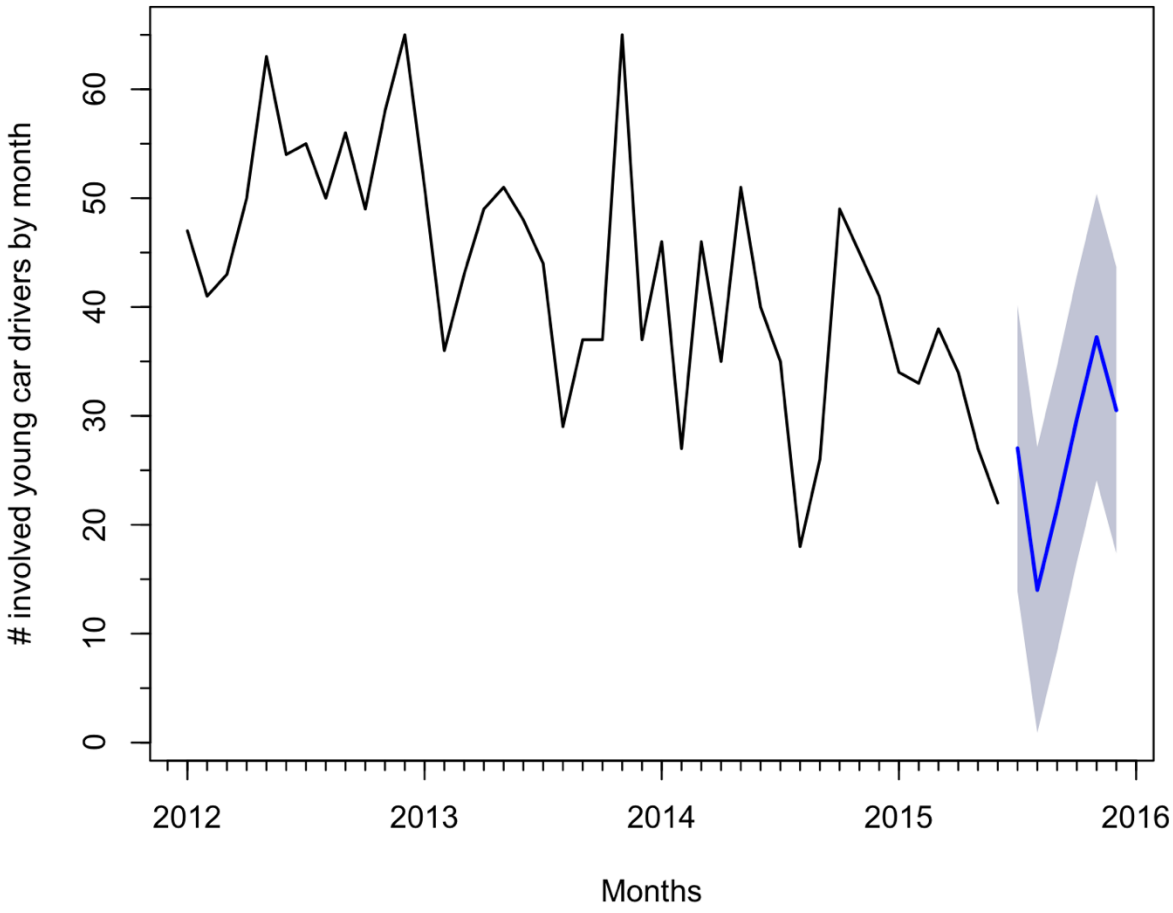

Figure 12 – NUMBER OF INVOLVED YOUNG CAR DRIVERS IN ROME BY MONTH [Years: 2012-2015]

The lowest and the highest value of the prediction interval can be used to refer to the worst and the best case.

Making use of the values contained in *Table 8* and considering the difference between the Staffordshire's trend and the trend of the rest of England as a consequence of the introduction of Pass Plus Extra, the following estimates can be obtained.

	# Involved Young Car Drivers	*Pass Plus Extra Effect*	*# Saved Young Car Drivers*
2012	631	+0%	0
2013	527	-4%	21
2014	459	-10%	46
2015 lowest value	269	-5%	13
2015 highest value	427	-5%	21

Table 21 – NUMBER OF POTENTIALLY SAVED YOUNG CAR DRIVERS IN ROME

The overall amount of potentially saved young car drivers in Rome in the last four years ranges from 80 to 88.
In the following paragraph a deep analysis on the costs and the benefits of a similar initiative in Rome is performed.

4.4.2 Costs

From the local authority's point of view, there are two types of cost: part-funding training costs and publicity costs. Regarding the first type of costs, in Staffordshire a complete Pass Plus Extra course average costs £ 130 and, having six practical modules for a total of at least six hours, the proportion is in line with the average cost of a one-hour practical lesson for driving license (£ 20). The client has only to pay £ 60 and the rest is funded by the County Council. There's no information about how much the local authority has to spend for publicity, but, looking at some documents about other counties that have been involved in the scheme, it seems that these costs have been estimated as £ 5 per client.
Converting Pounds to Euros, a value of € 165 could be assigned to the cost of a complete enhanced driving course in Rome and this would also be in line with the cost of a one-hour practical lesson for driving license (€ 30). The part-funding

training costs would be € 95 per client and each driver would only pay € 70. The publicity costs could be rounded to € 7 per client.

About the number of participants, two considerations have to be done. First, Rome has constantly had a population of about 2.5 times as much as the population of Staffordshire in the last ten years. Secondly, only data about all completions (including Staffordshire, Stoke-on-Trent, Derbyshire and Telford) are available in the Pass Plus Extra evaluation report, but, 77% of the applicants have been registered as resident in Staffordshire. For this reason, the number of participants who completed the scheme in the four examined years (*Table 9*) could be simply multiplied by 2 to get the number of potential completions in Rome.

What is obtained is described in the following table.

	# Clients Completing the Process	*Part-Funding Training Costs*	*Publicity Costs*	*Total Costs*
2012	354	€ 31,860	€ 2,478	€ 34,338
2013	720	€ 64,800	€ 5,040	€ 69,840
2014	734	€ 66,060	€ 5,138	€ 71,198
2015	928	€ 83,520	€ 6,496	€ 90,016

Table 22 – POTENTIAL COSTS FOR THE LOCAL AUTHORITY IN ROME

4.4.3 Benefits

The economic benefits that have been taken into account are represented by the savings on the social costs of road accidents. These costs are divided into two

categories: casualty related costs and accident related costs. The first category refers to lost output, medical care costs and other human costs about pain and sufferings caused by the crash. The second one concerns damages to vehicles and other property and administrative costs, which are in turn related to insurance claims processing, police costs and legal costs. Administrative costs represent only a minimum portion of the social costs. For a fatal crash, for instance, they are considered as 0.2% of the total amount. [9]

In order to calculate the social costs that would have been saved in Rome in the last four years, the following formula has been used.

$$TC = ACi * NI + ACd * ND + AAC * NA$$

in which:

- TC = total costs
- ACi = average cost per injured person
- NI = number of injured
- ACd = average cost per dead person
- ND = number of dead
- AAC = average accident related cost
- NA = number of accidents

Moreover, the following coefficients have been calculated and used to estimate the total amount of saved social costs. These represent the average number of accidents, injured and dead that correspond to a young car driver involved in a serious crash. The first coefficient, in particular, has to be computed because there are some cases in which two or more young drivers are involved in the same accident. The time period considered in these equations is from January 2012 to June 2015.

$$crashes\ per\ young\ driver = \frac{\#\ crashes\ with\ young\ car\ drivers}{\#\ involved\ young\ car\ drivers} = 0.982$$

$$injured\ per\ young\ driver = \frac{\#\ injured\ in\ crashes\ with\ young\ car\ drivers}{\#\ involved\ young\ car\ drivers} = 1.899$$

$$dead\ per\ young\ driver = \frac{\#\ dead\ in\ crashes\ with\ young\ car\ drivers}{\#\ involved\ young\ car\ drivers} = 0.008$$

In the table below, an estimate of the economic benefits about the last four years in Rome is provided. The number of potentially saved accidents, injured and dead are obtained multiplying the number of potentially saved car drivers (taken from *Table 21*) by the three coefficient calculated before. The saved costs are obtained multiplying the number of saved accidents, injured and dead by the average cost of an accident, injured and dead respectively. These values are taken from a study published by the Italian Ministry of Transport in 2010. [9]

	# Saved Young Car Drivers	*# Saved Accidents*	*Saved Accident Related Costs*	*# Injured*	*Saved Injured Related Costs*	*# Dead*	*Saved Dead Related Costs*	*Total Saved Social Costs*
2012	0	0	€ 0	0	€ 0	0	€ 0	€ 0
2013	21	21	€ 230,706	40	€ 1,688,760	0	€ 0	€ 1,919,466
2014	46	45	€ 494,370	87	€ 3,673,053	0	€ 0	€ 4,167,423
2015 worst case	13	13	€ 142,818	26	€ 1,097,694	0	€ 0	€ 1,240,512
2015 best case	21	21	€ 230,706	41	€ 1,730,979	0	€ 0	€ 1,961,685

Table 23 – POTENTIALLY SAVED SOCIAL COSTS IN ROME

Rounding to the nearest integer, there seems to be no effect on the yearly number of dead. Actually, if the decimal values related to the four years are summed up, both considering the worst and the best case for 2015, the result indicates that one life would have been saved, with the related saved cost of € 1,503,990.

4.4.4 Results and Considerations

Considering the annual amount of costs related to the local authority and benefits for Rome, this is the result.

	Total Costs	*Total Saved Social Costs*	*Saved Costs – Costs*
2012	€ 34,338	€ 0	- € 34,338
2013	€ 69,840	€ 1,919,466	€ 1,849,626
2014	€ 71,198	€ 4,167,423	€ 4,096,225
2015 worst case	€ 90,016	€ 1,240,512	€ 1,150,496
2015 best case	€ 90,016	€ 1,961,685	€ 1,871,669

Table 24 – ANNUAL POTENTIAL COSTS AND BENEFITS FOR ROME

Although a negative value is present for the first year, the economic advantages from the second year on are evident. For this reason, this enhanced course can be considered a "quick-fix action", as it quickly reaches the expected results, not only of economic type.

The costs and benefits analysed up to now are focused on the local authority's point of view. Actually, there are three more actors in this context and all of them could obtain specific benefits from the introduction of an enhanced driving course in Rome.

The first actor is represented by young drivers. Besides benefiting for the further training in special conditions and increasing their safety awareness, they could get a consistent discount on their car insurance, able to cover the expenses related to the completion of the course (estimated to be € 75 per young driver). Looking at the Staffordshire Pass Plus Extra evaluation report, in fact, the average insurance discount is about £ 100, which, converting to euros, approximately corresponds to € 125. [19]

The driving instructors are the second actor. They directly benefit from the enhanced course, because each complete course is estimated to cost € 165 on average and, in the United Kingdom, they have only to pay a fixed amount of £ 37 (about € 50) to register as Pass Plus approved instructors and £ 29 (about € 40) for each Pass Plus refill pack, which contains some guides and a list of participating insurers. [6]

Finally, the third actor is represented by the insurance companies. Their benefits are the saved costs related to the compensations and the handling of insurance claims, the costs are the total amount discounted on car insurances. In the United Kingdom not all the insurers participate to Pass Plus scheme, because no evident demonstration of the Pass Plus effectiveness has been provided yet. This work, however, has proven the effectiveness and the potentiality of Pass Plus Extra and, even if actual cost-benefit analyses aren't available about Staffordshire's insurance companies, it seems to be clear that also the third actor could take economic advantage of the introduction of a scheme similar to Pass Plus Extra.

5. CONCLUSIONS

The huge potential of open data can't be ignored anymore. They represent an essential tool available to private companies and governments for achieving economic and social benefits.

This work has demonstrated the power of open data if appropriately exploited by analyses based on data science techniques. In addition, it has proposed a modified business plan structure. Inside this innovative business plan, data science techniques have been used to precisely evaluate the effectiveness of a road safety intervention, the introduction of an enhanced driving course in this case, and provide a very accurate estimate of the social and economic advantages that a similar action would lead to. The obtained results could not have been achievable by a classic approach to the business plan, which is usually based on general statistics and experience and, thus, lacks the great accuracy provided by data science.

First, this work has revealed and described in detail a problem that affects Rome: the high number of young drivers involved in car crashes.
Then, it has found and evaluated a possible remedy, represented by an enhanced driving course called Pass Plus Extra, obtaining some results that have never been reached in previous studies. In fact, a precise quantification of the consequences that can be attributed to Pass Plus Extra has been provided and the identification of the most affected groups of young drivers has been performed.
Finally, an estimate of the impact that a similar course would have had in Rome in the last four years has been produced. This has been possible because data about Rome and Staffordshire, the county of the United Kingdom where Pass Plus Extra was introduced in April 2006, present similar features.
What has been obtained is a very precise description of the costs and the benefits related to the introduction of an enhanced driving course in Rome. The cost-

benefit analysis has been inserted in a business plan that is enriched and characterized by a deep open data analysis.

There are still some final considerations that have to be done about open data management.
First, concerning the UK dataset, as it has been also stated in previous works, the addition of the information about the completion of an enhanced driving course for each driver involved in serious crashes would help to obtain an even more accurate evaluation of the course.
Secondly, regarding the Italian open data portals, it is necessary to update and improve them by following the general guidelines about open data management. Rome's open data portal, even if it represents one of the best maintained portals in Italy, still has some shortcomings. Data about road accidents, for example, are provided in different file formats and some pre-processing operations, not so easy to perform for everyone as open data definition imposes, are necessary.

6. REFERENCES

[1] Cerasi, G. (2012, May 15). *Caro-benzina, incidenti stradali in calo del 20% ma resta il record delle vittime, 165 nel 2011.* Retrieved from Repubblica.it: http://ricerca.repubblica.it/repubblica/archivio/repubblica/2012/05/15/caro-benzina-incidenti-stradali-in-calo-del-20.html

[2] Cumbria Road Safety Partnership. (2016). *Register for Pass Plus.* Retrieved from http://www.crsp.co.uk/page/133/Register-for-Pass-Plus-.htm

[3] Delen, D., & Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems, 55*(1), 359-363.

[4] Department for Transport. (2011). STATS 20 - Instructions for the Completion of Road Accident Reports. UK.

[5] Government Digital Service. (2015, December 4). *Pass Plus*. Retrieved from GOV.UK: https://www.gov.uk/pass-plus

[6] Government Digital Service. (2015, January 8). *Pass Plus approved driving instructor (ADI) services.* Retrieved from https://www.gov.uk/pass-plus-approved-driving-instructor-services

[7] HM Government. (2009, December). Putting the Frontline First: smarter government.

[8] Ionta, F. (2015, July 22). *Quanto è "open data" la nostra pubblica amministrazione?* Retrieved from Wired.it: http://www.wired.it/attualita/politica/2015/07/22/open-data-pubblica-amministrazione/

[9] Ministero delle Infrastrutture e dei Trasporti. (2010). *Studio di valutazione dei Costi Sociali dell'incidentalità stradale.* Retrieved from http://www.mit.gov.it/mit/mop_all.php?p_id=12919

[10] News & Star. (2011, September 19). *Cumbria Pass Plus young drivers scheme survives spending cuts.* Retrieved from

http://www.newsandstar.co.uk/news/Cumbria-Pass-Plus-young-drivers-scheme-survives-spending-cuts-47223fd3-457b-425c-9e52-226fb72ff398-ds

[11] Obama, B. (2009, January 21). *Transparency and Open Government.* Retrieved from https://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment

[12] Open Knowledge. (n.d.). Retrieved March 17, 2016, from https://okfn.org/about/

[13] Open Knowledge. (n.d.). *The Open Data Handbook.* Retrieved March 17, 2016, from http://opendatahandbook.org/guide/en/

[14] Red Box Research. (2015, March). An Evaluation of Pass Plus Cymru on behalf of the Welsh Government.

[15] Stemwedel, J. D. (2008, January 29). *Basic concepts: the norms of science.* Retrieved from Adventures in Ethics and Science: http://scienceblogs.com/ethicsandscience/2008/01/29/basic-concepts-the-norms-of-sc/

[16] Tauberer, J. (n.d.). *The Annotated 8 Principles of Open Government Data.* Retrieved March 17, 2016, from https://opengovdata.org

[17] Ubaldi, B. (2013). Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. *OECD Working Papers on Public Governance*. Retrieved from http://dx.doi.org/10.1787/5k46bj4f03s7-en

[18] Viettone, E. (2012, December 18). *Niente auto e neanche la patente con la crisi gli italiani non guidano più.* Retrieved from inchieste.repubblica.it: http://inchieste.repubblica.it/it/repubblica/rep-it/2012/12/18/news/patenti_in_calo-47609542/

[19] Wilcox, K. (2010, December). Staffordshire County Council Pass Plus Extra Evaluation Report.