POLITECNICO
MILANO 1863

Master of Science in Management Engineering

AY. 2018/2019

MASTER THESIS

Big Data Quality:
Stock Market Data Sources Assessment

Supervisor: *BARBARA PERNICI*

Co-supervisors: *CINZIA CAPPIELLO*

VALTER BERNARDINI

Master's thesis of:

MOHAMED ELFAKHFAKH

Student ID: 903157

Acknowledgement

I would like to express my deepest gratitude to my supervisor, Prof. Barbara Pernici, for her valuable guidance, her continuous support and insightful knowledge and considerable encouragements throughout the whole period of this work. I really do appreciate your efforts. I would also like to thank Prof. Cinzia Cappiello for her useful and constant help, support and effort.

Additionally, I would love to extend my sincere thanks to Mr. Valter Bernardini for his continuous support and guidance. Mr. Marc Carazzato, thank you for doing everything to make this thesis work. The physical and technical contribution of 'Assioma.net' is truly appreciated.

A special thanks to my fiancée, Mariam Abdelaty, who is and has been the best partner I could ever ask for. Thank you for believing in me, your continuous support and help in the thesis. I could have never done this without you.

To my mother, your unconditional love has carried me through every moment of my life. Even though you are no longer with us, your soul has always guided me to be the best version of myself.

Finally, to all my family, friends and colleagues, I cannot thank you enough for everything you have done to me throughout the thesis. Thank you for your continuous support and encouragement whenever needed.

Mohamed Elfakhfakh

Abstract

Nowadays, the stock market data is used widely for various purposes. Companies, investors, and traders are very interested in understanding the stock prices trend. This data affects the decision-making process. Thus, high quality is essential. Since, there are many sources providing the stock market data. We proposed a model to assess the data quality of different stock market data sources, ranking them, and choosing the most reliable one. Moreover, we developed a predictive model using Long-Short Term Memory (LSTM) architecture to predict the missing values in that source, in order to enhance its quality. Three main dimensions were used to assess the quality, namely, **completeness, consistency and accuracy**. We introduced a Quality Indicator (QI) index to rank the sources. The data were collected from four sources: Yahoo Finance, MSN Money, Stooq and Tiingo. The collected data is focusing on 60 companies in NASDAQ stock market over a period of 10 months from January 2019 to October 2019. The quality glitches were mainly in the completeness and accuracy dimensions, and no glitches were found in the consistency dimension, resulting to choose Yahoo Finance as the most reliable source. In addition, we used the predictive model on a sample of three companies, to fill in the missing days in the chosen source.

Keywords: Data Quality, Data quality dimensions, Stock market, Quality indicator, Data quality assessment, LSTM, Data prediction

Table of Contents

Chapter 1: Introduction 1
1.1 Motivation 1
1.2 Aim of the Thesis 1
1.3 Thesis outline 2
Chapter 2: State of Art 3
2.1 Introduction to Big Data 3
2.2 Introduction to the concept of DQ 4
2.3 effect of poor-quality DQ 5
2.3.1 Disasters caused by poor DQ 5
2.3.2 How poor DQ affects businesses 7
2.4 Data Quality management (DQM) 8
2.5 Data Quality Dimensions 9
2.5.1 Intrinsic Dimensions 10
2.5.2 Contextual Dimensions 11
2.6 Approaches to Data quality Dimensions 11
2.6.1 Theoretical Approach 12
2.6.2 Empirical Approach 12
2.6.3 Intuitive Approach 13
2.7 Changes in Data quality 14
2.8 Data quality models 16
2.8.1 Total DQ Management (TDQM) 17
2.8.2 Information Integrity Methodology (IIM) 17
2.8.3 AIM Quality model (AIMQ) 17
2.8.4 DQ Management Maturity Model (DQMMM) 18
2.8.5 Complete DQ Management (CDQM) 18
2.9 Different techniques to deal with data 19
2.9.1 Quality Assessment using batch processing 20
2.10 Measuring data quality 22
2.10.1 Domains presented by the Data quality Literature 23
2.10.2 Measuring data quality and data quality dimensions 24
2.10.3 Data quality measurement in the financial domain 37

2.11 Deep learning in Stock market data forecast 48
Chapter 3: Methodology 51
3.1 The implemented model 51
3.1.1 DQ basics phase 54
DQ dimensions and metrics 54
3.1.2 Data preparation phase 59
Data collection step 59
Data pre-processing step 59
Chapter 4: Model Implementation 61
4.1 data preparation 61
4.2 DQ assessment phase 65
4.3 Sources evaluation phase 65
4.4 Predictive model 65
Chapter 5: Results 68
5.1 First scenario: TOP 20 COMPANIES in nasdaq 68
5.1.1 Completeness dimension 68
5.1.2 Consistency dimension 68
5.1.3 Accuracy dimension 69
5.1.4 Summary 74
5.2 Second scenario. MEDIUM SIZE COMPANIES 74
5.2.1 Completeness dimension 74
5.2.2 Consistency dimension 79
5.2.3 Accuracy dimension 79
5.2.4 Summary 86
5.3 Third scenario: SMALL AND MICRO COMPANIES 87
5.3.1 Completeness dimension 87
5.3.2 Consistency dimension 89
5.3.3 Accuracy dimension 89
5.3.4 Summary 96
5.4 The three scenarios comparison 96
5.5 Predictive model results 97
Chapter 6: Conclusion and future work 100
6.1 Conclusion 100
6.2 Future work 101

List of Figures

Figure 2.1: The Five V's of big data [9] 4
Figure 2.2: DQ word cloud 5
Figure 2.3: O-ring Leakage [2] 6
Figure 2.4: Number of correct data records for a study involving 75 executives [13] . 8
Figure 2.5: DQM process 9
Figure 2.6 Practical dimensions of DQ [15] 10
Figure 2.7: Data Quality architecture presented by [35] 21
Figure 2.8: Quality Rules Discovery Framework [37] 22
Figure 2.9: The domain of studies [31] 23
Figure 2.10: Mapping between dimensions of DQ and DQ assessment methods[44] 31
Figure 2.11 Networked Grouping of Information Quality Criteria [46] 33
Figure 2.12: The model used in [51] 37
Figure 2.13: The data quality axes [55] 40
Figure 2.14: RNN simple cell versus LSTM cell [66] 50
Figure 3.1: (a) Sources assessment model, (b) Predictive model 53
Figure 4.1: LSTM model architecture 66
Figure 5.1: The constraints types for all sources(Top 20 Companies) 69
Figure 5.2:The maximum error value for the three quarters in Stooq (Top 20 Companies) 70
Figure 5.3: The error distribution for the third quarter in Yahoo (Top 20 companies) 70
Figure 5.4:Error distribution across all quarters in Stooq (Top 20 Companies) 71
Figure 5.5: The error count for each company in Stooq (Top 20 Companies) 72
Figure 5.6: The error count across the whole period for all sources (Top 20 Companies) 73
Figure 5.7: The MSE for all quarters in all sources (Top 20 Companies) 73
Figure 5.8: Count of missing values for all sources (Medium companies) 75
Figure 5.9: missing in NEBU for the first and second quarters in Tiingo (Medium companies) 75
Figure 5.10: missing values in BMLP and NEBU across all quarters in Stooq (Medium companies) 76
Figure 5.11:The missing days in the MSN source for each company 77
Figure 5.12:Heatmap for Stooq and MSN sources (Medium companies) 78
Figure 5.13:Number of excess days reported in MSN (Medium companies) 79
Figure 5.14:Yahoo error distribution for the third quarter(Medium companies) 80
Figure 5.15:Tiingo error distribution for the third quarter(Medium companies) 81
Figure 5.16: Maximum error values for all the sources across the quarters(Medium companies) 81
Figure 5.17:The count of the error per company for Stooq is on the left side, while the mean error per each company is a on the right (Medium companies) 82
Figure 5.18: Error distribution in Stooq for all quarters(Medium companies) 83
Figure 5.19: Number of errors in MSN for each Company (Medium Companies) 83
Figure 5.20: Error distribution in MSN for all quarters(Medium companies) 84

Figure 5.21: The Error distribution in the first quarter of a sample company in MSN(Medium companies)....84
Figure 5.22:Error count for all sources in each quarter (Medium)....85
Figure 5.23: MSE for all sources across the quarters (Medium)....86
Figure 5.24:Number of missing values for each source (Small-Micro companies)....88
Figure 5.25:Reported days in Yahoo for ONTO Symbol (Small-Micro companies)..88
Figure 5.26:On the left side is the number of missing values per company while on the right side is the missing values for WSC across the year (Small-Micro companies)..88
Figure 5.27: Extra reported days in MSN for each company (Small-Micro companies)....89
Figure 5.28:Error Distribution in Stooq for each period (Small-Micro companies)...90
Figure 5.29: The number of errors in Stooq for each company (Small-Micro companies)....91
Figure 5.30:The number of errors in MSN for each company (Small-Micro companies)....91
Figure 5.31: Error distribution in Tiingo for the first and the second quarter (Small-Micro companies)....92
Figure 5.32: The number of errors in Tiingo for each company (Small-Micro companies)....92
Figure 5.33:Error distribution in MSN for each period (Small-Micro companies).....93
Figure 5.34:the number of errors in each source for each quarter (Small-Micro companies)....94
Figure 5.35: MSE in all sources for each quarter (Small-Micro companies)....95
Figure 5.36:The number of errors in Tiingo and MSE without the NTIC symbol (Small-Micro companies)....95
Figure 5.37: The predictive vs the real in Alcoa Corp....98
Figure 5.38: The predictive vs the real in Almaden Minerals Ltd....99
Figure 5.39:The predictive vs the real in GE....99

List of Tables

Table 2.1: Three V's definition 3
Table 2.2 Dimensions proposed in the empirical approach [17] 12
Table 2.3: Dimensions proposed in the intuitive approach [18] 13
Table 2.4:Ten potholes on Information Quality [22] 15
Table 2.5: Quality-in-use model for Big Data based on ISO 25012 [23] 16
Table 2.6 Strengths and Weaknesses of DQ [30] 19
Table 2.7: Dimensions measurements and weights [39] 25
Table 2.8: The metrics for semantic accuracy [43] 28
Table 2.9:The metrics for syntactic accuracy [43] 29
Table 2.10:The metrics for uniqueness [43] 29
Table 2.11:The metrics for consistency [43] 29
Table 2.12:The metrics for completeness [43] 30
Table 2.13: The summary for the DQ dimensions, the approach and the selected domain used for nonfinancial publications 35
Table 2.14: Three case studies in [54] 39
Table 2.15: Types of constraints proposed by [5] 42
Table 2.16: List of the attributes examined in[6] 44
Table 2.17: The summary for the DQ dimensions and the approach used for financial publications 47
Table 3.1: The rules the dataset should meet and the affected dimension 54
Table 3.2: The list of constraints to be calculated in each constraint type 57
Table 4.1: Companies classification by market capitalization provided by NASDAQ [77] (M: Million, B:Billion) 62
Table 4.2: The top 20 companies of the first scenario 62
Table 4.3: The second scenario list of companies 63
Table 4.4: The Third scenario list of companies 63
Table 4.5: The standard calendar quarters that make up the year [79] 65
Table 4.6: Sample data for GE used as input in the predictive model 66
Table 4.7: The final chosen hyperparameter for the LSTM model 67
Table 5.1: Statistical summary for error in all sources(Top 20 companies) 70
Table 5.2: Accuracy dimension values for each source (Top 20 companies) 74
Table 5.3: Dimensions summary for the first scenario 74
Table 5.4:The count of missing values in NEBU and BMLP companies in Stooq (Medium companies) 76
Table 5.5: The count of missing values in NEBU and BMLP companies in MSN source 77
Table 5.6:Statistical summary for the error in each quarter for all sources (Medium companies) 80
Table 5.7:Accuracy dimensions values for each source (Medium companies) 86
Table 5.8:Dimensions summary for the second scenario 87
Table 5.9: The statistical parameter of error distribution for all sources across quarters (Small-Micro companies) 90
Table 5.10: Accuracy dimensions for all sources (Small-Micro companies) 96
Table 5.11: Dimensions summary for the Third scenario 96

Table 5.12: The QI for all the sources in each scenario.....97

List of Terms

AIMQ	AIM Quality Model
ARCH	Autoregressive conditional heteroscedastic
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
CDQM	Complete DQ Management
DL	Consecutive Changes
CNN	Convolutional Neural Network
DMA	Data Management Association
DQ	Data Quality
DQES	Data Quality Evaluation Scheme
DQM	Data Quality Management
DQR	Data Quality Rules
DQMMM	DQ Management Maturity Model
EHR	Electronic Health Record
ETL	Extract Transfer Load
FIT	Feed Inspection Tool
Q1	First Quarter
Q4	Fourth Quarter
GQM	Goal Questions Metric
IIM	Information Integrity Methodology
IP	Information Protocol
IQ	Information Quality
ICs	Integrity Constraints
IOT	Internet of Things
LOD	Linked Open Data
LSTM	Long-Short Term Memory
Market Cap	Market Capitalization
MSE	Mean Squared Error
MLP	Multi-Layer Perceptron
NASDAQ	National Association of Securities Dealers Automated Quotations
NYSE	New York Stock Exchange
NMSE	Normalized Mean Squared Error
SP	Price Spread
PSP	Product and Service Performance Model
QI	Quality Indicator
RNNs	Recurrent Neural Networks
RIS	Research Information Systems
Q2	Second Quarter
STD	Standard Deviation
LG	Successive Trends
Q3	Third Quarter
TAR	Threshold Autoregression
TDQM	Total DQ Management

Chapter 1: Introduction

1.1 MOTIVATION

No doubt that the world we are living in today is being driven by data more and more every year. According to a report done by McKinsey Global Institute (MGI): The age of analytics: Competing in a data-driven world [1]; data volume continues to double every three years as information emanates from digital platforms, wireless sensors, applications for virtual reality, and billions of mobile phones. Capacity for data storage has increased while costs have dropped. Consequently, businesses do not have to go on gut instinct anymore; they can use data and analytics to make decisions quicker and accurate predictions backed by an enormous amount of evidence.

In order to exploit the power of data we must validate its quality. Poor Data Quality (DQ) can be the reason of tragic disasters e.g. space shuttle Challenger and the USS Vincennes/Iranian Airbus disasters [2]. Both businesses and economy can be affected by the quality of the data as well. IBM estimated a whopping 3.1 trillion US dollars as the cost of poor DQ on the US economy in the year 2016 alone [3]. It has also operational impacts like lowering customer satisfaction, typical impacts like difficulties to implement data warehouses, and strategic impacts like difficulties to set and implement the strategy [4]. As a result of the aforementioned reasons, DQ has been a thriving point of research lately with several applications in various domains.

1.2 AIM OF THE THESIS

The financial health of a company can be determined from its stock price and its trend. An increase in the company's profit may lead to the rising of its stock price, whereas a huge amount of debt may lead to the opposite. So, it is very important for the investors and the shareholders to monitor the stock prices for all the companies they are interested/invested in.

There is an enormous number of sources that provide stock prices. Some are offered free of charge, while others should be paid for to be granted access. The data is available in many formats as well, which make it more challenging to maintain its

quality. The DQ can be assessed minute by minute [5] or it can be assessed day by day [6], in this thesis end of the day data has been used to assess the DQ.

The main objective of the thesis is to assess the quality of different sources providing stock market data, more specifically NASDAQ (National Association of Securities Dealers Automated Quotations) stock market, to rank them and identify the most reliable source to be used in any application. The datasets gathered from these sources will be compared to the ground truth to evaluate its quality. The ground truth will be the dataset provided form the NASDAQ official website, which is the original source of the data and the only credible one. The ground truth is a single true value. The examined data is a 10 months period from January 2019 to October 2019. The quality assessment for the sources will be conducted using the proposed model in Chapter 3. It has four phases, namely, DQ basics phase, data preparation phase, DQ assessment phase, and sources evaluation phase.

DQ assessment phase is performed on three different scenarios, categorized based on the company size that can be expressed by its Market Capitalization (Market Cap). The three classification are Top 20 companies, medium companies, small and micro companies, with Market Cap of greater than 10B$, form 2B$ to 10B$, from 300M$ to 2B$ and less than 300M$ respectively. For the second and the third scenario the companies are chosen randomly from different geographical areas and different industries to ensure the presence of all possible ranges of companies.

A predictive model using Long-Short Term Memory (LSTM) is developed by using the historical data of the past years as input. This model is used to predict the stock price in a specific day which could be missing in the most reliable source to make the source more reliable in order to be used in any application.

1.3 THESIS OUTLINE

In Chapter 2, an overview on the DQ and its management along with the DQ assessments are presented. Afterward, the methodology used to assess the quality and its steps are illustrated in Chapter 3. Followed by the implementation steps of the model and the used tools are discussed in Chapter 4. Then in Chapter 5, the proposed model's results are shown and analysed. Finally, the conclusion and future work will be mentioned in Chapter 6.

Chapter 2: State of Art

2.1 INTRODUCTION TO BIG DATA

In the recent years, big data has emerged as one of the trendiest topics with numerous applications in a multitude of fields ranging from daily tasks to space exploration. The data analytics giant SAS defines big data as a "large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis" [7]. According to SAS, the term "big data" refers to data that is so large, fast or complex that it's difficult or impossible to process using traditional methods [7].

Big data has been introduced in terms of the three V's [8]: volume, velocity, and variety. The elaboration of each one is shown in Table 2.1.

Table 2.1: Three V's definition

The 3 V's	Definition
Volume	The amount of data needed to be stored compared to the possibility of storing and managing it.
Velocity	The calculation speed required to process the data relative to the rate of receiving the data.
Variety	The number of the different formats included in the data.

Most of the big data definitions focus on these three V's, although, lately they have added 2 more attributes to define data quality, namely, veracity and value [9] as shown in Figure 2.1. By analyzing big data, valuable information can be extracted, the results of such an analysis are hardly reliable unless well-defined. Moreover, proper verification and quality control mechanisms are applied to the data before it is used.

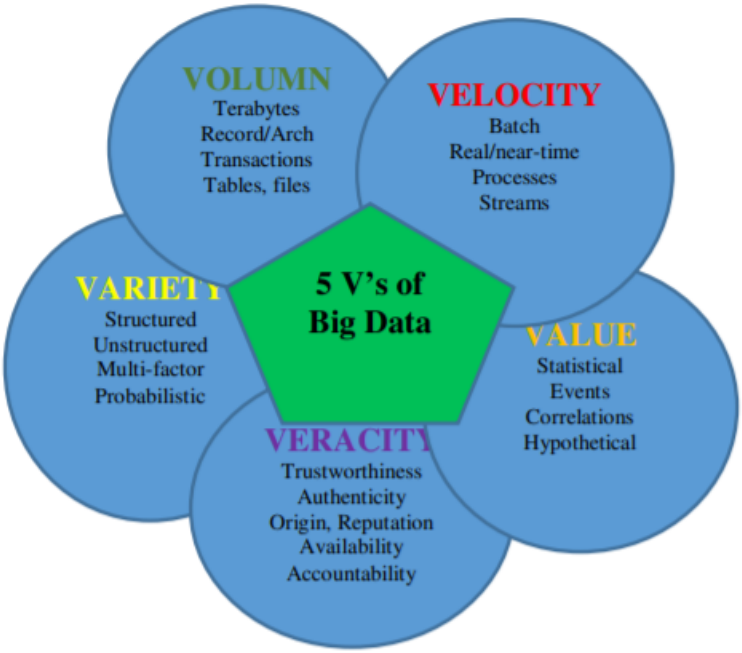

Figure 2.1: The Five V's of big data [9]

2.2 INTRODUCTION TO THE CONCEPT OF DQ

A term that is usually associated with big data is DQ. Data is considered of high quality if it correctly reflects the real-world status, allows the party using the data to effectively get useful insights that help to determine the clients' needs and to find the best ways to serve the clients.

DQ does not necessarily mean zero defects, but it is the conformance of data to valid requirements. Therefore, we must determine who sets the requirements, the rules by which the requirements are set, and the degree of conformance needed by these requirements [10]. In short, DQ is the assessment of how much the data is usable and fits its serving context. In Figure 2.2, the word cloud for the most common words related to data quality is presented.

Figure 2.2: DQ word cloud

2.3 EFFECT OF POOR-QUALITY DQ

DQ is crucial for business intelligence and other data analytics. It leads to a high operational efficiency and a better customer relationship. In the information technology field, it helps to increase the accuracy of the system or of the software used; the better the DQ the more reliable results you get. Basically, if the data is flawed the results will be flawed, hence, the decisions most probably will be flawed as well. But what could really happen if the quality of the data is not good? In this section we will discuss the consequences of poor DQ with a real-life disaster and how it can affect the businesses.

2.3.1 Disasters caused by poor DQ

When the data is flawed, which means that its quality is not good, it may lead to life-threatening decisions. These wrong decisions may lead to a serious disaster not just to the business, like losing profit and bad customer relationship but also could affect human lives.

There are two well-known disasters, namely, the space shuttle Challenger and the USS Vincennes/Iranian Airbus disasters. They have been mentioned in many newspapers as well as in the literature, however, not many focused on the problem of DQ which is one of the main reasons that caused these disasters. The authors of [2] they focused on these reasons from a DQ perspective.

NASA launched the space shuttle Challenger on 28 January 1986, however, there was an internal debate about the safety of the O-rings in cold temperatures. The committee that investigated this disaster reported that the main reason causing it was due to a flawed decision-making process. Mainly, allowing the rocket's launch while there was an evidence of a possible problem. This led to a leakage as shown in Figure 2.3. These flaws in the decision process are mainly due to the following quality issues: accuracy, completeness, consistency and relevance [2]. Not only the DQ issue was the reason for this disaster, but also other theories have highlighted narcissism and the organization decay, information format, interaction of images and technology as reasons for the disaster [4].

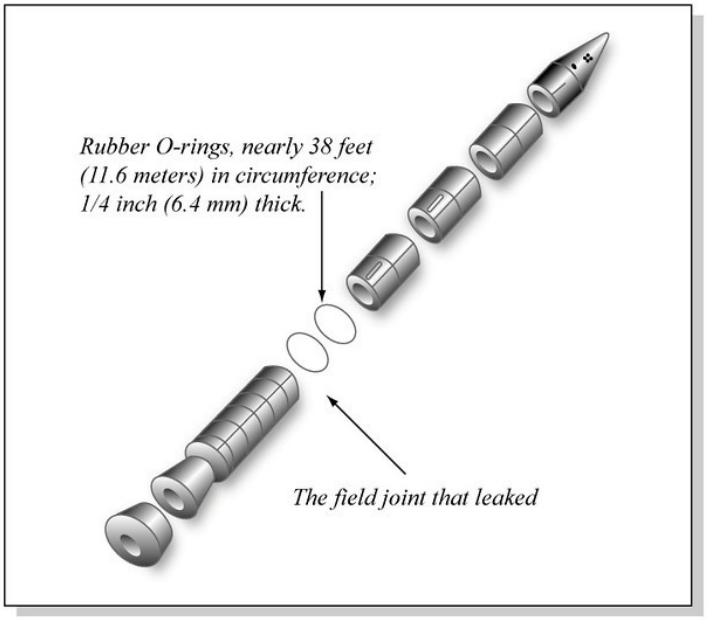

Figure 2.3: O-ring Leakage [2]

The accuracy problem was highlighted by the erroneous identification of the O-rings. As it was reported that one manager has declared that the problem of the O-rings was solved without any evidence or consent of doing that. On the other hand, the consistency problem was presented in the O-rings misclassification. In some cases, the equipment needed for the O-rings were marked as redundant, while in other cases they were not.

The USS Vincennes took down an Iranian Airbus on 3 July 1988 by mistake, results of killing 290 civilians. Several justifications have been given as a reason to mistake a civilian aircraft to a fighter like inexperienced crew having poor reaction to combat, insufficient time to verify data, incomplete training and hostilities in the area

that created an environment conductive to incorrect interpretation. Although the main reason was the poor DQ, specifically in these dimensions: accuracy, completeness, consistency and timeliness. [2]

It is obvious that the DQ was not an important aspect to get the attention of the decision-making board in both cases. As mentioned in [2] there were 10 glitches spread over five quality dimensions for Challenger case like the erroneous identification of the O-rings in the accuracy dimension and there were eight glitches spread over five quality dimensions for the USS Vincennes case like the error of the their system that said that the aircraft is in ascending mode while the crewmen operating on separate console reported that the aircraft is in descending mode and this affects the accuracy dimension. Given that it is difficult to believe that a proper decision could be made with the existence of glitches in the dataset.

2.3.2 How poor DQ affects businesses

Poor-quality data cause good decision making to be so much harder and a lot more costly to the business. Thomas C. Redman in his book "data driven" introduced the so-called "rule of 10", which provides a simple way to estimate the extra costs of bad decisions taken due to a DQ problem. Redman observed that it costs 10 times as much to complete a unit of work when the input data are defective as it does when they are perfect [11]. Thus, someone who's using a dataset with 80% good data will take good decisions with no added effort 80% of the time, but 20% of the time it will cost about 10 times as much to make corrections and to complete the work.

IBM estimated a whopping 3.1 trillion US dollars as the cost of poor-quality data on the US economy in the year 2016 alone [3]. The reason poor-quality data costs so much is that decision makers, managers, knowledge workers, data scientists, and others must accommodate it in their everyday work. And doing so is both time-consuming and expensive. The data they need has plenty of errors, and in the face of a critical deadline, many individuals simply make corrections themselves to complete the task at hand. They do not think to reach out to the data creator, explain their requirements, and help eliminate root causes [12]. Using poor-quality data can as well lead to some non-financial impacts such as the loss of credibility for your business, customer dissatisfaction, and increasing risk levels.

In a study that was done over a 2 years period and was published in the Harvard business review, involving 75 executives from different businesses and departments, only 3% found that their department fell within the acceptable range of 97 or more correct data records out of 100 (DQ score) [13]. The findings of the study can be seen in Figure 2.4. The study found that on average, 47% of the newly created data records have at least one critical (e.g., work-impacting) error [13].

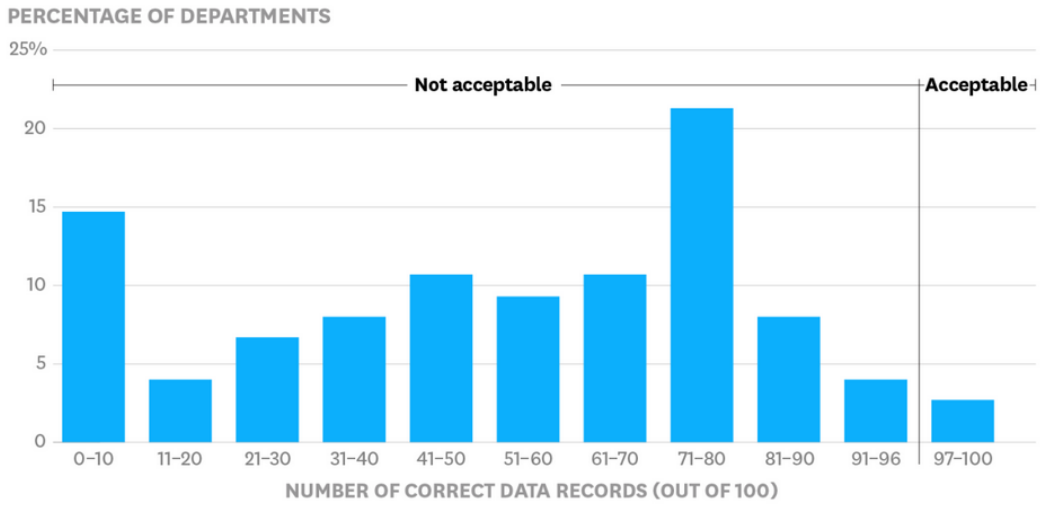

Figure 2.4: Number of correct data records for a study involving 75 executives [13]

From a managerial point of view, these results can be scary as whether, as a manager, you see it or not, most data are bad unless you take the right measures to make sure your data is of high quality.

2.4 DATA QUALITY MANAGEMENT (DQM)

The Data Management Association (DAMA) defines data management as "the business function that develops and executes plans, policies, practices, and projects that acquire, control, protect, deliver, and enhance the value of data" (DAMA, 2011, p. 78). From DAMA point of view, DQ management is a function within the overall scope of data management.

As mentioned in the previous section, poor-quality data can lead to poor decisions. Hence, DQM is an essential process for any business as it can save a lot of time and money. DQM is a set of practices that aim at maintaining a high quality of the

information and make sure the used data is relevant, reliant and accurate. A DQM program establishes a framework for all departments in the organization that provides, and sometimes enforces, rules for DQ [14]. For any set of data, according to [10], there's a four-phase process for achieving successful DQM. This process is shown in Figure 2.5.

Figure 2.5: DQM process

- Data profiling is the process of gaining an understanding of the existing data relative to the quality specifications. Data profiling determines if the data is complete and accurate.
- In the DQ step, we build on the information learned in data profiling to understand the causes of the problems.
- Data integration involves combining data residing in different sources and providing users with a unified view of them.
- Data augmentation involves combining internal data with data from external sources not related to the base data, to increase the level of understanding and gain insights.

2.5 DATA QUALITY DIMENSIONS

DQ dimensions are a useful measurement approach to compare DQ levels across different systems over time. There are many different sources describing and talking about the DQ dimensions; however, all of them almost followed the same approach of describing it, although it often refers to different levels and different data model elements. Below is the interpretation of one of these authors.

David Loshin in [15] said that "Different dimensions are intended to represent different measurable aspects of DQ and are used in characterizing relevance across a set of application domains to monitor against the specified organizational standard of DQ". Loshin categorized the practical dimensions of DQ into intrinsic dimensions and contextual dimensions; Intrinsic dimensions relate to the data values themselves out of

a specific data or model context, while the contextual dimensions look at the data element in relation with other data elements (driven by context). Figure 2.6 shows the practical dimensions of DQ as introduced in [15].

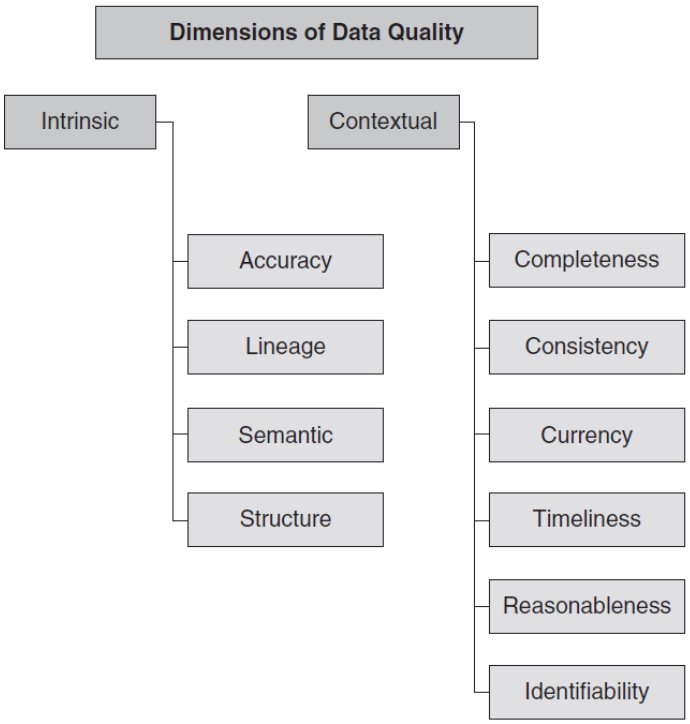

Figure 2.6 Practical dimensions of DQ [15]

2.5.1 Intrinsic Dimensions

The intrinsic dimensions focus on the values of data themselves, without necessarily evaluating the context of those values. These dimensions characterize structure, formats, meanings, and enumeration of data domains – essentially the quality of organizational metadata and how it is used [15]. The intrinsic dimensions introduced are:

- **Accuracy:** it refers to the degree to which data values agree with an identified source of correct information.
- **Lineage:** A dimension measuring the historical sources of data in order to have the ability to identify the source of any new or updated data element and hence measure the trustworthiness of the data.
- **Semantic consistency:** Semantic consistency refers to consistency of definitions among attributes within a data model, as well as similarly named

attributes in different enterprise datasets, and it characterizes the degree to which similar data objects share consistent names and meanings.

- **Structural consistency:** Structural consistency refers to the consistency in the representation of similar attribute values, both within the same dataset and across the data models associated with related tables.

2.5.2 Contextual Dimensions

The contextual dimensions provide a way for the analyst to review conformance with DQ expectations associated with how data items are related to each other [15]. The contextual dimensions introduced are:

- **Completeness:** refers to the expectation that certain attributes are expected to have assigned values in a dataset.
- **Consistency:** relevant to the different levels of the data hierarchy, within tables, databases, across different applications, as well as with externally supplied data which is in another words integrity constraints.
- **Currency:** refers to the degree to which information is current with the world that it models. Currency can measure how "up to-date" information is, and whether it is correct despite the possibility of modifications or changes that impact time and date values.
- **Timeliness:** refers to the time expectation for accessibility of information. Timeliness can be measured as the time between when information is expected and when it is readily available for use.
- **Reasonableness:** this dimension includes general statements associated with expectations of consistency or reasonability of values, either in the context of existing data or over a time series.
- **Identifiability:** refers to the unique naming and representation of core conceptual objects as well as the ability to link data instances containing entity data together based on identifying attribute values.

2.6 APPROACHES TO DATA QUALITY DIMENSIONS

This section describes three main different approaches adopted for addressing comprehensive sets of DQ dimensions definitions. There approaches are theoretical, empirical, intuitive. The theoretical approach it contains a formal model to define the dimensions. The empirical approach constructs the set of dimensions from experiments,

interviews, and questionnaires. The intuitive approach basically defines the dimensions based on common sense and practical experience.

2.6.1 Theoretical Approach

This approach to the definition of DQ is proposed in Wand and Wang [16], the identified dimensions are presented below quoted from [16]:

- Accuracy: "inaccuracy implies that the information system represents a real-world state different from the one that should have been represented."
- Reliability: "whether the data can be counted on to convey the right information; it can be viewed as correctness of data."
- Timeliness: "the delay between a change of the real-world state and the resulting modification of the information system state." Lack of timeliness may lead to a state of past data (out-of-date data)
- Completeness: "the ability of an information system to represent every meaningful state of the represented real-world system."
- Consistency: "inconsistency would mean that the representation mapping is one-to-many."

2.6.2 Empirical Approach

This approach is mentioned in Wang and Strong [17]. Through interviewing data customers, DQ dimensions have been chosen. Among of 179 DQ dimensions, the author focused on 15 of them (see- Table 2.2).

Table 2.2 Dimensions proposed in the empirical approach [17]

Category	Dimension	Definition: the extent to which
Intrinsic	Believability	Data are accepted or regarded as true, real and credible
	Accuracy	Data are correct, reliable and certified free of error
	Objectivity	Data are unbiased and impartial
	Reputation	Data are trusted or highly regarded in terms of their source and content
Contextual	Value-added	Data are beneficial and provide advantage for their use
	Relevancy	Data are applicable and useful for the task at hand
	Timeliness	The age of the data is appropriate for the task at hand
	Completeness	Data are of enough depth, breadth, and scope for the task at hand

	Appropriate amount of data	The quality or volume of available data is appropriate
Representational	Interpretability	Data are in appropriate language and the data definitions are clear
	Ease of understanding	Data are clear without ambiguity and easily comprehended
	Representational consistency	Data are always presented in the same format and are compatible with the previous data
	Concise representation	Data are compactly represented without being overwhelmed
Accessibility	Accessibility	Data are available or easily and quickly retrieved
	Access security	Access to data can be restricted and hence kept secure

Wang and Strong have classified the dimensions into four categories, which they added two more categories that were discussed in [15]:

- Representational DQ captures what is related to the data representation quality. (e.g. interoperability)
- Accessibility DQ can be interpreted from its name, is related to accessibility of data and the level of security.

2.6.3 Intuitive Approach

This approach is discussed by Redman [18]. DQ dimensions are classified into three categories, namely, conceptual schema, data values, and data format. As we are not concerned about the conceptual schema, so in Table 2.3 we present the definitions provided by Redman for data value and format dimensions only.

Table 2.3: Dimensions proposed in the intuitive approach [18]

Dimension Name	Type of dimension	Definition
Accuracy	Data value	Distance between the true value and the value, considered as correct
Completeness	Data value	Degree to which values are present in a data collection
Currency	Data value	Degree to which datum is up to date
Consistency	Data value	Coherence of the same datum, represented in multiple copies, or different data to respect integrity constraints and rules

Appropriateness	Data format	One format is more appropriate than another if it is more suited to user needs
Interpretability	Data format	Ability of the user to interpret correctly values from their format
Portability	Data format	The format can be applied to as a wide set of situations as possible
Format precision	Data format	Ability to distinguish between elements in the domain that must be distinguished by users
Format flexibility	Data format	Changes in user needs and recording medium can be easily accommodated
Ability to represent null values	Data format	Ability to distinguish neatly (without ambiguities) null and default values from applicable values of the domain
Efficient use of memory	Data format	Efficiency in the physical representation. As icon is less efficient than a code
Representation consistency	Data format	Coherence of physical of data with their formats

2.7 CHANGES IN DATA QUALITY

There are many changes in the past years that highly affected the DQ and how to deal with it. Changes like the possibility to increase the processing power, the increase of communication speed, the increase of physical storage room with a practically low cost and the apparition of ubiquitous devices [19][8]. In addition to that the availability of various cloud computing and associated commercial solutions, it is not required now to buy and deploy an IT infrastructure from scratch [20][21].

These new changes have dramatically affected the traditional vision of DQ. Several potholes in the path of information quality that affect the DQ dimensions. As shown in Table 2.4, the ten potholes and the affected dimensions. That leads to a transition from the stable and controlled solid ground to a dynamic unstable world, which the data is being received from different sources, velocity, sizes, format, and representations. In other words, quoted from [22], we are shifting from what they called “close world assumptions” to “beautiful and challenging chaos”.

Mainly the changes that are needed to be faced are known as the 3Vs: Velocity (e.g. data coming at real time or streaming data), Volume (e.g. data coming in a huge size or in tables or in files), Variety (e.g. data coming in unstructured way) [8]. Some

authors add a fourth V as well which is Veracity [22] so it becomes be the 4Vs. Others have added two to become five Vs [9] as mentioned before in Section 2.1.

Table 2.4:Ten potholes on Information Quality [22]

Potholes	Affected DQ dimension(s)
1. Multiples sources of the same information produce different values	Consistency and believability
2. Information is produced using subjective judgements, leading to bias	Objectivity and believability
3. Systemic errors in information production lead to loss of information	Correctness and completeness
4. Large volumes of stored information make it difficult to access the information in reasonable time	Concise representation, timeliness, value-added, and accessibility
5. Distributed heterogenous systems lead to inconsistent definitions, formats and values.	Consistent representation, timeliness, and value-added
6. Nonnumeric information is difficult to index	Concise representation, value-added, and accessibility
7. Automated content analysis across information collection is not yet available	Analysis requirements, consistent representation, relevance, and value-added
8. As information consumers' task and the organizational environment change, the information that is relevant and useful changes, the information that is relevant and useful changes.	Relevance, value-added, and completeness
9. Easy access to information may conflict with requirements for security, privacy, and confidentiality.	Security, accessibility, and value-added
10. Lack of sufficient computing resources limits access.	Accessibility, and value

Because of these changes some authors proposed some solutions for this issue. As for [23] they proposed a DQ model called the 3Cs, which is combination of Consistency, Temporal Consistency and Operational Consistency. Below are the three consistency types with their description:

Contextual Consistency refers to capability of datasets to be used within the same domain of interest of the problem independent from any format (e.g. structured vs unstructured), any size, or coming at different velocities.

Temporal consistency refers to the fact that dataset is generated throughout time. The time is used for performing analysis and understood data consistency.

Operational Consistency refers to the extent of which dataset can be included in the same analysis, from a technological point of view. Where basically means the data accessibility.

In [23], they are claiming that the main DQ dimension is consistency and all the types of consistency in order to assess the level of quality for big data project. As shown in Table 2.5, these three consistencies will affect most of the external DQ dimensions based on ISO 25012, which is the standard that can be used to establish DQ requirements, define DQ measures, and perform DQ assessments.

Table 2.5: Quality-in-use model for Big Data based on ISO 25012 [23]

DQ characteristics	Contextual consistency	Temporal consistency	Operational consistency
Accuracy	X		X
Completeness	X		X
Consistency	X	X	X
Credibility	X	X	
Currentness		X	
Accessibility			X
Compliance		X	X
Confidentiality	X		
Efficiency			X
Precision			X
Traceability			X
Understandability	X		
Availability		X	X
Portability			X
Recoverability			X

2.8 DATA QUALITY MODELS

In order to manage DQ dimensions and improve it, it is important to follow a systematic process to ensure better quality within the organization and to make sure of the presence of continuous quality check. Therefore, many researches have proposed models and methodologies for DQ management.

2.8.1 Total DQ Management (TDQM)

It is basically an extension of Total Quality Management (TQM) framework which is using for physical product quality. TDQM has been proposed to support the concept of "data as a product", that they used the same procedures used in TQM to achieve high quality [24].

The methodology starts with the information product (IP) concept. At this point, in order to achieve high quality state, the IP has its own characteristics and specifications. The information quality (IQ) metrics are then developed and used to calculate the IP. The outcome of the measurement is then analyzed using statistical control of the process, identification of trends and comparison map. Finally, improvement of TDQM.

Nonetheless, when comparing information output to physical production, there are several concerns. These included the ability to share data between users. Second, when needed, raw data may not arrive in time and several value measurements such as integrity are difficult to assign to physical production. TDQM has been designed to manage the quality of data in databases and current technologies, including big data, may limit their use. This is because of the variety of data types in big data available. The framework can be redesigned by incorporating other data sources into big data in future work.

2.8.2 Information Integrity Methodology (IIM)

This methodology has been introduced later and expressed the need to meet information integrity by focusing on the foundation of the data itself [25]. Information integrity considered the ability to meet strategic goals of the organizations. However, a requirement for information integrity should be met in order to achieve high data reliability. The framework contains data policy, capability of organization, data management, design, system, verification, interaction, and compliance with the framework. On the other hand, the proposed methodologies added another phase of DQ management to reassure the quality of the data after the process of improvement.

2.8.3 AIM Quality model (AIMQ)

It includes the Product and Service Performance model (PSP) for IQ [26]. In this model, a questionnaire is used to evaluate the quality of data. Additional statistical analysis is then used to classify the problem area of data reliability. The aim of PSP /

IQ is to obtain high-quality information based on the attributes of dimensions: intrinsic, descriptive, contextual and accessible [27]. comprises

2.8.4 DQ Management Maturity Model (DQMMM)

The establishment of this model is to improve information structure quality and as the result it would give high quality of information [28]. In this model, structure of incorporated databases being overseen by normalizing its metadata. Standardization of database metadata can be separated into a few phases. For example, intelligent, physical. Other information quality administration model and philosophies referenced before does not oversee information quality during the mix of different databases over the association.

This model focused on the necessities of information mix to upgrade information exactness and consistency. Besides, its capacity to guarantee high quality of information during database incorporation will be an additional worth.

2.8.5 Complete DQ Management (CDQM)

All the previous models and most of the researches were focusing on the structured data type, however this model can deal with structured, semi structured, and unstructured data type. CDQM proposes theoretical, empirical and intuitive approach to check the quality of the data [29]. It has three stages: state reconstruction, assessment and choice of optimal improvement process. As mentioned before, the model is flexible dealing with different kinds of data types, however, it does not have a clear measurement method or a way to calculate the quality dimensions which makes it difficult to apply it in the organization. Table 2.6 is a comparison done by Izham and Fatimah showing the strengths and the weaknesses for the different models [30].

Table 2.6 Strengths and Weaknesses of DQ [30]

Model/Methodology	Strengths	Weaknesses	Data Type
TDQM	Various choice of tool to analyses DQ such as statistical process control, pattern recognition and pareto chart.	Data can be shared among user whereas raw material assigned to a single product. -Timeliness raw material arrived at time. -Believability difficult to compare with physical products.	Structured
IIM	Reassurance phase helps organization to reevaluate DQ after appropriate DQ improvement process.	IIM required DQ policy creation and fulfilment. Thus, it takes more effort for the organization to create DQ policy.	Structured
AIMQ	Measure DQ dimensions in the attributes of intrinsic, representational, contextual and accessibility.	Limited tool to identify information quality problem areas.	Structured
DQMMM	Manage DQ during database integration process.	Suitable only for relational database.	Structured
CDQM	Support structured, unstructured, and semi-structured data type.	Unspecific. No DQ dimensions measurement and calculations defined in CDQM.	Structured, unstructured, and semi-structured

2.9 DIFFERENT TECHNIQUES TO DEAL WITH DATA

Based on the studies and the papers published so far, they followed three processing techniques to assess the data quality, namely, Stream processing, Batch processing, and a Hybrid one. It is basically divided according to how they process the

data in their model. Each type of the processing techniques has outlier detection, evaluation and cleaning phase [31].

Stream processing deals with continuous data, it is a way to turn the big data into fast data. It works with the data online, by feeding the data into an analytic tool in real-time. On the other hand, the batch processing deals with the data in an offline mode. Which the data points are been collected within a specific time interval, after that start to process the data. For example regarding the financial related papers, in [5] they used the stream processing but in [6] they used the batch processing. The model of this thesis is built based on batch processing technique and it focuses on the evaluation phase.

2.9.1 Quality Assessment using batch processing

DQ assessment methods using the batch processing techniques have been divided into two categories: schema and instance based [31]. Both will be explained below.

Instance based technique has been followed by three papers [32][33][34], all followed the same steps. In [32], they developed a model for improving the quality of open data. This model has four steps, the first step is to assess the DQ dimension, secondly the criteria of the DQ are defined. In the third step, the DQ index is calculated by weighing each dimension. Finally, the comments and the measures of the end users are collected in order to evaluate the quality level.

The quality of health data has been assessed in [33]. In this model the first step was to collect the data, and they assessed the quality of the data before and after the pre-processing phase. They choose dimensions like accuracy, completeness and correctness to evaluate before the pre-processing and measured the same dimensions after this phase in order to know the degree of quality improvement. The pre-processing phase is to work on the data before analyzing it, like filtering, transforming and other pre-processing steps. Electrical data has been evaluated in [34] with almost the same steps of [32] but with more dimensions like accuracy, consistency, integrity, redundancy, timeliness, and intelligence.

Schema based technique has been adopted by [35]. An architecture for data quality assessment has been proposed in [35]. They defined two modules that are considered as core of the architecture, the DQ Profiling module and the DQ Assessment

module. Profiling is responsible for measuring the metadata that defines the data source, while the DQ Assessment module is mainly to calculate the DQ dimensions.

The architecture proposed by [35] along with the steps based on their approach is shown in Figure 2.7. The dimensions are being selected based on the data source, as it differs from one application to another and to the interest of the final user. Hence, in the first step the dimensions are determined automatically by the Source Analyzer module. Then an initial profiling to the source is done. In step 3 the Data Quality Service Interface let allows the users to access the DQ service to collect metadata that define the quality level. In the next step the system gathers all the user's settings to build a configuration file which is used for execution of evaluation. And all the preferences are saved in Custom Settings. Finally, as soon as the confidence level is established, the DQ Assessment is performed.

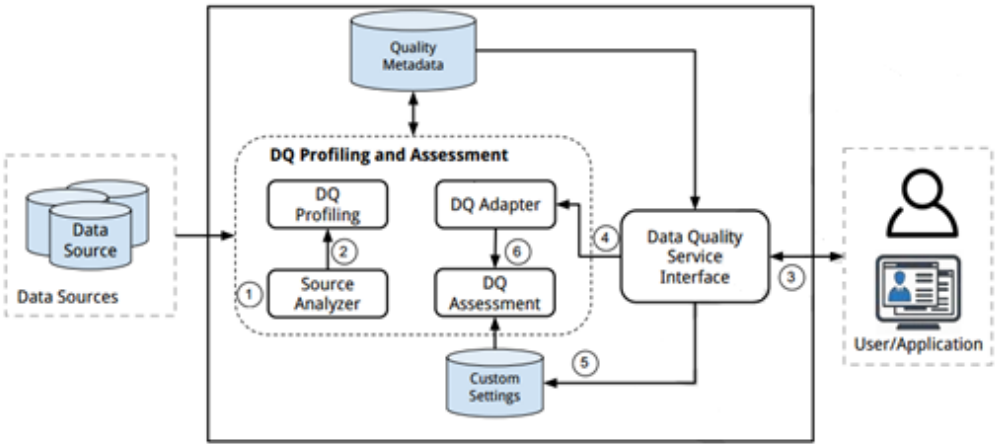

Figure 2.7: Data Quality architecture presented by [35]

Data extraction, data pre-processing, data processing, data analysis, data transformation, and data visualization are the big data management steps and has been followed by [36] in their model. They have tried to assess the data quality by metadata in each of the mentioned steps. They defined metadata as structured information that describes, explains or make it easier to use or manage an information source, and the Quality metadata: describes the quality attribute of the data and the metrics for each quality attribute.

A framework to find the data quality rules (DQR) has been introduced by [37]. The framework components are shown in Figure 2.8. It consists of five components as

follows: big Data sampling and profiling, big data quality mapping and evaluation, big data quality rules discovery, DQR validation and DQR optimization. In the first step, data sampling and profiling are carried out from a huge amount of data. Quality dimensions are defined after profiling and evaluating data characteristics, and analyzed data are evaluated using quality dimensions. Quality mapping is made between DQ dimensions and the targeted data attributes. The quality mapping produces a set of Data Quality Evaluation Scheme (DQES), each element is a quality score for a specific attribute. At the processing stage the DQES is applied on a set of samples, which result in a DQ dimensions quality scores for each attribute. These scores are analyzed against quality requirements. The quality rules are generated, and attributes fully violate these rules might be discarded. Then the rules apply to the data sampled and inspect the changes. If it is necessary to change the rules, they will be changed.

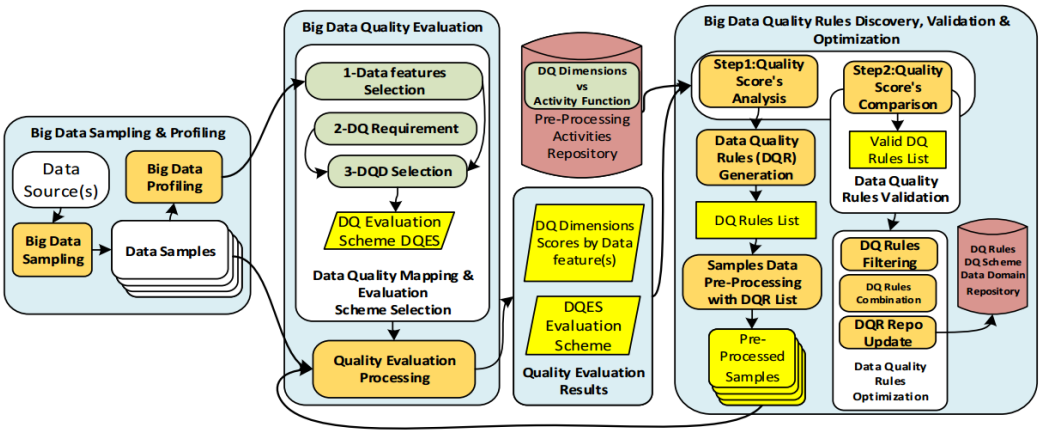

Figure 2.8: Quality Rules Discovery Framework [37]

2.10 MEASURING DATA QUALITY

The previous sections goal was to provide a general knowledge about DQ and its managements, in addition to defining the quality dimensions and their different approaches. Finally, we illustrated various model and different techniques used to assess the DQ. Now, as the general aspects and definitions about DQ have been presented we will focus on measuring DQ applied by various authors and what dimensions they used choose.

2.10.1 Domains presented by the Data quality Literature

Nowadays we are living in a data-driven world, and the big data is becoming more significant, which it rises the importance of DQ. A lot of researches have been done lately on DQ and DQM in various domains. The dispersal of the researches done form the perspective of their application domain is illustrated in Figure 2.9.

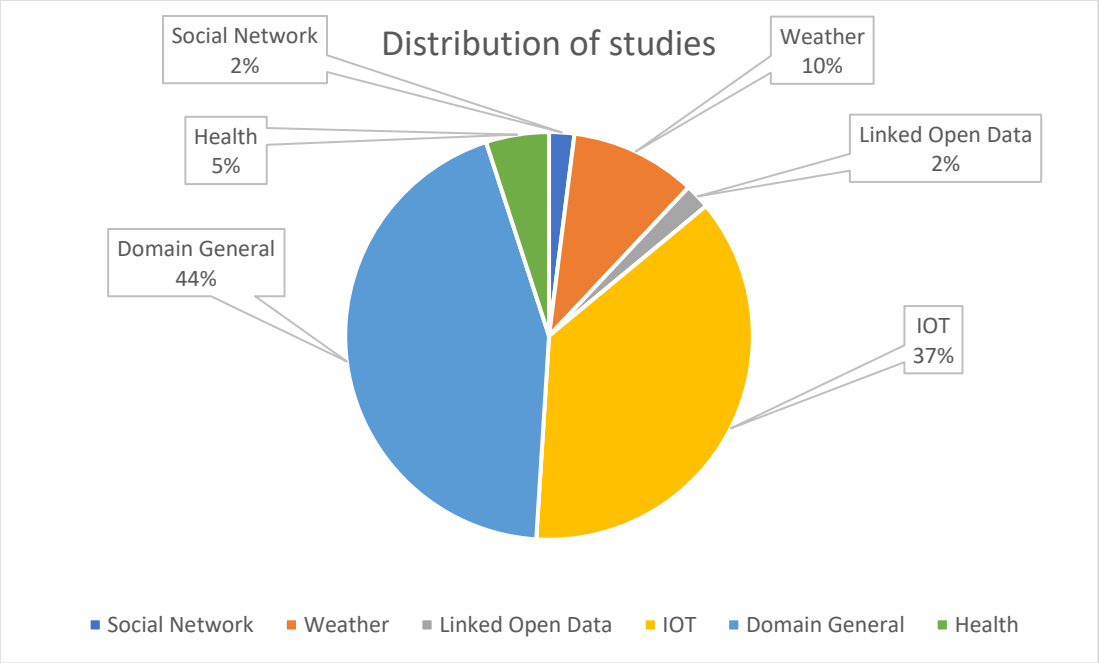

Figure 2.9: The domain of studies [31]

As shown in the Figure 2.9, 44% of the papers were about general domain which there is no specific domain is specified. In another studies, most of the papers were about internet of things (IOT) with 37%; this is because of the massive usage of IOT. The rest of the studies are spread among weather, health, social network, and linked open data, with 10%, 5%, 2%, and 2% respectively [31].

To the best of our knowledge, there are many studies done on DQ in various domains, but not many on the financial domain. Nevertheless, there are some papers that addressed the financial domain by assessing the quality of the relevant sectors like: banking, financial organizations, and insurance companies' datasets [51][54][55][56]. Other papers have addressed the financial domain by evaluating the stock market data [5][6].

2.10.2 Measuring data quality and data quality dimensions

In [39] they claim that defining the metrics is the most important step that should be done to assess the data quality and that not all the methodologies published in the DQ literature highlighted the importance of DQ metrics. They believe that the metrics usually do not take into consideration that data importance differ in the context of helping the company, and that metrics developments strategies are developed for a very specific case and lack the general point of view and sometimes it is difficult to apply as it will be too costly for the company.

They used a weighted criteria method to assess the quality, as if a data unit carries more weight in usage, it should play an important role in measuring the quality. Simple relevancy functions were used to determine the weight of each column, relevancy is defined as the extend of which the data is useful and helpful for the task in hand. Formula (1) [39] is an example of measuring the weight by the frequency of accessing. It could also be calculated using other criteria that suits the company like economic value, recency, and source reputation.

$$Relevancy(columnC) = \frac{Accesses\ to\ Column\ C}{Total\ accesses\ to\ the\ table\ that\ includes\ C} \quad (1)$$

Their application is a case study in an international seed trade company specifically their phone directory data set. The focus was on two dimensions which there are formulas to calculate them, namely, Completeness and Accuracy. The completeness dimension is calculated by a simple ratio between the missing values over the total number of data units then dimension value is calculated with respect of each column weight by formula 2 [39]. In this formula, CW_i is the column weight of the ith column and CC_i is the column completeness of ith column by simple ratio. This weighting criterion can be used in different levels such as cells, columns, tuples, tables and databases. Choosing the level used is totally based on the data and the methodology being used.

$$\sum_{i=1}^{n}(CW_i \times CC_i) \quad (2)$$

For the accuracy dimension there were two formulas one for the character values and other one for numerical values. Formula 3.a [39] is for the character values in which the r_i is the ith value of the tuple t, NED is normalised edit distance and $D(r_i)$ is the closest value in the domain. The function returns 1 if there is an exact match other and the 1-NED otherwise. On the other hand, formula 3.b [39] is for numerical values

in which it returns 1 in there in an exact match otherwise the difference is calculated mathematically and divided by the max of the two values. Both cases are followed by formula 3.c [39] to calculate the accuracy dimension in which $|t|$ is the number of values in the tuple. Once accuracy dimension values are available, weights can be used as in the completeness case.

$$(a)\ \ acc\big(r_i, D(r_i)\big) = \begin{cases} 1, & if\ r_i \in D(r_i) \\ 1 - NED\big(r_i, D(r_i)\big), & Otherwise \end{cases}$$

$$(b) acc\big(r_i, D(r_i)\big) = \begin{cases} 1, & if\ r_i \in D(r_i) \\ 1 - \frac{|\, r_i - D(r_i)\,|}{Max\big(r_i, D(r_i)\big)}, & Otherwise \end{cases} \qquad (3)$$

$$(c) Acc\ [t] = \frac{\sum_{i=1}^{t} acc\big(r_i, D(r_i)\big)}{|t|}$$

They claim that the results they got following this criterion brought objective and subjective measurements closer together. Furthermore, if the organization is aware of the significant dimensions, they can weight each dimension in order to combine them and to reach a comprehensive DQ value for the entire organization as shown in Table 2.7. In this paper they did not proposed a way to calculate the timeliness and consistency dimensions. In their application they just calculated the completeness and accuracy dimension as well as they did not consider a weight for each dimension.

Table 2.7: Dimensions measurements and weights [39]

Dimension	Measurement	Weight
Completeness	0.85	0.3
Accuracy	0.7	0.4
Timeliness	0.75	0.2
Consistency	0.65	0.1

Another approach of calculating data quality was in Research Information System (RIS) domain, which is defined as a central database that can be used to collect, manage and provide information on research activities and research results [40]. In this paper they considered four main dimensions, namely, completeness, correctness, timeliness and consistency because they were discussed widely in scientific publication and they play an essential role in practice. They also provide a general quantification definition for the metrics of a DQ dimension as follows [40]:

$$Rating\ score = 1 - \frac{Number\ of\ unwanted\ results}{number\ of\ all\ results}$$

For the completeness they considered two types of completeness, the value and the tuple completeness as these are what can be fit in the RIS domain. Following the same presented formula, the completeness dimension can be calculated using formula 4 [40]. Bearing in mind the availability of existence of a value that does not exist as discussed in [29] in these cases, it does not consider as incompleteness.

Timeliness is calculated based on how current a data value is. In order to limit the cost of active examination, an estimate approach been followed by defining some parameters like A is a data attribute, w is a suitable data value, $age(w, A)$ in the age of the data value and $decline(A)$ is an empirical ascertained value which describes the decay rate of the data value and the data attribute. Formula 5 [40] was proposed to measure the timeliness dimension for each attribute after that they applied the weighting criterion that was used in [39], except here was either 0 which means "not important" or 1 which means "important".

Correctness dimension has been considered using both syntactic and semantic correctness, formula 6 [40] is used to calculate this dimension. They used the Levenshtein distance to calculate this dimension. Levenshtein distance calculates the minimum number of insertions, deletions, substitutions, and match operations to convert a given string to a second string, as well as transform strings of unequal length or to measure the effort based on the minimum number of these operations. Similarly, the consistency dimension is calculated by the formula 7. In this paper, they managed to use the desired dimensions useful for their application and combine the weighting criterion as well, however, their weighting criterion lack flexibility as there may be cases that the attribute is not important yet it needed to be weighted more than 0. On the other hand, there may exist an attribute that is important but not that important to be weighted by 1.

$$Q_{completeness} = 1 - \frac{Number\ of\ incomplete\ data\ unit}{number\ of\ checked\ \ data\ unit} \quad (4)$$

$$Q_{timeliness}(w, A) = e^{(-decline(A) \times age(w,A))} \quad (5)$$

$$Q_{correctness} = 1 - \frac{Number\ of\ incorrect\ data\ unit}{number\ of\ checked\ \text{data}\ unit} \quad (6)$$

$$Q_{consistency} = 1 - \frac{Number\ of\ inconsistent\ units}{number\ of\ consistency\ checks\ performed} \quad (7)$$

In the industrial domain, where the focus is on the data generated from maintenance management system or warranty databases or data warehouse systems, the purpose is mainly to calculate the reliability [41][42]. The dimensions used here are interpretability, plausibility, timeliness and usefulness [42]. However, in [41] completeness, free-of-error, inconsistency, sample selection and substitution quality were added and used different names for interpretability and usefulness like richness of information. As the goal is to measure the reliability so selecting the sample form the dataset is important, therefore they introduced the sample selection. In which the sample should represent the population from which the sample is collected from [41][42].

Gitzel and Turrin [41] applied two steps in order to assess and improve the DQ of a dataset. The first step they identified the possible data quality dimensions to check if it is important or not. The second step is using a hierarchy approach based on different levels to identify the key problems in the dataset and the most important metrics. The different levels are critical, substitution for critical, subfleet, added value and unspecified. This classification is based on their importance to calculation of the reliability Both steps are discussed in the following paragraphs. They developed a software framework with the aim of calculating the metrics for a specific dataset. All the metrics range from 0 to 100% where the higher the percentage is the higher the quality.

Completeness dimension is typically calculated by checking the empty values or the unknown values, however, they did not specify the probability of misjudge the existence of a missing value. This metric reflects the percentage of properties which are not empty. Free of error is basically the accuracy dimension but they used different name. It has been distinguished between logical, set membership and syntactical error, in which for each rule there is a metric to measure it. The plausibility dimension is more or less like the free of error as it is based on several rules as well but more domain specific, it has some rules that could be defined and added to the inconsistency metric, this comment goes for the free of error metric also. Richness of information metric is also too specific for their domain in which it measures if there are not enough details in the information, they consider the information to lack of richness. Regarding the inconsistency metric, Gitzel and Turrin [41] claim that it has a lesser importance among the metrics, that may be because they just check for the format and the unit of the data

value. However, different unites may affect the results and the decision-making process so it should be as important as the other metrics [42]. The last metric is substitution quality, this is another very specific domain metric. The metric tracks the percentage of which use a certain value instead of its substitutes. Noticed here that many of the metrics considered here could be added to the consistency dimension, as well as most of the metrics were highly tailored to the industrial domain.

Other domains have been introduced in the DQ literature, linked Open Data (LOD) is one of them. In [43], the authors proposed an approach to measure the inherent data quality of LOD datasets. Their approach is a metrics-driven approach that developed based on Goal Question Metric (GQM) approach. This approach starts with defining the set of goals that reflects the management requirements, then these goals are progressively developed into different questions to break down the issue and for each question one or more metrics are associated with it to be measured. GQM was initially proposed in the software engineering field, since then it has been used and applied in a different of other domains.

Initially, the authors of [43] focused on three dimensions accuracy, inconsistency and completeness, the accuracy dimension has been divided into three parts, namely, semantic accuracy, syntactic accuracy and uniqueness. After applying the GQM approach, 20 metrics were proposed for all the dimensions. All the metrics have been derived in quantitative way and based on a ratio scale, most of them have been calculated by calculating the number of undesired outcomes divided by the total outcome then subtracted by 1. Subsequently, 1 will represent the most preferable score and 0 the least preferable one. The questions proposed by the authors as well as the relevant metrics can be shown in Table 2.8 to Table 2.12.

Table 2.8: The metrics for semantic accuracy [43]

Question	Metric
Are the entities described with the correct values?	M1. Ratio of triples contain missing objects
	M2. Ratio of triples with out-of-range objects
	M3. Ratio of triples contain misspelling data value
Do entities accurately represent the real world?	M4. Ratio of entities without correspondent in real-world

Table 2.9:The metrics for syntactic accuracy [43]

Question	Metric
Is the syntax of the RDF documents valid?	M5. Ratio of syntactically incorrect triples
Are the resources described with the appropriate properties?	M6. Ratio of triples with improper assignments of data types to literals
	M7. Ratio of instances using undefined classes/properties
	M8. Ratio of instances being Members of disjoint classes
	M9. Ratio of triples containing improper usage of vocabularies

Table 2.10:The metrics for uniqueness [43]

Question	Metric
What is the degree of redundancy in the context of classes?	M10. Ratio of redundant classes
What is the degree of redundancy in the context of properties?	M11. Ratio of similar properties
Does the dataset contain multiple representations for the same entity?	M12. Ratio of redundant instances
Does the dataset contain redundant values for the properties?	M13. Ratio of functional properties with different values

Table 2.11:The metrics for consistency [43]

Question	Metric
Is there any inconsistency in the schema of the dataset?	M14. Membership of disjoint classes
	M15. Invalid usage of inverse-functional Properties
	M16.Ratio of triples using similar properties
	M17. Heterogeneous data types
What is the degree of conflict in the context of data value?	M18. Inconsistent values of properties

Table 2.12:The metrics for completeness [43]

Question	Metric
Have all the resources been described with adequate number of properties?	M 19. Ratio of properties to class
Is all the required information for each entity presented?	M 20. Missing properties per instance

The authors of [33][44][45] have focused on the medical domain. In[33][45], they identified the quality dimensions intuitively for the general medical data. These dimensions are completeness, consistency, accuracy and timeliness. Relevancy has been used as an alternative for timeliness in [46]. Regarding measuring the data quality of the Electronic Health Record (EHR) domain five dimensions were identified and seven categories of data quality assessment methods [44]. These dimensions are completeness, correctness, concordance, plausibility and currency, for each dimension there are different synonyms used in the literature for measuring the DQ of EHR domain and they merge them to these broader categories. Examples of what the completeness dimensions has been mentioned in other papers are accessibility, accuracy, availability, missingness and validity. For the correctness dimension has been referred to as accuracy, errors, misleading and predictive value quality. On the other hand, the methods used to assess the dimensions of DQ are categorized into seven categories, namely, gold standard, data element agreement, element presence, data source agreement, distribution comparison, validity check and log review.

Based on the number of articles examined in [44], the authors found that the main dimensions used to measure the DQ of EHR are completeness and correctness with 64% and 60% of the articles respectively, following by concordance, plausibility and currency with 17%, 7%, 4% of the articles respectively. Which can also be represented by Figure 2.10, it represents the mapping between the dimensions of DQ and the assessment methods used to measure them. The dimensions are on the left side and the assessment methods are on the right side, both in descending order of frequency from top to bottom. The weight of the edge connecting a dimension and method indicates the relative frequency of that combination.

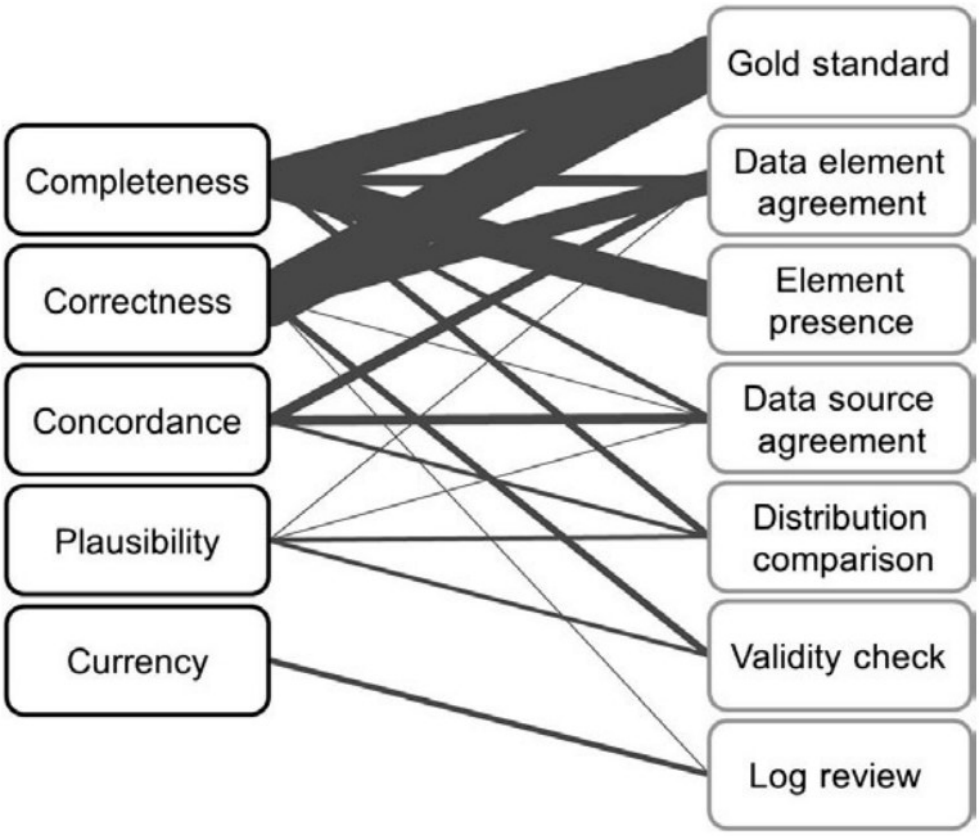

Figure 2.10: Mapping between dimensions of DQ and DQ assessment methods[44]

The authors of [36][46] tried to figure out what are the main dimensions required to be identified to measure the DQ for the social media. In [36], they have followed the four traditional dimensions mentioned in Wang and Strong [17]. While for [46], they categorized all the possible dimensions into five categories as shown in Figure 2.11, namely, syntactic, semantic, user-pragmatic, information-pragmatic and process-pragmatic. Based on the cumulative occurrence of the respective information quality criterion on their sample of Web 2.0 sites (they call it Presence) they choose the most important dimensions. They showed that the most used categories by all the evaluated websites are syntactics category specially consistency, semantics category especially cohesiveness and conciseness, process pragmatics category especially accessibility, latency, response, time Ease of Operation, availability and ease of Navigation. They discovered two new dimensions which are enjoyability and user-conformability, yet they were widely ignored and not used.

The authors of [47][48] have analysed the quality dimensions regarding the product and service quality and the authors of [49] focused on general domains. In

[48][49], they identified the DQ dimensions based on [17] as follows: Completeness, consistency and accuracy to assess the quality of the data. In [49], they identified more dimensions like currency, timeliness, and volatility dimensions as it is more general. In [47], the analysis is consisted of 4 steps in order to identify the proper dimensions shall be used from all the available dimensions in the literature and studies.

The first step they reviewed the latest literature and classified the DQ dimensions respecting some perspectives, and this is because of the diversity of the studies and their contextual nature which makes it hard to construct a unified understanding of the DQ measurements and dimensions. These perspectives are as follows industrial practitioners' perspective, market leaders of DQ management tools perspective, organizations that have developed their own framework to manage DQ perspective and academia with thorough research, while respecting the data quality standards as identified by ISO 8000. [47]

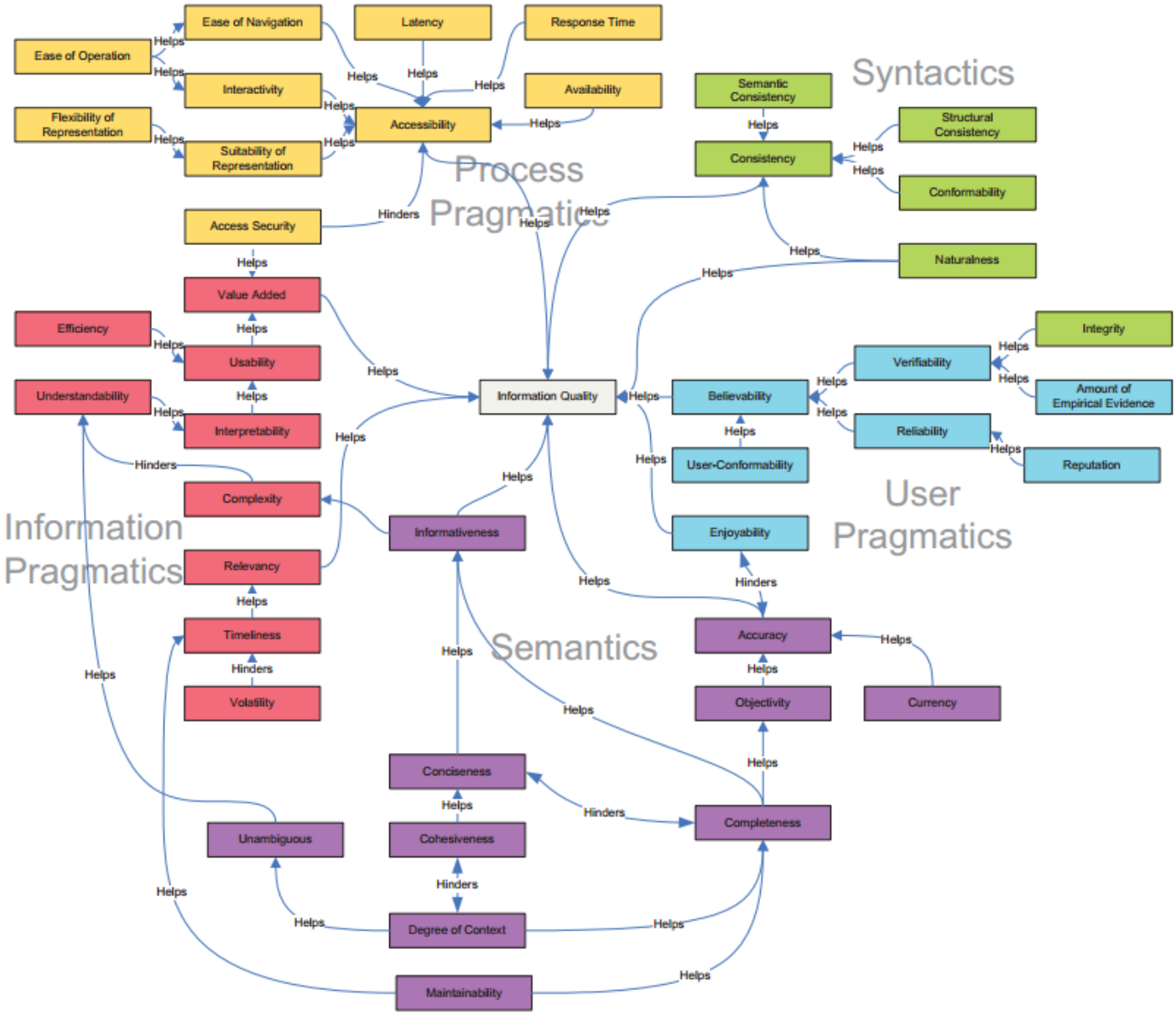

Figure 2.11 Networked Grouping of Information Quality Criteria [46]

In the second step, they uploaded the selected papers into NVIVO, which is a quantitative data analysis tool for analysing rich text-based and multimedia information, in order to differentiate between the dimensions and classify the alike dimensions into categories. Next stage they analyse the definition of each dimension with respect to two perspectives, declarative and usage perspectives. The main purpose of this step is to improve the selected dimensions by removing those that do not respect any of above-mentioned perspectives. Finally, the researchers analyse the output from the previous step and create a classification for the dominant dimensions. [47]

Sixteen sources of dimensions have been selected for this study that revealed around 127 dimensions. Following their model and clustering way, eight main categories were selected as the dominant dimensions, namely, completeness, availability & accessibility, currency, accuracy, validity, usability, reliability & credibility and consistency. [47]

In the previous part the measurement of DQ was discussed in different domains. Identifying the most convenient DQ dimensions helps in business process development and it helps to better assess the DQ [50]. In the literature, selecting the proper DQ dimensions is domain dependent. However, some authors explain the dimensions from a general point of view [15][16][18][23][39]. As mentioned above, some authors proposed a model with metrics and methods to measure the DQ dimensions [39][40][41][43][48]. Two of them introduced the weighting criteria while applying the DQ metrics [39][40]. Others did an analysis to determine what are the most excellent dimensions that should be used to assess the DQ depending on each domain of which they were representing [44][46][47][49] . Most of the papers are considering data to be both numerical and characters.

The summary of the DQ dimensions used in each paper as well as the approach used to identify them as well as the related domain is shown in Table 2.13. The common dimensions used in all the papers are **completeness, accuracy, consistency**. In some cases, they have different names, but the function and the goal are the same. For example, accuracy dimensions is called free of error in [41][48], while in [40][44][45][49] is called correctness. Another example in [23], they introduced three classification of consistency, then highlighted the affected known dimensions. A distinction between semantic and syntactic accuracy has been introduced by some authors [17][40][41][43][46].

Table 2.13: The summary for the DQ dimensions, the approach and the selected domain used for nonfinancial publications

Reference	Domain	Dimensions	Approach
[15]	General	intrinsic (accuracy, lineage, semantic and structural consistency) and contextual (completeness, consistency, currency, timeliness, reasonableness, identifiability)	Theoretical
[16]	General	accuracy, reliability, completeness, timeliness, consistency	Theoretical
[17]	Industry and government.	intrinsic (accuracy, objectivity, believability, reputation), contextual (value-added, relevancy, timeliness, completeness, appropriate amount of data), representational (interpretability, ease of understanding, representational consistency, concise representation), accessibility (accessibility, access security)	Empirical
[18]	General	accuracy, completeness, currency, consistency, appropriateness, interpretability, portability format precision, format flexibility, ability to represent null values, efficient use of memory, representation consistency	Intuitive
[23]	General	contextual consistency, temporal consistency, operational consistency (accuracy, completeness, consistency, credibility, currentness, accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability, availability, portability, recoverability)	Theoretical
[34]	Electrical	completeness, consistency, accuracy, timeliness, redundancy, integrity, intelligence	Empirical
[36][38][48]	Social media, sensors, product and services.	followed the four categories mentioned in [17]	N/A
[39]	General	completeness, accuracy	Intuitive
[40][44][45][49]	RIS, health and general	completeness, correctness, timeliness and consistency	Intuitive
[41]	Industrial	completeness, free-of-error, inconsistency, plausibility, richness of information, sample selection and substitution quality	Intuitive

[42]	Industrial	interpretability, plausibility, timeliness and usefulness	Intuitive
[43]	LOD	completeness, consistency, accuracy, uniqueness	Empirical
[47]	Product and service	completeness, availability & accessibility, currency, accuracy, validity, usability, reliability & credibility and consistency.	Theoretical

2.10.3 Data quality measurement in the financial domain

In this part, we will discuss the proposed DQ assessments in the financial domain. Financial domain has very wide and diverse data types, it could be data from banks, credits rating, financial service organizations, stock markets, insurance companies and many other examples. For each one of the mentioned examples and the others not mentioned has a different way in measuring the DQ, especially in choosing the suitable DQ dimensions used to measure it. This is because of the different data nature of each type. The techniques and the different dimensions used for DQ assessments in some types of the financial domain will be represented in this section.

The main aim of [51] is to validate a specific model for assessing the information quality for the banking industry, more specifically, the public banks. This study has been carried out on the public banks at a federal and state level in brazil. While for [52], the main goal was to identify the most important DQ dimensions in the banking industry as well. The authors of [51][52] have initially adapted a model to measure the information quality proposed by [53] that consist of 15 dimensions and 65 items, it is originally taken from Wang and Strong [17]. They have chosen this model because they are claiming that it has all the dimensions found in the literature. Hence, they proposed an initial research model to start with their study, shown in Figure 2.12.

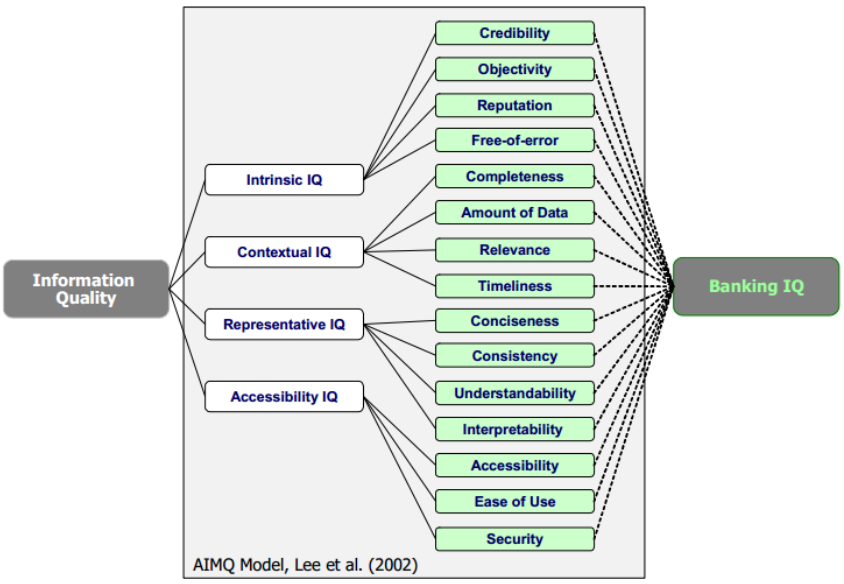

Figure 2.12: The model used in [51]

The steps followed by [51][52] are the same except for [52] they have provided an empirical calculations. The first step is pilot study while the second one is large scale study. The purpose of pilot study stage is to get a preliminary understanding of the measurement's properties of the scales. They conducted this step by sending around 170 questionnaires to executives from the selected banks, and then they start to analyse the replies using the exploratory factorial analysis. The target segments of this study were composed of branch-based management level executives who deal with a huge and intensive amount of data and information. In the second stage they collected more information and applied extra confirmatory techniques in order to develop and examine a first order measurement model and to test the second order measurement model, they have done this initially by sending 200 questionnaires and then by using statistical techniques in order to validate the first and the second order measurements model. This step mainly is to ensure the reliability and the quality of the final instrument selected.

The output model they have reached is composing of four dimensions, namely, accessibility, believability, contextuality and comprehensiveness, distributed into 12 items [51]. Dimensions like security, free of error and ease of use have been removed from the model which is weird as they appear to be important specially for this domain. However, it was claimed by [51] that it was found that such items are not considered significant by the respondents, as they are already taken into consideration in the informational culture of the banks. Some semantic redundancy was found in some of the original model like understandability dimension and interpretability. Therefore, it can be stated that all the dimensions in the resulted model are qualitatively and quantitively different form each other.

Adiska and Joris have presented a financial case study [54] for organizations that could be banks or insurance companies. The main aim of this study is to identify the antecedents of big data quality and the main dimensions that represent big data quality particularly in the financial domain.

The study of [54] starts with collecting data by searching through Google as it is the largest search engine. The search is converged to 10 largest banks and 10 largest insurance companies in Europe, and seven articles are observed as addressing relevant data which is in an adequate level of details regarding big data projects in financial industry. Then, they selected three case study organizations where each of these organization must meet the following three criteria:

1) Should be a bank or an insurance company.

2) Has been providing financial services for at least five years.

3) Has taken any big data initiatives and above all of these is willing to participate in the study.

A list of questions has been asked to the three-case study organizations to identify the big data quality issue. Summary of their cases is listed in the following Table 2.14, the institution names are not mentioned for confidentiality reasons. Afterwards, an analysis to pinpoint the antecedents of DQ is carried out.

Table 2.14: Three case studies in [54]

Institution	Information output	Information quality goal
Bank X	Package of mortgage files	Accuracy, Completeness, Currency, Consistency, Timeliness, Uniqueness, Validity, Traceability
Bank Y	Credit risk level, most suitable loan for customer	Believability, Comprehensiveness, Relevancy, Validity
Insurance Z	A single customer 360 profile	Uniqueness, Accuracy, Completeness, Currency, Timeliness, Validity Comprehensiveness

The results of this study are identifying 10 big data quality antecedents that can be grouped into five categories, namely, data, technology, people, organization, and external environment. However, it can be shown that not all of them are related to DQ and 11 essential dimensions of the big data quality have been identified, namely, accuracy, believability, relevancy, currency, completeness, comprehensiveness, consistency, uniqueness, timeliness, validity, and traceability. It has also been observed that not all the dimensions are important in all the different projects. Thus, the dimensions can be context independent or context dependent. [54]

Karel and Bart [55] developed a study that has been performed in order to manage the DQ, specifically to detect the DQ issues. This structure approach has been applied to the credit rating process of a company in the financial sector, more precisely, a large financial institution in Belgium (Europe).

Based on the literature review made by [55], they found five DQ dimensions: accuracy, comprehensibility, consistency, completeness, and time. Others choose just correctness and currency as in [56], which also examining the DQ in financial domain. Exploiting the knowledge of the domain experts, they made improvements on the aforementioned dimensions. Thus, they have proposed what they call "The data quality axes". The different axes can be shown in Figure 2.13.

The accuracy is described by syntactic and semantic accuracy. The definition of these two types have been illustrated before. Comprehensibility dimension is measuring to what extent the end user can understand the data. In the consistency dimension a distinguish between interrelational and intrarelational consistency has been made based on the nature of the constraints. Completeness dimension is the existence of missing values as known in most of the papers. They considered the uniqueness dimension as a supplementary dimension to completeness by assuring there are no doubles in the dataset. However, others consider it as a separate dimension [43].

Figure 2.13: The data quality axes [55]

Regarding the time dimension there are three sub dimensions, namely, volatility, timeliness and currency. Volatility sub dimension is for describing how the data is changing over time. Timeliness sub dimension shows how recent is the data in relation to their usage, it mainly depends on the selected domain. Finally, the currency sub dimension that focuses on how the data is immediately updated when a change happens. For the security dimension is of a vital importance for the financial organizations and this is because of the privacy and safety regulations. [55]

The methodology used in this study [55] to manage the DQ is consisted of smaller steps by adopting an ETL (Extract Transfer Load) approach wherein the process stages are identified where data is created or extracted from the source, where data is manipulated or transferred and last stage where data stored or loaded into another database. This approach has been followed as they believe that these parts of the process are considered the most critical parts where the DQ issues can be created. This is a generic view of the process of the data which is not related to the financial domain.

Dasu and Duan [5] developed a general framework for data quality assessment of temporal data with a dynamic behaviour. The temporal data they examined is stock market data. They refer to the data quality issue as “glitch” or “data quality anomalies”. Their framework has been applied on a commercially available stream of the NYSE (New York Stock Exchange) stock prices by collecting every minute over a single year starting from November 2011 to November 2012.

The data is structured and has seven attributes, each record in the dataset represents a different trading minute. It is important to highlight that if there is no trade, no record is generated for that stock. The attributes of the data are trading time, day of the month opening price, highest price, lowest price, closing price, trading volume. Which are the typical attributes you can get from the stock quotes. The highest and lowest prices are referring to the price observed during the relevant trading minute. The trading time which is also the timestamp, has the date and time of the data. The stock symbol is embedded in the file name. The stock symbol and the timestamp are used as a unique key. A sample of the data as follows:[5]

timestamp, Trading-Day, open, high, low, close, volume

2011-11-01 09:38, 1,21.75, 21.78, 21.75, 21.76, 1200

2011-11-01 09:39,1, 21.74, 21.75, 21.73, 21.74, 1481

2011-11-01 09:40, 1,21.75, 21.77, 21.74, 21.76, 6675

Dasu and Duan used four constraints to assess the quality of the data, which are basically rules the data should follow. In these rules all the dimensions will be considered like the completeness, accuracy, consistency and timeliness mainly. Same as [23], they defined three main categories for consistency that affects other dimensions. The four types of constraints are represented in Table 2.15. These types have been classified based on their dependence on the columns (attributes) and the rows (entities). Type 1 constraint: requires access to a single cell in the dataset as it depends just on one column and one row. For instance, this constraint could be checking if a single value is an integer or not. It is inexpensive and easy to be applied. Type 2 constraint: in this case other attributes are being considered for the same row. In which, two cells are being check for a specific rule for example if one value should be greater or less than the other one. Type 3 constraint focus is on one attribute across multiple rows. This is mainly to compare between the rows, especially of this column is a key attribute and all its value should be unique. The case of involving multiple rows as well as multiple columns has been called Type 4 constraint. [5]

Table 2.15: Types of constraints proposed by [5]

Type	Single column	Multi column
Single row	Type 1: Applies to a single column and single row	Type 2: Depends on multiple columns but only row
Multi row	Type 3: Applies to a single column but depends on multiple rows	Type 4: Depends on multiple columns and multiple rows

The authors are using the statistical distortion as a data quality metric, in which it measures the difference the perfect value that we expect to receive at time t, and the actual data that we observe at the same time t, so mainly the statistical distortion calculates the distance D between them by the following equation, where D_t^I is the ideal value and D_t is the current data:

$$SD(D_t) = D(D_t^I, D_t)$$

The framework followed by authors of [5] has some primary steps. The first stage is to identify and construct the rules that the data should meet, with some margin of error, in order to maintain the high quality. They call this the ideal D_t^I at time t. The

following step is to compare the actual data D_t with the ideal one for the unjustified violations and highlight them as potential glitches. Because sometimes these violations have some explanations which make it normal to happen, in that case these explanations could be used in order not to be flagged again. Third step is to check the deviation of the observed data from the actual one using the statistical distortion. Finally, used the statistical distortion as a data metric to continuously monitor the observed data.

They followed two approaches to detect the quality issues. The first one is Thomson Reuters (TR) that for example consider the pervious value is the ground truth which in this case they call it ideal D_t^I, this approach is used to measure the data quality. The other one is the Feed Inspection Tool (FIT) which is being used for monitoring streaming data. FIT calculates various of statistical summaries like mean, quantiles and counts and based on historical data in a sliding window to empirically estimate the ground truth. [5]

Dasu and Duan examined the completeness dimension and the timeliness in the data gathering phase and included them to the type 4 constraint since they use multiple stocks and multiple times. Then, the four types of constraints have been calculated. For instance, type 1 constraint they have defined two rules and the violations of these rules denoted by $g_1^1(t)$, $g_2^1(t)$ where the superscript refer to constraint type 1, and the subscript indexes the rules as 1 and 2 respectively. Then it can be aggregated for all the stocks across NYSE to get the total for of them regarding this type. The following equation is an example of calculating the type 1 constraint for rule number 1: [5]

$$G_1^1(t) = \sum_{NYSE} g_1^1(t)$$

Similarly, type 2 constraint can follow the same concept and equation of type 1. The rules specified in type 1 and 2 implies that they are measuring the consistency dimension. However, for type 3 constraint it will be more complex because of the more interaction and interrelationships in the data. Even though that it is not mentioned here explicitly, the accuracy dimension is measured in this type. They check for all the data attributes which are high, low, open, close price and volume. Three more derived variables have been introduced that depends on the original attributes. These variables are the lag between successive trades (LG), the price spread (SP) and consecutive changes (DL). These variables are expected to fall within the statistical range they have

specified. Finally, they defined the statistical distortion equivalent to a specific stock *S* to be the proportion of glitches of all types in all the stream. [5]

It can be concluded that Dasu and Duan [5] focused on four dimensions, namely, completeness, timeliness, consistency, and accuracy. However, they do not refer to them explicitly but included them within the types of constraints.

Another study has been made on the stock market domain along with the flight domain, which the authors classify them as Deep Web [6]. Deep Web data are the data stored in underlying databases and queried using Web forms. Their study initially starts with data collecting stage to collect the data from different sources.

The data collected by the authors of [6] were for 1,000 stocks, including the 30 symbols from Dow Jones Index and the 100 symbols form NASDAQ Index. The data were collected after the stock market closes by one hour on each day and it was for just one month. They collected the data from 55 sources, which provide attributes ranges from 3 to 71. Some of the attributes have the same semantic but with different name. After considering the popularity of the attributes and its stability they chosen 16 attributes shown in Table 2.16.

Table 2.16: List of the attributes examined in[6]

Last price	Market Cap	Dividend	EPS
Open price	Volume	Yield	P/E
Today's change (%)	Today's high price	52-week high price	Shares
Today's change ($)	Today's low price	52-week low price	outstanding
			Previous close

Four questions must be answered to assess the sources quality from the authors point of view [6], it is like what have been followed in [46]. The empirical approach has been adopted also by [51][54][55] but as a survey approach sent to the experts. These questions are as follows:

1) Are there many of redundant data on the web?
2) Are the data consistent?
3) What is the level of data accuracy in each source?

4) Are the sources copping from each other? The answer of each question of those leads to a dimension or metric to measure the source quality.

Regarding the first question, the redundancy is being checked. The redundancy is the percentage of sources that provide a certain object. For example, how many sources are covering the data of a certain stock. The second question is led to calculate the consistency dimension that they measure it by the entropy proposed by [57]. The higher the entropy the higher is the inconsistency. They also measured the deviation of the value from the true value and consider it as consistency. [6]

The accuracy dimension was the answer for the third question, which they calculate it as the percentage of the source provided true values among all its data items. The ground truth here in this case they call it the "dominant value". They get it by the voting results from the five popular financial websites, they vote only on the data item provided by at least three sources. They admit that this is a challenging and subjective to risk as they must trust some particular sources. In addition to that, they calculated dominance factor which is the percentage of sources that provide the true value among all providers of a specific value and precision of dominant values which is the percentage of data items on which the dominant value is true. However, the later two calculation they include it under the consistency part which can create a confusion as discussed in [47]. Finally, the potential copying has been evaluated by measuring the average accuracy and schema commonality. In which the schema commonality is measured as the average Jaccard similarity between the sets of provided attributes on each pair of sources. With the same concept they measure the Object and Value commonality. [6]

The authors of [6] have identified three main dimensions by answering the questions they proposed. These dimensions are redundancy, consistency, and accuracy. Noticed here that they did not consider the completeness dimension or asked a question regarding if the dataset is complete or having missing values. The only thing that could cover it partially is the redundancy but still not considered as assessment for completeness.

For both studies [5][6] despite of their different approaches used to assess the quality of the sources. Errors were found in the inconsistency dimension and the accuracy dimension.

The summary of the DQ dimensions used and the approaches followed by the financial domain publications are shown in Table 2.17. Noticed that an empirical approach has been followed by all of them [6][51][52][54][55] to identify the DQ dimensions. As all of them have conducted a questionnaire for the experts and the data users in order to identify the DQ dimensions and develop metrics for them. Most of them focused on the same dimensions as the other domains did, which they are **completeness, consistency, and accuracy.**

As will be discussed in the next chapter, the main goal of the thesis is to assess the DQ of different stock market data sources and rank them to identify the most reliable one. The selected source could be used in any downstream application; hence, a high quality is required. DQ issues found the selected source were mainly in the completeness dimension, showing the need of filling the missing values before using it. Therefore, a predictive model is introduced to estimate the missing days in the dataset to fill it. In the upcoming section, the different techniques for data predictions are discussed, along with the utilization of deep learning in data prediction.

Table 2.17: The summary for the DQ dimensions and the approach used for financial publications

Reference	Domain	Dimensions	Approach
[51]	Financial	initially followed the four categories mentioned in [17] and then choose accessibility, believability, contextuality and comprehensiveness	Empirical
[52]	Financial	followed the four categories mentioned in [17] and added these dimensions alignment, actionability and traceability	Empirical
[54]	Financial	accuracy, believability, relevancy, currency, completeness, comprehensiveness, consistency, uniqueness, timeliness, validity, and traceability.	Empirical
[55]	Financial	accuracy, comprehensibility, consistency, completeness, and time	Empirical
[56]	Financial	correctness and currency	Empirical
[5]	Financial	consistency (which affects different dimensions)	Empirical
[6]	Financial	redundancy, consistency, accuracy	Empirical

2.11 DEEP LEARNING IN STOCK MARKET DATA FORECAST

Stock time series forecast depends on the analysis of time series data and the ability to identify patterns, trends and periods existing in the data. The existing methods for stock price forecasting can be classified as follows [58]:

- Fundamental Analysis
- Technical Analysis
- Time Series Forecasting

Fundamental Analysis focuses on the company's earnings, revenues, sales and other economic factors to estimate the company's share value. This method is mostly convenient for long term forecasting. Technical analysis technique identifies the future price based on the historical stocks price and it is most suitable for short term predictions. Finally, time series forecasting focuses on the sole analysis of time series data. It can be divided into two main classes: linear and nonlinear models.

Autoregressive Moving Average (ARMA) [59] and Autoregressive Integrated Moving Average (ARIMA) [60] methods are the conventional statistical methods implemented in the linear model's class. However, the main disadvantages of these models are that, they typically assume that the linearity of the stock time series process and model the generation process for a latent time series to predict future stock prices. Moreover, they are unable to identify the interdependencies among the various stocks. Thus, these methods are not convenient for a dynamic nonlinear process such as stock time series prediction.

Non-linear models involve methods like autoregressive conditional heteroscedastic (ARCH), Generalized Autoregressive Conditional Heteroscedasticity (GARCH) [60], Threshold Autoregression (TAR) and Deep learning algorithms [61]. Deep learning models can be considered as non-linear function that has the ability to deal with data that is non-linear, non-parametric, chaotic or discontinuous for a stock time series. Many studies use deep learning techniques to predict financial time series such as Multi-Layer Perceptron (MLP), Recursive Neural Networks (RNNs), LSTM and CNN (Convolutional Neural Network) [62]. For example, Ding et al. used a deep convolutional neural network to forecast the effect of events on stock price movements [63]. Additionally, Baek et al. proposed a new forecasting framework for stock market

using LSTM modules [64]. This work explores the ability to utilize deep learning techniques specifically LSTMs in stock market forecasting.

RNNs are connectionist models with the ability to selectively transfer information through sequence steps, while processing sequential data one element at a time. Thus, they can model input and/or output consisting of sequences of non-independent elements. The reason why RNN is called recurrent neural network is that a sequence's current output is also related to the previous output. Particularly, the network stores the previous information and applies it to the current output calculation. In other words, the nodes between the hidden layers are no longer connect-less and the input of the hidden layer includes not only the output of the input layer but also the output of the hidden layer at the last moment

In practice, in order to reduce the complexity, it is often assumed that the current state is only related to the previous few states. The disadvantage of this algorithm is that, as time goes by and the number of network layers increases, problems such as gradient vanishing or gradient explosion can be caused. This makes the traditional RNN faces difficulties with long-term dependencies. Moreover, the architecture uses the same transition function with the same parameters at every time step and the learned model always has the same input size. In order to overcome the aforementioned shortcomings of the RNN, LSTM was introduced in 1997 by Hochreiter and Schmidhuber [65]. It is a kind of recurrent neural network that uses the accumulated linear form in processing the information of the sequence data to avoid the problem of gradient vanishing and to learn long-period information. Furthermore, it can learn long-term and short-term time-dependent information. Thus, it is a commonly used deep learning model for processing time series data. Regarding the LSTM architecture, the usual hidden layers are replaced with LSTM cells. The cells are composed of various gates that can control the input flow. An LSTM cell consists of input gate, cell state, forget gate, and output gate. The various gates and their functions are as follows:

- Input gate: Input gate consists of the input.
- Cell State: Runs through the entire network and has the ability to add or remove information with the help of gates.
- Forget gate layer: Decides the fraction of the information to be allowed.
- Output gate: It consists of the output generated by the LSTM.

The cell state is updated based on the outputs form the gates. Mathematically we can represent it using the following equations.

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \quad (8)$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \quad (9)$$

$$c_t = tanh(W_c.[h_{t-1}, x_t] + b_c) \quad (10)$$

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \quad (11)$$

$$h_t = o_t * \tanh(c_t) \quad (12)$$

where x_t: input vector, h_t: output vector, c_t: cell state vector, f_t: forget gate vector, i_t: input gate vector, o_t: output gate vector, σ : sigmoid activation function and W, b are the parameter matrix and vector. A diagram for both RNN and LSTM architectures is shown in Figure 2.14.

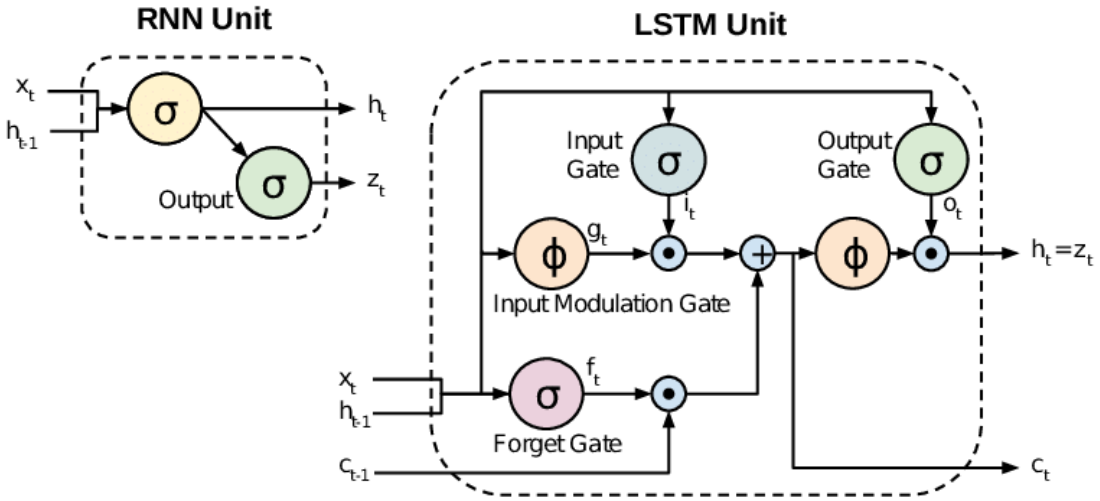

Figure 2.14: RNN simple cell versus LSTM cell [66]

Chapter 3: Methodology

The focus of the model is on the stock market data, which include the Market Cap, market prices and the volume of trades. The stock market is mainly where the stocks are bought and sold. The higher the quality is, the more the data will be useful and help on taking a proper decision. The stock market is the most effective channel for the company to raise its capital [67]. The investors and traders are interested in the stock because dividends, long-term growth of capital and to protect against the inflationary destruction of purchasing power [68]. Therefore, stock market data affects the investors decisions and the company decisions [69][70]. The stock market data is used in the literature for several purposes, like evaluating the chief executive officer and the company performance [71], the merges [72], the Market Cap influence on the real estate investment trust [73] and the effect of the foreign investors in the market [74].

The usage of stock market data is not only in real-time to make fast decisions about stocks trading, however, historical data can be used to project pricing trends, to evaluate the efficiency of the market [75] and to develop a predictive model to predict the stock price [76]. Based on the aforementioned, stock market data should be reliable and free of glitches in order to be useful for any application. Thus, the main goal of this model is to assess the DQ of different sources in order to rank them and to identify the most reliable source that could be used in various application. In order to reach this goal, we had to identify the proper DQ dimensions. Since there are many sources that provide the stock market data and it should be a trustworthy data, and there are many dimensions mentioned in the literature. The focus of the applied part of model is offline, using the stock market historical data and the batch processing technique [31]. While for second part it should be online, but it was not applied in this thesis.

3.1 THE IMPLEMENTED MODEL

In this chapter we discuss our model used to evaluate the DQ of different sources. There is an assumption that the examined data are numerical values and the ground truth is a single true value. The ground truth is the dataset provided from the NASDAQ official website [77]. We use ground truth and true value as substitutes. Other assumptions will be mentioned throughout the chapter when necessary. This

methodology will be applied on the stock market data in which we can evaluate different sources and define a rating for each source.

Our data is collected from different sources for companies in NASDAQ stock market. It has seven attributes, namely, symbol, timestamp, open price, low price, high price, close price and volume. “Symbol” and “timestamp” are the key attributes of the dataset. Symbol and company are used as substitutes. “Open price” is the price of the stock at the very beginning of that trading day but opening price does not need to be equal to the previous day’s closing price. while the “close price” is the price of the stock at closing time of that trading day. “Low price” and “high price” are the lowest and highest price the stock reached during the trading day. Finally, the “volume” is the number of the stocks that were traded during the trading day for the given stock. The examined data in [6] was one month in a specific year, however, we believe that the period should be longer. Because short periods could lead to misjudgments on data. Hence, we collected and examined the stock price data for the period of 10 months from January 2019 to October 2019. The year is not complete because the data was collected in November 2019.

The model is composed of two parts (a) and (b) as shown in Figure 3.1. The first part has four phases, namely, DQ basics phase, data preparation phase, DQ assessment phase, and sources evaluation phase. In DQ basics phase, the DQ rules, dimensions, and metrics are being identified. Concluded from the literature that defining the DQ dimensions and rules is domain dependent. While in the data preparation phase, the data is collected, and ground truth is identified. Data preprocessing step is very important before starting the following phases, as the data schema and structure are different from one source to another. Adjusting the schema and structure of the sources before the assessment phase will save time and effort. Next phase is DQ assessment phase, where the DQ of each source is examined by the pre-defined dimensions and visualized then stored to be used in the next phase. Finally, sources evaluation phase, where a comparison between the source is conducted and the final results are stored and presented.

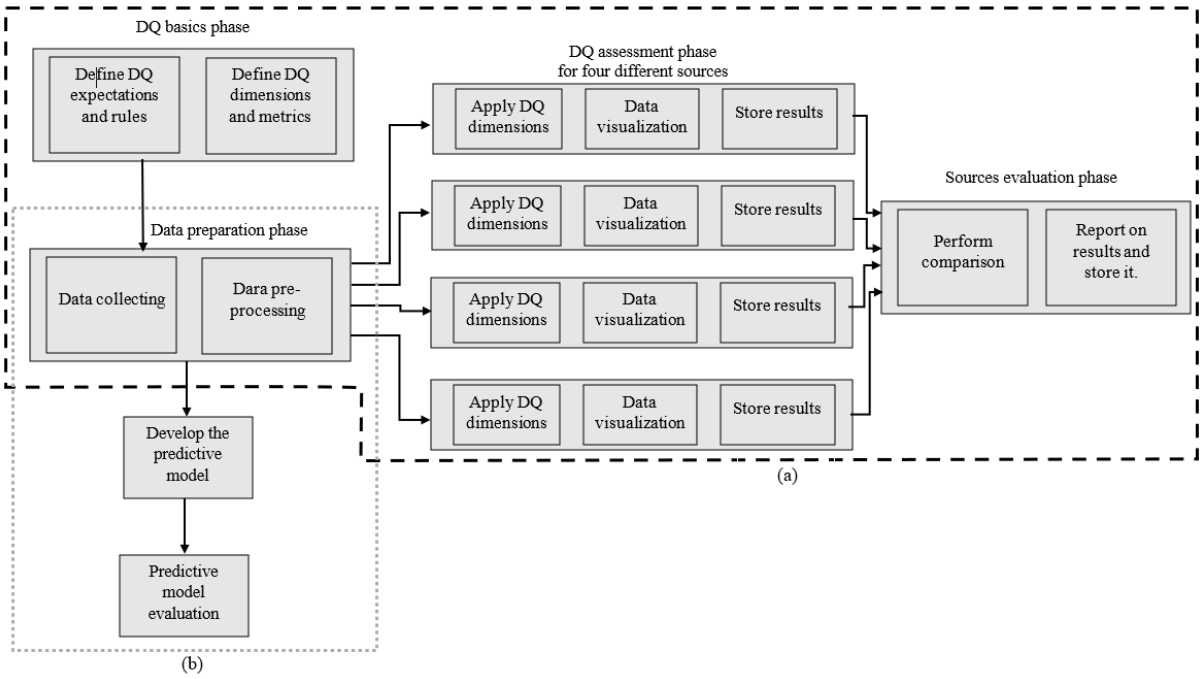

Figure 3.1: (a) Sources assessment model, (b) Predictive model

The main goal of the second part (b) is to predict the stock price in a specific day. The predictive model output will be used fill the missing values in two cases, missing days in the selected source and missing days in NASDAQ dataset. It will be shown in the results chapter, that for a specific source, the only problem found was the missing values. The predictive model is used to fill in these values before using the dataset, in order to be more reliable when using in any application. Although it is not common to find missing data collected form the official website of NASDAQ, missing values could still occur during rare exceptional events. For example, trying to get the data in the end of the current day. In this case, the predictive model could be used to fill in this missing day.

The first phase in the second part (b) is the same as the first phase in part (a), which is data preparation phase. In this phase, the input data of the predictive model is collected and pre-processed to be ready to use. Good preparation of the data will lead to more accurate results. The second phase is developing the predictive model. Finally, the model is evaluated.

3.1.1 DQ basics phase

The most important part in DQ assessment is to identify the rules the data should follow and the dimensions used to measure the quality [39]. In this phase the rules that our dataset should follow will be presented and the relevant dimensions and metrics to assess the quality.

Following the same concept of questionnaire followed by [6][17][43], we identified the rules that the dataset should meet. These rules are shown in Table 3.1 along with the affected dimension.

Table 3.1: The rules the dataset should meet and the affected dimension

Rules	Dimension
Data item should be equal to the true value	**Accuracy**
There should be no missing values in the dataset and all the symbols are represented.	**Completeness**
No duplication in the key attribute (Timestamp and symbol)	**Uniqueness**
The data item should not be negative value	**Consistency**
Close price should be within the high price and low price	**Consistency**
If the close price is reported the volume value should not be 0	**consistency**

DQ dimensions and metrics

Based on our research, understanding of the literature, the rules identified and the nature of our dataset, we choose three main dimensions to evaluate the sources in our case study. These dimensions are **completeness, consistency, and accuracy**, which they are commonly used across the publications in different domains.

The timeliness dimension is very important in our domain, but if we are evaluating the sources in real-time. In our case we are following the batching processing technique [31] that the timeliness is not considered. Regarding the rule that affect the uniqueness dimension mentioned in Table 3.1, it is covered in the consistency dimension. According to the literature in the financial domain, they added a dimension which is the comprehensibility dimension [54][55]. Which is the extent of understandability of the dataset, however, in our case we are nor dealing with complex

financial data that is hard to understand. Thus, we believe it is not relevant to consider this dimension.

It is important to mention that the same dimension could be presented by different authors, but in different names. Like the accuracy dimension has been mentioned by [48][51] as free of error and by [44] as correctness. Also some dimensions with the same name could be interpreted in different way, like uniqueness dimension has been considered as accuracy dimension in [43], while as completeness dimension in [54]. Hence, Some publications categorized all the dimensions is small groups from their point of view [44][46][47].

Completeness dimension is to measure the percentage of missing values in the dataset. So basically, it means that every data item should have a value, otherwise it will be considered as a glitch. However, in some cases it is normal to have data items without any values. For example, if there are no trades occurred in the day, the trading volume attribute is expected to be null with no closing price as well. It could be a rare case, as the stock market is always active, but it should be taken into consideration. Thus, while examining the completeness dimension this issue should be respected and not be mistaken as a glitch. The completeness dimension has been widely mentioned in all the publications expect [6]. This reflects its importance in assessing DQ, and it has been identified as the most important dimension by [44]. It has been calculated using ratios in [39][41], by dividing the number of missing values by the total number of values. Others used the same ratio but subtracted by 1 [40][43]. Considering that we want the rank to be 1 is the best and 0 is the worst, we will calculate the ratio and subtract it by 1. This will be followed in all the dimensions.

We denoted the number of data items that violates the completeness dimension per column by M_{col} where col is the selected attribute to evaluate. The dimension can be calculated using the equation (1) [40], where N_{row} is the number of rows in the dataset and N_{column} is the number of columns. It should be positive values and ranges from 0 to 1, where 1 means that the dataset is complete and no glitches for this dimension is existing, while 0 is the worst case that all the data items are missing.

$$Completeness = 1 - \left(\frac{1}{N_{row} * N_{column}} \sum_{col=1}^{N_{column}} M_{col}\right) \qquad (1)$$

Consistency dimension main objective is to detect the violation of semantic rules across the data items and to make sure that the constraints are not violated. According to this definition it can also be named Integrity Constraints (ICs). ICs are the rules that the dataset must meet, it differs from one domain to another. Hence, it is recommended to consult an expert of the domain in setting these rules and in improving them. Involving the domain experts is important as it could happen that a data item violates one constraint however not because of a quality issue but instead as a result of inadequately specified constraints. For instance, in the telecommunication domain, call volumes are higher in specific holidays like Mother's Day. Thus, this exception should be reflected on any constraints set by the user on the call volumes attribute.

Inspired from [5], we choose three types of constraints, namely, type 1, type 2 and type 3. However, the calculation has been followed the same criteria of [40]. Type 1 constraints focus on single row and column, e.g. single data item value. A constraint like "data item must be a positive number" can be checked by evaluating a single value without any extra involvement of other rows or columns. This type of constraints does not consume too much time or money since they access just one single data item. We denote the violation of the types by T_x^{col} for each column, where col is the selected column to check and x is the number of the type. The type rules are being referred to by C_x and their number by N_{rules}^x. Each rule has its own formula to be calculated as shown in equation (2). After checking all the rules, type 1 constraint can be calculated using the equation (5) in which all the outputs of the calculated rules are multiplied together.

Type 2 constraints checks different data items across attributes for the same row. As an example, in the stock market data, the "close price" attribute should lay between the "low price" and "high price" attributes. In this case we need to check single row for two different columns to validate the selected data item value. The rules of this type can be measured by the equation (3), then the type will be measured by the equation (5).

Type 3 constraints validate the data item value across rows for the same column. For example, if the first column is a key so all its values must be different and not duplicated. Type 3 rules will be calculated using the equation (4) and similarly, it will

be measured by equation (5). Finally, the consistency dimension can be measured using equation (6), by multiply all the types percentages calculated from equation (5). It is always positive, and it ranges from 0 to 1, where 1 means that the dataset has no glitches regarding consistency and 0 means that all the data items in the dataset are violating the consistency dimension. In Table 3.2, the list of constraints that should be checked in each type is presented.

$$Type\ 1_{C_1} = 1 - \left(\frac{1}{N_{row} * N_{column}} \sum_{col=1}^{N_{column}} T_1^{col}\right) \quad (2)$$

$$Type\ 2_{C_2} = 1 - \left(\frac{1}{N_{row}} \sum_{col=1}^{N_{column}} T_2^{col}\right) \quad (3)$$

$$Type\ 3_{C_3} = 1 - \left(\frac{1}{N_{row} * N_{column}} \sum_{col=1}^{N_{column}} T_3^{col}\right) \quad (4)$$

$$Type\ x = \prod_{C_x=1}^{N_{rules}^{x}} Type\ x_{C_x} \quad (5)$$

$$Consistency = Type\ 1 * Type\ 2 * Type\ 3 \quad (6)$$

Table 3.2: The list of constraints to be calculated in each constraint type

Constraint Type	Rules
Type 1	• All fields should not be negative value • All fields should not be missing (unless there is an explanation)
Type 2	• "Close price" value should be within the "High price" and "Low price" • If "Close price" is reported, the "Volume" should not be 0 • If "Close price: is not reported, "Open price" should be equal "Low price" and "High price"
Type 3	• "symbol" and "Timestamp" attributes should be unique and not repeated.

Accuracy dimension is to check how close the data item value is to the true value. We assume that the ground truth is already known (single value) and has numerical values, in our case it will be the data collected from the official website of NASDAQ as mentioned before. There is a distinction between semantic and syntactic

accuracy introduced in the literature [17][40][41][43][46][55]. However, in our case it is not required to make this distinguish because all our dataset has numerical values. Thus, two aspects have been taken into consideration while calculating the accuracy dimension, namely, the error count and the deviation from the true value. The error count part can be calculated by the equation (7) [41], where Acc^{col} is how we denote the violation of the accuracy dimension. It is considered violation if the data value is deviated from the ground truth by 0.1 and above, this value is chosen based on the domain and the sensitivity of the prices in stock market. The output is non-negative and ranges from 0 to 1 as the other dimensions.

$$Accuracy_{count} = 1 - \left(\frac{1}{N_{row} * N_{column}} \sum_{col=1}^{N_{column}} Acc^{col}\right) \quad (7)$$

In order to consider the deviation from the true value, Mean Square Error (MSE) is calculated. MSE measures the average squares of the errors, where the error is the difference between the source value and the actual value. It is always positive and the closer to zero is the better. MSE can be calculated by the equation (8), where n is the number of rows, v_i is the true value, and v_i' is the data item value. We now can calculate the deviation by equation (9), where NMSE is the normalized Mean Squared Error. NMSE is calculated by dividing the MSE by the maximum error in the dataset. In case there are no errors in the dataset with 1 $Accuracy_{count}$ and 0 MSE, there is no need to calculate $Accuracy_{deviation}$. The accuracy dimension is now can be calculated by equation (10). The output is non-negative and ranges from 0 to 1, 0 is the worst and 1 is the best.

Weighting criteria has been adopted by [39][40][52] to evaluate the DQ, as they weight the importance of each dimension to calculate its contribution in the final evaluation. However, concerning our data and its sensitivity, we believe that the dimensions we choose have equal importance. Hence, we did not use the weighting criteria introduced by some authors. Using the aforementioned dimensions, it is possible now to evaluate the dataset by equation (11). We denote it by Quality Indicator (QI), it is calculated by multiplying all the dimensions and used to rank the sources. QI gives us an output that ranges from 0 to 1 which 1 is the best and 0 is the worst.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(v_i - v_i')^2 \quad (8)$$

$$Accuracy_{deviation} = 1 - NMSE \quad (9)$$

$$Accuracy = Accuracy_{count} * Accuracy_{deviation} \quad (10)$$

$$QI = Completeness * Consistency * Accuracy \quad (11)$$

3.1.2 Data preparation phase

This is the second phase of our model with two steps; where the data is collected and then pre-processed. We believe that data pre-processing step is very important and should be done to all the available datasets. It helps to speed up the process and make the dataset easily to be processed in the next phases.

Data collection step

In this step, the data should be gathered in specific format from different sources. This step differs when the selected domain is changed. As the data could be a real-world measurement picked using sensors or power grids in the case of IOT domain. It could also be collected by a specific script design to crawl the web for social media domain. Within the enterprises the data could be gathered from the databases as well.

As the data is collected from different sources, the probability of facing a problem such as sources heterogeneity is very high. This heterogeneity could be found in different levels, schema, format, and values [6]. At the schema level, different sources may differ in the structure of the data and the name of the attributes. Regarding the format level, each source may present the data in different data type. Finally, they definitely may differ at the value level. Some of the values might be precisely the same as the ground truth, some might be slightly different from the true value, and could be totally different from it. Thus, the next step is required in order to solve the problem of sources heterogeneity.

Data pre-processing step

This is an important step as it helps to accelerate the upcoming phases of the model. It helps also to identify the potential copying sources, by comparing their structures and schema. We assume the data could be collected in unstructured way unlike what have been assumed in [39][40][41][43] and others that the data is always in structured format. Mainly in this step the problem of sources heterogeneity mentioned in the data collection step should be solved and all sources become

homogenous and have structured format. Mostly we focus on adjusting source schema, data format, data structure, and data selection.

In order to prevent any mistakes while processing and evaluating the quality of the data, it is essential to unify the sources data format. For instance, in the financial domain some sources can present the value of a company in a numerical format (e.g. 25,000,000) others may present it in different way (e.g. 25M). In addition to that, the index (key attribute) should be unified across the sources to access data records, which most probably will be the timestamp by default as we are dealing with temporal data, yet it should be checked.

Data structure is about the attributes in each source; the attribute name may be different among sources, besides each source could have a lot of attributes. Here we try to select the desired attributes only, then unify its name across sources. It has been done by defining a set of global attributes for the preferred attributes and then change the local attribute of each source to the global one. Assuming that the data collecting has no time interval constraints, it is better to select the time interval needed from the collecting data. These steps should speed up the processing time while evaluating the datasets. Finally, the datasets should be stored in a way to be accessible for the used tool. It could be a database or an online server as long as it is accessible by the tool and process in a smooth and fast way.

Chapter 4: Model Implementation

The implementation of the model in detail will be discussed in this chapter. We evaluate the quality of the data for different stock market data sources. The analysis calculations and visualizations have been made with Python, by using Google Colab [78]. Google Colab is free cloud service based on Jupyter Notebooks that supports free GPU provided by Google. Several libraries are used such as pandas, pandas_datareader.data, numpy, matplotlib, seaborn and datetime.

4.1 DATA PREPARATION

The first step in preparing the data is to collect it. As mentioned before the focus is on stock market data, more specifically NASDAQ stock market. NASDAQ is an American stock exchange located in New York City. It is ranked the second stock exchange by Market Cap of shares traded. It has more than 3,300 company listings, with approximately 1.8 billion trades per day. There are many sources providing the stock market data, some of them are for free and others require subscription with fees. In this thesis, the data were collected from free sources. As the main goal of the thesis is to assess the quality of different stock market data sources, and we believe that the paid sources have already high quality compared to the free ones.

The three popular free financial sources are Yahoo Finance, Google Finance, and MSN Money. The data were collected in November 2019; thus, the examined time interval is from January 2019 to October 2019. At this time, Google Finance stopped providing the data for free. Hence, we collect the data from four sources, namely, Yahoo Finance, MSN Money, Stooq, Tiingo. For Yahoo Finance and Stooq, the data were collected using "pandas_datareader.data" Python library. While for Stooq and Tiingo, Get method were used to collect the data. Get requests are the most common method in APIs, used to retrieve data from a server.

We focused on 60 companies divided in three scenarios, 20 companies in each scenario. The classification is made based on the Market Capitalization of the company, this classification is provided by NASDAQ [69] as shown in Table 4.1. The classification is conducted based on the Market Cap of December 2018, as it changes every month. The largest 20 companies are in the first scenario, the list of companies is

shown in Table 4.2. The second scenario includes 20 medium companies, these companies are selected randomly from different industries and different geographical locations. The second scenario list of companies is shown in Table 4.3. Finally, the third scenario has 20 small and micro companies, they are selected randomly as well. The list of companies in the third scenario is shown in Table 4.4.

Table 4.1: Companies classification by market capitalization provided by NASDAQ [77] (M: Million, B:Billion)

Classification	Market Capitalization
Mega	Greater than $200B
Large	From $10B to $200B
Medium	From $2B to $10B
Small	From $300M to $2B
Micro	Less than $100M

Table 4.2: The top 20 companies of the first scenario

Symbol	***Sector***	***Industry***	***Country***
AAPL	Technology	Computer Manufacturing	United States of America
AMZN	Consumer Services	Catalog/Specialty/Distribution	United States of America
ADBE	Technology	Computer Software: Prepackaged Software	United States of America
ASML	Technology	Industrial Machinery/Components	United States of America
AMGN	Health Care	Biotechnology: Biological Products	United States of America
AVGO	Technology	Semiconductors	United States of America
CELG	Pharmaceutical	Pharmaceutical	United States of America
ADP	Technology	EDP Services	United States of America
AMD	Technology	Semiconductors	United States of America
AMAT	Technology	Semiconductors	United States of America
BIIB	Health Care	Biotechnology: Biological Products	United Kingdom
ATVI	Technology	Computer Software: Prepackaged Software	United States of America
ADI	Technology	Semiconductors	United States of America
BIDU	Technology	Computer Software: Programming, Data Processing	China
ADSK	Technology	Computer Software: Prepackaged Software	United States of America
ALXN	Health Care	Major Pharmaceuticals	Switzerland
CDNS	Technology	Computer Software: Prepackaged Software	United States of America
ALGN	Health Care	Industrial Specialties	United States of America
BMRN	Health Care	Major Pharmaceuticals	United States of America
AAL	Transportation	Air Freight/Delivery Services	United States of America

Table 4.3: The second scenario list of companies

Symbol	***Sector***	***Industry***	***Country***
APO	Finance	Investment Managers	United States of America
NICE	Technology	Computer Software: Prepackaged Software	United States of America
BMLP	Finance	Banking services	canada
FBHS	Basic Industries	HomeBuilding	United States of America
CZR	Consumer Services	Hotels/Resorts	United States of America
VIAB	Media	Media	United States of America
NEBU	Finance	Business Services	United States of America
KRC	Consumer Services	Real Estate Investment Trusts	United States of America
JHX	Capital Goods	Building Materials	Ireland
CMA	Finance	Major Banks	United States of America
CY	Technology	Semiconductors	United States of America
COTY	Consumer Non-Durables	Package Goods/Cosemtics	United States of America
ANAB	Health Care	Major Pharmaceuticals	United States of America
BPY	Finance	Real Estate	Bermuda
WLK	Basic Industries	Major Chemicals	United States of America
DVN	Energy	Oil & Gas Production	United States of America
AEG	Finance	Life Insurance	Netherlands
LOGI	Technology	Computer peripheral equipment	United States of America
ARGX	Health Care	Biotechnology: Biological Products	Belgium
ABIL	Technology	Ability Computer & Software Industries	Israel

Table 4.4: The Third scenario list of companies

Symbol	***Sector***	***Industry***	***Country***
WSC	Industries	Portable storage unite and offices	United States of America
WD	Finance	Finance: Consumer Services	United States of America
DRH	Consumer Services	Real Estate Investment Trusts	United States of America
GEO	Consumer Services	Real Estate Investment Trusts	United States of America
MTSI	Technology	Semiconductors	United States of America
RWT	Consumer Services	Real Estate Investment Trusts	United States of America
BANF	Finance	Major Banks	United States of America
ONTO	Capital Goods	Industrial Machinery/Components	United States of America
USPH	Health Care	Medical/Nursing Services	United States of America
ABR	Consumer Services	Real Estate Investment Trusts	United States of America
CORT	Health Care	Major Pharmaceuticals	United States of America
DRNA	Health Care	Major Pharmaceuticals	United States of America
EGOV	Miscellaneous	Business Services	United States of America
HSC	Consumer Services	Diversified Commercial Services	United States of America
EPZM	Health Care	Major Pharmaceuticals	United States of America
TRVG	Technology	EDP Services	Germany
IMH	Consumer Services	Real Estate Investment Trusts	United States of America
NTIC	Capital Goods	Industrial Specialties	United States of America
TGB	Basic Industries	Precious Metals	Canada
TRX	Basic Industries	Precious Metals	Tanznia

The next step after collecting the data for the three scenarios from all the sources is data preprocessing, which we prepare the data in order to enter the DQ assessment phase. In this step, the data from each source is checked and unified based on three aspects, namely, schema, format and time interval. Sources could be heterogeneous regarding these three aspects, for this reason we tried to unify these aspects in order to facilitate the next steps. Many attributes are provided by the sources, which we call it local attributes. We identified seven global attributes as mentioned in Chapter 3 and select these attributes in sources. The data structure is heterogenous as well, as some sources provide the data per each company and other provide them all aggregated in one file. In order to compare the sources, we merge the data from all the sources in each scenario in one file.

Regarding the data format, different format is provided from different sources. For example, in the volume attribute some represent it with letters like “2M”, others represent it all with numbers like “2,000,000”. Hence, unifying the data format is required, we choose the numerical format. While collecting the data, it is not always possible to select the required time interval for some sources. Therefore, after collecting the data we had to make sure to select just the examined period, which is from January 2019 to October 2019.

Finally, as we are dealing with financial data, it is important to introduce the financial quarters. A quarter is a three-months period on a company’s financial calendar that acts as a basic for periodic financial reports and paying of dividends [79]. In Table 4.5 is shown the standard calendar quarters. Companies, investors, analysts, and traders are interested in the data quarterly to make comparison and evaluate trend. Therefore, the examined period is evaluated per each quarter. Since we do not have data after October 2019, the given period is classified into three quarter and the October month. This phase not only helped in facilitating and speed up the next phases, but also helped us to check for the potential copying between the sources. By checking the difference between their structures, schemas and data formatting.

Table 4.5: The standard calendar quarters that make up the year [79]

Quarter	Months
First quarter (Q1)	January, February and March
Second quarter (Q2)	April, May and June
Third quarter (Q3)	July, August and September
Fourth quarter (Q4)	October, November and December

4.2 DQ ASSESSMENT PHASE

In this phase we applied the equations introduced in Chapter 3, to calculate and assess the DQ of each source. Data and error have been visualized to analyse and better assess the sources. All these steps have been carried out by Python code using Google Colab as mentioned before. Finally, the results are stored, it is presented in the next chapter. After applying the model on several attributes, the results were almost identical. Hence, the focus of the presented results are on the "Close price" attribute.

4.3 SOURCES EVALUATION PHASE

A comparison between the sources is carried out, after receiving the results from the previous phase. The comparison is done based on the companies that have glitches within the dataset, not only by comparing the DQ dimensions. The results of this phase are presented as well in the next chapter.

4.4 PREDICTIVE MODEL

For predicting the stock market values, LSTM model was implemented using the Keras framework and trained using three different datasets that contained data about the price of the three companies. Furthermore, this section describes how the different hyperparameters of the LSTM model were chosen to enhance the model's accuracy.

Data collection

The data of three different companies were collected from NASDAQ official website. The datasets contained data about the prices from 01-03-1970, 01-03-2005 and 01-03-2007 to 31-12-2019 for Alcoa Corp, Almaden Minerals Ltd and General Electric (GE) respectively. Shown in Table 4.6 a sample of the stock prices for GE company.

Table 4.6: Sample data for GE used as input in the predictive model

Data	Open	High	Low	Close	Volume
03-01-2007	30.116	30.715	30.096	30.572	53632518
04-01-2007	30.572	30.594	30.150	30.392	38619002
05-01-2007	30.247	30.401	30.039	30.239	33199928
08-01-2007	30.150	30.328	29.989	30.230	29445866
09-01-2007	30.417	30.586	30.079	30.230	30551941

Data pre-processing

Before the collected dataset is used by the LSTM model, normalization was applied to all data in the features in order to improve the model's accuracy. The LSTM model was constructed and trained with Keras, a high-level neural network API written in Python with TensorFlow as the backend. Training was performed on a Tesla K80 GPU provided by Google Colab. The dataset is split into training and validation in an 80%-20% manner. For training the network, RMSprop optimizer was used. Moreover, a loss function of Mean-squared error was chosen and the weights were randomly initialized. The LSTM Model is trained to determine the price of the closing price for each training day.

Network's architecture

The LSTM model architecture is shown in Figure 4.1. It consists of three main layer types: LSTM cells hidden layers, Dense layer and dropout layer. As shown in Figure 4.1, the model implemented consists of 4 layers: 2 LSTM hidden layers followed by 2 dense layers. After each hidden layer a dropout layer is added for better generalization. A dense layer is a type of neural networks layers where each neuron in one layer has direct connections to the neurons in the next one. The dropout is an important technique that reduces overfitting by randomly choosing cells in a layer according to the probability chosen and set their output to 0. Overfitting occurs when the model memorizes the patterns and features of the training data and fails to apply the learnt attributes on the test data.

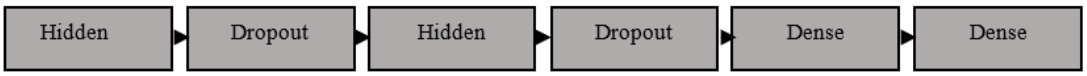

Figure 4.1: LSTM model architecture

While building the LSTM model, certain hyperparameters have to be adjusted properly to get an accurate prediction when testing the model. Batch size, number of epochs, learning rate, dropout rate, number of LSTM cells, number of hidden layers and time steps are considered to be the hyperparameters in our implemented model. These parameters are tuned by empirical testing to find their optimal values. Each hyperparameter is tested one by one and we try to find an optimal value to that specific parameter. The best value for a hyperparameter is found by evaluating the LSTM model by back testing it with the test data. The hyperparameter value that achieved the lowest MSE between the model's prediction on the closing price and the actual closing price for that day is the one implemented. In Table 4.7, an illustration of the final chosen hyperparameters for all datasets after various trials is demonstrated. The results of the model are shown in the next chapter.

Table 4.7: The final chosen hyperparameter for the LSTM model

Hyperparameter	Value
Batch size	20
Learning rate	$1x10^{-5}$
Number of epochs	300
Dropout rate	0.4
Number of LSTM cells	100,60
Number of hidden layers	2
Time steps	**GE**: 60
	Alcoa & Almaden Minerals :180

The mentioned hyperparameters and their functions are as follows:

- **Batch size**: The number of training examples utilized in one iteration.
- **Learning rate**: The rate by which the values of the learned parameters are updated.
- **Number of epochs**: The number of iterations obtained for training the data.
- **Dropout rate**: The probability of training a given node in a layer
- **Number of LSTM cells**: The number of LSTM cells in each hidden layer
- **Number of hidden layers:** The number of the layers containing LSTM cells
- **Time steps**: This is equivalent to the amount of time steps needed to be remembered by the LSTM model.

Chapter 5: Results

In this chapter, the results of implementing the model illustrated in Chapter 4 and the output of the predictive model are presented. These results are the output of first part which is DQ assessment phase. In the second stage, which is data preparation phase, we did not encounter any problems refining the datasets. When we check if the source is providing data for all the selected companies or not, no problems were detected as well. As mentioned before, there are three different scenarios used to assess the quality of the sources. The results of each scenarios are being presented below as well as the final output by merging all the scenarios. All the results shown in this chapter are for "close price" attribute, which is the closing price, as other attributes have been checked and the results were quite the same.

5.1 FIRST SCENARIO: TOP 20 COMPANIES IN NASDAQ

In this section, the results for all the sources are presented for the first scenario, that includes the top 20 companies in NASDAQ.

5.1.1 Completeness dimension

For all the sources, no missing values were observed. The sources were totally clean regarding the completeness dimension, which could be because of the importance of these companies as they are the top ones in the stock market. Therefore, the completeness dimension for all the sources will be equal to 1.

5.1.2 Consistency dimension

No violations were detected for all the types of consistency across all the evaluated sources. By checking all the single values for each attributes of the sources, the rules were all followed which lead to the value of 1 to type 1 constraint for all sources. The same case is applied on both type 2 and type 3, resulting the same value for both. As shown in Figure 5.1, on the x-axis are the different types of constraints and on y-axis is the value. It shows that all the sources are clean from the consistency dimension point of view.

The consistency dimensions are calculated for all the sources by multiplying all the types values for each source and the result was 1 for all the sources.

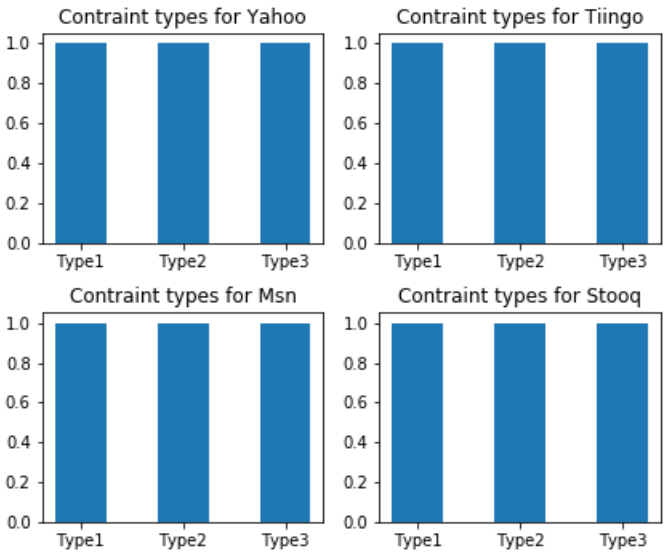

Figure 5.1: The constraints types for all sources(Top 20 Companies)

5.1.3 Accuracy dimension

Regarding the error which is the deviation of the data item value from the ground truth. As a primary step to analyze the error, the statistical parameters and the distribution of the error are presented in order to take an overall glimpse on all the sources. The visualization of error graphs is made in various types and presentations in order to choose the best visualization way to present the data to be easy understandable.

The statistical parameters of error distribution which include mean, Standard deviation (STD), maximum (Max) are shown in Table 5.1. It is shown that the error in most of the sources across the quarters is zero, except for Stooq. The Stooq source has an error in all the quarters with a noticeable high value of maximin error in the Q1 which is equal 8.15. Since the error is mainly in the source Stooq, so focusing on the maximum values in Stooq as shown in Figure 5.2. The error values across the quarters are relatively high, even though it is decreasing across the quarters but still high. The maximum error value for Yahoo is in Q2 which is 0.93. It is an outlier, which means that it is not a consistent trend as shown in Figure 5.3.

Table 5.1: Statistical summary for error in all sources(Top 20 companies)

	First quarter			Second quarter			Third quarter			October		
	Mean	**STD**	**Max**	**Mean**	**STD**	**Max**	**Mean**	**STD**	**Max**	**Mean**	**STD**	**Max**
Yahoo	0	0	0	0	0	0.159	0	0.02	0.93	0	0	0.04
MSN	0	0	0	0	0	0	0	0	0	0	0	0
Tiingo	0	0	0	0	0	0	0	0	0	0	0	0
Stooq	0.92	1.72	8.15	0.58	1.25	5.93	0.22	0.58	2.77	0	0	0

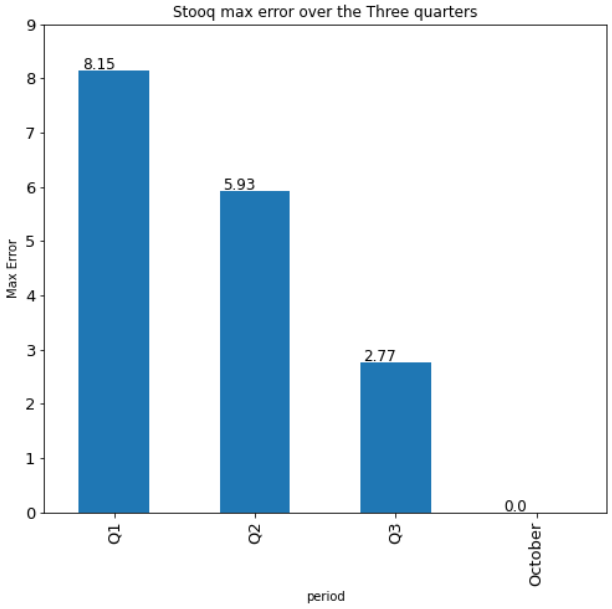

Figure 5.2:The maximum error value for the three quarters in Stooq (Top 20 Companies)

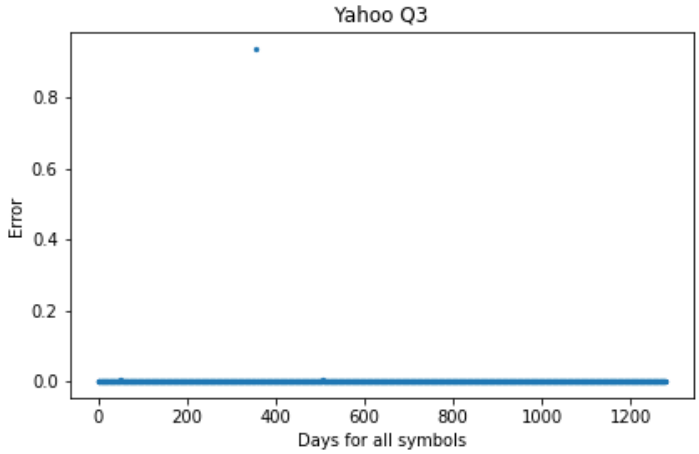

Figure 5.3: The error distribution for the third quarter in Yahoo (Top 20 companies)

From the statistical parameters of error distribution, it is clearly shown that Yahoo, Msn and Tiingo quality is good from the accuracy point of view. While Stooq has high error values which can be noticed from the mean and STD. The error distribution for all the quarters in Stooq can be shown in Figure 5.4, where the error are mainly in the first three quarters with no error in October month. Noticed from the patterns in the error distribution that the errors are not in all the companies, as this distribution is for all the companies in the dataset. Mainly the errors are in the companies where in the middle of the dataset with high difference in the error values.

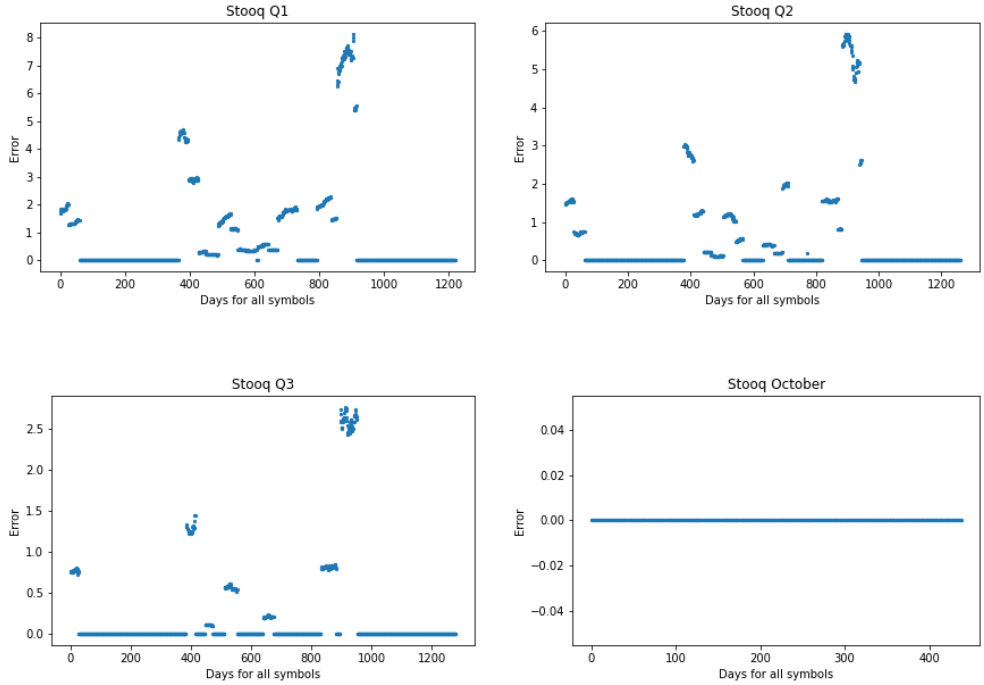

Figure 5.4:Error distribution across all quarters in Stooq (Top 20 Companies)

In order to take a closer look on which companies have the error, the error count for each company is visualized in Figure 5.5. It shows that the errors are mainly in the middle companies with a total of 1,273 incorrect values. Obviously, the error here is only in 10 companies out of the 20 chosen companies which is 50% of the companies presented in the dataset. This means that if another 10 companies are chosen other that the affected companies, the results could have been different.

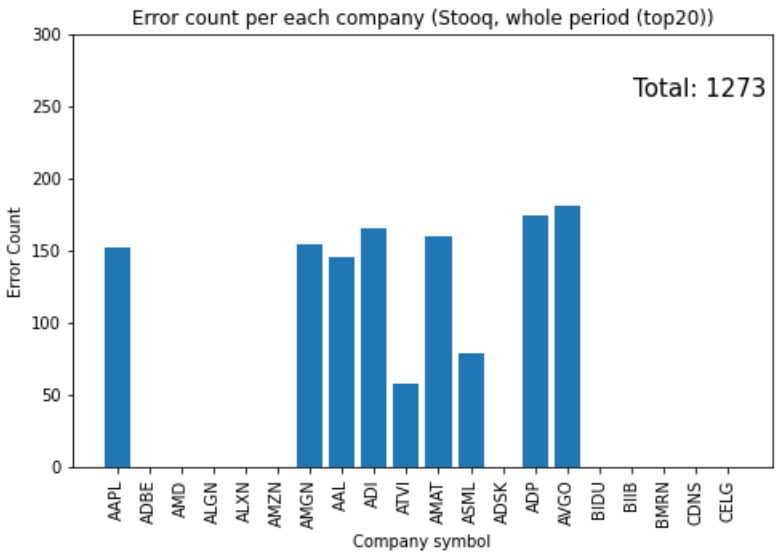

Figure 5.5: The error count for each company in Stooq (Top 20 Companies)

Finally, the number of errors and MSE are calculated to measure the accuracy dimension. The error count in each quarter for all the sources are shown in Figure 5.6. As mentioned before and can be shown here that there are no errors found in Tiingo, Yahoo, and MSN. On the other hand, Stooq has a large number of wrong values that deviated from the true value as shown in Figure 5.6 with 1,273 wrong values in total. There is a descending trend along the quarter until it reached zero in October. For the other part of the accuracy, which is the MSE calculations that is not affected by the error counts but the amount of deviation from the ground truth. MSE values for all quarters in all sources are shown in Figure 5.7, underlines the same results of the error count in Figure 5.6. Which is the error being just in Stooq and the descending trend in the error across the quarters. Comparing the decreasing trend in the error count and in MSE for Stooq, MSE is dramatically decreasing more than the error count.

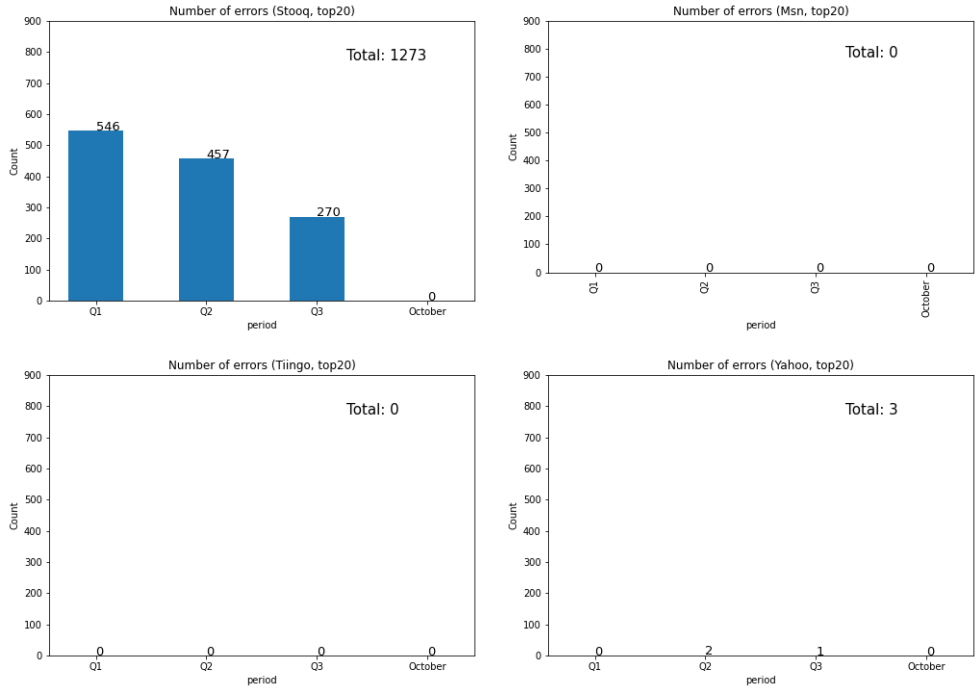

Figure 5.6: The error count across the whole period for all sources (Top 20 Companies)

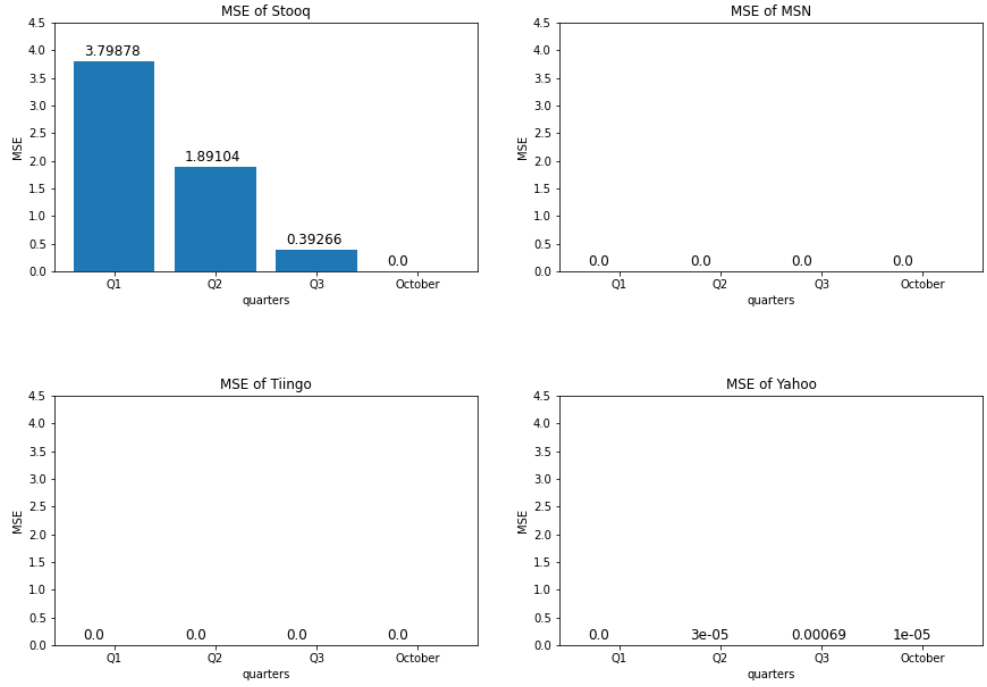

Figure 5.7: The MSE for all quarters in all sources (Top 20 Companies)

Following the equations mentioned in Chapter 3, $Accuracy_{count}$ and $Accuracy_{deviation}$ in order to measure the accuracy dimension. In Table 5.2, these two calculations and the overall accuracy dimensions for each source are shown.

Table 5.2: Accuracy dimension values for each source (Top 20 companies)

	Stooq	Msn	Tiingo	Yahoo
$Accuracy_{count}$	0.7	1	1	1
$Accuracy_{deviation}$	0.77	1	1	1
$Accuracy$	0.54	1	1	1

5.1.4 Summary

The summary for all the first scenario dimensions as well as the QI is presented in Table 5.3. As shown all the sources that have a QI of 1 which is the best possible case except for Stooq. The problem of stooq came from the accuracy dimension, which is 0.54. This was because the huge number of wrong values compared to the true value.

Table 5.3: Dimensions summary for the first scenario

	Completeness	Consistency	Accuracy	QI
Stooq	1	1	0.54	0.54
Yahoo	1	1	1	1
MSN	1	1	1	1
Tiingo	1	1	1	1

5.2 SECOND SCENARIO. MEDIUM SIZE COMPANIES

This scenario, as mentioned in the previous chapter, includes the medium companies, where the companies are selected randomly from different geographical locations and different industries.

5.2.1 Completeness dimension

The missing values in each source can be shown in the Figure 5.8. As shown the number of missing values varies between the sources, with a highest value is in MSN following by Stooq and finally Tiingo with no missing values in Yahoo. However, the number of missing values in Tiingo is 15, which is not huge number compared to the total number of values in the dataset. By taking a closer look at Tiingo source, it

can be shown that missing values are mainly in two symbols, namely, NEBU and BMLP. As shown in Figure 5.9, the missing values are especially for Q1 and Q2. The missing values can be shown as there are no reported values presented.

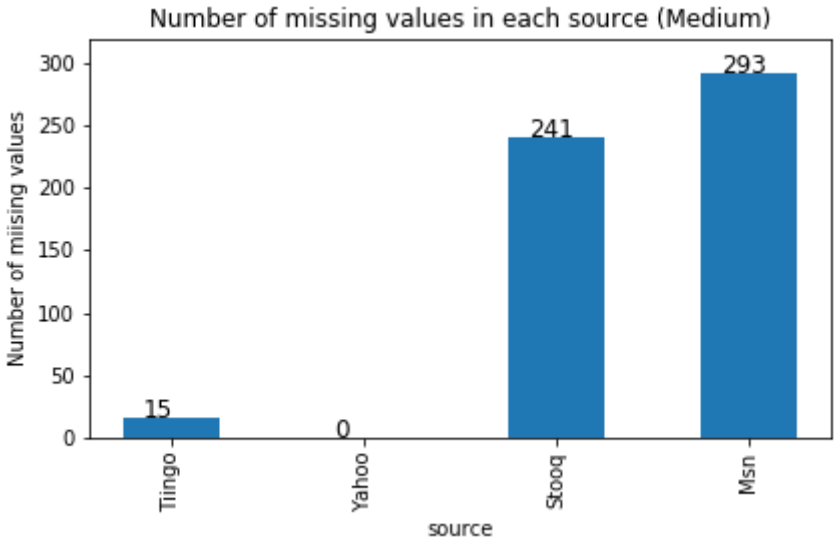

Figure 5.8: Count of missing values for all sources (Medium companies)

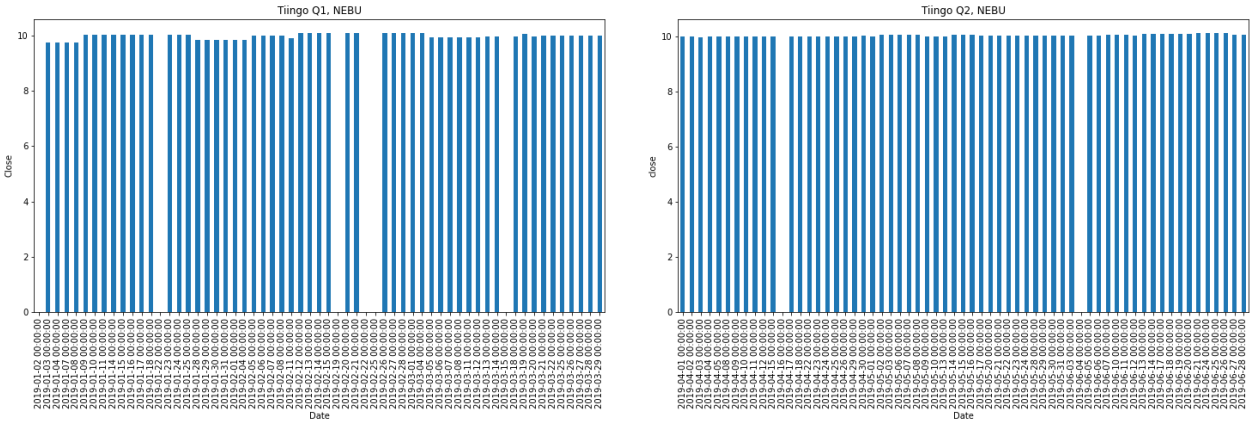

Figure 5.9: missing in NEBU for the first and second quarters in Tiingo (Medium companies)

For Stooq, similarly for Tiingo case; the missing values are just the same two companies. Shown in Figure 5.10, the missing days in the reported days of BMLP and NEBU. The dates are not clear in this presentation, because it is not the aim of the figure. In Stooq the missing values are spread across the whole period. It can be shown in Table 5.4, the number of missing values per company in each quarter in Stooq. Notice

that it has a descending behavior as it is getting lower across quarter, yet still has missing values.

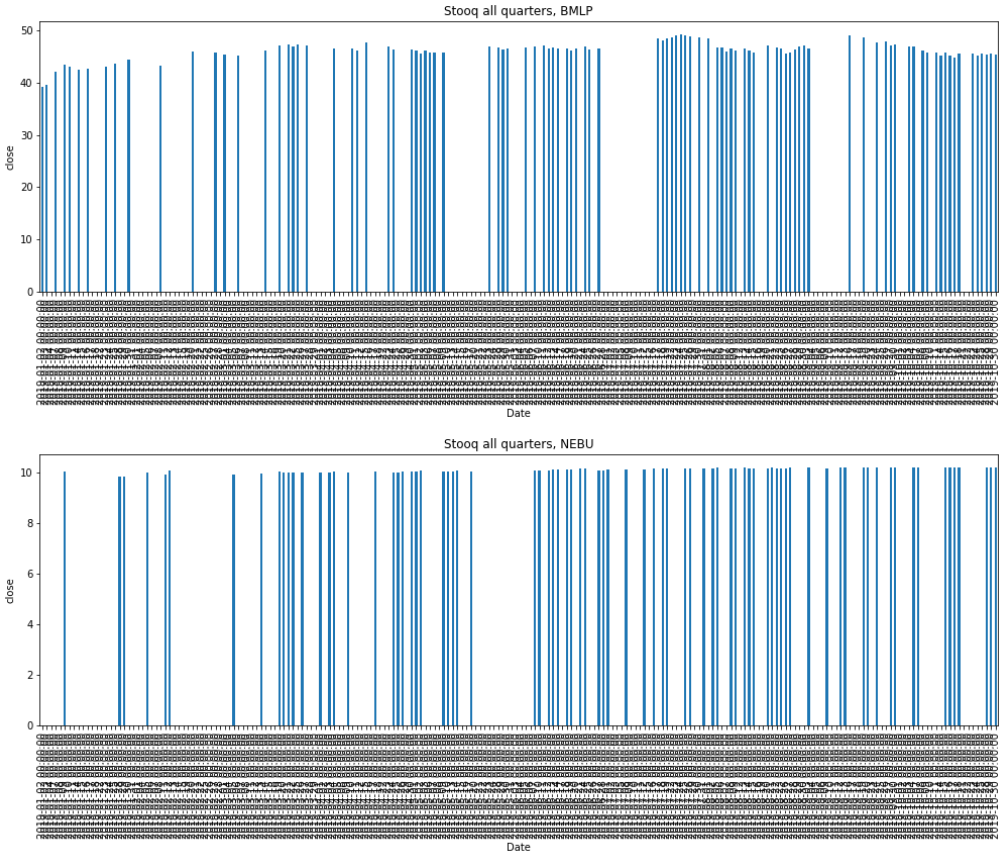

Figure 5.10: missing values in BMLP and NEBU across all quarters in Stooq (Medium companies)

Table 5.4:The count of missing values in NEBU and BMLP companies in Stooq (Medium companies)

Quarter (Stooq)	NEBU	BMLP
1st quarter	48	40
2nd quarter	36	34
3rd quarter	33	31
October month	13	6
Total	130	111

Surprisingly the missing values in MSN was not just missing values, it was missing to report the whole day. In other words, they are not reporting these days as a working day for the stock market. In addition to that, it was noticed extra days reported that is not reported by the official NASDAQ dataset. Regarding the missing days, as shown in Figure 5.11 the missing values are mainly in NEBU and BMLP with a very low percentage in the other companies.

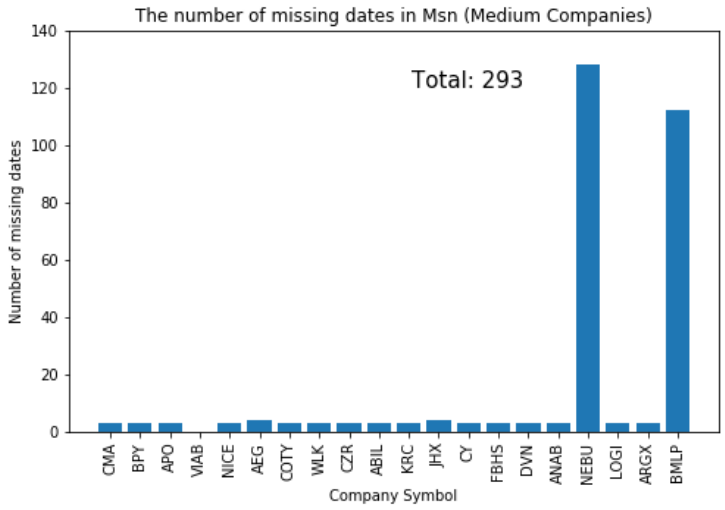

Figure 5.11:The missing days in the MSN source for each company

Table 5.5: The count of missing values in NEBU and BMLP companies in MSN source

Quarter (MSN)	NEBU	BMLP
1st quarter	48	41
2nd quarter	36	34
3rd quarter	33	31
October month	14	6
Total	128	112

Likewise, the Stooq case, the number of missing values for MSN can be shown in Table 5.5 with the same descending behavior across quarters. As it easily can be seen

that it is not just the behavior but also almost the same numbers, which raise a flag that they could be coping from each other. However, using the same concepts introduced in [6] by comparing the schema commonality while performing the data preparation phase, it was noticed that there was no potential coping. A heatmap between the attributes of Stooq and the attributes of MSN is being checked as shown in Figure 5.12, which can be seen that there is no correlation between both. The correlation between all the attributes of Stooq and of MSN are black which according to the graph scale it is almost 0.

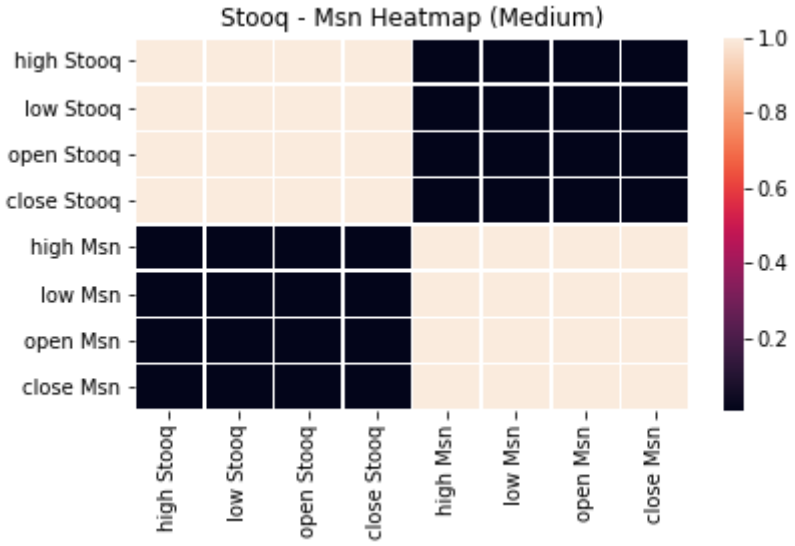

Figure 5.12:Heatmap for Stooq and MSN sources (Medium companies)

Regarding the excess days in MSN source, as shown in the Figure 5.13, mainly there are four common days are reported for all the companies except for three companies which they are VIAB, NEBU, BMLP. These four days are 10-3-2019, 17-3-2019, 24-3-2019, and 27-10-2019. By checking these days, it will be found that all of them are Sundays, which is a weekend and the stock market is not working. We believe that they are reporting the Fridays on Sundays with a delay for these specific Sundays. Randomly checking, it was found that for AEG, the reported close price on Sunday is 4.84 and the true value on Friday is 4.83. and for CMA it was the same case the difference between both values is very low.

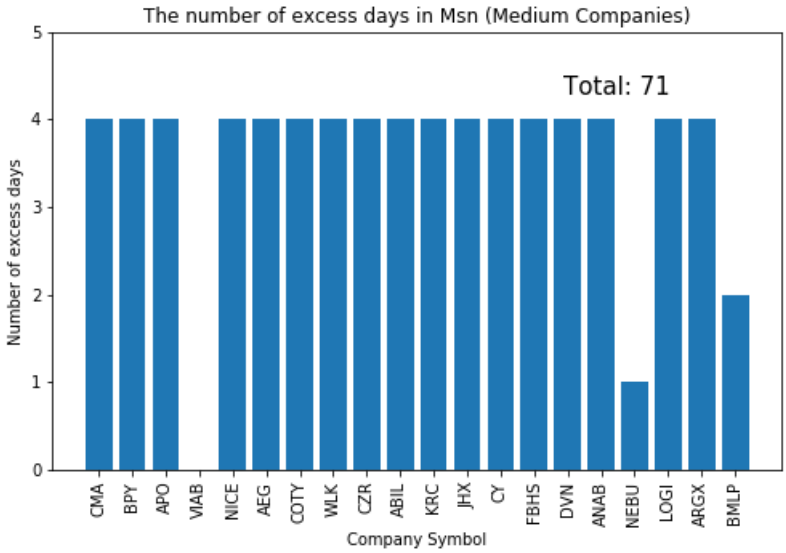

Figure 5.13:Number of excess days reported in MSN (Medium companies)

Now we can calculate the completeness dimension using the equation mentioned before in Chapter 3. The results values will be 0.94, 0.99, 1, and 0.93 for Stooq, Tiingo, Yahoo, and MSN respectively.

5.2.2 Consistency dimension

The same for the first scenario, there are no violation for all types of constraints. The three types constraints values are identical to the fist scenarios shown in Figure 5.1 with no changes. The consistency dimensions are calculated for all the sources by multiplying all the types values for each source and the result was 1 for all the sources.

5.2.3 Accuracy dimension

Considering the accuracy, the same preliminary steps will be followed here as well. The statistical summary for the error distribution is shown in Table 5.6. For Yahoo the error is almost zero across the quarters with STD of zero as well. The maximum values are null also except for Q3 which is 2.1. In Figure 5.14, it can be shown that this maximum value is just an outlier, which represents three consecutive days in September the 16th, 17th and 18th. The same case goes for Tiingo with a perfect no errors in all the quarters. It can be concluded from the mean and the STD that the maximum value in Q3 is just an outlier and it can be also shown in Figure 5.15. The circle in upper right corner of the graph is the outlier which is in the same month of Yahoo case but just in two days the 17th and 18th. Regarding MSN and Stooq, the mean error is more than zero

that means the error exists in a quite large portion of the dataset with a STD of 0.535 and 0.603 respectively. The trend of the maximum values for all the sources across the quarters can be shown in Figure 5.16. For Stooq, it can be shown that the maximum value is decreasing over time yet still existed, except for October month. In MSN, it does not have a trend across quarters. But focusing on the values, the maximum error value in the Q1 is 6.46 which is very high compared to the others considering the sensitivity of the stock prices.

Table 5.6:Statistical summary for the error in each quarter for all sources (Medium companies)

	First quarter			Second quarter			Third quarter			October		
	Mean	**STD**	**Max**	**Mean**	**STD**	**Max**	**Mean**	**STD**	**Max**	**Mean**	**STD**	**Max**
Yahoo	0	0	0.001	0	0	0.01	0.006	0.1	2.1	0	0	0.005
MSN	0.13	0.535	6.46	0	0	0.002	0.001	0.023	0.8	0.071	0.332	3.810
Tiingo	0	0	0	0	0	0	0.004	0.087	2.105	0	0	0
Stooq	0.529	0.603	2.542	0.375	0.416	1.677	0.165	0.221	0.805	0	0	0

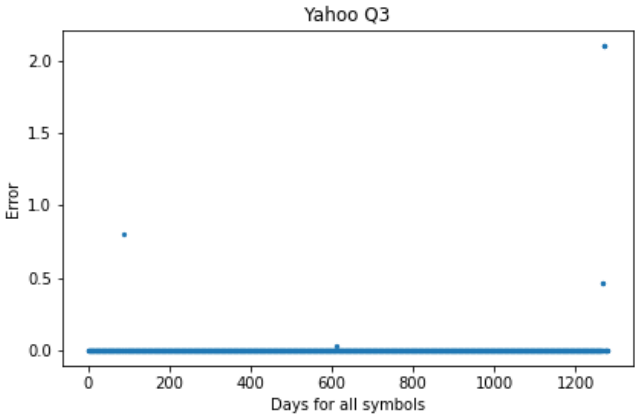

Figure 5.14:Yahoo error distribution for the third quarter(Medium companies)

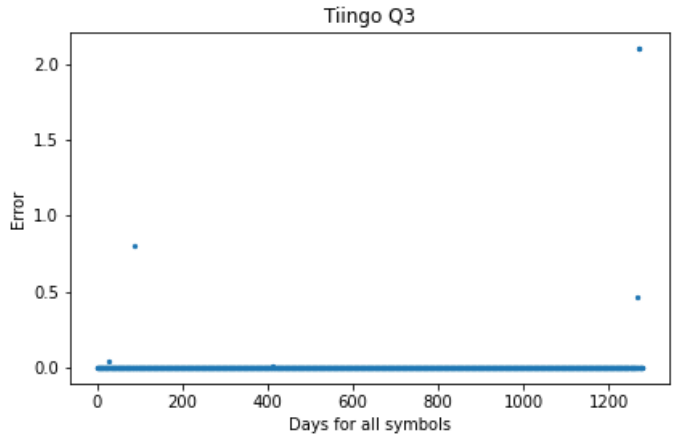

Figure 5.15:Tiingo error distribution for the third quarter(Medium companies)

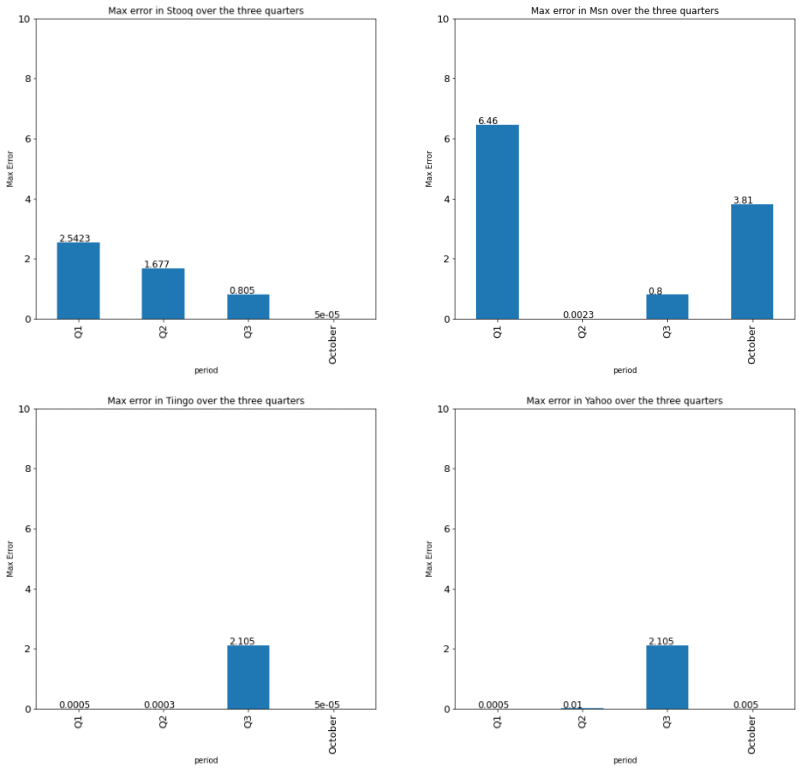

Figure 5.16: Maximum error values for all the sources across the quarters(Medium companies)

The error distribution visualizes the error across time for all the symbols in the dataset. The same remarks have been noticed, which is a huge amount of error in Stooq

across all the quarters except for October month and just in the Q1 and October month for the MSN source.

The distribution of errors in Stooq for all the quarters is shown in Figure 5.18. It can be noticed that the error has a random behavior with no noticeable patterns. By taking a closer look at the error per each company, the errors are mainly in 14 companies out of the 20 selected companies. On the contrary, for MSN despite that the number of errors is very low compared to Stooq, the errors are in most of the companies in the dataset as shown in Figure 5.19. In Figure 5.17, on the left side is the count of errors per company for Stooq in the overall dataset. On the right side of the same figure, is the mean error per company. As shown, having the highest number of errors does not mean the higher the deviation from the true value, as for the case of BMLP it has the lowest number of errors but the highest mean error.

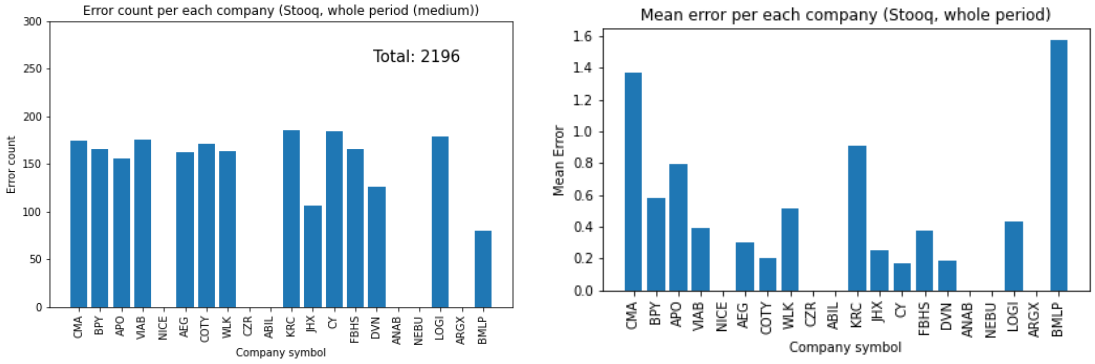

Figure 5.17:The count of the error per company for Stooq is on the left side, while the mean error per each company is a on the right (Medium companies)

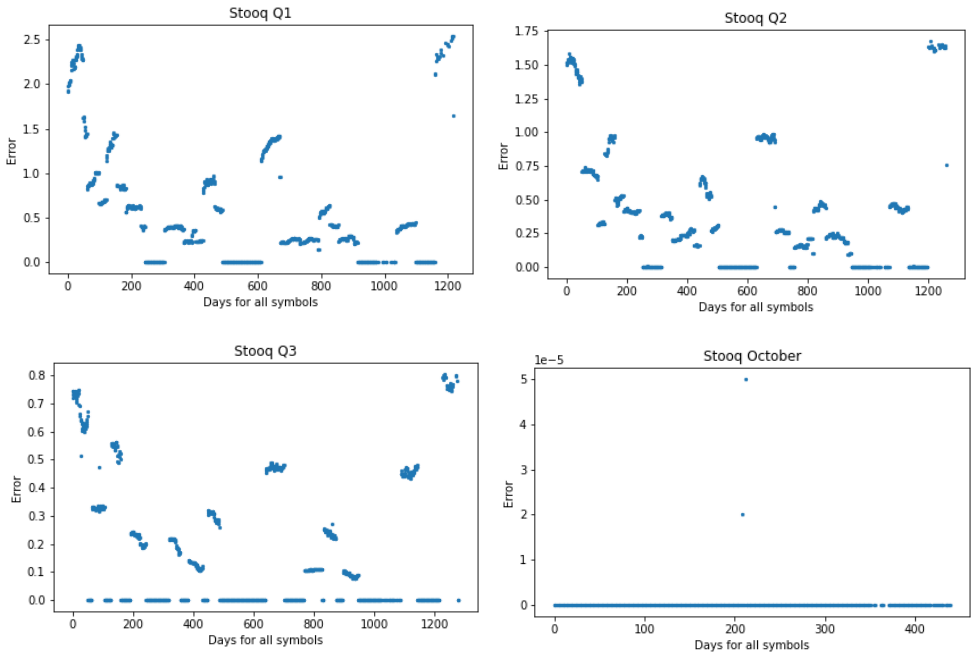

Figure 5.18: Error distribution in Stooq for all quarters(Medium companies)

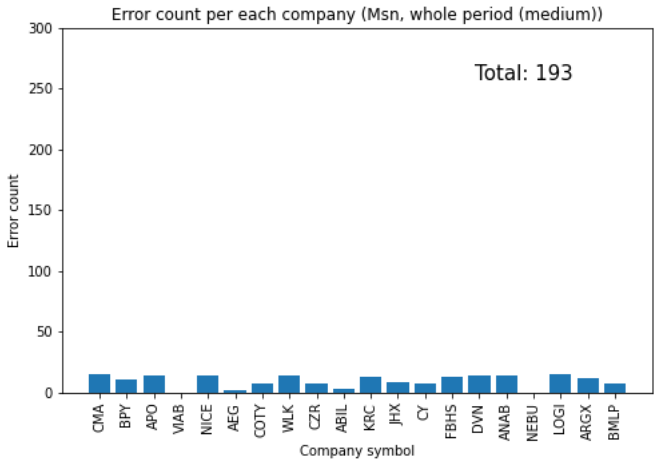

Figure 5.19: Number of errors in MSN for each Company (Medium Companies)

In Figure 5.20, the error distribution in MSN is visualized for all the quarters. As it can be shown in the graphs, the errors are mainly in the Q1 and the October month. A trend can be seen in the Q1 of no error in the beginning of the quarter but at the end the error starts to happen. This trend can be shown in Figure 5.21, this is a sample company (ANAB) to show the trend which applies on the other companies in the dataset as well.

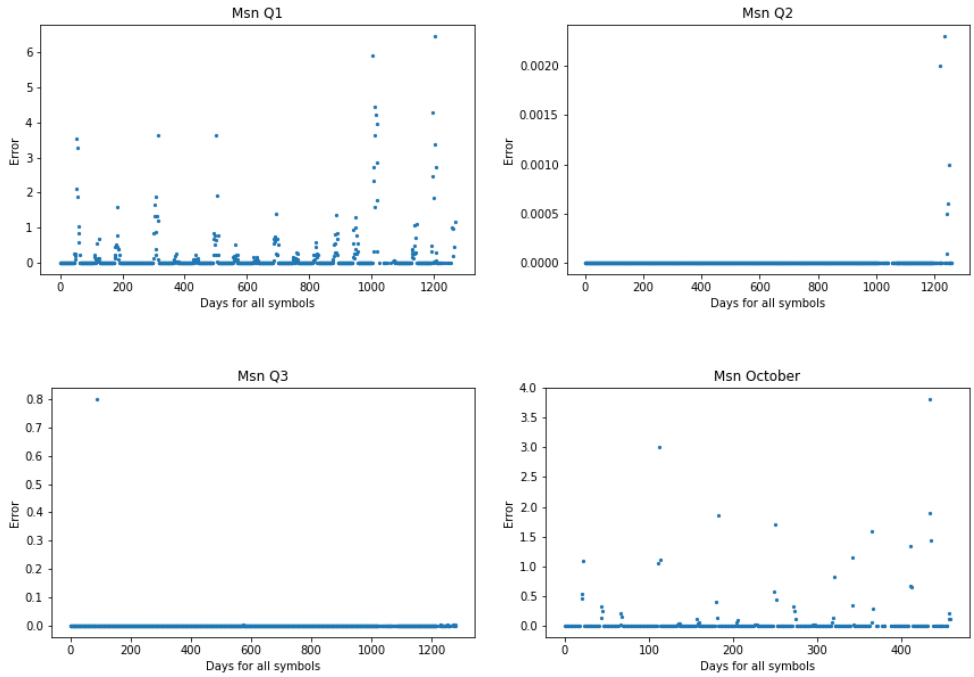

Figure 5.20: Error distribution in MSN for all quarters(Medium companies)

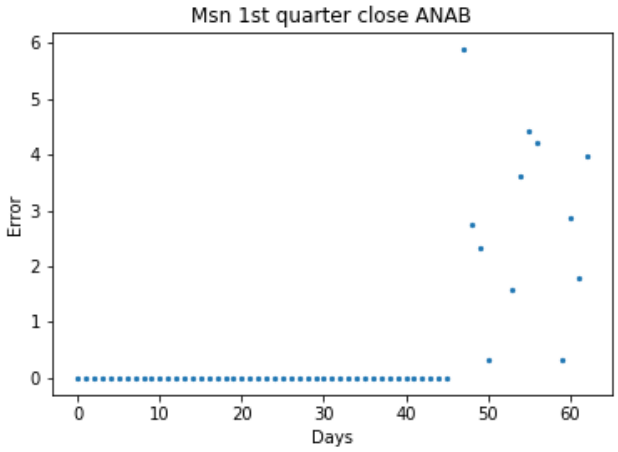

Figure 5.21: The Error distribution in the first quarter of a sample company in MSN(Medium companies)

The Number of errors for all the sources in each quarter is shown in Figure 5.22. The errors are mainly in Stooq and MSN with 2,196 and 193 errors respectively. The number of errors in Stooq is very high, almost half of the dataset and they are mainly in Q1 and Q2. While for October month, there are no errors. In MSN, the errors are mainly in Q1.

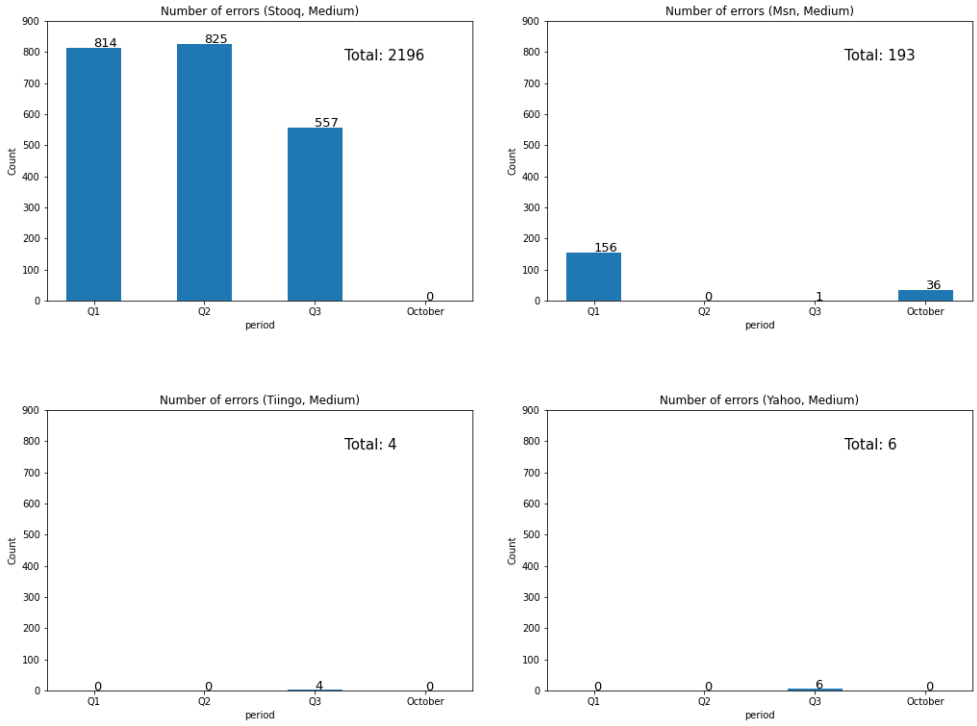

Figure 5.22:Error count for all sources in each quarter (Medium)

In Figure 5.23, the MSE for the examined sources across the quarters are shown. Yahoo and Tiingo have zero MSE, this is expected as they do not have errors. For Stooq source, it is decreasing over time until it reached zero in October. This also proofs that the number of errors it not correlated to the deviation from the true value. The overall value of MSE considering the whole period is 0.344 and 0.13 for Stooq and MSN respectively.

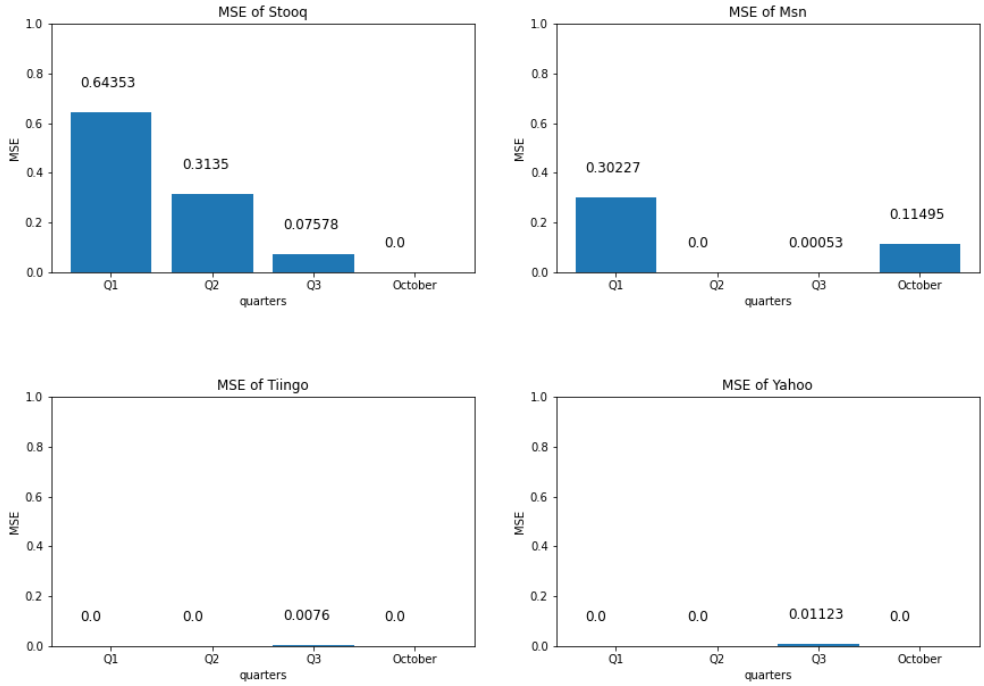

Figure 5.23: MSE for all sources across the quarters (Medium)

In Table 5.7, the accuracy parameter needed to measure the accuracy dimension and the accuracy dimension for all the sources are shown. Nevertheless, the number of errors in Stooq for Top 20 companies is lower than number of errors in the medium companies, the error deviation in Top 20 companies is higher than the error deviation in the medium companies.

Table 5.7:Accuracy dimensions values for each source (Medium companies)

	Stooq	Msn	Tiingo	Yahoo
$Accuracy_{count}$	0.48	0.97	1	1
$Accuracy_{deviation}$	0.87	0.99	1	1
$Accuracy$	0.42	0.96	1	1

5.2.4 Summary

Similarly, to the previous scenario, the dimensions summary and the QI are shown in Table 5.8. Yahoo is the only perfect source from QI perspective, the second-best source is Tiingo followed by MSN and finally Stooq. Tiingo is almost perfect, but the missing values affected its QI.

Table 5.8:Dimensions summary for the second scenario

	Completeness	Consistency	Accuracy	QI
Stooq	0.94	1	0.42	0.4
Yahoo	1	1	1	1
MSN	0.93	1	0.96	0.89
Tiingo	0.99	1	1	0.99

5.3 THIRD SCENARIO: SMALL AND MICRO COMPANIES

In this scenario, the small and micro companies are selected as mentioned before. The companies were selected randomly from different geographical locations and different industries.

5.3.1 Completeness dimension

Regarding the small and micro companies, there are quite large number of missing values in both Yahoo and MSN with a very small number in Stooq and no missing values in Tiingo as shown in Figure 5.24. By examining Yahoo dataset, it has been noticed that all the missing values are in one company, namely, ONTO. ONTO company name is Rudolph Technologies with a market cap of 1.8B$, which is about to be a medium company in the NASDQ classification. Yet, as shown in Figure 5.25 Yahoo is not reporting any days for it since the beginning of the year until the 28th of October. While on the other hand, MSN has missing values in all the companies but the majority of the missing values are in WSC symbol as shown in Figure 5.26. also noticed that the missing values are mainly in January with more than 20 days, which is almost the whole working days in the month.

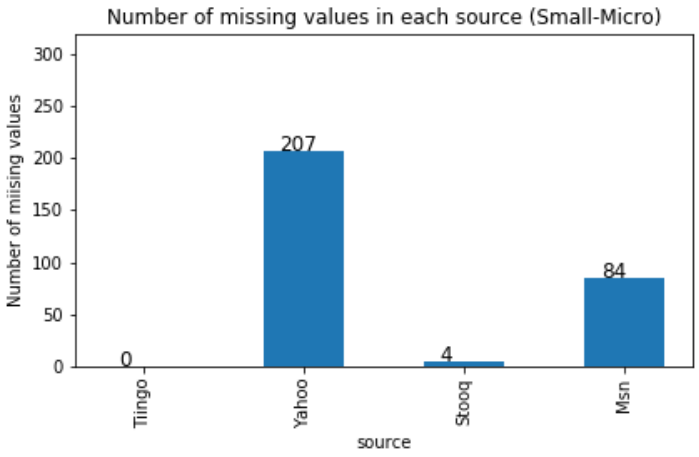

Figure 5.24:Number of missing values for each source (Small-Micro companies)

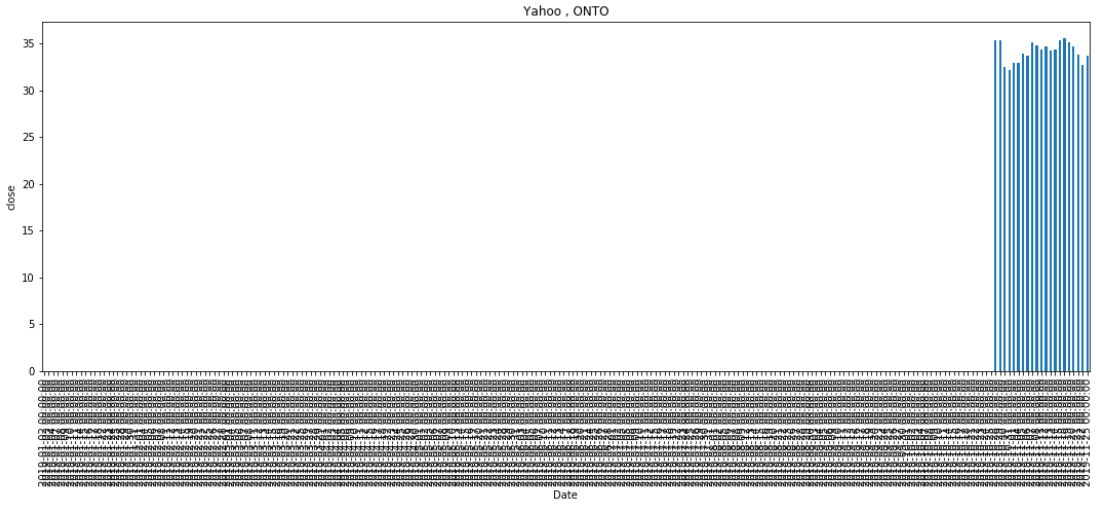

Figure 5.25:Reported days in Yahoo for ONTO Symbol (Small-Micro companies)

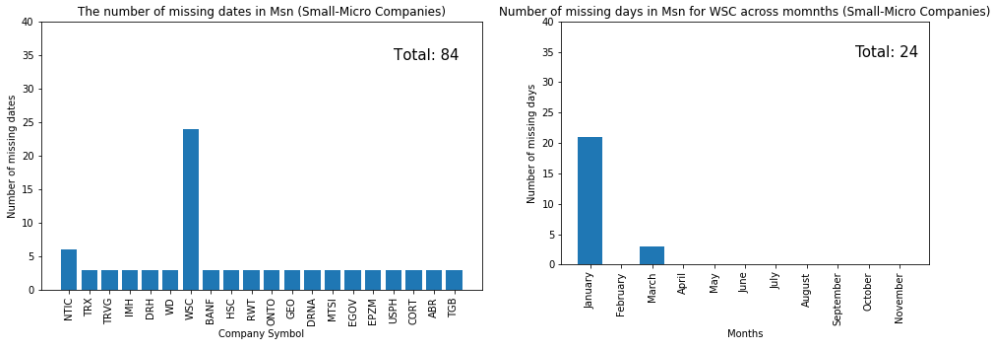

Figure 5.26:On the left side is the number of missing values per company while on the right side is the missing values for WSC across the year (Small-Micro companies)

The same weird behavior of MSN with the extra reporting days as the medium companies' case happened in the small and micro companies as well. As shown in Figure 5.27, for all the companies there are four extra days reported. By checking these days, they are all Sundays where the stock market is not working.

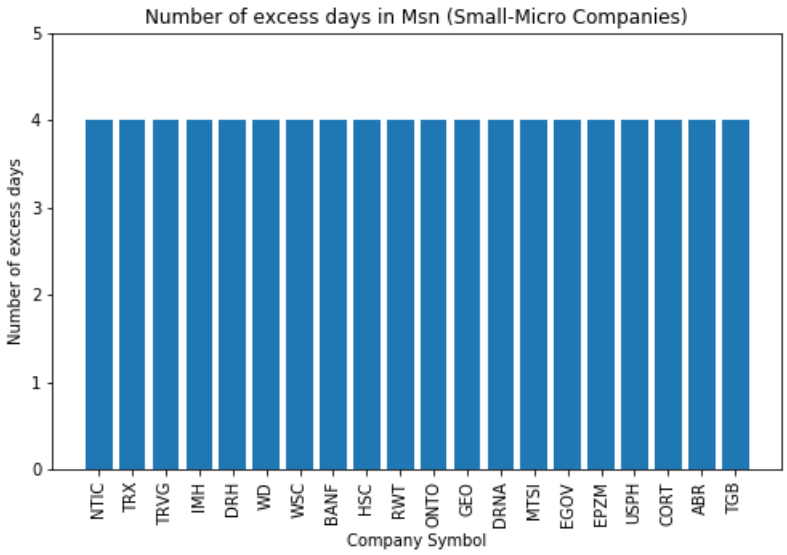

Figure 5.27: Extra reported days in MSN for each company (Small-Micro companies)

Calculating the completeness dimension giving the above-mentioned data. The completeness dimension values are 1, 0.99, 0.98, 0.95 for Tiingo, Stooq, MSN, and Yahoo respectively.

5.3.2 Consistency dimension

It will be the same as the previous two scenarios, where no violation found regarding any type of constraints. This will lead to the same values for all the sources which is 1.

5.3.3 Accuracy dimension

Similarly, to the other two scenarios, the starting point will be to analyse the statistical parameters of the error distribution as it gives an overview on the whole sources before going deeper. As shown in Table 5.9, Yahoo has no error in all the quarters.

Table 5.9: The statistical parameter of error distribution for all sources across quarters (Small-Micro companies)

	First quarter			Second quarter			Third quarter			October		
	Mean	**STD**	**Max**	**Mean**	**STD**	**Max**	**Mean**	**STD**	**Max**	**Mean**	**STD**	**Max**
Yahoo	0	0	0	0	0	0.118	0	0	0.115	0	0.08	1.75
MSN	0.072	0.305	4.62	0	0.004	0.116	0	0.004	0.115	0.05	0.238	2.8
Tiingo	0.736	3.212	15.938	0.635	2.77	13.99	0	0.003	0.114	0	0	0
Stooq	0.359	0.516	2.256	0.248	0.354	1.541	0.135	0.208	1.125	0.066	0.208	1.75

For Stooq, there is an error in every quarter with different values as the mean values in each quarter is greater than zero as well as the STD as shown in Table 5.9. It can be also demonstrated in Figure 5.28 that the errors are not in a single period. But, from the pattern of the error; it can be noticed that the error is not equally distributed between companies. The errors in Stooq are mainly in 9 companies out of the 20 selected ones with 1642 errors as shown in Figure 5.29. while in MSN the errors are in all the companies as shown in Figure 5.30.

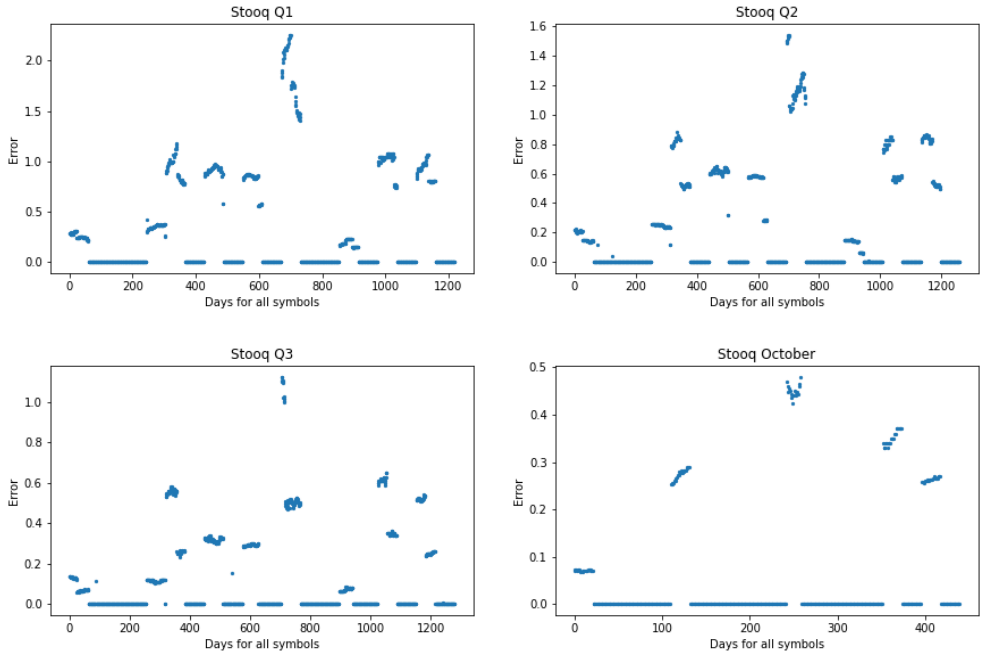

Figure 5.28:Error Distribution in Stooq for each period (Small-Micro companies)

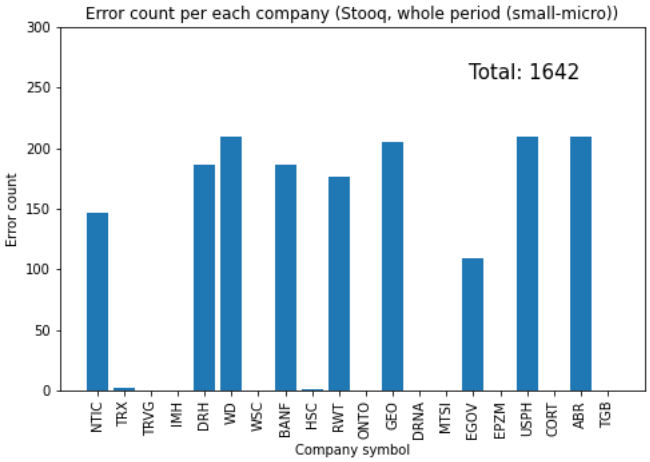

Figure 5.29: The number of errors in Stooq for each company (Small-Micro companies)

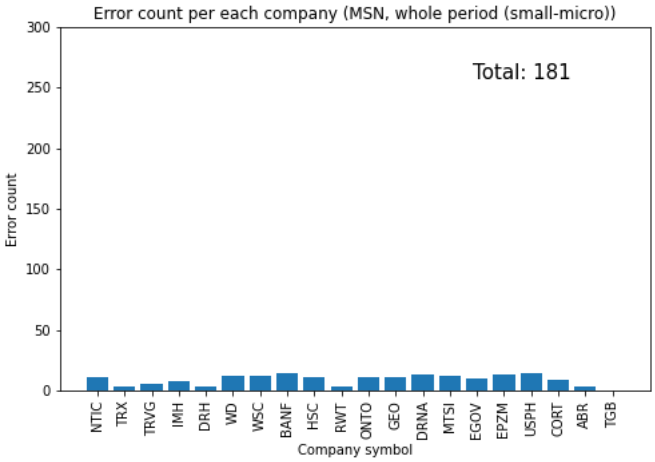

Figure 5.30:The number of errors in MSN for each company (Small-Micro companies)

Regarding Tiingo, Table 5.9 shows that the errors are mainly in Q1 and Q2 only, while Q3 and October month are free of errors with mean and STD almost equal to zero. In Figure 5.31, the error distributions in Tiingo for Q1 and Q2 are shown. It shows the presence of an error in the first part only, in addition, it shows that the high value of the maximum error is not just an outlier but a trend in these periods. Knowing that the x-axis is the days for all the companies in order, therefore from the figure we can say that the error is just in one or a couple of companies. In Figure 5.32, it can be shown that all the errors in Tiingo are mainly in one company and a neglected small error in another one. There is a total of 126 errors and no errors for the other companies. This

company is Northern Technologies International Corporation (NTIC), it is a micro company with 100M market cap. The size of the company could be the reason of this errors, as the traders may not be interested, hence, the sources do not focus on delivering a high quality for these kinds of companies.

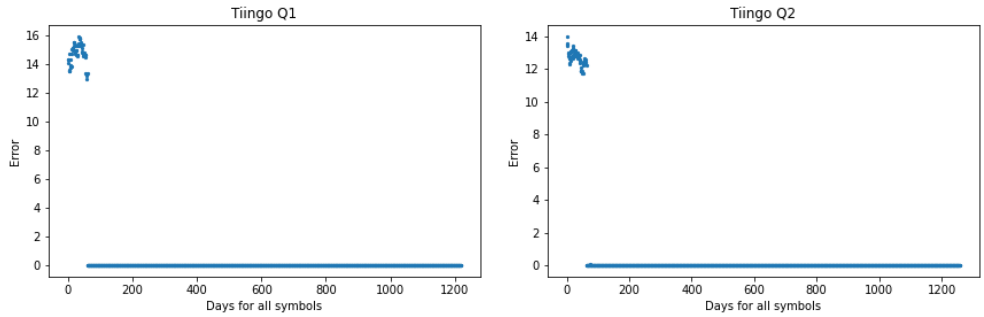

Figure 5.31: Error distribution in Tiingo for the first and the second quarter (Small-Micro companies)

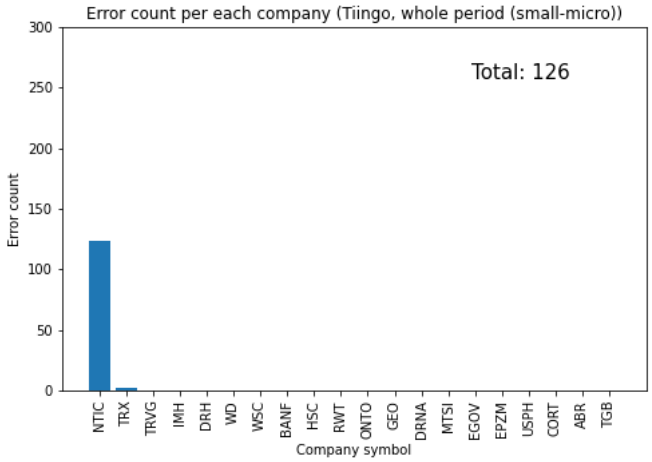

Figure 5.32: The number of errors in Tiingo for each company (Small-Micro companies)

In MSN, it can be determined from the statistical summary of the error in Table 5.9 that the errors are in Q1 and October month. Q2 and Q3 are free of errors with zero mean and STD, even the maximum error values were around 0.11 which is not high. Similarly shown in Figure 5.33 that the errors in MSN are mainly in Q1 and October month. It also shows the same pattern mentioned in the second scenario with the medium companies, where the errors are in the last portion of Q1.

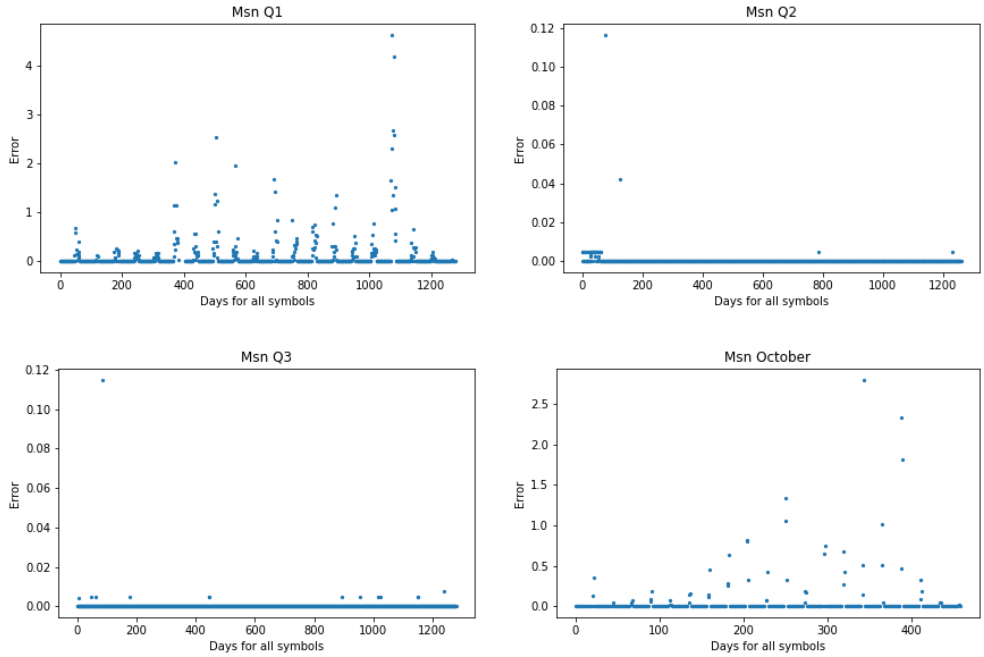

Figure 5.33:Error distribution in MSN for each period (Small-Micro companies)

The number of errors in each source can be represented in Figure 5.34. In Stooq, which is the largest number of errors among all the sources, has 1,642 errors. These errors are approximately equally distributed on the first three quarters. In the first 2 scenarios October month was free of errors, however, in this scenario it has errors. 126 errors were found in Tiingo and as discussed before all of them are mainly in one company. These errors are more or less equally distributed between Q1 and Q2. Regarding MSN, it has 181 errors. Most of them are in Q1. Yahoo has just four errors which is not a huge number that will not affect the accuracy dimension.

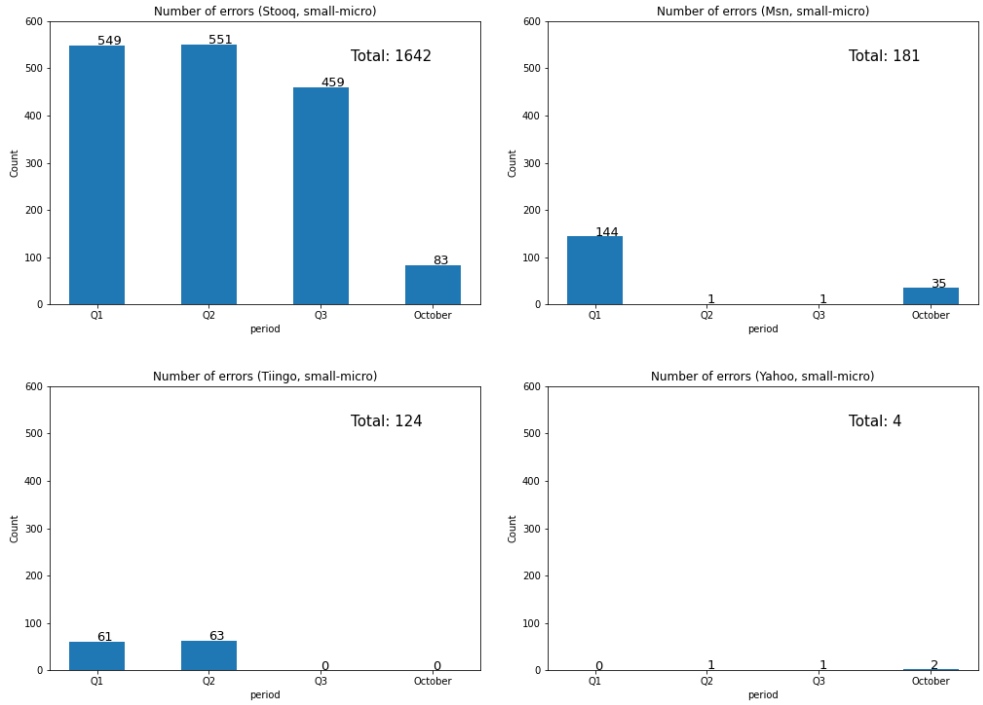

Figure 5.34:the number of errors in each source for each quarter (Small-Micro companies)

Calculating the MSE which is mainly affected by deviation from the true value as mentioned before. As shown in Figure 5.35, the MSE in Stooq is decreasing over the quarters. In MSN, the MSE value in Q1 is very low compared to the one on Stooq. This implies that the errors in MSN are less deviated from the true value than Stooq. The MSE values are zero in all quarters in Yahoo, this is expected as there were no errors in Yahoo dataset. Noticing the huge difference between Tiingo and the other sources. Therefore, the y-axis limit is different from the others. Nevertheless, Stooq has higher number of errors, Tiingo MSE is larger than Stooq. This is because the error values in Tiingo are huge as shown in Figure 5.31. As mentioned before, the errors in Tiingo is just in one company. Hence, if this company removed from the dataset, the number of errors will be zero as well as MSE as shown in Figure 5.36.

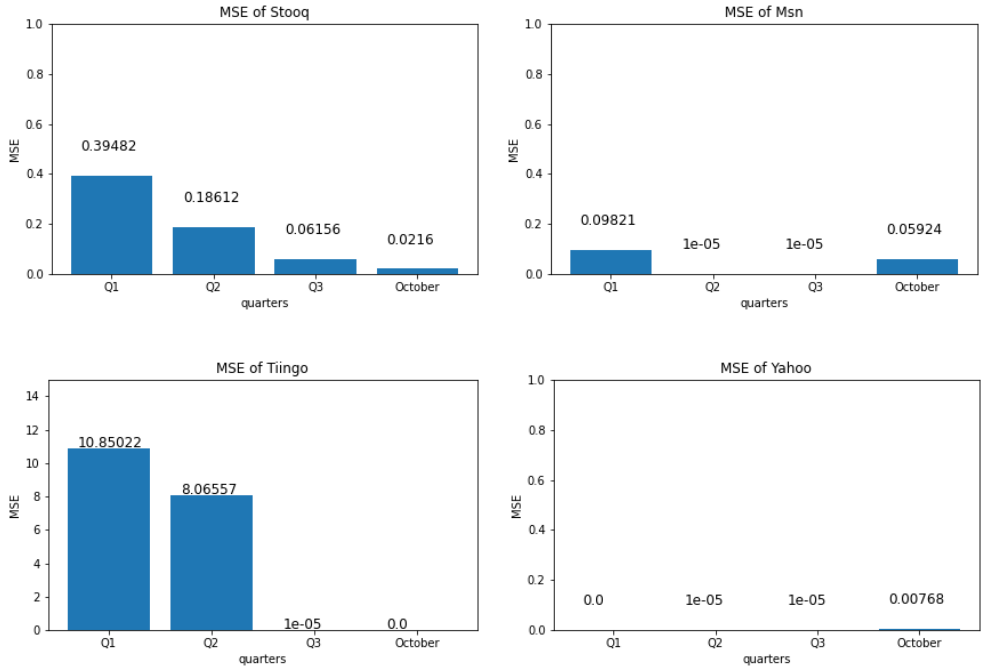

Figure 5.35: MSE in all sources for each quarter (Small-Micro companies)

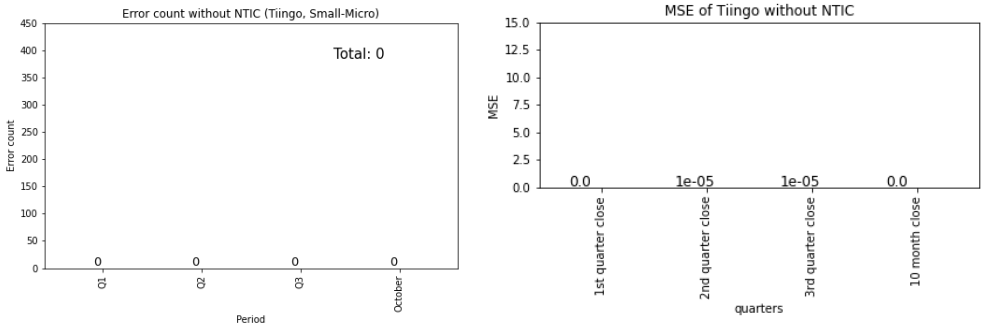

Figure 5.36:The number of errors in Tiingo and MSE without the NTIC symbol (Small-Micro companies)

The accuracy dimension parameters and the accuracy dimension are shown in Table 5.10. Despite the good quality seen in Tiingo by $Accuracy_{count}$, the overall accuracy dimension is not as good because of the $Accuracy_{deviation}$. On the contrary in Stooq, the accuracy dimension is not good because of $Accuracy_{count}$ not the error deviation.

Table 5.10: Accuracy dimensions for all sources (Small-Micro companies)

	Stooq	Msn	Tiingo	Yahoo
$Accuracy_{count}$	0.61	0.96	0.97	1
$Accuracy_{deviation}$	0.9	0.98	0.68	1
$Accuracy$	0.55	0.94	0.66	1

5.3.4 Summary

The dimensions summary for the third scenario can be shown in Table 5.11. The QI in this scenario did not reached the perfect value, all its values are below 1 as shown in the table. This may be because the chosen companies in this scenario were the small and micro ones, that do not have the attention of the data users which lead the sources to not focus on delivering a good quality for these companies.

Table 5.11: Dimensions summary for the Third scenario

	Completeness	Consistency	Accuracy	QI
Stooq	0.99	1	0.55	0.54
Yahoo	0.95	1	1	0.95
MSN	0.98	1	0.94	0.92
Tiingo	1	1	0.66	0.66

5.4 THE THREE SCENARIOS COMPARISON

In all scenarios, the consistency dimension for all the sources is free of glitches. While the completeness dimension has issues in the medium and small-micro companies with no missing values in the top 20 companies. The accuracy dimension has glitches in all the scenarios, especially in Stooq.

It has been noticed that the quality of the source depends on the size of the company from a market capitalization perspective and the companies selected in each scenario. Regarding the size of the company, it was noticed that the QI is getting worse when the company is not from the top companies of the stock market. This applies on all the sources except Stooq, where it has poor quality for all the scenarios.

It is also noticed that the selected companies in the data set affects its quality. NASDAQ stock market has more than 3,300 company listing. As discussed in this

thesis, we have three scenarios with 20 company in each scenario. The glitches found were not in all the companies, for example in Stooq in the first and third scenario the accuracy dimension was affected by 9 companies out of the 20 selected ones and in the second scenario was affected by 14 companies. Similarly, the completeness dimension of Yahoo was affected in the third scenario because of one company while the rest of the companies were of high quality.

Shown in Table 5.12 the QI in each scenario for sources and the average QI which is the final rank of the source. Yahoo is the best source in the available sources with QI of 0.98. It could be perfectly 1 if the ONTO company was not selected in the small and micro companies, which has the missing values that affected the QI of Yahoo. The worst source is Stooq with 0.49 QI. This is very bad value of QI compared to the other sources as the source above it in the ranking is Tiingo with 0.88 QI. The QI value in our case is mainly affected by the accuracy dimension. Because the number of glitches in the other dimensions are not as much high as the accuracy dimension. Not only the number of errors is high but also the deviation. As the lowest STD for the affected sources is 0.2 which is high. Because the stock market is a sensitive data and a value like this may affect making of the decision.

Table 5.12: The QI for all the sources in each scenario

	first scenario	second scenario	third scenario	QI (average)
Yahoo	1	1	0.95	0.98
MSN	1	0.89	0.92	0.94
Tiingo	1	0.99	0.66	0.88
Stooq	0.54	0.4	0.54	0.49

5.5 PREDICTIVE MODEL RESULTS

The predictive model consists of three-layer types: LSTM cells hidden layers, Dense layer and dropout layer. It consists of 4 layers: 2 LSTM hidden layers followed by 2 dense layers. After each hidden layer a dropout layer is added for better generalization. To choose the hyperparameters several trails have been made. We concluded the hyperparameters to be as follows: 20, 1x10-5, 300, 0.4, (100,60), 2 and 180 for batch size, learning rate, number of epochs, dropout rate, number of LSTM cells, number of hidden layers and time steps respectively for the three companies except for GE the time step is set to 60. The implemented model was tested on three

different companies as previously mentioned. These three companies were selected randomly from each scenario, to test the model. The results obtained using the aforementioned hyperparameters are discussed below. The model is evaluated based on the accuracy metric.

The model achieved an accuracy of 85.27%, 93.98% and 94.58% with MSE of 0.002, 0.00039 and 0.004 when applied on the dataset of Alcoa Corp, Almaden Minerals Ltd and GE respectively. The predicted output versus the ground truths for the three companies are shown in Figure 5.37, Figure 5.38, and Figure 5.39. As illustrated, the model was able to predict the underlying patterns and features in the three datasets after the hyperparameter tuning was obtained. The error values could be further improved by exploring different deep learning architectures. These results demonstrate the ability to exploit the model to enhance the data quality by filling in the missing data in the chosen source with the predicted values.

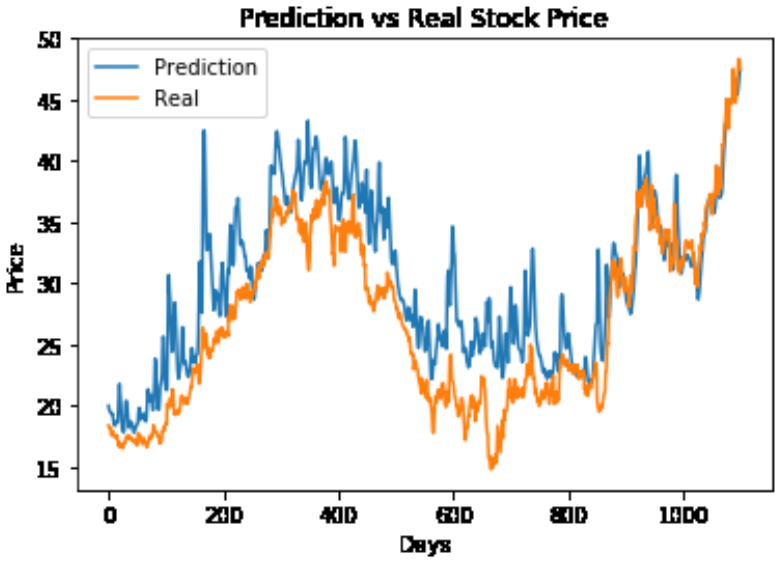

Figure 5.37: The predictive vs the real in Alcoa Corp

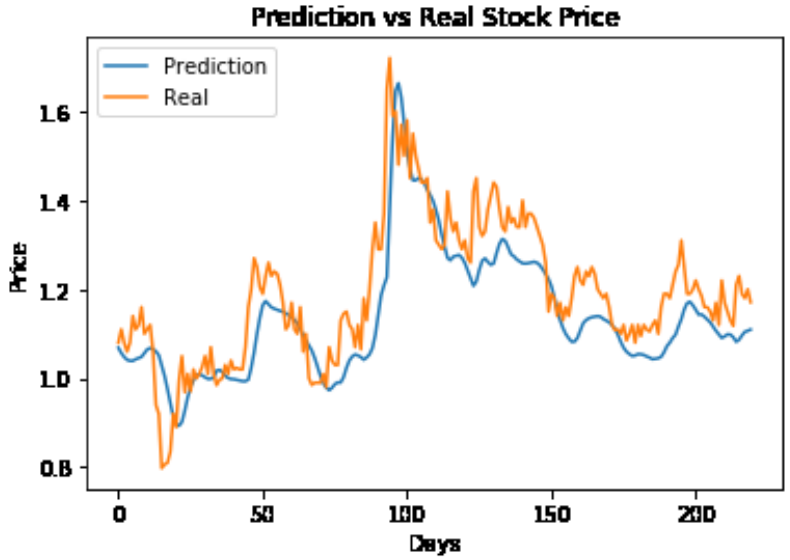

Figure 5.38: The predictive vs the real in Almaden Minerals Ltd

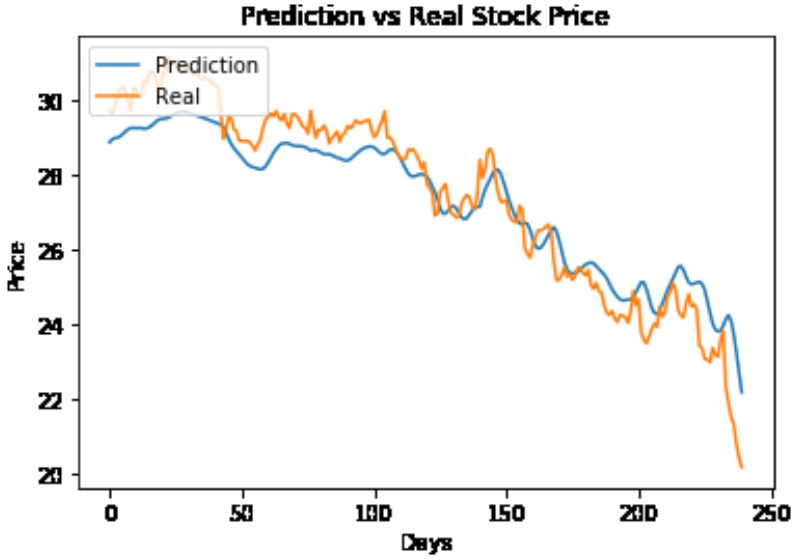

Figure 5.39:The predictive vs the real in GE

Chapter 6: Conclusion and future work

In this chapter we will discuss the conclusion of the thesis and the proposed future work.

6.1 CONCLUSION

Data quality is a crucial aspect in various domains due to its noticeable effect in the decision-making process and in any downstream application, especially in the financial domain. In this thesis, a model with the goal of assessing the DQ of different stock market data sources, ranking them, choosing the most reliable source and identifying the proper dimensions to be used along with the metrics for each dimension is implemented. Moreover, to further enhance the DQ of the chosen source, a predictive model is developed to predict the missing values in that source. The model consists of four main phases: DQ basics phase, data preparation phase, DQ assessment phase, and sources evaluation phase. The data were collected for 60 companies in NASDAQ stock market form four sources, namely, Yahoo Finance, MSN Money, Stooq and Tiingo for the period of 10 months starting from January 2019 to October 2019. These companies are classified into three scenarios, 20 company in each one based on the company's Market Cap in December 2018. The first scenario includes the top 20 companies, the second scenario has the Medium companies and finally the third scenario includes the small-micro companies. In the latter two scenarios, the companies were selected randomly from different geographical locations and different industries.

Based on the selected domain and the commonly used dimensions in literature, we chose three main dimensions to assess the data quality: **completeness, consistency and accuracy**. The consistency dimension has three different types as follows, Type 1, Type 2 and Type 3, classified based on their dependence on the columns and rows. Regarding the accuracy dimension, it consists of two parts; the first part is focusing on the number of incorrect values, while the second is concentrating on the deviation of the value from the ground truth. The dimensions are calculated by a ratio of glitches subtracted by 1. Furthermore, a Quality Index (QI) metric is introduced to rank the sources after measuring the DQ dimensions. It is calculated by multiplying all the dimensions' outputs. It ranges from 0 to 1, 0 is the worst and 1 is the perfect quality.

During data pre-processing phase we did not notice any potential copying from the sources. The quality issues were mainly in the completeness and accuracy dimensions with no glitches in consistency dimension. The final rank of sources was Yahoo Finance, MSN, Tiingo and Stooq with a QI of 0.98, 0.94, 0.88 and 0.49 respectively. It was noticed that the quality of the data is highly depending on the selected companies within the dataset. Additionally, the DQ is affected by the size of the company from the accuracy dimension perspective; as the company size decreases, the worse the DQ becomes. This result is consistent on all sources except for Stooq which has poor quality in all the three scenarios. It was noticed that for Yahoo Finance, the only quality issues were in the missing values especially in the third scenario.

Therefore, a predictive model is developed using LSTM architecture to estimate the stock market price. The aim is to fill in the missing values in two cases: the selected source and the ground truth. In the first case, filling the missing data is used to increase its reliability before using the source. While in the second case, filling in the data would be beneficial if the ground truth had missing values while collecting it. The predictive model consists of three main layer types: LSTM cells hidden layers, Dense layer and dropout layer. The model's input data is collected from 01-03-1970, 01-03-2005 and 01-03-2007 to 31-12-2019 for Alcoa Corp, Almaden Minerals Ltd and GE respectively. After several trials, the chosen hyperparameters identified as 20, $1x10^{-5}$, 300, 0.4, (100,60), 2 and 180 for batch size, learning rate, number of epochs, dropout rate, number of LSTM cells, number of hidden layers and time steps respectively for the three companies except for GE the time step is set to 60. The accuracy achieved by the model is 85.27%, 93.98% and 94.58% when applied on the dataset of Alcoa Corp, Almaden Minerals Ltd and GE respectively. Hence, the implemented model showed a promising capability to predict the stock market price and fill in the missing values in the selected source.

6.2 FUTURE WORK

More companies could be added to assess the DQ of the sources, if possible, include the 3,300 companies in NASDAQ stock market to enhance the ranking results. Regarding the predictive model, it could be developed for different companies. We recommend to initially use the proposed architecture and hyperparameters, then changed based on the input dataset.

References

[1] McKinsey Global Institute (MGI) Report: The age of analytics: Competing in data-driven world, https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world

[2] C. W. Fisher, B. R. Kingma: Criticality of DQ as exemplified in two disasters, Elsevier, vol.39, no. 2, pp. 109-116, December 2001

[3] https://www.ibmbigdatahub.com/infographic/four-vs-big-data

[4] T. C. Redman: "The Impact of Poor Data Quality on the Typical Enterprise", Communications of the ACM, vol. 41, no. 2, February 1998.

[5] T. Dasu, R. Duan, D. Srivastava: "Data Quality for Temporal Streams", in IEEE Data Engineering Bulletin Journal, Vol. 39, no. 2, pp. 78-92, June 2016.

[6] X. Li, X. L. Dong, K. Lyons, W. Meng, D. Srivastava: "Truth Finding on the Deep Web: Is the Problem Solved?", in Proceedings of the Very Large Data Base Endowment Journal, vol. 6, no. 2, pp. 97-108, December 2012.

[7] https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

[8] P. Russom: "Big Data Analytics." TDWI Best Practices Report, The Data Warehousing Institute (TDWI) Research, fourth quarter 2011.

[9] N. Abdullah, S. A. Ismail, S. Sophiayati, S. M. Sam: "Data quality in big data: A review", in International Journal of Advances in soft computing and its applications, vol. 7, no.3, pp. 16–27, November 2015.

[10] J. G. Geiger: "DQ Management The Most Critical Initiative You Can Implement", in SUGI 29, Montreal, May 2004.

[11] T. C. Redman: "Data Driven: Profiting from Your Most Important Business Asset", published by Harvard Business Press, September 2008.

[12] https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year

[13] https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards

[14] https://www.sas.com/en_us/insights/articles/data-management/data-quality-management-what-you-need-to-know.html

[15] D. Loshin: “The Practitioner’s Guide to DQ Improvement”, Elsevier, published by Morgan Kaufmann Publishers, October 2010.

[16] Y. Wand, R. Y. Wang: “Anchoring DQ dimensions in ontological foundations”, Communications of the ACM, vol. 39, no. 11, pp. 86-95, November 1996.

[17] R. Y. Wang, D. M. Strong: “Beyond accuracy: what DQ means to data consumers”, Journal of Management Information Systems, vol. 12, no. 4, pp. 5–33, March 1996.

[18] T. C. Redman: “DQ for the Information Age”, Published by Artech House Computer Science Library, 1997.

[19] A. McAfee, E. Brynjolfsson: “Big data: The management revolution”, Harvard Business Review, vol. 90, no. 10, pp. 60-68, October 2012.

[20] K. Kambatla, G. Kollias, V. Kumar, A. Grama: “Trends in big data analytics”, in Journal of Parallel and Distributed Computing - In Press - Corrected Proof, vol. 74, no.7, pp. 2561-2573, July 2014.

[21] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, T. Stoica, M. Zaharia: “A view of cloud computing”, in Communications of the ACM, vol. 53, no. 4, pp. 50-58, April 2010.

[22] J. Tee: “Integrate or disintegrate: How to keep your big data strategy from falling apart”, in The Server-Side website, http://-www.theserverside.com/feature/Handling the-four-Vs-of-big-data-volume-velocity-varietyand-veracity, 2013.

[23] I. Caballero, M. Serrano, M. Piattini: “A DQ in Use Model for Big Data”, in Future Generation Computer Systems, vol. 63, pp. 123-130, October 2016.

[24] R. Y. Wang: “A product perspective on total DQ management”, in Communications of the ACM, vol. 41, no. 2, pp. 58–65, Feb. 1998.

[25] C. Abrams, J. von Känel, S. Müller, B. Pfitzmann, S. R. Taylor: “Optimized enterprise risk management”, in IBM Systems Journal, vol. 46, no. 2, pp. 219-234, April 2007.

[26] Y. W. Lee, D. M. Strong, B. K. Kahn, R. Y. Wang, “AIMQ: a methodology for information quality assessment”, in Information & Management, vol. 40, no. 2, pp. 133–146, December 2002.

[27] F. Sidi, P. H. Shariat Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, A. Mustapha: “DQ: A survey of DQ dimensions”, in International Conference on Information Retrieval & Knowledge Management, pp.300-304, March 2012.

[28] J. S. Ryu, J. S. Park, and J. H. Park: “A DQ Management Maturity Model”, in ETRI Journal, vol. 28, no. 2, pp. 191–204, April 2006.

[29] C. Batini, C. Cappiello, C. Francalanci, A. Maurino: “Methodologies for DQ assessment and improvement”, in ACM Computing Surveys, vol. 41, no. 3, pp. 1–52, July 2009.

[30] I. Jaya, F. Sidi, L. S. Affendey: “A review of DQ research in achieving high DQ within organization”, in Journal of Theoretical and Applied Information Technology, vol. 95, no. 12, pp. 2647-2657, June 2017.

[31] M. Mirzaie, B. Behkamal, S. Paydar: “State of the Art on the Quality of Big Data: A Systematic Literature Review and Classification Framework”, in Arxiv, April 2019.

[32] M. Belhiah, “A User-Centered Model for Assessing and Improving Open Government Data Quality”, in MIT International Conference on Information Quality, October 2017.

[33] M. A. Serhani, H. T. El Kassabi, I. Taleb, and A. Nujum, “A Hybrid Approach to Quality Evaluation across Big Data Value Chain”, in 5th IEEE International Congress on Big Data, pp. 418–425, June 2016.

[34] T. Hongxun, W. Hong-gang, Z. kun, S. mingati, L. Hao-song, X. Zhong-ping, K. Taifeng, L. Jin, C. Ya-ai: “Data quality assessment for on-line monitoring and measuring system of power quality based on big data and data provenance theory”, in IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), pp. 248–252, April 2018.

[35] D. Ardagna, C. Cappiello, W. Samá, M. Vitali: “Context-aware data quality assessment for big data”, in Future Generation Computer Systems, vol. 89, pp. 548–562, December 2018.

[36] A. Immonen, P. Paakkonen, E. Ovaska: “Evaluating the Quality of Social Media Data in Big Data Architecture”, in IEEE Access, vol. 3, pp. 2028–2043, July 2015.

[37] I. Taleb, M. A. Serhani, “Big Data Pre-Processing: Closing the Data Quality Enforcement Loop”, in IEEE International Congress on Big Data (Big Data Congress), pp. 498–501, December 2017.

[38] A. Klein, W. Lehner: “Representing Data Quality in Sensor Data Streaming Environments”, in Journal of Data and Information Quality, article, vol, 1, no. 2, article No.: 10, September 2009.

[39] R. Vaziri, M. Mohsenzadeh, J. Habibi: “Measuring data quality with weighted metrics”, in Journal Total Quality Management & Business Excellence, Vol. 30, no.5-6, pp. 708-720, June 2019.

[40] O. Azeroual, G. Saake, J. Wastl: “Data measurement in research information systems: metrics for the evaluation of data quality”, in Scientometrics, vol. 115, no. 3, pp. 1271-1290, April 2018.

[41] R. Gitzel, S. Turrin, S. Maczey, S. Wu, B. Schmitz: “A Data Quality Metrics Hierarchy for Reliability Data”, in the 9th IMA International Conference on Modelling in Industrial Maintenance and Reliability, pp. 12-14, July 2016.

[42] M. Helfert, C. Herrmann.: “Proactive data quality management for data warehouse systems”, in Proc. of the 4th International Workshop on Design and Management of Data Warehouses, May 2002.

[43] B. Behkamal, M. Kahani, E. Bagheri, Z. Jeremic: “A Metrics-Driven Approach for Quality Assessment of inked Open Data”, in Journal of Theoretical and Applied Electronic Commerce Research, vol. 9, no. 2, pp. 806-816, May 2014.

[44] N. G. Weiskopf, C. Weng: “Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research”, in Journal of the American Medical Informatics Association, vol. 20, no. 1, pp. 144-151, January 2013.

[45] M. Bovee, R. P. Srivastava, B. Mak: “A conceptual framework and belief function approach to assessing overall information quality”, in International Journal of Intelligent Systems, vol. 18, no. 1, pp. 51–74, January 2003.

[46] M. Schaal, R. M. Mueller, R. MacLean: “Information Quality Dimensions for the Social Web”, in the International Conference on Management of Emergent Digital Ecosystems (MEDES), pp. 53-58, October 2012.

[47] V. Jayawardene, S. Sadiq, M. Indulska: “An Analysis of Data Quality Dimensions”, ITEE technical report, no. 2, 2015.

[48] B. K. Kahn, D. M. Strong, and R. Y. Wang, “Information quality benchmarks: product and service performance”, Communications of the ACM, vol. 45, no. 4, pp. 184–192, April 2002.

[49] C. Batini and M. Scannapieco: “Data Quality: concepts, methodologies and techniques”. Springer, 2006.

[50] P. H. S. Panahy, F. Sidi, L. S. Affendey, M. A. Jabar: “The Impact of Data Quality Dimensions on Business Process Improvement”, in 4th World Congress On Information And Communication Technologies, December 2014.

[51] L. F. Ramos-Lima, A. C. G. Maçada, X. Koufteros: “A Model for Information Quality in the Banking Industry - The Case of the Public Banks in Brazil”, in Proceedings of the 12th International Conference on Information Quality, November 2007.

[52] H. T. Moges, K. Dejaeger, W. Lemahieu, B. Baesens: “A multidimensional analysis of data quality for credit risk management: New insights and challenges”, in Information & Management, vol. 50, no. 1, pp. 43-58, January 2013.

[53] Y. W. Lee, D. M. Strong, B. K. Kahn, R. Y. Wang: “AIMQ: A Methodology for Information Quality Assessment. Information & Management”, in Elsevier, vol. 40, no. 2, pp. 133-146, December 2002.

[54] A. F. Haryadi, J. Hulstijn, A. Wahyudi, H. van der Voort, M. Janssen: “Antecedents of Big Data Quality; an Empirical Examination in Financial Service Organizations”, in IEEE International Confernce on Big Data, pp. 116-121, December 2016.

[55] K. Dejaeger, B. Hamersb, J. Poelmansa, B. Baesensa: “A Novel Approach to the Evaluation and Improvement of Data Quality in the Financial Sector”, in Proceedings of the 15th International Conference on Information Quality (ICIQ), November 2010.

[56] P. Missier, G. Lalk, V. Verykios, F. Grillo, T. Lorusso, P. Angeletti: “Improving Data Quality in Practice: A Case Study in the Italian Public Administration”, in Distributed and Parallel Databases, vol. 13, pp.135–160, March 2003.

[57] D. Srivastava, S. Venkatasubramanian: “Information theory for data management”, in Proceedings of the VLDB Endowment, vol. 2. pp. 1662-1663, August 2009.

[58] A. V. Devadoss, T. A. A. Ligori: “Forecasting of stock prices using multi-layer perceptron”, in International Journal of Computing Algorithm, vol. 2, pp. 440–449, December 2013.

[59] V. K. Menon, N. C. Vasireddy, S. A. Jami, V. T. N. Pedamallu, V. Sureshkumar, K. Soman: “Bulk price forecasting using spark over NSE data set”, in International Conference on Data Mining and Big Data, pp. 137–146, June 2016.

[60] G. E. Box, G. M. Jenkins, G. C. Reinsel, G. M. Ljung: “Time series analysis: forecasting and control”, published by John Wiley and Sons Inc., June 2015.

[61] G. Batres-Estrada: “Deep learning for multivariate financial time series”, ser. Technical Report, Stockholm, June 2015.

[62] Y. Bengio, I. J. Goodfellow, A. Courville, “Deep learning”, Nature, vol. 521, pp. 436–444, 2015.

[63] X. Ding, Y. Zhang, T. Liu, J. Duan, “Deep learning for event-driven stock prediction.”, in International Joint Conference on Artificial Intelligence, pp. 2327–2333, September 2015.

[64] Y. Baek, H. Y. Kim: “ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module”, in Expert Systems with Applications, vol. 113, pp. 457–480, December 2018.

[65] S. Hochreiter, J. Schmidhuber: “Long short-term memory”, in Neural computation, vol. 9, no. 8, pp. 1735–1780, November 1997.

[66] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell: “Long-term recurrent convolutional networks for visual recognition and description”, in IEEE conference on computer vision and pattern recognition, pp. 2625–2634, May 2015.

[67] O. Zuravicky: “The Stock Market: Understanding and applying ratios, decimals, fractions, and percentages”, published by The Rosen Publishing Group, Inc., August 2004.

[68] R. J. Teweles, E. S. Bradley: “The Stock Market”, Published by John Wiley and Sons, Inc., September 1998

[69] L. T. B. Ngoc: “Behavior Pattern of Individual Investors in Stock Market”, in International Journal of Business and Management, vol. 9, no. 1, December 2013.

[70] M. Campello, J. R. Graham: “Do stock prices influence corporate decisions? Evidence from the technology bubble”, in Journal of Financial Economics, vol. 107, no. 1, pp. 89-110, January 2013.

[71] Y. zhang, M. f. Wiersema: “stock market reaction to CEO certification: the signaling role of CEO background”, in Strategic Management Journal, vol. 30, no. 7, pp. 693-710, April 2009.

[72] T. Duso, D. J. Neven, L. H. Roller: “The Political Economy of European Merger Control: Evidence using Stock Market Data”, in Journal of Law and Economics, vol. 50, no. 3, pp. 455-489, August 2007.

[73] J. Clayton, G. Mackinnon: “The Relative Importance of Stock, Bond and Real Estate Factors in Explaining REIT Returns”, in Journal of Real Estate Finance and Economics, vol. 27, no.1, pp. 39–60, July 2003.

[74] H. Choe, B. C. Kho, R. M. Stulz: “Do foreign investors destabilize stock markets? The Korean experience in 1997”, in Journal of Financial Economics, vol. 54, no. 2, pp. 227-264, October 1999.

[75] M. J. Seiler, W. Rom: “A Historical Analysis of Market Efficiency: Do Historical Returns Follow a Random Walk?”, in Journal of Financial and Strategic Decisions, vol.10, no. 2, pp. 49-57, summer 1997.

[76] A. Porshnev, I. Redkin, A. Shevchenko: “Machine learning in prediction of stock market indicators based on historical data and data from Twitter sentiment analysis”, in IEEE 13th International Conference on Data Mining Workshops, pp. 440-444, December 2013.

[77] https://www.nasdaq.com/

[78] https://colab.research.google.com/notebooks/intro.ipynb

[79] https://www.investopedia.com/terms/q/quarter.asp