TÉCNICO LISBOA

A Method for Improving Business Data Analysis

Pedro Miguel Pires Torres

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Miguel Leitão Bignolas Mira da Silva
Prof. Cláudia Martins Antunes

Examination Committee

Chairperson: Prof. David Manuel Martins de Matos
Supervisor: Prof. Miguel Leitão Bignolas Mira da Silva
Member of the Committee: Prof. Henrique São Mamede

January 2021

Acknowledgments

To my three most important pillars, my deepest and most sincere thank you, for not giving up, for standing strong for me and for being my strength to propel me forward to finish this degree. It was not easy but you sure made it better with laughs and tears along the way. I love you all so much, thank for all the countless sacrifices you have made for me. If I could I'd give you the world. To Luísa, a thank you from the bottom of my heart, as she made me the man I am today, without her I would be nothing. To Margarida, the core of the family, the strongest of us all that always finds energy and love to raise all spirits. To Manel, for his kindness and support shown in his own way. This is the culmination of all our efforts. No thank you's will ever suffice.

To the love of my life, my Sophie, also without whom none of this would have been possible. She kept me strong and endured with me until the end. Thank you for all the love, silly and laughter which definitely helped in such an... intense experience. I love you more than anything.

To all my dear godchildren who helped me grow and develop more than I could have imagined.

To all my friends for the endless nights working, studying and developing projects.

To all my colleagues for being the best classmates one could ask for and with whom was a pleasure to pass my time in this great Institution.

This work could not have been completed without the crucial help of my advisers. As they provided essential help in the development of this work.

Also, a very special thank you to Luís who was also essential to the development and success of this work. Thank you for all the support and for believing in this new endeavour, I'm truly thankful.

To each and everyone that crossed my way these last few years, thank you.

Disclaimer: The views expressed in this Master Thesis are those of the author and may not necessarily represent the views of the involved institutions.

Abstract

Businesses are now, more than ever, facing unprecedented challenges. One of these challenges concerns the use of data inside organizations, many of which are known to use spreadsheet programs to manage, store, cleanse and create reports on. This has proven to be far from ideal when taking into account data validity and process efficiency. In this Master thesis we propose a generic solution to address this problem and apply it in a specific context. We followed a Design Science Research Methodology and performed two demonstrations (using two different approaches) to solve this problem. The first demonstration focused on descriptive analytics and the second one on predictive analytics. The evaluation of the system was done both with a DSRM evaluation framework as well as two metrics for measuring error (Mean Absolute Error and Weighted Absolute Percentage Error). We gathered that the two approaches yielded good results when compared to error-prone traditional spreadsheets and encouraging results in terms of the viability of predictions in this context. The predictions themselves yielded good to poor results depending on the technique used, which we believe this was due to the limited amount of data used. This work was developed in a Portuguese public finance organization and used datasets of the same domain.

Keywords

Data Analysis, Business Intelligence, Data Science, Forecasting, Public Finance

Resumo

A área de negócios, agora mais do que nunca, está a econtrar desafios inimagináveis. Um destes desafios é relativamente à forma como os dados são usados dentro das organizações, algumas das quais usam programas como folhas de cálculo para gerir, guardar, limpar e para criar relatórios. Isto tem se provado longe de ser ideal tendo em conta a validade dos dados e eficiência dos processos. Nesta Tese de Mestrado, propomos uma solução genérica para endereçar este problema aplicando-a a um contexto específico. Seguimos uma Design Science Research Methodology e efetuámos duas demonstrações (usando abordagens diferentes) para resolver este problema. A primeira foca-se em descriptive analytics e a segunda em predictive analytics. A avaliação deste sistema foi feita com ambas as técnicas de uma framework de avaliação DSRM e duas métricas para medições de erro (Mean Absolute Error and Weighted Absolute Percentage Error). Constatámos que as duas abordagens deram bons resultados em comparação com as tradicionais e falíveis folhas de cálculo e deram resultados promissores em termos da viabilidade de previsões neste contexto. As previsões produziram resultados bons a fracos dependendo da técnica usada. Este trabalho foi desenvolvido numa organização portuguesa de finanças públicas e os conjuntos de dados utilizados foram do mesmo domínio.

Palavras Chave

Análise de Dados, Business Intelligence, Ciência de Dados, Previsões, Finanças Públicas

Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence
AIC	Akaike information criterion
BI	Business Intelligence
BIC	Bayesian information criterion
ETL	Extract, Transform and Load
DSRM	Design Science Research Methodology
HR	Human Resources
IMF	International Monetary Fund
IS	Information Systems
IT	Information Technology
KDD	Knowledge Discovery in Databases
MAE	Mean Absolute Error
SCD	Slowly Changing Dimensions
SNS	Serviço Nacional de Saúde
WAPE	Weighted Absolute Percentage Error

1

Introduction

Contents

In business, like in many other fields, there is a high dependency on what technologies are available to us at a certain point in time. As such, and in the fields of work that are data driven development was always held captive by the limitations of the technologies of in use. One practical example of what we describe here happens with many companies where the use of spreadsheet programs such as Microsoft Excel is common practice. One of the focal points of this Master Thesis will be precisely addressing problems incurred in using tools like spreadsheet programs to do everything data-related. Data driven companies' biggest challenge is the transformation of great amounts of data into valuable insights or information to support decision. These needs led to the obligations of storing data to be of use in the future as well as the creation of a relatively new field of study. Concerning the importance that data has gained over the past few years, companies are doing more than ever with this data. Yet, this task may be harder than it appears at a first glance.

Companies, and in particular older companies are now, more than ever in need of keeping up with the ever-changing world of technology. Particularly in these last few months when an unpredictable event brought uncertainty on all levels. Made us adapt and challenged us in ways we didn't think possible. The digital era brought unprecedented potential but also brought the need for further knowledge when applying these new technologies [3].

With new technologies to support this data revolution we see companies facing the need to evolve to support this evolution. Often times, companies may struggle in this transition to a more digital world; and although already using fair technologies these are becoming more and more dated over time. Technologies like Microsoft Excel may be far from obsolete but have their own limitations. Nowadays we can find tools that outperform Excel in what companies need to do with their data with less errors and more efficiently.

More specifically one of the most prominent problems these companies face is the over-use of these spreadsheets. If containing a lot of data they can become almost impossible to read and even more so to find and correct errors. An alternative for tasks like those performed in Excel are relational databases which can be a more intelligent approach to problems like the one just mentioned. By having a database one could design the database schema and include integrity constraints or rules for the database to accept only correct values.

Despite the advantages an approach like relational data bases can bring this transition is never simple or easy, since it implies having costly qualified personal to design and manage these databases. However, the challenges this change may pose, the potential certainly seems to be worth the difficulties. These changes are what can be named as digital transformation and in implying great new opportunities it can also have some threats as evidenced by some studies [4] and [5]. This change requires that these older companies, that met their success in the pre-digital economy, rethink the way they are to compete in this new digital economy. Digital transformation aims to place these organizations back again as

leaders.

Further addressing the digital transformation that was involved in this work we briefly introduce the area as well as some key concepts used in this work. This is followed by the problem definition of the endless number of spreadsheets companies have with the most varied types of data. This leads us to believe there is a clear need for modernization of older technologies to fix current problems.

For a long time, and even still in some cases, spreadsheets have been the perfect tool for the jobs of crunching numbers and making simple calculations. Tools like Microsoft Excel have a lot of potential but also a lot of limitations, as is to be expected. Excel is far from being perfect for all the jobs that involve numbers and it's especially lacking when handling large datasets meant to be used for data visualization. Thus, tools like relational databases seem to be much more adequate for that job, for instance.

1.1 Research Context

In this section, we address the notion and theory behind digital transformation due to its relevance in the work we developed.

At a more abstract level, Digital Transformation, includes all the changes that happen in a company through the use of digital technologies. It can also be understood as "a process that aims to improve an entity by triggering significant changes to its properties through combinations of information, computing, communication, and connectivity technologies" [6]. In other words, digital transformation is the use of new technologies to enable the improvement of businesses on the most varied fronts of action.

According to David Rogers [7], a globally recognized expert on the subject, Digital Transformation is reshaping five key domains of strategy: customers, competition, data, innovation and value [7]. These domains describe the landscape of digital transformation in business nowadays.

One of the most important domains for this work is data and how businesses produce, manage and utilize it. Historically, data could be produced either out of customer surveys or inventories in the context of operations, sales or marketing (for example). This type of data was mainly used for evaluating, forecasting and decision making. In opposition, today, we are facing a tidal wave of data produced for the most varied purposes and reasons. The production of data at unprecedented rates is also likely to be unstructured having to be used incrementally with new analytical tools [7]. This subject addresses a similar problem to that of this Master Thesis, through this data, companies can make new kinds of predictions and extract new knowledge that was previously unavailable.

Data is progressively becoming the "lifeblood of every department and a strategic asset to be developed and deployed over time". It is a vital part of how businesses operate nowadays, and also help in their differentiation in the markets while generating value [7].

According to Rogers [7], and aligned with the wide availability and access of data nowadays, the

value of data is in how it can be turned into an asset [7]. To achieve this, there is a need to select the right data and use it correctly to generate the biggest business value possible.

Concerning this, we are currently living in what is often called the digital age of data. These new data oriented technologies are not only changing the way we live our daily lives but are also changing the dynamics of our businesses for organizations of any size [7]. The adoption of these changes requires intention and effort as a whole from the organization.

The author dwells even deeper on the potential of data and how to make use of it in the new digital era. With the vast amounts of data produced nowadays it is easier than ever before to access and to create new business value from that information. An example provided in the book, of a weather channel that was able to adapt to the digital era, explains the ways they are using old forms of data to capitalize it in new ways such as ads for medication, insurance companies and even food all based on weather data.

With this new data potential, there is a need to rethink the way data exist inside organizations since generating data is often the easiest part. We also know that a great amount of data is already produced from outside the organization. The real challenge lies in truly making data a powerful insight for the benefit of the organization. Also **"traditional analytics based on spreadsheets has given way to big data, where unstructured information joins with powerful new computation tools."** Yet for this, businesses need to change they way they think of data and turn it, truly, into a key strategic asset [7].

Moreover, Rogers defends that the cited principles must be followed to guide a companies' data strategy: gather diverse data types, use data as a predictive layer in decision making, apply data to new product innovation, watch what customers do, not what they say and combine data across silos. One should also bear in mind that what drove needs in the pre-digital era may not be what is needed now, in other words, datasets are fundamentally different from the ones in spreadsheets in the pre-digital era [7].

Other authors [8] also argue that digital transformation is not about the technology, in fact out of the $1.3 trillion, $900 billion are believed to have gone to waste. According to their experience Digital Transformation worked because they focused on the fundamentals, their priority was to change the mindset of its members, as well as the culture, to encompass this new strategy and only then decided on the digital tools and how to make use of them. Five lessons of success are presented on the paper to support their conclusions. They believe that what drove the technology was the vision people held of the future and not the the other way around.

Finally, one very interesting aspect of this master thesis is the polychotomy of disciplines and subjects due to the universality of its theme. These are at the top digital transformation due to all the changes it implies for organizations and their environments, fields like machine learning, database theory or visualization of information but also, the management of change that is involved in the process.

1.2 Research Problem and Proposal

Briefly put, the research problem we address aims to tackle ”spreadsheet hell”, meaning the endless, confusing, error-full spreadsheets that are used to store, treat, analyze, cross and visualize data. This problem affects many companies that are heavily dependent of spreadsheets as the lifeblood for their business.

Our research proposal, which will be detailed later on, is based on a full system architecture to solve all the implications of our research problem. This system encompasses the stages from the reception of the data all through to its deliverance and consumption by the target audience whether that is consumers or a board of directors.

In the following section we address the identification of the problem and motivation, a component of the chosen methodology - DSRM. In Chapter 4 we acknowledge the existence of a problem and what needs to be addressed.

1.3 Research Methodology

For this work we decided to follow the Design Science Research Methodology, DSRM. This choice was due to it being an iterative method adaptable to our research problem, helping us to develop a solution for the problem in question.

This methodology gives us a process model to apply in our work and a methodic way to guide or research. Hevner et al. [9], in 2004, proposed 7 guidelines to further understand the requirements for effective design-science research.

The reason for this choice was due to the popularity of DSRM in applied resource disciplines as the one of this work. As these disciplines apply theories from many areas the use of a carefully defined methodology was recommended by the literature [9].

It makes particular sense for this project, by belonging to the field of Information Systems and since we want to solve an Information Technology (IT) problem [1].

DSRM provides us with a form of creation and evaluation of artifacts, being an iterative process allows for the better accompaniment of our solution. It is comprised of 6 steps [1] that we will here briefly introduce and cover more extensively later:

1. **Problem identification and motivation** (definition of the problem to be approached and the value the proposed solution could have. This definition should be used to help develop and model the artifact that will eventually become the solution.)

2. **Definition of the objectives for a solution** (set the objectives for the solution based on the previous step, problem definition and motivation, and what can the solution provide in reality, based

on the analyzed related work)

3. **Design and development** (conceive the artifact and the functionality it is expected to have and then implement it, afterward create the actual artifact with the performed research in mind)

4. **Demonstration** (showcase the use of the artifact that has been developed to try and solve the problem, must include the way the artifact solves the problem)

5. **Evaluation** (measurement of how appropriate the developed artifact is to solve the proposed problem; compare the obtained results with the pre-defined objectives of the solution)

6. **Communication** (communication of the components of all the aforementioned steps such as the problem and its importance, proposed solution, artifacts, and their utility and novelty, the rigor of the design as well as the effectiveness of all the research)

In Figure 1.1 there is a DSRM process scheme adapted to our research.

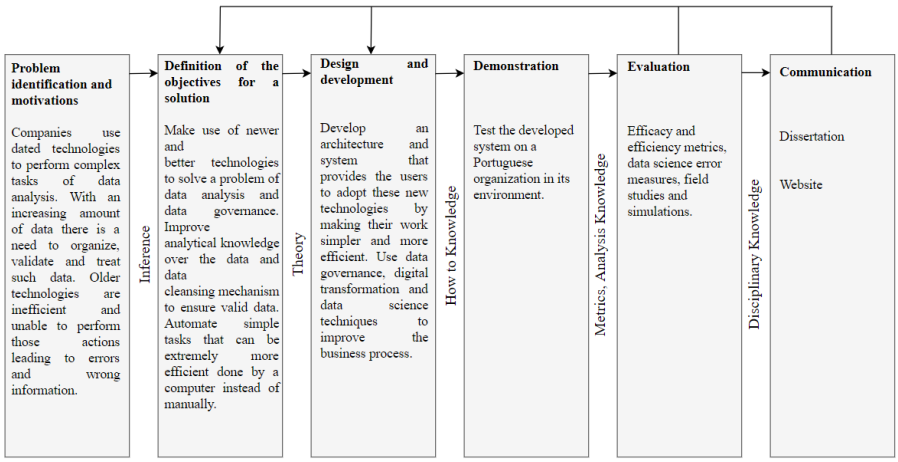

Figure 1.1: DSRM process (adapted from [1])

1.4 Organization of the Document

This Master Thesis is organized as follows: Chapter 2 presents the research problem, in Chapter 3 we described the background, afterwards, in Chapter 4 we have the research proposal and then the demonstration, in Chapter 5. Chapter 6 is dedicated to the evaluation of the work developed and in Chapter 7 we delve on the conclusions and future work.

2

Research Problem

Following the DSRM methodology, the research should start with the identification of the problem and the underlying motivation. As such we will do so in this section.

In the present Chapter we will be addressing in-depth the research problem of this Master Thesis.

Users in business make use of information that is stored in databases and use that data to extract knowledge to make business decisions as informed as possible [10]. Nowadays, databases are part of companies' enterprise architecture model [11]. In theory, these databases are generally simple and easy to use/maintain [12]. The term databases here is to be understood as a tabular design and not relational databases which we will dwell into further on.

This use of tables to store and use information has been a common practice for quite some time and even more the use of spreadsheet programs. A notorious example of a spreadsheet program is Microsoft Excel, in fact, it is considered to be the industry-leading spreadsheet program. It is known to be a number cruncher but is also often used for data visualization and analysis [13].

Spreadsheets are known as the most widely used programming systems in the world, these are used for businesses and personal use for a very wide variety of purposes. From simple calculations to complex financial models [14]. These types of tools although having quite a lot of potential are very error-prone and, depending on the case, the impact can vary from meaningless to being considered as one of the causes for the 2008 financial crisis [15] (although there were more substantial causes [16]). This has led to some research on the subject and the surge of possible solutions to avoid errors [14].

The problem we address is directed to spreadsheet programs, such as Microsoft Excel. When these are being used beyond their capabilities and possibly incur in errors or complications when, clearly, other tools would be more suitable.

The limitations of these tools has been widely studied [15], with countless examples [17] of error stories with unprecedented consequences. From governments to banks that were and still are very dependent on such fallible technologies. Yet these spreadsheets are still pointed out to be "integral to the function and operation of the global financial system".

In more detail, the main known risks of spreadsheets can include: human error, fraud, overconfidence, interpretation and archiving [15]. This article indicates that 90% of spreadsheets contain errors mainly because these spreadsheets are rarely tested, even recent studies point to, about 50% of spreadsheet models used in large companies, having defects. Due to the mix of program code and data, spreadsheets appear to be the perfect environment to perpetrate fraud. Once more, due to spreadsheets not being checked for errors these are not found/fixed. This can be due to the overconfidence employees place on it. The translation of a business problem into a spreadsheet can lead to issues regarding the interpretation decision-makers have on said data. An example of problems in archiving is the case of failed Jamaican commercial banks [18], poor archiving can lead to weakness in spreadsheet control which, in turn, can lead to operational risk [15].

”The use of spreadsheets in business is a little like Christmas for children. They are too excited to get on with the game to read or think about the ’rules’ which are generally boring and not sexy” (D. Chadwick, 2000). From this quote we can understand how people in business tend to handle the complex use of spreadsheets leading to the problems already mentioned. Moreover, these issues are widely known yet loosely ignored.

The issue we described is the focus of this Master Thesis, and can be summed up as the many issues spreadsheets pose for the current business models they are part of. The problem is aggravated when this is thew only program companies use to save, edit, manage and visualize data.

The problem addressed in this Master Thesis is present on almost any organization that uses spreadsheets for the mentioned tasks. This problem is not a new one [15], [18], [19]. Often referred to as “spreadsheet hell” among other names can be described as the issues that occur when trying to read, analyze or validate large amounts of data on spreadsheets [19].

As spreadsheets appear to be at the center for many companies’ operations being cited as the only true universal technology [20] their importance is definitely hard to ignore [15].

The problem is worsened when there is a need to analyze data contained in these spreadsheets, as they are very hard to read and validate. This analysis, which usually leads to the production of reports based on the information in those spreadsheets, can be faulty because there are no mechanisms to automatically ensure data quality.

It’s hard to deny that spreadsheet programs can be great tools it only depends on what they’re being used for. And here is where the core problem of this Master Thesis lies, when using technologies like spreadsheets for data analysis and validation we are faced with “spreadsheet hell”. The term can include poor data management and poor data quality [19]. It can also occur on two levels micro and macro. The micro level refers to ”Frankensheets”, these are big, ugly spreadsheet monsters that are hard to understand, hard to use and hard to test. On the other hand, at the macro level regardless of the quality (or lack thereof) the problem lies in the ways these spreadsheets are used, shared and replicated.

It is normal for small businesses to prefer simpler and more affordable technologies, this is possibly the reason why most of the spreadsheet-related issues occur and are kept for very long time [12].

One of the issues that tools like Microsoft Excel also have is related to collaboration. Excel is not natively a collaborative tool, and to make it so would imply use of external technologies and qualification for those tools [21].

As highlighted in the present chapter we provide the theoretical foundations to frame our research problem.

3

Research Background

Contents

In this chapter some basic definitions and concepts will be provided to be used throughout the document. This will also serve to explain some introductory concepts in the context of this project as well as the theoretical concepts we believe to be the most relevant for this master thesis.

In the study for the problem of this Master Thesis, described in the previous section, we came across a few disciplines that we believe to be essential for our work. We will approach Digital Transformation due to the process of adoption of new technologies to solve a problem, and Data Governance due to the large amounts of data that our problem implies and thus needing a management framework.

3.1 Data Governance

Following the statement of our research context in the field of digital transformation and in line with the scope of our work we believed there was a need to define and use data governance concepts.

Terms that are very closely related to Data Governance are: data management, enterprise information management and data information architecture. According to DMBOK, data governance is ”the exercise of authority, control, and shared decision making (planning monitoring and enforcement) over the management of data assets” [22]. The definition appointed by this author for consulting is ”the organization and implementation of policies, procedures, structure, roles, and responsibilities which outline and enforce rules of engagement, decision rights and accountabilities for the effective management of information assets.” [22].

These three definitions already help us draw a clear picture of what this encompassing term can mean to the areas where it is applied. But put simply it is the use of authority and policy to guarantee the correct management of information assets.

According to Ladley, data governance should not be seen as a function performed by those who manage information, describing a distinctive line of duties between those who manage and those who govern.

This may help understand what needs to be addressed first and how to best address it. We followed a framework developed by Khatri et al. [23]. We opted for this framework due to its adaptability and popularity in the field. From what we gathered there were many others yet we will use a couple more to help ensure the fit for this particular one.

The process of data governance is highly dependent on the type of environment to which it is applied, meaning there is no one-size-fits-all solution that can be implemented out of the box. The nature and characteristics of the environment need to be studied and the methods of the framework applied accordingly.

The mentioned framework was developed based upon an IT Governance framework. The authors [23] explain this adaption due to the similarities in the fields.

Due to the practical nature of this component we will be covering it more thoroughly in section 5.

3.1.1 Data inventory

As part of the implemented Data Governance solution we believed there was a need to study another very common technique when using strategies such as this one. A data inventory can be described as a list of datasets followed by a description for those datasets, their content, their licenses and other relevant information [2]. This is very pertinent when facing the issues covered by our research problem.

A data inventory may help users find, manage, use and share associated data. By providing an overview of data and by abstracting from the data it can help understand what data were compiled, what information that encompasses, how are the data to be managed and made available to its users.This type of inventory is usually applied in situations where there is a lot of data from many different sources.

According to the Open Data Institute a data inventory can be created for a different number of reasons, these can be to help with the data governance strategy in place, to help improve knowledge discovery, in the making of decisions of data management, in the creation of new products and to help comply with standards like GDPR. Overall, a data inventory can provide useful information on the location, quality, technical details and legal framework and how this data is managed and used across the organization.

The creation of a data inventory is an essential step in using data as an asset which is believed to be one of the most important elements in designing data governance [23].

Furthermore, a data inventory is much more valuable if it's published. It could even only contain non-sensitive information and follow privacy protection rules, but the ability to publish the data inventory created adds to its value. It helps other researchers/employees to find datasets or data in general that can be used to produce more knowledge. By doing so we are able to create more transparency regarding the data and more trustworthiness for the data used.

In terms of the steps for the creation of a data inventory these are: planning for the inventory, deciding on what information we are to collect, populating the inventory, publishing that inventory [2]. Following these steps will help in the creation of a data inventory, by including the biggest number of datasets we could obtain a more comprehensive inventory leading to better results in terms of knowledge discovery.

As the first step in the creation of this data inventory we decided to understand the real need for one. The main reasons being that the organization needed to be able to track how many datasets were in use, what kind of information they contained, how many data sources were there (and how reliable were they), how could they be related, and overall improve the knowledge discovery. A data inventory provides, in this way, an abstraction to the numbers and the ability to focus on the metadata to manage data in a more efficient way as it is needed for a successful data governance strategy.

To improve data management the data inventory should collect enough technical information so that

the data is understandable and easy to access. To improve the business use of the data we must study what rights of access does the organization have over the use and sharing of that data. It should also be noted that a data inventory is even more relevant and valuable if it's kept updated in the future. As such, after each project that involves the use of these datasets these should be updated accordingly.

The authors also noted that it is very important to establish a definition of dataset and follow it for the remainder of the inventory; in this first phase we should also decide on which data to include in the inventory. The authors suggests a definition of datasets to help the development of this work.

> " A dataset is a collection of data that relates to a common topic or was curated for a common purpose. A dataset has a consistent standard in terms of its format and structure. A dataset can contain 'raw data', analysed results or derived information. "

The first phase of the data inventory is also used to define what level of granularity the data included should have. It is important to know that it is better to have more data and be able to make a selection from such data than to start off with less data and having to go back to repeat the whole process if needed.

Attribute	**Description**
ID	Unique identifier for the dataset
Title	The name of the data asset
Description	A description of the data asset
Purpose	Why was the data collected or produced?
Data creator	Who created the data?
Data manager/owner	Who manages the data?
Subject/keywords	What subjects/topics does this dataset cover? This will help discovery for users searching for this data. It is recommended to use a controlled vocabulary for this attribute (and others where possible) to improve future search and data linking potential eg finding related datasets
Location	Where is the data located or stored?
Creation date	When was the data created?
Update frequency	How often is the data updated?
Type	What type of data is it? Text, numbers, statistics, images, a database?
Format	What format is the data in? Eg MS Excel, CSV, JPEG, SQL DB
Rights and restrictions	What are the access and usage rights and restrictions? If you are publishing the data, what can users do with the data? Include a link to the relevant licence for use of the data (eg Creative Commons or a bespoke licence)

Figure 3.1: Data inventory attribute table [2]

In Figure 2.1 we have a depiction of what the authors [2] believe to be a good starting point for a data inventory.

Concerning the phase of populating the inventory there is the option of delegating the work to the original data owners, conducting surveys, interviews or questionnaires, or even to create automated processes for these data inventories. These techniques are the ones cited by the authors to aid in the

populating the inventory with the relevant information. The last component of this stage is to plan how the inventory is to be kept up to date.

Lastly, the inventory should be published if only partially due to some sensitive information it may contain. This step is important, because as we have mentioned, it helps other people work on data they may need and provides transparency for the data being used by the organization.

3.2 Data Science

Data science can be defined not only as a concept to unify statistics, data analysis and its related methods it is also very result-driven [24]. Data science aims to analyze and reveal the features or the hidden structure of phenomena relating to the data, shining a different light from the established or traditional theory and method. By having this goal, data science, and the notion of discovering knowledge from data or finding useful patterns in data has leads us, historically, to names as data mining, knowledge extraction, information discovery, information harvesting or data pattern processing which can, in many cases, mean the same. This term also gained some notoriety in the database field leading to the coining of the of knowledge discovery in databases (KDD) [25] being used interchangeably with the term data mining by some authors [26].

Data science, in discipline form, is defined as as a mixture of statistics, mathematics, computer science, graphic design, data mining, human-computer interaction, and information visualization [27].

As the most prominent area for this work we will be dedicating a considerable amount of our attention to it. According to our study, data science, as a term, emerged in the preface of a book called ”Concise Survey of Computer Methods” by Naur in 1974 [28]. There, data science was defined as ”the science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences”. A similar notion of ”the science of data and of data processes” was called ”datology” [29]. These definitions may appear simpler than what we today view as data science today but it loosely inspired our current understanding of it.

Following history the origin of data science or its evolution from data analysis started in the statistics and mathematics community in 1962. This eventually led to the wish that data analysis could be ”to convert data into information and knowledge” [30]. Over 20 years later the first workshop on ”Knowledge Discovery in Databases” emerged. Eventually this would gain more notoriety and increased recognition in computer science and other disciplines. This turned data science into the fastest growing and most popular computing, statistics and interdisciplinary communities.

This new science involves not only core disciplines such as computing, informatics, and statistics but also the broad-based fields of business, social science, and health/medical science. This made data science the possible key to fit all locks and it is what connects it to its use and popularity in many

business related situations.

To further study data science use and its popularity among the field and search interest over time we searched Google Trends. The resulting graphic is depicted in the figure below.

Figure 3.2: Data science interest over time according to Google Trends (information obtained on june 25th of 2020)

In Figure 3.2, and according to Google Trends the numbers on the interest of data science over time are to be read as the ”search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term”. The start of the growth of interest was around 2013/2014. Some scholars argue that fields with a similar names as data analytics, data analysis or even big data have also notable numbers when compared to data science. In the case of data analytics these share a similar trend in number and evolution, although, in the case of big data we see a bigger boom which scholars attribute to a business related buzz [31].

As in many cases, one definition would not paint a very clear picture since the field itself ranges from many different areas of study. From it's origin was in statistics, data science was often suggested to be the new name for statistics. With scholars arguing that statisticians should also be renamed to data scientists. This would also imply that statistics would change its focus from “data collection, modeling, analysis, problem understanding/resolving, decision making” onto “large/complex data, empirical-physical approach, representation and exploitation of knowledge.” [32]. Another author suggested that statistics should be broadened “to enlarge the major areas of technical work of the field of statistics” taking inspiration from computing and computer scientists [33]. Also leading to suggestions of the need to shift from the ”exclusive dependence on data models (in statistics) and adopt a more diverse set of tools” [34].

The simplest way of formally defining data science is as the science of data or the science that focuses on the study of data. According to Cao [31], data science can be defined, from a disciplinary perspective, as ”a new interdisciplinary field that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments (including domains and other contextual aspects, such as organizational and social aspects) in order to transform data to insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology.”

In light of this definition a discipline-based data science would be the use of statistics, informatics, computing, communication, sociology, management conditionally with data and environment and thinking. According to this author the result of this would be a data product. A data product "is a deliverable from data, or is enabled or driven by data, and can be a discovery, prediction, service, recommendation, decision-making insight, thinking, model, mode, paradigm, tool, or system." In conclusion, these can be knowledge intelligence wisdom and decision [31].

Last but not least, Artificial Intelligence (AI) is also a very important discipline in the field of data science. AI can comprise algorithms to make use of data in order to come up with or help in making decisions or solve tasks. For instance in this field we can find scripting algorithms for data manipulation where the probability of human error is high, also due to the repetitiveness of tasks.

Due to the complexity and variety of what a human can do, we come across a term called narrow AI, meaning that the current state of the art for AI is highly specialized in a few, limited, set of tasks. As an example, there can be AI that is great at guessing the class of a certain point in a timeframe but terrible in guessing a point in the future [35].

There are specific AI's, for instance, designed for robots or for cleaning houses and others that are specialized in managing and easing the analysis of data (for instance in very large datasets).

In the particular case of data science, AI potentiates the work that has been done in traditional data science. Algorithms based on AI may bring better performing models (which are mathematical abstractions used to simulate the behaviors observed in data) [35]. In terms of business models these AI-powered advantages can be differentiating between competitors.

In the present day, and seeing the light and evolution that AI has endured, we have more generalized access to the algorithms and libraries so that the more common data scientist can use and learn from them.

All these descriptions allow us to paint a more complex picture of data science and what it came to be as it matured through the years as a discipline.

3.3 KDD

Knowledge Discovery in Databases (KDD) as part of data science is the more specific process of getting more information out of the data we already have. This can include finding patterns by making use of fields such as machine learning, statistics and database-like systems.

Data mining can also be seen as an analysis component inside KDD or performing "Knowledge Discovery in Databases" [25]. This can also involve databases, data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, Visualization, understandability and online updating [36].

It is also relevant to note that KDD is not the component of extracting data but the process of analyzing it and being able to take conclusions from it. Another distinction to be made is between data mining and data analysis, while the first focuses on the test of models and hypothesis for the dataset, the latter is more tied to the use of machine learning and statistical models to find less obvious patterns from the data.

In the same way that AI facilitates Data Science there is also another very important field for this: Data Warehousing. Believed to be born from the industry and later imported into the academy data warehousing and other techniques allowed for a different more organized view into the data and what could be done with it.

These notions allied with multi-dimensional modeling made on the data much more interesting and promising due to being organized in a way so that there would remain no data-related ambiguity.

The benefits that data warehousing provides are the fact that is centered on the users' needs and that is presented and built from a simple dimensional perspective [37]. This is unlike the modeling of traditional relational databases in which data is structured following normal forms to avoid repetition at all costs, in data warehouse modeling this is precisely avoided to be able to slice and dice data in every way possible, making all kinds of cross analysis easier.

Often mentioned alongside data science, KDD, or Knowledge Discovery in Databases is the process of discovering hidden and interesting knowledge from data [38]. To define data mining we must also define knowledge discovery in databases (KDD) since one is contained in the other.

According to Fayyad et al., KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data [25].

There are many more other steps in KDD such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining. All these steps contribute the correct extraction, and analysis, of information from the data being analyzed. The authors also advert that the blind applicaton of data-mining methods (also called data dredging in statistics) can be a dangerous activity that can lead to meaningless and invalid patterns.

According to Frawley et al., the in-deepth definition of KDD is "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." and here data is a set of facts (such as cases in a database) and a pattern is "an expression in some language describing a subset of the data or a model applicable to the subset" [38].

In this section it is also relevant to mention that data warehousing can be defined as the business trend of collecting and cleaning transactional data to make it available for online analysis and decision support. This technique helps "set the stage for KDD in two important ways: (1) data cleaning and (2) data access." [25].

These two approaches are very important since these are the foundations of our work, the first focus was on cleaning the data and having it be easily accessible. This comes from the need of companies to have their data organized and in a unified logical view of the wide variety of data and databases they have available. Essentially the step is the migration from a messy disperse system to a unified and unique one throughout the organization. There is a need to resolve mapping issues in terms of name conventions (since between departments could exist different names for the same attributes). This, when done correctly, leads to uniformly represented data and handling of missing information.

3.4 The KDD process

To accompany the correct development of this work we will be following the KDD process as suggested by some authors [25]. This process is interactive and iterative, implying many steps with many decisions from the user. We will be citing these steps according to Fayyad et al.:

The first step is to develop an understanding of the domain of the application and the prior relevant knowledge and to identify, from the customer's viewpoint, the goal of the work that is to be developed. Secondly a target data set should be created, or focusing on a subject of variables or data samples, on which there is a need for discovery. Third is the step of data cleaning and data preprocessing, elemental operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes. The fourth step is data reduction and projection: finding useful features to represent the data depending on the goal of the task with dimensionality reduction or transformation methods. Fifth is matching the goals to the ones defined in the first step to a specific data mining method, such as summarization, classification, regression, clustering and so on. Sixth is the exploratory analysis, model and hypothesis selection: choosing the data mining algorithms and the methods for searching the underlying patterns. Seventh is data mining: which means searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression and clustering. The success of this step can be helped by performing the prior ones correctly. Eight is interpreting the mined patterns returning to any of the steps 1 through 7 for further iterations. This step can also involve visualizations of the extracted patterns and models or visualization of the data given the extracted models. The ninth and last step of the process is to take action upon the discoveries that KDD can bring. This can be by using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This process also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge [25]. All these steps, tie back into the notion of data mining. According to the literature, this component of the KDD process is often an iterative process of

some data mining processes. The two main goals of data mining are verification and discovery [25]. Verification implies that the developed system is limited to verifying the users' hypothesis. With discovery, the system autonomously finds new patterns. Inside the discovery component there can also be two subsections: prediction and description. Where prediction is the system finding patterns to predict some future behavior and whereas description is the system finding patterns for presentation to a user in a human-understandable form.

3.5 Data Warehousing

According to some scholars, data warehouses are at the core of the development of Business Intelligence (BI) Systems [39] and as such we thought it to be essential to include it into the literature review for this master thesis.

Data warehousing is a system that can have many applications and definitions, these are: data warehouse as a collection of integrated databases to support the decision support system; data warehouse as a collections of data marts which are used for querying and reporting by connecting through conformed dimensions [39]. At its simplest definition, a data warehouse is a way of modeling and organizing data to make its analysis easier and more flexible.

A data mart is a physical and logical subset from the data warehouse's presentation area. It used to be defined as highly aggregated subsets of data specifically designed to answer desired business questions.Now, a data mart is seen as a flexible set of data that is based on the most atomic/granular data possible to obtain the most value from it. It should be presented in a dimensional model making it the most resilient it can be [37].

After defining a data warehouse it is important to note which of the goals should be attained with the adoption of this modeling technique. A data warehouse should make the information asset that an organization possesses easily accessible, it must be legible and be labeled meaningfully. The tools that are chosen to access the data warehouse must be simple and easy to use, as well as fast. The data that is presented in a data warehouse must be consistent and credible. This implies that this data is cleansed and has been through validation processes to ensure its quality and that no errors have been encountered.

A data warehouse should be designed to be resilient to change and changing user needs, since these are an inevitable aspect of businesses. This resilience translates into a data warehouse that can withstand change without having to be completely redesigned from the ground up. The work with data and applications already in use should not be disrupted when there is a need for new questions or there is new data added to the data warehouse.

The designed data warehouse must strive for the goal of the protection of data as an asset yet it

should still protect the privacy of the data in it contained, allowing access for only the people with the right to it.

This modeling technique should also serve as a foundation for the continuous improvement in the process of decision making. The only feasible output for a data warehouse are the decisions we are able to draw from the information that is provided.

And, as a last crucial goal, the data warehouse must be widely accepted in the business community that uses it. No matter how great and how polished the solution is, if it is not accepted and used by the community for which it was intended it will definitely be useless.

In terms of modeling, the star schema is widely recommended for reporting, it is recognized as the most appropriate design strategy for such tasks. A star schema is a model containing dimension tables and one or more fact tables. These dimensions have a key and other fields for the remaining attributes. Fact tables on the other hand are central tables containing transactional data as well as foreign keys that link to the previous dimension tables. These dimension tables only contain master data meaning that no physical transaction value should be stored there. This model, in terms of representation, is similar to a star shape, thus the name, it also has the benefit of ease of understanding and reduced number of joins needed for some queries. In the dimension tables the hierarchy is defined through attributes [39].

Another definition of fact tables is a set of tables that contain values or measures produced by some event in the real world [37]. This implies that a fact table row corresponds to a direct measurement event and these types of tables are usually designed taking into account a specific activity. This also means that the fact tables shouldn't depend on the desired reports. This concept is central to the conception of data warehouses since, according to the authors, the simplicity of the modeling is key to achieve ultimate success.

Other design strategies are the snowflake or galaxy schema, these should be considered as a separate dimensional modeling philosophy. The star schema is the most widely used in the industry and is often considered as the general dimensional approach for data warehouses. Although it is important to note that a galaxy schema or fact constellation is defined as some fact tables that share common dimensions, due to this reason it is seen as group of star schemas, hence the name.

Data warehouses are chosen over other types of databases usually due to their capacity to store and organize temporal information, usually for longer periods of time. As such, data warehouses require a precise planning and designing having their architecture be a crucial part of their creation [37].

Another important note is the definition of an enterprise bus matrix, it is an essential tool for creating and communicating the produced architecture. The rows of this matrix represent the business processes and the columns represent the dimensions. The cells that have been grayed out signal if a dimension is linked to the corresponding business process [37]. It can also happen for a dimension to related to multiple business processes.

Once having covered the basic structure with either one of the mentioned schemas, the bus dimension is what ties them all together leading to the data warehouse becoming a set of fact tables that can be cross analyzed as for the organization's needs. These are called conformed dimensions in which attributes in separate dimension tables have corresponding column names and contents. Conformed dimensions allow combining information from different fact tables in the same report [37].

3.6 Extract, Transform and Load (ETL)

Often seen as one of the most effort and time consuming stages required for creating a data warehouse and for business intelligence, there are many alien constraints that may impact this stage, nevertheless it can be difficult to appreciate its complexity [37].

It is also commonly referred to as the data staging area, and is defined as a selection of processes that aim to prepare the data for a data warehouse, being performed prior to the use of any queries of graphing [37]. Generally it consists of acquiring data from a source, transforming it, loading it and indexing it as well as guaranteeing its quality and its publication.

Traditionally, Extract, Transform and Load (ETL) is performed in three main stages following it's own name: Extract, Transform and Load. In the Extraction phase, to get a sense of the data, the datasets are obtained and analyzed by the naked eye just to evaluate the type of analysis to be performed. Usually this data comes from different sources and is aggregated in this stage. Once we have the data in the staging area there is quite a long list of transformations that are possible (cleansing the data, correcting names and spelling, dealing with missing values, parsing it to a different format, joining data from multiple sources, deduplicating data and assigning data warehouse keys, etc) [37]. It is in the transformation stage that a great amount of effort is needed considering all the labor-intensive manipulations that do occur.

According to other authors, the ETL stage can be greatly improved through the use of machine learning frameworks. These frameworks can help in the obtaining of data if coming from different sources, which is often the case in the industry. As previously noted this is a very time consuming process meaning that the use of certain frameworks, such as the one the Kimball Group presents [37], can be advantageous. These frameworks can also help reducing the extra need for data engineering since they already provide some ”out-of-the-box” data formatting. Of course, never disregarding the importance of doing some in-house data processing [35].

The final step of ETL is the loading of data, this implies loading quality ensured and correct data onto the data warehouse.

The Kimball Group also organized the ETL architecture into 34 subsystems: three focusing on the extraction of data, five to deal with cleansing and conforming data (which include dimensional structures

for error monitoring), 13 subsystems for the delivery of data as dimensional structures to the final BI layer and another 13 to help manage the production of the ETL environment.

Another term that we need to cover before delving into this work in the notion of data mining.

3.7 Data mining

In terms of data accessibility the use of uniform and clear methods helps the ease of access to the now clean and complete data. This type of transformation is needed in the event of having data in an organization that is of very difficult access (i.e. to find the value of a report there is a need to consult three different files).

The data mining step in the KDD process consists of using data analysis and discovery algorithms which under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data [25]. The KDD process itself involves:

- the use of the database along with any required selection, preprocessing, subsampling, and transformations of it;
- applying data mining methods (algorithms) to enumerate patterns from it;
- evaluating the products of data mining to identify the subset of the enumerated patterns deemed knowledge.

This component in the KDD process is concerned with the algorithmic means by which patterns are extracted and enumerated from data [25]. This process also involves the evaluation and possible interpretation of the mined patterns to understand which patterns can be seen as new knowledge.

Data mining implies fitting models to, or determining patterns from, observed data [25]. These models can bring interesting knowledge as part of the KDD process where human subjectiveness is needed, either to make decisions or take conclusions from the acquired knowledge. In this case two mathematical approaches can be used, these are statistical and logical approaches. The statistical one allows for nondeterministic effects in the model, on the other hand, the logical one is purely deterministic. The statistical approach tends to be the most widely used for practical data mining applications given the usual presence of uncertainty in real-world processes, which is the case for the work of this Master Thesis. Some examples are, classification, clustering, regression, etc. [25].

Classification is defined as learning a function that maps (classifies) a data item to one of the predefined classes [40] [41].

Regression is defined as learning a function that maps a data item to a real-valued prediction variable. Applications of this method can range from prediction of values given a set of other values to estimations (predictions) of time someone will survive given the results of a test [25].

Clustering is defined as a common descriptive task where the aim is to identify a finite set of categories or clusters to describe the data [42].

Summarization involves methods for finding a compact description for a subset of data. A simple example of this could be tabulating the mean and standard deviations for all fields. These techniques can be applied to interactive exploratory data analysis and automated report generation [25].

Pattern mining modeling consists of finding a model that describes significant dependencies between variables. These can be found on two levels: the structural level of the model specifies which variables are locally dependent on each other and the quantitative level of the model specifies the strengths of the dependencies on a numeric scale (probabilistic dependency using conditional independence) [25].

Change and deviation detection focus on discovering the most significant changes in the data from previously measured or normative values [43].

Followed by the definition of some data mining methods we must now delve into the components of the data mining algorithms that can be used. According to [25], there are three primary components in any data mining algorithm: model representation, model evaluation and search. Model representation is the language used to describe discoverable patterns. If these representations are too limited then no amount of training time or examples can produce an accurate model for the data being worked on. On the other hand, over exaggerated representational power for models extends the danger of overfitting the training data which can imply a reduction in the prediction accuracy [44]. Model evaluation criteria are quantitative statements (or fit functions) of how well a particular pattern (meaning its model and its parameters) meets the goals of the KDD process. Predictive models are evaluated based on their quality to predict upon a test set. Models can be assessed by their predictive accuracy, novelty, utility, and undertsandability of the fitted model. Finally, the search method can be subdivided into two components, parameter search and model search. After closing the model representation and the model evaluation the data mining problem is reduced to an optimization task [25]. This means that there is a need to find the criteria that optimizes the model-evaluation part.

Non linear regression and classification methods: These are composed of a family of techniques for prediction that fit linear and non linear combinations of basis functions (sigmoids, splines, polynomials) to the input variables. These range from feed-forward neural networks to adaptive spline methods or even projection pursuit regression [25]. Non linear regression methods despite being very powerful in representational power can be very difficult to interpret.

Example-based methods: The representation in example-based methods is simple, it uses representative examples from the database to approximate a model [25]. This means that predictions for new examples are based on the properties of models to which we already have predictions. Some techniques that use this are kNN classifications and regression algorithms [45]. A downside to this group of methods is that there is a need for a good distance metric to measure distances between data points

whereas in the tree-based methods that is not needed [25].

Probabilistic graphic dependency models: Graphic models specify probabilistic dependencies using a graph structure [46] [47]. In the most basic form these models specify which variables are directly dependent on each other. These tend to be used with categorical or discrete variables [25].

It was relevant to us to mention all these algorithm components because we believe that the present brief explanation can shed some light on how to better adapt to the problem that this work addresses. Taking into account the size and variety in a field such as data mining the present portion of this work could suffer from a deficit in terms of scope. There is a lot more we could have covered turning this work into an extensive view of the subject of data mining and knowledge discovery in databases (KDD). And, according to Fayyad et al. [25] in regards to the novelty of the area in general: "There are no established criteria for deciding which methods to use in which circumstances, and many of the approaches are based on crude heuristic approximations to avoid the expensive search required to find optimal, or even good, solutions."

3.8 Forecasting

As one of the most popular data science techniques in finding new information in data we present forecasting. These are often very interesting for organizations to apply in their day-to-day activities to aid in decision making.

Data mining and sophisticated forecasting techniques have been gaining critical amounts of respect, particularly in areas such as economic forecasting, fraud detection and risk analysis [37]. Given this promise we decided to dedicate a section to this field due to the potential it can have for our work.

Forecasting is understood as predicting the future in the most accurate way possible taking into account all the data available such as historical data and some knowledge of any future changes that can change the outcome of the forecasts [48]. This author makes and important distinction, in terms of business forecasting, where it can be mistaken for other concepts like goals and planning. Goals are what would be ideal to happen, these should be linked to forecasts and plans yet it does not always happen. Planning, on the other hand is an answer to the information of forecasts and goals, it involves choosing which actions are appropriate to make the forecasts align with the goals.

Forecasting should be taken as an integral part of the decision-making activities of management and an organization should develop a forecasting system that is tailored to their needs including different approaches for the prediction of uncertain events [48].

This author also mentions that a forecasting task usually involves five basic steps, these being: problem definition, gathering information, preliminary (exploratory) analysis, choosing and fitting models and using and evaluating the chosen forecasting model.

The problem definition, seen as the most difficult part in the forecasting process, requires a sound understanding of how the forecasts will be used, to which end, who needs the forecasts and how they fit inside the organization producing those forecasts. The forecaster needs to talk to everyone that collects, cleans, maintains the data and will be interpreting the produced forecasts.

In the preliminary analysis one should always start by visualizing the data on a graph first to see if there any visible underlying patterns, trends, seasonality, cycles or outliers that need to be justified by experts.

Choosing and fitting forecasting models depends highly on the availability of historical data, strength of variables' relationships and the way in which the forecasts are to be used. Usually two or three models can be compared. These models pose as artificial constructs based on a set o explicit and implicit assumptions involving the estimation of one or more parameters by using the available historical data. These models can be: regression models, exponential smoothing methods, ARIMA models, dynamic regression models and hierarchical forecasting. Other more advanced methods include neural networks and vector autoregression.

The final step is to use the chosen forecasting model and evaluating it. The performance of these models can only be tested after the data for the created forecast becomes available [48].

In the following sections we will cover the theory behind ARIMA and SARIMA.

3.8.1 ARIMA

ARIMA (Auto-regressive Integrated Moving Average) is known not only for its popularity but also for being known to work well for short term predictions. Which was also one of the reasons we chose to use this model. This model receives three parameters p, d and q. p is the number of lag observations to be set for the model. d is the degree of differencing for the raw observations. q is the size to be set for the moving average.

The auto regression component in the model is a specific type of regression where the attribute depends on its own past values. This, in turn, means that the present values have a correlation to past ones. As such, the value of p or the number of values to lag should be the number of observations that are relevant to be included in the model for the present observation.

The moving average models analyze the error associated of previous predictions to improve the current prediction. Then, the value of q represents the size of the moving window or the number of lagged observation errors that have meaningful impact over the current observation.

Last major component of the ARIMA model, the integrated (I) part. This is used when the series being studied is not stationary. And from theory, we know that to be able to use the AR and MA models the series needs to be stationary, when this condition is not met we simply transform the series according

to the following formula:

$$Z_t = Y_{t+1} - Y_t$$

This or the following derivatives (if needed) will eventually allow us to get a constant mean, standard deviation and no seasonality which are the conditions for a series to be stationary.

After doing so, the following step is to select the data, split into train and test data, which means to split the data in two portions, one for training the model and the other for testing the accuracy of the forecasts. Training the model can be interpreted as the act of teaching our model the behavior of data over a time period.

The fitting of the model is reached after finding which values or parameters yield the best results. for this techniques such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) are used, which stand for Akaike Information Criterion and Bayesian Information Criterion, respectively.

3.8.2 SARIMA

Another very popular technique is the SARIMA or Seasonal Auto-regressive Integrated Moving Average, the difference between this and the ARIMA model is the introduction of seasonality in new parameters as well as a new parameter, m. As such, we have seasonal P, D and Q as well as m, the new parameter for the number of time steps in a single period. SARIMA is defined as (p,d,q)(P,D,Q,m).

The choice for these parameters is made in a similar manner as for the ARIMA model.

The following steps, as described for the previous technique are, splitting data, fitting the model, generating the forecast and plotting them. The approach to these techniques will be covered in detail in Chapter 5.

To evaluate the accuracy of the forecasts we decided to use two metrics commonly used for the purpose, and which we believed to be more appropriate. These were Mean Absolute Error (MAE) and Weighted Absolute Percentage Error. MAE is the mean absolute error and it's "a risk metric corresponding to the expected value of the absolute error loss or -norm loss". This is a commonly used metric used for regression models, and since our data is in Euros we believed it was more adequate than other metrics such as the Mean Squared Error, which is also very common.

And WAPE which can also be referred to as weighted mean absolute percentage error which is sensibly the same as the previous measure (in percentage) although we thought it would give a more representative value of the error, eventually considering the diverse nature and value of the attributes.

The formulas used to calculate these errors were the following:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|A_i - F_i|$$

$$WAPE = \frac{\sum_{i=1}^{n} |A_i - F_i|}{\sum_{i=1}^{n} |A_i|} \times 100$$

The application of these formulas will be further explained upon its application in Chapter 5.

In conclusion to this chapter we understand that data science and data mining have enabled a ”new data economy” that is product-based and technology-driven. This may be why an ever increasing number of companies recognize the value that this data can have as a strategic asset. This leads them to invest in building infrastructures, resources and talent as well as teams to support this innovation. The aim is to lift competition and productivity these being core concepts on the work developed for this Master Thesis.

4

Research Proposal

Contents

The research proposal for this master thesis is defined in accordance with the the objectives and the design and development step from the chosen methodology. In this chapter we will describe the objectives and how to achieve them with our proposed solution.

4.1 Objectives

The main objective of this Master Thesis was the development of an architecture to solve the research problem we address. Essentially, the objective of this work is to provide an alternative to the problematic data management and analysis of data inside programs like Microsoft Excel. This implies the creation of an artifact - a system that comprises different tools and techniques to improve data analysis and intend to fix the problems associated with Microsoft Excel (regarding this subject).

To solve this problem we propose to make use of new digital technologies and align them in a system architecture so it becomes a complete and ready to use/implement system.

4.2 Pipeline

In order to help in the development of our solution we created a pipeline to organize our work and to keep it as close to the methods from the available literature.

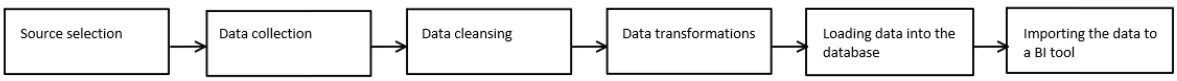

Figure 4.1: Pipeline designed according to the literature.

In Figure 4.1, we can see a simplistic design of the pipeline following the work we intend to perform in terms of the stages data passes through to ensure the most accurate use of data.

This pipeline is set to accompany the majority of our work to help ensure no steps are missed or performed in an incorrect order.

The particular stages of this pipeline are described in further detail in the following section.

4.3 Description

In order to accomplish the objectives set in the previous section we propose the creation of an artifact to explore the artifact itself and how well it suits the needs required. Many companies face these issues of using Excel beyond the program's capabilities and often to find quite some errors as well as coming

at great expense monetary and efficiency-wise. Very often Excel is used to crunch numbers but also to store, collect, cleanse, operate, visualize and even correct data. Anyone that has used Excel knows that it is simple and intuitive yet also very prone to errors, so, it is not the ideal tool to be using, specially solely for this task.

Our system will be comprised of many-step processes to get from the first stage where data is received and all the way to the last step which includes the analysis or production of reports on that data, never forgetting the validation component.

In practical terms, the system we propose can be separated into three logical parts categorized according to the academic area to which they belong to (from data science to data visualization and database management).

- ETL
- Analyzing the data on a BI tool
- Production of reports

From this list we can detail what our proposal is to include more specifically. The most time consuming, and possibly the most important part, is the ETL step of the process of our system. To solve the problem of timesheets we realized there was a need to look at data in a different way. We believe that the 34 subsystems of the ETL architecture proposed by the Kimball Group [37] is a great way to approach this step of our system .

The first component of the ETL, will consist in the extraction and aggregation of the data from the varied data sources available, subsequently comes the transformation stage of the ETL, in which we plan on spending the most time since transforming the data is very important to the theoretical concepts that accompany the design of a data warehouse. The transformation stage is also very crucial to an area such as the one of this Master Thesis due to some of the problems usually reported from Excel spreadsheets. The data is saved in a strange fashion, often doesn't comply to any specific rules, just the need for average human understanding. General activities here imply transposition of how the data is saved, dividing data into logical parts or dimensions (and ensure that it is at it's most granular level), removing aggregations from the data (since these usually mean redundant information), creation of hierarchies for attributes, establishing a prior-defined set of business rules to ensure the validity of the results, pivoting tables, etc. These are some of the transformations we expect to perform to ensure that the designed data warehouse is compliant with the rules of this modeling technique, almost guaranteeing the success of analysis later on.

The last component is fairly simple, it consists of the loading component of data. In our case, in a SQL server database, with the implemented business rules to ensure once more the quality and validity of the data inside the data warehouse.

According to the 34 subsystem ETL process [37] we should have four groups of subsystems, the first three regarding ETL, respectively and the former regarding the active management of the ETL environment.

In the first group, with respect to Extraction, there are three subsystems: data profiling, in which data sources are explored to determine fit as a source and there's a collection of cleaning and conforming requirements; data capture in which changes are isolated from the source system to reduce the process burden; and the extraction and loading of data (into the data warehouse) for further processing.

The second group, about Transformation, focuses on data cleansing and conforming: data is first cleansed and screened for quality, data quality processes are defined to check if business rules are being respected; cleaning control with error event schema and audit dimension is performed; deduplication of data, meaning elimination of redundant members of core dimensions (i.e. customers or products); and data conforming, ensures common dimension attributes in conformed dimensions and common metrics across related fact-tables.

The third group, regarding the preparation for presentation, including: implementation of logic for Slowly Changing Dimensions (SCD) attributes; production of surrogate keys that are independent between dimensions; hierarchy manager, delivering multiple simultaneous, embedded hierarchical structures in a dimensions; special dimensions manager that creates placeholders for repeatable processes supporting the multidimensional design characteristics; fact table builders create the three primary types of fact tables including transaction grain, periodic snapshots and accumulating snapshots; surrogate key pipeline replaces operational keys for the incoming fact table records with appropriate dimension surrogate keys; multi-valued bridge table builder creates and manages bridge tables for multi-valued relationships; late arriving data handler applies special changes to standard procedures due to late arriving fact or dimensional data; dimension manager is a centralized component that prepares and publishes the conformed dimensions to the data warehouse; fact table provider administrates one or more fact tables being responsible for its creation maintenance and use; aggregate builder builds and maintains aggregates for seamless use with navigation technologies for improved query performance; OLAP Cube builder uses data from the produced schema to populate the OLAP cubes; and data propagation manager prepares the conformed and integrated data from the data warehouse server to be delivered on other environments.

The last group focuses on the management of the ETL environment since the success of the data warehouse depends heavily on the quality of data present there loaded. To achieve this success the ETL system must aim to guarantee three criteria: reliability, for the processes from the system to run consistently to provide data on time and trustworthy at any level of detail; availability, ensuring the needs of its service implying the data warehouse must be available as needed; manageability, a data warehouse is always a work in progress constantly changing and growing with the business, this relies

on the correct adaption of the ETL process.

This last group of subsystems includes: a job scheduler that manages the ETL execution strategy; backup system that keeps a backup in the need for recovery restart or archival purposes; recovery and restart the actual process for recovery and restart in the event of failure; version control takes snapshots for archiving and recovering all the logic and metadata from the pipeline; version migration, migrating a complete version of the pipeline from development into test and the production; workflow monitor, guarantees the ETL processes operate efficiently and that the data warehouse is being loaded at the correct times; sorting, serves the ETL processing role; lineage and dependency, identifies the source of a data element and all transformations or vice versa; problem escalation supports the structure that aids the resolution of ETL problems; paralleling and pipelining enables the ETL system to automatically leverage multiple processors or grid computing that respect the schedule needs; security ensures authorized access to all ETL data and metadata by individual and role; compliance manager, supports the organization's compliance with the imposed requirements by maintaining the data chain of custody and by tracking who has authorized access to data; and metadata repository, captures the ETL metadata including the process metadata as well as technical or business metadata.

We decided to opt for a simple and intuitive BI tool so that the produced data warehouse could serve its purpose which, according to Kimball [37], means to have its data easy and fast to access, be labeled meaningfully and consistently.

The BI tool to be used should be able to explore the data warehouse extensively, making use of all the benefits a data warehouse provides, such as cross analysis, slicing and dicing or filtering. This tool should be easy to learn and use and it should facilitate the adoption of prior business uses. It should also prove to be simpler and far more efficient than the old way of doing things. It should have rapid increase learning curve so the users don't get discouraged upon adoption and it should also feel familiar to the user. This should also allow the users to have independence and mastery over the creation of reports without having to rely on IT or data warehouse managers.

In general, the architecture of our system can be described as a multi-platform solution that starts in the business process of an organization to gather the data it uses for analysis from the different data sources. From here, the data is set to follow the pipeline to ensure, completeness, correction and validity. This pipeline was defined with very low specificity to ensure it was adaptable to other contexts, if needed. There was also a strategic choice to not use just one specific tool, since this generic proposal can be applied with the any BI tool. There is also the option to chose whichever tool is preferred for storing the data, here we provided a particular example of Azure SQL Server tool, but this can also be changed.

The advantage of our system lies not only on **posing as a solution to eventually solve "spreadsheet hell" but also on the enabled cross-analysis capacity as well as how efficient the whole process becomes due to our complete end-to-end system**.

5

Demonstration

Contents

In this section, we will address the demonstration step of the methodology we chose, DSRM which follows the use of the designed artifact to solve the proposed problem. This will cover experiences and simulations to do so. It is important to know how an artifact is to be used to solve the problem at hand. In the previous chapter, we presented the research proposal, here we will demonstrate our solution as well as test it in a Portuguese finance organization.

Due to privacy concerns we will not be disclosing the name of the organization. Henceforth, we will the using the name Organization when mentioning the entity.

The organization where this work was developed it is an organization that specializes in analyzing public finance data being granted with the evaluation of the quality of fiscal policies and executions.

In the following section we will be addressing Public Finance to provide some context on the data used for the development of this work.

5.1 Public Finance

The present work depends greatly on the field of Finance and, in particular, of Public Finance. As such, this work would not be complete without a, even if brief, definition of Finances.

Public finance or public sector economics is **the study of government economic policies**.This means to determine if said policies promote societies' economic objectives which may not always prove to be so easy. There are two models for approaching public finance: the theory of public choice and behavioral economics. James Buchanan founded the theory of public choice earning him a Nobel Prize in Economics [49]. Behavioral Economics is a newer approach to be the mainstream theory. By applying psychological principles to understand behavior that is otherwise at odds with the mainstream assumption that people act to maximize their own self interests. [49]. This approach is gaining momentum in the different fields of economics.

Another similar view on public finance is **the analysis of government services subsidies and welfare payments as well as the methods by which the expenditures for these ends are included in taxes, borrowing, foreign aid and the creation of new money** [50]. Public finance handles the resource allocation system making significant use of government money. Police protection and health are examples of areas of public services that sometimes come under particular scrutiny. The focus of the present work is in the health sector of Portuguese public finance. Governments cover the cost for areas like the two we just mentioned, through taxes and typically the amount paid in taxes does not equal the amount of service received [50].

Another definition often used to objectively describe public finance is to put it as a branch of economics that studies the taxing and spending activities of the government [51]. According to the authors, the name can be a misnomer since the fundamental issues it addresses are not financial but relate to

the use of real resources. Public finance includes both the positive and normative analysis. Positive analysis deals with the issues of cause and effect. An example provided could be "if a government were to cut taxes on gasoline what would happen to its consumption?". On the other hand normative analysis refers to the ethical issues, for instance arguing about the fairness of taxing income or consumption [51].

The modern approach to public finance concerns the microeconomic functions of government, concretely with how governments allocate the resources already available. On the other hand, the remainder macroeconomic aspects of government management such as taxes, spending and monetary policies are covered in other areas [51].

In regards to public finance and its current context in a more technology driven world, some authors from the IMF [52] say that the effectiveness of fiscal policy depends crucially on the information and technologies available to the government and how these are exploited (which is the case for the organization this work was developed in). Governments support the economies in times of recession and use booms to economize using taxes to fund social safety nets, health and education services, infrastructures, etc. Given the way governments operate theoretically, the fiscal policies depend heavily and are shaped by the reliability, timeliness and detail of information that is available to the government about the economy and who impacts it.

This information ranges from the taxpayers' incomes and assets, data on social program beneficiaries, the employment status of workers, the size of the output gap and the magnitude and timing of government transactions. When changing the way governments collect, process and act on information, digital transformation (which will be covered in the next section) is reshaping the formulation and implementation of these policies.

The relevance of this Master Thesis lies in how it provides new opportunities for public finance and businesses in general (ie. better information, systems, policies, etc). However this has, of course, its challenges and limitations which we will also cover.

These new technologies, and digital transformation in particular, provide a plethora of new opportunities. Through the knowledge and reliability of information, governments can do what they have been doing before and even more with a guarantee of data quality. This way, governments can ensure better information, build better systems and implement better policies.

On the front of better information, some authors defend that out of all the benefits that digital transformation can bring, the benefits to the data may be the best one. By having the ability to gather information, process it, and share it in a more timely and accessible fashion which is a very big leap forward.

With the development of greater mass technologies (such as Intelligence tools) comes a greater storage and processing power. This power allows for governments to better collect and analyze to hopefully make the most informed decisions possible.

The improvement of IS can enhance the implementation of new tax and spending policies, which can

imply reduction in costs for tax collection and compliance, social programs and in the management of public finance.

Naturally, since the introduction of electronic taxes there has been a decrease in administration tax costs.

Finally, regarding the improvement in policies, greater access to information and better digital systems may also open up the possibilities for better policies. Being able to monitor and unify a greater amount of information provides the ability to rethink and adapt tax policies.

This greater availability of high frequency fiscal data gives policy makers a better margin for forecasting of revenues and budget preparation. This can in turn help stabilize the business cycle allowing governments to monitor economic activity in real time [52].

Authors furthermore defend that digital transformation can help the delivery of public service. The first and most obvious way is through the use of a widespread platform for the dissemination of important information.

There were many steps to this work which are presented and explained in detail in the current section. We first cover the datasets and explain the processes of data cleansing, followed by other data transformations needed culminating in the analysis or further operations performed on the data.

The organization where we were to test our solution had a problem aligned with the one described in this Master Thesis. This problem stemmed from an actual business need, in their case, considering the large amounts of data used within the organization, they needed a more simple and correct way of handling and managing that data but also more efficient ways of performing their business processes.

In other words, the main goal for them would be to not only fix the implications that spreadsheets are famous for but also to optimize processes and ease the analysis Another desired aspect would be to analyze data in ways that would allow for the forecast of attributes.

The solution proposed to solve this problem implies the use of fields such as analysis and integration of data, data science, big data but also digital transformation. This would follow digital transformation cues to use newer technologies that have shown to be promising. This potential and vision was one of the reasons for its deep exploration and use in this Master Thesis. The use of data science techniques was also highly anticipated since it is of great interest in the field of public finance.

In broad strokes, the purpose of this work is to apply data science and data mining techniques to a context where it had not been used before for this particular context and datasets. Due to the nature of data studied there were some fields of data science that would be interesting to explore, such as clustering (finding interesting patterns in the data and grouping data points by similarity) and forecasting (predicting how an indicator will evolve). As one of the most desired lines of work of the 21st century [53], data science has been used with success and great promises in many areas. The aforementioned potential that data science can have in treating and studying such data is evidenced by

some studies [31] [54] [53] [55] [24]. The buzz that data science has been amassing for quite some years now is worth studying. It has shown promise in fields such as BI (these terms will be covered in more detail further ahead in this chapter); for analytics has become more business-oriented [56]. Now it can cover a wide variety of data and domain-specific analytical tasks: business analytics, risk analytics, behavior analytics, social analytics, web analytics just to name a few. These domain-specific analytics often become the driving force to the use and success of data science. It is important to note that domain-specific and data specific analytics has become the cornerstone of data science [31].

As we have seen through some scientific works and from a wide array of disciplines, data science is changing the traditional way of using and analyzing data itself, particularly in data-oriented and engineering fields. This is possibly what motivates companies to take this step. Since this transformation involves a high degree of change in companies' status quo. It implies changing the way something has been done for years and to shift the employee's focus to pursue this new path and strategy; this is no easy feat.

5.2 Previous business processes

This section is dedicated to the description of how their business processes were before the implementation of our proposal. As the present work was tested in a public finance organization some aspects of the implementation of our solution were adapted to fit their particular business models. The Organization is a public institution and their main objective is to ensure the correct application of public finances.

This organization is particular in the sense that their high business value is not in terms of money but in terms of the impact it can have on its community. The goal is to provide a fair and just analysis of public finance.

The maximum value at the Organization is achieved by the accomplishment of its mission in providing reports on public finance in Portugal. Thus the best way to enhance their value is to optimize their business process, which means to optimize how their data is handled.

The work performed is very data-driven, which aligns perfectly with our field of study and our problem definition. The Organization is divided into two main categories of data-related business-roles, that is technical personal and technical coordinators, both of these operate the data directly; being the technical coordinators the ones to be held accountable for the quality of that data.

In the Organization, a usual data exchange starts with the technical staff asking public entities (such as public administrations) for specific data that is needed for analysis deemed necessary. This was the first opportunity we encountered where our solution could serve to optimize their business uses for data and to solve a problem related to the reception of that data.

Then, after asking for the data, typically via email, the technicians are sent a Microsoft Excel file

containing the requested data, afterwards the technicians validate the data manually, which can often contain errors, and then is followed another exchange of emails to fix certain errors or ask for clarification of some values that might seem off or raise questions.

After this sometimes lengthy email exchange process only to obtain the data, the technicians perform one last quality check to ensure the data is trusted. All these steps are part of the Organization's business process and they are crucial since the value of their work relies on the quality of the data they work with.

These steps are fundamental to their day to day work and this is where a proposal like ours can help. After these interactions the data is stored in spreadsheets that often may exceed their capacity.

Processes like these are very common across the fields of operation like this one making it an ideal place to test our proposal to solve that same problem.

According to the studied literature, this is somewhat of a common problem since many companies rely solely on Microsoft Excel for their day-to-day activities. Companies sometimes lack the fundamental theoretical concepts that would allow them to develop more complete solutions to somewhat complex problems.

On the subject of Microsoft Excel, just because it is a very powerful tool it doesn't mean it's a suitable tool for any task at hand. In fact, that has been proven time and time again. Despite being appointed as an easy program to use for finance it poses many risks and becomes unreliable when it's the only tool being used. The pros are being a low-cost alternative included in the software packages that companies might already have to use (for instance for word processors and email clients), it's fairly easy to use, it's intuitive and from its basic functions to some more complex formulas it's simple and direct to learn, it provides the users with templates and it provides effortless integration in the Office 365 already used by the majority of companies.

However, Excel lacks some useful features such as collaboration in documents, scalability for support of bigger data sets, and proper data visualization tools. Although having some basic features it lacks more complete capabilities competitors have. Situations like these capture precisely our research problem using Excel as the only tool to collect, analyze, validate, and visualize data. Considering the challenges and limitations it may pose, it becomes clear that using this program for data mining and data visualization is far from the ideal solution. In the present chapter, we present the artifact of this Master Thesis as a system to improve all the faults that the previous system had. While presenting better features in terms of knowledge discovery and potentiating machine learning to meet other business objectives the organization also desired to achieve.

5.3 Digital Transformation

All the work we have developed can be associated with the field of digital transformation since, as a whole, can be described as the adoption of new technologies and thus considered digital transformation.

Following a digital strategy thought out precisely for this organizations' particular context, we applied the theory studied from [57]. According to this article, the defined digital transformation strategy should be a blueprint that helps companies navigate the transformations and it can take the form of three main dimensions whichever is the domain to which is being applied. These dimensions are the use of technologies, changes in value creation, structural changes, and financial aspects.

The most important one in the case of the chosen organization was the change in value creation, this reflects the impact of the digital transformation techniques on the organization's value chains, meaning the difference between the new digital activities compared to the original ones.

Another important mention, according to this article, are the structural changes about an organization's setup, in particular with the placing of these new digital activities. We believed this was worth a mention because our proposal involves changes to the existing corporate structures.

To guarantee the successful adoption of these digital transformation strategies, the authors point to how essential is the alignment of the four dimensions: use of technologies, value creation, structural changes, and the financial aspects. [57].

The authors continue onto saying that an organization's digital transformation process is a continuous complex undertaking that can drastically change the organization in itself and its operations. It is extremely important to assign clear responsibilities for the digital transformation strategy undertaken, advising that in failing to so companies can lose their scope and miss the potential the strategy could've had. At the beginning with the initial planning phase, they also mention that top management support is essential throughout all the transformation process since these transformations can impact the entire organization. To ensure this process is as smooth as possible the active involvement is required from the different stakeholders that are affected by the transformation.

The authors also point that besides the correct choice of staff for the process of digital transformation there is also a need to find procedures for formulating, implementing, evaluating, and adapting to those specific digital strategies.

These digital strategies should be reassessed as needed to ensure no development block is struck and the strategy is still current and can suffice the companies' needs.

We believe that digital transformation does encompass a very large scope of technologies and techniques, as such, the following sections until the end of the present chapter include the efforts that were made in the sense of the digital transformation aligned with our proposal.

In addition, we will briefly describe the strategy adopted and elaborate upon it in the following sections.

In terms of ETL we used some different techniques to different problems, initially we designed what the ideal data warehouse would look like and worked up from there, planning the steps backwards to what would get us o that ideal data warehouse.

In terms of visualization, we Power BI which integrated already very well with the technologies (Office 365). This was used for the business intelligence portion of our work.

5.4 Tools

We decided to use the Python programming language due to our familiarity with it and the its potential and popularity. Python has become one of the most popular interpreted programming languages [58]. Python is often seen as a scripting language since it can be used to write small programs or scripts to automate other tasks, this was also one of the reasons we needed a language such as this one, to automate the tasks of data manipulation to replace the manual and error-prone Excel spreadsheet data manipulation. Python is also seen as one of the most important languages for data science, machine learning and general software development in academia and industry [58].

We will cover a few Python libraries that we found to be relevant for the development of this work. The first one is NumPy, this library, short for Numerical Python has been identified as one of the most important ones in terms of numerical computing in Python. This brings into Python data structures, algorithms and library elements essential for scientific computing for numerical data [58].

In terms of data analysis and data visualization, Python is often compared with other languages such as R, MATLAB, SAS, Stata among others. Recently Python has improved support for popular data analysis libraries such as pandas and scikit-learn making it an easy contender for these data-driven tasks. Considering this programming language's strength and the tools with which it can be paired with we know this makes it an easy choice for this field.

pandas is a Python library developed for data manipulation and analysis, and one we've made use to manipulate our data. This allows for fast, easy and expressive handling of structured and tabular data. The primary object to be used in pandas is the DataFrame, which is a tabular, column-oriented data structure with both row and column labels [58].

This library was chosen due to the ability to merge the capabilities of NumPy with the flexibility that spreadsheets and SQL queries in relational databases allow for. It allows for indexing, reshaping, slicing and dicing, aggregations, and selection of subsets of data. The stage in which we are applying these libraries is one the most important ones in data analysis, which is data preparation and cleaning, and for this we will be using pandas, also because it's an all-in-one tool. Since other tools had limitations we tested and this one proved to be able to do all the manipulations needed.

The reason pandas was created was precisely to be able to do all the data manipulation needed

which was not available in any other single tool available at the time. This included the need for data structures with labeled axes with data alignment; integrated time series; equal data structures for time series and non-time series; flexible handling of missing data; merge and other relational operations of databases like SQL. [58].

matplotlib is the most popular Python library for producing 2D-graphs and visualizations and integrates rather well with the rest of the Python ecosystem making it a sound choice for immediate visualization of the data.

To conclude this section of the tools used we describe the last tool used in our pipeline described in Chapter 4. To fulfill the visualization needs of our proposal we needed an appropriate BI tool to support the data warehouse also described in Chapter 4, in this tool, specialists would be able to analyze the data that has been prepared up to this point. **BI programs become particularly more interesting in cases where an organization already struggles with the problems that Excel spreadsheets pose and already has specialized technicians that don't have an IT background**.

After studying alternatives like Tableau and Qlikview we opted to use Microsoft Power BI since it feels very familiar and comprises quite an extensive list of features. This tool allows for flexibility and interactivity when it comes to data. By being as easy as drag-and-drop it takes away the step of having to check if all the data is there or if it is correct (this step is covered prior in our system architecture). By focusing on the data, the users can spend more time analyzing and trying to find interesting patterns in data. One of the most interesting aspects of BI tools is how much more efficient the processes become with these tools' help, as such, a tool like this would be essential for our work.

One of the reasons that motivated the choice for this BI tool was the fact that it was very simple to use and felt very familiar where users could be empowered to explore the data in whole new levels. This factor was essential in the choice for the tool since the performance of architecture depends highly on how the users develop mastery over it.

In Power BI, information can be imported from a variety of different sources from databases to data warehouses of all kinds to Microsoft Excel or can even be introduced manually. It is based on the creation of reports inside Power BI Desktop which can be later published to the web in Power BI Service or integrated into dashboards to aggregate all the important information or key performance indicators (KPIs) in one place to help in the process of decision making.

Inside Power BI there is a powerful scripting language Power Query/M which allows for simple and complex operations. It can handle very large datasets and the creation of calculated columns from other columns or condition-based. Power Query also allows to refresh data and impute values for example as well as dealing with problematic instances in columns.

The following two sections focus on two different types of demonstrations for the same research problem although with slightly different approaches. The first one focuses more on the conception of

an extensive multidimensional data warehouse, Business Intelligence and the production of reports to comply with the needs of the organization; while the second demonstration has a simpler data warehouse model and poses as a proof-of-concept for the application of some data science techniques. To make the body of work of this Master Thesis more interesting we decide to divide our work into these two demonstrations that are somewhat different but address the research problem defined previously.

5.5 First Demonstration

5.5.1 Description of the original datasets

After choosing where to test our solution and meeting with the organization we decided to use a representative sample of datasets, also in their interest to test with our proposal. The reason for this alignment in goals was due to their interest in developing the organization's use of digital technologies. With this project, they would be able to delve precisely into digital transformation and data mining.

The first datasets they wanted to bring into this new business model were of public finance and, to be more specific, relating to the National Health Service (SNS in its Portuguese abbreviation). This data ranged from the number of employees included in the SNS, to how much money was being spent on payments, whether to people, medicine, or technicians. It was a very rich dataset containing over 100 attributes relating to the Portuguese Hospitals that are part of the SNS.

The three datasets considered were:

1. SNS Accounting Information
2. Human Resources (HR) Information
3. Public Hospitals' Accounting Information

To describe the original datasets used in this master thesis we decided to divide them into logical categories of the field of public finance. These categories were: entity name and general information, users' reach, primary health care activity, assets and liabilities and estate, pharmaceuticals, human resources (both in number and in financial value), depreciation and EBITDA, and late payments. These datasets when cross-analyzed allow us to have a richer analysis. For instance, by having both the number of people working on a certain category and the value spent on their wages, we can have the average cost per employee, among other metrics.

The first dataset included attributes like the values for the SNS account, ranging from revenue and expense, balance as well as the particular values that allow for the breakdown of the revenue and expense. There are also some variations in this dataset (parallel to the real and observed values we have the variation and the value that was predicted to be spent).

The HR dataset is fairly simple and includes the information for each of the entities, the year, and the value by category. The category list is quite comprehensive, it ranges from interns and doctors to nurses, to all the technical staff that makes up a Hospital.

Entidade	Código	Ano	Médicos S/ Inter	Médicos Inte	Enfermeiros	Técnicos Sup	Técnicos de [	Assistentes T	Assistentes C	Técnicos Sup	Técnicos de I	Outros	Total Geral
Administração Central do Sistema de Saúde	1	2014	2	0	2	0	0	37	5	0	0	106	153
Administração Central do Sistema de Saúde	1	2015	3	0	2	0	0	34	5	0	0	24	174
Administração Central do Sistema de Saúde	1	2016	3	0	3	0	0	35	6	0	0	24	193
Administração Central do Sistema de Saúde	1	2017	2	0	4	0	0	34	6	0	0	26	194
Administração Central do Sistema de Saúde	1	2018	1	0	4	0	0	37	7	0	0	27	200
Administração Central do Sistema de Saúde	1	2019	1	0	4	0	0	36	7	0	0	27	198
Centro Hospitalar Universitário Cova da Beira, E.	1001	2014	108	53	398	53	53	170	296	53	53	87	1202
Centro Hospitalar Universitário Cova da Beira, E.	1001	2015	118	55	402	55	55	167	298	55	55	5	1222
Centro Hospitalar Universitário Cova da Beira, E.	1001	2016	116	52	387	52	52	166	290	52	52	8	1193
Centro Hospitalar Universitário Cova da Beira, E.	1001	2017	117	70	388	70	70	162	290	70	70	9	1208
Centro Hospitalar Universitário Cova da Beira, E.	1001	2018	113	71	386	71	71	161	303	71	71	9	1221
Centro Hospitalar Universitário Cova da Beira, E.	1001	2019	117	76	403	76	76	158	314	76	76	11	1261
Centro Hospitalar Médio Tejo, E.P.E.	1003	2014	134	59	610	59	59	178	450	59	59	59	1644
Centro Hospitalar Médio Tejo, E.P.E.	1003	2015	136	51	659	51	51	176	473	51	51	11	1719
Centro Hospitalar Médio Tejo, E.P.E.	1003	2016	138	59	699	59	59	173	515	59	59	14	1825
Centro Hospitalar Médio Tejo, E.P.E.	1003	2017	157	67	731	67	67	172	513	67	67	14	1886
Centro Hospitalar Médio Tejo, E.P.E.	1003	2018	154	68	747	68	68	176	532	68	68	20	1935
Centro Hospitalar Médio Tejo, E.P.E.	1003	2019	163	71	775	71	71	185	548	71	71	20	2008
Hospital Distrital Figueira da Foz, E.P.E.	1008	2014	73	39	183	39	39	56	122	39	39	27	555
Hospital Distrital Figueira da Foz, E.P.E.	1008	2015	77	44	191	44	44	60	123	44	44	2	575
Hospital Distrital Figueira da Foz, E.P.E.	1008	2016	85	40	205	40	40	61	123	40	40	4	598
Hospital Distrital Figueira da Foz, E.P.E.	1008	2017	87	52	203	52	52	59	130	52	52	4	614
Hospital Distrital Figueira da Foz, E.P.E.	1008	2018	89	45	207	45	45	61	137	45	45	7	631
Hospital Distrital Figueira da Foz, E.P.E.	1008	2019	92	57	226	57	57	62	147	57	57	7	676
Hospital Distrital S. Maria Maior, E.P.E.	1009	2014	44	32	169	32	32	41	134	32	32	21	465
Hospital Distrital S. Maria Maior, E.P.E.	1009	2015	52	29	178	29	29	43	140	29	29	5	495
Hospital Distrital S. Maria Maior, E.P.E.	1009	2016	60	32	174	32	32	43	140	32	32	6	502
Hospital Distrital S. Maria Maior, E.P.E.	1009	2017	59	40	180	40	40	49	128	40	40	4	507
Hospital Distrital S. Maria Maior, E.P.E.	1009	2018	63	35	184	35	35	50	131	35	35	5	516

Figure 5.1: Example of the original dataset for the HRinformation

The final dataset, and possibly the richest one, is the one regarding the accounting information of all the Hospitals that belong to the SNS. For each of the entities, we have information regarding the group of those entities, the year and 77 different attributes regarding the accounting information, the patient reach and types of primary health care activity performed by each entity.

Entidade	Sigla	ARS	Tipo	Grupo de financiamento	Grupo de benchmarking	Ano	População abrangida/ residente	N.º Primeiras consultas	N.º Consultas subsequentes	Internamentos (doentes saídos)	Episódios de GDH de ambulatório	Urgência (atendimentos)	Sessões em HD	Ativo	Ativo não corrente	Ativos fixos tangíveis - Quantia escriturada bruta [430+431+432+433+434+435+436+437]	Ativos fixos tangíveis - depreciações e imparidades acumuladas [438+439]
CENTRO HOSPITALAR BARREIRO MONTIJO, EP	CHBM	LVT	CH	C	C	2018	213 584	42 594	130819	13249	6 974	152122	16554	n/d	n/d	n/d	n/d
CENTRO HOSPITALAR DA COVA DA BEIRA,EPE	CHUCB	Centro	CH	C	C	2018	87 869	50 090	98657	9959	4 543	72335	18102	63 858 128,74	39 938 719,07	72 357 674,68	32 773 211,36
CENTRO HOSPITALAR DE ENTRE DOURO E VOU	CHEDV	Norte	CH	C	C	2018	274 856	84 768	187366	17176	15 644	210426	19189	52 892 949,60	8 757 365,31	51 528 688,00	44 854 629,48
CENTRO HOSPITALAR DE LEIRIA, EPE	CHL	Centro	CH	C	C	2018	363 404	86 882	180219	22984	12 634	196495	14170	80 228 510,72	19 504 382,67	55 979 914,52	37 151 283,45
CENTRO HOSPITALAR E UNIVERSITÁRIO LISBO	CHULC	LVT	CH	E	E	2018	327 416	189 989	535015	43279	23 775	247053	23150	242 485 291,54	86 346 262,20	295 408 554,33	211 400 375,21
CENTRO HOSPITALAR E UNIVERSITÁRIO LISBO	CHULN	LVT	CH	E	E	2018	248 120	181 841	535497	38473	19 960	247542	57581	n/d	n/d	n/d	n/d
CENTRO HOSPITALAR DE LISBOA OCIDENTAL,E	CHLO	LVT	CH	E	E	2018	257 372	114 390	347307	24409	23 185	158618	21833	136 707 522,56	53 516 428,56	180 340 844,96	129 817 871,65
CENTRO HOSPITALAR UNIVERSITÁRIO DE SAO	CHUSJ	Norte	CH	E	E	2018	330 386	196 068	548344	39912	36 929	255098	73087	233 234 692,26	81 468 763,18	199 253 221,87	128 686 546,32
CENTRO HOSPITALAR DE SETUBAL, EPE	CHS	LVT	CH	C	C	2018	233 516	77 384	168301	13949	8 343	149006	23780	n/d	n/d	n/d	n/d
CENTRO HOSPITALAR DO BAIXO VOUGA, EPE	CHBV	Centro	CH	C	C	2018	289 914	63 641	158608	16099	8 734	172321	13156	34 881 581,64	12 262 813,03	58 681 493,56	46 597 849,31
CENTRO HOSPITALAR DO MEDIO AVE,EPE	CHMA	Norte	CH	B	B	2018	244 361	47 000	122308	9959	6 394	134127	10092	35 299 157,43	9 082 522,92	36 561 899,22	27 479 376,30
CENTRO HOSPITALAR DO MEDIO TEJO, EPE	CHMT	LVT	CH	C	C	2018	182 067	75 828	105317	15614	9 184	145106	19103	84 035 786,26	54 251 568,64	117 425 946,75	63 315 861,41
CENTRO HOSPITALAR DO OESTE, E.P.E.	CHO	LVT	CH	B	B	2018	290 782	44 003	95135	13311	5 633	177654	11918	52 185 632,79	11 207 622,98	45 225 577,91	34 017 954,93
CENTRO HOSPITALAR E UNIVERSITÁRIO DO PO	CHUP	Norte	CH	E	E	2018	302 891	179 543	501655	30507	5 214	147526	25600	185 647 305,28	100 869 998,36	218 321 673,37	118 587 024,82
CENTRO HOSPITALAR DO TAMEGA E SOUSA,EI	CHTS	Norte	CH	C	C	2018	519 769	117 791	192455	20513	14 318	191444	33841	91 299 095,82	44 975 074,63	114 866 683,96	70 576 042,62
CENTRO HOSPITALAR E UNIVERSITARIO DE CO	CHUC	Centro	CH	E	E	2018	362 274	226 821	665913	58126	19 680	297654	52883	306 224 403,85	73 644 106,73	276 259 685,26	203 142 054,95
CENTRO HOSPITALAR POVOA DO VARZIM - VI	CHPVVC	Norte	CH	B	B	2018	142 941	29 660	61065	7298	2 330	74639	5824	12 297 920,39	3 592 378,78	12 619 283,22	9 200 370,75
CENTRO HOSPITALAR TONDELA-VISEU, EPE	CHTV	Centro	CH	D	D	2018	267 633	73 040	184539	20674	10 910	163531	57818	121 270 330,15	38 154 099,18	94 777 189,05	56 811 982,70
CENTRO HOSPITALAR TRAS-OS-MONTES E ALT	CHTMAD	Norte	CH	D	D	2018	273 263	77 922	227718	24138	34 924	182217	17062	116 383 870,92	47 285 814,41	126 186 763,05	81 831 505,58
CENTRO HOSPITALAR UNIVERSITÁRIO DO ALG	CHUA	Alentejo	CH	D	D	2018	451 006	79 331	202054	29265	17 674	348652	34152	135 121 617,92	68 393 790,22	173 952 569,73	106 155 359,08
CENTRO HOSPITALAR VILA NOVA DE GAIA/ES	CHVNGE	Norte	CH	D	D	2018	335 589	157 814	347609	22861	25 726	177768	31867	n/d	n/d	n/d	n/d
HOSPITAL DA SENHORA DA OLIVEIRA GUIMAR	HSOG	Norte	H	C	C	2018	256 660	64 398	188508	20790	12 274	138685	25635	38 792 449,79	8 856 449,83	70 704 805,14	62 045 133,83
HOSPITAL DISTRITAL DA FIGUEIRA DA FOZ, EP	HDFF	Centro	H	B	B	2018	107 541	32 938	58332	5640	5 499	74686	5932	19 175 263,00	9 631 316,58	23 853 274,32	14 652 760,60
HOSPITAL DISTRITAL DE SANTAREM, EPE	HDS	LVT	H	C	C	2018	196 620	40 908	102979	14939	9 341	132653	13496	97 316 544,57	50 609 514,73	78 263 593,15	29 574 669,19
HOSPITAL DO ESPIRITO SANTO DE EVORA, EPE	HESE	Alentejo	H	D	D	2018	166 726	55 824	133652	11359	10 243	74127	12013	41 598 691,98	18 134 258,88	57 526 599,60	42 060 752,85
HOSPITAL GARCIA DA ORTA, EPE - ALMADA	HGO	LVT	H	D	D	2018	332 299	89 076	202529	21360	25 328	169039	10425	n/d	n/d	n/d	n/d

Figure 5.2: Example of the original dataset for the Hospitals' accounting information

As evidenced by Figure 5.1 and Figure 5.2, these datasets were stored in an Excel file with the following structure: in the first column the names of all the entities, followed by a column with the year

to which that row information belongs to, and followed by several columns containing the values for the attributes in which the name is in the header. Naturally, this would pose a problem in terms of data representation since certain visualizations and programs require something similar to a star schema format. This even allows for an improved analysis of the data since the information is properly organized. This will be covered more in-depth in the ETL and Data Warehouse sections.

From an initial analysis, these datasets had many problems, and we found even more in the ETL stage of this work. The preliminary issues were the organization of the information, empty values marked as “NA”, “n/d” or “n.d.” and changes in representation of values from year to year.

5.5.2 Data Governance and data inventory

We followed a framework for a data inventory which suggested that a table like the one in Figure 2.1 should be created so that the data inventory would include at least the essential mentioned metadata. Following this suggestion, we conceived Table 5.1 according to the ODI [2].

Table 5.1: Data inventory table adapted from [2]

Atributo	**Descrição**
ID	Identificador único do dataset
Nome	Nome do dataset
Área de operação	Área de operação dos dados
Sub-Área de operação	Sub-Área de operação dos dados
Propósito	Razão de recolha dos dados/fim
Criador do Dataset	Quem criou o dataset
Origem dos dados	Entidade de onde originaram os dados
Natureza dos dados	Qual o tipo de dados?
Gestor dos dados	Pessoa responsável pela gestão dos dados
Palavras chave	Que tópicos é que estes dados cobrem?
Localização	Onde está a informação armazenada?
Data de criação	Data de criação dos datasets
Periodicidade (frequência de atualização)	Regularidade com que os dados devem ser atualizados
Tipo de dados	Tipo de dados incluídos no dataset em questão
Formato dos dados	Em que formato se encontram os dados?
Direitos e restrições	Que direitos se detêm sobre os dados?
Descrição	Descrição dos dados em questão

This data inventory initially started with public finance health data and should be extended to all the working areas of the organization. We are unable to share the complete data inventory due to privacy reasons.

5.5.3 ETL

After covering the theoretical aspects in our background chapter we will here mention the practical aspects of how we performed ETL on our data.

Possibly one of the most challenging stages of the work we developed was the data cleansing step in which we also defined rules for the database onto which data was uploaded. There was a prior analysis of how the data was represented on the Excel file, to understand in advance what kind of problems we could encounter, some of these, already mentioned, such as the null values and other values not marked as such (i.e. ”NA”, ”n.d”).

In the Extraction phase we first gathered the data and then took a first look at it, as proposed in the 34-step ETL process. This data came from different entities, yet it was partially prepared to be used in our proposal, as this data was already somewhat processed by the organization.

In the Transformation phase, we took the data in its original format and performed all the manipulation needed to obtain the final data structure, as required by the data warehouse we designed. These datasets had information stored simply in columns across the span of time, quite different from what a data warehouse looks like. As we can see in Figure 5.2 we realized that importing the data, even after cleansing, would mean a weaker analysis, since that data does not inherently allow for hierarchies. From the way the data was structured we would not be able to drill-down on the data, for instance. From a theoretical point of view, actions like drill-down would require a different logical organization of the data. To allow for this we would need something like a star schema, which we will cover in more detail further on this section.

Before we delve into the transformation of the table itself we had to upload the partially clean table into the database to ensure the quality of the data. This meant to obey a set of business rules that are essential in the field of finance.

These business rules were encoded onto the database using two different mechanisms: the first one through the design of these database tables by requiring that some values be either not null or bigger than 0; and the second one, through integrity constraints. This way we could ensure that all the values that entered the database would be correct by respecting all the business rules defined prior to the importation of the data.

One of the most simple examples is the Assets of an entity having to be equal to the sum of the Equity and the Liability, this was programmed as an integrity constraint in the database. Another example of an integrity constraint, which was essential to ensure the data was correct, is the sum of the current and non-current assets needing to be the same value as the Assets. In the design of our database we also defined what data type each field would be. This posed once again some issues considering that some random values on the original Excel spreadsheet did not respect all the data types we defined when creating the database.

In this part of the ETL we programmed our constraints in SQL and used SQL Server Management Studio to import our data onto the database (an Azure SQL Server database). After having designed a robust database onto which only values that respected all these integrity constraints would be loaded, we could have much more confidence on the values we would work with. We performed one final validation of these results against the values of previous years as well as some independent values. At this point, we had all our raw data validated inside a database and ready to be further transformed.

After this first validation process there was yet still a need to transform the data to unlock even more possibilities to extract more knowledge from the data, such as allowing for hierarchies and drilling down the data. In practical terms there were two main approaches for this phase of ETL.

The first one, would be to use a data manipulation tool like Power Query to try and unpivot the columns, and the second one being to perform a script in which we would manually manipulate the table. These two different approaches would yield the same results. We eventually tried both of these but decided to perform the first manipulations through a script written in SQL (a query language we were already familiar with) and then further manipulations would happen inside the chosen BI tool.

Essentially, the script written in SQL performed unions based on the values of the original table with the raw data. These SQL unions would take the values in that table and make it so it would respect the organization of the data warehouse we designed. Eventually creating the fact table to establish associations with the corresponding dimensions.

Taking into account that there are hierarchical relations between the attributes, these unions would take the information in its original format, keep elements such as date and the groups that each attribute belonged to and present them in a hierarchical view, typical of data warehouses. As such, the information instead of being displayed in columns with the values beneath them, as seen in Figure 5.2, it would be displayed in separate columns, the first representing the attribute names and the following representing its value.

This script was created to transform the data tables we had into a data warehouse, which we will describe thoroughly in the next section. After this stage our data would be cleansed, in the desired format and in agreement with all the required business rules.

The last component of the ETL was the actual loading of the data onto our SQL Server database. This step was performed by also creating a script that loaded the data provided to us in Microsoft Excel sheets onto the empty already created database.

There were some minor transformations performed after the load stage but these were made specifically for the exploration of the data inside the BI tool. These were defining the correct types of data or establishing specific connections in the models.

This stage, due to the data validations and the use of business rules made evident that there were many incorrect values that needed to be redone and double checked. To obtain these values we needed

to contact the entities responsible for providing this data and wait for the correct values to come in, which delayed our work to some extent.

5.5.4 Data Warehouse

In Chapter 3 regarding the background for this master thesis, we covered some basic concepts of how to approach data warehousing and how to do it correctly. According to the studied literature and taking into account the type of problem we had we realized that for our data the best course of action would be to design a well structured data warehouse.

From the first dataset, regarding the HR numbers, we had one type of data which was the number of people employed for each one of the categories. On the other hand, the second major dataset had financial information in € (Euros), such as how much was spent on each one of the listed categories.

This last dataset, being the biggest one, contained information from 2013 to 2019 having around 77 different financial attributes. From what we gathered, to construct a data warehouse we would have to design it according to our needs by defining which of those attributes would be dimensions and which would be facts. It is clear that all the numerical values had to eventually figure as facts in our data warehouse.

Starting with the simplest and smallest dataset, the number of human resources would have to appear on a table of their own since these were a different type of data.

In practical terms, the data warehouse was designed prior to the transformation but was mainly obtained from the ETL stages as described in the previous section. After all the manipulation from the ETL stages we obtained the following data warehouse model.

The produced model, due to its data and how it was designed conforms to a galaxy schema or fact constellation. This is a collection of fact tables that have some dimensions in common as is the case for our data. These are usually more complex models and harder to maintain although we opted for it because it is more flexible and able to adapt to the data we worked with.

Figure 5.3 highlights a clear separation between the dimensions in the top and the fact tables on the bottom, to make it easier to read.

In terms of the data present on this data warehouse we have in dim_mov_assistencial the types of consultations from the hospitals (whether medical, nursing or others) and whether it belongs to primary or secondary care which are two major groups. This dimension table for instance connects to two fact tables the ones associated to the cost of these operations/consultations and the cost of these health care professionals to the SNS. From the start this type of analysis becomes much more interesting because it can allow us to cross information to understand, for instance, which groups are more expensive to the system and which produce the biggest income. Data in terms of Euros is common to fact tables such as fact_balanco, fact_pagamentos and fact_custos_outros, these all pertain to financial data.

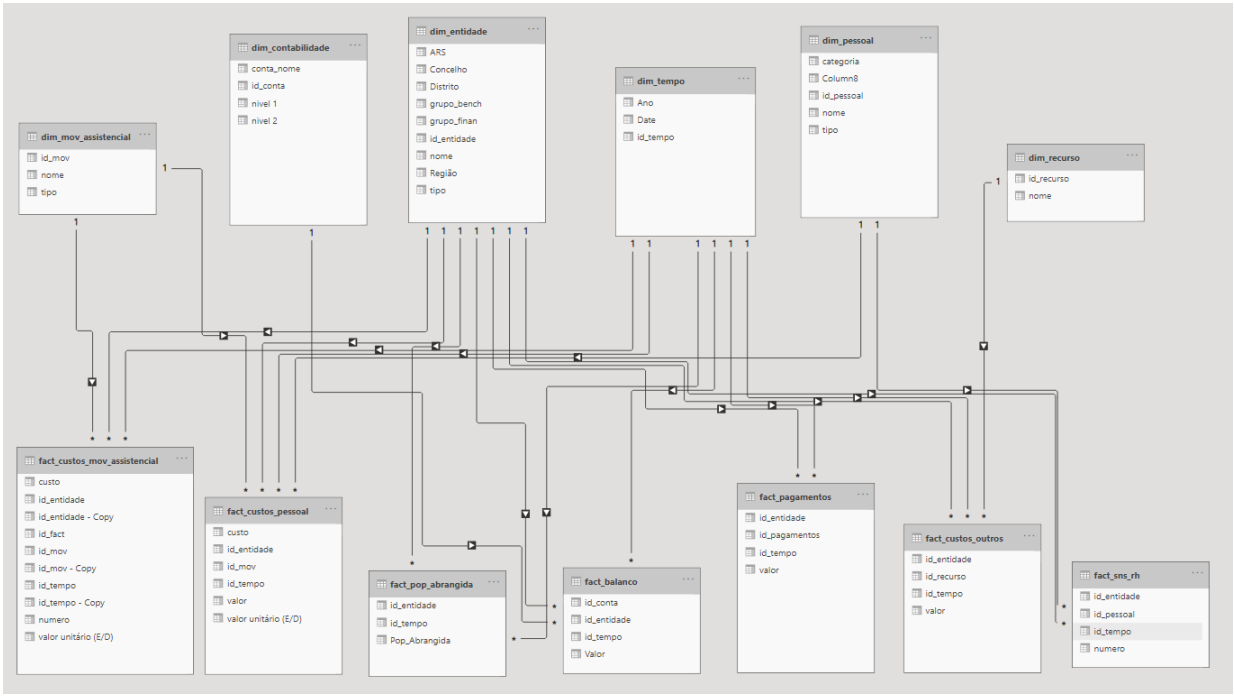

Figure 5.3: Representation of the multidimensional data warehouse model and its connections.

These dimensions and facts were chosen logically according to the needs of information.

Other examples interesting to analyze are the human resources numbers represented by categories in dim_pessoal and the corresponding values in fact_sns_rh. As in a true data warehouse this data can also be crossed with the amount paid to different categories to, say, calculate the average worker cost without the need to redesign or create a new table of information, all due to the connections established in this multidimensional model.

The last example just to help understand another interesting exploration of data is in terms of the entities that are part of the SNS, these can be crossed with facts such as the cost of primary or secondary care activity (painting a picture of value spent in assistance per hospital and category), the number of people that are served by a particular hospital, the balances of those entities (i.e. which ones have the biggest debt or make their payments later) and, for instance the evolution in human resource numbers, all these have the time dimension in common allowing to spot trends or understand evolution.

Besides the types of data, when describing a data warehouse it is also important to connect the dimensions to the business processes that need to happen. Table 5.2 is a representation of a bus matrix, the rows of this matrix represent business processes whereas the columns represent the dimensions involved, this allows us to better understand the multidimensionality of the model.

Concluding the data warehouse section of this demonstration we can get into report creation enabled by this multidimensional modeling architecture.

Table 5.2: Data warehouse bus matrix

business process\dimension	tempo	entidade	conta	pessoal	recurso	mov_assistencial
custos_mov_assistencial	X	X				X
pop_abragangida	X	X				
pagamentos	X	X				
balanco	X	X	X			
sns_rh	X	X		X		
custos_pessoal	X	X		X		
custos_outros	X	X			X	

5.5.5 Reporting

One essential part of data analysis is the communication and exploration of results. By having created the data warehouse presented in the previous section we made the data ”unbreakable”, being able to be sliced and diced every which way needed. This made our analysis component stronger and more flexible. This flexibility is key in terms of report creation as will be described further ahead.

This stage of our pipeline was highly influenced by the requirements of the organization where this work was developed, as such our work complies with their business needs.

The first challenge our system had to overcome was to be able to allow for the production all the reports that were being made in the prior business process. Due to the design of the data warehouse and logical structuring of the models we are able to see data in virtually any way possible.

Having all the data and the model connections established for the data warehouse we were able to start creating the visuals required to fulfill the organizations reporting needs.

As some brief examples of the data types were provided in the previous section we will now present some results that our multidimensional model was able to yield. Although from the theory of data warehouses these visualizations give almost endless possibilities for the exploration of data.

The first and most simple examples are just bar charts in which we used the columns of values from the fact tables and created a graph with the corresponding attributes from the dimensions columns. Doing so, and having the correct connections in the model we were then able to produce the following graphs.

Other simple examples include pie charts, visualizations made possible due to categories coded onto the data warehouse dimensions such as the following example, in Figure 5.6.

Another very interesting visualization that would have been impossible without a data warehouse multidimensional model are waterfall charts. These are often used for showcasing the variance, for instance, for one month to another where the detail is separated into what makes up the total value, as shown in the following figure. In figure 5.7, we see the decomposition of the elements that make up the value for the revenue, thus being titled revenue growth.

Other visualizations needed some further calculations based on the values which would be redun-

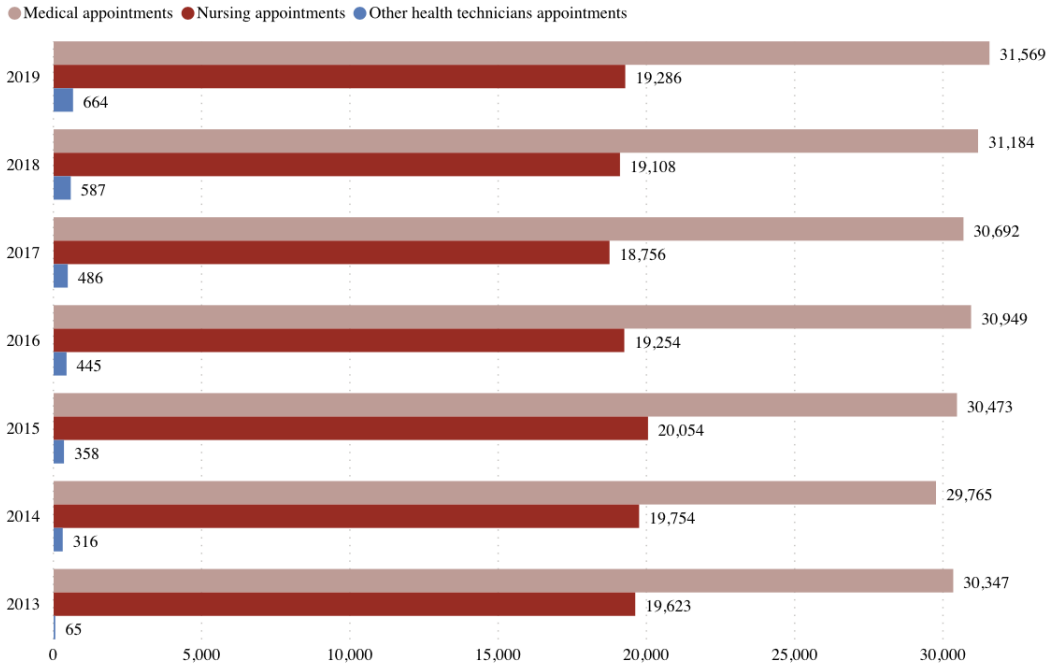

Figure 5.4: Primary care activity in thousands.

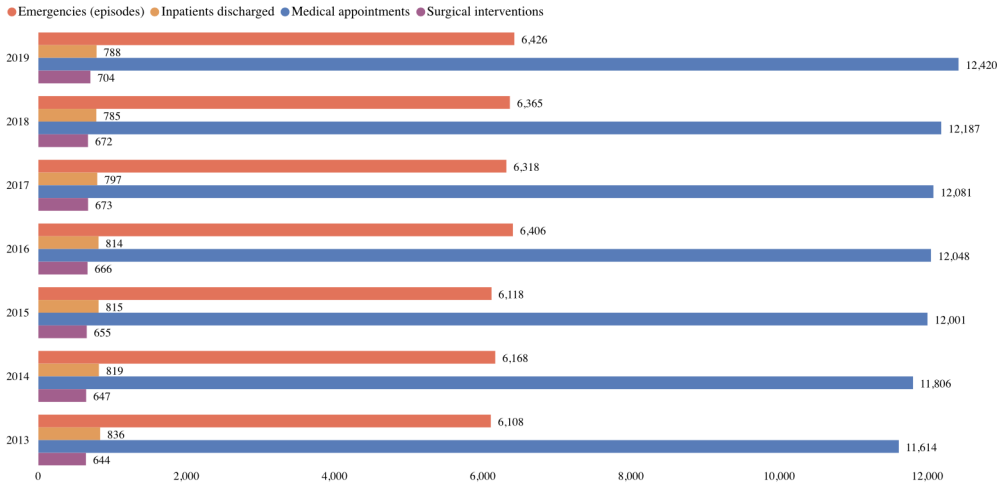

Figure 5.5: Secondary care activity in thousands.

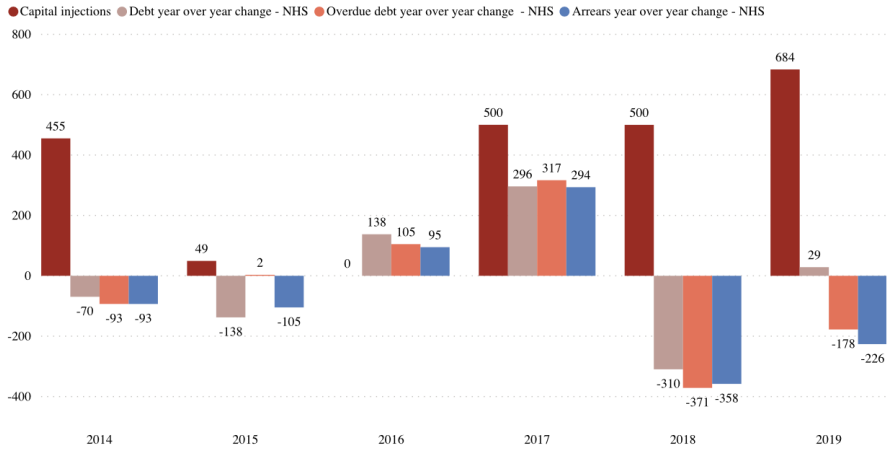

Figure 5.6: Capital injections in debt.

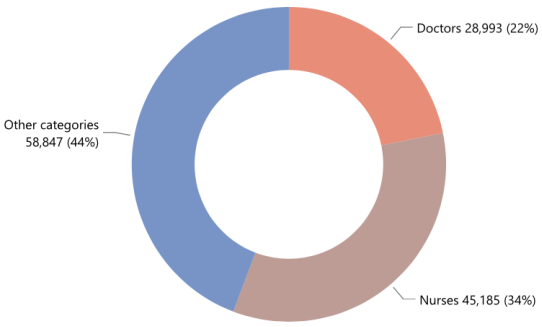

Figure 5.7: Major human resources categories for 2019.

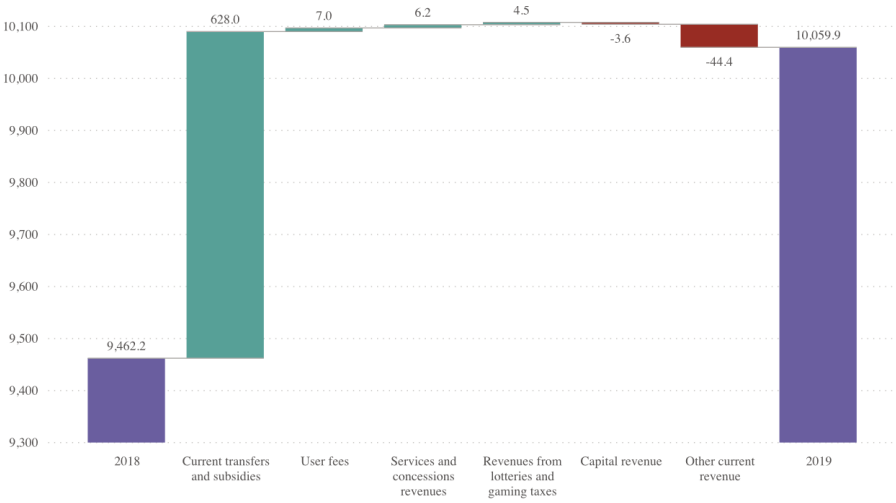

Figure 5.8: Contributions to revenue growth.

dant and unnecessary to be in the actual data warehouse. These calculated values we will be calling measures. These measures, for the particular case of this demonstrations, were created inside the BI tool we used and allowed for calculations of percentages or year-over-year changes and we can see them evidence in Figures 5.8 and 5.8.

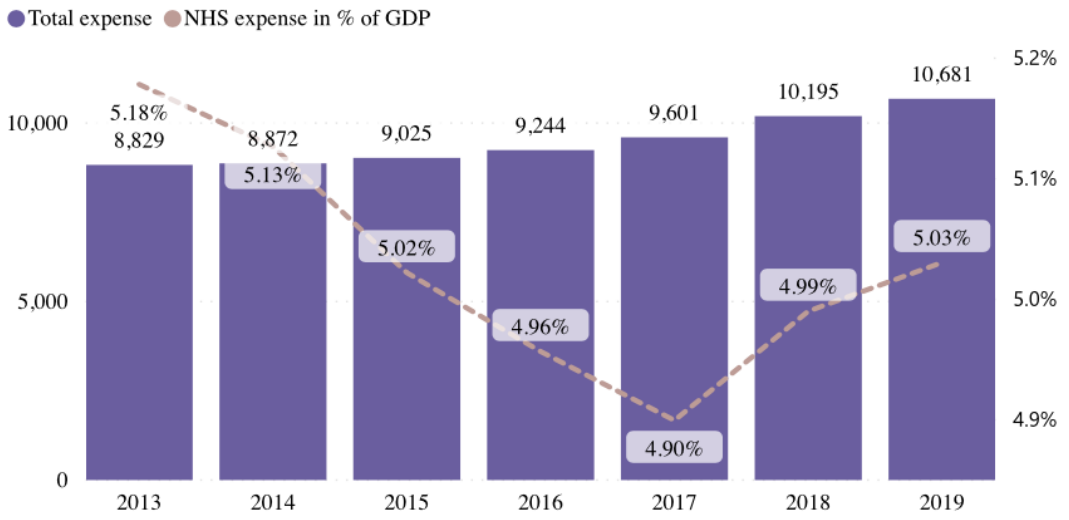

Figure 5.9: SNS fiscal context.

One last example of the use of conformed dimensions or information crossing from different facts is the ability to, taking the number of employees per category and the amount of Euros spent with those categories, calculate the average expense per employee. Which we can see in the following Figure 5.8. On the left-hand side we can also see the variation in human resources expenses through the years, also a calculated measure.

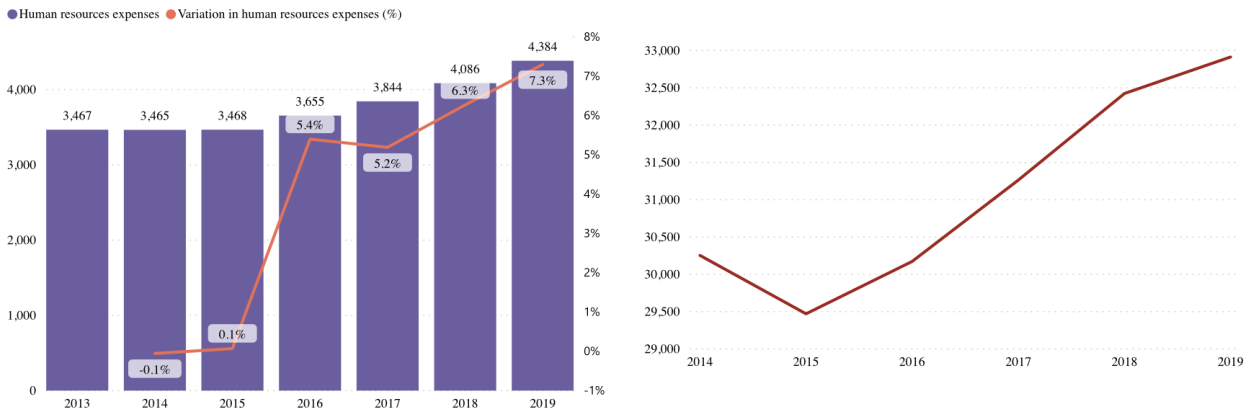

Figure 5.10: SNS fiscal context.

It would have been ideal to be able to compare the results from our solution with previous results of

similar implementations (of the same field). But, to the best of our knowledge, this was the first type of analysis with this data and domain inside the organization and, as such, we were not able to compare it to previous results. We were only able to compare it to the generality of their processes that use spreadsheets, which we described previously.

Due to the nature of the data in this demonstration, further analysis in terms of data science would yield very poor results due to how limited these time series were.

As such, we decided to continue to develop our work with other data to prove further on that our system can be a solution to the research problem addressed in this master thesis.

5.6 Second Demonstration

The second component of this demonstration focused less on the elaboration of a complex data warehouse schema and more on the exploration of data science techniques. Which was also one of the demands of the organization to justify its improvement over classic Excel spreadsheets.

5.6.1 Dataset

The dataset chosen for this second demonstration contained information from the Social Security accounting information. Its previous business process followed an Excel spreadsheet and the calculations made on it. As of our proposal described in Chapter 4, we intend to solve our research problem by creating a system architecture that comprises all the steps from data extraction all the way to report creation.

This dataset although more simple in terms of number of dimensions was more complex in its own structure, it was organized through different levels of specialization as can be seen in Figure 5.1.

5.6.2 ETL

As for the previous demonstration this was the most labor intensive section. We used a scripting language to perform the manipulation of our data which was python.

The extraction phase was quite simple containing just the collection of a large dataset from the Social Security Account. This dataset belongs to just one entity and its validation is more simple than that of the previous demonstration, since there are no business rules to check for, at least to be implemented onto the data warehouse.

This is also a particular dataset since some missing values can be the norm. This is due to some attributes of social security well fare having existed in the past but no longer exist in the present, yet these have to be maintained to preserve past information.

The last subsystem in the extraction phase, according to the Kimball Group, is importing the source data into the data warehouse environment for further manipulation.

The following step is the transformation phase in which we performed the majority of the manipulations from the ETL stage. For this transformation we created a python script that would read the Excel file provided by the public entity responsible for these data and the only source of data for this demonstration.

In python we made use of a python library called pandas, which is very popular and commonly used for data manipulation and wrangling. Some essential concepts were obtained from the book ”Python for Data Analysis” [58].

This script after reading the Excel provided by the public entity, which always follows a predetermined format, proceeds to convert the data into a pandas DataFrame. The rest of our work for this section revolves around this data structure. After creating it, we separate this accounting data into expense and revenue. These are the two separate DataFrames that we will be managing henceforth.

Parallel to this work we were given another Excel sheet that would be the target manipulation of this original dataset. This is the in-house format for the analysis of this data, whenever there is a need to produce a new report this manipulation needs to happen. This is also a different view on the data making use of only the needed attributes.

After analyzing this file, we noticed that there were quite a few attribute aggregations and quite a few hierarchies, also, one of the reasons for the creation of a data warehouse. From the theory we studied, we know that there is no need to store anything but the values at their most atomic levels since the rest are just aggregations of those values.

As such, from the excel file we decided to encode the logical hierarchies and, to later be able to construct our data warehouse, we decided to create a python dictionary with tuples for the keys and values. The reason behind this was to encode the hierarchies needed for this data and to later use that dictionary to build the hierarchies in our data warehouse.

After creating the dictionary with the connections we went back to the two DataFrames mentioned before regarding the expense and revenue for the SS account. The first step after separating the two datasets was removing the values that were not at the most atomic level, instead we decided to create a list of the name and level number for the most granular data. This meant to use the dictionary to check what were the values that respected this condition. We then created a new Dataframe (one for the expense and another for the revenue) that only contained these values. We used a function from the pandas library called melt. This transforms the data from a wide format (values encoded through the columns) to a long format (values encoded with two columns one for the name of the attribute and another for the corresponding value). This operation is also often called ”unpivoting” in the BI community. This long format, and theory and practice tell us so, will allow us in the future to perform drill downs or

rollups.

As of now, these manipulations have led us to having three columns one for the dates, another for the attributes of the most granular level and another one for the corresponding values.

After this we needed to construct the rest of the dataset from the information encoded in the dictionary. In terms of the revenue, as an example, the most granular level happened to be just up to level five, so to be clear we renamed the column attribute to level five which is the corresponding level in this case for the revenue. The remaining levels were created from the mapping of that level five to level four and then level three and so on.

Due to the way our script was written it makes it adaptable to any problem. Thus, being able to change the way data is organized to one that follows data warehousing modeling.

After all these manipulations we obtained our next-to-final versions of the data warehouse in which we had information of the dates, the levels and the corresponding values.

To ensure the validity of our data warehouse we performed cross validation with one report already produced by the organization. Here we found some issues but after backtracking the wrong attributes into the script we were able to quickly correct them. These were issues in the creation of the dictionary (the data structure where the hierarchies were encoded).

The final step in this ETL process was to load the data onto a data warehouse which was also included in our python script.

5.6.3 Data warehouse

Following the work in the previous section our design of the data warehouse was mainly put in practice by our ETL transformations. These have allowed us to transform data into a completely new organization since it changes the way values are stored columns (wide format) vs attribute and values column (long format). This transformation is also what allowed us to construct our data warehouse model, from the fact table to the associated dimensions.

This data warehouse had a more simple multidimensional design although it required quite a lot of transformations as described in the previous section. The design of this data warehouse and also due to the nature of the data was being just across one dimension and one fact table. The social security account only has one one time dimension table and one dimension corresponding to the hierarchies these share a one-to-many relation to the facttable corresponding to either the revenue or the income for this public entity. This implies that the data warehouse follows a star schema having only two dimensions.

This design is quite ideal because it reduces redundancy and repeated data, since the fact tables are in the third normal form (used to reduce duplication of data, avoid data anomalies and make managing simple) and the dimensions are de-normalized (this is to maximize performance of queries).

This type of schema is desired for a variety of reasons, whether for its simplicity, its faster queries

due to requiring less table joins or for its common acceptance in BI tools, which is a very important component of our system to ensure its long term effectiveness in the organizations it should be applied.

Due to the foreign keys shared in our model we are able to cross data and seamlessly use the many tables in the model, this is detrimental in systems where there is a scale up in the solutions.

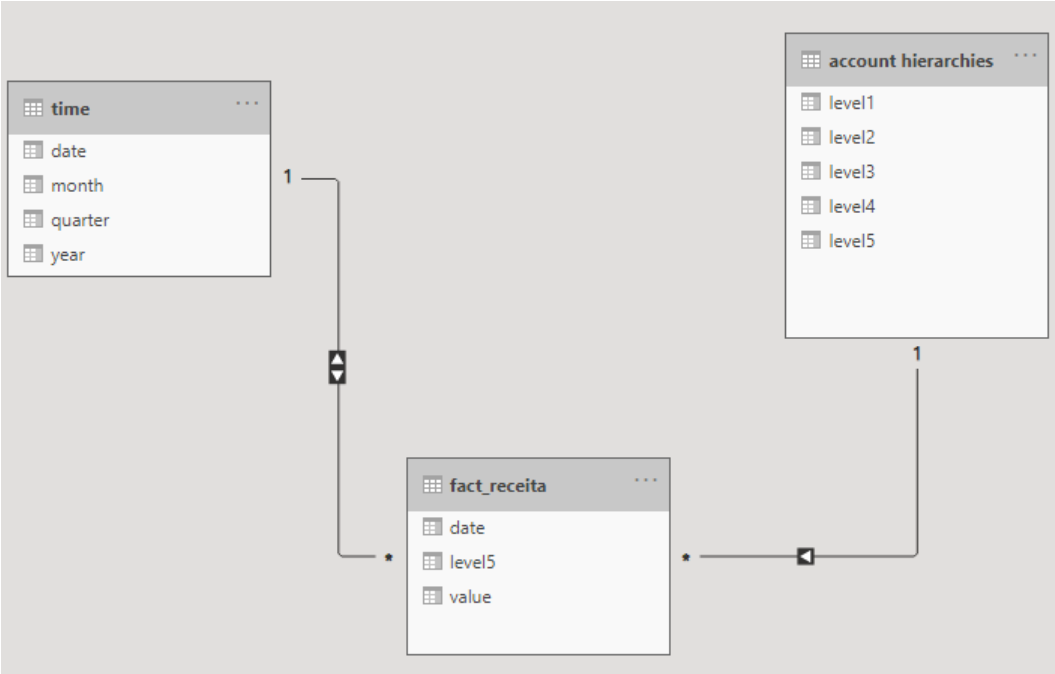

Figure 5.11: Representation of the Social Security account data warehouse model and its connections

After modeling the data into a warehouse multidimensionsal structure we could import it into the chosen BI tool although we decided, for variety and to continue its exploration using python and some specific data science techniques such as forecasting due to its interest in the field of finance.

5.6.4 Forecasting

To further test our system we decided to investigate a few of the most popular techniques in terms of forecasting,

Forecasting can be seen as a supervised leaning problem in the way that these predictions depend on past information to work.

To perform this we decided to look into a few of the most popular techniques for forecasting, these were ARIMA and SARIMA.

5.6.4.A ARIMA

To start with the ARIMA technique we diced our previously prepared data from our data warehouse, making only predictions for the most atomic values for simplicity. Forecasting attributes with null values was possible although this implied further treatment for the missing values, although this could affect the reliability of the forecasts even more. For this work, we opted to forecast only attributes that did not contain null values.

For this, the business explanation is that some attributes may have existed in some period in time but for several reasons (i.e. policy change) may have stopped being processed. In this particular case, for the social security account, there might have been support packages that might have been cut off in favor of other ones, for instance. And, for historic reasons, these must be maintained in the data warehouse. A practical example was the creation of complimentary subsidies to help people who were infected with the novel coronavirus (SARS-CoV-2).

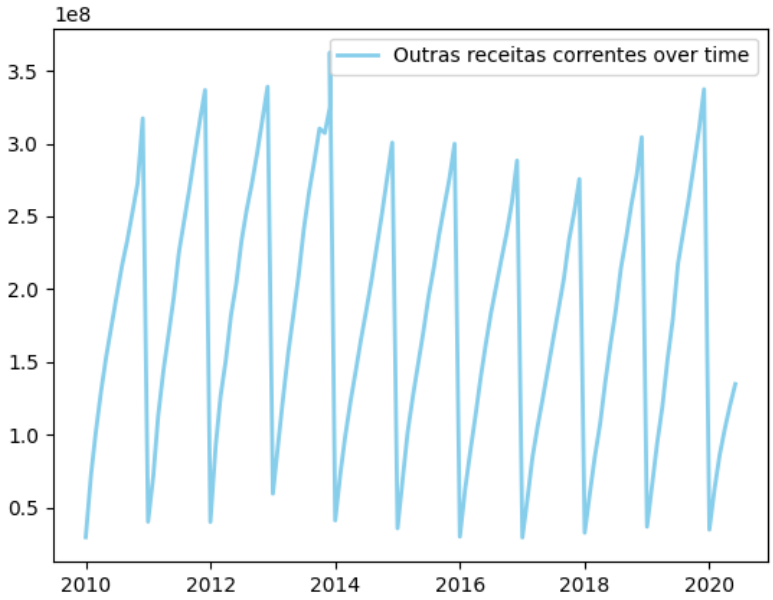

Figure 5.12: Representative plot of our time series (for the ”Outras receitas correntes”).

In figure 5.12 we find a representative example of our time series, series like this one were used to train our data (only the values from 2010 to 2019).

To proceed, we split our dataset into a bigger component for training and a smaller one for testing. After some trials we found that an adequate split would be to have 90% for training and 10% for testing, this roughly means to train with 9 years worth of data and to test for the last year of available data.

After this split we created a function to perform the fitting of the model to our data, generate and plot

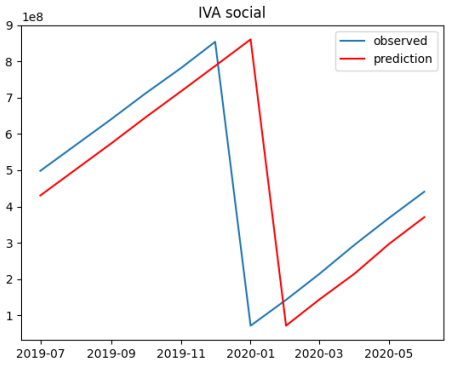

Figure 5.13: IVA Social (Social Tax) plot for the cumulative values using the ARIMA method with the specified parameters.

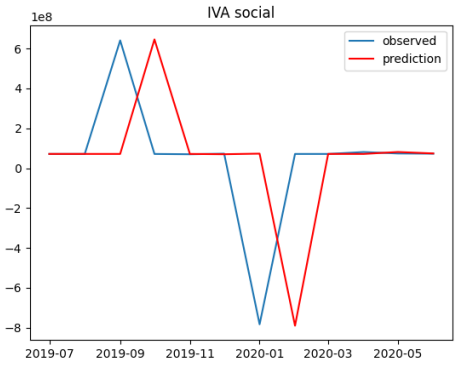

Figure 5.14: IVA Social (Social Tax) plot for the non-cumulative values using the ARIMA method with the specified parameters.

the forecasts and calculate the error associated to the forecasts which should be the difference between the observed value and the predicted one. These errors are also stored for later use.

The fitting of the model is done as previously described in Chapter 2 for which we tested a combination of values for p, d and q to understand which ones gave us the smallest error associated. The choices for the specific values were greatly influenced by the parameters that produced the lowest AIC and BIC. AIC and BIC are both estimators for information loss, the bigger the AIC and BIC, the more information the model lost and therefore the least quality it has. AIC is based on a frequentist inference, while BIC is based on Bayesian probabilities.

Using our evaluation metric which was the MAE (due to our predictions relating to a numeric value measured in Euros) and manually evaluating the forecasts we tried and tested with some changes to the parameters of the model, trying to reach the ones that were closer to the real observed values from the testing portion of our dataset.

The optimized parameters for the ARIMA model were ARIMA(0,1,0) and the produced plots were the following.

As stated before for the ARIMA method we will be analyzing two different plots, one that contains the cumulative values (current month plus previous month(s)) and the other the individual values for each month.

From these first plots regarding the social tax (IVA Social) we can tell immediately that, due to the values being summed in cumulative values (Figure 5.13), there is a little less variation yet both of them seem to present roughly the same amount of deviation from the observed values.

There are a few plots that behave in similar ways to these two, such as ”Receitas de Jogos sociais”, ”Contribuições e Quotizações”, ”Transf. do OE para cumprimento da LBSS”, ”Transf. do OE ...” and ”CGA - pensões unificadas”, etc. Here we can see a couple more as an example.

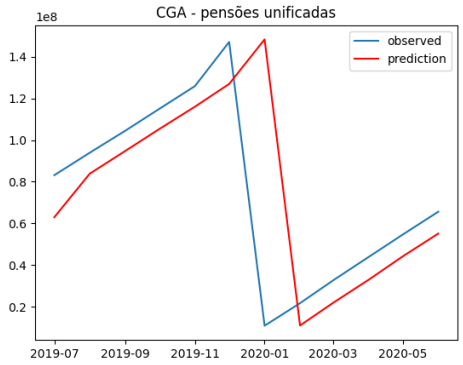

Figure 5.15: CGA - pensões unificadas plot for the cumulative values using the ARIMA method with the specified parameters.

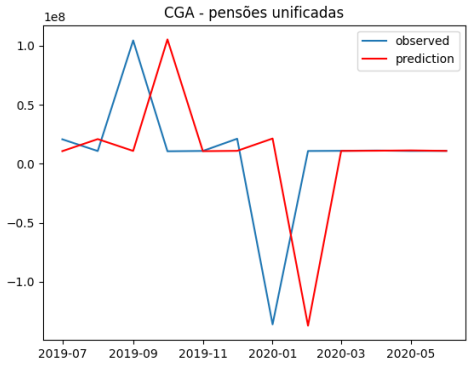

Figure 5.16: CGA - pensões unificadas plot for the non-cumulative values using the ARIMA method with the specified parameters.

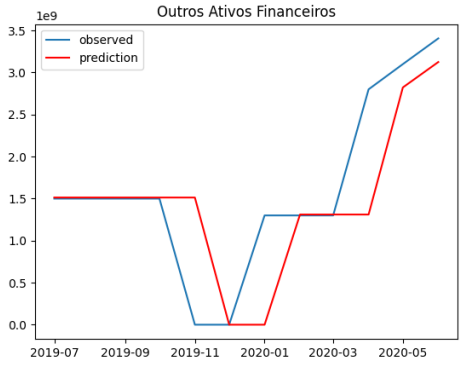

Figure 5.17: Outros Ativos Financeiros plot for the cumulative values using the ARIMA method with the specified parameters.

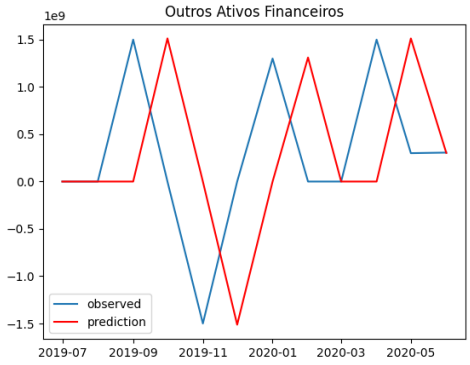

Figure 5.18: Outros Ativos Financeiros plot for the non-cumulative values using the ARIMA method with the specified parameters.

We can see that the model appears to be ”learning” these patterns correctly, since there seems to be a pattern where it is matching the peak or fall to the correct place, although there is a certain lag that we were not able to correct even by trying out with different parameters. The choice for these depended on which ones minimized AIC and BIC.

There are also other plots where our model and parameters seem to show slightly better results but yet there is always a variation that is failed in terms of the predicted values against the observed ones. Such is the case in plots like the following Figures 5.17 and 5.18.

Figure 5.17 seems to show a better prediction than 5.18 this is also due to the difference that the cumulative values seem to have in the training of our model.

The behavior of the model, concerning its accuracy of the forecast value, also depends on the domain and the difference of attributes. These domains vary from the type of data that attribute represents (e.g. from financial domains that originate from external entities and the notation rarely changes or is just

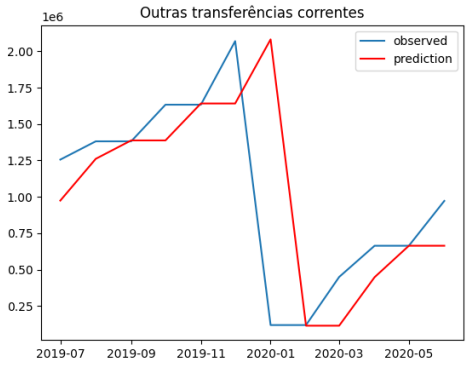

Figure 5.19: Outras transferências correntes plot for the cumulative values using the ARIMA method with the specified parameters.

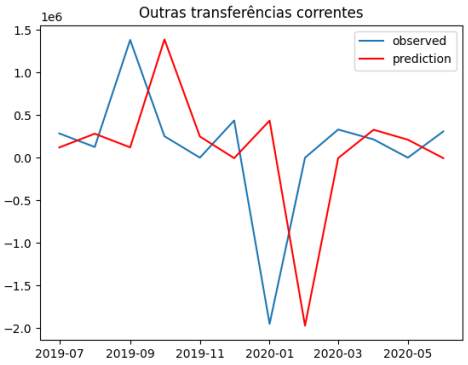

Figure 5.20: Outras transferências correntes plot for the non-cumulative values using the ARIMA method with the specified parameters.

slightly adjusted from year to year). For instance, there are some attributes that seem to have fairly steady movements such as mandatory transfers to the Social Security account that make up most of its revenue as a public institution.

There are also predictions where the trained model seems to struggle a little more, such is the case for attributes like ”Outras transferências correntes” as we can see in Figures 5.19 and 5.20. In these, particularly for the non-cumulative values, we see the model struggling with

Yet overall these results that the ARIMA method yielded seem to be fairly good results when considering how little data was used to train the model (only 9 years). Traditional models such as ARIMA seem to be able to pick up patterns quite easily having better results when there is a clear pattern. We can observe however that in both the cumulative and non-cumulative examples the predictions appear to be delayed by a month in comparison to the actual values.

As covered in Chapter 3, to measure the accuracy of our forecasts we will be using metrics such as the MAE and WAPE. This was because, after studying the MAE results, we noticed that it was not expressive enough as a metric, meaning the values were so different between themselves that it was impossible to compare them. We also noted that WAPE overcomes the ”infinite error” (when the error goes to infinity because there was a division by zero) which we observed for some of the attributes (such as ”Transferências de capital”, ”FEAC - POAPMC”, or ”Receita especial imposto jogo online”).

In Figures 5.21-5.24 we can see the plot of the above-mentioned metrics for their absolute value and the weighted percentage as well. We also note that we evaluated these metrics for the cumulative and non-cumulative values of the analyzed attributes.

Comparing the model in which we used the cumulative values with the one that had the non-cumulative values we can see that for the MAE metric the cumulative values’ model gave results 2 times better, roughly. Where for ”Contribuições e Quotizações” we observe the greatest error on 5e9

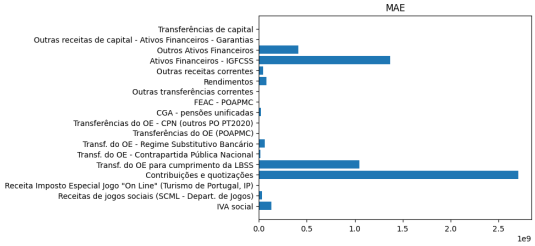

Figure 5.21: Plot of the MAE metric for the cumulative values using the ARIMA method with the specified parameters.

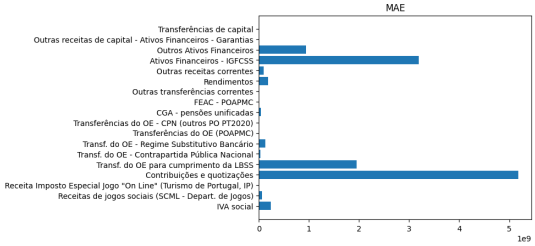

Figure 5.22: Plot of the MAE metric for the non-cumulative values using the ARIMA method with the specified parameters.

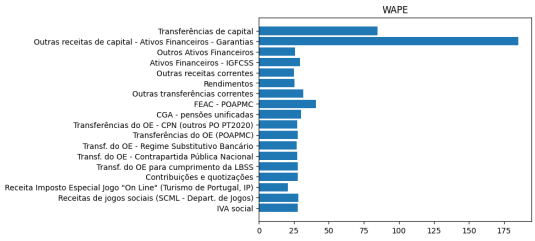

Figure 5.23: Plot of the WAPE metric for the cumulative values using the ARIMA method with the specified parameters.

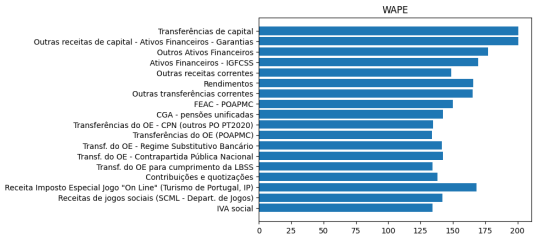

Figure 5.24: Plot of the WAPE metric for the non-cumulative values using the ARIMA method with the specified parameters.

(for the non-cumulative values) on the other we see it a 2.7e9€ (for the cumulative values).

Another example is ”Outros ativos financeiros” where for the cumulative values we observe an MAE of 0.5e9 and for the non-cumulative values we observe an MAE of 1e9.

Comparing the values for the WAPE we once again notice a fair percentage for the error in the model in which we used the cumulative values and a much worse WAPE percentage for the non-cumulative values. As an examples we have a WAPE around 25% for cumulative values of the ”FEAC - POAPMC” and around 600% for the non-cumulative values.

In general, this method proves to yield very poor results in terms of the forecast and we are led to conclude that ARIMA was not enough to make the predictions.

Naturally, ARIMA comes not without limitations, one of the most important ones is not supporting seasonal time series which means having repeating cycles for the data, for this reason we decided to test the SARIMA technique as well and compare the results.

5.6.4.B SARIMA

For the SARIMA technique we followed a similar process as we did for the ARIMA model.

We split our data into train and test and performed several tests to discover what parameters minimized AIC and BIC.

And as well as for the ARIMA model, SARIMA was also tested with both the cumulative values and

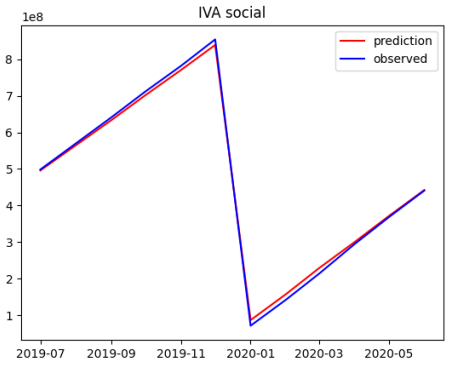

Figure 5.25: IVA Social (Social Tax) plot for the cumulative values using the SARIMA method with the specified parameters.

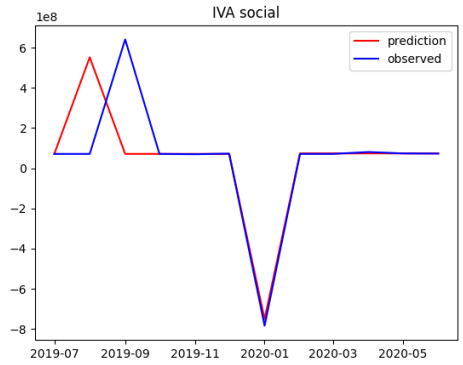

Figure 5.26: IVA Social (Social Tax) plot for the non-cumulative values using the SARIMA method with the specified parameters.

the non cumulative values.

From here, after defining the parameters we went on to fit the models for our training data, which was the same 90% for training and the remaining 10% for testing our forecasts. We adopted the same values as for the previous method to ensure we would be able to compare them fairly.

Having fit the model we were able to produce our forecasts for the SARIMA method. The following were the results obtained for the attributes we chose to forecast.

The optimized parameters for the SARIMA model were the following: SARIMA(0,1,0)(0,1,0,12) and the produced plots were the following.

As we have done for ARIMA we will be also analyzing two different plots for the SARIMA method one that contains the cumulative values (current month plus previous month(s)) and the other the individual values for each month.

Analyzing the first plots for the SARIMA method we can already tell that the results seem much more aligned with the observed values. We analyze here the same attributes as we did for the ARIMA method and we see the same behavior: the models that had the cumulative values seem to perform better than the ones with the non-cumulative values (Figure 5.25 against Figure 5.26, respectively).

Something important to note immediately is that there is no longer a lag for the time to which the values go up or down, at least for Figure 5.25 and 5.27. We can still observe that this lag exists, if only partially in the plot in Figure 5.26 and 5.28.

As with what happened with the ARIMA models, we observe behaviors similar to this one in some plots we did feel the need to display them all. An example is the following in Figure 5.26 and Figure 5.28. We can report the same behavior: missing the first increase in value but being very close to the first big decrease in Figure 5.28 for the non cumulative values plot.

Going against the trend that we were seeing for the generality of the SARIMA models, we see some

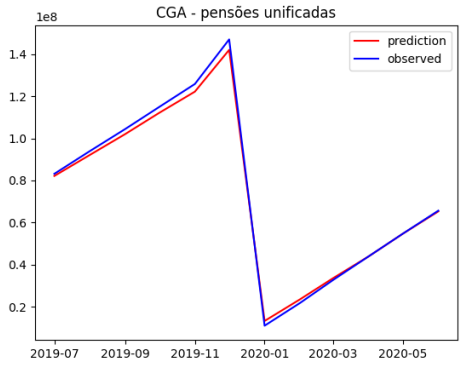

Figure 5.27: CGA - pensões unificadas plot for the cumulative values using the SARIMA method with the specified parameters.

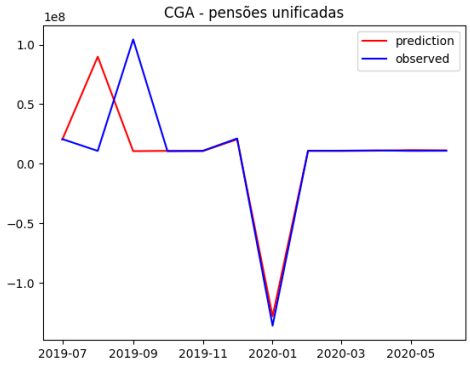

Figure 5.28: CGA - pensões unificadas plot for the non-cumulative values using the SARIMA method with the specified parameters.

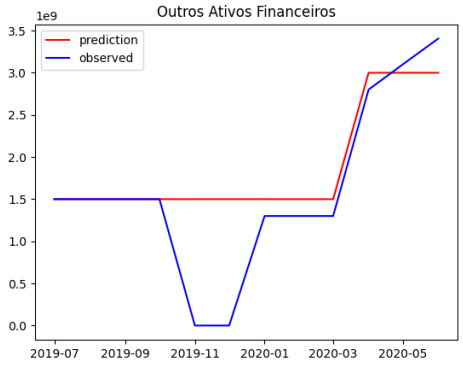

Figure 5.29: Outros Ativos Financeiros plot for the cumulative values using the SARIMA method with the specified parameters.

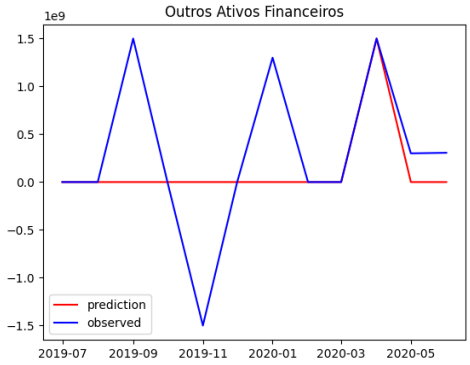

Figure 5.30: Outros Ativos Financeiros plot for the non-cumulative values using the SARIMA method with the specified parameters.

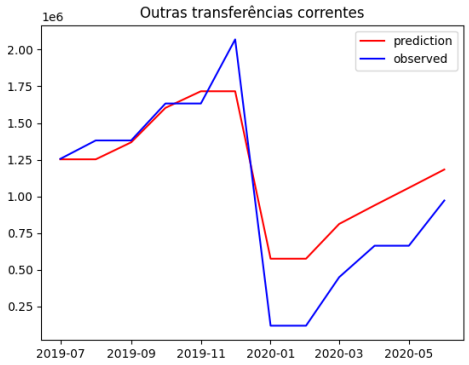

Figure 5.31: Outras transferências correntes plot for the cumulative values using the SARIMA method with the specified parameters.

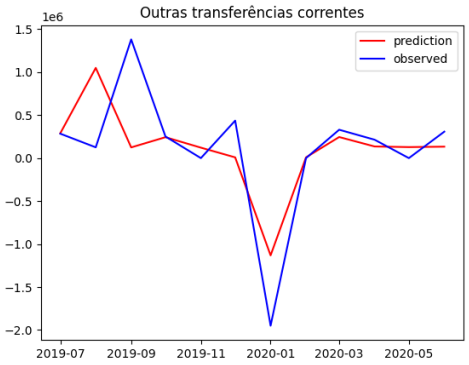

Figure 5.32: Outras transferências correntes plot for the non-cumulative values using the SARIMA method with the specified parameters.

plots that show the model making predictions that are more deviated from the observed value. For Figures 5.29 and 5.30 we see that the last three months of predictions are closer to the observed values, but these simply do not compare to the accuracy we saw for the Figures 5.25 and 5.27.

Similarly, we see other predictions that seem to deviate from the observed values, although appearing to follow the correct general trend, in Figures 5.31 and 5.32. Once again is not as accurate as the first set of predictions presented for the SARIMA models. Also in these figures (5.31 and 5.32) we see that the forecasts do not have such an expressive difference given that they are somewhat close to the observed value. Although we can still see a bit of a difference in a peak in Figure 5.32.

These changes in accuracy for the model can also be explained from the domain point of view. The values that we observe performing better must have a more constant or expectable behavior such as the case for ”IVA Social” (which finances the family protection subsystem), whereas there are other attributes that depend on external variables such as the FEAC - POAPMC which is a European fund that can change according to policy or needs.

The SARIMA plots show improved results over the previous ARIMA method. We confirm again that the model with the cumulative values seems to yield better results, once again.

At a first glance and without yet considering the chosen evaluation metrics we can observe that the SARIMA method appears to yield better results than the ARIMA method. This was already expected since from analyzing the data we could understand that there would certainly be some kind of seasonality. We saw this happening for both the cumulative values and the non-cumulative values.

As a whole the SARIMA method seems to provide results much closer to the actual observed values clearly showing that, in general, this appears to be a better forecast for these attributes, depending on the attribute.

Although, we should keep in mind that these greater results were obtained from training the model

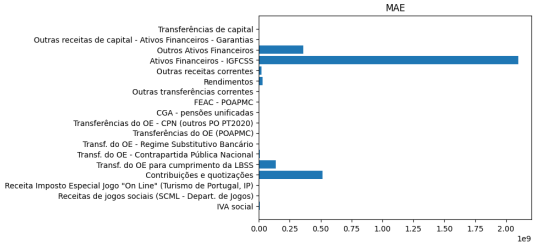

Figure 5.33: Plot of the MAE metric for the cumulative values using the SARIMA method with the specified parameters.

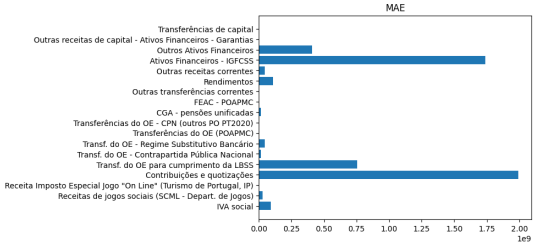

Figure 5.34: Plot of the MAE metric for the non-cumulative values using the SARIMA method with the specified parameters.

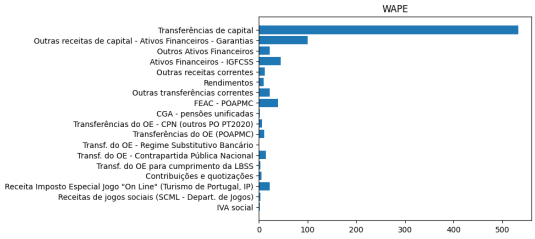

Figure 5.35: Plot of the WAPE metric for the cumulative values using the SARIMA method with the specified parameters.

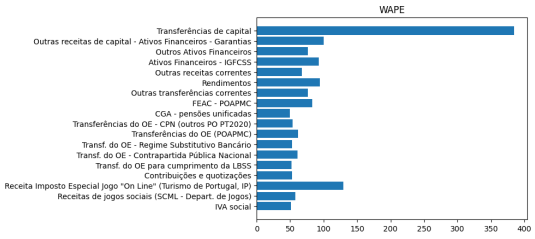

Figure 5.36: Plot of the WAPE metric for the non-cumulative values using the SARIMA method with the specified parameters.

using a very limited data set which included only nine years.

As we did for the previous method we will also be using two metrics commonly used, MAE and WAPE. The evaluation metrics for the SARIMA are represented in Figures 5.33-5.36.

In terms of the MAE metric we can see some differences between the cumulative and non-cumulative values, where, once again, the model with the cumulative values provided better results as we can see in Figure 5.33 and 5.34. Though there are a few attributes that present worse results for this metric (i.e ”Ativos Financeiros”, ”Transf do OE para cumprimento da LBSS” and ”Contribuicoes e quotizações”).

We also decided to use another metric for the evaluation of our models, WAPE, since we have many parameters with varying degrees of magnitude that this measure balances out through a weighted percentage (thus giving less impact to bigger variations in lower values).

Through this metric we can see the forecast associated with our errors for the SARIMA cumulative model and understand that these are better than the ones for their SARIMA counterpart (with the non-cumulative values), but also better than any of the forecasts created by any of the ARIMA models.

In general, SARIMA was slightly better than the ARIMA model, however it only had moderately good results for some of the attributes, whereas for some other attributes it had very poor results. This leads us to assume that for these attributes this model was not appropriate.

To better analyze the error associated with the best of our approaches, which was the cumulative values for the SARIMA method, we created another plot for the errors by removing the attributes ’Transferências de capital’ and ’Outras receitas de capital - Ativos Financeiros - Garantias’. As these had

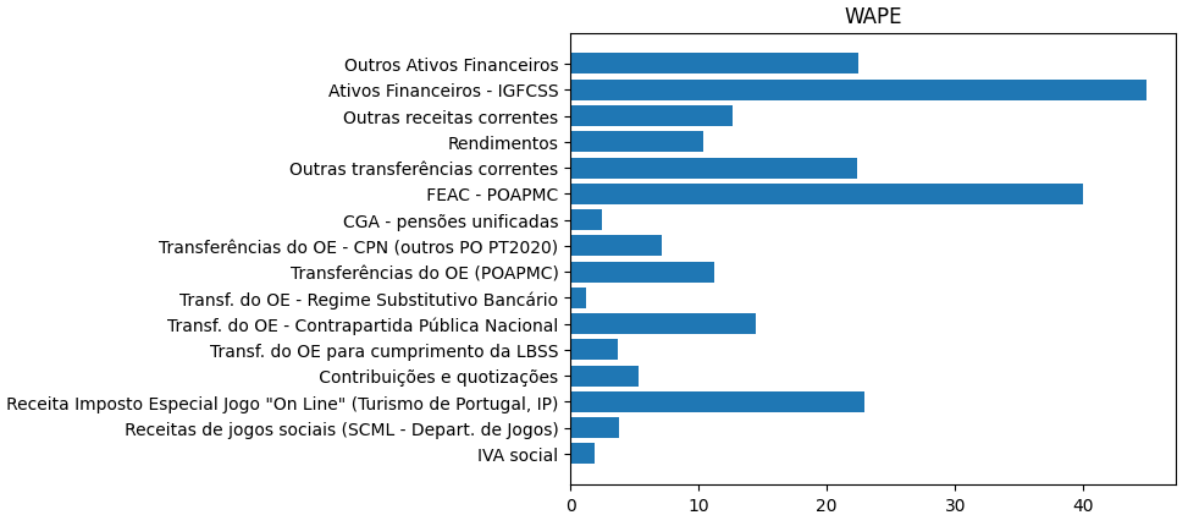

Figure 5.37: SARIMA-associated error plot for the WAPE measure having removed two outliers.

disappointing results and were impeding the analysis of the actual error values for the other attributes. In Figure 5.37 we can observe the percentage of error obtained using the WAPE measure. From this figure we can observe that this model performed rather well for a few attributes such as 'IVA social', 'Receitas de jogos sociais', 'Transf. do OE - Regime Substituivo bancário' and 'Transf. do OE para cumprimento da LBSS' which had an error percentage of under 5%, implying a forecast accuracy of 95%-99%.

On the other hand, for other attributes, the predictions were not so good, even reaching a 45% error in the worst case for Figure 5.37. There are also other attributes for which the predictions were quite poor. These are the two outliers removed from Figure 5.35 ('Transferências de capital' and 'Outras receitas de capital - Ativos Financeiros - Garantias') and also two other attributes, 'Ativos Financeiros - IGFCSS' and 'FEAC - POAPMC'.

As a whole we can gather that roughly, the values of WAPE for this method, SARIMA with cumulative values, are around 15% which implies an accuracy of 85% for the majority of the forecast attributes as shown in Figure 5.37.

The problems we eventually encountered with our results could possibly be fixed by training our model with more data, applying different forecasting techniques or even different parameters.

Summing up, we split our original dataset into train and test data, we used the first portion to train our model and we used the latter to evaluate the error between our forecast and the actual observed value. This, in practical terms, meant to use 9 years of monthly data to train our models and compare the produced model with one year. Yielding some different results between the produced forecasts in terms of the different techniques and attributes that we were able to forecast.

To measure the accuracy of the forecasts we decided to evaluate the error associated with the pre-

dicted values (when being compared to the real ones).

In the following Chapter we will be evaluating the system as a whole considering the produced artifact.

6

Evaluation

Contents

In this Chapter we present the evaluation of our system and it made sense to us to evaluate the two approaches separately.

In the context of our work we believe there should be two different instances of evaluation: one for the artifact we produced (which in this case is the system itself) and another one for the models we used for the data science aspects that have already been covered in the previous chapter with metrics like the error associated with the results. We believe these are subject-specific and adapted to the two main areas of this work.

6.1 Evaluation framework

The method we intend to use to evaluate the solution mentioned in Chapter 5 will be defined here.

One of the reasons for which we chose to follow the DSRM is due to the detailed way to evaluate the produced artifact. Jan Pries-Heje defined, in an article, the strategies for design science research [59]. There is a relevant distinction between two main types of evaluations artificial and naturalistic.

Artificial is for the evaluation of a solution in a non-realistic way whereas naturalistic is used for the evaluation of an artifact in its real environment (for instance within and organization). Artificial approaches may include laboratory experiments, field experiments, simulations, criteria-based analysis, theoretical arguments, and mathematical proofs, however, this type of evaluation may be unreal in some ways and therefore is a need to resort to naturalistic approaches.

The type of solution being developed in this work is, without question, naturalistic since we are going to be using it inside the actual organization for which it is destined. This means that it's going to be used for the environment it was developed for [59].

It is also interesting to note the differences between the evaluations ”ex ante” and ”ex post”. ”Ex ante” implies that evaluations take place before the system is constructed and ”ex post” implies that the evaluations are performed after the system is constructed. For the system we are developing we believe there is a mix of ”ex-ante” and ”ex post”.

As we are developing and improving a system that had a previous modus operandi we will simply be comparing the produced artifact with it.

As suggested by Pries-Heje for the actual evaluation of the product we can use an international standard (such as ISO 9126). This standard is divided into four parts that cover the quality of the model, external metrics, internal metrics and quality-in-use metrics following a set of characteristics.

These can include, for instance, the efficiency of the system, which is particularly interesting and relevant for all the work we have develop being one of the most important selling points.

According to Pries-Heje our evaluations is naturalistic and ex-post since we tested it inside of the organization for which it was designed as well as performed the evaluation after the system was com-

pleted.

The following sections will describe the evaluations for the work produced, specific for the two demonstrations described in the last chapter.

6.2 Artifact Evaluation

Artifacts in the Information Systems (IS) field are often considered systems, such as the case for our own artifact. These are created to solve a real life problem having functions and objectives for its purpose.

For the evaluation of the artifact we will be taking a IS approach following our choice for the methodology, using techniques commonly used to evaluate these artifacts. According to a paper on the subject [60] there are many ways to evaluate them, yet there is a need for adapting which ones make sense for the scope of the artifact itself.

Prat et al. suggest that the evaluation criteria should follow a hierarchy based on the theory of justification, conforming into three interrelated levels. These levels are system dimensions, evaluation criteria and sub-criteria.

According to their proposed hierarchy there are five system dimensions, 20 evaluation criteria and 12 sub-criteria [60]. From these we believe that our system relies most on two out of the five system dimensions, these being goal and environment.

Goal, as with any artifact from the Design Science Research Methodology (DSRM) being very goal-driven it was essential that our system were evaluated in terms of resolution of the problem for which it was created. For this Master Thesis, the goal was to solve the many problems associated on the use of using spreadsheet programs for storing, cleaning, shaping, transforming, visualizing and crossing data. There are technologies designed specifically for this which we have studied and used to solve our research problem. The fulfillment of the goals should be evaluated following three criteria efficacy, validity and generality [60].

Efficacy is the extent to which the goals were met, in this criteria our system was able to comply with all the demands of the organization in terms of structuring of the system and organization as well. The data was stored in a similar tabular form but, for our system it complied with the design of a data warehouse that allowed for more flexible manipulation of the data.

For this system we designed a more complex data warehouse that would encompass all the dimensions and facts relevant to the production of reports regarding the data being analyzed.

Validity means the degree to which the artifact is able to work correctly, or performs reaches the goal correctly. An example of how our system meets these goals is the implementation of the integrity constraints into our data warehouse that would only accept valid data, in terms of complying with prior defined and in place business rules. This can also encompass reliability which after validating the data

that went through our system we can confidently say that it is, indeed, reliable producing consistent and correct results.

The generality of an artifact implies how applicable it is to the broader problem. A broader goal for the artifact means a more general artifact. So much so, that we even applied our system to another problem different from the first one. Due to our research problem, and consequently our goal, being defined to solve the widespread problem of ”spreadsheet hell” we can say that it is also a general artifact.

As for the other system dimension relevant for our work, environment, Prat et al. suggest this should be evaluated using the following criteria: consistency with people, consistency with organization and consistency with technology.

This is a very relevant system because, the environment of IS artifacts includes people, organization and technology. It also made particular sense for our system because being an all-rounded solution for such a big problem its environment is also broad and need to be sound for all the smaller components to work correctly.

There are some limitations to this component of evaluation because our solution although complete is only partial to the whole transformation that is to happen inside the organization which was set in motion with some of the work produced for this Master Thesis.

The consistency of the environment for either people, organization and technology encompasses 10 sub-criteria. Utility measures the quality of the artifact in use, our system, for this criterion was met with excitement for this new technology as well as promises since after validating the data (common practice even in the previous business model) all that was left to do was to build reports and share the findings, which is part of the mission of the organization.

Understandability, can also mean ease of use of which our system can take a little time getting used to the learning curve is definitely worth the extra effort. This evaluation is also made in a point of view that the people who are going to be using this system have no previous knowledge of computer science fields such as databases or programming, this could be the reason why the system may seem a little less easy to use than a simple Excel spreadsheet. The only compromise to the ease of use of our system lies only on the multiple parts that compose it. From the data warehouse validation rules, to the process of loading and transforming the data for the data warehouse and also the learning curve for the new Business Intelligence software. Although we firmly believe that once we’re past this first period our system can become easier to use and even more reliable.

Ethicality means for the system to not put animals, people, organizations or the public at risk, this sub-criteria does not directly apply to our line of work since the only risk would be the leak of information that is in general, not the case since almost all of it is public domain or publicly accessible.

In terms of the fit of our system with the organization we believe that it is ideal since the work developed there benefices greatly of the use of our system. Considering the line of work of the organization

and since business intelligence and business analytics are at the core of its mission, our system was precisely designed to meet those needs.

When it comes to the criterion of consistency with technology the value of the produced artifact lies on it being a new layer built on new IT artifacts, which is also the case of our system. We make use of some of the most recent technologies to harness their potential and elevate our work.

According to the article [60], the evaluation should also consider the side effects that this system might have in its environment. From what we studied and, from a digital transformations stance, we know that these fundamental changes to an organization's business process can be met with some resistance from the people. Although from the literature this only happens when it is done incorrectly, there is a need to educate and empower the employees and helping them see the potential these new tools have for their work. By helping them achieve mastery we almost guarantee the success of our system in the long term, according to the literature.

As a matter of fact, our system not only meets the previous requirements it also surpasses some expectations in the field of technology particularly, the way the system is built allows to use the most modern technologies to treat, analyze and mine the data involved.

From the demonstration in the previous chapter, the interest behind some of the visualizations isn't on the visuals themselves but more on the efficiency that our system provides. By using our scripts, the work only needs to be done once and all the visualizations are readily available and ready to be updated with just one click with our artifact. Instead of having to manually use the values for the creation of new visualizations.

We believe this is a work with a lot of potential, that has been done in many other fields and is upcoming in many more. This digital transition has to be done mindfully and has to comply with norms and the evolution of the technologies themselves.

To further extend our evaluation we would also have compared our results to other works similar to our research problem, although we were not able to find any in the digital libraries we search on. This would have allowed us to study previous approaches to the same problem we address and maybe start from there. In the absence of such works we did all that was possible to ensure the completeness and correction of our system. One possible reason for the lack of articles in this area is that the approach of problems such as this one are usually addressed inside of companies and, as we know, these are not the most common authors of scientific articles on the processes of their own transformation. As such, the novelty of our work also lies here. Through our academic background in diverse fields such as Information Systems, Database Theory, Machine Learning and Visualization of Information we were able to develop this work from the ground up and establish a working pipeline and architecture for the implementation of a system like this one to solve our research problem.

7

Conclusion

Contents

In this chapter, we will be presenting the conclusions we were able to draw from this body of work.As we have shown in Chapter 6, our system architecture is able to reach its goals with the minor inconvenience of the learning curve for the organizations that use older technologies. Although we firmly believe the pros more than surpass the cons. Also with the extra added benefit of the improvement of process efficiency reducing pointless computations (as aggregations) already implement into our data warehouse's logical multidimensional model. Precisely due to this fact there is only the need to validate the finest granularity data once, at the moment they are uploaded onto the database.

7.1 Motivation

The motivation for this work lied deeply in the interest of solving a problem many companies still face nowadays - the incorrect use of spreadsheet programs beyond their capacity. Tools like Microsoft excel are very powerful although nowadays we have to our service tools that are more indicated for everyday data analysis tasks. As this type of work has been done for some years now, the novelty in our approach is in its simplicity and ability of adoption for companies/organizations starting their digital transformation.

At the core of this work is the will to solve a common, and also very costly, business problem as backed by our research. By taking a practical approach in this work we wished to demonstrate the essential steps for the implementation of a BI system from end-to-end.

7.2 Approach and Results

Following the DSRM, our approach was consummated with the realization of two different demonstrations. The first demonstration focused on following a pipeline and architecture defined precisely to solve the research problem at hands culminating in the use of a BI tool. The second demonstration also focused on following the same pipeline and architecture to solve the research problem, although the tools used to achieve this were different. We designed a simpler data warehouse so we could test a set of different tools for the BI component.

In the first demonstration, we used a very popular BI tool called Power BI, as requested by the organization in which this work was developed. For the second demonstration, we delved into business analytics using tools such as python, pandas, and matplotlib. For this, we covered what is commonly referred to as predictive analytics, also of the interest of the organization. From predictive analytics, we focused essentially on forecasting which we found interesting and also met the requirements of the organization.

More specifically, the first demonstration followed our defined pipeline all the way from source selection to cleansing, to load the data onto the warehouse and finally analyzing it or working with it inside the

BI tool of choice. For this demonstration to be successful we would have to be able to fulfill the process of the previous business model. We achieved just that and even more efficiently. With our data warehouse modeling we were able to introduce even more analysis that could have not been done before. According to DSRM, we produced an artifact to solve our research problem which materialized in the form of a system. To complement the demonstration we also produced reports and visuals for the data needs of the organization. These were also used in the official report published by the organization.

In terms of the second demonstration, we set to solve the same research problem, using slightly different techniques and converging on a pure data science technique - forecasting. The data went through all the steps in the pipeline that was entirely manipulated using python script, guaranteeing the automation of the whole process. Furthermore, after passing the data through all the ETL subsystems we used machine learning to make predictions for a few of the attributes of the dataset. In the process of developing this demonstration, we also found that it could be interesting to have the cumulative values (which is the normal form for the dataset) but also the non-cumulative values to see the real value of an attribute in a month. This could maybe allow us to see patterns/friends that are implicit to the data.

For the forecasts we used two different methods: ARIMA and SARIMA, which are, respectively, Autoregressive Integrated Moving Average and Seasonal Autoregressive Integrated Moving Average. These were chosen according to their popularity in the field. We tested different values to try to find which would be the best parameters for these two techniques (by trying to minimize AIC and BIC).

The results for this experience were measured using two metrics that are commonly used for this type of work, these are MAE and WAPE. These metrics yielded satisfactory results, although, one could argue these are not very reliable due to the interpretation of the metrics and the limited time series we had access to in our dataset.

Although the results of this particular experience were not the most encouraging (despite having some forecasts that appear to be very close to the observed values as shown in Chapter 5) the results of our artifacts were very promising indeed. With our work, we have proven that there can be a quick and easy implementation of our system to solve a problem many organizations face. With the added benefit of time and also cost-efficiency.

One takeaway here is that to ensure that a system is viable to make fair predictions we need a broader time series. This would allow to better train a model and possibly obtain better results.

This was a symbiotic experience since we needed to test our proposal somewhere and the organization also had intended to start delving into new technologies that could have a bigger impact on their business.

7.3 Limitations

The limitations to the present work are that a sound framework or architecture such as this one should have many iterations and evolve through many trial stages. Usually, when an organization initiates their digital transformation process there is a lot of planning and possibly an entire team of professionals from different areas involved. The novelty of our work is on the accessibility and the creation of a system that is easy to access, a practical approach to a change desired by many companies. Naturally, this Master Thesis had its limitations, specially taking into account that is was developed over the span of 10 months. Had we had more time, we could've developed more, gone through more iterations and tested on other datasets, etc. Yet, as an academic work, we believe to have developed a proposal capable of demonstrating its use and even to show the full potential that such a system can have.

There are also limitations pertaining to the datasets used. The dataset regarding the Portuguese NHS (SNS) is of a decent size and detail although it does not span over the course of many years. The information available to us is only from 2013 to 2019, one of the goals of this work was also to automate, even if partially, the process of data validation.

From the data science standpoint the results presented may not be very reliable since we used a very short time series of only 9 years. Results can suffer some bias and even incur in abnormal results as there are too many unknown variables at stake. Even more so when speaking of the nature of financial variables. As an example, these can be impacted due to financial crises, cuts in funding, changes in governments, changes in foreign policies, etc.

Although we understand the limitations that this data may have due to the very limited temporal instances, we believe, along with the first demonstration, this is a very important work to help organizations in the process of digital transformation and to prepare the data to be eventually used, more reliably, for data mining.

One very important conclusions was, in the adoption of these new technologies, organizations need to learn and understand the importance of data quality and the steps to ensure it by going through a few of the core fields of this master thesis (data governance, data mining, and data science).

We hope that with our system we can not only provide a working and adaptable artifact but also to shine a brighter light on the problem of “spreadsheet hell”.

7.4 Future Work

Following this works limitation's, we believe there is still an opening for some future work, even though we consider the goals were met and we even had a chance to use other new technologies as proof of concept. As stated before we could have continued this work and explore even more the technologies used for this Master's Thesis.

In the warehouse field, we could have designed more models, implementing different techniques or schemas and even used different datasets to test our proposal. We also would have liked to explore further data science techniques since we only covered two popular techniques of supervised learning (ARIMA and SARIMA) there was still a chance to continue exploring different variations of these methods.

The next step after these techniques would have been to test our proposal with yet another very popular, and promising, technique, neural networks. Particularly LSTMs (Long Short Term Memory), since these have also shown promising results in the field of forecasting which was of particular interest for the organization in which our work was developed. Although anticipating the limitations of this approach we know that neural networks need a lot of data to be trained which was not the case for either of our two demonstrations.

On another different line of work, it would also have been interesting to study other aspects of the work for this Master Thesis such as the approach of digital transformation less from the point of view of data science and the technologies involved but more from an IS perspective. This would be closer to the management component and more focused on the people/employees (such as studying the best way to tackle the issue) it could also have been interesting to test our proposal with other types of data or even BI tools, in other words, study the management of change. The business processes involved in the transition and the post business processes would also have been relevant to include in a Future Work.

Bibliography

[1] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee, "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007.

[2] T. Beale, "How to create a Data Inventory," Open Data Institute, Tech. Rep., 2018. [Online]. Available: https://www.researchgate.net/publication/327631764

[3] A. Andal-Ancion, P. A. Cartwright, and G. S. Yip, "The digital transformation of traditional businesses," *MIT Sloan Management Review*, vol. 44, no. 4, pp. 34–41, 2003.

[4] I. Sebastian, J. Ross, C. Beath, M. Mocker, K. Moloney, and N. Fonstad, "How big old companies navigate digital transformation," *MIS Quarterly Executive*, 2017.

[5] J. Reis, M. Amorim, N. Melão, and P. Matos, "Digital transformation: a literature review and guidelines for future research," in *World conference on information systems and technologies*, 2018, pp. 411–421.

[6] G. Vial, "Understanding digital transformation: A review and a research agenda," pp. 118–144, 6 2019.

[7] D. L. Rogers, "The Digital Transformation Playbook: Rethink Your Business for the Digital Age," Columbia Business School Publishing, Tech. Rep., 2016.

[8] B. Tabrizi, E. Lam, K. Girard, and V. Irvin, "Digital transformation is not about technology," *Harvard Business Review*, vol. 13, 2019.

[9] V. A. R Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS quarterly*, vol. 28, no. 1, pp. 75–105, 2004.

[10] D. M. Kroenke and D. J. Auer, *Database processing : fundamentals, design, and implementation.* Pearson, 2012.

[11] Y. B. Senichenkov, I. C. on Mathematical Models, R. Methods in Applied Sciences (2014 : Saint Petersburg, I. C. on Economics, and R. Applied Statistics (2014 : Saint Petersburg, "Recent advances in mathematical methods in applied sciences : Proceedings of the 2014 International Conference on Mathematical Models and Methods in Applied Sciences (MMAS '14) : Proceedings of the 2014 International Conference on Economics and Applied St," in *Proceedings of the 2014 International Conference on Mathematical Models and Methods in Applied Sciences (MMAS '14)*, 2014, p. 438.

[12] O. Y. Iliashenko and S. V. Shirokova, "Application of Database Technology to Improve the Efficiency of Inventory Management for Small Businesses," *WSEAS Transactions on Business and Economics*, vol. 11(1), pp. 810–818, 2014. [Online]. Available: http://www.isem-fem.spb.ru

[13] D. Z. Meyer and L. M. Avery, "Excel as a qualitative data analysis tool," *Field Methods*, vol. 21, no. 1, pp. 91–112, 2009.

[14] R. Abraham and M. Burnett, "Spreadsheet Programming," 2006.

[15] G. J. Croll, "Spreadsheets and the Financial Collapse," *arXiv preprint arXiv:0908.4420*, 2009. [Online]. Available: www.eusprig.org

[16] V. V. Acharya and M. Richardson, "Causes of the financial crisis," *Critical Review*, vol. 21, no. 2-3, pp. 195–210, 6 2009.

[17] European Spreadsheet Risk Interest Group, "Spreadsheet mistakes - news stories," 2020. [Online]. Available: http://www.eusprig.org/horror-stories.htm

[18] V. Lemieux, "Archiving: The Overlooked Spreadsheet Risk," *arXiv preprint arXiv:0803.3231*, 2008.

[19] S. Murphy, "Comparison of Spreadsheets with other development tools (limitations, solutions, workarounds and alternatives)," *European Spreadhseet Risk Interest Group*, 2005. [Online]. Available: http://arxiv.org/abs/0801.3853

[20] D. Durfee, "Spreadsheet Hell?" *Cfo-It*, vol. 20, no. 8, pp. 30–35, 2004.

[21] J. Freiheit, R. Görner, J. Becker, and F. Fuchs-Kittowski, "Collaborative Environmental Data Management Framework for Microsoft Excel," *Proceedings of the 28th EnviroInfo 2014 Conference, Oldenburg, Germany*, 2014.

[22] J. Ladley, *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program*, 2nd ed., Elsevier, Ed. Elsevier, 2012.

[23] V. Khatri and C. V. Brown, "Designing data governance," *Communications of the ACM*, vol. 53, no. 1, pp. 148–152, 1 2010.

[24] C. Hayashi, *What is Data Science ? Fundamental Concepts and a Heuristic Example*. Springer, 1998.

[25] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, p. 37, 1996.

[26] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[27] N. Yau, "Rise of the Data Scientist.(2009)," *URL http://flowingdata. com/2009/06/04/rise-of-the-data-scientist*, 2009.

[28] P. Naur, *Concise survey of computer methods*. Petrocelli Books, 1974.

[29] ——, "Software engineering-report on a conference sponsored by the NATO Science Committee Garimisch, Germany," *http://homepages. cs. ncl. ac. uk/brian. randell/NATO/nato1968. PDF*, 1968.

[30] IASC, "IASC (International Association for Statistical Computing)," 1977. [Online]. Available: http://iasc-isi.org/mission/

[31] L. Cao, "Data science: A comprehensive overview," *ACM Computing Surveys*, vol. 50, no. 3, 2017.

[32] J. Wu, "Statistics = Data Science?" 1997. [Online]. Available: http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf

[33] W. S. Cleveland, "Data science: an action plan for expanding the technical areas of the field of statistics," *International statistical review*, vol. 69, no. 1, pp. 21–26, 2001.

[34] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[35] Z. Voulgaris and Y. E. Bulut, *AI for data science: Artificial intelligence frameworks and functionality for deep learning, optimization, and beyond*. Technics Publications, LLC, 2018.

[36] S. Chakrabarti, M. Ester, U. Fayyad, J. Gehrke, J. Han, S. Morishita, G. Piatetsky-Shapiro, and W. Wang, "Data mining curriculum," *Intensive Working Group of ACM SIGKDD Curriculum Committee*, vol. 140, pp. 1–10, 2006.

[37] R. Kimball and M. Ross, *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.

[38] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge discovery in databases: An overview," *AI magazine*, vol. 13, no. 3, p. 57, 1992.

[39] N. Dedić and C. Stanier, "An evaluation of the challenges of multilingualism in data warehouse development," *Proceedings of the 18th International Conference on Enterprise Information Systems*, pp. 196–206, 2016.

[40] C. A. Kulikowski and S. M. Weiss, “Computer Systems That Learn-Classification and Prediction Methods from Statistics, Neural Nets,” *Machine Learning, and Expert Systems. Morgan Kaufmann, San Mateo, CA*, 1991.

[41] D. J. Hand, “Discrimination and classification,” *dicl*, 1981.

[42] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[43] D. J. Berndt and J. Clifford, “Finding patterns in time series: a dynamic programming approach,” in *Advances in knowledge discovery and data mining*, 1996, pp. 229–248.

[44] P. C. Cheeseman, “On the importance of evidence: a response to Halpern,” *Computational Intelligence*, vol. 6, no. 3, pp. 188–192, 1990.

[45] B. V. Dasarathy, “Nearest neighbor (NN) norms: NN pattern classification techniques,” *IEEE Computer Society Tutorial*, 1991.

[46] J. Whittaker, *Graphical models in applied multivariate statistics*. Wiley Publishing, 2009.

[47] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[48] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

[49] R. W. Tresch, *Public finance: A normative theory*. Academic Press, 2014.

[50] C. Shoup, *Public finance*. Routledge, 2017.

[51] H. S. Rosen, “Public finance,” in *The encyclopedia of public choice*. Springer, 2004, pp. 252–262.

[52] S. Gupta, M. Keen, and A. Shah, *Digital Revolutions in Public Finance*. International Monetary Fund Washington, DC, 2017.

[53] T. H. Davenport and D. J. Patil, “Data scientist,” *Harvard business review*, vol. 90, no. 5, pp. 70–76, 2012.

[54] L. Cao, “Data Science: Profession and Education.” *IEEE Intell. Syst.*, vol. 34, no. 5, pp. 35–44, 2019.

[55] V. Dhar, “Data Science and Prediction,” *Commun. ACM*, vol. 56, no. 12, p. 64–73, 12 2013. [Online]. Available: https://doi.org/10.1145/2500499

[56] R. Kohavi, N. J. Rothleder, and E. Simoudis, “Emerging trends in business analytics,” *Communications of the ACM*, vol. 45, no. 8, pp. 45–48, 2002.

[57] C. Matt, T. Hess, and A. Benlian, “Digital Transformation Strategies,” *Business and Information Systems Engineering*, vol. 57, no. 5, pp. 339–343, 10 2015.

[58] W. McKinney, *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. ” O’Reilly Media, Inc.”, 2012.

[59] J. Pries-Heje, R. Baskerville, and J. R. Venable, “STRATEGIES FOR DESIGN SCIENCE RESEARCH EVALUATION,” *ECIS 2008 Proceedings. 87*, 2008.

[60] N. Prat and J. Akoka, “ARTIFACT EVALUATION IN INFORMATION SYSTEMS DESIGN-SCIENCE RESEARCH-A HOLISTIC VIEW,” *PACIS*, pp. 23–undefined, 2014.