NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

REAL ESTATE: DEVELOPING A DATA MODEL FOR THE CITY OF LISBON TO IMPROVE INFORMATION SHARING

André Sevinate Sousa Rendas Pereira

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management with specialization in Knowledge Management and Business Intelligence

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

REAL ESTATE: DEVELOPING A DATA MODEL FOR THE CITY OF LISBON TO IMPROVE INFORMATION SHARING

by

André Sevinate Sousa Rendas Pereira

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management with specialization in Knowledge Management and Business Intelligence

Advisor*:* Professor Miguel de Castro Simões Ferreira Neto

July 2021

ACKNOWLEDGMENTS

I want to express my sincere gratitude to everyone that in some way has contributed to this project.

To my advisor, Professor Miguel de Castro Neto. This project would certainly not be possible without his vision, his guidance but also his constructive feedback.

To my family, parents and brother, for being always there for me. Thank you for your patience and understanding which turned this project a lot easier.

To Sabrina Oliveira and her family for not giving up on me and providing all the motivation that I needed. Their support and care kept me going in the most difficult times.

To my NOVA IMS colleagues, Gabriela Costa, Ines Vilas Boas and Miguel Alves dos Reis that provided both support and help necessary to conclude this project.

To all my friends, specially Diogo Silva, André Gomes, Sara Silva and João Faria for being very helpful during the development of this work. Their inputs were key to the development of this project.

To my work colleagues, Rita Borges, Laura Squarzon and Paraschiva Farcas-Ramos. Their understanding and companionship were very important during all times.

ABSTRACT

The digital and technological transformation throughout the 21st century proved to have interference in the overall functioning and modus operandi of areas such as the financial sector. The economic importance of real estate allied to the Information and Communication Technologies transformation may also alter the perception and the way the market faces and interacts with this sector. This work will try to bridge the real estate market to the world of information and communication technologies taking advantage of these to generate knowledge on this market. While there are many possibilities available to explore these two concepts, this project will use the available literature to gather insights on both and use them to develop a data model. This data model will be applied on the Portuguese real estate market, more specifically in the city of Lisbon and the districts of Oeiras, Amadora, Odivelas and Loures. It is aimed to understand in what way could the technological and communicational technologies that are nowadays accessible, bring benefits to its users in terms of speed, reliability, readability and convenience when consulting real estate information. It is expected that the search intensity can be increased while lowering its costs (time, opportunity) while also reducing information asymmetry. It will also be interesting to understand if the implementation of technological tools can act as a substitute or if they will be used as a complement of traditional methods in the real estate sector. This project is going to be applied to an actual real estate sector context.

KEYWORDS

Real Estate; *Proptech*; Technology; Economy; Information Asymmetry.

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AR	Augmented Reality
BI	Business Intelligence
BK	Business Keys
DSR	Design Science Research
DW	Data Warehouse
ETL	Extraction Transform and Load
FK	Foreign Key
FSBO	For Sale by Owner
GDP	Gross Domestic Product
ICT	Information and Communication Technology
IOT	Internet of Things
NAR	National Association of Realtors
SA	Staging Area
Saas	Software as a Service
SCD	Slowly Changing Dimension
SK	Surrogate Key
TAM	Technology Adoption Model
VR	Virtual Reality

1. INTRODUCTION

Real Estate Market has always been a key area in the economics of a country. This project will focus on the residential Real Estate Market and will benefit from the use of the existent Information and Communication Technology (ICT) tools to attempt the creation of an information sharing and visualization model that can provide advantage and optimized ways to analyze data while reducing information asymmetry.

The main goal is to develop a full stack process from the gathering of data to be used, to the implementation of a Business Intelligence (BI) model to monitor and create knowledge on the Portuguese real estate market. The final delivery will be a report that is expected to be relatively simple to use, but also a complete report on relevant real estate measures and indicators. It is anticipated that it also delivers an interactive way to surf trough the data according to the best of the spare literature found throughout the entire research.

2. MOTIVATION AND PROBLEM FRAMMING

The disruptive technologies are hereby described as the big 9 by F. Ullah et al., 2018. Technologies as Software as a Service (Saas), IOT and Cloud Technology, are within that group and are nowadays present in everywhere in our quotidian and in real estate as well.

As any other market, housing market also has its stakeholders. The four key players are: consumers, agents, associations and government and regulatory authorities. The interactions between these counterparties are, according to F. Ullah and his team, through dissemination mechanisms that include smartphone technology, websites and social media platforms. These mechanisms act on behalf of their needs while buying or selling properties, profit or tax-based issues and business factors.

On the consumer view of the real estate sector, there are information deficits that lead to market flaws and consequently may end with consumers regretting their post purchase or post rental decisions according to F. Ullah et al. in 2018.

On 2009 MK. Kolic and I.Vida suggested that when purchasing a house, consumers first define their needs and goals. Then, as they describe, consumers spend time gathering as much information of the possible different alternatives within the criteria fitting that which was defined in first place. According to Vida and Kolic, consumers tend to search for external information to increase their knowledge and reduce perceptions of risk. This external information can go from, *word-of-mouth*, to any other type of information that consumers find relevant in a pre-purchase our pre-sell situation. This information is obviously also obtained through the dissemination mechanisms highlighted before. Curiously, MK Kolic's and I.Vida´s study points to that potential buyers should consider less the skills and behaviour of the seller and base their decision on the direct comparison of the alternatives found, if they match the criteria defined at the start.

Real estate brokers, regardless of the consumer´s needs or goals, are still engaged nowadays. According to Zumpano et al. (1996), when hiring a real estate broker service, consumers are indeed buying a bundled good. They perceive that the cost of opportunity and information is high and that represents an important role on their decision. A real estate broker allows to increase the search intensity and lower these type of search costs. Interestingly, the help of professional real estate agents is less requested by inexperienced buyers, as his studies demonstrated. In a more recent study and according to NAR (2019), first time buyers constitute 33% of all home buyers on the market. This data also showed that 44% of the buyers took the first step of looking for properties online, while only 16% first contacted a real estate agent. Also 93% of the buyers felt that online information was useful. This study also indicates that the traditional buyer, the one who did not use internet for house consuming

needs, spent less time searching compared to the more sophisticated consumers. The consumers who used internet on the other hand also visited twice more houses and were not only subject to the properties listed and presented by that broker. In other words, they had a more diversified choice and more alternatives for their needs and goals. FSBO´s, for sale by owners, only represented 8% of recent home sales and typically sold for less than houses listed in real estate agents but sold more quickly.

As real estate agents are still very requested, according to CB-Insights (2008) global real estate industry is lagging the technology curve by five years, as data is normally managed on spreadsheets and innovative information technology is barely used.

However, the use of the computer processing power and internet to perform activities related to the housing market has been increasing throughout the years.

Proptech, as described by Baum in 2017, were the first signs of computed driven property research. The first sighting of Proptech (Properties + Technology) related movements took place in the US and UK in the mid-1980s, which took use of computing raw power for processing data. The invention of personal computers, knowing that computing existed since the 1940s, was what finally triggered a change on the properties market. Baum (2017) also describes that the growing of data availability and the globalization contributed for the need of new ways to manage properties and properties portfolio systems which became computer and technology based. In parallel, internet was becoming a thing in the late 1990s which also seized its place in what is known today of the online real estate market.

This project will focus Information and Communications Technology (ICT), on residential real estate, as it is a more homogenous asset type (Baum 2017), with more public information available like prices, rents and general details but also because of the importance and impact that the real estate sector has on people's lives.

As real estate is an information driven sector, searching, accessing and effectively managing information on housing can be a challenge because the housing alternatives are spread across several real estate agencies. In addition, unless the real estate agencies are willing to work towards that specific goal, some important information may not be directly given nor displayed to consumers. In most cases, simple but important metrics such as average price by area are even not possible to get from their websites for a more macro analysis perspective. This lack of information and difficulty in managing this type of information on the consumer side, brings inefficiencies and market flaws leading to consumers post regretting their decisions as concluded by F. Ullah in 2018.

3. IMPORTANCE AND RELEVANCE

Even though real estate is a much conservative sector, it is hard to believe that it will remain the same in the future. For Baum (2017), real estate is a huge sector and one of the last industries to adopt technological change and innovation, meaning that there is a big opportunity driven by the size of the industry and the lack of technology within.

Proptech is by any means no longer new in some countries and there is always room to new development, given that this industry just scratched the surface regarding tech, according to Baum (2017)

This project intends to show that technology can give access to information in an easier way, but also create more knowledge with the information that is available on the internet. This work can also be important to framework the current status of the Portuguese real estate market, given that it will work with data directly related with houses that are currently for sale in the city of Lisbon which is the largest in Portugal by population.

4. MAIN GOAL AND SPECIFIC OBJECTIVES

By reading the literature that was found on the different matters subjected to this project it was possible to support and strengthen this work with theoretical background.

The main goal is to build a real estate visualization tool concept model of housing market of the city of Lisbon while trying to present an innovative visualization that can improve market awareness, search efficiency and readability while making it also interactive.

To achieve this specific goal, it is needed to be define:

- Relevant and specific housing information to be measured.
- Calculated metrics.
- Selection of real estate agencies population to be included on the problem.
- Retrieval and data management practices.

Consequently, the following questions to be researched are:

i. Can Business Intelligence be used to create an interactive, convenient and user-friendly model to the city of Lisbon?
ii. How will real estate awareness and search efficiency be improved on real estate market using Business Intelligence to design a model?
iii. Is the model going to act as a support tool or can substitute entirely the role of a real estate agent?

5. METODOLOGY

The above specific goals will be tried and tested with resource to the Design Science Research (DSR) methodology. As Hevner describes in 2007, a good design science research is motivated by the desire to improve the environment by introduce new and innovative artifacts as well as their building process. This process often begins by exposing and representing opportunities and problems in an actual application environment to give context and providing requirements for the research, which in this case is being done by revisiting literature on the subjects.

Upon the definition of a problem, framing research questions with the given context, comes the design cycle in which, in an activity framework for DSR is described by Venable in 2006 as the enhancement of or creation of a method, product, system, practice, or technique.

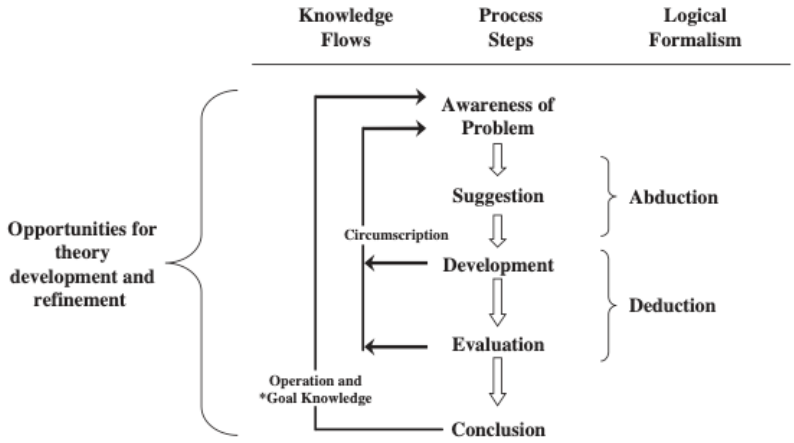


Figure 1 - DSR Process Model (DSR Cycle) (V. Vaishnavi e Kuechler, 2008)

When analyzing the work of Vaishnavi e Kuechler (figure1), they propose five steps method for describing DSR. This method starts with an awareness stage, equivalent to the environmental phase proposed by Hevner. This is followed by a cycle of three levels:

Suggestion: Incorporating a presentation of the tentative design or approach of the addressing of the problem considering the best practices given by the literature review.

Development: Implementation of the suggested tentative design. Definition of the software to anchor the designed solution and measurement practices to analyze the output results.

Evaluation: Measured results presentation and assessment. Subject the artifact to an extended evaluation to analyze the impact of its development on the problem at hands. The evaluation procedure may support or be disconfirmed by the theory. A revisit to the literature available may bring advancements or more awareness to lead to possible adjustments.

Conclusion: It is the final step. It is at this stage that the researcher makes the presentation of the results of his model. The artifact may or may not fill the requirements for the solution.

6. LITERATURE REVIEW

6.1. REAL ESTATE

Real estate is a very big market, if not one of the largest markets in the world but it is also far behind other industries regarding the adoption of technology and all the changes that come with it (Baum 2017).

According to Savills World Research at the end of 2017, the value of world´s real estate reached 280.6 trillion USD. Residential real estate accounted for the largest share, approximately 79% of the total value, followed by commercial real estate with almost 12%. Agricultural and forestry real estate make up for the rest. This research also shows that real estate is the most significant store of wealth, representing globally 3.5 times the total global GDP.

In a research published by KPMG (2019) on bank lending for the real estate sector in Europe, residential real estate is highlighted as the largest asset type of investment for 2019 as figure 2 confirms:

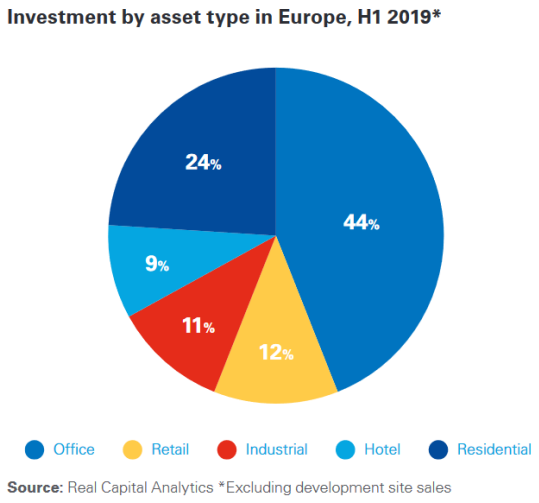


Figure 2 - Investment by asset type in Europe, H1 2019. Source: Real Capital Analytics "Excluding development site sales"

Not only residential real estate bank loaning is the largest, but also representing almost half of the bank lending business of this type.

Real estate, technologically speaking, is categorized by Baum in 2017 as a clunky market. He points various reasons that could be behind his theory that he acknowledges as the agency problem: "The professional advisors that dominate the transaction process clearly have an interest in protecting their income sources, so chartered surveyors, brokers and lawyers might all be expected to resist tech-driven innovations designed to 'disrupt' their work."

Real estate was always considered of extreme importance regarding political and economic issues and several debates on housing have been occurring through Europe regarding it. The strategic habitational policies are nowadays related with cities self-laws (Guerra, 2011). Housing is the main asset of households. Any fluctuation in its value may directly impact its consumption as well as the owners (Lourenço e Rodrigues, 2017). The differences between supply and demand, concerning globalization and sociodemographic transformations, the regulatory premise of the government in housing, are well known and interrelated. (Guerra, 2011). The real estate sector is highly heterogeneous. After Portugal has entered the European Community (1986), there has been throughout the years a development in real estate, transforming itself into one of the most prosper and influential of the national economy. (Guerra, 2011). The demand for houses fluctuates within several factors such as age, gender, habilitations and income. (Tavares e Pacheco, 2015). Regarding supply, there is a correlation with the availability of construction land, its cost and the cost of the very construction (Lourenço e Rodrigues, 2017). The preferential location of properties is related with the building type, the use of communal areas and with the positive or negative externalities that may grant or take pleasure to the residents (Tavares e Pacheco 2015). Negative externalities are perceived as urban constructions with negative impact, such as landfills or chemical refineries (Farber, 1998). On the other hand, positive externalities are seen as the quality of the green areas or the panoramic view and its surrounding neighborhoods (Bourassa, et al., 2003). Equally, the proximity to health services, transportation, schools and commercial areas, contribute to grow the value and the demand of a property (Seixas, et al., 2019). Lisbon and Porto, evolved rapidly in improving the number of tourists, becoming attractive to fresh startups and young professionals. Gold Visas have been and still are, a way in for the foreign investment, bringing more people with higher purchasing capacity, and consequently contributing to the raise of the general price of properties (Lourenço e Rodrigues, 2017). Information asymmetry impacts deeply on property businesses because the illiquidity and the waiting time on this market during the full length of a sale has costs. The fair value and the professional ethics that may benefit ones in detriment of others, can also be impacted by that asymmetry.

6.2. CONSUMER REAL ESTATE AND ICT

As post-purchase and post-rental regrets are increasing, ascending to 44%, according to F. Ullah et al. this is due to the lack of information about properties and to the complexity of the purchase process. Such regrets can be eliminated or reduced by dragging disruptive technologies to this sector furnishing consumers with sufficient and detailed information before they make any real decision.

Real estate industry has always been an “information driven business” with high transaction costs and high inefficiency because of such heterogenous and infrequent trading (Smullyan, 1994).

Property management systems continue to see information technology as competitive advantage in the housing industry, because firms that have successfully implemented it have achieved it (M. Kummerow et al., 2005). Amongst the ICT applications in real estate industry, the world wide web, commonly known as the internet has become a major advertising vitrine for leasing and home sales. Real time listings, virtual home tours and other features help the consumers to become better informed achieving greater awareness of the market regardless of their goals and needs. According to Kummerow et al., real estate agents were not replaced by websites but have been increasing their productivity and quality of services, on Zumpano´s perspective in 1996, they have been improving their bundled goods.

ICT speeds up and lowers search costs in the decision process particularly in the early stage of the awareness process. Not only that, but ICT also revolutionizes travel and tourism marketing costs, as local listings of bed and breakfast accommodation are now detailed with different layers of easily and available information.

As real estate is heavily turned to information, Kummerow et al. say that ICT can easily find different applications and can affect the overall quality and productivity of the sector services and the costs of the services itself. Also, he implies that the low-cost databases, wireless communications and other technologies such as the ability to represent such information in pictures and maps, are the new possibilities of ICT. For him it all has to do with the experience given to the end consumer and this experience must be apart from the raw use of text and numbers. Lai in 2003 already defined some of the benefits from the user point of view to be speed, reliability, readability and reachability, security, time saving and convenience in the form of instant connectivity, personalization and location. In 2000 Freedman said that consumers usually expected agents to locate the house, negotiate the price and take care of the bureaucracy for the financing and legal issues, in other words, they expect to use the expertise and local market knowledge acquired by the agents from the years of intermediation. More than for searching houses, consumers look for social capital and negotiation skills of the real estate agents. It is sure that a property is indeed a very complex product and cannot lent itself to pure e-commerce (Kummerow et al., 2005), however ICT and internet can help real estate to potentiate its agents in doing their job better, faster and easier.

John S. Baen and Randall S. Guttery understood that the rapid growth of both consumer and real estate provider business computing, with the growing availability of market information through the birth of new real estate databases would heavily affect the industry. However, gathering information in a timely manner is costly due to the decentralized nature of the proprietary nature of the data, as Y. Fu and Lilian K. describe in 2001. As real estate contains a wide set of data, F. Ullah says that agents and

professionals should focus their roles on social networking and market expertise and leave the analysis to technology. These analytics will only be achieved through data centricity which will place the relevant data at the core of the decision creating wiser consumer decisions. Some websites such Zillow.com and Domain.com provide core residential insights for consumers but also express some information regarding neighborhood, crime and average property sales or travel rates. For Kummerow et al., user-friendly supports on the real estate market sector see a clear opportunity on merging separate databases in order to achieve better market efficiency in detriment of a decrease transaction cost. However, for them, firms that do not develop a technology adoption model (TAM) can suffer and be pushed aside from the market itself. Competitive advantage is to use it to present to consumers an overall better service. F. Ullah et al., mentioned that sites such as Facebook, WhatsApp and Instagram can provide help in disseminate information in general but can also be used to gather other types of information like crime rates, walking scores and traffic scores which may be of interest to real estate intervenient. F. Ullah and his team also found that half of real estate websites also needed improvement. He claims that the lack development of interactivity and reachability of websites presents itself as business opportunities but also reveal the absence of innovation of the sector. To conclude their research, they also claimed that web-based programs, extensions and smartphone apps should be provided to consumers to fill their voided needs of information, equipping them with real time analytics and updates on the localities of their intended properties, adapting information to consumers needs and goals.

6.3. INFORMATION ASYMMETRY ON THE HOUSING MARKET

As suggested by Tavares et. al., in 2013, information asymmetry is of great importance on real estate industry. This asymmetry is later divided into four segments: information asymmetry and adverse selection, information asymmetry and predictability of returns, price distortion in housing market, and information asymmetry and real estate depreciation. Wong et., al in 2011 refer to information asymmetry as a kind of friction that deters buyers and sellers from making mutually beneficial trades in a free market. In 1970 Akerlof, suggested that information asymmetry is spread across the second-hand markets since sellers who have owned or used their goods for a time period are better informed about the state of their product than buyers. This sector is also dependable of subjective factors as buyers and sellers also make value judgements when adopting an intervening position on this sector. The counterparty which has the most information, will have the upper hand and will use it to its benefit.

Moreover, Tavares et al., (2013) state that duly informed sellers tend to purposely retract information from its customers. It is stated that the longer that a property is for sale, the higher the costs for the

selling agent are. This means that sometimes, the buyer will get only the information that the seller wants to tell, not the information he sought because there is a sense of urgency in selling the house. Other big factor of asymmetry of information is the proximity of the buyer to the specific property. There will always be information asymmetry between a local buyer and an outsider buyer, as the proximity factor will allow him to have more detailed and accessible information. According to Tavares et. Al (2013), it would be more expensive for an outsider to buy a house when competing with a local buyer.

Below are some of the flaws in information asymmetry in the Portuguese real estate market in terms of price distortion and according to Levitt and Syverson (2004):

- Professional real estate agents have better access to historic information of transaction prices and market performance.
- Professional real estate agents sell their properties 3,7% higher in price than their clients. Also, the amount of time that their houses spend on the market is generally higher 10 days in average.

And according to Tavares (2011):

- Real estate market is characterized by illiquidity and property idiosyncrasy is difficult to evaluate by outsiders.

Adverse selection can happen when a price for a good or a service with the lowest quality exceeds the market price for the respective good or service with the highest quality. In other words, there is a wrong of perception of the true characteristics of a specific good or service (Schettino, 2006).

For adverse selection, the following were highlighted:

- Adverse selection and information asymmetry problems can result in moral hazards, because of a wrong perception by a party with lower access to information (Stiglitz, 1975).
- It is difficult to obtain and evaluate positive and negative externalities regarding the selection of a house. (Tavares, Moreira, Pereira, 2010).

The perception of the market and of its underlying objects might differ from two different entities. There is also moral hazard risk resulting of information shortage. As Tavares et al., concluded in 2013 on a case study involving over than 330 commercial delegates of the three largest Portuguese real estate companies, more and more individuals depend on specialists to take decisions on real estate, as they have valuable information that could reduce market distortions. Also, the study claims that

real estate agents try to convey information that are convenient to them, meaning that they are more predisposed to show the positive externalities to a buyer to sell rapidly, and do not concern what is the worth that the buyer attributes to the negative externalities. It depends on the very individual himself to find ways to limit information asymmetry.

6.4. TECHNOLOGY AND PROPERTIES (PROPTECH)

Nowadays individuals usually dependent on technology and uses it in all kinds of contexts (at work, at home, within education purposes, in health, in finance and for entertainment). Technology has come to ease the access to information by disposing it in real time, allowing transactions through online channels and enabling remote control activities (Baum, 2017).

Since 2010, over 50 billion USD were spent in more than 2500 companies. These companies are revolutionizing the way to store, spare, lend, invest, move, spend and protect money. Fintech (Finance + Technology) is related to the innovation presented on the finance sector due to the creative evolution of processes. (Accenture, 2016). The online transactions, crowdfunding events, online funding platforms and stock exchange market are good examples. In that sense, Fintech is offering several solutions to enterprises of the real estate market, enabling the trading of property assets through web platforms (Feth and Grüneberg, 2018).

Proptech (Property+Technology), is a vast concept. Some of the main aspects of Proptech are Smart real estate, Real Estate Fintech and Shared Economy, (Figure 3). Broadly, smart real estate is the use of technology to raise the efficiency of the resources and space, such as heating and sustainable use of water and power. As described by Feth and Grüneberg in 2018, smart real estate is all about software-based platforms that facilitate the running and management of real estate properties. This information can be tracked by the owner in real time through a computer, smartphone or a tablet (Baum, 2017).

Shared Economy also uses of software-based platforms to facilitate the daily use of real estate properties. It mainly deals with solutions and logistics from the sharing of accommodation to offices, spaces and buildings (Feth and Grüneberg, 2018).

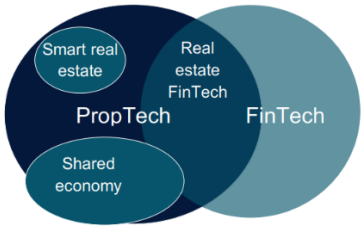


Figure 3 - Differences between Proptech and Fintech (Baum, 2017)

Real estate fintech is also a software-based platform, where buying and selling of real estate assets are optimized. It can go from the classic individual properties, to shares or debt and equity capital invested on real estate. One example of this type of platforms is the American Zillow Group that offers the possibility to consumers to obtain information on listed properties without having to contact a real estate agent.

The properties market is huge and represents more than half of the main assets globally (Savills, 2017).

Technology contributed to a greater efficiency of processes, automatization, uniformization and economies of scale (Blair, 2015). Specifically, on real estate, technologies such as Blockchain, Internet of Things (IoT), Cryptocurrency, Geolocation, Big Data and Augmented and Virtual Reality (AR and VR) are nowadays considered the most important.

For Baum (2017), proptech will eventually break through the real estate sector. While this is a huge sector and one of the last industries to adopt technological change and innovation, it means that there is a big opportunity driven by the size of the industry and the lack of technology within.

6.5. DATA VISUALIZATION

As Zheng explained in 2020, visualization is related to vision (seeing trough eyes). Visualization process may change the original form of things or create a new form for a better understanding and communication. While a report is a presentation in numbers and text of a direct detailed query data that usually involves simple analysis and data transformation (sorting, calculating, filtering, grouping, formatting), a dashboard tends to be interactive and visual but maintaining focus still on the detailed data (Zheng, 2020). Data visualization is also an important part of decision making as it is only natural to apply the rich communication techniques in the field of BI and analytics. Nevertheless, data visualization is also considered as a separate field from business intelligence, but visualization tools have become increasingly important to BI, given the visual prowess that they can supply over complex data sets.

Data visualization according to Zheng is a human physiological and psychological capability which plays an important role in behavior and decision making as it allows:

- To recall or memorize data more effectively.
- Fast perception.
- To extract additional (implicit) perspectives and meanings.
- To ease the cognitive load of information processing and exploration.
- An effective communication (story telling).

Chen et al., in 2009 propose the showcase the following figure to explain the visualization process at its core:

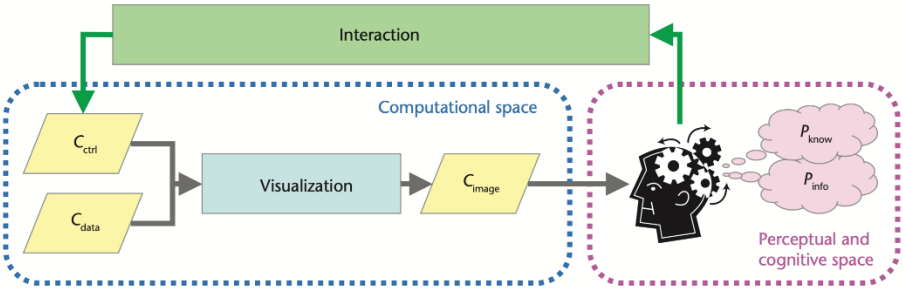


Figure 4 - Information-assisted visualization – Data, Information and Knowledge in Visualization (Chen et al., 2009)

A typical visualization process, where interaction provides the primary means for reducing the search space in visual exploration. The input data, control parameters and visualization results are stored in the computer memory, and the information and knowledge acquired by the user are created based on the interaction with the computer.

6.6. DASHBOARD DESIGN

Part of this project is focusing on the business intelligence capabilities to allow a better understanding of the real estate sector. The best and easiest way to provide a better access to information is without a doubt trough a Dashboard. Therefore, this chapter will present some guidelines to dashboard building based on the literature.

A dashboard is a visual display of the most important information needed to achieve one or more objectives, arranged on a single screen so the information can be monitored *at a glance* (Few, 2007). Dan Dubriwny and Kurt Rivards in 2004 talked about the ideal way to monitor business performance. They refer dashboards as they provide key performance indicators though simple visual graphics such gauges, charts and tables. They are appealing because they present a wide number of different metrics in a single consolidated view, roll up details into high level summaries and intuitive indicators by using color management.

When imagining a fully automated tool that can incorporate all these layers, Few, Dubriwny and Rivards´s dashboard definition easily explains why nowadays everyone wants one. However, Stephen Few in 2005 alerted for the fact that while it is true that dashboards are very powerful, they will pose a specific set of design challenges. Only those who really have a clue of what they want to build should pursuit to build a dashboard.

For Zheng (2020) the elements on a dashboard are reflected by on figure 5 on the elements of a dashboard:

Figure 5 - Elements of a dashboard – Data Visualization for Analytics and Business Intelligence – A Comprehensive View (Zheng, 2020)

In his model, data/information is the critical element, that is complemented by the visuals which provide an *at a glance* view. The User interface (UI) unifies all the elements, supporting interactions as needed. According to Zheng (2020) and Few (2004), the following points were adapted to describe the values of a Dashboard:

- can be used by non-technical users.
- compile data and enable trend visualization and implicit occurrences (high level summaries).
- provide a one place presentation of the most critical data.
- intuitive and concise display mechanisms.
- can be customized according to any division or departments.
- can integrate several different visual graphics to illustrate the data/information such as gauges, charts, tables.

In 2009 Abela (figure 6) presents a very concise and clear structure on charts.

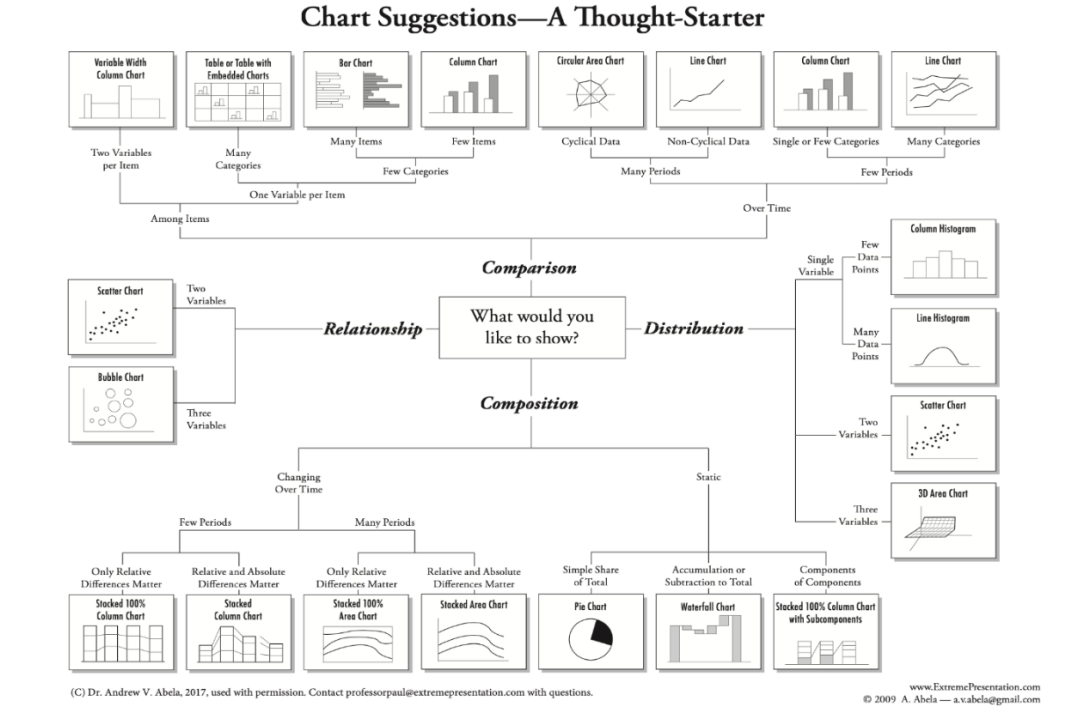


Figure 6 - Chart Suggestions - A Though-Starter (Abela, 2009)

Zheng in 2020 mentions that identifying structures and relationships that are hard to express on words are the reason that data visualization is so important when build a report of this kind. This figure contrasts very well with the goals that data visualization and report building techniques as the graphical presentation deals a great amount of perception in a more clear, concise and better way that if the same information was displayed in a text or table. More specifically, it could be difficult to explain in detailed all the intended message from the picture by only using words.

7. CONCEPTUAL MODEL PROPOSAL

This chapter will suggest a conceptual model. During this chapter, context regarding the creation of the artifact will be presented. At the end of this chapter the tentative design explanation will be complete and hopefully ready to be developed.

This model will have four main phases:

- Data gathering.
- ETL and Data Warehousing.
- Creation of a business analytics solution.

The goal is to academically prove that is possible to design a model that could facilitate the flow of information through Business Intelligence and use it to get insights for the current panorama of real estate in the designated areas.

Figure 7 presents the macro perspective surrounding the creation of the artifact. This artifact is in fact a process created from root which will enhance the chance to create a tool under a final dashboard form that can reach the proposed goals.

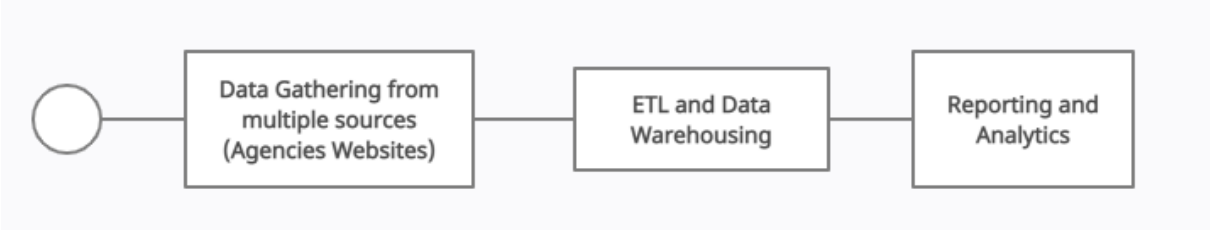


Figure 7 - Main Process (Macro Perspective)

7.1. FIRST STAGE – DATA GATHERING

This model´s main source of data will come from real estate agencies websites itself. Three websites from different agencies seem reasonable. The model needs to be robust so it can be extended to other agencies if desired. The data will be available on each agency website. To collect it, it will be created a web crawler. This is the first process which can be seen according to figure 7. The data will be gathered in several flat files for each agency. All the flat files need to have the same data structure.

7.2. SECOND STAGE – ETL AND DATA-WAREHOUSING

After the initial data is gathered from the websites, it will be sent to a Data Warehouse. Afterwards, an ETL process will run to keep the data integrity while loading it into a Data Warehouse (DW). As seen before, this market is inserted in a context of low liquidity, meaning that the information will change slowly across time. New houses may be added, others may be sold or eventually withdrawn by each website. Due to that, this first two processes will need to run in a time-based manner. The data model should express a periodic snapshot, where there is one row per as Kimball and Ross in 2002 described on Table 1. The renting market is not within the scope of this project.

Table 1 - Fact Table Comparison – The Data Warehouse Toolkit (Kimball and Ross, 2002)

CHARACTERISTIC	TRANSACTION GRAIN	PERIODIC SNAPSHOT GRAIN	ACCUMULATING SNAPSHOT GRAIN
Time period represented	Point in time	Regular, predictable intervals	Indeterminate time span, typically short-lived
Grain	One row per transaction event	One row per period	One row per life
Fact table loads	Insert	Insert	Insert and update
Fact row updates	Not revisited	Not revisited	Revisited whenever activity
Date dimension	Transaction date	End-of-period date	Multiple dates for standard milestones
Facts	Transaction activity	Performance for predefined time interval	Performance over finite lifetime

By collecting the data across these three different websites, it will not be possible to obtain all houses in the real estate market. Firstly, there are more agencies on the market and there are also some houses which are for sale by owners that will not be available to join the data model. The goal is to provide certainly a realistic overview of the market. The intent is to turn this artifact into a working concept proving that it is possible to make access to information easier and better and more importantly, different.

7.3. THIRD STAGE – DASHBOARD, BUSINESS ANALYTICS

The third phase of this project is to create a Dashboard with concise and intuitive and interactive information.

Some of the categorical and calculated measures that are the goal of this model are:

- A detailed and realistic background of information about the city of Lisbon and its housing panorama.
- House pricing and number of available houses, over a certain location, parish, neighborhood, or within a selected area radius.

- Storing of historical data, and continuity of data. It is intended to add the upcoming houses to the model and keep the information of the houses that taken from the websites.
- Updates on the existing properties that are already in the model and that received newer information.

7.4. OVERVIEW OF THE PROCESS (BPMN NOTATION)

The tentative design is ultimately the creation of a process. This chapter will explain the process using a BPMN (Business Process Modeling Notation). According to White on 2004 in “Introduction to BPMN”, the goal of this notation is to provide a clear and understandable notation to all business users. From the initial drafters to the technical developers responsible for implementing the technology and to sustain the processes, as well as the process managers who are to supervise and monitor them, all participants in the scope of a project need to able to understand the notation easily. BPMN is defined as the bridge for the gap between business process design and process implementation.

Figure 8 in page 26, expresses the BPMN process in order to provide a much more detailed and clear perspective of the proposed goal.

By analyzing the diagram, the three stages of the process reviewed on the previous chapters are presented into the different containers each with its different processes and sub processes.

This diagram intends to represent a time-based process in which there is a business flow to be respected. Each of these components will have its individual importance in order to successfully complete the process each day.

Following White’s perspective on his Process Management theory, the process after implementation, must be managed by the businesspeople, which means that monitoring and following the process rules is as important as creating it, otherwise there would be no reason to develop it in first place.

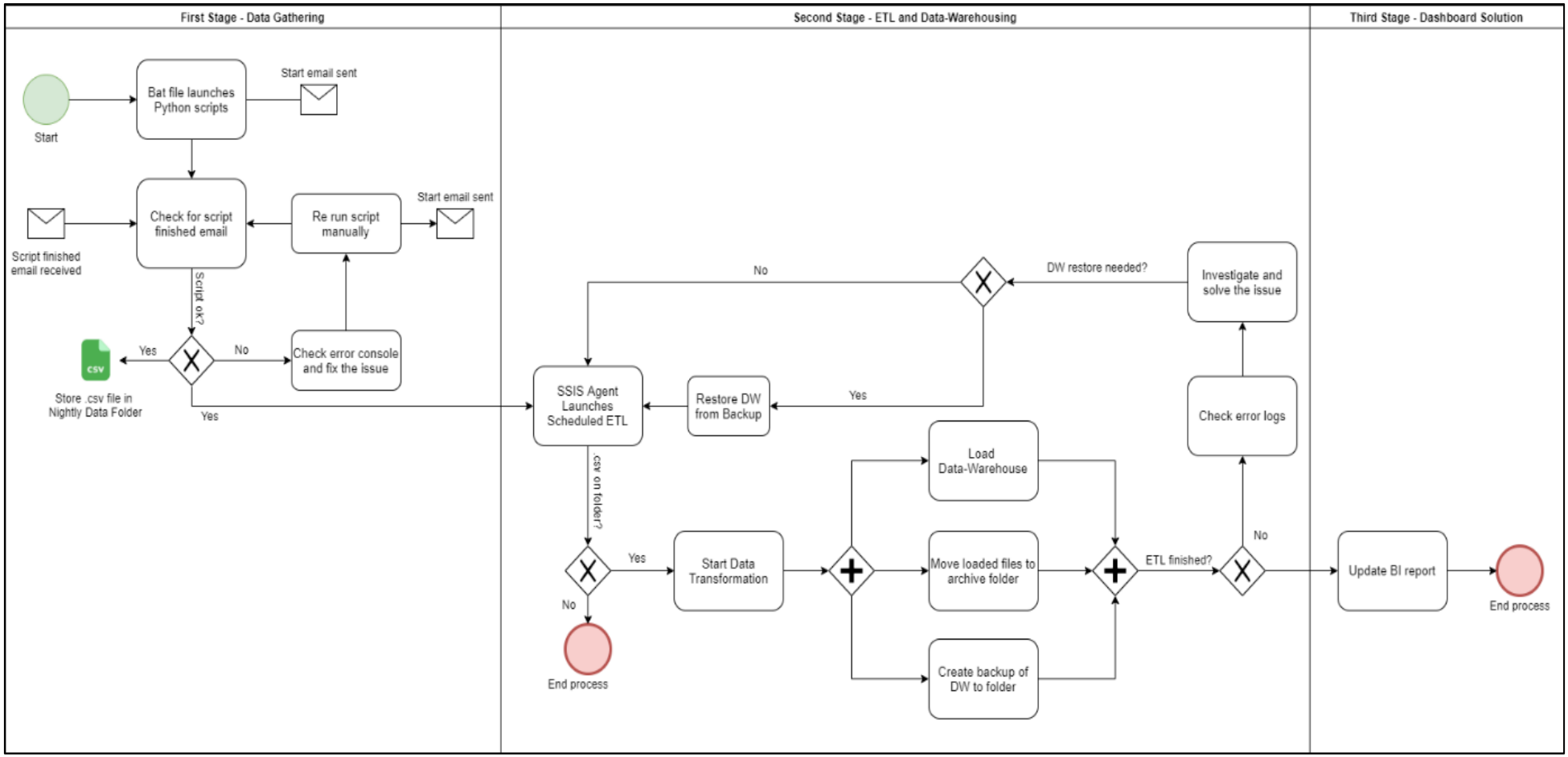


Figure 8 - Overview of the Process - BPMN Tentative Design

7.5. DESIGNED STRUCTURE

7.5.1. AVAILABLE DATA

Following the DSR approach methodology the conceptual model construction was initiated. Testing was taken to gather data from the websites. It was needed to verify which data could be extracted from the websites. If there was no way to retrieve the data from the websites, or to get it by another means, the concept would fall apart at its beginning.

The pilot test showed that it was possible to obtain the following data for the different websites:

- Listing reference (each house as its own reference)
- Sold Flag
- Type of house (apartment, dwelling)
- Listing Status (sold, not sold)
- District
- Parish
- Year of Construction
- Location
- Price
- Number of bedrooms
- Number of parking places/garage
- WC
- Full area
- Useful Area
- Energy Performance Certificate
- Approximate GPS coordinates
- Approximate Zip-Code
- URL
- Thumbnails/Images of the listed house (3 in total)
- Description

Obtaining this data was crucial to begin the process of visualizing the intended dashboard. By focusing on the effort to present the current real estate situation, also several metrics need to be defined, using the given data.

It is also going to be important to defining what kind of report will be created around the collected data from these websites.

7.5.1. MEASURES AND METRICS

Considering that the application of this project on real estate market is to provide a complete visualization on the proposed four parishes of Lisbon, Oeiras, Amadora and Odivelas, this chapter introduces the metrics and measures that will be presented in the final delivery.

On table 2 are presented the measures and metrics that incorporate the tentative design:

Table 2 - Measures and Metrics

Name	*Measure /Metric*	*Description*	*Observation*
Houses	Measure	**Houses Available**	The number of houses that are still available for sale
	Measure	**Houses Sold**	The number of houses sold that are still available in the report
	Measure	**Sold in the last 10 Days**	Number of houses that were sold in the last 10 days from current date.
	Measure	**Houses Reserved**	The number of houses that are still available for sale but with status 'Reservado'
Listing	Metric	**Listed On**	Date when the house arrived the model
	Metric	**Days for sale**	Total days in the house is in the market
	Metric	**Price Square Meter**	Worth of a square meter in EUR
	Metric	**Parish Average Price**	Average price of parish for total available houses
	Metric	**Above/Below price**	Price relation with the average price of the same parish
Market	Measure	**Housing Mkt Value**	Market Value of the available houses for sale
	Metric	**Average Price (m2)**	Average price for square meters for Parish
Prediction	Metric	**Price Prediction by Area**	Linear regression to predict house pricing according to gross area.

7.5.2. LINEAR REGRESSION AND HOUSING MARKET

From Bangdiwala's "Regression: simple linear" paper on the International Journal of Injury Control and Safety Promotion Linear in 2018, linear regression models assume that the relationship between a dependent continuous variable Y and one or more explanatory (independent) variables X is linear.

- **Simple Linear Regression Mathematical Model**

  The simple linear model is exactly that, a simple straight line that relates the one independent variable X to the dependent variable Y. It is given by the mathematical formula for a straight line where b0 is called the intercept and b1 is called the slope. In the standard x–y Cartesian plane, the intercept is the point on the y-axis that is intersected by the line, and the slope is the amount of change in the y-axis for a one-unit change in the x-axis (S. I. Bangdiwala, 2018).

  Fitting a Simple Linear Regression Model requires the statistical regression model to be estimated based on actual observations, obtained from a simple random sample of size n from some population of interest. To be able to incorporate this mathematical model on the artifact it was necessary to find in the literature a proof that it could be used on this project. From the website Towards data science article written by Venelin Valkov, it is observed that the Linear Regression Model can fit the tentative design. The article analyzes a training data set containing 1460 data points. On figure 9 are represented the descriptive statistics shown by the article concerning the test data:

```
count      1460.000000
mean     180921.195890
std       79442.502883
min       34900.000000
25%      129975.000000
50%      163000.000000
75%      214000.000000
max      755000.000000
Name: SalePrice, dtype: float64
```

Figure 9 - Descriptive Statistics – Source: Towards Data Science (Venelin Valkov, 2019)

Also, on figure 10 are represented the distribution/density of the data by price and the scattered chart of price versus living area (square feet):

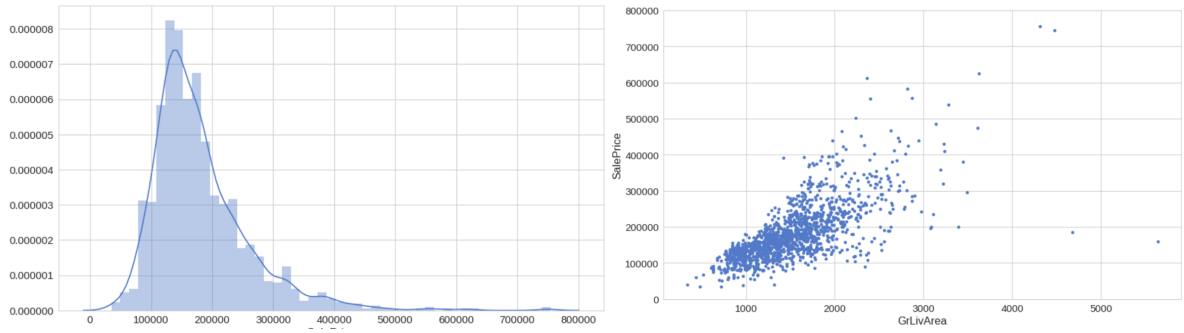


Figure 10 – Price Density and Living Area against Price Scatter - Towards Data Science(Venelin Valkov, 2019)

The distribution of the house pricing follows a density between 100k and 250k and the price versus living area chart confirms the expectations, the price increases with the size of the house. There are also some outliers that may bring additional precaution to deal with.

Although there are some features on these data set that cannot probably be captured by the data scraping of the websites of the agencies, there is a good correlation matrix to help understand if Linear Regression can fit our project while there are similar attributes that the future data set is expected to have. Below is the correlation matrix that includes the Price, Living Area Size, Garages, and Bedrooms which will naturally be also present in the project's dataset (figure 11):

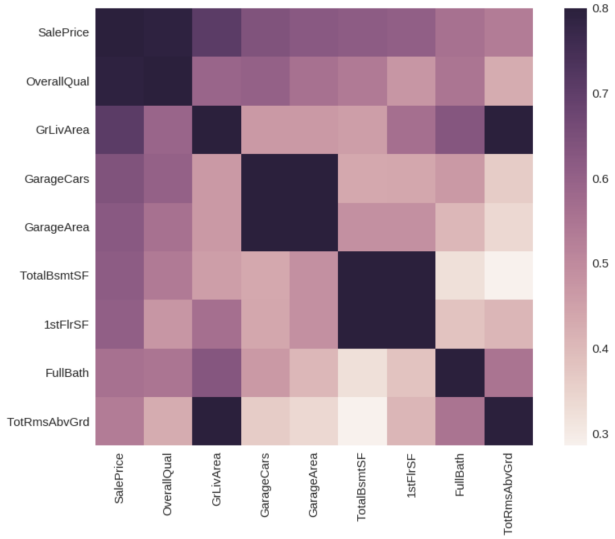


Figure 11 - Correlation Matrix - Towards Data Science - (Venelin Valkov, 2019)

The matrix is providing a diagnostic for the regression. The living area feature seems, according to the author at the end of the full analysis on the article, a good parameter to be used on the model to predict the house prices.

- **Simple Linear Regression formula**

$$Y = \beta_0 + \beta_1 X_1$$

- Y: Predicted value (calculated from β_0, β_1 and X_1)
- β_0 : Constant Interception

$$\beta_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

- β_1 : Coefficient (Slope)

$$\beta_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

- X_1 : Predictor

Since the Linear Regression to be used is simple, only one predictor will be used.
Based on the results of the study developed one Towards Data Science website article the tentative design will try to incorporate and adapt the predictive model of Simple Linear Regression although as the literature explains, there are more advanced and complex models that have a better fit to this case. However, to proof the concept, Linear Regression will be used to its lower complexity when comparing to other models.

7.5.3. REPORT CONTEXT

The goal of this chapter is to provide context to the final report creation. The literature really was important to clear the vision for the design of the conceptual model. As such, the analytics report should be predisposed to a fast perception of the data, promote an easy exploration of the information where implicit data is easily fetched or in the best-case scenario, *at a glance*.

The tentative design should also include an effective communication, which according to Zheng in 2020, must be able to present information in a story telling way. It is also very important that the final product can be used by non-technical users to receive high level summaries. It needs to place the information in a way that the most critical data can be accessed without any effort. Additionally, it must incorporate several different visual graphics that can make a good illustration of the most relevant information to the users.

Having in mind all these concepts it was envisioned to have four integrated pages in one report: Map and List, House View, Overview Dashboard, Pricing Prediction. Find below a small description of each, and a generic presentation of all, starting by figure 12:

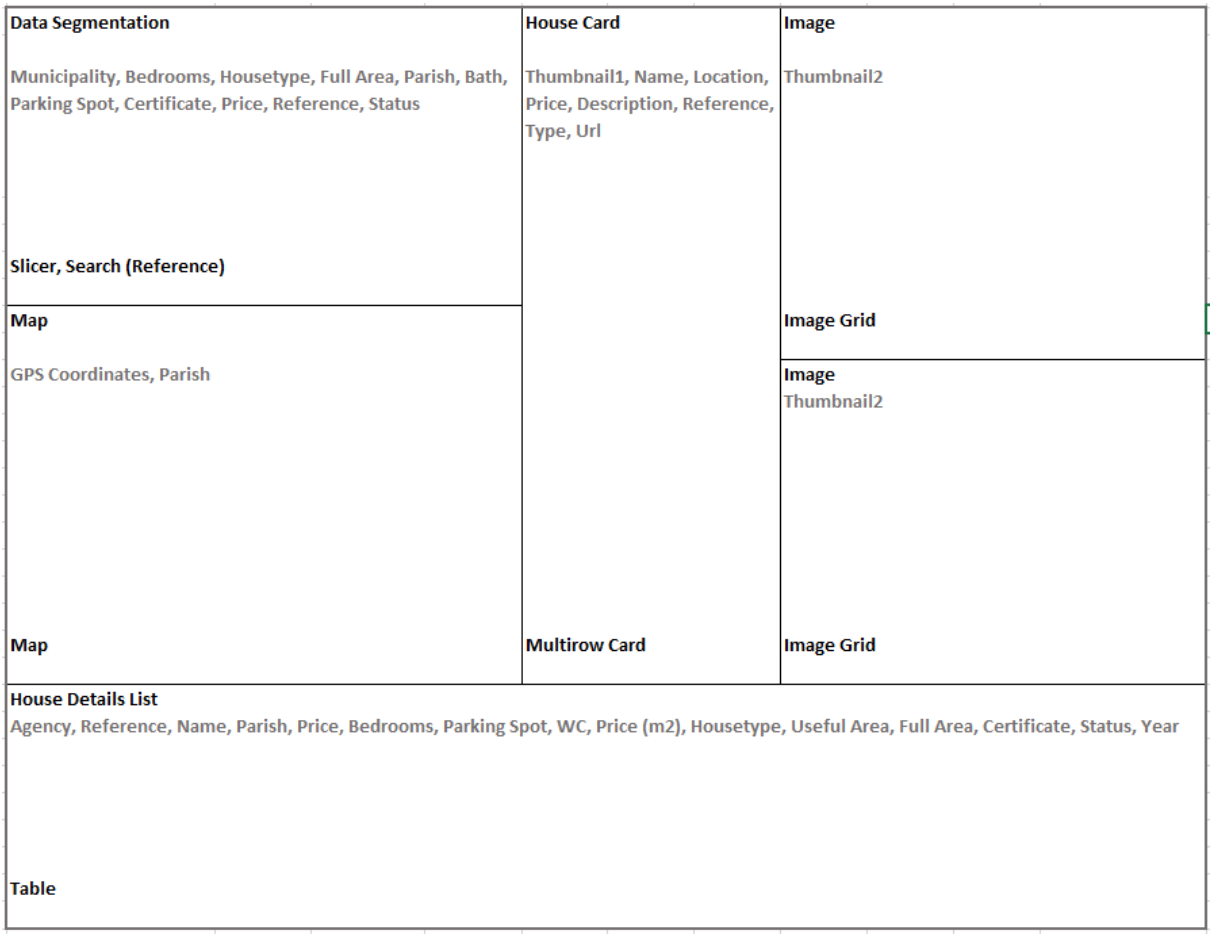


Figure 12 - Conceptual Model - Page Module - Map and List

The first page, “Map and List”, is supposed to let users roam through information on a map that is pinpointing an approximate address of the houses in the data model. Users also can interact with different data segmentation slicers. On the bottom, a list will try to encapsulate all the important information of houses within the searched criteria. On the other hand, the remaining visualizations should concentrate to present the information on a more website type of view. A house card displaying general information together with images will give a much more depth view of the searched houses.

The “House View” page (figure 13) tentative design module:

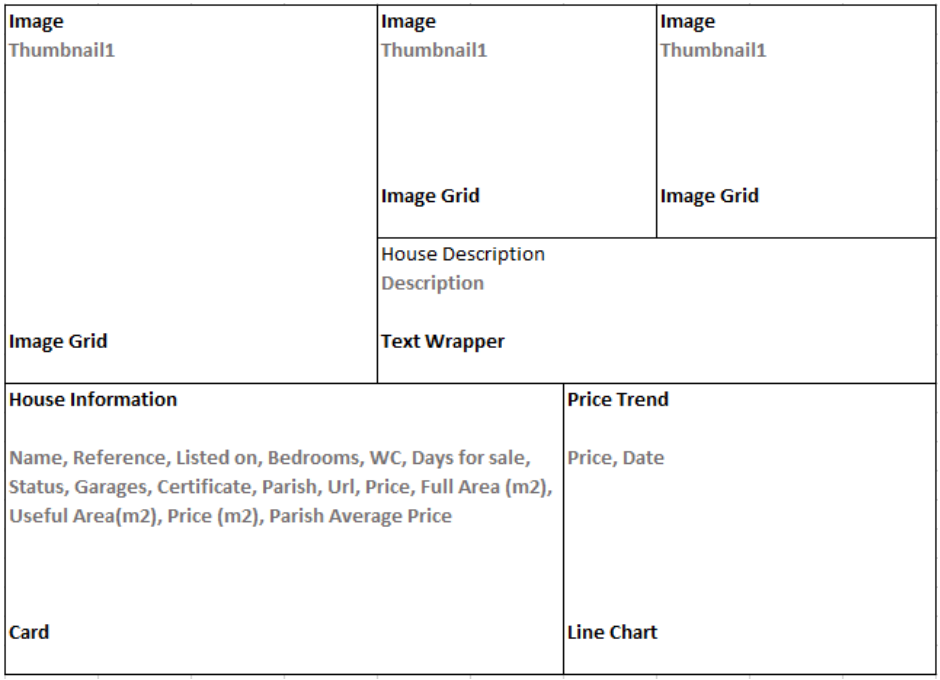


Figure 13 - Conceptual Model - Page Module - House View

This page will explicitly gather all the important information of the house in a unique place. It will display not only all the gathered images of the house but all the general information on the house. It is supposed to be comparable to a website type of report. It will also incorporate price trends and average price values concerning the parish and the typology of the house in one place. The full description and the URL of the house will be also available on this page.

On figure 14, is presented the "Overview Dashboard Module":

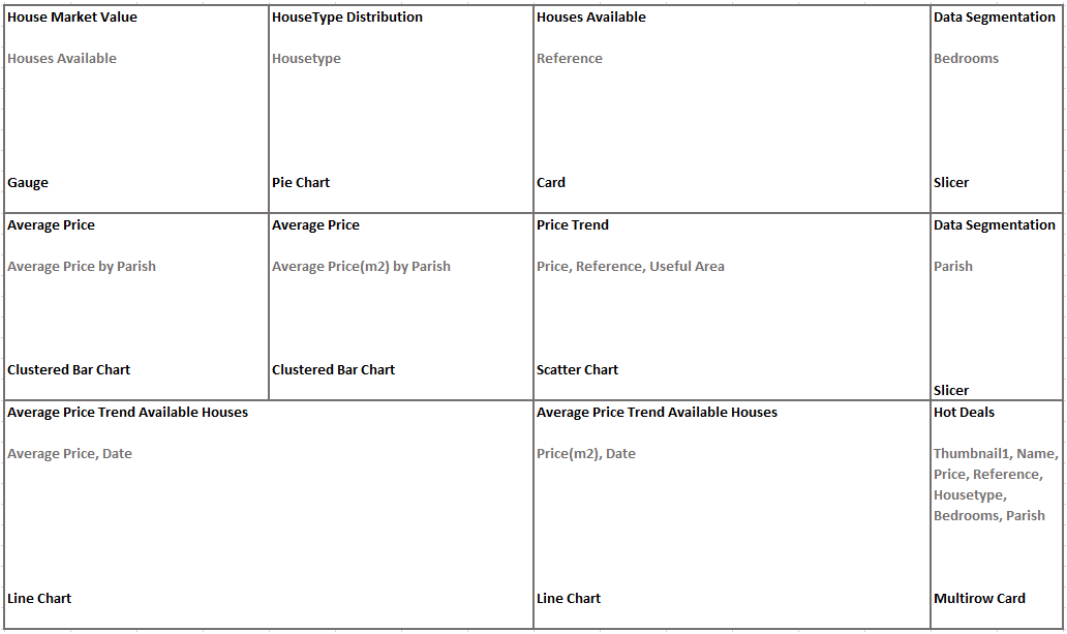


Figure 14 - Conceptual Model - Page Module - Overview Dashboard

The Overview Dashboard will provide lots of drill through chances to the user. In this page it will be possible to analyze the full market regarding the House Types, Market Value, current Average Price and Square Meter Price and its price trends. The user will have the chance to use the data segmentation to discover all the information above for each desired, Municipality, Parish and Typology.

The last page, Price Prediction (figure 15), will give the user the opportunity to choose a house segment between several options like, Parish, Bedrooms, Garages between others, and to frame all the houses within the same level in the search criteria to compare the results.

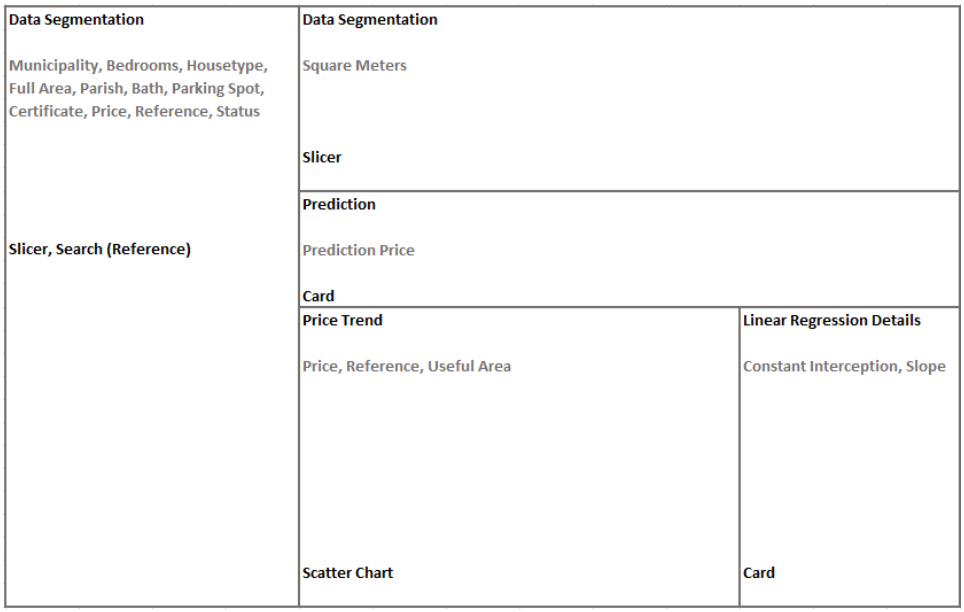


Figure 15 - Conceptual Model - Page Module - Price Prediction

The user will be able to Predict the price of a hypothetical house on any given location given its size in square meters. Also, the design was thought to enable using the available slicers, in order to be more specific and get better results. The more specific it gets, better will be prediction be, as longs as there is data to back up the analysis. In case there is insufficient data, it is believed that the prediction will not work.

8. DEVELOPMENT

8.1. TOOLS AND PLATFORMS

This chapter is all about the development of the tentative design. The tentative was conceived without any constraint regarding the tools to be used. However, the software chosen to materialize the project was elected according to what is perceived as the best tools to use based on the literature.

The software used to build the artifact was based of four main programs. Jupyter Notebook in the first phase, Microsoft SQL Server Management Studio and Microsoft Visual Studio (SSIS) for the second phase and Microsoft Power BI Desktop for the last.

Perkel in 2018 described Jupyter as a free, open-source, interactive web tool known as a computational notebook, which researchers can use to combine software code. Jupyter has exploded in popularity over the past couple of years since that easily allows the notebook to speak dozens of programming languages without much effort. Jupyter notebook was used as a bridge to use Python language to perform web scraping on real estate agencies websites.

Web scraping is described by Ryan Mitchell in 2015 on his book “Web Scraping with Python”, as the practice of gathering data through any means other than a human using a web browser, which is commonly accomplished by writing an automated program that queries the webserver, requests data (usually in the form of HTML), and then parses the data to extract the needed information.

If you can view it in your browser, you can access it via a Python Script. If you can access it in a script, you can store it in a database and if you can store it in a database, you can do virtually anything with that data (Ryan Mitchell, 2015). This sentence from his book described almost entirely the vision that was had in design of the tentative. Hence, the tools used in the first stage of the development were decided based on the literature available and found.

For the second phase and third phase of the development, a solution of ETL and Data Warehousing was developed Microsoft´s tools entirely. On figure 16 there are explicitly shown the players of the BI industry in the 2020’s Magic Quadrant for Analytics and BI Platform.

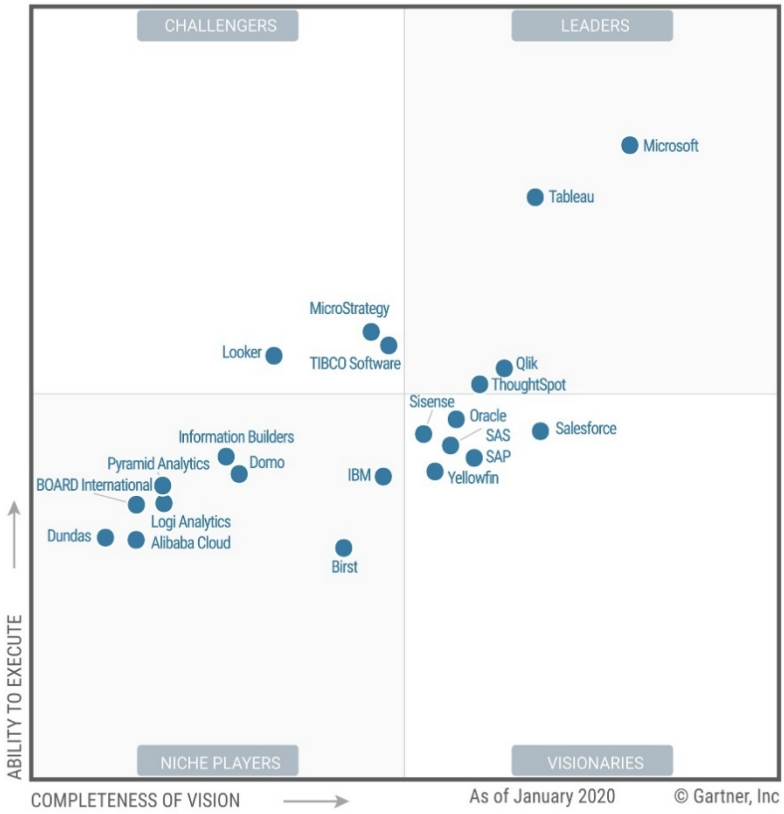


Figure 16 - Magic Quadrant for Analytics and BI Platforms (Gartner, 2020)

Howson et al., 2018, in Gartner Research, created what is called the Gartner Magic Quadrant for Analytics and BI Platforms. This report is produced every spring and is used to show how each key vendor of BI and Data Analytics platforms are performing across two key metrics, the completeness of vision and the ability to execute. By putting their analysis under a visual representation, it is possible to analyze what platforms are considered the best but also in which platforms should those vendors be investing.

As demonstrated, nowadays Microsoft is clearly ahead of all the other contestants in both dimensions, being Tableau a little behind. This makes Microsoft, in Gartner's report, to be classified as a Leader which means that Microsoft demonstrates a solid understanding of its products capabilities and commitment to customer success, that buyers in this market demand.

8.2. PYTHON WEB SCRAPING

The development level started by coding the Python web crawling scripts. The real estate agencies selected were Remax, ERA and Century 21. These agencies are franchises currently operating in Portugal and are according to the Portuguese Revista Imobiliária´s website, are the largest real estate agencies in the country. The decision to include these agencies was to provide to the model a great

sample of data of the Portuguese real estate market using the least number of agencies. This way it would be possible to have a great set of data that could trully represent the market.

As seen described above by Ryan Mitchell (2015), obtaining data by any means other than a human manually using a web browser, can be considered web scraping. However, the process and the tasks performed by the web crawler can be very similar to what a human would do. It is demonstrated that there can be several ways to develop a script, however, the idea behind code is to tell Python to perform as a human would, if given the task. By predicting and handling all the exceptions that may occur when navigating through the web pages the software can know what to do in specific situations, for instance, not finding a webpage due to bad link or connection.

If a user were to gather data from a website in mass, he could start by opening the website and navigating through each page, identifying the desired information, copy it and store it elsewhere. The same applies when collecting data through a script. The difference is that for the script to run, it needs to know precisely which data to get and where to get it from. The websites chosen are HTML based, meaning that they follow a structure that can be used advantageously to retrieve the data.

With an automated process there are adverse situations that can occur. These challenges would easily be solved by a human, but using a script, they need to be predicted and handled accordingly to prevent running errors. For example, if the website is down or if there is a webpage that no longer exists, if the exception is not handled, the script will stop running and will print an error on the Python console. It is very important to predict most of these situations in order to guarantee a smooth experience. To achieve a certain level of reliability and robustness the code was tested several repeated times until all the exceptions captured could be handled thus resulting in data being gathered successfully. However, it is always possible that changes done by the owners to the website, may cause some negative impact on the code, hence, when the script fails to do its job, it must be conducted an analysis to the errors that contributed to the failure and the script must be readjusted.

There is in total three scripts, one for each agency. However, in Remax's and Century 21's case the script was divided in four equal scripts, only changing by municipality. The script is the same but was separated into different files, each dedicated to its region. The reason behind is because that not only by separating into different files it is possible, yet in some cases not advisable, to run the scripts in parallel to collect data faster. This enables also to diversify the risk of a script failing and losing all the collected data until the failure. To provide a concrete example, if a script had already collected all houses from Lisbon in a total of 180 pages each page containing 20 houses, failing in the middle of scrapping another region, all that data would be lost, and the script would need to run again.

Contrarily to ERA, which is entirely HTML structured website, Remax and Century 21 have embedded some Javascript coding that needs to be executed first for each page of these websites. To briefly explain, to be able to present the website fully in HTML, the script needs to handle the Javascript first. For that, it was used an open-source tool named Selenium. This tool that can be imported directly into Python environment through code and will act as part of the script.

All the scripts have the main structure and logic expressed on figure 17:

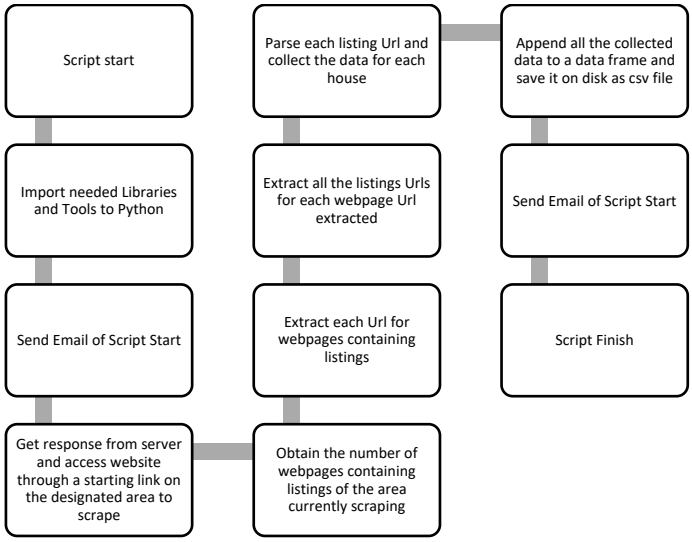


Figure 17 - Web Scraping Scripts, Logic and Structure

Although the scripts essentially follow the main process above, Century 21 and Remax as explained before, will have to deal with the Javascript execution for every Url opened at least once, so, every time that an Url containing house listings, or an Url of a specific house is opened, the script will run the Javascript code using Selenium to properly display the HTML structure of the. After that, the Python can treat the result as it would normally do in ERA's case and retrieve all the data normally.

8.2.1. HANDLING EXCEPTIONS

For handling exceptions that may prevent the script to run successfully to end, the scripts were written mostly on a try/except basis. To properly explain this definition, the python documentation on w3schools.com website was most useful (figure 18):

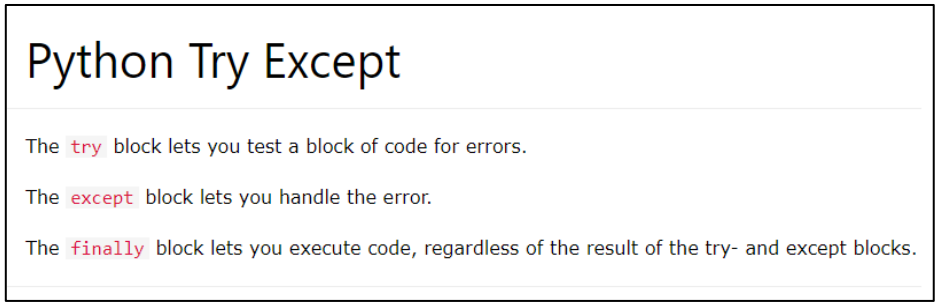


Figure 18 - Python Try Except (source: www.w3schools.com)

The above figure is explaining that by writing some code and inserting it on a try clause, it is possible to test it and if all goes well, the code will normally run. If the code does not work the try block will raise an exception automatically.

Of course, by predicting all the outcomes and exceptions within the code it is possible to prevent the script to fail, or at least minimize greatly the chances to it. As seen on the figure 19, there may exist several exceptions.

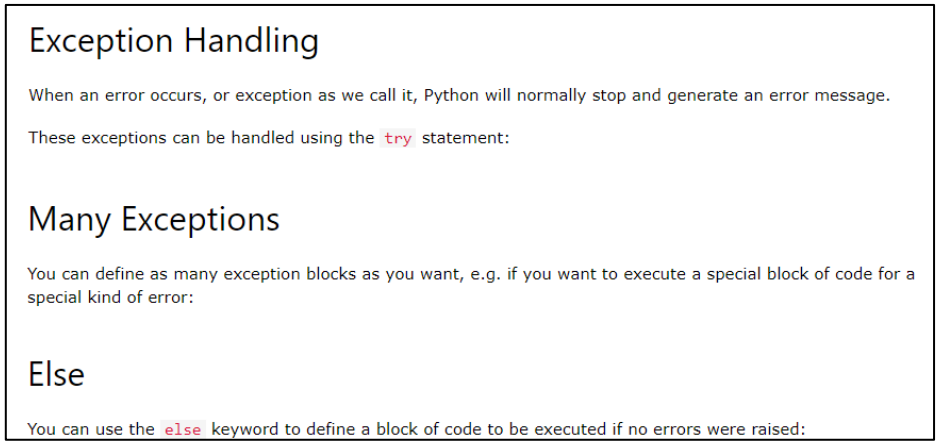


Figure 19 - Exception Handling (from the website: www.w3schools.com)

Running the script several times can catch most of the exceptions and with the necessary adjustments the script will become more robust. The principal concern is if the website remains the actual structure throughout time. Any change on the website can invalidate the script, the introduction of a complete new site by the agencie will force the creation of an entire new script.

8.3. DATA SOURCES

The data that will feed the model is the result of the extraction performed on the first stage of the process. The data must follow some criteria to simplify the incoming processes and to be standardized between all agencies.

The data set will be composed by several flat files according to the number of agencies and areas scrapped. Any flat file from any website scrapped in the future may be added if the file keeps the data structure.

The attributes in table 3 were defined to be the necessary standard features of the flat files:

Table 3 - Elected Source Attributes

Source Attributes	**Agency**; **Agency_ID**; **Reference**; **Housetype**; **Sold**; **Zone_ID**; **Zone**; **Parish**; **Year of Construction**; **Location**; **Price**; **Number_of_bedrooms**; **Number_of_garages**; **WC**; **Fullarea (m2)**; **Usefull_area**; **Certificate**; **Gps**; **Zip_code**; **Url**; **Thumbnail1**; **Thumbnail2**; **Thumbnail3**; **Name**; **Status**; **Description**.

At this stage, all values are normally will be stored in a csv file.

There are some rules necessary to be taken to standardize all data between file sources:

- Agency ID needs to be according to the below:
    - ERA - Agency ID 1
    - Century 21 – Agency ID 2
    - Remax – Agency ID 3
    - Should a new agency be added, the ID will be incremented
- Zone ID needs to be according to the below:
    - Lisboa - 1
    - Oeiras - 2
    - Amadora - 3
    - Odivelas- 4

Regarding the gathered data here are some more assumptions that were taken according to each agency in order to present the best possible fitting model.

For **ERA** agency here are the necessary assumptions taken:

- **Year of Construction** - this attribute is never provided by ERA.
- **Price** - When price equals to “Preço sob consulta” on any listing, the price will be set to 0 by the script.

- **Housetype** – When Housetype has a data point “Outros” then it means that the row is a terrain, a commercial space or a building.
- **Full area/Useful area** - If full area is not provided by the website the value for useful area will be used to fill the value of the full area. The same happens if useful area is not found. If neither are found default value becomes a default value “99” meaning it will be disposable at a later stage.
- **Number of bedrooms/Number of garages/WC** – Will be set to 1 if the site does not provide the information.
- **Remaining attributes** – the remaining attributes when not provided by the website will present the character “- “.
- **Reference** – It was decided to add a prefix “ERA- “to the references of this agency. This is to prevent possible repeated values. Also, Microsoft Visual Studio software is adding leading zeros on the left or right of these values. The prefix solves that bug.
- **Description** – Has to be less than 3000 characters.

For **Century 21** agency here are the necessary assumptions taken:

- **Price** - When price equals to “Preço Sob Consulta” on any listing, the price will be set to 0 by the script.
- **Full area/Useful area** - The same happens if useful area is not found. If neither are found default value becomes a default value “99” meaning it will be disposable at a later stage.
- **Number of bedrooms** – If not provided then it means that the row is a terrain, a commercial space or a building.
- **Number of garages/WC** – Will be set to 0 if the site does not provide the information.
- **Remaining attributes** – the remaining attributes when not provided by the website will present the character “- “.
- **Description** – Has to be less than 3000 characters.

For **REMAX** agency here are the necessary assumptions taken:

- **HouseType** - When HouseType equals to “Duplex” or “Estúdio” on any listing, the HouseType applied will be “Apartamento”.
- **Number of garages** - Sometimes number of garages can have designations such as “3+”, “Sim” or “Não”. These values when caught by the script will be converted respectively to 3, 1 and 0.
- **WC** – Sometimes WC is designated as “-2” or “-1”. This value will be changed by the script to its integer.

- **Full area/Useful area** - The same happens if useful area is not found. If neither are found default value becomes a default value “99” meaning it will be disposable at a later stage.
- **Description** – Has to be less than 3000 characters.

8.4. SCRIPT TASK SCHEDULER

For feeding the database it will be necessary to fetch the data in a timely basis. The frequency may be daily, every two or three days, or even weekly. The big difference is that the facts table will be as granular, in terms of date, as the script is running since there is a periodic facts table. If the script is running daily the lowest granularity will be day. If the script runs weekly the granularity is obviously limited to week. The model is prepared for presenting daily factual data, which is the lowest granularity. If the Python scripts only run weekly, there can exist for some houses a weekly lag, meaning that the presented values can be already outdated. However, in market liquidity a weekly lag will not affect drastically the numbers in the model due to the slow changing pace of this type of sector, but it can affect the time for any given change in price or other information to be visible.

It was necessary to find a way to let Windows know that the code on the script is indeed a program written in Python. Python is installed on C:\Users\andre\anaconda3 and to be able to activate the scripting in Python the folder C:\Users\andre\anaconda3\Scripts must also be accessed.

To all the scripts, stored on the folder G:\Github\masterthesis, it was created a bat file using Notepad forcing Windows to run the code as a program. For every script, the same code was written on Notepad, and only the path that leads to each script was changed. On figure 20 there is an example of Century 21’s Amadora Script call written on Notepad:

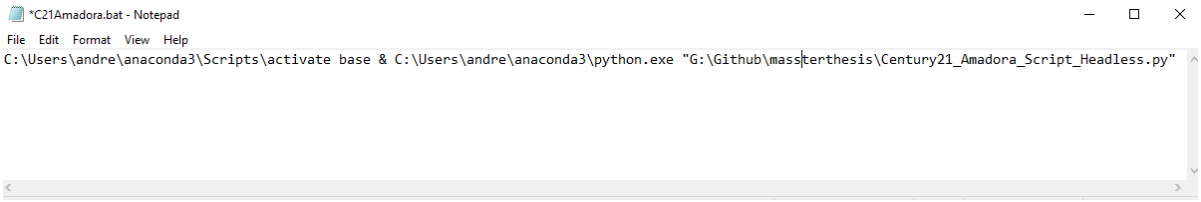


Figure 20 - Notepad Bat Files

Additionally, it was also necessary to find a way to run the bat files, and if possible, benefit from automation. Recurring to Windows Task Scheduler it was possible to create tasks which launch the bat files with the defined schedules and launch recurrency. Figure 21 resumes all the tasks that were created and the trigger time and recurrency of each task. In the picture all the scripts will run in a daily basis and will mostly run nightly and until dawn.

Name	Status	Triggers
Century 21 Odivelas	Ready	At 07:30 every day
Century21 Amadora	Ready	At 08:30 every day
Century21 Oeiras	Ready	At 08:00 every day
Remax Oeiras	Ready	At 06:00 every day
Remax Odivelas	Ready	At 07:00 every day
Remax Amadora	Ready	At 06:35 every day
Remax Lisboa	Ready	At 03:00 every day
Run Century 21_Lisboa	Ready	At 23:00 every day
Run ERA	Ready	At 23:00 every day

Figure 21 - Scheduled Tasks

8.5. DIMENSIONAL MODEL

In the previous chapter there was an introduction to the data sources and attributes constituting those data sets. These flat files that are in csv format, so the data will be categorized at a later stage.

It was necessary to create a dimensional model from root that could attain the specific results in terms of measures and metrics proposed on the beginning. But it also needs to be design having in mind the possibility to store the data while creating an historical database. This dimensional model was designed having in mind the possessed data, gathered by the Python scripts until this phase. The star schema can be seen in figure 22:

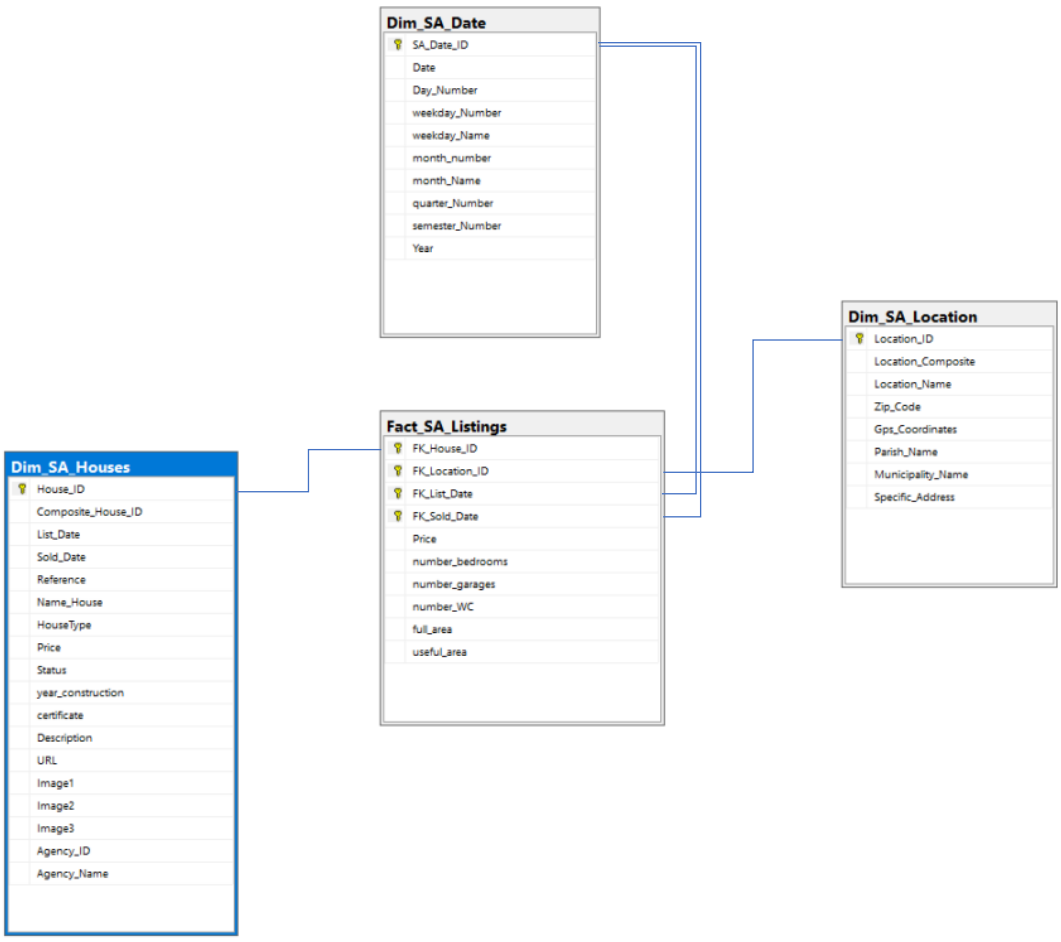


Figure 22 - Listed Houses Dimensional Model

The star schema methodology is often used for data modeling, and it is introduced in Ralph Kimball´s book “The Data Warehouse Toolkit” as an easy-to-use model for data query performance. In a relational DBMS environment, the base for this development, a fact table is constructed with one record for each discrete measurement and then surrounded by a set of dimension tables that describe precisely what is known in each context of that measurement. Due to the nature of this dimensional model and its disposition it is often called star schema. Also, it is used to lower the complexity of the model, which is centered around the fact table.

After conceiving the dimensional model, it was essential to focus on database building. Following to Kimball´s methodology, it was required the creation of a staging area which would be responsible to act as an intermediary on the ETL process until the data reaches the data warehouse. This SA is a landing zone where the source files are loaded, treated and transformed so they can be properly loaded into the destined DW.

Figure 23 is showing the different data types that will be used to categorize the data when entering the ETL phase:

House_ID	int	number_bedrooms	int		
Composite_House_ID	bigint	number_garage	int		
List_Date	date	number_WC	int		
Sold_Date	date	full_area	int	Location_ID	int
Agency_ID	int	usefull_Area	int	Municipality_ID	int
Agency_Name	nvarchar(50)	year_construction	nvarchar(50)	Municipality	nvarchar(100)
Reference	nvarchar(50)	certificate	nvarchar(50)	Parish	nvarchar(100)
Name_House	nvarchar(100)	Description	nvarchar(3000)	Location	nvarchar(100)
HouseType	nvarchar(50)	URL	nvarchar(700)	Location_Composite	int
Status	nvarchar(50)	Image_URL_1	nvarchar(300)	Zip_Code	nvarchar(100)
Price	int	Image_URL_2	nvarchar(300)	Gps_coordinates	nvarchar(100)
		Image_URL_3	nvarchar(300)		

Figure 23 - Data Types of Source Attributes

Some of the attributes as Composite_House_ID and Location Composite are derivatives of transformations of primary attributes. These were created to add context to the dimensional model and will be essential to the ETL process which is the next stage.

8.6. ETL PROCESS AND DATA-WAREHOUSING

8.6.1. STAGING AREA

This chapter joins two great BI concepts which are ETL and Data-Warehousing. Following the approach presented by and Kimball, R., & Ross, M. (2002), the development starts with the creation of a staging area. It has the utmost importance since the staging area represents the link between the operational source systems and the data presentation area. The data will not be ready to consumption or to be queried by users until it suffers the necessary transformations. After that it will be loaded into the warehouse. The staging area, as seen in the image (figure 24) contains 8 tables. It features one Facts

table, surrounded by four Dimension tables, Date, Location, Houses and Parishes. It also has an ETL log table and an Error log table.

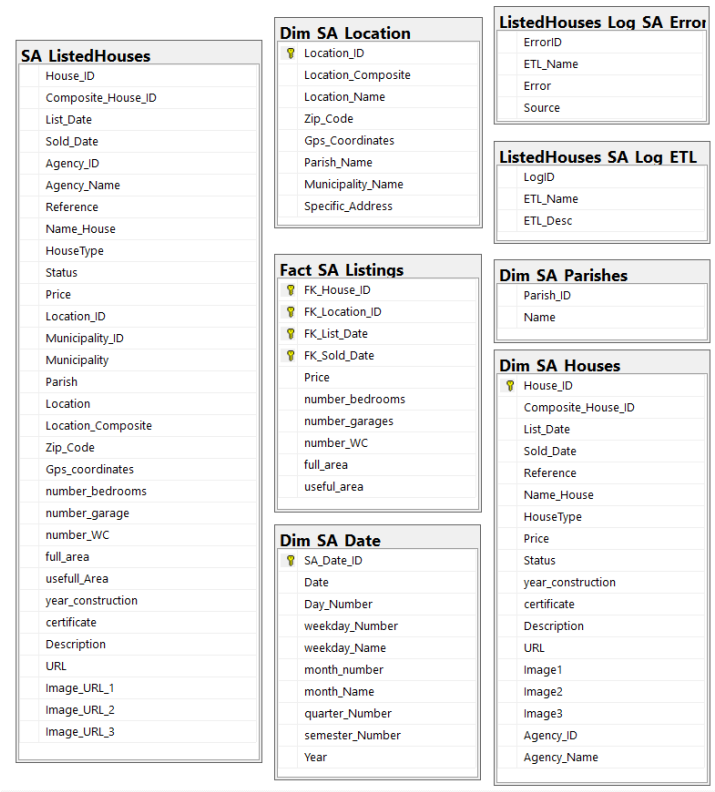


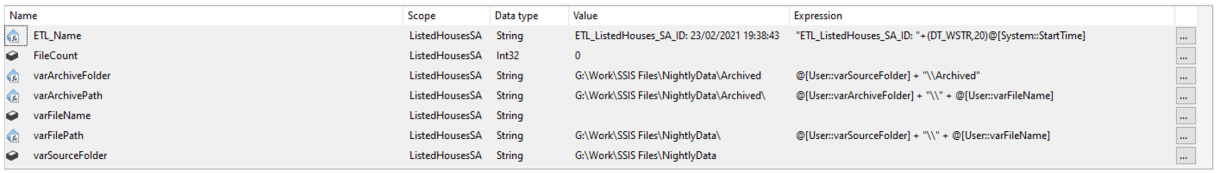
Figure 24 - Listed Houses Staging Area Tables

The table named SA ListedHouses exists for the purpose of receiving and storing the raw data. The dimension tables, act as a pedal point to create all the primary keys necessary to the dimensional model and to automatically and incrementally fill the data with those keys. For the purpose of simplifying the model and to more easily query the data, the dimension tables were also used to update the main table SA ListedHouses contributing actively to a tabular display of information which helped to turn the analysis of data easier.

Figure 25 is showing all the connections used in the staging area package. These will be explained along this chapter.

Figure 25 - Staging Area Connections

On figure 26, there are seven variables, five of which are associated with file task processes, varArchivePath, varArchiveFolder, varFilePath, varSourceFolder and varFileName and other two necessary to the ETL process naming and management.



Name	Scope	Data type	Value	Expression
ETL_Name	ListedHousesSA	String	ETL_ListedHouses_SA_ID: 23/02/2021 19:38:43	"ETL_ListedHouses_SA_ID: "+(DT_WSTR,20)@[System::StartTime]
FileCount	ListedHousesSA	Int32	0	
varArchiveFolder	ListedHousesSA	String	G:\Work\SSIS Files\NightlyData\Archived	@[User::varSourceFolder] + "\\Archived"
varArchivePath	ListedHousesSA	String	G:\Work\SSIS Files\NightlyData\Archived\	@[User::varArchiveFolder] + "\\" + @[User::varFileName]
varFileName	ListedHousesSA	String		
varFilePath	ListedHousesSA	String	G:\Work\SSIS Files\NightlyData\	@[User::varSourceFolder] + "\\" + @[User::varFileName]
varSourceFolder	ListedHousesSA	String	G:\Work\SSIS Files\NightlyData	

Figure 26 - Variable Creation

These variables were created for automation purposes. ETL_Name is a string variable and was used to automatically name the intended processes when needed. When called, will give a name composed by the name of the package and the date and the time for logging purposes. FileCount was created as an int variable. It will act as a count start value, and it will allow the system to know how many files exist to be integrated. As it will be further explained, this variable will be necessary to handle exceptions that may occur while running the ETL process. For instance, if the Python scripts fail to run overnight or are not terminated, the Nightly Data folder, created to aggregate all data sources that are created from the scripts may remain empty. This will lead to the ETL process to fail because it will not find any files when scheduled to run. If the ETL fails to run the log will be available on the log table.

Regarding the connections, in figure 27, the Dynamic flat file connection was used to point the connection to the path “G:\Work\SSIS Files\NightlyData\” which leads to the folder Nightly Data. This folder was designated to store all the file sources that resulted from running the scripts. By defining the Connection String Expression as the variable created by user, varFilePath, and the Flat File Connection String to the path of the folder this one connection becomes dynamic and can load several files from the source folder.

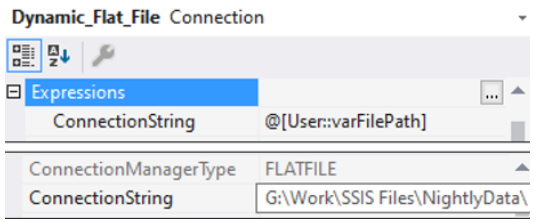


Figure 27 - Dynamic Flat File Connection

For the ETL process to know which files are to be integrated, the below instructions were applied inside the For Each Container Loop (figure 28):

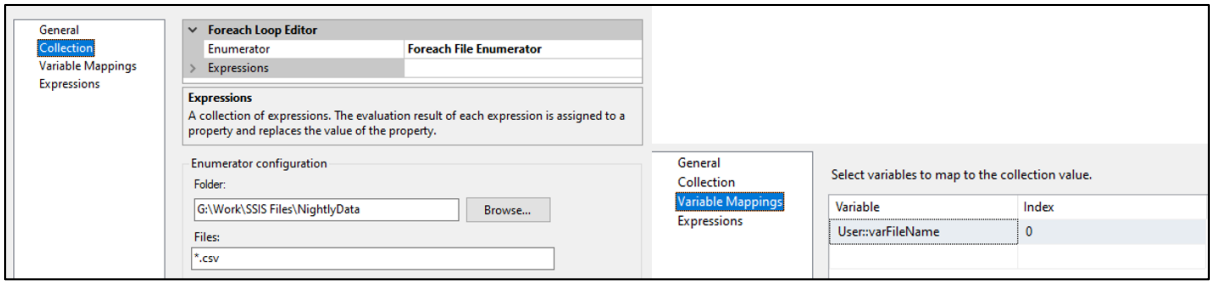


Figure 28 - For Each Container Loop Configuration

This enables the container to recognize the csv and use them to load the data into the SA.

The ETL process starts with a sequence container to clear the previous data from all the tables from of the SA. As the SA contains no relationships, all the data is truncated. Truncate will reset every counter to default while erasing all the data from the tables. The container also triggers logging if any error and creates text logs of the current process. Both types of logs will be inserted into to the Log_SA_Errors and Log_SA_ETL respectively.

Data loading process follows up with a For Each Loop Container. This container is useful because it allows the system to loop through the several data source files of the different agencies and apply the sub process within (figure 29).

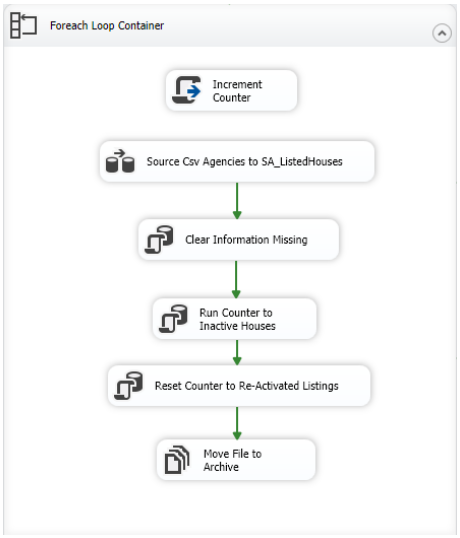


Figure 29 - For Each Container Loop Sequential Process

This sub process starts with a Script Task (figure 30) called Increment Counter which will add a counter by the intermediary of the variable FileCount based on the number of files that are found on the NightlyData folder.

```
public void Main()
{
    // Get value of counter variable and increment with 1
    Dts.Variables["User::FileCount"].Value = Convert.ToInt32(Dts.Variables["User::FileCount"].Value) + 1;

    Dts.TaskResult = (int)ScriptResults.Success;
}
```

Figure 30 - Increment Counter Code

This way, and by using another figure 31 Script task code after the Container Loop, the model will know exactly how many files will be loaded on the current ETL process.

```
public void Main()
{
    // Check if counter is zero
    if (Dts.Variables["User::FileCount"].Value.ToString() == "0")
    {
        // Throw error event and fail Script Task
        Dts.Events.FireError(-1, "Foreach Loop", "The For Each File enumerator is empty. The For Each File enumerator did not find any files that matched the file pattern, or the specified directory was empty.", String.Empty, 0);
        Dts.TaskResult = (int)ScriptResults.Failure;
    }
    else
    {
        // Files where found so no error
        Dts.TaskResult = (int)ScriptResults.Success;
    }
}
```

Figure 31 - Error Handler Code

Should the validation Script task after the For Each Container Loop find that the counter is equal to zero, the script will throw an exception and stop the ETL Process, which will be logged on the log tables created on the SA.

The remaining of the sub process inside the For Each Loop Container in the control flow is rather objective and has a specific goal that is to load the data sources and categorize the data types of the different attributes, deleting all the rows that have information missing and that will not be taken into the model. All the duplicated rows if any exist are also eliminated.

It was decided that all rows which fail to contain critical information should be automatically deleted. The information that is strictly necessary and cannot assume empty values is on the attributes of Reference, Parish, Zip Code and Location. This decision was taken after analyzing the data from the different data sources and will have a reasonably low impact on the model. This is because the listings that are missing those specific attributes, are houses that were already sold and the listing is currently being dismantled, or listings that are not apartments or dwellings, like land, stores, garages and buildings which do not belong to the target of this project. However, it was necessary to remove them, so the data is properly clean and ready to be loaded.

All the loaded data is being stored in the table SA_ListedHouses. For now, the dimension tables are still empty. They will be later used for acquiring the primary keys automatically and to feed the SA_ListedHouses table with the new information.

After all the information is loaded, the source csv files are moved automatically to an archive folder using a File System Task with the configuration visible on figure 32.

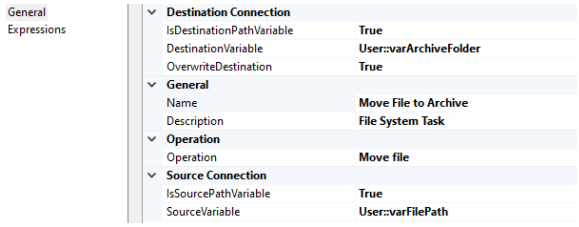


Figure 32 - Archive Folder Path File Task Configuration

8.6.1.1. THE EXPIRED LISTINGS PROBLEM

This model is prepared to deal with listings that are sold and listings that simply expire or are removed by the agencies. The listings that are sold have a sold flag captured by the Python script if existing, but the same does not happen with the listings that naturally expire or are removed by the agencies.

The data loaded will be essentially the same every day for every agency, except the new listings that are added or the updates that the listings suffer throughout the time of their existence. By using a SQL task to make the model find and match the rows already loaded in the DW with the rows being currently loaded using their references, the ETL process will be able to put a counter on the houses that were not matched. This indicates that the houses that were not matched are either not available anymore or the script could not scrape them. As there is always a remote chance that the listing is not scraped by the script for the reasons pointed on the web scraping it was decided to use a counter. By adding a counter using a SQL task, the system will increment one unit to a specific column designed on the DW specifically. Hence, for every reference not caught by the find/match SQL task, the model will be able to exact which houses are not for sale anymore by fine tuning the number of counters needed to exclude the house of the available houses. This implementation was named Expiry Counter Trigger (figure 33).

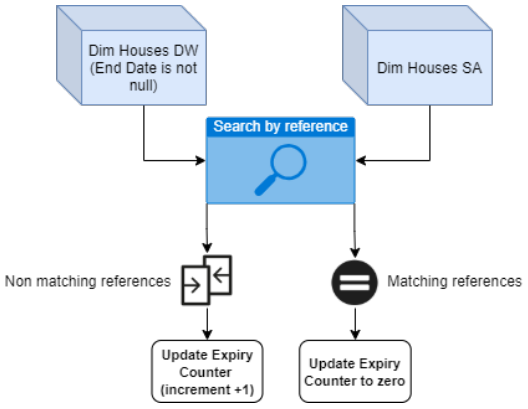


Figure 33 - Expired Listings Problem Scheme

Another SQL task was added to reset the counter if the listing reappears on the model. This implementation was called Expiry Counter Reset. This operation will reset the expiry counter for the houses that reappeared on the market, for instance, the owner decided that he wants to sell the house again. It will also work on the houses that were missed through the script phase by any reason. By running the ETL process in daily the counter will remain for the truly expired since it will not be reset.

The number of tokens that will be taken in consideration for an expired listing to be considered as such will be defined at a later stage since the separation will be done in the report and analytics design stage.

8.6.1.2. LOADING DIMENSION AND FACTS TABLES

For loading the dimension and facts tables a Sequence Container was added. The main subprocess for the dimension tables can be found on figure 34:

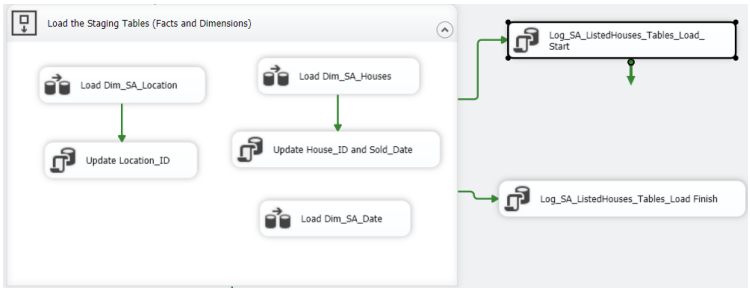


Figure 34 - Loading Staging Tables Container

By looking to the above image, it is noticeable that only Date, Location and Houses dimension tables are participating at the core of the ETL process. The reason behind this is that Parishes table is static and was loaded once at the start of the implementation of the model. While containing all parishes for all the four municipalities proposed in the beginning of this project it has a total of thirty-nine parishes. To fit the designed model, it was determined that the parish table would be static. Some of the parish's names came outdated from the websites and it was needed to harmonize all the parish names among all agencies. For those cases, the Python script will write the name of the outdated parish according to the current one, and the Parish ID will be given at the ETL phase using the Parish table. Also, it is important to notice that the model keeps its future scalability if wanted.

For the date dimension, it was vital to create a flat file source to serve as a calendar (table 4).

PK_Date	Date	DayNumber	WeekDayNumber	WeekDayName	MonthNumber	MonthName	QuarterNumber	SemesterNumber	Year
19000101	01/01/1900	0	0	NoDate	0	NoMonth	0	0	0
20110101	01/01/2011	1	6	Saturday	1	January	1	1	2011
20110102	02/01/2011	2	7	Sunday	1	January	1	1	2011

Table 4 - Date Structure and Granularity

The first date on the figure is called Zero Date. This date was established to be a null value for a date representing date datapoints that are still waiting for an update. For instance, houses that were not sold yet, will have sold date filled with 01/01/1900.

For the date dimension, the same data source file is repeatedly loaded in every ETL process. The calendar may be extended if the structure of the file is maintained, implying that if the rules of the file are satisfied it is possible to upload future dates if desired. Currently, the calendar covers the timeline between 01/01/2011 and 31/12/2030.

Regarding House and Location dimensions, the source used is the data from the SA_ListedHouses table and the destinations are the respective House and Location Staging Area tables. The dimension tables are being used as pedal points to the creation of the primary keys. The system will automatically spring the primary keys Location and House and then the created ID's will be loaded into the SA_ListedHouses table. For Location primary key, a composite Location ID was created by removing all non-numeric characters from the zip code and concatenating it to the Parish ID. This entailed using the Data Conversion and Derived Column modules as seen on figure 35.

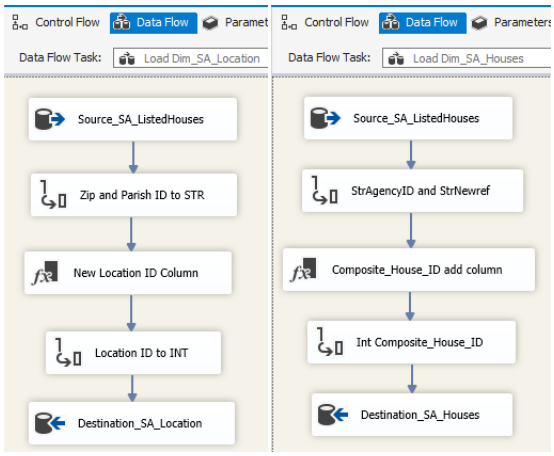


Figure 35 - Location and Houses Data Flow Schemes

The primary keys of Houses table used the same method above to be created. By aggregating the agency ID and reference it was created a composite key which will perfectly fit into the model. In this case the below call to the source had to be done through an SQL command, figure 36, which allowed to manipulate the data to our advantage right from the source:

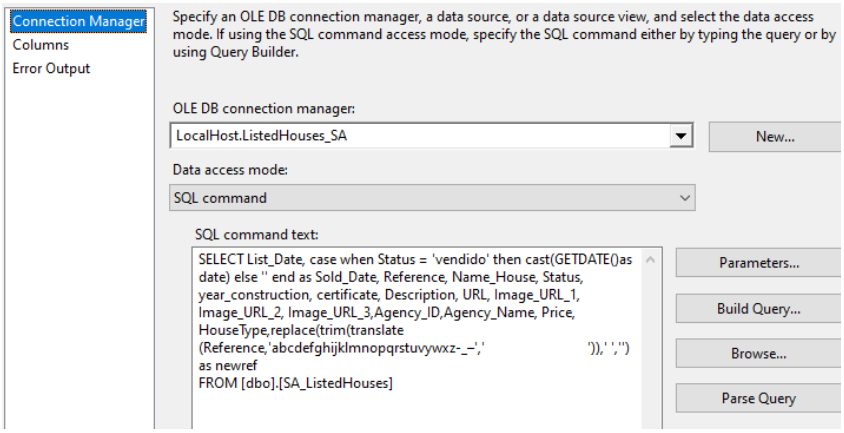


Figure 36 - Source Connection Configuration for Dimension Houses Data Flow

The newref column was created thought the SQL command, ensuring that the Composite House ID could be created as an integer, by removing the alphanumeric digits since some agencies contain letters on their reference. The model needed that the Composite House ID could be an integer to serve as ID. The query also detects if the listing is cast as “vendido” from the status column, meaning that it was already sold. Since there are no details of when the house was really sold on the websites, the houses that entered the model with a sold flag will have Sold Date equal to the Listed Date. The houses that have status “disponível” that means available, if sold after being loaded into the database, the model will be able to change their status automatically to sold when the flag is captured by the python script. All the records will also be kept of when the house was still available for sale.

After the House and Location tables are loaded and the primary keys generated, an SQL task was added to each Data Flow task in order to take the primary keys and the rest of the created columns important to the model to update the SA_ListedHouses table.

To load the facts table Facts_SA_Listings the source used was the table SA_ListedHouses. The process was relatively straightforward since the only flow that was additionally performed was to created List Date and Sold Date as intelligent keys (figure 37) so they could be used as Foreign Keys (FK).

Derived Column Name	Derived Column	Expression	Data Type
IntelligentList_Date	<add as new column>	YEAR(NewList_Date) * 10000 + MONTH(NewList_Date) * 100 + DAY(NewList_Date)	four-byte signed integ...
IntelligentSold_Date	<add as new column>	YEAR(NewSold_Date) * 10000 + MONTH(NewSold_Date) * 100 + DAY(NewSold_Date)	four-byte signed integ...

Figure 37 - Facts Table Intelligent Keys

This way the FK for the Sold and List dates could be used to fetch information from the data dimension. As it is visible on figure 38, the data flow uses SA_ListedHouses as source to load the data, this time in the facts table (Price, number of bedrooms, number of garages, useful area, full area).

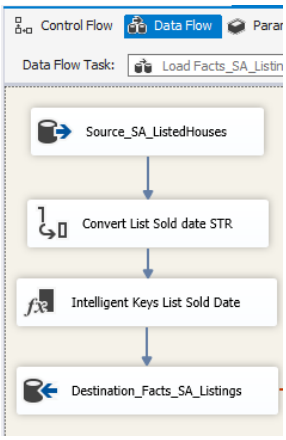


Figure 38 - SA Facts Table Data Flow

8.6.2. DATA-WAREHOUSE

The DW was designed to supply the needs of the model. As much as there are similarities between SA and DW, there are also several real differences. While is out of the production phase, the DW is supposed to let the analyst to extract and analyze the data contrarily to the SA. Another big difference is that while there are no relationships between the tables of the Staging Area, there are multiple relationships on the Data Warehouse. On figure 39 is the Data Warehouse for this project:

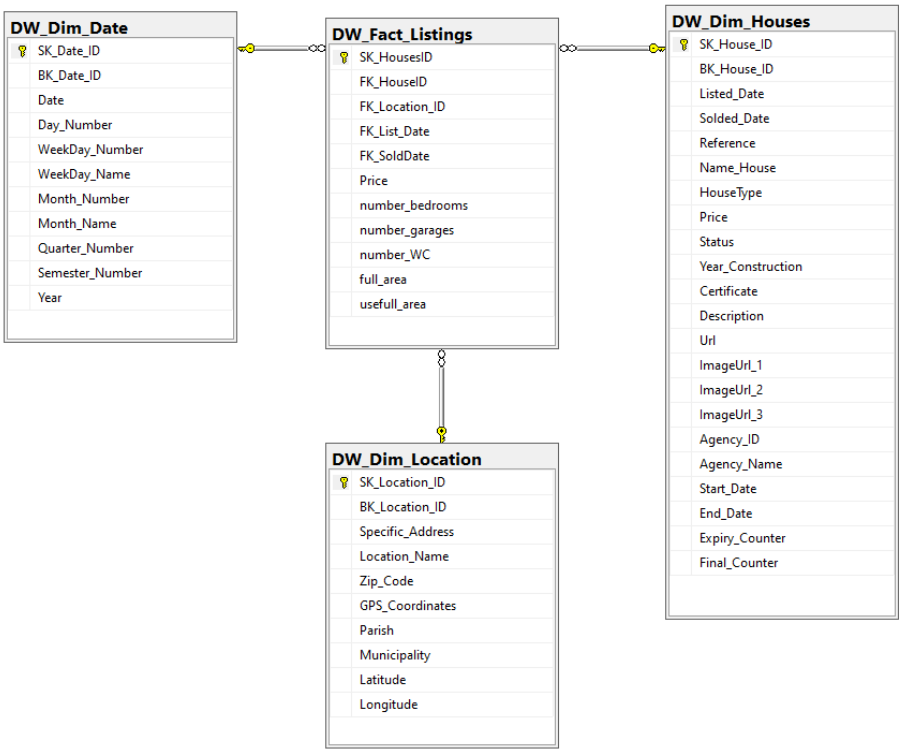
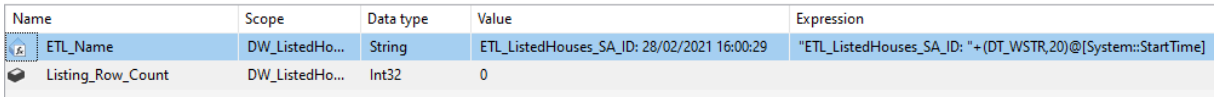


Figure 39 - Listed Houses Data Warehouse

The solution implemented in terms of ETL for the DW is also different although is follows the same structure used on the SA.

The chosen method to load the DW was incremental. This method means that the data will be loaded incrementally by running the ETL process in a periodical basis. No data should be deleted or reloaded in the DW. This is valid for both DW's dimension and facts tables. As seen before, all the data that is loaded is pretty much the same apart from updates on listings and new house listings that may be added. Other than that, all the loaded data will already within the DW, which implies that the dimension tables will remain almost the same in between ETL processes in terms of size. On the other hand, the Facts table will be growing rapidly. This is because the model is keeping track of the houses as they were an inventory. The same house can be for sale for several days implying that the house being in the market on one day is a fact, and the same house still being on the market on the following day is also a fact. As seen in the chapter four, Conceptual Model Proposal, and as Kimball and Ross in 2002 depicted, the fact table structure to achieve this is called periodic snapshot, meaning one row per period.

On figure 40 are the variables created within the ETL process for the DW:



Name	Scope	Data type	Value	Expression
ETL_Name	DW_ListedHo...	String	ETL_ListedHouses_SA_ID: 28/02/2021 16:00:29	"ETL_ListedHouses_SA_ID: "+(DT_WSTR,20)@[System::StartTime]
Listing_Row_Count	DW_ListedHo...	Int32	0	

Figure 40 - Data Warehouse Variable Creation

Listing_Row_Count will be used to obtain the rows that are added to the DW Facts table while ETL_Name maintains the job that had in the SA package.

Again, to continue with the ETL process, and in line with what was said earlier, the package runs without deleting any of the information previously loaded on DW. That is the reason why it starts directly in the sequence container named Load the DW Dimensions Tables.

8.6.2.1. LOADING THE DIMENSION TABLES

The connections created for the scope of the DW can be seen in the figure 41, taken from the project's connection manager:

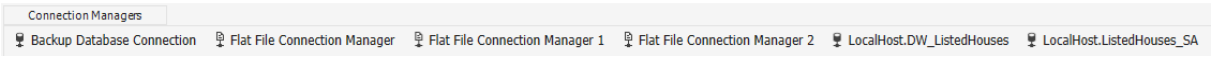


Figure 41 - Data Warehouse Connections

There are some flat file connection managers visible on figure 41 that were used to debug and help finding some errors preventing the DW to be successfully loaded. Those connections were proposedly left there to show that the process to create a robust ETL was a hit and miss and preventing that the

process fails to run due to error. It will also be useful if new error events uprise. Any error log can be followed through the logs of the ETL after the package successfully runs. As the data that may fail to follow the transformations and by consequence is not loaded into the model, it will be loaded into text files providing a much more detailed perception of the data points that were responsible for the error, leaving only to analyze and rectify whatever may be causing the error.

Advancing to the proper dimensions process, the Sequence Container on figure 42 was used:

Figure 42 - DW Dimensions Tables Load

The connection source that is being used inside the data flow for each dimension is LocalHost.ListedHouses_SA. Evidently as the DW is the database being loaded this time, it will be the destination connection.

To load each dimension table a similar method was used. The Houses dimension used a Slowly Changing Dimension (SCD) type 2, after defining the data source through the source connection. The configuration was defined to identify as business keys the composite house ID as well as the Agency ID. The SCD will parse all rows using these two attributes and will change, update or create historical attributes following the schema on figure 43:

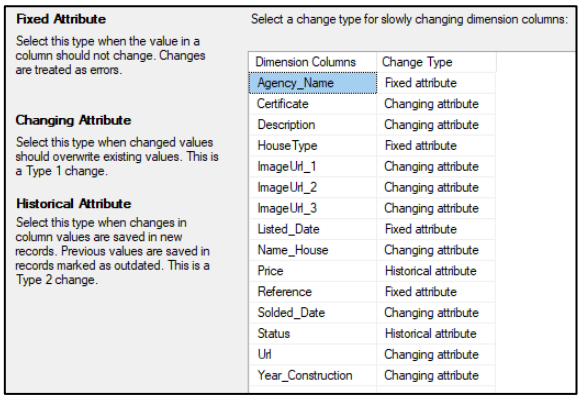


Figure 43 – Houses Slowly Changing Dimensions Configuration

As the Slowly Changing Dimension parses trough the data, the attributes Status and Price while being type 2, will save changes by ending older records and create new. The way of doing this is by adding a Start Date and an End Date to the dimension (figure 44). The records that have End Date equal to null are the current rows, while the records that contain a value on the End Date column will be historical.

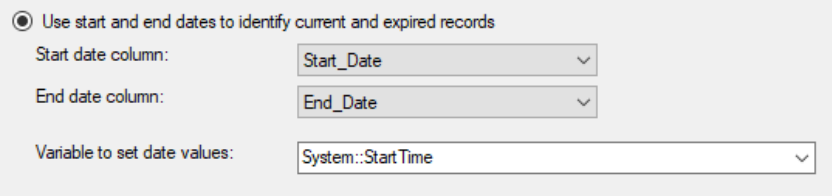


Figure 44 - Start Date and End Date variable set

To shorten, the variable that sets the date values of the SCD is the system start time. This output is afterwards converted to a date in the format (dd-mm-yyyy) with a derived column node, to be able capture the date when the record became historical while also maintaining the desired data format.

The Location dimension data flow process may be seen on figure 45:

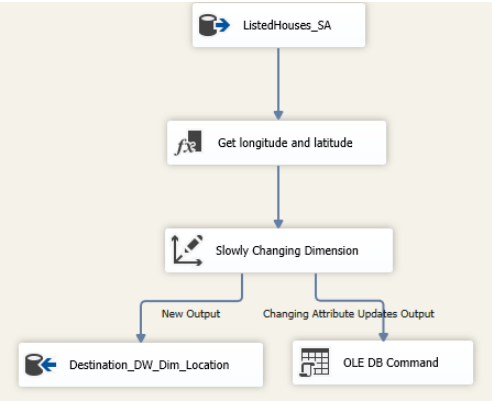


Figure 45 - Location Latitude and Longitude new columns

The derived column is separating the Gps coordinates attribute and transforming into two new columns, latitude and longitude. Then the slowly changing dimension carries out the work needed to parse the data rows and change all the needed attributes. The configuration of the SCD for location dimension used Specific_Adress attribute as a business key. The Specific_Adress is an attribute created on the transforming step of the data through the SA and it is the compound of the Zip Code and the name of the location of the listing. For the remaining configuration of the SCD, all the attributes were classified as changing attributes although is not really expected for them to be altered unless they are updated by the website, for instance, a listing that did not contain a Zip Code at start but after an update from the website the Zip Code can now be found.

The Date dimension being the simpler of the three dimensions, was also configured with an SCD. This configuration will parse the data using as business key the attribute Date. All the remaining attributes were defined as fixed. The calendar will then continue immutable until new rows are added.

8.6.2.2. LOADING THE FACTS TABLE

The data flow used to load the Facts Table DW_Facts_Listings can be seen in figure 46:

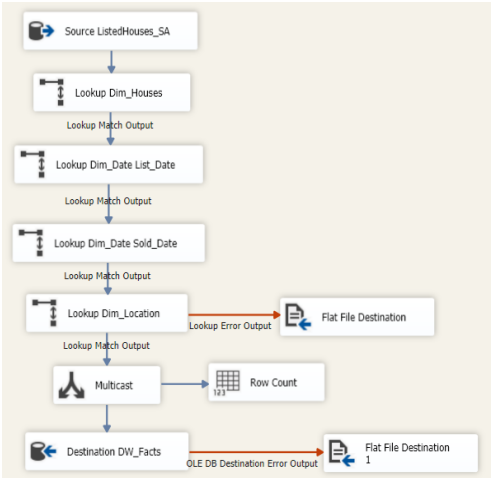


Figure 46 - Facts Table Loading Scheme

This data flow contains four lookup tasks and a multicast. The data source is the SA´s Fact table, Fact_SA_Listings, and the destination is the DW´s dw_Fact_Listings. Once again, flat file connections were left by the same reasons stated above on the dimensions chapter. A sequence of Lookup tasks was used to load the Facts table in order to establish a connection between the FK (Foreign Keys) of the Facts Table of the SA, the source, and the BK (Business Keys) from all the dimension tables of the DW.

The Lookup Dim_Houses task is calling the source DW_Listed_Houses dimension where the End_Date attribute is null. The reason behind this is that the lookup will only search for current data by the means of selecting the rows that have no End_Date.

Finally, and to serve the purpose of SK (Surrogate Key) the facts table will use the ticked check mark on lookup task as per figure 47:

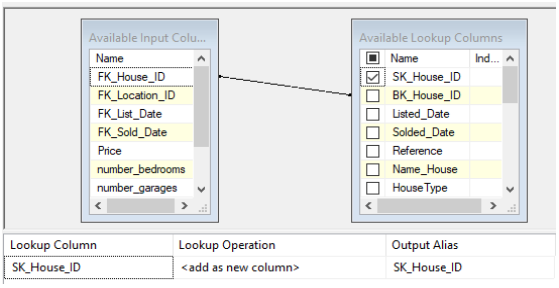


Figure 47 - SK Lookup Configuration

By adding as a new column with the alias Sk_House_ID the surrogate key will be automatically created by the system. All the consecutive lookups followed the same logic however, the sources will differ between each dimension.

That is the case of List_Date and Sold_Date FKs. The source for both attributes is DW_Dim_Date, which is the DW dimension for Date. By using the FKs for List Date and Sold Date from the SA and inner joining with the Dimension table of the Data Warehouse resourcing to the dimension tables BKs, the DW Facts table will be able to automatically obtain the SKs for each of the rows just as it was done with the Houses dimension.

For Location, the process is the same as described above, but this time the source used for the lookup task was the Dimension Location table from the DW. The linking of the SA Facts table FKs against Location dimension table BK's is done the same way. Also, the SK's will be automatically obtained using this method on both processes.

Using a Multicast after all the Lookup process in the data flow was a matter of also directing the outputs to a row count task while the process concludes by inserting the data into the destination DW_Facts_Listings. The row count task aggregated to one end of the multicast will give the many records changed and loaded during the process. All the process of the facts table ETL is also recorded using the logs introduced on the start of the ETL chapters.

8.6.3. ETL PROCESS CONCLUSION AND BACKUP

At the end of the ETL there is a log who is generated when the process is over. Also, a backup of the full DW database is done every time the process of loading it comes to an end.

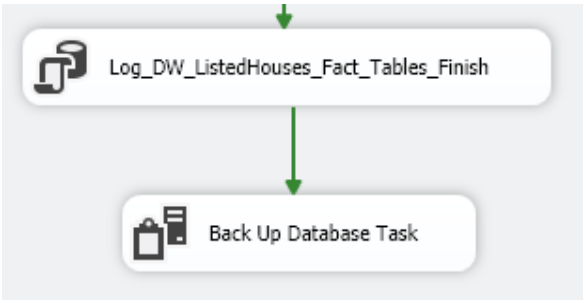


Figure 48 - Backup Database Task Creation

By employing a Back Up Database Task (figure 48) at the end of the process it is ensured that, for each iteration of the ETL process, the data is duly secured. To prevent major issues with data if the process goes wrong and the data is lost this task has the job of creating a backup of the DW. If something happens to the DW while running the ETL process it will be very easy to wipe the damaged DW and replace it with the last backup. This task allows that the .bak files that are created to be erased after 14 days preventing cluttering in the disk. The path leading to these backups is "G:\Work\ETL\ListedHouses_ETL\BackupDW".

8.6.4. LOGGING

Through the ETL process all the log files, were being stored and sent to the loggings table ListedHouses_SA_Log_ETL. This is valid for the SA and DW ETL process. This is important since it provides a possibility to control the ETL flow for the duration of the process. By querying the table ListedHouses_SA_Log_ETL on Microsoft SQL Server Management Studio the following table will display all flows represented on figure 49:

LogID	ETL_Name	ETL_Desc
1	ETL_ListedHouses_SA_ID: 30/06/2021 09:29:23	ETL start: loading Staging Area...
2	ETL_ListedHouses_SA_ID: 30/06/2021 09:29:23	Start of ETL Task: Truncate SA Tables
3	ETL_ListedHouses_SA_ID: 30/06/2021 09:29:23	SA Tables Truncated
4	ETL_ListedHouses_SA_ID: 30/06/2021 09:29:23	Files Loaded
5	ETL_ListedHouses_SA_ID: 30/06/2021 09:29:23	Start of ETL Task: Load Files
6	ETL_ListedHouses_SA_ID: 30/06/2021 09:29:23	SA Tables Loaded
7	ETL_ListedHouses_SA_ID: 30/06/2021 09:29:23	Start of ETL Task: Load SA Tables
8	ETL_ListedHouses_SA_ID: 30/06/2021 09:29:23	SA Fact Tables Loaded
9	ETL_ListedHouses_SA_ID: 30/06/2021 09:29:23	Start of ETL Task: Load SA Fact Tables
10	ETL_ListedHouses_DW_ID: 30/06/2021 09:32:01	DW Dim Tables Loaded
11	ETL_ListedHouses_DW_ID: 30/06/2021 09:32:01	Start of ETL Task: Load DW Dim Tables
12	ETL_ListedHouses_DW_ID: 30/06/2021 09:32:01	Listings updated:1
13	ETL_ListedHouses_DW_ID: 30/06/2021 09:32:01	DW Fact Table Loaded
14	ETL_ListedHouses_DW_ID: 30/06/2021 09:32:01	Start of ETL Task: Truncate DW Fact Table

Figure 49 - Log Tables Sample

The same happens when capturing possible errors that can possibly occur during the ETL process. By employing an *Execute Task* in the Events Handlers tab with an SQL statement loading the errors that can happen, will enable to get the origin of the error, the type and where it occurred during the ETL process.

Resourcing to the variable ETL_Name created in the beginning of the Visual Studio project once more it will be able to capture the name of the package as well as the time when the error occurred if any exists. Figure 50 shows the code used to make the task do its job.

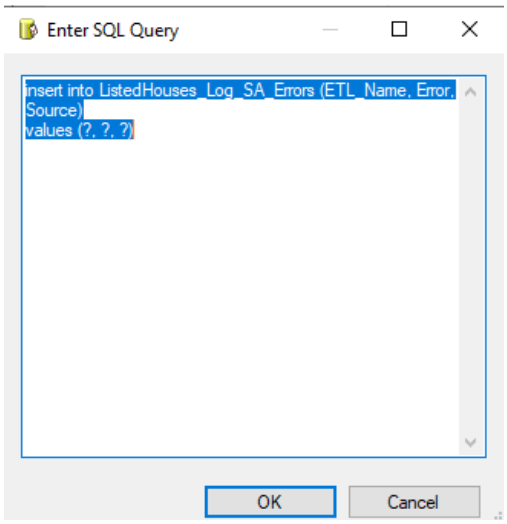


Figure 50 - Error Handler SQL Task Configuration

8.6.5. DEPLOY

In order to simplify the ETL deploy which is composed by the nodes of SA and DW's packages, it was also created a third package in Microsoft Visual Studio, where all the packages are called and are executed sequentially. With an Execute Package Task referencing the SA and other referencing the DW respectively, the process will execute in the desired order. To simplify and automate all the process of ETL, I was also necessary to create a catalog on Integration Services of Microsoft SQL Management Studio as shown on the figure 51:

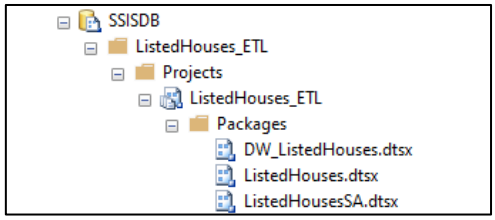


Figure 51 - SSIS Catalog Creation

Also, a Job was created (figure 52) to run two times a day, automatically, one in the morning and other at night, in order to run all the landing csv files containing the data gathered through the first step of the entire process.

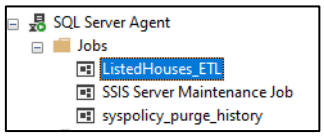


Figure 52 - SSIS Job Creation

This way the system will be auto sufficient and will run by itself since the beginning to the end of all process as seen on figure 53.

Schedule list:

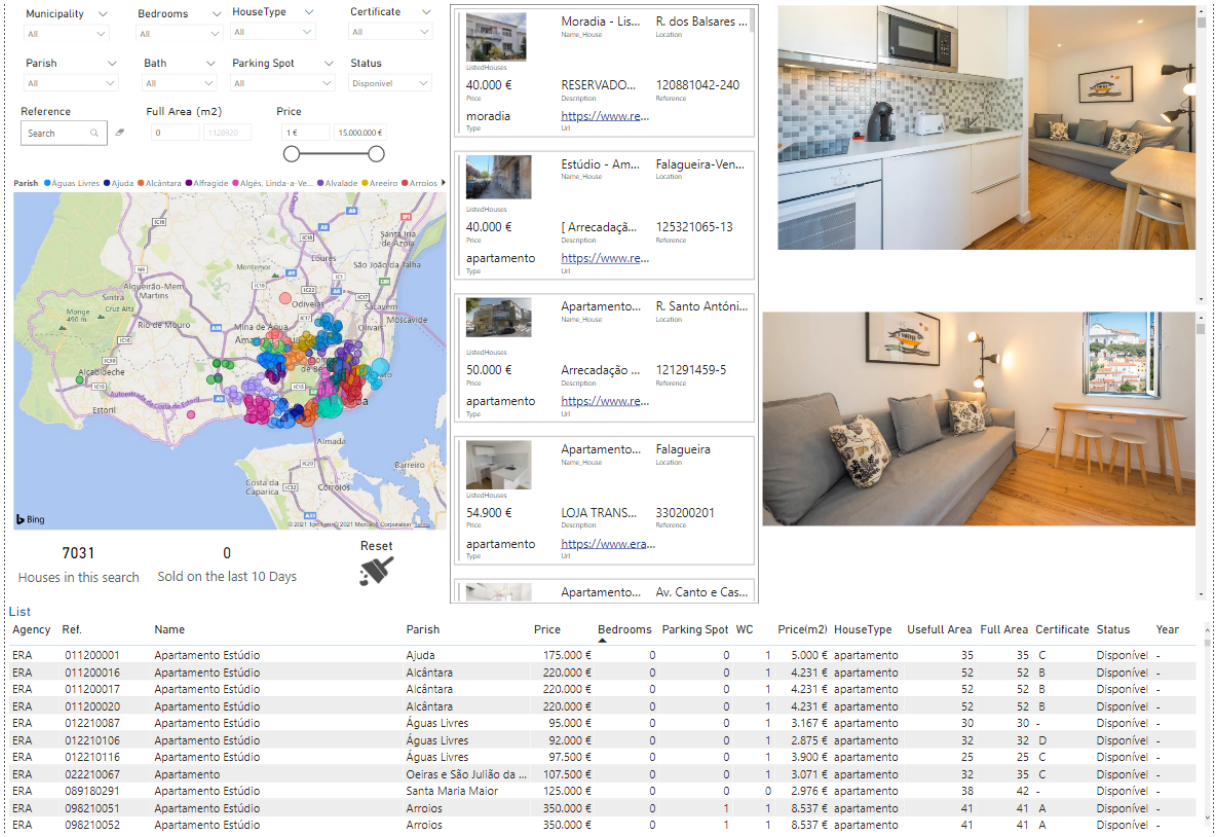
ID	Name	Enabled	Description
12	ETL Run Schedule	Yes	Occurs every day every 11 hour(s) between 10:00:00 and 23:59:59. Schedule will be used starting on 27/12/2020.

Figure 53 - ETL Job Schedule

There is some maintenance needed since there is always the chance that the data scraped from the websites may bring some new error cases that up until now weren´t present. If this happens the system will be hedged since there is already a backup from the last batch of data and, all the logs created during the process will specify when and how was the process interrupted.

8.7. DASHBOARD

The Map and List module (figure 54), as expected, presents a good overview of the distribution of houses for sale between the different areas of the Zone of Lisbon. It helps to identify and segmentate the information that the user is looking for, since it gathers all the data points of a determined area and distributes them through map. This way of presenting the information, already benefits the user since it provides a good idea of the available places and houses for sale on the city of Lisbon.



Agency	Ref.	Name	Parish	Price	Bedrooms	Parking Spot	WC	Price(m2)	HouseType	Usefull Area	Full Area	Certificate	Status	Year
ERA	011200001	Apartamento Estúdio	Ajuda	175.000 €	0	0	1	5.000 €	apartamento	35	35	C	Disponível	-
ERA	011200016	Apartamento Estúdio	Alcântara	220.000 €	0	0	1	4.231 €	apartamento	52	52	B	Disponível	-
ERA	011200017	Apartamento Estúdio	Alcântara	220.000 €	0	0	1	4.231 €	apartamento	52	52	B	Disponível	-
ERA	011200020	Apartamento Estúdio	Alcântara	220.000 €	0	0	1	4.231 €	apartamento	52	52	B	Disponível	-
ERA	012210087	Apartamento Estúdio	Águas Livres	95.000 €	0	0	1	3.167 €	apartamento	30	30	-	Disponível	-
ERA	012210106	Apartamento Estúdio	Águas Livres	92.000 €	0	0	1	2.875 €	apartamento	32	32	D	Disponível	-
ERA	012210116	Apartamento Estúdio	Águas Livres	97.500 €	0	0	1	3.900 €	apartamento	25	25	C	Disponível	-
ERA	022210067	Apartamento	Oeiras e São Julião da ...	107.500 €	0	0	1	3.071 €	apartamento	32	35	C	Disponível	-
ERA	089180291	Apartamento Estúdio	Santa Maria Maior	125.000 €	0	0	0	2.976 €	apartamento	38	42	-	Disponível	-
ERA	098210051	Apartamento Estúdio	Arroios	350.000 €	0	1	1	8.537 €	apartamento	41	41	A	Disponível	-
ERA	098210052	Apartamento Estúdio	Arroios	350.000 €	0	1	1	8.537 €	apartamento	41	41	A	Disponível	-

Figure 54 - Final Report - Map and List Page

According to the number of houses available for each zone, and by using GPS coordinates with Parish attributes, the Map automatically clusters the house samples and sizes each area with a larger indicator when more houses are included in a determined spot and a point, smaller, if only one house is present on the following location as seen on figure 55.

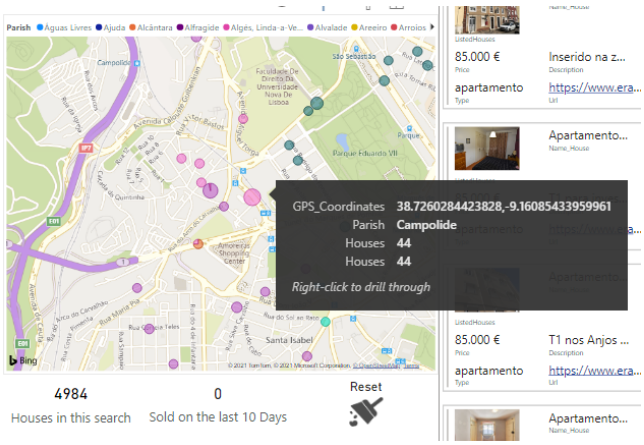


Figure 55 - Map Visualization and Tooltip Configuration

Also, as all the included visualizations are interactive with each other, by utilizing either of them, it gives agility to the user to not only analyze the information from the geolocation point of view but also

to quickly have a glance on the houses and its innate information as well as their images. Quickly the user can start to get important insights about the houses and compare them. Since the websites that provided the data are Portuguese, the values for attributes such as Housetype (Moradia, Apartamento) and Status (Disponível, Reservado, Vendido).

There are 2 measures and 1 metric on this page, given respectively by the Houses on this search, which was created to assess how many houses are in the designated criteria, Houses Sold in the last 10 days, and Price Square Meter which comes from a calculated column between the divide of Price and Full Area. By presenting the information in one page the user can quickly navigate between these three ways to search for the desired houses while doing it in a report which gathers the information from three different websites that capture the majority extent of Lisbon's housing market as we studied on the chapters before.

By selecting the Status "Vendido", users will discover how many houses were sold that are still in the model, meaning that are still visible on the agencies websites and do not have expired. Although it could be very interesting to focus the report on the Properties Sold analysis, this report was not created having that in mind. This is partly because this work focus more on the current data since the gathered data is not sufficiently historic yet to perform such an extended analysis on the past. However, it can be very interesting to check the type of houses and were and which properties have been sold in a recent or not so recent past. Several insights can be taken by using historical data such as where have more houses been sold, which typology is more wanted, and the prices practiced on those sold houses. These prices may or may not be accurate since negotiation can occur, but it gives already a good perspective on the market flow. This type of summarized module on the sold properties is not available in any website currently live for the Portuguese market and demonstrates the power that this model can contain in the future. This report has been predefined to show information of houses with price above 1 €. This is mainly because several websites use "Preço Sob Consulta" instead of advertising the real price. This model captures those houses and can display them in the map also, but they were not considered when calculating any measures. That said, if the user still wants to see them represented, he can. This page also enables direct access to the agency website directly from the URL of the house.

One of the major opportunities given by this page is to act as drill through page for the second page of this report. This feature is done using Power BI's capabilities to drill through, and the process is done as per figure 56:

Figure 56 - Drill Through to House View

This takes us to the second page of this report, “House View”. This report acts in a sort of storytelling of the searched houses on the Map and List report.

Resuming, after searching for a house that is potentially interesting, all that the user needs to do is to right click and follow the drill trough, leading to the “View House” page (figure 57).

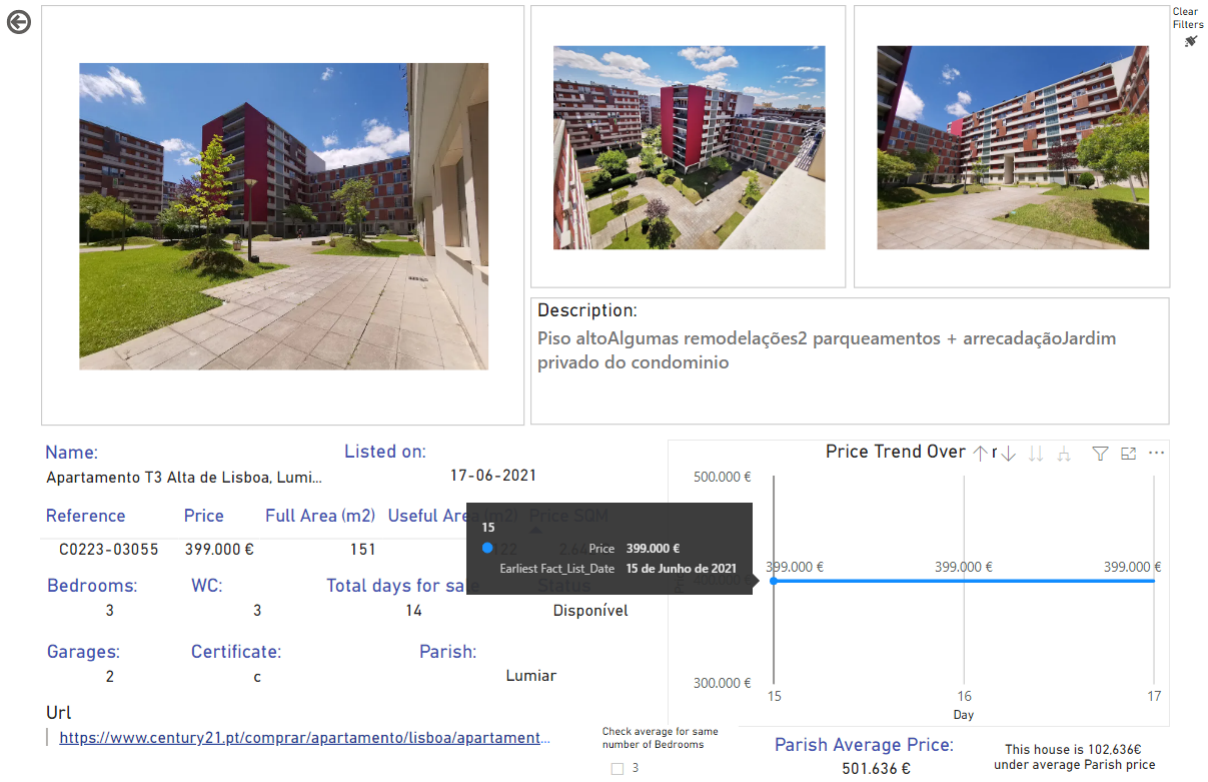


Figure 57 - Final Report - View House page

This report summarizes all the attributes of the selected house and was created to give a more in-depth perspective of a specific property. In this page the user can easily read the description of the house while looking through the available pictures of it. This page was also intended to provide information that is relevant to someone who is really interested in getting more information from this specific house and that is why that, beyond the usual attributes, there are present other important attributes and measures like the List Date, provided by the “Listed on” visualization and the “Total days

for sale" that calculates the number of days that the specific house is on the market. More importantly and because it is the main goal of this project, this page can provide several relevant information to be used to compare this house with the rest of the market. On figure 57 the house is listed on the market for 399 000 €. Immediately, the graph is already showing that the house price has maintained itself stable for the three days, being the first available data of this house 15 of June of this year and the last 17 of June of the same year. Should the price change over time, the graph will register those variations. The user can also drill down or up concerning the desired time frame (Years, Months, Days) by using the arrows on the top right of the graph and use it to check for oscillations in price max of two years back providing the desired interactivity. A recommendation for the future would be to create a metric to calculate the minimum and the maximum price that this house was ever listed.

In one look, the user can also compare the current price of the house with the average price for parish. In the picture the average price for parish is 501 636 €. The page automatically displays if the price of the house is under or above the average. The metric "Parish Average Price" highlights the price concerning all houses inserted in the same parish. The average price a very good indicator of the overall perspective of the market in a determined location which is handy for an immediate global perspective, however, it is not the best metric to be used in an individual comparison. Therefore, the price of this house, should be compared to average price of houses with the same typology explaining the creation of the filter button "Check average for the same number of bedrooms" (figure 58).

Figure 58 - Average Values for same Typology

By checking the box, the values presented will be automatically adjusted to the average price of the same typology, thus presenting a much more realistic summary. In the graphic above the average price for parish increases to 517 230 € meaning that the average price for houses with three bedrooms in this special case is higher than the complete average for this parish. Also, the metric that calculates the difference between the two prices (Current - Average) changes, suggesting that the analyzed house inserted in Lumiar parish with three bedrooms is indeed 118 230 € under the average price for this parish.

The third page of the report is “Overview Dashboard” and can be found on figure 61.

In this page there is a broad but drillable perspective of the real estate sector for Lisbon. This page reflects the market current situation and presents several measures and metrics to help getting insights of it.

Firstly, we have the measure of the “Housing Mkt Value”. It compounds the current global business volume for all the four proposed municipalities. In the image, at current date, the user could easily discover that the real estate sector business volume is approximately 2,3 billion €. This value will be automatically adjusted the more deep the drill is done.

For instance, when segmenting the data to the area of Lisbon municipality this value drops to 1,8358 billion €. Hence, 80% of the total housing market value resides in the municipality of Lisbon. This value is followed by the municipality of Amadora with 176,238 million €, Odivelas with 145,180 million € and Oeiras with 136,307 million € (figure 59). This can mean two things: either the municipality of Lisbon has highest number of houses for sale or the most expensive houses. The first condition is already met since the indicator “Houses in this search” is also available on this page. When comparing the number of houses between the four municipalities at current date in first place comes again Lisbon, with 3360 houses available, against 838 in Amadora, 500 from Odivelas and 284 from Oeiras municipalities. When consulting the drillable visualization “Average Price by Zone”, the second condition is also validated because it shows that the houses in Lisbon are more expensive with an average of 571,4m €. Right behind comes Oeiras with an average of 516,7m €, in third Odivelas 277,6m € and last is Amadora with 230,1€.

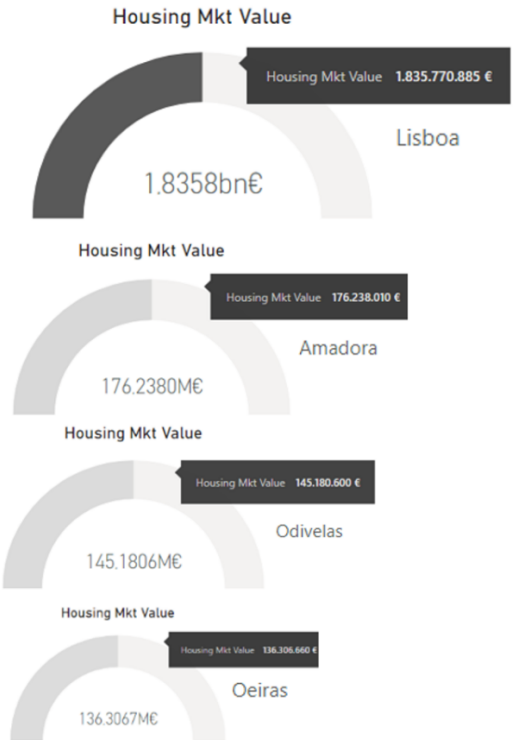


Figure 59 - Municipalities Market Share

When comparing Odivelas and Oeiras, Odivelas has approximately more 43% of available houses than Oeiras, but the Market Values do not differ as much in proportion, 145 million of Odivelas against 136 million against Oeiras. This low difference in Market Value for both, is explained by the fact that the municipality of Oeiras, containing five parishes against four parishes of Odivelas, has the average price of 516,7m €, against 277,6m € of Odivelas, once again available on the bar chart Average Price by Area (prints em anexo). Ultimately, Oeiras has less houses available but is more expensive than Odivelas. Also, it is important to notice that Oeiras contains, in percentage, more dwells (Moradias) available for sale than Odivelas, about 16,2% of dwells against 5%. By highlighting dwells on the "Housetype" pie chart, it is observed that the gap of prices between these two municipalities is explained by the high price of the dwells. These are much higher priced above the average that the apartments for the same area as it is visible on figure 60.

Figure 60 - Dwells versus Apartments, Distribution and Average Price

When observing the "Average Price by Zone", filtered to the parish granularity only, it is observed that Belém is the most expensive parish with an average of 976,3m €. It is not a coincidence that the following higher priced parishes are from Lisbon municipality too. As observed before this is the most expensive municipality at the time. In fact, that does not mean that the best or worst business, in terms of price exclusively, is given by the highest or lowest average price. When simultaneoulsy analyzing the bar chart of "Average Price(m2)" and "Average Price by Zone", it is observed that even though Belém can be the most expensive zone, it is not the zone that charges more euros for a single square meter. That is the case of Alcântara that prices 6737€ in average for a square meter. This may be an indicator that that although Belém is more expensive, tendentially the houses in Alcântara are smaller in size

compared to Belém. This is verified by the scattered chart of "Distribution of Price by Area (m^2)". By observing the picture below, containing the distributions for both parishes, Alcântara and Belém, it is noticeable that the majority of the points is located in different places in the graphic between them. In one hand Alcântara has most of its data points situated in the left side, between 50 and 80 square meters of area, and has a lower scale for size in meter when comparing to Belém (figure 62). Also the houses for Belém are more dispersed to the right side of the chart. When looking at two points with the same approximate area for both Parishes (70/71 m^2), it is noticeable that whilst the prices for Belém are lower, the majority of the houses are much larger in size, since contrarily to Alcântara, it has more houses concentrated between the interval of 100 m^2 to 200 m^2.

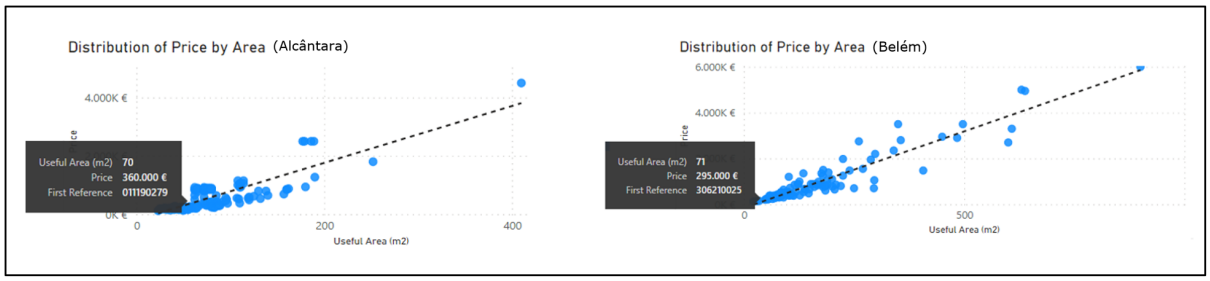


Figure 61 – Overview Dashboard - Comparison of Price by Area Distribution between Alcântara and Belém

These graphics were easily obtained using the available segmentation slicers for the field values of Municipality, Parish and Bedrooms. The bedrooms segmentation was not used in this analysis but it can be used in other analysis such as the one performed on the "House View" page, in order to limit the conclusions to a more specific typology of houses.

This dashboard also includes time based graphs such as "Price Trend Over Time" and "Square Meter Price Trend". The first analyzes the trend of prices, that can be drillable. Supposing that there is no filtering or slicing of the data, the graph will ensemble the trend over time for the global market, meaning that it will showcase the oscilations between all four municipalities. This graph will obviously vary with the number of houses available in the market as well as their prices. For instance, if more houses are sold than houses are added to the model this metric will oscillate. But will also be dependant of the prices of the houses that were sold or added to the model. In any way it will always be a good indicator of the trend of the market globally or individually since it will register the variations in price that can be drilled by parish, municipality and by typology.

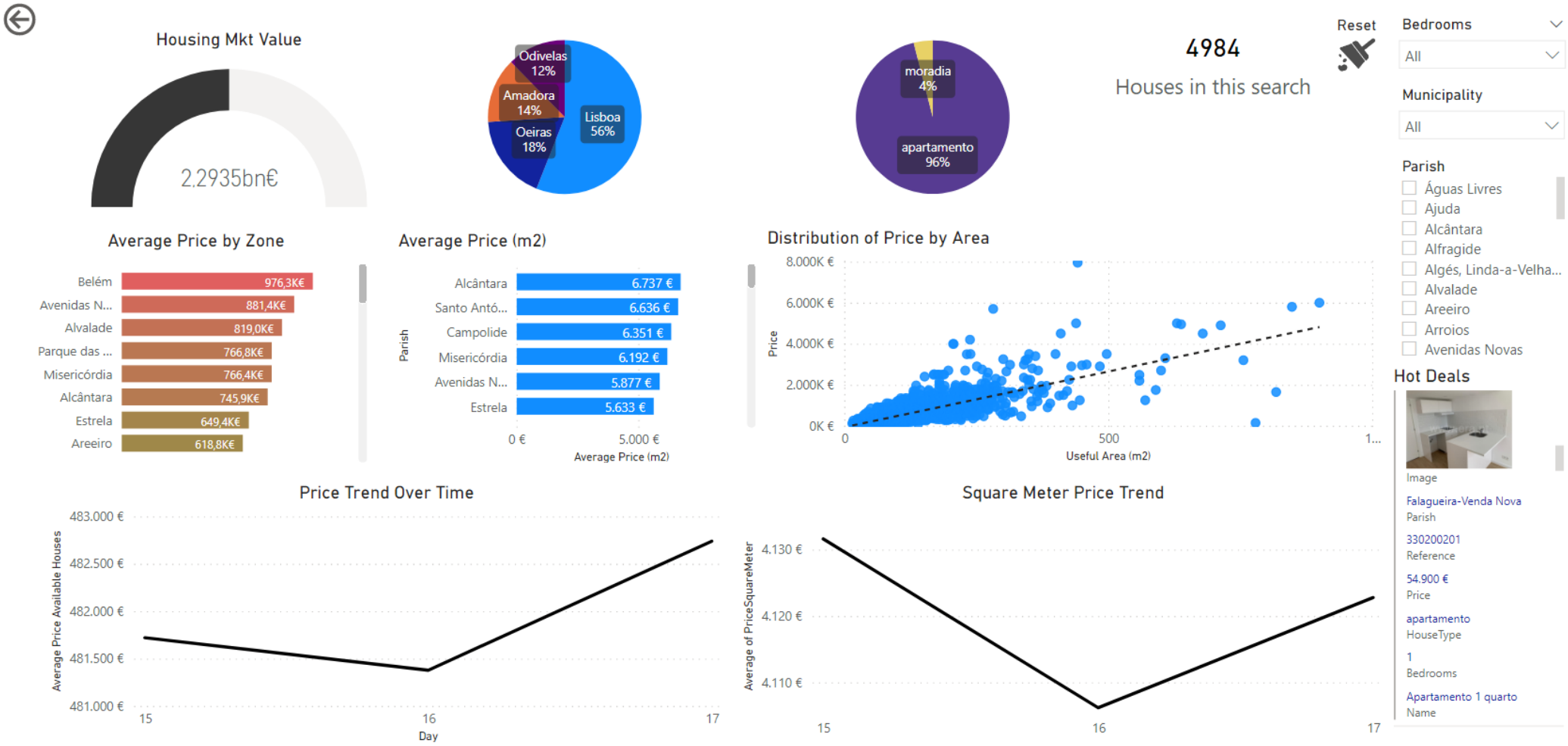


Figure 62 - Final Report - Overview Dashboard

For instance, if the user wants to know how is Benfica's parish average price trend by the typology of houses that have two bedrooms, he can use the data segmentation to filter the results. After that, it is instantly provided that between the dates of 15 and 17 of June of 2021 the average price has been in a positive trend for the parish of Benfica (even that only three days are being analyzed due to the fact that the data does not have yet historical importance). As visible on figure 63 , the price has increased from approximately 260m €

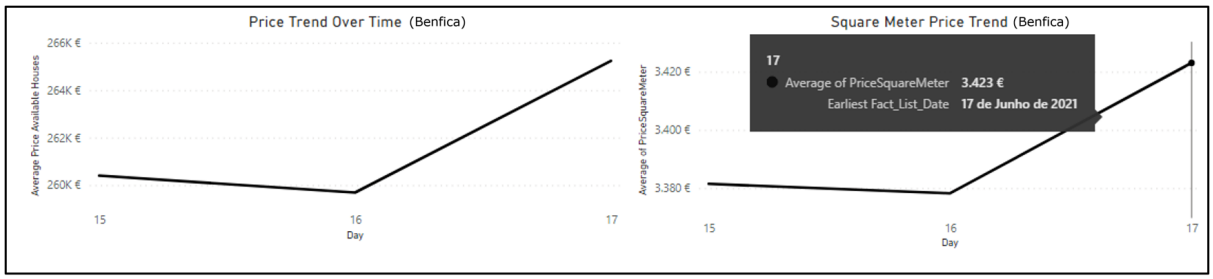


Figure 63 - Overview Dashboard - Price Trend Analysis

to 265m €. Or the general house pricing has increased or there are more houses available on the market for Benfica's parish. Also on the Square Meter Price trend graph, it is observed that the price has also increased leading to the fact that, the new houses added to the model are more expensive that the houses that already were incorporated, concerning the size in meters of all. This is may be related with the behaviour of recently listed houses that are usually listed it above the price that are really worth. The seller wants to have a margin to negotiate the price and to lower it if the house remains too much time without any real interested people. This metrics are really important because they will track down the historical trends of the direct relation between the price and the remaining attributes of the houses. Lastly, to this page it was added a small summary that gathers the top ten smallest priced houses. This summary was called 'Hot Deals' since it will display the lowest bottom ten houses in terms of price. This summary will obviously vary with the applied data segmentations available. For the creation of this summary a multi-row card was added similarly to what had been done on the first page of this report, "Map and List". In this page it was also added a reset button to use to deafult all values possibilitate a new quick search.

On figure 64 is visible the "Price Prediction" module:

Figure 64 - Final Report - Price Prediction page

“Price Prediction” can be one of the most interesting pages of this report. It is composed by a predicting model adequate for house pricing. Although there exist better house pricing models than the one incorporated in this report, the main goal in this project is to create a funcional, realistic and fit model working towards to concept that is to be proven. To create a new house pricing model would be in a entirely different scope and not within the range of this project. In this sense, and by the knowledge acquired regarding the fitness of Linear Regression on house pricing, it was decided to incorporate it in a full page.

The idea behind is to give the user the ability to choose between several housing attributes such as Location from Municipality to Parish while using categories like Bedrooms, Wc and Garages to help describing a specific house. By using the slider or manually inserting the desired value in the field for the area in m^2, the model wil be able to provide an expected price concerning all the available houses for sale at the current date. For analysis purposes the formula and components were purposly left on the page (Slope, Constant Interception). The report is utterly useful if the user has an ideia of the type of house that he desires to look for. But can also be very useful to obtain a general ideia if a conception is yet to be form regarding the market value of a hipotetical house. Equally to the “Overview Dashboard” this page performs a very good job in showing the distribution of prices concerning the size in m^2 for the houses displayed.

One of the deficits of this model is that it cannot accuratly predict the price for houses that are inexistent in model or, in other words, that there is not any similar house available . For instance a house with two bathrooms, three bedrooms and 45 m^2 will not have an accurate pricing. However it is more than likely that such a house does not exist given that this are some irrealistic values. More importantly the model will work as much better as more data points exist in the database, meaning that the highest the sample of houses with the specific attributes is, the more the accuracy will have this report. The contrary is also true. The model will not allow to predict any price for a house with a given attribute, unless there exist in the database a house with the same attribute searched. For instance, the price for a house with zero bathrooms is not possible to predict since in the model there is no house with zero bathrooms.

Has it was observed in the Linear Regression chapter, there are other types of factors that weigh much on the house pricing equation like the condition of the house itself, but it is not possible with the given dataset to bring those aspects into the model.

So, if a user would want to know how much would be spending in a house with 3 bedrooms and approximately 120 m^2 in size in the municipality of Lisbon, he would find that would need to spend about 664 267€. If he wanted to limit that choice to the parish of Avenidas Novas, the price would

increase to 715 824 €. If the user was still deciding wether he wanted two or three bathrooms he could quickly check that he would pay more if he picked three instead of two, 703 706€ and 677 630€ respectively.

As usual a reset button was added to facilitate the task of defaulting all values, adding a touch of usefullness.

9. RESULTS AND DISCUSSION

This chapter attends to present and to evaluate the results for this project. It is in this chapter that the outputs of the complete work are analyzed and if the project meets the expectations. If not, it means that there are more iterations to improve the current model.

9.1. EVALUATION OF THE RESULTS

The conceptual model of a real estate sector was pictured, designed and developed according to the literature research available and found, trying to reach the best solution for the proposed problem.

Through public data, that in fact is limited by its non-centricity and by its banal use on the existent agencies websites in the Portuguese real estate sector, it was developed a model that can capture several implicit knowledges while also easily sharing these within a properly built dashboard.

This dashboard allows a direct access to the data and benefits from the so needed interactivity in order to deliver easy exploitation to its users. The report is built with information to satisfy the specific goals proposed on the beginning of this work and delivers clear insights about the real estate sector in an informed, clear and organized way. It has created ways to analyze and support the real estate users to explore and understand the data.

It conserves important aspects and features that the users already are accustomed when searching for information on the real estate market, while bringing new challenges on the way that the used metrics were employed, and the measures incorporated through the different pages that compose the complete report.

It is considered that the entire report has a good quantity of interactions that can be used to explore the data, contributing to the generation of knowledge on real estate sector of Lisbon. Also, it does it in a very concise, summarized and clear way, turning the overall appreciation, convenient and user-friendly way to navigate.

This tool is perceived to increase the awareness and search efficiency of the market for its users due to all the reason stated above.

Additionally, it was presented as a scalable and expandable solution, meaning that the only requirement is to insert additional data following the correct structure. This solution was not intended to be a one-time analysis, instead, this solution achieved a level of automation and continuity, which

means that this model can be continuously updated, receiving, if necessary, data in an automated way and in a daily basis if desired.

More importantly, this model was designed to store data and conserve it through time since the created dimensional model allows to keep records of outdated lines. Having created a process to generate and store important historical data is a great achievement since it will provide the advantage, in the future, to also focus the analysis on the factual past. As soon as time passes the outdated data will share its implicit secrets and more analysis can be driven.

Regarding the three main questions asked in the beginning of this project, the results obtained answer positively to the first two questions, confirming that it is possible to use BI to create an interactive, convenient and user-friendly model to the city of Lisbon, while improving awareness and search efficiency from a user's perspective. The last question as studied on the literature, the model will not suffice to substitute entirely the role of an agent, as these agents' product is in fact a bundled good. This bundled good is in form of services but also in knowledge and expertise of the market. Also, real estate stakeholders want to use the social networks that these housing intermediaries acquired during their years of experience working inside this market.

Following the DSR approach methodology it is considered that the artifact developed during this project is suited for the proposed problem and its resolution, meaning that other developments won´t be taken.

10. CONCLUSIONS

The artifact is the result of the creation of a full stack ICT process developed within a DSR approach. The development of this process led to the creation of a summarized and brough information sharing benefits. This tool serves as part of a solution, for Lisbon´s real estate market, which is to reduce information asymmetry, improve market awareness, increase search efficiency and reduce costs, while providing an understandable, readable and interactive way of doing so. This tool is eventually one of the best ways to quickly get awareness of the reality of Lisbon housing market since it organizes the available information concisely and in a way that is not seen yet available in traditional websites. Not only provides those possibilities but also invites its users to do explore the data in several different ways. In this report it is possible to search for houses through a very familiar way, since it was designed a way for the users to control it as they were navigating through a website. It brings also a more detailed and specific housing report merging all the relevant information in one place.

With this tool, the users can quickly gain knowledge of the market and use the artifact to compare real estate realities, looking for housing opportunities and evaluate the trends in an endless number of ways. By drilling through the housing attributes like the number of bedrooms or the number of WC and use the data segmentation provided users can compare the data from a municipal level to a more specific location level. While exploring this tool the user will have a very easy task of formulating insights about the market, not only but also as a whole, and by using the various attributes available, the user can get detailed answers regarding the quantity, type and value of the houses inserted. To actively do this kind of analysis nowadays is difficult, given that there are current available tools on the traditional agencies websites that can sustain this type of conclusions. That provided, it makes this type of data visualization a strength of this project.

Even though this artifact was developed to target only three agencies and four municipalities, it is prepared to receive more data from multiple sources.

This project has been an enriching experience. It allowed to deepen the knowledge on a large scope of data science tools as well as to learn several skills that had never been explored.

The final balance concludes that the created artifact accomplished the proposed goals and that it was a success. This artifact is the proof that although there are currently amazing tools to explore different kinds of markets or businesses, there is always something that can be improved.

11. LIMITATIONS AND RECOMENDATIONS

Following, are presented the limitations and difficulties occurred during the development of this project:

- The process of gathering of data was challenging since there was no idea how to get the data at the start of the project.
- Due to the disparity of the data between the agencies websites there was an enduring process of trial and error for each website, until the data was being collected correctly. Due to that it was not possible to have more historical data since the oldest data sources did not meet the requirements to fit the model.
- The lack of a specific remote web server to run the Python scripts makes necessary to have a computer working at night dedicated to it.
- There is the risk of failure when running more than one script at the same time. This is due to the limited frequency of requests to the Geopy and Google maps libraries.
- Some of the scripts (Remax) used APIs from Google Cloud's service to provide geodata. This service is free if the requests done do not exceed the requests allowed with a free account. There was one month that the limit was largely exceeded, due to extensive script running, leading to a heavy invoice.
- The data gathering is heavily dependent on the structure of the websites. It happens that during the development of this work that, Remax completely changed their website and another program had to be written from scratch.
- The data is in some ways inserted manually by the real estate agencies. Sometimes it can assume unexpected and unpredictable values. In this sense in this model's level of development it is necessary maintenance to assure that the integrated data is not disrupting the report.
- The data can be integrated all together if the list date attribute of the data sources for all agencies is the same. If not, the integration needs to be done in daily batches.
- As a periodic facts table there is some concern to the storing of data through the time if the model runs in a daily basis.
- The expired or delisted houses will have a higher delay to be categorized as such.
- The real sold dates of sold houses will only be settled with time since the sold date being used now is the list date. When the model reaches its maturity, the sold dates will start to become more realistic.

- The map visualization may exhibit some incorrect location data points in terms of geography. This has to do with the tool itself since it may display an incorrect placing of the given coordinates.
- The map visualization has a limited number of data points. If all parishes all select it will not display all the values.

After the development of this project, the recommendations for future works are:

- Presenting more historical visualizations and timed based comparisons.
- Presenting a specified report on the sold houses through average days in market and other measures such houses sold by location, typology and house type.
- More detailed descriptive and predictive analytics to improve synergies within the already given data enabling a faster perception.
- The introduction of a ranking score for the location. By accessing and integrating demographic data such as the occurrences and intervention requests of the Portuguese portal “Na Minha Rua” it would be interesting to create a ranking score of all the parishes. With this it would be expectable that it was provided an incredible number of insights in terms of life quality of the surrounding area searched. This would unquestionably be acting as one of the most complete and complementing information that could be added to this data model since it could label the different ranks through color visualization within the already created map.
- The creation of a population density around the location. This could be possible by accessing the Portuguese “Censos 2021” demographic information to generate a populational density rank. This would add a great layer of perspective in population density to this model by also showing through the map the different densities between zones.

12. BIBLIOGRAPHY

Abela, A (2009). Chart Suggestions – A Thought-Starter, (\), 2009.

Accenture. (2016). Fintech and the evolving landscape: landing points for the industry. Available in: https://www.accenture.com/t20160427T053810__w__/us-en/_acnmedia/PDF-15/Accenture-Fintech-Evolving-Landscape.pdf#zoom=50.

Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In Uncertainty in economics (pp. 235-251). Academic Press.

As 4 Maiores Agências Imobiliárias em Portugal, retrieved on November of 2020, from the website: https://revistaimobiliaria.com.pt/agencias-imobiliarias/

Bangdiwala, S. I. (2018). Regression: simple linear. International journal of injury control and safety promotion, 25(1), 113-115.

Baum, A. (2017). PropTech 3.0: the future of real estate.

Blair, W. (2015). Financial Technology Sector – First Quarter 2015 Update. 42.

Bourassa, S. C., Hoesli, M., & Sun, J. (2004). What's in a view? *Environment and Planning A*, *36*(8), 1427-1450.

Chen, M., Ebert, D., Hagen, H., Laramee, R. S., Van Liere, R., Ma, K. L., ... & Silver, D. (2008). Data, information, and knowledge in visualization. IEEE computer graphics and applications, 29(1), 12-19.

Farber, S. (1998). Undesirable facilities and property values: a summary of empirical studies. *Ecological Economics*, *24*(1), 1-14.

Feth, M., & Gruneberg, H. (2018). Proptech-The Real Estate Industry in Transition. *Available at SSRN 3134378*.

Few, S. (2005). Dashboard design: Beyond meters, gauges, and traffic lights. Business Intelligence Journal, 10(1), 18-24.

Few, S., & Edge, P. (2007). Dashboard confusion revisited. Perceptual Edge, 1-6.

Fu, Y., & Ng, L. K. (2001). Market efficiency and return statistics: Evidence from real estate and stock markets using a present-value approach. Real Estate Economics, 29(2), 227-250.

Guerra, I. (2011). As políticas de habitação em Portugal: à procura de novos caminhos. *Cidades, Comunidades e Territórios*, (22).

Kimball, R., & Ross, M. (2002). The data warehouse toolkit: The complete guide to dimensional modeling. New York: Wiley.

KPMG (2019). Property Lending Barometer 2019: A survey of banks on the prospects for real estate sector lending in Europe. Available in: KPMG (2019). Property Lending Barometer 2019 : A survey of banks on the prospects for real estatesector lending in Europe. Available in: https://assets.kpmg/content/dam/kpmg/nl/pdf/2019/sector/property-lending-barometer-2019.pdf.

Levitt, S. D., & Syverson, C. (2008). Market distortions when agents are better informed: The value of information in real estate transactions. The Review of Economics and Statistics, 90(4), 599-611.

Lourenço, R. F., & Rodrigues, P. M. (2017). Preços da habitação em Portugal-uma análise pós-crise.

Mitchell, R. (2018). Web scraping with Python: Collecting more data from the modern web. " O'Reilly Media, Inc.".

Muther, C. (2013). The growing culture of impatience, where instant gratification makes us crave more instant gratification. Available in: https://www.bostonglobe.com/lifestyle/style/2013/02/01/the-growing-culture-impatience-where-instant-gratification-makes-crave-more-instant-gratification/q8tWDNGeJB2mm45fQxtTQP/story.html.

Perkel, J. M. (2018). Why Jupyter is data scientists' computational notebook of choice. Nature, 563(7732), 145-147.

Predicting House Prices with Linear Regression | Machine Learning from Scratch (Part II). Retrieved on March 12 of 2021, from the website: https://towardsdatascience.com/predicting-house-prices-with-linear-regression-machine-learning-from-scratch-part-ii-47a0238aeac1.

Python Try Except. Retrieved on January 25 of 2021, from the website https://www.w3schools.com/python/python_try_except.asp.

Rivard, K., & Cogswell, D. (2004). Are you drowning in BI reports? Using analytical dashboards to cut through the clutter. Information Management, 14(4), 26.

Savills World Research (2018). 8 things to know about global real estate value. Available in: https://www.savills.com/impacts/market-trends/8-things-you-need-to-know-about-the-value-of-global-real-estate.html.

Savills. (2017). Around the World in Dollars and Cents. Disponível em: https://pdf.euro.savills.co.uk/global-research/around-the-world-in-dollars-and-cents-2016.pdf.

SCHETTINO, T. (2006). Inclusão social e “assimetria de informação” no sistema de ensino superior brasileiro: uma análise comparativa. Universidade Federal do Rio de Janeiro–UFRJ. Dissertação de Mestrado em Sociologia com concentração em Antropologia.

Schwab, K. (2017). *The fourth industrial revolution*. Currency.

Seixas, J., Tulumello, S., & Allegretti, G. (2019). Lisboa em transição profunda e desequilibrada. Habitação, imobiliário e política urbana no sul da Europa e na era digital. *Cadernos Metrópole.*, *21*(44), 221-251.

Stiglitz, J. E. (1975). The theory of" screening," education, and the distribution of income. The American economic review, 65(3), 283-300.

Tavares, F. O., Moreira, A. C., & Pereira, E. T. (2010). Avaliação imobiliária sob a perspectiva das externalidades: uma revisão da literatura.

Tavares, F. (2011). Avaliação imobiliária: entre a ciência da avaliação e a arte da apreciação (Doctoral dissertation, Tese de Doutoramento, Universidade de Aveiro, DEGEI).

Tavares, F. O., & Pacheco, L. M. (2015). Fatores determinantes na escolha de apartamentos: estudo empírico em Portugal.

Tavares, F. O., Moreira, A. C., & Pereira, E. T. (2013). Assimetria de informação no mercado imobiliário em Portugal.

White, S. A. (2004). Introduction to BPMN. Ibm Cooperation, 2(0), 0.

Wong, S. K., Yiu, C. Y., & Chau, K. W. (2012). Liquidity and information asymmetry in the real estate market. The Journal of Real Estate Finance and Economics, 45(1), 49-62.

Zheng, Jack (2020). Data Visualization IT 7113 Lecture Notes