NOVA IMS Information Management School

MGI

Master Degree Program in

Information Management

A PILOT BUSINESS INTELLIGENCE SOLUTION FOR AIRBNB

Analyzing Accommodations in Portugal

Ana Lúcia Martins Massano

Project Work

Presented as partial requirement to obtaining the Master´s degree in Information Management

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

A PILOT BUSINESS INTELLIGENCE SOLUTION FOR AIRBNB

Analysing Accommodations in Portugal

by

Ana Lúcia Martins Massano

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

Supervised by

João Bruno Morais de Sousa Jardim, PhD, NOVA Information Management School

Miguel de Castro Neto, PhD, NOVA Information Management School

November, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 30 of November of 2024

Ana Lúcia Martins Massano

DEDICATION

I dedicate my project thesis work to my family. A special feeling of gratitude to my loving parents, Manuel and Maria Massano whose words of encouragement helped me a lot. I dedicate project thesis work also to my boyfriend, Vasco Silva, who has never left my side is very special, and who has supported me throughout the process.

ACKNOWLEDGEMENTS

I want to offer my acknowledgments to my family, colleagues, CGI company, and professors for all their support during this period of my life. In particular, I want to say thank you to my advisor for the time spent with my work, and to professors Sara Ribeiro and Dra. Sónia Nunes by their support and comprehension in some of the most important moments of this project thesis year. Lastly, I want to present my sincere acknowledgments to one of the most important people in my life, my boyfriend, for his unconditional support during my master's. For all these people I want to say many thank you because without them this achievement wouldn´t be possible.

ABSTRACT

Nowadays, massive amounts of data are generated by activity in the hospitality industry every day and strong processes are needed to handle all this new information. The purpose of this study is to establish a general prototype for Business Intelligence and Data Analytics Environments, enabling end-to-end development and data modeling efforts that support insightful decision-making in the short-term rental market. The methodology chosen is based on the Kimball Life Cycle approach, which takes into account current concepts of data warehousing as applied to a Medallion Architecture. Big Data and Cloud Computing are modern concepts that disrupted the traditional notion of BI and data warehousing. Based on data sourced from Inside Airbnb, the project provides guidelines covering the entire process, from data extraction to visualization, using tools like Databricks and Power BI. Results demonstrate that BI tools can effectively enhance performance analysis, improve customer satisfaction, and support strategic decision-making in a short-term rental market that can enhance performance accuracy, and heighten customer satisfaction and strategic decision on business. They also hint at the potential broader applicability of using more advanced data models with Airbnb data (i.e., perhaps in different regions, or against traditional hospitality).

KEYWORDS

Business Intelligence; Big Data Analytics; Cloud Computing; Data Warehouse; Hospitality Services

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ACID	Atomicity, Consistency, Isolation, Durability
API	Application Programming Interface
BI	Business Intelligence
CSV	Comma-Separated Values
DB	Database
DSS	Decision Support Systems
ELT	Extract, Load, Transform
ERP	Enterprise Resource Planning
ETL	Extract, Transform, Load
GDP	Gross Domestic Product
ICT	Information and Communication Technologies
KPI	Key Performance Indicator
MDSS	Marketing Decision Support System
ML	Machine Learning
NLP	Natural Language Processing
NoSQL	Not Only SQL
OLAP	Online Analytical Processing
OSM	OpenStreetMap
RDBMS	Relational Database Management System
SME	Small and Medium Enterprise
SQL	Structured Query Language

1. INTRODUCTION

1.1. MOTIVATION

Nowadays, technological developments are increasingly rapid and have a significantly positive impact on various aspects of different sectors, particularly in the industry (Statista, 2024). Hospitality and tourism are the most advanced industries in the world so their crucial role in the worldwide or international economy is beyond doubt (Madyatmadja et al., 2021) The advancement is evident, as shown by the increase in worldwide tourists, from over 400 million in 1990 to 1.3 billion in 2017 (Statista, 2024). A large amount of generated data (e.g., hotel reservations, flights) makes it challenging for the industry to manage it to get the information with quality and consequently perform decision-making as there is no time for the industry to consume all the data and it is expensive also (Alaei et al., 2019). The huge amount of data needs to be converted to information on time before the data gets old and useless (Hočevar & Jaklič, 2010). Business intelligence and Big Data Tools can play an important role in solving those problems and answering the needs of the Hospitality Services and Tourism industries (Madyatmadja et al., 2021).

As part of Hospitality Services, Airbnb is the world's leading project in providing a mediation service between hosts and travelers for informal accommodation, generally for tourism purposes. It provides relevant data on establishments such as the price, evaluations, and location (Bustamante et al., 2020). In this project, we want to investigate the impact of the implementation of a Business Intelligence modern architecture in Airbnb Portugal data generated to centralize the data and provide it quality for decision making.

1.2. OBJECTIVES AND METHODOLOGY

The objective of this research is to develop a prototype of an end-to-end BI (Business Intelligence) and Data Analytics process regarding customer experience Airbnb data in Portugal. The data will be obtained from the Inside Airbnb web page and will contain data from the main Portugal regions: Lisbon and Porto. The research question that conducts all this project is “How can a BI and Data Analytics prototype enhance insights into customer experience and inform strategic decisions in the short-term rental market in Portugal?”.

This project follows the Kimball Lifecycle methodology for data warehouse development, initially proposed by Ralph Kimball in the 1980s. The main methodological approach to be adopted will be Design Science(Hevner A. R., 2007); (Hevner & Chatterjee, 2010), making use of Microsoft Azure Cloud BI tools to create the structure of the data model - collect the data about Airbnb in Portugal, process and transform them and load them in a centralized data repository that will feed future analytic applications that will support the decisions. Additionally, based on the “comments” field from the reviews source file was performed a small sentiment analysis to understand if the customer's reviews were good or not. After the

data modeling phase, a Dashboard will be developed, through the BI data visualization tools, to allow insights from the data.

With this project, is expected to provide a replicable methodology about how to build a data warehouse for Airbnb that can become public and allow people to get insights about Airbnb in Portugal. For example, allow identifying the main Airbnb accommodations in the main regions of Portugal, Lisbon, and Porto; which ones have the best reviews; where they are located; the type of room; the average number of nights the person stays there; availability in days throughout the year and so on. These insights can support future decisions about which hotels to choose or even allow an understanding of where the best location is to start a new business renting apartments or rooms for tourism. Besides the customer experience perspective, is expected to contribute to identifying the key factors that contribute to customer preferences by Airbnb Hotels.

This project work is organized into six main chapters: The first chapter, Introduction, where is explains the motivation for this work and its main goals and methodology. The second chapter, Literature Review, corresponds to the State of the Art and the Related Work where is presented some similar studies. The third and fourth chapters, Data and Methodology, describe the original data source files, the methodology, architecture and ETL processes adopted for the project development. The fourth chapter, Results, and Discussion, presents the main results of the developed solution. In the fifth chapter, Conclusions, are exposed to the main conclusion obtained with the development of this project. The last chapter, Limitations and Recommendations for Future Works, presents the main limitations faced with this work and some recommendations for future works.

2. LITERATURE REVIEW

2.1. HOSPITALITY SERVICES INDUSTRY

The hospitality industry has been growing over the past few years, with a compound annual growth rate of 4.7% from 2016 to 2021, and has perspectives of keep to grow, making the role of data-driven decision-making increasingly important (Mnyakin, 2023).

The international hospitality and tourism industry is one of the most important, largest, and fastest-growing industries in the world (Limna, 2023). Globally, it generates $7.6 trillion in revenue and employs 292 million people, accounting for nearly 10% of global GDP (Gross Domestic Product) and one out of ten jobs, with more jobs expected to be generated in the imminent years due to its projected steady growth (Ruel & Njoku, 2021).

In many countries, the hospitality industry is critical to economic success (Martinez-Martinez et al., 2019). Hospitality refers to the practices of welcoming and entertaining customers (Limna, 2023). Hospitality services are mainly focused on the provision of food, drinks, and lodging and can take place in both commercial and non-commercial establishments (Naumov, 2019). The hospitality industry encompasses hotel and tourism sectors, food and beverage sectors, and meeting and event sectors (Sisson & Adams, 2013).

This industry is considered a vast sector that encompasses a wide range of businesses, including hotels and other establishments that provide services to customers, and plays a significant role in the global economy (Mnyakin, 2023). It is one of the fastest-growing industries in the world and is known for its ability to create a personalized experience for each customer (Barrows et al., 2011; Jones & Pizam, 1994). The hotel sector is a significant component of the hospitality industry, that is a highly competitive, with a focus on customer satisfaction, comfort, and convenience. In recent years, there has been an increase in the number of hotels, particularly in emerging economies, due to the growth of the tourism industry (Mnyakin, 2023) . According to this author, the travel and tourism industry is an important driver of Hospitality Industry. This industry is extremely competitive, with a focus on providing unique experiences to customers. The growth of the travel and tourism industry has led to an increase in the demand for hotels, and other hospitality services worldwide. The hospitality industry plays a critical role in ensuring the satisfaction of tourists and travelers, creating positive experiences that contribute to the growth of the industry (Mnyakin, 2023).

2.1.1. HOSPITALITY SERVICES INDUSTRY – TECHNOLOGICAL POINT OF VIEW

These days hospitality industry faces a highly competitive environment, which is saturated with new technologies, customers who expect superior service, serve as a significant source of innovation, and are constantly confronted with the challenges of rising costs (Limna, 2023). The survival of these businesses is often determined by their overall financial performance,

ability to adapt to changing environments, and how they transform and expand their services to meet the needs and expectations of their customers (van Niekerk, 2016), (Wikhamn, 2019). Many hospitality businesses have been investing more in technology in last years to increase revenues and growth (Loureiro et al., 2021). As result the hospitality and tourism industry is leveraging cutting-edge technologies to enhance customer service and experience (Limna, 2023). According to this author innovative technology can put the customer's point of view at the center of operations. The technological advancements have been transformed into smart tools for providing customer service, and they are being used to improve the customer experience (Goel et al., 2022). Furthermore, the latest development and application of advanced technology and information and communication technologies (ICT) have transformed and automated every aspect of the tourist experience resulting in tremendous changes in the tourism and hospitality industry (Kumar et al., 2021). The hospitality industry, especially the hotel businesses, is a data-intensive industry that collects massive amounts of data in various forms (Limna, 2023).

2.2. SHARING ECONOMY

The information technologies area has been growing with big velocity until today´s, which allowed people to start to practice goods sharing in Online platforms (Lee et al., 2020). That activity is encompassed in the concept of Sharing Economy. R. Belk (2007) sees the Sharing Economy as being an alternative to the private ownership that is emphasized in both marketplace exchange and gift giving, while Botsman (2015) defines it as being an economic system that has its basis in sharing underused assets or services for free or with payment, directly from the persons. In Schor (2014) perspective Sharing Economy activities are classified into four big categories: increased utilization of durable assets, recirculation of goods, exchange of services, and sharing of productive assets. Barnes & Mattsson (2016) argues that a Sharing Economy or collaborative consumption is “the use of online market places and social networking technologies to facilitate peer-to-peer sharing of resources (such as money, space, goods, skills and services) between individuals, who may be both suppliers and consumers.”. For Hamari et al. (2016) Sharing Economy is an umbrella concept that encompasses several ICT developments and technologies, among others collaborative consumption, which endorses sharing the consumption of goods and services through online platforms. In his publication “The Sharing Economy – Consumer Intelligence Series” PWC (2015) defines Sharing Economy as being an economic movement where common technology allows people to get the services and goods, they need from each other, peer to peer, instead of buying from established corporations” (PWC, 2015). In the Cambridge Dictionary (2023) Sharing Economy is defined as an economic system that is based on people sharing possessions and services, either for free or for payment, usually using the Internet to organize this.

2.2.1. SHARING ECONOMY BUSINESS MODELS

The sharing economy supported by technological advancement promoted the growth of new business models (Šepeľová et al., 2021). These new platforms excluded traditional markets, decomposed sector categories, and maximized the use of scarce resources (Schor, 2014). For Timmers (1998) a business model is defined as being an architecture for the product, service, and information flows, including the description of the various business actors and their roles, potential benefits for the several business actors and sources of revenues. According to Fraiberger & Sundararajan (2017) the business models of the sharing economy distinguish three important components characterized as suppliers, consumers, and platforms, which take on the role of markets. The suppliers or providers are the individual micro-entrepreneurs with small businesses that deliver goods or services in per-to-per transactions that generate the supply of goods and services on the platform.

Šepeľová et al. (2021) defines suppliers or providers as being the individual micro-entrepreneurs with small businesses that deliver goods or services in per-to-per transactions that generate the supply of goods and services on the platform; the Consumers as the ones that buy, acquires, or rents goods and services from a supplier, dealing with peer-to-peer transactions on online platform due to access to less expensive goods and services; and the Platforms (marketplaces, intermediaries, sharing economy companies). Peer markets facilitate, organize, and mediate connections between suppliers and consumers to simplify their mutual transactions. According to Jevons (2015) to Sharing Economy platforms represent online platforms that coordinate a group of individuals (or peers) to allow the sharing of an asset or resource, including physical assets or skills. In that platforms people can share, rent, exchange, or donate goods and services.

Unlike traditional business models based on ownership, the sharing economy is a business model that is based on sharing underutilized assets from spaces to skills to stuff for monetary or non-monetary benefits (Botsman, 2013). The fundamental notion is to leverage information technology to empower each party with information that enables distribution, sharing and reuse of excess capacity in goods and services (Heinrichs, 2013). The rise of the sharing economy can be attributed to the paradigm change in consumer behavior (Puschmann & Alt, 2016). Consumers are moving away from owning goods to temporarily using goods and making goods available to strangers online (Bardhi & Eckhardt, 2012; R. W. Belk, 2013).

2.3. AIRBNB

Airbnb is a short-term rental platform created in 2008 and its purpose is to facilitate short-term lodging rentals mainly for vacation periods. That platform acts as a facilitator where individuals rent their rooms, homes, or apartments to the guests, that is, the hosts publish their properties on Airbnb platform by providing details, images, house rules, prices, and other accommodation characteristics, and the Guests can search for the accommodation that best

satisfies them needs and then communicate with the Hosts through the platform. Due to its wide range of lodging options, personalized and local offers Airbnb has gained popularity and advantage when compared to traditional Hotels. As it is a platform that operates in many countries around the world, it connects hosts and guests from several backgrounds enriching customer and host experience.

2.3.1. BUSINESS INTELLIGENCE AND SHORT-TERM RENTAL PLATFORMS

In the digital era, social and collaborative data sources, like Airbnb platform, have been gaining high importance in several domains. There is happening a shift of roles from internet users as consumers to information creators and sharing. That shift highlights the growing relevance of user-generated content (tourist-generated content), mainly in the tourism sector where traveler's opinions are considered reliable and influential. According to Bustamante et al., (2020) tourist's comments on the web, particularly on platforms like Airbnb, are perceived as reliable, credible, and influential in decision-making. This makes that data a good basis to analysis the tourism sector because users increasingly rely on online reviews before making decisions, such as choosing hotels or restaurants.

Business Intelligence solutions and collaborative data sources have seen increased use, providing various benefits such as operational optimization, improved customer relationships, and competitive advantage (Peng et al., 2017; Castellanos et al., 2012; Radhakrishna et al., 2015).

Business Intelligence a decision support, process improvement, and performance management tool (Williams, 2016). Represents a concept for extracting and analyzing business data for decision-making. As two main concepts of business intelligence, data warehousing and data mining became essential elements of any computer strategies, yet very few companies succeeded to set up such a system that aims to centralize and organize all the company data in a perspective of discovering unexpected information but could help in decision making (El Moukhi et al., 2015). Inmon and Kimball provided the most important definitions of the concept of Data Warehouse. A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process (Inmon, 1992). Kimball (1996) defined the data warehouse concept as a copy of transaction data specifically structured for query and analysis.

Several sectors make use of Business intelligence techniques and methods like data warehouses, OLAP (Online Analytical Processing) systems and dashboards for monitoring policies; spatial data warehouses that seek to take advantage of people information; and use of data mining techniques to create profiles of people and communities (Berndt et al., 2000; Musa et al., 2013; Rizi & Roudsari, 2013; Mooney et al., 2015; Wisniewski et al., 2003). BI architectures are increasingly used in tourism management and development. In tourism sector are adopted techniques like OLAP and data mining for understanding tourist behavior

(Miah et al., 2017), sentiment analysis for tourist opinions (Krawczyk & Xiang, 2016 and Alaei et al., 2019), creating indicator systems (Höpken et al., 2013) and (Shayegh & Daneshpour, 2015) and, linked data for data retrieval and integration (Sabou, Onder, et al., 2015).

For Chen (2004) and Baggio & Caporarello (2005) BI represents one of the facets of Decision Support Systems (DSS). DSS are systems that help managers to optimize the process of making decision; by giving the possibility to quickly retrieve efficient information from multiple blocks of data (El Moukhi et al., 2015). Example of this type of system on tourism industry are TourMIS (The Tourism Management Information System) and ETIHQ (Exposing Tourism Indicators as High Quality Linked Data) (Sabou, Brașoveanu, et al., 2015), both supported by official tourism data sources for tourism analysis. ETIHQ employs semantic technologies and opinion mining for processing data, facing challenges in data integration due to different syntactic formats and semantic difficulties (Bustamante et al., 2020).

In Thailand, Vajirakachorn & Chongwatpol (2017) applied BI to the tourism industry, integrating data about products purchased by tourists, experienced services, evaluated destinations and data about accommodations for understanding tourists' behavior and enhancing satisfaction and profits.

The above examples make basis for assertion of studying application of BI within the tourism industry from several contexts. For example, in Thailand, (Vajirakachorn & Chongwatpol, 2017) integrated BI tools to assess the products bought, experienced services, rated places and lodging facilities of visitors. This concept helped understand the tourist's movement which helped in formulating a strategy that improved customer satisfaction and increased revenues. This demonstrates the power of Business intelligence in converting data into useful information and solving day-to-day and long-term problems within the tourism sector.

Integrating collaborative data in data-driven BI system brings an opportunity to foster the decision-making process towards improving tourism competitiveness (Bustamante et al., 2020). The tourism competitiveness improvement is linked to understanding the sector through data analysis. Collaborative data sources, such as Twitter, Airbnb and OpenStreetMap (OSM) are valuable for understanding user behavior and extracting several types of information (Diakopoulos et al., 2010) . For example, Twitter and Airbnb can be used for sentiment analysis, determining the image tourists have of a destination, identifying tourists and residents, and extracting geographic information for route and concentration analysis (Bustamante et al., 2019).

That authors work highlights the growing importance of BI and collaborative data in various sectors, focusing on tourism and health as significant domains benefiting from these technologies and methodologies.

2.4. RELATED WORK

In this section are presented the similar studies developed in the same area of our research – Airbnb and Business Intelligence or Analytics and will be discussed the approaches and methods applied.

As hospitality industry is based on human services, it is strongly reliant on representation and customer reviews (Limna, 2023).

A study developed by Wöber (2003) addressed the information supply in tourism management by marketing decision support systems like TourMIS, MDSS (Marketing Decision Support System) and ETIHQ. TourMIS is a decision support system developed according to specific requirement of tourism managers (Bustamante et al., 2020) and is financially supported by Austrian National Tourist Office and the European Travel Commission. The main goal of this system is to provide an integrated view of several data sources that will be visualized and analyzed through a graphical interface. TourMIS stores official data from Eurostat, the Federal Statistical Office and local and national tourism data supplied by the respective tourism organizations. That data allows get information about trends of occupancy rates, number of visitors, host destinations. In addition, it showcases statistics from TourMIS as Linked Data (LD), allowing tourism practitioners connect to other sources of indicators and explore linked data archives (Wöber, 2003). This study alerts to the relevance of information and explains that due to the evolution of new technologies and high-capacity storage media the efficient information management is steadily increasing and also due the growing market dynamics raise information needs. A MDSS can be of particular importance because it supports organizations in collecting, storing, processing, and disseminating information and in the decision-making process by providing forecasts and decision models (Little, 1970).

Another system referred by Wöber (2003) was the ETIHQ related to the exposing tourism indicators as High Quality Linked Data. According to Bustamante et al. (2020) ETIHQ is a tourism DSS which draws upon TourMIS and allows visualizing and analyzing statistical indicators from different data sources and from different domains (tourism, economics, environment). ETIHQ utilizes advanced semantic technologies and opinion mining methods to process gathered data and derive practical insights from its repositories. However, ETIHQ encountered challenges in integrating data due to the diverse syntactic formats of most open data, demanding significant efforts for harmonization. From a semantic perspective, the obstacles arise from variations in terminologies for identical entities, differences in geographic granularity, and variations in measurements across distinct time intervals. It's worth noting that both TourMIS and ETIHQ heavily rely on official data sources.

Vajirakachorn & Chongwatpol (2017) studied the application of business intelligence in the tourism industry. This case study is based on a local food festival in Thailand, in which the authors tried to applicate Business Intelligence to Tourism Industry by integrate data from several data sources. The system integrates a massive volume of data about products

purchased by tourists, experienced services, destinations reviews done by the tourist, as well as data about accommodation, and translates such data into a meaningful information to allow the event organizers understand the behavior of tourists in order to increase their satisfaction and boost their revenues and profits. The framework relies on an architecture composed of a database management system, business analytic, business performance management, machine learning techniques, and data visualization to guide the analyze.

Lee et al. (2020) analyzed online reviews with the goal to investigate the customers behavior in the sharing economy. In this study was performed a customer's online reviews analysis from London Airbnb with the purpose of identify customers experience attributes that impacts customers choices in Airbnb; and improve the service offers and customers' expectations in shared economy. The adopted approach was Text Mining to identify a set of big points thought the text reviews and, was applied to 169.666 reviews from London Airbnb users between 2011 a 2015. In order to group the reviews, the author used Hierarchical Cluster method that allowed clustering similar words based on co-occurrence. This study make use of longitudinal analysis was used for best understanding of Airbnb customer behavior. The main identified attributes were: cleanness, location and communication.

Bustamante et al. (2020) developed a BI Platform for Tourism Analysis (BITOUR) where were integrated data from 4 platforms: Twitter, Openstreetmap, Tripadvisor and Airbnb. That platform followed a classical BI architecture and provides functionalities for data transformation, data processing, data analysis and data visualization. BITOUR platform allows interactive destination analysis, data loading, routine execution, and visualization. That platform was designed to analyze tourist activities, opinions, attractions, and seasonal trends based on nationality. It facilitates dynamic data manipulation for analyzing tourist trends and optimizing responses.

In respect data processing, the BITOUR mechanisms identify the tourists on Twitter thought tweets and cross the information to the local attractions and Airbnb and TripAdvisor accommodation and makes a sentiment analysis based on tourist's opinions using geolocation objects on Openstreetmap. The expected results were enabling data analysis and data visualization to answer questions like the average of the most frequent places; the average accommodation duration; or travelers opinion regarding specific destination places (Bustamante et al., 2020).

Table 2.1 - Comparing Airbnb Studies

Study	Goal	Approach	Results
Wöber, K.W. (2003)	• Provide an integrated view of several data sources that will be visualized and analyzed through a graphical interface • Allowing tourism practitioners to connect to other sources of indicators and explore linked data archives	• Build a centralized system based on official data provided my tourism entities. • Analyze customers reviews	• Tourism Decision Support System (DSS) • Based on the built centralized system get information about trends of occupancy rates, number of visitors, host destinations
Bustamante et al. (2020)	• Visualizing and analyzing statistical indicators from various data sources across different domains such as tourism, economics, and the environment. •	• ETIHQ employs advanced semantic technologies and opinion mining techniques for processing the collected data. • Based on official data sources • The use of these methods is intended to extract practical insights from the repositories, enhancing the decision-making process in the tourism sector. • Analyze customers reviews	• Tourism Decision Support System (DSS) • Effective Visualization and statistical Analysis of indicators • Actionable Insights • Improved Decision Support • Enhanced Semantic Integration
Thanathorn Vajirakachorn and Jongsawas Chongwatpol (2017) .	• Apply a Business Intelligence decision support system (DSS) to Tourism Industry.	• Build a data warehouse with data provided by a food festival in Thailand. • Analyze customers reviews	• Tourism Decision Support System (DSS) • Allow to the organizers extract meaningful information to boost customers satisfaction and increase revenues.
Carmen Kar Hang Lee, Ying Kei Tse and Minhao Zhang and Jie Ma (2020)	• Identify customers experience attributes that impacts customers choices in Airbnb • Improve service offers and customers' expectations in shared economy	• Customers online reviews analysis from London Airbnb • Text Mining to identify a set of big points through the text reviews • 169.666 reviews from London Airbnb users between 2011 a 2015 • Hierarchical Cluster used to clustering similar words based on co-occurrence • Longitudinal analysis for best understanding of Airbnb customer behavior	• Important identified attributes - cleanness, location, location and communication
Alexander Bustamante, Laura Sebastia and Eva Onaindia, (2020)	• Development of a BI Platform for Tourism Analysis based on data from 4 platforms - Twitter, OpenStreetMap, Tripadvisor and Airbnb. • ETL and Visualization	• The BITOUR platform mechanisms identify the tourists on Twitter thought tweets and cross the information to the local attractions and Airbnb and TripAdvisor accommodation and makes a sentiment analysis based on tourists' opinions using geolocation objects on OpenStreetMap	• Data visualization to answer questions as: • Average of the most frequent Places; • Average accommodation duration; • Travelers' opinion regarding specific destination places

3. DATA AND METHODOLOGY

In this chapter is explained the followed methodology to develop this thesis project. It will start by providing a clear overview of the main methodologies adopted and ends with the description of the main steps followed to develop the Data Analytics/Business Intelligence solution that addresses the purpose of this thesis project. The chapter is organized in three sub-chapters: the first one will describe the methodology adopted, the second one describes the approach and architecture adopted to perform the ETL (Extract, Transform and Load) work in order to load, clean, transform and aggregate the data that will feed the last layer of this project solution - the dashboard and report. The third subchapter presents the adopted approach at the reporting level – from the cleaned and transformed data to the report and dashboard development.

3.1. THE KIMBALL LIFE CYCLE APPROACH

Implementing Business Intelligence (BI) involves various steps, including gathering data, integrating it, analyzing the results, and generating reports. To ensure a successful solution implementation, several critical factors must be taken into account, such as the quality of the data, governance, user adoption, scalability, security, flexibility, integration, reporting capabilities, visualization, and cost-effectiveness. A properly executed BI solution should offer organizations accurate, real-time insights, enabling more informed decision-making, enhancing operational efficiency, and driving revenue growth (Moitas et al., 2023).

Ralph Kimball´s methodology emphasizes dimensional modelling focusing on the business needs, infrastructure and phased project delivery in order to improve customer satisfaction, report accuracy and processing times acting as a decision-making facilitator and bringing organizational data managements effectiveness (Moitas et al., 2023).

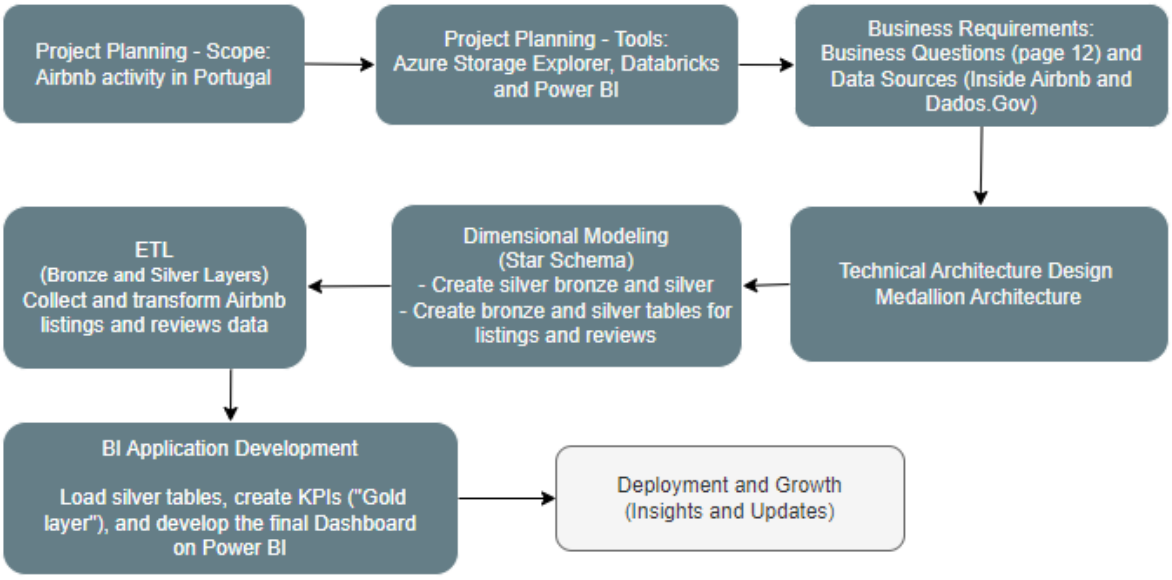

Figure 3.1 – Project phases based on Kimball life cycle

In this project was followed the Kimball lifecycle approach (figure 4.1) that is frequently used to design, develop and maintain data warehouses and business intelligence systems. Ralph Kimball was one of the pioneers in the data warehousing and refers the importance of a bottom-up approach when building data warehouses, focusing on incremental development of data marts integrated into an enterprise data warehouse. That methodology involves project planning, business requirements gathering, technical architecture design, dimensional modelling, physical design, ETL development, BI application development, deployment, maintenance and growth. According to Chávez et al., (2021) focus on business, adequate information infrastructure and advanced deliveries in significant stages are basic principles to create a data warehouse and is must follow the methodology based on project planning and definition of business needs.

That project followed the first seven steps as the purpose of this project is to present a prototype of a Business Intelligence end-to-end solution for academic purposes.

This master thesis project prototype started by defining planning, defining the scope and trying to understand the structure of the dashboard – Airbnb Listings and Reviews. The next step was defining the business requirements mainly define the data sources of the data (Inside Airbnb1) and define the data modelling processes. Next was chosen the technical architecture, that is, where chosen the tools to use in the solution development – Databricks to data modelling, python for local processing reviews dataset for sentiment analysis purpose (section 3.2.3 of this thesis project) and Power BI do dashboard development. The most important stage according to Kimball is the data modelling. For Kimball there is four steps in the dimensional design process – select the business purpose, declare the grain of the data, identify the dimensions and the facts. In that case was used the star schema to design the data model with four dimensions – date, listings, host and geography; two fact tables – listings and listings reviews because the business purpose is to analyze the Airbnb in Portugal. The lowest level of detail of the data is day.

To implement the physical database design and build the ETL processes Databricks Community platform is used in this kind of project usually the datasets are large and being permanently updated and Databricks is a better tool in terms of processing large datasets. Depside the goal of this project to be develop only a small prototype, it can be applied to the real world of hospitality services, that have a different data volume dimension.

The ETL phase involved extract the data from the data sources, Inside Airbnb (listings, calendar and reviews data) and *dados.gov*[2] (geographical data) csv files, and ingest them into a bronze layer (raw data), clean and transform (dimensional data modelling) and persist them into a silver layer (cleaned and transformed data). The csv source files and the delta tables were stored in azure storage blob container. That was the last phase of this solution, where were

[1] https://insideairbnb.com/
[2] https://dados.gov.pt/

aligned the business requirements, queried and displayed the data stored in the built data warehouse. The approach adopted in the data warehouse development was the Imon Top-Down approach that defends the creation of a centralized data repository that will serves data to the listings and reviews data marts. In that project the Gold Layer was applied at the Power PI level due time limitations and only to create the Key Performance Indicator (KPI´s) to be used in the Dashboard (BI application management phase).

Kutay (2021) provided a comprehensive overview of three prominent cloud data storage patterns: Data Warehouse, Data Lake, and Data Lakehouse, offering valuable insights for organizations navigating cloud-based data management. Armbrust et al. (2021) introduced the concept of "Lakehouse," representing a new generation of open platforms aiming to unify data warehousing and advanced analytics, addressing the limitations of traditional data storage and analytics platforms (Ravat & Zhao, 2019). The authors analyzed the metadata management for data lakes, exploring strategies for effectively organizing, storing, and leveraging metadata to ensure data quality and usability within data lake environments (Ganesan & Kalavathy, 2020). Azeroual et al. (2019) addressed the challenges associated with research information heterogeneity during integration processes, contributing to the development of strategies for effectively integrating diverse research data sources (Sahoo & Das, 2019). This author conducted a comprehensive study on the performance of MapReduce-based big data processing frameworks, offering valuable insights for optimizing big data processing pipelines in distributed computing environments (Gallinucci et al., 2019). This author highlighted the importance of considering data variety in analytics processes, aiming to enhance the analytical capabilities of document-oriented databases, particularly in scenarios involving diverse data types and structures (Rodriguez & Garcia-Valls, 2018). This author evaluated the performance of Medallion Architecture in real-time embedded systems, providing insights into its effectiveness in meeting the performance requirements of such systems (Kumar et al., 2016). (H. Wang et al., 2017) focused on the performance evaluation of traditional ETL systems in cloud environments, offering valuable insights for optimizing data integration workflows in cloud-based environments (Alqarni & Pardede, 2012).

Patel & Jones, (2017) benchmarked the performance of Medallion Architecture against traditional data processing systems, providing comparative insights into their efficiency and scalability (Smith & Johnson, 2014). (Kim & Park, 2015) conducted a comparative analysis of Medallion Architecture and MapReduce for large-scale data processing, exploring their respective strengths and weaknesses (Patel & Jones, 2017). (Shenoy & Babu, 2008) developed a performance model for automated transaction processing systems, providing valuable insights for designing and optimizing transaction processing systems for high-performance applications (Wang, 1998). Lup Low et al. (2001) proposed a knowledge-based approach for duplicate elimination in data cleaning processes, aiming to improve the accuracy and efficiency of duplicate detection and elimination in large datasets Wang et al. (2017). Alqarni & Pardede (2012) discussed the integration of data warehouses and unstructured data and suggested strategies for effective integration to improve the usability and value of integrated

data for business intelligence and decision-making (Smith & Johnson, 2014). Wang (1998) provided a product perspective on total data quality management, emphasizing the importance of data quality in information systems and discussing strategies for ensuring data quality throughout the data lifecycle (Shenoy & Babu, 2008).

Traditional data storage and processing technologies, although popular and well-established, exhibit shortcomings in meeting contemporary performance requirements (Ganesan & Kalavathy, 2020). The advent of cloud-based solutions and distributed computing has attempted to address these limitations but introduces challenges related to data consistency, security, and scalability (Ravat & Zhao, 2019). Not Only SQL (NoSQL) databases, such as MongoDB, have emerged as alternatives, providing improved flexibility but posing challenges in terms of data consistency and complex querying (Sahoo & Das, 2019). According to Armbrust et al. (2021) the existing work in cloud data storage patterns highlights the need for comprehensive solutions that balance performance, scalability, and consistency in modern data management. A Lakehouse architecture was suggested by Armbrust et al. (2021) to help to unify data warehouses and advanced analytics. While innovative, the conceptual focus leaves gaps in practical implementation guidelines, scalability considerations, and real-world use cases (Ravat & Zhao, 2019). Metadata management for data lakes, as explored by Ravat & Zhao (2019), sheds light on challenges but lacks detailed strategies for addressing interoperability and standardization issues. Azeroual et al. (2019) work on research information heterogeneity Ganesan & Kalavathy, (2020) provides solutions but may lack specificity in adapting to diverse research domains. (Gallinucci et al., 2019) focus on OLAP for document-oriented databases, but the study's practical applicability to real-world scenarios remains unexplored. (Alqarni & Pardede, 2012) explored the integration of data warehousing with unstructured data (Smith & Johnson, 2014). Kandel et al. (2011) research on data wrangling identified key directions but lacks concrete guidelines for implementing visualizations and transformations in diverse data environments. The knowledge-based approach for duplicate elimination by Lup Low et al. (2001) focused on theory, potentially overlooking adaptability and effectiveness across various datasets. (R. Y. Wang, 1998) work on total data quality management (H. Wang et al., 2017) offered high-level strategies but may lack granular insights into addressing specific data quality challenges across diverse organizational contexts.

Shabu et al. (2024) proposed system, grounded in the Medallion Architecture represents a strategic response to the limitations observed in current data management paradigms. For this author the Medallion Architecture seamlessly intertwines traditional data storage technologies with contemporary cloud-based solutions, presenting a holistic approach to data management. The system proposed by this author was based on Medallion Architecture and integrates traditional and modern data storage technologies to address interoperability challenges, introducing key modules such as Smart Metadata Management for standardization and Adaptive Research Information Integration for handling diverse research data. According to this author the Next-Gen Online Analytical Processing (OLAP) module

improves performance for document-oriented databases while scalability challenges in unstructured business documents are addressed through dedicated processing mechanisms. The results of the study of this author shown that Medallion Architecture is shown to outperform traditional ETL and MapReduce in terms of execution speed, CPU utilization, memory efficiency, and scalability and the statistical analysis highlighted its optimized resource use, faster processing, and better data partitioning, making it superior to traditional Relational Database Management System (RDBMS). The system leverages advanced OLAP techniques, parallel processing, and efficient indexing to improve query speeds and overall data processing performance. The project addressed practical challenges such as scalability, usability, and data quality management with dynamic duplicate elimination and customizable frameworks. Its adaptable approach makes it a pioneering solution, revolutionizing data-driven decision-making by providing a scalable, domain-aware, and efficient data management system that is ready for future needs.

Today, are produced many data in many sectors, and Hospitality Services Sector is no exception. As a global company, Airbnb as to deal every day with many data from hosts, customers, communication between the both. This way the medallion architecture proposed by Shabu (2024) was also adopted by us as a purpose to be the bases from our solution data modelling and help to deal with data efficiently.

3.2. Business Understanding and Data

3.2.1. Business Understanding

Airbnb is one essential company in the accommodation sector allowing users to look up and book short-term bases and sometimes even long-term stays. The site brings together the owners of properties who are referred to in this case as hosts and people on visit who are buddies and facilitates their communication with a view of addressing accommodation issues among themselves or even making bookings. For the purpose of this thesis, the data are oriented toward Lisbon and Porto districts which contains primarily important data sets like listings, calendar and reviews. Learning these data is important as it helps to establish trends, patterns, and insights that are beneficial in making decisions for the host and the guests.

3.2.2. Project Goals

The main goal of this work is to utilize the Airbnb statistics in order to get insights that can help the hosts, potential investors and political leaders in making informed choices. This project draws upon data from listings, calendars and reviews to demonstrate key trends including those of occupancy levels, demand fluctuations across seasons, pricing trends and levels of guest satisfactions. These insights can be used for enhancing the rentals of the hosts,

the experience of the guests as well as for the understanding of the Airbnb business in those cities.

3.2.3. Understanding Airbnb Listings, Calendar, and Reviews Data

3.2.3.1. Listings Data

This dataset includes information about each property listed on the Airbnb platform, such as location, price, amenities, and property type (APPENDIX A). Analyzing listings data helps you understand the supply side of the market, focusing on aspects such as property distribution, price variations and amenity offerings in Lisbon and Porto.

3.2.3.2. Reviews Data

Reviews provide qualitative feedback from guests about their stay experiences (APPENDIX B). This feedback can be invaluable for analyzing guest satisfaction, identifying common issues, and exploring areas where hosts can improve. Additionally, sentiment analysis on reviews can offer deeper insights into guest perceptions of listings and areas for improvement.

3.2.3.3. Calendar Data

Calendar data records daily availability and prices for each listing. This dataset provides critical insights into the demand side, including occupancy rates and price adjustments over time (APPENDIX B). Analyzing this data can reveal trends in booking patterns, seasonal demand fluctuations, and price sensitivity.

3.2.4. Business Questions

Business Objectives

The following business questions were formulated to address key aspects of Airbnb's performance in Portugal, focusing on insights that are crucial for decision-makers. To guide analysis and ensure relevance to stakeholders (accommodations owners), this project focuses on the following key business questions (BQ):

1. How many listings are available, and what are the main averages for capacity, bedrooms, beds, and minimum/maximum nights?
2. How does the available listings forecast changes through time for Lisbon, Porto and generally?
3. How does the forecast for unavailable listings change in Lisbon, Porto, and overall, in the future?
4. What is the geographic distribution of the listings?
5. Which users are ranked in top 10 and have maximum number of listings? To add, are they super hosts too?
6. How are the available listings distributed across property types?

7. What's the total count of listings made by hosts who uploaded their profile pictures?
8. How are the listings distributed across different room types?
9. How are the listings dispersed in terms of capacity?
10. What is the number of listings that have been posted based on the number of minimum nights that are required?
11. What are the average review scores across different categories?
12. How have the number of reviews evolved over time from January 2023 to August 2024?
13. How do reviews vary by month?
14. How are the number of reviews distributed by price and room type?
15. How do positive sentiments in Airbnb listings reviews fluctuate over time, and are there identifiable seasonal patterns that could influence customer satisfaction or operational strategies?
16. How do neutral sentiments in Airbnb listings reviews vary over time, and are there seasonal patterns that may indicate periods of mixed or neutral customer feedback?
17. How do negative sentiments in Airbnb listings reviews evolve over time, and are there specific periods with higher negative feedback that may require operational improvements or customer support adjustments?
18. What are the most used words in comments?

3.2.5. BUSINESS OBJECTIVES

1. Support Hosts in Decision Making: By analyzing data on occupancy, pricing and guest feedback, this project aims to provide actionable insights that can help hosts maximize their revenue, identify optimal pricing strategies and understand guest expectations.
2. Support Strategic Planning for Investors: For potential investors, the analysis can highlight profitable investment opportunities and demand trends in the Lisbon and Porto market, helping them make informed property investment decisions.
3. Inform Public Policy: Findings from this project may also be useful to policymakers who want to understand the impact of the short-term rental market on the local economy and housing availability.

3.2.6. DATA

The practical component of this master thesis project will be supported by a set of CSV files extracted from Inside Airbnb – Adding data to the debate[3] published on September 2024, used as main data source in this project, and by csv files from dados.gov used to create geography dimension. The Airbnb CSV files contain data regarding Airbnb Listings, customers Reviews and Airbnb Listings Calendar. The focus of that data are the major Portugal cities/districts – Lisbon and Porto that are a significative representation of the Portugal geographic area. The data are relative to the period comprehended between 2023 and 2024. To facilitate the data interpretation from a geographical point of view, was added a flag (FLAG_DISTRICT field of string data type) in the bronze layer of our architecture to all tables of this layer. In that layer also were added some metadata fields (audit fields in our case) as will be used a data load logic – ID_INPUT of the integer data type and that will provide us the number of the load, DT_INSERT of timestamp data type that will provide us the insert date of the load and the USR_INSERT of string data type that will identify the user that performed the load. In table 3.1 is presented a summarized table of the source files.

Table 3.1 – Data sources summary

Source	File Name	Country/ City	Nrº of rows	Nrº Of columns	Size Of File	First Date	Last Date
Insideairbnb	Listings.csv	Portugal/ Lisbon	24204	75	53.65 MB	14-09-2023	14-09-2024
Insideairbnb	Calendar.csv	Portugal/ Lisbon	8834345	7	351.49 MB	14-09-2024	14-09-2025
Insideairbnb	Reviews.csv	Portugal/ Lisbon	1547005	6	481.73 MB	14-09-2023	14-09-2024
Insideairbnb	Listings.csv	Portugal/ Porto	14446	75	32.26 MB	14-09-2023	14-09-2024
Insideairbnb	Calendar.csv	Portugal/ Porto	5272692	7	209.11 MB	14-09-2024	14-09-2025
Insideairbnb	Reviews.sv	Portugal/ Porto	907029	6	264.01 MB	14-09-2023	14-09-2024
Dados.gov	concelhos-metadata.csv	N/A	308	7	0.06 MB	N/A	N/A
Dados.gov	freguesias-metadata.csv	N/A	3092	6	0.23 MB	N/A	N/A
Dados.gov	distritos-metadata.csv	N/A	29	5	0.00 MB (4 KB)	N/A	N/A

[3] http://insideairbnb.com/

The Listings CSV´s data source files contain 38.650 records in total, of which 24.204 records are relative to Lisbon local rental accommodations and 14.446 records are relative to Porto local rental accommodations.

Regarding schema, the Listings CSV data source has 75 attribute columns regarding Listings, Listings Reviews, and Listings Hosts. The statistics summary of that raw data shows that it has some quality problems that need to be solved during the cleaning process that will be done in the silver layer of the ETL process. On table 3.2 can be seen an example of this data problems (the remain are described on Appendix A due them size):

Table 3.2 - Listings Statistics problems

summary	id	scrape_id	name	host_id	host_name	host_since
count	38650	38650	38650	38650	38650	38650
mean	4.39E+17	2.02E+13	30	2.07E+08	57	null
stddev	4.88E+17	31.44755274	38.18376618	1.94E+08	null	null
min	1000044	2.02409E+13	! 10 minutes to Lisbon airport and city center	100001921	(Elke &) Dieter	22/04/2009
max	9.99949E+17	2.02409E+13	ðŸ‡´Flores Studio, Best Location & Relaxing Terrace	99997584	é™ˆ	11/09/2024

Table 3.3 - Listings CSV Schema

id: string (nullable = true)
listing_url: string (nullable = true)
scrape_id: string (nullable = true)
last_scraped: string (nullable = true)
source: string (nullable = true)
name: string (nullable = true)
description: string (nullable = true)
neighborhood_overview: string (nullable = true)
picture_url: string (nullable = true)
host_id: string (nullable = true)
host_url: string (nullable = true)
host_name: string (nullable = true)
host_since: string (nullable = true)
host_location: string (nullable = true)
host_about: string (nullable = true)
host_response_time: string (nullable = true)
host_response_rate: string (nullable = true)
host_acceptance_rate: string (nullable = true)
host_is_superhost: string (nullable = true)
host_thumbnail_url: string (nullable = true)

host_picture_url: string (nullable = true)
host_neighbourhood: string (nullable = true)
host_listings_count: string (nullable = true)
host_total_listings_count: string (nullable = true)
host_verifications: string (nullable = true)
host_has_profile_pic: string (nullable = true)
host_identity_verified: string (nullable = true)
neighbourhood: string (nullable = true)
neighbourhood_cleansed: string (nullable = true)
neighbourhood_group_cleansed: string (nullable = true)
latitude: string (nullable = true)
longitude: string (nullable = true)
property_type: string (nullable = true)
room_type: string (nullable = true)
accommodates: string (nullable = true)
bathrooms: string (nullable = true)
bathrooms_text: string (nullable = true)
bedrooms: string (nullable = true)
beds: string (nullable = true)
amenities: string (nullable = true)
price: string (nullable = true)
minimum_nights: string (nullable = true)
maximum_nights: string (nullable = true)
minimum_minimum_nights: string (nullable = true)
maximum_minimum_nights: string (nullable = true)
minimum_maximum_nights: string (nullable = true)
maximum_maximum_nights: string (nullable = true)
minimum_nights_avg_ntm: string (nullable = true)
maximum_nights_avg_ntm: string (nullable = true)
calendar_updated: string (nullable = true)
has_availability: string (nullable = true)
availability_30: string (nullable = true)
availability_60: string (nullable = true)
availability_90: string (nullable = true)
availability_365: string (nullable = true)
calendar_last_scraped: string (nullable = true)
number_of_reviews: string (nullable = true)
number_of_reviews_ltm: string (nullable = true)
number_of_reviews_l30d: string (nullable = true)
first_review: string (nullable = true)
last_review: string (nullable = true)
review_scores_rating: string (nullable = true)

review_scores_accuracy: string (nullable = true)
review_scores_cleanliness: string (nullable = true)
review_scores_checkin: string (nullable = true)
review_scores_communication: string (nullable = true)
review_scores_location: string (nullable = true)
review_scores_value: string (nullable = true)
license: string (nullable = true)
instant_bookable: string (nullable = true)
calculated_host_listings_count: string (nullable = true)
calculated_host_listings_count_entire_homes: string (nullable = true)
calculated_host_listings_count_private_rooms: string (nullable = true)
calculated_host_listings_count_shared_rooms: string (nullable = true)
reviews_per_month: string (nullable = true)

During the ETL process will be removed all unnecessary columns, clean the data, remove the records without quality, create the tables necessary to the technical design and structure architecture that have been purposed.

The calendar CSV´s data source files contain 14.107.037 records in total, of which 8.834.345 are relative to Lisbon local rental accommodations and 5.272.692 records are relative to Porto local rental accommodations.

Regarding schema, the Calendar CSV data source has 7 attribute columns mainly regarding listings prices. That table present the list of listings, their availability in a specific date, the original price and the adjusted price as well the minimum and maximum number of nights that the listing is available in a date.

Table 3.4 - Calendar CSV Statistics

summary	listing_id	date	available	price	Adjusted price	Minimum nights	Maximum nights
count	14107037	14107037	14107037	14107037	1825	14107035	14107035
mean	4.39E+17	null	null	null	null	4.788316822	633.3586087
stddev	4.88E+17	null	null	null	null	20.32535279	519.6012786
min	1000044	14/09/2024	f	$0.00	$1,000.00	1	1
max	1E+18	14/09/2025	t	$999.00	$950.00	999	9999

4.

Table 3.5 – Calendar CSV Schema

listing_id: string (nullable = true)
date: string (nullable = true)
available: string (nullable = true)
price: string (nullable = true)
adjusted_price: string (nullable = true)
minimum_nights: string (nullable = true)
maximum_nights: string (nullable = true)

Finally, the reviews CSV´s data source files contain 2.454,034 records in total, of which 1.547.005 are relative to Lisbon local rental accommodations and 907.029 records are relative to Porto local rental accommodations. Regarding schema, the reviews CSV´s data source has 5 columns mainly regarding listings_id, reviwer_id, reviewer_name, date and comments about the listing and the owner.

Table 3.6 - Reviews CSV Statistics

summary	id	listing_id	date	reviewer_id	reviewer_name	comments
count	1773286	1773286	1773286	1773286	1773283	1773286
mean	5.07E+17	9.68E+16	null	1.33E+08	NaN	null
stddev	4.52E+17	2.60E+17	null	1.43E+08	NaN	null
min	100001545	1000044	2010-07-24	1	!!!!!!!!!!!!!!!!	
max	99999252	999949488041968405	2024-09-14	99999656	“Dexter” Janet	

5.

Table 3.7 - Reviews CSV Schema

id: string (nullable = true)
listing_id: string (nullable = true)
date: string (nullable = true)
reviewer_id: string (nullable = true)
reviewer_name: string (nullable = true)
comments: string (nullable = true)

All this data, listings, calendar and review will be cleaned and modeled in silver and gold layer described in the next chapter.

3.3. ETL PROCESS

3.3.1. MEDALLION ARCHITECTURE

Once the initial data exploratory analysis of the data is finished, the next phase of the practical part of this project is the ETL process to prepare the data that will be the basis for the reporting level. To develop this phase, was chosen the Medallion architecture, also known as Multi Hop architecture, to deal with the CSV´s data files extracted from Inside Airbnb and dados.gov. This architecture consists in the progressive and incrementally enrichment of the data structure and quality layer to layer until the final state of the data where is expected to have the data cleaned, transformed and aggregated. According to Databricks company[4] the Medallion architecture consists in a data pattern used to logically organize the data into a Lakehouse that is the "new open data management architecture that combines flexibility, cost-efficiency, and scale of data lakes with the data management and ACID transactions of data warehouses, enabling business intelligence (BI) and data engineering on all data"[5] (Databricks Company).
The Medallion architecture is divided into three layers with a self-purpose:
In the gold layer the data are ingested into the schema (database in the Databricks platform context) in it´s raw form from external source systems, that means that here the data will be stored as it is in the data source. Can be applied small transformations, but just to add some metadata columns where is stored information like the data capture date/time or process ID. On the other side, in the silver layer will be feed by the data stored in the bronze layer, that will be matched, merged, conformed and cleansed (just-enough) according to the necessity to serve several destinations purposes: self-service analytics for ad-hoc reporting (data analysis), advanced analytics (data engineering) or ML (data scientists). The raw data from several data sources is joined and instead ETL, in Lakehouse, is applied ELT (Extract, Load, Transform) methodology, that means that the after extract raw data, this data are loaded and only the necessary data will be transformed according to the destination purpose. From data modelling perspective, this layer can have the models similar to the third normal form. That layer will provide an "*Enterprise view of all its key business entities, concepts and transactions, and enables self-service analytics for ad-hoc reporting, advanced analytics and ML*."[6]. According to the Databricks company web page (2024), "*Speed and agility to ingest and deliver the data in the data lake is prioritized, and a lot of project-specific complex transformations and business rules are applied while loading the data from the Silver to Gold layer*"[6]. In this project the silver layer was used to clean, transform and join the data received the sources.

[4] https://www.databricks.com/glossary/medallion-architecture
[5] https://www.databricks.com/glossary/data-lakehouse

The last layer of Medallion architecture - gold layer, will feed the silver layer, where after the necessary clean and transformations, will be added business aggregations, business and quality rules and, the final transformations. That layer will make use of de-normalized and read-optimized data models with fewer joins, and will present small data models similar to Data Marts that will serves for example marketing, finance, human resources, between other customer data consumer departments. In Gold layer can be find several data models that follow the start schema Ralth Kimball style data models or Inmon style Data marts. This layer will feed the reports and dashboards. In that project the Gold Layer was applied at the Power BI level due time limitations and only to create the KPI´s to be used in the Dashboard.

Figure 3.2 - Architecture Medallion[6]

To build this Lakehouse architecture, will be make use of Delta Lake storage framework "*with compute engines including Spark, PrestoDB, Flink, Trino, and Hive and APIs for Scala, Java, Rust, and Python*"[7]*:*

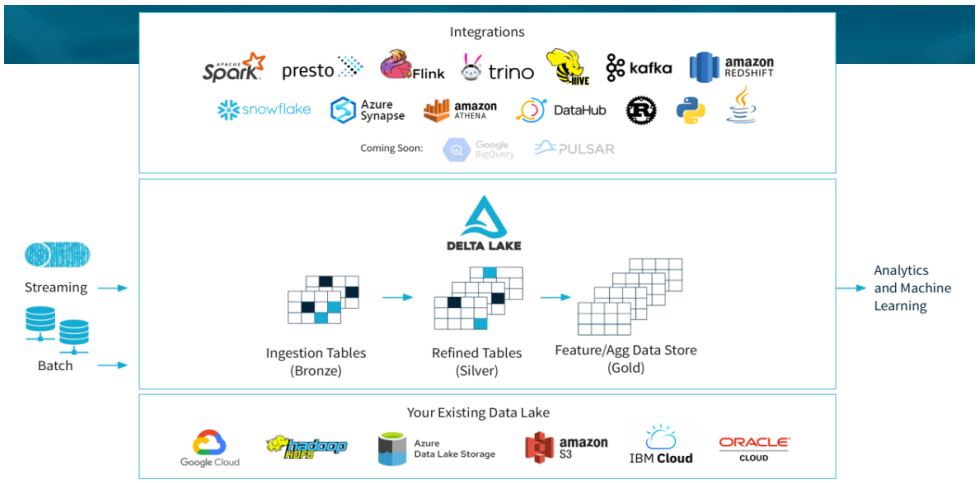

Figure 3.3 - Delta Lake Storage framework

[6] Source: https://blog.bismart.com/en/medallion-architecture

[7] Source: https://delta.io/

All this process will provide reliability data, simply data models, time travel, tables recreation from raw data any time, performance, simply implementation and understanding when compared to the traditional approaches like the ones used by the authors presented in the Literature Review chapter[3].

3.3.2. MEDALLION ARCHITECTURE – APPLICATION TO THE PRESENT MASTER THESIS PROJECT

Before starting explaining the Medallion architecture applied in this project to perform the data modelling and cleaning, will be explained small ad hoc steps: after performing the exploratory data source analysis, was made a technical design of the entities identified with the raw data (bronze layer), the expected entities that will store the cleaned, transformed, and joined data (silver layer), and the final entities that will form the final data model in Power BI level (gold layer) in the context of this project. In this technical design are specified information's like the sources of the data, primary keys of the tables, the hierarchies, ETL/ELT rules, data quality rules or tables names as well acronym list. That technical design has the goal of organize the data model building phases and facilitate the implementation. This way, bellow is described some important acronyms and definitions as well the architecture schema purposed in this project (table 3.8 and figure 3.4).

Table 3.8 - Nomenclatures – Acronyms and definitions

Acronym	Description
UDM	Unique Data Model
UDM_D_table	Data Mart
UDM_F_table	Unique Data Model - Fact table
Dim	Dimension table
Fact	Fact table
ID_x	Key field without associated description
COD_x	Key field with associated description
DESC_x	COD field description
LT	Last
M	Months
D	Days
VW	View
BRONZE	bronze layer
SILVER	silver layer
GOLD	gold layer
VAL	Value/Rate
QTY	Quantity

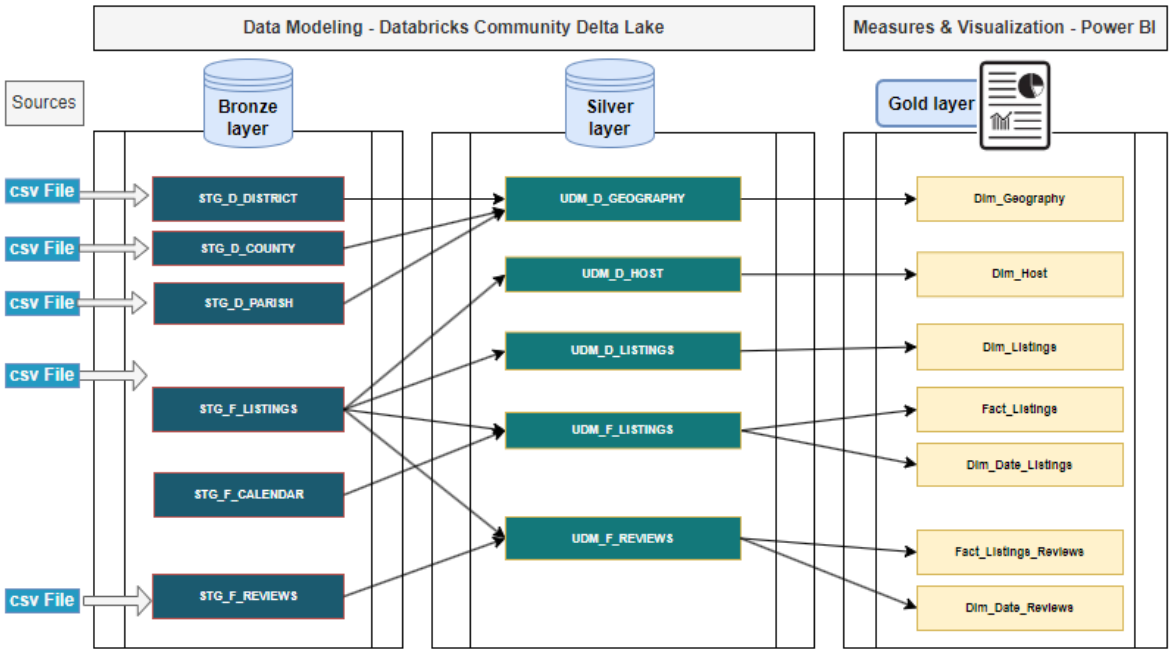

Figure 3.4 - Medallion architecture application purpose to this master thesis project

3.3.2.1. BRONZE AND SILVER LAYERS

Before processing the data in Databricks Community, were downloaded the listings, reviews and calendar September 2024 data source files from Inside Airbnb to Lisbon and Porto regions. Also were downloaded from dados.gov the Portugal lists of districts, county and parish to build the geography dimensions in silver layer and relate them to the fact tables. That files were stored on Azure container blob storage.

After upload the source files into azure blob storage, were created the schemas on Databricks (bronze and silver) and the delta tables for both layers – bronze and silver. To the bronze layer were created STG_D_COUNTY, STG_D_DISTRICT, STG_D_PARISH, STG_F_CALENDAR, STG_F_LISTINGS and STG_F_REVIEWS tables. On the other side, to the silver layer were created the UDM_D_HOST, UDM_D_GEOGRAPHY, UDM_D_LISTINGS, UDM_F_LISTINGS_REVIEWS and UDM_F_LISTINGS tables.

Implemented the tables, was developed the ETL for bronze layer, that reads the CSV´s files stored in the Azure container, proccess them with the correct options and write them into each respective delta table. In that layer, the data stays in the raw format, that means that no transformations are done.

The next step was developing the ETL process to clean, transform and join the source data, bronze tables, and store the data into the respective silver delta tables. The process starts by reading STG_F_LISTINGS and extract the relevant fields to build the UDM_D_HOST dimension. Then were deleted the duplicate rows and created the primary key made up by "host_id", "scrape_id" fields. Additionally, was created a new field “host_country” based on “host_location” source filed using split function, were removed the white spaces using trim

function and values were replaced by the country acronyms to stay homogeneous. After that, all the fields were renamed to the final name and where added the COD_ to the fields that only had the descriptions. Additionally, was read the source file “UDM_R_COUNTRY.csv” that contains the list of countries and respective codes to get the COD_COUNTRY_HOST description and store into the DESC_COUNTRY field. These were the main transformations done to build UDM_D_HOST dimension.

The geography dimension was fed by STG_D_DISTRICT, STG_D_COUNTY, and STG_D_PARISH bronze tables. Firstly, were renamed the codes and descriptions fields to the most user-friendly and joined the three tables and was created the key to the geography dimension made up district, county and parish codes. After the parish description field values were corrected according to the official parish names. To finish that dimension were selected only the most relevant fields and the data was written in the UDM_D_GEOGRAPHY delta table.

The next process to be developed was the process to UDM_F_LISTINGS starting by reading STG_ F_LISTINGS and renaming some fields to distinguish them later when joining listings table with calendar table in order to get the date and available fields from the STG_F_CALENDAR table. This join was made by listing_id and scrap_year_month taken from the file names. Using initcap and split functions were got the county descriptions and after both descriptions county and parish were replaced by the correct ones. After this step was done a left join with the geography dimension to get the district, county and parish codes as the best practices in a fact table is to keep the code instead descriptions. The next steps were to select the relevant fields, rename, convert to the correct data types and write the data into the destination delta table.

For fed the listings reviews delta table was used UDM_F_REVIEWS and the score from fields UDM_F_LISTINGS, joined by the listing_id. Were selected the relevant fields and created the Listings_Key made up my listing_id and id_date, that will allow the relation to the listing dimension. Next the field were renamed and converted to the correct data types. Note that for that fact table were created three new field related to the “comments” field sentiment analysis. The process for that sentiment analysis will be explained in a specific subsection (3.2.2.2).

The last table to be developed was the listings dimension that will contain the attributes for filter the listings and reviews fact tables. The ETL process for that table was made up by the Listings Key creation (listing_id and id_date), property type field content uniformization, adding any descriptions and the necessary data type conversions.

In short, the data source files downloaded from data.gov and Inside Airbnb were loaded into bronze layer delta tables that maintain the row format, the original format of the data. The next steps were to build a silver layer where the tables were joined; the data was cleaned and any transformations. After these ETL processes the silver tables were loaded into bi power where only the required fields were chosen and renamed for visualization purposes. In the

context of this project due to time limitations, Power BI worked not only for the report building but also as the gold layer where the necessary measures were created to feed the visuals. The next sub chapter explains deeply the sentiment analysis and respective ETL process.

3.3.2.2. REVIEWS ETL – SENTIMENTS ANALYSIS BASED ON "COMMENTS" FIELD

In the context of listings reviews, a sentiment analysis was performed based on "Comments" field present on the reviews source file. The goal was to analyze the evolution of the customers satisfaction regarding the listings and identify the most cited words.

Since reviews has a large dataset and the Databricks Community has some processing limitations causing cluster shutdown when processing a huge number of records, for sentiment analysis purposes, the reviews dataset was cleaned and dealt with locally making use of Python language and loaded back to Databricks later. For this process, the review files from September 2024 were used for Lisbon and Porto regions, totaling 2.6 million of reviews.

The transformations applied to the "Comments" field included:

- Reading reviews CSV files using Pandas library;
- Remove duplicates;
- Remove tags html using BeautifulSoup library;
- Remove some identified spam records such as # or $ using regular expressions
- Remove white spaces in the beginning and end of each comment.
    - Identification of the language of each comment using langdetect library, since those the sentiment calculation algorithms are optimized to english comments. The behavior is: If the comment is in english then save, otherwise translate to english.

The next step was to perform a sentiment analysis to determine the emotional tone of each comment. Three sentiment analysis algorithms were used: TextBlob, VADER and Flair.

3.3.2.3. TEXTBLOB

According to Loria et al, (2024) TextBlob is a rule-based sentiment analysis tool built on the Natural Language Toolkit (NLTK). It uses a lexicon-based approach by assigning predefined sentiment scores to words. The overall sentiment is calculated as the average polarity of all words in the text.

Sentiment Scores:

- **Polarity**: Ranges from -1 (very negative) to +1 (very positive). A score close to zero indicates a neutral sentiment.

3.3.2.4. VADER (VALENCE AWARE DICTIONARY AND SENTIMENT REASONER)

It is a lexicon- and rule-based sentiment analysis tool designed specifically for analyzing social media text, but works well in a variety of domains. It is particularly good at detecting word polarity in context, such as accounting for the intensity of a word by considering punctuation, capitalization, and degree modifiers (e.g., "very", "extremely") (Malde, 2020).

Sentiment Scores:

- **Compound**: A normalized score ranging from -1 (extremely negative) to +1 (extremely positive). This score aggregates the positive, negative, and neutral scores into a single metric.
- **Positive, Negative, Neutral**: Scores representing the proportion of the text that falls into each category.

3.3.2.5. FLAIR

For Pankaj (2023) Flair algorithm is a neural network-based model for Natural Language Processing (NLP) that uses deep learning techniques to access the sentiment of text. According to Devlin et al. (2019) unlike rule-based systems like TextBlob and VADER, Flair relies on pre-trained language models (like BERT) to understand the context and meaning of words, allowing for more nuanced analysis.

Sentiment Scores:

- **Binary Classification**: Flair typically provides a "positive" or "negative" label based on the overall sentiment of the text.

Terminated the sentiment analysis the reviews data were loaded into Databricks and stored in a delta table for later feed the Power BI “Listings - Reviews View – Sentiments” page charts. This will be explained in Results and Discussion section.

Table 3.9 – Sentiment analysis algorithms

Algorithm	Description & Calculation Method	Scoring, Advantages, & Disadvantages
TextBlob	A Python NLP library using simple rules for sentiment analysis. It performs polarity analysis, with scores ranging from -1 (negative) to +1 (positive).	**Scoring**: -1 (very negative) to +1 (very positive), with 0 as neutral. **Advantages**: Easy to implement, fast, works well for short texts. **Disadvantages**: Limited in

		detecting complex sentiments in longer texts.
VADER	Designed for social media text sentiment analysis. It computes a composite score using a lexicon and intensity analysis, with scores ranging from -1 to +1.	**Scoring**: -1 (very negative) to +1 (very positive), with intensity intervals. **Advantages**: Accurate for short, informal text like tweets or reviews; captures sentiment intensity. **Disadvantages**: Less effective for long or formal texts.
Flair	Deep learning-based, using contextualized word embeddings and sequence labeling to analyze sentiments. Scores can be binary or probabilistic.	**Scoring**: Binary or probabilistic (positive/negative). **Advantages**: High accuracy for complex texts, captures context and nuance better than rule-based models. **Disadvantages**: Requires more computational resources, slower compared to simpler models, may need fine-tuning for specific data.

3.4. REPORTING AND DASHBOARD

Implemented the data structure and data modeling phase, Power BI (Data visualization tool) will be connected to Databricks Delta Lake in order to access the gold layer of the implemented medallion architecture and get the developed data marts. Here will be chosen the import mode as we are dealing with a low data volume. This mode allows store the data into Power BI in contrast to direct query mode that only read the data from the source each time it is requested. The data refresh will be daily after publish it to the Power BI Service in the end of the Dashboard and Report development.

After importing the data from Databricks, will be provided an user-friendly names to the tables and fields, so the end user can get a good understanding of the fields used. Also, will be developed some simple measures that weren´t built in the gold layer phase to feed some visualizations. In that phase will be added the date and time dimensions that will allow filter the dashboard data in a temporal perspective and, the filter fields will be sorted based on the ID´s, COD´s or integer fields so the data will appear in the correct order when used in the

visualizations. After all this steps, it will be ready to build the visualizations that will allow us to get insights try to answer to the start questions.

4. RESULTS AND DISCUSSION

This chapter presents and discusses the results of the analysis built around the Airbnb dataset for the cities of Lisbon and Porto made available in September 2024. The listings and reviews refer to the last 12 months of data and the calendar addresses 365 days in the future.[8]

The efficiency of the analysis carried out was assessed from different angles: by exploring listings, reviews and seeking in particular, a more in-depth understanding of hosts, properties and their reviews interactions. Several techniques of visual representation were used to demonstrate the important aspects and dependencies of the data given: the number of listings per host, types of properties, types of rooms and the reviews' distribution.

The results are analyzed to be able to explain the significance behind the observed tendencies such as the effects of the changing property room types on the popularity of the listings and the number of reviews. In addition, was explored how listing features including the minimum number of nights and prices affect user and review scores engagement.

Lastly, the chapter presents a comparative analysis of the listings and reviews of Airbnb from the investigated cities with an emphasis on their specific contradictory or corresponding points. Similarities and differences were noted in order to shed more light on the Airbnb industry in these areas, which made great recommendations for managing operators and their target audience.

The output dashboard from this project BI solution is presented and discussed now, as well as the respective data model. The process to achieve this data model followed the Four-Step Dimensional Design Process methodology based on the dimensional modeling of Ralth Kimball, that address the process of designing and creating a data warehouse (Kimball & Ross, 2013). This process starts by identifying the business activity that generates the data, in this project, the Portugal Airbnb activity. After this step, was necessary to define the grain of the data, the level of detail of the data, that in that project is the day for date and parish for geography. Based on the file source data, were identified four main groups of attributes that allowed to provide context to our business process and will allow the end-users to filter, group, and drill in the data for a more detailed analysis. Those attributes made up the four dimensions of our model: the date dimension allows a temporal point of view of the data and situates the data on time (when?); the geography dimension provides the localization of the data (where?) and the host dimension provides information regarding the host characteristics; and the Listings dimension is made up of listing attributes that allow for example characterize listings from availability or listings content perspective. This was the third step of the process of defining the data model. Lastly, the fourth step followed was to identify the facts that allow determining the measurable metrics or quantities associated with the business activity, here, the Fact_Listings has all transactional data that will allow performing the listings analysis from

[8] https://insideairbnb.com/data-assumptions/

a property and host point of view. On the other side, Fact_Listings_Reviews has data that provides information about the reviews and that will allow an analysis of the customer's satisfaction with the Airbnb listings.

To wrap it up, the final data model is made up of two fact tables - Fact_Listings and Fact_Listings_Reviews and four dimensions - Dim_Host, Dim_Listings, Dim_Geography, and Dim_Date (figure 4.1).

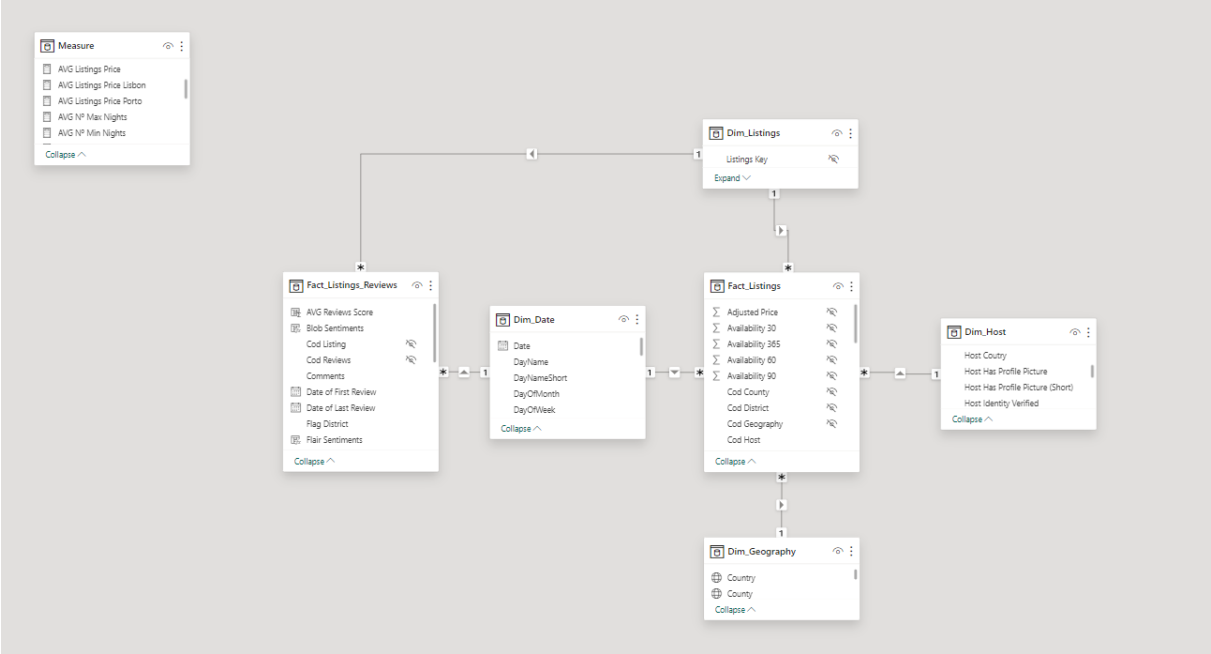

Figure 4.1 – Power BI final data model

4.1. REPORT MAIN MEASURES – "GOLD LAYER"

The table 5.1. describes the main measures used to feed the report visuals. The number of the business questions defined on sub chapter 4.1.3, and that are answered by the measures is presented on the "Related business question No." column in the table 5.1.

Table 4.1 – Measures description

Measure	Description	Related business question No.
AVG Bedrooms	Calculates the average number of rooms per listing. This metric helps you understand the average housing capacity of listed properties.	1

AVG Beds	Calculates the average number of beds per listing, providing additional insight into the average accommodation capacity offered.	1
AVG Listing Capacity	It determines the average accommodation capacity of listings, offering an estimate of the number of guests that can be served.	1
AVG Listings Price	Calculates the average price per ad. This average value gives an indication of the average cost of available properties.	9
AVG Listings Price Lisbon	Calculate the average price of ads in Lisbon. It helps to compare the average price of ads in Lisbon with other regions.	9
AVG Listings Price Porto	Calculates the average price of ads in Porto. Compare average ad prices in Porto with other regions	9
AVG Nº Max Nights	Calculates the maximum number of nights that, on average, a listing allows to stay. This metric indicates the flexibility of hosts regarding the length of stays	1, 5
AVG Nº Min Nights	Calculates the minimum number of nights a guest should book, on average. This metric helps you identify hosts' minimum stay restrictions.	1, 10
Total Listings	It counts the total number of distinct advertisements, providing an overview of the property supply in the database.	2, 3, 5, 6, 7, 8, 9, 10, 14
Total Listings Lisbon	Counts the total number of distinct ads in Lisbon, allowing regional comparisons.	2, 3, 4
Total Listings Porto	Counts the total number of different ads in Porto, helping to compare the number of offers with other areas.	2, 3, 4
Total Negative Blob	Counts the number of reviews classified as negative by the TextBlob algorithm. Allows an analysis of the amount of negative feedback to assess customer satisfaction.	17
Total Negative Flair	Counts the number of reviews classified as negative by the Flair algorithm. It provides an alternative perspective on negative feedback based on another algorithm.	17
Total Negative Vader	Counts the number of reviews classified as negative by the Vader algorithm. Complements sentiment analysis by	17

	offering additional insight into negative feedback.	
Total Neutral Blob	It counts the number of reviews considered neutral by TextBlob, giving an indication of moderate feedback.	16
Total Neutral Vader	It counts the number of reviews considered neutral by Vader, helping to verify the amount of feedback without a strong emotional charge.	16
Total Positive Blob	It counts the number of reviews classified as positive by TextBlob, helping to assess the level of guest satisfaction.	15
Total Positive Flair	It counts the number of reviews rated positive by Flair, providing a second opinion for customer satisfaction.	15
Total Positive Vader	Counts the number of reviews classified as positive by Vader, completing the satisfaction analysis with another perspective.	15
Total Reviews	It counts the total number of reviews, giving you an overview of the volume of guest feedback.	11, 12, 13, 14

4.2. REPORT RESULTS

The listings data used to feed this dashboard are distributed by Lisbon and Porto. The figure 4.2 presents the first page from the report and analyzes the listings as to location, capacity, average of number of nights and availability. This page can be filtered by date and region.

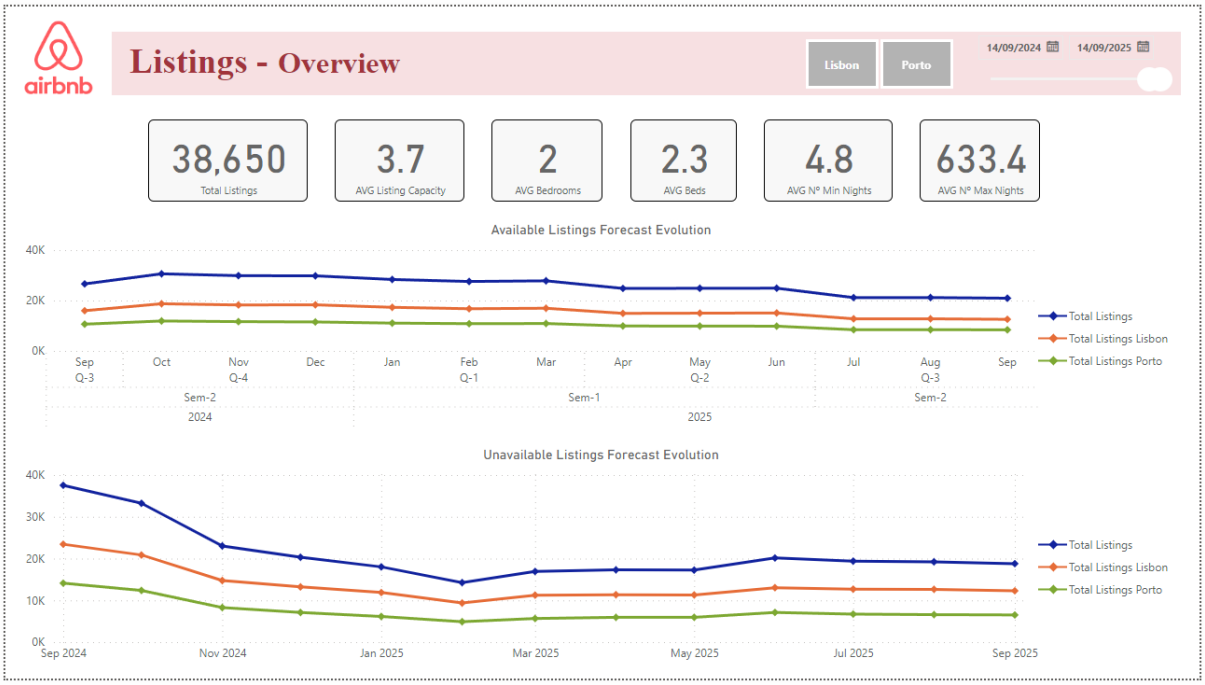

Figure 4.2 – Listings Overview

In the figure 4.3 are presented six key metrics that allow get information to answer the business question “How many listings are available, and what are the main averages for capacity, bedrooms, beds, and minimum/maximum nights?”. The card format was chosen to highlight key metrics clearly. This type of visualization draws attention to the main numbers, making it easy to read and understand the most important information at a glance. This visual allow conclude that the scope of the data are made up by a total of 38,650 listings. In average each listing accommodates 3.7 people, there are 2 bedrooms and 2.3 beds. The minimum and maximum number of nights that the people can rent the listings is in average 4.8 nights and 633.4 nights respectively. That suggests that Airbnb listings are available for short and long rental.

Figure 4.3 – Listings Key Metrics

Figure 4.4 and figure 4.5 allow get information to answer business questions “How does the available listings forecast changes through time for Lisbon, Porto and generally?” and “How does the prognosis for unavailable listings change in Lisbon, Porto, and overall, in the future?”.

The line chart has been utilized in depicting the changes over a period, and this is the best fit in showing trends and variations. The use of different colours makes it possible to understand total listings and lists for each city on a seasonably easier basis to see patterns and trends within the seasons. The blue curve shows the total number of available listings which settles almost rigidly with a small dip at the earlier part of 2025. Lisbon (orange line) and Porto (green line) exhibit similar behaviour, but with lower volumes than the overall numbers.

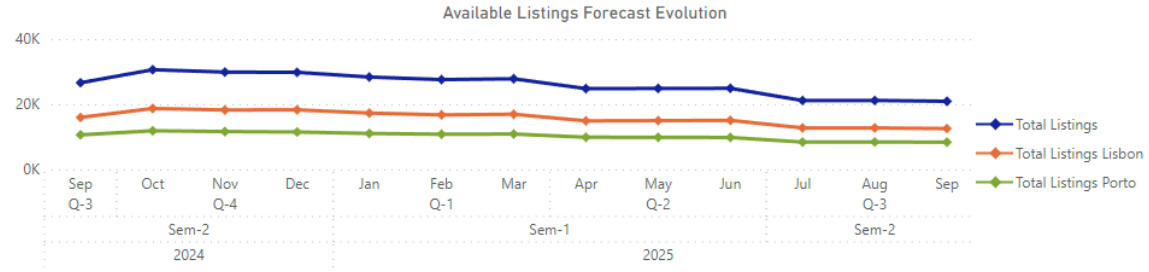

Figure 4.4 – Available listings forecast evolution

A graph that exhibits the passage of time was used once again to represent the changes of unavailable listings over a certain period. As already pointed out, this type of graph is appropriate for displaying and analyzing the trends and variations over time in order to establish seasonal variations or other reasons that may contribute to the availability of listings. The number of listings that are unavailable for booking tends to reduce with time and more so towards the end of 2024. The pattern is the same across the cities but Lisbon has more variations than Porto.

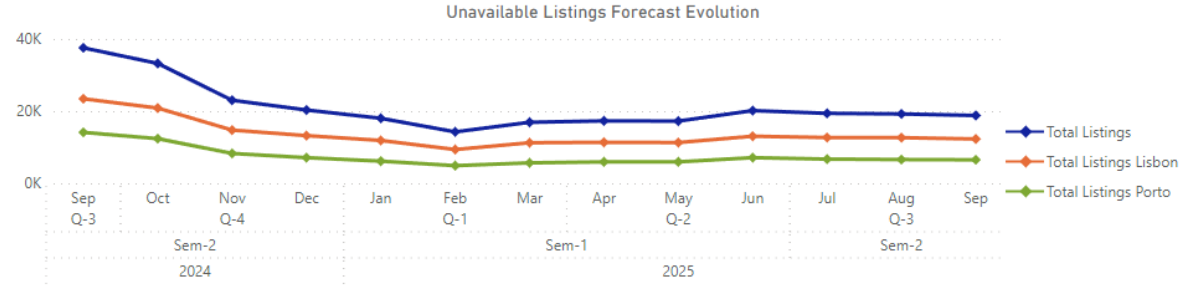

Figure 4.5 - Unavailable listings forecast evolution

Regarding figures 4.4 and 4.5, note that to achieve these results was taken into account that is only necessary to exist one day in a specific month with a record value equal to 1 or equals to 0, in the field available, respectively to be considered to the count available or to the count unavailable. This suggests that, unlike what can be thought, is not necessary for the listing to have 1 or 0 in the field available from the calendar table to be considered available or unavailable.

Figure 4.6 analyzes listings as properties and their hosts. That view can be filtered by date, district, host and host country. This page allows analyze listings top 10 hosts, listings by property type, capacity, host picture, room type and short vs long rental.

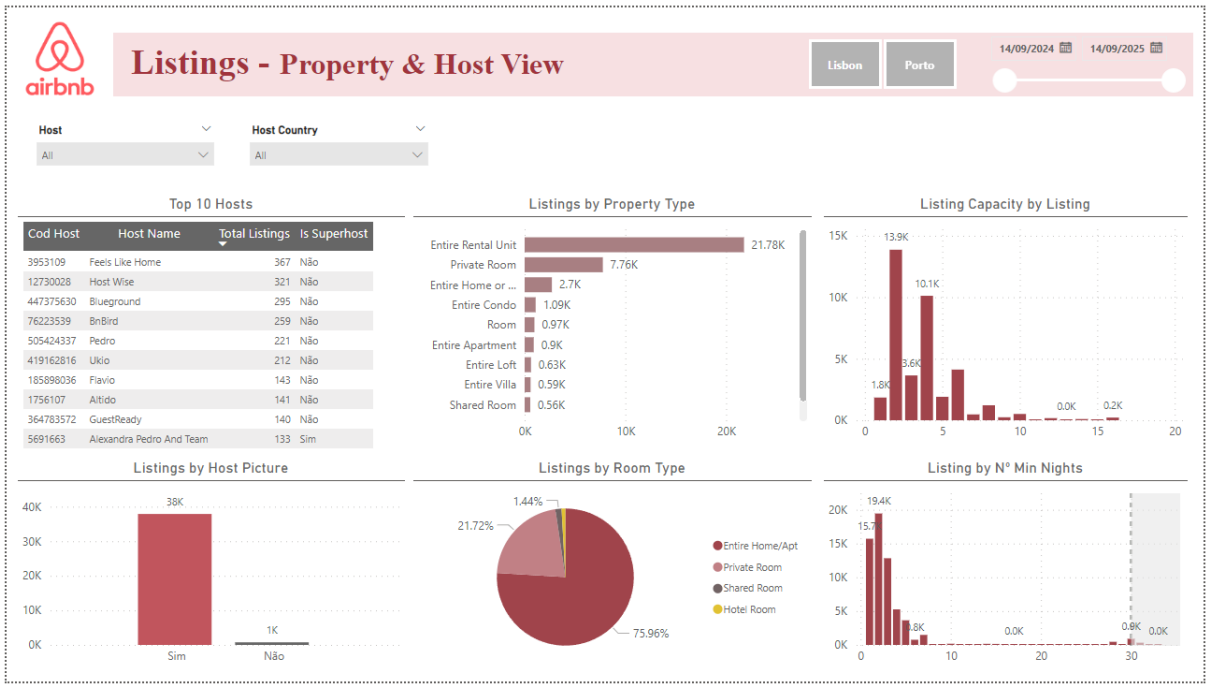

Figure 4.6 – Listings Host View

For answer the business question “Which users are ranked in top 10 and have maximum number of listings? To add, are they super hosts too?” was built a table with the top ten hosts (figure 4.7). These data were presented in the form of a simple table because it provides the best clarity and understandability in ranking the hosts by total listings and their super host status which is best suited for tabular data. The highest-ranked host is "Feels Like Home" with 367 listings, while the next in line is "Host Wise" with 321. Outsider attaches any of the top hosts except "Alexandra Pedro and Team", who has 133 listings as a super host. Unlike what was expected, the hosts with a higher number of listings are not super hosts, which suggests that are not the ones with best performance.

Top 10 Hosts

Cod Host	Host Name	Total Listings	Is Superhost
3953109	Feels Like Home	367	Não
12730028	Host Wise	321	Não
447375630	Blueground	295	Não
76223539	BnBird	259	Não
505424337	Pedro	221	Não
419162816	Ukio	212	Não
185898036	Flavio	143	Não
1756107	Altido	141	Não
364783572	GuestReady	140	Não
5691663	Alexandra Pedro And Team	133	Sim

Figure 4.7 – Top 10 Hosts

Figure 4.8 presents listings by type of property allowing answer to the business question “How is the available listings distributed across property types?”. The horizontal bar of chart is such that it is easy to rank and draw comparisons different types of property where length is used to show the amount and it is easy to see that the entire rental units are the highest. The chart allows conclude that the most common listings type is ‘Entire Rental Units,’ totaling to 21.78K,

followed by Private Rooms at 7.76K. The other categories include much fewer units such as Entire Home or Apartment at 2.7K and Shared Rooms at 0.56K.

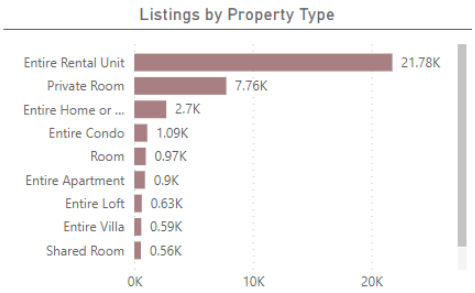

Figure 4.8 – Listings by property type

For answer the business question "How are the listings dispersed in terms of capacity?" was built the bar chart present in figure 4.9. This chart is most appropriate because it is a frequency distribution and each bar represents the number of the listing at that particular capacity. The majority of the listings are geared toward 1-5 people, with more listings concentrated at 2 and 4 guests, while any capacity of 15 and above is negligible.

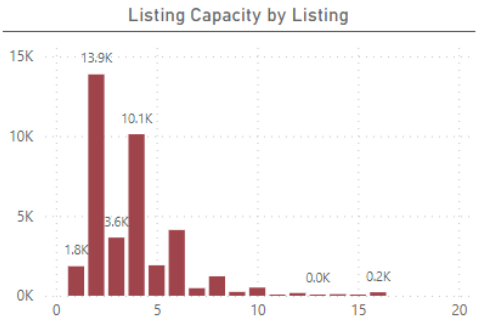

Figure 4.9 - Listing capacity by listing

Figure 4.10 shows the listings distribution by host picture allowing answer to the business question "What's the total count of listings made by hosts who uploaded their profile pictures?". The bar graph shows the difference between two types of listings very well, which makes for a very nice presentation of this type of two-dimensional data. The majority of photographs hosts have 38K listings in total while non-photos have only 1- 1K.

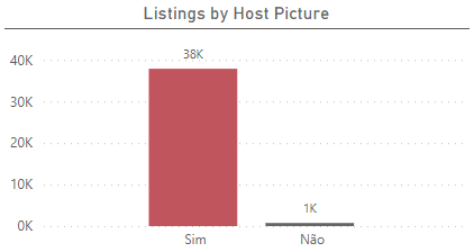

Figure 4.10 – Listings by host picture

Looking in figure 4.11 it´s possible conclude that the share of “Whole House/Apartment” listings is 75.96%, while that of “Private Rooms” is 21.72%. Less than 2% belong to “Shared Rooms” (1.44%) and “Hotel Room” (0.87%). This chart was built to answer the business question “How are the listings distributed across different room types?” and was used because it was easy to discern the relative size of each segment within the unit and the fact that whole homes or apartments occupied the greatest portion in all listings, a pie chart was deployed.

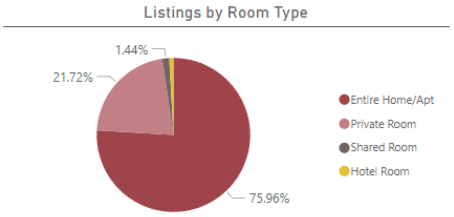

Figure 4.11 - Listings by room type

Figure 4.12. represents the minimum number of nights that the customers rent the Airbnb listings and answer to the business question “What is the number of listings that have been posted based on the number of minimum nights that are required?”. An A histogram has been employed in order to portray this information out in the best possible manner, and such information is the distribution of the minimum night’s requirement across the listing. There are many listings which have a minimum requirement of 1 to 5 nights, where 19.4K listings have a stay of 1 night, followed by 15.7K who have a stay of 2 nights minimum. However, the trend drops significantly for minimum stays of 10 nights and above.

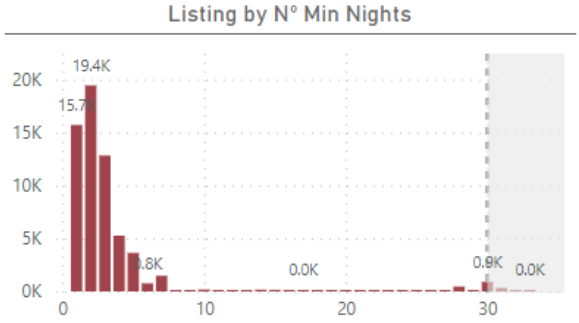

Figure 4.12 - Listings by minimum number of nights

Figure 4.13 analyses the listings reviews trough the representation on the scores given by the customers, the reviews evolution along the time and taking in account factors like the price, room type and number of reviews.

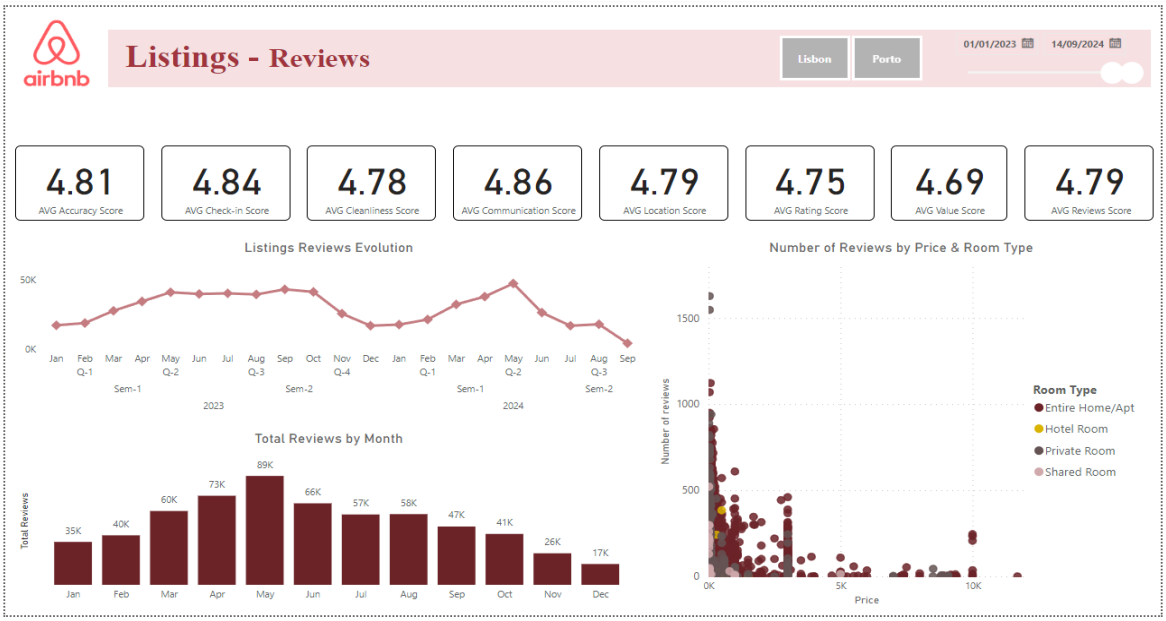

Figure 4.13 - Listings Reviews View

Analyzing figure 4.14 can be concluded that the factor that the customers most valorize in the listings is the communication, followed of check-in and accuracy. The card visual was chosen to display these key performance indicators (KPIs) for each review score. It's a clear and direct way to showcase important metrics in a dashboard without overwhelming the viewer. That visual allows answer to the business question “What are the average review scores across different categories?”.

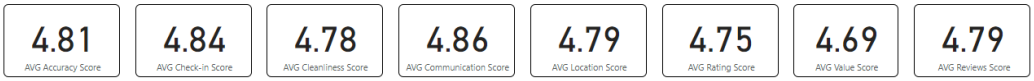

Figure 4.14 - Listings average scores

Figure 4.15 analyses how have the number of reviews evolved over time from January 2023 to August 2024. The line chart helps to visualize the trend in review activity over time, making it easy to observe changes in review numbers across the months. It is the ideal type of chart for demonstrating temporal data. There is a steady increase in the number of reviews until June 2024, with a peak around May and a noticeable decline afterwards, particularly from July to August.

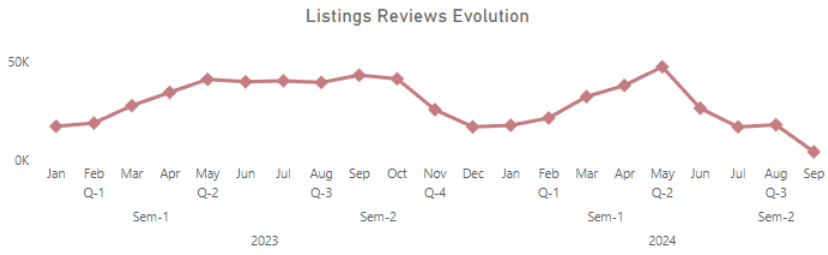

Figure 4.15 - Listings reviews evolution

Figure 4.16 allows analyze how do reviews vary by month. A bar chart was chosen because it illustrates the comparison of aggregate reviews by month in an easy way. It helps to easily identify the months with the greatest and least number of reviews, thus showing sharp peaks and deep troughs in activity. In terms of number of reviews, May 2023 emerged at the top with 89K reviews, followed by April with 73K reviews. The least number of reviews was registered in December standing at 17K.

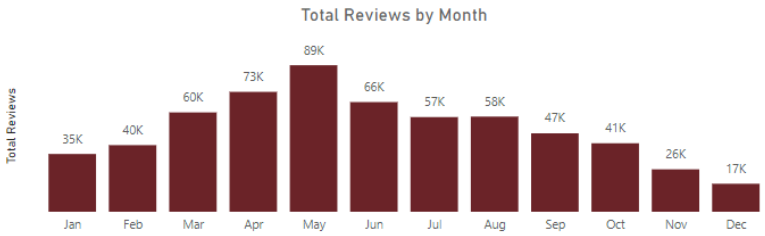

Figure 4.16 - Total reviews by month

To analyze how are the number of reviews distributed by price and room type, was chosen a scatter plot that contrasts Price with Number of reviews and room type. The purpose of the scatter plot was to show the price and the number of reviews for each of several room types. It makes sense for the audience because it shows the relationship between price and the number of reviews, while differentiating room types with different colours. The majority of the reviews are focused on the listings that price below 5K especially for listings tagged as "Entire Home/Apt" with very few reviews for listings priced higher. "Hotel Room" and "Shared Room" categories have very few reviews as compared to "Entire Home/Apt" and "Private Room" reviews (figure 4.17).

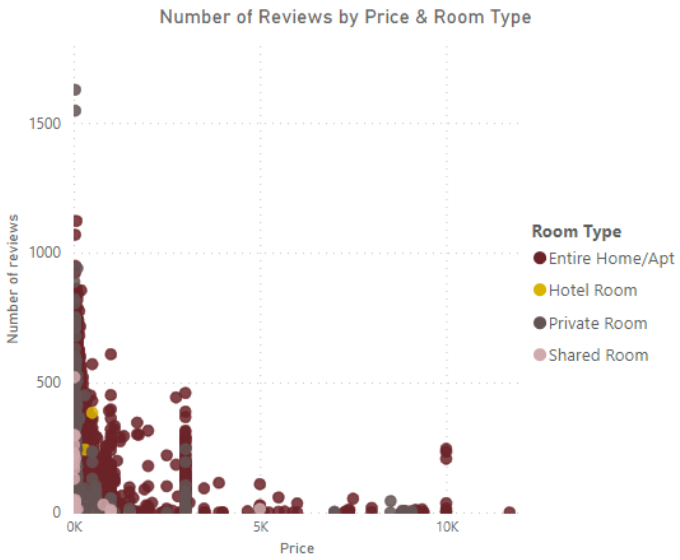

Figure 4.17 – Number of reviews by price and room type

The last page of the output report of this master thesis project makes use of a sentiment analysis to analyze the listings reviews along the time (figure 4.18).

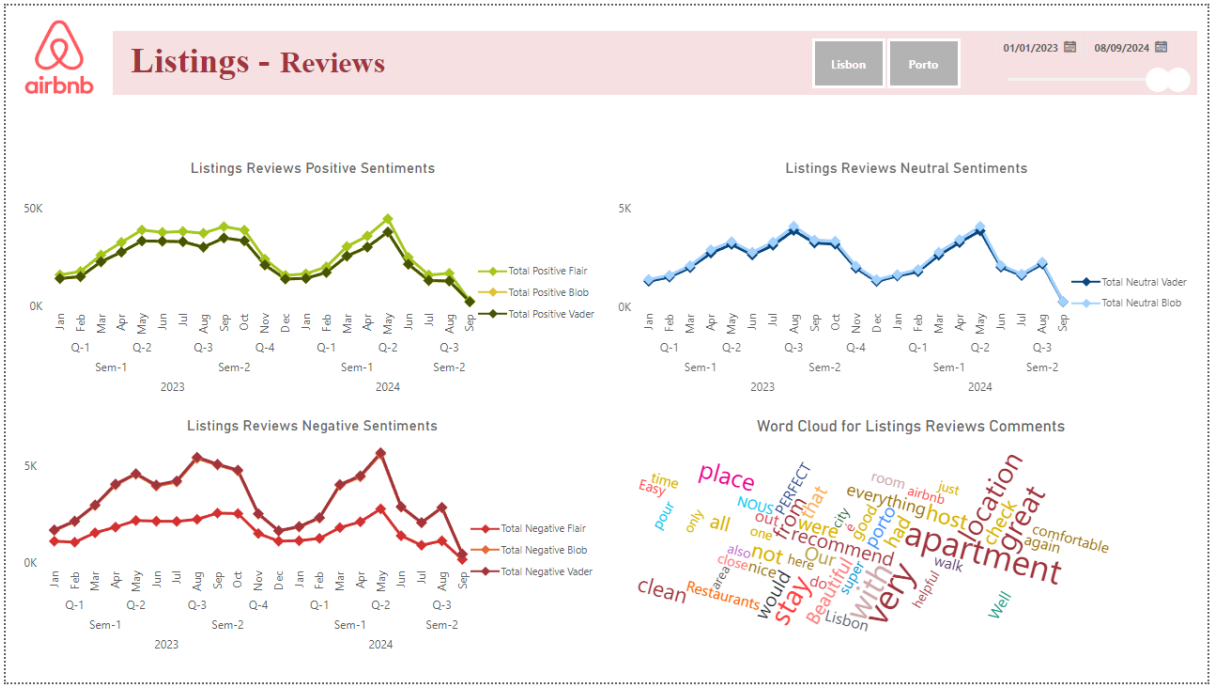

Figure 4.18 - Listings Reviews View – Sentiments

The figure 4.19 allows to analyze listings reviews positive sentiments by comparing 3 sentiment algorithms allowing answer to the business question “How do positive sentiments in Airbnb listings reviews fluctuate over time, and are there identifiable seasonal patterns that could influence customer satisfaction or operational strategies?”. The line chart effectively compares the three sentiment analysis methods over time, showing how each tracks positive sentiment trends. The inclusion of quarterly (Q1, Q2, etc.) and semiannual (Sem-1, Sem-2) divisions aids in identifying potential seasonal patterns in sentiment. The use of different colors for each line allows for easy distinction between analysis methods, making it straightforward to observe consistency or variation among them. The graph displays the total positive sentiments in Airbnb listing reviews over time, from 2023 to September 2024.Three sentiment analysis methods—Flair, Blob, and Vader—are represented, each tracking positive review sentiments. There is a clear upward trend in positive sentiments during the first half of 2023, peaking in June and remaining high through October. In 2024, positive sentiments rise again, peaking in May, followed by a noticeable decline through September.

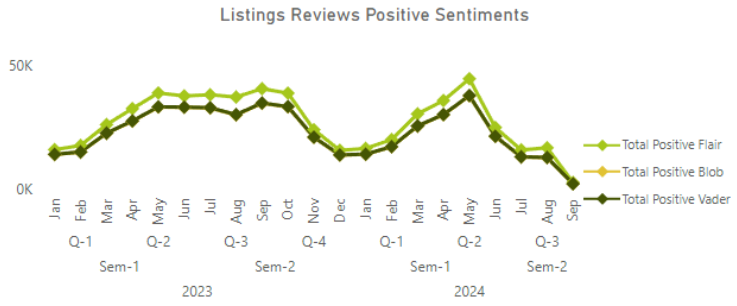

Figure 4.19 – Listings reviews positive sentiments

For analyze listings reviews neutral sentiments and answer to business question “How do neutral sentiments in Airbnb listings reviews vary over time, and are there seasonal patterns that may indicate periods of mixed or neutral customer feedback?” was built another line chat (figure 5.20). The line chart effectively compares the two sentiment analysis methods, illustrating how each one tracks neutral sentiment trends over time. Quarterly (Q1, Q2, etc.) and semi-annual (Sem-1, Sem-2) divisions provide a clear view of potential seasonal changes in customer feedback. Distinct colours and line styles for each analysis method make it easy to distinguish between the two, helping identify periods where neutral feedback is consistently high or low. The graph shows the total neutral sentiments in Airbnb listing reviews over time, from 2023 to September 2024. Two sentiment analysis methods—Vader and Blob—are tracked, each monitoring neutral review sentiments. In 2023, there is a steady increase in neutral sentiments, reaching a peak around July, followed by a gradual decline until November. In 2024, neutral sentiments rise again, peaking in May, before sharply decreasing through September.

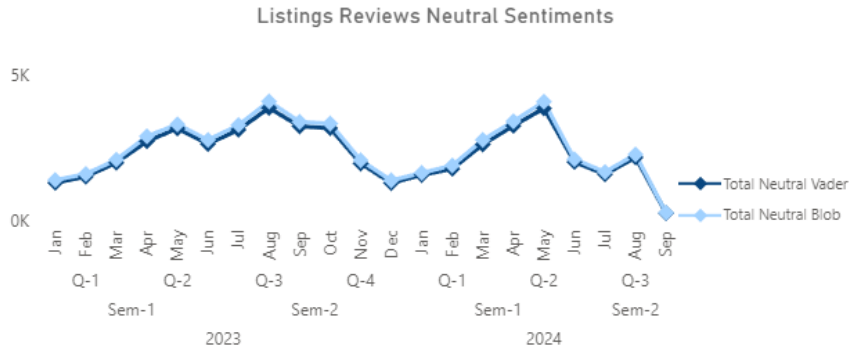

Figure 4.20 – Listings reviews neutral sentiments

Lastly, figure 4.21 shows the evolution of the negative sentiments regarding listings, allowing to answer to business question “How do negative sentiments in Airbnb listings reviews evolve over time, and are there specific periods with higher negative feedback that may require operational improvements or customer support adjustments?”. The line chart effectively highlights trends in negative sentiments over time, comparing three sentiment analysis methods for a comprehensive view. The inclusion of quarterly (Q1, Q2, etc.) and semiannual (Sem-1, Sem-2) markers helps identify patterns and seasonal shifts in negative feedback. Distinct colors for each sentiment method make it easy to differentiate between them, helping pinpoint months with notable increases or decreases in negative reviews, which could indicate areas for potential service improvement. The graph displays total negative sentiments in Airbnb listing reviews from 2023 to September 2024. Three sentiment analysis methods—Flair, Blob, and Vader—are tracked, each measuring negative sentiments in the reviews. Negative sentiments increase steadily in early 2023, peaking around June, and then gradually decline until October. In 2024, negative sentiments rise sharply again, reaching another peak in May, followed by a decline through September.

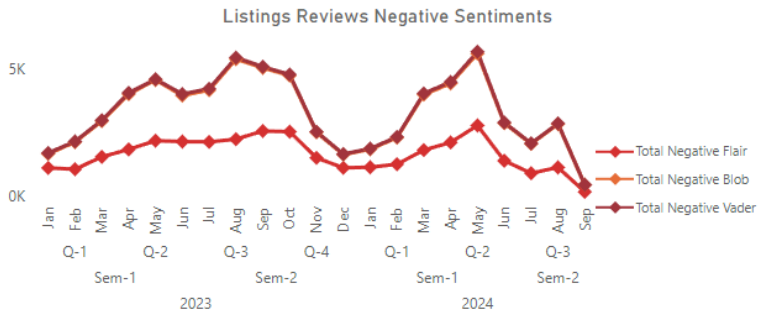

Figure 4.21 – Listings reviews negative sentiments

Figure 4.22 allows get insights to answer business question “What are the most used words in comments?”. Using a word cloud give us a Quick Overview of Key more used. This kind of visual help us exploring sentiments. This is useful to find or identify trends. Looking into the word cloud it´s possible can identify the main themes: larger words generally indicate the most recurring themes in the text. This can help to understand the central theme of a set of texts or opinions. In that case can be seen that the most cited words in the comments field were “location”, “apartment”, “perfect”, “beautiful”. That words can suggest that the customers are satisfied with the location of the Airbnb apartments in Lisbon and Porto.

Figure 4.22 - Word cloud for listings reviews “Comments”

Lastly, the table 4.2 present a summary table of the business questions defined for this project and the respective obtained conclusions after the project development. That table is organized by business question number, the process followed and the main conclusions.

Table 4.2 – Business questions conclusions

Business Question (BQ)	Figure	Insights/Conclusions
BQ1	Figure 4.2 - Power BI - Statistics calculated from the listing fact table.	Lisbon: 24,204 listings, Porto: 14,446 listings. Averages: 4 people capacity, 2 bedrooms, minimum of 2 nights.
BQ2	Figure 4.2 - Power BI - Time-series forecasting.	Lisbon: decrease; Porto: decrease by 2025.
BQ3	Figure 4.4 - Power BI - Forecasting with availability data (Calendar).	Lisbon and Porto show stability in seasonal periods.

BQ5	Figure 4.7 - Power BI - Rankings using Listings table.	Top host in Lisbon: 295 listings, isn't super host. Porto: Top host has 316 listings, also isn't a super host.
BQ6	Figure 4.8 Power BI - Distribution charts by property type.	Majority are apartments in both Lisbon (75%) and Porto (70%).
BQ7	Figure 4.10 - Power BI - Filter on profile pictures.	Lisbon: 24k listings, Porto: 14.2k listings.
BQ8	Figure 4.11 - Power BI - Charts segmented by room type.	Lisbon: 60% entire homes, 30% private rooms. Porto: 65% entire homes, 25% private rooms.
BQ9	Figure 4.9 - Power BI - Capacity analysis from Listing's data.	Lisbon: Most listings accommodate 2-4 people. Porto: Similar pattern.
BQ10	Figure 4.12 - Power BI - Filters on minimum nights field.	70% of listings in both cities require a minimum of 2 nights.
BQ11	Figure 4.14 - Power BI - Averages calculated on Reviews data.	Lisbon: 4.5 stars average; Porto: 4.6 stars average across categories.
BQ12	Figure 4.15 - Power BI - Trend analysis over time.	Lisbon: Steady growth; Porto: Significant increase during summer months.
BQ13	Figure 4.16 - Power BI - Monthly analysis of Reviews.	Peaks in May for Lisbon and May, June, July and August for Porto.
BQ14	Figure 4.17 - Power BI - Scatter plots combining price and reviews.	Lower-priced entire homes get more reviews.
BQ15	Figure 4.19 - Python (VADER, TextBlob, Flair) - Sentiment analysis.	Positive sentiments peak during summer months in both cities.
BQ16	Figure 4.20 - Python (VADER, TextBlob, Flair) - Sentiment analysis.	Neutral sentiments peak during May in Lisbon and between May and August for Porto.
BQ17	Figure 4.21 - Python (VADER, TextBlob, Flair) - Sentiment analysis.	Negative sentiments are slightly higher during winter months, often related to heating or weather.
BQ18	Figure 4.22 - Power BI - Word cloud analysis.	Common words: "apartment", "location," "clean," "host,", "comfortable", "recommend". Positive attributes dominate.

4.3. DISCUSSION

This work developed a prototype for an end-to-end Business Intelligence and Data Analytics process, focusing on the short-term rental market in Portugal, using data from the Inside Airbnb platform. This model, based on Kimball methodologies and Medallion architecture, demonstrates how modern tools like Databricks and Power BI can be used to extract valuable insights from large volumes of data.

The results achieved offer a significant contribution to the area of data analysis applied to tourism, with an impact on three main points: Decision-making support; Replicable methodology; and Sentiment analysis integration. The developed dashboard provides insights into the performance of listings on Airbnb, allowing you to identify customer behavior patterns, such as location preferences, room type, and seasonality of demand. These results can be used by owners to adjust pricing, availability, and marketing strategies. The model created can be adapted to other regions or industries, serving as a practical guide for organizations interested in implementing BI and data analysis systems. The inclusion of sentiment analysis based on customer comments adds a new layer of understanding about the customer experience, something still little explored in the literature analyzed.

The work directly relates to the studies analyzed, such as that of Chen (2004) and Baggio & Caporarello (2005), which identify BI as an essential component of decision support systems (DSS). The prototype developed in this project reinforces this perspective, demonstrating how BI systems can be used to transform raw data into useful information in the tourism sector. Additionally, the challenges faced in this project, such as the integration of data from different formats and the standardization of textual data, are in line with the problems described by Bustamante et al. (2020) in the development of systems such as ETIHQ. However, the use of modern technologies, such as Databricks, proved to be efficient in overcoming many of these obstacles, something that could be further explored in future studies.

Finally, studies such as that by Vajirakachorn & Chongwatpol (2017) reinforce the importance of the practical application of BI in tourism. This work advances by demonstrating how insights into customer behavior and preferences can be obtained from open data, highlighting the model's potential to promote customer satisfaction and more informed strategic decisions.

5. CONCLUSIONS

This project proposed the development and implementation of an end-to-end Business Intelligence (BI) solution applied to the hospitality industry. That solution included an analysis of Airbnb data in Portugal, Lisbon and Porto regions. This project was motivated by the increased reliance on data-driven decisions in our modern digital economy, particularly with regards to businesses in hospitality and tourism. Moreover, the use of advanced BI tools in order to enable structured data processes show how valuable insights can be drawn from large datasets that commitment to more strategically informed choices.

The BI architecture developed in this project utilized the Kimball Life Cycle methodology focusing on dimensional modeling and phased project delivery. Medallion Architecture was selected to provide a scalable and stable solution to transform raw data ingestion to final analytic reports/dashboards. Bronze (initial): Raw data not cleaned or transformed. Silver: Cleaned and organized in a way from which you can make sense of its. gold — processed to be in the format that is optimal for analysis at time phase. In this project was implemented the bronze layer to ingest and store the raw data from csv files stored on azure storage container; and the silver layer to perform the joins between bronze tables and perform small transformations. Due time limitation, was not implemented the gold layer and the Key Perform Indicators KPIs were developed in Power BI.

This analysis utilizes data from the Inside Airbnb datasets, which include key characteristics of listings and reviews in addition to neighborhood metrics. The data processing and transformation steps took place in Microsoft Azure environment using Databricks, where Power BI was used for the final visualization of the data. This combination enlisted us to deliver an integrated end-to-end BI solution, which would not only be real-time but also host great insight. The study further evaluated key metrics like — Property availability within Airbnb trends, Pricing, Customer reviews and High demand Host performance to enable customer as well business data driven decisions.

For businesses, key findings include where Airbnb is most active in Lisbon to Porto property issues and what guests seek from different types of accommodation; through average lengths of stay. This added layer of intelligence has the potential to be helpful for hosts, policymakers and even investors in gauging consumer sentiment towards ridesharing services, as well as satisfaction with respect to an entire city. This knowledge can direct decisions make — for example, how to price properties, what types of services should be offered or even where the best locations are for new investments in short-term rentals.

Moreover, the project adds to scholarly knowledge of advanced application of BI systems in tourism and hospitality industry. Operational data sources such as reviews, location and user generated content are now likely sharing the same cloud environment for a holistic view of customer behavior and market dynamics. Such a methodology and architectural approach

might be copied elsewhere, or possibly used to scale it up for comparing Airbnb data against traditional hotel market data.

However, the project had some limitations in spite of favorable results. Some features were less well developed, due to time limits and the slower access to cloud resources. Also, the exclusive consideration for counts from only two Portuguese cities: Lisbon and Porto mean that the results may not apply to other contexts. Future research could examine additional regions or the replication of these findings across different European markets. Additional future work could also explore the utilization of machine learning tools to improve advanced prediction-analytics, such as market intent forecasting or trend-predictions for new markets.

To conclude, this thesis explores how BI systems can be a transformative force in the hands of hospitality practitioners to convert raw data into knowledge. This solution opens up the door to creating a robust, durable process for working with Airbnb data at scale and also highlights how cloud computing, data warehousing, and analytical capabilities will become table stakes when building an unfair advantage in your short-term rental business. BI will enable the stakeholders across different spheres in an Airbnb ecosystem to focus and streamline their decision-making capabilities so that they can impart high-quality customer satisfaction, great business operating efficiencies which are prime extractors of tourism industry sacrosanct — Growth & Sustainability. Future studies could explore additional machine learning models to refine sentiment accuracy, particularly in multilingual contexts.

6. LIMITATIONS AND FUTURE RESEARCH

6.1. LIMITATIONS

Although a complete BI solution was developed successfully for Airbnb data in Portugal, some limitations were faced while implementing the project which needs to be considered.

The first limitation was the limited dataset. The analysis only used Airbnb data from Lisbon and Porto. Even so, these places are large Portuguese cities; as such can be only concluded for this scope (meaning no comparison in other regions/cities). That means the results may not capture all of Airbnb activity from the country, yet it can still serve as a roughly accurate gauge on nationwide trends for spring season rentals.

The second limitation faced is related to Listings ETL process. After the listings ETL process development was identified an issue related to some of geographical fields transformations. Due to time constraints was not possible to correct the process and it made it unfeasible to get an accurate geographical representation of the data.

Technical and Resource Limitations is the third limitation. Since the project ran on free-tier cloud platforms like Databricks Community and Microsoft Azure it had limits. For example, constraints on the size of databases that could be processed and stored meant it was not possible to exploit its full capacity, while only having intermittent access to computational resources presented a practical difficulty. In addition, there was a lot of tables or processes that needed to be recreated due cluster resets in Databricks Community which slowed down progress even more.

The fourth limitation is related to updates of the Data. In this project the data are static and accessed manually from Inside Airbnb web page. That it's not practice since the data are made available quarterly and someone needs to download the csv´s, store it in the storage, run the ETL process and refresh the dashboard. In future developments would be interesting to develop a notebook that calls Airbnb API and runs it into a pipeline triggered quarterly to get the updated data. Consequently, would be needed adjust the processes by creating pipelines to run the ETL automatically and schedule the dashboard refresh.

Lastly, the limited focus on comparative analysis demonstrated another limitation. The project still much focused on Airbnb data and not included some good comparison with traditional hoteling or other short renting platform. While this would give great information on the competitive situation, this was not included in order to finalize it faster with less resources.

6.2. FUTURE RESEARCH

Given the limitations outlined above, several opportunities for future research and system enhancements have been identified.

Future research could expand geographic scope, that is, could address the issue of regional bias by broadening the geographic area to other regions or cities in Portugal, such as Algarve, Madeira and some inland smaller scale region. Secondly, extension to other countries may yield a more complete understanding of Airbnb's global foot print and how region-specific trends differ.

Incorporation of if machine-learning techniques to improve predictive power of the BI solution. This could mean making better predictions on booking trends, price changes or customer preferences with the help of forecasting models fitted to past data. In the case of customer reviews, sentiment analysis could be performed to understand guest satisfaction better.

This solution processes automatically update data in the source (data ingestion and analysis) which can offer more instant insights to the hosts/property managers. This would entail the incorporation of streaming pipelines and richer clouds that could process automatically updates — allowing users to respond faster to market changes or customer movement.

Similar analysis between the data of Airbnb listings and traditional hotels might reveal some interesting facts about types of customers & market competition. In addition, future work might explore how Airbnb competes or complements the traditional hospitality industry by using data on Pricing, Occupancy Rates, and Guest satisfaction as proxies(factors).

Further iterations of this project could include building web-friendly interfaces or interactive dashboards that future Airbnb hosts, policy-makers or even potential investors can use to analyze in comprehensive ways all of it strands together. Different user groups may also benefit from more tailored insights, for instance through scenario analysis or custom filters.

Another suggestion is cloud infrastructure optimization. One of the outcomes from this study is that moving to a more robust cloud infrastructure (higher tiers in newer cloud services, Databricks or Azure) would address imposed technical limitations. Such a system would be scalable, process data more quickly and easily lend itself to extensibility for doing things like advanced analytics or building tailored machine learning models.

Perform a longitudinal analysis. Future work may leverage a time-series Airbnb data set over many months or years to examine long-term trends in the business as well as how external factors such as the COVID-19 pandemic, changes in tourism demand etc. affect listing performance on Airbnb platforms (Supply and Price effect).

Overall, this project shows that a BI solution on Airbnb data can be developed in Portugal successfully but there is still more can be done to make the tool serve us even better such as expanding features or adding advanced analytics while improving technical infrastructure for all sources and return higher insights.

BIBLIOGRAPHICAL REFERENCES

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment Analysis in Tourism: Capitalizing on Big Data. *Journal of Travel Research*, *58*(2), 175–191. https://doi.org/10.1177/0047287517747753

Alqarni, A. A., & Pardede, E. (2012). Integration of Data Warehouse and Unstructured Business Documents. *2012 15th International Conference on Network-Based Information Systems*, 32–37. https://doi.org/10.1109/NBiS.2012.59

Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021). *Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics*.

Azeroual, O., Saake, G., Abuosba, M., & Schöpfel, J. (2019). Solving problems of research information heterogeneity during integration – using the European CERIF and German RCD standards as examples. *Information Services & Use*, *39*(1–2), 105–122. https://doi.org/10.3233/ISU-180030

Baggio, R., & Caporarello, L. (2005). *Decision Support Systems in a Tourism Destination: Literature Survey and Model Building*. https://www.semanticscholar.org/paper/Decision-Support-Systems-in-a-Tourism-Destination%3A-Baggio-Caporarello/8889d0fa135266d7b079efb2ae6ead1a965509de

Bardhi, F., & Eckhardt, G. M. (2012). Access-Based Consumption: The Case of Car Sharing. *Journal of Consumer Research*, *39*(4), 881–898. https://doi.org/10.1086/666376

Barnes, S. J., & Mattsson, J. (2016). Understanding current and future issues in collaborative consumption: A four-stage Delphi study. *Technological Forecasting and Social Change*, *104*, 200–211. https://doi.org/10.1016/j.techfore.2016.01.006

Barrows, C. W., Powers, T., & Reynolds, D. R. (2011). *Introduction to the Hospitality Industry*. John Wiley & Sons.

Belk, R. (2007). Why Not Share Rather Than Own? *The ANNALS of the American Academy of Political and Social Science*, *611*(1), 126–140. https://doi.org/10.1177/0002716206298483

Belk, R. W. (2013). Extended Self in a Digital World. *Journal of Consumer Research*, *40*(3), 477–500. https://doi.org/10.1086/671052

Berndt, D. J., Hevner, A. R., & Studnicki, J. (2000). Hospital discharge transactions: A data warehouse component. *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, 10 pp. vol.1-. https://doi.org/10.1109/HICSS.2000.926791

Botsman, R. (2013, November 21). *The Sharing Economy Lacks A Shared Definition*. Fast Company. https://www.fastcompany.com/3022028/the-sharing-economy-lacks-a-shared-definition

Botsman, R. (2015, May 27). *Defining The Sharing Economy: What Is Collaborative Consumption–And What Isn't?* Fast Company. https://www.fastcompany.com/3046119/defining-the-sharing-economy-what-is-collaborative-consumption-and-what-isnt

Bustamante, A., Sebastia, L., & Onaindia, E. (2019). Can Tourist Attractions Boost Other Activities Around? A Data Analysis through Social Networks. *Sensors (Basel, Switzerland)*, *19*(11), 2612. https://doi.org/10.3390/s19112612

Bustamante, A., Sebastia, L., & Onaindia, E. (2020). BITOUR: A Business Intelligence Platform for Tourism Analysis. *ISPRS International Journal of Geo-Information*, *9*(11), Article 11. https://doi.org/10.3390/ijgi9110671

Cambridge Dictionary. (2023). *Sharing economy*. https://dictionary.cambridge.org/dictionary/english-spanish/sharing-economy

Castellanos, M., Gupta, C., Wang, S., Dayal, U., & Durazo, M. (2012). A platform for situational awareness in operational BI. *Decision Support Systems*, *52*(4), 869–883. https://doi.org/10.1016/j.dss.2011.11.011

Chávez, A., Banda-Barrientos, J., & Cabanillas-Carbonell, M. (2021). *Business Intelligence, Based on the Ralph Kimball Methodology, for Decision-Making in General Management*. 643–646. https://doi.org/10.1109/ISKE54062.2021.9755430

Chen, K. C. (2004). Decision Support System for Tourism Development: System Dynamics Approach. *Journal of Computer Information Systems*, *45*(1), 104–112. https://doi.org/10.1080/08874417.2004.11645822

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Diakopoulos, N., Naaman, M., & Kivran-Swaine, F. (2010). Diamonds in the rough: Social media visual analytics for journalistic inquiry. *2010 IEEE Symposium on Visual Analytics Science and Technology*, 115–122. https://doi.org/10.1109/VAST.2010.5652922

El Moukhi, N., El Azami, I., & Mouloudi, A. (2015). Data warehouse state of the art and future challenges. *2015 International Conference on Cloud Technologies and Applications (CloudTech)*, 1–6. https://doi.org/10.1109/CloudTech.2015.7337004

Fraiberger, S. P., & Sundararajan, A. (2017). *Peer-to-Peer Rental Markets in the Sharing Economy* (SSRN Scholarly Paper 2574337). https://doi.org/10.2139/ssrn.2574337

Gallinucci, E., Golfarelli, M., & Rizzi, S. (2019). Approximate OLAP of document-oriented databases: A variety-aware approach. *Information Systems*, *85*. https://doi.org/10.1016/j.is.2019.02.004

Ganesan, D., & Kalavathy, S. (2020). *Comparative analysis of traditional RDBMS and NoSQL databases for big data applications. 12(2)*, 235–245.

Goel, P., Kaushik, N., Sivathanu, B., Pillai, R., & Vikas, J. (2022). Consumers' adoption of artificial intelligence and robotics in hospitality and tourism sector: Literature review and future research agenda. *Tourism Review*, *77*(4), 1081–1096. https://doi.org/10.1108/TR-03-2021-0138

Hamari, J., Sjöklint, M., & Ukkonen, A. (2016). The sharing economy: Why people participate in collaborative consumption. *Journal of the Association for Information Science and Technology*, *67*(9), 2047–2059. https://doi.org/10.1002/asi.23552

Heinrichs, H. (2013). Sharing Economy: A Potential New Pathway to Sustainability. *GAIA - Ecological Perspectives for Science and Society*, *22*. https://doi.org/10.14512/gaia.22.4.5

Hevner, A., & Chatterjee, S. (2010). *Design Research in Information Systems: Theory and Practice* (Vol. 22). Springer US. https://doi.org/10.1007/978-1-4419-5653-8

Hevner A. R. (2007). *A three cycle view of design science research. Scandinavian journal of information systems*. *19(2)*, 4.

Hočevar, B., & Jaklič, J. (2010). Assessing benefits of business intelligence systems—A case study. *Management*, *15*, 87–119.

Höpken, W., Fuchs, M., Höll, G., Keil, D., & Lexhagen, M. (2013). Multi-Dimensional Data Modelling for a Tourism Destination Data Warehouse. In L. Cantoni & Z. (Phil) Xiang (Eds.), *Information and Communication Technologies in Tourism 2013* (pp. 157–169). Springer. https://doi.org/10.1007/978-3-642-36309-2_14

Inmon, W. H. (1992). *Building the Data Warehouse*. John Wiley & Sons, Inc.

Jevons, D. (2015). *A fair share? The economics of the sharing economy*. Oxera. https://www.oxera.com/insights/agenda/articles/a-fair-share-the-economics-of-the-sharing-economy/

Jones, P., & Pizam, A. (1994). The International Hospitality Industry: Organizational and Operational Issues. *Journal of Travel Research*, *32*(3). https://doi.org/10.1177/004728759403200351

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., Weaver, C., Lee, B., Brodbeck, D., & Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, *10*(4), 271–288. https://doi.org/10.1177/1473871611415994

Kim, D., & Park, H. (2015). *Comparative analysis of Medallion Architecture and MapReduce for largescale data processing. Information Systems*. *60*, 1–15.

Kimball, R. (1996). *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*. Wiley.

Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. John Wiley & Sons.

Krawczyk, M., & Xiang, Z. (2016). Perceptual mapping of hotel brands using online reviews: A text analytics approach. *Information Technology & Tourism*, *16*(1), 23–43. https://doi.org/10.1007/s40558-015-0033-0

Kumar, Gupta, M., & Mittal, N. (2016). *Performance evaluation of MapReduce framework on different Hadoop distributions. 85*, 109–116.

Kumar, S., Kumar, V., & Attri, K. (2021). IMPACT OF ARTIFICIAL INTELLIGENCE AND SERVICE ROBOTS IN TOURISM AND HOSPITALITY SECTOR: CURRENT USE & FUTURE TRENDS.

Administrative Development 'A Journal of HIPA, Shimla', *8*, 59–83. https://doi.org/10.53338/ADHIPA2021.V08.Si01.04

Kutay, J. (2021, November 15). Data Warehouse vs. Data Lake vs. Data Lakehouse: An Overview of Three Cloud Data Storage Patterns. *Striim*. https://www.striim.com/blog/data-warehouse-vs-data-lake-vs-data-lakehouse-an-overview/

Lee, C. K. H., Tse, Y. K., Zhang, M., & Ma, J. (2020). Analysing online reviews to investigate customer behaviour in the sharing economy: The case of Airbnb. *Information Technology and People*, *33*(3), 945–961. Scopus. https://doi.org/10.1108/ITP-10-2018-0475

Limna, P. (2023). Artificial Intelligence (AI) in the Hospitality Industry: A Review Article. *International Journal of Computing Sciences Research*, *7*, 1306–1317. https://doi.org/10.25147/ijcsr.2017.001.1.103

Little, J. D. C. (1970). Models and Managers: The Concept of a Decision Calculus. *Management Science*, *16*(8), B466–B485.

Loria, S., Keen, P., Honnibal, M., & Yankovsky, R. (2024, November 1). *TextBlob: Simplified Text Processing—TextBlob 0.18.0.post0 documentation*. TextBlob: Simplified Text Processing. https://textblob.readthedocs.io/en/dev/

Loureiro, S., Ashfaq, M., & Berga Rodrigues, M. (2021). AI Meaning and Applications in the Consumer Sector of Retailing, Hospitality, and Tourism. In *Handbook of Research on Applied Data Science and Artificial Intelligence in Business and Industry* (pp. 291–303). IGI Global. https://doi.org/10.4018/978-1-7998-6985-6.ch013

Lup Low, W., Li Lee, M., & Wang Ling, T. (2001). A knowledge-based approach for duplicate elimination in data cleaning. *Information Systems*, *26*(8), 585–606. https://doi.org/10.1016/S0306-4379(01)00041-2

Madyatmadja, E., Adiba, C., Sembiring, D., Pristinella, D., & Putra, A. (2021). The Positive Impact of Implementation Business Intelligence and Big Data in Hospitality and Tourism Sector. *International Journal of Emerging Technology and Advanced Engineering*, *11*, 59–71. https://doi.org/10.46338/ijetae0621_07

Malde, R. (2020, July 6). *A Short Introduction to VADER*. Medium. https://towardsdatascience.com/an-short-introduction-to-vader-3f3860208d53

Martinez-Martinez, A., Cegarra-Navarro, J.-G., Garcia-Perez, A., & Wensley, A. (2019). Knowledge agents as drivers of environmental sustainability and business performance in the hospitality sector. *Tourism Management*, *70*, 381–389. https://doi.org/10.1016/j.tourman.2018.08.030

Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2017). A Big Data Analytics Method for Tourist Behaviour Analysis. *Information & Management*, *54*(6), 771–785. https://doi.org/10.1016/j.im.2016.11.011

Mnyakin, M. (2023). Big Data in the Hospitality Industry: Prospects, Obstacles, and Strategies. *International Journal of Business Intelligence and Big Data Analytics*, *6*(1), Article 1. https://orcid.org/0000-0003-3052-3112

Moitas, J., Albuquerque, J., & Mano, R. (2023). *Business Intelligence Implementation and its Impact on Decision-Making*. 20–23. https://doi.org/10.23919/CISTI58278.2023.10211744

Mooney, S. J., Westreich, D. J., & El-Sayed, A. M. (2015). Epidemiology in the Era of Big Data. *Epidemiology (Cambridge, Mass.)*, *26*(3), 390–394. https://doi.org/10.1097/EDE.0000000000000274

Musa, G. J., Chiang, P.-H., Sylk, T., Bavley, R., Keating, W., Lakew, B., Tsou, H.-C., & Hoven, C. W. (2013). Use of GIS Mapping as a Public Health Tool—From Cholera to Cancer. *Health Services Insights*, *6*, 111–116. https://doi.org/10.4137/HSI.S10471

Naumov, N. (2019). The Impact of Robots, Artificial Intelligence, and Service Automation on Service Quality and Service Experience in Hospitality. In S. Ivanov & C. Webster (Eds.), *Robots, Artificial Intelligence, and Service Automation in Travel, Tourism and Hospitality* (pp. 123–133). Emerald Publishing Limited. https://doi.org/10.1108/978-1-78756-687-320191007

Pankaj. (2023, November 18). The Power of NLP with Flair: A Comprehensive Guide and Comparison with other libraries. *Medium*. https://medium.com/@pankaj_pandey/the-power-of-nlp-with-flair-a-comprehensive-guide-and-comparison-with-other-libraries-d99875595396

Patel, S., & Jones, L. (2017). *Benchmarking the performance of Medallion Architecture against traditional data processing systems. 21(4)*, 498–510.

Peng, M. Y.-P., Tuan, S.-H., & Liu, F.-C. (2017). Establishment of Business Intelligence and Big Data Analysis for Higher Education. *Proceedings of the International Conference on Business and Information Management*, 121–125. https://doi.org/10.1145/3134271.3134296

Puschmann, T., & Alt, R. (2016). Sharing Economy. *Business & Information Systems Engineering*, *58*(1), 93–99. https://doi.org/10.1007/s12599-015-0420-2

PWC. (2015). The Sharing Economy: Consumer Intelligence Series. *PricewaterhouseCoopers LLP*. https://www.pwc.com/us/en/technology/publications/assets/pwc-consumer-intelligence-series-the-sharing-economy.pdf

Radhakrishna, V., Kumar, G. R., & Aljawarneh, S. (2015). Strategic Application of Software Process Model to Optimize Business Intelligence Results. *Proceedings of the The International Conference on Engineering & MIS 2015*, 1–6. https://doi.org/10.1145/2832987.2833053

Ravat, F., & Zhao, Y. (2019). Metadata Management for Data Lakes. In T. Welzer, J. Eder, V. Podgorelec, R. Wrembel, M. Ivanović, J. Gamper, M. Morzy, T. Tzouramanis, J. Darmont, & A. Kamišalić Latifić (Eds.), *New Trends in Databases and Information Systems* (pp. 37–44). Springer International Publishing. https://doi.org/10.1007/978-3-030-30278-8_5

Rizi, S. A. M., & Roudsari, A. (2013). Development of a public health reporting data warehouse: Lessons learned. *Studies in Health Technology and Informatics*, *192*, 861–865.

Rodriguez, P., & Garcia-Valls, M. (2018). *Performance evaluation of Medallion Architecture in real-time embedded systems. 14(3)*, 1–10.

Ruel, H., & Njoku, E. (2021). AI redefining the hospitality industry. *Journal of Tourism Futures*, *7*(1), 53–66. https://doi.org/10.1108/JTF-03-2020-0032

Sabou, M., Brașoveanu, A. M. P., & Önder, I. (2015). Linked Data for Cross-Domain Decision-Making in Tourism. In I. Tussyadiah & A. Inversini (Eds.), *Information and Communication Technologies in Tourism 2015* (pp. 197–210). Springer International Publishing. https://doi.org/10.1007/978-3-319-14343-9_15

Sabou, M., Onder, I., Brasoveanu, A., & Scharl, A. (2015). Towards Cross-Domain Decision Making in Tourism: A Linked Data Based Approach. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2580242

Sahoo, B., & Das, A. (2019). *A comprehensive study on the performance of MapReduce-based big data processing frameworks. 6(1)*, 1–21.

Schor, J. (2014, May 2). *Debating the Sharing Economy*. Great Transition Initiative. https://www.greattransition.org/publication/debating-the-sharing-economy

Šepeľová, L., Calhoun, J. R., & Straffhauser-Linzatti, M. (2021). Sharing Economy Business Models: Informational Services Innovation and Disruption in Uber and Airbnb. *Studies in Systems, Decision and Control*, *376*, 521–540. Scopus. https://doi.org/10.1007/978-3-030-76632-0_18

Shabu, J., Kumar, N., Kumar, R., Maheswari, M., & Refonaa, J. (2024). Analysis of Semiconductor Chip Performance using Medallion Architecture. *Proceedings of the 7th International Conference on Inventive Computation Technologies (ICICT 2024)*. https://doi.org/10.1109/ICICT60155.2024.10544873

Shayegh, P., & Daneshpour, N. (2015). *Using a Data Warehouse to improve analyzing Tourism Data*.

Shenoy, P., & Babu, S. (2008). *Performance modeling of automated transaction processing systems. 33(3)*, 15.

Sisson, L. G., & Adams, A. R. (2013). Essential Hospitality Management Competencies: The Importance of Soft Skills. *Journal of Hospitality & Tourism Education*, *25*(3), 131–145. https://doi.org/10.1080/10963758.2013.826975

Smith, J., & Johnson, R. (2014). *Understanding the performance characteristics of Medallion Architecture for data analytics. 8(2)*, 147–159.

Statista. (2024). *International tourist arrivals worldwide 1950-2023*. Statista. https://www.statista.com/statistics/209334/total-number-of-international-tourist-arrivals/

Timmers, P. (1998). Business Models for Electronic Markets. *Electronic Markets*, *8*(2), 3–8. https://doi.org/10.1080/10196789800000016

Vajirakachorn, T., & Chongwatpol, J. (2017). Application of business intelligence in the tourism industry: A case study of a local food festival in Thailand. *Tourism Management Perspectives*, *23*, 75–86. https://doi.org/10.1016/j.tmp.2017.05.003

van Niekerk, M. (2016). Business, Technology, and Marketing Trends Influencing the Financial Performance of The Hotel Industry. *Faculty Scholarship and Creative Works*, *24*(2). https://doi.org/10.1080/10913211.2016.1236582

Wang, H., Zhu, M., & Li, Y. (2017). *Performance evaluation of traditional ETL systems in cloud environments. 6(1)*, 1–13.

Wang, R. Y. (1998). *A product perspective on total data quality management. Communications of the ACM. 41*(2), 58–65. https://doi.org/10.1145/269012.269022

Wikhamn, W. (2019). Innovation, sustainable HRM and customer satisfaction. *International Journal of Hospitality Management*, *76*, 102–110. https://doi.org/10.1016/j.ijhm.2018.04.009

Williams, S. (2016). *Business Intelligence Strategy and Big Data Analytics: A General Management Perspective*. Morgan Kaufmann.

Wisniewski, M. F., Kieszkowski, P., Zagorski, B. M., Trick, W. E., Sommers, M., Weinstein, R. A., & for the Chicago Antimicrobial Resistance Project. (2003). Development of a clinical data warehouse for hospital infection control. *Journal of the American Medical Informatics Association*, *10*(5), 454–462. https://doi.org/10.1197/jamia.M1299

Wöber, K. W. (2003). Information supply in tourism management by marketing decision support systems. *Tourism Management*, *24*(3), 241–255. https://doi.org/10.1016/S0261-5177(02)00071-7

APPENDIX A - LISTINGS

Name	Description	Python Data Type	SQL Data Type	In Use?
id	Unique identifier for each listing	int	INTEGER	x
listing_url	URL link to the Airbnb listing	str	VARCHAR(255)	
scrape_id	Identifier for the scraping instance	int	BIGINT	
last_scraped	Date when the listing was last scraped	datetime	DATE	
source	Source of the data (e.g., city scrape)	str	VARCHAR(50)	
name	Title of the listing	str	VARCHAR(255)	x
description	Detailed description of the listing	str	TEXT	
neighborhood_overview	Overview of the neighborhood	str	TEXT	
picture_url	URL of the main listing picture	str	VARCHAR(255)	
host_id	Unique identifier for the host	int	INTEGER	x
host_url	URL to the host's Airbnb profile	str	VARCHAR(255)	
host_name	Name of the host	str	VARCHAR(100)	x
host_since	Date the host joined Airbnb	date	DATE	x
host_location	Location of the host	str	VARCHAR(100)	x
host_about	Description provided by the host	str	TEXT	
host_response_time	Average host response time	str	VARCHAR(50)	
host_response_rate	Percentage of times the host responds	str	VARCHAR(10)	
host_acceptance_rate	Percentage of requests the host accepts	str	VARCHAR(10)	
host_is_superhost	Whether the host is a superhost (t/f)	bool	BOOLEAN	x
host_thumbnail_url	URL to host's thumbnail image	str	VARCHAR(255)	
host_picture_url	URL to host's full-size image	str	VARCHAR(255)	

host_neighbourhood	Neighborhood of the host	str	VARCHAR(100)	x
host_listings_count	Number of listings the host has	int	INTEGER	
host_total_listings_count	Total number of listings the host has	int	INTEGER	
host_verifications	Types of verifications the host has	list	TEXT	x
host_has_profile_pic	Whether the host has a profile picture (t/f)	bool	BOOLEAN	x
host_identity_verified	Whether the host's identity is verified (t/f)	bool	BOOLEAN	x
neighbourhood	Neighborhood of the listing	str	VARCHAR(100)	x
neighbourhood_cleansed	Cleaned neighborhood name	str	VARCHAR(100)	x
neighbourhood_group_cleansed	Cleaned neighborhood group name	str	VARCHAR(100)	x
latitude	Latitude coordinate of the listing	float	DECIMAL(10, 8)	x
longitude	Longitude coordinate of the listing	float	DECIMAL(11, 8)	x
property_type	Type of property (e.g., apartment, house)	str	VARCHAR(50)	x
room_type	Type of room (e.g., entire home, private room)	str	VARCHAR(50)	x
accommodates	Number of people the listing can accommodate	int	INTEGER	x
bathrooms	Number of bathrooms	float	DECIMAL(3, 1)	x
bathrooms_text	Textual description of bathrooms	str	VARCHAR(50)	x
bedrooms	Number of bedrooms	int	INTEGER	x
beds	Number of beds	int	INTEGER	x
amenities	List of amenities provided	list	TEXT	x
price	Price per night	str	DECIMAL(10, 2)	x
minimum_nights	Minimum number of nights allowed	int	INTEGER	x
maximum_nights	Maximum number of nights allowed	int	INTEGER	x

minimum_minimum_nights	Minimum of the minimum nights allowed	int	INTEGER	
maximum_minimum_nights	Maximum of the minimum nights allowed	int	INTEGER	
minimum_maximum_nights	Minimum of the maximum nights allowed	int	INTEGER	
maximum_maximum_nights	Maximum of the maximum nights allowed	int	INTEGER	
minimum_nights_avg_ntm	Average minimum nights (next 12 months)	float	DECIMAL(5, 1)	
maximum_nights_avg_ntm	Average maximum nights (next 12 months)	float	DECIMAL(7, 1)	
calendar_updated	When the calendar was last updated	str	VARCHAR(50)	
has_availability	Whether the listing has availability (t/f)	bool	BOOLEAN	x
availability_30	Number of available days in next 30 days	int	INTEGER	
availability_60	Number of available days in next 60 days	int	INTEGER	
availability_90	Number of available days in next 90 days	int	INTEGER	
availability_365	Number of available days in next 365 days	int	INTEGER	
calendar_last_scraped	Date the calendar was last scraped	date	DATE	
number_of_reviews	Total number of reviews	int	INTEGER	x
number_of_reviews_ltm	Number of reviews in last 12 months	int	INTEGER	
number_of_reviews_l30d	Number of reviews in last 30 days	int	INTEGER	
first_review	Date of the first review	date	DATE	x
last_review	Date of the last review	date	DATE	x
review_scores_rating	Overall review score	float	DECIMAL(3, 2)	x
review_scores_accuracy	Review score for accuracy	float	DECIMAL(3, 2)	x
review_scores_cleanliness	Review score for cleanliness	float	DECIMAL(3, 2)	x

review_scores_checkin	Review score for check-in	float	DECIMAL(3, 2)	x
review_scores_communication	Review score for communication	float	DECIMAL(3, 2)	x
review_scores_location	Review score for location	float	DECIMAL(3, 2)	x
review_scores_value	Review score for value	float	DECIMAL(3, 2)	x
license	License or registration number	str	VARCHAR(50)	x
instant_bookable	Whether the listing is instantly bookable (t/f)	bool	BOOLEAN	x
calculated_host_listings_count	Number of listings the host has in this area	int	INTEGER	
calculated_host_listings_count_entire_homes	Number of entire home listings by this host	int	INTEGER	
calculated_host_listings_count_private_rooms	Number of private room listings by this host	int	INTEGER	
calculated_host_listings_count_shared_rooms	Number of shared room listings by this host	int	INTEGER	
reviews_per_month	Average number of reviews per month	float	DECIMAL(4, 2)	x

APPENDIX B – REVIEWS

Field	Type	Description
listing_id	BigInt	Listing Id
id	BigInt	Review unique id
reviewer_id	BigInte	Reviewer ID
reviewer_name	Varchar	Name of the reviews
comments	Varchar	Comment of the reviewer

APPENDIX C – CALENDAR

Field	Type	Description
listing_id	BigInt	Listing Id
date	Datetime	Date of calendar
available	VarChar	T = true, F = false related to available of the listing
price	Decimal	Price of the listing in date day
adjusted_price	Decimal	Adjusted price
minimum_nights	Integer	Minimum of nights
maximum_nights	Integer	Maximum of nights

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa