Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

QUANTITATIVE METHODS FOR ECONOMIC AND BUSINESS DECISION-MAKING

MASTER FINAL WORK

PROJECT

ANALYZING THE PERFORMANCE OF AIRBNB LISTINGS IN LISBON USING MACHINE LEARNING TECHNIQUES

XIN CHEN

SUPERVISION:

PROF. DR. CARLOS J. COSTA

OCTOBER – 2024

Acknowledgments

I would like to express my sincerest gratitude to my supervisor, Professor Carlos Costa, for his continuous availability, his expert guidance, and his priceless support. His insightful suggestions and thoughtful feedback have been of great help in the development and improvement of this study, and I am truly grateful for his mentorship.

I extend my heartfelt thanks to all my Master's professors, whose dedication to teaching and commitment to academic excellence have provided me with a solid foundation throughout my academic journey. Their passion for their subjects and encouragement have greatly contributed to my personal and academic growth.

I am also sincerely grateful to my colleagues in CGD, Alcindo Neves, Diana Pereira, and Joana Silvestre, for their generosity in sharing both their professional expertise and personal experiences. Their continuous support during the final year of my Master's program has been a source of motivation.

To my friends, I owe a special thanks for their unwavering support, encouragement, and understanding during this journey. Your patience and kindness, particularly in tolerating my absence from many social events, have meant more to me than words can express.

Special thanks go to Maria Fernanda and Xia Qiang for their constant encouragement, advice, and support, which have been a continual source of motivation for me throughout this process.

In addition, I am immensely grateful to my parents and siblings, whose continuous support and encouragement have been instrumental in my success throughout this endeavor. Your presence during both the difficult and rewarding times has been invaluable. Thank you.

Abstract

Since its opening in 2008, Airbnb has expanded rapidly to over 220 countries and regions over these years. In order to understand the performance of its listings and predict future trends, it has become essential to study performance metrics in light of this global growth. The objective of this research is to identify the key drivers of performance metrics, specifically occupancy and review ratings, and to compare different machine learning models for performance prediction.

The data set used in this study focuses on Lisbon, Portugal, and contains more than 20.000 different advertisements and 75 variables in total. After the cleaning and pre-processing of the dataset, the Least Absolute Shrinkage and Selection Operator were applied to select the most relevant variables to be included in the models. The results show that host-related factors, such as superhost status, acceptance rate, and response rate, are the features that most influence the listing performance. In addition, the number of amenities also plays a key role in success. Among the predictive models, tree-based ensemble methods, namely Random Forest, outperformed other approaches and provided the most accurate predictions.

Keywords: Airbnb, Machine Learning, Performance Metrics, Lisbon

Resumo

Desde o seu lançamento em 2008, a *Airbnb* expandiu-se rapidamente para mais de 220 países e regiões ao longo destes anos. A compreensão do desempenho dos seus anúncios e a previsão de tendências futuras tornam essencial o estudo das métricas de desempenho à luz deste crescimento global. O objetivo desta investigação é identificar os principais fatores que influenciam as métricas de desempenho, nomeadamente, as taxas de ocupação e de avaliação dos comentários, e comparar diferentes modelos de aprendizagem automática para a previsão do desempenho.

A base de dados utilizada neste estudo é constituída por dados referentes a anúncios publicados no *Airbnb* da cidade Lisboa, Portugal, e comtempla mais de 20.000 anúncios diferentes e 75 variáveis no total. Após a limpeza e o pré-processamento do conjunto de dados, foi aplicado o operador de seleção e redução mínima absoluta para selecionar as variáveis mais relevantes a incluir nos modelos. Os resultados mostram que os fatores relacionados com o anfitrião, como o estatuto de *superhost*, a taxa de aceitação e a taxa de resposta, são as variáveis que mais influenciam o desempenho da propriedade. Além disso, os números de comodidades oferecidas também desempenham um papel fundamental nas métricas de sucesso. Entre os modelos preditivos, o método de ensemble, nomeadamente o *Random Forest,* superaram outras abordagens e forneceram as previsões mais exatas.

Palavras-chave: Airbnb, **Aprendizagem Automática, Métricas de Desempenho, Lisboa**

Index

List os Tables

List of Figures

Acronyms

B2C	Business-to-Customer
BLUE	Best Linear Unbiased Estimator
CRISP-DM	Cross Industry Standard Process for Data Mining
DT	Decision Tree
EDA	Exploratory Data Analysis
GB	Gradient Boosting
IV	Instrumental variables
KNN	K-Nearest Neighbours
LASSO	Least Absolute Shrinkage and Selection Operator
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Squared Error
P2P	Peer-to-peer
RF	Random Forest
RMSE	Root Mean Squared Error

1. INTRODUCTION

The collaborative economy, also known as the sharing economy, is defined as an economic model that facilitates the exchange of goods and services, often provided by private individuals offering their shared assets or services, either for free or for a charge, through digital platforms that create an accessible marketplace for everyone. (European Commission, 2016) The rise of this collaborative economy worldwide resulted from the combination of the technological revolution and an unprecedented global economic crisis. During the financial crisis of the early 2000s, consumers faced reduced purchasing power and a significant decline in their confidence in a globalized corporate world. (Guardado, 2022) In this context, platforms like Airbnb provide travelers with affordable accommodation options while offering hosts a new source of income.

The sharing economy also reduces costs by providing underutilized resources, promoting traction across its network, and enabling significant scalability. Additionally, it has the potential to lower resource consumption in the current economic model by reducing the need for individual ownership, thus increasing the efficient use of products. Also, this model promises to contribute to sustainable economic and social development. (Simic and Liem, 2023)

However, while the sharing economy has introduced positive effects, this business model has also emerged as a disruptive force in industries traditionally operated in a business-to-customer (B2C) environment (Moon et al., 2019), representing a significant threat to established traditional businesses. This disruption extends beyond the accommodation sector, impacting the transportation sector with companies like Uber, Bolt, and inDrive and the office sector with coworking spaces and office-sharing platforms. (Giana M. Eckhardt, 2015)

This study will focus on one type of sharing economy, a prominent example of the peer-to-peer (P2P) business model, Airbnb, which has transformed the hospitality industry by enabling individuals to rent out their private spaces to travelers.

1.1 Sharing Economy: Airbnb

In 2008, Airbnb started as a single-room rental in San Francisco, California, United Stated. Initially, it was created to offer an alternative lodging option for attendees who had difficulty finding hotel rooms during the Democratic National Convention. And taking this opportunity, Airbnb entered the market and has since expanded its operations to over 220 countries and regions worldwide. (Airbnb, 2024)

According to Airbnb's official website, by the end of 2023, the platform had over 5 million hosts and offered more than 7.7 million active listings, with approximately 4 million located in Europe. The global phenomenon of short-term rentals is particularly popular in major urban centers worldwide. The local accommodation sector has seen significant growth, with Lisbon becoming the European capital with the highest number of local accommodations per capita, surpassing cities such as Paris, Rome, and Amsterdam. The market for local accommodation continues to expand, reaching new cities and attracting clients globally. Airbnb listings are primarily concentrated in regions like Western Europe, North America, East and South Asia, and Pacific Asia. At the national level, the United States has the most listings, followed by France, Italy, Spain, and the United Kingdom (Ke, 2017). Within cities, these listings are often clustered in central areas (Gyódi, 2017). For example, in the Lisbon Metropolitan Area, 70% of Airbnb listings are concentrated in the central area, particularly within Lisbon municipality.

Airbnb, which operates as a short-term rental platform utilizing a peer-to-peer business model, enables direct interactions between individuals through its website. The emergence of Airbnb during the 2008 economic crisis led to a radical shift worldwide, impacting urban planning, real estate markets, employment, the environment, culture, and social life (O'Regan & Choe, 2017), and especially transforming the hospitality industry as it was known until then (Guttentag, 2017). This business model offers financial support during economic downturns, reduces waste, and enhances resource utilization, which are crucial for sustainability and environmental well-being.

In Portugal, academic research on Airbnb hosts and the professionalization of short-term rentals (AL) is beginning to gain prominence (Cocola-Gant et al., 2021). However, the media consistently portrays AL activity in a negative light, frequently associating it with gentrification, the loss of character in historic neighborhoods, rising housing prices in city centers, rent increases, declining quality of life for residents, and the displacement of locals to the periphery. (Fernandes et al., 2019; Capineri & Romano, 2021).

On the other hand, there are positive factors associated with the concept of short-term rentals that are rarely mentioned in public discourse. These include the rehabilitation of Lisbon's deteriorated and abandoned properties, the revitalization of local businesses that were severely impacted by the 2008 financial crisis, urban renewal, and the upgrading of historic neighborhoods with new facilities and infrastructure. Furthermore, short-term rentals contribute to reducing the cost of tourist accommodation, enabling landlords to become independent entrepreneurs and, through self-employment, contribute to reducing the monopoly of tourist revenues. This has helped democratize the tourism accommodation industry, promoted resource efficiency, supported job creation and competition, and stimulated innovation, boosting the economy of both the city and the country (Gil & Sequera, 2020; Lopes, 2018).

1.2 Objectives

In today's competitive environment, understanding the factors that contribute to performance is crucial for organizations aiming to achieve excellence. There is a large range of accommodation choices on Airbnb, from private rooms in shared houses to luxury mansions. The diversity of accommodation options caters to different budgets and preferences. The prices offered by Airbnb are often more affordable compared to traditional hotels, making it particularly suitable for group travel or long-term stays (Thaichon et al., 2020).

Airbnb's influence on urban environments like Lisbon is multifaceted. On one hand, it has revitalized historic neighborhoods, creating new income streams for local hosts. On

the other hand, it has contributed to rising rental prices, gentrification, and displacement of local populations (Lee and Kim, 2023). Therefore, analyzing the performance of Airbnb listings is critical not only for hosts but also for policymakers seeking to balance the economic benefits with social implications.

The occupancy rates are a direct measure of how frequently a listing is booked, serving as a strong indicator of its financial success since the higher the occupancy rate, the more the host can earn. In contrast, the guest reviews reflect customer satisfaction and the quality of the hosting experience. Positive reviews can boost a listing's visibility on the platform and attract more bookings, as potential guests often rely on past feedback when selecting a property.

This study aims to identify the key factors that influence the performance of Airbnb listings, focusing on metrics such as occupancy rate and guest reviews.

In addition, another objective of the research is to utilize machine learning techniques to forecast the performance of these listings and to conduct a comparative analysis of different models to determine which is the most effective. Traditional data analysis methods are often insufficient to address the complexity and volume of data generated by platforms such as Airbnb, making machine learning an ideal approach.

This study contributes to the existing literature by offering a data-driven analysis of Airbnb's performance, particularly in Lisbon. The study applies machine learning models to assess and predict the performance of listings, with a focus on occupancy rates and guest satisfaction. By doing so, it aims to discover key predictors of Airbnb's success, actionable insights that allow both hosts and policymakers to react to past performance and forecast future outcomes, enabling smarter decision-making and more sustainable growth.

1.3 Dissertation Outline

This dissertation is structured into five chapters, with the first chapter introducing the sharing economy and a particular emphasis on Airbnb as a case study. It presents the

research objectives, explaining the study's goals and rationale, along with an overview of the dissertation structure.

The second chapter reviews the existing academic literature on Airbnb and examines key performance metrics used to evaluate Airbnb listings, establishing the foundation for the analysis in later chapters.

The third chapter describes the methodology employed. This comprises a concise account of the dataset's provenance and the transformations that have been applied to it, followed by an exposition of the data preparation techniques employed. The chapter concludes with a description of the modeling approaches employed to analyze the dataset.

The fourth chapter presents the research findings. It includes an Exploratory Data Analysis (EDA), offering insights into occupancy rates and guest reviews. This chapter also covers the results of the models developed to predict occupancy and review performance. Furthermore, this chapter ends with a discussion.

Finally, the last chapter summarizes the main findings, draws conclusions from the research, and suggests recommendations for further studies and practical applications within the Airbnb context.

2. LITERATURE REVIEW

2.1 An Overview of Primary Studies Related to Airbnb

The rapid growth of Airbnb has attracted significant attention from researchers, leading to a wide variety of studies across multiple domains. For instance, Lee (2016) examined its effect on property values and housing availability, while Gurran and Phibbs (2017) focused on urban planning and the impact of short-term rentals on tourism employment. Additionally, studies such as Wachsmuth and Weisler (2018) have explored the social and environmental impacts and dynamics of short-term rentals. Horn and Merante (2017) investigated the influence of the platform on housing affordability and neighborhood dynamics. Research developed is often related to the shared economy, but it is often viewed as a proxy for the study of real estate (Samandani and Costa, 2021)

Some researchers have focused on consumer behavior and trust in P2P platforms. Ert et al. (2016), for example, explored how host photos influence the perceived trustworthiness of a listing. They concluded that hosts who are perceived as more trustworthy based on their photos tend to have higher prices and a greater likelihood of bookings. Similarly, trust has been linked to other factors, such as the quality of the information provided, perceived social capital Chen et al. (2015), and host-related attributes like reservation confirmation speed, acceptance rate, and the presence of a complete profile page Wu et al. (2017). Yang et al. (2019) referred the importance of cognitive trust and identity attachment in the sharing economy, particularly in Airbnb transactions, where both hosts and guests face risks.

Trust remains a central theme in research on peer-to-peer economic transactions. Yang et al. (2019) highlight factors such as security, privacy, and Airbnb-specific traits that significantly contribute to building user trust. Studies such as Fagerstrøm et al. (2017) and Sparks and Browning (2011), trust is influenced by early negative reviews and positively framed reviews with numerical ratings, highlighting the role of review content and structure in shaping consumer choices in a hotel context.

Moreover, customer characteristics have emerged as another key area of research. Guttentag (2016) identified five distinct customer segments within the Airbnb market, noting that, aside from home-seekers, low cost remains the primary motivation for Airbnb users. Building on this, Chiappa et al. (2020) segmented Italian Airbnb customers into three clusters: enthusiastic Airbnb lovers, pragmatic Airbnb users, and pragmatic authenticity seekers, distinguishing these groups based on marital status and education level.

Further research by Festila and Müller (2017) examines the relationship between consumers and access objects, finding that meaningful consumption experiences arise when consumers identify with the object using a psychological dimension by categorizing Airbnb customers as either 'extrovert' or 'introvert', and as individuals who 'go to see' versus those who 'go to feel." Each of these customer groups has different satisfaction criteria, highlighting the complexity of consumer preferences. Similarly, Lutz and Newlands (2018) compared customers who prefer shared rooms with those who choose entire homes, revealing differences in preferences linked to gender and socio-economic status.

Research on regulations has also been prominent, with Maciel (2019) exploring the challenges of regulating platforms like Airbnb by examining the legal aspects of sharing economies, focusing on how the principle of the social function of property and the strengthening of objective good faith can guide legal analysis. In contrast, Bei and Celata (2023) compare regulations in 16 cities in different countries to assess whether these cities regulate such platforms and evaluate the effectiveness of their approaches.

Another important aspect examined is the impact of the sharing economy on local economies and labor markets, which varies across regions. For example, Basuroy et al. (2020) examined the impact of Airbnb on restaurant revenues in Texas, finding a notable economic boost. Similarly, Suciu (2016) examined the impact on employment in Germany, highlighting how traditional businesses face challenges from emerging business models such as Airbnb, reshaping the local labor market.

2.2 Performance Metrics in the Airbnb Context

Some research has focused on identifying factors that determine Airbnb listing performance. Chen and Chang (2018) studied the factors influencing the purchase intentions of Airbnb users, focusing on five key elements: rating, rating volume, reviews, information quality, and media richness. Using ANOVA and path analysis, they examined how these factors impact perceived value, satisfaction, and, ultimately, purchase intention. The conclusion achieved indicates that perceived value and satisfaction are critical determinants of purchase intention. Similarly, Zhao and Peng (2019) studied how the quality of online reviews influences purchasing decisions, finding that high-quality reviews significantly impact purchasing behavior. These insights suggest possible improvements for rental platforms to enhance user experience and trust.

Building on this, Kirkos (2022) investigates the performance of Airbnb listings in Thessaloniki, Greece, focusing on key metrics such as occupancy rate, number of bookings, and revenue. The study aims to identify the strongest determinants influencing customer purchase intentions and to develop predictive models for listing performance using data mining and machine learning techniques.

Meanwhile, Liang et al. (2017) focus on the impact of Airbnb's "Superhost" badge, a gamification feature, on review volume, ratings, and pricing in the vacation rental industry. Focusing on accommodations in Hong Kong, the researchers used a negative binomial model and a Tobit model to analyze data, controlling for various accommodation characteristics. The conclusion reveals that "Superhost" accommodations are more likely to receive higher review volumes, better ratings, and higher prices from guests. The study provides insights for both academic research and practical recommendations for Airbnb hosts.

Tussyadiah (2016) investigated the factors influencing guest satisfaction and found that enjoyment, monetary benefits, and accommodation amenities are key drivers of satisfaction, an important performance metric of Airbnb. Furthermore, cultural differences can also impact customer satisfaction and behavior. Suh et al. (2022) investigated the social and economic benefits prioritized by Airbnb guests from the U.S.

and China, finding that both benefit types significantly influenced satisfaction and behavioral intentions. The study concluded that cultural differences played a moderating role in these relationships, highlighting the importance of considering cultural context in peer-to-peer accommodation experiences.

Additionally, at the end of year 2020, the COVID-19 pandemic spread globally, leading to widespread lockdowns. Despite these challenges, research by Farmaki et al., (2020) revealed that hosts still expressed intentions to continue using P2P platforms. (Filieri et al.,2023) further noted a shift in consumer preferences post-pandemic, with travelers preferring entire homes or apartments over shared accommodations and showing a greater inclination toward higher-priced accommodations in rural areas. Towards the end of the pandemic, Jang and Kim (2022) suggested that city-specific strategies and the use of local resources can help P2P accommodation hosts and policymakers effectively address pandemic-related challenges.

3. METHODOLOGY

The objective of the research reported here is to identify the primary factors that influence the performance of Airbnb listings, focusing on metrics such as occupancy rate and guest reviews. In order to reach this purpose, a research approach based on CRISP-DM was followed (Shearer, 2000; Costa & Aparicio,2020, 2021; Aparicio et al., 2019). It considers the following Phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This research only includes model evaluation, not the evaluation of the CRISP-DM approach. On the other hand, Deployment is also outside the analysis produced here. Nevertheless, it is possible to consider this study in the context of design science (Aparicio et al., 2023).

3.1 Data Understanding

The primary dataset used in this study is sourced from the platform Inside Airbnb[1], which aggregates the information from the public listings on Airbnb's official website as a snapshot, providing a detailed and more comprehensive view. This independent, non-commercial website aims to provide transparency about the impact of Airbnb on housing markets, particularly in urban areas.

The data available on this website are updated periodically for 33 countries and multiple cities within these countries. This study focuses on Lisbon, Portugal's geographical area, covering the period from March 18th, 2023, to March 18th, 2024. The dataset is composed of more than 20.000 different rows representing the active listings within the Lisbon metropolitan area.

The dataset comprises 75 variables, which can be broadly categorized into different sections. Identification variables, such as *'id'*, *'listing_url'*, and *'scrape_id'*, provide foundational information about the listing's identity and its source. Descriptive variables, including '*name*', '*description'*, and *'neighborhood_overview'*, offer insights into the

[1] https://insideairbnb.com/

property and its surrounding area. Host-related information, such as *'host_id'*, *'host_response_time'*, *'host_is_superhost'*, and *'instant_bookable'*, captures details about the property owner or manager, while location-based variables like '*latitude*', '*longitude'*, and *'neighborhood'* specify the geographical positioning of each listing. Property characteristics, including *'property_type'*, *'room_type'*, and *'amenities'*, provide information on the physical attributes of the accommodations.

Pricing and stay-related variables *'price', 'minimum_nights', and 'maximum_nights'* define the cost structure and rental terms. Availability and booking details are captured through variables such as *'availability_30'* and *'has_availability'*, which give an overview of the listing's availability over time. Review metrics, including *'number_of_reviews'*, *'review_scores_rating'*, and *'reviews_per_month'*, reflect customer feedback and satisfaction.

3.1.1 Location

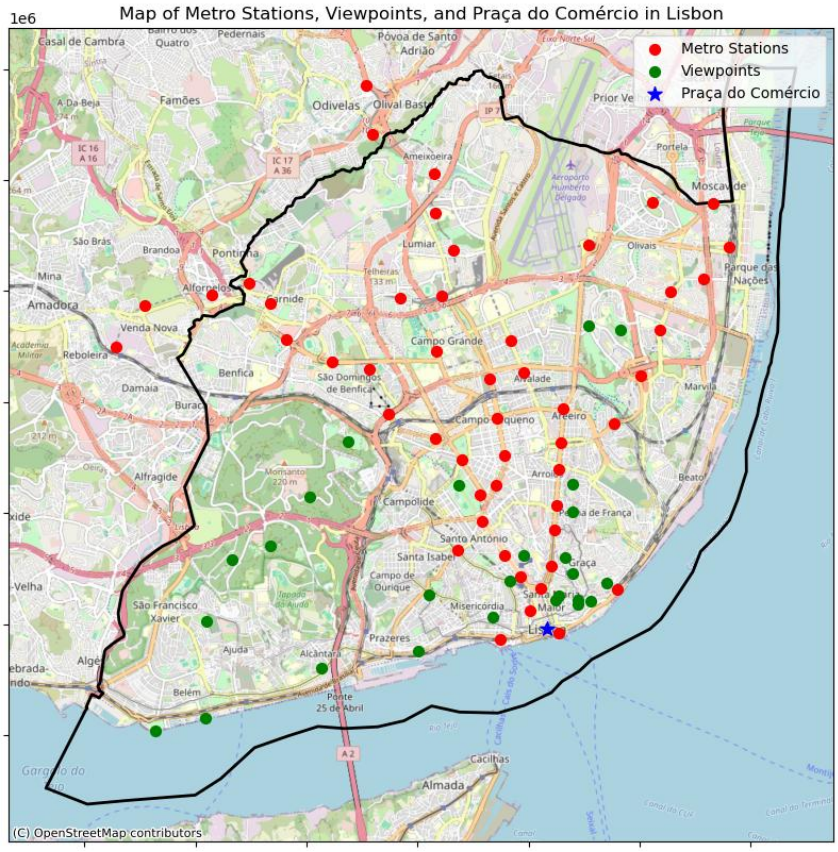
As highlighted by Constantin and Dardala (2015), location is a fundamental factor in the analysis of the tourism sector due to the industry's dependency on the natural, cultural, social, and built environments of a particular region. In the context of Airbnb, a platform that primarily provides short-term rental services to tourists, the importance of location becomes even more pronounced. The performance of Airbnb listings, which are to be measured through metrics such as occupancy rates and guest reviews, can be influenced by the proximity of these listings to subway stations and attractions. Therefore, analyzing the role of location in shaping Airbnb's performance is a better understanding of how spatial characteristics contribute to the success of listings.

For this study, properties outside the municipality of Lisbon were excluded. Because including the listings outside this area could have introduced a high correlation with those within the municipality, potentially distorting the analysis. By excluding properties from neighboring areas, the results remain specific to the dynamics of the Lisbon municipality, minimizing external noise and ensuring a more accurate focus on the region's unique market conditions.

In the present dataset, the only location-related variables are *'neighborhood'*, *'longitude*', and *'latitude*'. However, by utilizing additional data from the *Geodados* platform[2], it was possible to collect coordinates for key points of interest within the Lisbon municipality, including subway stations, viewpoints, and a center point of the city.

Figure 1 presents a visual representation of the geographical features mentioned above. The map illustrates the spatial distribution of these amenities across the Lisbon metropolitan area, highlighting areas with greater accessibility to public transportation and tourist viewpoints.

Figure 1Map of Lisbon



[2] https://geodados-cml.hub.arcgis.com/

The coordinates from *Geodados* were used to create new location-related variables,

Variable	Description
metro_station_count_500m	Total number of subway stations within a 500 meter radius.
viewpoint_count_1km	Total number of viewpoints within a 1km radius.
baixa_distance_km	Distance, in kilometers, from the Praça do Comércio to listings.

with the detail presented in Table 1.

Table 1 New variables of location

3.1.2 Occupancy rate

In the context of understanding Airbnb's dynamics, analyzing occupancy rates can provide significant insights and a direct influence on revenue and overall success. Furthermore, it reflects the proportion of time units that are rented, which can facilitate the management and measure the efficiency of property utilization. However, due to Airbnb's policy of not providing exact occupancy information for individual listings, some methods are adopted to estimate this metric.

It is important to mention that the variable 'availability_365' is important when considering occupancy, as it reflects the number of days a listing is available for booking throughout the year. (Kirkos, 2022) However, one challenge is that this column does not clarify whether the property was actually booked by the guests or simply not made available by the host on the remaining days. Since the availability for a year does not indicate whether the host chose not to rent the house or if it was already booked, this variable cannot be reliably used as a proxy for calculating occupancy when studying performance, adding complexity to the analysis.

Given that, the dataset does not provide information about the number of room reservations. One approach to estimating occupancy rates involves using the number of reviews as a proxy for room reservations. The model is called the "San Francisco Model," and it was released in a report by Inside Airbnb. Another model used by Shen, Lily,

Wilkoff, and Sean (2020) also relies on the number of reviews and the number of stays to calculate occupancy rates, with slight differences in the time frame used for the calculation.

The exact formulas for these two different methods are presented as follows:

1. San Francisco Model: This model estimates the occupancy rate using monthly variables.

$$occupancy\ rate = \frac{number\ of\ reviews\ per\ month * average\ stay * 12}{365} * 100$$

2. According to the author Shen, Lily, and Wilkoff, Sean (2020), the formula for the calculation is the space of a year:

$$occu.\ rate = \frac{2 * number\ of\ annual\ comments * number\ of minimum\ nights}{365} * 100$$

Since the occupancy rate obtained is only an estimation, some values exceed 100%. The San Francisco Model estimated 35 listings with an occupancy rate greater than 100%, while the model used by Shen, Lily, and Wilkoff, Sean (2020) identified 468 such listings. Therefore, for further analysis, we will consider the values obtained by the San Francisco Model. This choice is based on its more conservative estimation, which aligns better with realistic occupancy scenarios.

3.2 Data Preparation

In this section, data processing is carried out, which includes the treatment of the missing value, the exclusion of the outliers, and the transformation and organization of the raw data into a structured format that can be effectively analyzed. This is a fundamental step of the overall process, as it directly impacts the quality and reliability of the results. Without proper data processing, models may be built on inaccurate, incomplete, or irrelevant data, leading to biased or misleading conclusions. Furthermore, well-processed data ensures that insights derived from analysis are valid, enhances model

performance, and supports the overall goal of making data-driven decisions with precision and accuracy.

In the initial data processing phase of a study examining the performance of Airbnb listings, it is crucial to eliminate certain variables in order to make the analysis more efficient and improve the overall performance of the model. Variables such as *URL links*, *'scrape_id'*, *'calendar_updated'*, '*source*', '*last_scraped*', and '*calendar_last_scraped'* provide information about the timing and source of the dataset rather than contributing any meaningful insights into the listings' performance. As a result, these variables are typically irrelevant to the core research objectives.

The variable *'has_availability'* was excluded because other variables already provide a more detailed and comprehensive measure of future availability. Similarly, *'neighbourhood_cleansed'* was removed due to redundancy, as more detailed and informative geographic data was retained. Variables such as *'neighborhood'*, *'neighborhood_overview'*, and *'host_neighbourhood'* were eliminated due to having over 40% missing data, where eliminating these variables was deemed preferable to deleting rows or imputing missing values using methods such as the mean or median. Furthermore, the variable *'neighbourhood_group_cleansed_code'* had already been transformed using categorical encoding, as detailed in the previous subsection, where only listings within the Lisbon Municipality were selected for analysis, so this variable became irrelevant due to its uniformity.

Certain host-related variables, such as *'host_about'*, *'host_location'*, *'host_verifications'*, *'host_name'*, and *'host_since'*, although relevant to the host's profile, have limited direct influence on the performance of a listing.

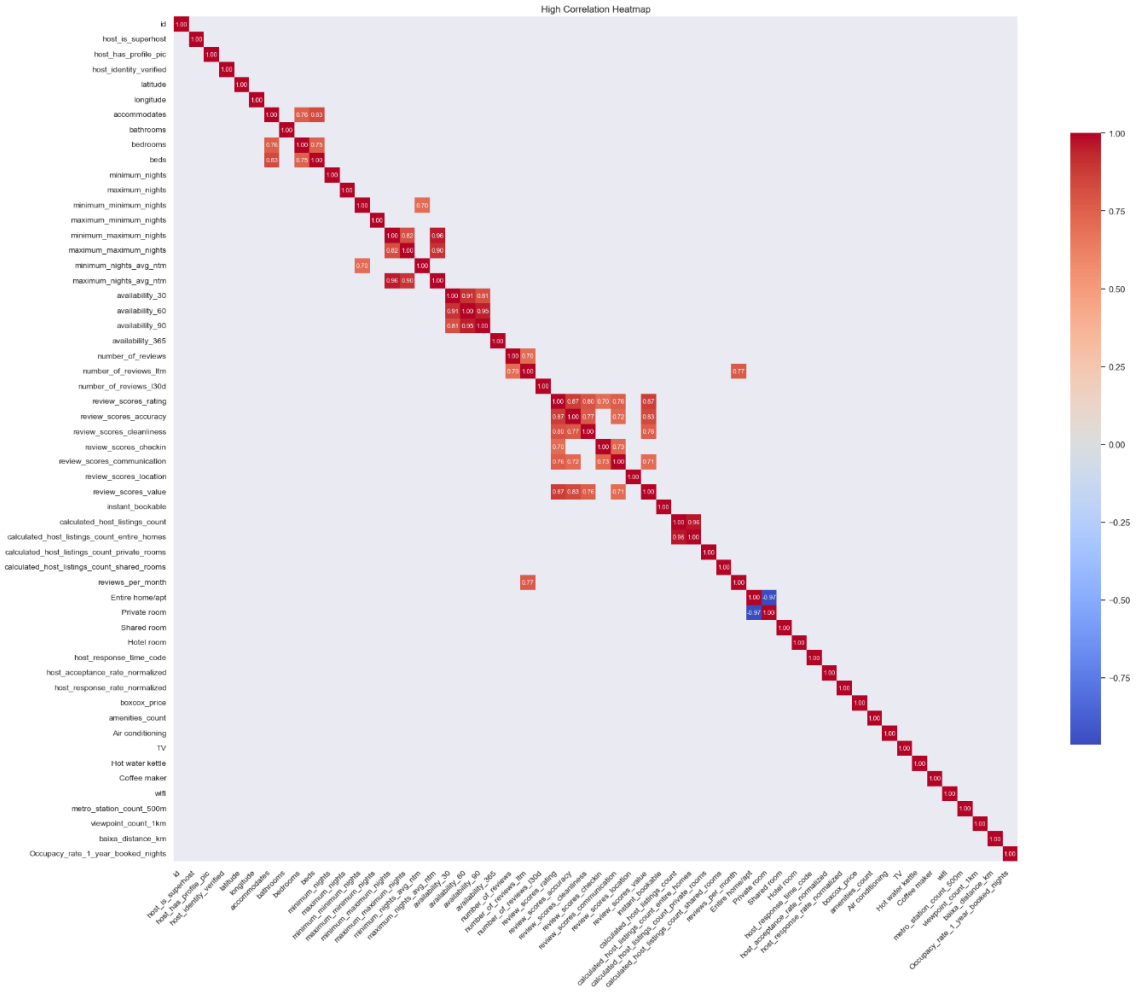
Additionally, fields such as *'host_total_listings_count'* and *'host_listings_count'* are redundant when more precise alternatives, such as *'calculated_host_listings_count'*, are already present in the dataset. The latter provides a reflection of the host's current listings within the specified geography at the time of data collection, making the other variables unnecessary.

The variable *'bathrooms_text'* is redundant as it contains the number of bathrooms in a string format, while the *'bathrooms'* variable provides this information in a numerical

format, making *'bathrooms_text'* unnecessary. Lastly, the *'license'* variable, which indicates whether a listing holds a license, is not directly relevant to the performance analysis and is thus eliminated. Similarly, variables such as *'first_review'* and *'last_review'* are excluded, as they primarily provide information about the timing of reviews rather than directly contributing to the analysis of a listing's performance.

Manually eliminating these variables is a necessary step in the feature selection process, which enhances model interpretability, reduces computational complexity, and focuses the analysis on the variables most directly related to Airbnb listing performance. This approach improves model accuracy and also prevents overfitting.

Figure 2 Correlation Matrix



After this initial elimination of variables, some adjustments to some categorical variables and clearing the outliers along with the missing values are made. A correlation matrix was used to test for multicollinearity among the remaining variables. All variables with a correlation coefficient of 0.7 or higher were, as strong correlations above 0.7 indicate that the variables are highly related. The correlation of variables is presented as Figure 2. Multicollinearity poses a significant issue in model development, as it can distort the interpretation of results, leading to unreliable estimates and reduced model accuracy. By addressing multicollinearity, the analysis becomes more robust, and the model's interpretability is enhanced.

Afterward, the method of Least Absolute Shrinkage and Selection Operator (LASSO) is used to select different variables for different metrics of performance (occupancy rate and guest reviews) for Airbnb.

Lasso is a regularization technique that not only addresses multicollinearity but also performs variable selection by penalizing less significant coefficients, shrinking them to zero. This ensures that the final model focuses on the most predictive variables.

The final variables selected for occupancy rate and review rate are as following table:

Table 2 Variables Selected for Occupancy Rate

Variables - Occupancy rate	Description
host_is_superhost	Whether the host is a superhost or not
host_identity_verified	Whether the host is verified or not
bathrooms	The number of bathrooms in the listing x
beds	The number of beds in the listing x
minimum_nights	Minimum number of night stay for the listing
maximum_ minimum_nights	The biggest minimum_night value from the calender
availability_365	The availability of the listing 365 days in the future as determined by the calendar.
review_scores_rating	The rating scores of overall
review_scores_location	The rating scores of locations
instant_bookable	Whether the listing is instant bookable or not

calculated_host_listings_count	The number of listings the host has in the current scrape, in the city/region geography.
calculated_host_listings_count_private_rooms	The number of Private room listings the host has in the current scrape, in the city/region geography
calculated_host_listings_count_shared_rooms	The number of Shared room listings the host has in the current scrape, in the city/region geography
entire home/apt	Whether the listing is entire home/apartment or not
shared room	Whether the listing is shared room or not
hotel room	Whether the listing is hotel room or not
host_response_time	The host response time
host_acceptance_rate	The host acceptance rate
host_response_rate	The host response rate
price	Price of the listing x
amenities_count	Number of amenities in the listing x
air conditioning	Whether the listing x contain air conditioning or not
TV	Whether the listing x contain TV or not
hot water kettle	Whether the listing x contain hot water kettle or not
coffee maker	Whether the listing x contain coffee marker or not
wifi	Whether the listing x contain wifi or not
metro_station_count_500m	Total number of subway stations within a 500-meter radius in listing x.
viewpoints_count_1000m	Total number of viewpoints within a 1km radius in listing x
baixa_distance_km	Distance, in kilometers, from the Praça de Comércio to listing x

Table 3 Variables Selected for Review Rating

Variables - Review rating	Description
host_is_superhost	Whether the host is a superhost or not
host_identity_verified	Whether the host is verified or not
minimum_nights	Minimum number of night stay for the listing
maximum_ minimum_nights	The biggest minimum_night value from the calender
availability_30	The availability of the listing in 30 days in the future as determined by the calendar.

availability_365	The availability of the listing in 365 days in the future as determined by the calendar.
number_of_reviews	The number of reviews the listing x has
review_scores_location	The rating scores of locations
instant_bookable	Whether the listing x is instant bookable or not
calculated_host_listings_count	The number of listings the host has in the current scrape, in the city/region geography.
calculated_host_listings_count_private_rooms	The number of Private room listings the host has in the current scrape, in the city/region geography
calculated_host_listings_count_shared_rooms	The number of Shared room listings the host has in the current scrape, in the city/region geography
entire home/apt	Whether the listing x is entire home/apartment or not
shared room	Whether the listing x is shared room or not
hotel room	Whether the listing is hotel room or not
bathrooms	The number of bathrooms in the listing x
beds	The number of beds in the listing x
host_response_time	The host response time
host_acceptance_rate	The host acceptance rate
host_response_rate	The host response rate
price	Price of the listing x
amenities_count	Number of amenities in the listing x
air conditioning	Whether the listing x contain air conditioning or not
TV	Whether the listing x contain TV or not
hot water kettle	Whether the listing x contain hot water kettle or not
coffee maker	Whether the listing x contain coffee marker or not
wifi	Whether the listing x contain wifi or not
metro_station_count_500m	Total number of subway stations within a 500 meter radius in listing x.
viewpois_count_1000m	Total number of viewpoints within a 1km radius in listing x
baixa_distance_km	Distance, in kilometers, from the Praça de Comércio to listing x

3.3 Modeling

For the study of Occupancy Rate and Review Rate, two distinct datasets were selected due to the differences in variables influencing these metrics. With the final

dataset, the analysis begins with Ordinary Least Squares (OLS) regression, which is a common approach due to its simplicity and interpretability.

The OLS model is expressed as follows:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

Where:

- y_i, represents the dependent variable for observation i;
- α is the intercept, represents the expected value of Y when all the predictor variables are zero;
- β_k are the coefficients corresponding to the independent variables;
- x_{ki} are the values of the independent variables for observation i;
- ε_i is the error term, capturing the difference between the observed and predicted values of the dependent variable for each observation.

The OLS regression model is a widely used statistical method for estimating the relationship between observations in the training data set and the values predicted by the linear model. It aims to minimize the sum of the squared differences between the observed and predicted values of the dependent variable, thus producing estimates of the model's coefficients. These coefficients are a measure of the impact of each independent variable on the dependent variable.

However, in the real world, do not conform to an ideal experimental setup making it difficult to meet the Gauss-Markov assumptions, which guarantee that the OLS estimator is BLUE (Best Linear Unbiased Estimator). For the OLS estimator to be considered BLUE, the following ideal conditions must be met as the follow table:

Table 4 Assumption of OLS Regression

Conditions	Description
Linear in parameter	The relationship between the dependent variable and the independent variable must be linear in the parameters.
Zero mean of the error term	The expected value of the error term should be zero, meaning that, on average, the model's predictions are correct.

Homoscedasticity of the error term	The variance of the error term must be constant across all observations, indicating that the model's uncertainty is consistent for all predictions.
Independence of errors	The error terms should be independently distributed and not exhibit autocorrelation, meaning that errors in one observation should not affect errors in another.
No multicollinearity	The independent variables should not be highly correlated with each other, ensuring that each variable uniquely contributes to explaining the variation in the dependent variable.

Violating these assumptions can result in biased, inefficient, or unreliable estimates. For example, non-linear relationships, heteroscedasticity, and autocorrelation undermine the accuracy of the model.

While OLS provides a useful baseline, it has limitations when dealing with complex data structures and non-linear relationships. To improve prediction accuracy and handle more complex data, we employ four well-known machine learning models for comparison: Random Forest (RF), Gradient Boosting (GB), K-Nearest Neighbours (KNN), and Multilayer perceptron (MLP). These models are better suited for capturing intricate patterns and interactions between variables that OLS may not account for.

Random Forest is an ensemble learning method that builds multiple decision trees and merges their results to improve accuracy and avoid overfitting. Each tree is uncorrelated and trained on a random subset of data, and the final prediction is the average for regression or majority vote for classification of all trees. It is highly robust and handles non-linear relationships well.

Gradient Boosting is another ensemble method that builds decision trees and train sequentially, where each new tree corrects the errors of the previous ones. It gradually minimizes a loss function by iteratively improving the model. Gradient Boosting is effective for both regression and classification tasks, providing high accuracy but requiring careful tuning to avoid overfitting.

KNN is a simple, non-parametric algorithm that makes predictions using proximity based on the "k" closest data points in the feature space. For regression, it predicts the average of the "k" nearest neighbors' values. KNN is easy to implement but can be computationally expensive for large datasets and may struggle with noisy or high-dimensional data.

A multilayer perceptron is a type of artificial neural network used for both classification and regression. It consists of multiple layers of neurons: an input layer, one or more hidden layers, and an output layer. Each neuron in a layer is connected to every neuron in the next layer by weighted connections. MLP uses non-linear activation functions, allowing it to capture complex relationships in the data. During training, MLP learns by adjusting these weights using backpropagation, a process that minimizes the error between predicted and actual outcomes. This makes MLP particularly effective for modeling complex patterns and interactions in data.

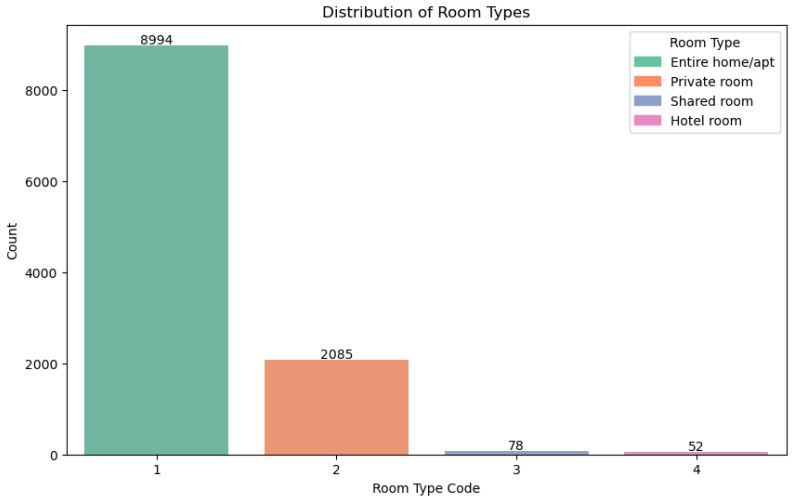
4. RESULTS AND DISCUSSION

4.1 Exploratory Data Analysis

To better understand the characteristics of the variables, using Exploratory Data Analysis (EDA) is essential, as it provides clarity and insight into the dataset. EDA permits better decision-making for feature selection, helps address data issues like missing values or multicollinearity, and guides the choice of appropriate modeling techniques.

The analysis of occupancy rates examines the estimated rates for different types of Airbnb units, including entire homes/apartments, private rooms, shared rooms, and hotel rooms. Figure 3 illustrates the distribution of room types in the dataset. More than half of the listings are entire homes or apartments, with 8,994 units. Private rooms are the second most common type, with 2,085 units. Shared rooms and hotel rooms make up the smallest portion of the dataset, with 78 and 52 units, respectively.

Figure 3 Distribution of Room Types



The table 5 and the discussion that follows aim to illustrate the distribution of occupancy rates within these categories:

Table 5 Estimated Occupancy by Type of Unit: Lisbon. Feb.23 – Feb.24

Type of unit	Number of Units	25th Percentile	Median	75th Percentile	90th Percentile	95th Percentile
Entire home/apt	8994	4.66%	13.70%	27.40%	40.27%	49.32%
Private room	2085	1.64%	5.75%	14.25%	27.95%	36.16%
Shared room	78	0.54%	2.19%	9.24%	15.48%	21.85%
Hotel room	52	0.54%	1.64%	3.42%	4.77%	6.16%

Table 5 shows significant differences in occupancy rates across room types. Entire homes/apartments tend to have higher occupancy rates across all percentiles compared to private rooms, shared rooms, and hotel rooms. This trend indicates a higher demand for entire homes/apartments, possibly due to the privacy and amenities they offer.

Private rooms also show relatively higher occupancy rates, particularly in the upper percentiles, suggesting a strong demand within this segment as well. Shared rooms and hotel rooms, however, have much lower occupancy rates, especially in the lower percentiles, which may reflect their smaller appeal or the niche markets they cater to.

Figure 4 Evolution of the number of comments per month from Feb.23 to Feb.24

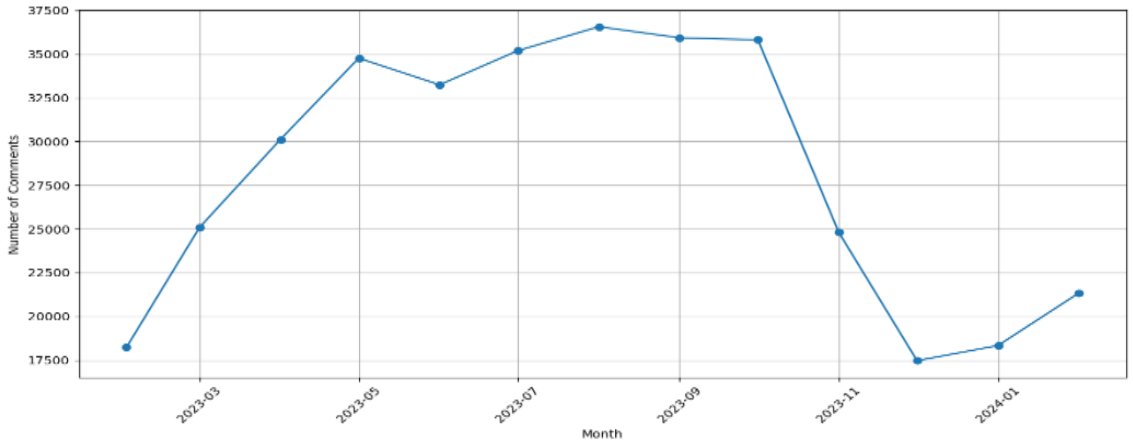
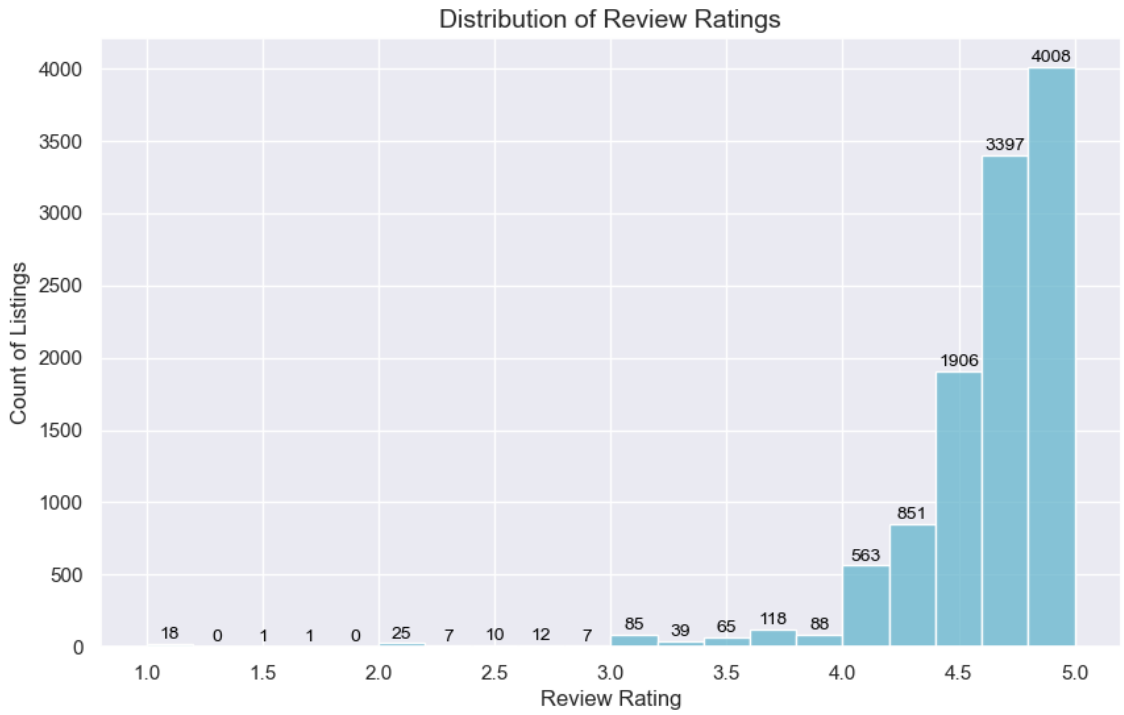


Figure 4 shows the number of comments of all listings from Airbnb in Lisbon during the period of February 2023 until February 2024. It is visible that the number of

comments is much higher in the months of summer, and with lower levels during the months of winter.

Regarding the variable review scores, figure 5 below illustrates the distribution of review ratings for Airbnb listings. From the graph, it is evident that most of the listings have high review scores, with the largest concentration around 4.5 and 5.0. Specifically, 4,008 listings have a perfect rating of 5.0, followed by 3,397 listings with a rating of 4.5. Lower ratings are far less common, with a noticeable drop-off in listings rated below 4.0. The data suggests that most guests tend to leave high ratings, and low-rated listings are relatively rare.

Figure 5 Distribution of Review Ratings



4.2 Occupancy Rate

As demonstrated in the previous chapter, the occupancy rate is calculated based on the number of reviews and the minimum stay requirements. Although the variable '*minimum_night'* is directly related to occupancy rate, its correlation with the target

variable is moderate, at 0.59. Additionally, the Variance Inflation Factor (VIF), which assesses the degree of multicollinearity in regression models, is 1.29 for this variable is well below the threshold of 5, indicating that multicollinearity is not a significant concern. While '*minimum_nights*' could potentially show endogeneity, the results of the instrumental variables (IV) regression do not support this. Therefore, it is reasonable to include '*minimum_nights*' in the final model as one of the selected variables. For the rest of the variables, the VIF has been verified as acceptable. In the end, 28 variables were selected.

Using all the variables selected by Lasso in the OLS model, we can obtain a final result with 12 variables statistically significant (p-value < 0.05), presented in Table 6. The R^2 of 0.404 and adjusted R^2 of 0.401, which means that approximately 40.4% of the variance in the occupancy rate can be explained by the independent variables included in the model.

The result identifies several key variables that significantly influence occupancy rates in short-term rental properties. Firstly, the host acceptance rate and response rate emerge as critical factors. Interestingly, while hosts are generally expected to respond quickly to ensure a positive guest experience, this study shows a negative correlation with occupancy rates, suggesting that slower communication may be a discouraging factor for potential bookings.

The analysis also reveals that the number of amenities offered positively correlates with occupancy rates. In contrast, overall ratings and specific amenities appear to have less impact on driving guest bookings.

Location is another essential determinant, as properties situated in well-rated areas and within proximity of viewpoints (within 1 kilometer) tend to attract greater occupancy. This emphasizes the important role of location in attracting potential guests.

Furthermore, the findings indicate a positive correlation between occupancy rates and a higher minimum night's requirement, suggesting that guests tend to be attracted to properties that are suitable for longer stays.

The variables that are statistically significant (p-value < 0.5) are presented in the table below:

Table 6 Variables of Occupancy Rate

Variables	Coefficient
host_is_superhost	0.0446
minimum_nights	0.0550
maximum_minimum_nights	-0.0028
availability_365	-0.0139
calculated_host_listings_count	-0.0382
amenities_count	0.1825
host_response_rate	-0.2630
host_acceptance_rate	0.2373
host_response_time	-0.0801
viewpoint_count_1km	0.0091
hot water kettle	-0.0401
review_scores_location	0.0464

The application of various machine learning techniques, including Random Forest, K-Nearest Neighbors, Gradient Boosting, and Multilayer Perceptron, produced distinct results in predicting the occupancy rate, with the dataset split into 80% for training and 20% for testing to evaluate the model's performance.

It is important to highlight that the Multilayer Perceptron, Random Forest, and Gradient Boosting regressors were optimized to achieve the best performance. Specifically:

- The Multilayer Perceptron was configured with two layers (100, 50) and used the logistic activation function.
- The Random Forest model was trained using 300 trees.
- The Gradient Boosting model was trained using 100 trees.

The performance of each model is assessed using four key metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2), as presented in the table:

Table 7 Machine Learning Model's Performance Comparison of Occupancy Rate

Model	Random Forest	Gradient Boosting	K-Nearest Neighbors	Multilayer Perceptron
RMSE	0.16	0.23	0.48	0.15
MAE	0.07	0.09	0.10	0.10
MSE	0.02	0.05	0.23	0.02
Train R^2	0.92	0.94	0.46	0.92
Test R^2	0.94	0.88	0.47	0.93

The results indicate that the Random Forest model is the most robust and accurate technique for predicting occupancy rates in this dataset, closely followed by Multilayer Perceptron and Gradient Boosting, both of which also demonstrate strong predictive power. In contrast, K-Nearest Neighbors shows significantly weaker performance, indicating its limitations in dealing with the underlying patterns and complexity of the dataset.

Random Forest emerged as the most effective model, achieving a low RMSE of 0.16 and MAE of 0.07, accompanied by a very high R^2 of 0.94. This indicates that the Random Forest model is highly accurate in predicting occupancy rates and is able to explain almost 0.94 of the variances in the data. The low RMSE and MAE values confirm that prediction errors are minimal, making this the most appropriate model for this dataset.

The Multilayer Perceptron also performs well, with the lowest RMSE of 0.15 and an R^2 of 0.93, which shows that it is effective in capturing the structure of the data. Meanwhile, Gradient Boosting shows reliable performance, with an RMSE of 0.23 and an R^2 of 0.88, indicating strong, albeit slightly lower, accuracy compared to the Random Forest and Multilayer Perceptron models.

On the other hand, K-Nearest Neighbors has the highest error rates, with an RMSE of 0.48 and an R^2 of only 0.47, highlighting its struggle to generalize well with this data.

4.3 Guest reviews

In the second study, the model predicts review scores, which is important because it can lead to improvements in guest experiences, increased host revenue, enhanced platform quality control, and insights into the factors driving customer satisfaction. 30 variables were selected using Lasso regression. The results show that 20 variables are statistically significant (p-value < 0.05), as shown in Table 8. The R^2 value of 0.519 and adjusted R^2 of 0.518 indicate that the model explains 51.9% of the variance in review scores, demonstrating a moderate ability to capture the factors influencing guest satisfaction.

The results highlight the importance of host reliability. The superhost variable has a positive and significant impact on review scores, suggesting that properties hosted by superhosts tend to receive higher ratings, likely due to superior guest experiences. Similarly, the verified variable host identity has a positive effect on review scores, further reinforcing the value of trust and transparency in the hosting process. In addition, higher response rates are associated with higher ratings, suggesting that guests value quick communication from hosts. Interestingly, this is the opposite of the occupancy rate. On the other hand, a higher acceptance rate has a negative effect on review scores, suggesting that hosts who accept all bookings may not maintain the same standards as those who are more selective.

The variable Review Scores for Location has the strongest positive effect on review scores, emphasizing the critical role that location plays in guest satisfaction. Properties situated closer to the tourist center (Praça do Comércio) receive higher ratings, further highlighting the preference for centrally located accommodation. The number of amenities offered also correlates positively with review scores, confirming that guests favor properties that offer a wider range of services and facilities.

Interestingly, entire homes or apartments negatively correlate with review scores, possibly reflecting higher guest expectations for such properties. Price slightly but positively affects review scores, suggesting that guests associate higher prices with better quality experiences.

However, increased availability, especially over longer periods, shows a small but significant negative effect on review scores. This could imply that properties with high availability might be perceived as less desirable or less well-maintained.

Lastly, variables such as coffee maker, wifi, and the number of reviews do not show significant relationships with review scores, indicating that these factors are less critical in determining guest satisfaction in this dataset.

Overall, the findings provide key insights for hosts and platforms to optimize their listings by focusing on factors like host reliability, location, and amenities to improve guest satisfaction and ratings.

The variables are statistically significant (p-value <0.05) are presented in the table below:

Table 8 Statistically Significant Variable of Review Rating

Variables	Coefficient
constant	1.1739
host_is_superhost	0.1531
host_identity_verified	0.0681
minimum_nights	0.0013
maximum_minimum_nights	-0.0002
availability_30	-0.0013
availability_365	-0.0117
instant_bookable	-0.0288
calculated_host_listings_count	-0.0446
calculated_host_listings_count_private_rooms	-0.3071
calculated_host_listings_count_shared_rooms	-0.2837
entire home/apt	-0.0508

amenities_count	0.3184
host_response_rate	0.2566
host_acceptance_rate	-0.0848
price	0.0323
metro_station_count_500m	-0.0082
baixa_distance_km	0.0164
air conditioning	0.0358
TV	0.0195
review_scores_location	0.6409

Table 9 presents the performance metrics of various machine learning models, such as Random Forest, K-Nearest Neighbors, Gradient Boosting, and Multilayer Perceptron, which are applied to predict review score ratings. For the prediction, the dataset is split into 80% for training and 20% for testing. The metrics evaluated are RMSE, MAE, MSE, and R^2, the same as the occupancy rate. These metrics provide insight into the accuracy and predictive power of each model.

It is noteworthy that the Multilayer Perceptron, Random Forest, and gradient-boosting regressors were optimized to achieve the best performance. Specifically:

- The Multilayer Perceptron was configured with two layers (100, 50) and used the logistic activation function.
- The Random Forest model was trained using 300 trees.
- The Gradient Boosting model was trained using 100 trees.

The results highlight differences in model performance, suggesting that some models are more suitable than others for predicting review scores in this context. And the results are presented in the table below:

Table 9 Comparison of Machine Learning Model Performance of Review Rating

Model	Random Forest	K-Nearest Neighbors	Gradient Boosting	Multilayer Perceptron
RMSE	0.25	0.34	0.25	0.28
MAE	0.15	0.23	0.15	0.18
MSE	0.07	0.12	0.07	0.07
Train R^2	0.94	0.41	0.59	0.49
Test R^2	0.53	0.16	0.54	0.49

Overall, the comparative analysis of these machine learning models highlights the better performance of Gradient Boosting and Random Forest in predicting review scores. Both models achieve low error rates, with RMSE values of 0.25 for each, and relatively high explanatory power, with R^2 values of 0.54 for Gradient Boosting and 0.53 for Random Forest. These figures suggest that they are the most reliable choices for this dataset. However, while Random Forest achieves an impressively high Train R^2 of 0.94, this is significantly higher than its Test R^2 of 0.53, indicating some degree of overfitting. The model fits the training data extremely well but does not adapt as well to the unseen test data. However, their ability to capture complex, non-linear relationships within the data allows them to outperform simpler models, such as K-Nearest Neighbors, which struggle to achieve comparable accuracy.

K-Nearest Neighbors shows weaker performance, with the highest error rate with RMSE of 0.34 and a low R^2 of 0.16, reflecting its limited ability to explain the variance in review scores effectively. This poor adaptability suggests that K-Nearest Neighbors may not be suitable for this problem.

The Multilayer Perceptron also performs less effectively than Random Forest and Gradient Boosting, with an RMSE of 0.28 and a moderate R^2 of 0.49. While this suggests

that the MLP captures some aspects of the data structure, its performance still lags behind the tree-based models.

In conclusion, tree-based ensemble methods - particularly Gradient Boosting and Random Forest - are the most suitable for predicting review scores in this context, as they provide high accuracy and superior generalization across the dataset.

4.4 Discussion

It is important to compare our results with those of other authors, as this helps connect our findings to existing studies and demonstrates how our research contributes to the broader field. Such comparisons clarify whether our results align with or diverge from previous research, providing a deeper understanding of our study's relevance and impact.

From the results achieved, we observe that factors such as superhost status, host acceptance rate, and the number of amenities are the variables with more weight, which is consistent with the findings of Kirkos (2022). Kirkos highlights that host-related factors like the Superhost badge, a well-presented host profile, and quick response times are key determinants of Airbnb's success. Additionally, the provision of amenities and high guest ratings also emerged as strong predictors of performance in both studies. Moreover, in terms of predictive models, Kirkos found that Random Forest was the best-performing model, and in our study, we achieved similar results with Random Forest and Gradient Boosting as the top-performing models. This further supports the robustness of tree-based ensemble methods in predicting Airbnb performance.

Both studies highlight that Superhost status is a key feature influencing performance in Airbnb listings. According to Airbnb's official criteria, achieving Superhost status requires meeting specific standards, including 1) Hosting at least ten reservations or three reservations totaling 100 nights; 2) Maintaining a 90% or higher response rate; 3) Keeping a cancellation rate below 1%, with exceptions for major disruptive events or valid reasons; and 4) Maintaining an overall rating of 4.8 or higher.

A review counts toward Superhost status when both the guest and the host have submitted a review or when the 14-day review window closes.

These criteria ensure that superhosts demonstrate consistent occupancy, commitment, responsiveness, and high guest satisfaction, all factors that contribute to higher occupancy rates and better review ratings. As a result, the Superhost badge, as Liang referred to in his study, reflects a high level of reliability and quality, making it a crucial determinant of success in Airbnb performance metrics.

5. CONCLUSIONS

Airbnb, a primary platform in the context of the sharing economy, enables hosts to offer short-term rentals, ranging from entire homes to private rooms. With its rapid global expansion, understanding guest feedback has become increasingly important. This study focuses on evaluating the success of Airbnb listings in Lisbon, Portugal, by analyzing key performance metrics such as occupancy rates and review ratings using machine learning techniques with the objective of building models capable of predicting listing performance and identifying the factors that strongly influence the occupancy rate and the review rating.

In terms of predicting occupancy rates, the Random Forest and Multilayer Perceptron algorithms emerged as the most optimal algorithm, with Gradient Boosting also showing strong predictive power. However, K-Nearest Neighbors struggled to achieve accurate predictions, highlighting its limitations in capturing complex patterns in this dataset. In terms of review ratings, Gradient Boosting and Random Forest models provided the highest accuracy, highlighting their robustness to this task. The results indicate that host-related factors, such as superhost status and the number of amenities, are crucial for both occupancy rates and review ratings. In contrast, variables like price had a smaller impact on the performance metrics.

It is important to highlight some limitations in this research. First, Airbnb's policy of not disclosing exact occupancy rates prevents the use of an absolute performance measure, necessitating the use of approximations. Second, the occupancy rate was calculated at a specific point in time without considering the seasonal effects, limiting the ability to account for fluctuations in seasonal attributes and occupancy rates. Additionally, despite the dataset's size, it lacked key variables commonly used in other studies, such as cancellation policies, extra guest fees, etc. Including these variables could potentially improve the model's explanatory power, likely resulting in higher R^2 values and a more accurate representation of the factors influencing occupancy rates and review ratings.

Further research is needed to develop a more accurate model that better captures the factors influencing occupancy rates in Lisbon. This could involve revising the

occupancy rate formula and identifying more context-specific predictors relevant to the Lisbon market. Additionally, using datasets from platforms like AirDNA could provide valuable insights, as it would introduce new variables for studying performance metrics. It would also be beneficial to analyze the influence of major events in Lisbon on Airbnb ratings, exploring how local events impact both occupancy and guest satisfaction. These adjustments could offer a more comprehensive understanding of the dynamics affecting the city's short-term rental performance.

REFERENCES

Airbnb. About us. Retrieved September 1st, 2024, from https://news.airbnb.com/about-us/

Aparicio, S., Aparicio, J. T., & Costa, C. J. (2019)"Data Science and AI: Trends Analysis," 2019 14th Iberian Conference on Information Systems and Technologies (CISTI), Coimbra, Portugal, pp. 1-6, https://doi.org/10.23919/CISTI.2019.8760820.

Aparicio, J.T., Aparicio, M., Costa, C.J. (2023). Design Science in Information Systems and Computing. In: Anwar, S., Ullah, A., Rocha, Á., Sousa, M.J. (eds) Proceedings of International Conference on Information Technology and Applications. Lecture Notes in Networks and Systems, vol 614. Springer, Singapore. https://doi.org/10.1007/978-981-19-9331-2_35

Basuroy, S., Kim, Y., and Proserpio, D. (2020). Estimating the impact of Airbnb on the local economy: Evidence from the restaurant industry. *Available at SSRN 3516983.*

Bei, G. and Celata, F. (2023). Challenges and effects of short-term rentals regulation: A Counterfactual assessment of European cities. *Annals of Tourism Research, 101:103605.*

Chen, C.-C. and Chang, Y.-C. (2018). What drives purchase intention on Airbnb? perspectives of consumer reviews, information quality, and media richness. Telematics and Informatics, 35(5):1512– 1523.

Chen, D., Lou, H., and Van Slyke, C. (2015). Toward an understanding of online lending intentions: Evidence from a survey in China. *Communications of the Association for Information Systems*, 36(1):17.

Chiappa, G., Sini, L., and Atzeni, M. (2020). A motivation-based segmentation of Italian Airbnb users: an exploratory mixed method approach. *European Journal of Tourism Research*, 25:2505–2505.

Cocola-Gant, A., Jover, J., Carvalho, L., and Chamusca, P. (2021). Corporate hosts: The rise of professional management in the short-term rental industry. *Tourism Management Perspectives*, 40:100879.

Costa, C. J., & Aparicio, J. T. (2020). POST-DS: A methodology to boost data science. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-6). IEEE. https://doi.org/10.23919/CISTI49556.2020.9140932

Costa, C.J., Aparicio, J.T. (2021). A Methodology to Boost Data Science in the Context of COVID-19. In: Arabnia, H.R., et al. Advances in Parallel & Distributed Processing, and Applications. *Transactions on Computational Science and Computational Intelligence*. Springer, Cham.

Constantin, D.-L. and Dardala, A. E. (2015). A spatial analysis of tourism activity in Romania. In 55th Congress of the European Regional Science Association: "World Renaissance: Changing roles for people and places", 25-28 August 2015, Lisbon, Portugal. European Regional Science Association (ERSA).

Ert, E., Fleischer, A., and Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism management*, 55:62–73.

European Comission (2016). A European agenda for the collaborative economy. *In A European agenda for the collaborative economy*, pages 1–12.

Fagerstrøm, A., Pawar, S., Sigurdsson, V., Foxall, G. R., and Yani-de Soriano, M. (2017). That personal profile image might jeopardize your rental opportunity! on the relative impact of the seller's facial expressions upon buying behavior on Airbnb™. *Computers in Human Behavior*, 72:123–131.

Farmaki (2020). Impacts of covid-19 on peer-to-peer accommodation platforms: Host perceptions and responses. International journal of hospitality management, 91:102663.

Festila, M. and Müller, S. D. (2017). The impact of technology-mediated consumption on identity: The case of Airbnb. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, pages 54–63.

Filieri, R., Milone, F. L., Paolucci, E., and Raguseo, E. (2023). A big data analysis of covid-19 impacts on airbnbs' bookings behavior applying construal level and signaling theories. International Journal of Hospitality Management, 111:103461

Giana M. Eckhardt, F. B. (2015). The sharing economy is not about sharing at all. Harvard business review, 28(1), 881-898

Gil, J. and Sequera, J. (2022). The professionalization of Airbnb in Madrid: Far from a collaborative economy. *Current Issues in Tourism*, 25(20):3343–3362.

Guardado, M. I. A. (2022*). Airbnb em Lisboa: da partilha à profissionalização do alojamento local*. Master's thesis, Instituto de Geografia e Ordenamento do Território da Universidade de Lisboa

Gurran, N. and Phibbs, P. (2017). When tourists move in: how should urban planners respond to Airbnb? *Journal of the American planning association*, 83(1):80–92.

Guttentag, D. (2016). *Why tourists choose Airbnb: A motivation-based segmentation study underpinned by innovation concepts*. Doctor's thesis. University of Waterloo.

Guttentag, D. (2017). Regulating innovation in the collaborative economy: An examination of Airbnb's early legal issues. *Collaborative economy and tourism: Perspectives, politics, policies and prospects*, pages 97–128.

Gyódi, K. (2017). Airbnb and booking. com: Sharing economy competing against traditional firms. *WorNing Paper DELab UW*, 3.

Horn, K. and Merante, M. (2017). Is home sharing driving up rents? Evidence from Airbnb in Boston. *Journal of housing economics*, 38:14–24.

Jang, S. and Kim, J. (2022). Remedying Airbnb covid-19 disruption through tourism clusters and community resilience. Journal of Business Research, 139:529–542.

Ke, Q. (2017). Sharing means renting? an entire-marketplace analysis of Airbnb. *In Proceedings of the 2017 ACM on web science conference*, pages 131–139.

Kirkos, E. (2022). Airbnb listings' performance: determinants and predictive models. *European Journal of Tourism Research*, 30:3012–3012.

Lee, D. (2016). How Airbnb short-term rentals exacerbate Los Angeles's affordable housing crisis: Analysis and policy recommendations. *Harv. L. & PoI'y Rev.*,10:229.

Lee, S. and Kim, H. (2023). Four shades of Airbnb and its impact on locals: A spatiotemporal analysis of Airbnb, rent, housing prices, and gentrification. *Tourism Management Perspectives*, 49:101192.

Liang, S., Schuckert, M., Law, R., and Chen, C.-C. (2017). Be a "superhost": The importance of badge systems for peer-to-peer rental accommodations. *Tourism management*, 60:454–465.

Lutz, C. and Newlands, G. (2018). Consumer segmentation within the sharing economy: The case of Airbnb. *Journal of Business Research*, 88:187–196.

Maciel, F. K. (2019). *Os impactos jurídicos da economia compartilhada: uma análise do Airbnb.* Master thesis, Faculdade de Direito da Universidade de Lisboa

Moon, H., Miao, L., Hanks, L., and Line, N. D. (2019). Peer-to-peer interactions: Perspectives of Airbnb guests and hosts. International Journal of Hospitality Management, 77:405–414.

Regan, M. O. and Choe, J. (2017). Airbnb and cultural capitalism: enclosure and control within the sharing economy. *Anatolia*, 28(2):163–172.

Samadani, S. and Costa, C. J. (2021). "Forecasting real estate prices in Portugal : A data science approach," 2021 16th Iberian Conference on Information Systems and Technologies (CISTI), Chaves, Portugal, pp. 1-6, https://doi.org/10.23919/CISTI52073.2021.9476447.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. Journal of data warehousing, 5(4), 13-22.

Shen, L. and Wilkoff, S. (2022). Cleanliness is next to income: The impact of covid-19 on short-term rentals. Journal of Regional Science, 62(3):799–829.

Simic, V. and Liem, A. (2023). The sharing economy's success: Advantages, drawbacks, and applications. *Proceedings of the Design Society*, 3:3493–3502.

Sparks, B. A. and Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism management*, 32(6):1310–1323.

Suciu, A. M. (2016). The impact of Airbnb on local labour markets in the hotel industry in Germany. Available at SSRN 2874861.

Suh, J., Tosun, C., Eck, T., and An, S. (2022). A cross-cultural study of value priorities between us and Chinese Airbnb guests: An analysis of social and economic benefits. Sustainability, 15(1):223.

Thaichon, P., Surachartkumtonkun, J., Singhal, A., and Alabastro, A. (2020). Host and guest value co-creation and satisfaction in a shared economy: The case of Airbnb. *Journal of Global Scholars of Marketing Science*, 30(4):407–422.

Tussyadiah, I. P. (2016). Factors of satisfaction and intention to use peer-to-peer accommodation. *International Journal of Hospitality Management*, 55:70 80. Wachsmuth, D. and Weisler, A. (2018). Airbnb and the rent gap: Gentrification through the sharing economy. *Environment and planning A: economy and space*,50(6):1147–1170.

Wu, J., Ma, P., and Xie, K. L. (2017). In sharing economy we trust: The effects of host attributes on short-term rental purchases. *International Journal of Contemporary Hospitality Management*, 29(11):2962–2976.

Yang, S.-B., Lee, K., Lee, H., and Koo, C. (2019). In Airbnb we trust: Understanding consumers' trust-attachment building mechanisms in the sharing economy. *International Journal of Hospitality Management*, 83:198–209.

Zhao, J. and Peng, Z. (2019). Shared short-term rentals for sustainable tourism in the social-network age: The impact of online reviews on users' purchase decisions. *Sustainability*, 11(15):4064.

APPENDICES

Figure 6: Distribution of the Variable host is superhost

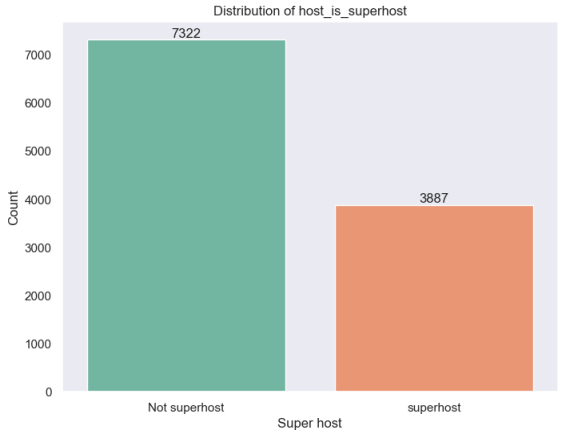


Figure 7: Distribution of the Variable instant bookable

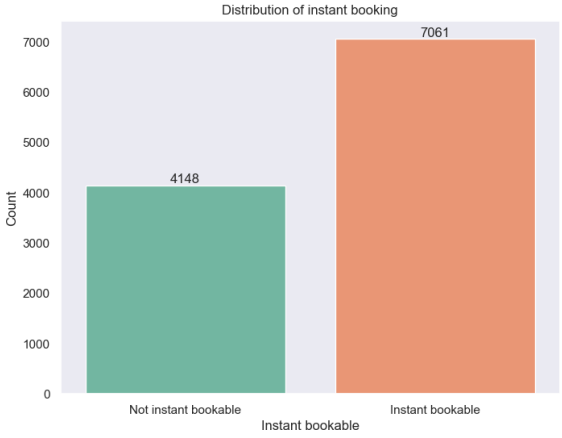


Figure 8: Distribution of the Variable host identity verified

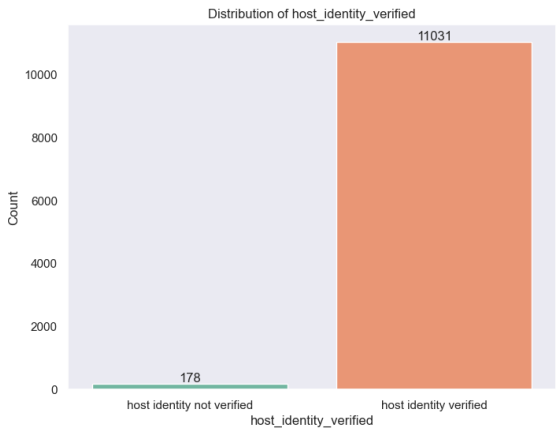


Figure 9: Distribution of the Variable host has a profile picture

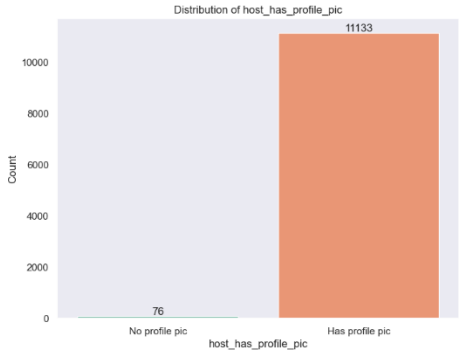


Figure 10: Distribution of the Variable host acceptance rate

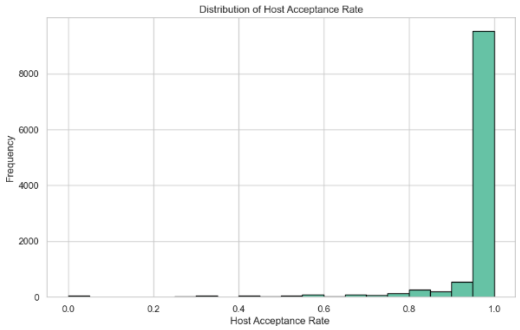


Figure 11: Transformation of Price

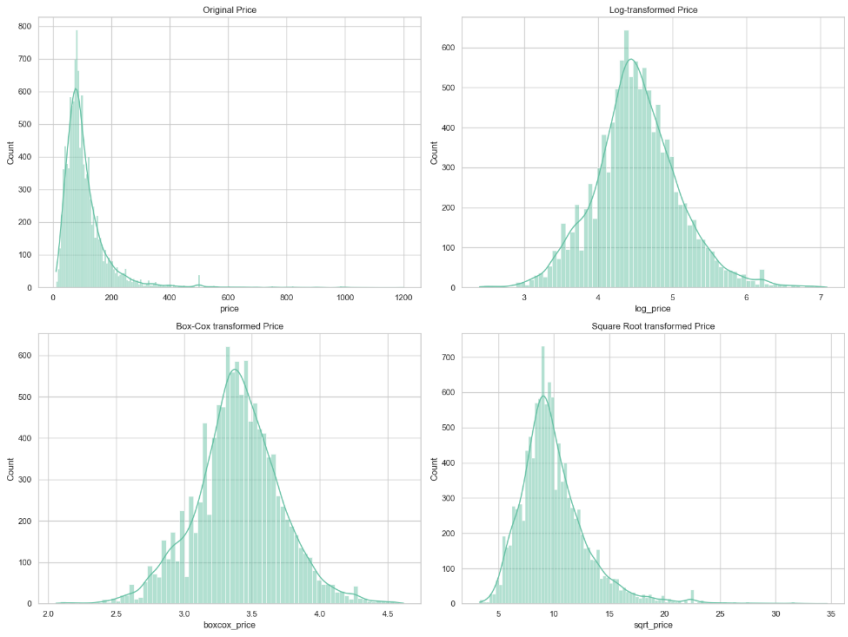


Figure 12: Count of Most Commonly Used Words in Amenities

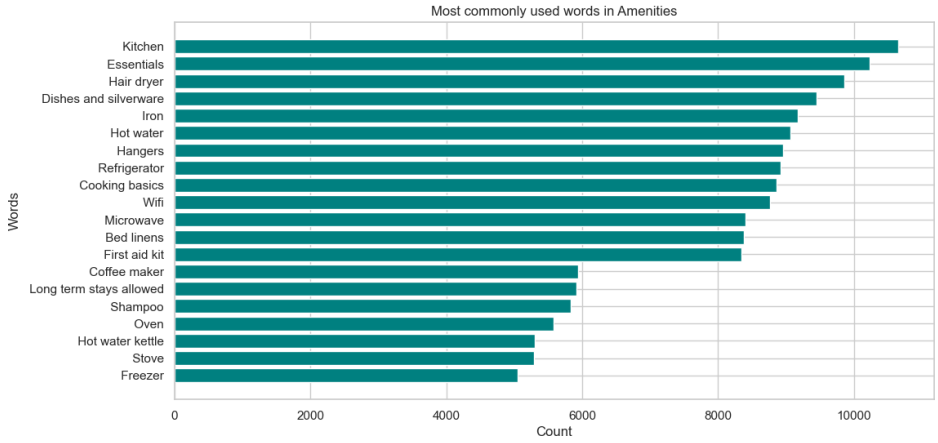


Figure 13: Gradient Boosting Regressor Feature Importance (Review Rating)

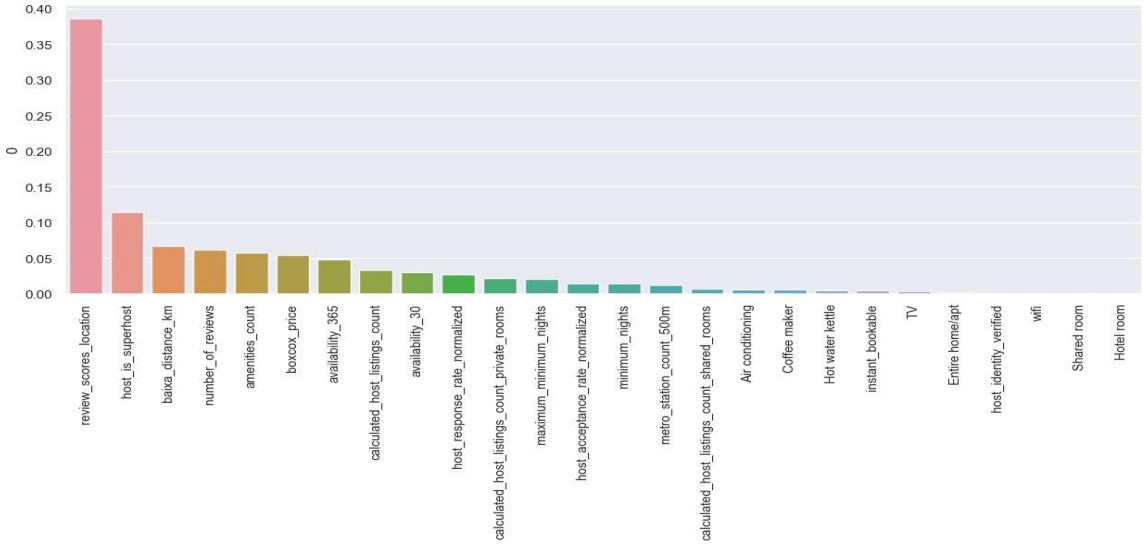


Figure 14: Random Forest Regressor Feature Importance (Occupancy Rate)