NOVA IMS Information Management School

MEGI

Master Degree Program in

Statistics and Information Management

Exploring Airbnb Price Determinants in the Porto Metropolitan Area

A Geographically Weighted Regression Analysis

Ana Catarina Nunes Albasini

Project Work

presented as partial requirement for obtaining the Master Degree in Statistics and 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

Exploring Airbnb Price Determinants in the Porto Metropolitan Area

A Geographically Weighted Regression Analysis

by

Ana Catarina Nunes Albasini

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

Supervised by

Miguel de Castro Neto, Phd, NOVA Information Management School

Bruno Jardim, Phd, NOVA Information Management School

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or 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, 13th July 2024

ACKNOWLEDGEMENTS

I want to thank my supervisors, Professor Miguel de Castro Neto and Professor Bruno Jardim, for their expertise and guidance throughout this journey. I'm also very grateful to my colleagues for their valuable insights and knowledge, and to my family and friends for always supporting me and believing in me.

ABSTRACT

The sharing economy has transformed traditional industries by enabling peer-to-peer exchanges of goods and services, with Airbnb being a prominent example in the accommodation sector, allowing property owners to rent out their spaces to travelers. This study aimed to explore the determinants of Airbnb pricing in the Porto Metropolitan Area, Portugal, and to understand how these determinants vary across different neighborhoods, by employing both Ordinary Least Squares (OLS) regression and Geographically Weighted Regression (GWR), where the GWR performed better than the OLS in capturing the spatial variations in pricing determinants. By identifying the key determinants that influence Airbnb pricing in Porto, this research provides valuable insights into the factors affecting rental prices, such as listing type, amenities, location, host characteristics, and proximity to attractions. The findings highlight the importance of spatial heterogeneity in understanding local market dynamics and can inform pricing optimization strategies for Airbnb hosts, policy development for local authorities, and future research on the economic impacts of the sharing economy. This study contributes to a deeper understanding of how Airbnb listings are priced and the spatial variations that affect these prices within an urban context.

KEYWORDS

Airbnb; OLS; GWR; Porto Metropolitan Area; Price Determinants

Sustainable Development Goals (SDG):

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CRISP-DM	Cross-Industry Standard Process for Data Mining
EDA	Exploratory Data Analysis
GWR	Geographically Weighted Regression
OLS	Ordinary Least Squares

1. INTRODUCTION

In recent years, Airbnb has revolutionized the hotel industry by redefining how people worldwide find and offer accommodation. Created in 2008, Airbnb has rapidly grown into a significant player in the hospitality sector, operating in numerous locations globally. The platform provides a wide range of accommodation options, including personalized experiences in addition to entire homes and rented rooms. This creative approach to housing not only gives tourists affordable and distinctive substitutes for conventional hotels, but also enables people to monetize their assets and support the thriving sharing economy.

As Airbnb continues to transform the hospitality and travel sectors, it has become the subject of various research studies. Researchers have examined a range of aspects of Airbnb's influence, such as how it has impacted local communities' economies and cultures, impacted traveler movements, and changed the competitive environment of the accommodation sector (Jiménez et al., 2022; Kuhzady et al., 2022; Quattrone et al., 2022).

The goal of this study is to advance our knowledge of the dynamics of Airbnb pricing, using as a case study the Porto Metropolitan Area, in Portugal. The choice to employ a Geographically Weighted Regression (GWR) for this study is based on its ability to capture spatial variations in price determinants. Traditional regression models may overlook subtle regional differences influencing Airbnb pricing, whereas a GWR allows for a more nuanced analysis by accounting for spatial heterogeneity within Porto.

The findings of this study hold significant implications for researchers, policymakers, and Airbnb hosts in the Porto's Metropolitan Area short-term rental market. By identifying the determinants of Airbnb pricing and uncovering subtle regional differences, this study can contribute to a better understanding of the market dynamics and inform strategies for pricing optimization, policy development, and host decision-making. The insights gained from this analysis can also serve as a foundation for future research on Airbnb pricing in other cities or regions.

To achieve this goal, the study addresses the following research questions:

- RQ1: What are the most important determinants of Airbnb prices in the Porto Metropolitan Area?
- RQ2: Does the proximity of Airbnb listings to attractions affect the price?
- RQ3: How do the impacts of these determinants on Airbnb prices vary across different neighborhoods?

2. LITERATURE REVIEW

2.1. AIRBNB AND THE SHARING ECONOMY

In recent years, Airbnb has emerged as a prominent player in the sharing economy, reshaping global accommodation dynamics. By seamlessly connecting hosts and guests worldwide, Airbnb not only provides travelers with a wide range of cost-effective options and personalized experiences but also empowers individuals to monetize their assets (Guttentag, 2015). The platform's innovative model allows users to rent out their homes, spare rooms, or even couches, presenting a distinctive and economical alternative to traditional lodging models such as hotels, motels, and B&Bs (Hall et al., 2022).

This breakthrough in the sharing economy has not only revolutionized the hospitality sector but also posed a formidable challenge to the established dominance of traditional hotel companies, a narrative underscored by the comprehensive study conducted by Quattrone et al. (2022), emphasizing Airbnb's competitiveness with conventional accommodation firms and its profound impact on the movement of tourists and visitors.

Adding complexity to this transformative landscape, Zervas et al. (2017) emphasize the conflict between Airbnb and the hotel industry, which is particularly noticeable in Texas. Their research uncovers a concrete detrimental effect on hotel revenue, particularly on budget accommodations. The attractiveness of Airbnb to cost-conscious tourists, based on informal accommodations, strongly corresponds to the clientele of more affordable hotels. Nevertheless, this competition is complex, especially for hotels that do not specifically serve business tourists. In contrast to conventional hotels, Airbnb hosts may not offer amenities such as conference rooms, making these hotels more liable in specific market niches. The presence of business travelers, who are generally less sensitive to price and have specific amenity preferences, adds complexity to the competitive landscape.

Huang et al. (2023) examined paradoxes in the sharing economy, specifically focusing on Airbnb. The Conformity-Distinctiveness Paradox argues that economic success is enhanced by differentiating from hotels, but only up to a certain point. Beyond this point, the advantages begin to decline. On the other hand, the Economic-Social Paradox suggests that deviating from established hotel standards can improve social performance, but it may not necessarily lead to the best economic results. Their study revealed the complex balance between host uniqueness and economic success, contributing to the discussion on sharing economy platforms.

Furthermore, the integration of the sharing economy in Airbnb facilitates the development of a communal atmosphere and fosters cultural exchange, as guests frequently engage with their hosts and gain insights into the local culture and traditions. The focus on community involvement is consistent with Tussyadiah's (2016) research, which examines the factors that influence guest satisfaction and the likelihood of returning to peer-to-peer lodgings. The study

reveals that enjoyment, monetary value, and social advantages are important factors that strongly influence guest satisfaction. Specifically, social benefits have a particularly strong impact on guests who stay in private rooms with hosts. This cultural exchange dynamic further enhances Airbnb's diversified influence, demonstrating its capacity to not only reinvent economic transactions but also reshape the fundamental nature of hospitality and vacation experiences. This underscores the platform's revolutionary significance that extends beyond economic factors.

2.2. AIRBNB'S IMPACT ON PROPERTY MARKETS AND REGULATORY CONSIDERATIONS

Regarding the influence of Airbnb on property values and rental prices, these effects of Airbnb vary depending on their features. The biggest influence on rent, housing value, and gentrification comes from multi-unit hosts' entire home listings (Lee & Kim, 2023). Also, Franco et al. (2019) provide significant insights into the influence of Airbnb on property values and rental prices, using Portugal as a case study. Their research reveals a significant surge in property values and rental prices after Airbnb's introduction, especially in locations that are popular among tourists. This study emphasizes the economic consequences of sharing economy platforms on local property markets. In Sydney, Airbnb has a generally favorable effect on house values, with a 1% rise in Airbnb density translating into a 2% increase in the price of properties, with these prices varying geographically (Thackway et al., 2022).

The study conducted by Jiménez et al. (2022) explores the effects of Airbnb on the tourism industry in Spain, focusing on the specific characteristics of different regions, and concluding that it has a positive impact on local tourism markets. The study highlights the need to understand the local dynamics within the larger context of the sharing economy. The systematic review conducted by Kuhzady et al. (2022) provides additional evidence to support the notion that Airbnb has played a key role in transforming the way lodging operates and influencing consumer choices.

Examining the regulatory landscape, Estevens et al. (2023) highlight the importance of establishing a balance between promoting innovation and guaranteeing consumer protection. This emphasizes the complex interaction between technical progress, regulatory structures, and the development of the sharing economy. (Luo, 2023) offers a macroeconomic perspective on the sharing economy's growth, emphasizing the need for regulatory interventions.

2.3. AIRBNB PRICE DETERMINANTS AND PREDICTION

Bode et al. (2021) examined the factors influencing Airbnb listing prices in Porto's Cedofeita area using a Weighted Least Squares (WLS) regression technique. This study used a dataset of 4 467 Airbnb listings to focus on studying the factors influencing prices before and during the COVID-19 outbreak. The factors that are included in the analysis are host reputation, qualities, location, and rental policies, among others. The regression model used the WLS in the event

of heteroskedasticity and assumes linearity, constant variance, normality, and independence of the disturbances. The study's conclusions provide insight into guest preferences, showing that regardless of the pandemic's effects, guests are consistently prepared to spend more for features like quality, privacy, and space. Interestingly, the size of the property had less of an impact because there was less demand, and the pandemic's increased focus on security and safety minimized the importance of rental regulations. During the COVID-19 epidemic, trust, reputation, and superhost status became important factors in the selection process for guests. Moreover, it highlights shifts in the significance of specific listing attributes, emphasizing the complex dynamics of the Airbnb market amid the extraordinary pandemic events.

In Fernandes' (2019) study, the author specifically examined the process of predicting the prices of Airbnb rentals in Lisbon by applying local spatial regressions. The study applied an OLS Regression and a GWR to consider spatial differences in listing prices, where the second model outperformed the first. The parish location is a significant factor, as some areas—like Parque das Nações and Santo António—have greater prices than others. The type of listing—private or shared room or complete home/apartment—has a big influence on price, with entire home/apartment listings often being more expensive. The results suggested that the prices of various types of Airbnb listings are affected by specific circumstances, highlighting the importance of considering local dynamics when developing pricing strategies.

Zhang et al. (2017) analyzed key factors influencing Airbnb prices in Nashville using a General Linear Model (GLM) and a GWR, finding that proximity to the convention center significantly affects pricing. Once again, GWR outperforms GLM in accuracy and variable selection, with price sensitivity to location varying across different areas. There is a negative correlation between prices and reviews and rating scores, indicating that lower expectations can result in higher satisfaction and ratings.

Gyódi & Nawaro (2021) investigated the factors that influence Airbnb prices in 10 major European cities by also employing a spatial econometrics methodology. The study found that the size, quality, and location characteristics had a substantial impact on Airbnb costs in European cities. Listings close to restaurants, tourist destinations, city centers and metro stations typically obtain higher rates. Moreover, a noteworthy pattern is observed among experienced hosts in charge of several listings: those in charge of over four listings typically charge more for their accommodations. This implies that the features of hosts and the surrounding area have a big impact on how the Airbnb ecosystem sets prices.

Based on two analytical techniques, Hong & Yoo (2020) made various conclusions on Airbnb pricing in Los Angeles and New York. The study highlights the significance of geographical characteristics in explaining location-specific variations in pricing tactics. It has been observed that the OLS is inadequate in explaining spatial variance, which could result in errors when determining meaningful associations. By taking into account spatial variance both within and between cities, the study presents the Multiscale Geographically Weighted Regression (MGWR), a more effective technique that provides a better understanding of pricing

strategies. MGWR is especially commended for its ability to illustrate the relationship between price variables and geographic location, which offers a useful tool for examining spatially varied interactions. Factors such as the number of guests a listing can accommodate, the number of bedrooms and bathrooms, and whether the rental is shared, or an entire unit/house were recognized as key influencers. Additionally, considerations like the cancellation policy, the location's proximity to tourism destinations, the density of nearby Airbnb listings, the poverty ratio of the area, and the reputation of the host, including superhost status and service duration, were all acknowledged as variables that may impact pricing.

Alharbi (2023) constructed an enduring price forecasting framework for Airbnb rentals by employing various Machine Learning (ML) techniques. The study revealed key factors that impact price forecasts, such as the sentiment polarity derived from reviews and the qualities of the property. On the other hand, the influence of amenities on price prediction was minimal. The Lasso and Ridge regression models demonstrated a higher performance, highlighting the efficacy of machine learning in forecasting Airbnb listing values.

Liu (2021) constructed resilient machine learning models to predict Airbnb pricing in Holland, such as K-Nearest Neighbors (KNN), Multiple Linear Regression (MLR), Lasso Regression, Ridge Regression, Random Forest (RF), Gradient Boosting, with Extreme Gradient Boosting (XGBoost) being determined as the most effective model. The study focused on the constraints of prior research and highlighted the significance of thorough exploratory data analysis and parameter modification to improve model performance.

Moreover, H. Wang (2023) sought to forecast the values of Boston Airbnb listings by employing RF, Linear Regression, KNN, and Gradient Boosting models, with Gradient Boosting demonstrating superior performance compared to the other models. The findings yielded valuable insights into the determinants that impact the pricing of Airbnb accommodations in Boston.

In their study, Nunes (2023) employed geographical characteristics and implemented the OLS and Gradient Boosting techniques to forecast and elucidate the pricing of Airbnb accommodations in Lisbon. The research not only enhanced the performance of the model but also made a significant contribution to the literature by highlighting the significance of model interpretability. The study's results outperformed previous research by Liu (2021).

Chattopadhyay & Mitra (2019) take a comparative approach, analyzing a considerable dataset from 11 U.S. cities employing the OLS, RF, and Decision Tree approaches. In terms of prediction accuracy, the RF model performs better than the OLS and Conditional Inference Trees (CTree) models. Factors such as the number of bathrooms, accommodations, and positive review ratings, alongside specific amenities like elevators and complimentary parking, consistently hold significant influence on pricing.

Also, Wang & Nicolau (2017) examined a total of 180 533 rental listings in 33 different cities using the OLS. The study finds that a variety of factors, including customer evaluations, rental policies, property features, amenities, and host traits, interact to affect pricing. Specifically, having more listings, being a superhost, and having identities that can be verified are all linked to higher rates; on the other hand, host profile photos are linked to lower rates.

This thorough summary of many studies offers a varied range of insights into the factors influencing Airbnb pricing. The intricacy of pricing decisions in the Airbnb ecosystem is clear, ranging from the macro-level influence of geographic and economic factors to the micro-level considerations of property qualities. The growing popularity of sophisticated spatial regression models, like MGWR and GWR, highlights how crucial it is to take spatial differences into account when analyzing pricing dynamics. To effectively navigate the complex world of Airbnb pricing methods, hosts, researchers, and legislators must all consider these diverse views as the short-term rental market develops.

Table 1 summarizes the key findings from these studies, providing an overview of the methodologies used and the primary conclusions drawn.

Table 1: Summary of Findings from Studies on Airbnb Price Determinants and Prediction

Study	Location	Methods	Key Factors	Findings
(Bode et al., 2021)	Porto	WLS	Host reputation, qualities, location, rental policies	Guests value quality, privacy, space. Size of property less important. Safety and trust crucial during COVID-19.
(Fernandes, 2019)	Lisbon	OLS, GWR	Parish location, listing type	Location and type of listing crucial - entire home/apartment listings are more expensive. Parque das Nações and Santo António have higher prices.
(Zhang et al., 2017)	Nashville	GLM, GWR	Proximity to convention center	GWR outperforms GLM. Negative correlation between prices and reviews/ratings.
(Gyódi & Nawaro, 2021)	10 European cities	Spatial Econometrics	Size, quality, location, host characteristics	Listings near amenities and by experienced hosts charge more.

Study	Location	Methods	Key Factors	Findings
(Hong & Yoo, 2020)	Los Angeles, New York	OLS, MGWR	Number of guests, bedrooms, bathrooms, listing type, cancellation policy	MGWR better explains spatial variance. Key factors include location, density of listings, and host reputation.
(Alharbi, 2023)	Barcelona	Lasso, Ridge, Bayesian, KNN, SVR, Decision Tree, Sentiment Analysis	Sentiment polarity, property qualities, amenities	Sentiment polarity important for price prediction. Amenities have minimal impact. Lasso and Ridge regression perform better.
(Liu, 2021)	Holland	KNN, MLR, Lasso, Ridge, RF, XGBoost	-	XGBoost most effective. Emphasizes thorough data analysis and parameter tuning.
(H. Wang, 2023)	Boston	RF, Linear Regression, KNN, Gradient Boosting	-	Gradient Boosting outperforms other models on price prediction.
(Nunes, 2023)	Lisbon	OLS, Gradient Boosting	Accommodates, number of reviews, type of listing, geographical characteristics	Gboost outperforms. Emphasizes model interpretability.
(Chattopadhyay & Mitra, 2019)	11 U.S. cities	OLS, RF, Decision Tree	Number of bathrooms, accommodations, positive reviews, amenities	RF performs better. Consistent influence of bathrooms, accommodations, and positive reviews.
(D. Wang & Nicolau, 2017)	33 cities	OLS	Customer evaluations, rental policies, property features, amenities, host traits	Verified identities and superhost status linked to higher rates. Host profile photos linked to lower rates.

3. DATA & METHODOLOGY

To approach the analysis of Airbnb pricing in the Porto Metropolitan Area, we adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology (Schröer et al., 2021). CRISP-DM provides a structured and comprehensive framework that guides the process of data mining and analysis through six distinct phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, which are described in detail below:

1. **Business Understanding**: This initial phase focuses on understanding the project objectives and requirements from a business perspective, where the research gap and question are formulated. For this study, the primary goal is to identify the determinants of Airbnb pricing in the Porto Metropolitan Area and understand the spatial variations in these determinants. This phase was already employed in the previous sections.
2. **Data Understanding**: The second phase involves collecting initial data and familiarizing with it, which includes data description and data exploration. For this investigation, the dataset was sourced from Inside Airbnb[1], providing detailed information on 13 601 Airbnb listings in Porto, including attributes like price, location, amenities, host information, and reviews.
3. **Data Preparation**: In this phase, the data is prepared for modeling by performing tasks such as data cleaning, data transformation, and data reduction. For our analysis, data preparation steps included cleaning inaccuracies, handling missing values, and refining the dataset by removing irrelevant attributes and creating new relevant variables. This step ensures that the dataset is of high quality and relevant to our research questions.
4. **Modeling**: This phase involves selecting and applying various modeling techniques and calibrating their parameters to optimal values. In this study, the OLS was employed to identify the most influential features on the Airbnb pricing and then we utilized the GWR to capture spatial variations.
5. **Evaluation**: In this phase, the models are evaluated to ensure they meet the business objectives and to determine the best model. For this study, we evaluated the GWR model's ability to explain the spatial variations in Airbnb pricing and compared its performance with the OLS model.
6. **Deployment**: The final phase involves applying the research findings in a practical context. For this research, the deployment phase involved summarizing the results and conclusions in this thesis. This will help inform future research and provide insights into optimizing Airbnb pricing strategies.

[1] https://insideairbnb.com/

This methodology ensures a thorough and iterative approach to data analysis, enhancing the reliability and validity of the results.

3.1. STUDY AREA: PORTO METROPOLITAN AREA

The study area for this research is the Porto Metropolitan Area, located in the northwest of Portugal. Porto is the second-largest city in Portugal and is known for its rich history, cultural heritage, and economic significance. The metropolitan area encompasses not only the city of Porto but also several surrounding municipalities, making it a diverse and dynamic region.

Porto's vibrant tourism industry, historical landmarks, and scenic beauty attract millions of visitors annually. The metropolitan area includes a variety of neighborhoods, each with distinct characteristics that can influence Airbnb pricing. From the bustling city center with its array of attractions and amenities to quieter suburban areas, the diversity within the Porto Metropolitan Area provides a rich context for examining spatial variations in Airbnb pricing. The map in Figure 1 shows the neighborhoods of the Porto Metropolitan Area.

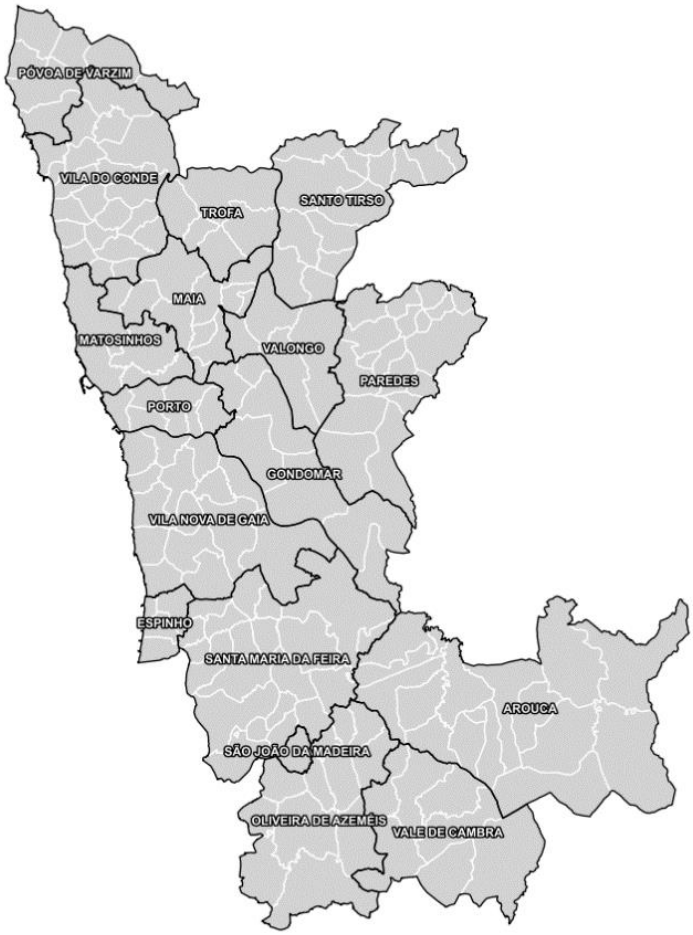


Figure 1: Porto Metropolitan Area - Neighborhoods

3.2. DATA COLLECTION

For this investigation, the dataset was sourced from Inside Airbnb, an independent initiative that compiles and disseminates data on Airbnb listings from various global cities. Serving as the analytical foundation, this dataset contains 13 601 rows, each row corresponding to a different listing and 75 columns corresponding to various aspects such as price, location, amenities, host information, and reviews, among others. It was scrapped on December 17, 2023.

3.3. DATA PREPROCESSING

This phase started by removing variables that are not relevant and uninformative such as features like textual descriptions and URLs, for example, description, neighborhood_overview, host_about, name, host_url, picture_url, license and many others since they brought no relevant information to the analysis and our models are focused on numerical data. Regarding data transformation, boolean categories like host_is_superhost and host_identity_verified were turned into binary variables. Also, the following numerical variables were binned into groups, turning into categorical variables (Table 2). Before modeling, these variables were one hot encoded in order to be included in the regression analysis.

Table 2: Categories for Categorical Variables

Variable	Categories
review_scores_rating review_scores_accuracy review_scores_cleanliness review_scores_checkin review_scores_communication review_scores_location review_scores_value	○ 0-4/5 stars ○ 4-4.5/5 stars ○ 5/5 stars ○ no reviews
host_response_rate host_acceptance_rate	○ 0-49% ○ 50%-89% ○ 90%-99% ○ 100% ○ unknown
time_since_first_review time_since_last_review	○ 0-2 weeks ○ 2-8 weeks ○ 2-6 months ○ 6-12 months ○ 1+ year
host_response_time	○ within an hour ○ within a few hours

Variable	Categories
	○ within a day ○ a few days or more ○ unknown

Porto is a major tourist destination with numerous historical and cultural landmarks, such as the Lello Bookshop, Luís I Bridge, Clérigos Tower, Bolhão Market, São Bento Station, as well as Francisco Sá Carneiro Airport serving as a major gateway for international visitors. Previous research has demonstrated that proximity to tourist attractions significantly impacts Airbnb prices, as accommodations closer to these attractions tend to command higher prices due to the added convenience for tourists (Xu et al., 2020). Additionally, accessibility to transportation hubs such as airports is crucial. Properties located nearer to airports generally attract higher prices because of the convenience they offer to travelers who prioritize easy access to transportation (Chang & Li, 2021). Given the importance of these locations, we created two new distance variables and added them to the dataset: the distance to the nearest tourist attraction (dist_nearest_attraction) and the distance to the airport (dist_airport). These distances were calculated using the geodesic distance formula from the geopy Python module, which employs the algorithm given by (Karney, 2013). This approach provides a comprehensive understanding of how location and accessibility contribute to price variations, addressing the research question more effectively.

A detailed overview of the variables remaining, including their respective descriptions are identified in Table A 1 in the Appendix.

3.4. EXPLORATORY DATA ANALYSIS (EDA)

The purpose of this chapter is to perform a comprehensive EDA on the dataset that was obtained from Porto Airbnb listings. EDA is a crucial early stage of the research process that aims to clarify and condense the primary characteristics of the dataset. The importance of this investigation resides in its ability to highlight minor details and patterns in the data that are not readily apparent through conventional modeling or hypothesis testing.

3.4.1. TARGET VARIABLE: LISTING'S PRICE

The focus of this analysis is the pricing of Airbnb listings for overnight stays. Originally recorded in dollars, prices were converted to euros, reflecting the primary currency in Porto, the area of study. The average listings' price for a night stay in Porto is approximately 85€. An examination of the dataset reveals notable variability in listing prices, ranging from 9€ to 3 615€.

To evaluate the distribution of listing prices, both a histogram and a boxplot were created (Figure 2). The histogram revealed a strongly right skewed distribution, indicating a presence of outliers—common in pricing data, where a minority of listings are priced significantly higher

than the majority. This observation was further supported by the boxplot, suggesting that adjustments, such as transformations, might be necessary to mitigate bias in the results of subsequent analyses.

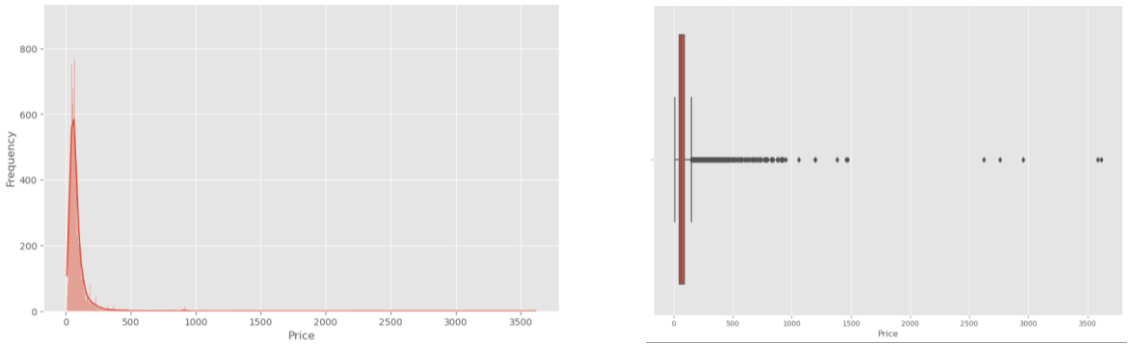


Figure 2: Price distribution before removing outliers and log-transformation

Addressing the issue of skewed distribution requires first dealing with missing values, which account for approximately 5,78% of the dataset. The median was selected as the imputation method for its resilience to outliers. Unlike the mean, which can be heavily influenced by extreme values, the median provides a more representative figure of the dataset's central tendency, ensuring a balanced approach to filling in missing data without distorting the original distribution. This method is particularly effective in maintaining the integrity of the dataset when preparing for further analysis.

Concerning outliers, the flexibility Airbnb hosts have in setting prices for their properties introduces significant variability in listing prices, often resulting in outliers. After evaluating various scenarios, a decision was made to cap the listing prices at 500€, thus excluding all observations with prices above 500€. This action resulted in discarding less than 1% of the data, leaving 13 512 listings in the dataset. Consequently, the adjusted mean price is now 78€, with a standard deviation of 54€, offering a more standardized basis for further analysis.

An additional strategy to address skewness involves applying a logarithmic transformation to the price variable. By converting the dependent variable to its logarithm, the distribution of prices improved markedly. The second histogram (Figure 3), illustrating the logarithm of price, demonstrates that the distribution has shifted from a right skew to a form closely resembling a normal distribution.

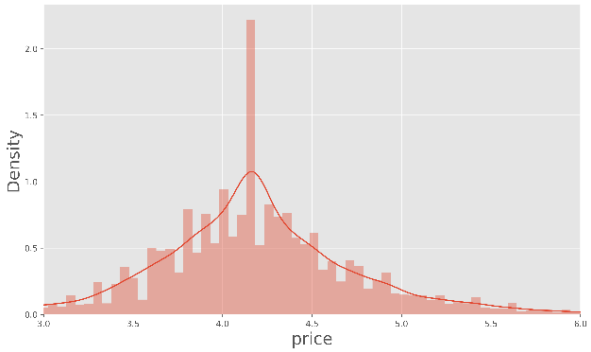


Figure 3: Price distribution after removing outliers and log-transformation

This transformation, coupled with the exclusion of listings priced above 500€, significantly enhanced the dataset's suitability for further analysis. The combined effect of these adjustments was a notable reduction in skewness, bringing the price distribution into closer alignment with the assumptions required for many statistical modeling techniques.

3.4.2. EXPLANATORY VARIABLES

When analyzing the distribution of listings across neighborhoods (Figure 4), a significant concentration was observed in the central neighborhood of Cedofeita, Ildefonso, Sé, Miragaia, Nicolau, and Vitória, accounting for 6 947 listings. The next substantial cluster was found in Bonfim, with 1 386 listings. These areas, pivotal to Porto's urban center, showcased average nightly prices of 75€ and 69€, respectively. This spatial distribution highlights the prominence of certain neighborhoods in the Airbnb market, potentially influenced by their central location and appeal to visitors.

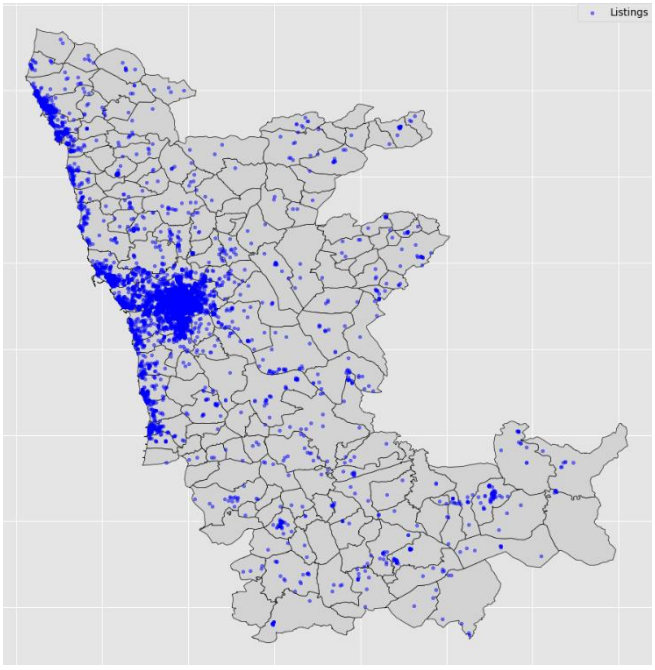


Figure 4: Airbnb Listings' distribution across Porto's Neighborhoods

In analyzing the growth and development of the Airbnb market in Porto, one pivotal aspect examined was the variation in the number of listings per host over the years (Figure 5). A boxplot visualization was utilized, mapping the year each host joined Airbnb against the log-transformed count of their listings. This methodological choice, particularly the log-transformation, was instrumental in mitigating the skewness inherent in the data, where a minority of hosts possessed a significantly higher number of listings. The transformation also rendered the data variance more uniform, thus facilitating a more nuanced interpretation. The resultant plot underscores not just the growth in listings but also how the hosting dynamic within Porto's Airbnb market has evolved, highlighting years of particular expansion or stabilization. This analysis offers a macroscopic view of the market's maturation over time, illustrating trends that underpin both market-wide shifts and individual hosting behaviors.

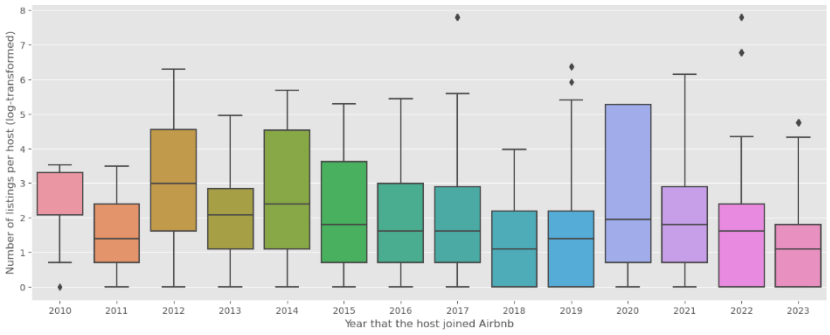


Figure 5: Change per year in the number of listings per host

Unsurprisingly, it's evident that properties with the capacity to accommodate more guests tend to command higher median prices per night (Figure 6). This trend continues up to a point where the price increase starts to taper off, which in this dataset, occurs around the threshold of accommodating 10 guests. Beyond this number, while prices continue to rise for properties that can host more guests, the rate of increase is less pronounced, suggesting a plateau in the price versus capacity curve. This pattern might be attributed to the practical limitations of guest accommodations or a smaller market for such large properties, thus leading to the phenomenon of diminishing returns.

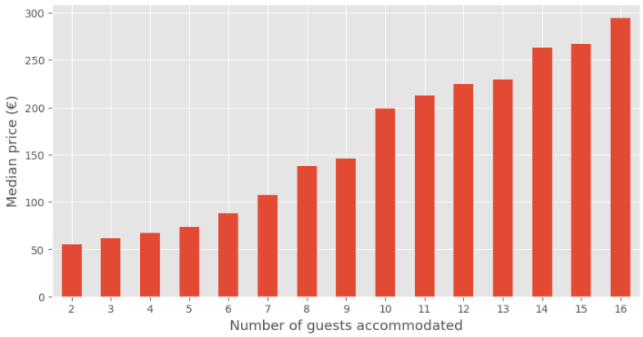


Figure 6: Median price of Airbnb accommodations for varying numbers of guests

The dataset indicates that a substantial majority of the listings, approximately 81%, are categorized as entire homes (Figure 7), meaning guests have the entire property to themselves

during their stay. Most of the remaining listings are private rooms, which typically include a bedroom and possibly a private bathroom while sharing the property with others. Shared rooms, where guests might share space with the property owner or other guests, make up less than 2% of the listings. Hotel rooms, offering a more traditional hospitality experience within the platform, constitute less than 1% of the listings. This distribution suggests that Airbnb users in Porto have a strong preference for the privacy and space offered by renting entire homes.

Analyzing the mean prices for each listing type (Figure 8), we find that entire homes/apartments have a mean price of approximately 82,52€, while hotel rooms are the most expensive, with a mean price of approximately 99,64€. Private rooms have a mean price of about 58,75€, and shared rooms are the least expensive, with a mean price of approximately 24,34€. This pricing information highlights the premium associated with entire homes and hotel rooms, reflecting their higher levels of privacy and amenities. Conversely, shared rooms are more budget-friendly, likely appealing to cost-conscious travelers.

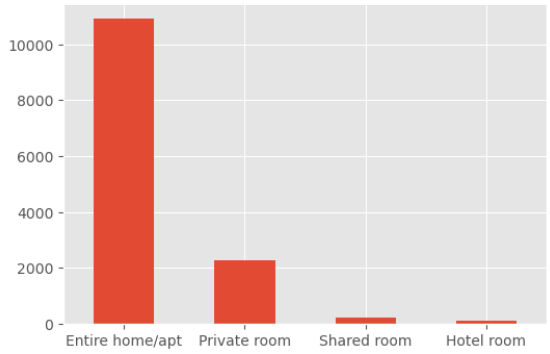


Figure 7: Listing Type Distribution

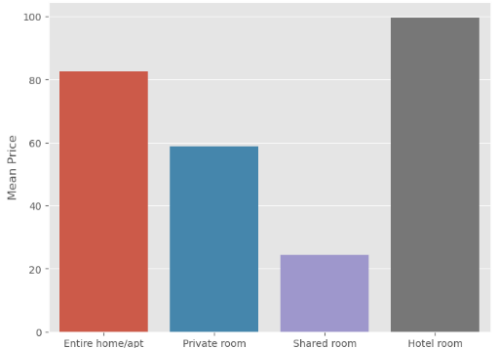


Figure 8: Mean Price per Listing Type

The review data (Figure 9) shows a clear trend towards high satisfaction among guests, with many listings receiving a perfect score of 5 out of 5 across various review categories. Instances of ratings at 4 or below are uncommon. Particularly noteworthy is guests' positive feedback regarding the ease of communication with hosts, the check-in process, and the accuracy of the listings' descriptions. This pattern of high ratings suggests that hosts in this market are attentive to guests' needs and accurate in the portrayal of their offerings, contributing to an overall positive guest experience.

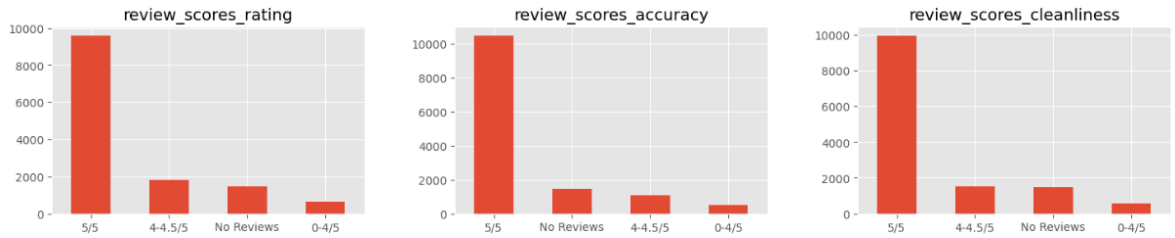


Figure 9: Distribution of Review Scores

The data reveals that a significant proportion of the Airbnb listings have a longstanding presence on the platform, with the majority having received their initial review more than 4 years ago (Figure 10). This indicates that many properties have been available for booking and actively reviewed for a considerable duration. Concurrently, the most frequent timespan for the most recent review falls between 2 to 8 weeks, suggesting that a number of listings have not accrued new reviews in the near past. This could imply either a temporary dip in occupancy or a delay in guests submitting their reviews.

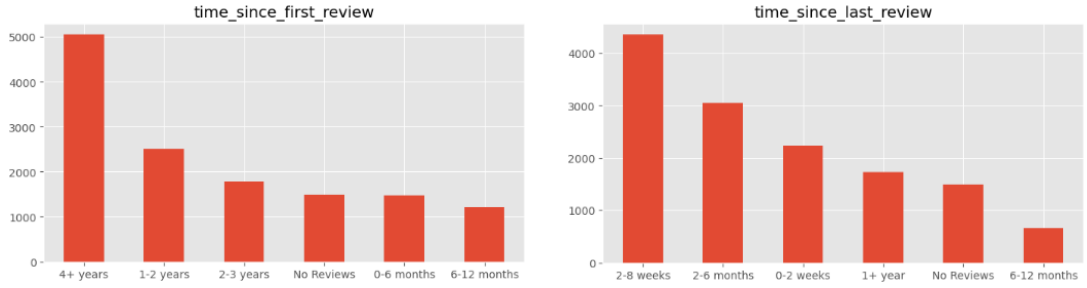


Figure 10: Frequency of Airbnb Listings by Time Elapsed Since First/Last Review

Demand for Airbnb has shown a notable increase since 2012 (Figure 11). The most significant surge in new hosts joining the platform occurred between 2014 and 2015, likely influenced by Airbnb's growing popularity for short-term leases and the platform's adaptation to local legislation and taxation. Although there was an increase in the number of hosts in 2022, it did not surpass the peak seen in 2015. The time series analysis also reveals clear seasonality, with higher numbers of hosts joining during the middle of each year—typically the summer months—and lower numbers at the start and end of each year.

Figure 11: Time Series Decomposition of New Airbnb Hosts Joining Each Month

Before performing the OLS, the correlation matrix (Figure A 10) was analyzed and variables that were highly correlated with each other, mainly some of the review's variables, were dropped for the model. The categorical variables previously identified were one-hot encoded in order to be included in the model. Also, several numerical variables exhibited right-skewed distributions, which can negatively impact the performance of regression models, so the log-transformation was performed just like it was done with the price variable. Figure 12 shows the distribution of these log-transformed variables:

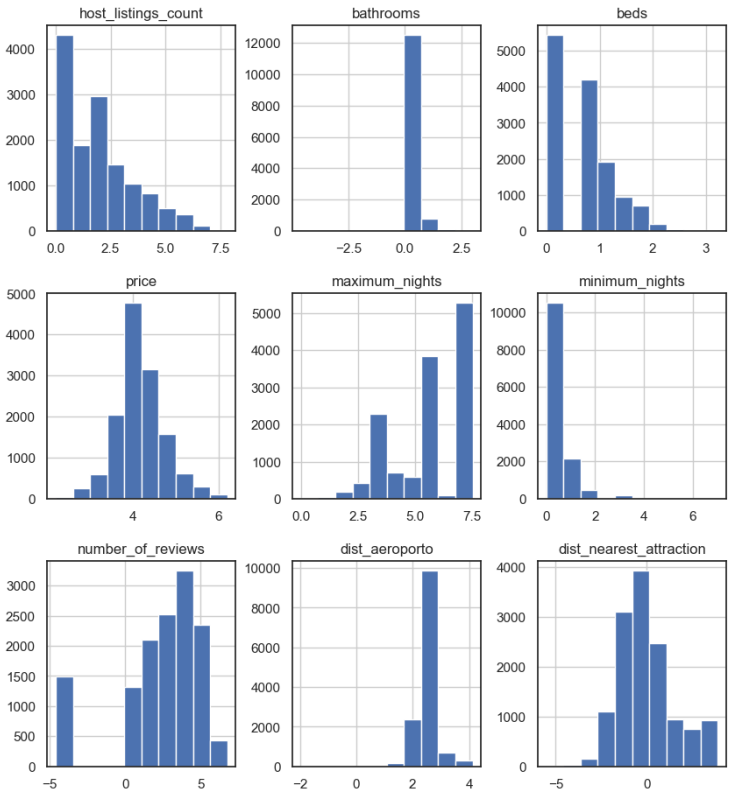


Figure 12: Log-transformed variables' distribution

3.5. MODELS

In transitioning to the modeling chapter, the focus shifts from exploratory data analysis to the application of regression techniques, which are instrumental in interpreting the spatial patterns within the Airbnb dataset. While the GWR is acknowledged for its robustness in handling spatially correlated data, yielding superior results, it is a standard practice to first establish an optimal OLS model.

The rationale for beginning with OLS lies in its utility for pinpointing an effective combination of explanatory variables. By identifying the most predictive variables and refining the model to mitigate multicollinearity—an issue where predictor variables are highly correlated, thus potentially confounding the effects of individual predictors—this step lays a strong foundation. The selected variables can then be transferred to the GWR model, where the spatial dimension of the data is more thoroughly accounted for, ensuring that the local variations and geographical context of the data are captured.

This staged approach helps to verify that the variables hold consistent explanatory power across different modeling techniques and mitigates multicollinearity as much as possible before applying the GWR. With the best possible OLS model identified, we can proceed with greater confidence in applying the GWR to explore the localized patterns and influences on Airbnb pricing in the Porto Metropolitan Area. Both regressions were made at the level of each Airbnb listing.

3.5.1. ORDINARY LEAST SQUARES (OLS)

The specification of the OLS model is an integral preliminary step in regression analysis. The model is formally defined as follows:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + e_i$$

where:

- y_i denotes the i-th observation of the dependent variable;
- x_{ij} represents the j-th explanatory variable for the i-th observation;
- α (the intercept) is the predicted value of y when all explanatory variables are zero;
- β_j are the coefficients of the explanatory variables, reflecting the change in y for a one-unit change in x_j, holding all other predictors constant;
- e_i is the error term for the i-th observation, representing random fluctuations not explained by the model.

3.5.2. GEOGRAPHICALLY WEIGHTED REGRESSION (GWR)

The GWR model represents an evolution of traditional regression analysis, such as the OLS model. It is designed to explore spatial heterogeneity by allowing the relationship between the dependent variable and a set of explanatory variables to vary across different locations. Unlike OLS, which provides a global model of the data, GWR acknowledges that parameters might change over space (Brunsdon et al., 1996).

In GWR, for each location i, a localized regression is fitted using a subset of data weighted by proximity. This is achieved through a spatial weights matrix, which is constructed using an adaptive kernel method, typically applying a near-Gaussian weighting function to prioritize nearby observations.

In the Python ecosystem, libraries like PySal (Python Spatial Analysis Library) can be used to perform GWR analysis. These libraries offer functionalities to calculate the spatial weights matrix and can incorporate various selection criteria, such as the Akaike Information Criterion (AIC), to optimize the bandwidth of the kernel. This optimized bandwidth is crucial for ensuring that each local regression model includes a suitable number of nearby observations, balancing the risk of overfitting against the need for sufficient data to estimate local model parameters.

The general form of the GWR model is:

$$y_i = \alpha(u_i, v_i) + \beta_1(u_i, v_i) x_{i1} + \beta_2(u_i, v_i) x_{i2} + \cdots + \beta_j(u_i, v_i) x_{ij} + e_i$$

where:

- y_i represents the observed value of the dependent variable at location i;
- x_{ij} denotes the value of the j-th explanatory variable at location i;

- $\alpha(u_i, v_i)$ is the local intercept, or the predicted value of y at location i when all explanatory variables are zero;
- $\beta_j(u_i, v_i)$ are the local regression coefficients for the explanatory variables, indicating the influence of each predictor at location i;
- (u_i, v_i) are the geographic coordinates for location i;
- e_i is the local error term at location i.

By fitting a local model at each location, GWR provides insights into the spatial variation of the relationship between the variables, which can be crucial for understanding complex regional dynamics and for informing local policy decisions.

4. RESULTS AND DISCUSSION

4.1. OLS RESULTS

Following a series of model iterations, a more simplified model with 36 variables was found to have an Adjusted R^2 of 0,399, meaning that, after correcting for the number of predictors, the model explains around 39,9% of the variance in the dependent variable. Even though a larger model with 65 variables had an Adjusted R^2 of 0,413, which is a little higher explanatory power, it also added more complexity and the possibility of multicollinearity problems. The chosen model balances between effectively capturing the key price determinants and maintaining a level of simplicity that ensures clarity in interpretation and practical application. The summary results of the OLS model are detailed in Table A 2 of the Appendix, while for simplicity, Table 3 presents the 10 most significant variables alongside the intercept for comprehensive interpretation:

Table 3: OLS Results - 10 most important features

Variables	Coefficient	Std. Error
intercept	4,186	0,004
room_type_Entire home/apt	0,281	0,010
accommodates	0,230	0,007
number_of_reviews	-0,191	0,008
room_type_Private room	0,185	0,010
host_response_time_unknown	0,079	0,017
bathrooms	0,062	0,004
review_scores_rating_No Reviews	-0,059	0,006
host_is_superhost	0,055	0,004
dist_nearest_attraction	-0,054	0,005
minimum_nights	-0,045	0,004
R^2	0,400	
Adjusted R^2	0,399	

Moving forward, to confirm the presence of spatial autocorrelation, the Moran's I test was employed. The computed Moran's I Value (0,2481) exceeded the expected value (-0,000074) under the null hypothesis, indicating a significant positive spatial autocorrelation. In simpler terms, nearby locations exhibit similar values to a greater extent than expected if the spatial distribution were random. Additionally, the low p-value (0,001) further supports these findings, suggesting that the observed spatial pattern is highly improbable to occur by chance alone. Specifically, the likelihood of this spatial pattern resulting from random chance is less

than 0,1%, providing strong evidence for non-random spatial clustering in the analyzed variable.

4.2. GWR RESULTS

After finding the 10 most important features identified in the OLS model previously, the GWR and a new OLS model based on those features were employed. As anticipated, the GWR model demonstrated superior performance relative to the OLS model, with an Adjusted R^2 of 0,608, a lot higher than 0,377. These results are compiled in Table 4.

Table 4: GWR vs OLS Results

	GWR		OLS	
Variables	**Mean Coefficient**	**Std. Error**	**Coefficient**	**Std. Error**
intercept	0,165	3,483	-0,000	0,007
accommodates	0,381	0,152	0,494	0,018
number_of_reviews	-0,334	0,232	0,440	0,008
room_type_Entire home/apt	0,284	3,293	-0,322	0,013
room_type_Private room	0,149	3,127	0,322	0,018
dist_nearest_attraction	-0,220	1,904	0,147	0,007
bathrooms	0,197	0,206	0,106	0,008
review_scores_rating_No Reviews	-0,156	0,194	-0,128	0,013
host_response_time_unknown	0,130	0,124	0,120	0,007
host_is_superhost	0,107	0,132	-0,065	0,007
minimum_nights	-0,038	0,127	-0,066	0,007
R^2	0,667		0,378	
Adjusted R^2	0,608		0,377	

For the variable accommodates, the mean coefficient is 0,381. This means that for each additional person that a listing can accommodate, the price increases by 38,1%, on average. This strong positive relationship suggests that larger listings can command significantly higher prices, likely due to their ability to cater to larger groups or families, which are often willing to pay more for additional space. The variable number_of_reviews, which is log-transformed, has a mean coefficient of -0,334. This indicates that a 1% increase in the number of reviews is associated with a 0,334% decrease in the price, on average. This negative relationship may indicate that listings with a higher number of reviews tend to have lower prices, possibly due to increased competition or a perception of diminished exclusivity. The variable room_type_Entire home/apt has a mean coefficient of 0,284, indicating that listings categorized as Entire home/apt have, on average, about 32,8% ($e^{0,284}-1 \approx 0,328$) higher prices

compared to shared or hotel rooms, holding all other variables constant. This significant positive impact reflects the premium associated with entire home/apartment listings, which is consistent with expectations due to the increased privacy and space offered by these types of accommodations. For the variable room_type_Private room, the mean coefficient is 0,149, suggesting that listings categorized as Private room have, on average, about 16,1% ($e^{0,149}$-1≈0,161) higher prices compared to other types of listings like shared or hotel rooms, holding all other variables constant. This reflects the higher value placed on private rooms, though they are still priced lower than entire homes/apartments. When comparing the effects of room_type_Private room and room_type_Entire home/apt, we find that Entire home/apt listings are about 14,4% more expensive than Private room listings. This comparison is based on the ratio of their exponentiated coefficients $e^{0,284}/e^{0,149}$≈1,144–1=0,144. This means that, on average, Entire home/apt listings command a higher price premium over Private room listings, which aligns with the expectation that entire homes/apartments offer greater space and privacy. The variable dist_nearest_attraction, which is log-transformed, has a mean coefficient of -0,220. This means that a 1% increase in the distance to the nearest attraction is associated with a 0,220% decrease in the log-transformed price, on average. This significant negative impact indicates that proximity to attractions is a major driver of higher prices, as guests are often willing to pay more for convenient access to points of interest. The variable bathrooms, which is log-transformed, has a mean coefficient of 0,197. This indicates that a 1% increase in the number of bathrooms is associated with a 0,197% increase in the log-transformed price, on average. This positive relationship highlights the value of additional bathrooms in enhancing the appeal and price of a listing, as more bathrooms typically offer greater convenience and comfort. Regarding the variable review_scores_rating_No Reviews has a mean coefficient of -0,156, suggesting that listings with no reviews have, on average, about 14,4% ($e^{-0,156}$-1≈-0,144) lower prices compared to those with reviews. This negative impact underscores the importance of guest feedback in establishing trust and perceived value in the market, as potential guests often rely on reviews to assess the quality and reliability of a listing. The mean coefficient for the variable host_response_time_unknown is 0,130, implying that listings with an unknown host response time have, on average, about 13,9% ($e^{0,130}$-1≈0,139) higher prices compared to those with known response times. This could imply that such listings may be perceived as more exclusive or less frequently booked, allowing for higher pricing. For the variable host_is_superhost, the mean coefficient is 0,107, indicating that listings hosted by superhosts have, on average, about 11,3% ($e^{0,107}$-1≈0,113) higher prices compared to those hosted by non-superhosts. This premium likely reflects the higher trust and better service associated with superhosts, which can justify higher prices. Finally, the variable minimum_nights, which is log-transformed, has a mean coefficient of -0,038. This suggests that a 1% increase in the minimum nights required is associated with a 0,038% decrease in the log-transformed price, on average. This indicates that listings with higher minimum stay requirements tend to be priced slightly lower, possibly to attract longer-term bookings.

The local R^2 values in the Porto Metropolitan Area vary significantly across different neighborhoods (Figure 13). The highest local R^2 values, ranging from 0,70 to 0,90, are predominantly observed in the northwestern part of the region, particularly in neighborhoods such as Póvoa de Varzim and Vila do Conde. These high R^2 values indicate that the GWR model effectively explains the variability in Airbnb listing prices in these areas, capturing the underlying spatial relationships well. On the other hand, the model's lowest performance is registered in the regions of Gondomar and Santa Maria da Feira having local R^2 values ranging from 0,20 to 0,50.

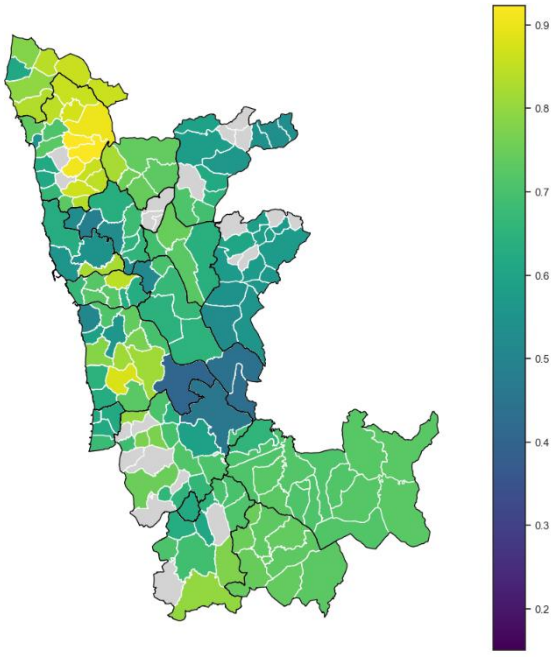
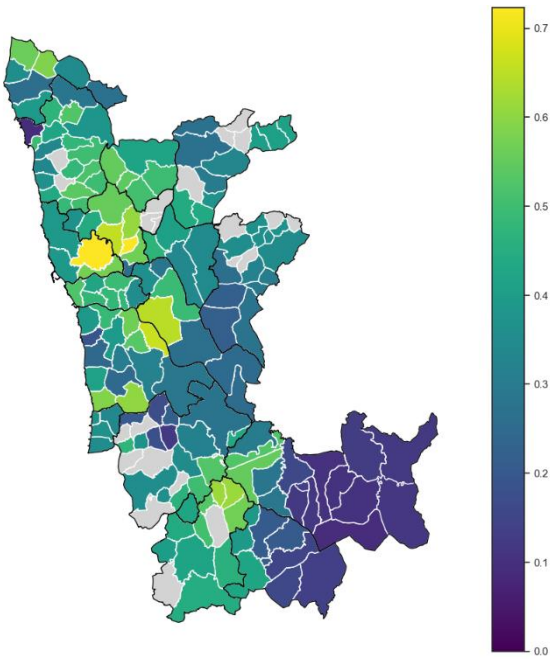
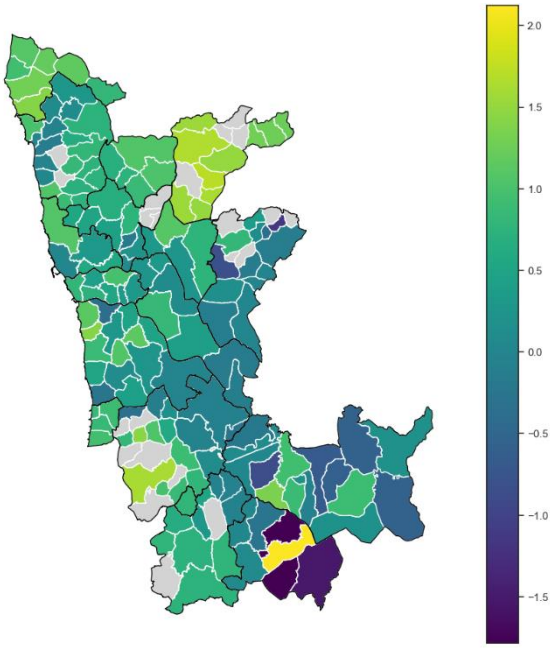


Figure 13: Spatial Variation of Local R^2

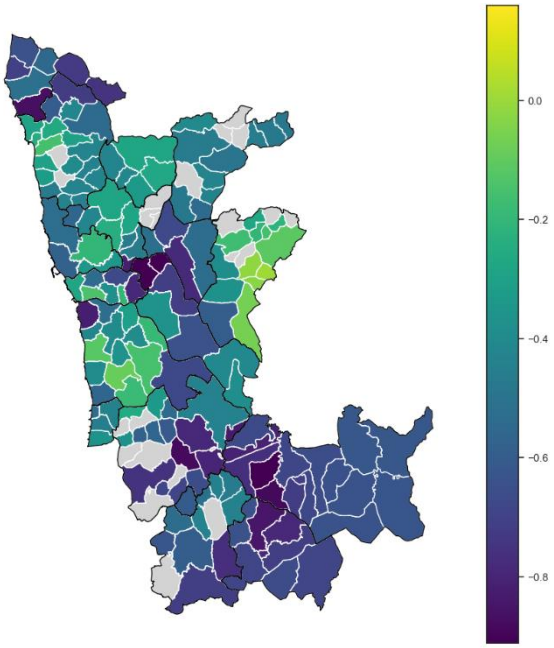
Next, having analyzed the mean coefficients of the GWR model and the distribution of the local R^2 results, we now turn to examining the spatial variability of the coefficients. This spatial analysis is essential for understanding how the impact of each variable varies across different geographic locations within the study area. By doing so, we can gain insights into localized effects that a global model, such as OLS, would fail to capture. To analyze the spatial distribution, the mean per parish of each coefficient was calculated and is illustrated in Figure 14.



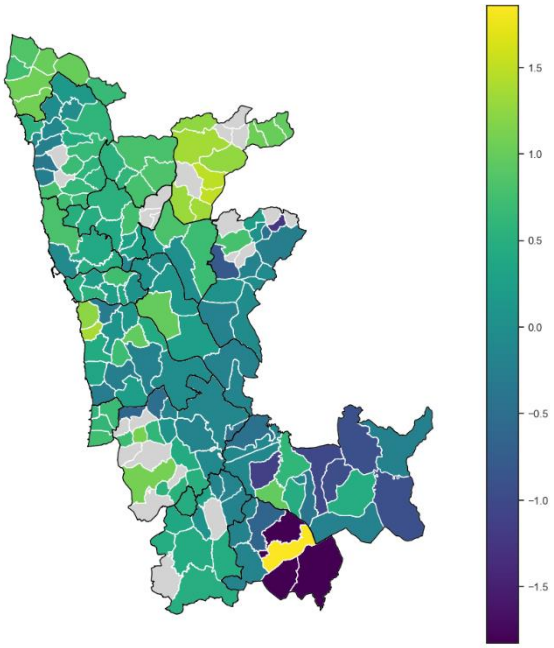
(a) accomodates



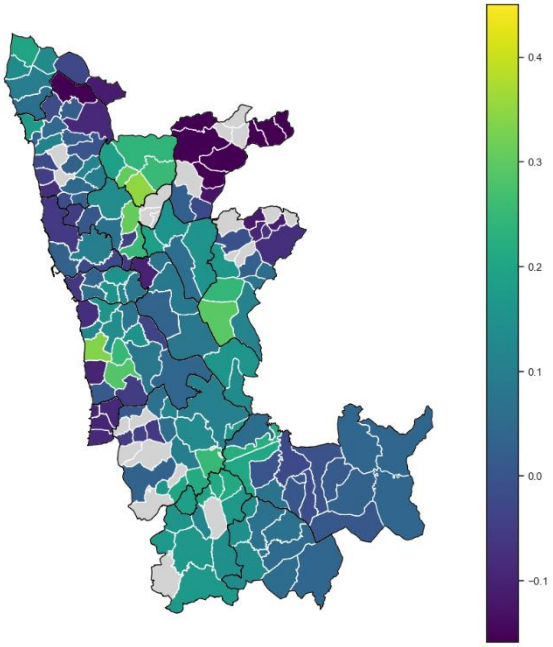
(b) room_type_Entire home/apt



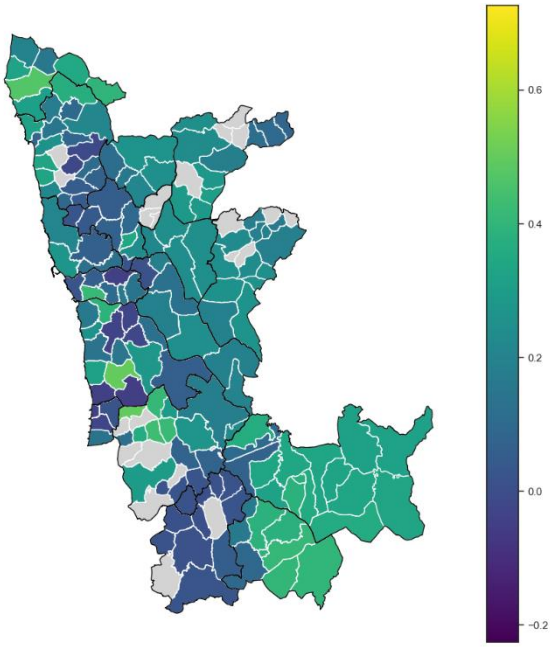
(c) number_of_reviews



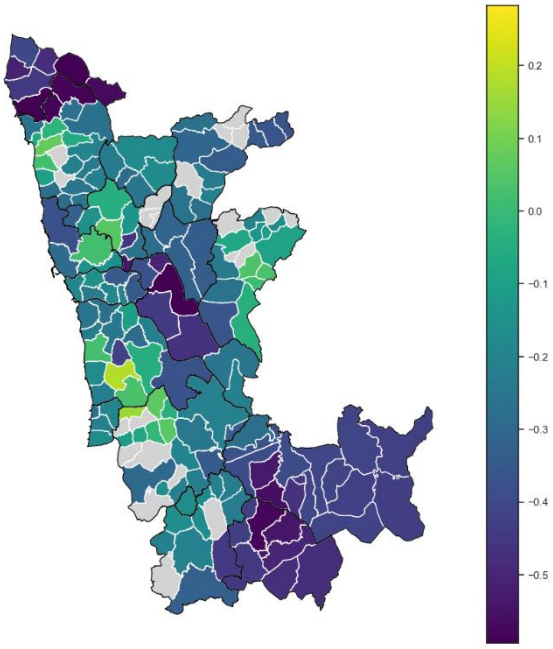
(d) room_type_Private room



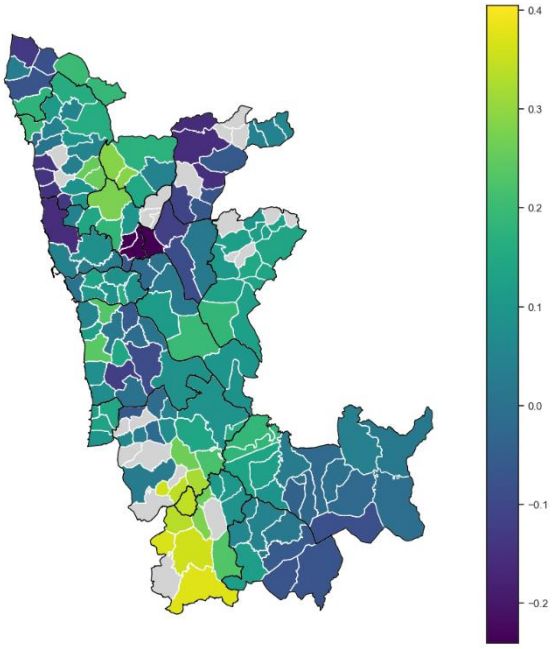
(e) host_response_time_unknown



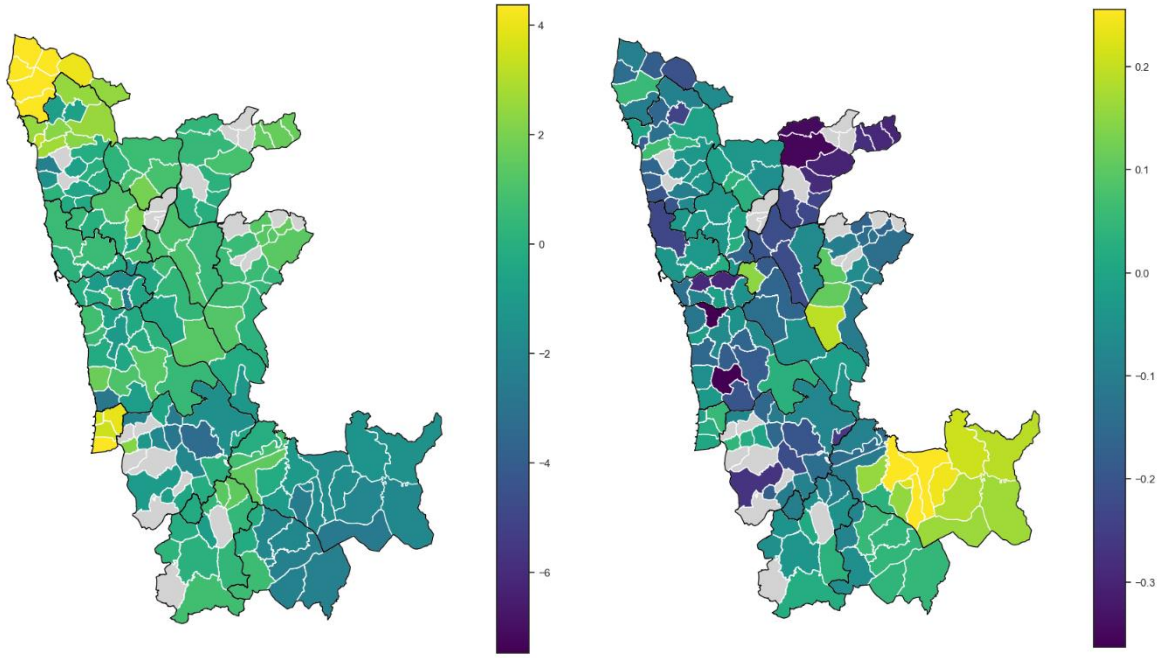
(f) bathrooms



(g) review_scores_rating_No Reviews



(h) host_is_superhost



(i) dist_nearest_attraction (j) minimum_nights

Figure 14: Spatial Distribution of Coefficients

Regarding the number of accommodates (Figure 14(a)), which is the variable that has the biggest overall influence on the Airbnb prices in Porto according to the GWR results, the map shows a generally positive relationship across most of the Porto Metropolitan Area, with coefficients ranging from 0,3 to 0,7. The highest positive impacts are observed in the Matosinhos neighborhood, where coefficients range from 0,6 to 0,7. This suggests that properties in these neighborhoods that can accommodate more guests tend to command higher prices. The attractiveness of these areas for larger groups could be due to their amenities, proximity to the coast, or other tourist attractions that appeal to families or larger groups of travelers. In contrast, areas such as Arouca, Vale de Cambra located further south and far away from the central neighborhood Porto, have lower positive impacts, with most of the coefficients ranging from 0 to 0,3. Other areas, including central Porto and Vila Nova de Gaia, display moderate positive impacts, with coefficients around 0,4 to 0,6. These neighborhoods likely balance demand between smaller and larger groups, leading to a consistent but not extreme increase in prices with the number of guests accommodated. The coefficients for room_type (Figure 14 (b) and (d)) display both positive and negative relationships. Vale de Cambra registers both the highest and lowest coefficients. Other regions such as Santo Tirso and Póvoa de Varzim also display high coefficients, indicating that these types of listings in these neighborhoods are relatively more expensive. This could be due to higher demand for privacy or limited availability, driving up prices. For the number_of_reviews variable (Figure 14(c)), the map shows a negative relationship with Airbnb prices across most of the area, with coefficients ranging from -0,8 to 0. The most substantial negative impacts are seen in neighborhoods located further south such as Arouca, Oliveira de Azeméis, Vale de Cambra, where coefficients range from -0,8 to -0,6. This indicates that a higher number of

reviews are associated with lower prices in these areas, possibly reflecting that more established listings (with more reviews) tend to be priced lower due to competition. Regarding the impact of an unknown host response time (Figure 14(e)), neighborhoods like Trofa, Maia and Vila Nova de Gaia exhibit the highest positive coefficients, reaching up to 0,4. This indicates that in these locales, an unknown host response time may not significantly deter potential guests, possibly because other factors, such as the attractiveness of the area or the quality of the listings, outweigh concerns about host responsiveness. On the other hand, in Santo Tirso and Espinho tends to be negative, with coefficients nearing -0,1. This negative relationship suggests that listings with unclear host response times in these regions may be perceived as less reliable or trustworthy, potentially leading to lower prices. Guests in these areas might place a higher value on the assurance of prompt communication from hosts. In general, the number of bathrooms (Figure 14(f)) has a positive impact on Airbnb prices in several areas within the Porto Metropolitan Area. Neighborhoods such as Póvoa de Varzim exhibit higher positive coefficients from 0,4 to 0,6, suggesting that an increase in the number of bathrooms is associated with higher Airbnb prices in these areas. Conversely, areas like São João da Madeira and Oliveira de Azeméis show lower or even negative coefficients ranging from -0,2 to 0,2, indicating that an increase in the number of bathrooms has a lesser or negative impact on Airbnb prices in these regions. The review_scores_rating_No Reviews (Figure 14(g)) in neighborhood groups such as Vila Nova de Gaia, Matosinhos, display higher positive coefficients, indicating that properties without reviews are associated with higher prices in these areas up to 0,2. On the contrary, neighborhood groups like Póvoa de Varzim, Gondomar and Vale de Cambra have lower coefficients, suggesting a lesser or negative impact of no reviews on Airbnb prices. Regarding the host being superhost (Figure 14(h)), higher positive coefficients are prominently seen in areas such as Oliveira de Azeméis and São João da Madeira. These regions display values closer to 0,4, suggesting a significant positive impact of a superhost on Airbnb prices, meaning that listings managed by superhosts command higher prices in these neighborhoods. On some part of Maia and Valongo this relationship can be negative. In these neighborhoods, the superhost status does not significantly boost prices and may even correspond with lower prices. For the variable distance to the nearest attraction (Figure 14(i)), coefficients range from approximately -6 to 4. Higher positive coefficients are found in Póvoa de Varzim and Espinho, suggesting that being farther from attractions is associated with higher prices, possibly due to a preference for quieter areas. Lower negative coefficients are observed in southern and eastern regions, such as Arouca, Oliveira de Azeméis, and São João da Madeira, where proximity to attractions significantly boosts prices. For the variable minimum nights (Figure 14(j)), coefficients range from approximately -0,3 to 0,2. Higher positive coefficients are seen in Arouca, indicating that longer minimum stays are associated with higher prices. Negative coefficients are observed in Santo Tirso and parts of Vila Nova de Gaia, where increased minimum nights tend to lower prices. This variation highlights the local context's importance in Airbnb pricing strategies.

5. CONCLUSION

This study aimed to explore the determinants of Airbnb prices in the Porto Metropolitan Area, understand the spatial variations in these determinants, and examine the impact of proximity to attractions on pricing. Overall, the analysis revealed that the listing's type emerged as a crucial factor, with entire homes and apartments commanding higher prices compared to other types of listings. Also, the presence of more bathrooms and the capacity to accommodate more guests positively influences pricing, as well as having a superhost status. As expected, if a listing is close to tourist attractions, it's associated with higher prices. On the other hand, listings that have a higher number of reviews typically reduces prices, possibly indicating increased competition among well-reviewed listings. A listing having no reviews is also associated with lower prices, since guests might be less confident in booking the property, prompting hosts to lower prices to attract bookings and compensate for the uncertainty. Additionally, new listings without reviews are often priced more competitively to quickly build up reviews and establish a positive reputation on the platform. Listings with higher minimum stay requirements tend to be priced slightly lower, possibly to attract longer-term bookings.

Regarding spatial variation, the GWR successfully captured these differences, reaching a higher Adjusted R^2 than the OLS. In general, the regions that registered the highest Adjusted R^2 were Póvoa de Varzim and Vila do Conde, where the model explained about 70% to 90% of the price's variability, in contrast to Santa Maria da Feira and Gondomar, where the model only explained between 20% to 50%. This study highlights the complexity of Airbnb pricing in the Porto Metropolitan Area, where various determinants interact in spatially distinct ways. The findings emphasize the importance of considering local context when developing pricing strategies for Airbnb listings. By accounting for the spatial variations and key determinants identified, hosts and policymakers can better navigate the short-term rental market to optimize pricing and improve regulatory frameworks.

6. LIMITATIONS AND FUTURE WORK

A limitation that can be addressed is that the dataset used is specific to a single point in time, which may not capture seasonal variations or long-term trends in Airbnb pricing. Additionally, there may be other important variables influencing Airbnb prices that were not included in the dataset, such as local events, economic conditions, or specific property features (e.g., luxury amenities) that were not captured.

For future work, incorporating a temporal dimension to analyze how Airbnb prices fluctuate over time could capture seasonal trends and long-term changes, by employing a longitudinal study design. Also, employing more sophisticated spatial models, such as the MGWR, spatial lag models or spatial error models, could better address spatial dependency issues and improve the robustness of the findings. Including more detailed variables, such as the presence of specific amenities (e.g., swimming pools, rooftop terraces), proximity to public transportation, and local economic indicators, could provide a more comprehensive understanding of the determinants of Airbnb prices. Investigating the impact of external factors such as major events, regulatory changes, and economic shifts on Airbnb prices could offer valuable insights into how these elements influence the short-term rental market.

By addressing these limitations and pursuing the suggested future research directions, the understanding of Airbnb pricing determinants can be significantly enhanced, leading to more effective and tailored pricing strategies.

BIBLIOGRAPHICAL REFERENCES

Alharbi, Z. H. (2023). A Sustainable Price Prediction Model for Airbnb Listings Using Machine Learning and Sentiment Analysis. *Sustainability*, *15*(17), Artigo 17. https://doi.org/10.3390/su151713159

Bode, O. R., Ferreira, F. A., Rus, V., & Toader, V. (2021). Price determinants of Porto's airbnb listings. *Proceedings of the 4th International Conference on Tourism Research*, 76–83. https://doi.org/10.34190/IRT.21.096

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, *28*(4), 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Chang, C., & Li, S. (2021). Study of Price Determinants of Sharing Economy-Based Accommodation Services: Evidence from Airbnb.com. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(4), Artigo 4. https://doi.org/10.3390/jtaer16040035

Chattopadhyay, M., & Mitra, S. K. (2019). Do airbnb host listing attributes influence room pricing homogenously? *International Journal of Hospitality Management*, *81*, 54–64. https://doi.org/10.1016/j.ijhm.2019.03.008

Estevens, A., Cocola-Gant, A., López-Gay, A., & Pavel, F. (2023). The role of the state in the touristification of Lisbon. *Cities*, *137*, 104275. https://doi.org/10.1016/j.cities.2023.104275

Fernandes, I. S. C. (2019). *Modelling the Airbnb listings' price in Lisbon using local spatial regressions* [masterThesis]. https://run.unl.pt/handle/10362/74240

Franco, S. F., Santos, C. D., & Longo, R. (2019). *The Impact of Airbnb on Residential Property Values and Rents: Evidence from Portugal* (SSRN Scholarly Paper 3387341). https://doi.org/10.2139/ssrn.3387341

Guttentag, D. (2015). Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector. *Current Issues in Tourism*, *18*(12), 1192–1217. https://doi.org/10.1080/13683500.2013.827159

Gyódi, K., & Nawaro, Ł. (2021). Determinants of Airbnb prices in European cities: A spatial econometrics approach. *Tourism Management*, *86*, 104319. https://doi.org/10.1016/j.tourman.2021.104319

Hall, C. M., Prayag, G., Safonov, A., Coles, T., Gössling, S., & Naderi Koupaei, S. (2022). Airbnb and the sharing economy. *Current Issues in Tourism*, *25*(19), 3057–3067. https://doi.org/10.1080/13683500.2022.2122418

Hong, I., & Yoo, C. (2020). Analyzing Spatial Variance of Airbnb Pricing Determinants Using Multiscale GWR Approach. *Sustainability*, *12*(11), 4710. https://doi.org/10.3390/su12114710

Huang, Y., Shen, Y., Lai, F., & Luo, X. (Robert). (2023). The categorical paradoxes in the sharing economy: Empirical evidence from Airbnb. *Production and Operations Management*, *n/a*(n/a). https://doi.org/10.1111/poms.13974

Jiménez, J. L., Ortuño, A., & Pérez-Rodríguez, J. V. (2022). How does AirBnb affect local Spanish tourism markets? *Empirical Economics*, *62*(5), 2515–2545. https://doi.org/10.1007/s00181-021-02107-2

Karney, C. F. F. (2013). Algorithms for geodesics. *Journal of Geodesy*, *87*(1), 43–55. https://doi.org/10.1007/s00190-012-0578-z

Kuhzady, S., Seyfi, S., & Béal, L. (2022). Peer-to-peer (P2P) accommodation in the sharing economy: A review. *Current Issues in Tourism*, *25*(19), 3115–3130. https://doi.org/10.1080/13683500.2020.1786505

Lee, S., & 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*. Scopus. https://doi.org/10.1016/j.tmp.2023.101192

Liu, Y. (2021). Airbnb Pricing Based on Statistical Machine Learning Models. *2021 International Conference on Signal Processing and Machine Learning (CONF-SPML)*, 175–185. https://doi.org/10.1109/CONF-SPML54095.2021.00042

Luo, H. (2023). The Rise of the Sharing Economy. *BCP Business & Management*, *44*, 94–98. https://doi.org/10.54691/bcpbm.v44i.4798

Nunes, M. R. dos S. P. (2023). *Predicting and explaining Airbnb prices in Lisbon: Machine learning approach* [masterThesis]. https://repositorio.ucp.pt/handle/10400.14/41431

Quattrone, G., Kusek, N., & Capra, L. (2022). A global-scale analysis of the sharing economy model – an AirBnB case study. *EPJ Data Science*, *11*(1), Artigo 1. https://doi.org/10.1140/epjds/s13688-022-00349-3

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Thackway, W. T., Ng, M. K. M., Lee, C.-L., Shi, V., & Pettit, C. J. (2022). Spatial Variability of the 'Airbnb Effect': A Spatially Explicit Analysis of Airbnb's Impact on Housing Prices in Sydney. *ISPRS International Journal of Geo-Information*, *11*(1). Scopus. https://doi.org/10.3390/ijgi11010065

Tussyadiah, I. P. (2016). Factors of satisfaction and intention to use peer-to-peer accommodation. *International Journal of Hospitality Management*, *55*, 70–80. https://doi.org/10.1016/j.ijhm.2016.03.005

Wang, D., & Nicolau, J. L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. *International Journal of Hospitality Management*, *62*, 120–131. https://doi.org/10.1016/j.ijhm.2016.12.007

Wang, H. (2023). Predicting Airbnb Listing Price with Different models. *Highlights in Science, Engineering and Technology*, *47*, 79–86. https://doi.org/10.54097/hset.v47i.8169

Xu, F., Hu, M., La, L., Wang, J., & Huang, C. (2020). The influence of neighbourhood environment on Airbnb: A geographically weighed regression analysis. *Tourism Geographies*, *22*(1), 192–209. https://doi.org/10.1080/14616688.2019.1586987

Zervas, G., Proserpio, D., & Byers, J. W. (2017). The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. *Journal of Marketing Research*, *54*(5), 687–705. https://doi.org/10.1509/jmr.15.0204

Zhang, Z., Chen, R. J. C., Han, L. D., & Yang, L. (2017). Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach. *Sustainability*, *9*(9), Artigo 9. https://doi.org/10.3390/su9091635

APPENDIX A

Table A 1: Variables and Description

Variable	Description
host_response_time	The average amount of time the host takes to respond to inquiries and booking requests
host_response_rate	The percentage of inquiries and booking requests the host responds to
host_acceptance_rate	That rate at which a host accepts booking requests
host_is_superhost	Indicates whether the host is recognized as a Superhost, which is a status Airbnb awards to hosts who have met certain criteria of excellent service
host_listings_count	The number of listings the host has on Airbnb, according to Airbnb's calculations
host_identity_verified	Whether or not the host's identity has been verified with an ID
neighbourhood	The neighborhood in which the listing is located
latitude	The latitude coordinate of the listing
longitude	The longitude coordinate of the listing
room_type	The type of listing, such as an entire home/apt, private room, shared room, or hotel
accommodates	The maximum number of guests the listing can accommodate
bathrooms	The number of bathrooms available in the listing
beds	The number of beds available in the listing
price	Price (in €) for a night stay
minimum_nights	The minimum number of nights required for a booking at the listing
maximum_nights	The maximum number of nights a guest can stay at the listing
availability_30	The availability of the listing for the next 30 days as determined by the Airbnb calendar
availability_60	The availability of the listing for the next 60 days as determined by the Airbnb calendar
availability_90	The availability of the listing for the next 90 days as determined by the Airbnb calendar
availability_365	The availability of the listing for the next 365 days as determined by the Airbnb calendar
number_of_reviews	The total number of reviews the listing has received

Variable	Description
review_scores_rating	The overall review score rating of the listing
review_scores_accuracy	The accuracy score from reviews, indicating how accurately the online listing represents the actual space
review_scores_cleanliness	The cleanliness score from reviews, indicating the guests' satisfaction with the listing's cleanliness
review_scores_checkin	The check-in score from reviews, reflecting guests' satisfaction with the check-in process
review_scores_communication	The communication score from reviews, reflecting how well the host communicates with guests
review_scores_location	The score from reviews relating to the listing's location
review_scores_value	The score from reviews relating to the value for money of the listing
instant_bookable	Indicates whether or not the listing can be booked instantly without waiting for the host's approval
time_since_first_review	The amount of time since the listing received its first review
time_since_last_review	The amount of time since the listing received its most recent review
host_days_active	The number of days the host has been active on Airbnb
dist_airport	The distance (km) to the Francisco Sá Carneiro Airport
dist_nearest_attraction	The distance (km) to the nearest tourist attraction

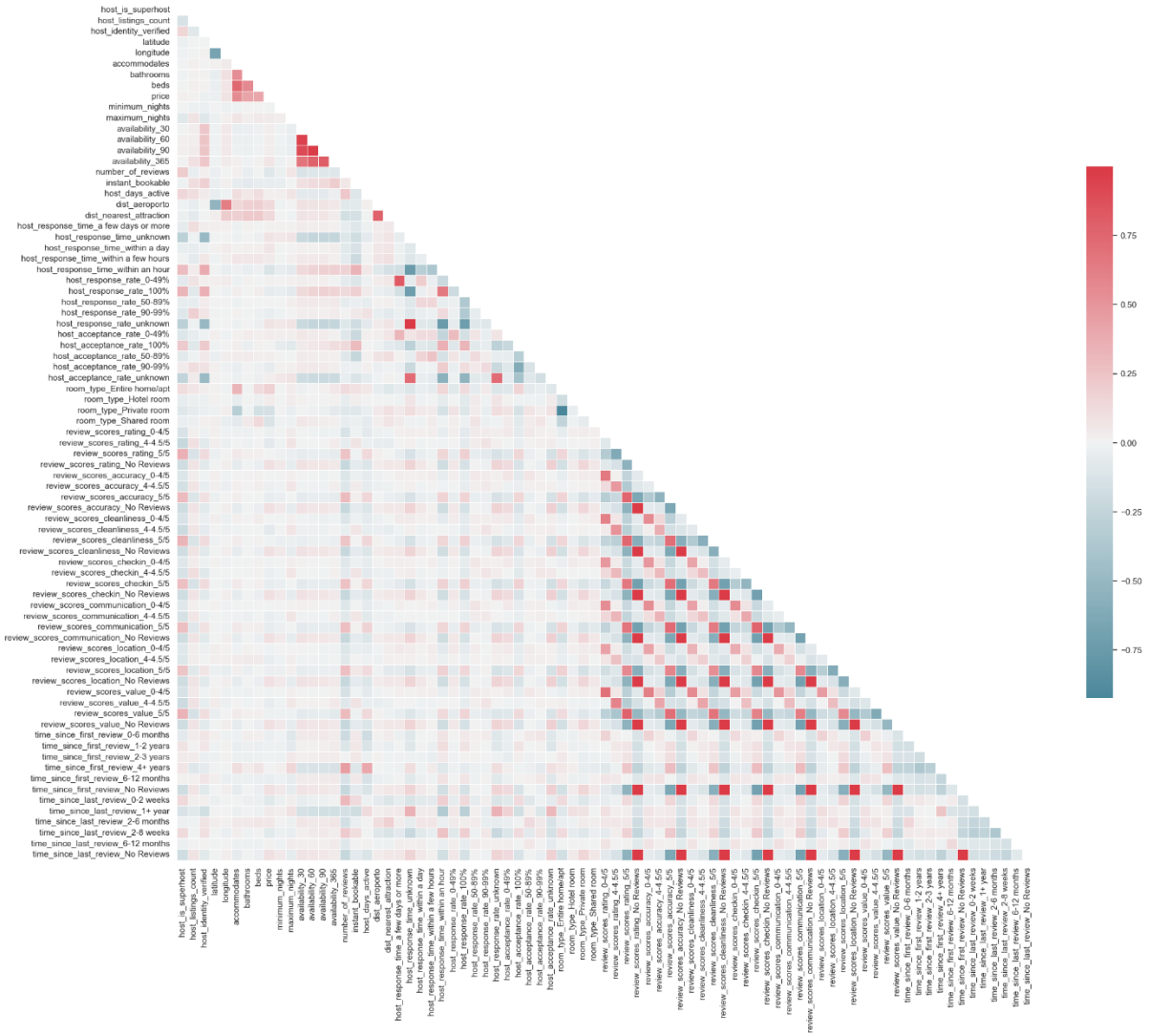


Figure A 1: Correlation Matrix

Table A 2: OLS Results

Variables	Coefficient	Std. Error
intercept	4,186	0,004
room_type_Entire home/apt	0,281	0,010
accommodates	0,230	0,007
number_of_reviews	-0,191	0,008
room_type_Private room	0,185	0,010
host_response_time_unknown	0,079	0,017
bathrooms	0,062	0,004
review_scores_rating_No Reviews	-0,059	0,006
host_is_superhost	0,055	0,004

Variables	Coefficient	Std. Error
dist_nearest_attraction	-0,054	0,005
minimum_nights	-0,045	0,004
host_response_rate_100%	-0,043	0,011
review_scores_rating_5/5	0,041	0,003
host_response_time_within a day	0,040	0,012
host_response_rate_90-99%	-0,039	0,006
availability_365	0,037	0,005
host_listings_count	-0,034	0,005
host_response_time_within a few hours	0,024	0,016
host_days_active	0,023	0,004
dist_airport	0,020	0,005
host_acceptance_rate_unknown	-0,017	0,005
instant_bookable	0,014	0,005
host_acceptance_rate_100%	0,011	0,003
beds	0,011	0,006
host_identity_verified	-0,009	0,005
host_acceptance_rate_0-49%	-0,009	0,004
maximum_nights	0,009	0,004
host_response_time_within an hour	0,009	0,031
review_scores_rating_0-4/5	-0,009	0,003
host_response_rate_50-89%	-0,008	0,006
review_scores_rating_4-4.5/5	0,005	0,003
host_acceptance_rate_90-99%	0,002	0,003
host_acceptance_rate_50-89%	0,002	0,004
longitude	-0,001	0,005
host_response_rate_0-49%	0,001	0,005
latitude	0,001	0,005
availability_30	0,000	0,005

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

Universidade Nova de Lisboa