Determinants of Airbnb Prices in Paris: Analyzing Property Characteristics, Geographic Clustering, and the Impact of the 2024 Olympics

Tristian Kao* Shane Gao* Kian Ghaffari† Nolan Tanaka†

December 9th, 2024

1 Introduction

The rise of the sharing economy has revolutionized how people travel and use accommodations, with Airbnb emerging as one of the most significant players in this space. Understanding the determinants of Airbnb prices has become a critical area of research, providing insights for hosts, potential guests, and policymakers. While previous studies have highlighted factors such as property characteristics, location, and customer reviews as key determinants, several questions remain unanswered such as the interplay of spatial clustering and major global events on Airbnb pricing.

In this study, we aim to address three research questions. First, we investigate the role of property characteristics in determining Airbnb listing prices. Second, we explore the presence and significance of spatial clustering in Airbnb prices, focusing on the extent to which geographical dependencies shape pricing patterns. Finally, we examine the impact of the 2024 Paris Olympics on Airbnb prices, using a Difference-in-Differences (DiD) approach to quantify the influence of this major global event.

Using publicly available data from Inside Airbnb, our analysis reveals several important findings. Random Forest and Gradient Boosting Machine models highlight the non-linear and complex relationships between property characteristics and Airbnb prices, with variables such as `accommodates`, `bedrooms`, and `review_scores_rating` being the most significant predictors. Geographically Weighted Regression (GWR) indicates that the effect of density on log prices varies spatially, with notable negative impacts in central Paris and small positive clusters in the north. The DiD analysis shows that while treated and control groups had similar trends in December 2023 and March 2024, the treated group exhibited a substantial increase in prices by June 2024, aligning with the heightened demand brought about by the Olympics.

*Department of Economics, University of California, Davis
†Department of Statistics, University of California, Davis

By combining predictive modeling, spatial analysis, and causal inference, this study provides a comprehensive understanding of Airbnb pricing dynamics. Our findings not only reinforce the importance of property characteristics but also highlight the role of spatial dependencies and temporal factors, offering valuable insights for stakeholders navigating the rapidly evolving short-term rental market.

2 Literature Review

The determinants of Airbnb pricing have been extensively studied within the context of the sharing economy, with researchers employing statistical and machine learning models to identify key predictors. Herrgård and Flöjs [2] conducted a multilinear regression analysis of Airbnb listings in Spain, identifying characteristics such as the type of lodging, number of accommodations, and number of bathrooms as significant factors influencing prices. Their findings revealed that property size and exclusivity accounted for approximately two-thirds of the price variation, underscoring their critical role in determining pricing.

Building on this, Choi [1] applied machine learning techniques, including linear regression and root mean square error evaluations, to create a reliable pricing model for Airbnb properties. This research highlighted the significance of room features, the number of beds, and other property characteristics as key predictors, demonstrating the effectiveness of advanced modeling techniques in optimizing pricing strategies for both hosts and guests.

Similarly, Prajapati and Suthur [5] explored various machine learning algorithms, such as tree-based models and support vector machines (SVM), for predicting Airbnb prices. Their study incorporated property characteristics alongside customer review data to enhance the accuracy of pricing models. The inclusion of customer feedback provided valuable insights into user preferences and refined the understanding of price determinants.

Despite these contributions, several gaps persist in the literature. Many existing studies focus on general pricing trends across multiple cities, which limits insights into the pricing dynamics of specific tourist destinations like Paris. While property characteristics and host status (e.g., superhost designation) have been thoroughly examined, spatial clustering of prices—such as the geographic grouping of high-priced listings—remains underexplored. Additionally, the effect of major global events, like the Olympics, on Airbnb pricing has not been rigorously analyzed. Although general demand fluctuations are acknowledged, there is limited research on how short-term international events directly influence local pricing trends.

To address these gaps, our study focuses on Paris, one of the most visited tourist cities globally and a prime destination for Airbnb listings. We examine not only the impact of property characteristics on pricing but also investigate geographic clustering using spatial statistical methods. Furthermore, we assess the influence of the 2024 Paris Olympics on Airbnb prices, analyzing the economic effects of this major global event. By employing multiple statistical models, our research provides a comprehensive understanding of both temporal and spatial factors affecting Airbnb prices in Paris.

3 Sample and Data

Our analysis utilizes publicly available data from Inside Airbnb. The website is a mission-driven project that provides data and advocacy on Airbnb's impact on residential communities. Inside Airbnb empowers communities with information to better understand and manage the role of short-term rentals in their neighborhoods. The platform periodically scrapes Airbnb data from various cities. However, the data collection is limited to a few specific days in the beginning of every quarter each year. Consequently, the dataset is not continuous, as it only reflects snapshots of Airbnb listings at irregular intervals [3].

For our study, we used Paris Inside Airbnb data from December 2023, March 2024, and June 2024. However, due to computational constraints and the need to minimize potential time effects, we focused primarily on the June 2024 dataset for most of our analysis. This approach allows us to reduce noise and ensure consistency in our results when examining the specific variables affecting Airbnb prices and the geographical clustering of prices.

For our analysis of the 2024 Paris Olympics' impact on listing prices, we incorporated data from all three time points: December 2023, March 2024, and June 2024. This broader timeframe enabled us to perform a Difference-in-Differences (DiD) analysis to assess the effect of the Olympics on Airbnb prices while accounting for temporal variations.

The original dataset included numerous duplicate and irrelevant variables that were removed during the data cleaning process to focus on the core variables essential for our analysis. The final dataset consists of eight numeric variables: `price`, `accommodates`, `bedrooms`, `bathrooms`, `beds`, `number_of_reviews`, `review_scores_rating`, and `host_response_rate`; and two categorical variables: the neighborhood of the listing (`neighbourhood_cleansed`) and whether the host is designated as a superhost (`host_is_superhost`). To address skewness in the price distribution and enhance the robustness of our statistical analyses, we applied a logarithmic transformation to the listing prices, creating the variable `log_price`, which serves as the dependent variable in our models. This transformation improves the interpretability of our results and ensures consistency across the various modeling approaches employed in the study.

The data for this study was sourced from Inside Airbnb's publicly available datasets, accessible at `https://insideairbnb.com/get-the-data/` [3].

4 Methods

In this section, we present the approaches used to address the research questions posed in this study. We employed a combination of predictive modeling, spatial analysis, and causal inference to explore the determinants of Airbnb prices, assess geographic clustering, and evaluate the impact of the 2024 Paris Olympics on prices. Each subsection corresponds to one of the three research questions, detailing the models, assumptions, and strategies applied.

4.1 Predictive Modeling

The analysis to identify key determinants of Airbnb prices utilized the following model specification:

$$\begin{aligned} \text{log\_price} = \beta_0 &+ \beta_1 \text{accommodates} + \beta_2 \text{bedrooms} \\ &+ \beta_3 \text{bathrooms} + \beta_4 \text{beds} \\ &+ \beta_5 \text{neighbourhood\_cleansed} + \beta_6 \text{number\_of\_reviews} \\ &+ \beta_7 \text{review\_scores\_rating} + \beta_8 \text{host\_is\_superhost} \\ &+ \beta_9 \text{host\_response\_rate} + \epsilon \end{aligned} \tag{1}$$

Where:

- log_price is the log-transformed Airbnb listing price.
- Independent variables include property characteristics (`accommodates`, `bedrooms`, `bathrooms`, `beds`), host characteristics (`host_is_superhost`, `host_response_rate`), and neighborhood effects (`neighbourhood_cleansed`).
- ϵ represents the error term.

This model specification was applied consistently across four predictive models: Ordinary Least Squares (OLS), Random Forest (RF), Gradient Boosting Machine (GBM), and k-Nearest Neighbors (KNN). Employing multiple models provides several advantages, enhancing the robustness and credibility of our findings. Each model offers unique strengths and captures different aspects of the relationships between variables, allowing for a more comprehensive understanding of the determinants of Airbnb prices.

OLS serves as a baseline model and assumes a linear relationship between the predictors and the outcome variable. Its simplicity and interpretability make it a great starting point for comparison. However, its reliance on linear assumptions limits its ability to capture the complexities of Airbnb price determinants.

In contrast, tree-based methods such as Random Forest and Gradient Boosting Machine excel at capturing non-linear interactions and variable importance. Random Forest, with its averaging mechanism across multiple decision trees, provides robust predictions and is less sensitive to overfitting. Gradient Boosting Machine builds trees sequentially, correcting errors iteratively, and often achieves high predictive accuracy for complex datasets. These models complement each other by offering insights into both global patterns and nuanced interactions within the data.

k-Nearest Neighbors adds a layer of diversity to the analysis by focusing on local relationships in the feature space. Although its performance is highly dependent on the choice of hyperparameters and is often limited in high-dimensional datasets, its inclusion allows for a valuable benchmark comparison to the more complex models.

By implementing these diverse models, we mitigate the risk of relying on a single methodological approach, which may oversimplify or misrepresent the data. Consistent results across

multiple methods increase confidence in our conclusions, while discrepancies provide an opportunity for deeper investigation. This multi-model strategy ensures that our findings are both robust and comprehensive, thus balancing interpretability, predictive power, and the ability to capture complex relationships within the data.

4.2 Spatial Analysis

To assess geographic clustering and spatial dependencies in Airbnb prices, we employed two spatial models. Each model is tailored to capture different aspects of spatial relationships in the data.

Geographically Weighted Regression (GWR): The GWR model allows for the examination of localized relationships between predictors and the dependent variable (`log_price`) across different neighborhoods in Paris. Unlike global models, GWR estimates coefficients that vary spatially. Thus, the model provides insights into how the effects of predictors such as `density`, `accommodates`, `bedrooms`, and `bathrooms` change over geographic space. The model specification for GWR is as follows:

$$\text{log\_price}_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)\text{density}_i + \beta_2(u_i, v_i)\text{accommodates}_i + \cdots + \epsilon_i \tag{2}$$

Where:

- (u_i, v_i) are the geographic coordinates of observation i, determining the local parameter estimates.
- density_i, accommodates_i, and other predictors are the characteristics of listing i.
- ϵ_i is the local error term.

The GWR model uses an adaptive bandwidth to account for variations in data density, ensuring that neighborhoods with sparse listings still provide meaningful estimates. By capturing the spatial variation of coefficients, GWR reveals localized effects, such as how listing density impacts prices in central versus peripheral neighborhoods of Paris.

Spatial Lag Model (SLM): To account for spatial dependencies and autocorrelation in Airbnb prices, we applied a Spatial Lag Model. The SLM incorporates a spatially lagged dependent variable (`log_price`) to quantify the influence of nearby prices on a given listing. The model is specified as:

$$\text{log\_price} = \rho W\text{log\_price} + \beta_0 + \beta_1\text{accommodates} + \beta_2\text{bedrooms} + \cdots + \epsilon \tag{3}$$

Where:

- ρ is the spatial lag coefficient, indicating the strength of spatial autocorrelation.
- W is the spatial weights matrix, derived from k-nearest neighbors ($k = 5$) in this study.

- The remaining terms are consistent with the predictors in the equation (1).

The SLM explicitly models the spatial influence of surrounding listings on prices, allowing us to quantify spillover effects in the Airbnb market.

Residual Spatial Autocorrelation: Residual spatial autocorrelation was evaluated using Moran's I test to ensure robustness. Moran's I is a measure of the overall clustering of spatial data and is defined mathematically as:

$$I = \frac{N}{W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{4}$$

Where:

- N is the number of spatial units indexed by i and j.
- x is the variable of interest.
- $\bar{x}$ is the mean of x.
- w_{ij} are the elements of a matrix of spatial weights, with zeroes on the diagonal (i.e., $w_{ii} = 0$).
- W is the sum of all w_{ij}, defined as $W = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$.

This definition follows the standard formulation of Moran's I, as described by [4]. Moran's I is widely used in spatial analysis for detecting spatial autocorrelation and assessing clustering in geographical data. By incorporating Moran's I, the study accounts for residual spatial autocorrelation, which strengthens the validity of the findings.

This test ensures that spatial clustering effects and dependencies are appropriately captured and interpreted. By incorporating both GWR and SLM, the analysis provides a comprehensive understanding of spatial patterns and dependencies in Airbnb prices across Paris neighborhoods.

4.3 Causal Inference

To estimate the causal effect of the 2024 Paris Olympics on Airbnb prices, we applied a Difference-in-Differences (DiD) approach. The identification strategy relied on comparing treated listings in Paris to control listings in similar markets over three time periods: December 2023, March 2024, and June 2024.

The DiD model included interaction terms between treatment status (`treated`) and time periods (`time_fe`). Control variables, such as `accommodates` and `beds`, were included to account for potential confounders. The model specification is as follows:

$$\text{log\_price}_{it} = \beta_0 + \beta_1 \text{treated}_i + \beta_2 \text{time\_fe}_t + \beta_3 (\text{treated}_i \cdot \text{time\_fe}_t) + \mathbf{X}_{it}\gamma + \epsilon_{it} \tag{5}$$

Where:

- log_price_{it} is the log-transformed price for listing i at time t.
- treated_i indicates whether listing i is in Paris (treated group).
- time_fe_t represents time period fixed effects.
- $(\text{treated}_i \cdot \text{time\_fe}_t)$ captures the interaction effect of treatment and time, estimating the causal impact of the Olympics.
- $\mathbf{X}_{it}$ is a vector of listing characteristics, including `accommodates` and `beds`.
- ϵ_{it} is the error term.

The validity of the DiD approach relies on the parallel trends assumption, which posits that treated and control groups would have followed similar trends in the absence of treatment. However, due to the limited data availability (only three non-consecutive time points), the pre-treatment period is short, and parallel trends could not be fully verified. This limitation, combined with potential noise in the data, may affect the precision of our estimates.

The results of the DiD analysis were visualized to compare trends in predicted log prices between the treated and control groups over time. These findings provide insights into how the 2024 Paris Olympics influenced Airbnb pricing in Paris.

5 Results

In this section, we present the findings from our analysis, addressing the three primary research questions: (1) the impact of property characteristics on Airbnb prices, (2) the presence and significance of geographical clustering in Airbnb prices, and (3) the effect of the 2024 Paris Olympics on listing prices. Using a combination of statistical models and spatial analysis methods, we evaluate the relationships between the dependent variable, `log_price`, and the independent variables, including property characteristics and location. The results provide insights into the key determinants of Airbnb pricing, the extent of spatial patterns, and the influence of major global events on the short-term rental market.

5.1 Impact of Property Characteristics on Airbnb Prices

To evaluate the impact of property characteristics on Airbnb prices, we compared the prediction accuracy of four different models: Ordinary Least Squares (OLS), Random Forest (RF), Gradient Boosting Machine (GBM), and k-Nearest Neighbors (KNN). Table 1 summarizes the R^2 and Root Mean Squared Error (RMSE) values for each model.

Table 1: Prediction Accuracy of Different Models

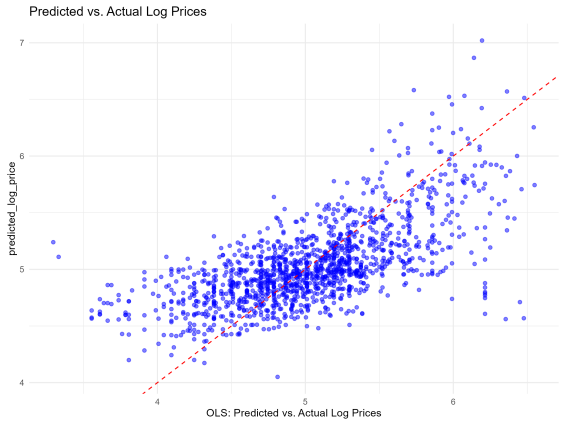
Model	$\mathbf{R^2}$	**RMSE**
OLS	0.4368	0.4100
Random Forest	0.8235	0.2295
GBM	0.6328	0.3111
KNN	0.4747	0.3960

The results demonstrate that the Random Forest model achieved the highest prediction accuracy ($R^2 = 0.8235$, RMSE = 0.2295), followed by the GBM model ($R^2 = 0.6328$, RMSE = 0.3111). These outcomes highlight the effectiveness of tree-based ensemble methods in capturing complex, non-linear relationships within the dataset. The OLS model, while offering greater interpretability, underperformed with $R^2 = 0.4368$ and RMSE = 0.4100, reflecting its inherent limitations in modeling interactions and non-linear effects. The KNN model exhibited the poorest performance, with $R^2 = 0.4747$ and RMSE = 0.3960, emphasizing its inadequacy for this particular dataset.

The underperformance of KNN can be attributed to several inherent challenges. First, KNN is particularly sensitive to high-dimensional datasets. The presence of numerous continuous and categorical predictors creates a feature space where the concept of "closeness" between data points becomes less meaningful, a phenomenon known as the curse of dimensionality. This reduces the model's ability to effectively identify relevant neighbors for prediction. Second, the dataset's mix of categorical and continuous variables, such as `neighbourhood_cleansed` and `accommodates`, complicates the distance calculations that KNN relies on. While distance metrics can be adapted for mixed data types, their application often struggles to balance the influence of these heterogeneous predictors. Third, KNN does not explicitly model non-linear relationships or interactions between variables. Instead, it assumes that similar feature vectors correspond to similar target values, which is insufficient for datasets like this one, where nuanced interactions among property characteristics and host attributes influence Airbnb prices. Lastly, the choice of the hyperparameter k, which specifies the number of neighbors considered, significantly impacts KNN's performance. A smaller k risks overfitting to noise, while a larger k oversmooths predictions. The chosen k in this analysis may not have been optimal for the dataset.

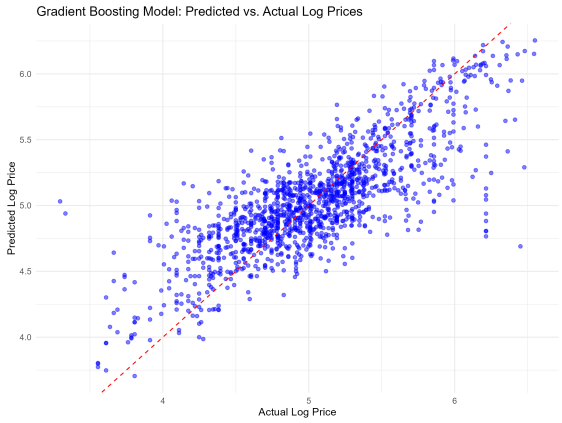
In contrast, Random Forest and GBM demonstrated superior performance due to their ability to handle the dataset's complexity. Random Forest's averaging mechanism across multiple trees mitigated overfitting and provided robustness in identifying influential predictors. GBM's sequential tree-building process, where each tree corrects the errors of its predecessor, allowed for high predictive accuracy by capturing intricate patterns in the data. However, GBM performed slightly worse than Random Forest, likely due to its sensitivity to hyperparameter tuning. Unlike Random Forest, which is robust to default parameter settings, GBM requires careful adjustment of parameters such as learning rate, tree depth, and number of iterations to achieve optimal performance. Additionally, GBM's sequential nature makes it more susceptible to noise in the data, as errors can propagate across iterations. Despite these limitations, both models outperformed OLS and KNN, reflecting their

ability to effectively model non-linear relationships and interactions in the dataset.

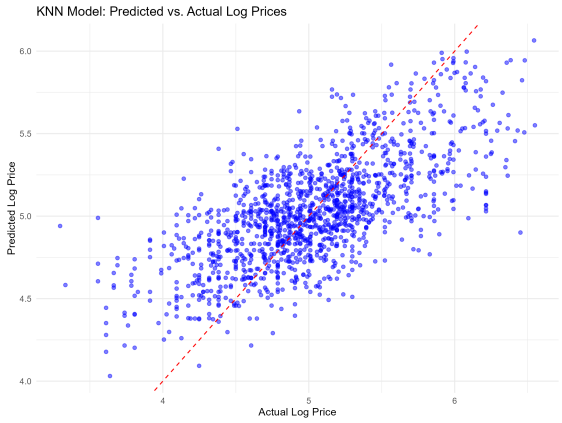


(a) OLS

(b) Random Forest



(c) GBM



(d) KNN

Figure 1: Predicted vs. Actual Prices for Different Models: (a) OLS, (b) Random Forest, (c) GBM, and (d) KNN.

Using the two most accurate models, Random Forest and GBM, we explored the relative importance of property characteristics. The importance scores for each variable are summarized in Tables 2 and 3.

Table 2: GBM Variable Importance

Feature	**Gain**	**Cover**	**Frequency**
Accommodates	0.3338	0.1146	0.1420
Bedrooms	0.2608	0.0696	0.0976
Review Scores Rating	0.1095	0.2721	0.2431
Bathrooms	0.0937	0.0922	0.0524
Number of Reviews	0.0913	0.3044	0.2600
Neighbourhood Cleansed	0.0908	0.1098	0.1526
Beds	0.0201	0.0373	0.0524

Table 3: Random Forest Variable Importance

Feature	**%IncMSE**	**IncNodePurity**
Accommodates	51.6741	77.5363
Review Scores Rating	46.4236	52.9470
Bedrooms	43.3312	64.9167
Beds	31.5636	31.0357
Bathrooms	29.8276	30.0611
Number of Reviews	25.5024	47.9313
Neighbourhood Cleansed	19.9554	28.2817
Host Response Rate	19.0009	19.3465
Host is Superhost	14.6632	9.9385

Figures 2 and 3 provide visual representations of variable importance for the Random Forest and GBM models, respectively.

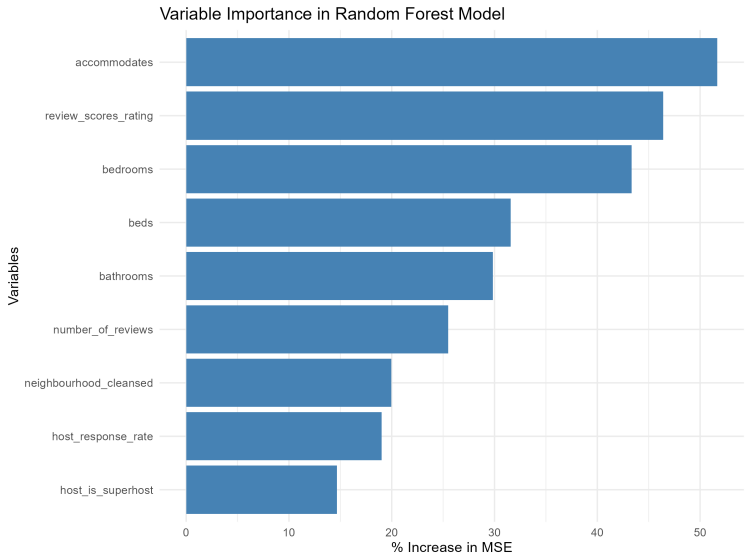


Figure 2: Random Forest Variable Importance

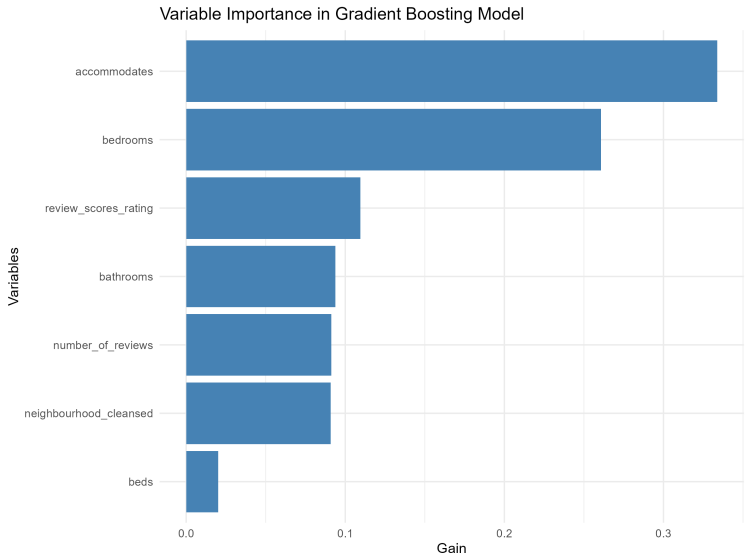


Figure 3: GBM Variable Importance

The analysis confirms that key predictors of Airbnb prices include `accommodates`, `bedrooms`, and `review_scores_rating`, with `accommodates` being the most influential variable across both models. These results highlight the critical role of property characteristics in determining pricing.

5.2 Geographical Clustering of Airbnb Prices

To examine the geographical clustering of Airbnb prices in Paris, we conducted spatial analyses to investigate the spatial distribution of log-transformed prices (`log_price`) and

the impact of density and spatial dependencies on pricing.

5.2.1 Geographical Distribution of Log Price

The geographical distribution of average log prices across the Paris arrondissements was visualized by grouping listings based on their neighborhoods and mapping their mean log prices. Figure 4 illustrates this distribution, with a clear gradient of higher log prices concentrated in the central and western arrondissement, reflecting proximity to tourist attractions and affluent areas.

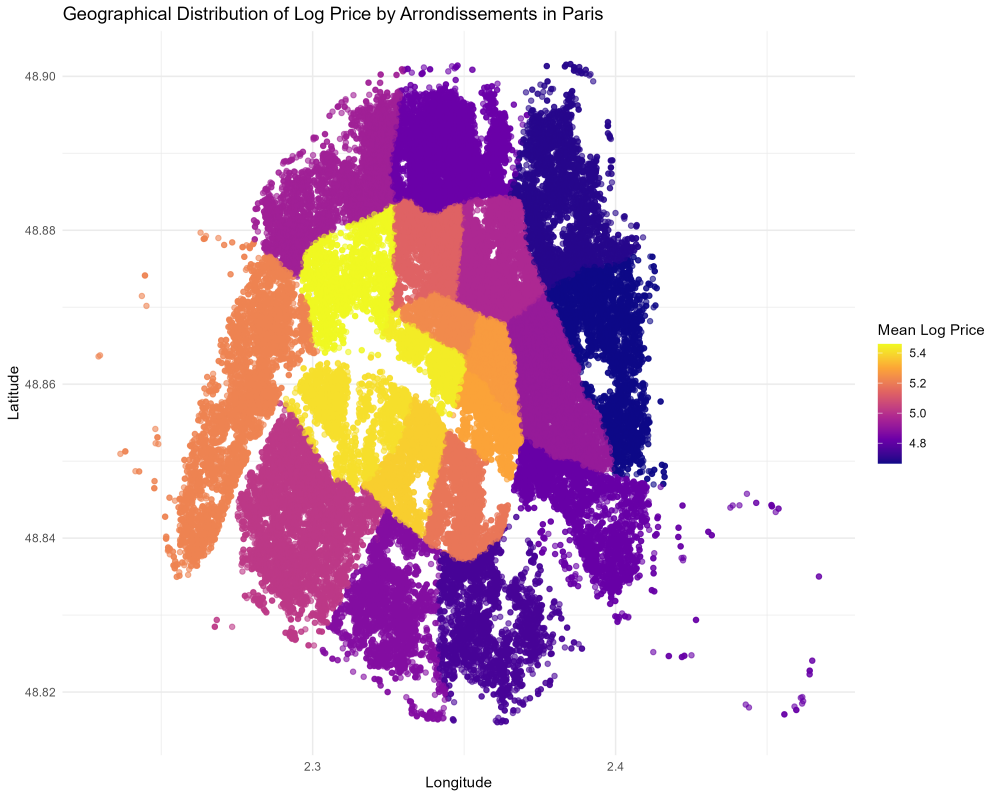


Figure 4: Geographical Distribution of Log Price by Arrondissements in Paris.

5.2.2 Local Impact of Density on Log Price

To further explore spatial clustering, we analyzed the local impact of listing density on log prices using Geographically Weighted Regression (GWR). Density was calculated as the number of listings within a 400-meter buffer around each listing. The GWR model revealed spatially varying effects of density on log price, which are visualized in Figure 5.

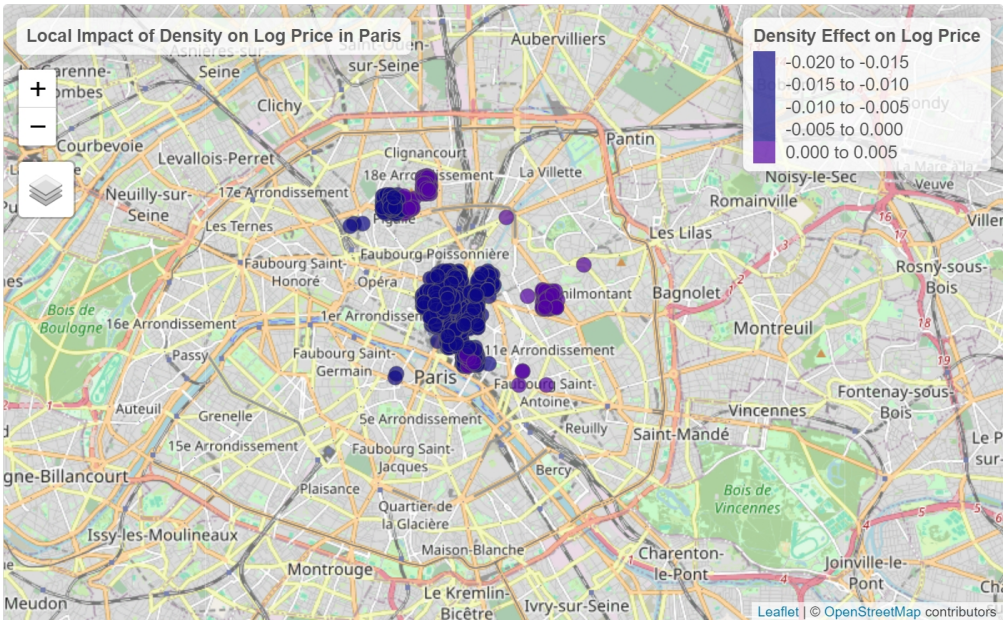


Figure 5: Local Impact of Density on Log Price in Paris.

The results show that density has a heterogeneous effect on log prices, ranging between -0.020 and 0.005. In central Paris, the density effect is predominantly negative, suggesting that higher listing density slightly suppresses prices in these areas, possibly due to increased competition among listings. Conversely, in northern Paris, two clusters exhibit positive density effects, although the magnitudes are very small. These findings indicate that while local listing density does influence pricing patterns, the overall impact is minimal, highlighting a limited role of density in shaping Airbnb prices across different neighborhoods.

5.2.3 Spatial Lag Model (SLM) Analysis

To account for spatial dependence, we implemented a Spatial Lag Model (SLM), incorporating spatial weights derived from a k-nearest neighbor (k = 5) approach. Table 4 presents the model coefficients.

Table 4: Spatial Lag Model Coefficients

Variable	**Estimate**	**Std. Error**	**p-value**
Intercept	1.5340	0.2048	< 0.001
Accommodates	0.1207	0.0120	< 0.001
Bedrooms	0.1514	0.0205	< 0.001
Bathrooms	0.1505	0.0342	< 0.001
Beds	-0.0377	0.0184	0.040
Neighbourhood Cleansed	-0.0011	0.0025	0.657
Number of Reviews	0.0001	0.0001	0.618
Review Scores Rating	0.2507	0.0284	< 0.001

The spatial lag coefficient (`Rho`) was estimated at 0.2818 and was statistically significant ($p < 0.001$), indicating that log prices are spatially autocorrelated. Additionally, the AIC value for the SLM (1512.9) was lower compared to the standard linear model (1623.4), confirming the improved fit provided by incorporating spatial dependencies.

5.2.4 Spatial Autocorrelation Test

Finally, Moran's I test was conducted to assess residual spatial autocorrelation. The test yielded a significant Moran's I statistic of 0.0462 ($p < 0.001$), indicating a spatial dependence in the data.

5.3 Effect of the 2024 Paris Olympics on Listing Prices

To evaluate the impact of the 2024 Paris Olympics on Airbnb listing prices, we conducted a Difference-in-Differences (DiD) analysis. This method compares changes in log-transformed Airbnb prices (`log_price`) between treated listings (in Paris) and control listings (from similar markets) across three time periods: December 2023, March 2024, and June 2024.

5.3.1 Boxplot Analysis

Figure 6 provides a visual comparison of `log_price` distributions for treated and control groups across the three time periods. In December 2023 and March 2024, the treated and control groups exhibit differences in log prices, but the magnitude of these differences remains relatively stable. However, in June 2024, the difference between the groups increases substantially, with the treated group showing a significantly higher log price compared to the control group. This sudden divergence highlights a potential impact on prices for the treated group (Paris) as the Olympics approach.

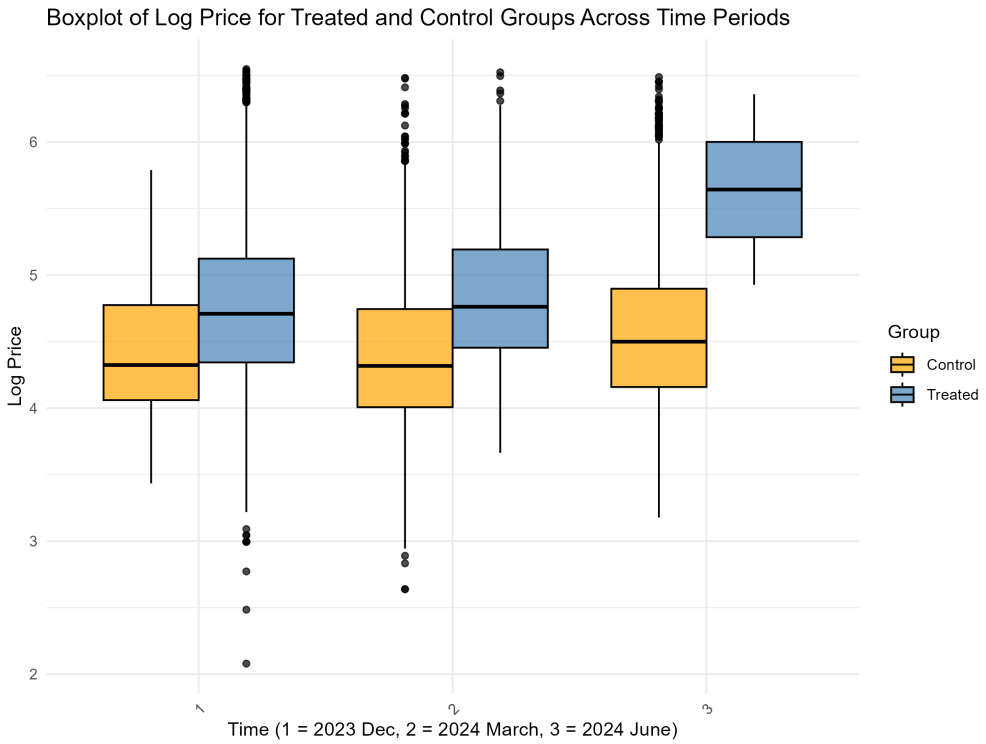


Figure 6: Boxplot of Log Price for Treated and Control Groups Across Time Periods.

5.3.2 Difference-in-Differences Model Results

The DiD model was specified to include interaction terms between treatment status (`treated`) and time periods (`time_fe`), along with control variables for `accommodates` and `beds`. Table 5 summarizes the key coefficients from the regression model.

Table 5: Difference-in-Differences Model Coefficients

Variable	**Estimate**	**Std. Error**	**p-value**
Intercept	4.1601	0.0635	< 0.001
Treated (Paris)	0.1976	0.0637	0.0019
Time: March 2024	-0.0815	0.0646	0.2068
Time: June 2024	-0.0806	0.0644	0.2107
Accommodates	0.1198	0.0045	< 0.001
Beds	0.0049	0.0068	0.4671
Treated: March 2024	0.1611	0.0691	0.0198
Treated: June 2024	0.6353	0.3850	0.0990

The interaction term `Treated: March 2024` (estimate = 0.1611, $p = 0.0198$) indicates a statistically significant increase in log prices for treated listings in March 2024 compared to control listings. While the interaction term `Treated: June 2024` shows a larger positive

effect (estimate = 0.6353), it is only marginally significant ($p = 0.099$). This result could be influenced by the limited nature of our dataset, which only includes data from December 2023, March 2024, and June 2024. Additionally, these data points are not time-consecutive, and we lack a sufficiently long pre-treatment period to establish baseline trends fully. This limitation may contribute to the high p-value observed for the June 2024 interaction term, as the variability in log prices could not be adequately accounted for with the available data.

5.3.3 Predicted Log Prices

Figure 7 visualizes the predicted average log prices for treated and control groups across the three time periods. The treated group exhibits a noticeable upward trend as the Olympics approach, consistent with the model's coefficients.

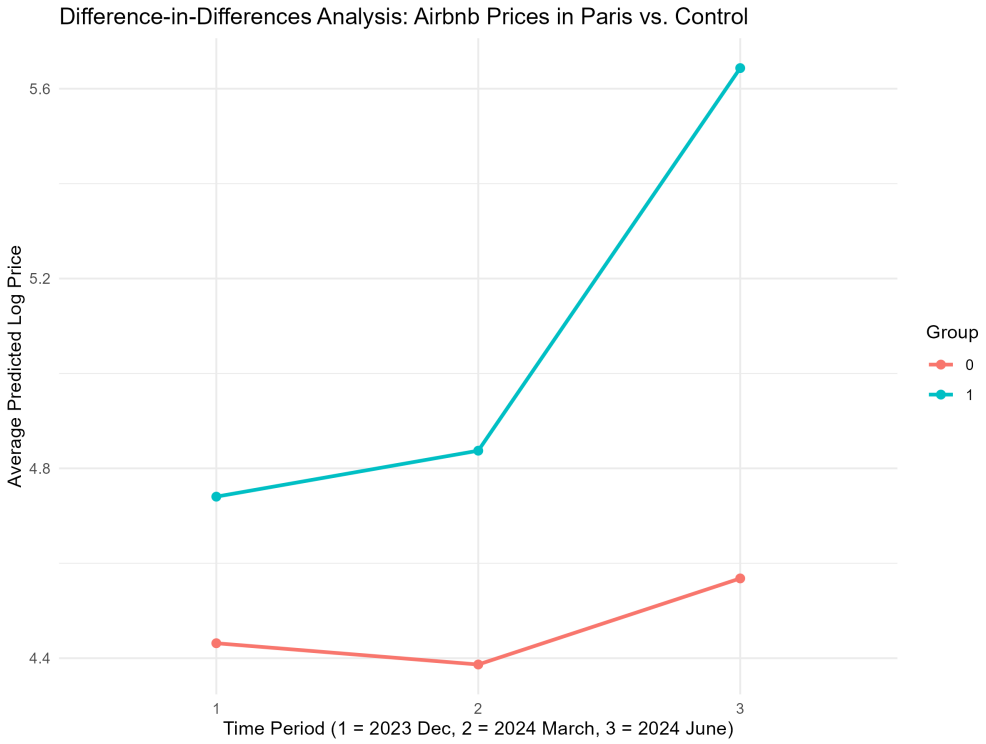


Figure 7: Difference-in-Differences Analysis: Predicted Average Log Prices for Treated and Control Groups Across Time Periods.

5.3.4 Key Findings

The DiD analysis highlights a significant increase in Airbnb prices in Paris (treated group) compared to the control group, particularly in March 2024. This suggests that the 2024 Paris Olympics exerted upward pressure on listing prices as the event approached. However, the effect in June 2024 shows high variability, warranting further investigation into possible confounding factors.

6 Discussion and Conclusion

The findings of this study provide valuable insights into the determinants of Airbnb prices in Paris, highlighting the influence of property characteristics, geographic clustering, and the impact of the 2024 Paris Olympics. By employing multiple models and spatial analysis techniques, we ensured robust and comprehensive conclusions, while aligning our results with the broader literature on short-term rental markets.

6.1 Discussion

Our analysis revealed that Random Forest and Gradient Boosting Machine (GBM) models significantly outperformed other predictive models, such as Ordinary Least Squares (OLS) and k-Nearest Neighbors (KNN). These results underscore the ability of tree-based ensemble methods to capture non-linear interactions and complex relationships within the data. The superior performance of Random Forest, in particular, reflects its robustness in identifying key predictors, such as `accommodates`, `bedrooms`, and `review_scores_rating`, which consistently emerged as the most influential variables. These findings align with previous studies, such as Herrgård and Flöjs [2], which emphasized the critical role of property size and quality in determining Airbnb prices.

Spatial analysis further elucidated the geographic patterns underlying Airbnb pricing. Geographically Weighted Regression (GWR) demonstrated that the effect of listing density on `log_price` varied significantly across neighborhoods, with negative effects in central Paris and small positive clusters in the northern regions. However, the magnitude of these effects was minimal, suggesting that while density influences pricing, its impact is highly localized and dependent on neighborhood characteristics. This nuanced spatial heterogeneity aligns with findings by Choi [1], who highlighted the importance of local features in price determination.

The Spatial Lag Model (SLM) quantified the spatial dependency of Airbnb prices, showing that nearby listings exert a statistically significant influence on the pricing of individual properties. This spillover effect reflects the competitive dynamics within the short-term rental market, as listings in proximity often compete for similar groups of travelers. Residual spatial autocorrelation, evaluated through Moran's I, further validated the presence of clustering, confirming that Airbnb prices exhibit significant spatial patterns. These results align with the theoretical expectations of spatial competition and provide empirical evidence for neighborhood effects in short-term rental markets.

Finally, the Difference-in-Differences (DiD) analysis revealed that the 2024 Paris Olympics had a substantial impact on Airbnb prices, particularly in the months leading up to the event. While treated and control groups exhibited similar pricing trends before March 2024, the difference in log prices significantly increased in June 2024. This suggests that the Olympics heightened demand for Airbnb listings in Paris, consistent with the broader literature on the effects of major international events on local rental markets. However, the limited temporal scope of the dataset constrained the ability to establish a longer pretreatment period, potentially affecting the robustness of these causal estimates.

6.2 Conclusion

This study contributes to the growing body of literature on short-term rental markets by providing a comprehensive analysis of Airbnb prices in Paris. Key determinants, including property characteristics, neighborhood effects, and external events like the 2024 Paris Olympics, were identified and analyzed using advanced modeling and spatial techniques. The findings emphasize the importance of combining multiple models to capture the complexities of pricing determinants and leveraging spatial analysis to account for geographic heterogeneity.

Future research could address the limitations of this study by incorporating more extensive time-series data to establish stronger pretreatment periods for causal analyses. Additionally, integrating alternative data sources, such as local economic indicators or tourism statistics, could further enhance the understanding of factors influencing Airbnb prices. As the short-term rental market continues to evolve, these insights will remain critical for hosts, policymakers, and researchers seeking to optimize strategies and mitigate the broader impacts of short-term rentals on residential communities.

Overall, this study underscores the dynamic and multifaceted nature of Airbnb pricing, providing actionable insights for stakeholders and a strong foundation for future research.

References

[1] J. W. Choi. Recommendation of price on airbnb using machine learning. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9 (2):1–10, 2019. URL https://www.ijitee.org/wp-content/uploads/papers/v9i2/B6445129219.pdf.

[2] J. Herrgård and J. Flöjs. Driving factors behind airbnb pricing: A multilinear regression analysis. *Diva Portal*, 2023. URL https://www.diva-portal.org/smash/get/diva2:1823748/FULLTEXT01.pdf.

[3] Inside Airbnb. Get the data, n.d. URL https://insideairbnb.com/get-the-data/. Accessed October 2024.

[4] Patrick A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2): 17–23, 1950. doi: 10.2307/2332142. URL https://doi.org/10.2307/2332142.

[5] P. Prajapati and M. Suthur. A survey on price prediction model for airbnb listing using machine learning. *International Journal of Scientific Research in Science Engineering and Technology*, 9(2):1–8, 2022. URL https://www.researchgate.net/publication/360441197_A_Survey_On_Price_Prediction_Model_for_Airbnb_listing_using_Machine_Learning.

A Code Appendix

Load Libraries

```r
library(dplyr)
library(tidyr)
library(readr)
library(naniar)
library(ggplot2)
library(jsonlite)
library(stringr)
library(forcats)
library(scales)
library(reshape2)
library(GGally)
library(kableExtra)
library(car)
library(broom)
library(randomForest)
library(sf)
library(forecast)
library(trend)
library(spgwr)
library(ggmap)
library(tmap)
library(spdep)
library(xgboost)
library(FNN)
library(spatialreg)
```

Import Data

```r
# Import data from multiple compressed CSV files and combine into
    a single dataset
listings_1 <- read.csv("listings_1.csv.gz")
listings_2 <- read.csv("listings_2.csv.gz")
listings_3 <- read.csv("listings_3.csv.gz")
listings_4 <- read.csv("listings_4.csv.gz")

# Combine all datasets row-wise
listings <- rbind(listings_1, listings_2, listings_3, listings_4)

# Create a backup of the original dataset for reference
listings_original <- listings
```