Final

**Wichita State University**

**Barton School of Business**

Wichita, Kansas

****

**BSAN 775:** Introduction to Business Analytics

**Project Report on:**

Predictive Pricing Dynamics for Short-Term Rent: An Integrated Approach to Host Clustering and Price Prediction

| **Submitted by: Group 3**  Anjali Khanal  Coto Nagamine  Ha Nguyen  Samip Khand Thakuri | **Submitted to:**  Prof. Zahra Ziaei |
| --- | --- |

**December 5, 2024**

**Table of Content**

[I. Introduction: 2](#_lgk2h6ww6a5c)

[II. Problem Statement: 2](#_yayabb2dn0ik)

[III. Objectives: 2](#_alleio54055p)

[IV. Literature Reviews: 3](#_nwy6r5801etf)

[1. Price Prediction: 3](#_j8n1q13w8nw3)

[2. Geographic Clustering: 3](#_401v6mfp77iq)

[3. Gap in Literature: 3](#_mkbu5jsulqt2)

[V. Methodology: Model, Data Source, and Techniques of Analysis 4](#_gxlg54hidgnm)

[1. Data Collection: 4](#_8l7f3j77jfjk)

[2. EDA: 4](#_d0op5htq3l4e)

[a. Data Visualization: 4](#_btf7fzwqbgm6)

[b. Handling Missing Data: 4](#_wmsbugu9130b)

[c. Data Type Conversion: 5](#_tx7ae7c2thed)

[d. Outlier Detection and Treatment: 5](#_vwrdcf7ogmzg)

[e. Feature Engineering and Scaling: 5](#_ibxt4yktjc9t)

[VI. Results: 6](#_nxas5yxtlzg)

[VII. Discussions: 9](#_szsgkgif9ocp)

[VIII. Future Work and Limitations 11](#_6rm64lu2m27t)

[Future Work 11](#_8f8indfjw51o)

[APPENDIX: 13](#_bt8wzrgi2u4t)

[References 28](#_csqhhturfqc)

# **Introduction:**

Airbnb among other platforms has opened a wide avenue for homeowners to generate an income stream from their spare rooms or their homes. With the rising popularity of the platforms, there is a growing interest in understanding the factors that contribute to pricing variations. Accurate predictions of rental prices, based on essential features, empower hosts to optimize their listings and remain competitive in a dynamic market while providing potential renters with better value propositions.

This study aims to analyze an Airbnb listing dataset available in New York and identify the most significant predictors among a variety of features such as neighborhood, room type, number of reviews, and listing availability. The paper seeks to reveal how these features influence rental pricing by applying clustering and predictive modeling techniques.

# **Problem Statement:**

With the rising number of players in the AirBnB market, many hosts are seeking ways to competitively price their property without making unnecessary sacrifices on profits. However, numerous factors –ranging from location to user reviews– may influence the price of listings. The goal of the study is to address the following questions by employing different machine-learning models and techniques:

1. What are the key features that significantly impact the price of short-term rental listings?
2. How can insights from the model's predictions inform hosts' pricing strategies?
3. To what extent does machine learning provide answers to these critical questions?

# **III. Objectives:**

T​he goal of this study is to investigate some of the factors that determine Airbnb prices in New York and provide a predictive model that will assist the hosts with price setting. The objectives are the following:

1. Determine and analyze the most pronounced determinants of price of Airbnb listings.
2. Assess the performance of the formulated models in predicting Airbnb prices, as well as investigate the effectiveness of incorporating clustering compared to the pure regression model in price prediction prowess.
3. Provide actionable insights to allow hosts to optimize their price-setting

# **IV. Literature Reviews:**

## **1. Price Prediction:**

Price prediction has always been one of the most popular research topics in the field of short-term rentals. One example is a study by Wang and Nicolau in 2017 that used Ordinary Least Squares (OLS) and Quantile Regression analysis to conclude that all 25 variables in 5 categories – host attributes, site and property attributes, amenities and services, rental rules, and online review ratings– are good predictors of price among Airbnb listings across 33 cities.

In a different study, Alhabri employed more complex and extensive models such as Lasso, Ridge, KNN, decision tree regression, and other models, to analyze Airbnb listings in 191 countries using three sets of features: property attributes, listing specific, and reviews characteristics (2023). The study concludes that regardless of the models used, the number of beds and number of guests accommodated consistently ranked among the most important pricing factors.

In 2019, Lawani et al. used textual mining to demonstrate that a more comprehensive approach combining both numeric and sentiment analysis can significantly improve the performance of price-prediction models, thus demonstrating a need for a more comprehensive approach to price-predicting.

## **2. Geographic Clustering:**

A 2018 paper by Forgosh, Nuru, and Ubiera applied clustering in mining for latitude and longitude data points for AirBnB listings in New York City to examine the frequency distribution of the price within each cluster and generalize pricing patterns among the identified groups, demonstrating the effective use of clustering in the AirBnB problem space.

Another study by Tang and McNicholas in 2017 introduces penalty parameters into clustering techniques to perform AirBnB reviews sentimental analysis and revealed two latent traits (concerning the property and the host) and four cluster of reviews (compound, neutral, positivity, and negativity). The model effectiveness suggests a significant opportunity to incorporate clusterings algorithms for deeper insights into the short-term rental domain.

## **3. Gap in Literature:**

It is important to recognize that the boroughs and neighborhoods in NYC often fall short in making accurate predictions because of their ambiguous nature, leading to potential inaccuracies in data and modeling. While there is a significant amount of literature on AirBnB price predictions using regression models and the distribution of price among geographical clusters, further research need to be conducted to determine the role of clustering on predictive models in short-term rental pricing in New York. Therefore, this study aims to investigate the impact of geographic characteristics on the pricing prediction power of machine learning models. By exploring this possibility, this study aims to improve the estimated price of AirBnB in New York City.

# **V. Methodology: Model, Data Source, and Techniques of Analysis**

## **Data Collection:**

This project used an Airbnb listing dataset for New York City, obtained from Inside AirBnB, a third-party site that provides quarterly Airbnb data. The dataset contains over 37,000 data points and many features such as room type, neighborhood, number of reviews, etc. The data contain both statistical and categorical variables, which are predictors of rental prices. Figure 1 in the Appendix shows the summary of the dataset. We focus on 16 independent variables (e.g., room type, location, reviews) and one dependent variable (price).

## **EDA:**

The analysis will start with a visual examination to evaluate the information traits (i.e mean, median, and widespread deviation of key variables) and fashion analysis to observe how condo fees vary across neighborhoods. Once the missing values and outliers are detected, appropriate investigation and techniques are applied to process these data points before applying machine learning models to ensure the models’ performance.

The average price of each neighbourhood is as follow: Bronx has the mean price of $ 112.495070, Brooklyn with $169.821215, Manhattan with $273.910879, Queens’ average price among the listing is $118.125743, Staten Island commands $127.064384 as the average price.

### Data Visualization:

Preliminary data visualization is carried out to assess the relationship between independent variables and the dependent variables. Figure 8 shows that location, represented by latitude and longitude, is a key driver of Airbnb prices, with more upscale locations commanding higher prices. Host experience, measured by calculated\_number\_of\_host\_lists, also plays an important role, as hosts with multiple features tend to charge more. Additionally, availability\_365 shows a positive correlation with price, indicating that properties that are available all year round attract higher demand and higher prices. In contrast, features such as reviews\_per\_month and minimum\_nights have lower correlations, suggesting that they have limited impact on price compared to location and availability.The graph in Figure 5 shows that listings with Entire Homes/Apartment dominates the room\_type feature

### Handling Missing Data:

The dataset contained missing values, which were addressed using imputation techniques. Numerical features with missing values were filled using the median to reduce the impact of outliers, while categorical variables were filled using the mode, preserving the most frequent category. Missing values in the target variable (price) were excluded, as they are essential for model accuracy.

### Data Type Conversion:

Certain columns initially contained text or non-numeric values. For proper model compatibility, categorical columns like neighbourhood\_group and room\_type were transformed using one-hot encoding, converting them into binary features. Numerical columns were verified and corrected to ensure they were compatible with machine learning algorithms.

### Outlier Detection and Treatment:

Outliers in the price, minimum\_nights, and number\_of\_reviews columns were identified using boxplots. Extreme outliers, particularly listings with nightly rates exceeding $100,000, were removed to prevent skewing the model. Overall, the values over the 99th percentile in the mentioned columns are investigated and are capped to the main model reliability, with the exception of calculated\_host\_lisitngs because verification shows that the same host\_id appears on 876 different listings, making it a valid data point, and removing 876 listings from the dataset could result in information loss.

### Feature Engineering and Scaling:

Figure 9 shows that the target variable price is extremely skewed to the left because of the presence of very high-priced listings that do not constitute the outliers in the dataset. To address this problema and preserve the performance of the predictive machine learning model, log transformation is applied to the price, transforming it to a normal distribution bell curve as shown in Figure 10. Thus, the target variable is price\_log.

The timeliness of reviews plays an important role in the pricing of AirBnB because they suggest that the host is actively managing the listing and that the property is in good condition. This can increase demand and potentially justify higher pricing. Listings with older reviews or no reviews may signal lower demand or decreased host activity, leading to lower prices. Thus, time\_from\_last\_review was created by calculating the days since the last review for each listing. For listings with missing reviews, a placeholder value of 9999 was used. This feature captures the effect of review recency on the price of the listings.

Although Random Forest models are insensitive to feature scaling, StandardScaler was used for consistency across numerical features. This also ensures that if other models requiring scaling, like Linear Regression, are considered, they will be compatible.

# **VI. Results:**

**Model Selection**

Several machine learning models were tested to predict Airbnb prices, including Linear Regression, Random Forest, Gradient Boosting, and ElasticNet and Ridge and Lasso. These models were evaluated based on their Mean Squared Error (MSE) and R-squared (R²) values, with Random Forest emerging as the most effective. A summary of the model performance is shown in Feature 12 in the Appendix.

Linear Regression Model was first explored due to its simplicity and interpretability, while it was not selected, this model provided the baseline to understand the relationship between features and pricing. There were two versions of the Linear Regression Model, one with only the numeric features which yield an R² of 0.204, while the latter included one-hot encoding of the neighbourhood\_group and room\_type yields and R² of 0.405. The summary of this OLS model is included in Figures 3 and 4 in the Appendix.

Next, Gradient Boost is explored because this model can handle non-linear relationships and due to its ability to correct weak tree models into a powerful final model. However, for this particular dataset and application, the Gradient Boost model has an R² of 0.641, significantly lower than that of the Random Forest model.

The LASSO, Ridge, and Elastic Net were fitted because they can reduce overfitting and handle feature selection and multicollinearity, which might be present due to the high number of features (such as one-hot encoded variables). The RMSE and Ridge Regression had the best performance among the three, with R² of 0.435, but still underperformed compared to Random Forest.

Based on its superior R², ability to generate feature importance, and to handle complex relationships, Random Forest was selected as the final model. Its high performance, combined with its interpretability through feature importance, makes it the best choice for predicting Airbnb pricing.

**Hyperparameter Tuning:**

The Random Forest model underwent hyperparameter tuning using GridSearchCV to optimize its parameters. The selected configuration—359 trees, a maximum depth of 40, and at least 4 samples per split—resulted in an R² of 0.748 and an RMSE of 0.122, significantly improving model performance compared to its initial settings (R² of 0.746, RMSE of 0.121).

**Feature Importance**

Figure 13 shows the ranking of the most important features according to the Random Forest model. The high score of both longitude (0.335) and latitude (0.165), and the relatively low scores of the neighbourhood\_group features suggests the application of clusterings to uncover more a nuanced and appropriate geographical feature for the predictive model.

**Cluster Analysis:**

Two types of clustering techniques were applied: K-means for its simplicity and DBSCAN because it does not assume the spherical shape of the data. Figures 14 and 16 show the visualization of the clustering on a graph where longitude is the horizontal axis and latitude is the vertical axis.

. Based on the Elbow Curve shows in Figure 11, k=5 was selected as the targeted number of clusters. For this value of K, the K-means cluster has an silhouette score of 0.4118. With trial-and-error, K=5 is determined as the best number of clusters for the geographical purpose.

As K-Means clustering identified five geographic clusters with distinct pricing behaviors as show in Figure 15:

* **Central High Price Cluster (Cluster 5)**: Average price\_log of **5.505788 (log)** or $246.11
* The other clusters do not demonstrate significant variation in mean price, but has potential for further pricing characteristics exploration.

The DSCAN clusters in Figure 16 involve a major cluster -1 and the minor clusters closer to the border of the graph, suggesting the model is susceptible to noise. Therefore, the final selection was the geo\_cluster generated by the K-means algorithm for a more stable, interpretable, and the potential to generate more actionable insights.

**Applying Clusters to Random Forest**

The geo\_cluster generated from K-means clustering is used as a feature in Random Forest to evaluate the impact of geographical clustering to Random Forest predictive power. The goal is to determine if including the regional patterns helps the model explain more variation in price and make better predictions. The resulting R² is 0.5889, and the MSE is 0.1973, which are lower than the R² and MSE of the base model. The feature importance analysis is used as the benchmark to measure its effectiveness as shown in Figure 17. Even though the feature contributed minimally to the model, it still provides valuable insights on the model’s ability to interpret geographical features, and at the same time highlight the role of non-geographical features in price predictions.

**Applying Random Forest to Clusters**

The goal of this section is to understand the performance of Random Forest model on each cluster. Random Forest’s efficiency in capture non-linear relationship makes it an ideal candidate for capture the interplay of variables in the clusters. The performance of Random Forest on each cluster is detailed in Figure 18 in the Appendix, with Cluster 3 having the highest R², suggesting a good model fit and MSE of 0.1131, which characterize a stable model. On the other hand, Cluster 2 was the weakest performing model, with R² of 0.4692 and MSE of 0.1859.

Figure 19 shows the most important features identified by Random Forest algorithm in each cluster. Here, longitude appears in the top three features for 4 out of 5 models and latitude appears in 3 out of 5 models, noting the importance of the exact location of the listings, as well as its closeness to key amenities as an important factor in price determining. However, there are also varying features in the top 3 for each model, suggesting model’s specific price-influencing factors.

**Applying Linear Regression to Clusters**

Linear Regression was applied to the different clusters identified by K-means clustering to evaluate the features in a more interpretable manners using their coefficients magnitude and p-values. The purpose of this step was to examine whether applying linear regression to these distinct clusters could provide more meaningful coefficients for the relationships between the features and prices in each specific geographic area. The key goal was to understand how prices in different clusters could be explained by distinct sets of features, making linear regression more effective when applied to individual clusters rather than the entire dataset.

Figures 20 to 24 summarize each model performance. Base on the key metrics, it can be concluded that Cluster 3 with the highest R² value presents the best model fit, while Cluster 4 with the lowest R² suggests variations in price is not fully captured by the model. It is worth noting that latitude’s strong influence on most models reveals that for most clusters, moving higher in the latitude coordinates positively influence the price of the listings. Other features like Room Type also see the positive dominance of Hotel Room when compared to Shared room or Private room in the same clusters.

**Model Evaluation Using OOB (Out-of-Bag) Score**

To evaluate the final performance of the Random Forest model and assess its generalization ability, the Out-of-Bag (OOB) score was used. The OOB score is a built-in cross-validation method within Random Forest that allows for an internal estimation of the model's accuracy by using each tree’s out-of-bag samples, or in other words use the samples that have not been trained for the particular tree to test the performance of that tree. The OOB score of the Random Forest is provided in Figure 25, with an OOB score of 0.7076, which indicates a strong ability to generalize and predict unseen data. Additionally, the Mean Squared Error (MSE) was 0.1228, and the R-squared (R²) value was 0.7441. These results confirm that the Random Forest model performed well, and the OOB score aligns closely with the training set’s performance, suggesting that the model is not overfitting and is capable of making reliable predictions.

# **VII. Discussions:**

The study attempts to answer the questions on price determinants of AirBnB listings, then compare the performances of Linear Regression and Random Forest model in predicting short-term rental rate as well as evaluating the impact of clusterings on each model. Finally, the study delivers actionable insights to hosts of each cluster.

The most prominent price predictors are determined based on the feature importance analysis of the base model, combined with further exploration by cross applications of other models: With their high frequency appearance, longitude and latitude are two most important factors in determining pricing of a short-term rental. Cluster 1 shows that listings belonging to geographic clusters enjoyed significantly higher prices than the rest of the clusters. Room type is the next feature to influence pricing with Hotel rooms and Entire homes/Apartments having consistently higher ranks and coefficients than Private room and Shared Room, confirming space as one of the key players in pricing determinants. Another important feature is Calculated\_host\_listing\_counts, especially when analyzing geographic clusters, suggesting the positive influence of managing multiple properties to pricing commands. Listings that are available year round also enjoy better pricing than limited opening listings, hinting at the importance of a dedicated AirBnB space as well as the consistency of the listing. Minimum nights is one of the features that hold significant impact on the price, with high minimum nights correlate with high rental price,

The Random Forest model demonstrates superior performance, with an R-squared of 0.746 when applied to the entire dataset. When applied to individual clusters, the R-squared ranges from 0.6000 to 0.6713, depending on the cluster. This indicates that Random Forest effectively captures non-linear relationships and interactions between features, which leads to better performance compared to Linear Regression models. The clustering process further enhances Random Forest's predictive power, particularly for clusters with clear, consistent pricing patterns, such as Cluster 3.

In contrast, when Linear Regression is applied to the full dataset, the model achieves an R-squared of 0.204 (with numeric features only), which increases to 0.405 when incorporating one-hot encoded features like neighbourhood\_group and room\_type. However, when applied to individual clusters, the performance of Linear Regression varies. While Cluster 4 achieves an R-squared of 0.205, which is relatively low, reflecting the complexity of this cluster’s pricing, Cluster 0 and Cluster 1 show moderate performance, with R-squared values of 0.356 and 0.377, respectively. Cluster 3 achieves a higher R-squared of 0.384, indicating that the linear relationship in this cluster is more consistent.

Applying clustering to Linear Regression allows the model to focus on more localized pricing patterns, improving performance over the full model, but it still struggles in capturing more complex, non-linear patterns compared to Random Forest. However, by sacrificing some predictive and explanatory features in both models, the important pricing characteristics of each model can be made clear for more actionable insights.

The following actionable insights are recommended to hosts in each cluster:

Cluster 4 includes listings in the most desirable neighborhoods, especially near midtown Manhattan, and tends to have significantly higher prices. These high prices are largely due to the area’s prime location, close to major tourist attractions, transportation hubs, and high-demand neighborhoods. Additionally, experienced hosts or hosts with multiple listings often leverage their reputation and expertise to capitalize on this premium market and charge higher prices.

Another important aspect of this cluster is its year-round availability. Listings that are listed as available 365 days a year consistently attract demand from both tourists and business travelers, and command higher prices. Your latitude also plays a role, as listings further north often charge higher prices. This is due to the northern borough’s attraction to parks, cultural attractions, and universities, all of which attract steady traffic.

In Cluster 4, room type is also an important pricing factor. Private rooms, in particular, are more expensive than shared spaces because they offer privacy and added amenities that guests appreciate. Longitude also plays a role, as properties near affluent areas like the West Beach and Upper West Side are often overpriced. Together, these features make Cluster 4 one of the most profitable segments for Airbnb hosts.

# **VIII. Future Work and Limitations**

**Limitations** This model might not represent other regions of the USA as this is based on a dataset from New York city. As the dataset used here is specific to New York City so the ability of the model is limited. House pricing in other cities like San Francisco, Los Angeles, Wichita or Miami may differ because of various factors like more tourism areas, bigger market, or local economic condition. Since the model is trained for New York only, it might not be accurate in other cities.

Another limitation is the model's ability to account for dynamic market changes may limit due to lack of real-time data. This model relies more on historical data that was collected over a period which does not contain real-time data. This does not support real-time trends or dynamic changes in price. During seasonal times tourism goes on peak and at that time prices tend to be higher, or on holidays rental prices become higher while in off season it goes downward. Such seasonal change causes change in price so this model may not be able to predict it. In some areas in some big events like concerts, or sports can make slight changes in price. Lastly, border economic changes, political changes, inflation and other major changes also impact consumer behaviour and rental prices so such sudden occurring factors may not be reflected in our dataset.

Detailed property amenities, quality of guest review and weather conditions are excluded. Some factors like internet speed, air conditioner or swimming pool availability can also cause price changes. Recent google reviews, comments and ratings are also another reason for changes in prices which might not be included in our dataset.

Host behaviour may not fully be captured by K-means clustering. Hosts have different factors like responsiveness, listing updates, or price negotiation, which affect pricing strategies but all hosts are grouped together like in room type or geographic locations by K-mean clustering. For k-mean predefining of clusters number is necessary

Missing data and outliers are handled well but it still might introduce biases in the results. Median or mode assumes missing data while filling it, and for more missing data it could be misleading or biased. Even if outliers are capped, they still affect models and be less reliable.

## **Future Work**

**Incorporating Real-Time Data:** We could integrate real-time data, market trends, and guest behaviour in future studies for improving model’s accuracy and adaptability changing market. The model would adapt current market trends, or guest behaviours and change prices rapidly.

**Cross-Regional Analysis:** Explore other regions of the country and expand the analysis of other major Airbnb markets. It could uncover region-specific pricing determinants. Different regions have different features like culture, rules and regulations, unique market, tourism area, stadium, etc., if these features are adapted by model then it would improve model’s ability.

**Enhanced Feature Engineering:** Exploring additional features like detailed amenities, host reputation, and review sentiments analysis which would enhance predictive power of the model. Renting areas pool access, wifi speed, google reviews, ratings affects the pisces, if model could capture these features then that would improve model accuracy.

**Advanced Clustering and Segmentation:** For segmenting hosts in a better way and uncovering more patterns, we would like to explore other clustering methods like DBSCAN (Density-Based Spatial Clustering of Applications with Noise). This method could identify individual clusters with its similar behaviour.

**Integration of External Data:** Including local events, weather and economic indicators for enhancing the model's predictive capabilities. Real-time events, or predicting upcoming events like concerts, festivals, peak vacation seasons, inflation, political changes, or boarder economic change, could be adjusted by model then the model would be more reliable and accurate.

**Price Optimization Framework:** We could use reinforcement learning or other advanced learning that would optimize prices automatically based on market conditions. Model could be set automatically to change according to market conditions which will adapt price recommendation, continuously learning and the model is updated.

# **APPENDIX:**

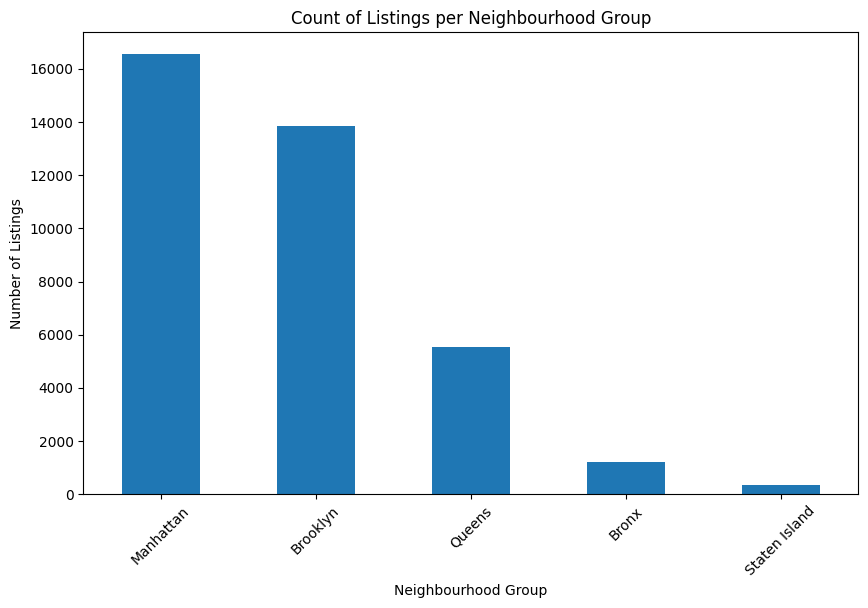


Figure 1. Distribution of Neighborhood Group

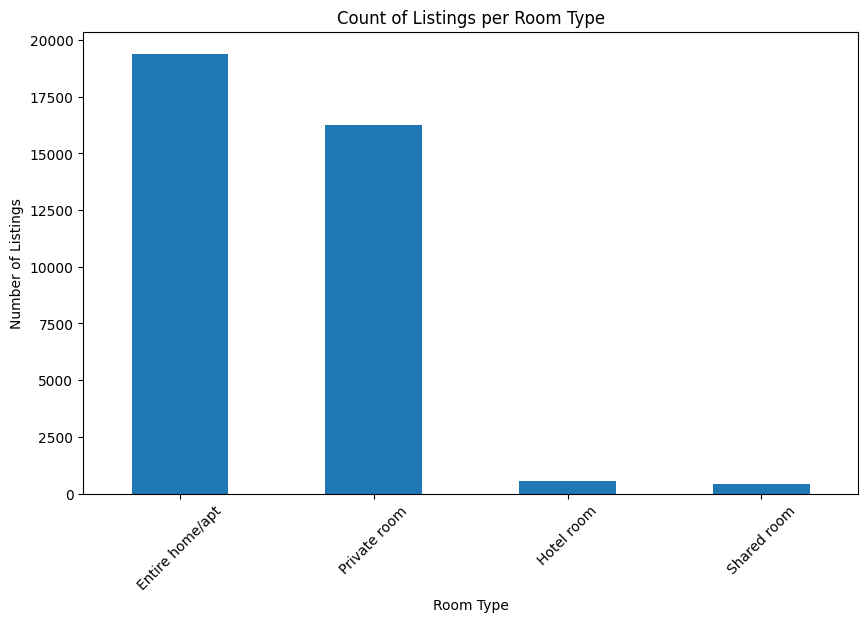


Figure 2. Distribution of Room Type

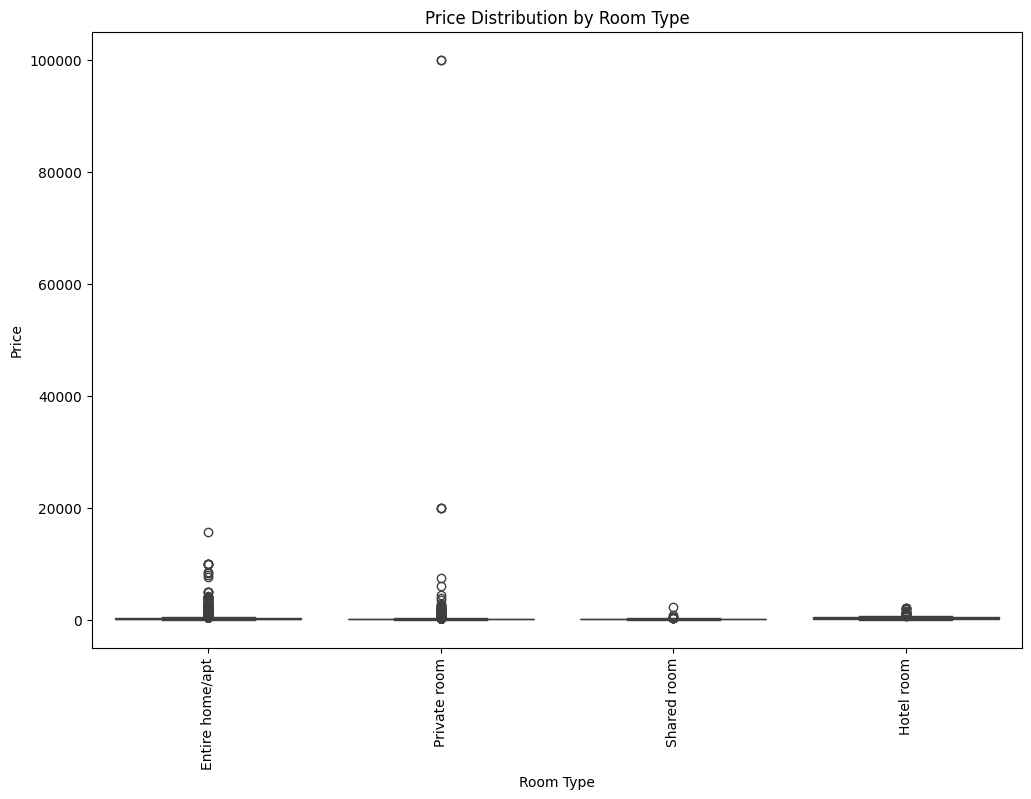


Figure 3. Room Type vs Price

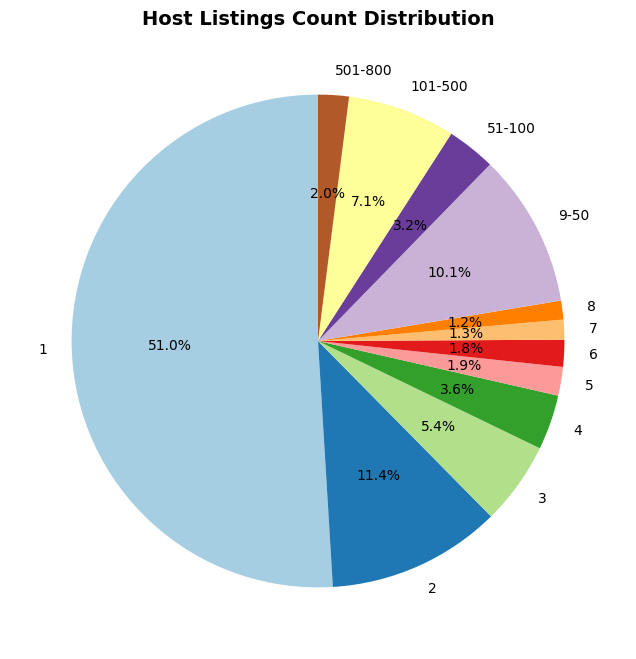


Figure 4. Distribution of calculated\_host\_listings\_count

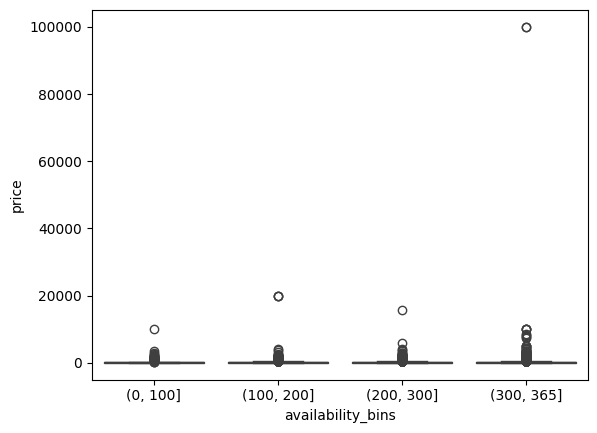


Figure 5. Availbility\_bins vs price

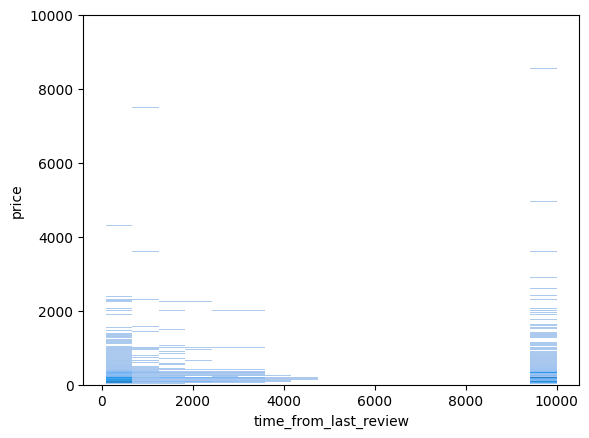


Figure 6. Time\_from\_last\_review vs Price

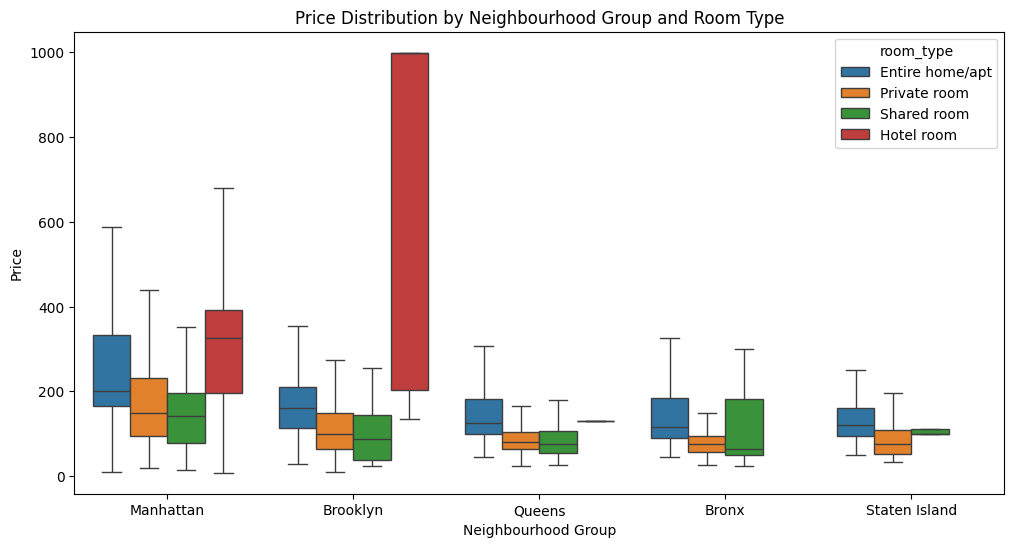


Figure 7. Interaction between room type, neighbourhood groups, and price

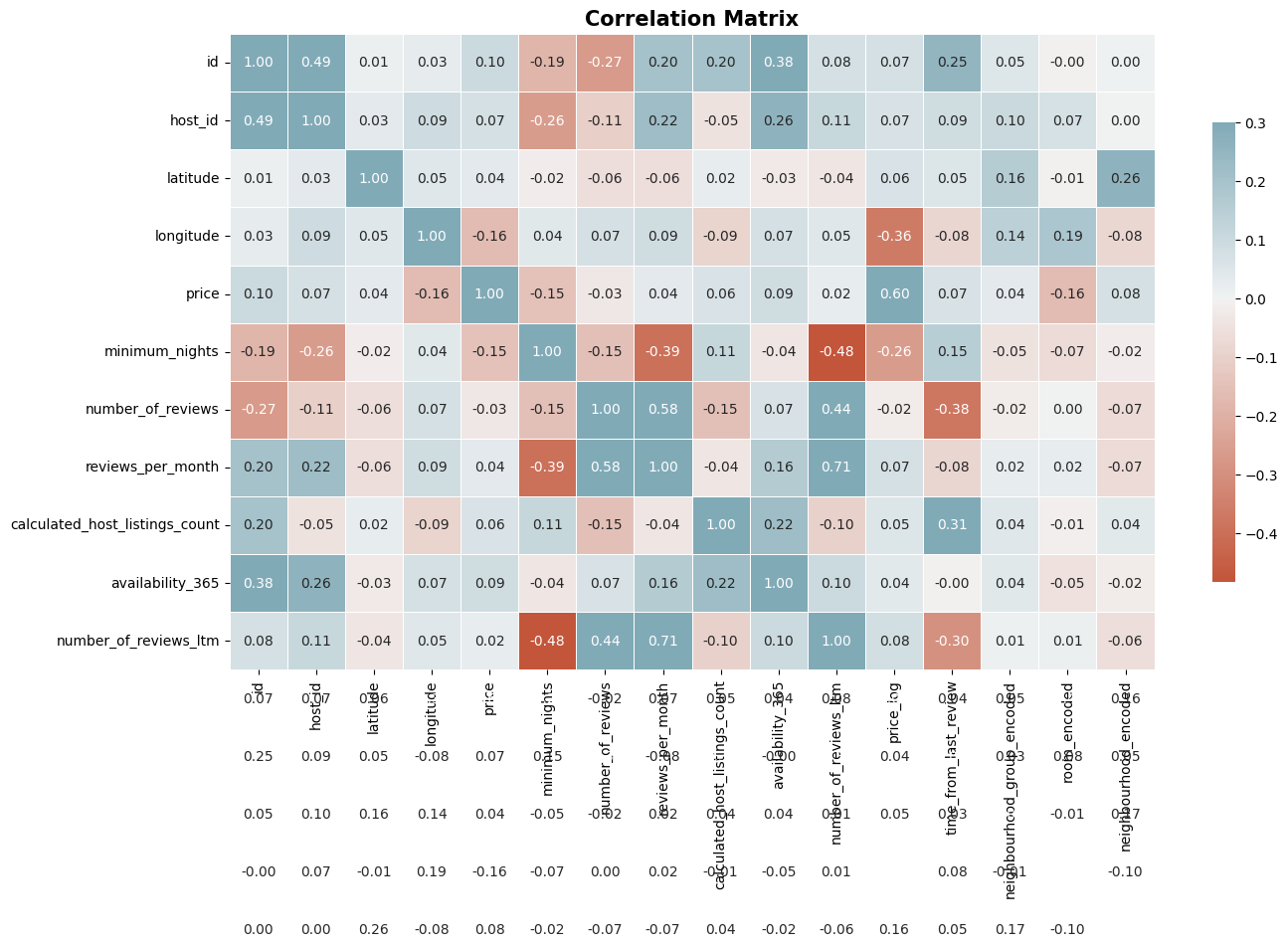
****

Figure 8. Correlation Matrix

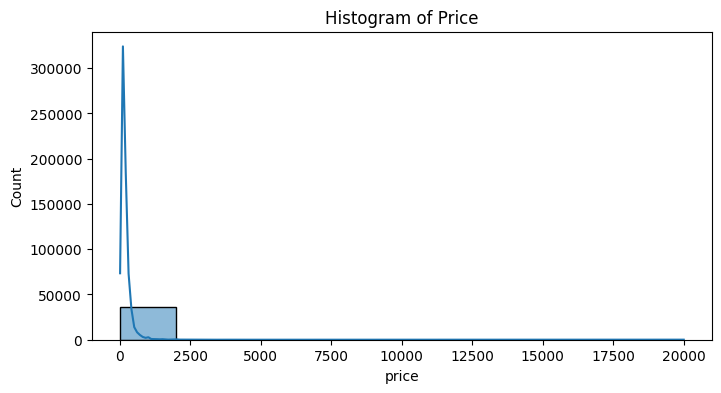


Figure 9. Skewness of Price

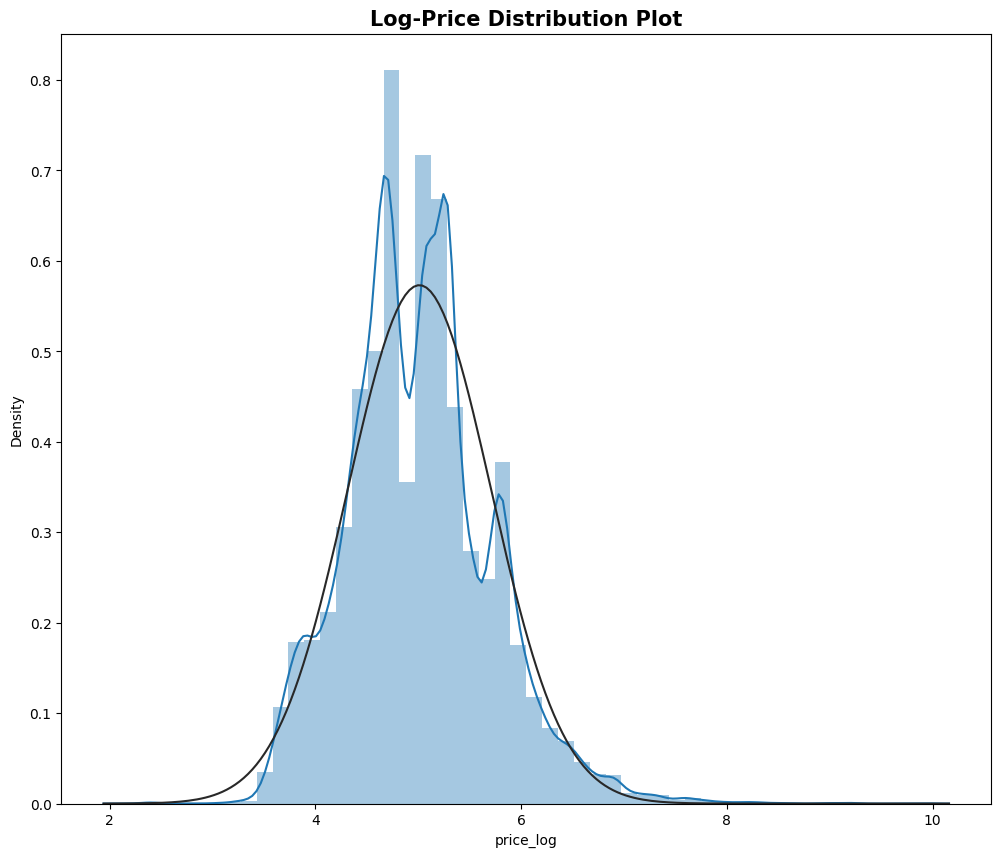


Figure 10. Price Normalization

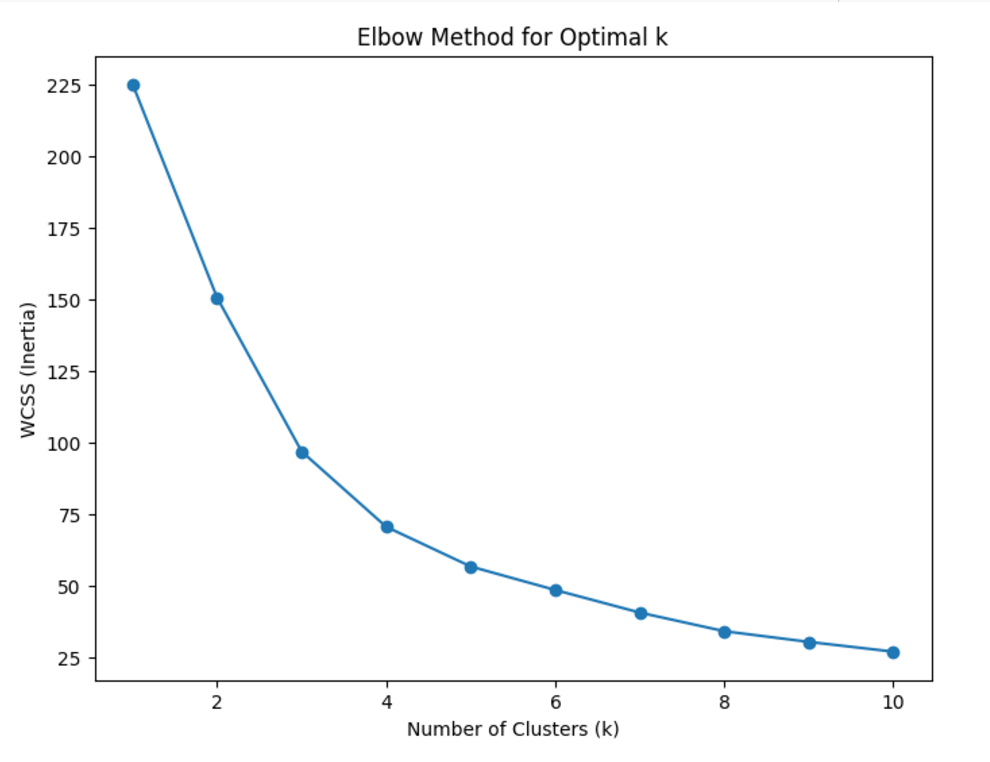


Figure 11. Elbow Method for Optimal K

| **Model Name** | **RMSE** | **R-squared** |
| --- | --- | --- |
| Numeric Only Linear Regression |  | 0.204 |
| Numeric and Categorical Linear Regression |  | 0.405 |
| Random Forest | 0.121 | 0.746 |
| Gradient Boosting | 0.172 | 0.641 |
| LASSO | 0.317 | 0.338 |
| Ridge | 0.271 | 0.435 |
| Elastic Nets | 0.294 | 0.387 |
| Random Forest with Hyperparameter Tuning | 0.122 | 0.748 |

Figure 12. Initial Fitting Models’ Performance

| **Ranking** | **Feature Name** | **Score** |
| --- | --- | --- |
| 1 | longitude | 0.3350 |
| 2 | latitude | 0.1639 |
| 3 | calculated\_host\_lisitngs\_count | 0.1112 |
| 4 | availability\_365 | 0.1046 |
| 5 | room\_type\_Private room | 0.1046 |
| 6 | minimum\_nights | 0.0059 |
| 7 | time\_from\_last\_review | 0.0418 |
| 8 | reviews\_per\_month | 0.0279 |
| 9 | number\_of\_reviews | 0.0239 |
| 10 | number\_of\_reviews\_ltm | 0.0108 |
| 11 | room-type\_Shared room | 0.0107 |
| 12 | neighbourhood\_group\_Brooklyn | 0.0018 |
| 13 | neighbourhood\_group\_Queens | 0.0016 |
| 14 | room\_type\_Hotel room | 0.0015 |
| 15 | neighbourhood\_group\_Staten Island | 0.0004 |
| 16 | geo\_cliuster | 0.00 |

Figure 13. Random Forest Feature Importance

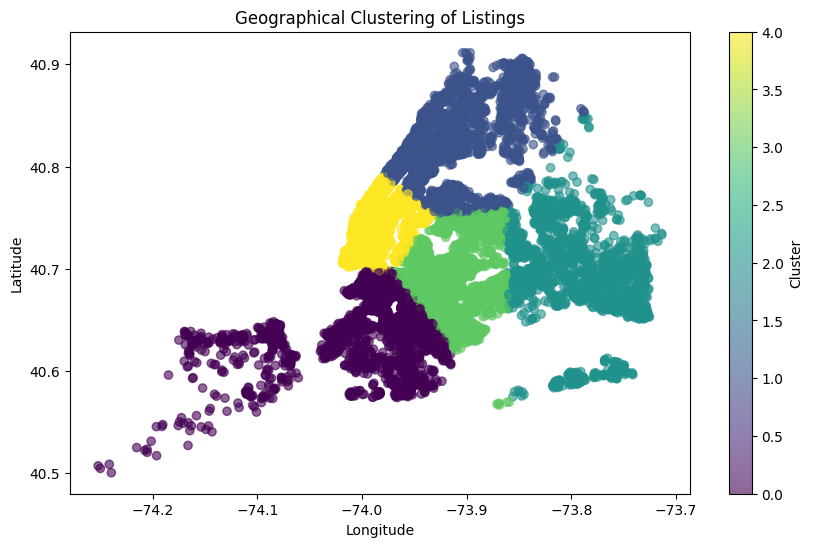


Figure 14. K-means geo\_clusters on latitude and longitude graph

| **Geo\_cluster** | **Description** | **Mean Price (log)** | **Mean Price (USD)** |
| --- | --- | --- | --- |
| 0 | Urban Core - Mid-Priced | 4.855197 | 128.405 |
| 1 | Affordable Urban | 4.778366 | 118.90 |
| 2 | Suburban Budget Friendly | 4.653657 | 104.97 |
| 3 | Transitional Zone | 4.705453 | 110.55 |
| 4 | Central - High Price | 5.505788 | 246.11 |

Figure 15 . Geographic Clusterings using K-means

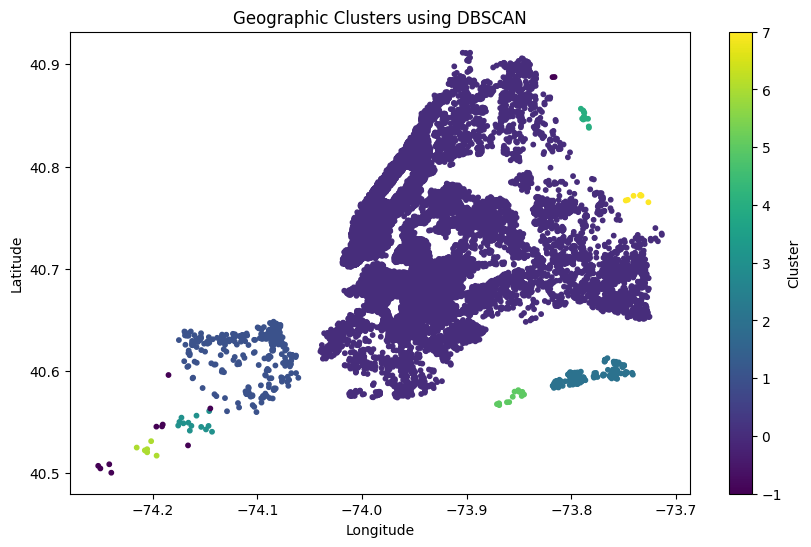


Figure 16. Geographic Clusterings using DBSCAN

| **Ranking** | **Feature Name** | **Score** |
| --- | --- | --- |
| 1 | room\_type\_Private room | 0.1906 |
| 2 | availability\_365 | 0.1707 |
| 3 | calculated\_host\_lisitngs\_count | 0.1469 |
| 4 | minimum\_nights | 0.1199 |
| 5 | time\_from\_last\_review | 0.1121 |
| 6 | neighbourhood\_group\_Manhattan | 0.0921 |
| 7 | reivews\_per\_month | 0.0612 |
| 8 | number\_of\_reviews | 0.053 |
| 9 | number\_of\_reviews\_ltm | 0.0181 |
| 10 | neighbourhood\_group\_Brooklyn | 0.0150 |
| 11 | room-type\_Shared room | 0.0139 |
| 12 | neighbourhood\_group\_Queens | 0.0032 |
| 13 | neighbourhood\_group\_Queens | 0.0013 |
| 14 | neighbourhood\_group\_Manhattan | 0.0011 |
| 15 | room\_type\_Hotel room | 0.0011 |
| 16 | neighbourhood\_group\_Staten Island | 0.0011 |
| 17 | geo\_cluster | 0.00 |

Figure 17. Applying Clusters to Random Forest Feature Importance Analysis

| **Cluster** | **RSME** | **R-Squared** |
| --- | --- | --- |
| 0 | 0.1607 | 0.6000 |
| 1 | 0.1336 | 0.6093 |
| 2 | 0.1210 | 0.6456 |
| 3 | 0.1131 | 0.6713 |
| 4 | 0.1859 | 0.4692 |

Figure 18. The results of applying Random Forest to Clusters

| **Cluster** | **Feature 1** | **Feature 2** | **Feature 3** |
| --- | --- | --- | --- |
| 4 | Calculated\_host\_  listings\_count | longitude | availability\_365 |
| 0 | latitude | Room\_type  \_Private room | longitude |
| 1 | room\_type\_Private room | latitude | longitude |
| 3 | room\_type\_Private room | Calculated\_host\_  listings\_count | availability\_365 |
| 2 | latitude | room\_type\_Private room | longitude |

Figure 19. The Clusters' Feature Importance

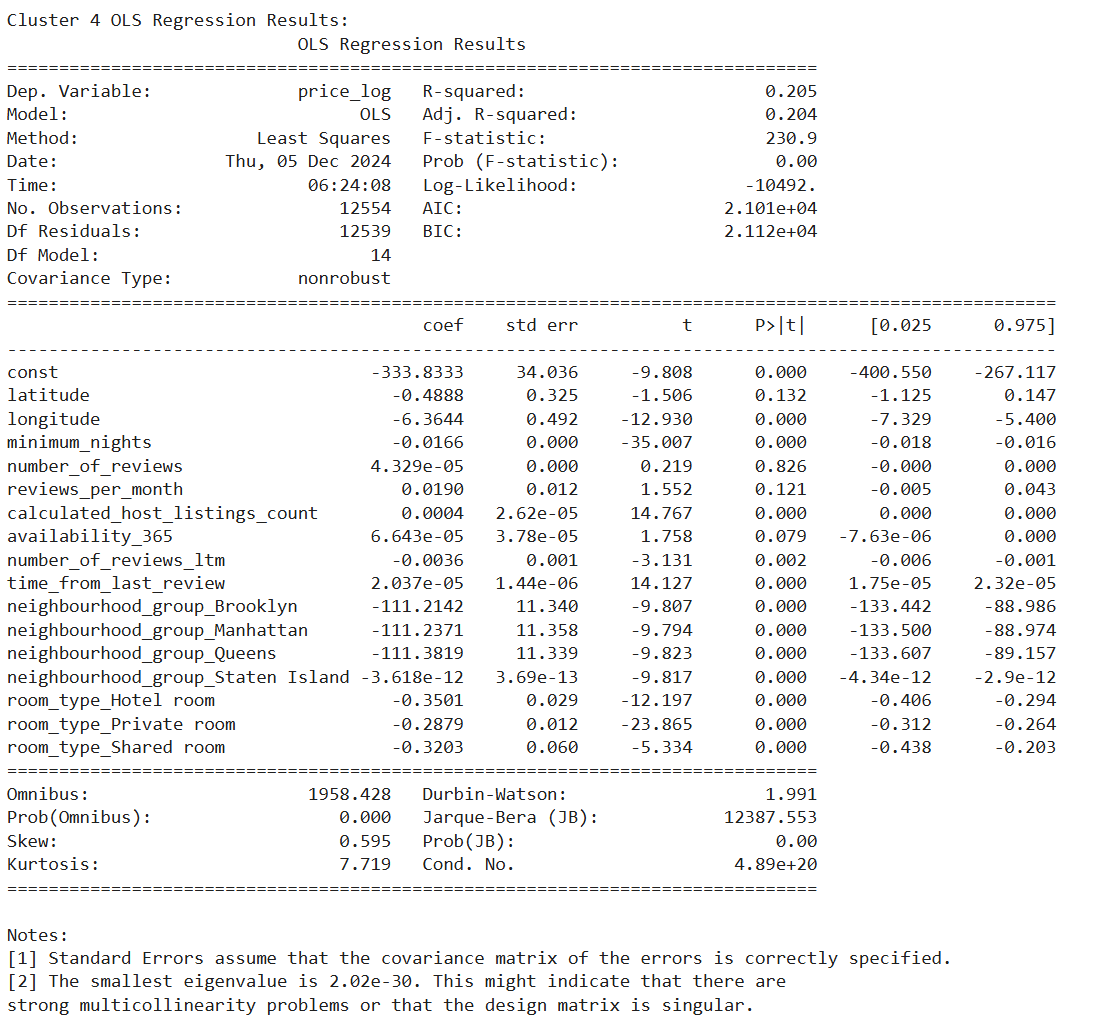


Figure 20. Cluster 4 OLS summary

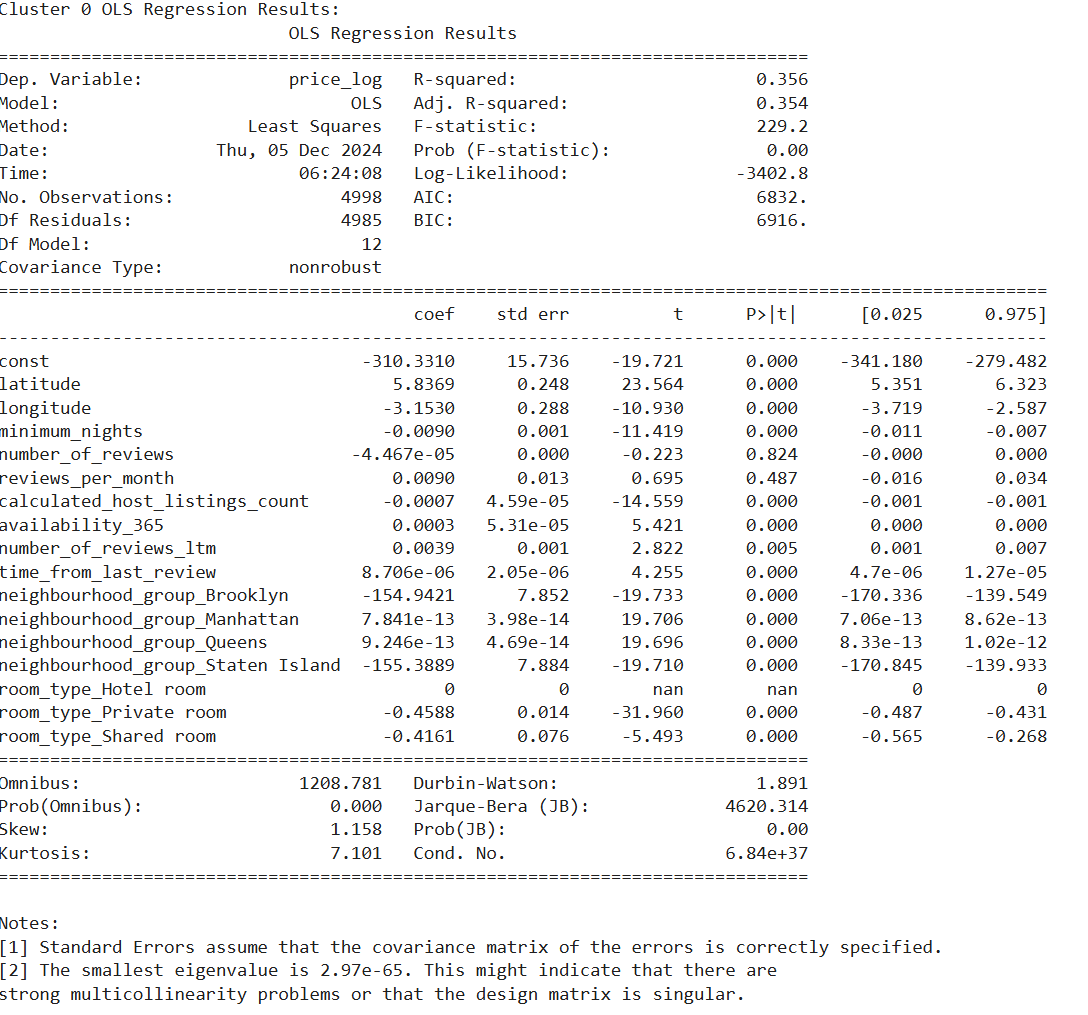


Figure 21. Cluster 0 OLS summary

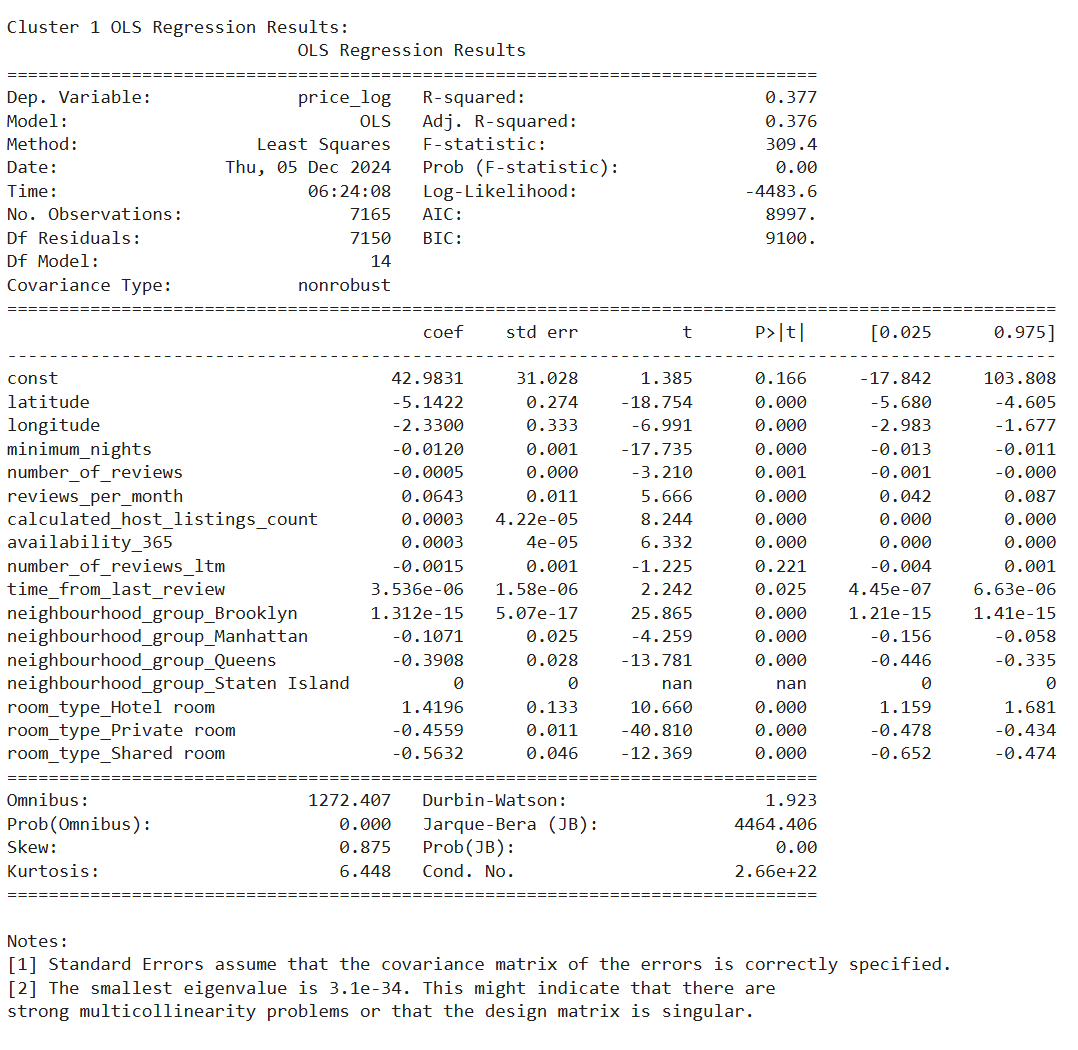


Figure 22. Cluster 1 OLS summary

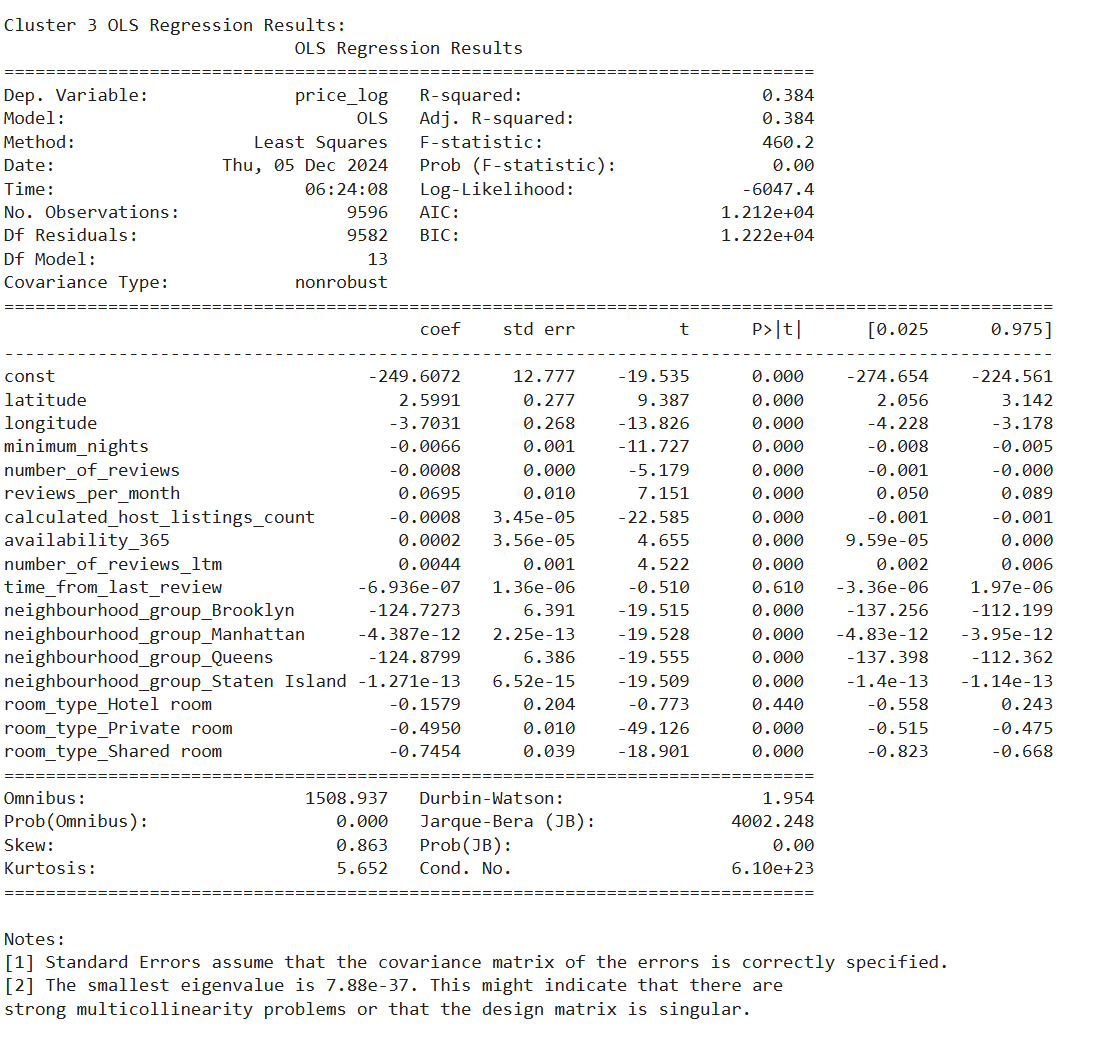


Figure 23. Cluster 3 OLS summary

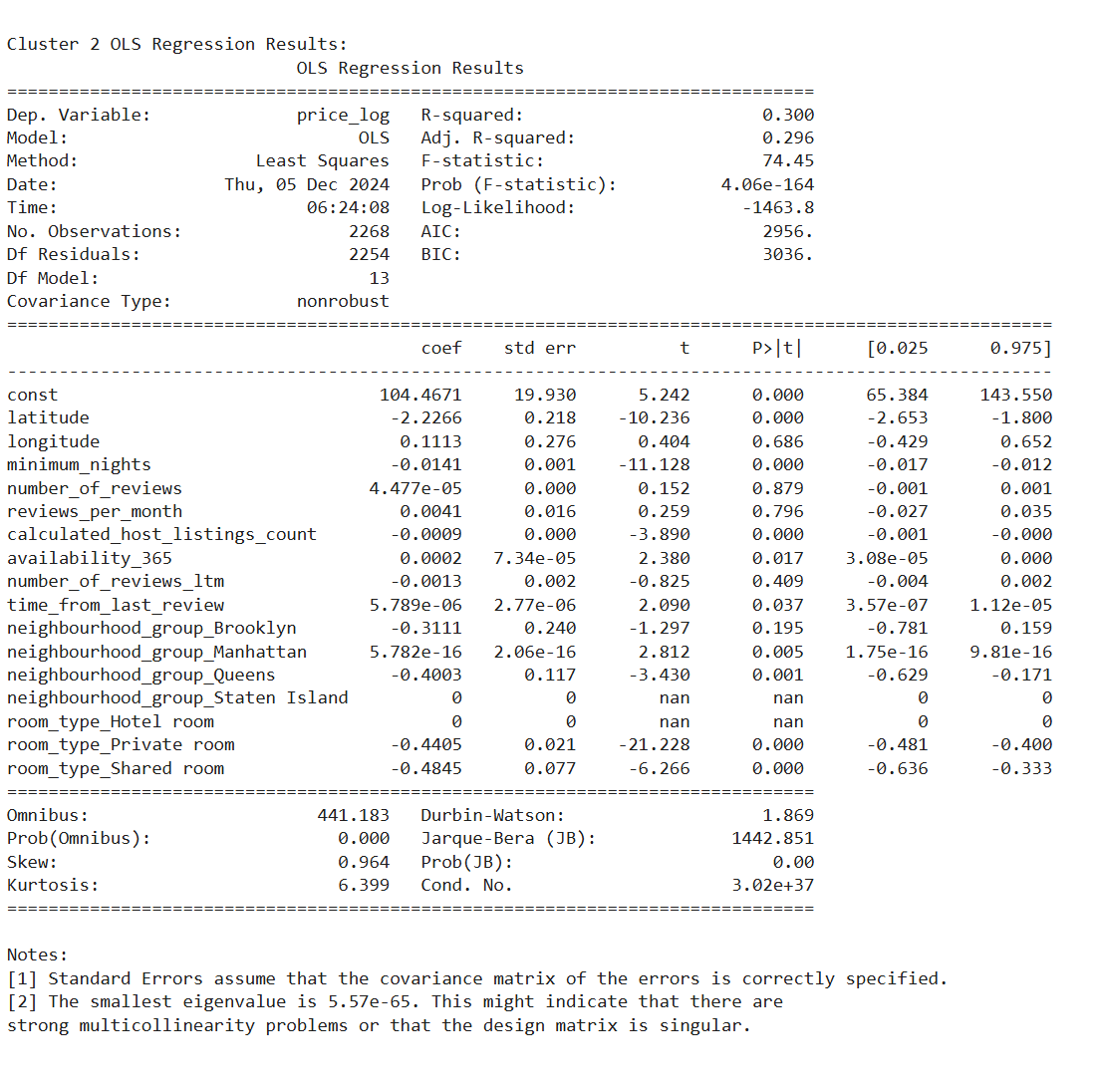


Figure 24. Cluster 2 OLS summary

| **OOB Score** | **MSE** | **R²** |
| --- | --- | --- |
| 0.707559 | 0.122798 | 0.744 |

Figure 25. OOB Results

# **References**

1. Alharbi ZH. A Sustainable Price Prediction Model for Airbnb Listings Using Machine Learning and Sentiment Analysis. *Sustainability*. 2023; 15(17):13159. <https://doi.org/10.3390/su151713159>
2. Forgosh, D., Brandon, U., & Nuru, B. (2018). *Airbnb Clustered Price Ranges*. danielfforgosh.com. /https://danielforgosh.com/papers/AirbnbFinal.pd
3. Lawani, A., Reed, M. R., Mark, T., & Zheng, Y. (2019). Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston. Regional Science and Urban Economics, 75, 22-34.
4. Tang, Y., & McNicholas, P. D. (2017). Clustering Airbnb Reviews. *ArXiv*. <https://arxiv.org/abs/1705.03134>
5. Verma, A., Deepu, S., & Bajaj, P. (2023). Real Estate Price Predictor – A Literature Review. 2023 11th International Conference on Intelligent Systems and Embedded Design (ISED). <https://www.semanticscholar.org/paper/203fdb2f69ca1a9e3fdd93e9b3fab4adb2cdd4cb>
6. 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.

Progress

This is what I have done so far

I have cleaned the data set:

1. Fill in the missing values
   1. Using Imputations (median) for the following columns: price, service fee, minimum nights, availability 365
   2. Use Neighborhood to fill out missing data in Neighbourhood group because they are very likely to be confounding so I decided to use Neihgbourhood group which is easier to interpret
2. Outliers:
   1. There are no outliers because there are no outliers shown in the box plot
3. Drop columns that provide little to no info: license
4. Created a new feature called building\_type\_new from Construction year to use in the full linear regression model:
   1. If the building time is less than 10 years then it is 1
   2. Else it is 0
   3. Fill out the missing values using the mode of the neighborhood because it does not make sense that a building does not have a construction year -> cannot just leave it
5. EDA:
   1. Balance?
      1. Price: it is balanced among both price range and neighbourhood group
      2. Room type:
      3. Neighborhood group: Not balanced -> might have to incorporate interaction terms
   2. Correlation among features
      1. None of the features highly correlated with price
      2. Numbers of reviews and reviews per month are highly correlated -> multicollinearity
6. Multiple Linear Regression Model:

Features used in the first iteration:

categorical\_features = ['host\_identity\_verified', 'neighbourhood group', 'neighbourhood', 'instant\_bookable', 'room type', 'cancellation\_policy', ‘building\_type\_new]

numerical\_features=['minimum nights', 'number of reviews', 'reviews per month', 'calculated host listings count', 'availability 365']

The first iteration was really bad with R^2 ~ 0 and F stat ~0 => I will perform PCA and

Feature Selection using Recursive Elimination method and Elastic Net to select which features should be included in the REGRESSION MODEL

1. The significance of feature selection (doing)
2. K-means clustering (done):

Based on the following evidence, it is concluded that there are no meaningful host clusters in this dataset:

1. Low Sihouette score (0.26 for all the identifiable K’s using Elbow Method)
2. Even with PCA, the silhouette score remains below 0.5
3. The average price between host groups are very similar
4. Even with different set of features, the clusters’ silhouette score remains to be very low
5. With different clustering methods (dbscan), the clusters are not significantly different from each other

=> Exhaustive trial and error effort confirmed that there are no meaningful host clusters

=> Shift focus to interaction terms

* "While host clusters did not show a significant impact on price, interaction terms such as neighbourhood group x room type and availability\_365 x reviews\_per\_month demonstrated significant relationships with price. These findings suggest that specific combinations of host behaviors and listing characteristics drive pricing variations."

### **Next Steps**

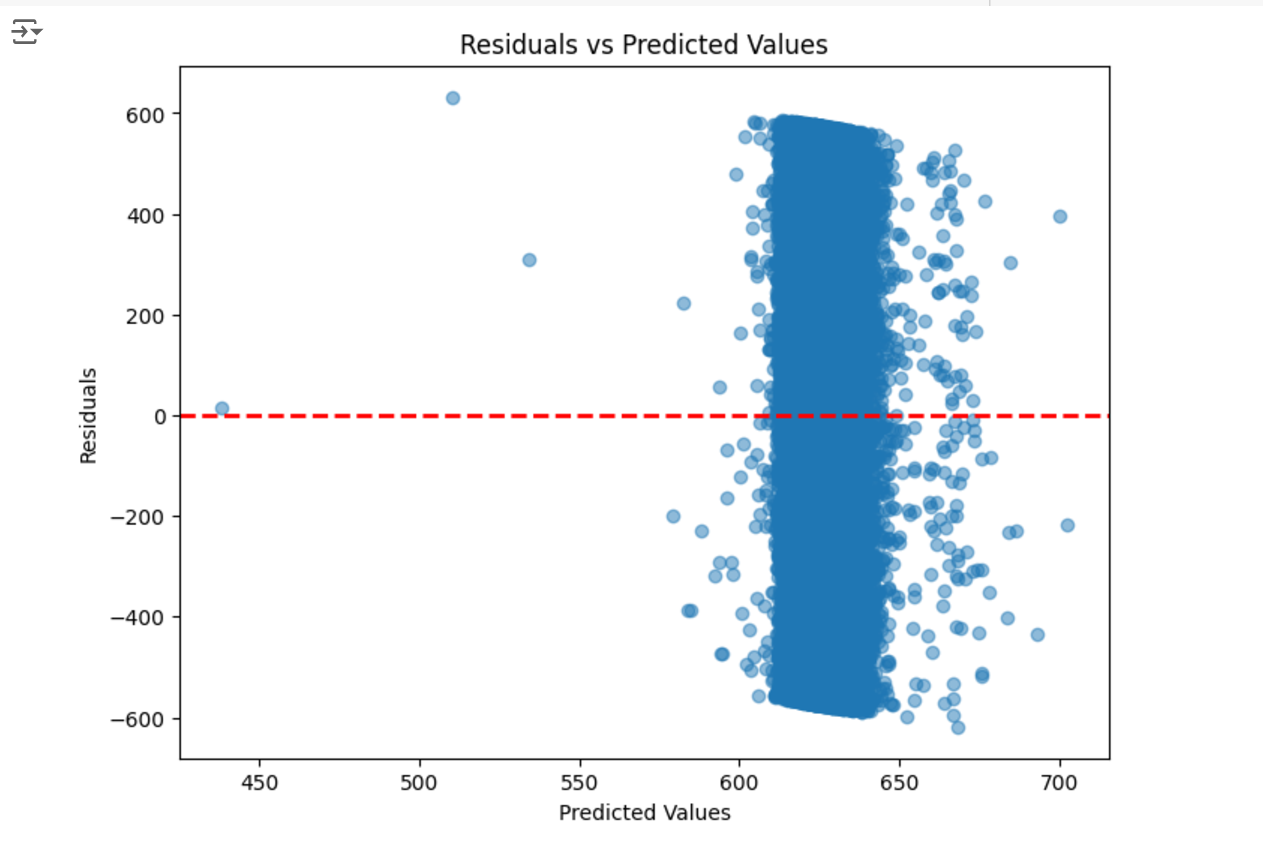
1. **Feature Engineering Refinement:**
   * Explore additional interaction terms beyond room type x neighbourhood group (e.g., availability x reviews or room type x cancellation policy).
   * Consider creating new features, such as luxury index (based on reviews or property descriptions).
2. **Residual Analysis:**
   * Continue to assess residuals for patterns that may indicate missing predictors or model misspecifications.
   * If heteroscedasticity persists, consider weighted regression or transforming target variables (e.g., log-transformed price).
3. **Model Comparison:**
   * Compare Random Forest and Multiple Regression using cross-validation to assess robustness on unseen data.
   * For better interpretability, stick to linear regression for research insights, but use Random Forest if prediction accuracy is the main goal.
4. **Reporting and Interpretation:**
   * Focus your discussion on how interaction terms, clustering, and feature importance analyses provide nuanced insights into pricing.
   * Emphasize that the interplay between room type and location is as important as overall service fees in determining Airbnb prices.
5. **Address Limitations:**
   * Acknowledge that certain features (e.g., luxury amenities) or external factors (e.g., seasonal demand, tourist attractions) may not be captured in your current dataset.

### **Findings and Insights on Airbnb Pricing**

#### **Limitations of Multiple Linear Regression**

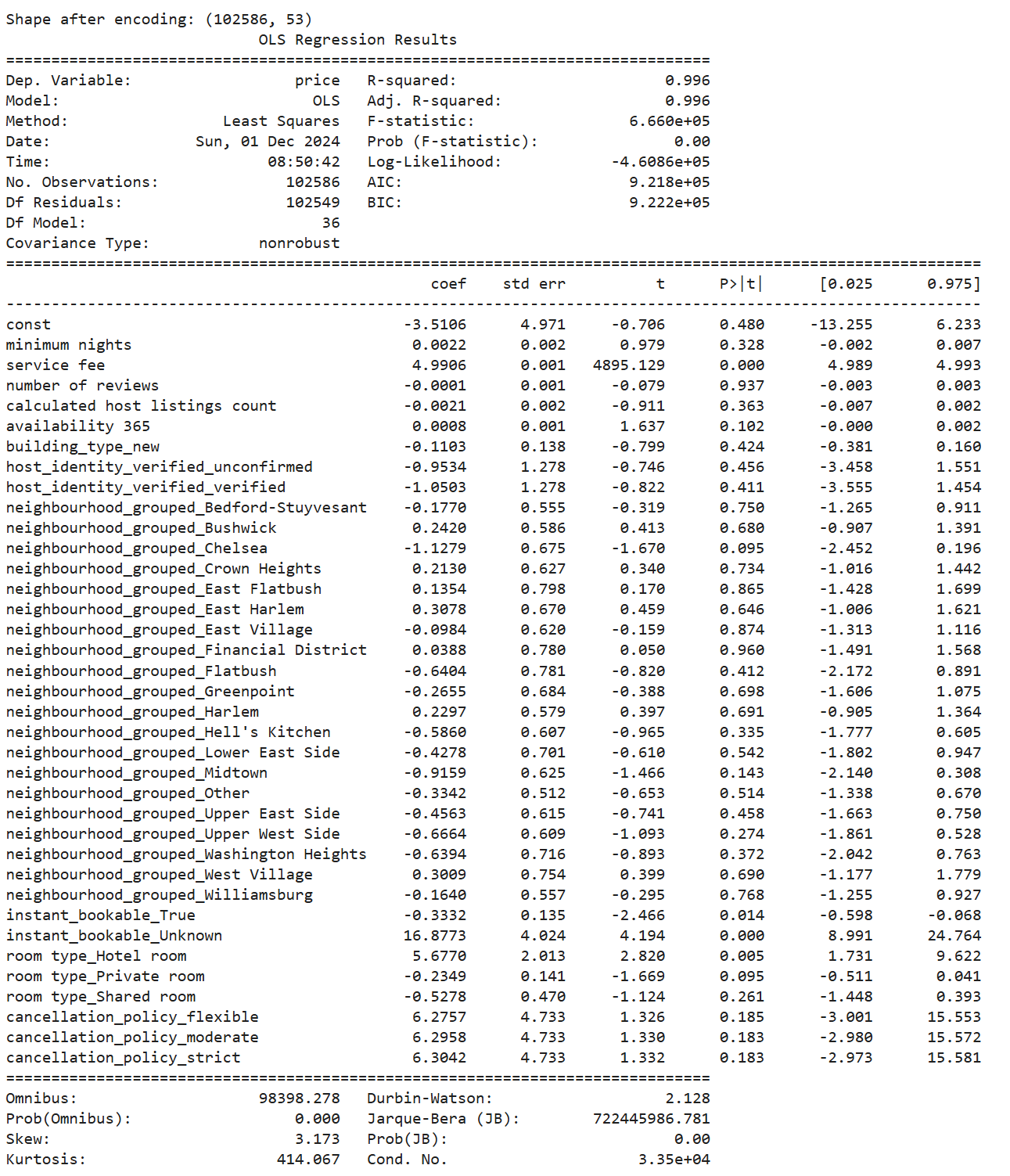
Initially, **multiple linear regression (MLR)** was explored as a potential modeling approach for Airbnb pricing due to its interpretability and simplicity. However, its application faced several critical challenges that rendered it unsuitable for this dataset:

1. **Violation of Assumptions**:
   * The model exhibited **heteroscedasticity**, as residuals showed non-constant variance, violating a key assumption of MLR.

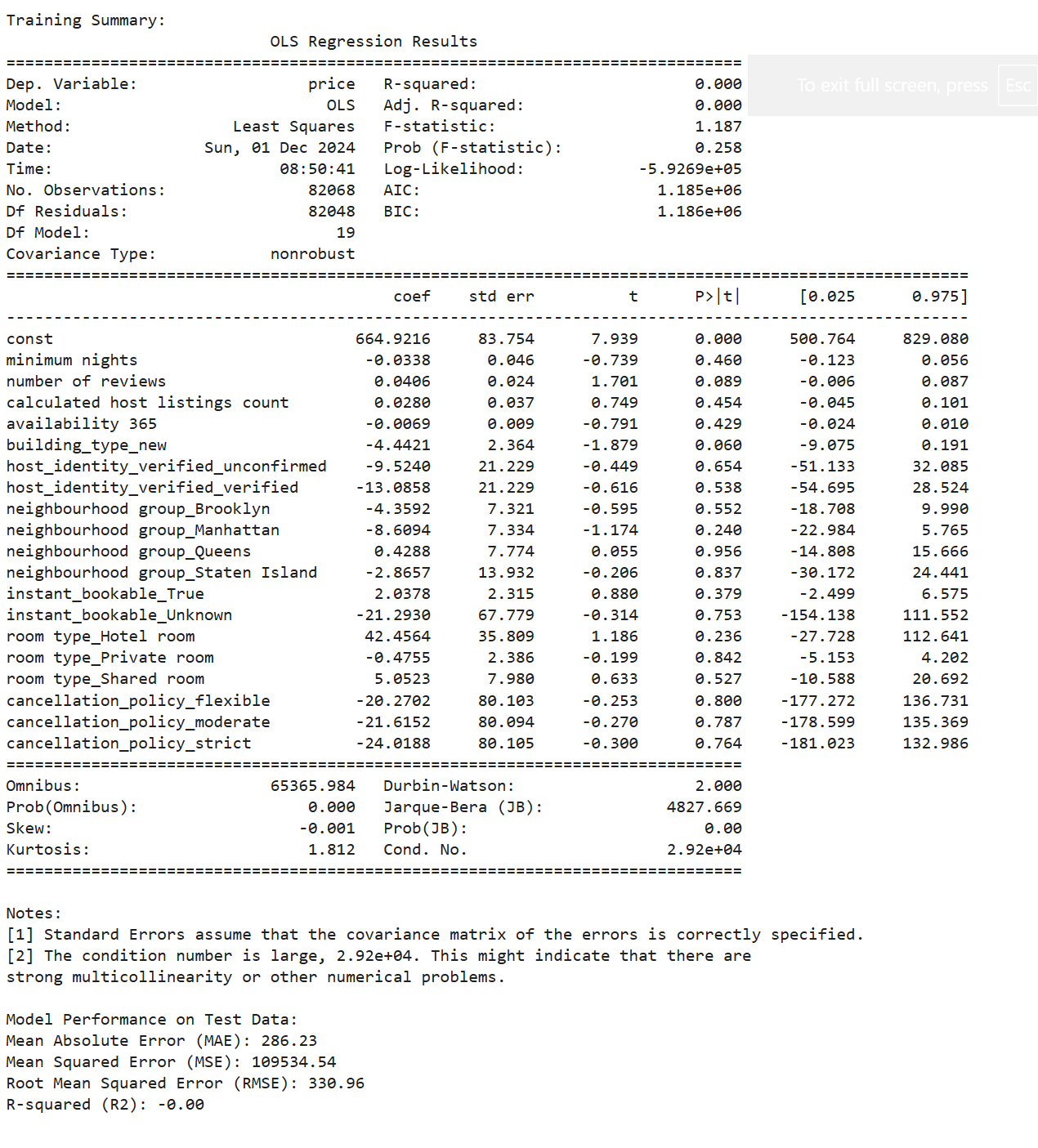


* + **Non-linear relationships** between predictors and price were evident, which MLR struggled to capture effectively.
  + **Multicollinearity** among predictors caused instability in coefficient estimates, further complicating the analysis.

1. **Overwhelming Dependence on Service Fee**:
   * The **service fee** explained nearly all the variance in price (R2=0.996R^2 = 0.996R2=0.996). When removed, the model's explanatory power dropped to near zero, highlighting its inability to utilize other predictors effectively.



1. **Failure Without Service Fee**:
   * Without the service fee, MLR performed poorly (R2≈0R^2 \approx 0R2≈0), emphasizing its limitations in modeling the complex relationships present in the dataset.



Given these limitations, **random forest regression** was chosen as an alternative. It offers the ability to handle non-linear relationships, complex interactions, and multicollinearity without the assumptions required by MLR.

### 

### **Random Forest Model: A Robust Alternative**

The random forest model emerged as a superior choice for modeling Airbnb pricing due to its ability to handle complex relationships and interactions between features, overcoming the limitations of multiple linear regression.

Why Random Forest Was Chosen

1. **Non-Linearity:**
   * Random forests can naturally capture non-linear relationships between predictors and the target variable (price), which multiple linear regression failed to model effectively.
2. **Handling Multicollinearity:**
   * Unlike regression models, random forests are robust to multicollinearity, meaning highly correlated features do not adversely impact model performance.
3. **No Strict Assumptions:**
   * Random forests do not require assumptions about residual distribution, homoscedasticity, or independence, making them ideal for datasets where these conditions are violated.
4. **Feature Importance Analysis:**
   * A key advantage of random forests is their ability to rank features by importance, providing insights into which factors most influence pricing. This enabled the identification of the service fee as the dominant predictor, explaining 99% of the variance in price.

### **Random Forest Model Performance**

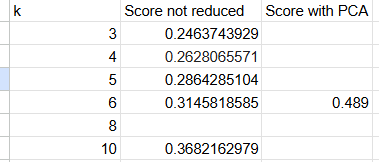
The random forest model demonstrated strong performance compared to multiple linear regression, particularly when interaction terms were included:

1. **Baseline Performance (Without Service Fee)**:
   * **R2=0.1R^2 = 0.1R2=0.1**: When the service fee was excluded, the random forest initially struggled to explain price variance, as the model relied heavily on this feature.
2. **Impact of Interaction Terms**:
   * Adding **interaction terms** (e.g., neighbourhood group x room type) significantly improved the model’s performance, raising R2R^2R2 from **0.1 to 0.6**. This underscores the importance of capturing nuanced relationships between features, which random forests excel at.
3. **Feature Importance Beyond Service Fee**:
   * After removing the service fee, the model identified other significant predictors:
     + **Availability (availability\_365)**: Listings with higher availability had higher pricing potential.
     + **Room Type**: The type of room (e.g., private room, entire home) remained influential, particularly in combination with location.
4. **Comparison with Multiple Linear Regression**:
   * Multiple linear regression failed to perform effectively without the service fee (R2≈0R^2 \approx 0R2≈0), whereas the random forest model leveraged interactions and non-linear patterns to significantly improve predictive accuracy.

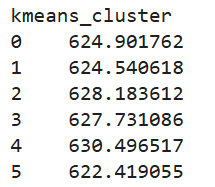
#### **Host Clustering Analysis**

**Host clustering** was thoroughly tested as a potential approach to segment listings into distinct groups based on host-related features. However, the analysis confirmed that clustering did not yield meaningful results:

1. **Low Silhouette Scores**:
   * Across various kvalues (determined via the Elbow Method), the **silhouette score** remained consistently low (0.260.260.26), indicating poor cluster separation.
   * Even after applying **PCA** or testing alternative clustering methods (e.g., DBSCAN), silhouette scores did not exceed 0.5.



1. **Minimal Variance in Average Prices**:
   * The average price across clusters was highly similar, suggesting that the clustering features did not differentiate hosts in ways that impacted pricing.



1. **Exhaustive Testing with Alternative Features**:
   * Testing different combinations of host-related features (e.g., availability, reviews, listings count) and clustering methods failed to produce meaningful groupings.

**Conclusion**: Clustering host-related features did not provide actionable insights or improve the model’s predictive power. Host clusters were therefore excluded from further analysis.

#### **Success with Interaction Terms**

To capture nuanced relationships between features, **interaction terms** were introduced. These terms significantly enhanced the model’s performance:

* Including the interaction term **neighbourhood group x room type** improved the random forest model’s R2R^2R2 from **0.1 to 0.6** (without the service fee).
* Interaction terms allowed the model to account for location-specific demand for different room types, which MLR failed to capture.

Further exploration of additional interaction terms (e.g., **availability x reviews**) is expected to yield even greater insights.

**Comparison of Models**:

* **Random Forest**: Outperformed MLR in handling complex relationships and interactions, explaining more variance in price.
* **MLR**: Unable to account for non-linear relationships or interaction effects, performing poorly without the service fee.

### **Actionable Insights for Hosts**

1. **Leverage Key Determinants of Price**:
   * **Service Fee**: Remains the most influential predictor. Hosts should factor this into pricing to align with competitive trends.
   * **Availability**: Listings with higher annual availability tend to achieve better pricing outcomes. Maximize the availability of listings to improve revenue potential.
   * **Reviews**: Encourage guests to leave reviews to improve listing visibility and perceived value.
2. **Optimize Room Type Based on Location**:
   * The interaction between neighbourhood group and room type has a significant impact on pricing. For instance:
     + Entire homes tend to command premium prices in urban centers like Manhattan.
     + Private rooms may perform better in residential areas such as Brooklyn.
   * Hosts should tailor their offerings to meet location-specific demand.
3. **Adapt Pricing to Local Competition**:
   * Compare pricing and performance with similar listings in the same neighborhood. Adjust prices dynamically to stay competitive.
   * For example, Staten Island listings may benefit from lower pricing strategies, whereas Manhattan hosts can adopt premium pricing.
4. **Enable Instant Booking**:
   * Instant Booking can attract more spontaneous bookings and improve occupancy rates, particularly in high-demand locations.
5. **Capitalize on Seasonality and Events**:
   * Adjust pricing dynamically for peak seasons or during local events to maximize revenue.
6. **Enhance Amenities**:
   * Highlight key amenities such as Wi-Fi, parking, and pet-friendly features, which can significantly influence demand and pricing.

### **Final Summary**

This analysis highlights the limitations of multiple linear regression for modeling Airbnb prices due to its inability to handle non-linear relationships, heteroscedasticity, and reliance on the service fee. While host clustering proved ineffective, interaction terms provided meaningful improvements in model performance, particularly when applied in a random forest model.

By focusing on interaction effects, key determinants of pricing, and actionable insights, hosts can optimize their pricing strategies and enhance revenue potential. Future work may involve testing additional interaction terms and refining dynamic pricing strategies.