Airbnb Listing Analysis

– Prudhvi Chekuri, Sihang Liu, Deepika Reddygari

Abstract

In recent years, Airbnb has become a major player in the hospitality industry, offering travelers a wide range of accommodations beyond traditional hotels and resorts. Airbnb's pricing model is a crucial part of its operation, which adjusts listing rates based on various factors. Our study focuses on key listing attributes such as room type, host attributes, room configuration, amenities, and geographical location, which contribute to the pricing dynamics within the Airbnb ecosystem. We have conducted a comprehensive analysis of these variables to better understand the Airbnb accommodation market. Our research aims to provide valuable insights that can help hosts and travelers make informed decisions to optimize their Airbnb experiences. Our goal is to enrich our understanding of Airbnb's market dynamics and empower Airbnb users to make the most of their stays.

2. Overview of the Project

a) Why did we choose this project?

When searching for datasets, we wanted to find a large one with many columns to apply the knowledge we gained from the class. We came across this dataset with many missing values, variables, and improper presentation of types which might lead to several questions. This is why we chose this dataset - to work with real-world data that is not straightforward to work with.

b) About the Data

We got the data from the official AIRBNB website. We scraped the data for all the provided cities in the USA. But is there a need for scraping? Yes, because we couldn't find a dataset that covered all the listings that were presented on the website. There are a lot of datasets on the internet for only a single city. We did find a dataset on Kaggle that includes all the cities in the USA, but it only has 21 columns, because that dataset was generated from summary data files and not the detailed data files (Which have 76 columns) provided on the website. There are other useful data related to reviews, which can be used to understand the listings and hosts better. But this requires natural language processing and more time, so we just used listings data for our analysis.

We decided to scrape the official website as we don't have a comprehensive dataset for our analysis. We used the 'requests' and 'BeautifulSoup' libraries in Python to extract the links for all the data files related to listings in the USA. Then we used the 'pandas' library to pull the data stored at those URLs. The listing data is provided for 34 cities in the USA, which covers 20 states. Our dataset has around 288K observations and 76 variables.

2.1 SMART Questions

- What features significantly influence the Airbnb listing price?
- Does the location have more impact on the listing price?
- Can we predict the price of a listing based on its latitude, longitude, and other relevant variables?
- How can techniques like Model-based Imputation, outlier detection and removal, and hyperparameter tuning improve the performance of the models?

3. Preprocessing the Data

Preprocessing is a crucial step in data preparation that involves cleaning and transforming raw data into a format suitable for modeling. Here's a brief overview of each aspect:

3.1 Dealing with duplicate records

- Duplicate listings are eliminated to ensure data integrity and prevent skewing of analysis results.
- By employing the drop_duplicates() function, duplicate listings based on the 'id' column are identified and removed from the dataset.
- This ensures that each listing is unique, maintaining data integrity and preventing redundancy in subsequent analyses.
- A starter dataset is constructed by selecting the most relevant columns for initial analysis.

3.2 Data Type Conversion

- By standardizing data types and performing necessary transformations, the dataset is prepared for robust analysis and modeling.
- Categorical variables such as host_response_time, host_is_superhost, and host_identity_verified are converted to numerical representations or categories, to make them suitable for modeling.
- Numerical variables like host_response_rate and host_acceptance_rate are converted from string representations to numerical types, ensuring consistency and enabling mathematical operations.

3.3 Handling Missing Values

- We used the implementation of a simple imputation technique to address missing values in the dataset.
- Numerical variables are imputed with the mean value, while categorical variables are imputed with the mode.
- Before applying imputation, listings with extreme price values (≥ 1200) are removed to enhance the accuracy of the imputation process.

- By applying imputation techniques, the dataset becomes more complete and suitable for subsequent analysis without the need for extensive data loss.

4. Exploratory Data Analysis

We attempted to comprehend the impact of different aspects on Airbnb listing costs as part of our EDA. We started by visualizing the data in order to get a sense of the price and regional distributions.

4.1 Visualization of Price and Geographic Distribution

Using heat maps as our visualization tool, we looked into the spatial impact on price. Heatmaps are an effective tool for displaying complex data in an understandable way so that patterns and trends can be quickly identified.

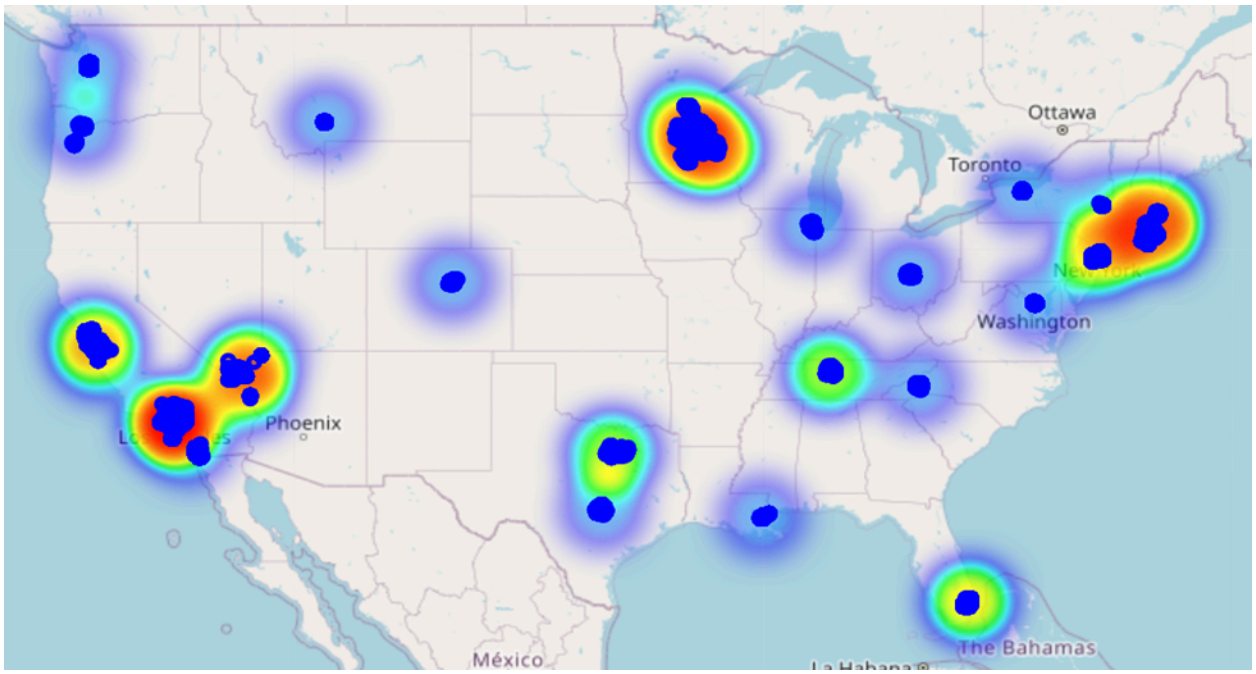


Figure 1. National heatmap of Airbnb listing prices.

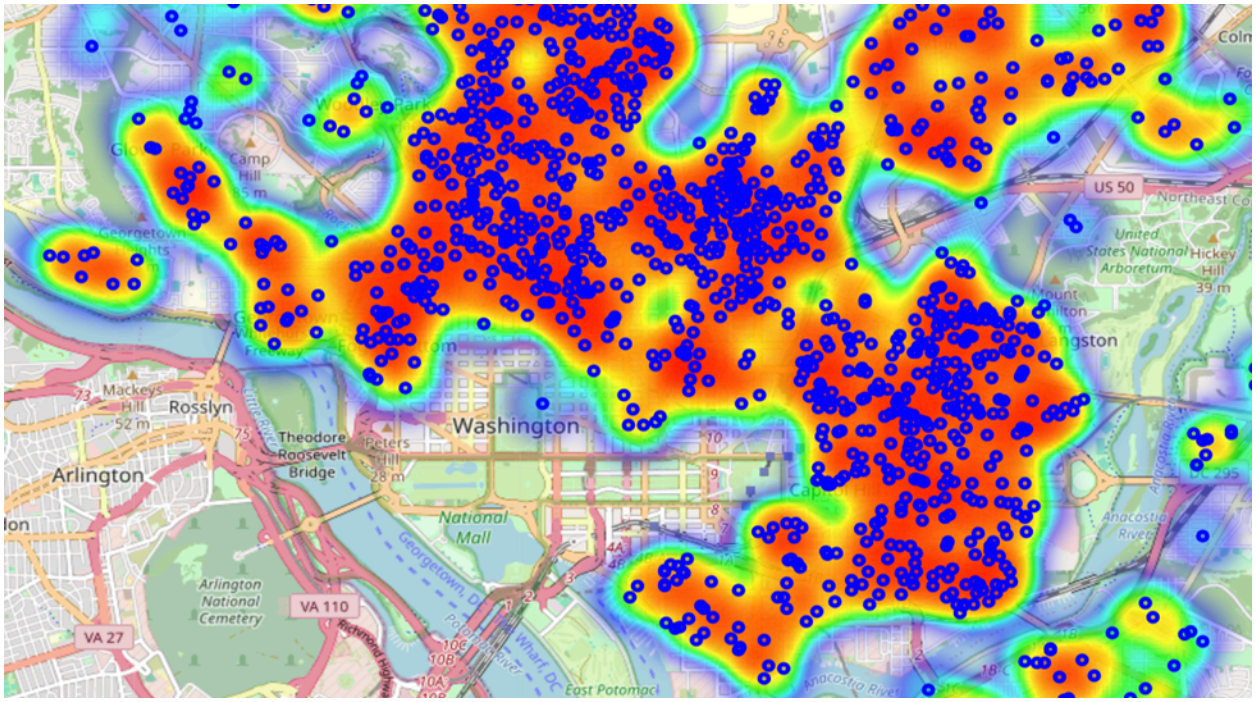


Figure 2. Heatmap of Airbnb listing prices in Washington, D.C.

The first heatmap (Figure 1) provides a macroscopic view of the United States, which shows the relative cost and concentration of Airbnb listings in various urban regions. Cooler hues (blues) denote more reasonably priced offerings, while warmer hues (oranges and reds) imply higher pricing. We can see from this graphic that some metropolitan areas, like Los Angeles, New York, and Washington, D.C., have a dense concentration of more expensive rents.

With a closer look, the second heatmap (Figure 2) provides a view of Washington, D.C. Here, we have a more detailed distribution of listings with price data overlayed throughout the city. Red points are concentrated in these places because they have a higher average listing price, which could indicate that these communities have greater rental prices.

Heatmap analysis demonstrated that, as one might expect, location has a big influence on listing pricing. Higher prices are associated with urban centers and particular neighborhoods within

cities, which is consistent with the real estate maxim "location, location, location." But these images also prompt other queries: What particular aspects of the community drive up prices? Are these places closer to tourist sites, have superior amenities, or are they easier to access?

4.2 Price & Rating Distribution

We continued our investigation into the distribution of prices and ratings for Airbnb listings as part of our exploratory data study. Since these variables frequently represent underlying factors related to customer satisfaction and the economy, it is imperative to understand their dispersion and key patterns

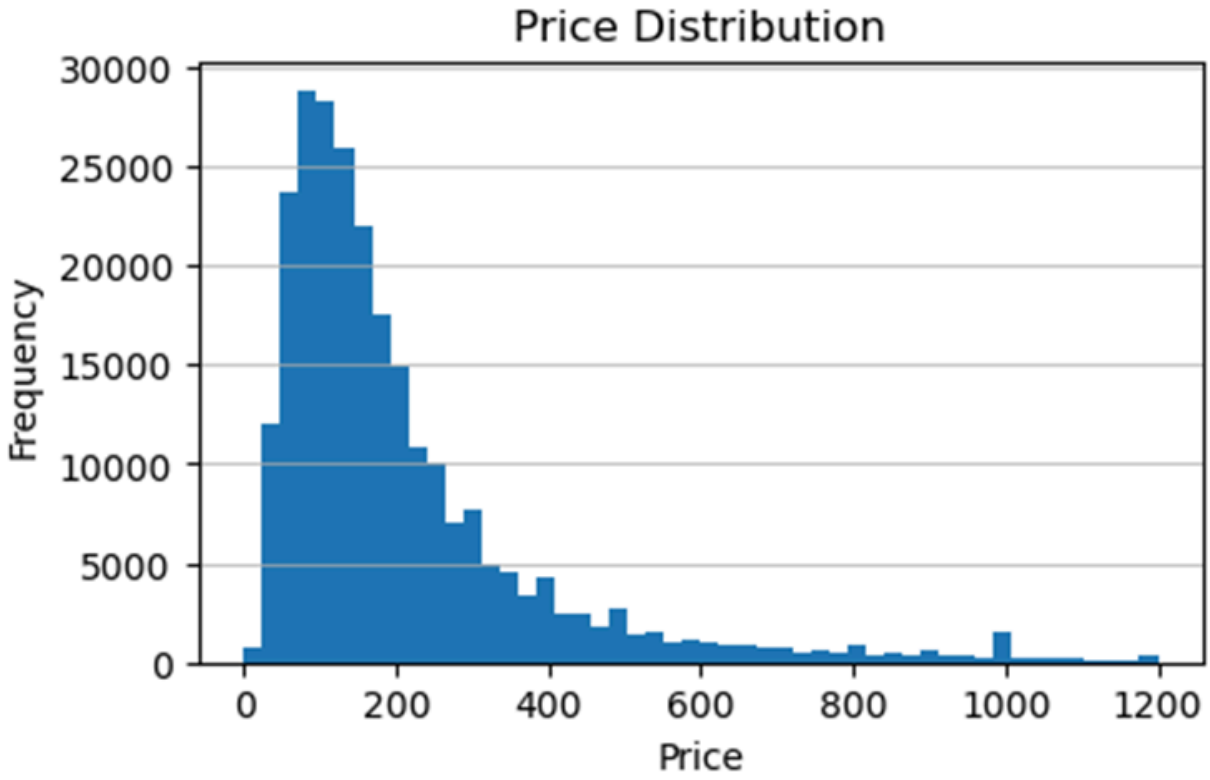


Figure 3. Distribution of Airbnb listing prices.

According to Figure 3, the price distribution of Airbnb listings is right skewed. This skewness suggests that there is a long tail of higher-cost listings even while the majority of listings are priced at the lower end of the spectrum.

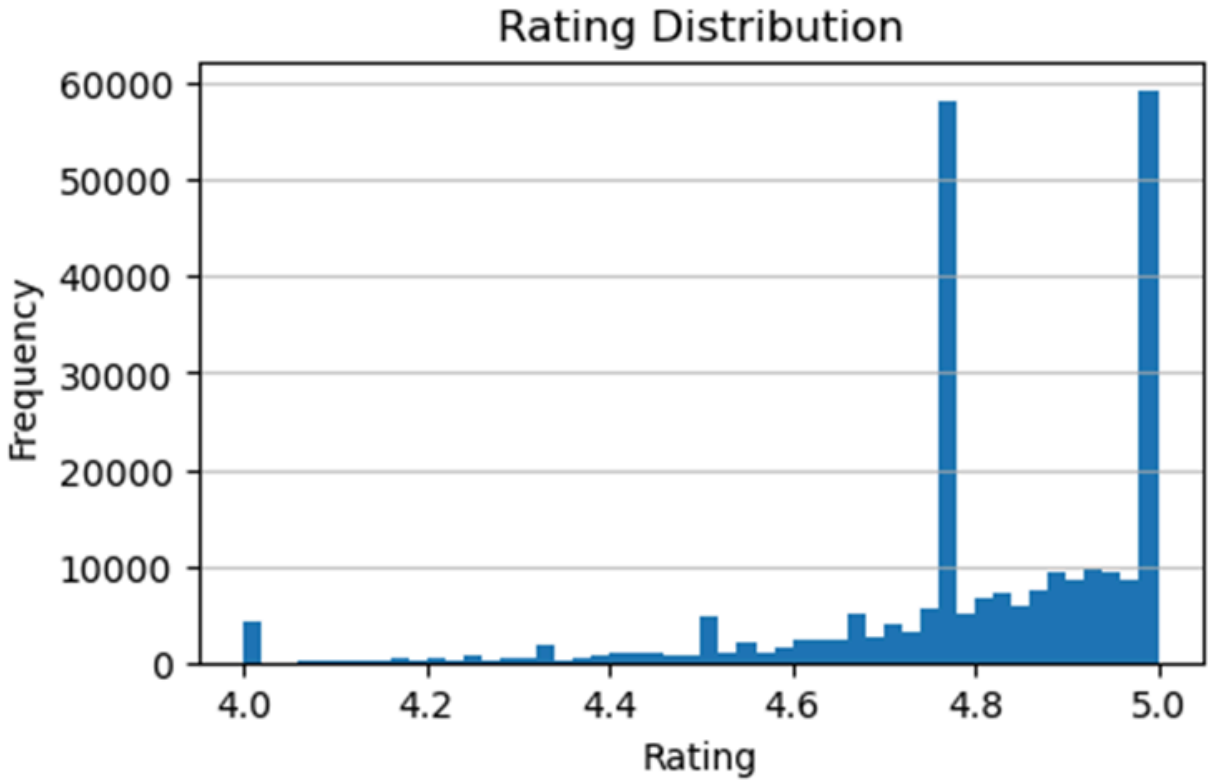


Figure 4. Distribution of Airbnb listing ratings.

The rating distribution for Airbnb listings is seen in Figure 4, where most listings are grouped toward the higher end of the rating scale, suggesting a left-skewed distribution. This preference for higher ratings may be due to the way that Airbnb handles reviews, or it may just be the case that satisfied customers are more inclined to write reviews. There are two separate peaks that may represent default scores that Airbnb's rating algorithm assigns to listings with comparable review statistics.

4.3 Geographical Insights on Airbnb Listings

Location is still a major factor for rentals on Airbnb. In order to identify the spatial economics in operation, we examined the distribution of listings among different cities and areas.

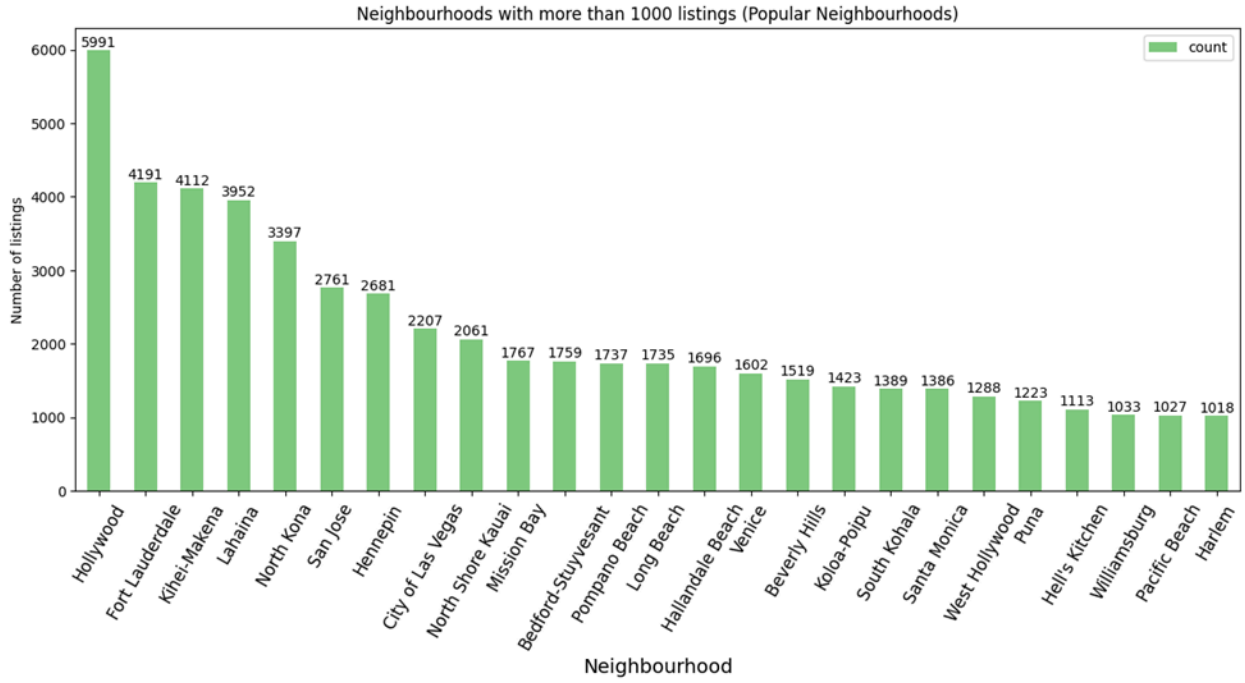


Figure 5. Bar chart showing the number of listings in popular neighborhoods.

The neighborhoods that have more than 1,000 entries, or what we consider to be "popular," are shown in Figure 5. Due to their close proximity to tourist destinations, downtown districts, or other sites of interest, these regions are assumed to have a higher demand. Because of the increased demand brought about by these districts' attractiveness, prices are predicted to rise. Hollywood has the largest number, indicating that it is a popular destination for Airbnb rentals, probably due to its popularity among tourists and its desirable location.

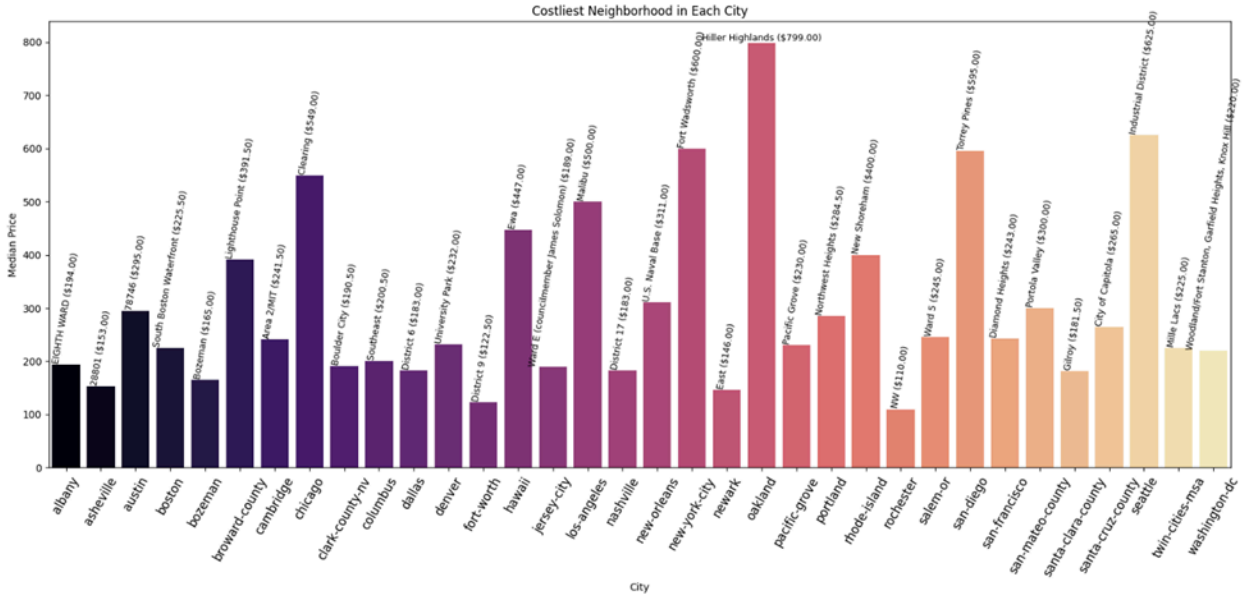


Figure 6. Bar chart illustrating the average price of the costliest neighborhood in each city.

Figure 6 illustrates the most expensive neighborhoods within several metropolitan areas. Cities such as San Francisco, New York, and Oakland have very different peak prices. These differences can be ascribed to local economic conditions, wealth concentration, and tourist attractiveness.

4.4 Numerical Variables Analysis

We divided the variables into numerical and categorical categories and examined their connections to listing prices to determine the factors that influence Airbnb pricing.

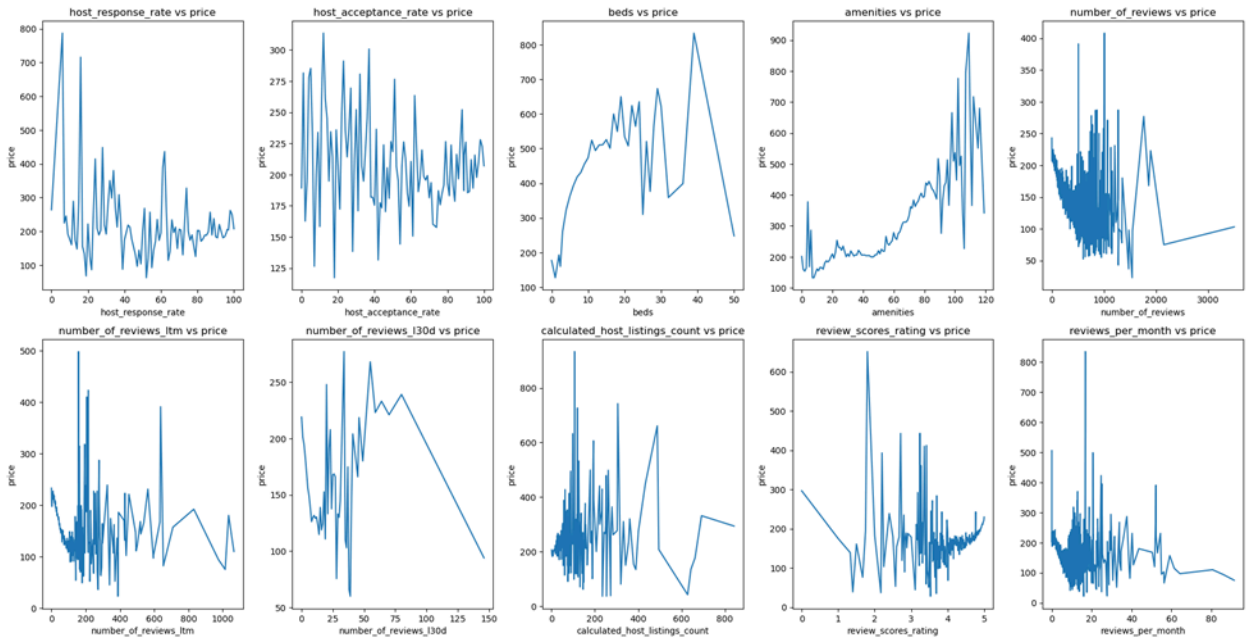


Figure 7. Line Charts for Numerical Variables.

The figures above show the average listing price as a function of multiple numerical variables, such as the number of beds, amenities, host acceptance rate, response rate, and several metrics based on reviews. Because of the presence of outliers, the relationship between these variables and price may appear unclear at first look, but closer inspection reveals some intriguing patterns.

The relationship between the selling price and the facilities and number of bedrooms is particularly noteworthy. Both have a positive link within a specific range: there is a clear upward trend in pricing for listings with 0 to 20 beds and 5 to 80 amenities. This implies that prices rise in proportion to the number of beds and facilities offered, which is consistent with the notion that larger, better-equipped listings fetch greater prices. According to the charts, this upward tendency continues until the data reaches what might be referred to as "extreme" numbers, at which point it becomes irregular and sparse. We concentrate on this interval because the majority of listings have a number of beds and amenities that fall within the defined range. The variation seen in

listings outside of these parameters could be the result of special or unusual properties that don't accurately reflect the overall market.

As long as outliers are taken into consideration and their effects are lessened, the positive patterns that have been discovered indicate that the number of beds and facilities are reliable indicators of Airbnb listing prices. These results are crucial for both hosts and guests in helping them determine the worth of possible rentals and in helping them set competitive and market-aligned prices.

4.5 Categorical Variables Analysis

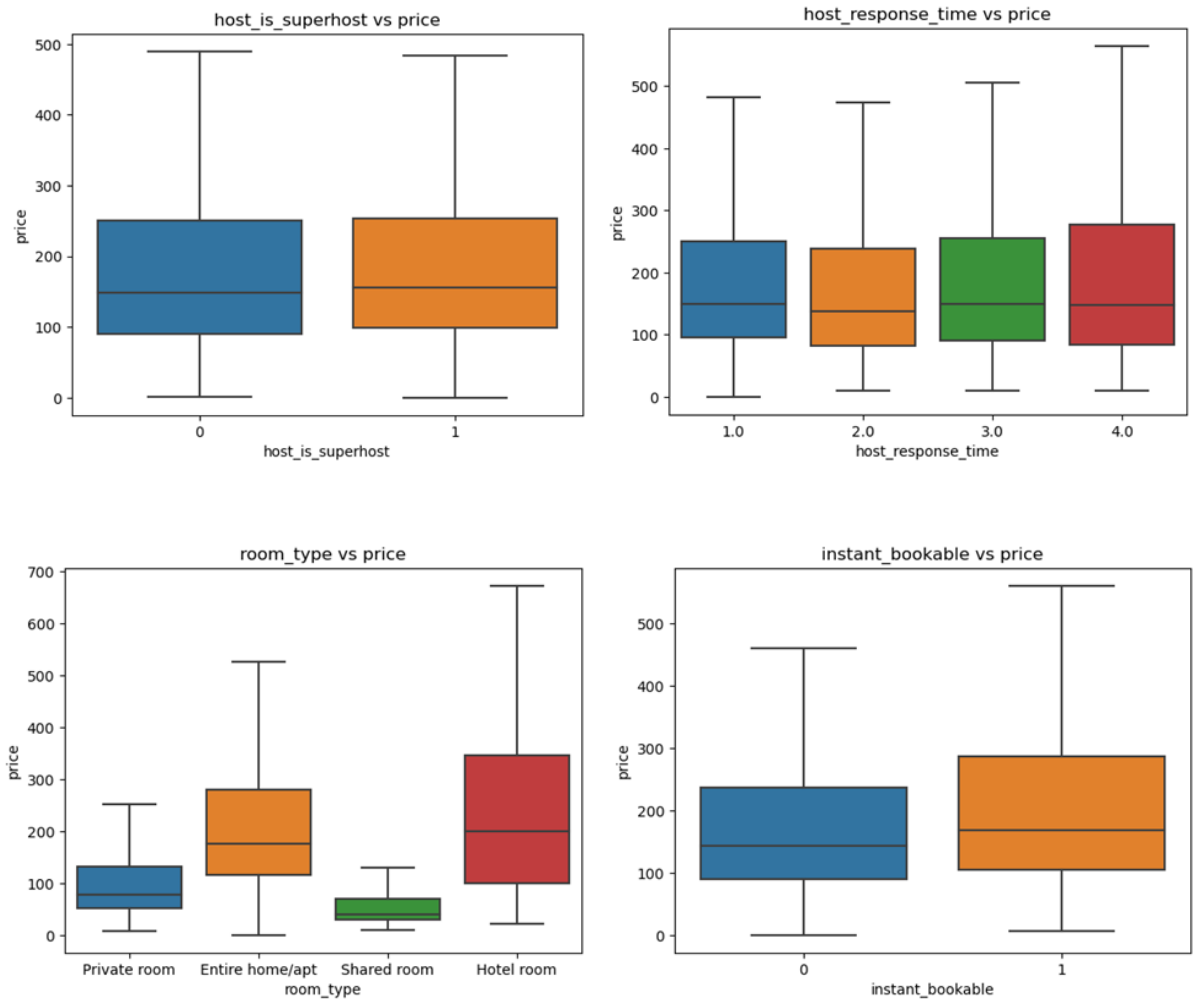


Figure 8. Boxplot of price based on superhost status.

Figure 9. Boxplot of price based on host response time.

Figure 10. Boxplot of price based on room type.

Figure 11. Boxplot of price based on instant bookability.

The boxplot in Figure 8 illustrates the distribution of prices based on whether the host has been designated as a 'superhost'. There is no discernible pricing difference between superhost and non-superhost listings, despite the assumption that superhost designation would attract higher costs due to perceived quality and service.

In Figure 9, the host response time that is ranked from 1 (fastest) to 4 (slowest), also shows no discernible pattern with price. This implies that timeliness has little bearing on pricing, even though it could be crucial for customer pleasure.

By contrast, as Figure 10 illustrates, room style has a significant impact on cost. Compared to private or shared rooms, hotel rooms, and complete homes/apartments are significantly more expensive. This conclusion makes sense—private lodgings can usually be more expensive since they provide more space, facilities, and solitude.

Finally, Figure 11 compares instantly bookable and non-instantly bookable listings, which shows a marginal price rise for the latter type of offering. This might be an indication of how important it is to visitors' convenience to be able to reserve a rental without having to wait for host clearance.

4.6 Property Type on Listing Prices

Our investigation of the impact of property type on listing prices was a crucial component of our

EDA. Pricing can be greatly influenced by the variety of experiences, amenities, and spaces that different property types offer.

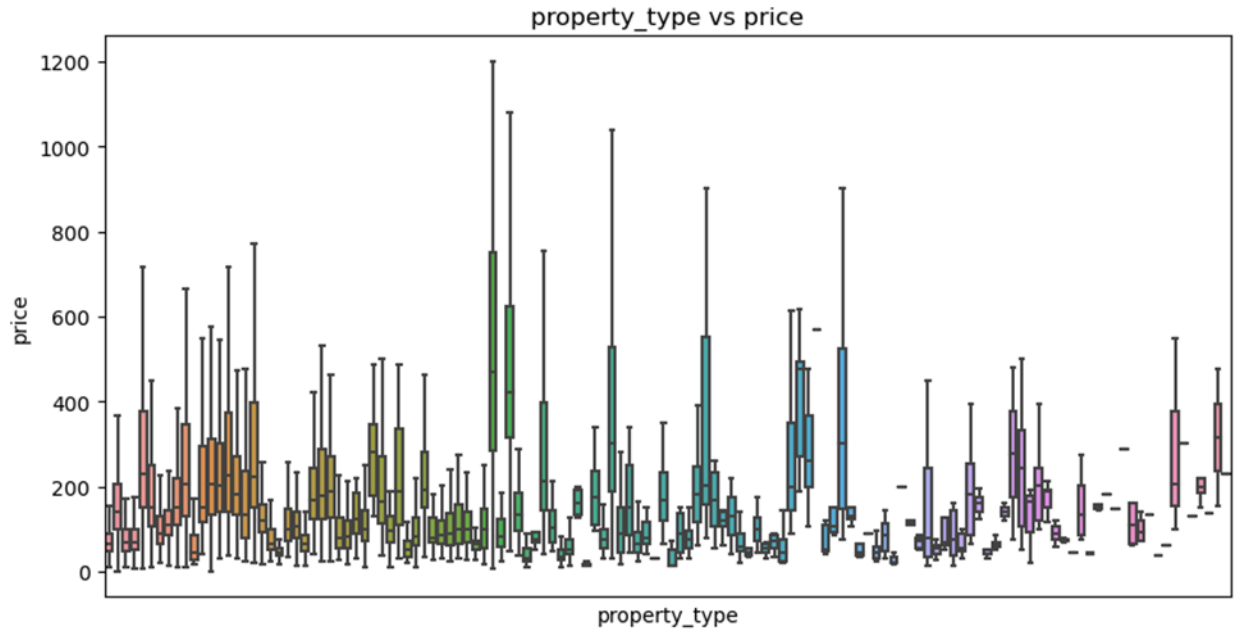


Figure 12. Boxplot of price distribution across different property types.

In the figure above, the boxplot examines the range of pricing for various property kinds. The graphic depiction explains the significant variance in cost related to every kind of property. Certain property categories have larger medians and broad interquartile ranges, indicating that they can fetch higher prices—possibly because they are special or luxurious. Because there are so many categories, examining different property kinds can be complicated. Although certain property types undoubtedly have higher price points, the overlap in price ranges and the existence of outliers make the precise impact of each on pricing less obvious. Including this variable in predictive modeling could improve the accuracy of price estimates because of the variety of property kinds and their corresponding price points. The large number of categories, however, poses a problem for the construction of the model and may need the combination of

different property types into larger categories or the use of sophisticated modeling methods that can handle high cardinality categorical data. It will take more model testing to fully utilize this variable's predictive potential in a pricing model.

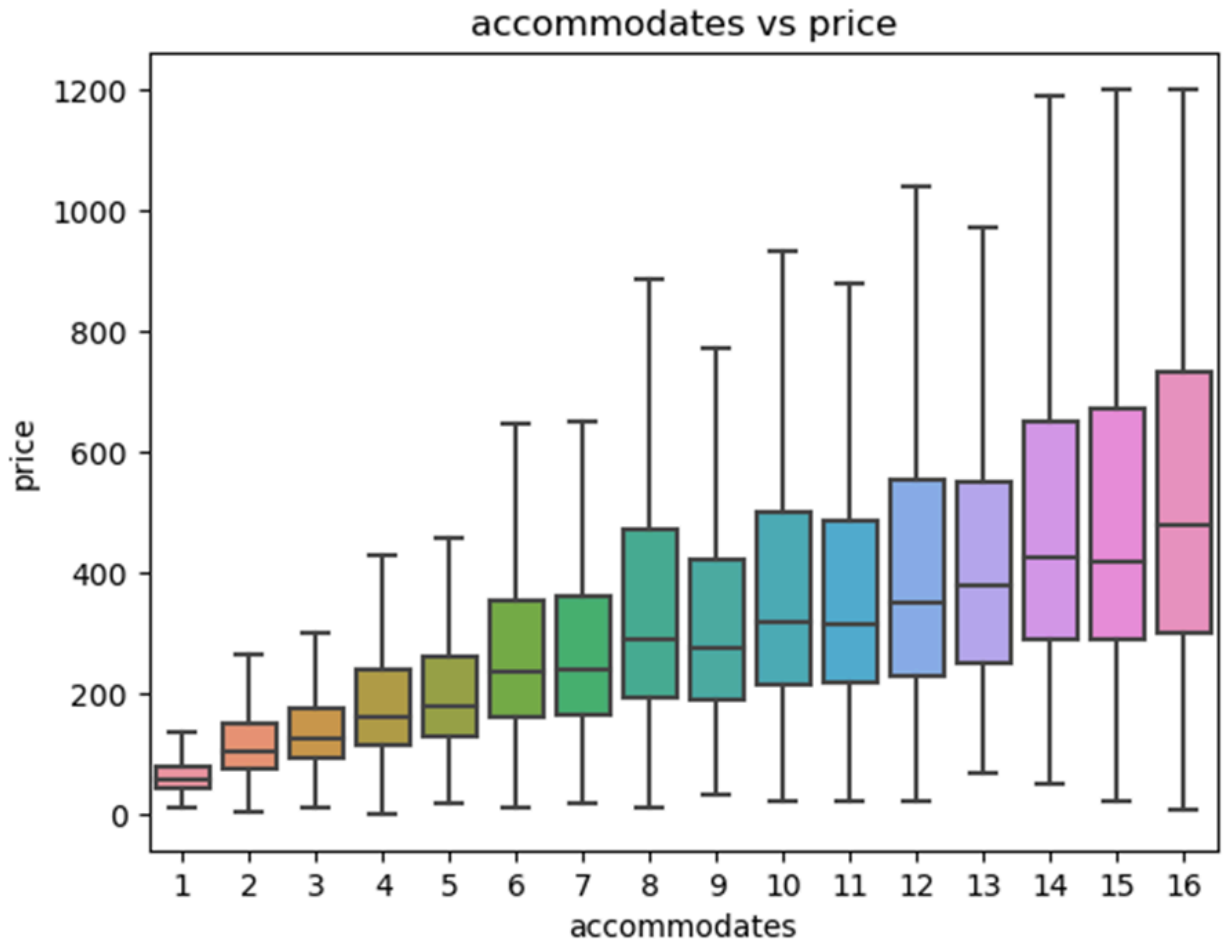


Figure 13. Boxplot of price distribution across different values of accommodates variable.

This boxplot analyzed how the quantity of accommodations affects pricing. Larger properties are likely to have higher costs as the third quartile of prices rises in tandem with an increase in accommodation capacity. It's interesting to note that the first quartile is steady across the range, suggesting that there might be a starting pricing for Airbnb listings. Although it seems that the

median price increases with the size of the lodging, more statistical research is required to determine the rate of this increase and its implications for pricing strategy.

4.7 Geographical Variables

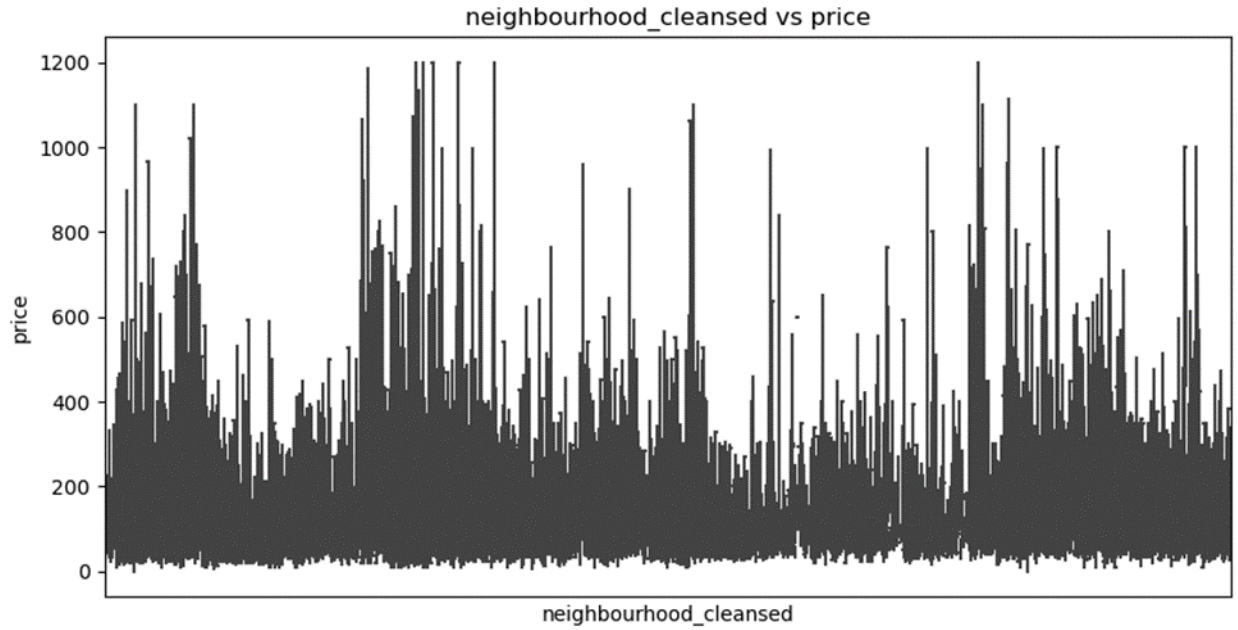


Figure 14. Boxplot of price distribution across different neighborhoods.

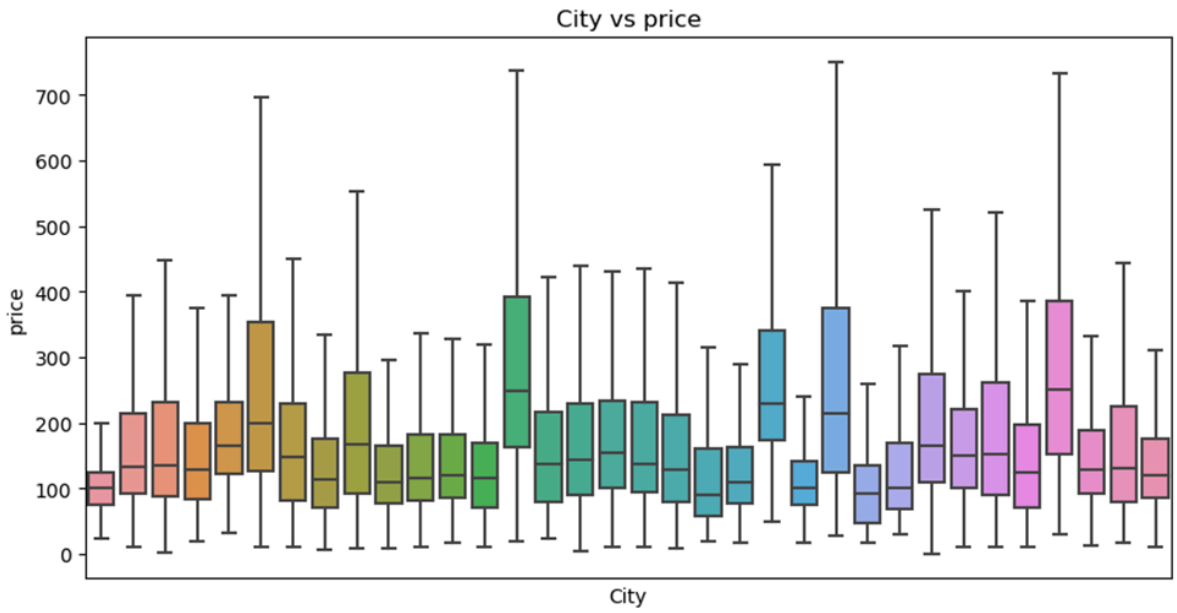


Figure 15. Boxplot of price distribution across different cities.

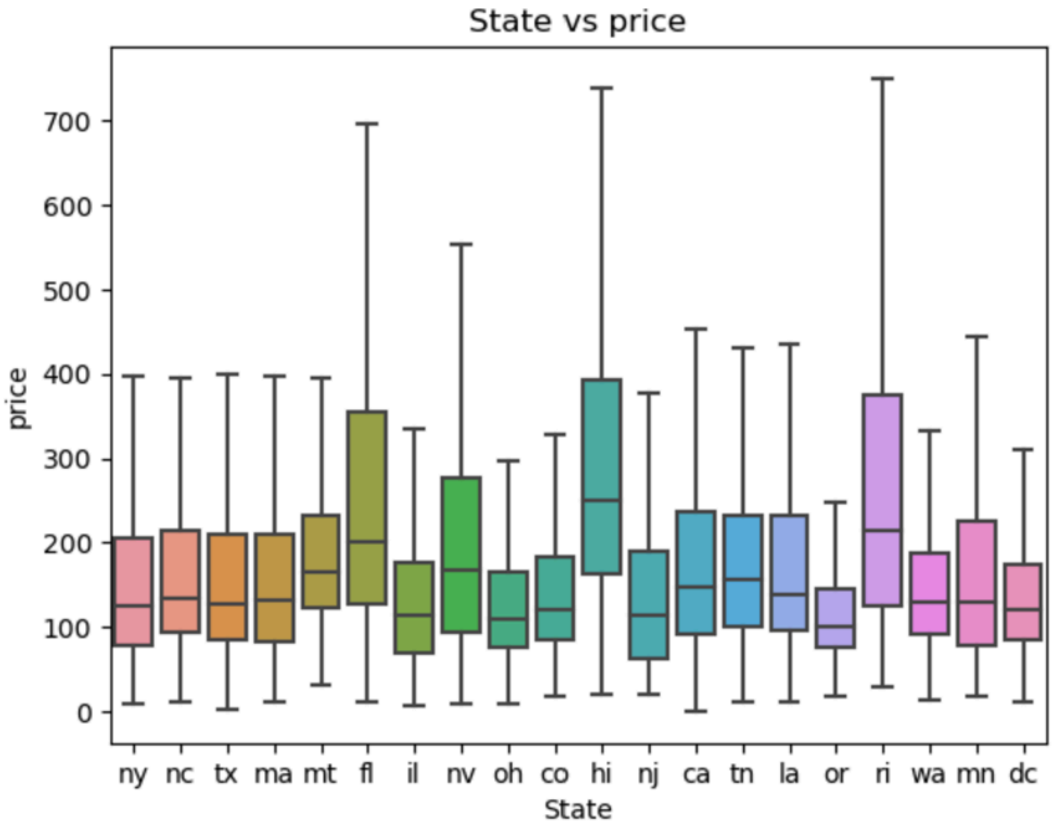


Figure 16. Boxplot of price distribution across different states.

The three boxplots above compared prices across various geographical variables illustrate a wide range of median values, which indicates that location plays a major factor in pricing. The fact that there are outliers in almost every neighborhood/city/state indicates that some postings do not follow the general trends in pricing. These anomalies can be expensive luxury products or exclusive lodgings that are in high demand. We can deduce from comparing data at the neighborhood, city, and state levels that, in some states, some cities or neighborhoods may have a substantial influence on average pricing because of their notoriety, tourist attractions, or commercial centers. However, other locations in the same state can have more affordable options available, adding to the overall variation in cost.

4.8 Correlation Matrix

We also need to look at the correlation heat map to identify any possible correlation between variables.

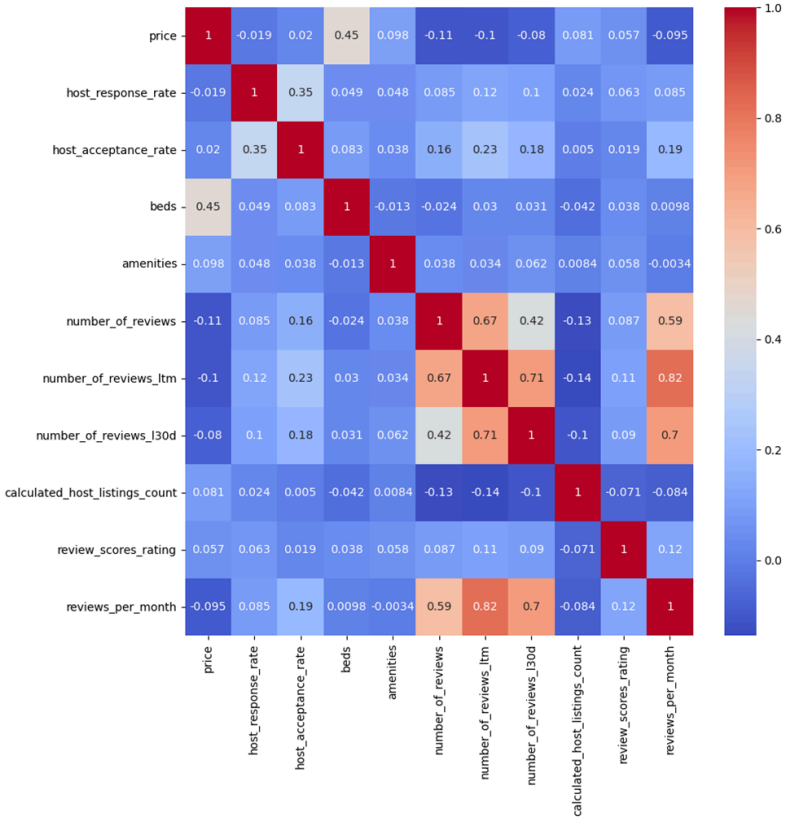


Figure 17. Heatmap of the correlation matrix.

It is obvious that the number of reviews, number of reviews per month, number of reviews each term, and number of reviews in 130 days are highly correlated. This result is in line with reasonable assumptions. Properties having a large total number of reviews are probably going to continue having reviews for an extended period of time. Regularly evaluated properties indicate a higher turnover rate, which could be a sign of increased popularity or a longer duration on the site.

The absence of other notable correlations suggests that a wide range of complicated elements that are not fully captured by the variables taken into consideration influence Airbnb's pricing approach. The heatmap indicates that a variety of factors, including location, property type, seasonal demand, and even less tangible elements like host hospitality and experience uniqueness, could contribute to the fluctuation in price. Possible relationships may be less visible due to the significant price variance. When two factors are taken into consideration at a time, it's possible that more complex linkages exist but are hidden.

4.9 Feature Selection

We used the information gain criterion to assess the feature importance of different variables in order to improve our predictive modeling. From the figure below, geographical coordinates were found to be the most significant predictors. They are particularly important for comprehending price fluctuations. While this feature importance plot provides an initial guide for selecting variables for modeling, it is imperative to verify each variable's true impact on the model's performance. Notably, when paired with other variables, variables that don't seem to be as important separately can nevertheless have a big predictive power. Therefore we need to apply

these variables in models to truly see the impact of variables on model accuracy.

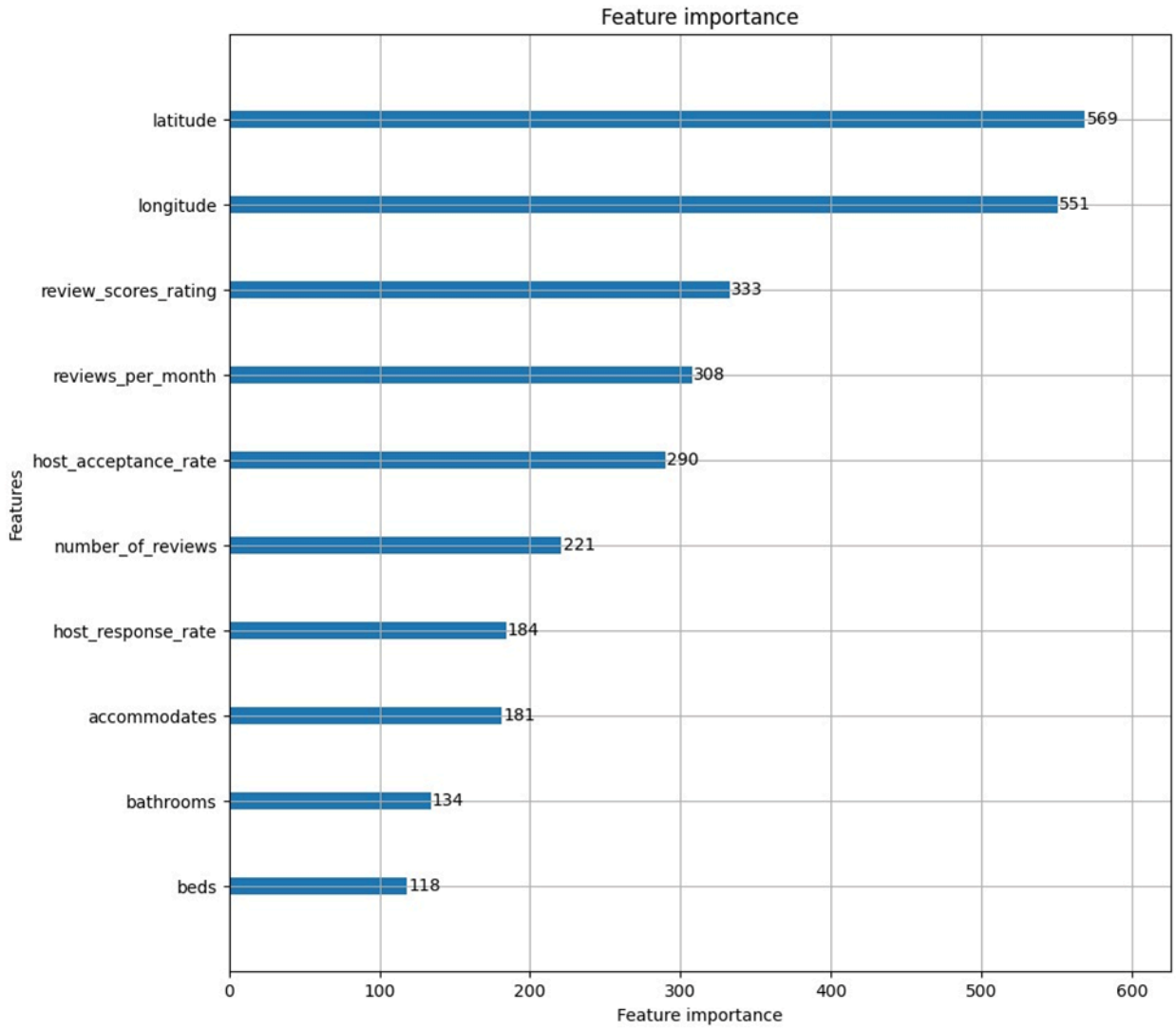


Figure 18. Information Gain

5. Modeling using Sklearn

- The variables we used for modeling are host_response_rate, host_acceptance_rate, latitude, longitude, accommodates, beds, number_of_reviews, number_of_reviews_ltm,

review_scores_rating, reviews_per_month, bathrooms, price. Where price is the target variable and all other variables are predictor variables.

- We've used the Mean Absolute Error and Root Mean Squared Error for evaluating our models.
- We have used train and test splits to evaluate the performance of all the models

5.1 Baseline Modeling

5.1.1 Linear Regression

Our initial attempt was to create a benchmark using the Linear Regression model. However, this model did not produce good results as the data has a complex, nonlinear pattern.

```
Results for LinearRegression:
Experiment: Linear Regression
Mean Absolute Error for Training data: 96.98994406998825
Root Mean Squared Error for Training data: 150.98518074295237
Mean Absolute Error for Test data: 96.52169943401792
Root Mean Squared Error for Test data: 151.53116914903325
```

Based on the metric values shown in the image, it is clear that the model is underfitting. As the metric values are higher, we have decided to increase the model's complexity instead of tuning the model.

5.1.2 Decision Tree

To increase the complexity of the prediction algorithm and capture the non-linearity in our data, we then used the Decision Tree algorithm.

```
Results for DecisionTreeRegressor:
Experiment: Decision Tree Regression
Mean Absolute Error for Training data: 0.48209903668818554
Root Mean Squared Error for Training data: 7.0477477638819535
Mean Absolute Error for Test data: 85.74403584595717
Root Mean Squared Error for Test data: 154.4938080593782
```

According to the results, we can infer that the model performs very well on the training data. However, when it comes to the test set, the model's performance is poor. This suggests that the model is overfitting on the training data.

To create a balance between the train and test performance, we regularized the model by changing the default parameters of the model like max_depth, min_samples_split, min_samples_leaf (Just hand-picked values based on their significance). This is the result of the model:

```
Experiment: Decision Tree Regression Tuned
Mean Absolute Error for Training data: 68.42112964693702
Root Mean Squared Error for Training data: 113.00513635766617
Mean Absolute Error for Test data: 75.92046078309431
Root Mean Squared Error for Test data: 127.13960401631445
```

After a bit of regularization, we were able to make the model generalizable. The model performance is balanced between train and test data. The scores are still not that good, so we decided to explore ensemble models.

5.1.3 Random Forest

The first ensemble model we tried is the Random Forest Regressor model. First, we trained the model with default parameters. Results:

```
Results for RandomForestRegressor:
Experiment: Random Forest Regressor
Mean Absolute Error for Training data: 24.699564814940146
Root Mean Squared Error for Training data: 42.55617137166669
Mean Absolute Error for Test data: 65.21986164782653
Root Mean Squared Error for Test data: 111.57220447517444
```

Looks like the model is overfitting a bit on the training data. This could be caused by the default value of the num_estimators parameter, which is set to 100, as well as other significant parameters like max_depth, min_samples_split, and min_samples_leaf that determine the complexity of the model.

For generalization, again we change the same set of parameters, and this is the result:

```
Results for RandomForestRegressor:
Experiment: Random Forest Regressor Tuned
Mean Absolute Error for Training data: 66.23799366354145
Root Mean Squared Error for Training data: 109.81436916197137
Mean Absolute Error for Test data: 71.5981232807644
Root Mean Squared Error for Test data: 119.5458941777922
```

Random Forest Regressor is taking more time than usual for training on our data with just 100 estimators. So we explored gradient-boosting frameworks like XGBoost and LightGBM.

5.1.4 XGBoost

We trained the XGBoost Regressor with default parameters, and it outperformed the tuned Random Forest model while taking less time to train. Result:

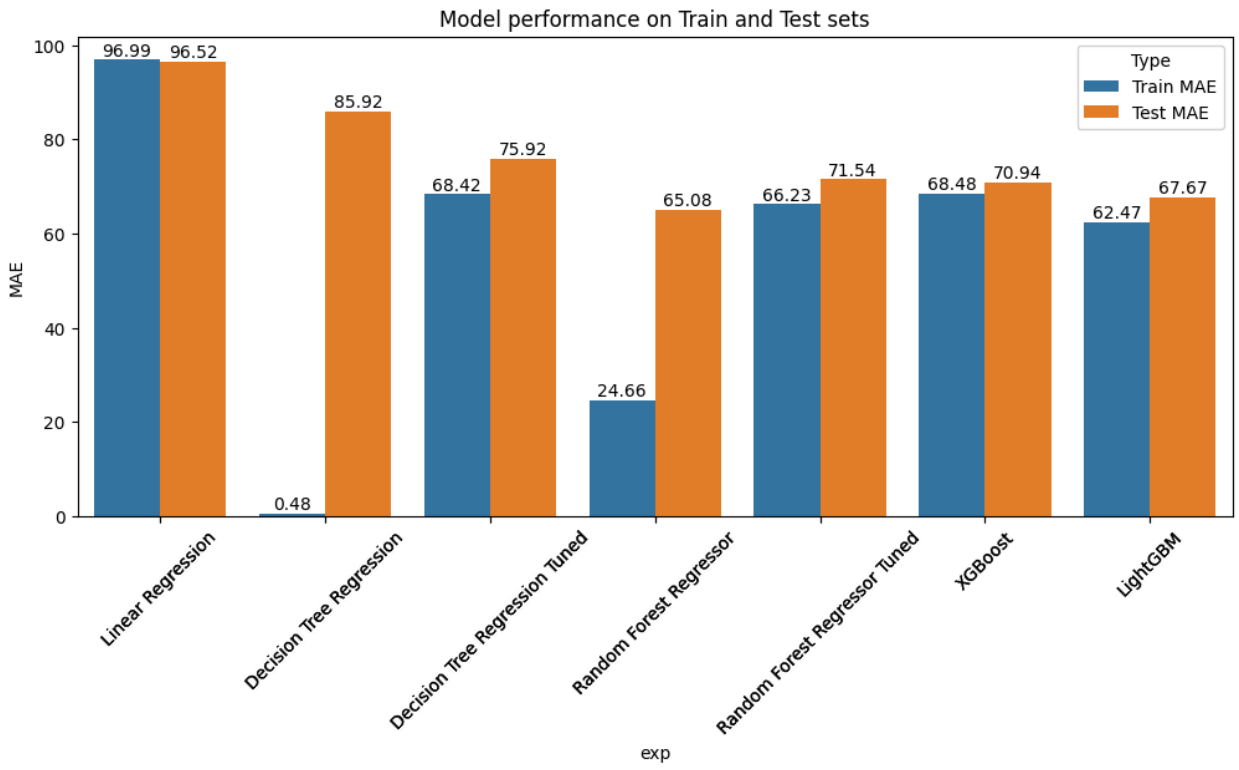
```
Results for XGBRegressor:
Experiment: XGBoost
Mean Absolute Error for Training data: 68.48331940377535
Root Mean Squared Error for Training data: 111.82330613427105
Mean Absolute Error for Test data: 70.94348557915295
Root Mean Squared Error for Test data: 117.32826736524217
```

5.1.5 LightGBM

We then tried LightGBM Regressor with its default parameters. The results show that the model outperformed all the previous models. The performance is balanced between both train and test data.

```
Experiment: LightGBM
Mean Absolute Error for Training data: 62.47449897109193
Root Mean Squared Error for Training data: 101.27313024706014
Mean Absolute Error for Test data: 67.66559344076553
Root Mean Squared Error for Test data: 112.34979929177572
```

Comparing all the baseline models



In the baseline model experiments, we learned how these different algorithms work and found how the gradient-boosting concept improves the performance of the price prediction.

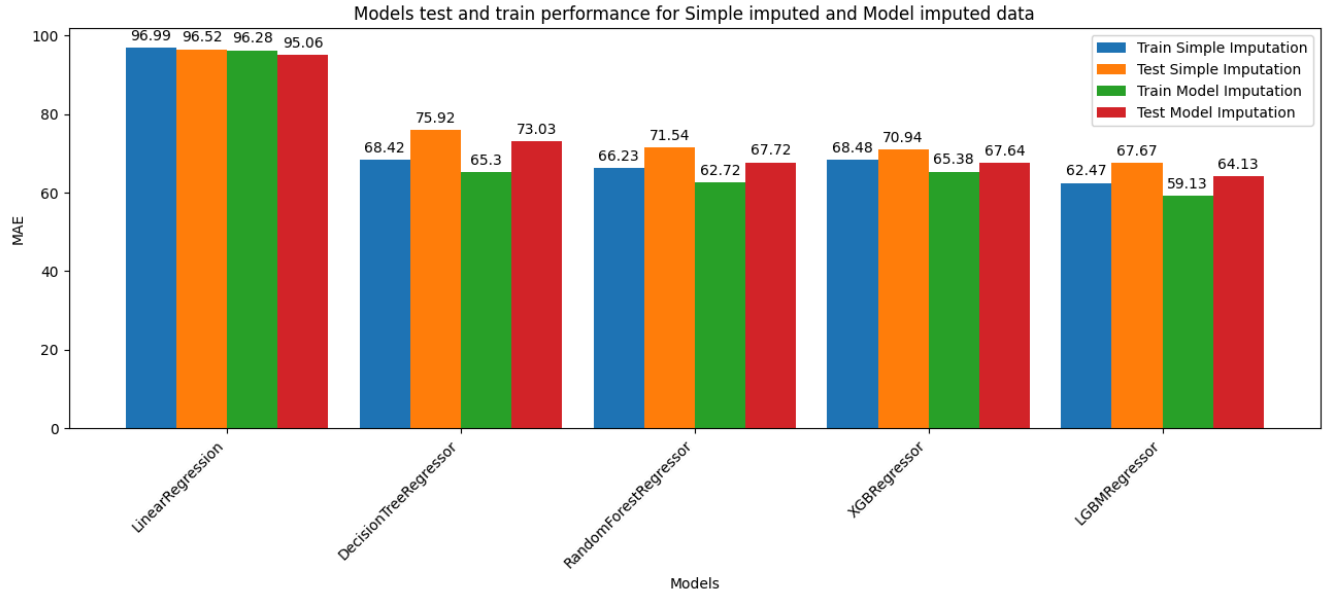
5.2 Improving the performance of the model using various techniques.

5.2.1 How can Model-based Imputation affect the performance?

In the first part of our analysis, we trained our models on data that had missing values which were imputed using simple methods such as Mean and Mode. However, in the second part of our analysis, we wanted to see if using model-based imputation would improve the performance of our models. Given that there were many missing values spread across different variables,

removing the records with missing values would have resulted in losing almost all of the records in our dataset. Therefore, we used Histogram Gradient Boosting models which can handle null values in the data to impute some of the most important variables. Each variable that needed to be imputed was considered as the target variable, and the remaining variables were considered as predictors.

We repeated the same baseline modeling with the model-imputed data. And the result is:

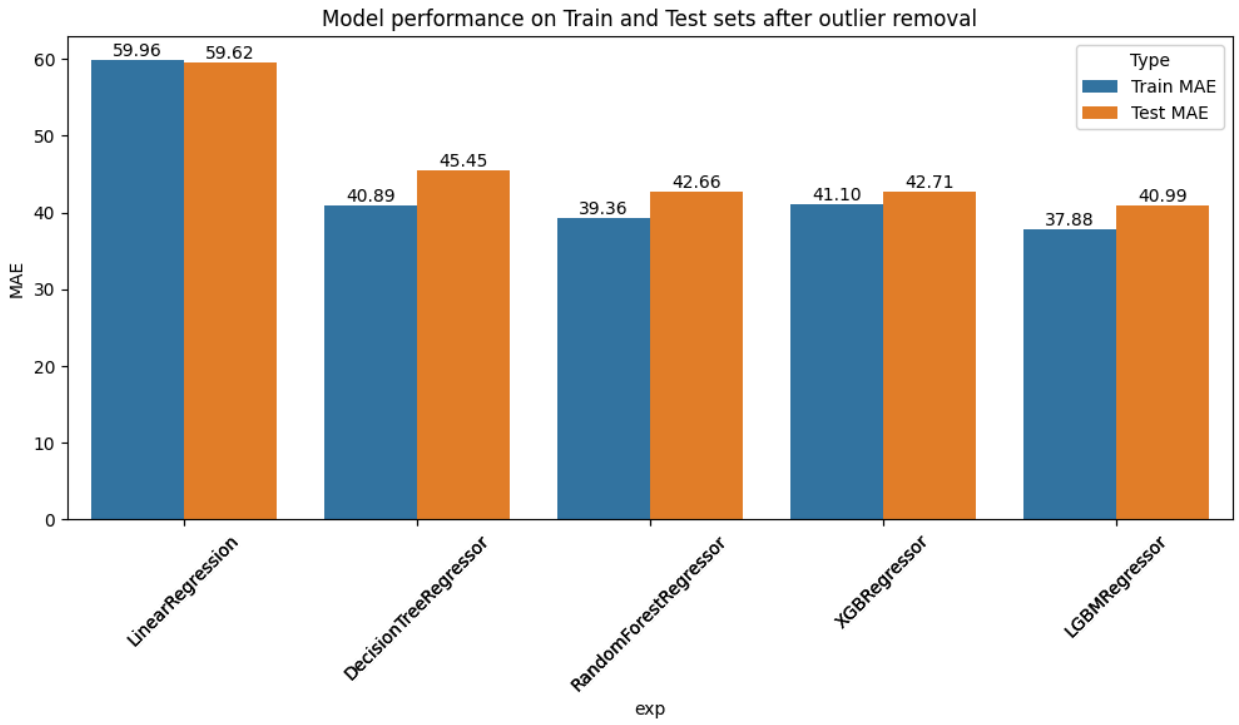


After analyzing the results, we have concluded that incorporating model-based imputation does enhance the performance of the models. However, the improvement is only marginal, with an increase of just 3 units in the MAE. It's not costly for us to run this step and we acknowledge that the model-based imputation should only be used when resource (Time, Compute, and Storage) consumption is not a concern and even small improvements in performance matter

5.2.2 Outlier Detection and Removal

Outlier detection and removal is a part of preprocessing, but we wanted to see to what extent the outliers impact the performance of the model. In the preprocessing step, we removed the records with extreme values of price. In this part, we removed the records further with “price” > 450 and “price” < 10. Since AIRBNB allows only a minimum price of 10$, it’s better to remove the records with “price” < 10.

Results:



From this experiment, we understood that removing outliers is one of the most important steps in modeling.

5.2.3 Hyperparameter Tuning using Optuna

Hyperparameters play a major role in the convergence of a machine-learning model, so we wanted to use this technique to improve the model's performance. Since we found that LightGBM is the best model among the others we experimented with, we want to tune the parameters of the model to improve its performance. Although Sklearn provides GridSearch for hyperparameter tuning, we chose Optuna because it uses algorithms to find the optimal parameters instead of searching through all the values in the parameter space. This makes it faster than grid search and enables us to search through a large parameter space in less time.

We focused on tuning the most important parameters of LightGBM, including learning_rate, max_depth, min_child_weight, subsample, colsample_bytree, num_parallel_tree, and gamma.

After running fifty iterations, the best parameters we found for our model are:

```
params = {'subsample': 0.992396132180261,
          'colsample_bytree': 0.9644764422219713,
          'num_parallel_tree': 1,
          'min_child_weight': 32,
          'gamma': 35.0147935853841,
          'max_depth': 10,
          'learning_rate': 0.19732349325787155}
```

5.2.4 K-FOLD Cross Validation

Our model has shown good performance on both the training and test data. This indicates that the model is generalizing well based on the results. However, we want to ensure that this

performance is not simply due to chance and that the model is truly generalized. To achieve this, we conducted a 5-fold cross-validation test to assess the model's ability to perform well on new and unseen data. The results of this test are:

```
Fold0 Val MAE: 40.31284974352994
Fold1 Val MAE: 40.58297363016835
Fold2 Val MAE: 40.47928754781962
Fold3 Val MAE: 40.59022479826402
Fold4 Val MAE: 40.40309679494194
5 Fold Average Val MAE: 40.47368650294477
```

From the results, we conclude that our model is generalized and can be used for predictions on unseen data points.

FINAL Model Results

After combining all techniques, we trained a final LightGBM model with the best parameters found during tuning. Here are the results:

```
Results for LGBMRegressor:
Experiment: FINAL LIGHTGBM MODEL
Mean Absolute Error for Training data: 35.84595032697756
Root Mean Squared Error for Training data: 49.55709259384933
Mean Absolute Error for Test data: 34.59071162553534
Root Mean Squared Error for Test data: 47.44763713978928
```

After analyzing the results, it appears that the model performs slightly better on the test data than on the training data. We were able to achieve this within the given time frame, but based on the final results, we believe that the model's complexity can be increased further.

6. Summary and Conclusions

This project has led us on a lengthy journey through data mining. The goal of our project was to decode the complexities of real-world data, and we have gained important insights from our research that broadened our understanding of not only this dataset but the techniques of data mining as well.

Diving deeply into both numerical and categorical variables, we were able to reveal how amenities, host interaction, types of properties, location, and more that affect listing. Heatmaps and box plots, two intuitive visualization tools, helped us convey these complex relationships in an understandable way.

Our modeling efforts ultimately improved our forecasts by using a variety of algorithms, from basic linear regression to complex methods like Random Forest, XGBoost, and LightGBM. Our models were refined and made robust every time through the use of critical techniques such as hyperparameter tuning, outlier detection, cross-validation, and model-based imputation. The combination of these methods produced a final model with good predictive power that generalizes well to new data, demonstrating the rigorous methodological approach used in this project.

Our research shows the power of data mining tools, while also adding to the theoretical and practical understanding of Airbnb's pricing systems. This project can help hosts strategically determine their listing pricing by analyzing the value that different qualities have from the viewpoint of the traveler. It provides tourists with an early look at the variables they may take

into account in order to locate listings that fit their tastes and price ranges. Additionally, Airbnb may also use these insights to optimize its platform.