Navigating New York's Airbnb Landscape: Data Insights for Hosts and Travelers

Rutuja Abhijit Kale, Sailendra Akash Bonagiri, Aishwarya Dwivedi

1 Introduction

Airbnb, an outstanding representation of modern hospitality, has grown to become one of the biggest providers of lodging in the world without ever owning a hotel room. Since its launch in 2008, Airbnb has revolutionized the lodging sector by enabling interactions between guests looking for lodging and hosts providing a range of rental choices. Because of Airbnb's Inside Airbnb & its openness with its operational data, which offers a wealth of information, the platform is a valuable source for insights based on data. The dataset used for this study includes information on over 38,000 listings in New York City, one of Airbnb's busiest cities. The most populated metropolis in the country and a major center of media, finance, and culture, New York metropolis welcomes more than sixty million visitors each year. The city's importance as a top Airbnb destination was highlighted by the $61.3 billion economic impact of tourism in 2014, which peaked in the city.

To achieve these objectives, we will:

1. Acquire and explore the dataset to understand its structure and the scope of data available.
2. Clean the dataset to ensure accuracy and usability in our analysis.
3. Conduct a detailed analysis to draw meaningful conclusions about the most popular room types, pricing trends, and strategic locations.

2 Background and Motivation

Airbnb has revolutionized tourism by allowing homeowners to turn their unused rooms into short-term lodging options. This new model offers travelers a wide range of choices, from single rooms to entire houses, catering to various preferences and budgets.

This project serves two main purposes. First, it helps hosts understand which features make their listings more attractive. This knowledge can lead to better pricing strategies, enhanced guest experiences, and, ultimately, more bookings and higher earnings. For hosts in a competitive market, insights into what travelers prefer and current trends are incredibly valuable. Second, for tourists, the sheer number of options on Airbnb can seem daunting. Our goal is to analyze data to uncover patterns in what guests look for and enjoy, helping tourists make choices that meet their expectations and enhance their travel experience.

We're diving into "Inside Airbnb's" comprehensive New York City dataset, which includes details on over 38,000 listings. This isn't just a pile of data; it's a key to understanding the complexities of one of the world's top tourist destinations.

Our analysis aims to offer practical insights that improve decision-making for both hosts and tourists. By pinpointing the key factors that boost a rental's appeal and guest satisfaction, we'll not only help hosts optimize their properties but also guide tourists through the vast array of choices more effectively.

3 Research Questions and Assumptions

1. What are the key factors that hosts should consider to enhance the appeal of their Airbnb listings?
    - What specific factors are most influential in determining the price of an Airbnb listing?
    - How do seasonal trends affect the demand and pricing of Airbnb listings in New York City?

- To what extent does the popularity of a neighborhood influence the demand and pricing of Airbnb listings within it?
- Can optimal pricing strategies for Airbnb listings be predicted using current data on listing attributes and market conditions?

2. How do tourists decide which Airbnb listings to choose based on available data?
   - Assuming that reviews significantly influence listing prices, how are pricing strategies affected by significant reviews?
   - Are there observable changes in listing prices before and after receiving significant reviews compared to control groups that did not receive any reviews during the same period?
   - How do the volume and quality of reviews affect potential guests' decisions when choosing Airbnb listings?

4 Dataset Overview

The Inside Airbnb dataset, sourced from insideairbnb.com, is a comprehensive collection designed to provide data and support advocacy regarding Airbnb's impact on communities. This dataset encompasses over 38,199 listings located within New York City. It includes a wide array of information, capturing details such as pricing, location, room types, host details, and reviews across various time 'snapshots' of listings throughout the city. The dataset is organized into several files: 'listings.csv' contains detailed information on individual listings; 'calendar.csv' offers data on availability and pricing over time; 'reviews.csv' provides detailed reviews linked to specific listings; 'neighbourhoods.csv' includes a list of neighborhoods useful for geographic filtering; and 'neighbourhoods.geojson' details the geographic boundaries of these neighborhoods. 'Airbnb NYC Data' includes all needed information to find out more about monthly demand, daily demand, necessary metrics to make predictions and draw conclusions. This structured array of data provides a rich foundation for analyzing the dynamics of Airbnb's marketplace in New York City.

The features in the dataset are organized in the following table

Category	Features
Host Information	'host_id', 'host_name', 'host_since', 'host_response_time'
Property Information	'Property_type', 'room_type', 'accommodates', 'bedrooms', 'bathrooms', 'beds', 'amenities'
Booking Details	'price', 'minimum_nights', 'maximum_nights', 'number_of_reviews', 'reviews_per_month'
Location Details	'latitude', 'longitude', 'neighborhood_cleansed'
Review Scores	'review_scores_rating', 'review_scores_cleanliness', 'review_scores_checkin', 'review_scores_communication', review_scores_location', 'review_scores_value'
Seasonal Trends	'Date', 'Demand', 'Easter', 'Thanksgiving', 'Christmas', 'Temperature', 'Marketing'

Figure 1: Category & Features

The features not included in the analysis: Id, name, description, instant_bookable, neighborhood_overview.

Column	Non-Null Count	Dtype
id	38199 non-null	int64
name	38197 non-null	object
longitude	38199 non-null	float64
latitude	38199 non-null	float64
description	37168 non-null	object
instant_bookable	38199 non-null	object
neighborhood_overview	21588 non-null	object
neighborhood_cleansed	38199 non-null	object
host_id	38199 non-null	int64
host_name	38194 non-null	object
host_since	38194 non-null	object
host_response_time	22494 non-null	object
review_scores_rating	26526 non-null	float64
property_type	38199 non-null	object
room_type	38199 non-null	object

Figure 2: Category & Features

accommodates	38199 non-null	int64
bathrooms	23634 non-null	float64
bedrooms	32494 non-null	float64
beds	23420 non-null	float64
reviews_per_month	26526 non-null	float64
amenities	38199 non-null	object
number_of_reviews	38199 non-null	int64
price	23634 non-null	object
maximum_nights	38199 non-null	int64
minimum_nights	38199 non-null	int64
host_listings_count	38194 non-null	float64
review_scores_checkin	26494 non-null	float64
review_scores_cleanliness	26504 non-null	float64
review_scores_communication	26498 non-null	float64
review_scores_location	26487 non-null	float64
review_scores_value	26488 non-null	float64

Figure 3: Category & Features

Attribute	Count	Unique	Top	Freq
name	38197	36460	Water View King Bed Hotel Room	30
description	37168	31264	Keep it simple at this peaceful...	106
instant_bookable	38199	2	f	30465
neighborhood_overview	21588	16042	This furnished apartment...	97
neighborhood_cleansed	38199	225	Bedford-Stuyvesant	2754
host_name	38194	8664	Blueground	837
host_since	38194	4979	2016-12-16	843
host_response_time	22494	4	within an hour	13788
property_type	38199	81	Entire rental unit	15734
room_type	38199	4	Entire home/apt	20187
amenities	38199	31212	['Smoke alarm', 'Air conditioning'...	182
price	23634	990	$150.00	618

Figure 4: Categorical variable Data Summary

Attribute	count	mean	std	min	25%	50%	75%	max
id	38199	...	...	...	...	...	...	...
longitude	38199	-73.94	0.05	-74.2	-74.0	-73.9	-73.9	-73.7
latitude	38199	40.73	0.05	40.5	40.7	40.7	40.8	40.9
review_scores_rating	29210	4.68	0.52	0.0	4.5	4.9	5.0	5.0
accommodates	38199	3.28	1.84	1	2	2	4	16
bathrooms	37624	1.25	0.65	0.0	1.0	1.0	1.5	8.0
bedrooms	38199	1.34	0.88	0.0	1.0	1.0	2.0	8.0
beds	38199	1.67	1.02	0.0	1.0	1.0	2.0	16.0
reviews_per_month	22947	1.28	1.47	0.0	0.0	0.72	1.95	17.61
number_of_reviews	38199	21.38	40.29	0.0	2.0	8.0	26.0	642.0
price	23634	151.93	175.16	0.0	65.0	100.0	180.0	10000.0
maximum_nights	38199	112.82	201.71	1	30	30	112.0	1125
minimum_nights	38199	4.48	11.58	1	1.0	2.0	3.0	1000
host_listings_count	38199	2.34	11.09	1	1.0	1.0	2.0	327
review_scores_checkin	29123	4.88	0.39	0.0	4.7	5.0	5.0	5.0
review_scores_cleanliness	29223	4.77	0.51	0.0	4.5	5.0	5.0	5.0
review_scores_communication	29215	4.85	0.42	0.0	4.7	5.0	5.0	5.0

Figure 5: Numerical variable Data Summary

5 Methods

5.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was essential to our project for several key reasons. Firstly, we aimed to identify if there are common features shared among Airbnb listings that span a variety of price points, which could help in understanding what attributes influence pricing strategies. Secondly, we sought to determine if specific neighborhoods in New York City consistently appeared across different pricing tiers, potentially revealing the impact of location on rental prices. Lastly, we were interested in exploring the relationship between the descriptions provided in listings and their associated prices, to ascertain whether more detailed or specific descriptions correlate with higher pricing. This analysis is intended to uncover underlying patterns that could inform both hosts and tourists in making better-informed decisions.

1. **Analyzing the listings based on room types:** Figure 6 reveals that the majority of listings are for entire homes or apartments and the least common type is Hotel room. This provides a brief overview of the kinds of listings available and their quantities.

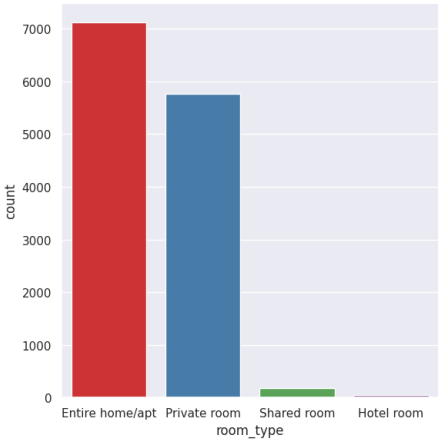

Figure 6: listing based on room types

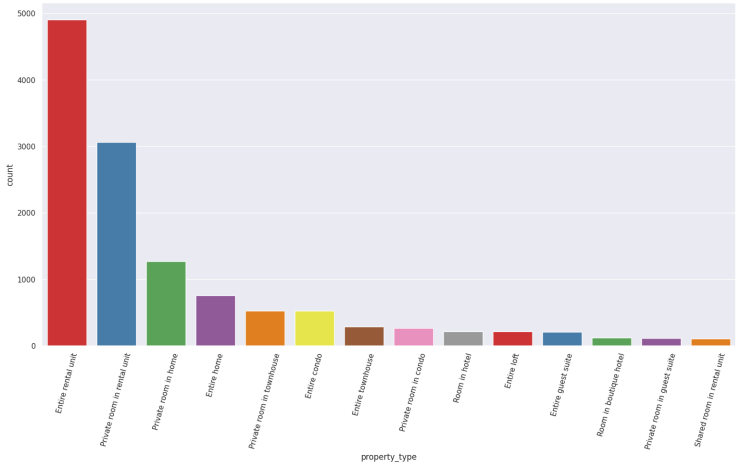

Figure 7: listing based on property types

2. **Analysis of Airbnb property types:** From figure 7 we can see that there are a lot more listings of apartments and full houses than any other property type in New York. Together with the earlier discovery that hosts prefer to list their full property than just a room or shared room, it can be inferred that most listings in New York are entire apartments or entire houses. Now let's analyze if these listing types have anything to do with the prices of the listings.

3. **Analyzing Mean Prices by Property type and Room type:**

   From the below heatmap, with lighter color representing lower price and darker representing higher price, we can see that shared rooms have the lightest color hence cheapest. Private rooms have a slightly darker color so they are in the middle, and entire houses are the darkest thus the most expensive. It is also important to note that the highest number of listings which were houses and apartments actually have very similar prices for each of the room_type categories. All of this tells us that the room_type and property_type both play a very important role in the final price of the listing.

Figure 8: Mean Prices by Property type and Room type

4. **Analyzing Mean Prices by Property type and number of Bedrooms:**

   From the heatmap and boxplot below, we can see that unsurprisingly, the price of listings increases with the number of bedrooms. Only the listing of a full house with 7 bedrooms does not follow this trend.

So far, we can see that room type, property type and number of bedrooms have some effect on the price of a listing. We will now analyze if any specific amenity in the property results in higher prices.

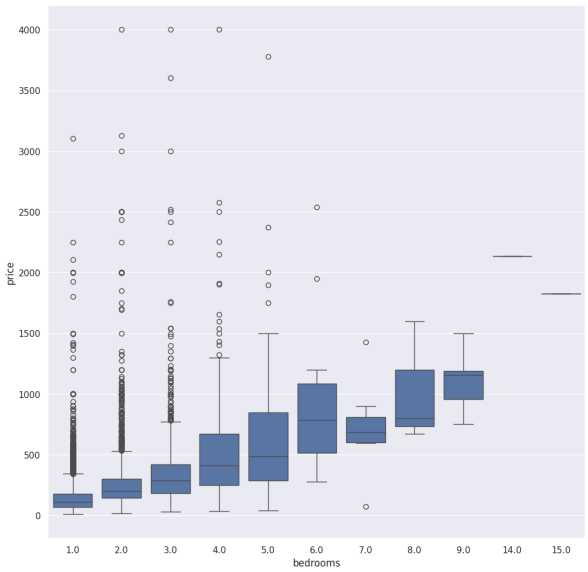

Figure 9: Boxplot: Mean Prices by Property type and number of Bedrooms

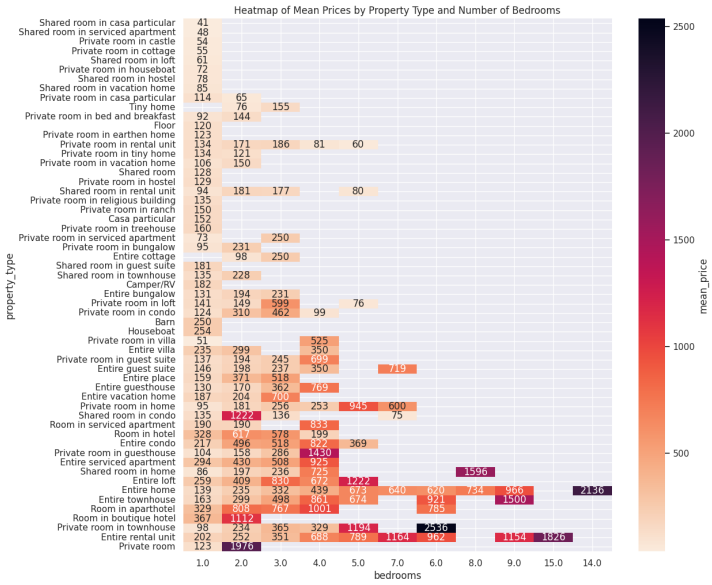

Figure 10: Heatmap: Mean Prices by Property type and number of Bedrooms

5. **Analyzing Top Amenities in High-End Airbnb Listings:**

   There are certain amenities such as Security and convenience features (e.g., CO alarms, air conditioning, hot water, coffee maker etc.) that most expensive listings also provide, and are prominently mentioned, indicating their importance. (Figure 11)

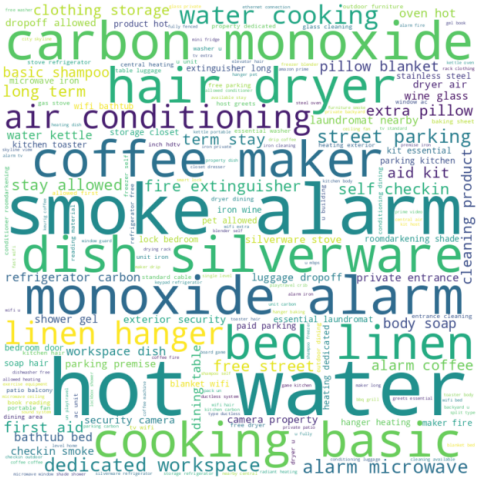

Figure 11: Top Amenities in High-End Airbnb Listings

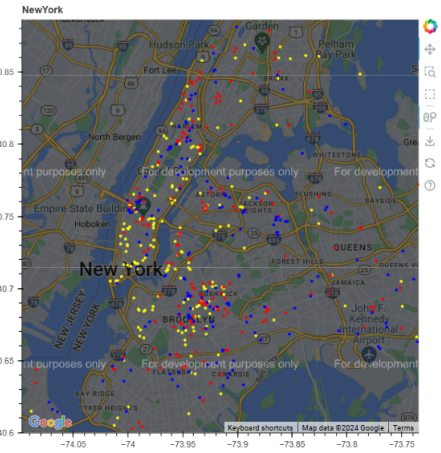

Figure 12: Number of listings of each room type in the neighborhoods

6. **Analyzing number of listings of each room type in the neighborhoods:**

   Yellow circles in figure 12, represent entire houses/apartments, red circles represent private rooms and blue circles represent shared rooms. From the map above, we can see that most of the yellow circles are concentrated in central New York. That is, most of the listings that list the entire house/apartment are concentrated in central NewYork. Since our problem was to identify factors that make a listing more expensive, we can infer that these neighborhoods tend to have more expensive listings.

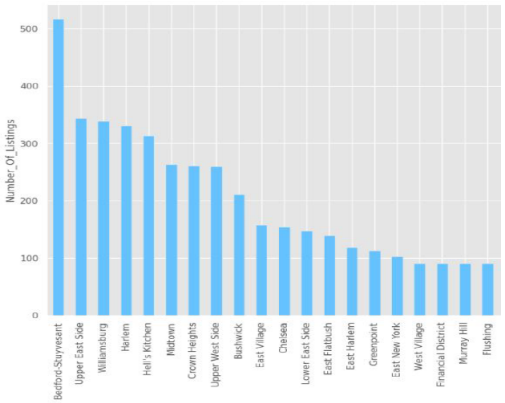

Figure 13: no. of listings for each neighbourhood

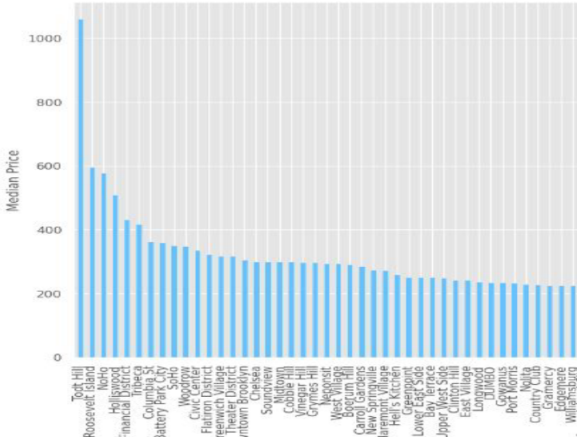

Figure 14: mean prices for each neighbourhood

From graph, we can see that most of the listings appear in 'Broadway', 'Bedford-Stuyvesant', 'Upper East Side', 'Williamsburg' etc. This gives us a good insight into the potential neighborhoods where there are a high number of listings. We found that prices are high for listings near popular neighborhoods.

7. **Analysis to find common words in the description of expensive listings:**

   The common words show that expensive listings have words focused on luxury and proximity to prime locations. (Figure 15) & The common words show that cheapest listings have words focused on commute, usually for locations far away from central. (Figure 16)

Figure 15: Common words in the description of expensive listings

Figure 16: Common words in the description of cheapest listings

5.2 Regression Models:

Regression models are used to target a prediction value based on independent variables used for finding out the relationship between variables and prediction/forecasting. Predictor Variables are Room_type, Accomodates, Bedrooms, minimum_nights, Number_of_Reviews, Property_type, Amenities, and Response Variable is Price. The following regression models will be carried out: Linear Regression, Random Forest Regression, XGBoost, CatBoost.

1. **Linear regression:**

   Linear Regression is a machine learning algorithm that is based on supervised learning. It performs the regression task to predict a dependent variable value (in this case, price) based on given independent variables (in this case, the identified predictor variables). It then tries to find a linear relationship between the variables and predicts the price based on the linear line.

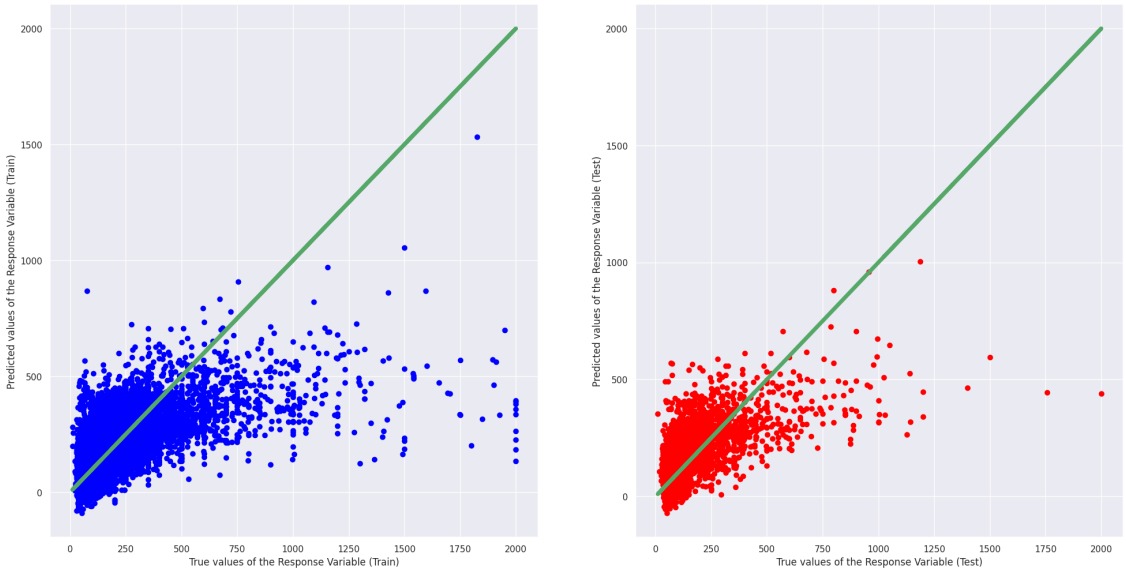

Figure 17: Linear regression

2. **Random Forest Regression:**

   Random Forest is an ensemble technique that is able to perform both Regression and Classification tasks with the use of multiple decision trees and a technique that is called Bootstrap Aggression. The idea behind this technique is to combine multiple decision trees in its prediction rather than relying on individual decision trees.Here, we use the RandomForestRegressor to help predict the price. To optimize the parameters used in the Random Forest Regression modeling algorithm, we first tune the parameters - in which GridSearchCV was used to optimize the parameters to determine the values that impact the model in order to enable the algorithm to perform at its best. Points that lie on or near the diagonal line means that the values predicted by the Random Forest Regression model are highly accurate. If the points are away from the diagonal line, the points have been wrongly predicted.

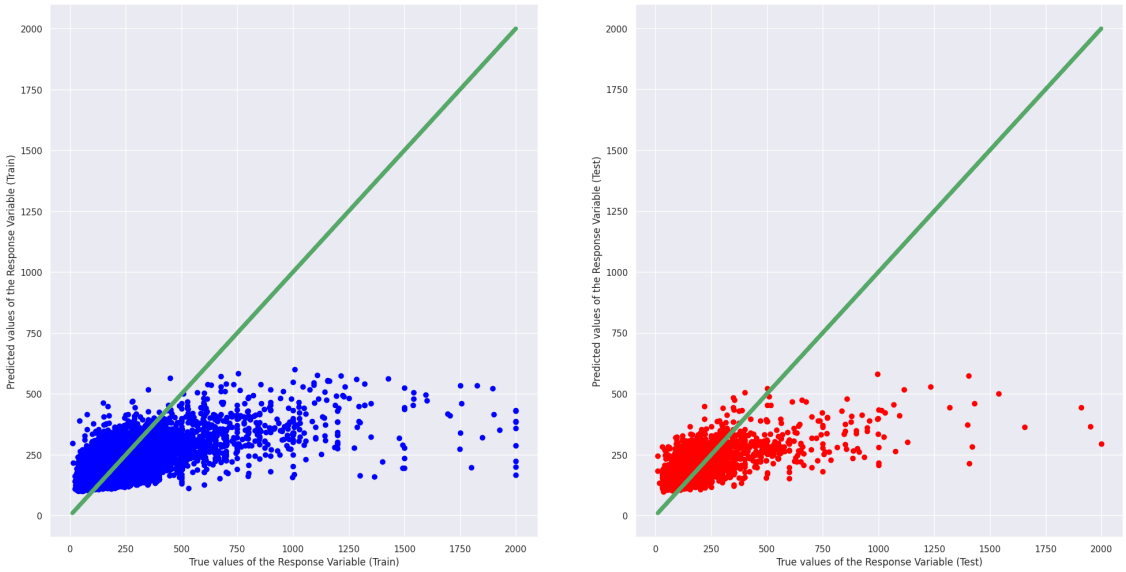

Figure 18: Random forest regression

3. **XGBoost:** XGBoost provides a high-performance implementation of gradient boost decision trees. The key idea of Gradient Boosted Decision Trees is that they build a series of trees in which each tree is trained so that it attempts to correct the mistakes of the previous tree in the series. To optimize the parameters used in the XGBoost modeling algorithm, we first tune the parameters - in which Grid-

SearchCV was used to optimize the parameters to determine the values that impact the model in order to enable the algorithm to perform at its best. Points that lie on or near the diagonal line means that the values predicted by the XGBoost Regression model are highly accurate. If the points are away from the diagonal line, the points have been wrongly predicted.

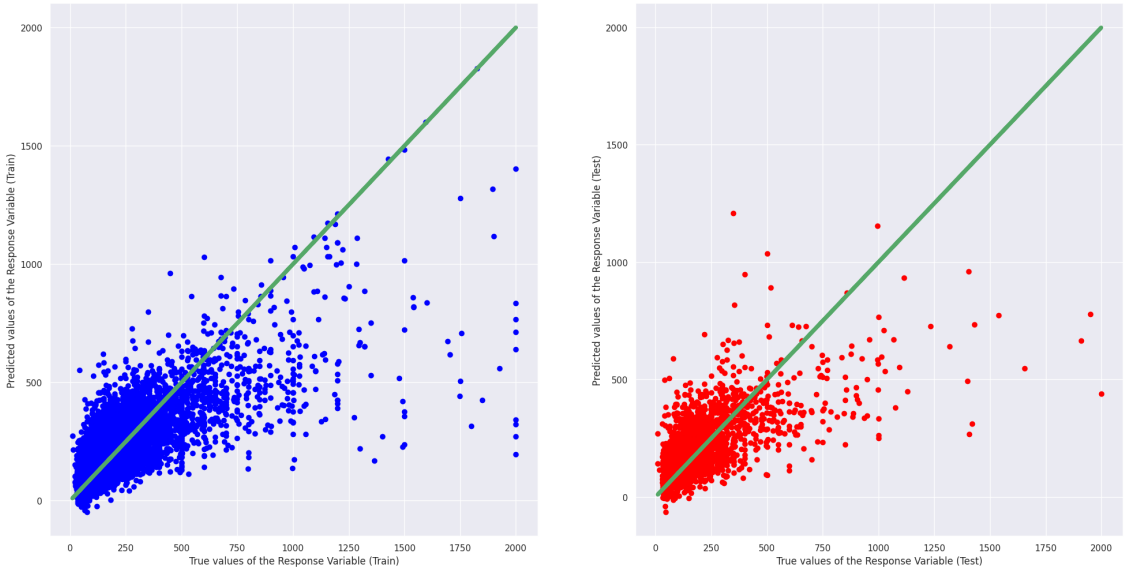

Figure 19: XGBoost

4. **CatBoost:** CatBoost is high performance gradient boosting on decision trees. Points that lie on or near the diagonal line means that the values predicted by the CatBoost Regression model are highly accurate. If the points are away from the diagonal line, the points have been wrongly predicted.

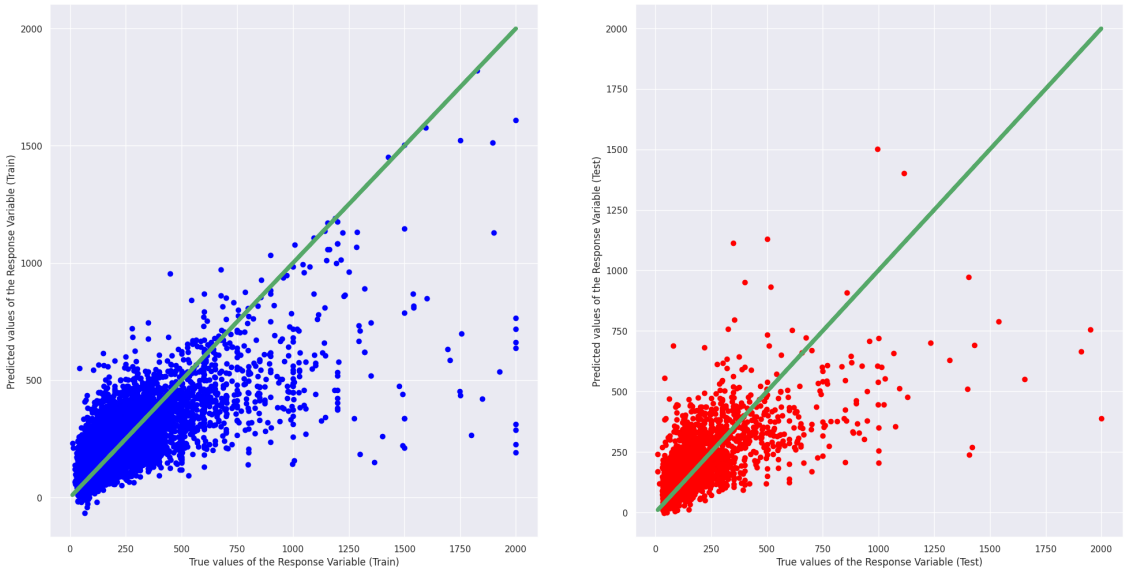

Figure 20: CatBoost

Model Comparison: Evaluation of Models Train Test- Split Validation of model performance is done using Train/Test Set Split in which the data set is split into 80% : 20%.

```
Linear Regression R^2: 0.3538

Random Forest R^2: 0.3309
XGBooost R^2: 0.4175
CatBoost R^2: 0.4141

Linear Regression MSE: 24015.2717

Random Forest MSE: 24921.7257
XGBooost MSE: 21628.1207
CatBoost MSE: 21753.6829

Performance Metrics for Train Set
-----------------------------------
Linear Regression (R^2):  0.3704

Random Forest Regression (R^2): 0.364
XGBoost (R^2): 0.5739
CatBoost (R^2): 0.5634
```

Figure 21: K Fold cross validation

```
Performance Metrics for Test Set
--------------------------------
Linear Regression (MSE): 47887.3145
Linear Regression (R^2):  0.3902

Random Forest Regression (MSE): 22449.515
Random Forest Regression (R^2): 0.3483

XGBoost (MSE): 19281.2912
XGBoost (R^2): 0.4403

CatBoost (MSE): 19556.4369
CatBoost (R^2): 0.4323
```

Figure 22: Model Comparison

However, Random Train/Test Set Splits may not always be enough as it can be subjected to selection bias during the split process (even if it's randomly split). This is especially so if the dataset is small. Train/Test Set Splits can also cause over-fitted predicted models that can also affect its performance metrics. As such, to overcome the pitfalls in Train/Test set split evaluation, k-fold Cross Validation is also performed. Here, the whole dataset is used to calculate the performance of the regression models. This validation method is more popular simply because it generally results in a less biased or less optimistic estimate of the model.

5. **K-fold Cross Validation:** K-Fold Cross Validation is where the dataset will be split into k numbers of folds in which each fold is used as a testing point. Here, k=10 is used as it is a value that has been found to generally result in a model skill estimate with low bias and a modest variance and we can see performance metrics comparison between (before K-Fold) and (after K-fold) and we can see the improvement in after-K-fold metrics.

XGBoost leads in performance among the evaluated models with the highest R^2 value, indicating the best fit, while Linear Regression showed the lowest MSE, suggesting strong predictive accuracy. Random Forest showed the least effective performance with the highest mean squared error and lower R^2

5.3 Backward Elimination and Selection

To validate our initial feature selection derived from exploratory data analysis (EDA), we implemented backward stepwise selection as a feature selection technique. This approach allowed us to quantitatively justify the feature set by starting with all possible predictors and systematically removing the least significant ones, ensuring each feature's relevance and statistical significance. By comparing the outcomes from this method to our initial EDA, we aimed to confirm the robustness of our hypotheses and address any discrepancies with sound qualitative reasoning, thus enhancing the empirical rigor of our analysis.

Data Preprocessing We began by removing outliers in the `price` variable using the interquartile range (IQR) method, ensuring data integrity. Amenities were transformed into binary variables to facilitate their inclusion in the regression analysis. Further, we performed log transformations on `price` and converted categorical variables like `instant_bookable` and `host_is_superhost` into numerical formats. The `host_since` was also converted to the number of days to quantify experience.

Feature Selection Using Backward Stepwise Selection To refine our predictor set, we implemented backward stepwise selection, guided by criteria such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and adjusted R-squared. Starting with all potential predictors, we iteratively removed the least significant variable based on the chosen criterion. This process was repeated until no further improvement in model performance was detected.

This methodical approach allowed us to effectively identify and retain the most significant predictors, ensuring our model's parsimony and enhancing its explanatory power.
Below is a concise summary of the variables selected by each model criterion:

AIC Model	BIC Model	Adjusted R-squared Model
accommodates	accommodates	accommodates
bathrooms	bedrooms	bathrooms
bedrooms	minimum_nights	bedrooms
beds	maximum_nights	beds
...	...	...
host_since_days	host_since_days	host_since_days
room_type_Private room	room_type_Private room	room_type_Private room
room_type_Shared room	room_type_Shared room	room_type_Shared room

Table 1: Variables selected by AIC, BIC, and Adjusted R-squared criteria

1. **AIC model:**

   The AIC model tends to include a broader array of variables, prioritizing model fit over simplicity. This approach captures more nuances and potential interactions within the data, which is reflected in the smaller p-values of the coefficients, indicating strong significance. However, it risks overfitting, potentially reducing the model's ability to generalize beyond the training dataset.

2. **BIC model:**

   By imposing a stricter penalty on the number of parameters, the BIC model includes fewer variables compared to the AIC model. This model is generally preferred for predictive purposes where the primary concern is avoiding overfitting. The coefficients in the BIC model may exhibit larger magnitudes for the same variables, indicating a reluctance to attribute effects to noise, thus favoring a simpler, more generalizable model.

3. **Adjusted R-squared Model:**

   This model strikes a balance between the complexity of the model (number of variables) and its ability to explain the variance in the response variable. It selects variables that significantly contribute to the model's explanatory power while adjusting for the number of predictors used. The result is a model that shares similarities with both the AIC and BIC approaches but is optimized for both explanation and prediction.

Variable Significance Across Models: Certain variables, such as 'accommodates', 'bedrooms', and 'review_scores_rating', have shown consistent significance across all models, underscoring their direct impact on rental pricing. Conversely, more nuanced features like 'dining table', 'shower gel', and 'oven' appear significant in the AIC model but are often excluded in the BIC and Adjusted R-squared models. This highlights how the stringency of variable inclusion criteria affects model composition and interpretation. In contrast, essential amenities such as 'wifi', 'air_conditioning', and 'kitchen' remain significant across all models, emphasizing their importance in rental pricing.

5.4 Validation of EDA through Backward Stepwise Selection

We employed backward stepwise selection which validated the set of features we initially pinpointed during our exploratory data analysis (EDA). By methodically removing the least impactful predictors, this stepwise process reinforced the significance of the EDA-highlighted features. This alignment supports our initial feature selections and hypotheses, confirmed through additional statistical tests.

Model Preference Based on Project Goals:

1. **Explanatory Focus:**

   If the goal is to understand all potential influences on pricing comprehensively, the AIC model is preferable due to its broader inclusion of variables.

2. **Predictive Focus:**

   The BIC model is ideal for scenarios requiring robust performance on unseen data, as it minimizes overfitting through a simplified model structure.

3. **Balanced Approach:**

   The Adjusted R-squared model is recommended for projects that seek a balance between complexity and simplicity, ensuring robustness in both predictions and explanations.

5.5 Time Series Analysis:

Our primary goal was to analyze how the demand for Airbnb listings changes over time and to develop a predictive model that reflects these trends. We employed various techniques such as differencing to ensure stationarity, and moving average smoothing to clarify underlying trends. Additionally, we compared ARIMA models and regression on moving averages to pinpoint the most accurate forecasting methods. This analysis helps us identify seasonal trends and demand patterns, crucial for setting optimal pricing strategies and predicting future demand. These insights into the temporal dynamics of Airbnb demand are integral to boosting listing appeal and aiding tourists in making informed decisions.

1. **Stationarity and Differencing in Time Series for stationarity:**

   Stationarity is crucial in time series analysis because most forecasting models assume the statistical properties of the series (mean, variance, autocorrelation) are constant over time. It is essential for reliable and meaningful time series analysis, ensuring that the model parameters are consistent over time. We applied the Augmented Dickey-Fuller (ADF) test to check for stationarity. Differencing is a common method to stabilize the mean of a time series by removing changes in the level of a time series, and thereby eliminating trend and seasonality. The ADF test showed a significantly low p-value and an ADF statistic much lower than the critical values, indicating the series is stationary after differencing.

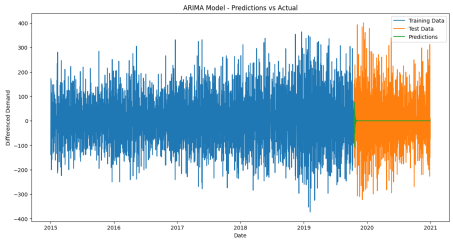

Figure 23: ARIMA Model - Predictions vs Actual

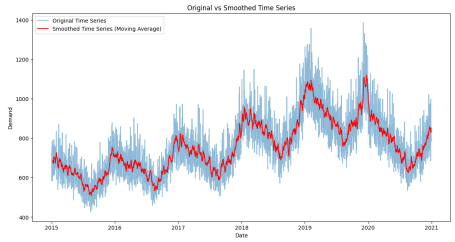

Figure 24: Original vs Smoothed Time Series

2. **ARIMA model:**

   In our analysis, we found that the average demand for Airbnb listings stands at approximately 756.06. There is considerable variability in demand, as indicated by a standard deviation of 152.14. The Root Mean Square Error (RMSE) of our predictive model is 148.49, which constitutes about 19.64% of the mean demand. This level of error suggests that, while the model's predictions are generally close to the actual values, there is a moderate prediction error. An RMSE of 148.49, in relation to a mean demand of 756.06, indicates a moderate level of accuracy.

3. **Moving Average (for smoothing):** Smoothing is essential for better visualization and understanding of the time series, making it easier to identify trends and cyclical patterns.

   We applied a moving average with a window size of 7 (weekly smoothing) and the smoothed time series highlighted the underlying trend by reducing the noise present in the original series.

4. **Linear Regression**

   We performed linear regression on the smoothened data and found an RMSE value of 88.78. The regression model using moving averages has a significantly lower RMSE (88.78) compared to the ARIMA model (148.49). This indicates that the regression model provides better predictions for this dataset. This suggests that the moving average feature captures important patterns in the data more effectively than the ARIMA model's approach.

5.6 Review Analysis:

Our review analysis aimed to bridge the quantitative data from Airbnb listings with qualitative feedback from guests, enhancing our understanding of factors that influence both pricing strategies and guest satisfaction. We focused on comparing the quantity and sentiment of reviews across differently priced listings to determine how they affect listing appeal and pricing dynamics.

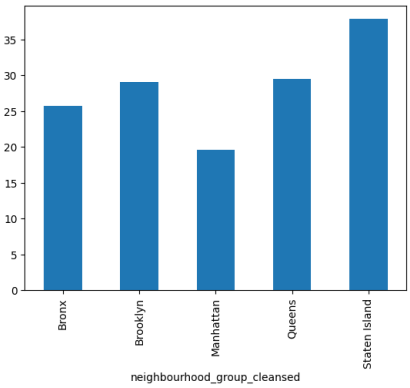

Figure 25: Avg. Reviews vs Neighbourhood

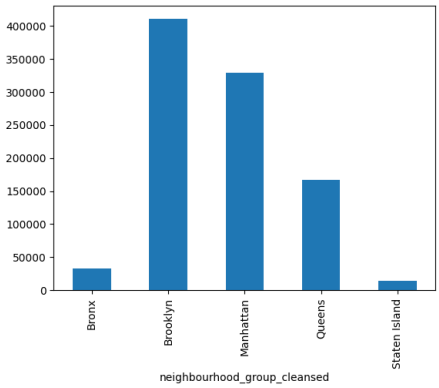

Figure 26: Total Reviews vs Neighbourhood

(a) **Relation with neighbourhood** Firstly, we tried to answer the question, can the popularity of a neighborhood be determined by the average and total number of reviews? The total number of reviews provides a direct measure of how many guests have stayed in a neighborhood, indicating guest satisfaction and return visits.

(b) **Relation with price** We wanted to examine if the

From the plot above we can see that more affordable listings have more number of reviews but to check this we see the correlation matrix as below.

The number of reviews a listing does not have much of an impact on the price as per the correlation matrix. Next, we checked the following:

- Are feedbacks more in pricier places or in cheaper places?
- Are there good or bad reviews in these places?

We analyzed the total number of reviews collected by both high-priced (greater than or equal to $200) and low-priced (less than or equal to $200) listings. This analysis found that low-priced listings accumulate more reviews than high-priced ones. This trend suggests that more affordable listings are more frequently chosen or have a higher turnover, possibly due to their accessibility and broader appeal among diverse guests

Figure 27: Number of reviews vs price

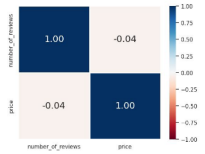

Figure 28: Correlation matrix reviews vs price

By generating word clouds for both high and low-priced listings, we visualized the most frequent words in reviews for each category. The word clouds revealed that terms like "apartment," "comfortable," "nice," and "clean" are commonly used across different price ranges. This suggests that regardless of price, guests value certain attributes like comfort and cleanliness in their accommodations.

Figure 29: Word Clouds for High Price vs Low Price

Sentiment analysis Using TextBlob, a sentiment analysis was conducted to classify the reviews into positive, neutral, and negative categories. The results indicated a higher proportion of positive reviews for both high and low-priced listings, with high-priced listings having a slightly higher positivity rate. This implies that while guests in higher-priced listings are slightly more satisfied, positive experiences are not exclusive to them.

Figure 30: Sentiment Distribution by Price Range

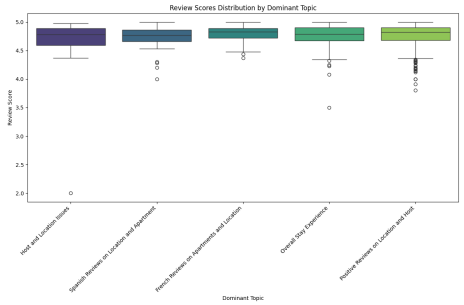

Figure 31: Review Scores Dist. by Dominant Topic

Boxplot Analysis

The boxplot analysis of Airbnb reviews reveals distinct correlations between specific listing aspects and guest satisfaction scores. Positive attributes such as favorable host interactions and desirable locations are consistently linked to high review scores, underscoring their importance in guest satisfaction. Conversely, issues related to hosts and locations contribute to a broader distribution of lower scores, highlighting these as critical areas for improvement to avoid negative guest experiences. Additionally, the analysis notes significant variability in reviews written in Spanish and French, suggesting that cultural and regional differences may influence guest expectations and perceptions.

Causal Analysis

In this causal analysis, we employed a difference-in-differences approach to investigate how significant reviews impact Airbnb listing prices over time. By comparing listings that received significant reviews with a control group that did not receive any reviews during the same period, we aimed to address several pertinent questions:

- Impact of Positive Reviews: We analyzed whether positive reviews (ratings above 4.5 stars) influence listing prices.

- We examined how negative reviews (ratings below 3 stars) affect listing prices.
- Is there a measurable difference in price changes between listings that received significant reviews and those that did not during the same period (control group)?

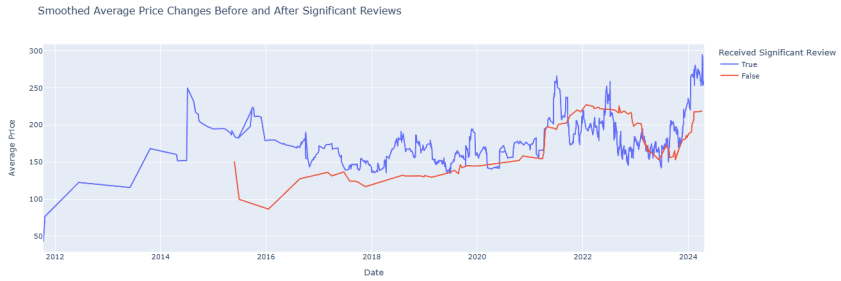

Figure 32: Smoothed average price changes before and after significant changes

Listings that received significant reviews exhibited a clear upward trend in prices, particularly noticeable around key time points such as the years 2020 and 2023. The smoothed average price changes graph illustrates that these listings tend to experience greater price increases following significant reviews. Listings that did not receive significant reviews (False) show a more stable trend with fewer fluctuations.

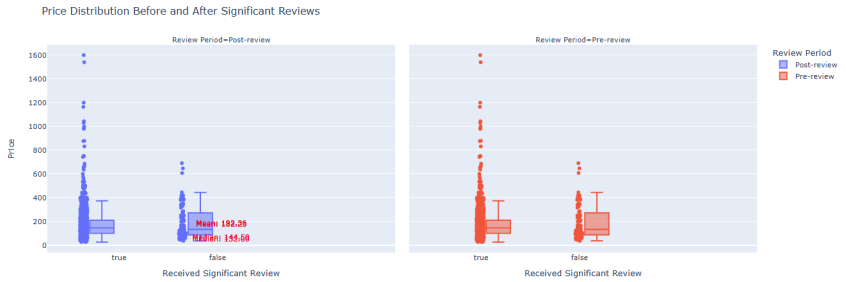

Figure 33: Price distribution before and after significant reviews

From the above plot of price distribution before and after significant reviews we see that for listings that received significant reviews, the post-review prices are generally higher and more spread out compared to the pre-review prices. The control group (False) shows less variation in prices, indicating that significant reviews have a notable impact on the pricing dynamics.

Our analysis suggests that reviews do not have a significant impact on property listing prices. Despite examining the effects of positive reviews (ratings above 4.5 stars) and negative reviews (ratings below 3 stars), the fluctuations in listing prices remained marginal. Additionally, when comparing listings that received a significant number of reviews to a control group with fewer or no reviews, we observed no substantial differences in price changes. This indicates that while reviews may offer insights into customer satisfaction, they do not markedly influence the pricing dynamics within the examined property market.

6 Results

For enhancing Airbnb Listing Appeal our analysis revealed several key factors that significantly enhance the appeal of Airbnb listings:

Property Characteristics: The number and quality of bedrooms and bathrooms directly impact desirability and can justify higher pricing.
Accommodation Size: Larger properties that accommodate more guests tend to attract higher prices, highlighting the importance of capacity in rental pricing.
Location: Listings in high-demand neighborhoods like Broadway and the Upper East Side command

premium prices due to their desirability.
Seasonal Timing: Adjusting prices according to peak tourist seasons enhances competitiveness and attractiveness by aligning with increased demand.

Tourist Considerations in Selecting Airbnb Listings For tourists, several factors are critical in choosing an Airbnb listing: **Reviews:** Positive reviews and price adjustments following such feedback indicate higher guest satisfaction and are reliable indicators of quality.
Seasonal Pricing: Awareness of seasonal price fluctuations aids in finding better deals and avoiding overpayment during high-demand periods.
Location and Neighborhood: Properties in well-reviewed neighborhoods offer better amenities and safety, significantly enhancing the stay experience.
Property Features: Essential features such as the appropriate number of bedrooms and bathrooms are crucial for ensuring comfort.

7 Conclusion

The exploration and analysis of over 38,000 Airbnb listings in New York City have yielded substantial insights into factors that influence both host profitability and guest satisfaction. Our findings underscore the importance of property characteristics, accommodation size, amenities, location, and strategic pricing in maximizing the appeal of listings. For tourists, selecting a listing based on understanding of pricing dynamics, and essential property features can significantly enhance their travel experience.

Our study not only assists hosts in optimizing their properties but also aids tourists in navigating the extensive options available on Airbnb. By applying a robust analytical approach, including regression models and time-series analysis, we have developed a comprehensive understanding that stakeholders in the Airbnb ecosystem can leverage to make informed decisions. This research underscores the dynamic interplay between host offerings and guest preferences.

GitHub repository Link

8 References

https://www.kaggle.com/datasets/sukanyabag/airbnb-nyc-data

https://insideairbnb.com