*Influential Factors that Determine the Price of an Airbnb Listing*

Parth Dalal   
*Data Analytics Engineering  
George Mason University*Fairfax, VA USA  
[pdalal5@gmu.edu](mailto:pdalal5@gmu.edu)

Sneha Kumaran  
*Applied Information Technology  
George Mason University*Fairfax, VA USA  
[skumara2@gmu.edu](mailto:skumara2@gmu.edu)

Moazzam Mansoob  
*Data Analytics Engineering  
George Mason University*Fairfax, VA USA  
[nfnu@gmu.edu](mailto:nfnu@gmu.edu)

Tanmaya Rodda  
*Applied Information Technology  
George Mason University*Fairfax, VA USA

[trodda@gmu.edu](mailto:trodda@gmu.edu)

*Abstract*—This study explored the dynamics of the Airbnb ecosystems. The goal was to determine the factors that are most influential in guiding a host in deciding the price point of their Airbnb listing and the most influential factors in helping a traveler book with Airbnb. The New York Dataset on Airbnb listings was examined. Variables examined include neighborhoods, room types, listing numbers, guest reviews, etc. The prediction models created used algorithms such as logistic classification, random forest regressor, and linear regression. The R-squared score of the linear regression model was -2.246558204332949e-05. The model was enhanced with the SelectKBest algorithm along with the selected features. Through this process, the result for this model obtained an R-squared score of 0.004056206511657501. Feature selection and hyperparameter tuning helped with the performance of the gradient-boosting regression model. In conclusion, the data exploration and models created provided some insight into the following: the features that determine the price of an Airbnb listing, the relationship between the price of a listing and customer reviews, and the neighborhoods in NYC with the highest number of customer reviews. Better results can be obtained with more complete datasets.

Keywords—AirBnb, Price, Hospitality, Factors, Reviews, Hosts, Neighborhood, Modeling, Prediction, Vacation, Property, Room, and Travelers

# Introduction

In the rapidly evolving world of modern travel, the rise of short-term vacation rentals has transformed the way people experience domestic and international travel. Alternative accommodations, like Airbnb, offer travelers a personalized experience compared to traditional hotels. As a result, the demand for Airbnb listings has skyrocketed, with millions of hosts worldwide seeking to attract guests and maximize their earnings. In this dynamic environment, the key factors that influence travelers in choosing an Airbnb and the determinants of the listing price have become paramount in the success of both hosts and guests.

The decisions made by hosts, such as the showing images and pricing of their property, can significantly impact bookings and generate income. Likewise, travelers have increasingly complex and diverse expectations when it comes to their Airbnb experiences. This intersection of host strategy and guest preferences has given rise to a complex and multifaceted problem.

This study explored the influential factors that affect both sides of the Airbnb marketplace. By delving into the dynamics of this shared economy, we hoped to shed light on how hosts can optimize their listings to attract guests while maximizing their returns and how travelers can select accommodations that align with their unique preferences and priorities for travel. We considered elements such as location, property characteristics, guest reviews, and pricing models to understand the decision-making process. In doing so, we sought to contribute to a deeper understanding of the Airbnb ecosystem and provide valuable insights for hosts, travelers, and researchers interested in the intersection of hospitality, economics, and technology.

# Motivation

Our motivation for this topic has come from a deep curiosity about the dynamics of technology, economics, and the hospitality industry. Airbnb has not only revolutionized the way of traveling but has also created a fascinating marketplace where hosts and guests engage in a unique form of economic exchange. Several factors have contributed to choosing this subject:

## Interest in Travel

As avid travelers and frequent users of Airbnb, we have always contemplated the factors that contribute to the perfect accommodation. Travel, for many, is an enriching and transformative experience. Airbnb emphasizes offering a traveler-centric approach, where guests can immerse themselves in local culture and choose accommodations that align with their unique preferences. This motivated us to understand how travelers make decisions that significantly impact their travel experiences.

## Economic Significance

Airbnb has emerged as a formidable player in global tourism and hospitality. The economic implications on local economies and the income potential for hosts have intrigued us. Understanding the underlying factors that influence pricing and booking decisions on Airbnb can offer valuable insights into this economic landscape.

## Technological Innovation

Airbnb relies on cutting-edge technology to connect hosts and guests, and we have been passionate about exploring how this technology is leveraged to enhance user experiences, streamline transactions, and influence decision-making.

## Academic Interest

We have been drawn to this topic because it lies at the intersection of multiple academic disciplines, including economics, psychology, marketing, and hospitality management. Analyzing these factors has allowed us to explore and integrate insights from various fields.

In summary, the motivation for selecting this topic lies in the desire to unravel the intricate web of decision-making in the Airbnb ecosystem, which can shed light on the factors that drive hosts and travelers in their quest for value and satisfaction. By delving into the motivations, preferences, and strategies of both hosts and guests, we aimed to contribute to a better understanding of the sharing economy and offer practical guidance for those engaging in the Airbnb marketplace.

# Research Questions and Significance

The following are the research questions we posed:

* What features determine the price of an Airbnb listing (factors such as reviews, room type, geography, etc.)?
* Is there any relation between the price of the listing and the customer review?
* Which neighborhood has the highest number of customer reviews?

The significance of the problem statements in studying the influential factors in the Airbnb marketplace lies within a few key aspects:

## Market Dynamics

Exploring factors influencing both hosts and guests within the Airbnb Marketplace offers a deeper understanding of the economy's functionality.There are many valuable insights to be attained. Understanding the market dynamics will help shape the decisions made by the host to place a price point of the listings and how the travelers make choices on various factors on their specific stay.

## Host Optimization

Identifying the factors that influence Airbnb listing prices can help the host optimize strategies to attract guests and maximize their earnings. Understanding factors such as reviews, room types, geography, etc. can allow the host to make more informed decisions on picking a price point and advertising their listing to potential guests/travelers.

## Enhancing Guest Experiences

The exploration of the relationship between the listing prices and the customer reviews can aid or help in understanding the impact of pricing on guest satisfaction. This will provide insight into adjusting and fluctuating pricing strategies to fit guest expectations and enhance their experience.

## Travelers Insights

Based on travelers, understanding which neighborhood has the highest number of customer reviews, it could help with crucial decision-making. It can play a role in safety and affordability. This can lead to a smooth vacation for any guest or traveler. More information about the listing will attract more guests to book a specific listing.

## Business and Economic Impact

Figuring out the influential factors holds strong economic significance just like how it is stated in the motivation section. It can play a pivotal role in understanding the local economies and broadening the tourism industry.

## Academic and Research Contributions:

An academic understanding of this study can provide knowledge on how this field is evolving and developing to satisfy all perspectives of the fields. The research contribution add more knowledge to developing and understanding the sharing economy, hospitality management, and consumer behaviors.

Overall, the problem statement and research questions are significant as they provide insights that can benefit all perspectives of the marketplace such as hosts, travelers, local economies, businesses, and the tourism industry.

# Timeline

Figure 1 shows the timeline for our project.

A timeline of milestones

Description automatically generated

Figure 1. Project Timeline

# Dataset

Our dataset was the Airbnb Open Dataset of New York City. It is a csv file and has 26 columns. The attributes are id, Name, host id, host\_identity\_verified, hostname, neighborhood group, neighborhood, lat, long, country, country code, cancellation\_policy, room type, construction year, price, service fee, minimum nights, number of reviews, last review, reviews per month, review rate number, calculated host listings count, availability 365, house\_rules, and license (Azmoudeh, 2022).

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| id | Nominal – Categorical | Unique id per listing |
| name | Nominal – Categorical | Name of the listing |
| host id | Nominal – Categorical | Host’s unique id |
| host\_identity\_verified | Nominal – Categorical | Tag of verified or unconfirmed host |
| host name | Nominal – Categorical | Name of the host |
| neighborhood group | Nominal – Categorical | City where listing is located |
| neighborhood | Nominal – Categorical | Neighborhood in the listing’s city |
| lat | Ratio – Continuous | Latitude |
| long | Ratio – Continuous | Longitude |
| country | Nominal – Categorical | Country where listing is located |
| country code | Nominal – Categorical | Unique code of the listing’s country |
| cancelation\_policy | Ordinal – Categorical | Consequences of canceling a reservation |
| room type | Nominal – Categorical | Type of space (entire place, private room, shared room, etc.) |
| construction year | Ratio – Continuous | Year listing was built |
| price | Ratio – Continuous | Price per night |
| service fee | Ratio – Continuous | Extra fees |
| minimum nights | Ratio – Discrete | Minimum number of nights on a reservation |
| number of reviews | Ratio – Discrete | Total number of reviews for a listing |
| last review | Ratio – Discrete | Date of the latest review on listing |
| reviews per month | Ratio – Continuous | Average percentage of reviews per month |
| review rate number | Ordinal – Categorical | Overall rating of the listing |
| calculated host listing | Ratio – Discrete | Number of listings the host owns |
| availability 365 | Ratio – Discrete | Number of days in a year the listing is available |
| house\_rules | Nominal- Categorical | Description of maintenance of the property |

Table 1. Description of the columns in the Airbnb Open Dataset of New York City from Kaggle

# Literature Search

In the article “Exploring Sources of Satisfaction and Dissatisfaction in Airbnb Accommodation Using Unsupervised and Supervised Topic Modeling,” Ding and his colleagues studied the key attributes affecting a user’s experience with Airbnb through the analysis of online reviews. They collected a large number of reviews from Airbnb for various locations and properties, and then they used sentiment analysis to check for positive or negative reviews. This helped to identify a user experience as being satisfactory or dissatisfactory. The results have shown users prefer places that are well-maintained, neat, and have comfortable beds. Users also prefer the hosts to be responsive and the location of the Airbnb listing to be convenient. Additionally, users do not like listings that are noisy, unclean, and in remote areas (Ding, K., 2021, April 21). Ding and his team’s research shows the importance of user reviews in helping hosts improve their listings. Using the results of this study as a reference point, we were curious to see if our analysis would find similar features to be important.

Additionally, Kanakaris and Karacapilidis investigated how to predict Airbnb pricing via graph neural networks and document embedding. In Airbnb, hosts set the prices of their listings based on demand, the price of neighboring listings, the popularity of the location, and the characteristics of the listing. The caveat to this autonomy is the hosts must have domain expertise to maintain profitable listings. Existing pricing models that aid hosts in determining the price of their listing include zip codes in their formulas. This poses an issue as these models are biased against specific destinations, do not apply to areas in which the price drastically changes within a zip code, and discount listings that are nearby but part of different zip codes. The use of graph neural networks (GNN) and document embeddings to predict Airbnb pricing mitigates these problems. This adapted method allows users to define a unique neighborhood for each listing, accounts for price variations in the same zip code, and takes into consideration features of a listing and the neighborhood to determine the price of an Airbnb listing. Kanakaris and Karacapilidis conducted their study on Airbnb listings on the island of Santorini, Greece which is all one zip code. The results of their evaluation show that “graph representation accompanied by a graph GNN model enhances the ability of modern ML-based pipelines to suggest accurately attractive (for the guests) and highly profitable (for the hosts) prices for a given Airbnb listing,” (Kanakaris & Karacapilidis, 2023). This study is important as the researchers demonstrated how helpful complex machine-learning techniques can be to hosts and guests. We aim to demonstrate the impactfulness of less complex machine-learning techniques. Additionally, Kanakaris & Karacapilidis’ study was conducted in Santorini, Greece. We are curious to see if similar results will be found in U.S. locations.

The study, “Airbnb Price Prediction Using Machine Learning and Sentiment Analysis” was conducted by Pouya Rezazadeh Kalehbasti, Liubov Nikolenko, and Hoormazd Rezaei. The authors used various machine learning algorithms and techniques to predict rental prices in New York City. Their dataset contains about 50,000 listings and 96 features. In this comprehensive study, the authors use natural language processing for sentiment analysis of the reviews. Their machine-learning algorithms were linear regression, ridge regression, gradient boosting, k-means clustering, support vector regression, and neural networks. These techniques were used to find the rental prices. Many feature selections were applied for this analysis. The Lasso regularization method was used to reduce the overfitting models. Also, the variance found from the model was needed to use the Lasso method. Using the SVR algorithm, the best performance acquired a test R2 of 0.69. This means it can grasp non-linear trends, patterns, and relationships between the hospitality features vs. the price (Rezazadeh Kalehbasti et al., 2021).

Similarly, in their study “Property Rental Price Prediction Using the Extreme Gradient Boosting Algorithm,” Marco Febriadi Kokasih and Adi Suryaputra Paramita forecasted property rental prices using the Extreme Gradient Boosting (XGBoost) algorithm. The researchers used Airbnb listing data sourced from Singapore and extracted features such as listing address, location, details of the listing, reviews, dates, and host information. Once the data was cleaned and preprocessed, 47 key features were selected to train an XGBoost model. This model later underwent hyperparameter tuning. Employing a 10-fold cross-validation technique, the model achieved an average Root Mean Square Error (RSME) of 10.86. There was a 13.3% error rate on the rental price prediction. The testing on sample data showed an error rate between 1 and 38%, with 5 to 10% having the most occurrences (Kokasih & Paramita, 2020).

Both Kalehbasti’s and Kokasih’s studies helped us in designing our model. Kalehbasti’s study showcases the possibilities of how to approach and predict prices using various models in machine learning. It also gives insight into how sentiment analysis and natural language could be great tools for us to use in our model. Kokasih’s Singapore dataset had many similarities to the New York dataset we used. Like Kokasih’s study, we applied gradient boosting along with hyperparameters to finetune the model created within the random forest regressor and linear regression machine learning algorithms. We also utilized amenities data and location attributes such as the neighborhood to find influential factors to validate our problem statement.

# Research Approach / Methods

We applied a comprehensive method that made use of data analysis, and machine learning algorithms, to precisely anticipate the elements that affect the pricing of Airbnb listings. Our approach comprised several key steps:

## Data Collection and Preprocessing

We compiled a complete dataset of Airbnb listings, including specific details about rates, reviews, property characteristics, client feedback, neighborhood and geography, and host-related variables. The data was processed before analysis. Processing was carried out meticulously using the Python programming language. The initial steps involved data cleaning, which included the removal of null values and the identification and elimination of duplicate entries. Additionally, unnecessary columns were systematically excluded to streamline the dataset. This thorough data preprocessing lays the foundation for robust and reliable analysis, ensuring the integrity and quality of the dataset for subsequent tasks." Compiling and cleaning comprehensive data ensured a thorough analysis.

## Feature Selection and Engineering

We selected appropriate features that could have an impact on the pricing domain. Using statistical methods, we determined their relevance to different elements such as property size, listing location, etc. Not every feature had the same weight when it came to pricing. Therefore, we focused on the most important qualities to reduce noise in the models. We used random regression for feature selection, and based on the feature importance, we selected important features. Selecting the optimal features was further refined using the SelectK method, focusing on the best features. Subsequently, we utilized these selected features to train and predict the model. It was found that a few important features, such as the number of reviews and availability in a year, played a significant role in influencing pricing.

## Machine Learning Models

To forecast Airbnb pricing, we created a prediction model that utilized the following machine-learning techniques: gradient descent, logistic classification, random forest regressor, and linear regression. To make a prediction based on data, machine learning models are important. Different algorithms take various approaches to analyzing data and thereby form various perspectives on the data.

## Evaluation and Validation

To evaluate the effectiveness of our predictive models, we used rigorous cross-validation techniques and metrics including RMSE, MAE, and R-squared. This evaluation was necessary to make sure our models worked properly and output reliable and accurate results/predictions for practical situations.

## Continuous Improvement

With the iterative nature of our approach, we were able to improve our model as we accumulated more data and insights. Continuous monitoring and feedback mechanisms ensured the adaptability and relevance of our predictive framework.

As a team, we utilized various data analytics techniques in the project. It involved quantitative research by handling different types of variables and manipulating the features to address the problem. Eventually, we developed an exploratory analysis to present information on factors influencing the choice of a particular Airbnb and the elements impacting the listing price. The team demonstrated leadership in the project development process, with each member presenting various ideas for enhancing the project. The objectives were clearly defined, and the attempt to find results was clearly stated and resolved. Each member cleaned the data, created visualizations, incorporated algorithms, and analyzed the data to understand the whole objective of the project.

In particular, we as a team incorporated various data analytical tools like Python, R, and Google Collab to help clean the data, create visualizations, and use algorithms for price prediction. We added more literature reviews to understand what factors play a significant role in choosing the right Airbnb and why Airbnb is listed at that exact price, shaping our understanding. We utilized the opportunity to demonstrate the cost perspective and the qualities of user decision-making for the selection of an Airbnb. This was our team’s plan to provide proper outcomes from the analysis of the data.

# Exploratory Data Analysis and Further Analysis

***In our first phase of analysis,*** we conducted an exploratory data analysis (EDA) on our dataset to understand its complex nature. This stage was crucial in revealing important insights and patterns in our dataset. EDA entails examining the structure, distributions, and connections between variables as well as searching for potential outliers. In addition to giving a general overview of our dataset, it also directed the subsequent phases of data cleaning and modeling.

A key component of EDA is data cleaning. In our research, data cleansing entailed two main steps:

1. Columns with High Null entries Removed: The "license" column, which contained 102,594 null entries, and the "house\_rules" column, which included many null values, were both eliminated. Columns containing a disproportionate number of null values might cause bias or confusion in any analysis or model, therefore taking this action was important to maintain the dataset's integrity.
2. Data transformation for "reviews per month": We multiplied the "reviews per month" data by 100 to make it more comprehensible. The ' reviews per month' variable is now easier to understand thanks to this change.

Data cleaning is important since it improves the quality of the data and gets the dataset ready for insightful analysis. It guarantees that there are no mistakes, discrepancies, or missing numbers in the dataset. Clean data is a requirement for precise modeling and insightful analysis.

The following is a summary of our initial findings:

## Host Identity Verification

The distribution of host identity verification between "unconfirmed" and "verified" categories is shown in Figure 2. This knowledge is important for estimating Airbnb pricing because homes hosted by verified individuals might have more credibility and favorable, which could have an impact on pricing.

A graph of blue squares

Description automatically generated

Figure 2. Bar graph of the number of unconfirmed and verified hosts in the Airbnb dataset

## Top 10 Communicities with the Most Hosts

For pricing forecasting, knowing which areas have the most hosts could be useful. To estimate pricing changes across different locations, it is crucial to comprehend the dynamics of supply and demand within particular communities. Figure 3 shows Bedford-Stuyvesant as the neighborhood with the most number of hosts, followed by Williamsburg, and the least was Crown Heights.

A graph of a number of neighborhood's

Description automatically generated

Figure 3. Bar graph of the number of hosts in the top ten neighborhoods in New York City

## Cancelation Policies

Examining how cancellation policies are distributed enables one to comprehend the degree of freedom provided to clients. More lenient cancellation policies may result in higher property rates, which can directly impact pricing. Figure 4 shows there is a higher number of Airbnb listings in our dataset with a moderate cancellation policy than those with a strict or flexible cancelation policy.

A graph of a number of cancellation policy

Description automatically generated

Figure 4. Bar graph of the number of listings with various cancelation policies

## Room type Distribution

It's crucial to understand how different room kinds, including "Entire home/apt," "Private room," "Shared room," and "Hotel room," are distributed to forecast Airbnb prices. The price ranges and demand for various room kinds vary. Figure 5 illustrates entire apartments are preferred over shared rooms or hotel rooms.

A graph of a room type distribution

Description automatically generated

Figure 5. Bar graph of the number of listings with various room types

These observations served as the basis for predictive modeling to calculate Airbnb pricing. For instance, characteristics like "room type," "monthly reviews," "cancellation policies," and "neighborhood" were anticipated to have a big impact on property pricing. More precise and reliable pricing models for Airbnb could be created by adding these variables into prediction models, taking into account various market situations and user preferences. This EDA method added a great deal of value to the field of predicting Airbnb prices by acting as a launchpad for later data analysis and modeling projects.

***In our second phase of analysis***, we delved into predicting listing prices using a Random Forest Regressor. Here are the key updates:

## Data Cleansing

We meticulously cleaned the dataset, addressing missing values and ensuring the integrity of numeric variables. Notable transformations include converting construction year, minimum nights, and other relevant columns to their appropriate numeric formats.

## Numeric Analysis

We processed numeric variables, ensuring data integrity by converting relevant columns to the appropriate numeric types. Figure 6 shows the summary statistics calculated on the first eight columns in our dataset after this cleaning was done. The preliminary numeric summary statistics provide insights into the central tendencies and variability of essential attributes such as construction year, minimum nights, and pricing.

A table with numbers and text

Description automatically generated

A table with numbers and text

Description automatically generated

Figure 6. A snippet of the first eight columns in our dataset after cleaning and summary statistics were calculated

## Price Prediction Model

Employing a Random Forest Regressor model with 100 estimators and a max depth of 7, we predicted listing prices. The model's accuracy, measured by the coefficient of determination (R-squared), stands at approximately 0.36%. This metric indicates the proportion of price variability captured by our model.

## Categorical Analysis

In addition to numeric variables, we explored categorical features, uncovering insights into host verification, neighborhood distribution, booking policies, cancellation preferences, and popular room types. Figure 7 displays summary statistics for the categorical variables we used.

A screenshot of a number

Description automatically generated

Figure 7. Categorical variables summary statistics

***In our final phase of analysis and model building,*** we implemented a random forest regression model to predict Airbnb prices. Random forest regression is a machine learning algorithm that ensembles multiple decision trees to make predictions. Decision trees are simple but powerful models that can learn complex relationships between features and the target variable.

We encoded categorical variables in the dataset using one-hot encoding. This is essential because machine learning algorithms can only work with numerical data.

Next, the code split the dataset into training and testing sets. The training set was used to train the random forest model, while the testing set was used to evaluate the model's performance on unseen data. We used 75-25 to split the data into training and testing.

The random forest model was trained with a hyperparameter of 100 trees and a maximum depth of 7. These hyperparameters were chosen through experimentation to achieve the best possible performance on the training set.

Once the model was trained, it was used to predict Airbnb prices on the testing set. The R-squared score of 0.003627 indicates that the model was able to explain a small but statistically significant portion of the variance in the data. As illustrated in Figure 8, the number of reviews, availability in a year, and construction year of the listing features are the most important factors in predicting Airbnb prices. This is reasonable, as these features provide valuable information about the Airbnb property, such as its availability, length of stay, and age.

A graph with blue and white bars

Description automatically generated

Figure 8. Variables that influence Airbnb pricing

The random forest regression model is a simple and effective approach to Airbnb price prediction. It can achieve reasonable accuracy on the testing set, even with a relatively small number of trees. However, it is important to note that the model is limited by the quality and completeness of the training data.

The linear regression model achieved an R-squared score of -2.246558204332949e-05, indicating a very poor fit to the data. The mean absolute error was 288.1124924853822, which means that the model's predictions were, on average, off by about $288. The model's coefficients indicate that the most important factors in predicting Airbnb prices are the neighborhood group, construction year, minimum number of nights, number of reviews, reviews per month, calculated host listings count, and availability.

To further enhance performance, feature selection using the SelectKBest algorithm was employed. The model trained on the six most crucial features attains an R-squared score of 0.004056206511657501 and a mean absolute error of 287.42077335789713.

A screenshot of a black and white screen

Description automatically generated

Figure 9. Linear regression with and without SelectKBest. The MAE for the two models is very similar. This suggests that the difference in accuracy is not very large.

The gradient-boosting regression model achieved an R-squared score of 0.0014630444715314495, indicating a slightly better fit to the data than the linear regression model. The mean absolute error was 287.74364625638776, which is very similar to the linear regression model. The model's coefficients are more difficult to interpret due to the nature of the gradient-boosting algorithm.

Feature selection and hyperparameter tuning improved the performance of the gradient-boosting regression model. The model trained on the selected features achieved an R-squared score of 0.004056206511657501 and a mean absolute error of 287.42077335789713.

In this study, 5-fold cross-validation was employed to evaluate the performance of gradient-boosting regression models. Grid search further improved the performance of the gradient-boosting regression model. The best model achieved an R-squared score of 0.012213845759609376 and a mean absolute error of 285.8843609616842. Figure 10 shows these statistics where Gradient Boosting Regressor (Original) shows the statistics after first tuning and Gradient Boosting Regressor (Tuning) shows results after the second tuning.

A screenshot of a graph

Description automatically generated

Figure 10. Comparison of linear regression, gradient boosting regressor, and tuned gradient boosting regressor

# Results, Conclusion, and Future Works

Our first research question states, what features determine the price of an Airbnb listing? Although our models did not perform strongly. Figure 8 indicates the following factors that could contribute to the price of a listing: the number of reviews left on a listing, the number of days the listing is available in a year, and the construction year of the listing. These features can be used to direct future studies.

Our second research question states, is there a relation between the price of a listing and the number of customer reviews? Analyzing the correlation between rental property costs and monthly review counts could help estimate Airbnb pricing. It demonstrates the relationship between pricing and user involvement. Figure 11 shows the correlation between price, reviews per month, and review rate number. The coefficient between price and reviews per month is 0.0042, indicating a very weak positive correlation. The correlation coefficient between price and review rate number is -0.0046, suggesting a very weak negative correlation. Additionally, the correlation coefficient between reviews per month and review rate number is 0.038, indicating a very weak positive correlation. The correlation matrix suggests overall weak relationships between price, reviews per month, and review rate number. Consequently, these variables are not strongly related and are not good predictors of one another. Considering Figure 11, we cannot there is a strong relationship between the price of a listing and the number of reviews.

A screenshot of a graph

Description automatically generated

Figure 11. Correlation matrix between the features price, reviews per month, and review rate number

Our last research question states, which neighborhood has the highest number of customer reviews? Based on our analysis, we found Brooklyn emerged as the leading neighborhood, boasting the highest number of Airbnb listings, followed closely by Queens. The abundance of Airbnb offerings in these areas suggests a high demand or popularity among users.

We pose the following suggestions for future studies. Our first recommendation would be to find a more complete dataset. A dataset with a minimal number of null values and more records would have performed better with our models. Future studies could also collect data from other major cities in the United States to see if various features are more impactful in certain cities. Data from rural areas can also be compared to data from urban areas. Our second recommendation is about machine-learning techniques. Future studies could perform more advanced techniques like sentiment analysis or time-series analysis. Sentiment analysis on customer reviews could uncover more accurate results on the exact features that travelers favor. Time series analysis could uncover the most profitable periods for hosts to make their listings available. Overall, big data analysis on Airbnb data can be beneficial for aiding hosts in improving their properties and aiding travelers in enhancing the value of their travel experience.

*References*

[1] Azmoudeh, A. (2022, August 1). Airbnb open data. Kaggle. <https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata?resource=download&select=Airbnb_Open_Data.csv>

[2] Ding, K., Choo, C.W., et al., (2021, April 21) Exploring sources of satisfaction and dissatisfaction in Airbnb accommodation using unsupervised and supervised topic modeling <https://doi.org/10.3389/fpsyg.2021.659481>

[3 ] Kanakaris, N., & Karacapilidis, N. (2023). Predicting prices of Airbnb listings via graph neural networks and document embeddings: The case of the island of Santorini. Procedia Computer Science, 219, 705–712. <https://doi.org/10.1016/j.procs.2023.01.342>

[4] Kokasih, M. F., & Paramita, A. S. (2020). Property Rental Price Prediction Using the Extreme Gradient Boosting Algorithm. *International Journal of Informatics and Information Systems*, *3*(2), Article 2. <https://doi.org/10.47738/ijiis.v3i2.65>

[5] Rezazadeh Kalehbasti, P., Nikolenko, L., & Rezaei, H. (2021, August). Airbnb price prediction using machine learning and sentiment analysis. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 173-184). Cham: Springer International Publishing.