Airbnb NYC 2019 Data Analysis Report

Group-1

Shridhi Patel- U02007154

Krushi Patel - U02007786

Ansh Kevadiya - U02007203

Smit Meghapara - U02007203

Lakshmi Nadella - U01950459

CS-661 - Python Programming

Professor Kshitij Sharma

December 19, 2025

Table of Contents

Abstract

This report presents an exploratory data analysis and machine learning study of the Airbnb NYC 2019 dataset to understand pricing and demand patterns across New York City listings. The dataset was examined to determine the main variables—such as location, room type, availability, and review-related characteristics—that affect Airbnb rates. Significant price differences between neighborhoods were found by exploratory data analysis; Manhattan and Brooklyn had higher average prices, whereas lower-priced listings had higher booking demand. Several machine learning models, such as Random Forest Regression, Decision Tree Regression, and Linear Regression, were used to forecast listing prices. RMSE, MAE, and R-squared metrics were used to assess the model's performance. The findings showed that Random Forest had the best predicted accuracy of all the models evaluated, while Linear Regression performed poorly since pricing is non-linear. Overall, the study shows that a variety of interrelated factors affect Airbnb pricing and emphasizes the significance of sophisticated machine learning methods for precise price prediction. The results offer a solid basis for predictive and recommendation-based analysis in the future.

Introduction

The global short-term rental sector has seen a dramatic transformation due to the explosive expansion of internet lodging platforms like Airbnb. Because of the fierce competition that Airbnb hosts face in large cities like New York City, pricing strategy is essential to increasing occupancy and income. Both hosts and platform analysts must comprehend the variables that affect listing prices and booking demand.

Location, room type, availability, seasonal demand, and customer ratings are just a few of the many interrelated factors that affect Airbnb pricing. Because Airbnb listings differ greatly from standard hotel pricing, it is difficult to forecast prices accurately. Because of this, straightforward pricing techniques frequently fall short of capturing the actual market dynamics.

The purpose of this project is to investigate how machine learning and data analysis methods might be applied to gain a deeper understanding of Airbnb pricing behavior. This analysis attempts to provide insights that can help hosts make better decisions and enhance predictive modeling techniques by detecting important price patterns and demand trends.

To find trends in pricing, location, and demand, this research initially conducts exploratory data analysis. The pricing of Airbnb listings are then predicted using machine learning models such as Random Forest Regression, Decision Tree Regression, and Linear Regression. To determine the best strategy, model performance is assessed and contrasted. Sections on dataset description, data preprocessing, exploratory analysis, machine learning models, results, conclusions, and future work comprise the remainder of this study.

Dataset Description

The Airbnb NYC 2019 dataset, which was acquired from Kaggle, was utilized in this investigation. The five boroughs of New York City—Manhattan, Brooklyn, Queens, the Bronx, and Staten Island—are represented in this dataset, which includes comprehensive data regarding Airbnb listings. This information is frequently utilized for scholarly research on short-term rental markets and reflects listing activity in 2019.

Each record in the dataset, which comprises over 48,000 Airbnb listings, represents a distinct property listing. It contains a mix of textual, category, and numerical qualities that indicate host details, listing features, and booking-related elements. These characteristics offer a thorough understanding of the New York City Airbnb economy.

Listing price, room type, area, necessary minimum number of nights, year-round availability, number of reviews, and monthly reviews are some of the dataset's key variables. Geographic analysis of price patterns is made possible by the inclusion of location-based data such neighborhood group, latitude, and longitude. Further feature research is also possible with descriptive parameters like listing name and host details.

This dataset covers both pricing and demand-related aspects in a fiercely competitive urban market, making it ideal for exploratory data analysis and machine learning tasks. It is a suitable option for examining Airbnb pricing behavior and developing predictive models because of its scale and diversity, which enable significant pattern discovery and model evaluation.

Data Preprocessing

To guarantee quality and consistency, the Airbnb NYC dataset was cleaned and preprocessed prior to exploratory data analysis and machine learning modeling. Because raw datasets frequently contain missing values, errors, and noise that can impair analysis and model performance, data preparation is an essential step.

Initially, all attributes in the dataset were checked for missing values. To preserve data integrity, columns containing null or missing values were either removed or subjected to the proper modifications. In order to avoid bias and redundancy in the analysis, duplicate records were also found and eliminated.

Data types were then examined and adjusted as needed. In order to facilitate statistical analysis and machine learning algorithms, numerical attributes including pricing, minimum nights, availability, and review counts were transformed into proper numerical representations. To guarantee uniformity among records, categorical data like room type and neighborhood group were standardized.

We looked at outliers because they might affect the model's performance, especially in the price attribute. To lessen skewness and increase the dependability of both visual analysis and predictive modeling, extreme price values were handled. Following these procedures, a final cleaned dataset was produced and utilized for all ensuing machine learning and exploratory data analysis tasks.

Exploratory Data Analysis (EDA)

The Airbnb NYC 2019 dataset was subjected to exploratory data analysis (EDA) in order to identify underlying patterns, trends, and relationships. Before deploying machine learning models, EDA assists in summarizing the key features of the data using statistical measures and visuals, enabling well-informed judgments. This section focuses on aspects relating to demand, listing features, and price behavior.

Price Distribution Analysis

Most Airbnb listings are priced at lower ranges, while a smaller percentage are priced much higher, according to the distribution of listing prices, which exhibits a right-skewed trend. This skewness draws attention to the existence of expensive luxury listings, which have a significant impact on summary statistics like the mean. Determining pricing variability and handling outliers during modeling requires an understanding of this distribution.

Room Type Analysis

The Airbnb market in New York City is dominated by listings for complete homes or apartments, followed by private rooms; shared rooms make up a very small percentage of listings, according to an analysis of room categories. This pattern indicates that consumers strongly favor private accommodation. Furthermore, whole house ads are typically more expensive than private and shared rooms, suggesting that room type has a big impact on pricing.

Neighborhood-Based Analysis

An investigation based on neighborhoods reveals significant differences in Airbnb rates between various boroughs. The average price of listings in Queens, the Bronx, and Staten Island is often lower than that of listings in Manhattan and Brooklyn. These variations can be explained by things like access to transit, business hubs, and tourist attractions. The findings highlight the significance of location as a major factor influencing Airbnb rates.

Demand Indicators

Demand-related attributes, such as the number of reviews and reviews per month, were examined as indicators of booking activity. According to the data, items with cheaper prices frequently garner more reviews, indicating greater demand. Higher-priced listings, on the other hand, typically have fewer reviews, which suggests fewer bookings. The trade-off between pricing strategy and occupancy is shown by the inverse relationship between price and demand.

Outlier Analysis

The price characteristic was the main area where outliers were found, with a few items having incredibly high costs. Because these outliers can skew visualizations and have a detrimental impact on machine learning models, they were thoroughly analyzed. The dependability of later analysis and modeling was enhanced by addressing or reducing the impact of such extreme numbers.

Machine Learning Methodology

Machine learning techniques were used to forecast Airbnb listing prices and assess the performance of several models following the completion of exploratory data analysis. The general approach for feature selection, data partitioning, and model evaluation is explained in this section. Building predictive algorithms that could identify price behavior based on listing attributes was the aim.

Feature Selection

To find suitable characteristics for price prediction, feature selection was carried out. To make machine learning models compatible with regression techniques, only numerical features were chosen. These features included availability all year round, the minimum number of nights, the number of reviews, and the number of reviews every month. In order to preserve model simplicity and consistency across several methods, categorical variables like room type and neighborhood were eliminated at this point.

Train–Test Split

The cleaned dataset was split into training and testing subsets to assess the model's performance. The models were trained using an 80–20 split, with 20% of the data set aside for testing. This method ensures a fair evaluation of predicted performance by enabling the models to learn patterns from the training data while being assessed on unseen data.

Evaluation Metrics

The accuracy and dependability of the model were evaluated using a variety of measures. Prediction error was measured using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), where lower values denoted greater performance. The degree to which each model accounted for the variation in listing prices was also assessed using the R-squared statistic. A more thorough examination of the model's efficacy was made possible by the use of several metrics.

Machine Learning Models and Results

The machine learning models used in this study are presented in this part along with an analysis of how well they forecast the prices of Airbnb listings. Three regression-based models—Random Forest Regression, Decision Tree Regression, and Linear Regression—were assessed. To provide fair comparison, all models were trained using the same feature set and assessed using identical performance criteria.

Linear Regression

The baseline model for price prediction was linear regression. The target variable, price, and the chosen numerical features are assumed to have a linear relationship by the model. Following training and testing, the model yielded a very low R-squared score and comparatively high error values. These findings show that the intricacy of Airbnb's pricing pattern was beyond the scope of linear regression. Despite its poor predicted accuracy, the model was a useful benchmark for assessing the efficacy of more sophisticated models.

Decision Tree Regression

Non-linear correlations between the input characteristics and listing prices were modeled using decision tree regression. The model was able to identify more intricate pricing patterns than Linear Regression by recursively dividing the data according to feature values. With lower RMSE and MAE values and a marginally higher R-squared score, the Decision Tree model performed better. But even with these enhancements, the model's performance was still constrained by its sensitivity to changes in the data and possible overfitting.

Random Forest Regression

Several decision trees are used in Random Forest Regression, an ensemble learning technique, to increase prediction accuracy and generalization. Out of the three versions that were tested, this one performed the best overall. It got the highest R-squared score and the lowest error values, suggesting a better capacity to forecast Airbnb prices. Random Forest's improved performance indicates that complex and non-linear interactions between features—which ensemble models are better suited to capture—have an impact on Airbnb price.

Literature Review

Prior research has used machine learning and exploratory data analysis to investigate Airbnb pricing trends in great detail. Location, room type, availability, and host-related characteristics all have a big impact on listing pricing, according to research. Neighborhood features and closeness to prominent areas are frequently found to be significant price drivers in metropolitan markets like New York City.

To forecast Airbnb prices, several academics have used conventional statistical models, such as linear regression. Although these models offer interpretability, they frequently fall short of capturing the intricate, non-linear correlations found in actual price data. Because they can simulate interactions between several variables, tree-based models like Decision Trees and ensemble techniques like Random Forest have become more popular.

According to recent research, when using Airbnb datasets, Random Forest models typically perform better than more straightforward regression models in terms of prediction accuracy. The diversity in pricing behavior between neighborhoods, non-linearity, and feature importance are all well handled by these models. Building on this earlier work, the current study analyzes and compares the performance of baseline and advanced machine learning models for Airbnb pricing in NYC.

Architecture / Methodology

This study's methodology includes model training, performance evaluation, exploratory data analysis, and data preprocessing. Prior to modeling, the Airbnb NYC 2019 dataset was cleaned by managing missing values, eliminating duplicates, and choosing pertinent numerical features.

Following preprocessing, an 80/20 ratio was used to divide the dataset into training and testing sets. Three machine learning models were used: Random Forest Regression, Decision Tree Regression, and Linear Regression. To create a baseline model and performance reference, linear regression was employed. While Random Forest Regression was employed as an ensemble technique to enhance generalization and prediction accuracy, Decision Tree Regression was utilized to identify non-linear pricing trends.

Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) were used to assess the model's performance. Python packages such as pandas, NumPy, scikit-learn, and matplotlib were used to implement each model.

Results

The findings of the experiment show that the applied machine learning models function differently. With a relatively low R2 value, linear regression yielded the worst results, suggesting that it was unable to account for a significant portion of the variation in Airbnb prices. This demonstrates that there isn't a straightforward linear relationship with Airbnb pricing.

With reduced RMSE and MAE values, Decision Tree Regression outperformed Linear Regression. Overfitting and generalization limitations continued to limit the model's ability to capture some non-linear interactions.

Out of all the models examined, Random Forest Regression performed the best overall. It demonstrated greater prediction accuracy with the greatest R^2 score and the lowest RMSE and MAE values. These findings support the idea that complex price behavior in Airbnb listings is better modeled by ensemble-based models.

Conclusion

Using exploratory data analysis and machine learning approaches, this study examined Airbnb NYC 2019 data to comprehend pricing behavior. The findings indicate that room type and location have a significant impact on listing pricing, with higher price levels found in areas like Manhattan and Brooklyn.

The comprehension of Airbnb pricing trends was greatly improved by machine learning models. Although it provided a helpful starting point, linear regression was not enough to predict prices accurately. By successfully capturing intricate and non-linear relationships in the data, Random Forest Regression fared better than previous models.

All things considered, this investigation shows the value of sophisticated machine learning models in practical pricing applications and offers a solid basis for next prediction research.

References

Inside Airbnb. (2019). Airbnb New York City listings dataset. https://www.kaggle.com/datasets/shivamb/netflix-shows

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning (2nd ed.). Springer.

Scikit-learn Developers. (2023). Scikit-learn: Machine learning in Python. https://scikit-learn.org/stable/