SEMMA Analysis of New York City Airbnb Listings

Abstract

This research paper presents a comprehensive analysis of Airbnb listings in New York City using the SEMMA (Sample, Explore, Modify, Model, Assess) methodology. The study aims to provide insights into the short-term rental market dynamics and develop predictive models for listing prices. By leveraging a dataset of 48,895 Airbnb listings, we explore various features influencing pricing and availability, identify patterns, and evaluate the performance of different machine learning models.

1. Introduction

The rise of the sharing economy has significantly impacted the hospitality industry, with Airbnb emerging as a major player in the short-term rental market. New York City, being one of the world's most popular tourist destinations, presents a unique and dynamic Airbnb ecosystem. This study applies the SEMMA methodology to analyze Airbnb listings in NYC, aiming to uncover insights that could benefit hosts, guests, and policymakers.

2. Methodology

2.1 Sample

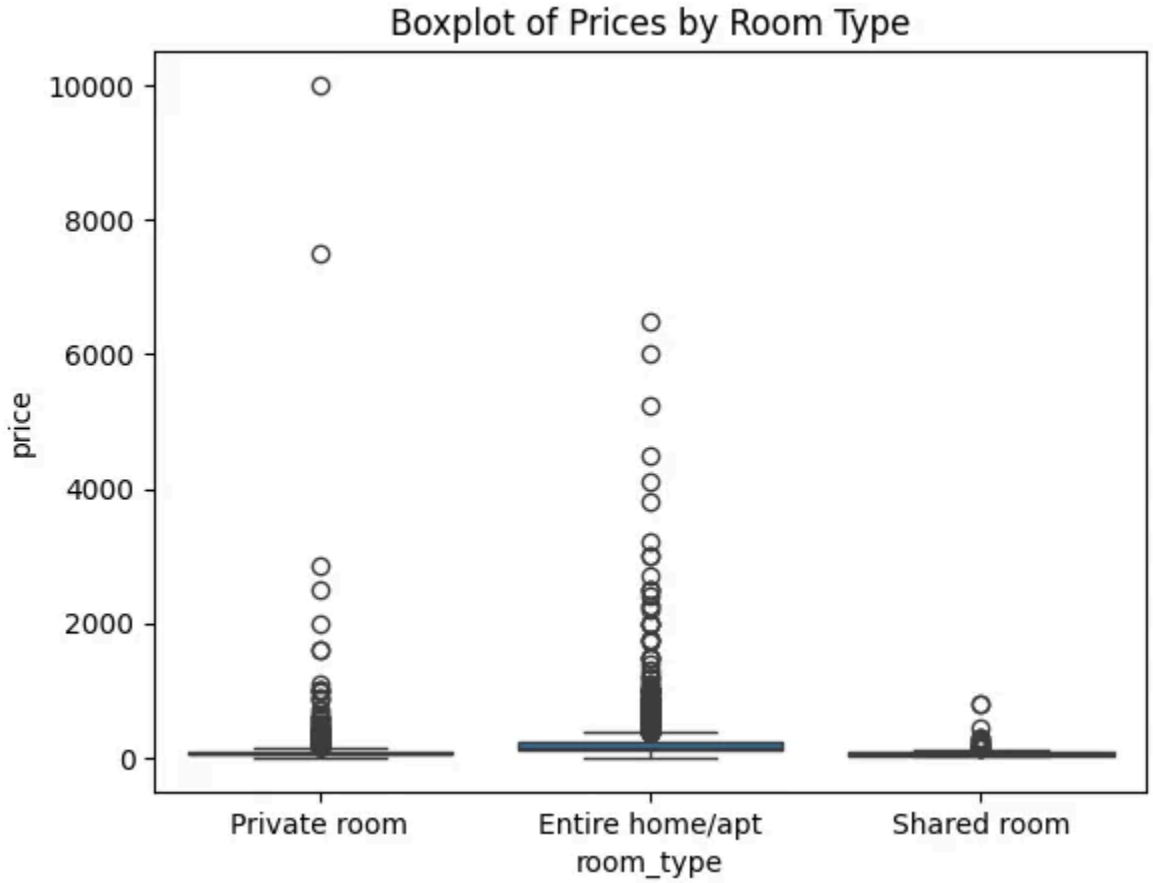
The dataset comprises 48,895 Airbnb listings in New York City, covering all five boroughs. It includes various features such as listing ID, host information, location details, room type, price, minimum nights required, number of reviews, and availability[1].

2.2 Explore

2.2.1 Distribution of Listings

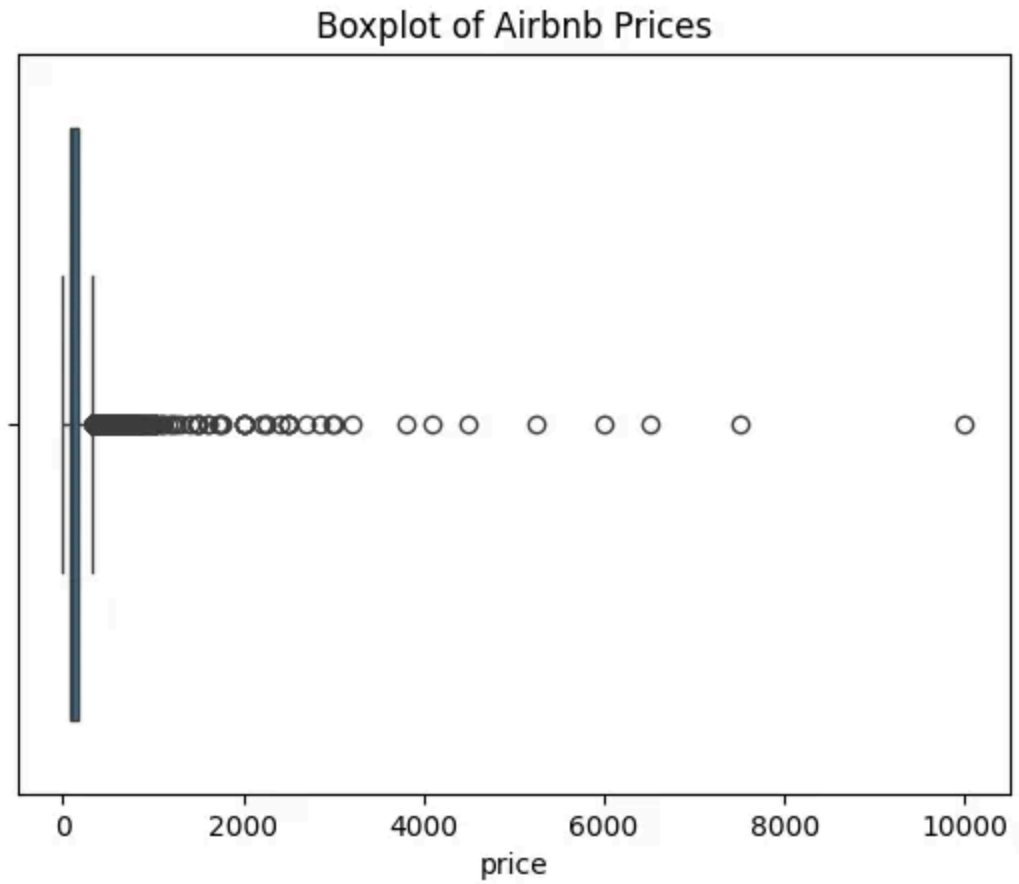
The listings are categorized into three main room types:
- Entire home/apartment: 25,409 listings

- Private room: 22,326 listings
- Shared room: 1,160 listings



2.2.2 Price Analysis

- Average price: $152.72 per night
- Median price: $106 per night
- Price range: $0 to $10,000 per night



2.2.3 Geographical Distribution

Listings span across various neighborhoods in New York City's five boroughs:
- Latitude range: 40.49979 to 40.91306
- Longitude range: -74.24442 to -73.71299

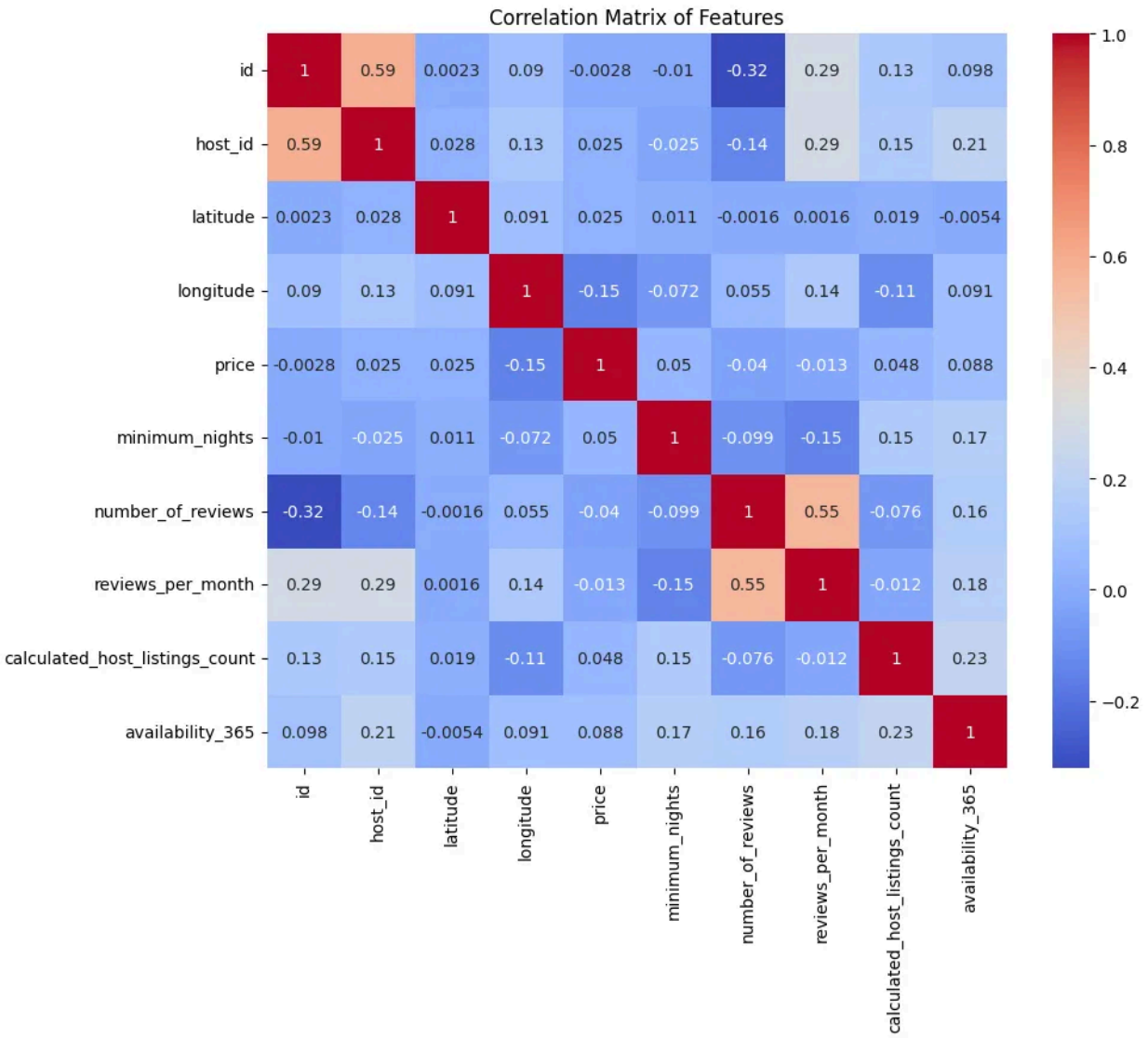
Notable neighborhoods featured in the dataset include Manhattan (Midtown, Harlem, East Harlem), Brooklyn (Kensington, Clinton Hill, Bedford-Stuyvesant), and Queens (Elmhurst).

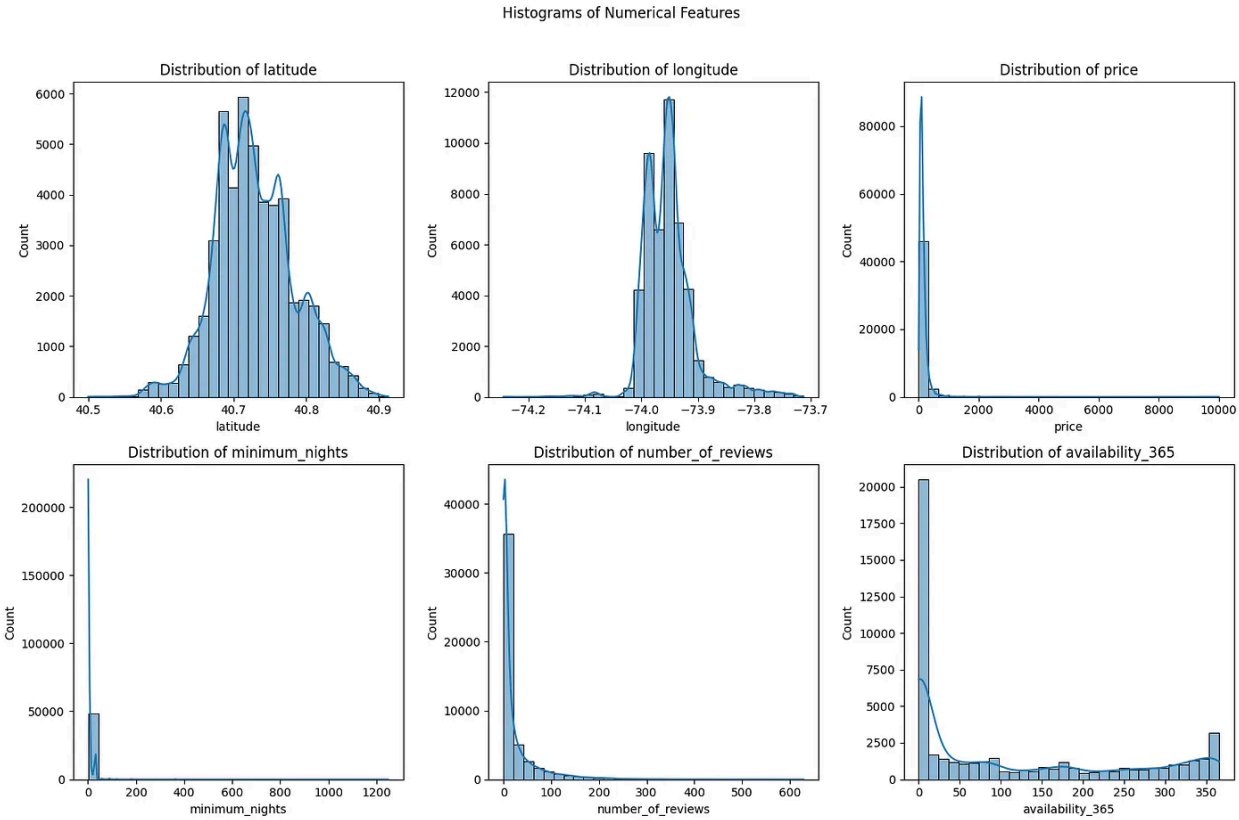
2.2.4 Booking Patterns and Availability

- Average minimum stay: 7 nights
- Maximum minimum stay: 1,250 nights (likely an outlier)
- Average number of reviews: 23
- Maximum number of reviews for a listing: 629
- Average reviews per month: 1.37
- Average availability: 113 days out of 365

2.2.5 Visualizations

Several visualizations were created to better understand the data distribution:
- Histograms of numerical features (price, minimum nights, number of reviews, availability)
- Box plots for outlier detection
- Scatter plots showing relationships between variables
- Heatmaps displaying correlations between numerical features


Correlation Matrix of Features
id 1 0.59 0.0023 0.09 -0.0028 -0.01 -0.32 0.29 0.13 0.098
host_id 0.59 1 0.028 0.13 0.025 -0.025 -0.14 0.29 0.15 0.21
latitude 0.0023 0.028 1 0.091 0.025 0.011 -0.0016 0.0016 0.019 -0.0054
longitude 0.09 0.13 0.091 1 -0.15 -0.072 0.055 0.14 -0.11 0.091
price -0.0028 0.025 0.025 -0.15 1 0.05 -0.04 -0.013 0.048 0.088
minimum_nights -0.01 -0.025 0.011 -0.072 0.05 1 -0.099 -0.15 0.15 0.17
number_of_reviews -0.32 -0.14 -0.0016 0.055 -0.04 -0.099 1 0.55 -0.076 0.16
reviews_per_month 0.29 0.29 0.0016 0.14 -0.013 -0.15 0.55 1 -0.012 0.18
calculated_host_listings_count 0.13 0.15 0.019 -0.11 0.048 0.15 -0.076 -0.012 1 0.23
availability_365 0.098 0.21 -0.0054 0.091 0.088 0.17 0.16 0.18 0.23 1
id
host_id
latitude
longitude
price
minimum_nights
number_of_reviews
reviews_per_month
calculated_host_listings_count
availability_365
1.0
0.8
0.6
0.4
0.2
0.0
−0.2



2.3 Modify

The modification phase involved several data preprocessing steps:

2.3.1 Handling Missing Values

Missing values were identified and addressed in the following columns:
- 'reviews_per_month': 10,052 missing values (filled with 0)
- 'last_review': 10,052 missing values (filled with 0)
- 'host_name': 21 missing values (filled with 0)
- 'name': 16 missing values (filled with 0)

2.3.2 Encoding Categorical Variables

One-hot encoding was applied to the following categorical features:
- 'neighbourhood_group'
- 'room_type'

2.3.3 Standardizing Numerical Variables

The following numerical features were standardized using StandardScaler:
- 'latitude'
- 'longitude'
- 'price'
- 'minimum_nights'
- 'number_of_reviews'
- 'reviews_per_month'
- 'availability_365'

2.4 Model

Several regression models were implemented to predict Airbnb listing prices:

1. Linear Regression
2. Decision Tree Regressor
3. Random Forest Regressor
4. Gradient Boosting Regressor (XGBoost)

The dataset was split into training (80%) and testing (20%) sets for model evaluation.

2.5 Assess

The performance of each model was evaluated using various metrics:

1. Linear Regression:
   - Mean Squared Error (MSE): 0.6746
   - Mean Absolute Error (MAE): 0.2948
   - R-squared: 0.1205

2. Decision Tree Regressor:
   - MSE: 1.7540
   - MAE: 0.3474
   - R-squared: -1.2866

3. Random Forest Regressor:
   - MSE: 0.6011
   - MAE: 0.2545
   - R-squared: 0.2163

4. Gradient Boosting Regressor (XGBoost):
   - MSE: 0.7448
   - MAE: 0.2735
   - R-squared: 0.0290

3. Results and Discussion

3.1 Exploratory Data Analysis Insights

1. The distribution of room types shows a preference for entire homes/apartments and private rooms, with shared rooms being less common.

2. Price distribution is right-skewed, with most listings clustered in the lower price range and a long tail of high-priced rentals.

3. There's a weak negative correlation between price and number of reviews, suggesting that more affordable listings tend to have more reviews.

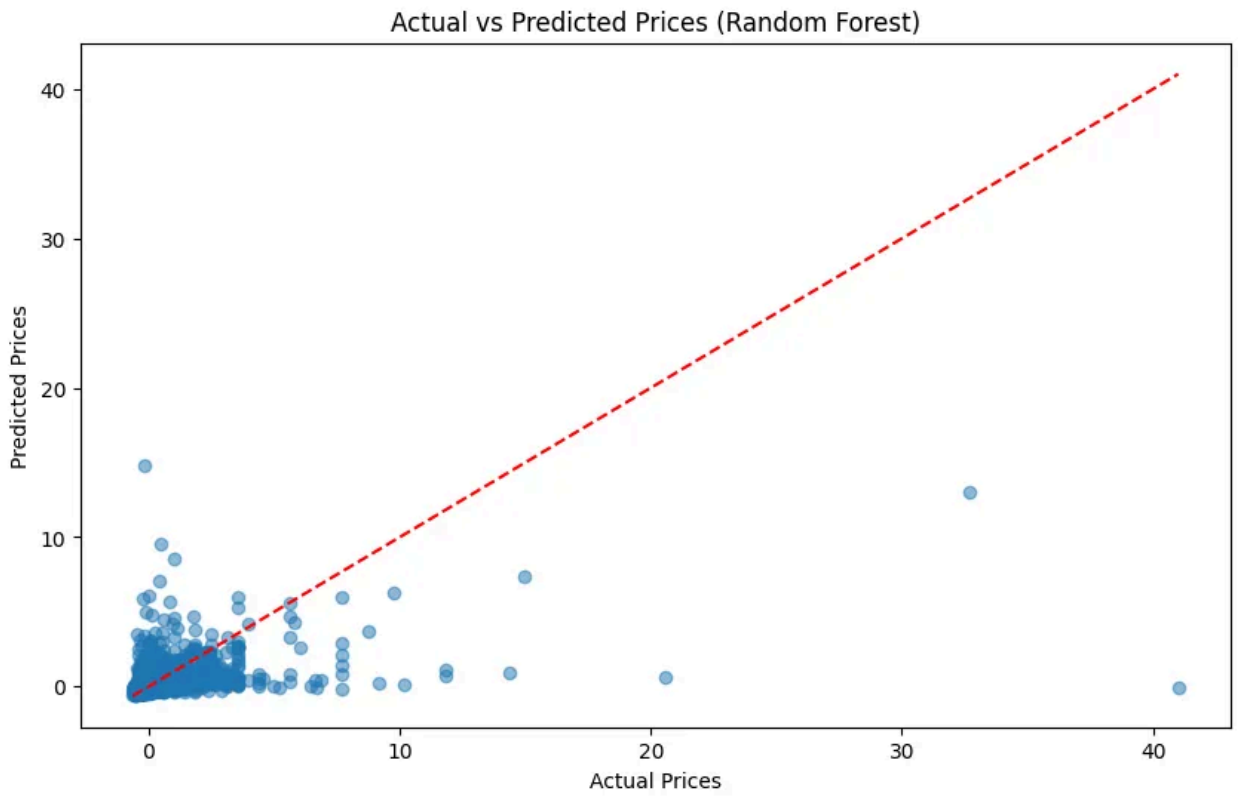
4. Manhattan and Brooklyn have the highest concentration of listings, reflecting their popularity among tourists and business travelers.

5. Entire homes/apartments generally command higher prices compared to private or shared rooms.

6. There's significant variation in minimum night requirements, which could reflect different host preferences or local regulations.

3.2 Model Performance

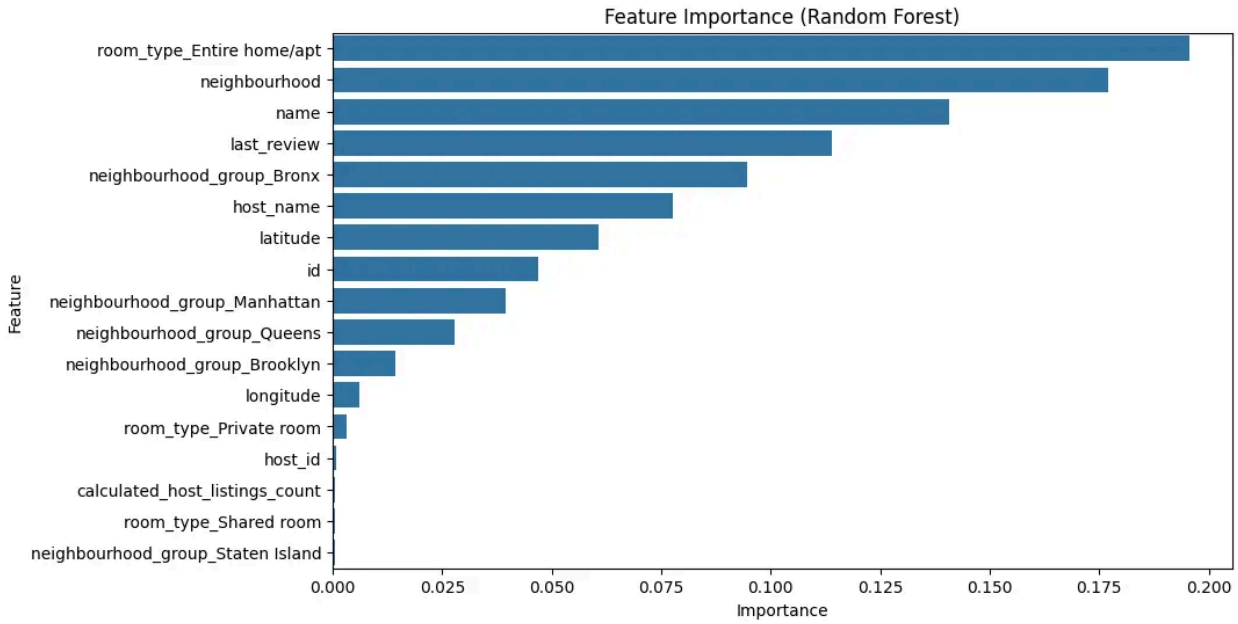
The Random Forest Regressor demonstrated the best overall performance among the models tested:
- Lowest Mean Squared Error (0.6011)
- Lowest Mean Absolute Error (0.2545)
- Highest R-squared value (0.2163)



The superior performance of the Random Forest model suggests that ensemble methods are well-suited for capturing the complex relationships within the Airbnb dataset. However, the relatively low R-squared values across all models indicate that predicting Airbnb prices is a challenging task, likely due to factors not captured in the dataset.

3.3 Feature Importance

A feature importance analysis conducted using the Random Forest model revealed the most influential factors in determining Airbnb listing prices. This analysis provides valuable insights for both hosts and guests in understanding price determinants.



4. Conclusions

This SEMMA analysis of New York City Airbnb listings has provided valuable insights into the short-term rental market dynamics:

1. The New York City Airbnb market offers a wide range of accommodations, catering to diverse traveler needs and budgets.

2. Location plays a significant role in pricing, with certain neighborhoods commanding higher prices.

3. Room type significantly impacts pricing, with entire homes/apartments generally fetching higher prices.

4. The average availability of 113 days per year suggests that many hosts use Airbnb as a part-time income source rather than a full-time business.

5. The Random Forest model's superior performance highlights the effectiveness of ensemble methods in capturing non-linear relationships and interactions between features.

6. The relatively low R-squared values across all models suggest that predicting Airbnb prices is a complex task, with factors not captured in the dataset likely playing a significant role.

5. Future Directions

1. Feature Engineering: Develop more sophisticated features, such as proximity to tourist attractions or public transport, to improve model performance.

2. Temporal Analysis: Incorporate time-based features to capture seasonal trends and price fluctuations throughout the year.

3. Advanced Modeling Techniques: Explore more advanced machine learning algorithms, such as neural networks or stacked models, to potentially improve predictive accuracy.

4. Geospatial Analysis: Conduct a more in-depth analysis of geographical factors, possibly incorporating external data sources on neighborhood characteristics.

5. Market Segmentation: Perform cluster analysis to identify distinct segments within the Airbnb market, which could lead to more targeted insights and strategies for hosts and platform managers.

6. Sentiment Analysis: Analyze review text data to extract additional insights about guest experiences and their relationship to pricing and booking patterns.

This comprehensive SEMMA analysis of New York City Airbnb listings provides a foundation for understanding the complex dynamics of the short-term rental market in a major urban center. The insights gained can inform decision-making for hosts, guests, and policymakers, while the modeling approach demonstrates both the potential and challenges of predicting prices in this dynamic marketplace.

Citations:

[1]
https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/31097934/5b9722ca-30ed-42b6-b3f7-567eb5ef1049/SEMMAColab2.pdf
[2]
https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/31097934/fa12e871-956b-4bb3-90b1-e8a0d5ff191d/SEMMAColab1.pdf