Analyzing Airbnb Pricing Dynamics in New York City

Using Clustering and Regression Techniques

Abrar Altaf Lone

Data Mining Fall 2025 72958

MS in Data Science

Pace University

Dec 16 2025

Abstract

This study analyzes Airbnb listings in New York City to understand pricing behavior using data mining techniques. Exploratory data analysis, clustering, and regression models are applied to identify pricing patterns and predict listing prices based on location, demand, and availability characteristics. Results show that non-linear models outperform linear baselines, highlighting the complexity of Airbnb pricing dynamics.

Contents

1 Introduction

Airbnb pricing in New York City varies significantly due to factors such as location, room type, demand, and regulatory constraints. Understanding these factors is important for hosts, guests, and policymakers. This project applies data mining techniques to explore pricing patterns and build predictive models using real-world Airbnb data.

2 Problem Statement

The objective of this project is to analyze Airbnb listings in New York City to identify pricing patterns, segment listings using clustering techniques, and predict listing prices using supervised learning models.

3 Dataset Description

The dataset used in this study is obtained from Inside Airbnb [1] and contains summary-level information on Airbnb listings in New York City. The file `listings.csv` includes listing characteristics related to price, location, demand, and host activity.

After preprocessing, the dataset contains approximately 20,900 listings with 10 predictor variables and one target variable.

4 Data Preprocessing

4.1 Missing Values

Listings with missing price values were removed, as price is the target variable. Missing values in review-related features were imputed with zero to indicate no recent review activity.

4.2 Feature Selection

Text-heavy attributes, identifiers, and metadata not directly related to pricing were excluded to improve interpretability and reduce dimensionality.

4.3 Target Transformation

Price was log-transformed to reduce skewness and stabilize variance for regression modeling.

4.4 Outlier Handling

An interquartile range (IQR)–based approach was applied to the log-transformed price. Approximately 1.8% of extreme observations were removed, retaining realistic price ranges between USD 16 and USD 1,500.

5 Exploratory Data Analysis

5.1 Price Distribution

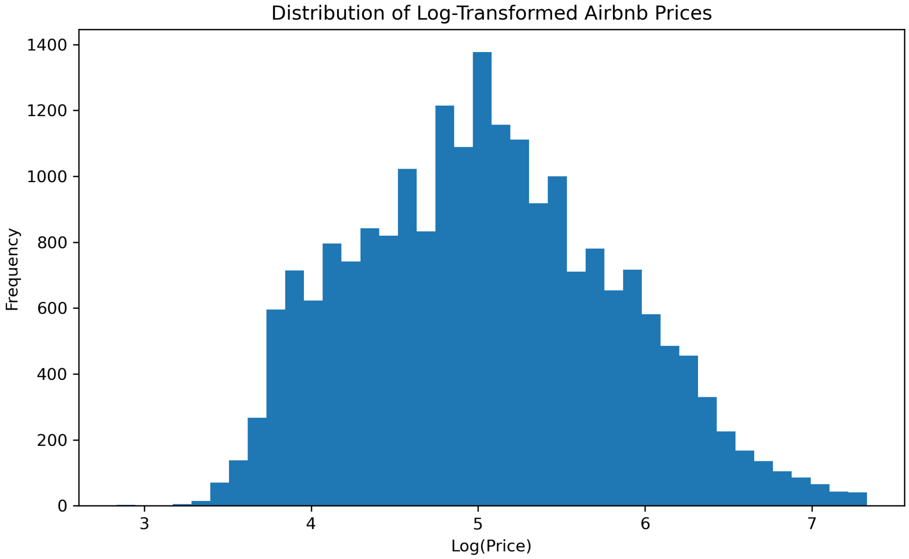

Figure 1: Distribution of Log-Transformed Airbnb Prices

Figure 1 shows that the log-transformed price distribution is approximately symmetric and unimodal, indicating that the transformation effectively reduced skewness.

5.2 Price by Borough

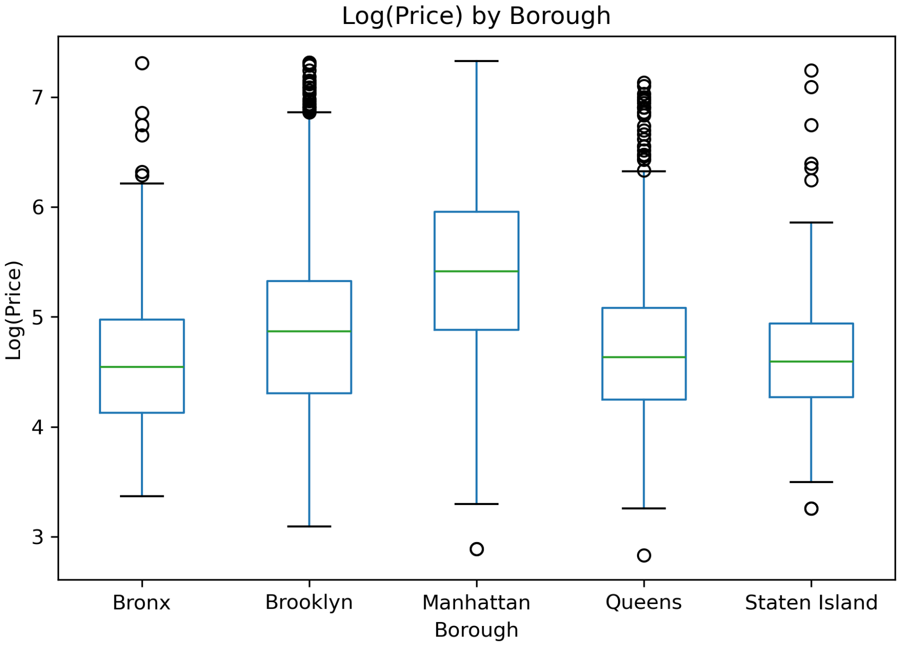

Figure 2: Log-Transformed Airbnb Prices by Borough

Figure 2 highlights substantial pricing differences across boroughs. Manhattan listings exhibit the highest median prices and greatest variability, while the Bronx has the lowest median prices.

5.3 Price by Room Type

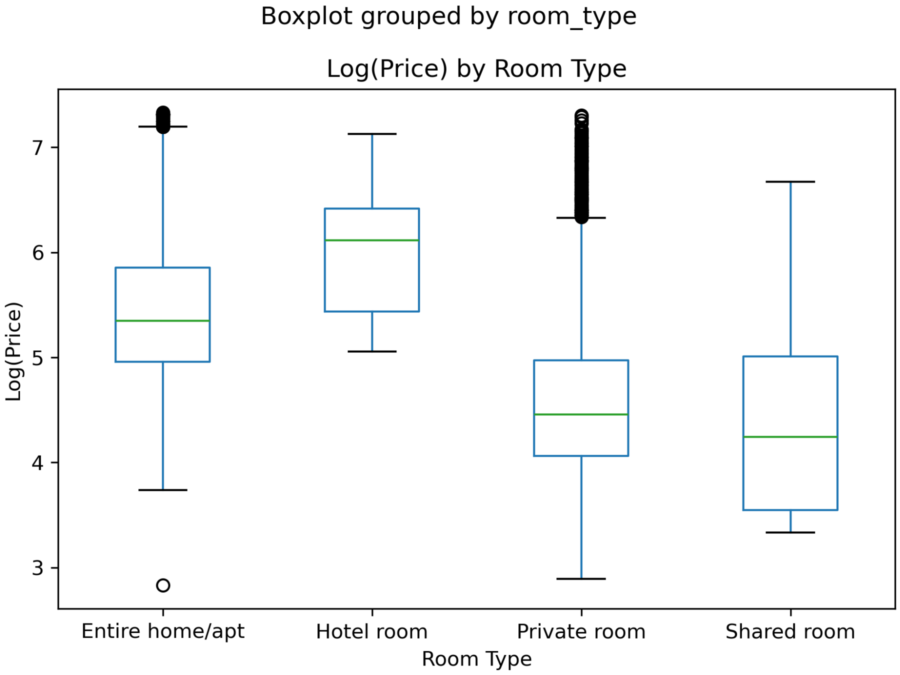

Figure 3: Log-Transformed Airbnb Prices by Room Type

Figure 3 shows that entire homes and hotel rooms command higher prices than private and shared rooms, indicating room type as a major determinant of pricing.

5.4 Correlation Analysis

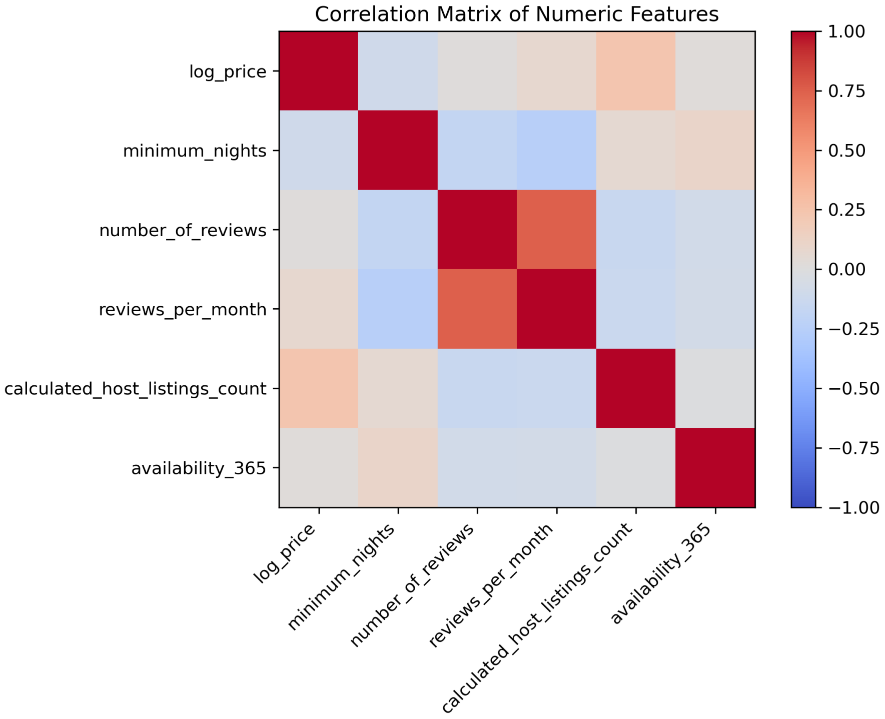

Figure 4: Correlation Matrix of Numeric Features

Figure 4 indicates weak linear correlations between most predictors and price, motivating the use of multivariate and non-linear modeling techniques.

6 Methods

6.1 Clustering

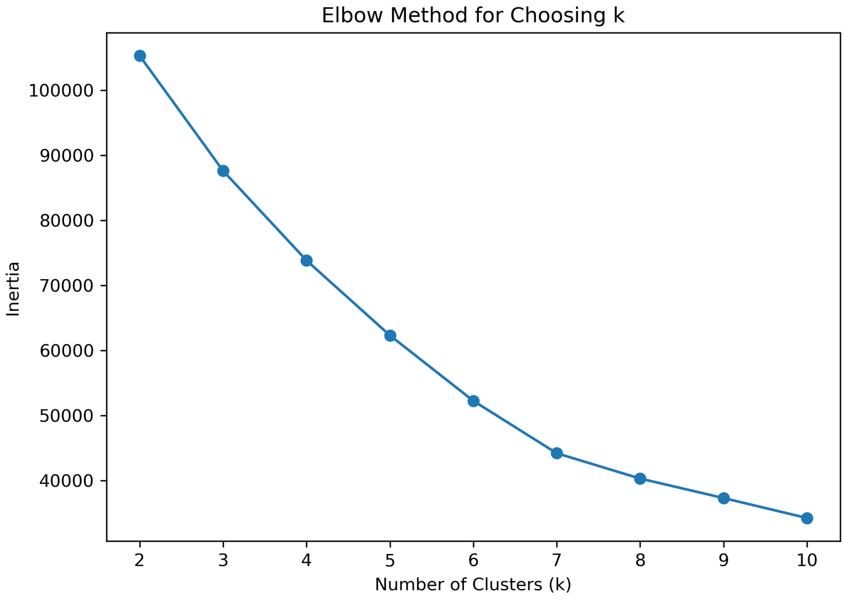

Figure 5: Elbow Method for Selecting the Number of Clusters

K-Means clustering was applied to standardized numeric features. As shown in Figure 5, a noticeable reduction in marginal gains occurs beyond $k = 5$, which was selected as the optimal number of clusters. The silhouette score of approximately 0.33 indicates reasonable cluster separation.

6.2 Regression Models

Linear regression was used as a baseline model, followed by a Random Forest Regressor to capture non-linear relationships and feature interactions.

7 Results

7.1 Clustering Results

Five distinct clusters were identified, representing typical listings, budget listings, high-demand listings, professional hosts, and regulation-driven long-term listings.

Table 1: Cluster Summary Statistics

Cluster	Size	Avg Log Price	Avg Min Nights	Avg Reviews	Avg Reviews/Month	Avg Availability
0	1029	6.06	31.24	0.17	0.01	224.99
1	5873	4.92	25.08	26.86	0.64	115.70
2	11910	5.00	28.80	16.59	0.38	321.48
3	2068	5.22	9.49	204.43	4.26	236.69
4	54	5.11	370.93	17.56	0.18	302.57

Table 1 summarizes the characteristics of each cluster. Clusters differ notably in terms of price, minimum stay requirements, review activity, and availability. These differences enable meaningful interpretation of listing segments, such as high-demand short-stay listings, long-term rental listings, and professionally managed properties.

7.2 Regression Results

Table 2 compares regression model performance.

Table 2: Regression Model Performance

Model	R^2	RMSE (log price)
Linear Regression	0.50	0.55
Random Forest Regressor	0.72	0.41

Table 2 shows that the Random Forest Regressor achieves substantially stronger predictive performance than the linear regression baseline. The linear model explains about 50% of the variance in log-transformed prices ($R^2 = 0.50$), while the Random Forest improves explanatory power to about 72% ($R^2 = 0.72$) and reduces prediction error (RMSE from 0.55 to 0.41 on the log-price scale). This improvement suggests that Airbnb pricing is not well-approximated by a purely linear relationship and instead depends on non-linear effects and interactions among location, room type, demand indicators, and availability constraints.

7.3 Feature Importance Analysis

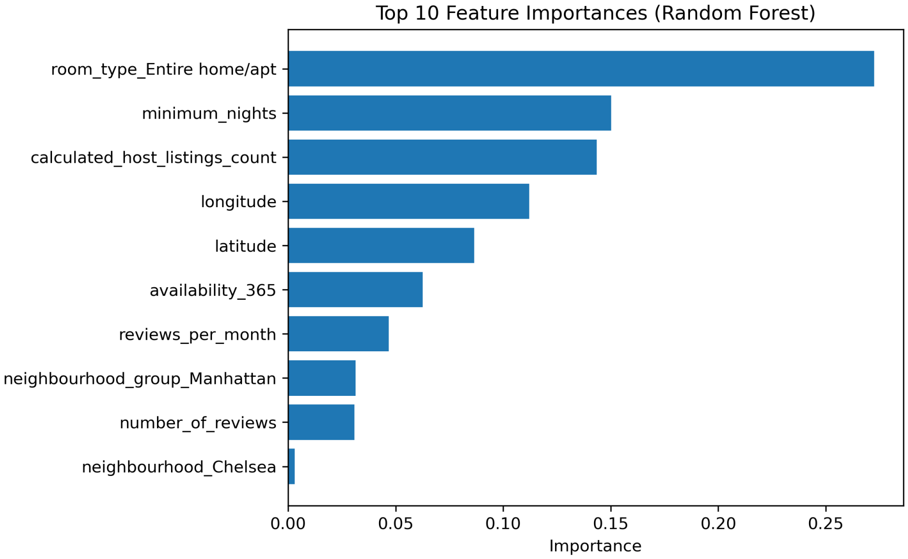

Figure 6: Top 10 Feature Importances from the Random Forest Regressor

Figure 6 summarizes the most influential predictors in the Random Forest model. Room type (especially entire home/apartment) emerges as a dominant driver of price, reflecting the premium associated with full-unit rentals. Minimum night requirements and host listing count also contribute meaningfully, indicating that stay-length policies and professional hosting behavior are related to pricing strategies. Geographic variables (latitude and longitude) rank highly, supporting the strong role of location within NYC. Overall, the feature importance results align with the EDA findings and provide an interpretable explanation for why the non-linear model outperforms the linear baseline.

8 Discussion

Results indicate that Airbnb pricing is influenced by complex interactions among location, room type, availability, and host behavior. The superior performance of the Random Forest model confirms the presence of non-linear pricing dynamics in the New York City Airbnb market.

9 Limitations and Future Work

This study is limited by the absence of amenities, textual descriptions, and seasonal effects. Future work could incorporate natural language processing, time-series analysis, and classification-based pricing strategies.

10 Conclusion

This project demonstrates the application of data mining techniques to analyze Airbnb pricing in New York City. Clustering reveals meaningful market segments, while regression models highlight the importance of non-linear relationships in price prediction.

11 Tools Used

The following tools and technologies were used in this project:

- Python 3.12
- Jupyter Notebook (Anaconda distribution)
- pandas and NumPy for data manipulation
- matplotlib and seaborn for visualization
- scikit-learn for machine learning and model evaluation
- LaTeX (Overleaf) for report preparation
- Google Slides for project presentation

12 References

[1] Inside Airbnb. *New York City Airbnb Data*. Available at: `https://insideairbnb.com/get-the-data/`

[2] Scikit-learn Documentation. `https://scikit-learn.org`