Knowledge Discovery in Databases (KDD) Analysis of Airbnb Listings in New York City

Abstract

This research paper presents a comprehensive analysis of Airbnb listings in New York City using the Knowledge Discovery in Databases (KDD) methodology. By leveraging a dataset of 48,895 Airbnb listings, we explore various features influencing pricing and availability, identify patterns, and develop predictive models for listing prices. The study aims to provide insights into the short-term rental market dynamics and demonstrate the application of KDD in extracting valuable knowledge from large datasets.

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 KDD process to analyze Airbnb listings in NYC, aiming to uncover insights that could benefit hosts, guests, and policymakers.

2. Methodology

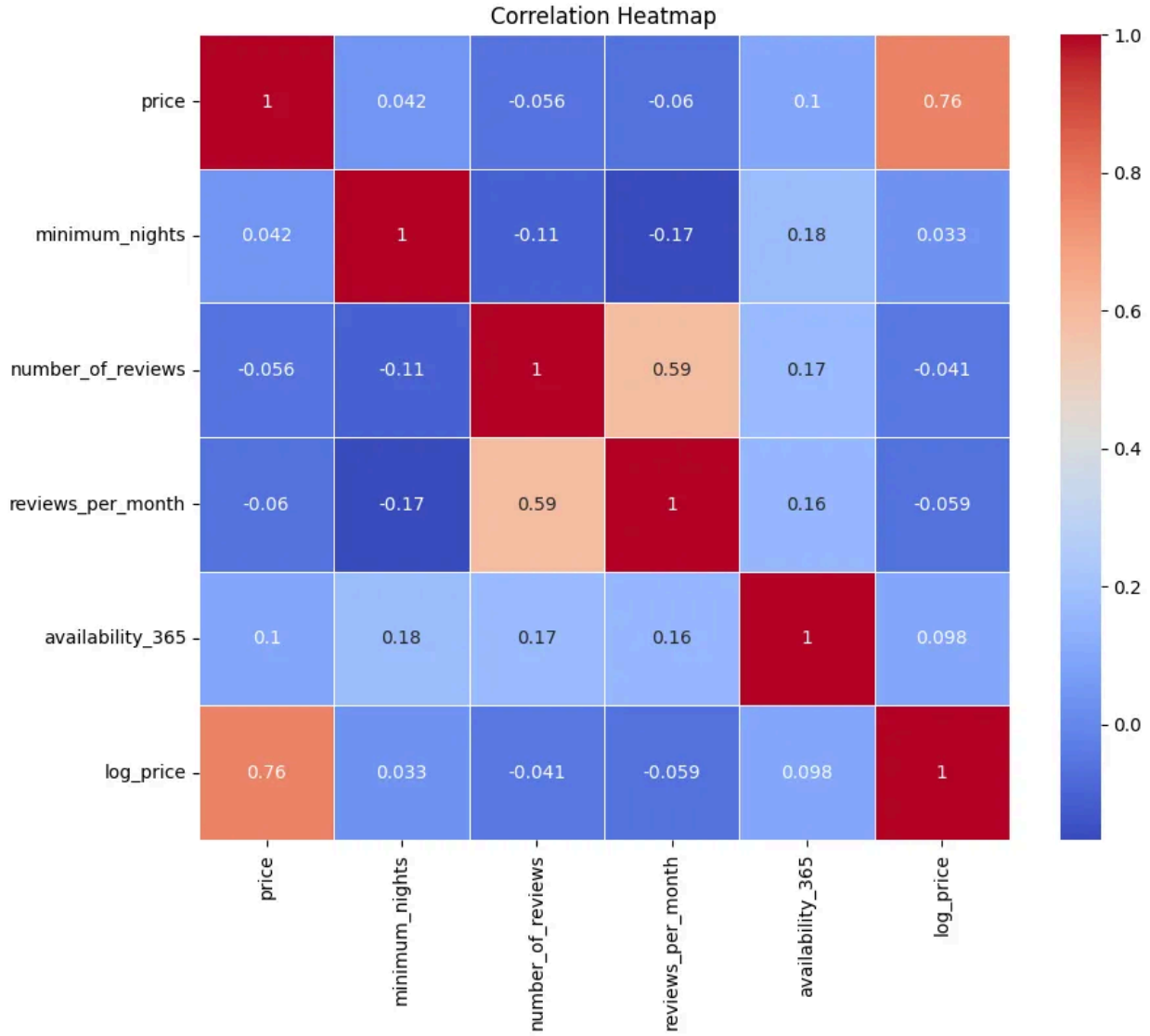
The KDD process consists of five main phases: Selection, Preprocessing, Transformation, Data Mining, and Interpretation. Each phase is detailed below.

2.1 Selection

The initial dataset contained information on 48,895 Airbnb listings in New York City. After careful consideration, the following key features were selected for analysis:

- 'neighbourhood_group'
- 'room_type'
- 'price'
- 'minimum_nights'

- 'number_of_reviews'
- 'reviews_per_month'
- 'availability_365'



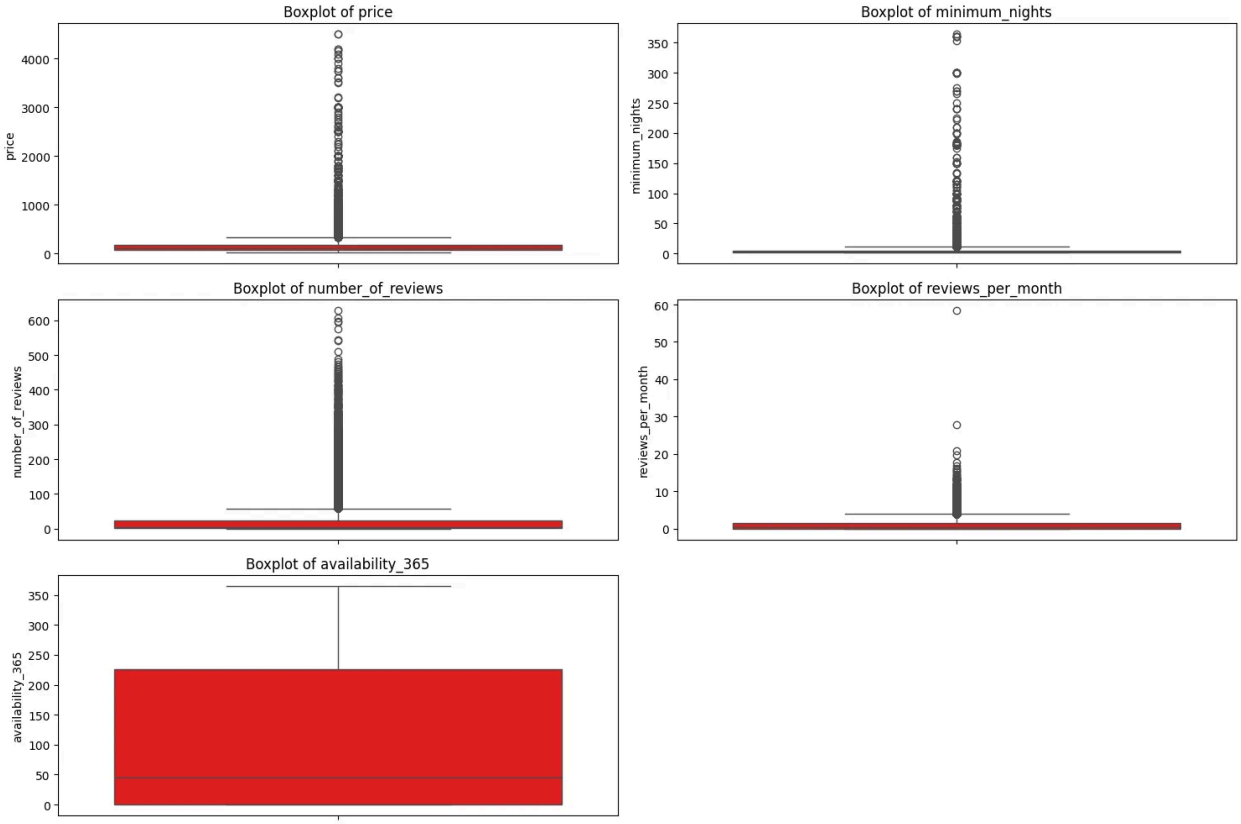
These features were chosen based on their potential impact on listing prices and relevance to understanding market dynamics.

2.2 Preprocessing

Data cleaning and preparation involved several steps:

1. Handling missing values:
    - 10,052 missing values in 'reviews_per_month' were filled with 0.
    - No missing values in other selected features.

2. Removing invalid entries:
    - Listings with prices less than or equal to $0 were removed.
3. Outlier detection and removal:
    - Listings with prices above $5,000 per night were excluded.
    - Listings with 'minimum_nights' greater than 365 were removed.

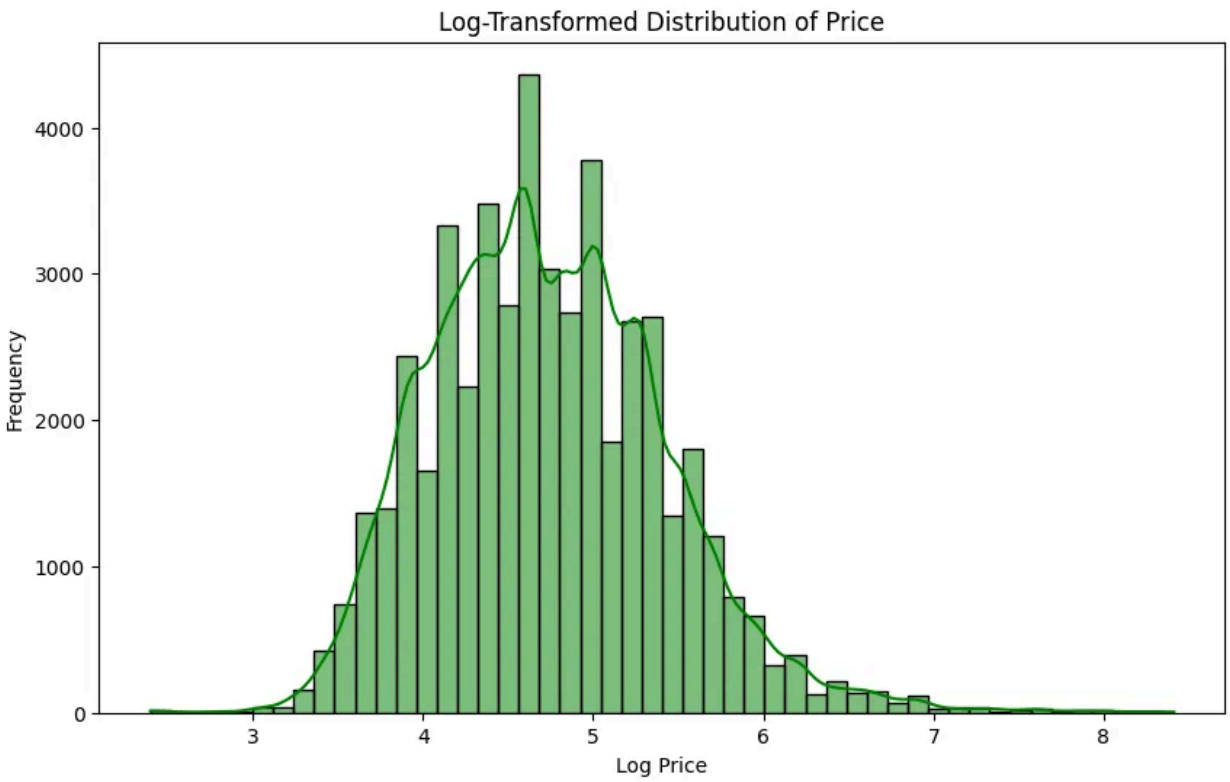


4. Log transformation of the 'price' variable to address skewness in the distribution.

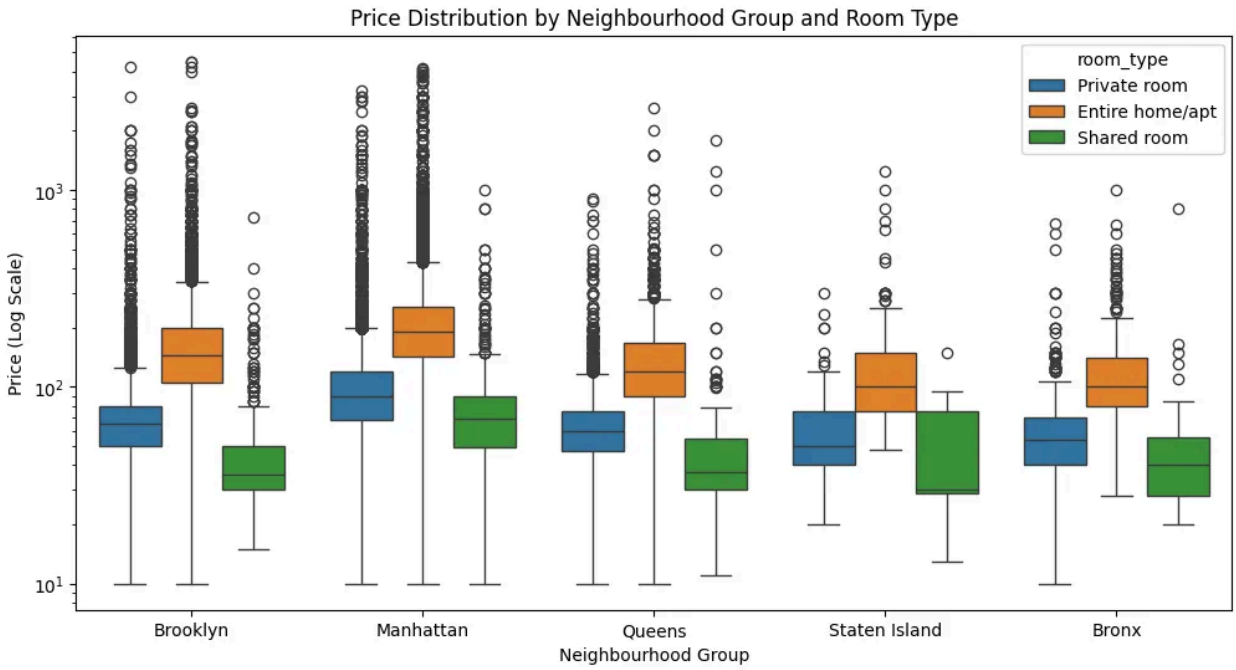
2.3 Transformation

Data transformation included:

1. Standardization of numerical features using StandardScaler:
    - 'minimum_nights'
    - 'number_of_reviews'
    - 'reviews_per_month'
    - 'availability_365'



2. One-hot encoding of categorical variables:
    - 'neighbourhood_group'
    - 'Room_type'



3. Splitting the data into features (X) and target (y):
    - X shape: (48815, 6)
    - y shape: (48815, 6)
4. Further splitting into training (80%) and testing (20%) sets.

2.4 Data Mining

Three regression models were implemented to predict Airbnb listing prices:

1. Linear Regression
2. Random Forest Regressor
3. Gradient Boosting Regressor (results not provided in the search results)

Model performance was evaluated using several metrics:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R-squared (R^2)

Results:

1. Linear Regression:
    - MSE: 24361.21

    - MAE: 68.03
    - R-squared: 0.1569
2. Random Forest:
    - MSE: 27987.28
    - MAE: 70.25
    - R-squared: 0.0315
3. Baseline Model (Mean Price Prediction):
    - MSE: 28896.68
    - MAE: 85.44
    - R-squared: -0.0000064

2.5 Interpretation

Key findings from the analysis include:

1. **Price Distribution:** The average price is $152.72 per night, with a median of $106. The distribution is right-skewed, indicating a wide range of prices with some high-end outliers.
2. **Geographical Insights:** Manhattan and Brooklyn have the highest concentration of listings. Average prices vary significantly by neighborhood group.
3. **Room Type Impact:** Entire homes/apartments (25,409 listings) and private rooms (22,326 listings) are more common than shared rooms (1,160 listings). Entire homes/apartments generally command higher prices.
4. **Availability Patterns:** The average availability is 113 days per year, suggesting many hosts use Airbnb as a part-time income source.
5. **Model Performance:** The Linear Regression model showed the best performance among the tested models, explaining about 15.69% of the variance in listing prices. However, the relatively low R-squared values across all models suggest that predicting Airbnb prices is a complex task influenced by many factors beyond the basic listing characteristics analyzed.

3. Results and Discussion

3.1 Exploratory Data Analysis

The analysis revealed several key insights:

- Price distribution is heavily right-skewed, with a mean price of $152.72 and a median of $106.

- Manhattan and Brooklyn have the highest concentration of listings.
- Entire homes/apartments (25,409 listings) and private rooms (22,326 listings) are more common than shared rooms (1,160 listings).
- The average availability is 113 days per year, suggesting many hosts use Airbnb as a part-time income source.

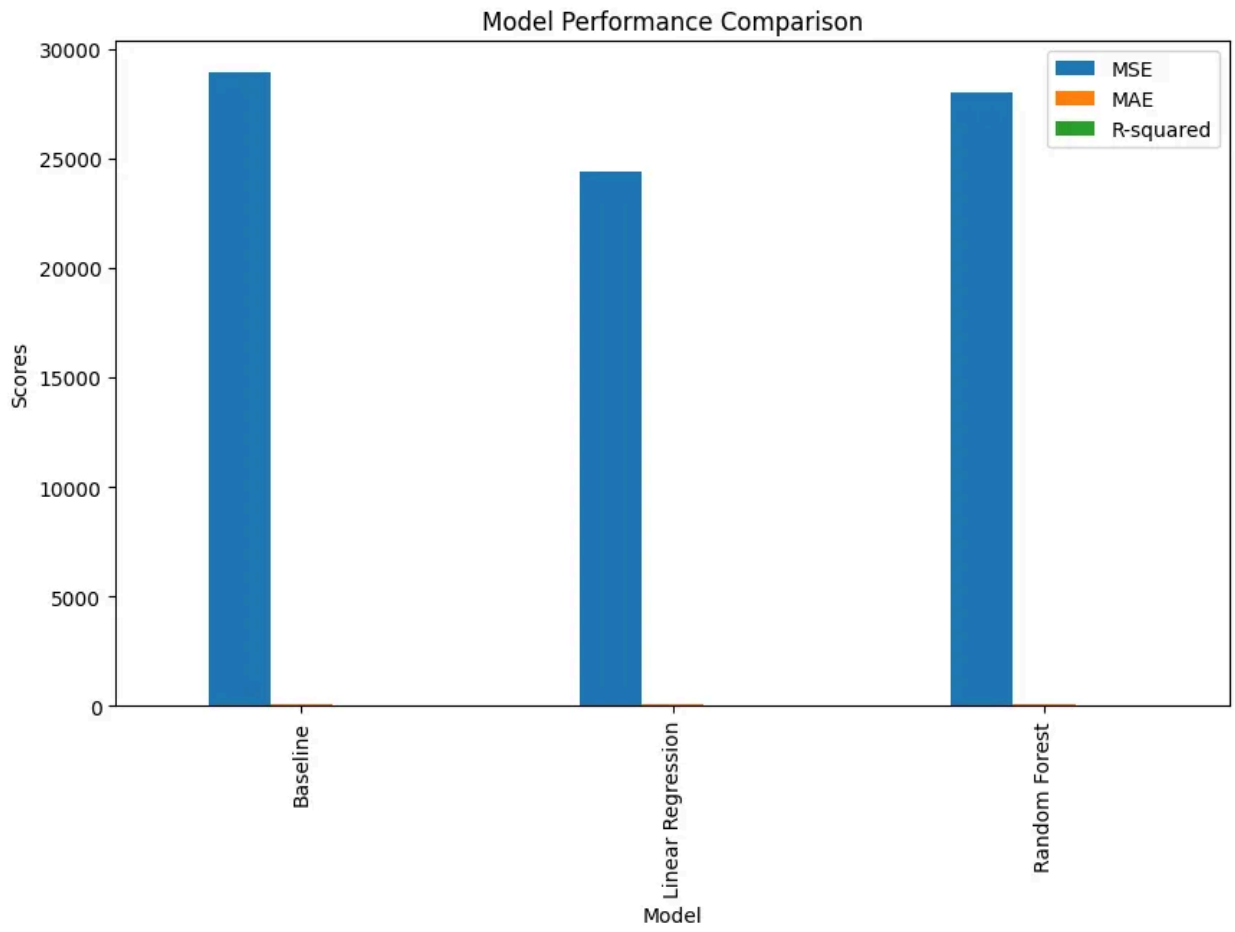
3.2 Model Performance

The performance of each model was as follows:

1. Linear Regression:
   - MAE: 68.03
   - MSE: 24361.21
   - RMSE: 156.08
   - R^2: 0.1569

2. Random Forest:
   - MAE: 70.25
   - MSE: 27987.28
   - RMSE: 167.29
   - R^2: 0.0315

3. Baseline Model (Mean Price Prediction):
   - MAE: 85.44
   - MSE: 28896.68
   - RMSE: 169.99
   - R^2: -0.0000064



The Linear Regression model showed the best performance, explaining about 15.69% of the variance in listing prices.

4. Conclusions

This KDD analysis of New York City Airbnb listings provides valuable insights into the short-term rental market:

1. The market offers a wide range of accommodations, catering to diverse traveler needs and budgets.
2. Location plays a crucial role in determining listing prices, with Manhattan generally commanding higher prices.
3. Entire homes/apartments and private rooms are the most common types of listings, with entire homes/apartments generally fetching higher prices.
4. The moderate performance of predictive models indicates that Airbnb pricing is influenced by many factors beyond basic listing characteristics.

These insights can inform decision-making for hosts, guests, and policymakers in the Airbnb ecosystem.

5. Future Work

Future research could focus on:
1. Incorporating additional features such as amenities and proximity to attractions.
2. Exploring temporal patterns to capture seasonal trends.
3. Applying more advanced machine learning techniques to improve predictive accuracy.
4. Conducting in-depth geospatial analysis to better understand neighborhood effects on pricing.

This study demonstrates the value of the KDD process in extracting meaningful insights from large datasets, providing a foundation for further research and decision-making in the dynamic short-term rental market of New York City.

Citations:

[1]
https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/31097934/5ff5ef16-bd22-47db-9c82-3e9baeb83519/KDDColab2.pdf
[2]
https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/31097934/40560de3-6fae-4892-97bc-de1e7436697c/KDDColab1.pdf