Analyzing Airbnb Listings in New York City: A CRISP-DM Approach

Abstract

This research paper presents a comprehensive analysis of Airbnb listings in New York City using the Cross-Industry Standard Process for Data Mining (CRISP-DM) 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. Business Understanding

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 aims to:

1. Understand the factors influencing Airbnb listing prices in New York City
2. Develop predictive models to estimate listing prices
3. Provide insights that could benefit hosts, guests, and policymakers

2. Data Understanding

2.1 Data Collection

The dataset used in this study contains 48,895 Airbnb listings in New York City. It includes various features such as:

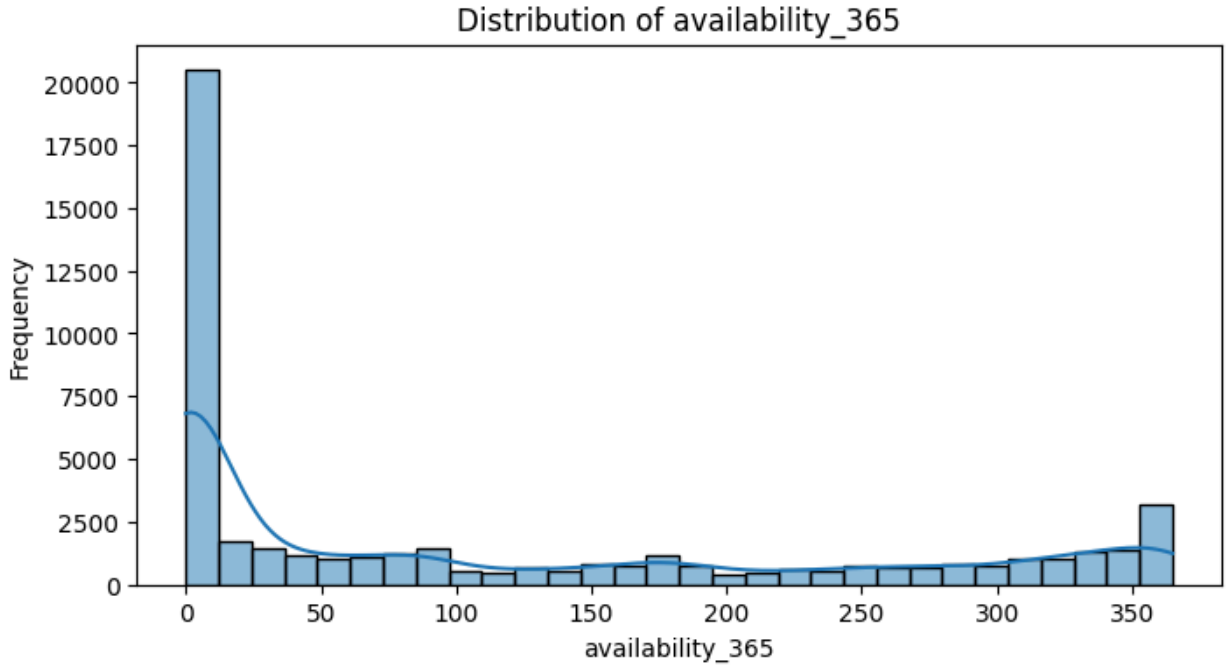
- Listing ID
- Host information
- Location details (neighborhood, latitude, longitude)

- Room type
- Price
- Minimum nights required
- Number of reviews
- Availability

2.2 Data Exploration

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

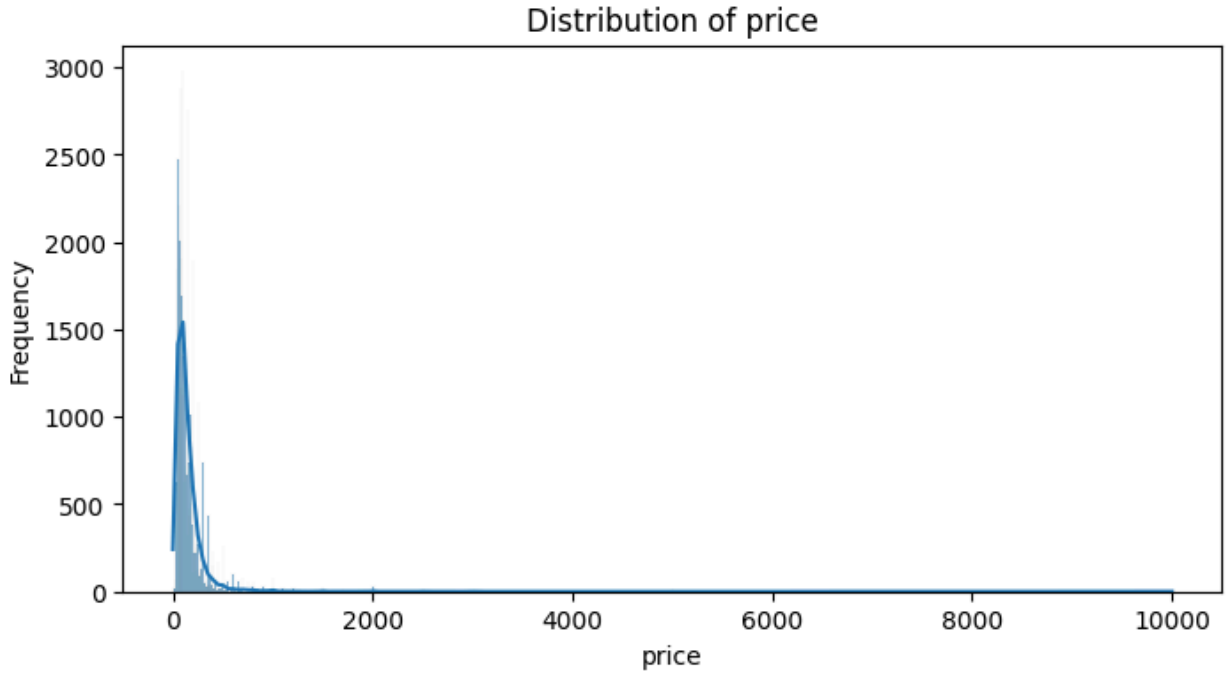


This distribution indicates a preference for entire homes and private rooms, with shared rooms being a less common option.

2.2.2 Price Analysis

- Average price: $152.72 per night
- Median price: $106 per night

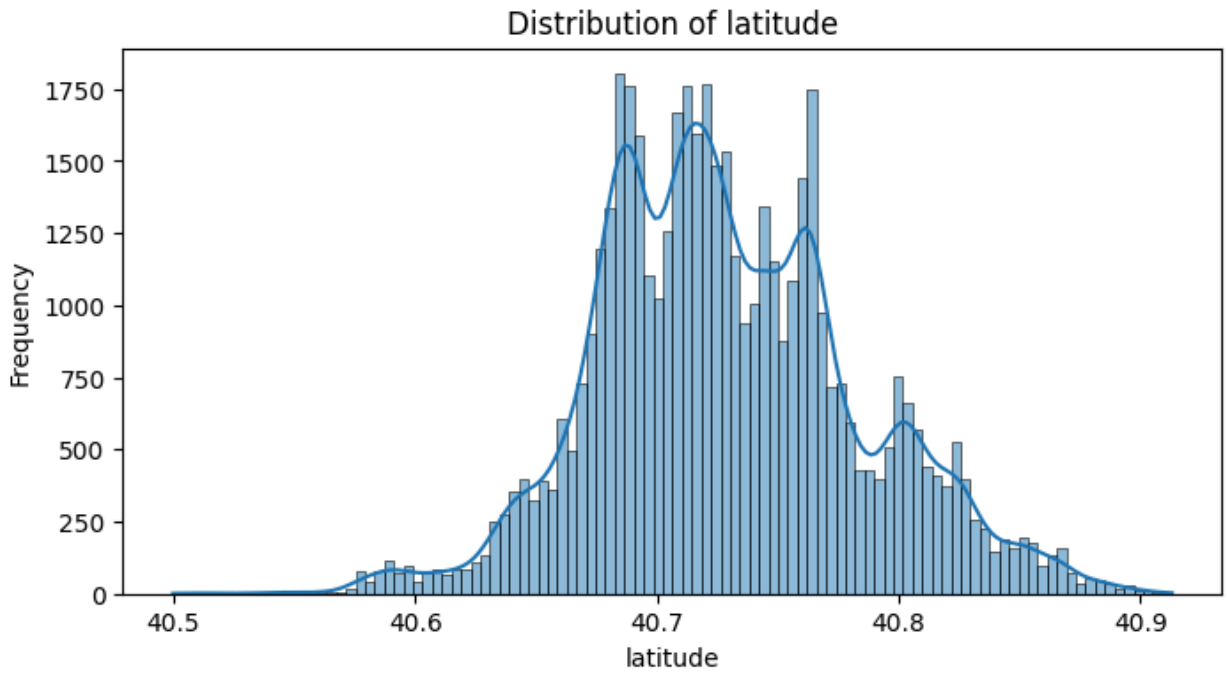
- Price range: $0 to $10,000 per night

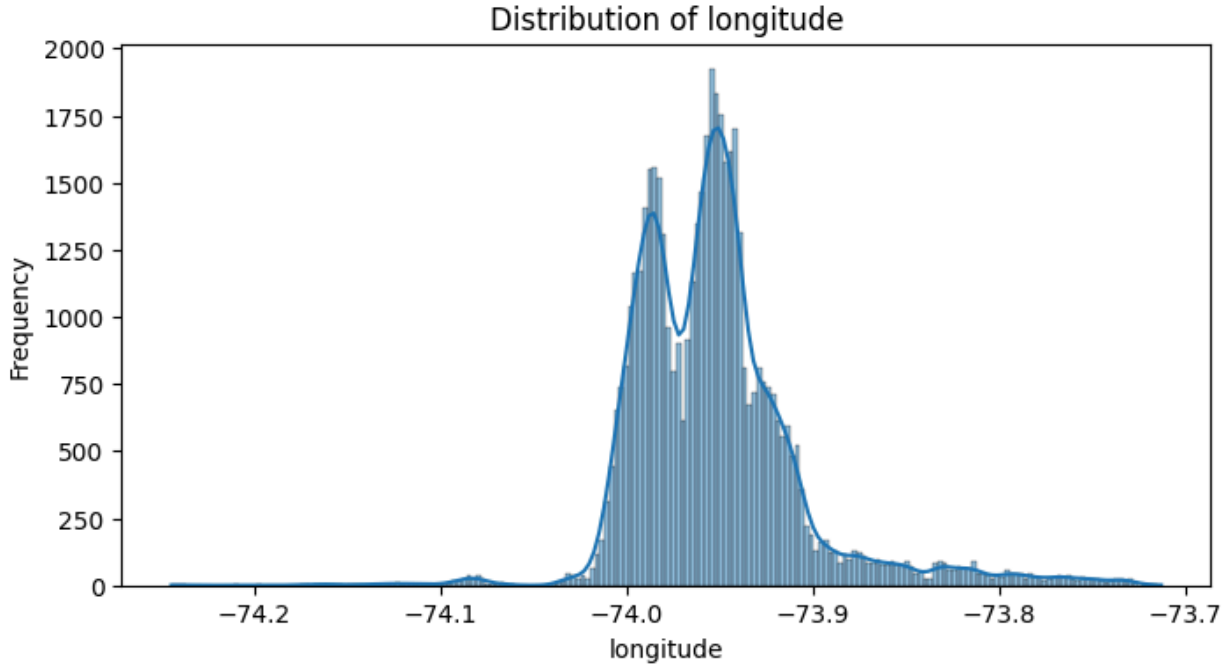


The wide range in prices reflects the diversity of accommodations available, from budget-friendly options to luxury rentals.

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
- Queens: Elmhurst

2.2.4 Booking Patterns and Availability

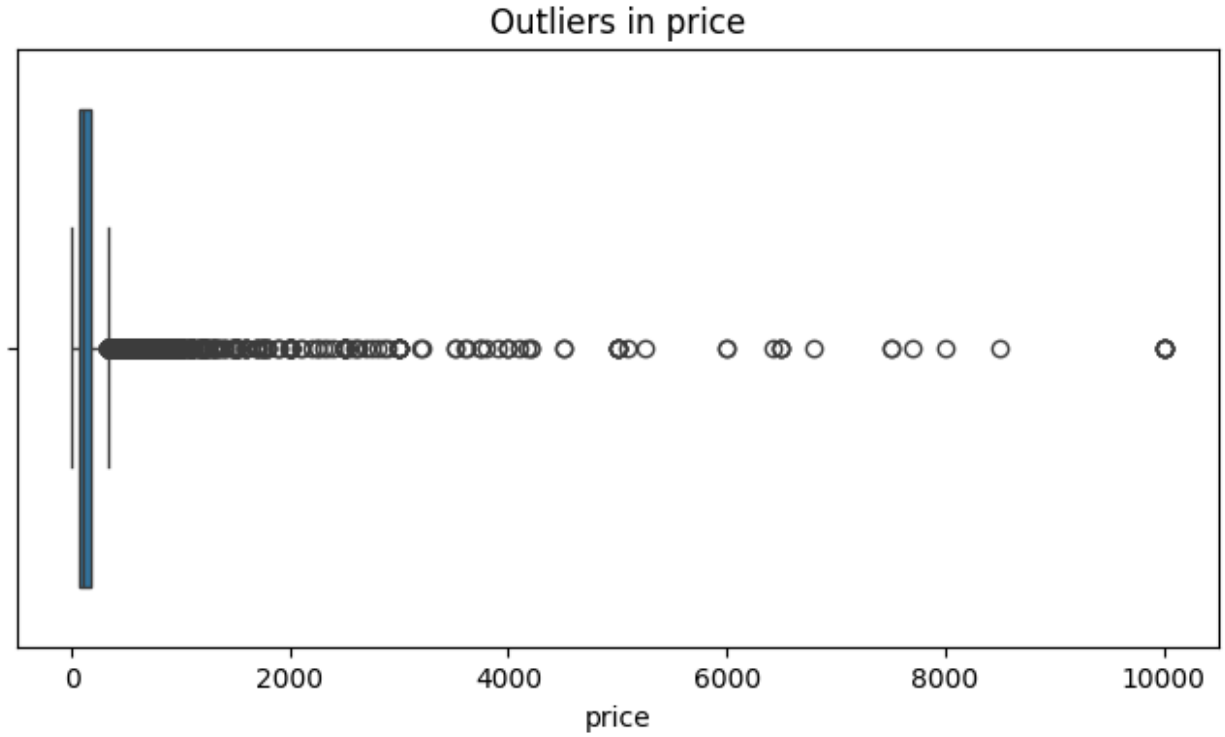
- Average minimum stay: 7 nights
- Maximum minimum stay: 1,250 nights (likely an outlier)
- Average number of reviews: 23
- Maximum number of reviews for a listing: 629
- Average reviews per month: 1.37
- Average availability: 113 days out of 365

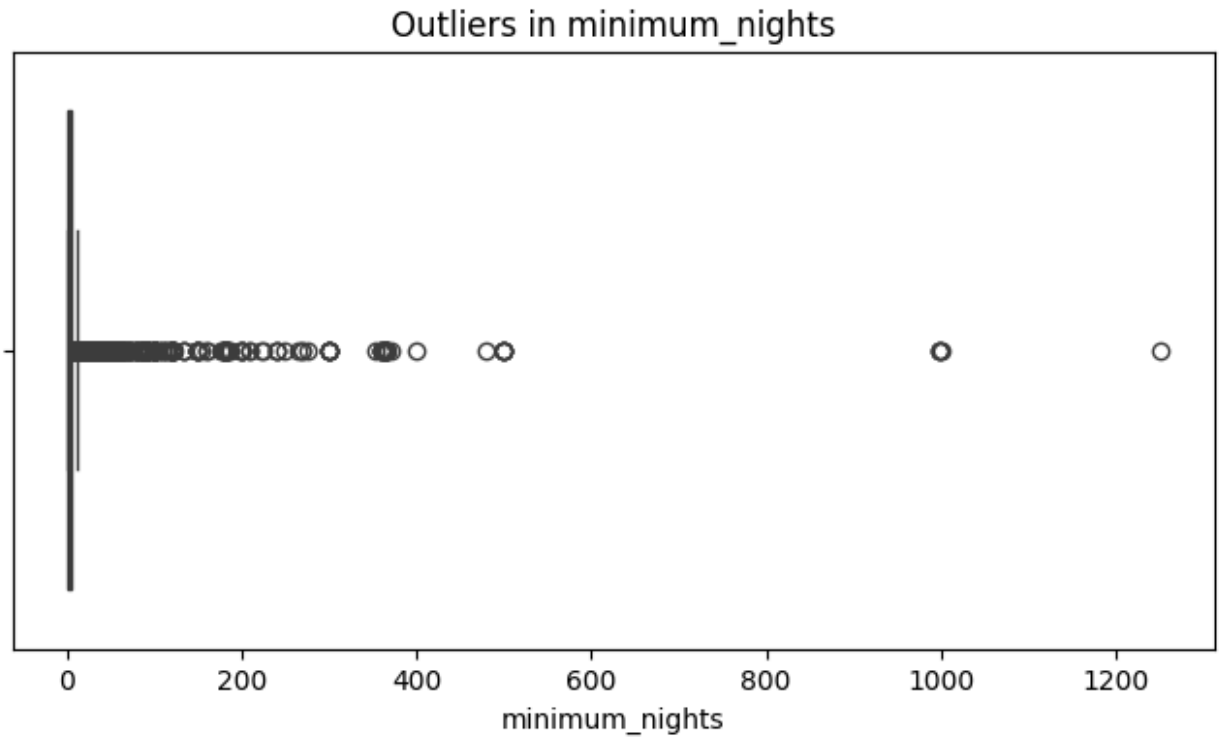
These metrics provide insights into booking frequency, guest satisfaction, and host management practices.

2.3 Data Quality Verification

Missing values were identified in the following columns:
- 'name': 16 missing values
- 'host_name': 21 missing values
- 'last_review': 10,052 missing values
- 'reviews_per_month': 10,052 missing values

Potential outliers were observed in several numerical columns, particularly in 'price' and 'minimum_nights'.





3. Data Preparation

3.1 Data Cleaning

To address data quality issues, the following steps were taken:

1. Missing values in 'reviews_per_month' were filled with 0, assuming no reviews for those listings.
2. Rows with missing values in 'name' or 'host_name' were dropped, as these are crucial identifiers.
3. The 'last_review' column was converted to datetime format.

3.2 Data Transformation

3.2.1 Feature Scaling

Numerical features were standardized using StandardScaler:
- 'latitude'
- 'longitude'
- 'minimum_nights'

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

3.2.2 Categorical Encoding

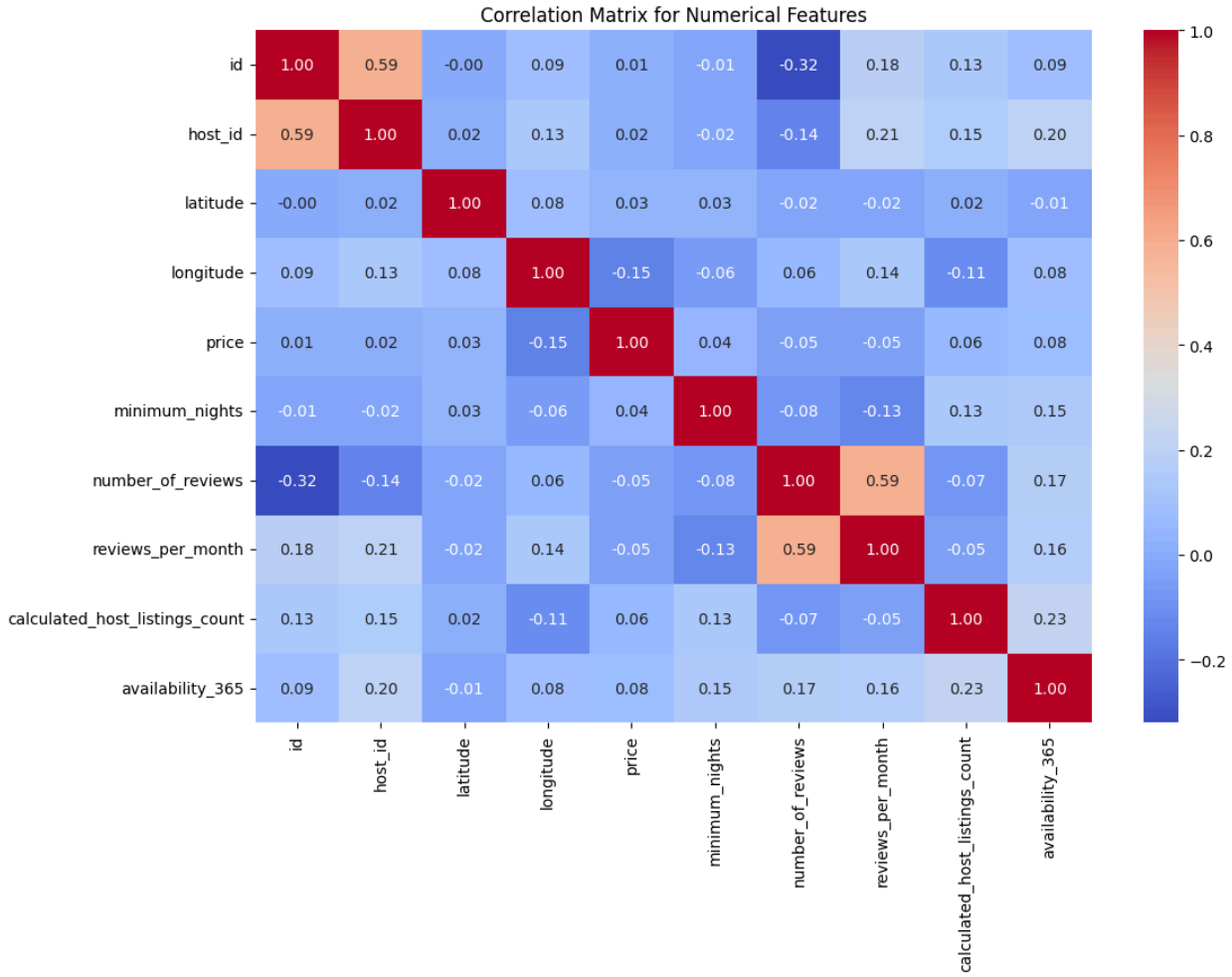
One-hot encoding was applied to:
- 'neighbourhood_group'
- 'room_type'

3.3 Feature Selection

Based on correlation analysis and domain knowledge, the following features were selected for modeling:

- 'latitude'
- 'longitude'
- 'minimum_nights'
- 'number_of_reviews'
- 'reviews_per_month'
- 'calculated_host_listings_count'
- 'availability_365'
- 'neighbourhood_group_Brooklyn'
- 'neighbourhood_group_Manhattan'
- 'neighbourhood_group_Queens'
- 'neighbourhood_group_Staten Island'
- 'room_type_Private room'
- 'room_type_Shared room'

The target variable was set as 'price'.



4. Modeling

4.1 Model Selection

Three regression models were implemented to predict Airbnb listing prices:

1. Linear Regression
2. Random Forest Regressor
3. Gradient Boosting Regressor

4.2 Model Training and Evaluation

The dataset was split into training (80%) and testing (20%) sets. Each model was trained on the training set and evaluated using the following metrics:

1. Mean Absolute Error (MAE)
2. Mean Squared Error (MSE)
3. Root Mean Squared Error (RMSE)
4. R-squared (R^2)

Initial results showed:

1. Linear Regression:
   - MAE: 72.9533
   - MSE: 45401.3477
   - RMSE: 213.0759
   - R^2: 0.1004

2. Random Forest:
   - MAE: 68.1695
   - MSE: 48337.0689
   - RMSE: 219.8569
   - R^2: 0.0422

3. Gradient Boosting:
   - MAE: 66.4096
   - MSE: 42906.5154
   - RMSE: 207.1389
   - R^2: 0.1498

4.3 Hyperparameter Tuning

To improve model performance, hyperparameter tuning was performed using RandomizedSearchCV for both Random Forest and Gradient Boosting models.

4.3.1 Random Forest Tuning

The best parameters found for Random Forest were:
- n_estimators: 50
- min_samples_split: 2
- min_samples_leaf: 4
- max_depth: None

Best score (MSE): 58452.77502965988

4.3.2 Gradient Boosting Tuning

The best parameters found for Gradient Boosting were:
- n_estimators: 50
- min_samples_split: 5
- max_depth: 3
- learning_rate: 0.1

Best score (MSE): 58532.44497642151

5. Evaluation

5.1 Model Comparison

After hyperparameter tuning, the models were re-evaluated on the test set. The Gradient Boosting model showed the best overall performance:

- MAE: 68.6852
- MSE: 44972.1174
- RMSE: 212.0663
- R^2: 0.1089

5.2 Feature Importance

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

5.3 Model Performance Visualization

Scatter plots comparing actual vs. predicted prices were created to visualize the models' performance. These plots help identify areas where the models perform well and where they struggle in price prediction.

6. Deployment

6.1 Model Saving

The final chosen model (Gradient Boosting Regressor) was saved using joblib for future use:

```python
import joblib
joblib.dump(optimized_gb, 'final_airbnb_price_model.pkl')
```

This allows for easy loading and deployment of the model in production environments.

7. Conclusions and Insights

1. Market Diversity: The New York City Airbnb market offers a wide range of accommodations, from budget-friendly options to luxury rentals, catering to diverse traveler needs.

2. Geographical Influence: Location plays a significant role in pricing, with certain neighborhoods commanding higher prices. Manhattan and Brooklyn emerge as hotspots for Airbnb listings.

3. Room Type Impact: Entire homes/apartments generally fetch higher prices compared to private or shared rooms, reflecting traveler preferences for privacy and space.

4. Review Dynamics: There's a slight negative correlation between price and number of reviews, suggesting that more affordable listings tend to accumulate more reviews.

5. Availability Patterns: The average availability of 113 days per year indicates that many hosts use Airbnb as a part-time income source rather than a full-time business.

6. Modeling Challenges: The relatively low R-squared values across all models suggest that predicting Airbnb prices is a complex task. Factors not captured in the dataset (e.g., property amenities, local events, seasonality) likely play a significant role in price determination.

7. Ensemble Methods: The superior performance of the Gradient Boosting model highlights the effectiveness of ensemble methods in capturing non-linear relationships and interactions between features.

Citations:

[1]
https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/31097934/366cc4c7-2ebe-4bd0-95e6-72f1aecdf7e3/CRISP-DMColab2.pdf
[2]
https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/31097934/1917f9cd-98b7-4f13-afee-a8fd6aba8d92/CRISP-DMColab1.pdf