New York City Airbnb Data Analysis Report

1.Introduction

This report presents an analysis of the New York City Airbnb dataset to explore how the number of reviews relates to price, availability, and other listing characteristics. The project follows a structured approach, including data cleaning, exploratory data analysis (EDA), and predictive modeling.

2. Data Overview

- **Dataset**: NYC Airbnb Open Data (2019)
- **Total Listings**: 48,895
- **Key Variables**: Price, Availability (days/year), Number of Reviews, Reviews per Month, Room Type, Neighborhood Group

3. Data Cleaning & Preprocessing

- **Missing Values**:
    - Filled missing values in name and host_name with "Unknown"
    - Replaced missing reviews_per_month values with 0
    - Converted last_review to datetime
- **Outlier Removal**:
    - Used IQR method to remove extreme values in price, minimum_nights, and reviews_per_month
    - Reduced max price from **$10,000** to **$334**
    - Limited minimum_nights to **11** (previously 1,250)
    - Capped reviews_per_month at **4.61**

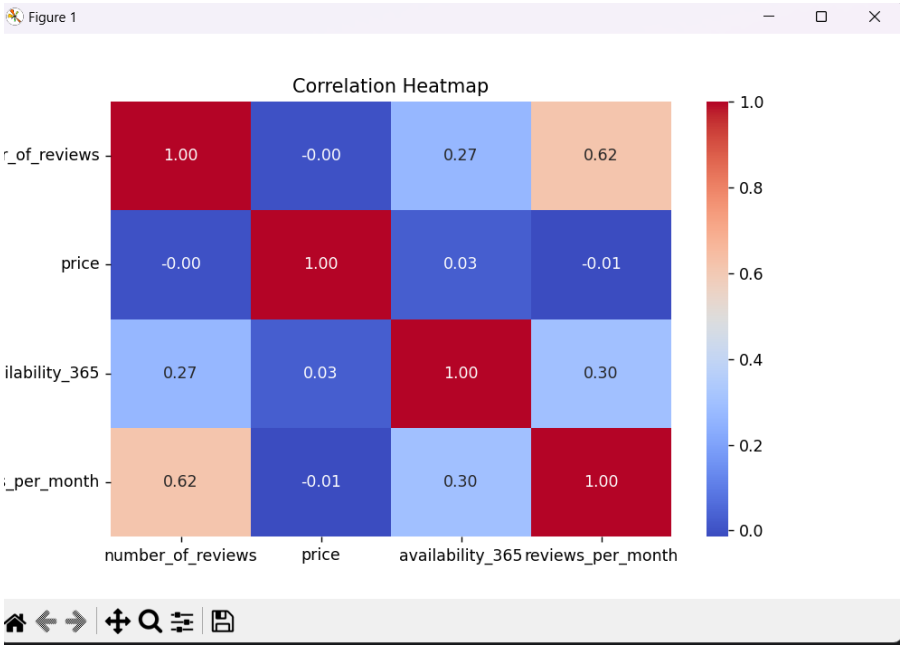
4. Exploratory Data Analysis (EDA)

- **Correlation Analysis**:

- number_of_reviews has a weak negative correlation with price (-0.05)
- Moderate positive correlation between availability_365 and number_of_reviews (~0.33)

- **Visualizations & Insights**:

Correlation Heatmap: Highlights relationships between key variables.

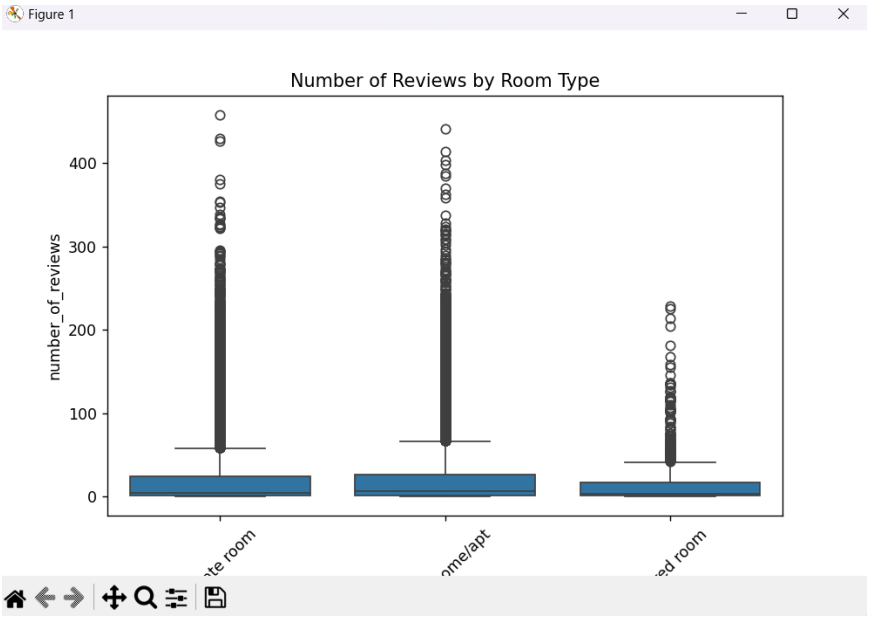


Scatter Plot: Reviews vs Price: Shows no strong relationship between price and reviews.

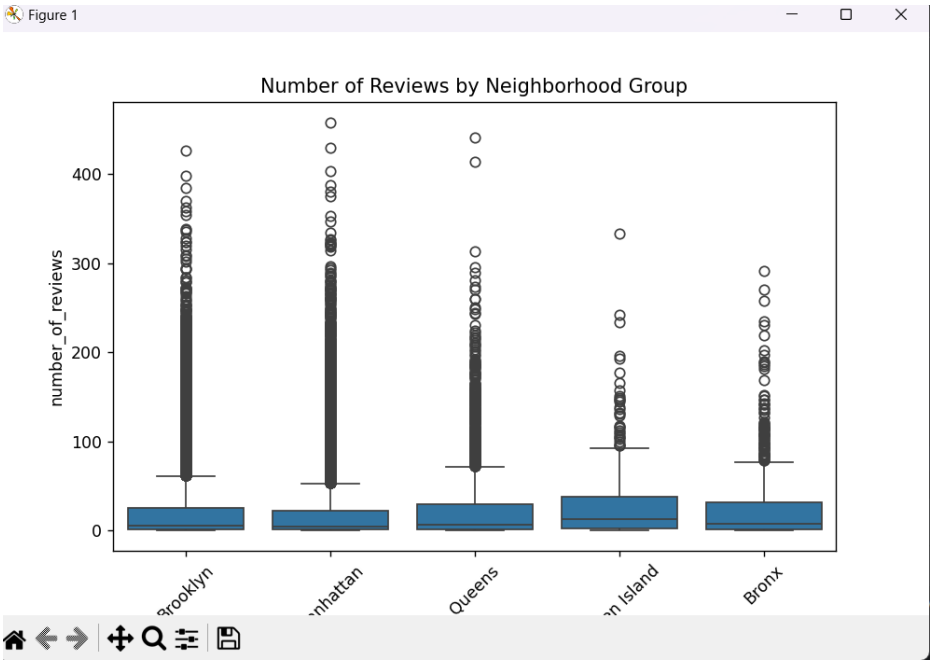
Scatter Plot: Reviews vs Availability: Confirms that listings with more availability receive more reviews.

- **Boxplots**:

Room Type vs Reviews: Private and shared rooms tend to have more reviews than entire apartments.



Neighborhood Group vs Reviews: Listings in Brooklyn and Manhattan receive the most reviews.



5. Predictive Modeling

- **Model Used**: Random Forest Regressor
- **Features Considered**: price, availability_365, reviews_per_month, calculated_host_listings_count

- **Model Performance**:
    - **MAE (Mean Absolute Error)**: 15.73
    - **MSE (Mean Squared Error)**: 969.68
    - **R^2 Score**: 0.45 (Moderate Predictive Power)
- **Key Findings**:
    - Price does not significantly impact the number of reviews.
    - Listings with higher availability and frequent reviews per month attract more guests.

6. Conclusion & Recommendations

- Listings with **higher availability** tend to receive more reviews.
- **Private and shared rooms** attract more reviews than entire apartments.
- **Brooklyn and Manhattan** have the most active listings.
- **Future Work**:
    - Improve predictive accuracy by including additional variables (e.g., sentiment analysis from reviews).
    - Consider seasonal effects on Airbnb bookings.
    - Optimize pricing strategies based on guest engagement data.

7. Reflection

- **Lessons Learned**:
    - Data cleaning is crucial for accurate analysis.
    - Visualizations help uncover trends that raw data does not easily reveal.
    - Feature selection impacts model performance.

- **Challenges**:
    - Handling missing values and extreme outliers.
    - Improving the predictive accuracy of the model.
    - Understanding external factors influencing reviews.