CAPSTONE PROJECT

Airbnb Booking Analysis

CONTENT

- Problem statement
- Data summary
- Data cleaning
- Visualization
- Conclusion

PROBLEM STATEMENT

- This EDA project aims to uncover valuable insights from Airbnb booking data, enabling informed decisions for the client.

DATA SUMMARY

- **Dataset name:** Airbnb_NYC_2019.csv

- **Dataset shape:** 48895 rows, 16 columns

- **Columns:** ‘id’, ‘name’, ‘host_id’, ‘host_name’, ‘neighbourhood_group’, ‘neighbourhood’, ‘latitude’, ‘longitude’, ‘room_type’, ‘price’, ‘minimum_nights’, ‘number_of_reviews’, ‘last_review’, ‘reviews_per_month’, ‘calculated_host_listings_count’, ‘availability_365’

DATA CLEANING

- **No duplicate rows**

- **Excluded columns:** 'name','host_name','last_review’

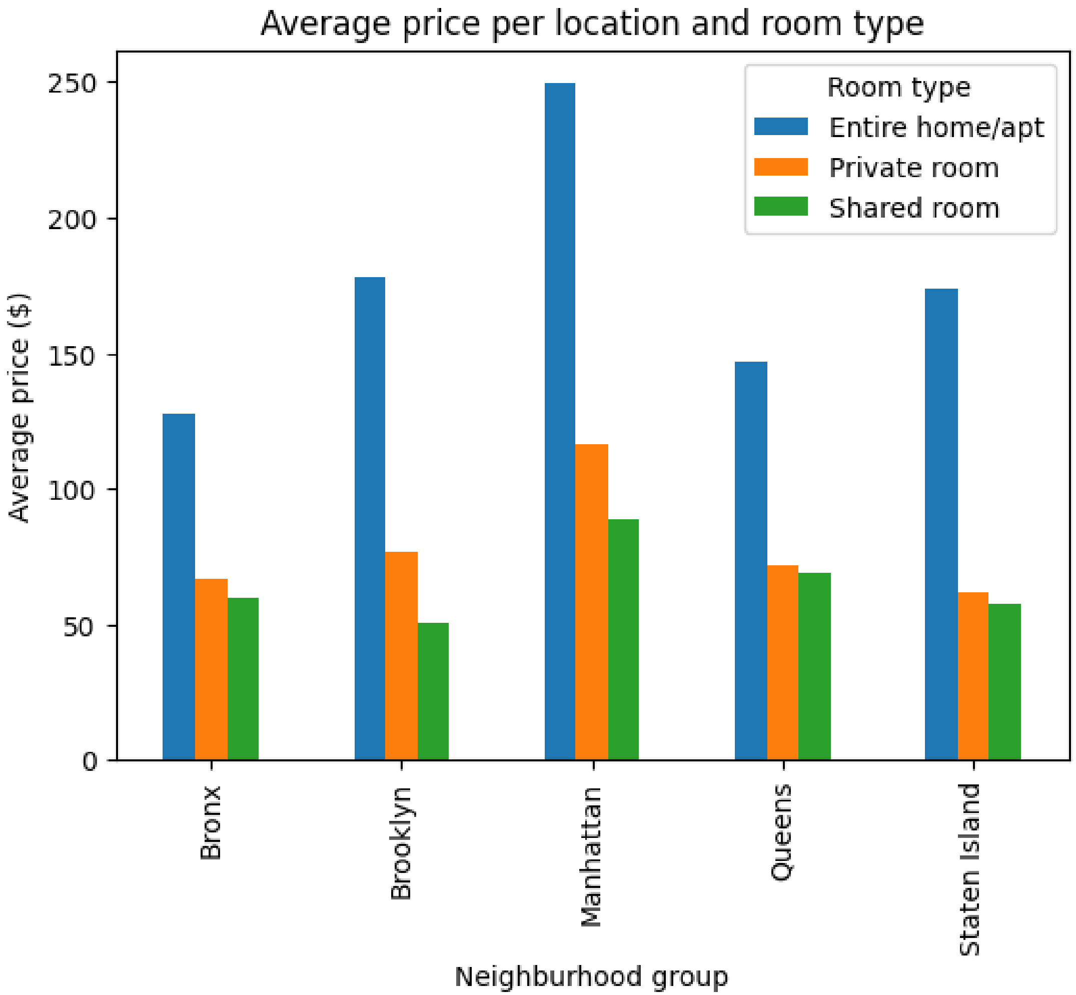
- **Excluded rows with zero price**

- **Missing/Null values:** 20.56% missing values in ‘reviews_per_month’ are replaced with zero.

- **Final dataset shape:** 48884 rows and 13 columns

VISUALIZATION

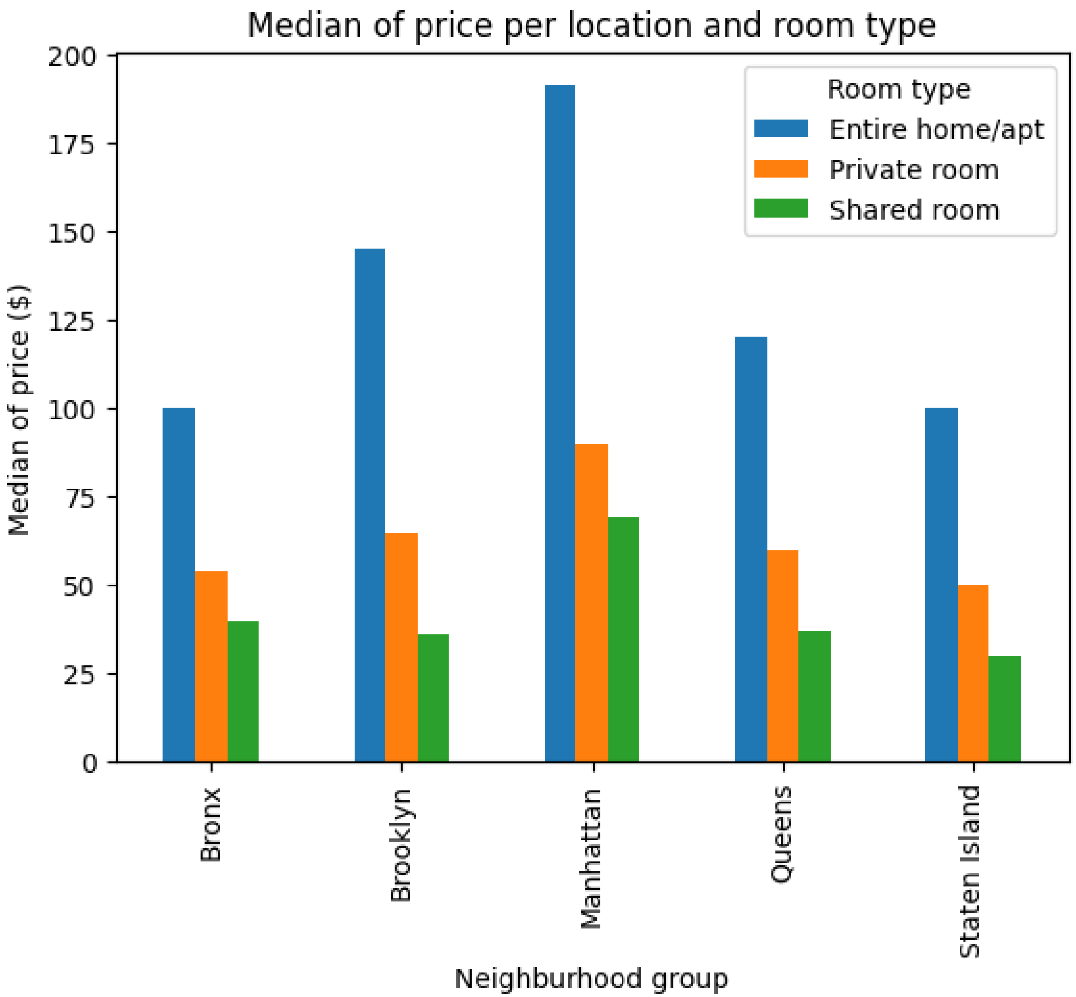
1. Mean price for all room type across various location- Bar Plot



- Mean price for entire home/apartment type room services is consistently higher across all locations
- Manhattan exhibits higher mean prices for all types of room services.

VISUALIZATION

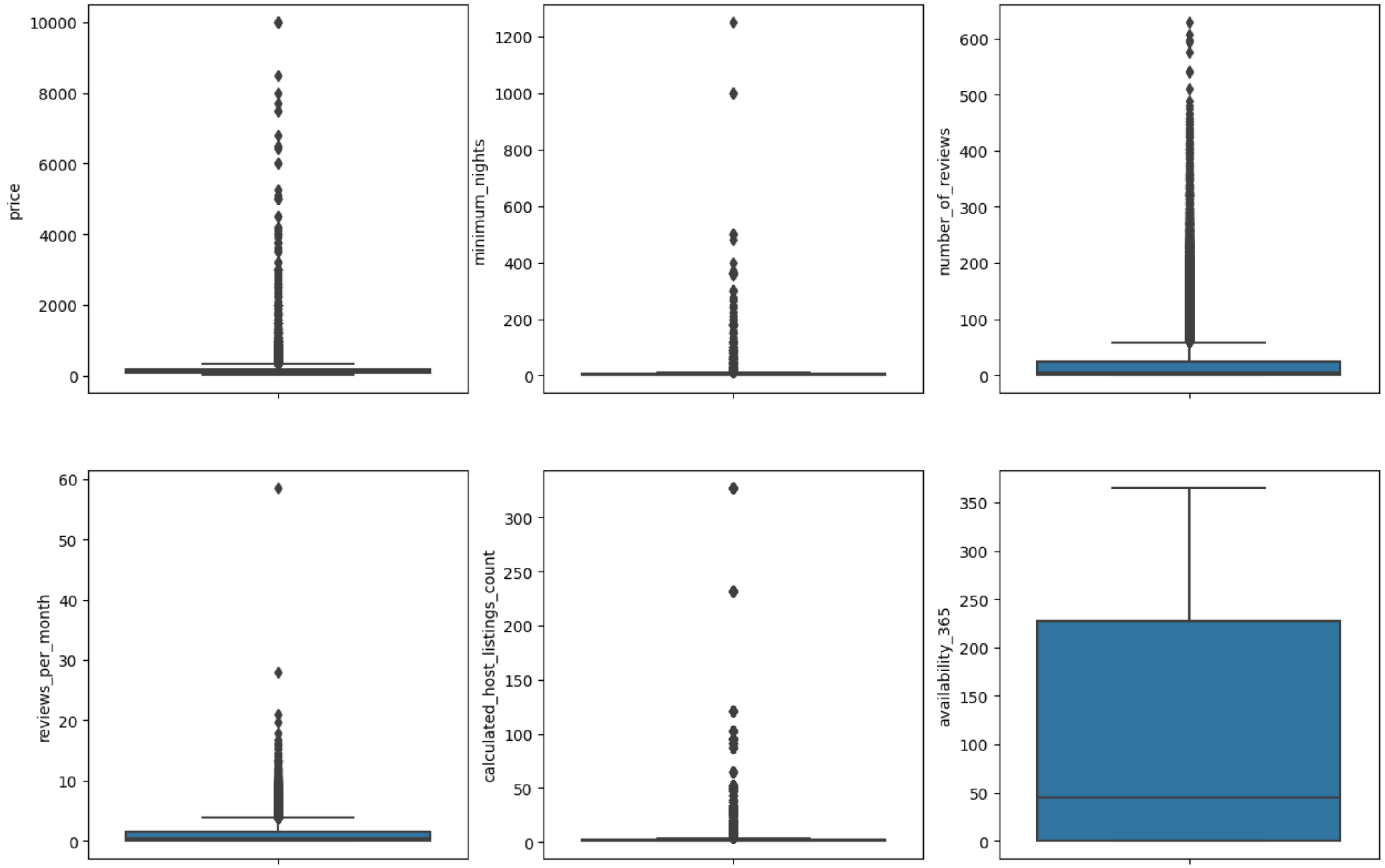
2. Median price for all room type across all locations- Bar plot



- Median price for entire home/apartment type room services tends to be higher across all neighborhood groups
- Median price is lower for shared rooms across the different neighbourhood groups

VISUALIZATION

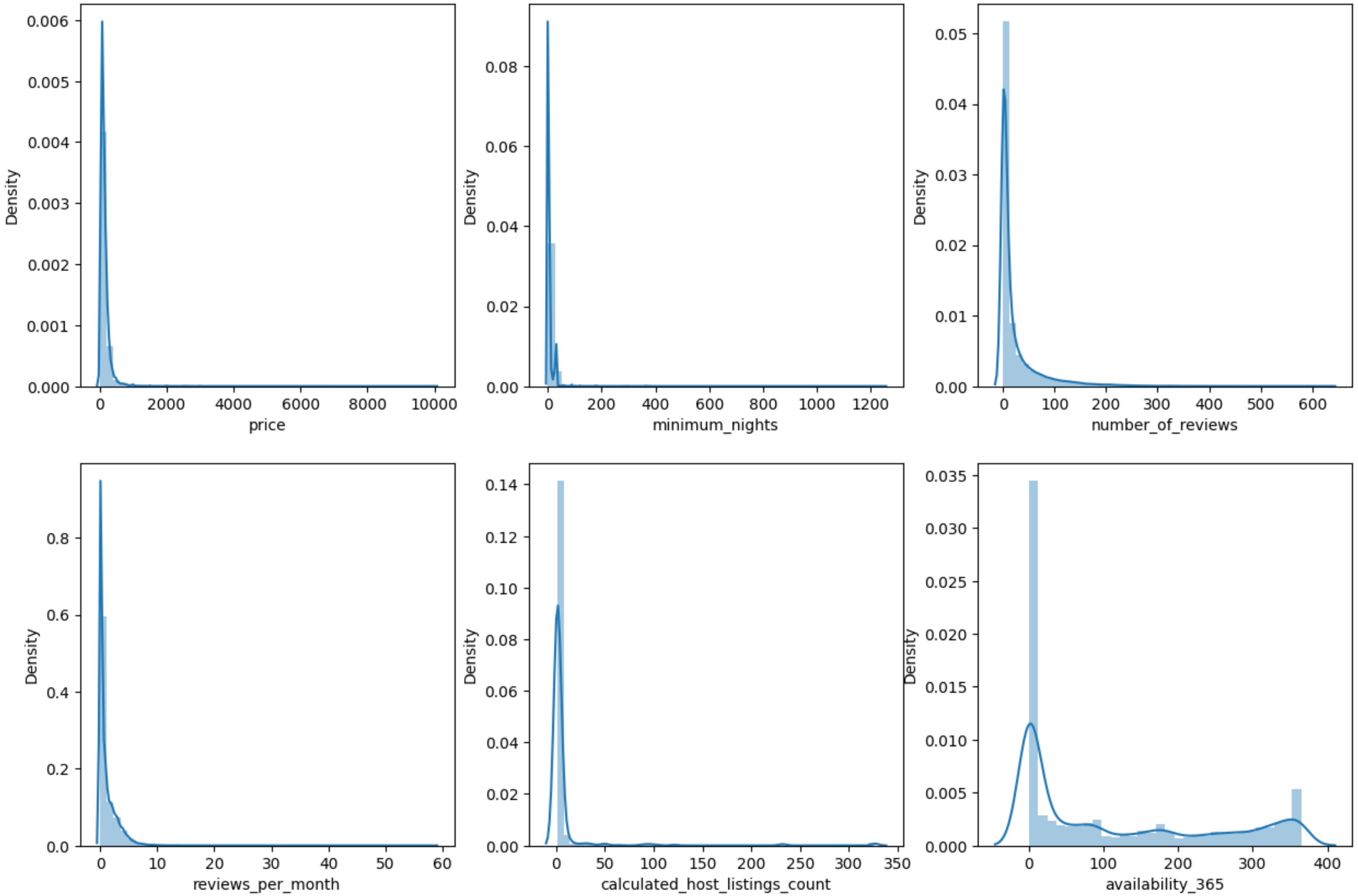
3. Distribution of numeric features in the dataset - Box plot



- Most of the values are far from each other except the variables 'availability_365’.
- Dataset needs pre-processing before applying model

VISUALIZATION

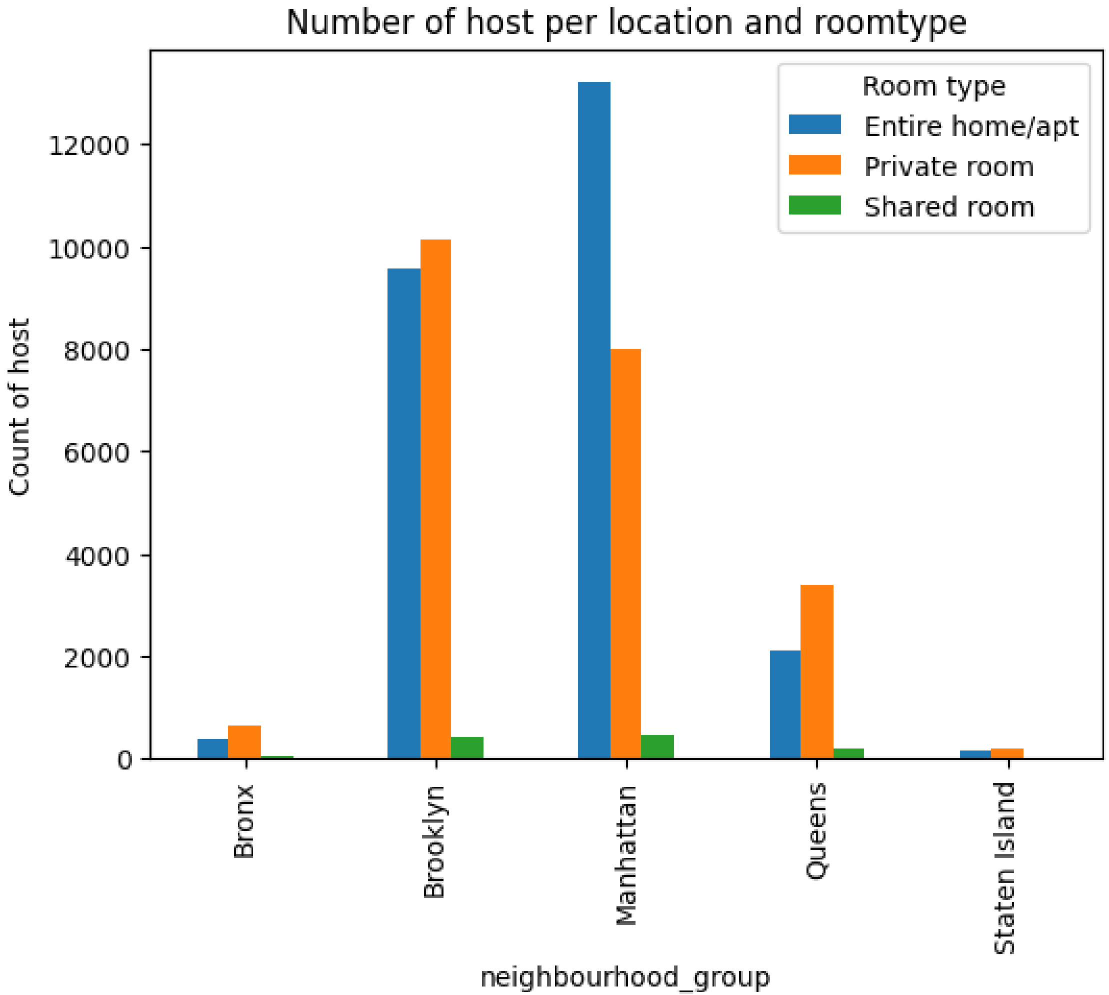
4. Distribution of numeric features in the dataset- Distribution plot



- Variables exhibit right-skewed distribution with concentration towards lower values and some extreme values.
- Majority of the values concentrate towards the lower end of the scale, with a few extreme values present in the dataset.

VISUALIZATION

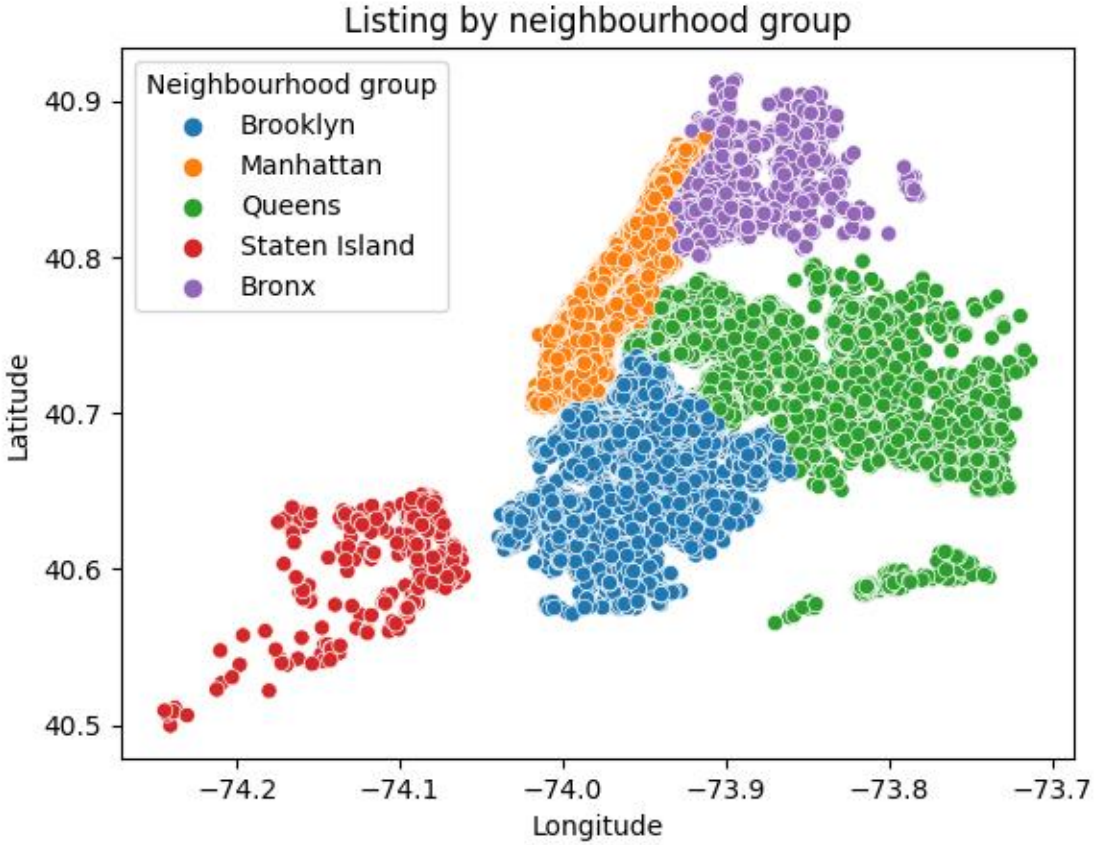
5. Number of hosts for all locations and room type - Bar plot



- Manhattan has the highest active host for entire home/apartment type.
- Bronx, Brooklyn, Queens, and Staten Island have a higher prevalence of private rooms.
- Shared room type services are less common overall

VISUALIZATION

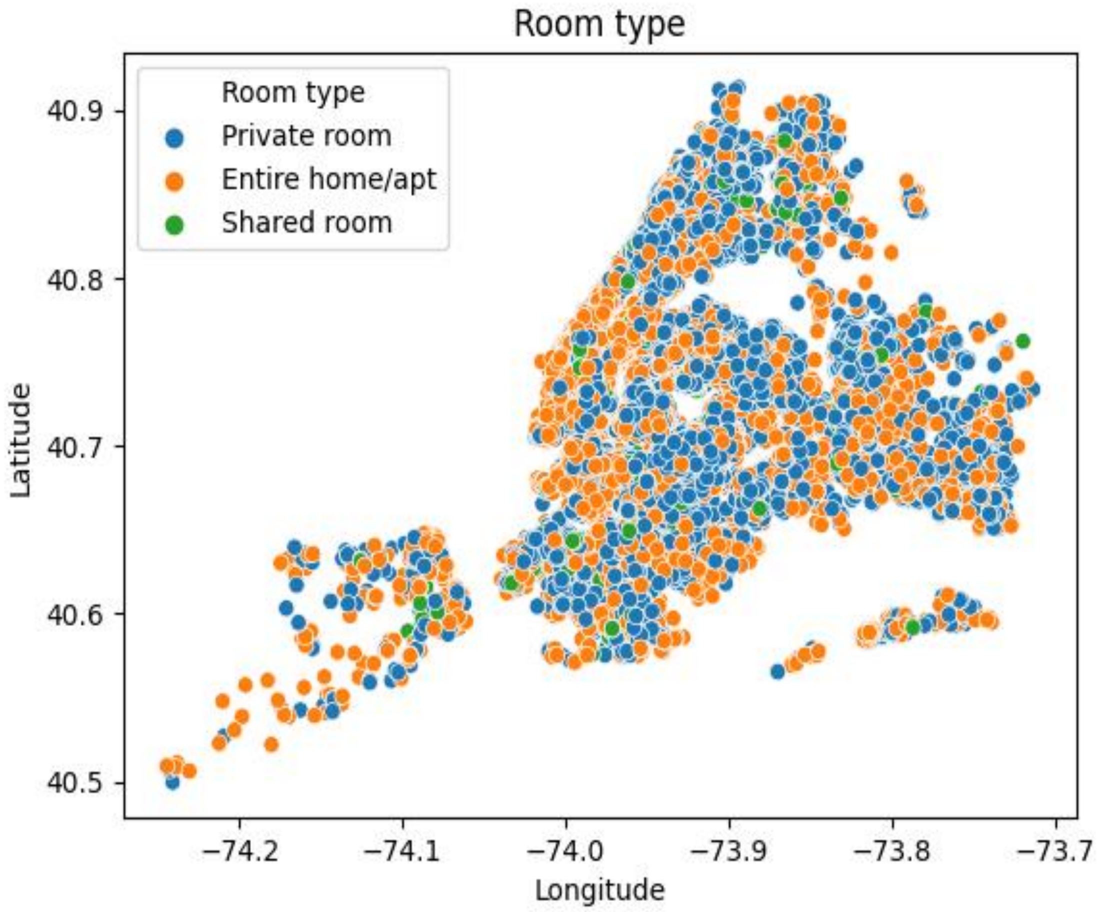
6. Geographical distribution of services among neighbourhood- Scatter plot



- Understanding the spatial distribution of Airbnb listing
- Identifying the patterns or variations across neighbourhood

VISUALIZATION

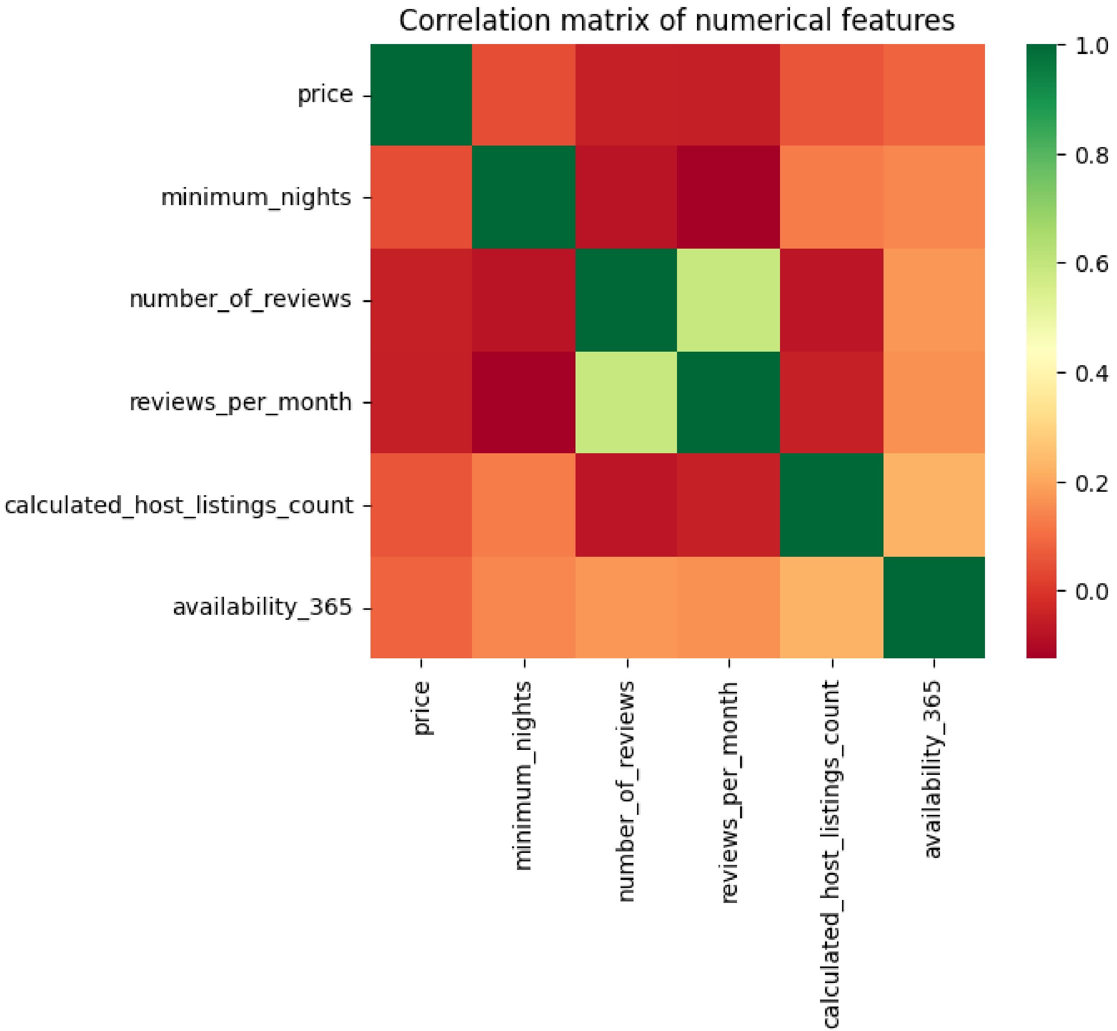
7. Geographical distribution of room type - Scatter plot



- When viewing private rooms and entire home/apartment type room services, the points seem scattered at some location and dense at other locations.

VISUALIZATION

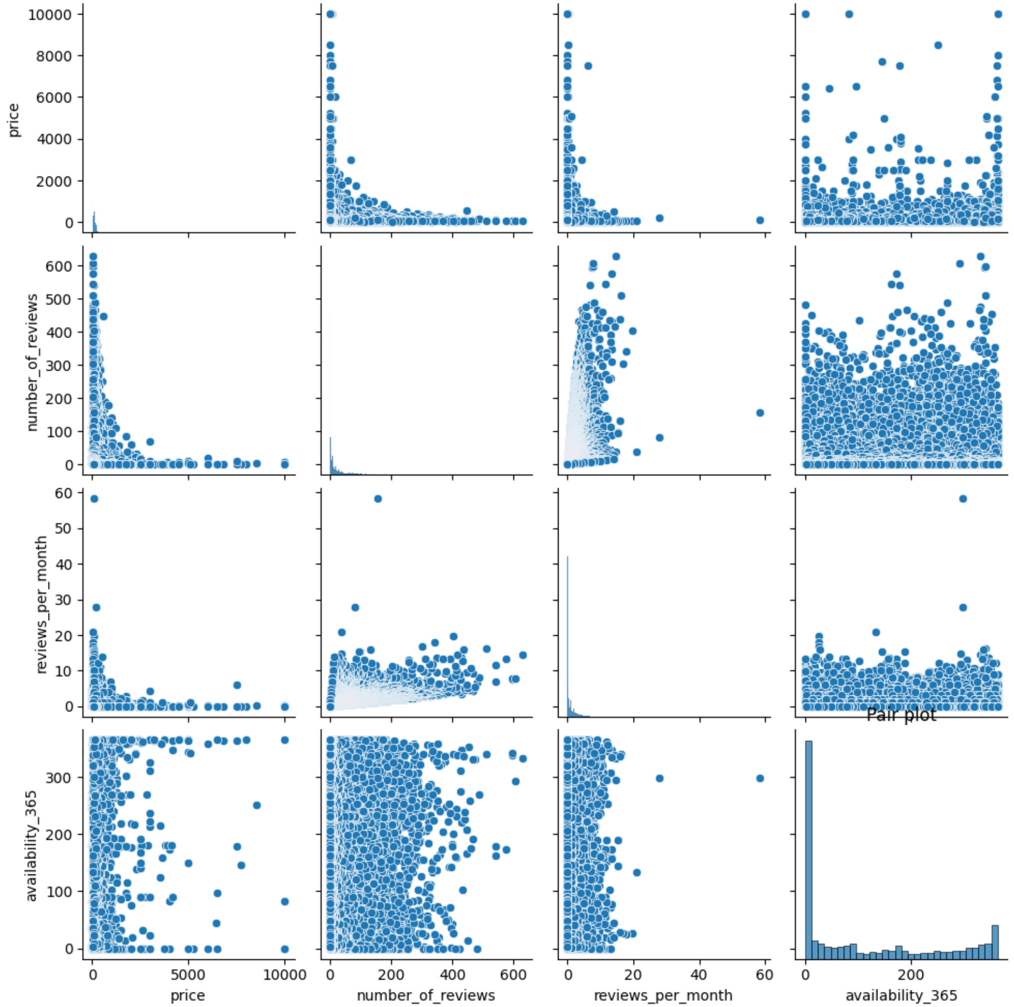
8. Correlation matrix



- It reveals a negative correlation between:
  - Calculated host listing count and number of reviews
  - Reviews per month and minimum nights
- Positive correlation is found between number of reviews and reviews per month

VISUALIZATION

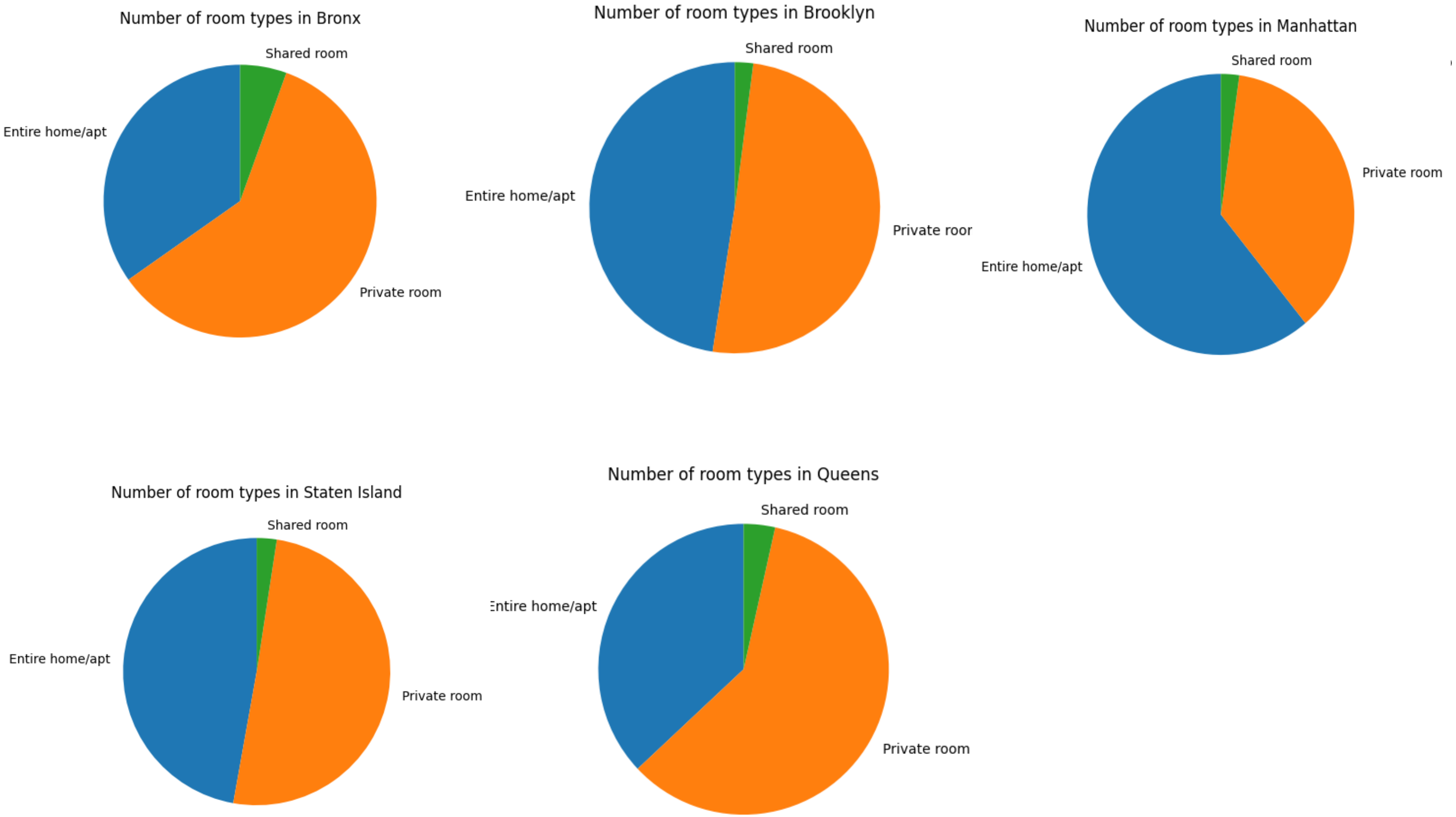
9. Relationship between different numeric features - Pair plot



- Price decreases upon increase in number of reviews and reviews per month, and vice versa.
- Reviews per month and number of reviews exhibits a positive relations ship.
- There is no strong relationship of variable availability.

VISUALIZATION

10. Number of room types in each neighbourhod group - Pie Chart



Number of room types in Bronx

Number of room types in Brooklyn

Number of room types in Manhattan

Number of room types in Staten Island

Number of room types in Queens

For all neighbourhood groups, the share of entire home/apartment and private rooms services are higher compared to shared room type services.

CONCLUSION

- Manhattan and Brooklyn have a noticeably higher number of services compared to other neighbourhood groups. This insight allows us to pinpoint potential expansion opportunities in neighbourhood groups that currently have fewer services.
- Price and number of reviews, and price and reviews per month are inversely proportional to each other, indicating the more customers might be preferring to opt for room services at lower price. By monitoring and responding to customer feedback, customer satisfaction can be enhanced and client can drive positive growth for the business.
- The box plot analysis revealed outliers in the variables minimum nights. It seems unrealistic that the minimum night for the booking is 1250. It is important to eliminate the anomalies and outliers.
- The correlation matrix reveals the positive correlation between number of reviews and reviews per month.

THANK YOU