Capstone Project – 1

EDA on Airbnb Bookings

By

Team Alma Phoenix

Azhar Ali

Mohd Taufique

Aishwary Sharma

Pushpam Raghuvanshi

Content

1. Problem Statement

2. Airbnb Booking Exploratory Data Analysis
    2.1. Importing Necessary Libraries
    2.2. Understanding Data
    2.3. Data Cleaning
    2.4. Data Analysis and Visualization

3. Conclusions

Problem Statement

- Airbnb (ABNB) is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales.

- The company has come a long way since 2007, when its co-founders first came up with the idea to invite paying guests to sleep on an air mattress in their living room. According to Airbnb's latest data, it has in excess of six million listings, covering more than 100,000 cities and towns and 220-plus countries worldwide.

- Data analysis on millions of listings provided through Airbnb is a crucial factor for the company.

- For the Exploratory Data Analysis, we are using Airbnb’s New York City Booking data of 2019.

Problem Statement(cont.)

AI

- Using EDA techniques we will explore and visualize the datasets and our focus will be on finding out the key factors that influence, analyzed and can be used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

- As part of the analysis, we will attempt to answer the following questions for the Airbnb data set:

1. What types of accommodation/room type are offered?

2. How many properties are in different neighbourhood groups and what proportions of accommodation/room type are available?

Problem Statement(cont.)

3. Who are the top 10 hosts along with their IDs and Name?
(Top 10 hosts with respect to maximum number of Properties listed)

4. What price the guest mostly prefer while Booking?

5. Where is the most popular and the most expensive area?

6. What are the top 20 keywords used in listing names?

7. What are the top neighbourhoods with highest listings?

8. Density and price distribution across different neighbourhood groups/Boroughs.

Steps involved in our EDA-

1. Importing Necessary Libraries – NumPy, Pandas, Matplotlib, Seaborn.

2. Importing Airbnb Booking csv file in Google Collab

3. Understanding the Data

4. Data Cleaning

5. Data Analysis and Visualization

Understanding the Data

- Airbnb dataset is huge with around 49000 row entries and 16 columns.

- Different columns are of various datatypes.

- There are significant NaN values in some columns.

- Some columns are not significant for more in depth analysis .

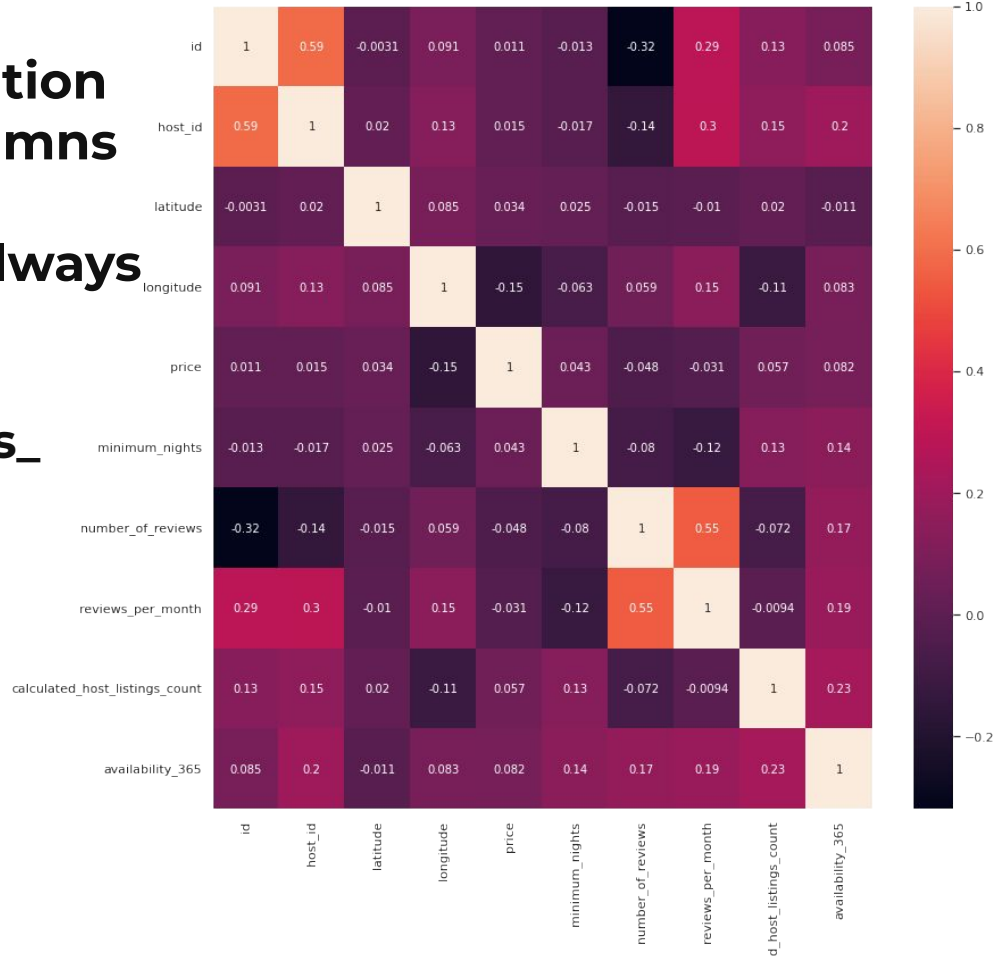
```
# Summary of Dataset using the info() funtion.
airbnb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              48895 non-null  int64  
 1   name                            48879 non-null  object 
 2   host_id                         48895 non-null  int64  
 3   host_name                       48874 non-null  object 
 4   neighbourhood_group             48895 non-null  object 
 5   neighbourhood                   48895 non-null  object 
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object 
 9   price                           48895 non-null  int64  
 10  minimum_nights                  48895 non-null  int64  
 11  number_of_reviews               48895 non-null  int64  
 12  last_review                     38843 non-null  object 
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64  
 15  availability_365                48895 non-null  int64  
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

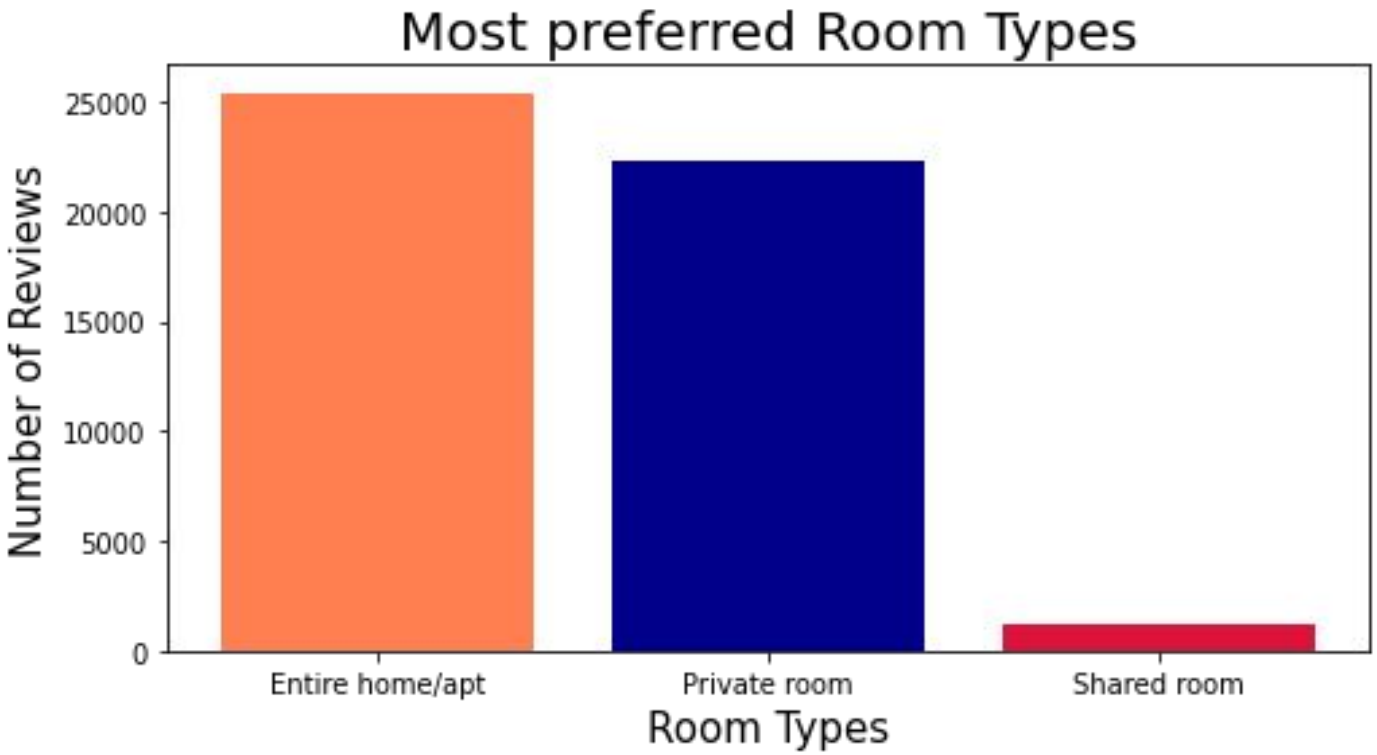
Data Analysis and Visualization

- There is no any strong correlation between any two features/columns in this dataset. The correlation between the same column is always equal to 1.

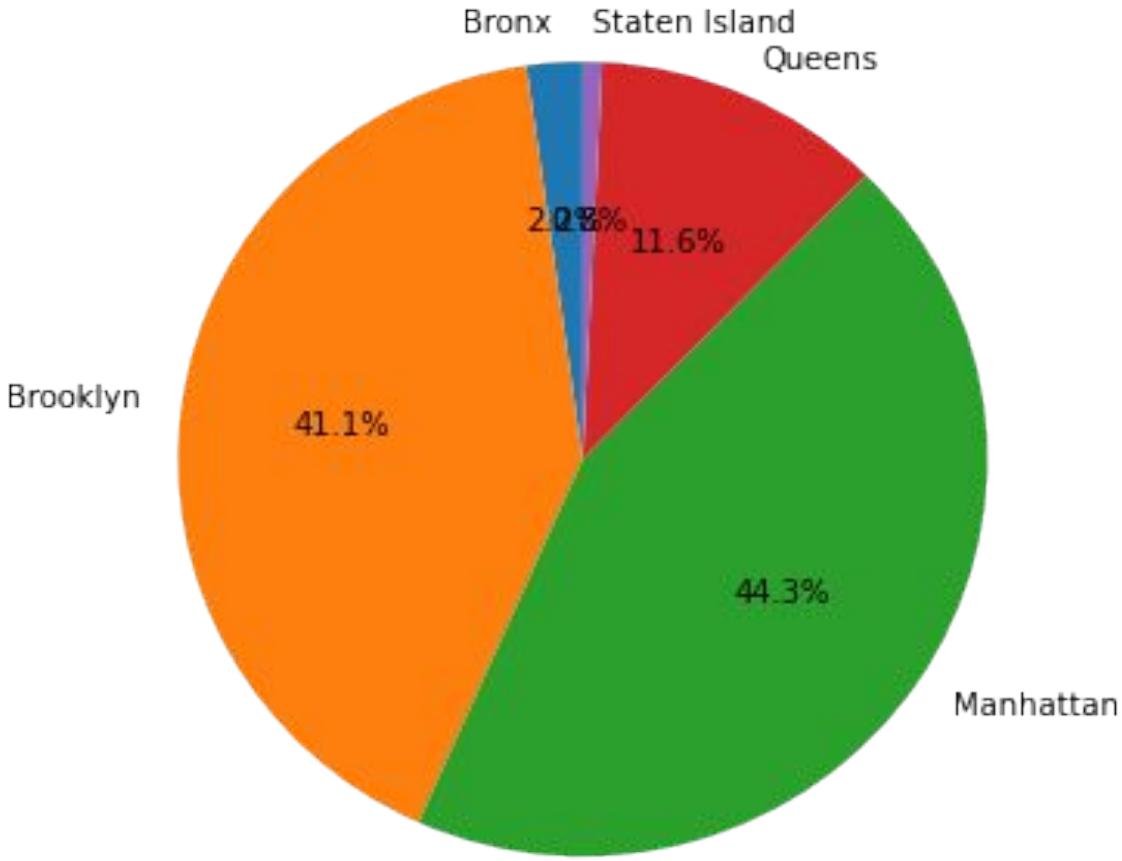
- number_of_reviews and reviews_per_month has high correlation.



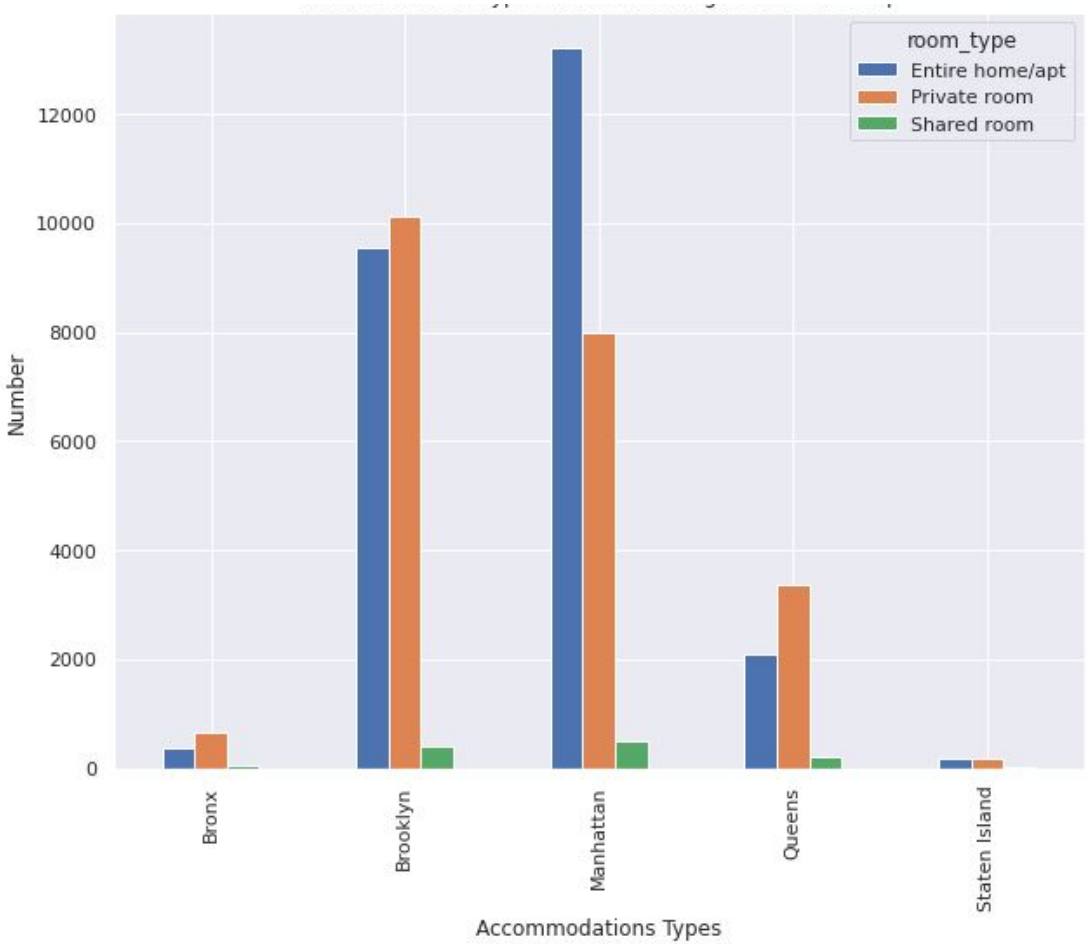
What are different accommodations/room types are preferred ?



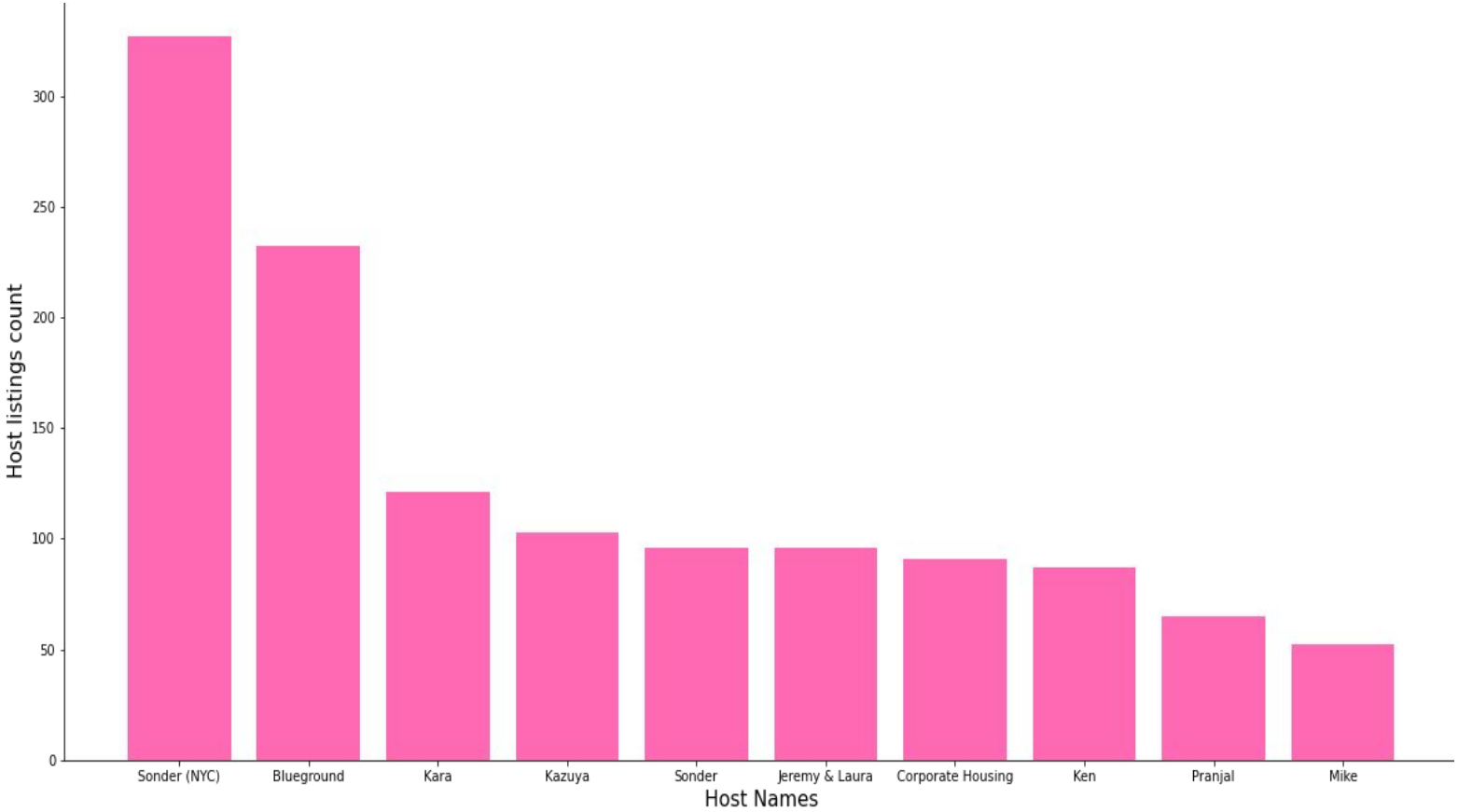
Percentage of Properties Listed in Different Neighbourhood Groups



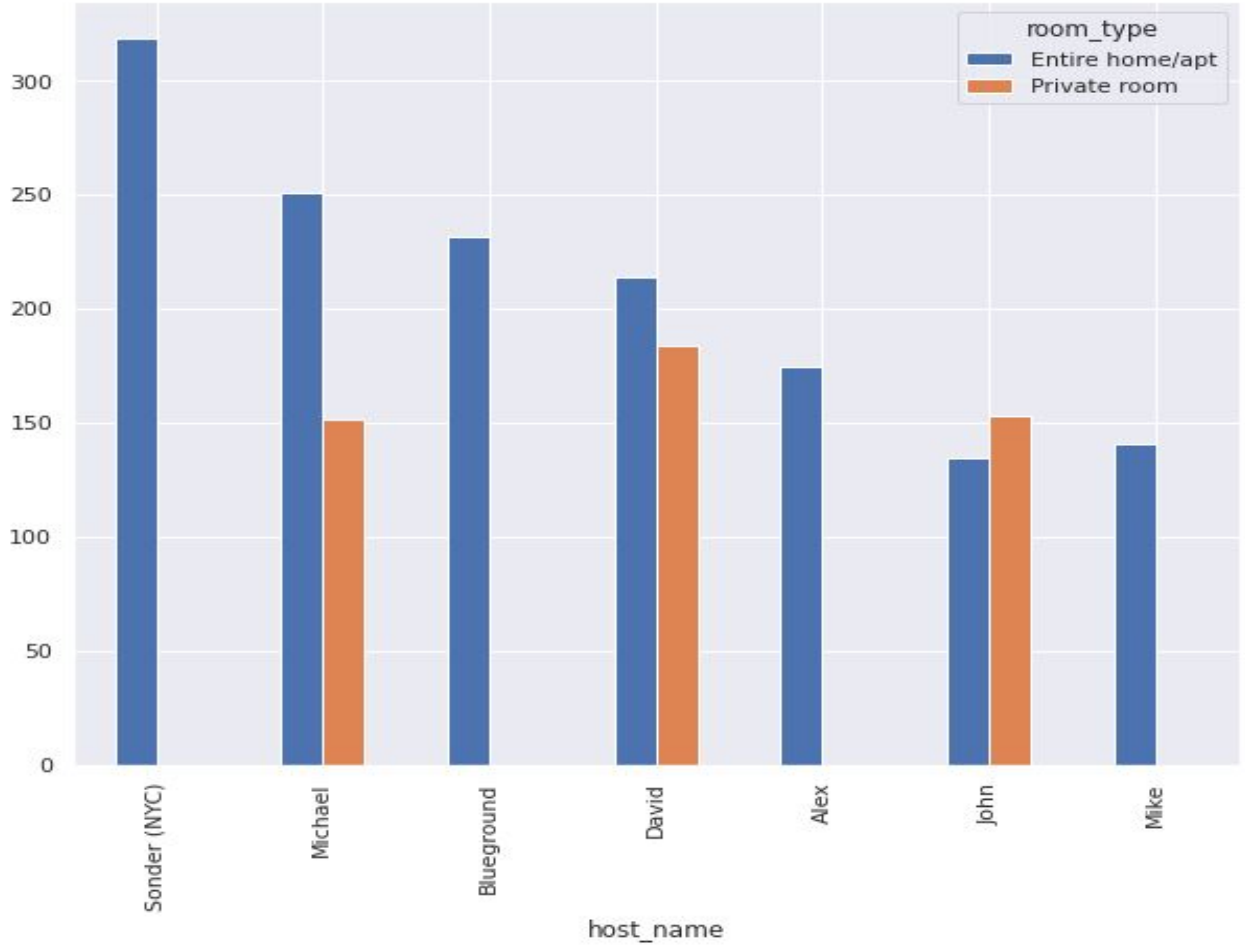
Number of Room Types in different Neighbourhood Groups



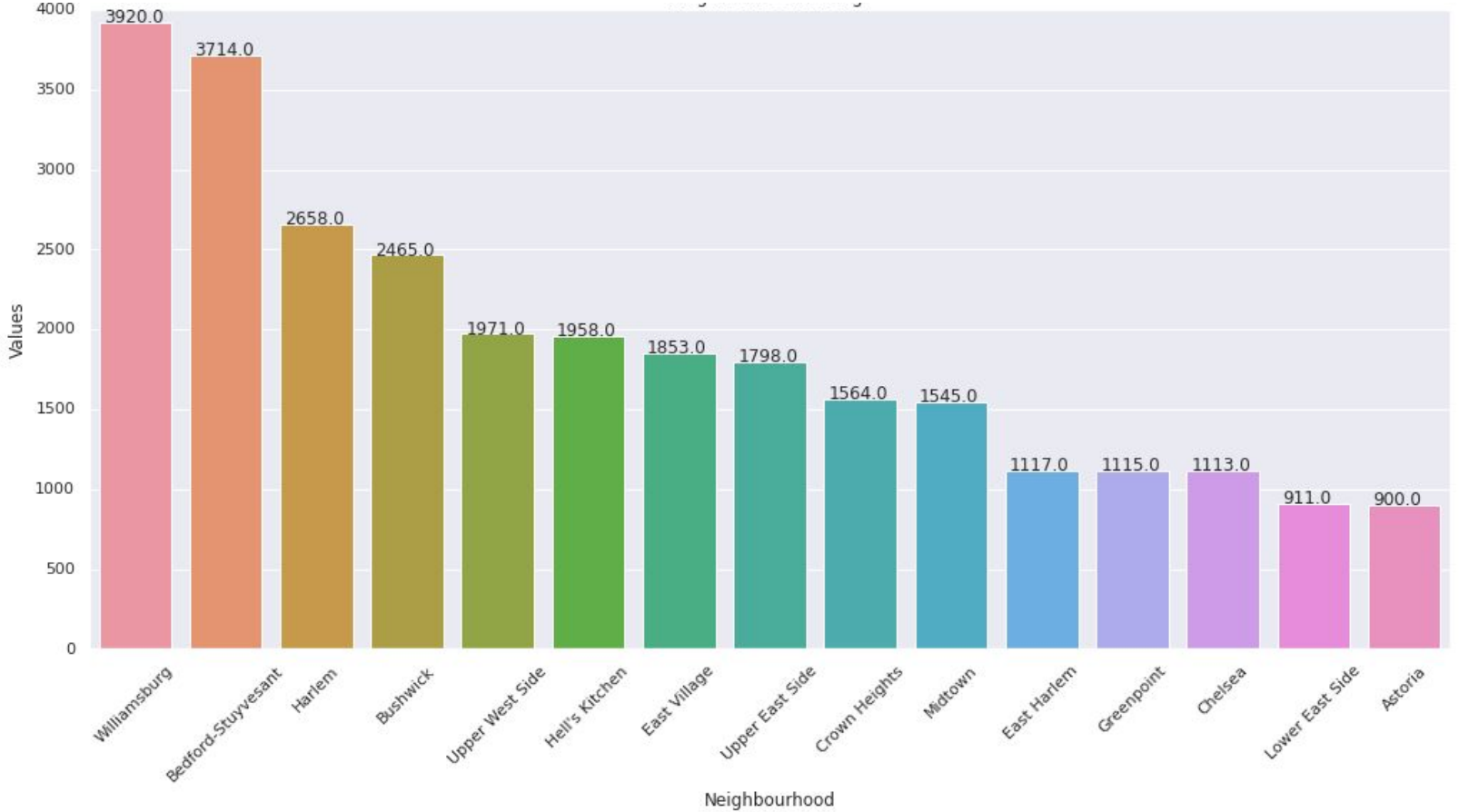
Who are the top 10 hosts among all neighbourhood groups? (Top 10 hosts with respect to maximum number of Properties listed)



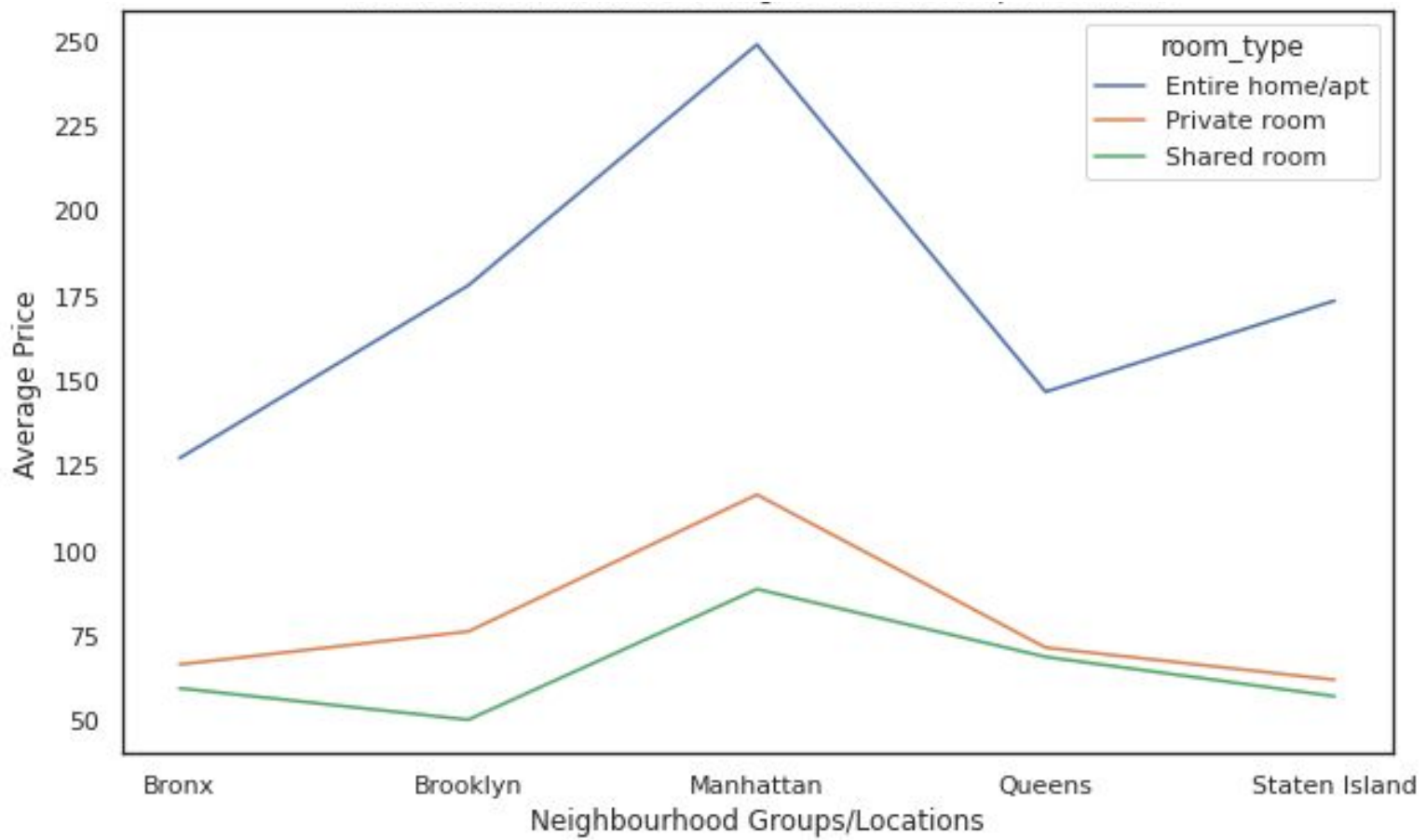
Room Types offered by Top Hosts



Top Neighbourhood with highest number of Listings Across all Boroughs.

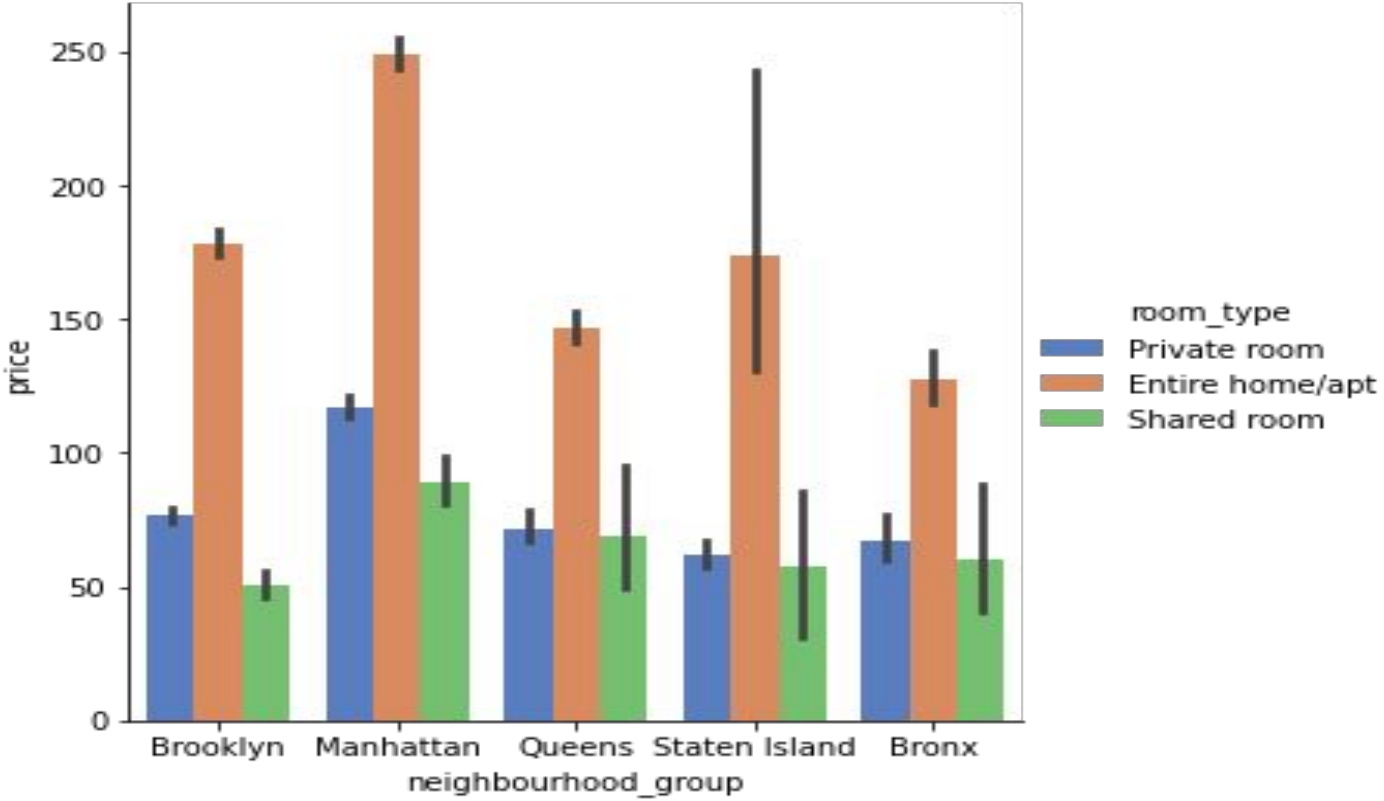


Price Variation in different Neighbourhood Groups/Locations



Most Expensive Neighbourhood group/area

- Manhattan among all five neighbourhood group is the most expensive area for booking of all types of room.



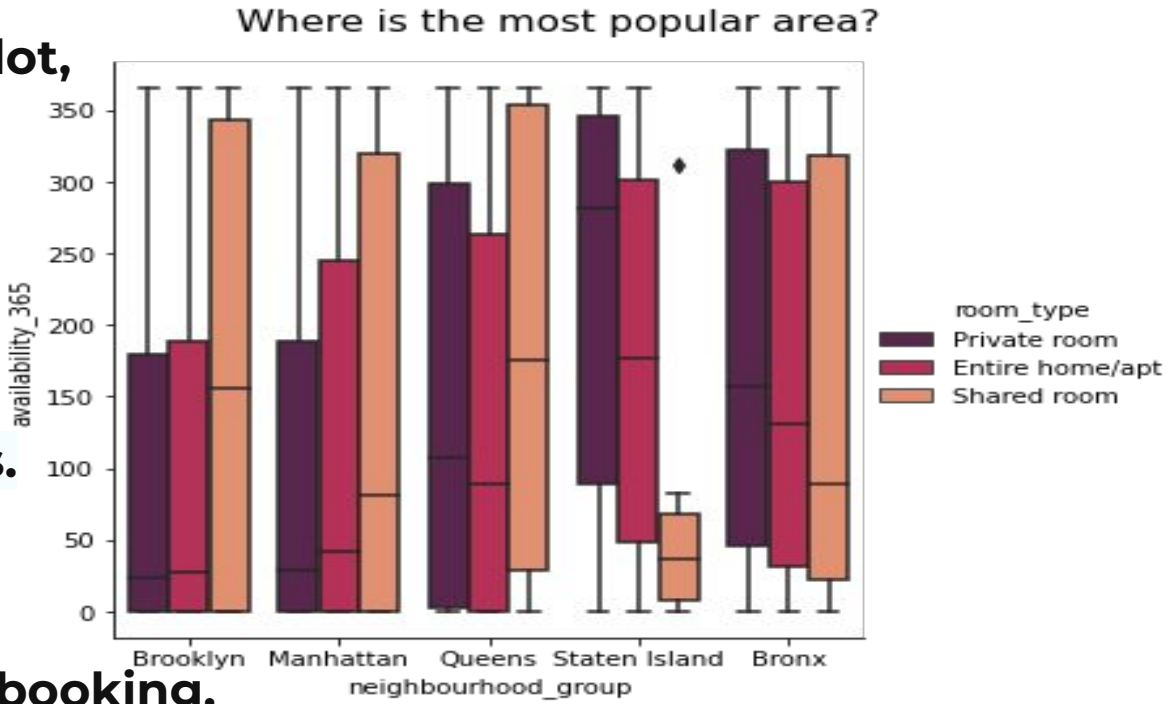
Availability_365 vs Neighbourhood Group

- Popularity of neighbourhood group is inversely proportional to availability. A smaller availability_365 is ideal for an Airbnb host.

- From the clustered box plot, Brooklyn and Manhattan are the most popular.

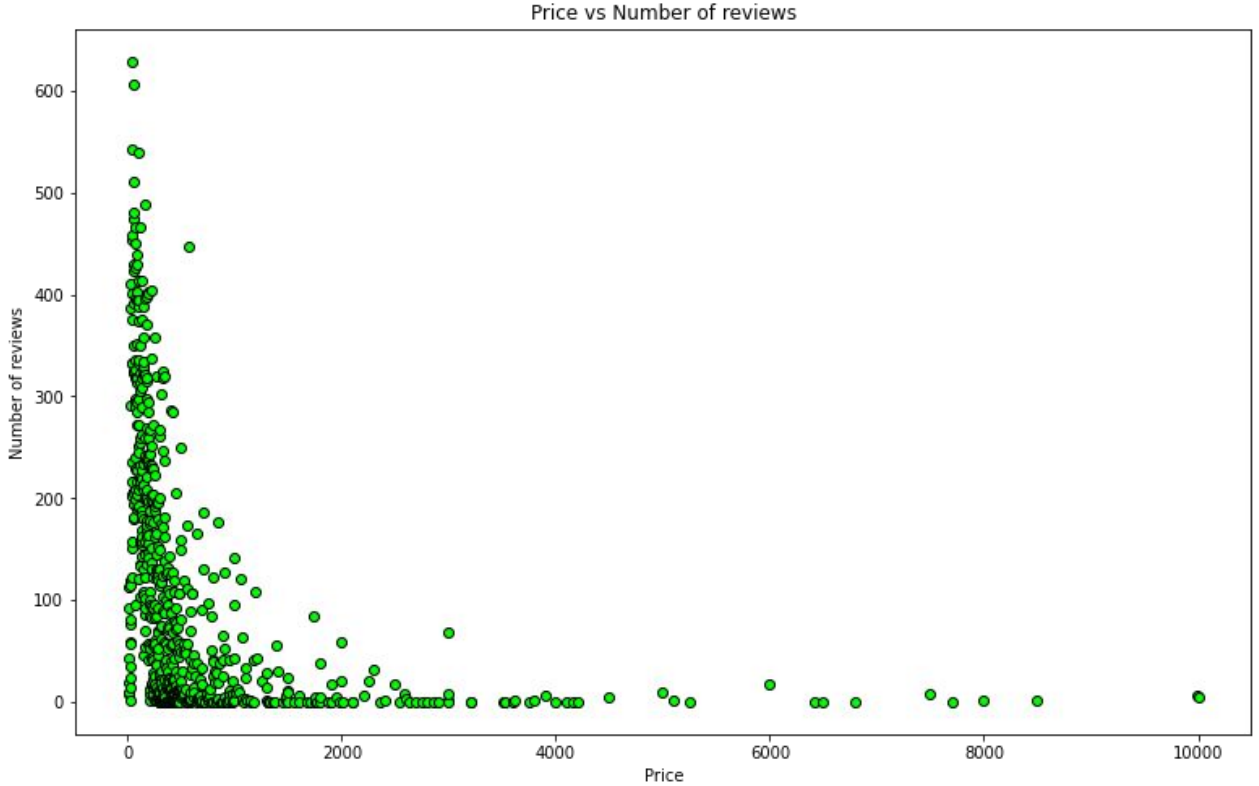
- Within these two areas entire home/apt and private room are the most preferred room types.

- On the other hand, Staten Island is the most preferred for shared room booking.



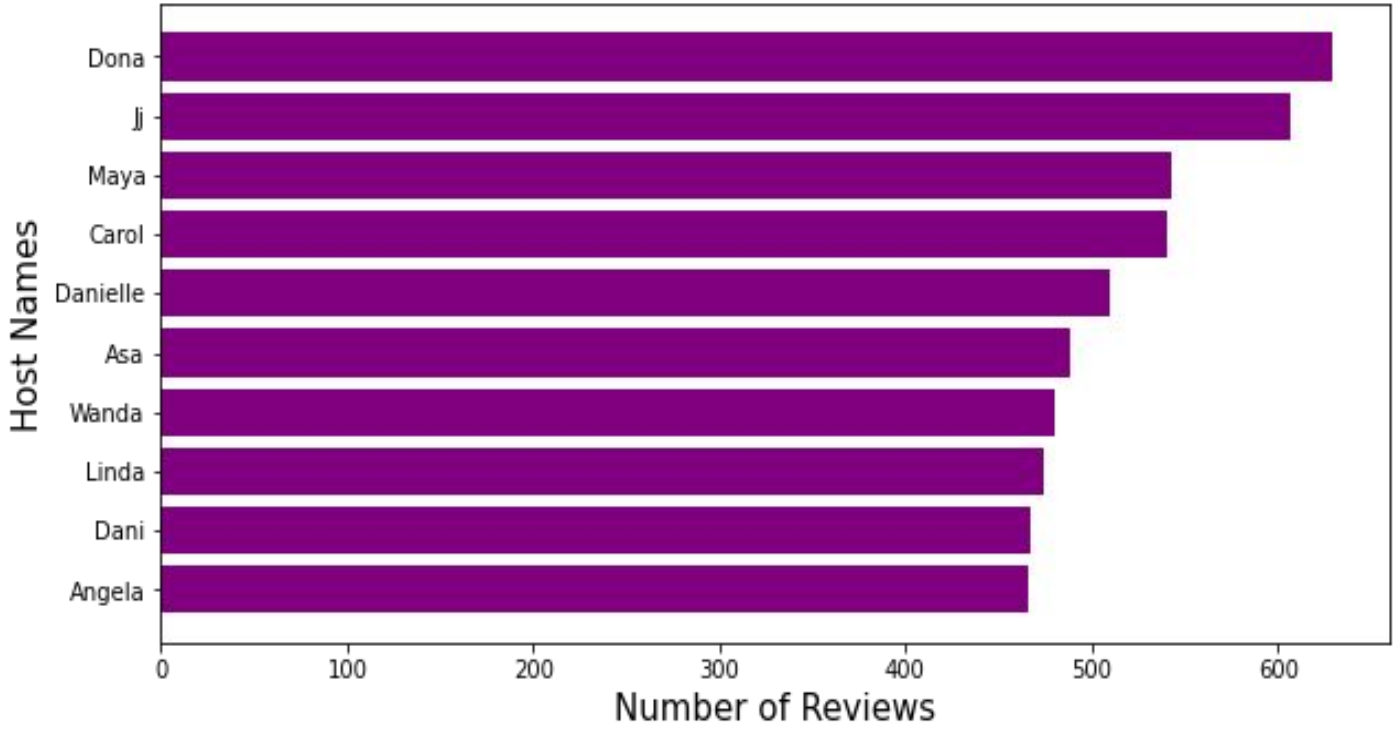
Price Vs Number of reviews

- The scatter plot clearly shows that most number of reviews are received for the listing having low price. It means guests prefer budget price.



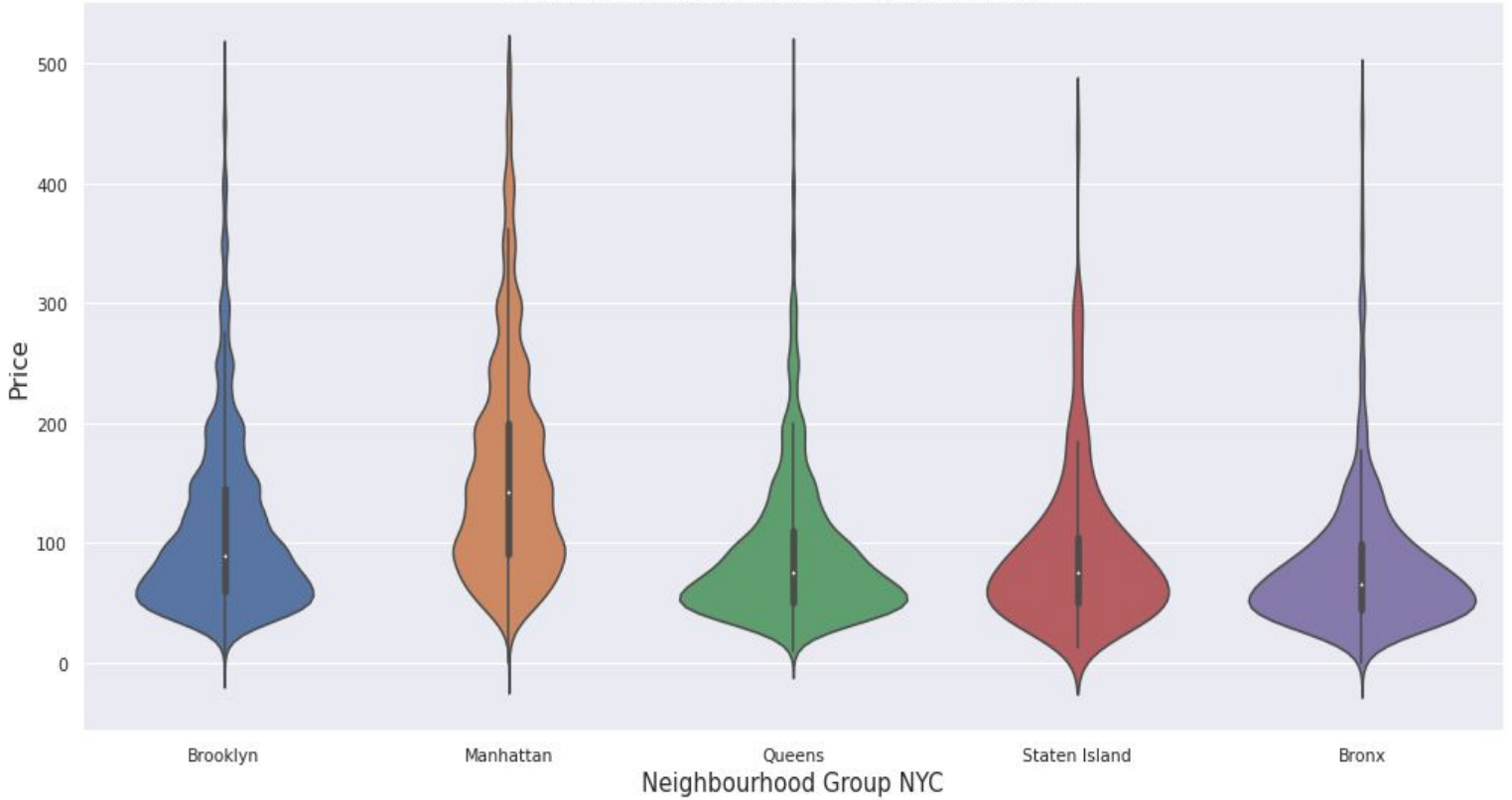
Busiest Hosts

- The host with name Dona of Queens Neighbourhood group is the most busiest and preferred host with highest number of reviews by guest among all five neighbourhood group.



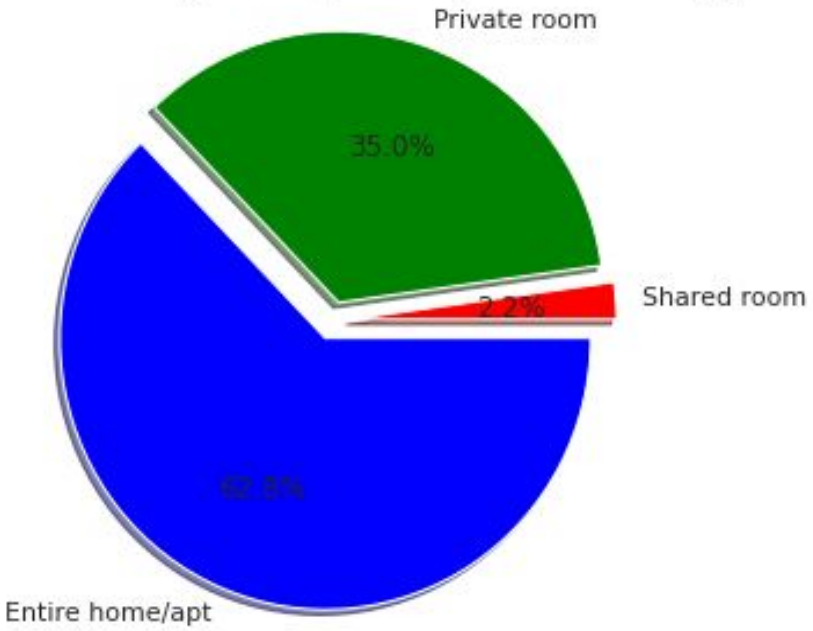
Density and Price Distribution of Neighbourhood Groups

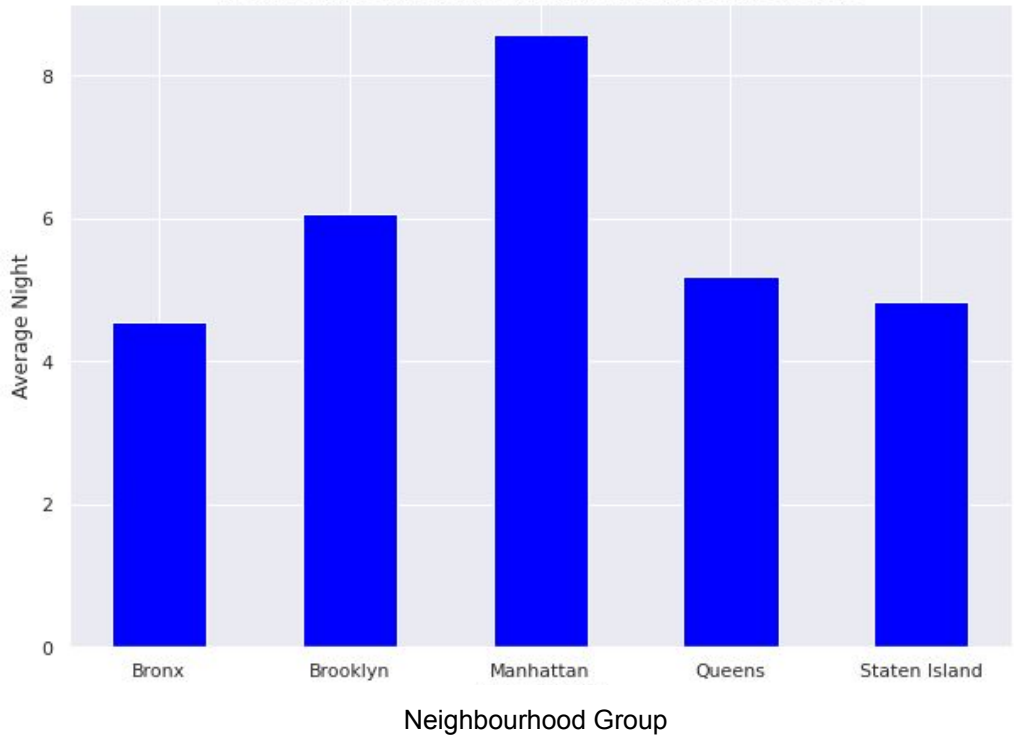
- Violin plot shows that Manhattan has the highest price range and distribution than other neighbourhood groups.



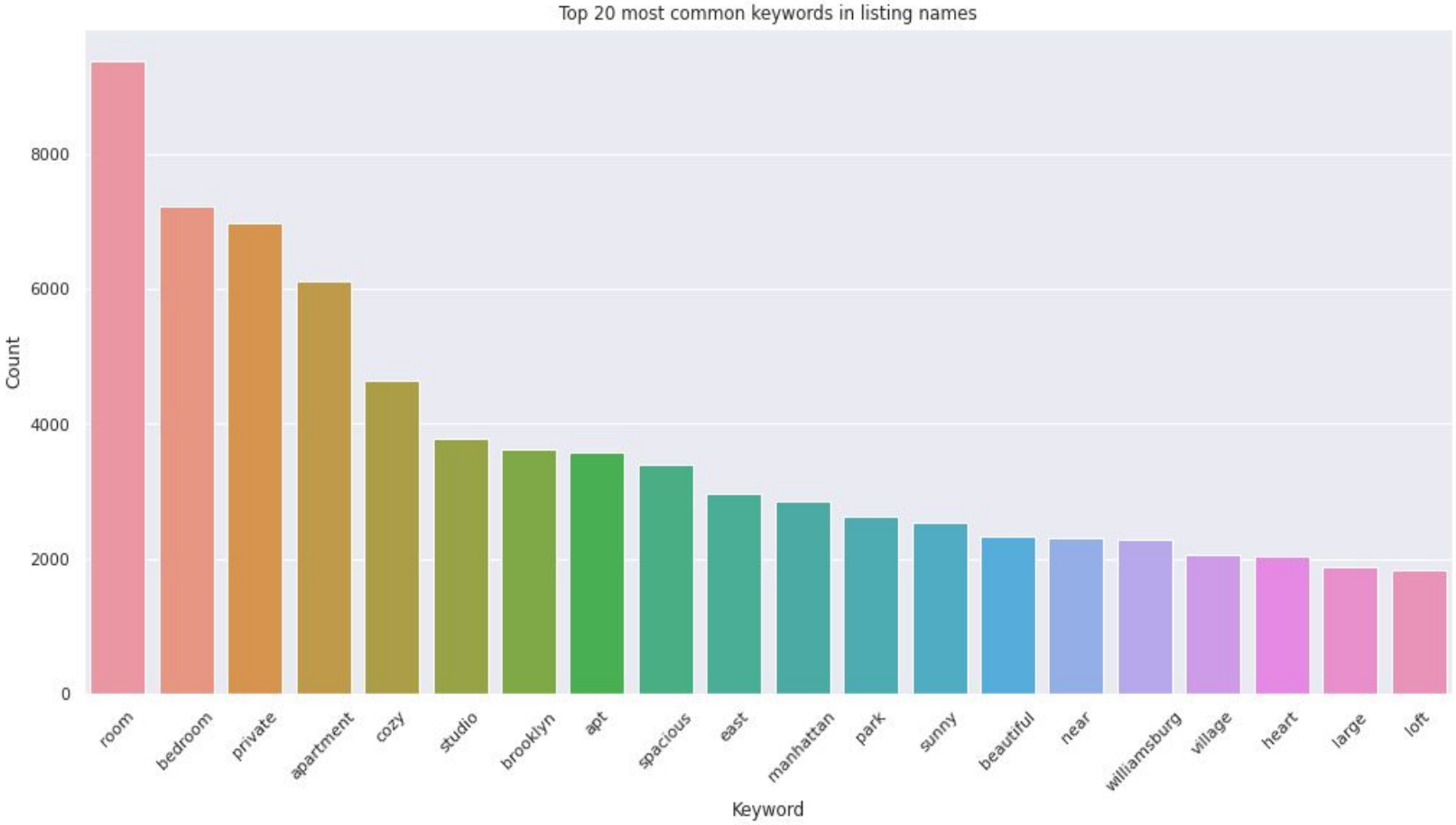
Total Number of Nights Spend per Room Type & Average Night spent per Neighbourhood Group

- Guests mostly prefer entire house/apt followed by private room.
- Very few guests prefer shared room.





Top 20 Most Common Keywords in Listing Names



Word Cloud of The Most Used Keywords

- Here we tried to show top used keywords in the form of New York map.
- Keywords like 'room', 'private' and 'studio' are used most number of the times by the hosts.
- Neighbourhood group like Manhattan and Brooklyn are also present in this list among the top 20 keywords, this implies that these two are most popular.

Challenges Faced

- Understanding the meaning of some columns.

- Dealing with Null values and duplicates.

- Understanding the business model of Airbnb that how they work.

- Also, forming different graphs to show insights from the dataset and to summarize the information and communicate the results and trends to the reader successfully.

Conclusions

1. Manhattan and Brooklyn are the two most popular, expensive & posh areas of NY city. These are the most focused place for hosts to do their business.

2. Most visitors don't prefer shared rooms, they tend to visit private room or entire home/apt. The obvious some reasons are privacy, hygiene & more space.

3. Williamsburg is the neighbourhood with highest number of listings. It means this place is famous for tourists attraction.

4. 'room', 'private', and 'studio' are used most number of the times by the hosts. Also Neighbourhood group like Manhattan and Brooklyn are also present in this list among the top 20 keywords.

5. Queens is the neighbourhood which has received maximum number of reviews while bronx has received least number of reviews.

Conclusions(cont.)

6. Around 44.3% properties are listed in Manhattan followed by 41.1% in Brooklyn. Staten island has minimum number of listed properties.

7. Percentage of nights spend is maximum for entire home/apt with 62.8% followed by private room with 35%. Average number of nights spend is maximum for Manhattan island which is more than 8 nights and minimum is 4 nights for Bronx.

8. Dona and ji are the most busiest hosts with maximum number of reviews from the visitors/guests.

9. Most guests prefer budget price for their stay.

10. Sonder (NYC) has highest number of listings (327) in the entire NYC. Listings by top 10 hosts is almost 2.5% (1270 listings) of whole dataset.

Thank You

Team Alma Phoenix