CAPSTONE PROJECT

Exploratory Data Analysis
OF
Airbnb Bookings Analysis

Submitted By:-

Umang Goel

Cohort:- Ottawa

(AlmaBetter)

Introduction to Airbnb

❑The San Francisco-based startup Airbnb was founded in 2007 when two hosts opened their residence to three visitors. Since then, the firm has expanded to over 4 million hosts and hosted over 1.5 billion guests in practically every nation.

❑Airbnb serves as an online accommodation marketplace that links customers searching for short-term lodging with those wishing to lease out their properties. Airbnb gives hosts a comparatively simple method to make a little money from their homes.

❑The basic premise of Airbnb is to connect locals with those who have extra space in their homes or apartments that they can rent out to tourists. With the security of knowing that a large firm will manage payments and provide other assistance, hosts utilizing the platform get to market their rentals to millions of individuals around the world.

Purpose of Exploratory Data Analysis

❖The main goal of this analysis is to understand the pattern and to provide useful insights into Airbnb to improve its functionality and serviceability.

❖First data is to be explored and cleaned using different data-wrangling techniques. The missing or null values are to be handled carefully as they will create problems during data analysis and can affect statistics and computations. Duplicates and outliers are also dealt with in this cleaning process only as they can cause errors in data collection, data entry, or data merging and can lead to false analysis results.

❖After cleaning the data, the identification of different patterns and trends present inside the data can be done using various data visualization techniques. Various graphs and charts can be created to visualize the data, and record observations and insight which will be very useful for future analysis and decision-making related to Airbnb.

Problem Statement

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb has become a one-of-a-kind service that is used and recognized by the whole world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - data that can be analyzed and 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. This dataset has around 49,000 observations in it with 16 columns and it is a mix of categorical and numeric values. Explore and analyze the data to discover key understandings.

Objectives

1. Which neighbourhood group has the maximum number of reviews?

2. What is the average price for each room type available for booking?

3. Most reviewed room types per month in neighbourhood groups.

4. How does the availability of the room vary throughout neighbourhoods & room type?

5. What is the relation between the number of reviews and price for different room types?

6. What is the location of availability of rooms according to their given latitudes & longitudes?

7. The relation between average minimum nights with neighbourhood groups and room type.

Contd.

Contd.

8. The price distribution between different neighbourhood groups

9. The maximum number of hosts present in different neighbourhood groups

10. The top 10 hosts according to their listings.

11. The correlation between different variables.

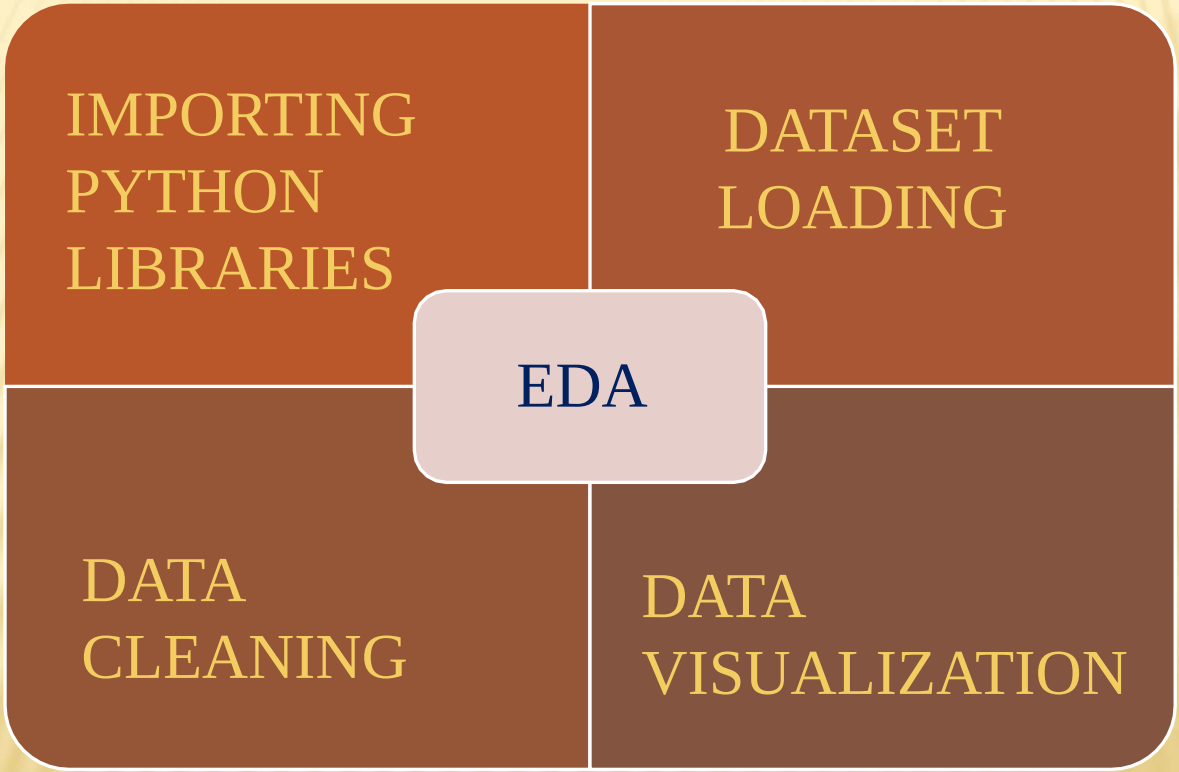
Dataset Attributes

1. id: (Unique id of listing of all the dataset)
2. name: (Name of the property)
3. host_id: (Unique id for each host in the dataset)
4. host_name: (Name of the host)
5. neighbourhood_group: (Name of the group in the city such as- Bronx, Brooklyn, Manhattan, Queens, Staten Island)
6. neighbourhood: (Specific name of the neighbour where the listing is located)
7. latitude: (Latitude of the location, continuous from 40.49979 to 40.90804)
8. longitude: (Latitude of the location, continuous from -74.24285 to -73.71299)
9. room_type: (Type of room such as- Entire home/apt, Private room, Shared room)
10. price: (Price of the stay per night in USD)
11. minimum_nights: (Minimum number of nights, that a guest will stay at the listing)
12. number_of_reviews: (Total number of reviews the property has recieved)
13. last_review: (Date of the last review recieved by the property)
14. reviews_per_month: (Average number of reviews property recieved per month)
15. calculated_host_listings_count: (Overall number of listing that host has on airbnb)
16. availability_365: (Number of days out of 365 days, the property is available for booking)

Different Data Type in the data

Date Type	Column
Numeric - Int64	id host_id price minimum_nights number_of_reviews calculated_host_listings_count availability_365
Numeric – Float64	latitude longitude reviews_per_month
String - Object	name host_name neighbourhood_group Neighbourhood room_type last_review

Roadmap


IMPORTING PYTHON LIBRARIES
DATASET LOADING
EDA
DATA CLEANING
DATA VISUALIZATION

Importing Libraries In Python

Necessary python libraries are:-

```
[2]  import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     %matplotlib inline
     import seaborn as sns
```

Dataset Loading

```
[4]  path_name = "/content/Airbnb NYC 2019.csv"
     airbnb_df = pd.read_csv (path_name)
```

Data Cleaning

➢Removing Duplicates

```
[10] # Dropping the duplicate values of rows using drop_duplicates() function
     airbnb_df = airbnb_df.drop_duplicates()
     airbnb_df.count()  #count() function will count the number of rows after dropping duplicates
```

➢Checking Null values

```
[11] # Check for null values in each column
     # isnull() function returns value
     airbnb_df.isnull().sum()
```

Result

```
id                                  0
name                               16
host_id                             0
host_name                          21
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
last_review                     10052
reviews_per_month               10052
calculated_host_listings_count      0
availability_365                    0
dtype: int64
```

➢Handling missing data of name & host_name by filling missing data with “No name”

```
[12] # Filling missing values in name & host_name with "No name"
     airbnb_df["name"].fillna ('no name', inplace = True)
     airbnb_df["host_name"].fillna ("no name", inplace = True)
     airbnb_df.isnull().sum()
```

➢Handling missing data of reviews_per_month by replacing null values with 0

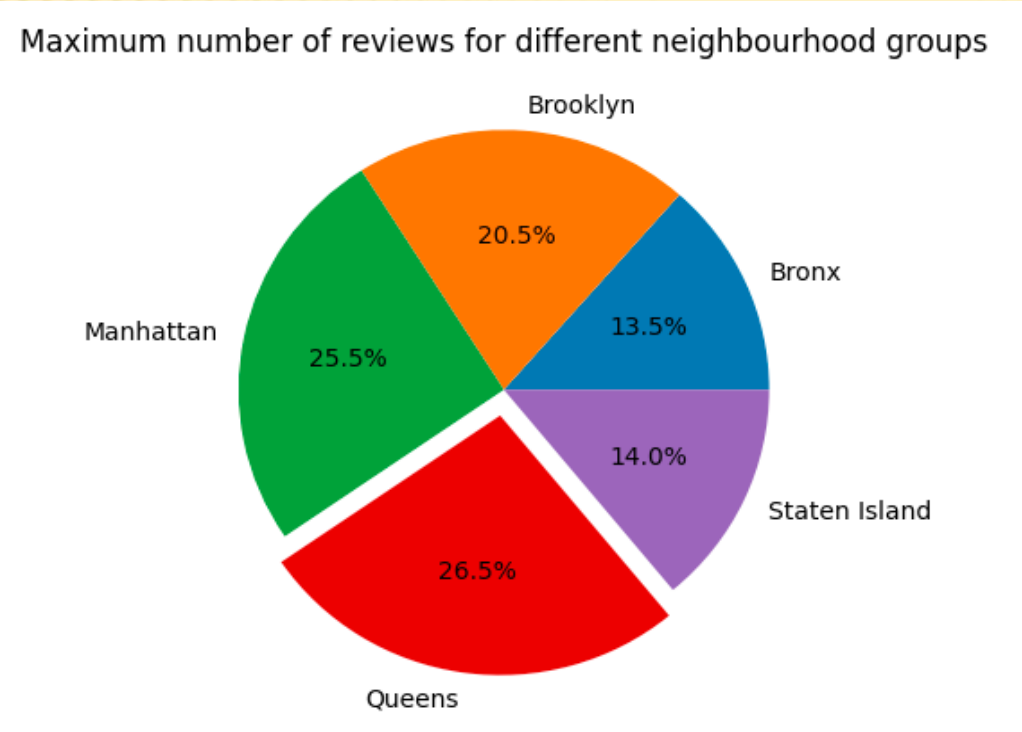
```
[14] # Replacing all the null values in review_per_month column by 0
     airbnb_df["reviews_per_month"] = airbnb_df["reviews_per_month"].replace(np.nan, value=0)
     airbnb_df.isnull().sum()
```

➢Handling missing data of last_review by dropping column

```
[15] #Dropping last_review column using drop() function
     airbnb_df = airbnb_df.drop(['last_review'], axis=1)
     airbnb_df.head().T
```

Data Visualization

1. Which neighbourhood group has the maximum number of reviews?

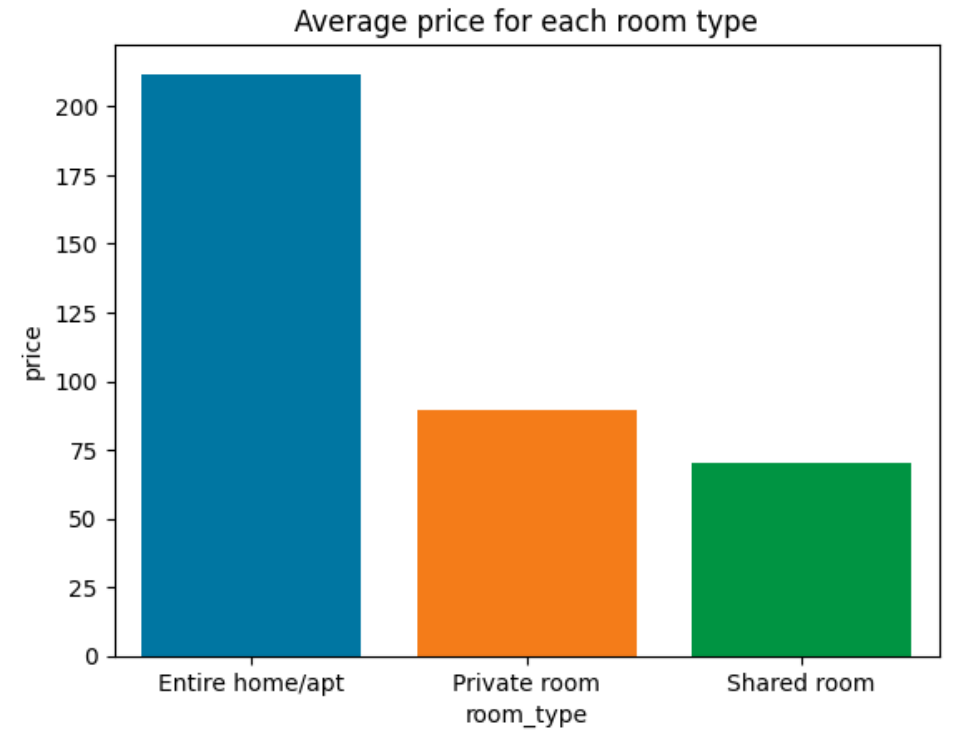


Insights:

•From the graph, it is clearly shown that **Queens** (26.5%) and **Manhattan** (25.5%) neighbourhood groups are most popular among customers.

•**Queens** has the highest percentage of reviews which made it the most popular location for travelers.

2. What is the average price for each room type available for booking?

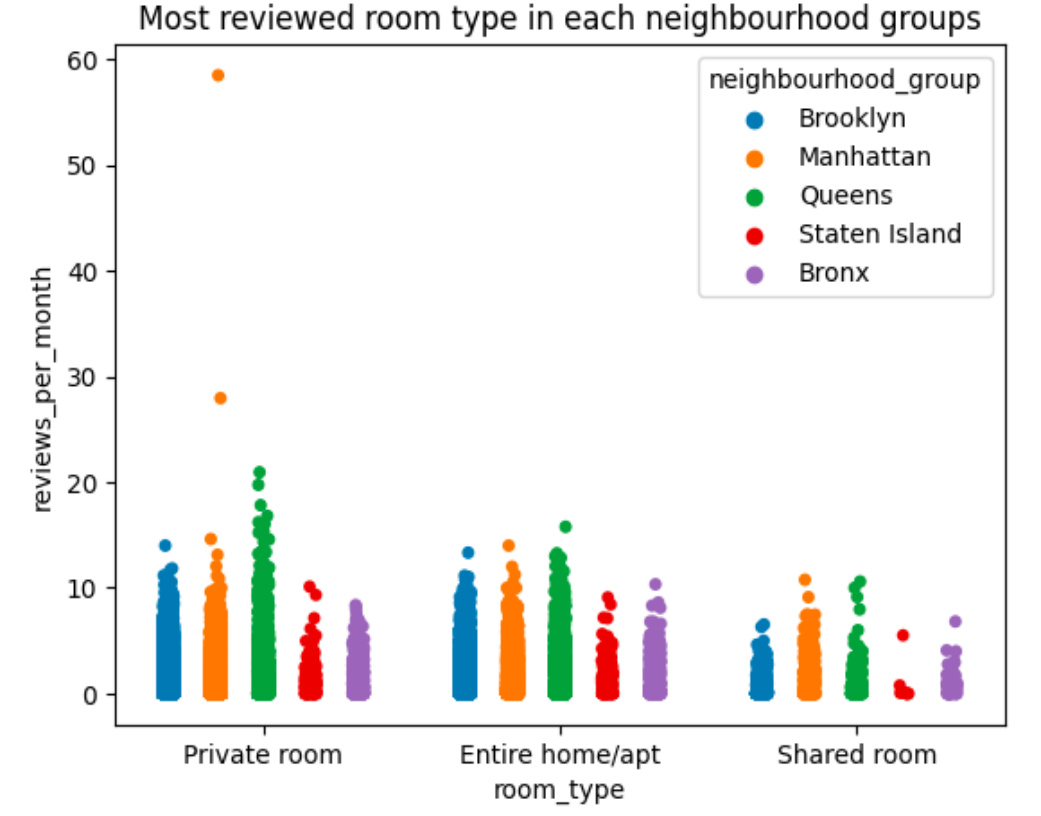


Insights:

•From the chart above, it is clearly visible that the **Entire home/apt** category of the room type is the most expensive category with an average price of 211.79 USD.

•**Shared rooms** are the cheapest room category with the average price of 70.13 USD.

3. Most reviewed room type per month in neighbourhood groups.

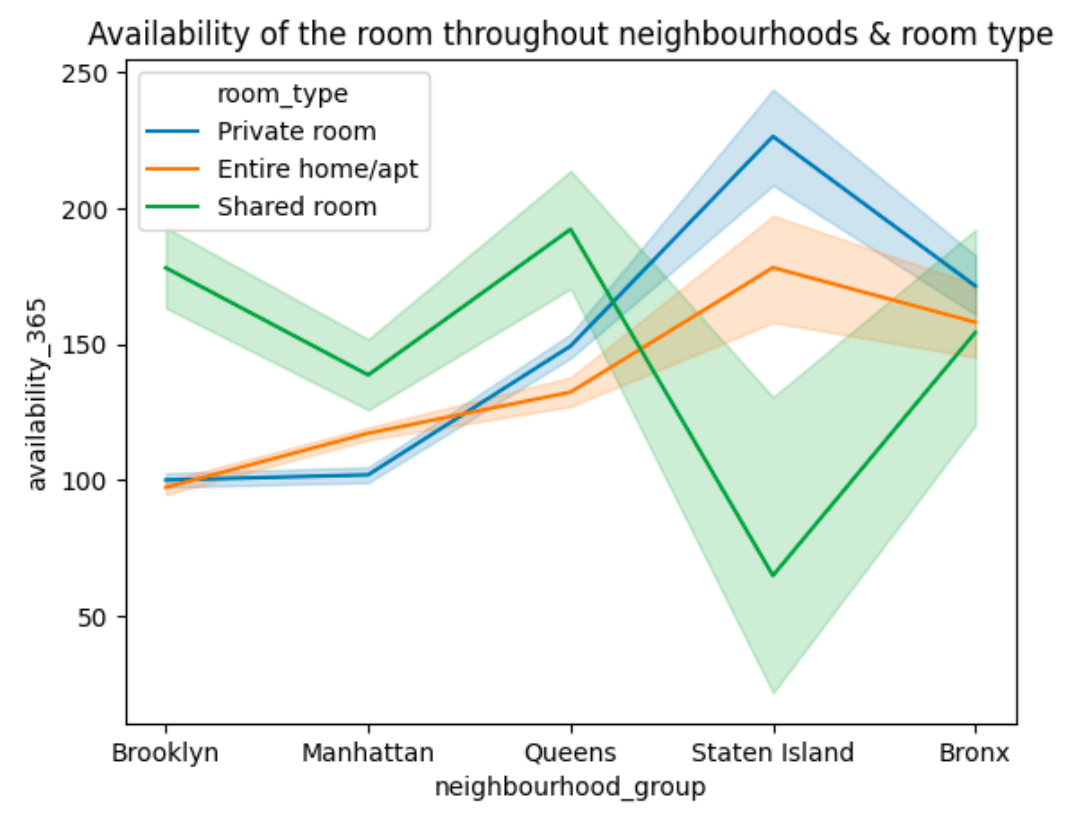


Insights:

•**Manahattan** neighbourhood group have maximum reviews per month for **Private** room type.

•**Staten Island** recieved lowest reviews for their **Shared** room type as compared to the other room type.

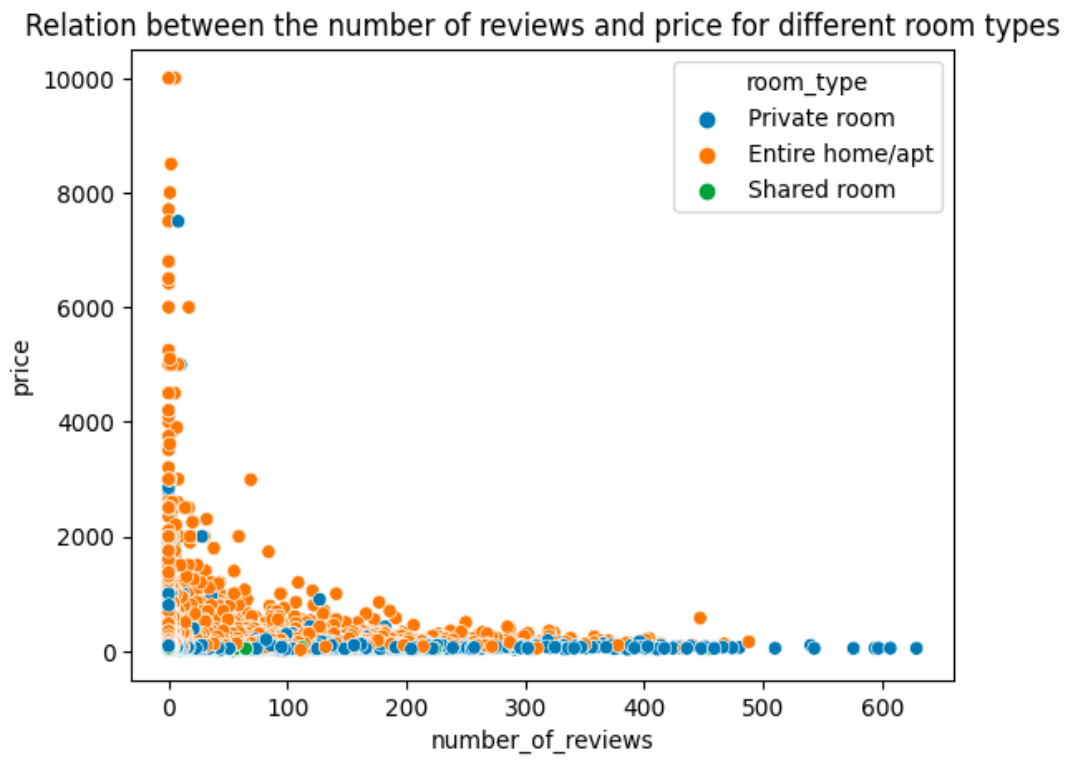
4. How does the availability of the room vary throughout neighbourhoods & room type?



Insights:

The **Private** room type for **Staten Island** are most available room type whereas the **Shared** room type for **Staten Island** is most busiest room type.

5. What is the relation between the number of reviews and price for different room types?

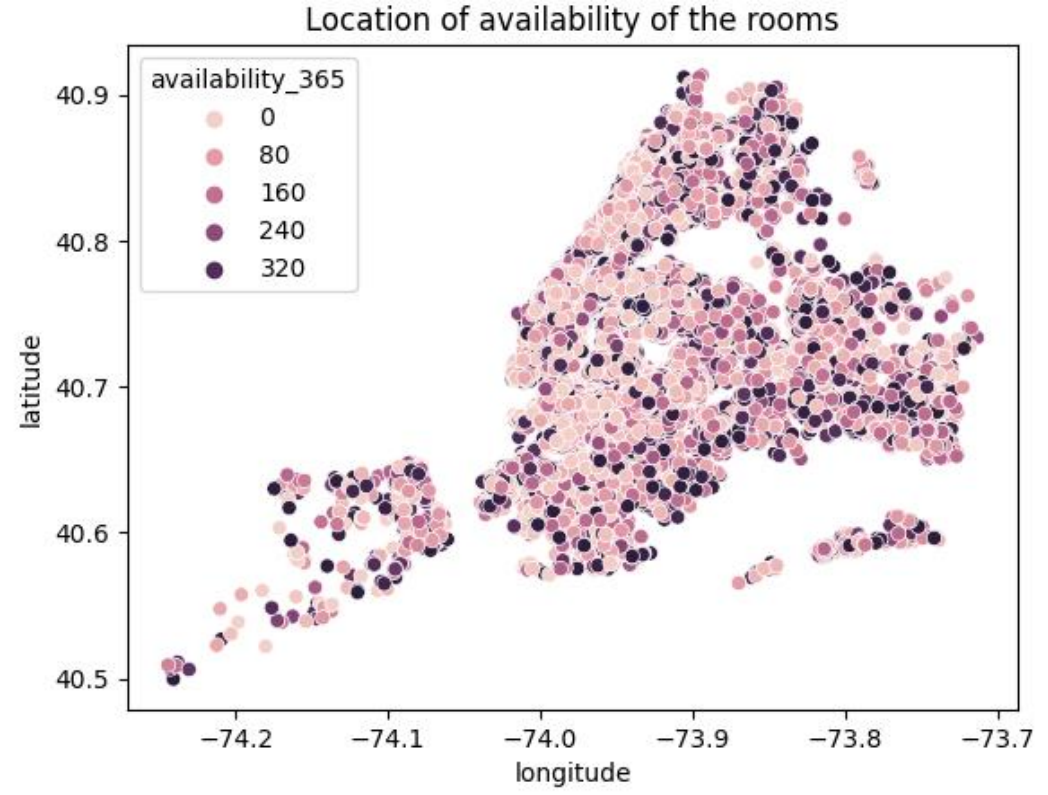


Insights:

•The **number of reviews** given by the customers decreases when the **price** of the room type is huge. Similarly, the **number of reviews** increases when the **price** of the room gets decreased.

•This trend shows that cheaper rooms have more guests as compared to the expensive rooms.

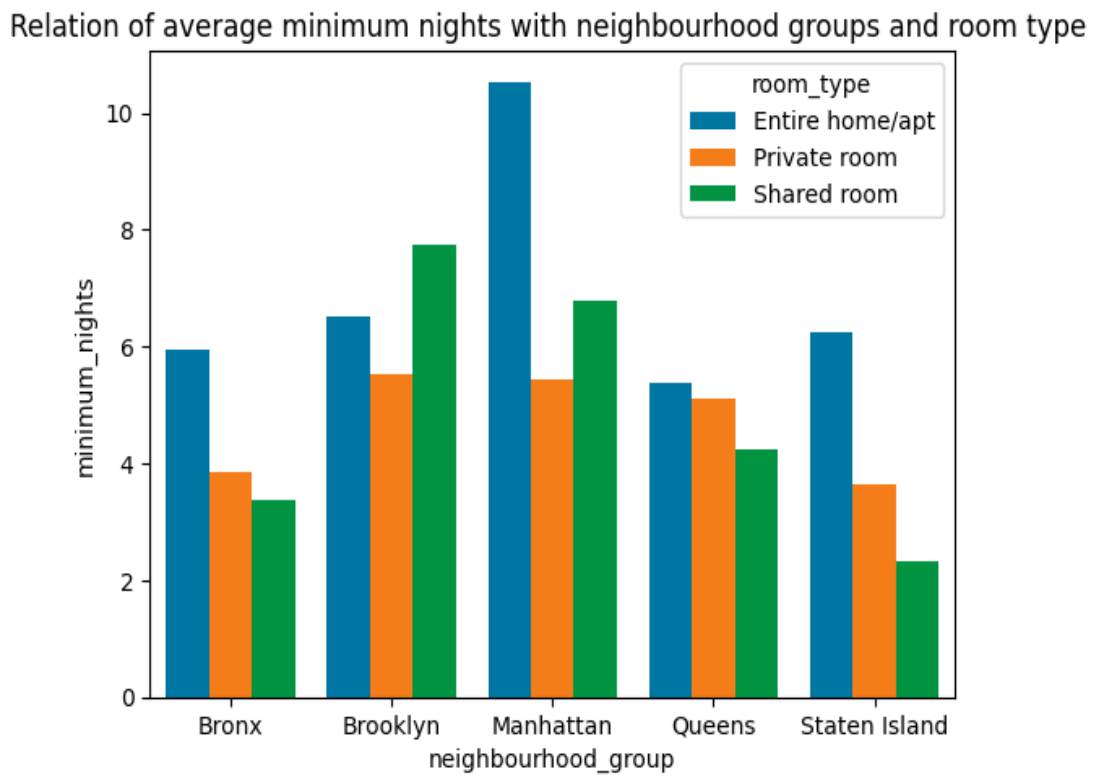
6. What is the location of availability of rooms according to their given latitudes & longitudes?



Insights:

From the chart, it is clearly visible that the center location of the city remains the busiest as the availability of the rooms as very less whereas the outer portion of the city remains mostly available throughout the year.

7. The relation between average minimum nights with neighbourhood groups and room type.

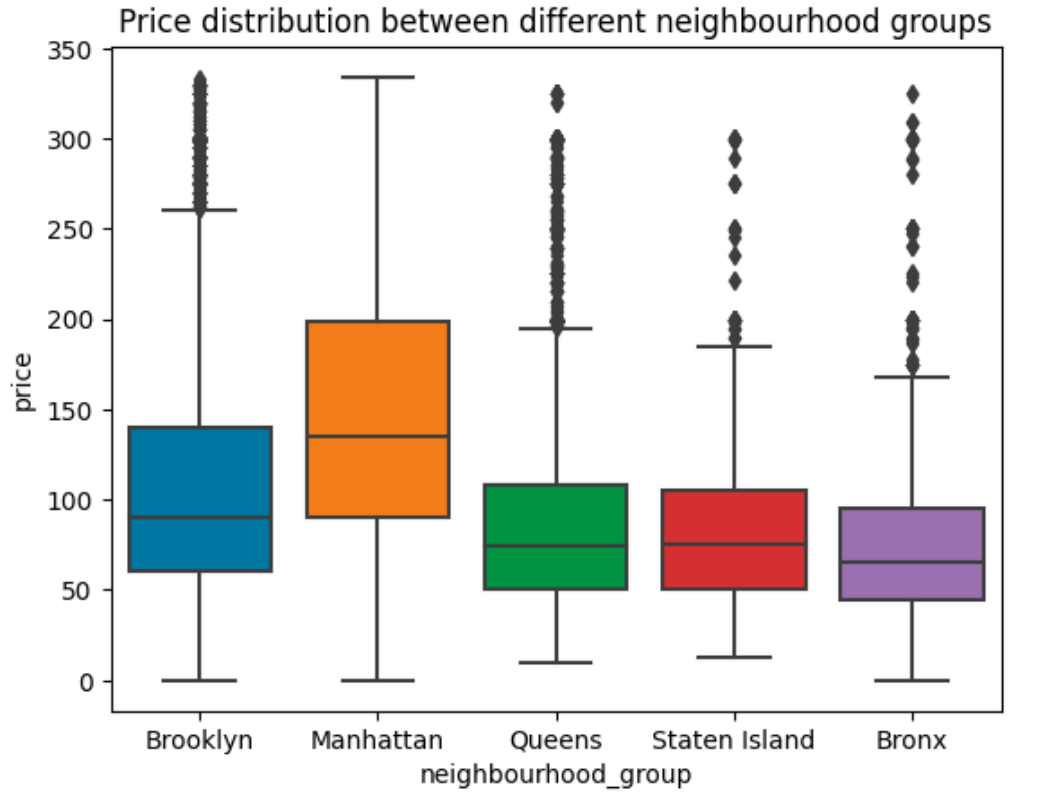


Insights:

•From the chart, it is clearly visible that **Manhattan** have maximum average minimum night (10.54) restriction policy for their **Entire home/Apt** room type as compared to other groups.

•**Staten Island** have less average minimum night restriction policy (2.33) for their **Shared** room type as compared to other groups.

8. The price distribution between different neighbourhood groups

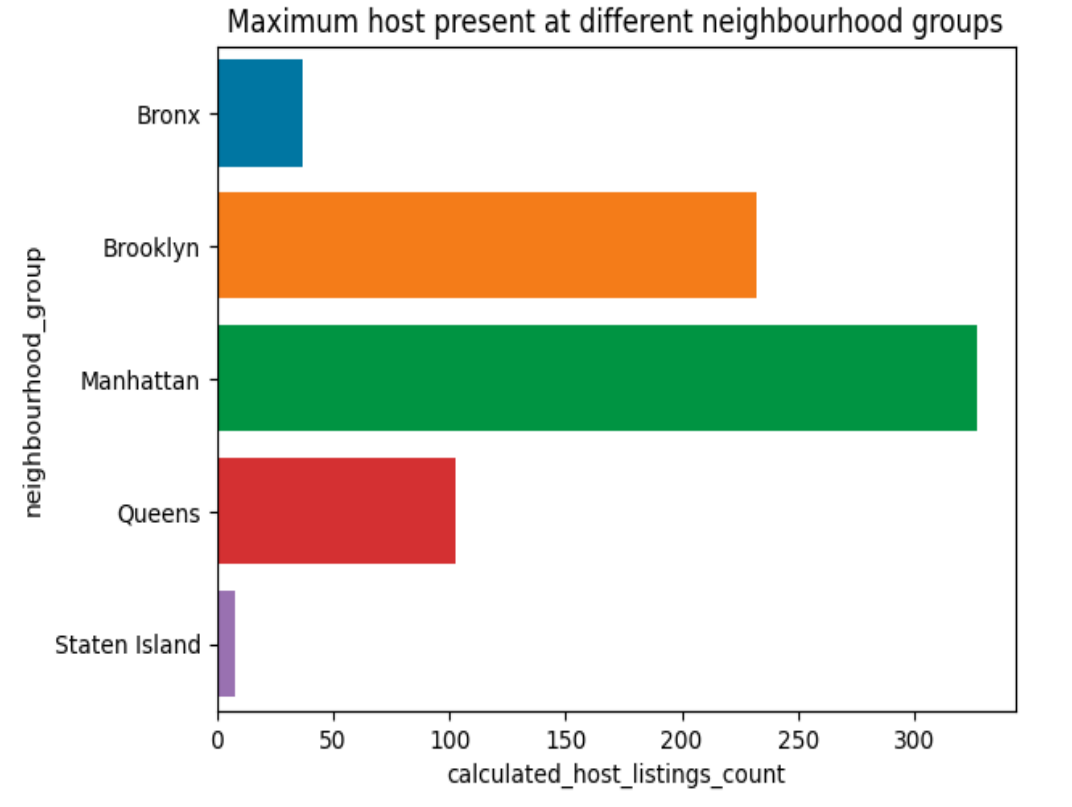


Insights:

•The median of the price in **Manhattan** group is maximum as compared to other groups. The median value is denoted by the center line of each boxplot.

•The **Manhattan** group also has the maximum price denoted by the top line of the boxplot whereas **Staten Island** has the minimum price denoted by the bottom line of the boxplot.

9. The maximum number of hosts present in different neighbourhood groups.

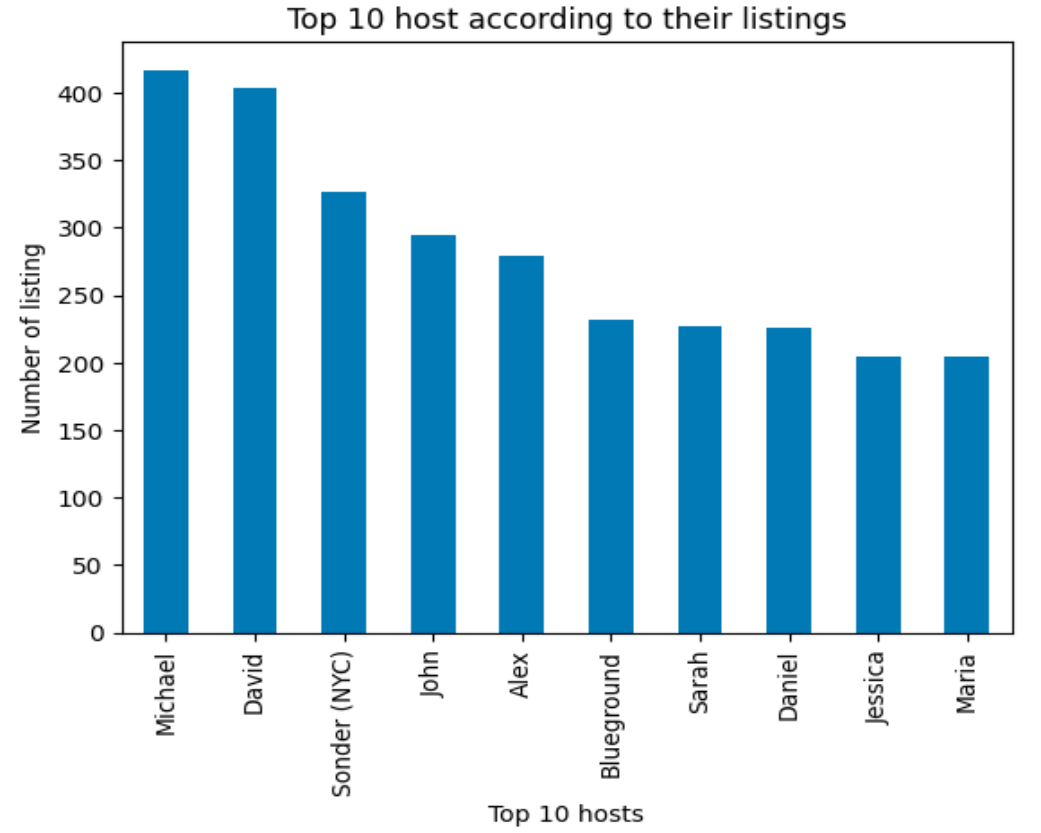


Insights:

•The chart shows that the maximum number of hosts are present in the **Manhattan** neighbourhood group i.e. 327

•The minimum number of hosts are present in the **Staten Island** neighbourhood group i.e. 8

10. The top 10 properties according to their listings.



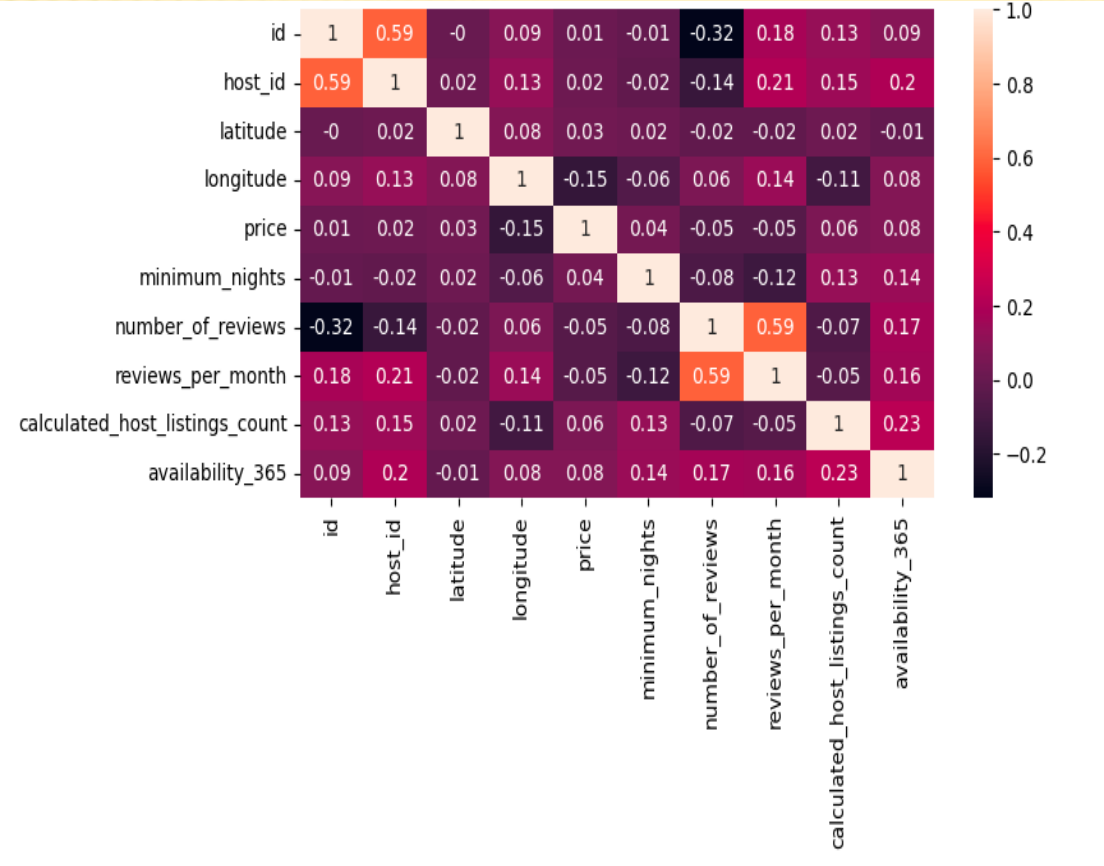
Insights:

•The top 2 host according to the listings are **Michael** (417) & **David** (403).

•The top 2 host mostly have their maximum properties in **Manahattan**, **Brooklyn** & **Queens** a reas, which suggest that these areas are the mostly demanded and popular group.

11. The correlation between different variables.

.



Insights:

•The **price** and **calculated host listings count** has a positive relation i.e. (0.17), which suggests that as the number of listing increases, the price may also increase.

•The **number of reviews** and **reviews per month** has moderate postive relation i.e. (0.59).

•The **minimum nights** and **reviews per month** have a negative correlation i.e. (-0.13), which suggests that if minimum nights increase, then reviews per month decrease.

Conclusion

➢**Queens** and **Manhattan** have the highest percentage of reviews which made them the most popular location for travelers. This also suggests that purchasing more and more properties in these areas and renting them to tourists will be more beneficial in terms of money as customers usually prefer to visit these groups.

➢The **Entire home/apt** category of the room type is the most expensive category with an average price of 211.79 USD whereas the **Shared rooms** are the cheapest room category with an average price of 70.13 USD. Also, the **number of reviews** given by the customers decreases when the **price** of the room type is large, this suggests that people usually prefer to stay at the location where the price is less.

➢The center location of the city remains the busiest as the availability of the rooms is very low whereas the outer portion of the city remains mostly available throughout the year. This suggests that the heart of New York City lies at the center portion, which is mostly **Manhattan**, **Brooklyn** and **Queens** neighbourhood groups.

Contd.

Contd.

- The maximum number of hosts are present in the **Manhattan** neighbourhood group i.e. 327. This suggests that **Manhattan** is the most popular neighbourhood group in New York City. So, the lifestyle and various local tourist place in Manhattan makes it best suitable for traveling in the city.
- **Staten Island** is the least preferred neighbourhood group in the city as per the reviews and price. When tourists travel to such places, they prefer to invest in **Shared** room type as compared to **Private** or **Entire home/Apt**.
- According to the analysis, Airbnb hosts face tough competition, with a small number of hosts controlling a significant portion of the market. To stand out from the competition, hosts might want to think about purchasing property in places where there are comparatively fewer listings.