Capstone Project

Exploratory Data Analysis

Airbnb Booking Analysis

(individual)

Gaurav Jindal

Content

- About Airbnb
- Problem Statement
- Data Understanding and Cleaning
- Exploratory Data Analysis(EDA)
- Challenges
- Conclusion

About Airbnb

What Is Airbnb?

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.

How Airbnb Works

Airbnb has revolutionized the hospitality industry. Prior to 2008, travelers would have likely booked a hotel or hostel for their trip to another town. Nowadays, many of these same people are opting for Airbnb.

The idea behind Airbnb is simple: Find a way for local people to make some extra money renting out their spare home or room to people visiting the area. Hosts using this platform get to advertise their rentals to millions of people worldwide, with the reassurance that a big company will handle payments and offer support when needed. And for guests, Airbnb can offer a homey place to stay that has more character, perhaps even with a kitchen to avoid dining out, often at a lower price than what hotels charge.

Problem Statement

In this project we are analyzing Airbnb's New York City(NYC) data of 2019.
Our main objective is to explore and analyze the data to discover key understandings about listing of properties on the platform. We will perform basic Exploratory Data Analysis(EDA). We will find out key metrics that influence every Airbnb listing based on their location, different hosts and areas, prices, reviews, room type, listing name, traffic and other related factors.

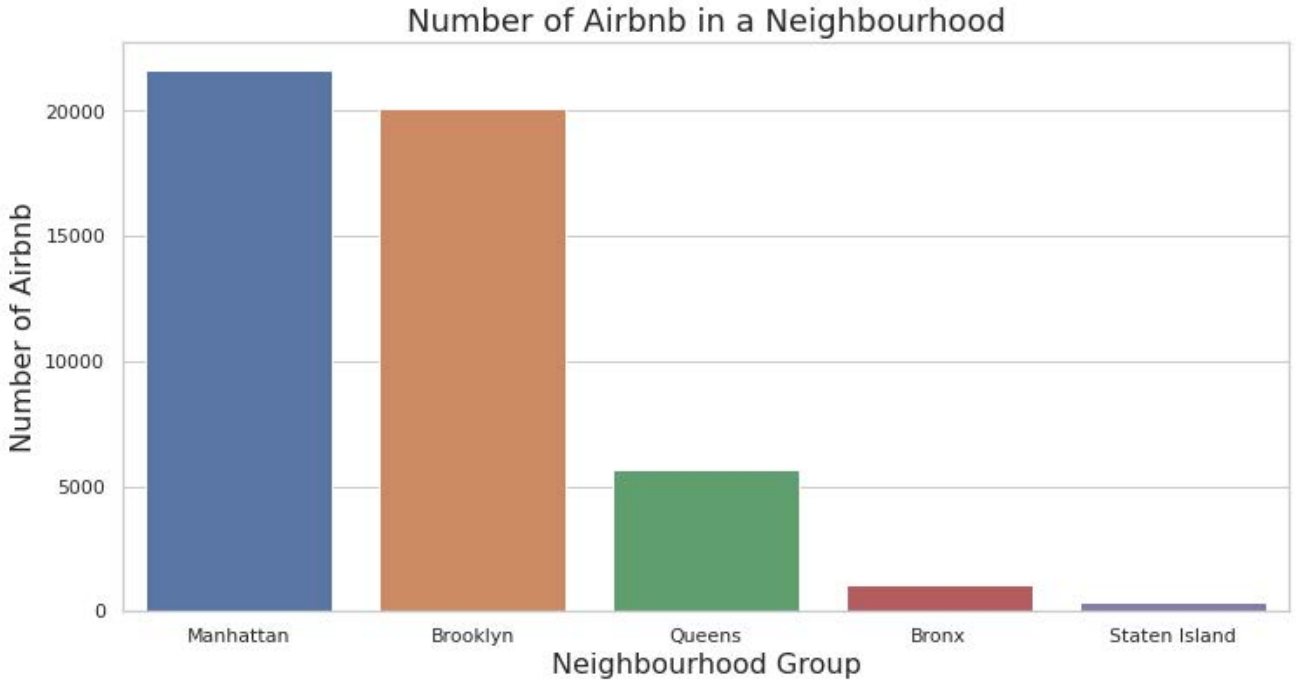
Data Understanding and cleaning

- We have removed the 'id' column as dataframe already have 'host_id' column to identify different hosts
- There were null values in 'name', 'host_name', 'reviews_per_month', 'last_review' columns that were replaced with appropriate values

```
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 15 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   name                            48895 non-null  object
 1   host_id                         48895 non-null  int64
 2   host_name                       48895 non-null  object
 3   neighbourhood_group             48895 non-null  object
 4   neighbourhood                   48895 non-null  object
 5   latitude                        48895 non-null  float64
 6   longitude                       48895 non-null  float64
 7   room_type                       48895 non-null  object
 8   price                           48895 non-null  int64
 9   minimum_nights                  48895 non-null  int64
 10  number_of_reviews               48895 non-null  int64
 11  last_review                     48895 non-null  object
 12  reviews_per_month               48895 non-null  float64
 13  calculated_host_listings_count  48895 non-null  int64
 14  availability_365                48895 non-null  int64
dtypes: float64(3), int64(6), object(6)
memory usage: 5.6+ MB
```

Exploratory Data Analysis(EDA)

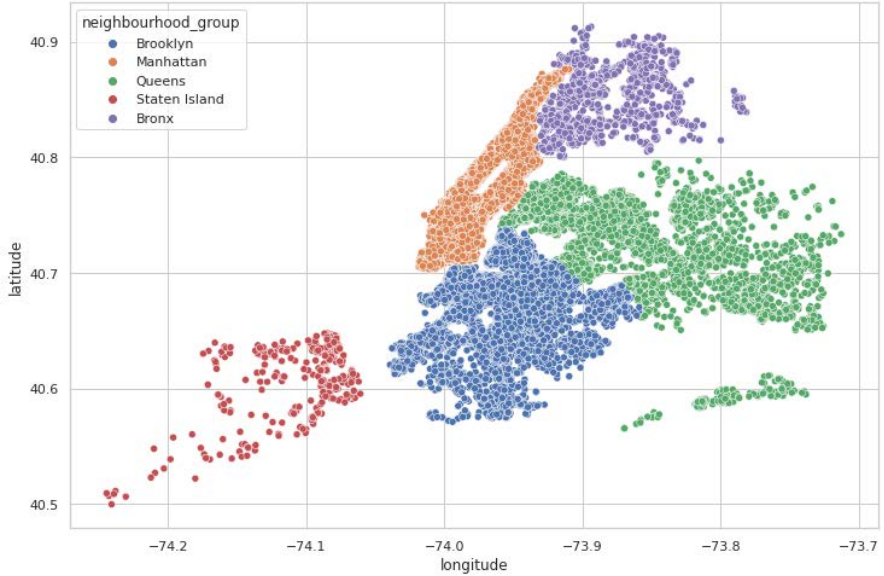
Number of listed Airbnb properties in different neighbourhood groups



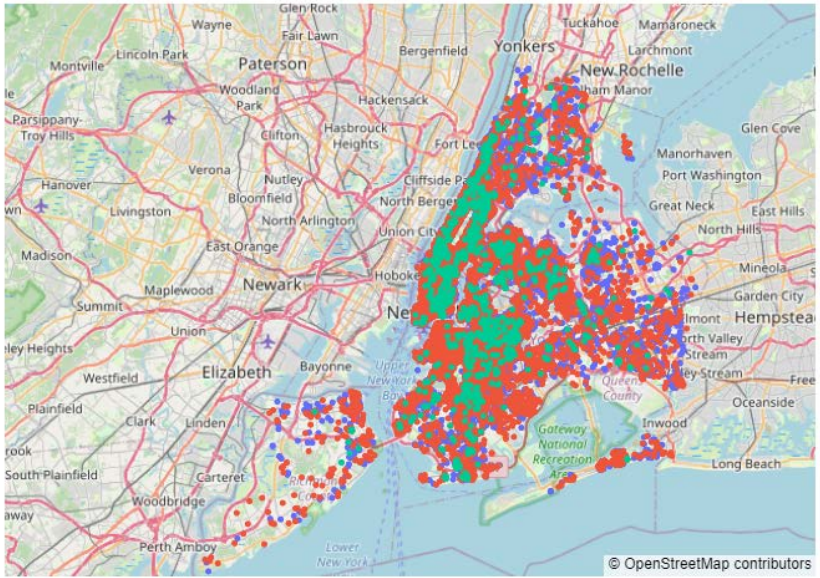
We can observe from the bar plot that Manhattan has the most number of listed Airbnb followed by Brooklyn.

Staten Island has the least number of listed Airbnb

Locations of listed Airbnb properties in different neighbouhood groups of NYC

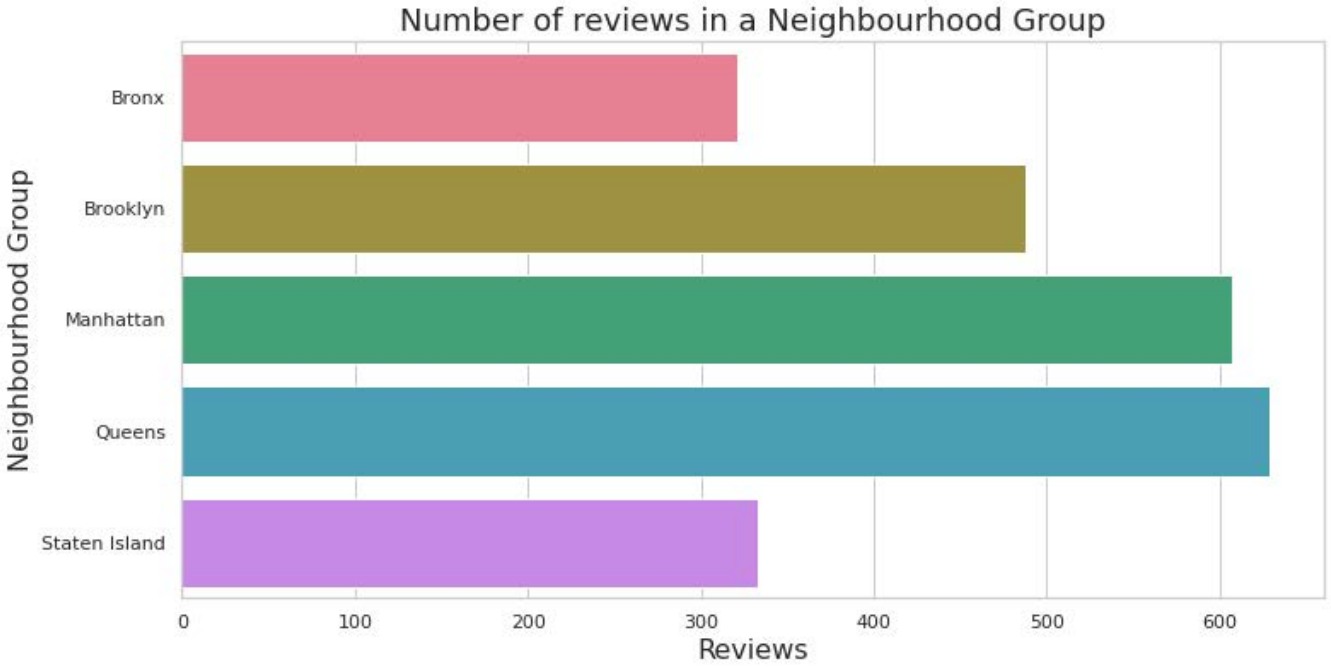


Scatter Plot



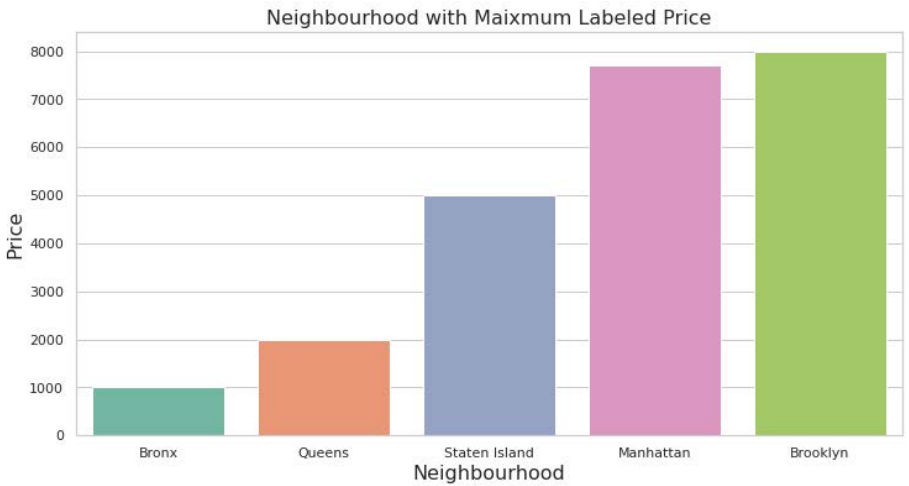
Open Street Map Screenshot

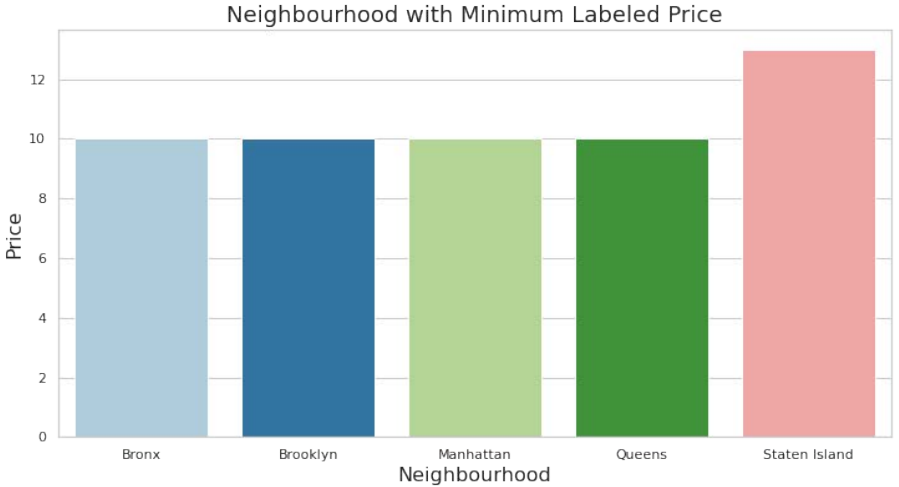
Number of reviews for listed Airbnb properties in different neighbourhood groups



We get the number of reviews in each, from this, we get the idea that as Queens has most reviews so most visitors are in Queens followed by Manhattan.

Maximum and Minimum Price of Airbnb properties in different neighbourhood groups

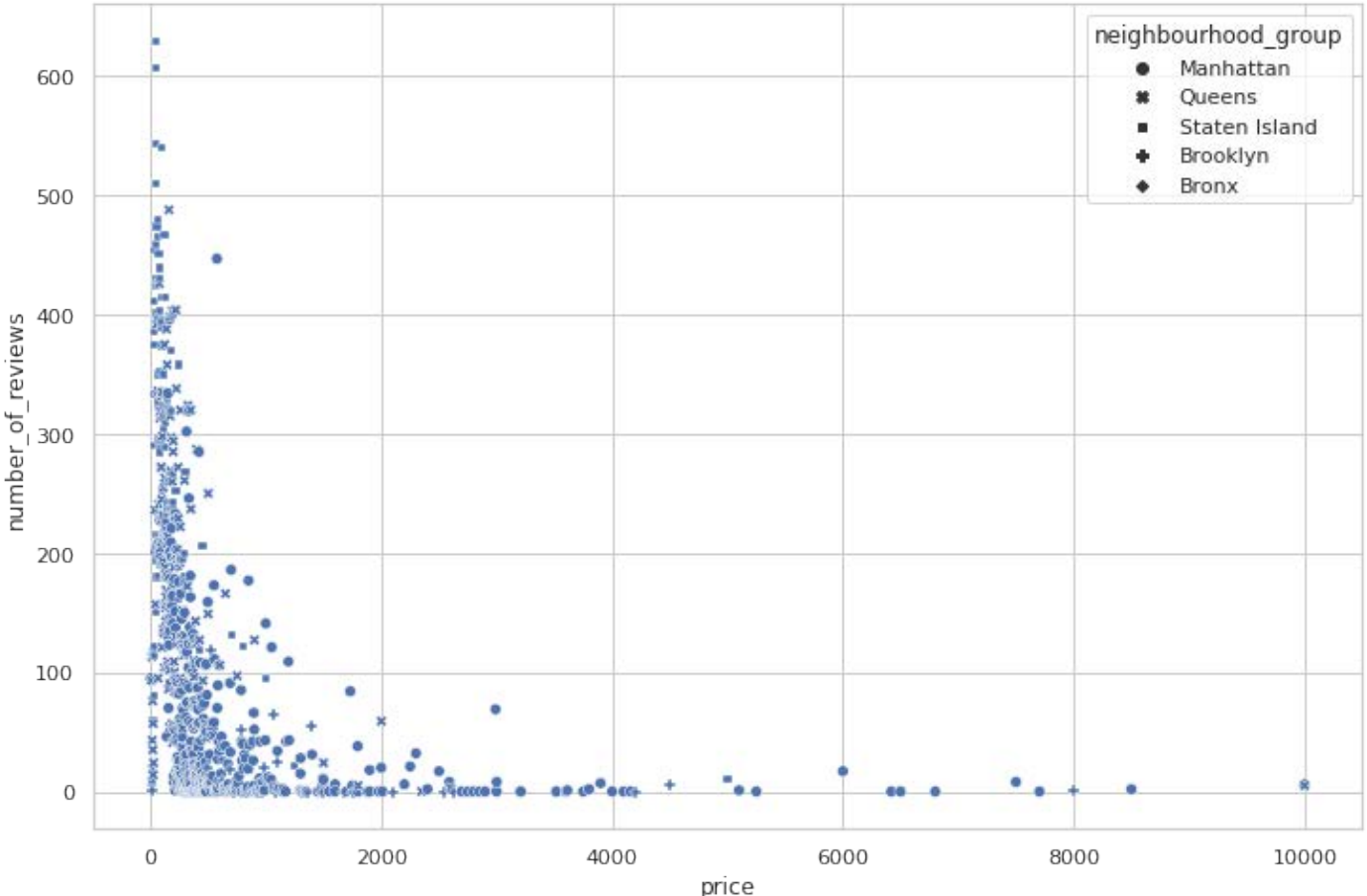




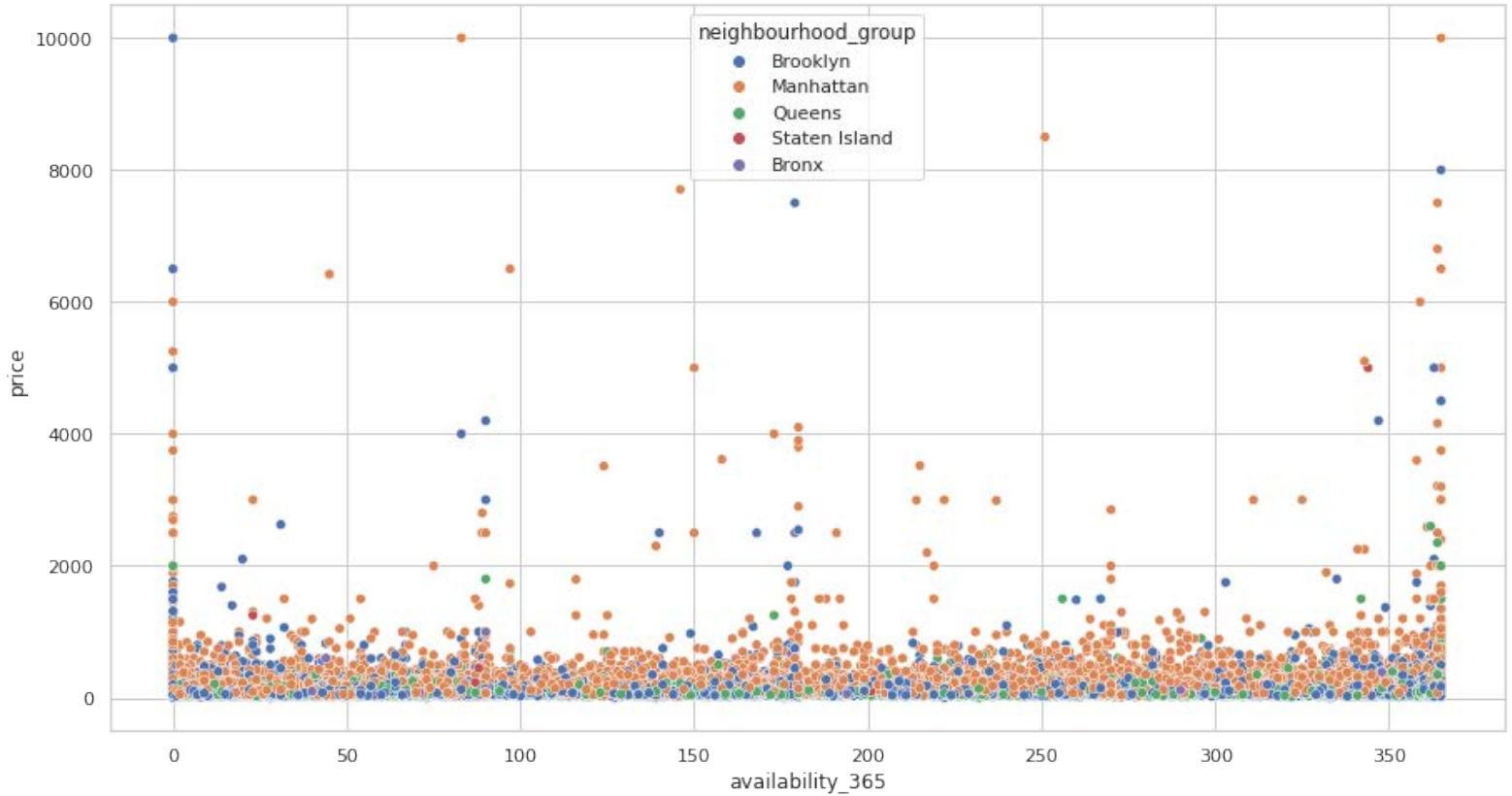
We can observe that Brooklyn and Manhattan neighbourhood groups has maximum price tag on Airbnb and the staten island has low price as compared to Brooklyn and Manhattan.

Staten Island has the highest minimum price among all and rest have the equal minimum price

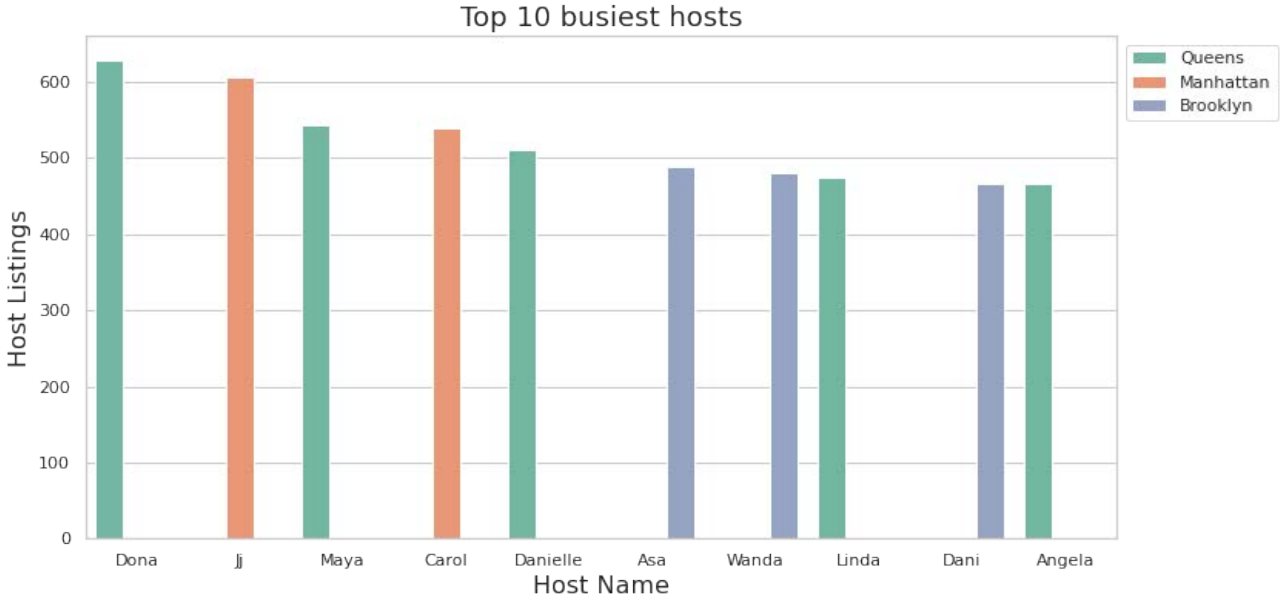
Number of Reviews VS Price comparison



Price VS Availability in a year in different neighbourhood groups comparison

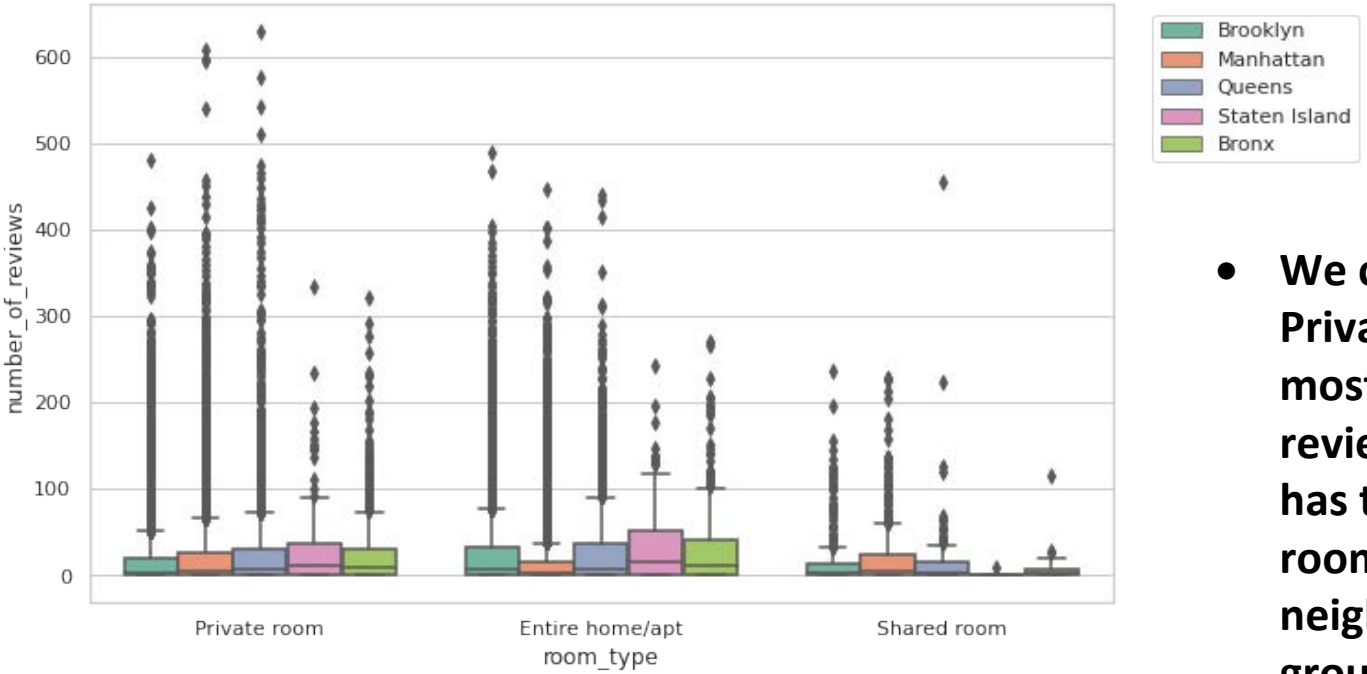


Busiest Hosts among the listed



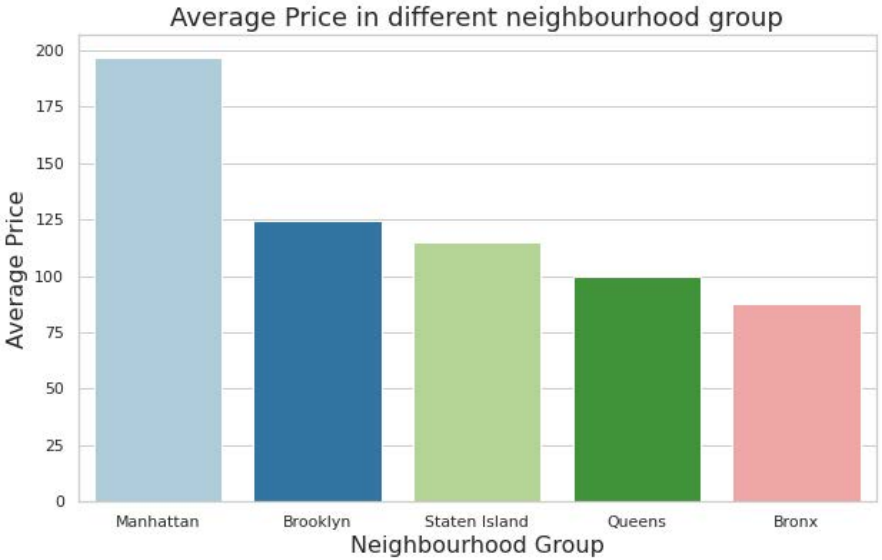
- **We can observe from the plot that Dona is busiest Host and this is in Queens, followed by Jj in Manhattan and so on.**

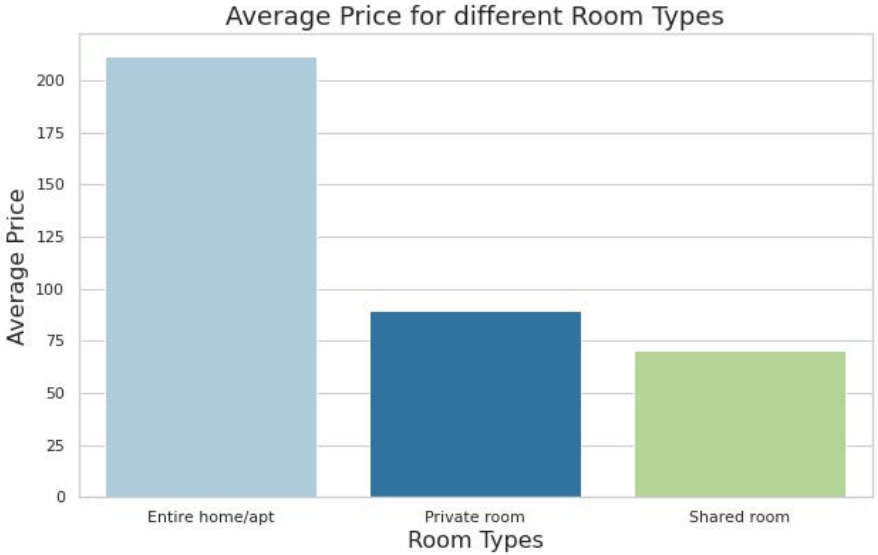
Number of Reviews VS Room Type Comparison



- **We can observe that Private Rooms has the most number of reviews and Queens has the most private rooms among all the neighbourhood groups.**

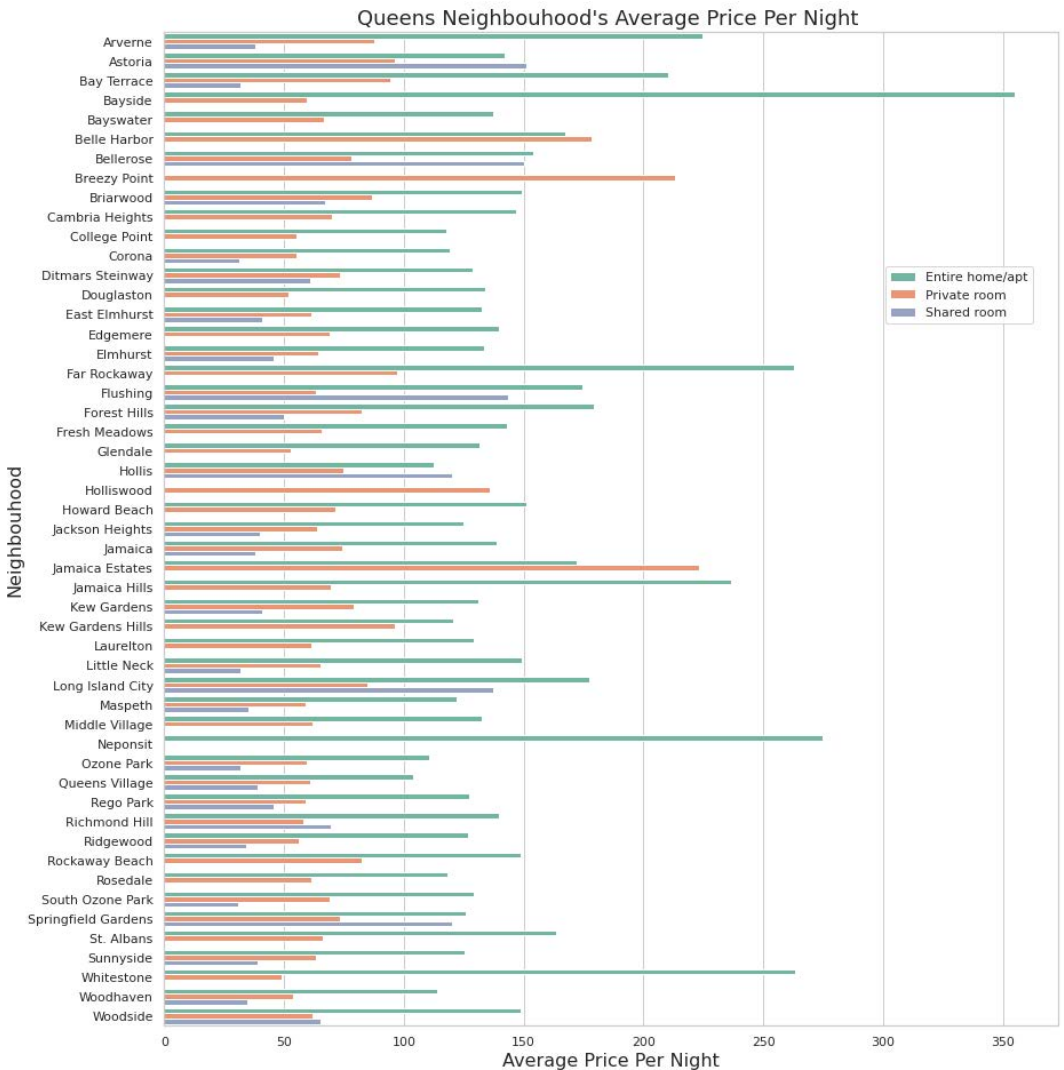
Average Price of listed Airbnb properties in different neighbourhood groups and for different Room types





We can observe that Manhattan has the most expensive listed Airbnb properties

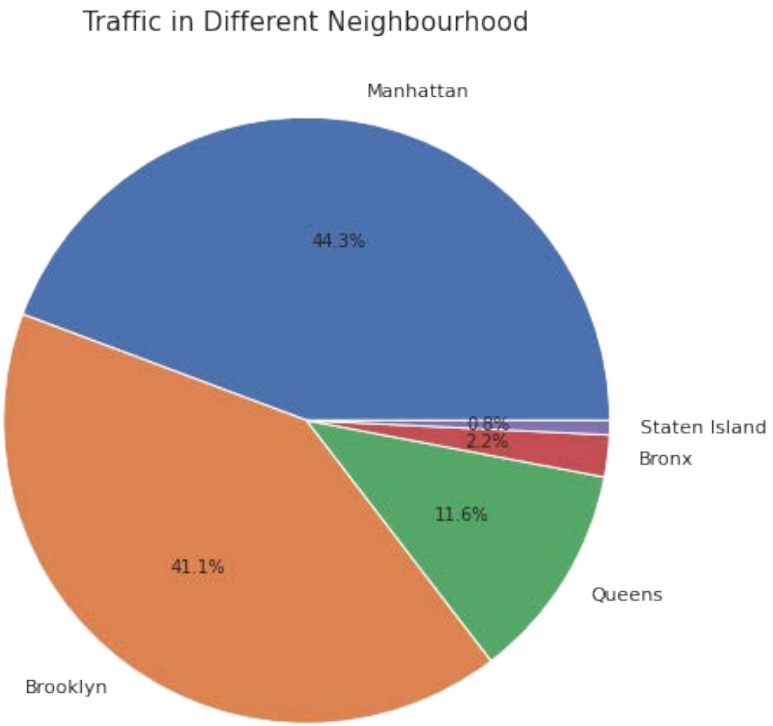
We can observe that Entire home/apt is the most expensive Airbnb Room Type



Average Price per night of listed Airbnb property of different areas in Queens neighbouhood

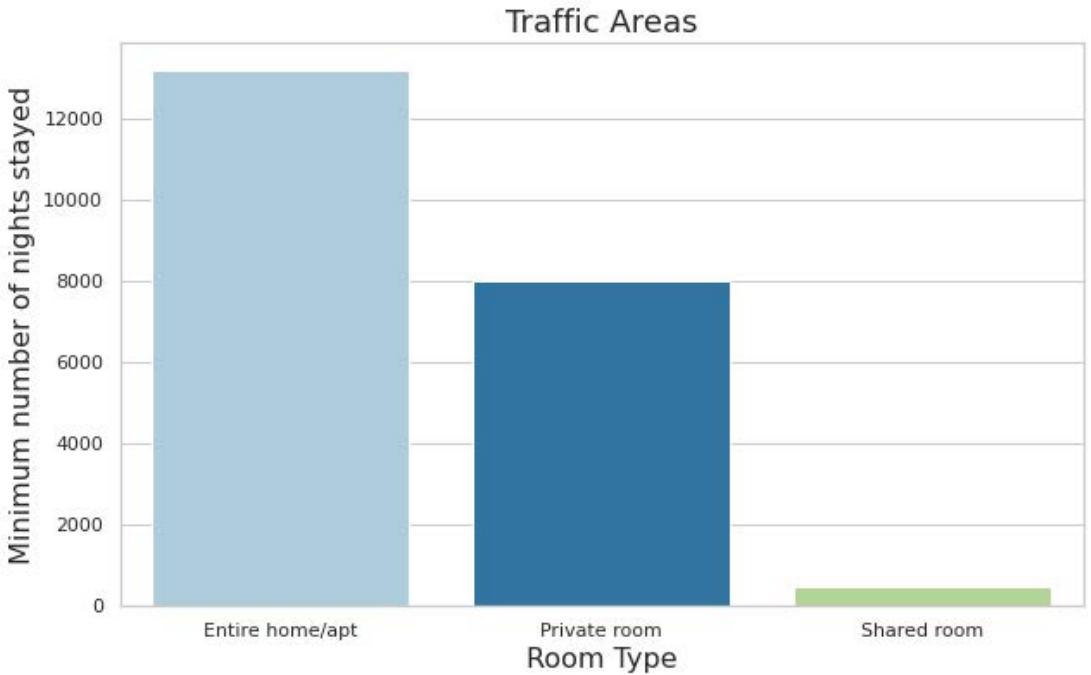
- This bar plot shows average price per night of listed Airbnb properties for different areas in Queens neighbourhood
- We can observe that 'Bay Terrace' is the most expensive

Traffic Details in different neighbourhood groups



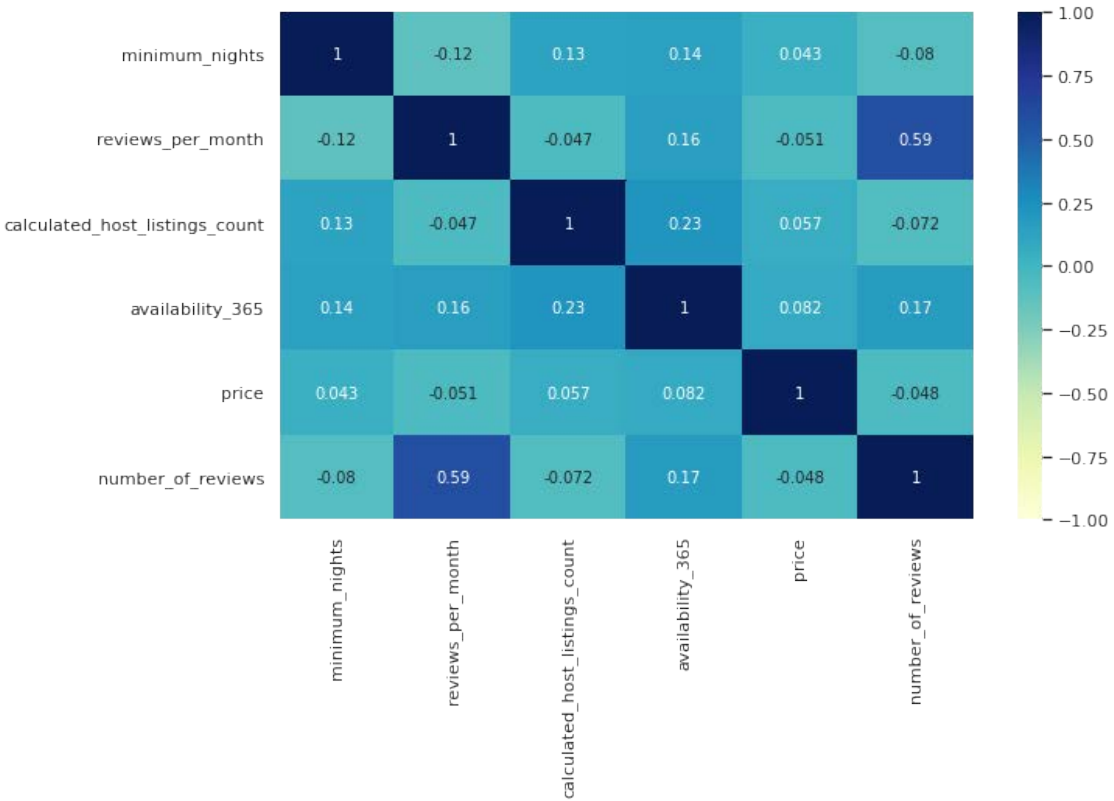
- **We can observe that Manhattan is the most Traffic area since Manhattan has the most number of minimum night stays**

Minimum number nights stayed VS Room Type comparison



- **We can observe that Entire home/apt is the most Traffic room type since Entire home/apt has the most number of minimum night stays**

Correlation check



- **no_of_reviews and reviews_per_month columns has high co-relation for obvious reason.**
- **The price column has very low corelation with other features.**

Challenges

- Handling NaN values, null values and duplicates
- Removing the outliers for some data set Finding and sorting few impossible dataset
- Computation Time
- For answering some of the questions we had to understand the business model of Airbnb that how they work

Conclusion

In this EDA project, different use cases are analysed for the given dataset to make better business decisions and help analyse customer trends and satisfaction, which can lead to new and better products and services.

Few key points:

- **Manhattan is the most focused place in New York for hosts to do their business**
- **Host Sonder(NYC) have the most listings and these listings are in Manhattan area.**
- **Manhattan has the most expensive Airbnb properties**
- **Dona is busiest Host and this is in Queens, followed by Jj in Manhattan and so on.**
- **Manhattan has the most number of minimumm night stayed, so Manhattan is the most Traffic area.**