Capstone Project - 1

Airbnb Bookings Analysis

Name: Debrani Priya
Cohort: Durban

Let’s Explore:

Defining the problem statement

Defining the data

Data Cleaning

EDA

Insights about hosts

Insights about neighbourhood groups

Insights about neighbourhoods

Insights about room types

Conclusion

Dataset Definition:-

1) **id:** It is an unique id given to the property listed in airbnb NYC which is a numerical variable.

2) **name:** It represents the name of the airbnb listed property which is a categorical variable.

3) **host_id:** This is an unique id given to the host of the property which is a numerical variable.

4) **host_name:** The name of the host of the property listed which is a categorical variable.

5) **neighbourhood_group**:This represents a big neighborhood inside which there are many mini neighborhoods which is a categorical variable. There are 5 neighborhood groups in the data:
   i) Manhattan
   ii) Brooklyn
   iii) Staten Island
   iv) Queens
   v) Bronx

Defining the problem statement

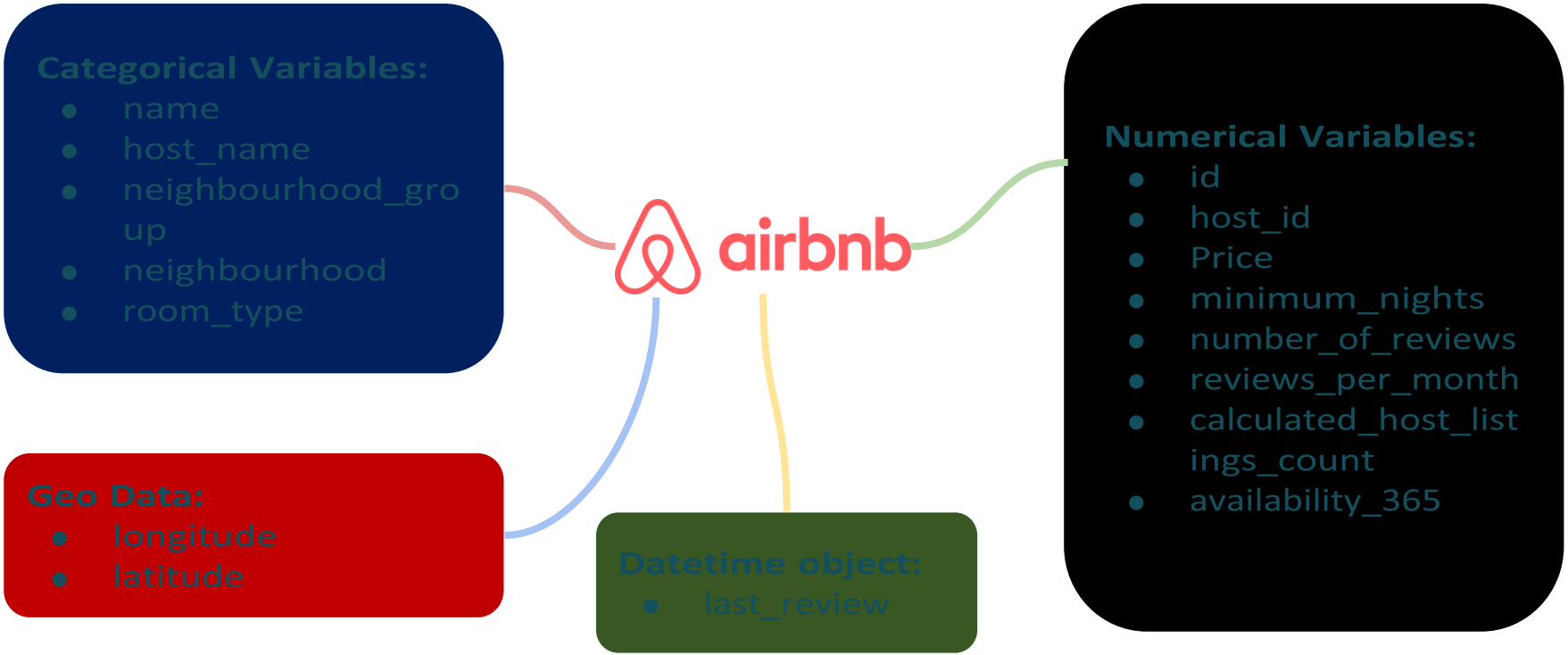
Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Based in San Francisco, California, the platform is accessible via website and mobile app.

In this EDA project, a dataset of 49000 Airbnb listings in NYC is provided and my task is to explore and find out a few interesting insights which could be helpful in decision making. The main purpose of EDA is to detect any errors, outliers as well as to understand different patterns in the data.

To accomplish this, the task was divided into 2 parts as follows:

- **Data pre-processing**: The first step consists of looking out for duplicates, missing values and outliers. There were a few missing values present which were dropped or imputed according to the variable at hand.
- **EDA**: Performed some univariate, bivariate and multivariate analysis to uncover insights about hosts, neighbourhood groups, neighbourhoods and room types in Airbnb NYC.

Data Summary



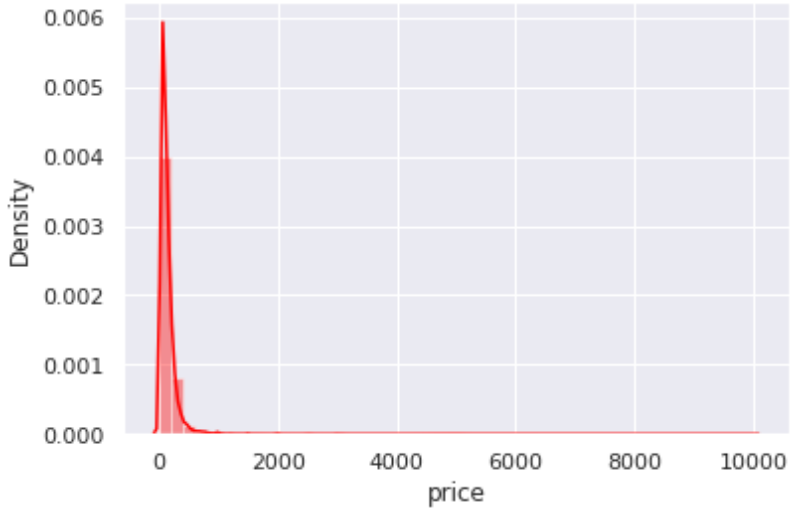
Data Cleaning

- The first step is to check for duplicate values. As we can observe that there were no duplicate values present in the dataset.
- The second step was to find out missing values in the dataset. We found 4 columns in which there were missing values present.
- We dropped id, name and last_review as these were irrelevant for our analysis.
- We imputed all NaN values in host_name by ‘No Name’ and reviews_per_month by ‘0’.

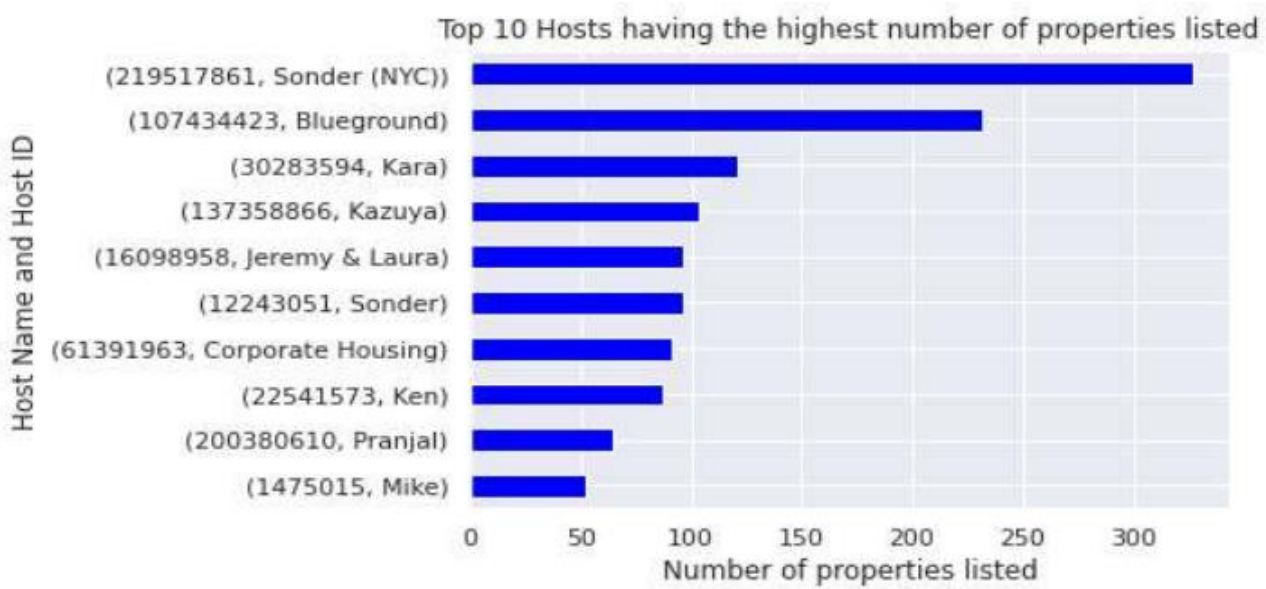
```
df_airbnb.info()

<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
```

Exploratory Data Analysis

- The price column was heavily skewed.
- Most of the prices was in between 10 to 200$.
- We didn’t treat the high values as outliers because there are few observations in minimum_nights which are high as 1250 and for those few observations the price can naturally go high upto 10000$.

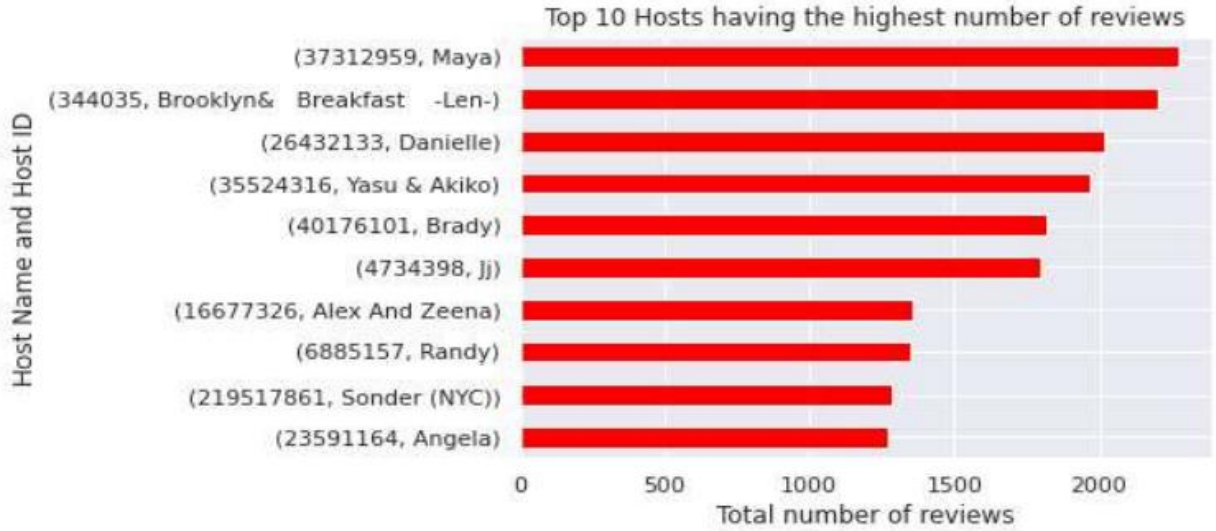


Exploratory Data Analysis



As we can infer from the above graph
Sonder (NYC) has the most properties listed

Exploratory Data Analysis



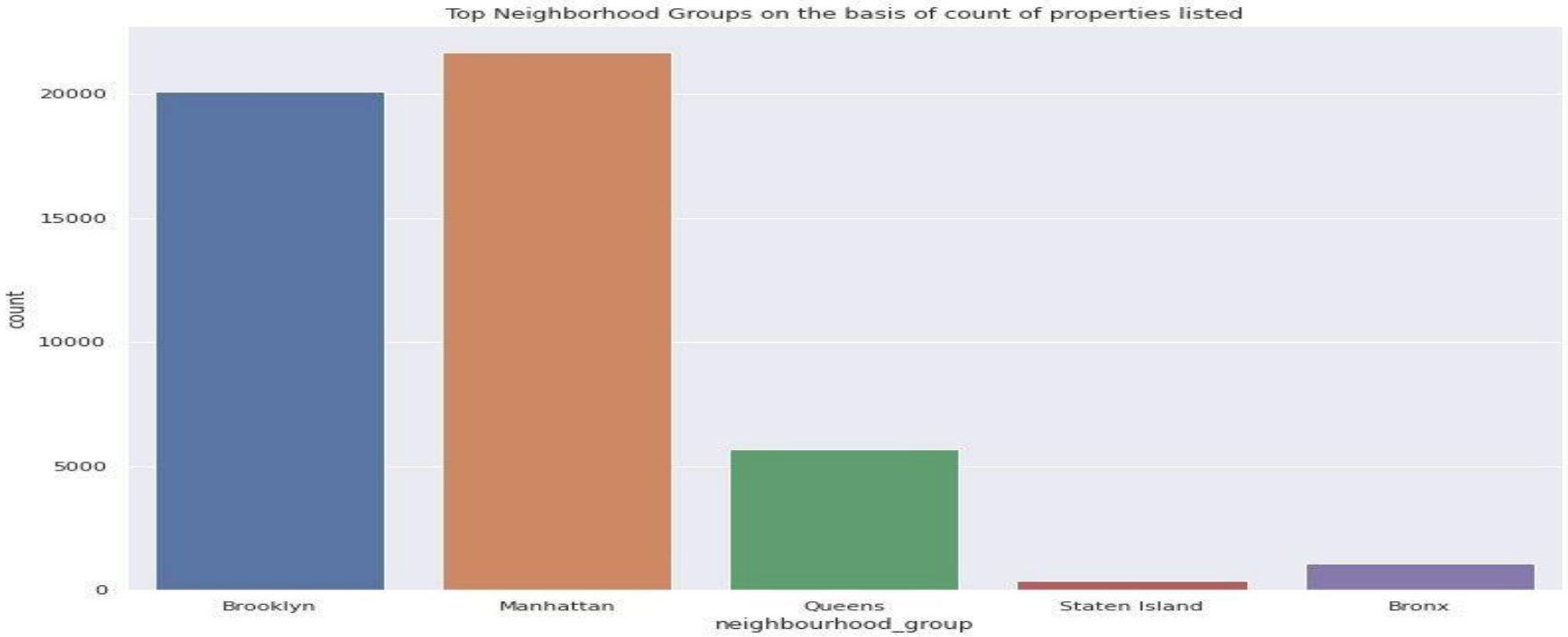
We can clearly see that Sonder (NYC) is not the top host who has received the most number of reviews. As host Maya receives the most number of reviews we can infer that she gets the most number of customers.

Maya receives the most number of customers and there are multiple reasons behind it:

- The price at which she offers her properties is less than the average of all neighbourhood groups
- The condition for minimum nights is 1 which is way less than many other properties
- She is available for quite a healthy number of days
- As she receives the most reviews, customers will naturally turn their heads towards her properties

Hence, we can conclude that host Maya is the busiest in NYC.

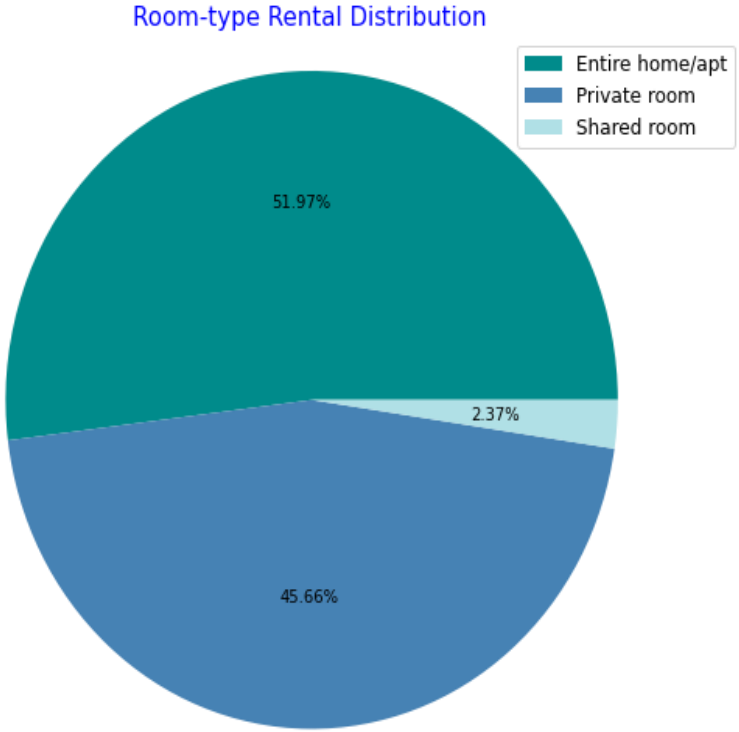
Exploratory Data Analysis



Manhattan has the most number of listings followed by Brooklyn

Exploratory Data Analysis

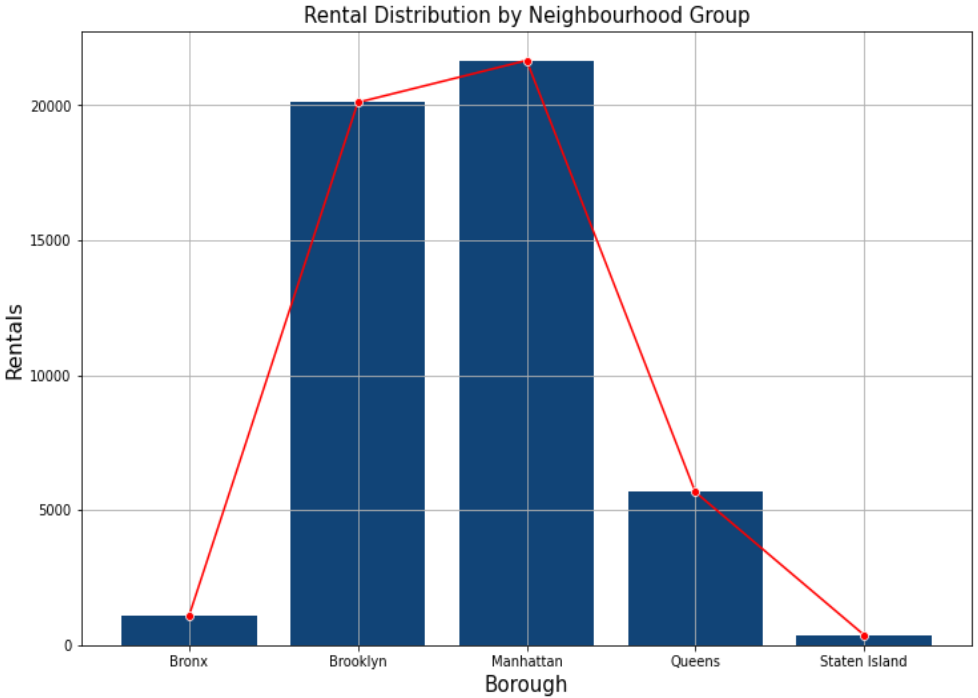
From the whole rentals available in the dataset,
52% of them correspond to entire-home apartments,
46% to private-room rentals and the minority
remaining corresponds to shared rooms with
a 2% of the sample.



Exploratory Data Analysis

As it's reflected in the visualization, Brooklyn and Manhattan boroughs concentrate the majority of the listed rentals on Airbnb, adding up more than 40,000 rentals between the two of them. This means that the bulk of visitors of New York stay in properties, rooms or residencies located in these areas.

rental distribution by neighbourhood

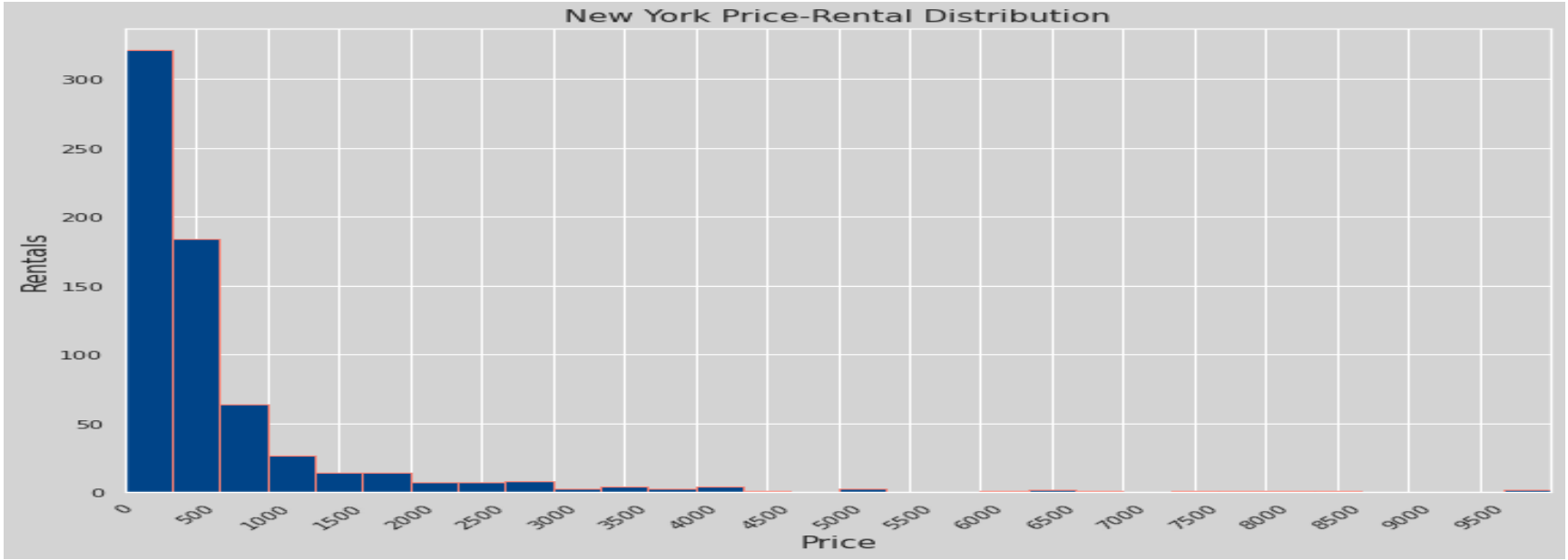


Exploratory Data Analysis

We can infer that the listings in Staten Island seems to be more available throughout the year to more than 300 days. On an average, these listings are available to around 210 days every year followed by Bronx where every listings are available for 15 0 on an average every year.

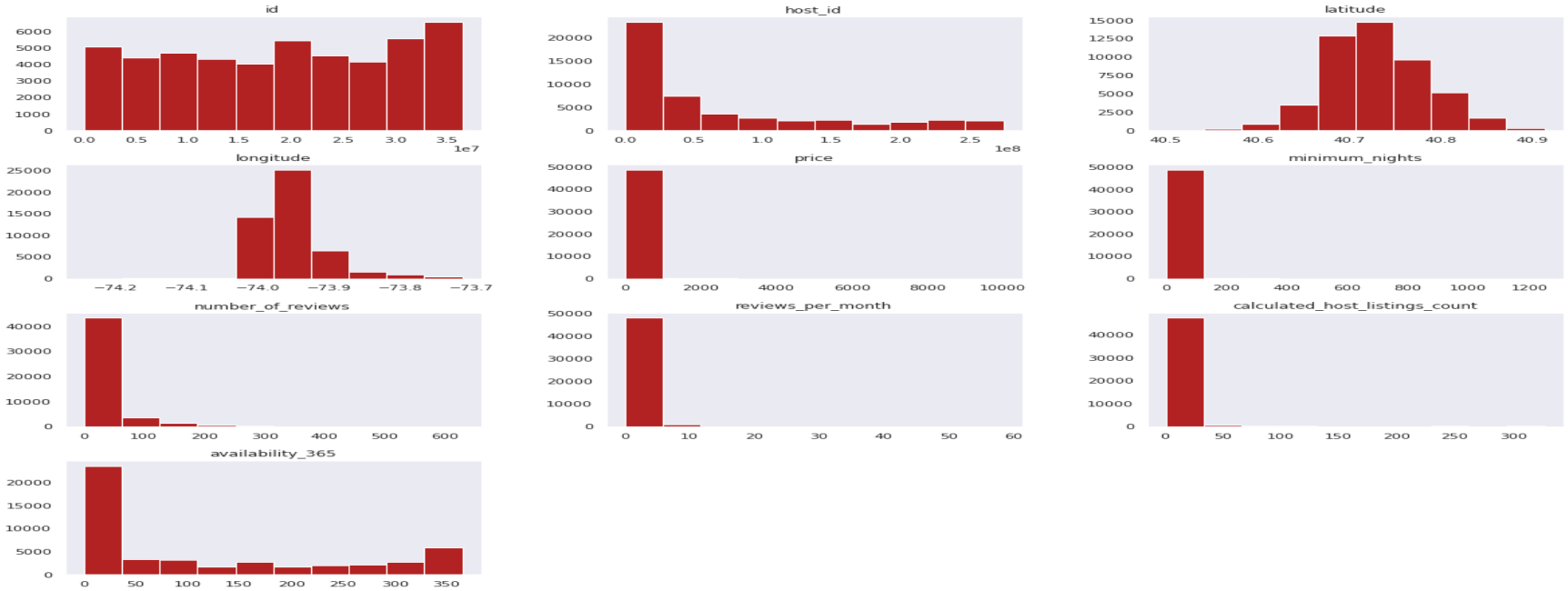
Exploratory Data Analysis

New york price rental distribution



From the above Graph we can say that Price distribution is focused around the $ 300–400 prices with few observations that present higher prices.

Exploratory Data Analysis

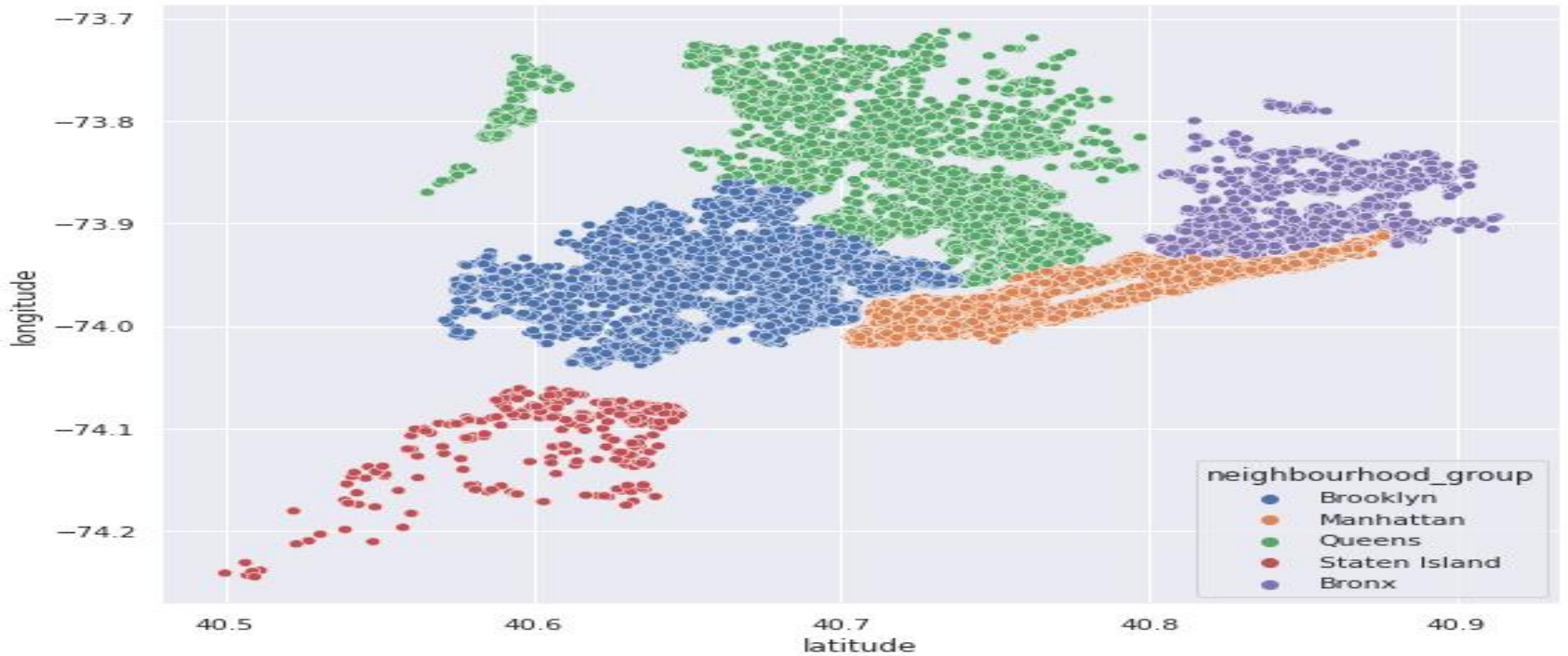


Some of the numerical values are positively skewed; Therefore, for the analysis, I chose to focus on the median values, because it is less susceptible to skewed data.

Exploratory Data Analysis

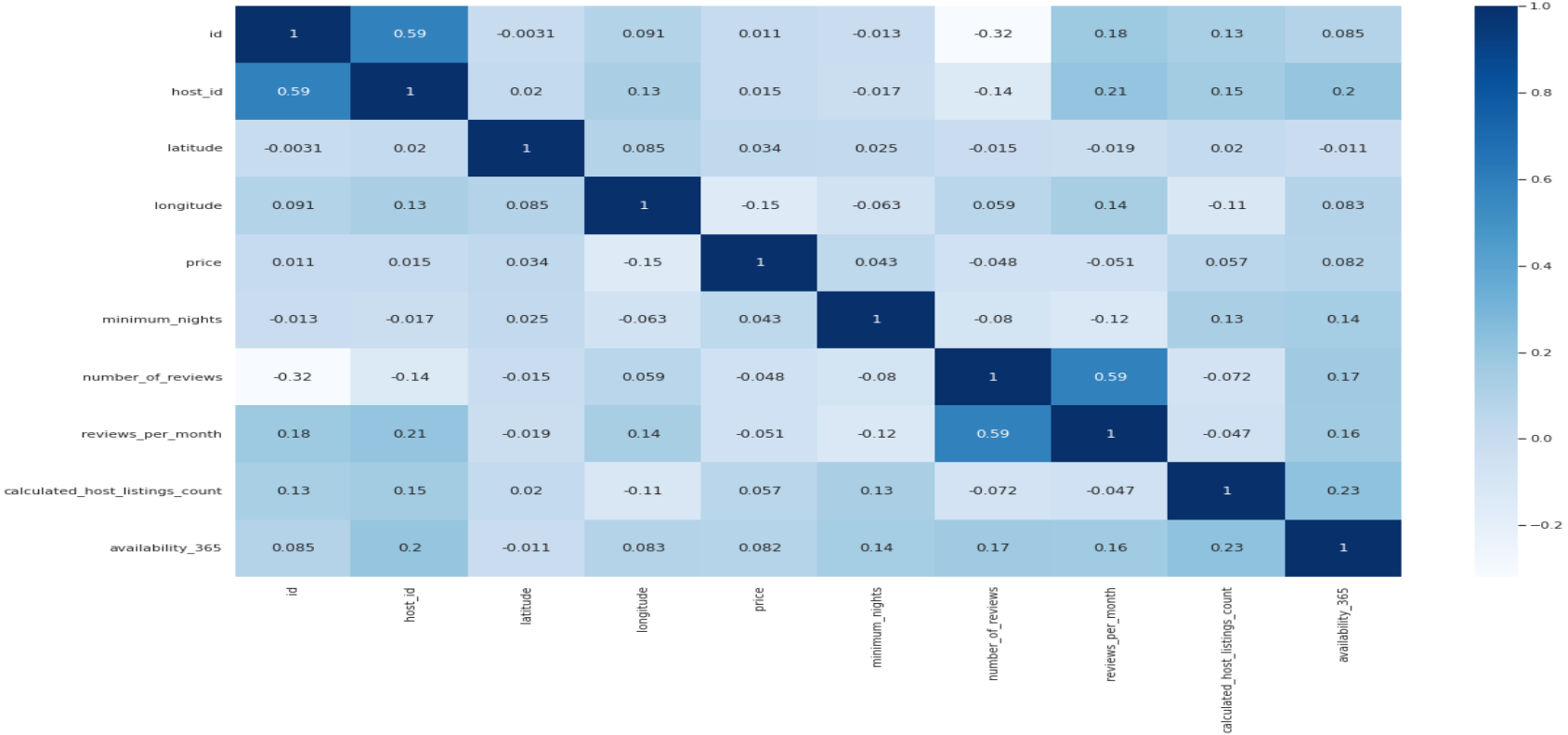
This graph below is a plot between room types and the number of airbnbs that are of that type.From this graph we i nfer the maximum number of airbnbs in the whole of NYC are that of entire room type.

Exploratory Data Analysis



The above resemble the map of NYC and shows the various neighbourhoods and the properties listed in each neighbourhood.

Exploratory Data Analysis



From the above Heatmap graph, we can infer that There's correlation among host_id to reveiws_per_month & availability_365. Also there's noticiable correlation between min_nights to no_of_listings_count & availability_365. Price also shows some correlation with availability_365 & host_listings_count.

Conclusion

We have reached the end of our analysis of Airbnb listings in NYC. Let us now summarize few of the important insights we gathered:

1) Manhattan and Brooklyn are the most expensive neighborhoods and they receive the most traffic as well. Due to many tourist attractions and the number of properties available, people tend to visit these two areas comparatively more than other ones.
2) Host Maya is the busiest host in NYC and there are multiple reasons in favor of it like price, minimum nights, availability and number of reviews. She has a total of 5 properties listed in the same neighborhood.

3) Entire Home/Apt is the costliest room type available but still the most preferred ones for the customers. Entire Home/Apt and Private Rooms

receive way more traffic than Shared Rooms and as a result Shared Rooms stay available for most of the time out of the 365 days.

4) The average price for Private Rooms in Staten Island is the least and has a good availabi lity out of 365 days which makes a good choice for customers seeking low cost accom modations.

5) Bronx is the least expensive neighbourhood and very less preferred by customers.

Thank You!