Capstone Project

Airbnb Booking Analysis

By –SACHIN YALLAPURKAR

AirBedandBreakfast

- Airbnb is an online marketplace for short-term homestays and experiences
- Founded in 2008, it acts as a broker and charges a commission from each booking.
- It is available in 65000 cities and over 191 countries around the world.
- In 2021, Airbnb generated $5.9 billion revenue.
- There are over seven million listings on Airbnb, run by four million hosts.

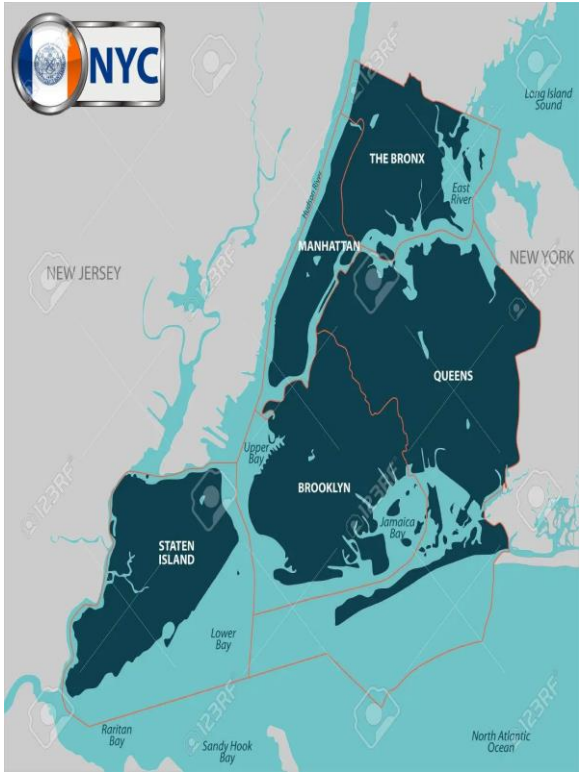
Problem Statement

Explore and analyze the data to discover key understandings (not limited to these) such as :

1. What can we learn about different hosts and areas?
2. What can we learn from predictions?
    - 2.1 Type of room
    - 2.2 locations,
    - 2.3 prices,
3. Which hosts are the busiest and why?
4. Is there any noticeable difference of traffic among different areas and what could be the reason for it?
5. What is the percentage of listings owned by Airbnb in different neighbourhoods?

City : New York City

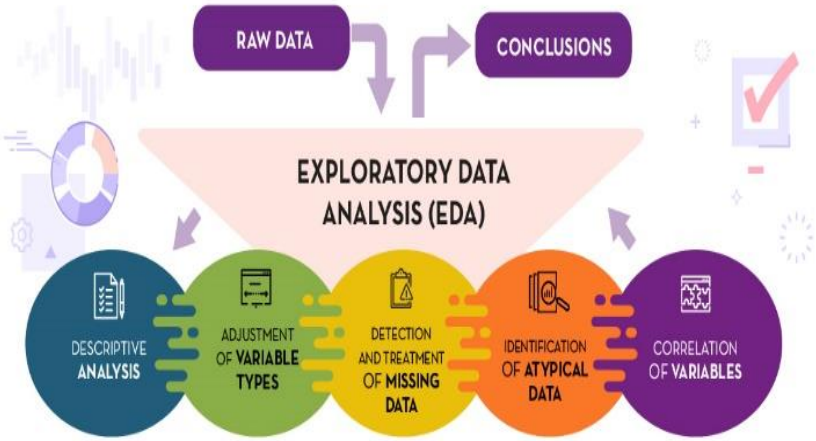
- Dataset provided for Analysis is of New York City.
- The City is divided into 5 Neighbourhood groups namely **Manhattan, Bronx, Queens, Brooklyn, and Staten Island.**
- These Boroughs are further divided into distinctive neighborhoods.
- Today New York is USA’s largest Short-term Rental market, with >30K hosts.



What is Exploratory data analysis (EDA)?

Exploratory data analysis (EDA) is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods.

Steps in Exploratory Data Analysis



Step 1: Descriptive analysis

Synthesis of the information provided by the dataset, extracting its most representative characteristics.

Step 2 : Adjustment of Variable types

Verify that the variables have been stored with the appropriate corresponding value type.

Step 3 : Detection and treating of Missing Data

Identify some of the missing data in the variable.

Steps in Exploratory Data Analysis

Step 4: Detection and treatment of atypical Data.

To identify data with values significantly different from those of the variable.

Step 5: Correlation of variables

Analyzing the relationship between two or more variables.

Step 1 - Descriptive Analysis

The given dataset has 48,895 observations and 16 different features. Let us look what each feature is all about.

1. id id given to listings
2. name name of the listing
3. host_id unique host ids
4. host_name Gives host name
5. neighbourhood_group It contains 5 neighbourhood groups namely : Brooklyn,Manhattan,Queens,Staten island,Bronx.
6. neighbourhood There are total of 221 different neighbourhoods.
7. latitude It gives the latitude of house listing.It helps in getting the location.

Step 1 - Descriptive Analysis

1. longitude It gives the longitude of the house listing. It helps in getting the location.
2. room_type There are total of 3 different types of rooms available on Airbnb i.e. Private room, Entire home or apartment and shared room.
3. price It tells about the price of each listing
4. minimum_nights It tells about minimum nights spent by people in listing
5. number of reviews It gives the total number of reviews
6. last_review It tells about when the last review was given
7. reviews_per_month It tells about review got by listing per month
8. calculated_host_listings It tells about the number of times a host was listed or booked by people
9. availability_365 It tells about availability of listing out of 365 days

Step 2 - **Adjustment of Variable types**

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 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: 6.3+ MB
```

We can see that datatype of columns is same as what is expected.

For example

- **id,host_id,price,number_of_reviews,calculated_host_listings_count,availability_365** are supposed to be integer datatypes and they are in actual int64.So they are compliant.
- **name,host_name, neighbourhood_group, neighbourhood, room_type, last_review** are supposed to be characters and they are "object" datatype.
- **latitude, longitude** are supposed to be floats and in the given dataset they are float64.

So the dataset doesn't need to be adjusted for variable types.

Step 3: Detection and treating of missing data

```
#find if any feature has null value
df.isnull().sum()
```

```
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
```

name and **host_name** column have 16 and 21 null values respectively.
last_review and **reviews_per_month** each has 10,052 observations as null.

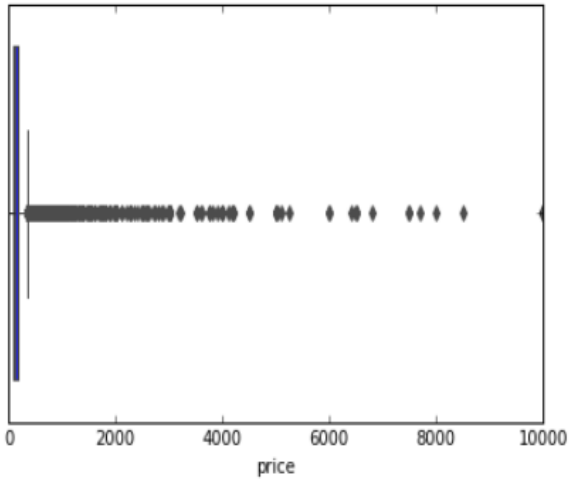
•In this case we observe that "id","name " and "host_name" are redundant for us as we will be referring to listings based on unique **host_id**.So we will be dropping "name " and "host_name" features.

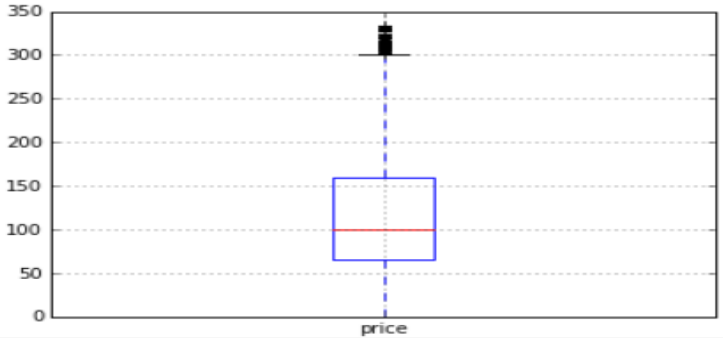
•"last_review" feature depicts the date on which last review was given for the listing,it is irrelevant here.So we will be getting rid of this feature.

Step 4: Detection and treating of atypical data

```
[ ]  #using boxplot to visualize outliers
     sns.boxplot(df["price"])
```

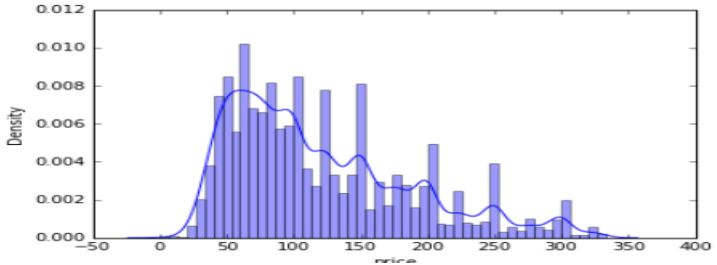
```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:
  FutureWarning
<matplotlib.axes._subplots.AxesSubplot at 0x7f888c8db490>
```



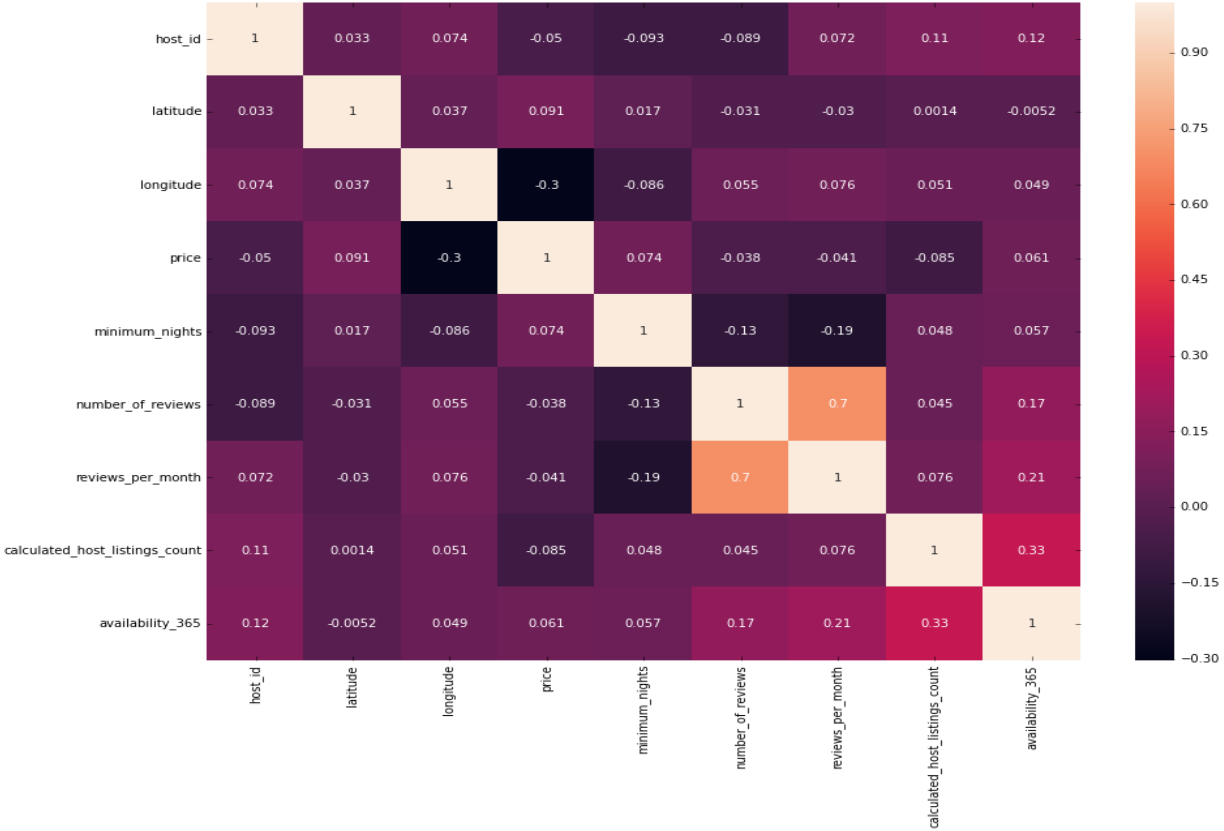



```
#distribution plot after handling outliers
sns.distplot(df_price["price"])
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributi
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f888d18a0d0>
```

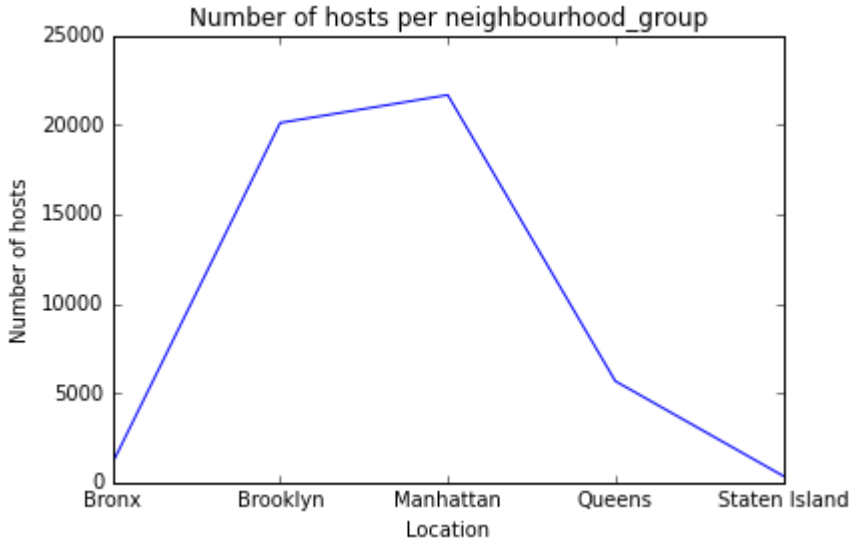


Step 5: Correlation of variables



- Features 'reviews_per_month' and 'number_of_reviews' have a positive correlation with a value of 0.7.So they almost give the same information.
- Features 'reviews_per_month' and 'minimum_nights' have a negative correlation with a value of -0.19.

1.What can we learn about different hosts and areas?

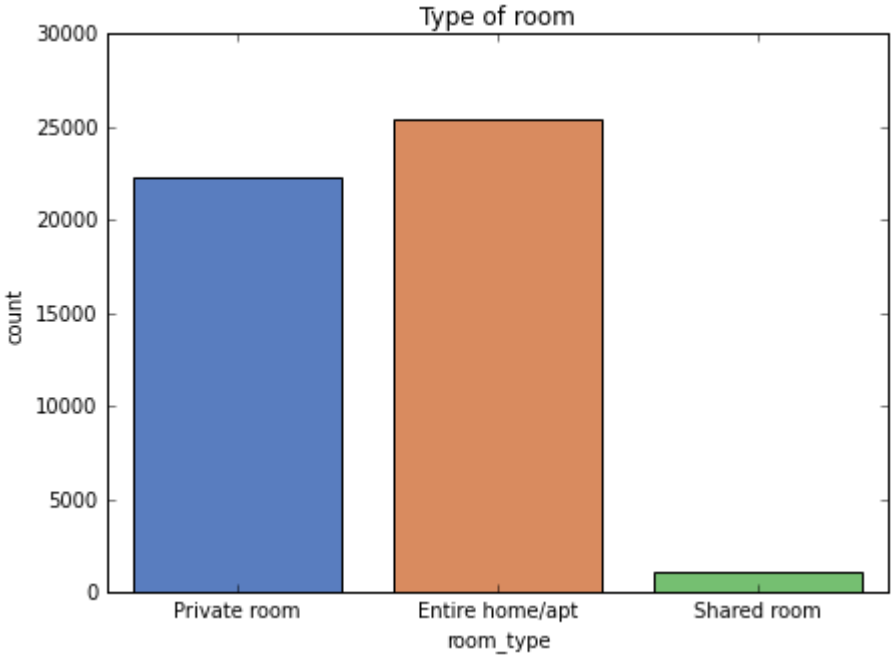


Observation

Most of the hosts are located in Manhattan.i.e.,about 21661 hosts.

Least number of hosts are in Staten island i.e., about 373 hosts.

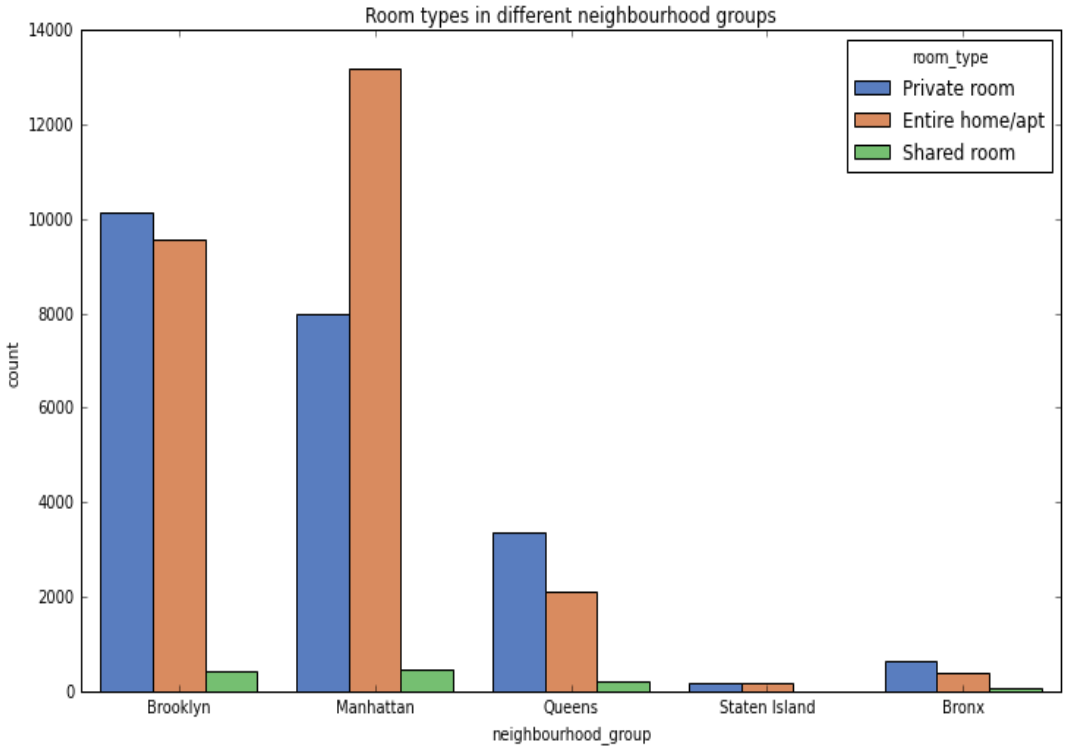
2. What can we learn from predictions?- Type of room



- **Observation** :
- On Airbnb 3 different types of rooms are available for booking.

- They are **Private room,Entire home/apartment and Shared rooms**

2. What can we learn from predictions?- Type of room



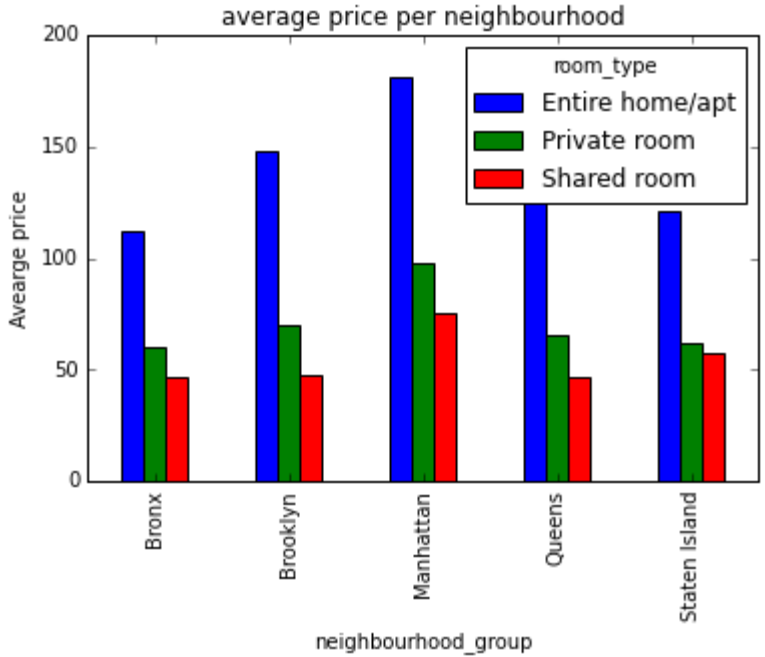
Observation

Most people opt for Entire home/apartment type of listing.

Shared rooms are the least sought out option on Airbnb.

Manhattan has most sought out option as Entire home or apartment,contrary to this in Brooklyn most sought option is private rooms.

2.1. What can we learn from predictions?- Price



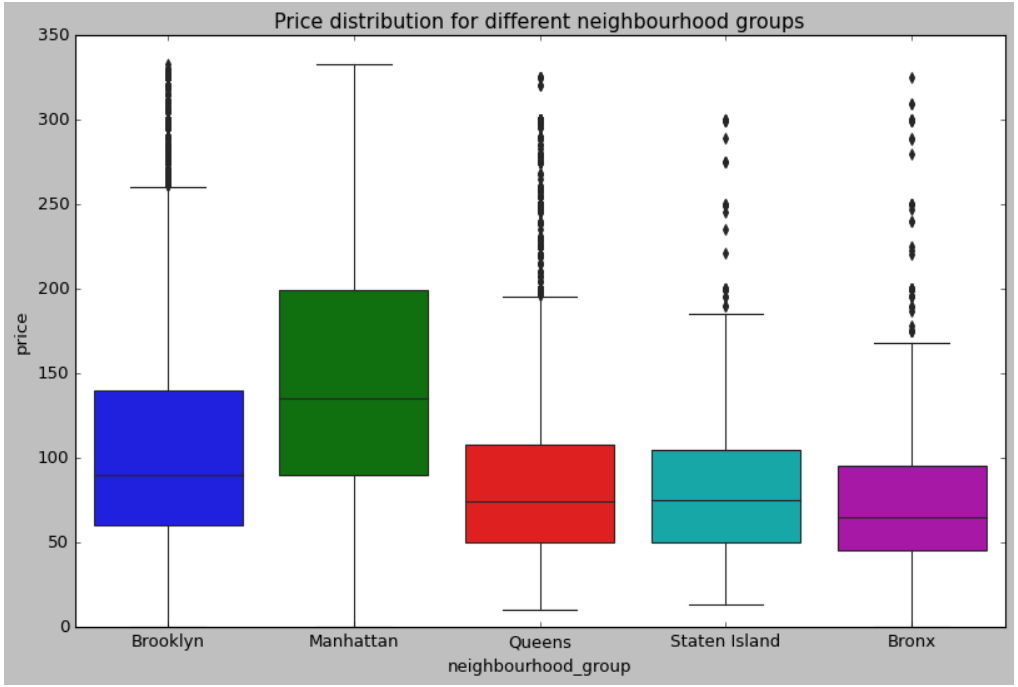
Observations:

Average price is highest for Entire home or apartment in Manhattan.

Among all 5 neighbourhood_groups, highest price is for Entire home or apartment.

Among all 5 neighbourhood_groups , lowest price is for Shared rooms.

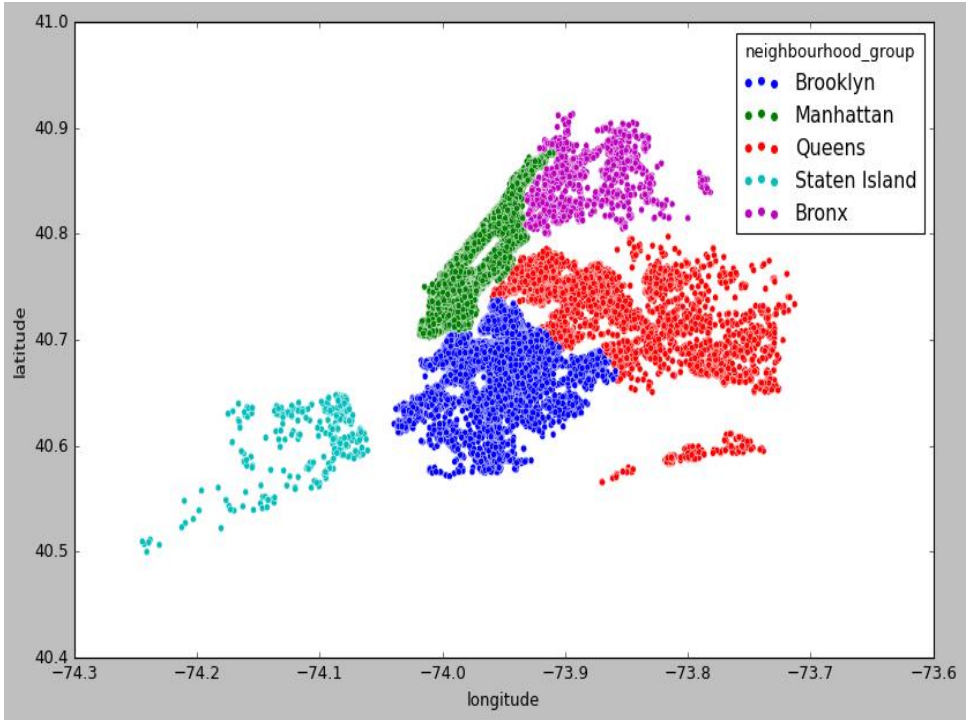
2.2. What can we learn from predictions?- Price



Observations:

- Using boxplot price distribution in all 5 neighborhood groups is observed.
- Range of price is highest in Manhattan.

2.3. What can we learn from predictions?- Location

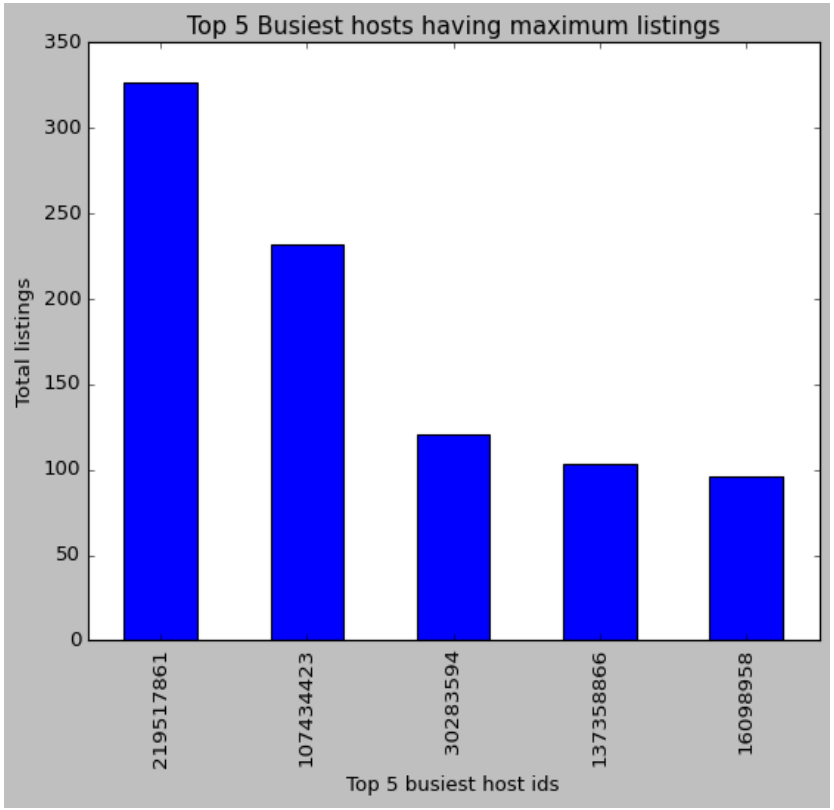


Observations:

- From the location scatterplot we can see that area occupied by Airbnb in Queens is highest and Manhattan is lowest. But still maximum of hosts are located in Manhattan.

- Using scatterplot for latitude and longitude we can map how the listings are located.

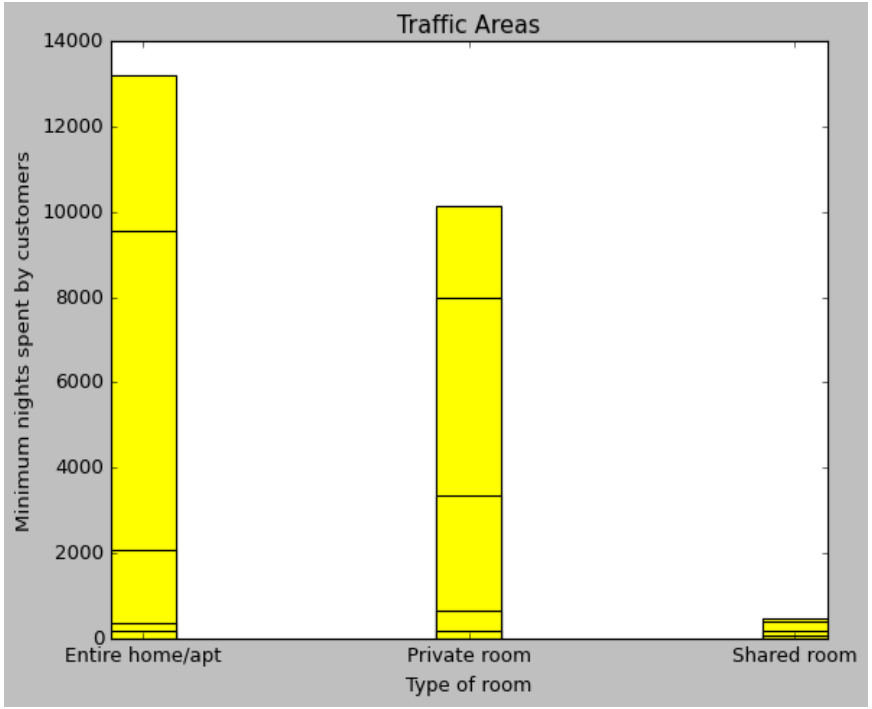
3. Which hosts are the busiest and why?



Observation:

- **host_id 219517861** is the busiest host with total of **327 listings** in Manhattan.

4. Is there any noticeable difference of traffic among different areas and what could be the reason for it?

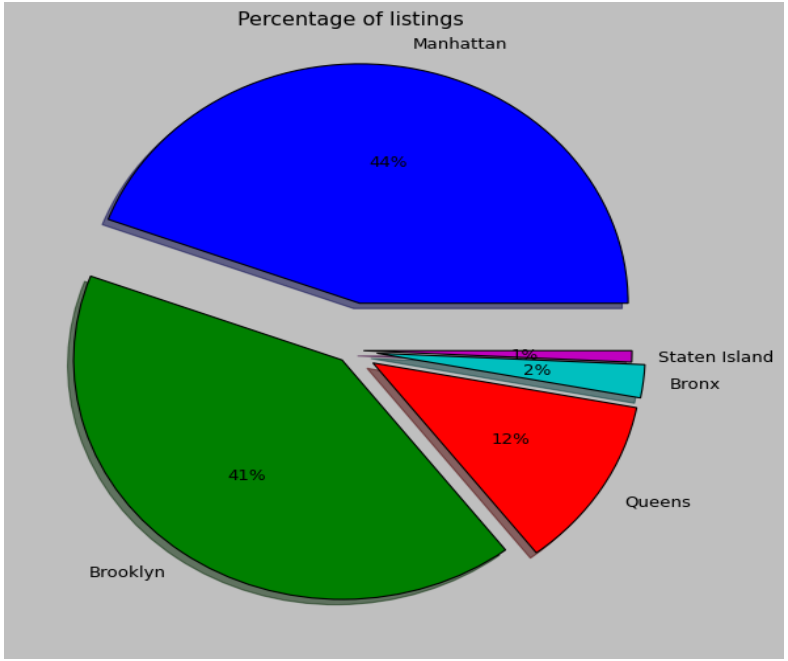


Observation:

- Traffic is mainly in Manhattan and Brooklyn.As Bronx and staten island are away from city center we see less traffic over there.

- People who are staying in Apartment or entire home are staying for longer duration compared to people staying in private room and shared rooms.

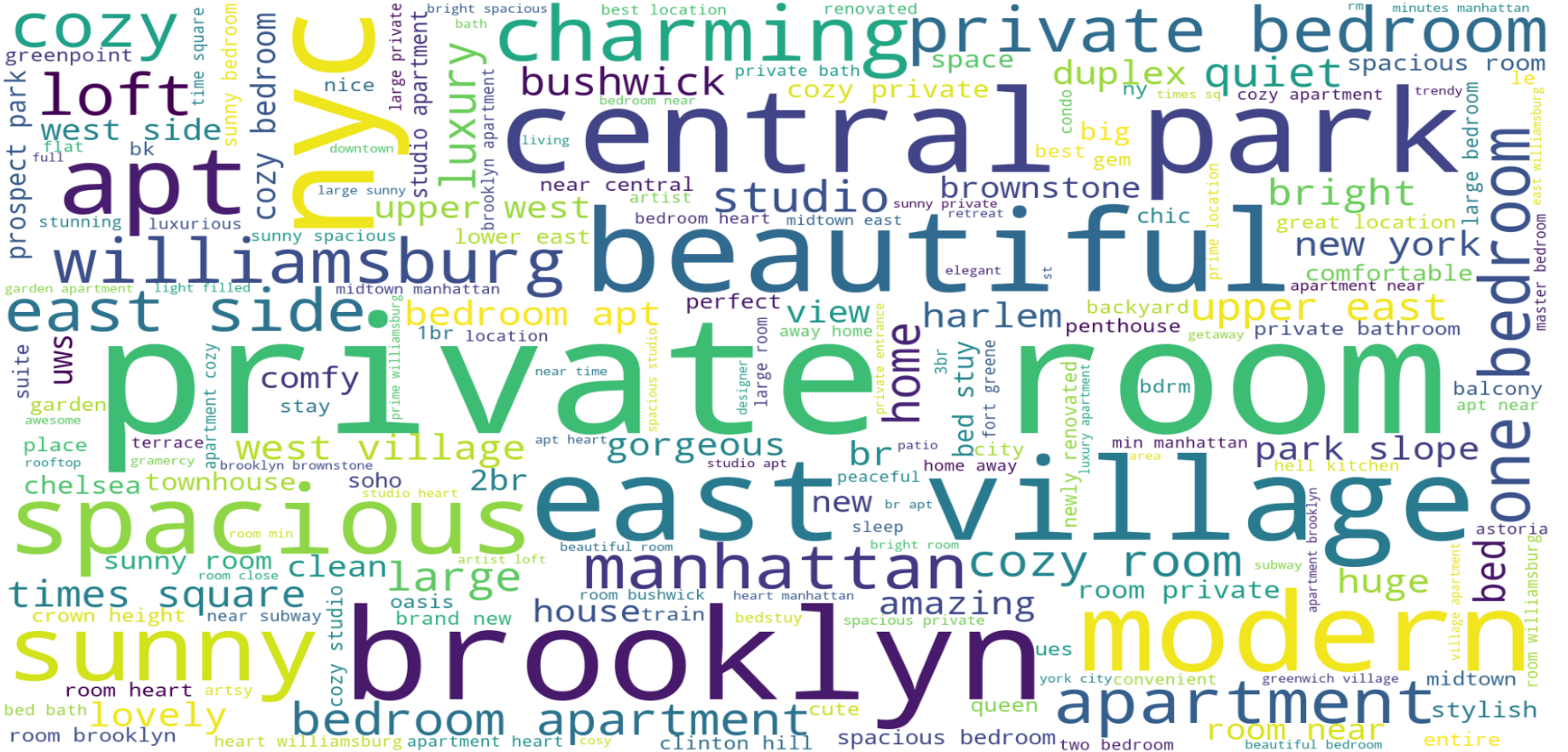
5. What is the percentage of listings owned by Airbnb in different neighbourhoods?



Observation:

- Percentage of listings in Manhattan is 44% and then followed by Brooklyn 41%
- Percentage of listings in in Staten island(1%).

Listing names in each Neighborhood



Conclusion

Observation:

- Most of the hosts are located in Manhattan.i.e.,about 21661 hosts.
- Least number of hosts are in Staten island i.e., about 373 hosts. Average price is highest for Entire home or apartment in Manhattan.
- Among all 5 neighbourhood_groups ,highest price is for Entire home or apartment.
- Among all 5 neighbourhood_groups , lowest price is for Shared rooms. On Airbnb 3 different types of rooms are available for booking.They are Private room,Entire home/apartment and Shared rooms

Conclusion

Observation:

- Traffic is mainly in Manhattan and Brooklyn. As Bronx and Staten island are away from city center we see less traffic over there.
- People are preferring mainly Entire home or Private room than shared rooms.It is due to privacy preference. People are ready to pay more for this.

- From the location scatterplot we can see that area occupied by Airbnb in Queens is highest and Manhattan is lowest.But still maximum of hosts are located in Manhattan.
- People who are staying in Apartment or entire home are staying for longer duration compared to people staying in private room and shared rooms.

Thank You