Capstone Project
Title: Airbnb Booking analysis

Presented by: Surabhi Mali
Ruchi Yadav

Index

- Problem Statement
- Data Summary
- Data Pipeline
- Data Cleaning
- Exploratory Data Analysis(EDA)
- Conclusion

Problem Statement

Since 2008,guest and host have used Airbnb to expand on travelling possibilities and present a more unique, personalised way of experiencing the world. Today, Airbnb become one of kind service that is used and recognised by the whole world. Data analysis on millions of listings provided through Airbnb is crucial factor for the company. These millions of listings generate a lot of data that can be analysed and used for security, business decision, understanding of customers and providers(hosts) behaviour 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 between categorical and numerical values.

How to explore and analyze the data to discover key understandings so that Airbnb buisness can expand?

Data Summary

This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

Host id: It is id given to specific host and there are in given dataset

Neighbourhood group: It represent location in given dataset.
There are 5 different neighbourhood groups :- 'Brooklyn' , 'Manhattan' , 'Queens' , 'Staten Island' and 'Bronx'

Neighbourhood: It represent specific areas where the listings are located in 5 different neighbourhood groups.

Room type: It represent category of room type being listed as:-
'Private room' , 'Entire home/apt' , 'Shared room'.

Minimum nights: It represents number of nights spend by customer in given listing.

Number of reviews: It represents the number of reviews for listings.

Availability 365: It represents number of days in year for which given property is available for rent.

Price: It represent rate for given room type in given location for one night.

Data Pipeline:

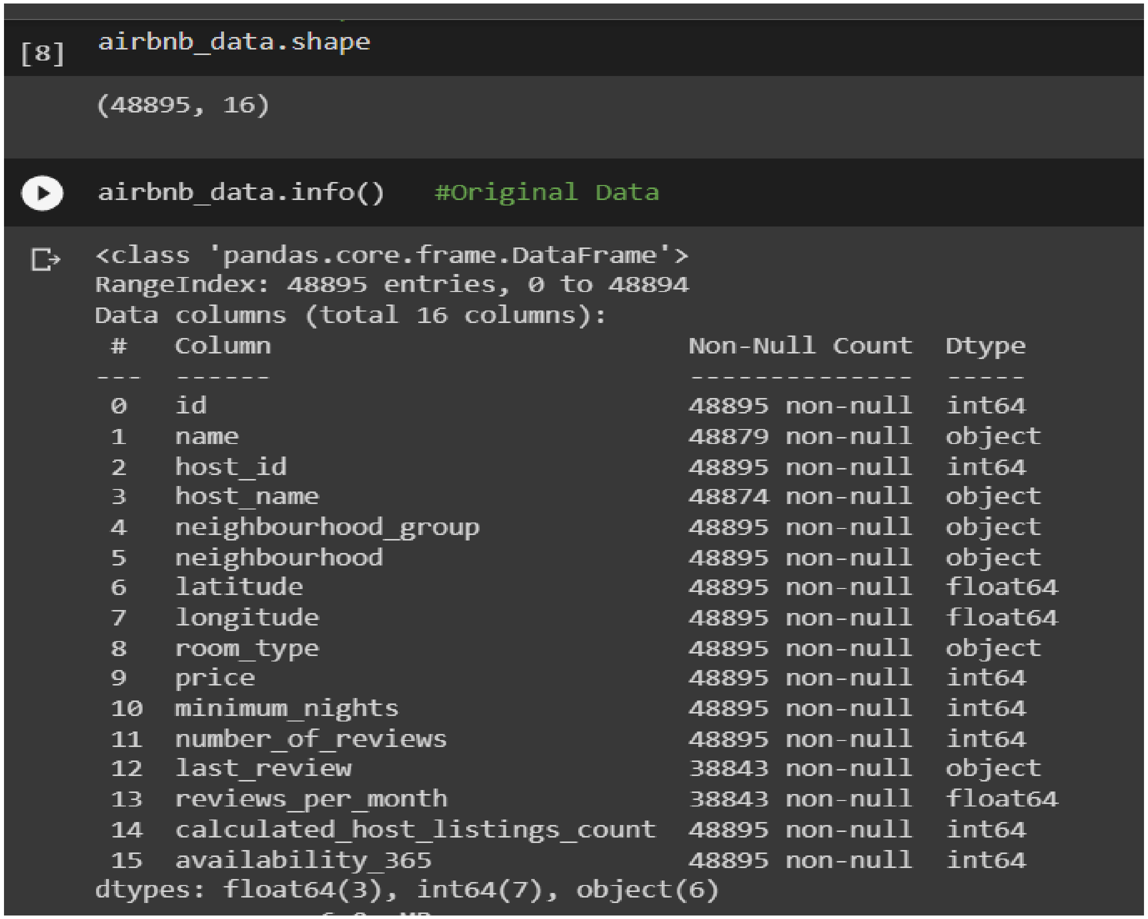
Data Pre-processing: In this part we have checked the data and its all features to get better understanding of the data

Data Cleaning: In this part we have checked for Nan values, missing values and duplicate observations and done the cleaning of dataset.

Exploratory Data Analysis(EDA): In this part we have done some exploratory data analysis on selected important features to get the insights of the dataset

Data Visualization: To visualize the data, we used different plots for distinct features of the data and tred to analyse the relationship between the features of data.

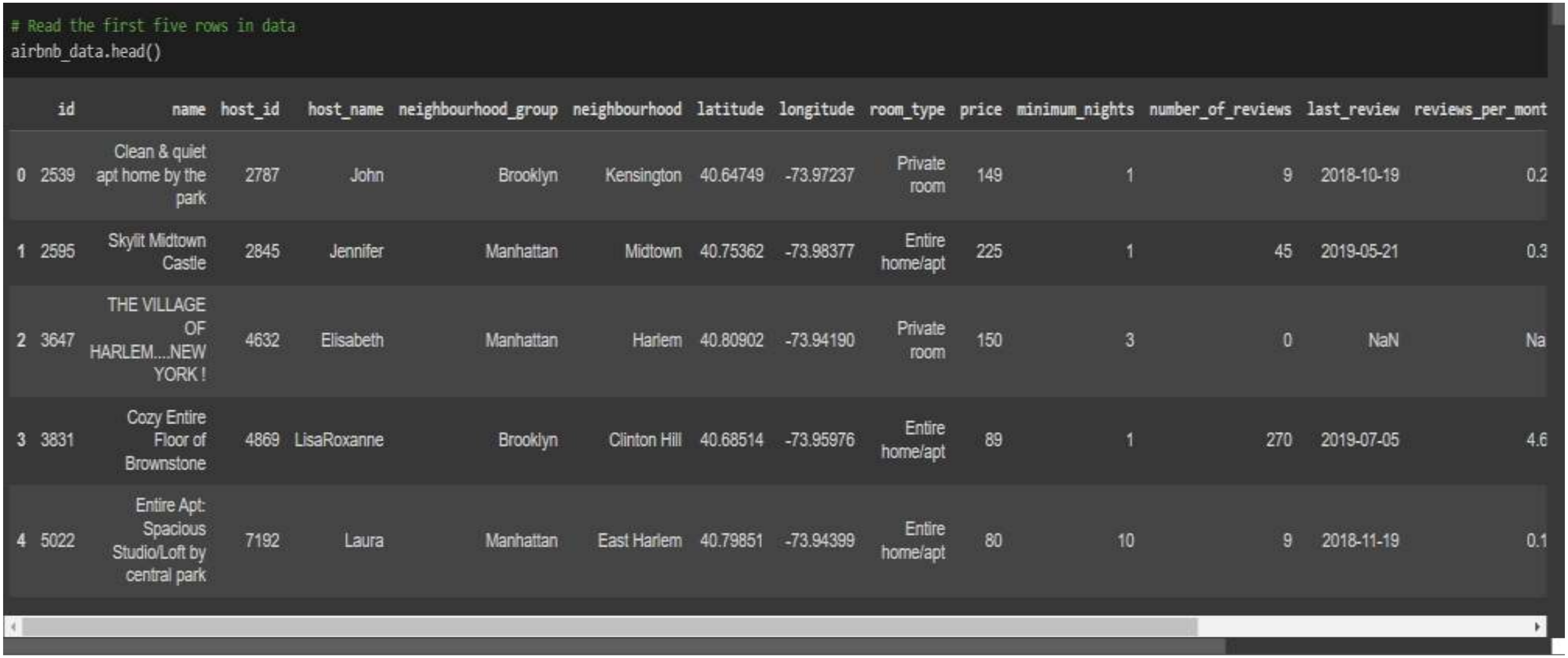
Data Pre-processing:-



```
[8]  airbnb_data.shape

     (48895, 16)

     airbnb_data.info()   #Original Data

     <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)
```

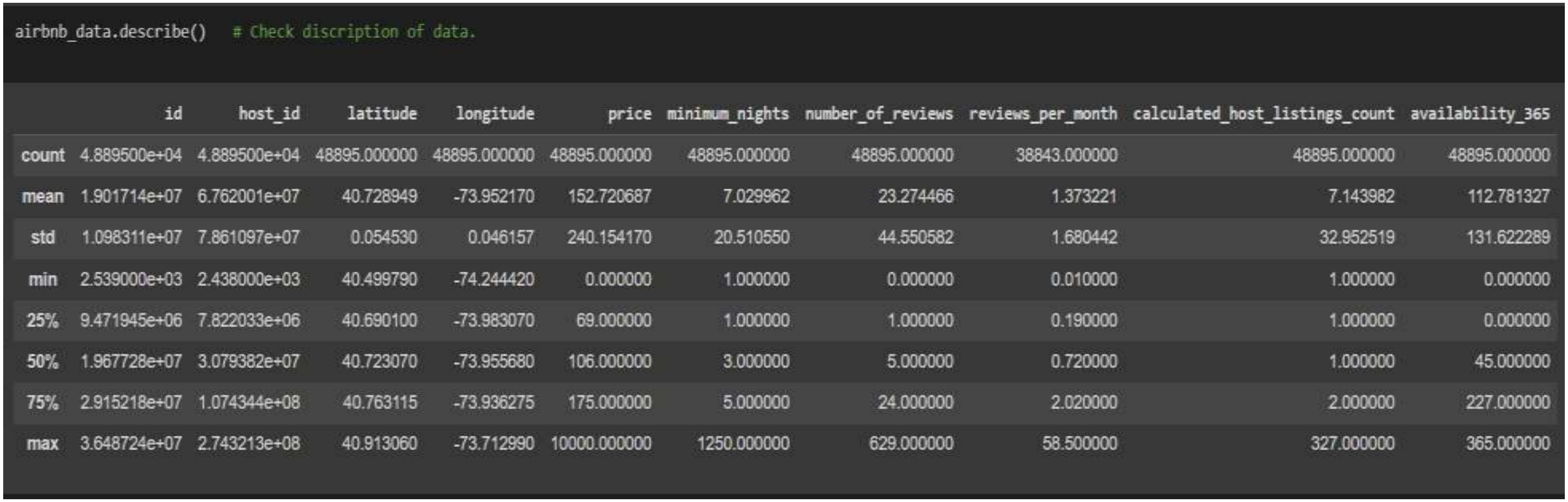
Observations:
Dataset contains around 49000 rows and 16 columns.

Host_name,last_review and reviews_per_month have some missing values.



```
# Read the first five rows in data
airbnb_data.head()
```

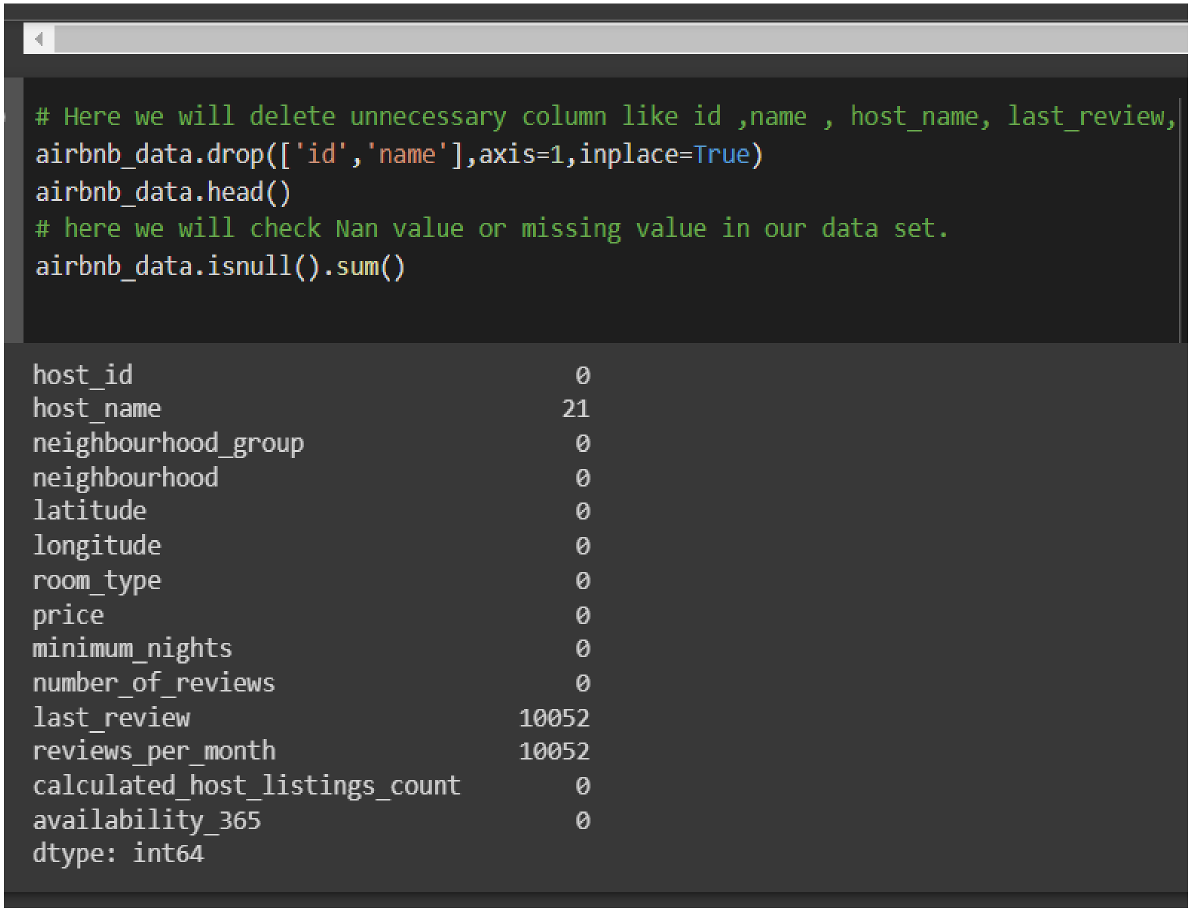
	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_mont
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.2
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.3
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	Na
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.6
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.1



```
airbnb_data.describe()   # Check discription of data.
```

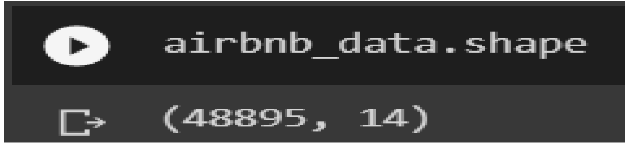
	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

Data Cleaning (Missing, Nan Values)



```
# Here we will delete unnecessary column like id ,name , host_name, last_review,
airbnb_data.drop(['id','name'],axis=1,inplace=True)
airbnb_data.head()
# here we will check Nan value or missing value in our data set.
airbnb_data.isnull().sum()
```

```
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
```

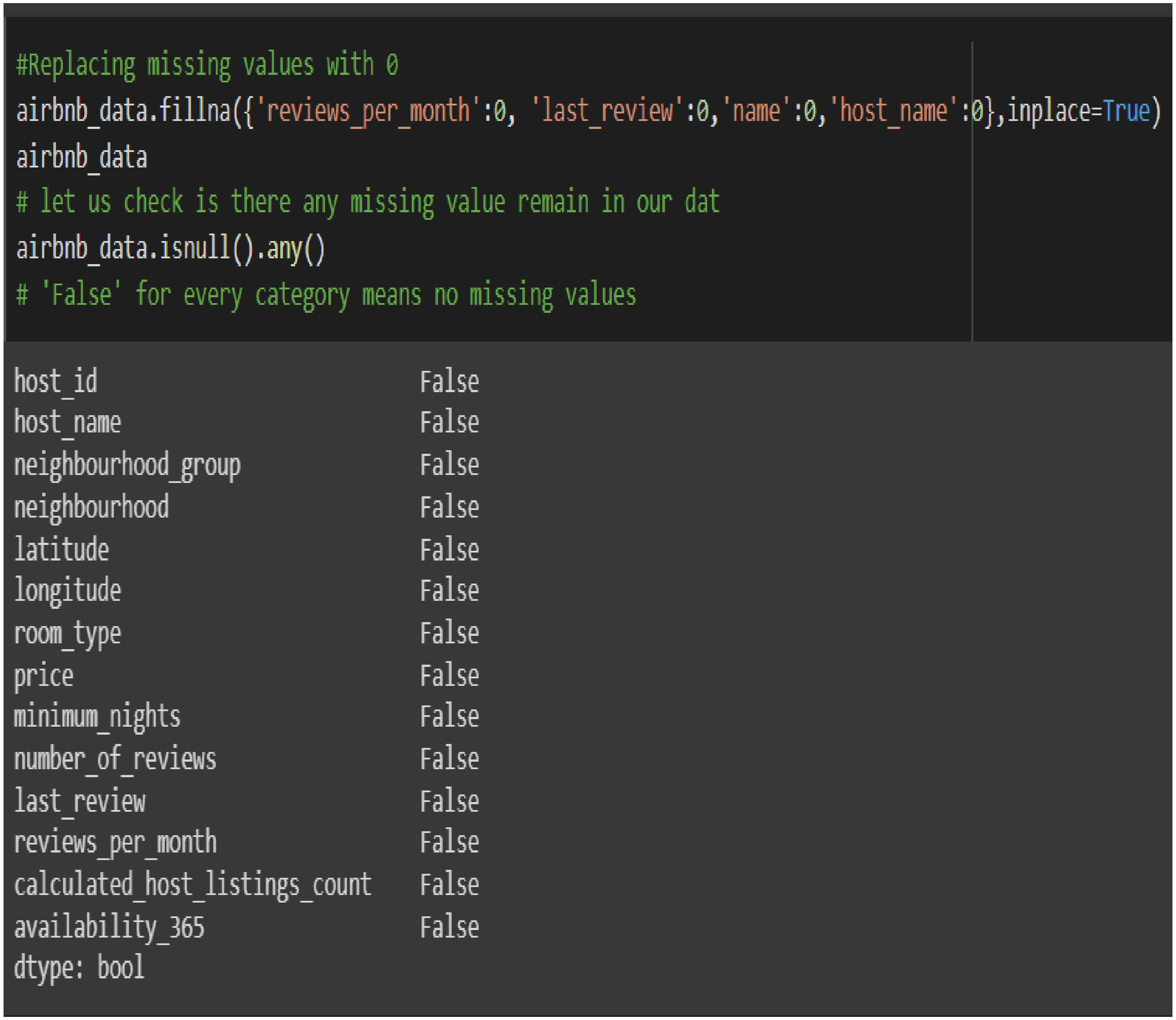


```
airbnb_data.shape
```

```
(48895, 14)
```

Observations:
Here we dropped unnecessary columns called ‘id’ and ‘name’ to make clean dataset.
Now we have 48895 rows and 14 columns in the new dataset

Host_name,last_review and reviews_per_month have some missing values,we will replace them with ‘0’ to get no inconsistencies



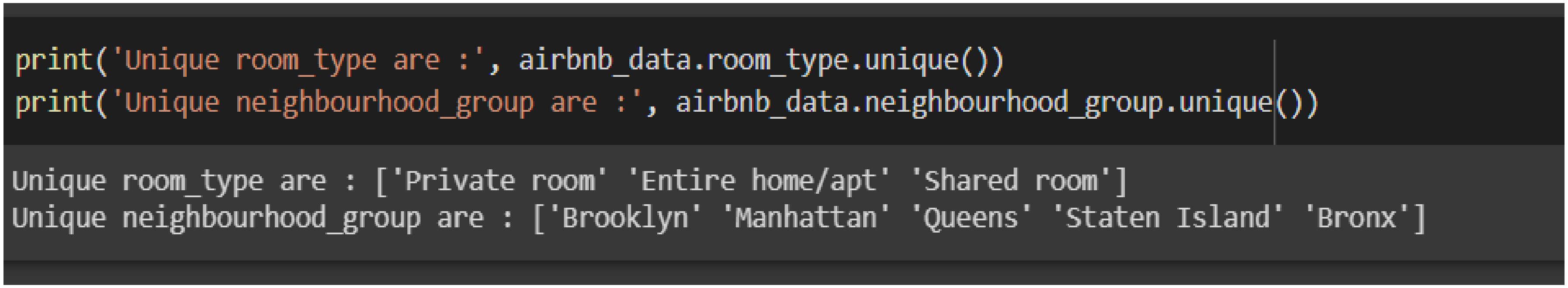
```
#Replacing missing values with 0
airbnb_data.fillna({'reviews_per_month':0, 'last_review':0,'name':0,'host_name':0},inplace=True)
airbnb_data
# let us check is there any missing value remain in our dat
airbnb_data.isnull().any()
# 'False' for every category means no missing values
```

```
host_id                           False
host_name                         False
neighbourhood_group               False
neighbourhood                     False
latitude                          False
longitude                         False
room_type                         False
price                             False
minimum_nights                    False
number_of_reviews                 False
last_review                       False
reviews_per_month                 False
calculated_host_listings_count    False
availability_365                  False
dtype: bool
```

Observations:
Missing values will be replaced with ‘0’ with the help of “fillna” function

We got all false values after checking “isnull().any()” that means now there are no missing values in data.

EDA(Exploring Neighbourhood-Groups and Room-types)



```
print('Unique room_type are :', airbnb_data.room_type.unique())
print('Unique neighbourhood_group are :', airbnb_data.neighbourhood_group.unique())
```

```
Unique room_type are : ['Private room' 'Entire home/apt' 'Shared room']
Unique neighbourhood_group are : ['Brooklyn' 'Manhattan' 'Queens' 'Staten Island' 'Bronx']
```

Observations:

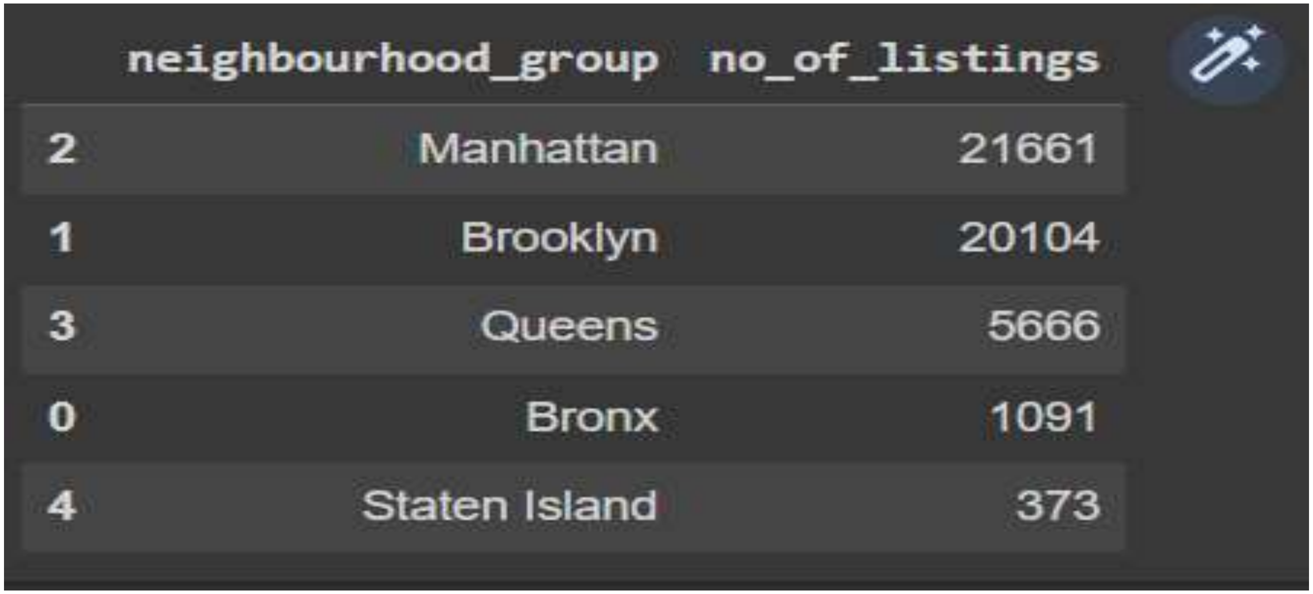
- There are 3 types of rooms: 1.Private room 2.Entire home/apt 3.Shared room

Majority of entire home/apartment are located in Manhattan and majority of private rooms are located in brooklyn

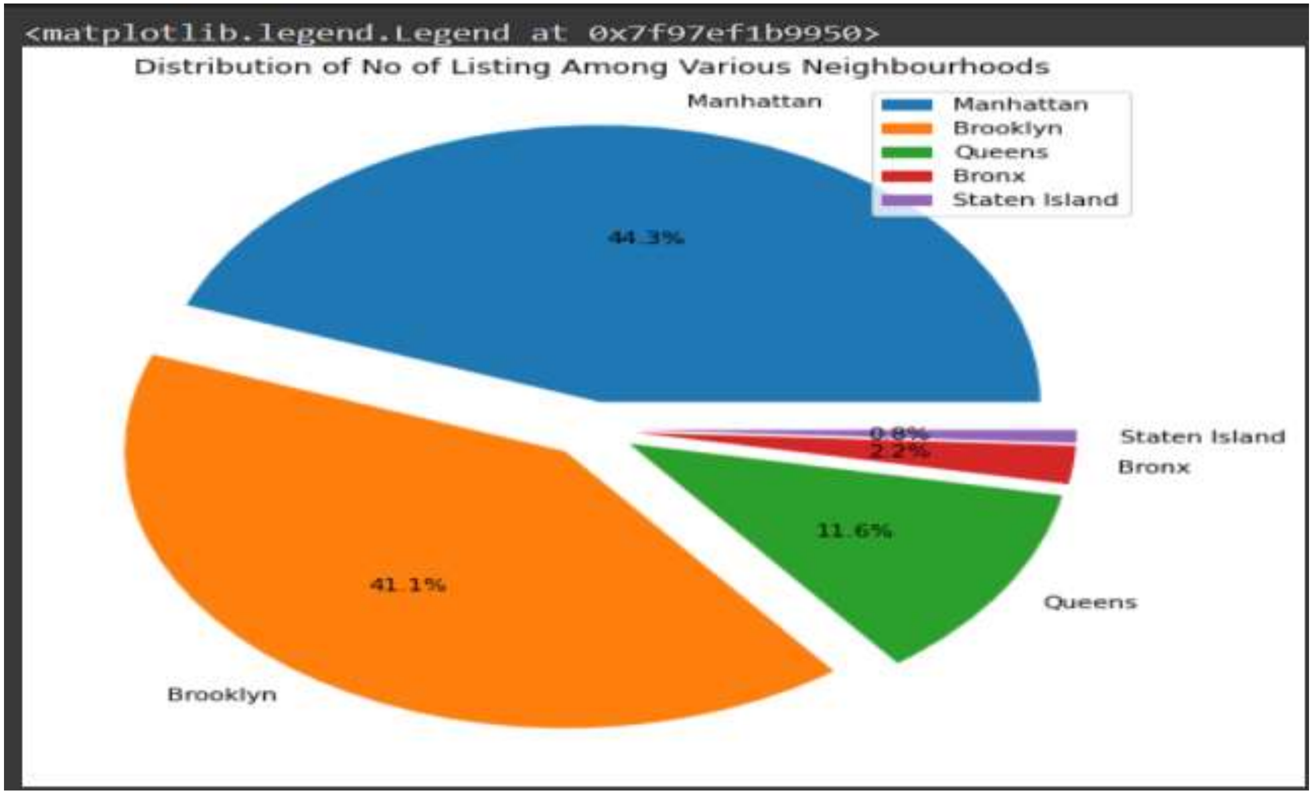
- There are 5 Neighbourhood-Groups called

1.Manhattan 2.Brooklyn 3.Queens 4.Bronx 5.Staten Island

Which neighbourhood group has most number of grouping?



	neighbourhood_group	no_of_listings
2	Manhattan	21661
1	Brooklyn	20104
3	Queens	5666
0	Bronx	1091
4	Staten Island	373

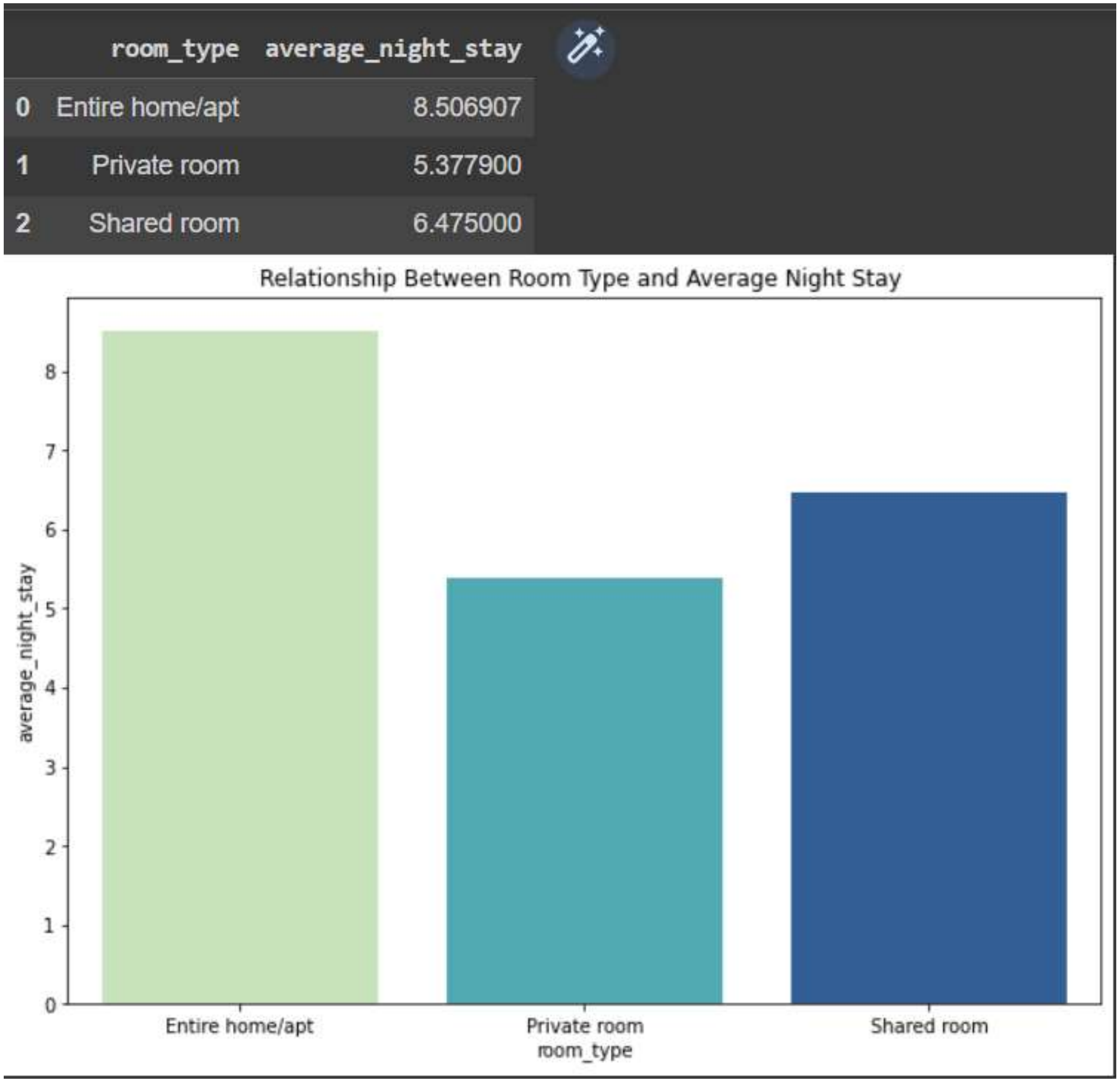


```
<matplotlib.legend.Legend at 0x7f97ef1b9950>
```

Observations:

we can conclude that we have highest number of listing i.e. 44.3%

How many minimum_nights people stayed in each room_type?

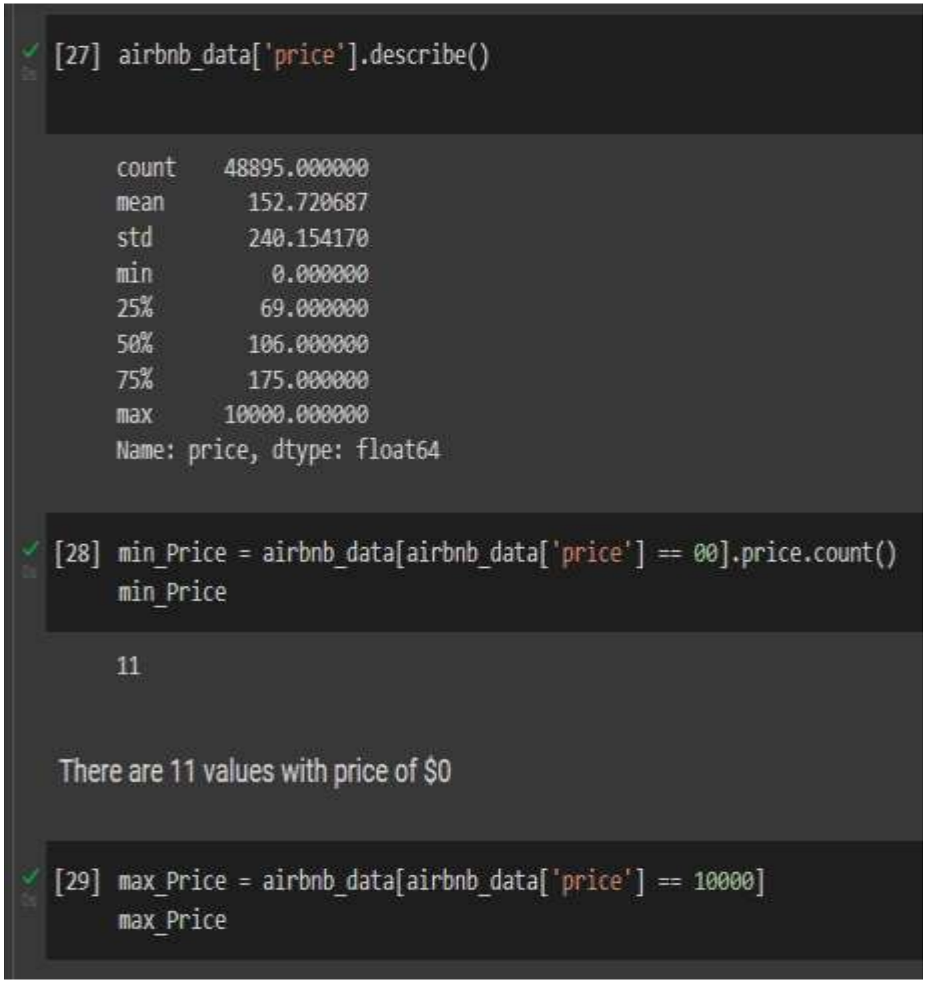


	room_type	average_night_stay
0	Entire home/apt	8.506907
1	Private room	5.377900
2	Shared room	6.475000

Observations:

If someone is booking Entire home/apt , They tend to stay for longer duration on an average 8 to 9 days. For private room on an average of 5 to 6 days. For shared room on an average of 6 to 7 days.

EDA(Price Exploration)



```
[27] airbnb_data['price'].describe()

     count    48895.000000
     mean       152.720687
     std        240.154170
     min          0.000000
     25%         69.000000
     50%        106.000000
     75%        175.000000
     max      10000.000000
     Name: price, dtype: float64

[28] min_Price = airbnb_data[airbnb_data['price'] == 00].price.count()
     min_Price

     11
```

There are 11 values with price of $0

```
[29] max_Price = airbnb_data[airbnb_data['price'] == 10000]
     max_Price
```

Observations:

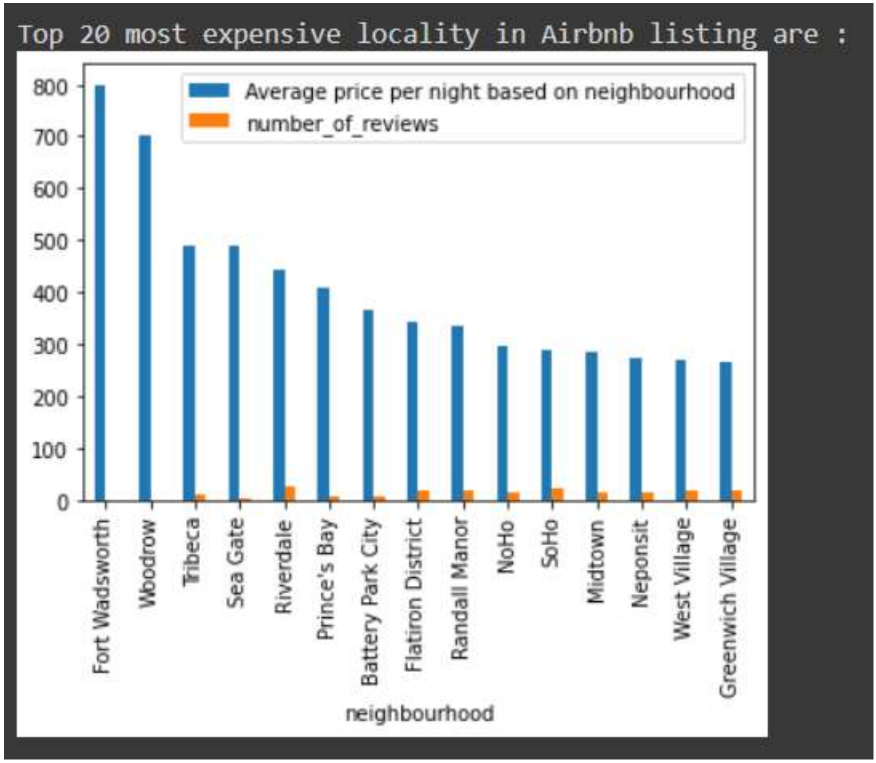
According to the statistics it is clear that the 75% of the listing’s Price ranges from 0− 200. But there are also 3 Airbnb with maximum price of $10000.

There are 11 values with price of $0, which can be due to dynamic pricing or the willingness of not to share the price with the Airbnb.

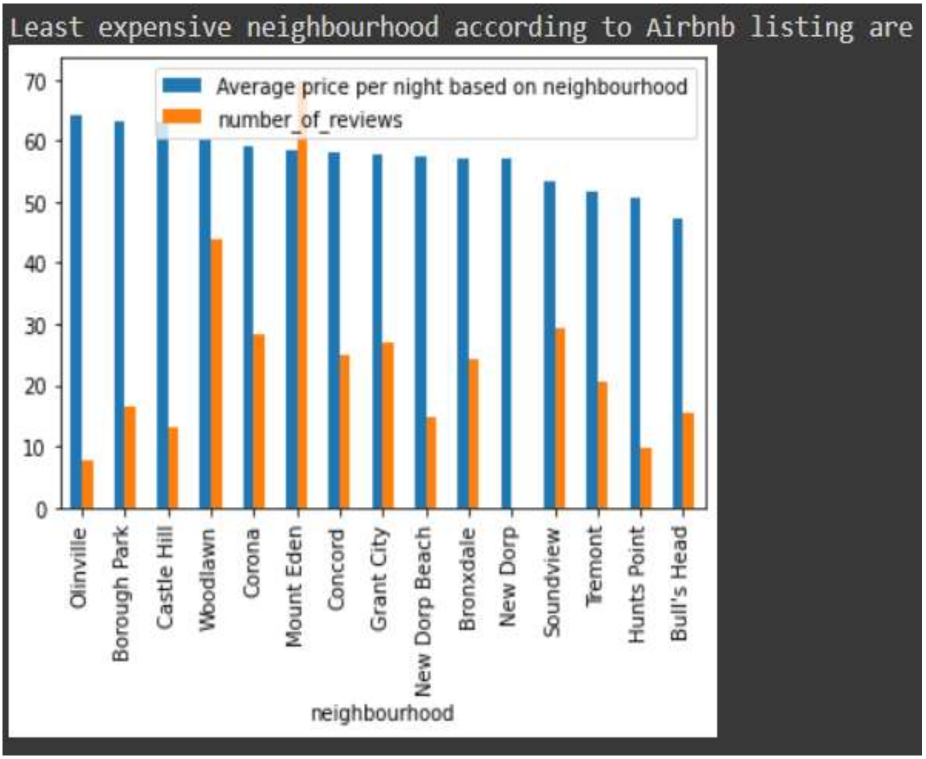
Which neighbourhood has the highest and lowest price?

We have plotted the most expensive and least expensive neighbourhood, and we will plot only Top 20 neighbourhood and Bottom 20 with respect to average price. This will help a traveller to choose the appropriate neighbourhood based on his budget and the number of reviews.

So according to the below plots Fort Wadsworth is the most expensive in terms of neighbourhood with 0 number of reviews on an avarage. Whereas Bull's head locality is the least expensive to stay with 15 number of reviews on an avarage.



```
Top 20 most expensive locality in Airbnb listing are :
```

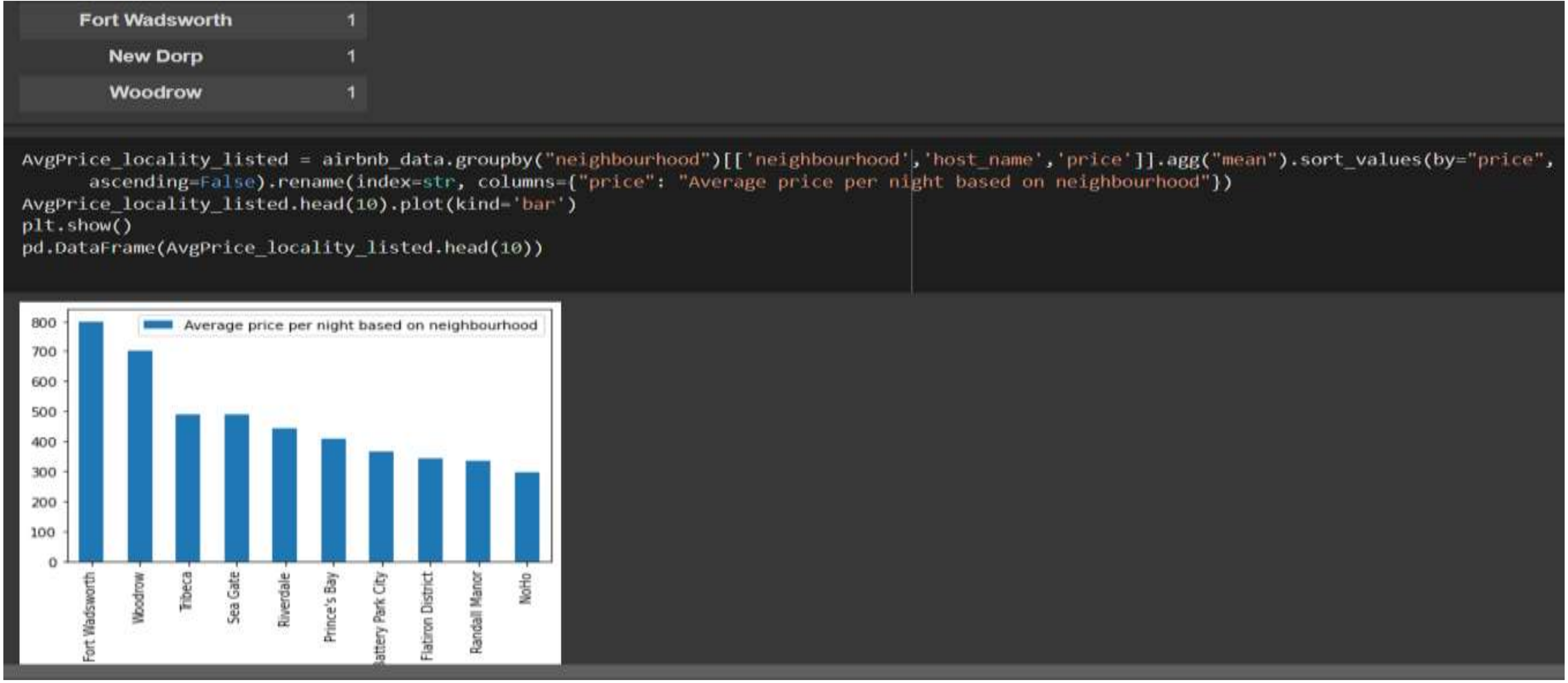


```
Least expensive neighbourhood according to Airbnb listing are
```

How many number of locality listed and how price is getting affected by listings?

Williamsburg has most number of listing count which is '3920' whereas Fort Wadesworth,Rossville,Richmondtown,Willowbrook,Fort Wadsworth,New Dorp,Woodrow has one of the least listing which is '1'.

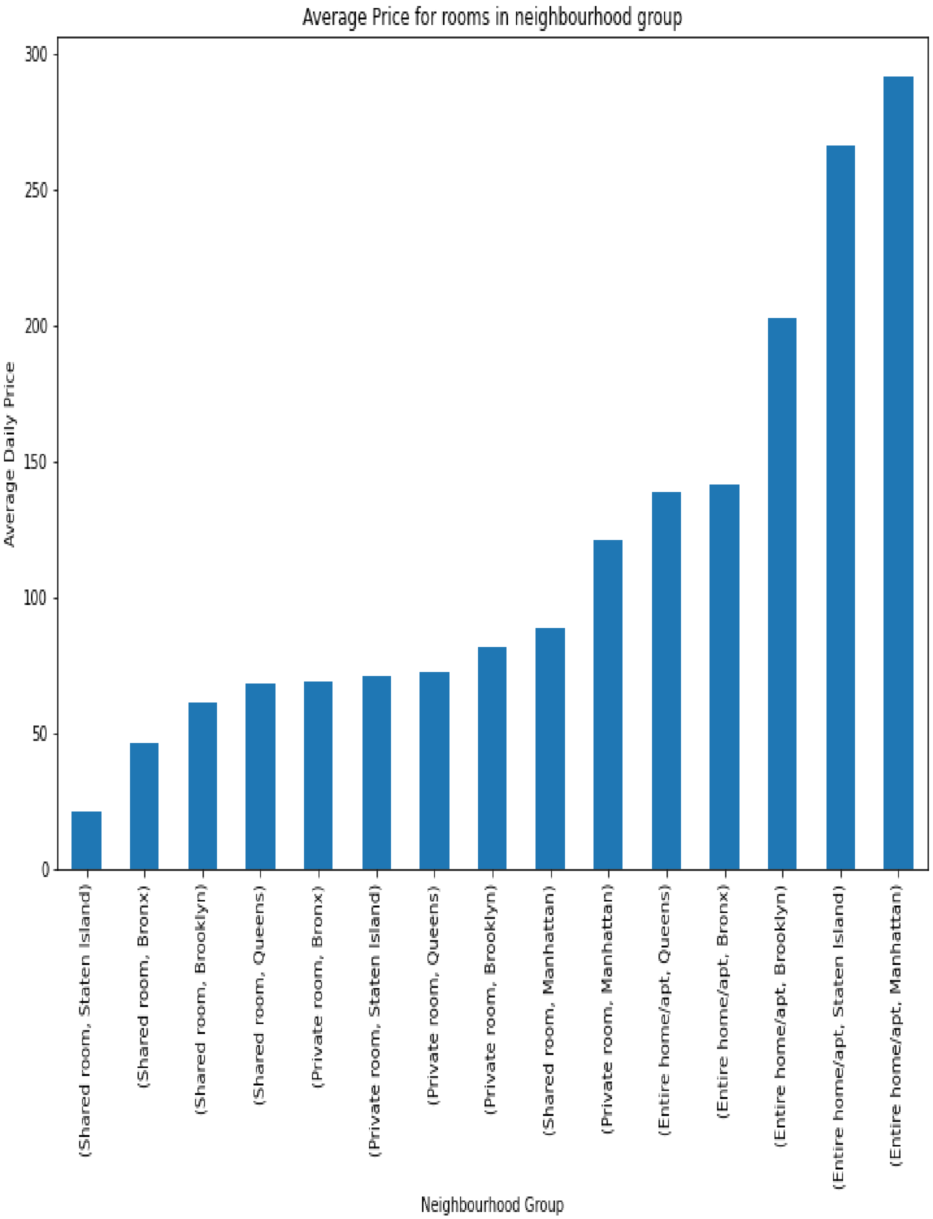
We can see this neighbourhood(Fort Wadsworth ,Woodrow) are one of the highest stay based on price the reason that the price is high in this neighborhood is due to the less number of listings.



Fort Wadsworth	1
New Dorp	1
Woodrow	1

```
AvgPrice_locality_listed = airbnb_data.groupby("neighbourhood")[['neighbourhood','host_name','price']].agg("mean").sort_values(by="price",
      ascending=False).rename(index=str, columns={"price": "Average price per night based on neighbourhood"})
AvgPrice_locality_listed.head(10).plot(kind='bar')
plt.show()
pd.DataFrame(AvgPrice_locality_listed.head(10))
```

How much will you spend on an average for room?



By looking at the plots, we can say that :

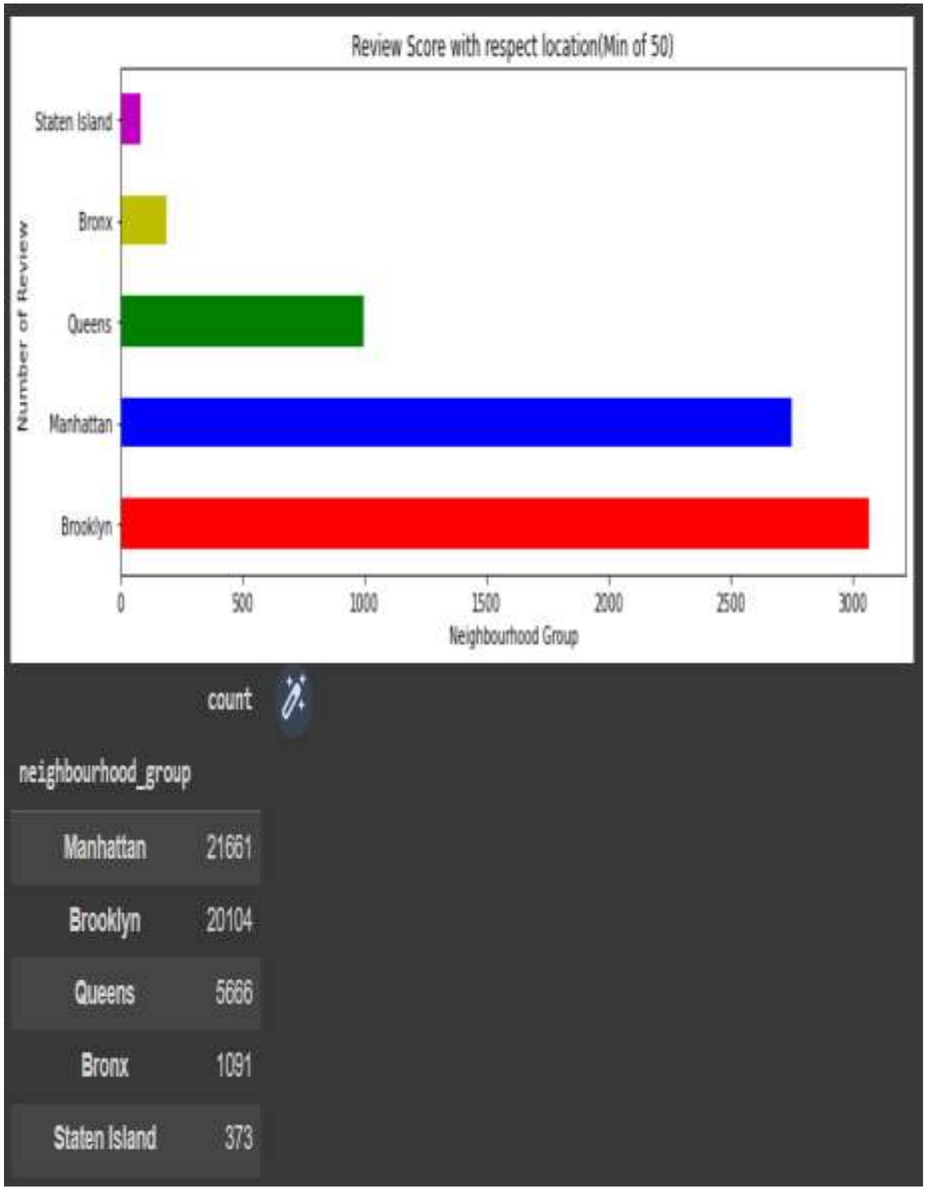
a. Shared room at staten Island is the most cheapest stay per night whereas Renting a Entire apartment/Home at Manhattan per night is the most expensive.

b. Average price for Private room is also considerably expensive at manhattan so is the shared room at Manhattan is expensive than other private rooms of the neighbourhood. This clearly states that Manhattan is the expensive stay than any other locality.

c. Bronx is the most cheapest stay in terms of neighbourhood group comparison in respect to room type.

d. Though Shared room at Staten Island is the cheapest whereas Apartment renting is not cheapest at Staten Island. This can be due to the location of a perfect gateway from the rush of the city for a quality time with family get together.

Which Neighborhood Group have good number of reviews?



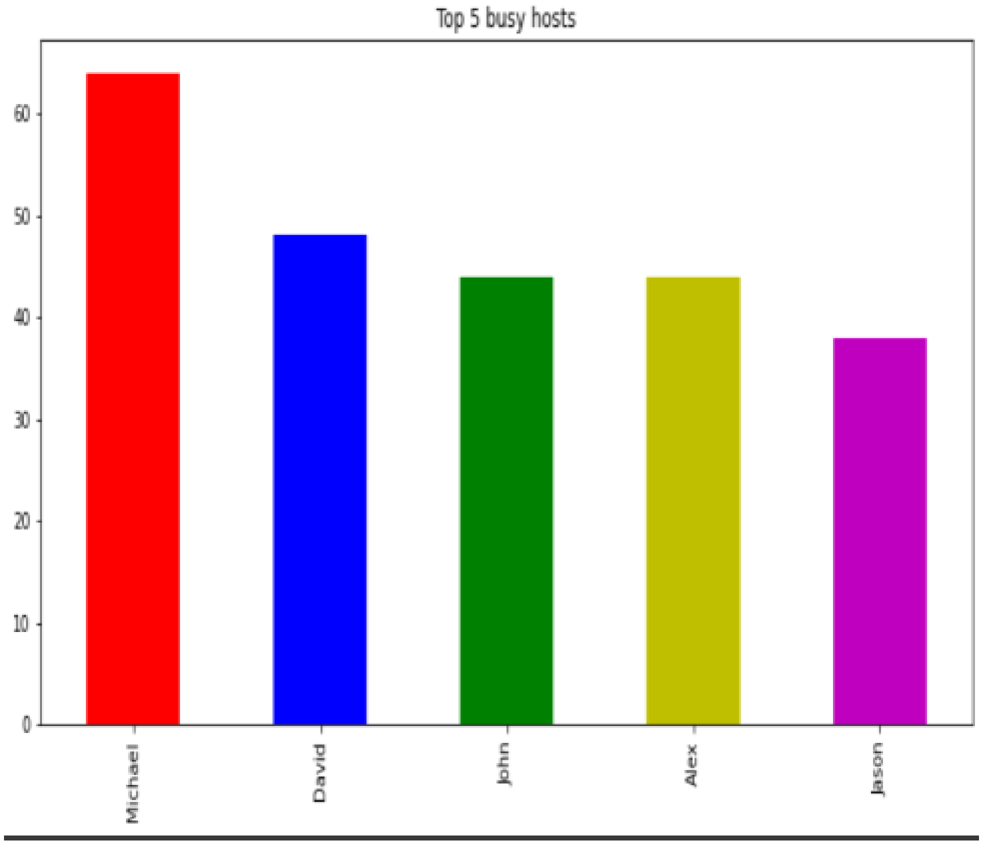
	count
neighbourhood_group	
Manhattan	21661
Brooklyn	20104
Queens	5666
Bronx	1091
Staten Island	373

- We have considered reviews more than 50, so that we can have an good insight of the data.

- So we can see from the plot, Brooklyn got most review around 3000 in comparison to Manhattan around 2700, even if the listings for Manhatten is more which is 21661 than Brooklyn is 20104. Also Staten Island which is cheaper has less review than the other neighbourhood group also less number of listings.

- From this we can say that if we have more listings in neighbourhood_groups, the tourists can have more options to try which will leads to more number of reviws after visiting the place.

Which 5 host has most number of reviews?

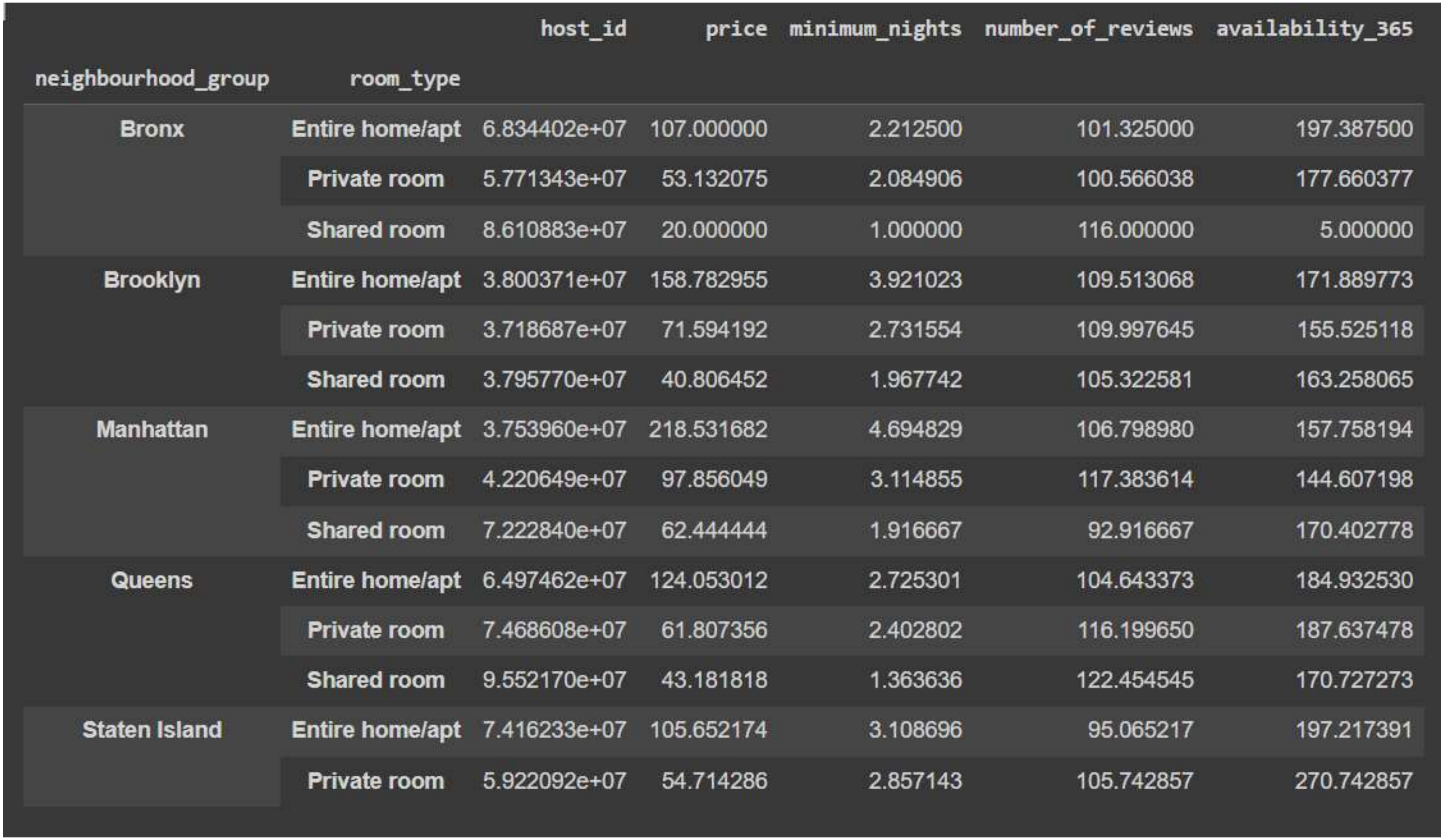


Based on the review score(Minimum 50) we will see who is our top 5 Host, this increases the confidence of tourist before booking.

We have total 48874 hosts out of them 7075 hosts have review count greater than 50.Now we will get Top 5 host who have most number of reviews.

We can say that Michael has received most number of reviews after David.

This is like a final conclusion for tourists to compare the average Listing of Airbnb which has more reviews for each Neighbourhood group in each category of room types.



		host_id	price	minimum_nights	number_of_reviews	availability_365
neighbourhood_group	room_type					
Bronx	Entire home/apt	6.834402e+07	107.000000	2.212500	101.325000	197.387500
	Private room	5.771343e+07	53.132075	2.084906	100.566038	177.660377
	Shared room	8.610883e+07	20.000000	1.000000	116.000000	5.000000
Brooklyn	Entire home/apt	3.800371e+07	158.782955	3.921023	109.513068	171.889773
	Private room	3.718687e+07	71.594192	2.731554	109.997645	155.525118
	Shared room	3.795770e+07	40.806452	1.967742	105.322581	163.258065
Manhattan	Entire home/apt	3.753960e+07	218.531682	4.694829	106.798980	157.758194
	Private room	4.220649e+07	97.856049	3.114855	117.383614	144.607198
	Shared room	7.222840e+07	62.444444	1.916667	92.916667	170.402778
Queens	Entire home/apt	6.497462e+07	124.053012	2.725301	104.643373	184.932530
	Private room	7.468608e+07	61.807356	2.402802	116.199650	187.637478
	Shared room	9.552170e+07	43.181818	1.363636	122.454545	170.727273
Staten Island	Entire home/apt	7.416233e+07	105.652174	3.108696	95.065217	197.217391
	Private room	5.922092e+07	54.714286	2.857143	105.742857	270.742857

Conclusion:

- The given dataset appear to be very rich dataset with a variety of columns that allowed us to do deep exploration on each significant column presented.
- After cleaning the data we had 48895 rows and 14 columns in the new dataset.
- There are 3 different types of rooms and 5 different Neighbourhood-Groups.
- we can conclude that highest number of listing i.e. 44.3% in Manhatten.
- To conclude we can say,people stay for private room on an average of 5 to 6 days and shared room on an average of 6 to 7 days.
- Statistics states that the 75% of the listing’s Price ranges from 0− 200. But there are also 3 Airbnb with maximum price of $10000.
- Fort Wadsworth is the most expensive in terms of neighbourhood with 0 number of reviews on an avarage. Whereas Bull's head locality is the least expensive to stay with 15 number of reviews on an avarage.

- We can see this neighbourhood(Fort Wadsworth ,Woodrow) are one of the highest stay based on price the reason that the price is high in this neighborhood is due to the less number of listings.
- Manhattan is the expensive stay than any other locality.Bronx is the most cheapest stay in terms of neighbourhood group comparison in respect to room type
- Brooklyn got most review around 3000 in comparison to Manhattan around 2700, even if the listings for Manhatten is morewhich is 21661 than Brooklyn is 20104.
- Staten Island which is cheaper has less review than the other neighbourhood group also less number of listings.
- From this we can say that if we have more listings in neighbourhood_groups, the tourists can have more options to try which will leads to more number of reviws after visiting the place.Airbnb needs to expand the listings in Bronx and Staten Island.
- Michael has received most number of reviews after David.

Future Scope:

- This data analytics will be very useful for higher level on Airbnb Data/Machine Learning team for better business decision, control over the platform, Marketing initiatives, implementation of new feature and much more.

- Airbnb buisness can be more focused on areas that are lacking after getting the useful insights from this exploaratory data analysis

Thank You!!!