Capstone Project

Airbnb Bookings Analysis (EDA)

Submitted by

Bhavesh Amol Amre

Objective of This Project

- To deliver insights from 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.
- The dataset is chosen from ALMA BETTER TEAM CAPSTONE PROJECT.
- Airbnb is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities.
- This dataset contains details of different neighbourhood,hosts,types of rooms,price and reviews from different users.

Problem Statement

- What can we learn about different hosts and areas?
- What can we learn from predictions? (ex: locations, prices, reviews, etc)
- Which hosts are the busiest and why?
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?

Features Description

The features in the dataset can be described as follows:

1. id - This is the identity number of the property listed by a particular host.
2. name - It stands for the name of the property listed by the host.
3. host_id - It is the identity number of the hosts who have registered on Airbnb website.
4. host_name - These are the names of the hosts who have listed their properties.
5. neighbourhood_group - These are the names of the neighbourhood groups present in the NYC.
6. neighbourhood - These are the names of the neighbourhood present in the neighbourhood groups in NYC.

7. latitude - These represent the coordinates of latitude of the property listed.
8. longitude - These represent the coordinates of longitude of the property listed.
9. room type - This represent the various types of room listed by host.
10. price - This is the rent of the property listed in USD.
11. minimum nights - This represent the minimum number of nights customer rented the property.
12. Number_of_reviews - This represent the number of customers reviewed the property.
13. last_review - This represent the date when the property was last reviewed.
14. reviews_per_month - It is the count of reviews per month which the property received.
15. calculated_host_listings_count - It is the number of listings done by a particular host.
16. Availability_365 - This represent the number of days the property is available among 365 days.

Work Flow

- So we will divide our work flow into following 3 steps.

Data Collection and Understanding:

1. Loading the data into data frame.

2. Data Summary.

3. Extracting statistics from the dataset.

1.Loading of dataframe

As shown in the image we importing the data set on Google Colab.

```
[ ]  # Importing Data Set
```

```
[ ]  from google.colab import drive
     drive.mount('/content/drive')
```

```
     Mounted at /content/drive
```

```
[ ]  airbnb_df = pd.read_csv('/content/drive/My Drive/Colab Notebooks/Project 1/Airbnb NYC 2019.csv')
```

As shown in the output image we load the required data set to the system with the help of head() method for the data set of top order and tail() method for the data set of bottom order

Top order Dataset loading

```
airbnb_df.head().T
```

	0	1	2	3	4
id	2539	2595	3647	3831	5022
name	Clean & quiet apt home by the park	Skylit Midtown Castle	THE VILLAGE OF HARLEM....NEW YORK !	Cozy Entire Floor of Brownstone	Entire Apt: Spacious Studio/Loft by central park
host_id	2787	2845	4632	4869	7192
host_name	John	Jennifer	Elisabeth	LisaRoxanne	Laura
neighbourhood_group	Brooklyn	Manhattan	Manhattan	Brooklyn	Manhattan
neighbourhood	Kensington	Midtown	Harlem	Clinton Hill	East Harlem
latitude	40.64749	40.75362	40.80902	40.68514	40.79851
longitude	-73.97237	-73.98377	-73.9419	-73.95976	-73.94399
room_type	Private room	Entire home/apt	Private room	Entire home/apt	Entire home/apt
price	149	225	150	89	80
minimum_nights	1	1	3	1	10
number_of_reviews	9	45	0	270	9
last_review	2018-10-19	2019-05-21	NaN	2019-07-05	2018-11-19
reviews_per_month	0.21	0.38	NaN	4.64	0.1
calculated_host_listings_count	6	2	1	1	1
availability_365	365	355	365	194	0

Bottom order dataset loading

```
airbnb_df.tail()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_cou
48890	36484665	Charming one bedroom - newly renovated rowhouse	8232441	Sabrina	Brooklyn	Bedford-Stuyvesant	40.67853	-73.94995	Private room	70	2	0	NaN	NaN	
48891	36485057	Affordable room in Bushwick/East Williamsburg	6570630	Marisol	Brooklyn	Bushwick	40.70184	-73.93317	Private room	40	4	0	NaN	NaN	
48892	36485431	Sunny Studio at Historical Neighborhood	23492952	Ilgar & Aysel	Manhattan	Harlem	40.81475	-73.94867	Entire home/apt	115	10	0	NaN	NaN	
48893	36485609	43rd St. Time Square-cozy single bed	30985759	Taz	Manhattan	Hell's Kitchen	40.75751	-73.99112	Shared room	55	1	0	NaN	NaN	
48894	36487245	Trendy duplex in the very heart of Hell's Kitchen	68119814	Christophe	Manhattan	Hell's Kitchen	40.76404	-73.98933	Private room	90	7	0	NaN	NaN	

2.DATA SUMMARY

```
airbnb_df.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
```

- There are 48895 observations from 16 columns in the data set.

- There are 3 types of data type in the data set.

  Float - 3

  Integer - 7

  Object - 6

3.Extracting statistics from the dataset

```
airbnb_df.shape
```

```
(48895, 16)
```

```
[ ] airbnb_df.columns
```

```
Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365'],
      dtype='object')
```

```
[ ] airbnb_df.describe()
```

	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 Preparation and Cleaning:

- Data preparation is the process of cleaning and transforming raw data prior to processing and analysis.
- It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data.
- Removing extraneous data
- Filling in missing values.
- Conforming data to a standardized pattern.

- The columns like last_review and reviews_per_month have largest number of null values.
- The columns like name and host_name contain fewer number of null values.

```
airbnb_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
```

Taking Necessary Columns Only

```
df = airbnb_df[[ 'id','name','host_id','host_name','neighbourhood_group','neighbourhood','room_type','price','minimum_nights',
                 'number_of_reviews','calculated_host_listings_count','availability_365']]
```

Exploratory Data Analysis :

After Analyzing the dataset we have got answers to some of the serious & interesting questions which any of the android users would love to know.

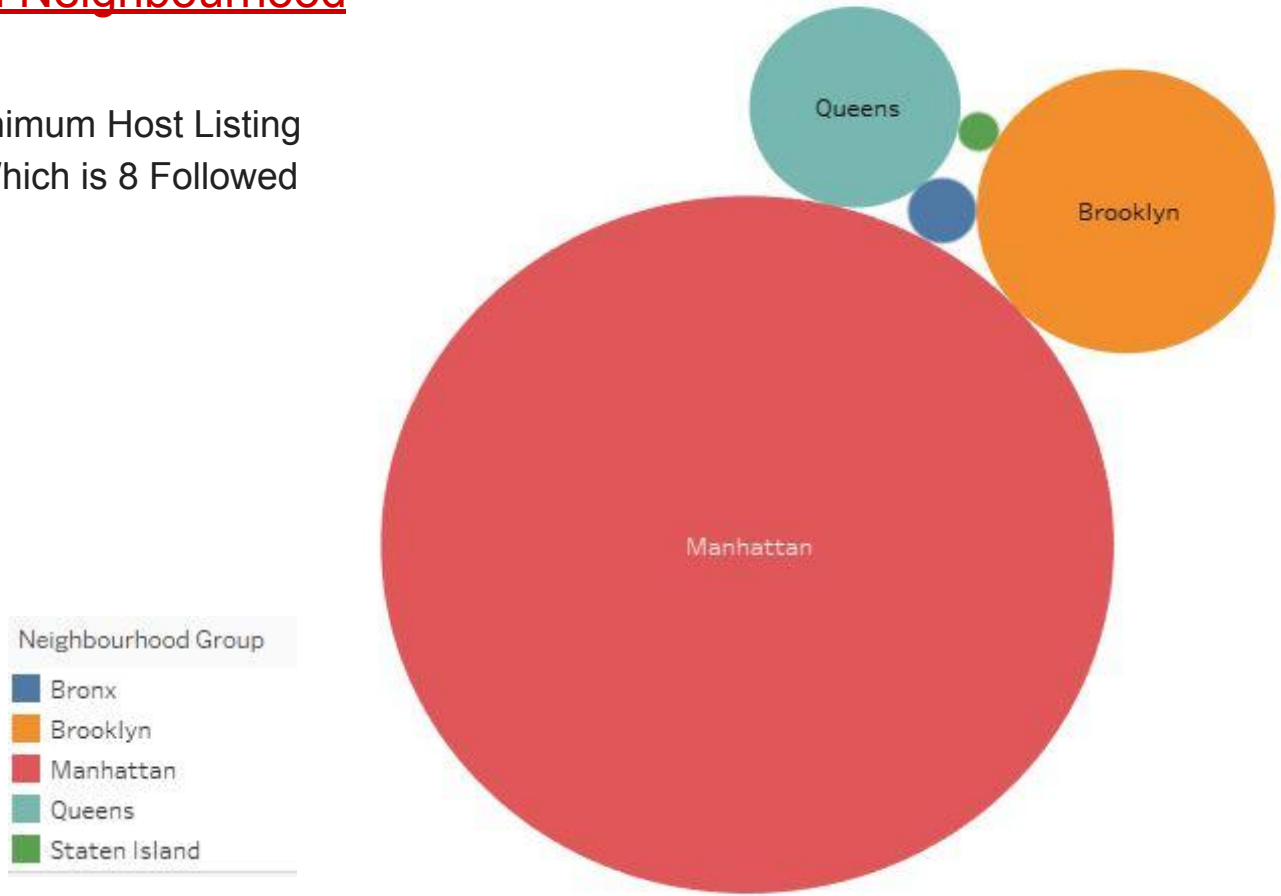
1) Maximum Host Listing in Neighbourhood ?
2) Minimum Host Listing in Neighbourhood ?
3) Which Hosts are Busiest ?
4) Hosts with maximum listings ?
5) Which Room Type Hosts Preferred The Most and Least ?
6) Room Types Count in different Neighbourhood ?
7) What can we learn from locations, prices and reviews in Data?

Maximum Host Listing in Neighbourhood

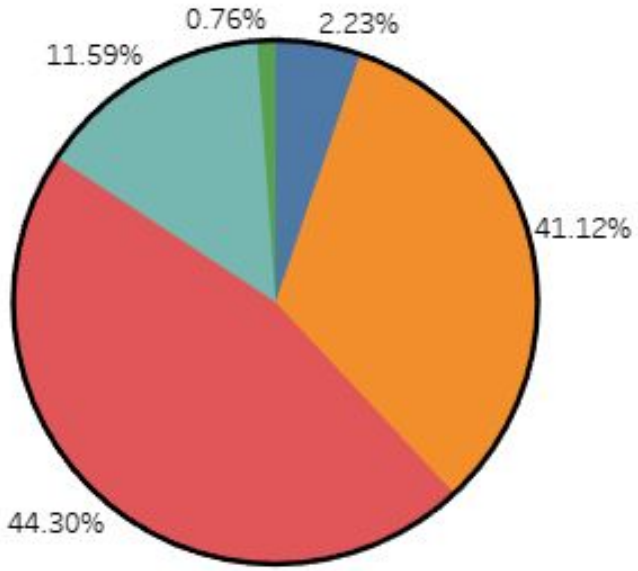
- Here we can see the Maximum Host Listing are from Manhattan Which is 327 Followed by Brooklyn Which is 232.

Minimum Host Listing in Neighbourhood

- Here we can see the Minimum Host Listing are from Staten Island Which is 8 Followed by Bronx Which is 37.

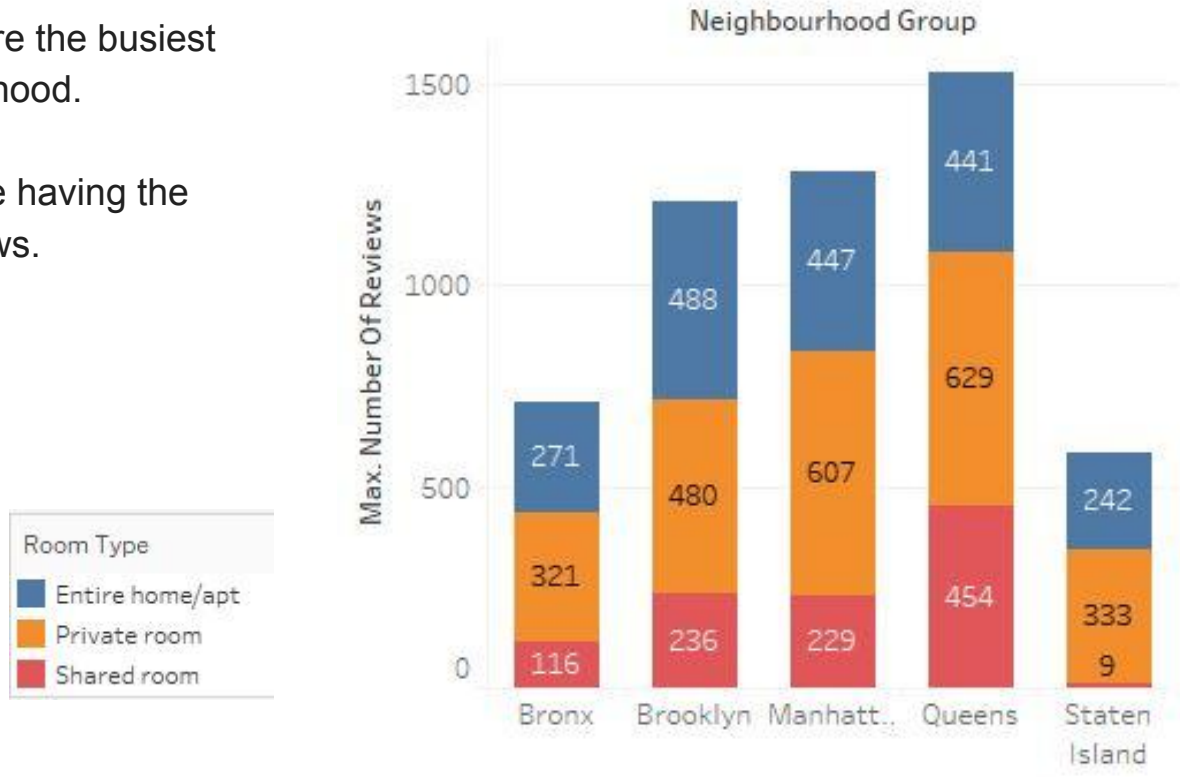


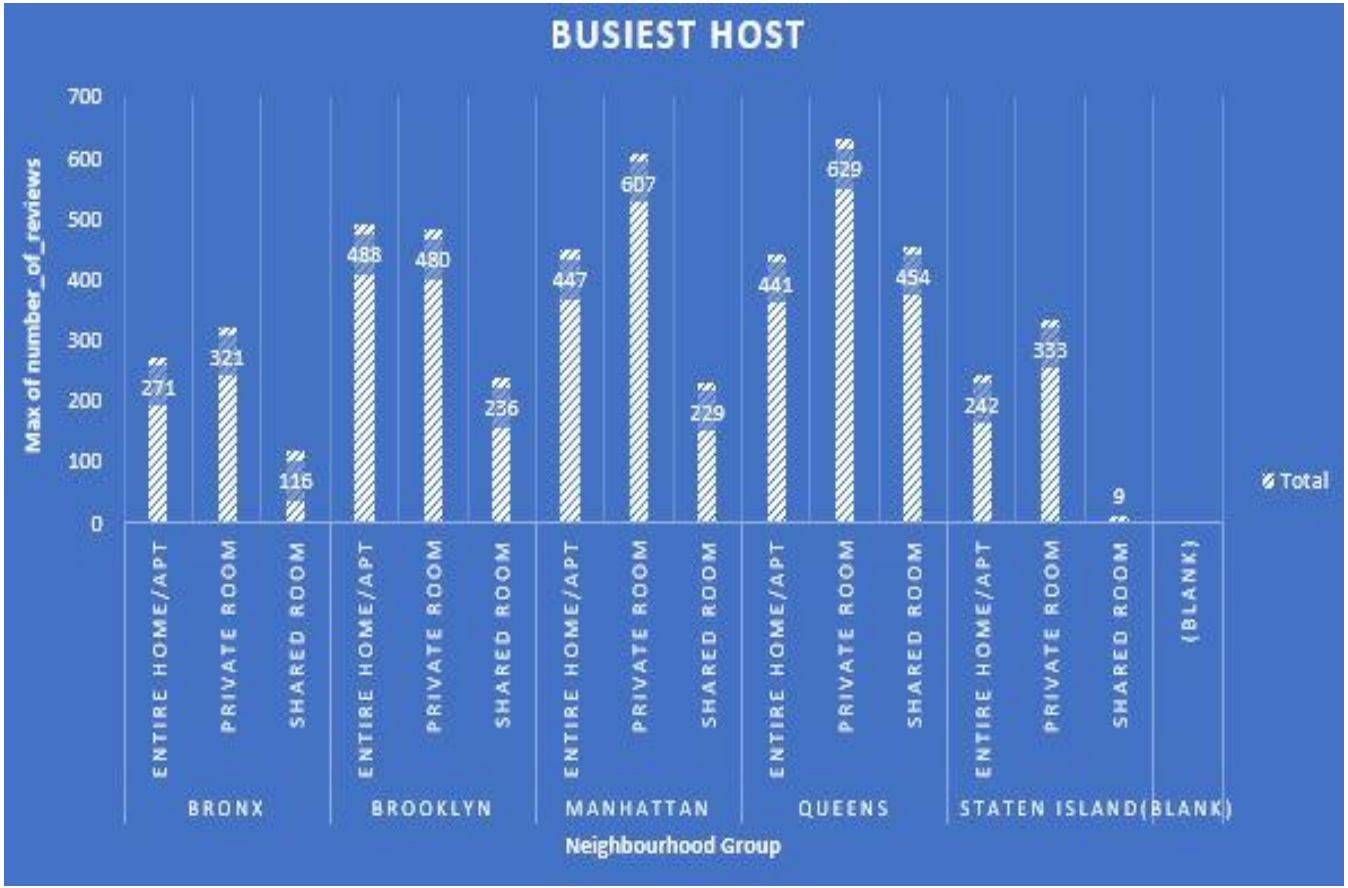
- Out of the total host listing Manhattan has 44.30%.
- Staten island has the least listing of 0.76% followed by Bronx which is having 2.33%.
- While, Brooklyn has 41.12% and Queens has 11.59% of Maximum Host Listing.



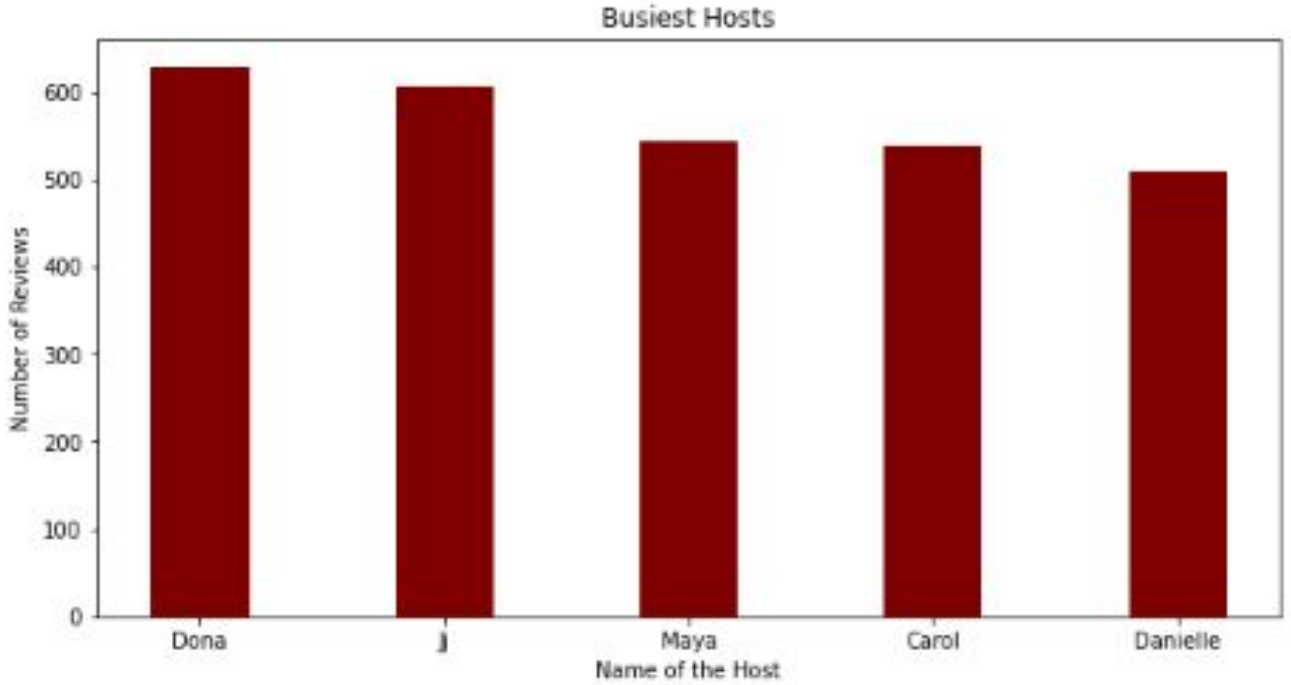
Which Hosts are Busiest and Why

- The host listed in Queen are the busiest among all listed neighbourhood.
- Because Private rooms are having the maximum number of reviews.



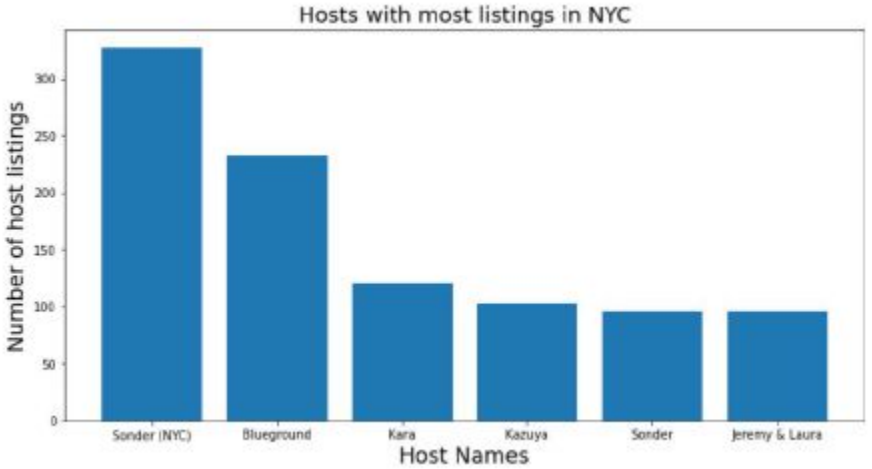

BUSIEST HOST
Max of number_of_reviews
700
600
500
400
300
200
100
0
271
321
116
488
480
236
447
607
229
441
629
454
242
333
9
ENTIRE HOME/APT
PRIVATE ROOM
SHARED ROOM
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND
(BLANK)
Neighbourhood Group
Total

- The Top 5 Busiest Hosts are Dona, Ji, Maya ,Crol , Danielle by Number of Reviews.



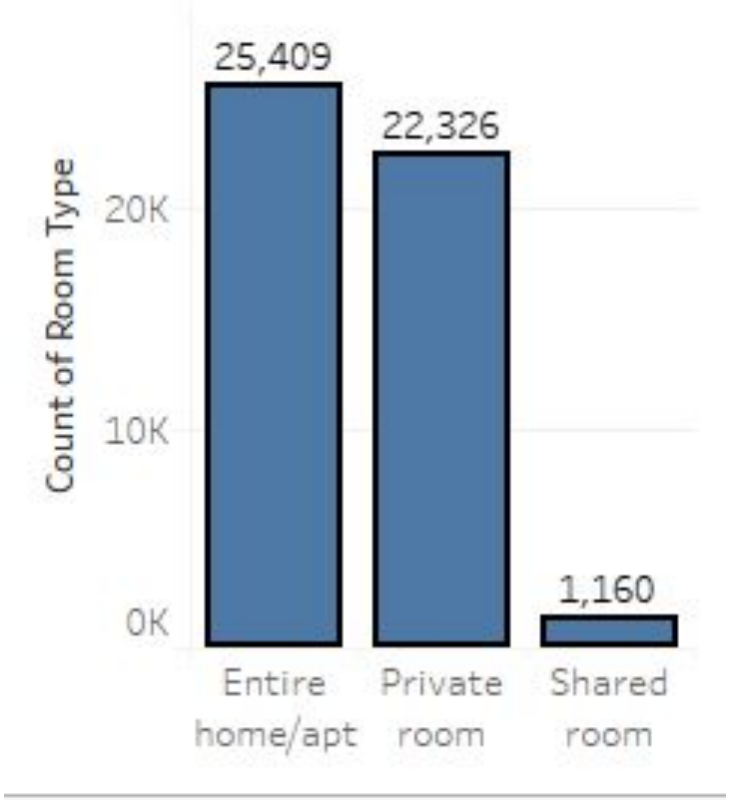
Hosts with maximum listings

- As shown in the adjacent bar chart, we can see there is a good distribution among the top 6 hosts.
- The host named Sonder(NYC) has highest number of listings of 327 in Manhattan neighbourhood group.
- The host named Bluegorund has 2nd highest listings of 232 in Manhattan Neighbourhood group.
- The host Blueground also has 232 listings in Brooklyn.



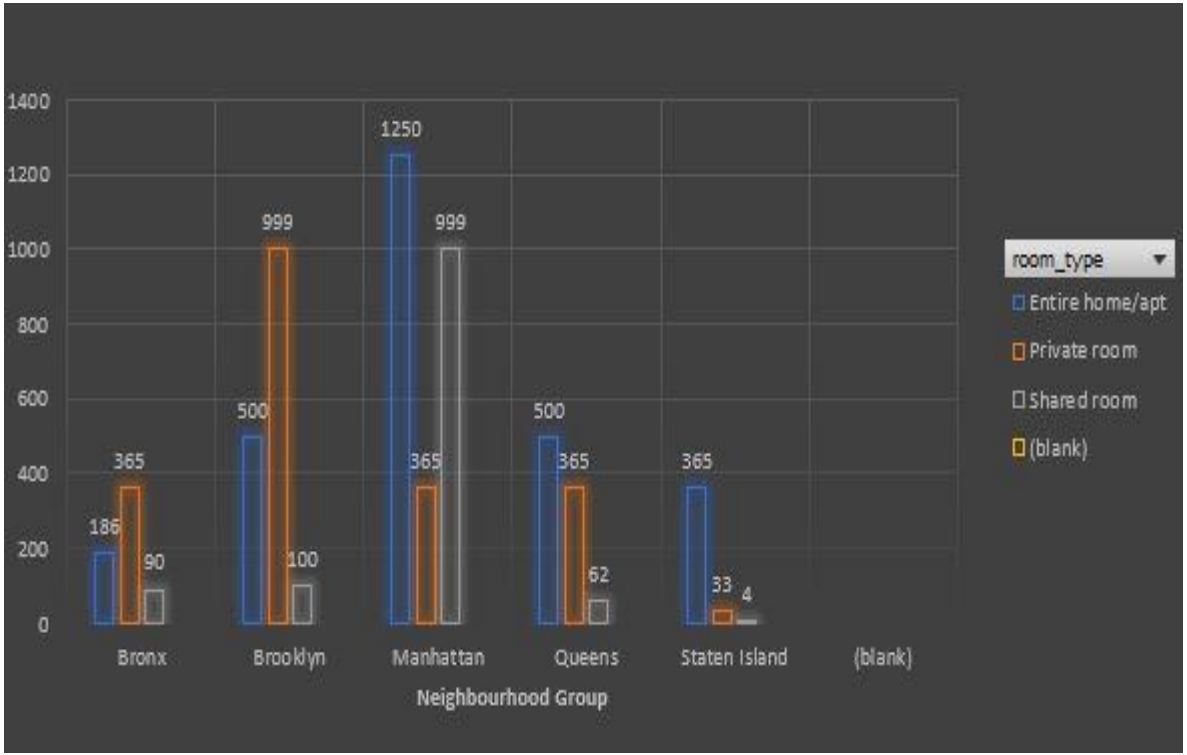
Which Room Type Hosts Preferred The Most and Least

- Looking at the Data, we can say that there are 3 room types listed in the entire dataset namely Private room,Entire home/apt,Shared room.
- Among this types the most preferred room type is Entire home/apt as well as private Room.
- Shared room is least preferred by people.



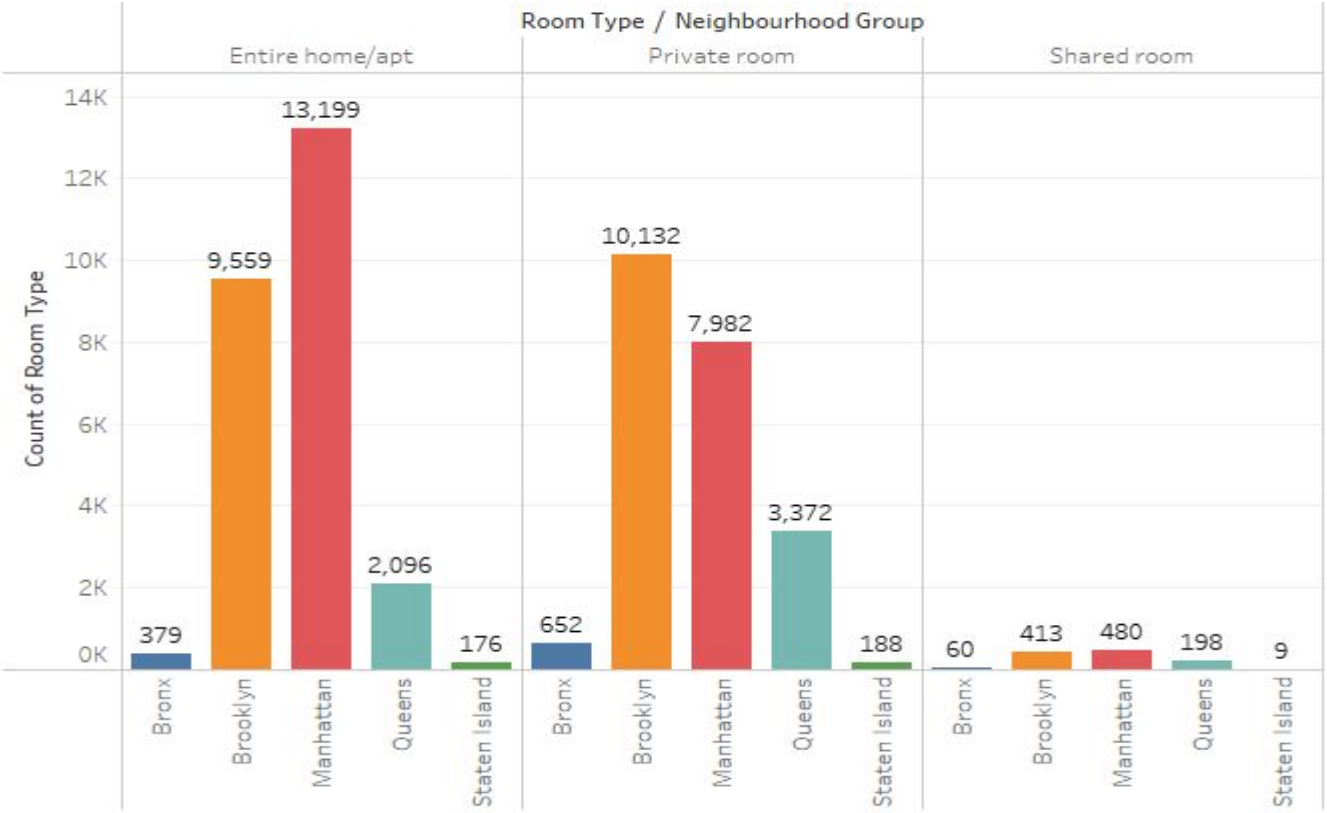
- Entire home/apt is been Preferred by Hosts in Neighbourhood Group Except in Brooklyn and Bronx.

- Shared Rooms Type is the Least Preferred In every Neighbourhood Group



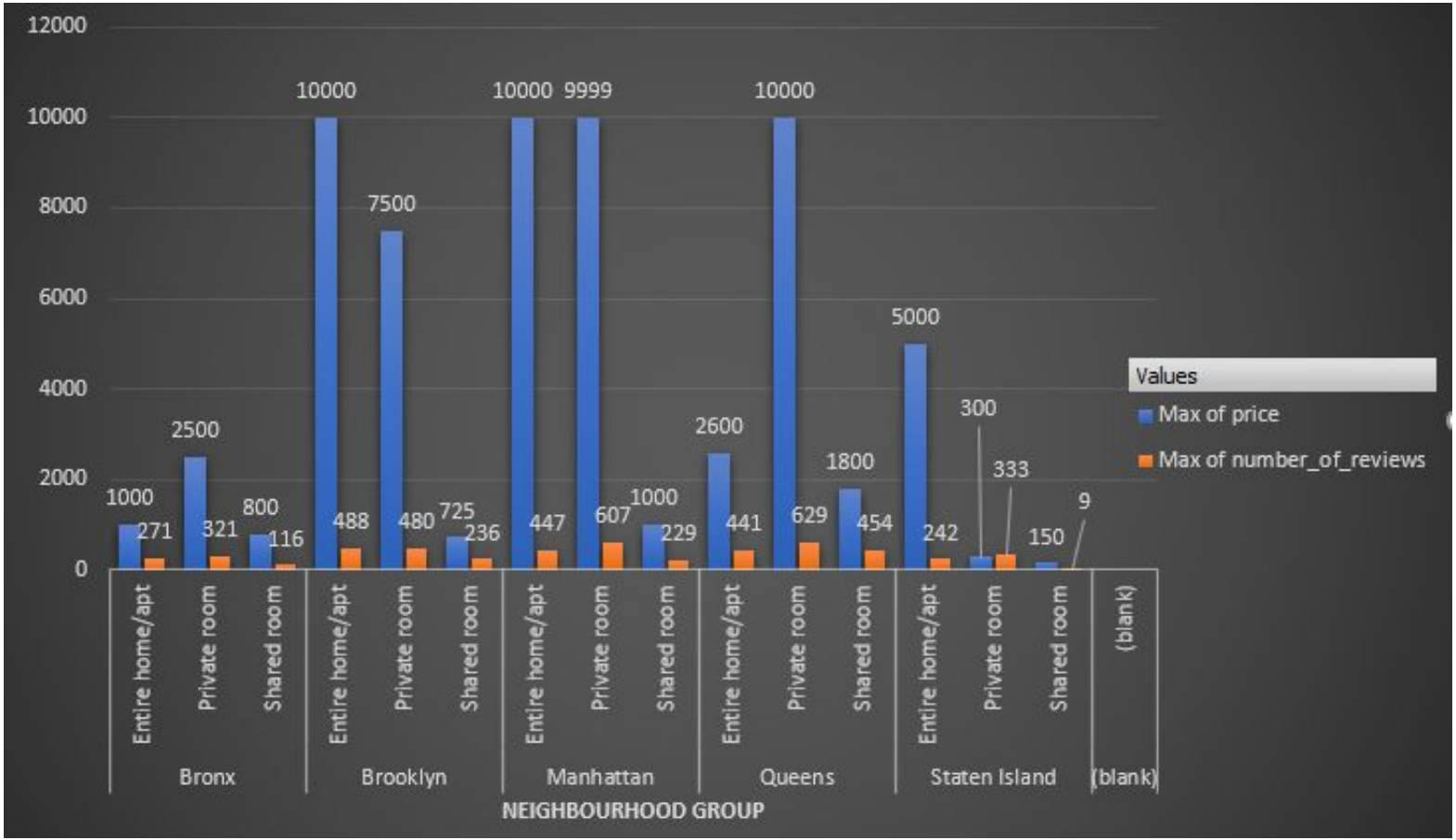
Room Types Count in different Neighbourhood Group

- We Can see in Each Neighbourhood The Shared Rooms are the least Of all Room Type.
- Whereas the Entire room/apt And Private Room are Available in Most Neighbourhood.

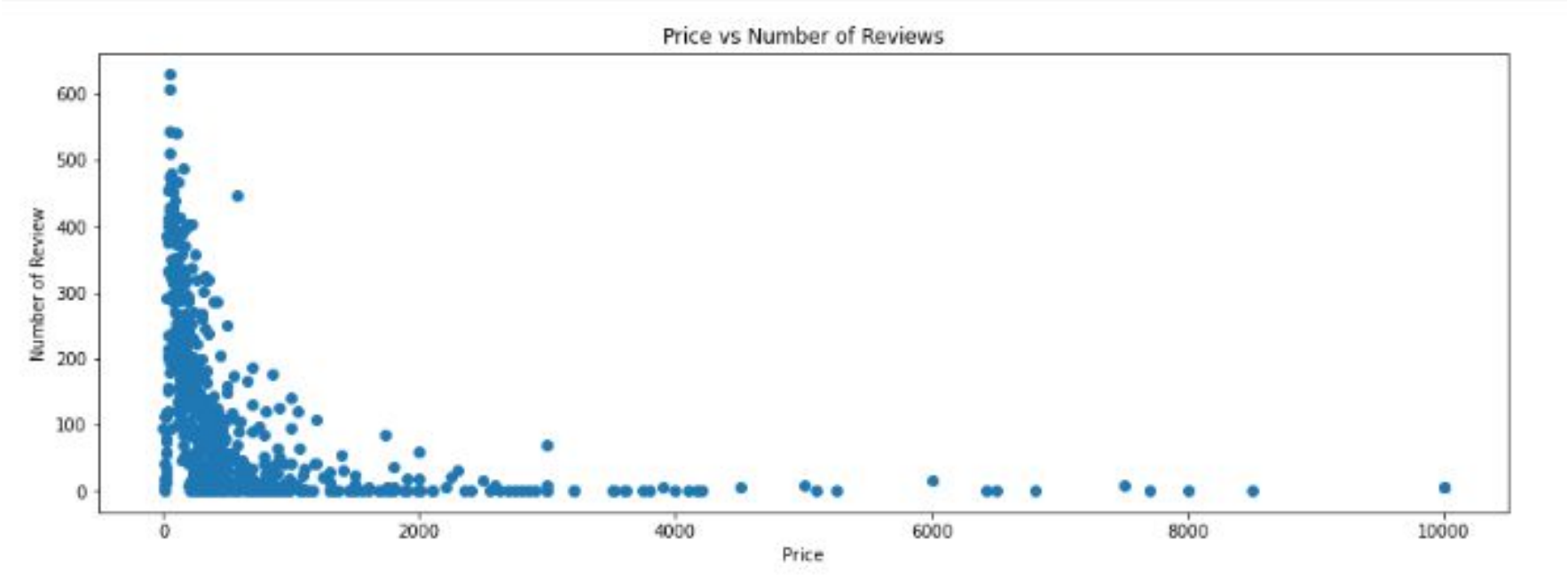


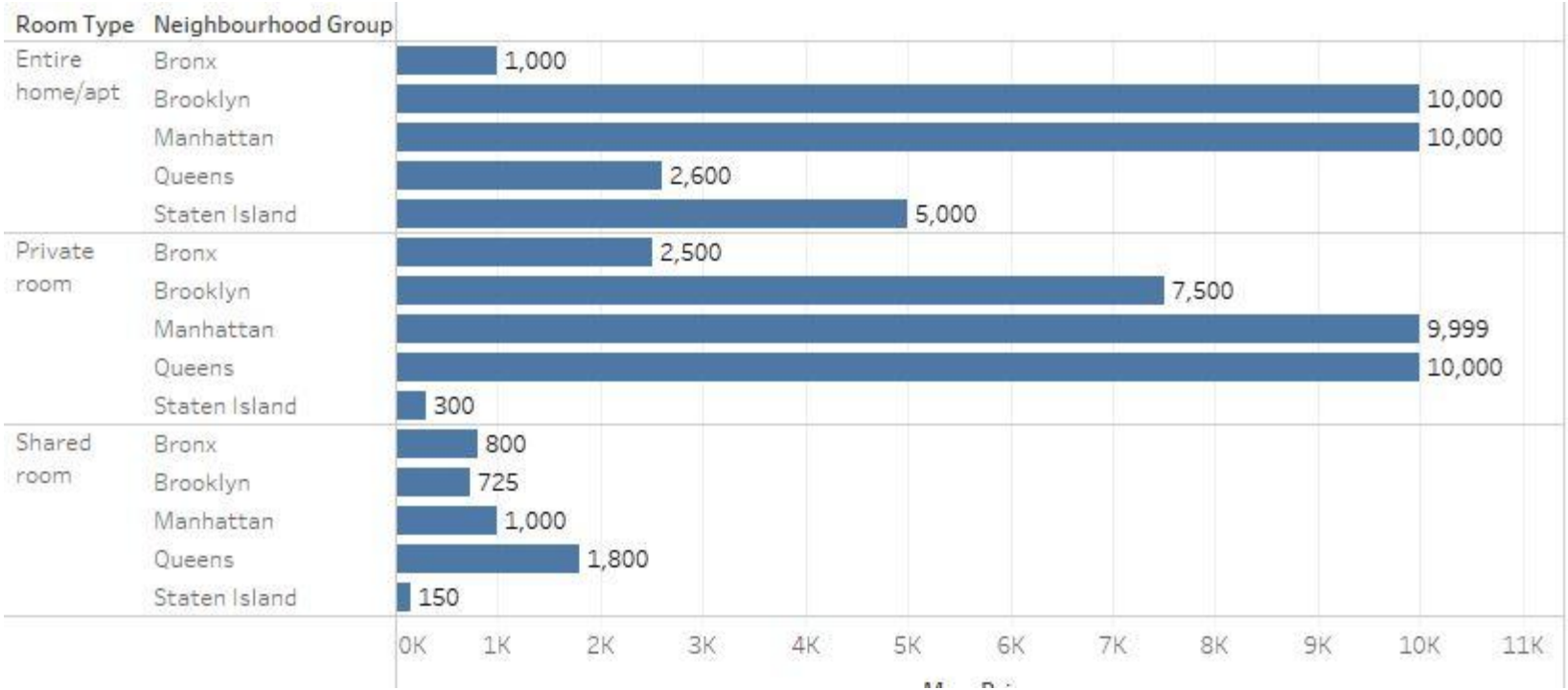
What can we learn from locations, prices and reviews in Data

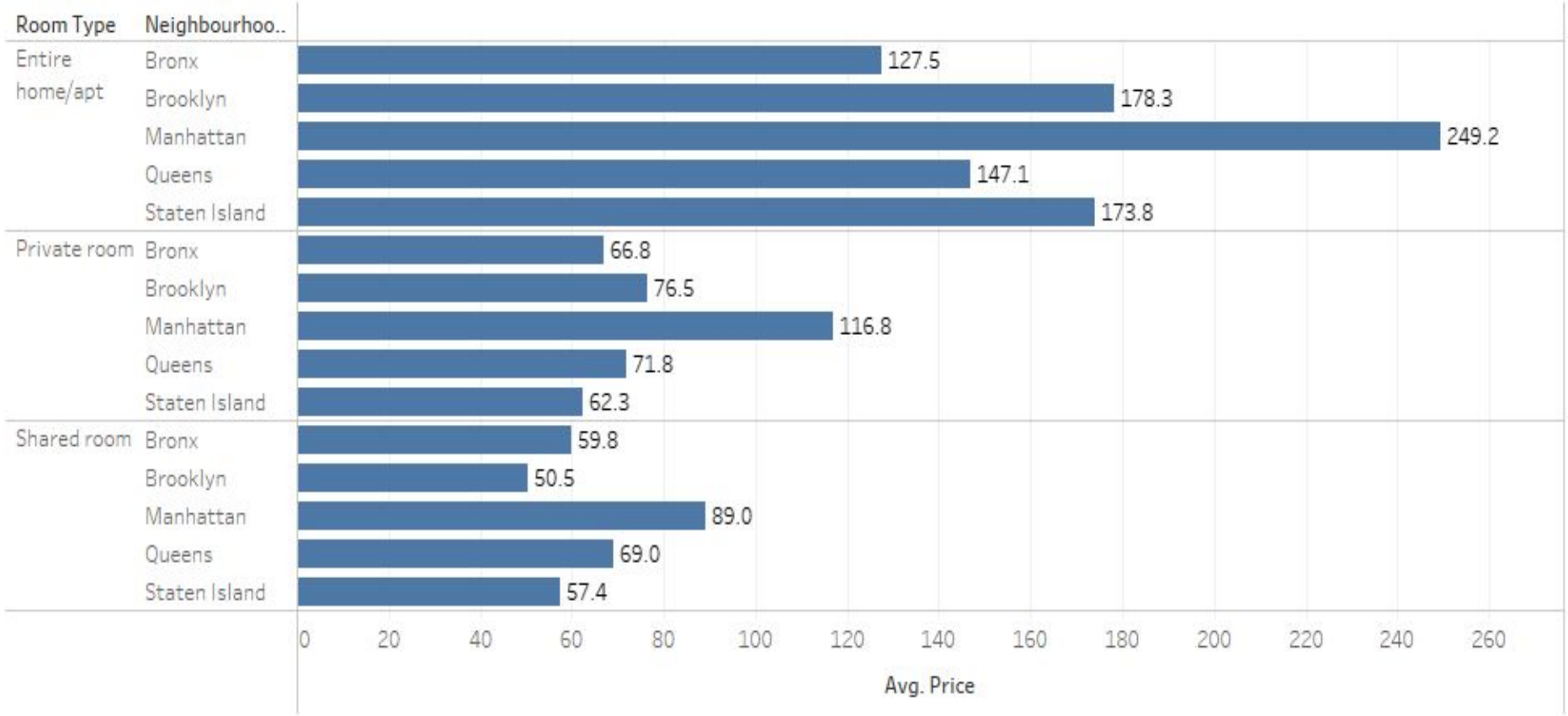
- We can learn that the Queens neighbourhood has maximum prices and maximum review.
- We can learn that the Staten Island neighbourhood has minimum prices and minimum review.
- We can also learn about the maximum Price of Different Room Type in Different Neighbourhood.


12000
10000
8000
6000
4000
2000
0
1000
271
2500
321
800
116
10000
488
7500
480
725
236
10000
447
9999
607
1000
229
2600
441
10000
629
1800
454
5000
242
300
150
333
9
Values
Max of price
Max of number_of_reviews
Entire home/apt
Private room
Shared room
Bronx
Brooklyn
Manhattan
Queens
Staten Island
(blank)
NEIGHBOURHOOD GROUP

- The Below Data tells that the People likes to Stay Where the Price is Low.




Room Type
Neighbourhood Group
Entire home/apt
Bronx
Brooklyn
Manhattan
Queens
Staten Island
1,000
10,000
10,000
2,600
5,000
Private room
Bronx
Brooklyn
Manhattan
Queens
Staten Island
2,500
7,500
9,999
10,000
300
Shared room
Bronx
Brooklyn
Manhattan
Queens
Staten Island
800
725
1,000
1,800
150
0K
1K
2K
3K
4K
5K
6K
7K
8K
9K
10K
11K


Room Type
Neighbourhoo..
Entire home/apt
Bronx
Brooklyn
Manhattan
Queens
Staten Island
127.5
178.3
249.2
147.1
173.8
Private room
Bronx
Brooklyn
Manhattan
Queens
Staten Island
66.8
76.5
116.8
71.8
62.3
Shared room
Bronx
Brooklyn
Manhattan
Queens
Staten Island
59.8
50.5
89.0
69.0
57.4
0
20
40
60
80
100
120
140
160
180
200
220
240
260
Avg. Price

Conclusion

- Manhattan has most number of listings, followed by Brooklyn and Queens. Staten Island has least number of listings.
- Manhattan and Brooklyn make up for 87% of listings available in NYC.
- Brooklyn and Manhattan are most liked neighbourhood groups by people.
- Queens has significantly less host listings than Manhattan. So, we should take enough steps to encourage host listings in Queens as there is decent demand in the neighbourhoods of Queens.
- The maximum demand is for private rooms and entire home/apartment. People are more interested in cheaper rentals.
- The top 10 neighborhoods with the most listings are located either inManhattan or Brooklyn, with Harlem and Williamsburg presenting leading numbers in each borough, respectively.

- Manhattan is the top neighbourhood group in terms of listings as well as highest price range. It was assumed that Brooklyn might have most number of listings as it is a quite popular place.

- Given that Manhattan is world-famous for its museums, stores, parks, and theaters, also its substantial number of tourists throughout the year, it makes perfect sense that prices are much higher in this neighbourhood group.

- Brooklyn comes in second with significant number of listings and cheaper prices as compared to the Manhattan. With most listings located in Williamsburg and Bedford Stuyvesant — two neighborhoods Strategically close to Manhattan — tourists get the chance to enjoy both boroughs equally while spending less.

THANK YOU