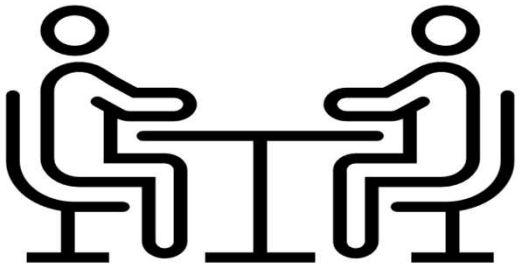
Capstone Project-1

Airbnb Bookings Analysis

Prepared By

- Name : Sandipan Das
  Email : sandipan.das202@gmail.com
- Name : Kashif Kamran
  Email : kashifkamran639@gmail.com
- Name : Raja Chowdhury
  Email : rajachowdhury2468@gmail.com
- Name : Aman Jain
  Email : amanjn932@gmail.com

Points of Discussion



- 1.Understanding the Airbnb
- 2.Defining Problem Statements
- 3.Data Exploration
- 4.Data Preparation
  - Data Cleaning
  - Handling Outliers
  - Understanding Correlation
- 5.Exploratory Analysis
  - 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?
- 6.Key Findings & Conclusion

UNDERSTANDING THE AIRBNB

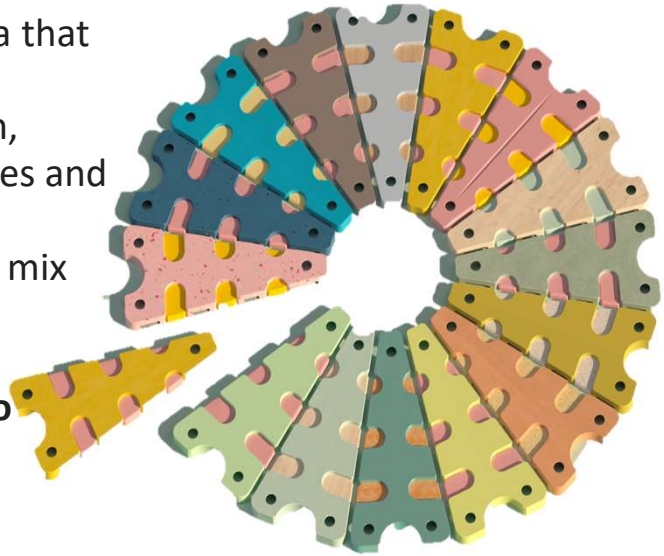
Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Based in San Francisco, California.

The platform is accessible via website and mobile app.

In general, Airbnb is cheaper than hotels because Airbnb does not have to pay for the overhead costs of a hotel or the general management of such a large operation.

DEFINING PROBLEM STATEMENT

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one of a kind service that is used and recognized by the whole world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - 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.
This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.



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

- 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?

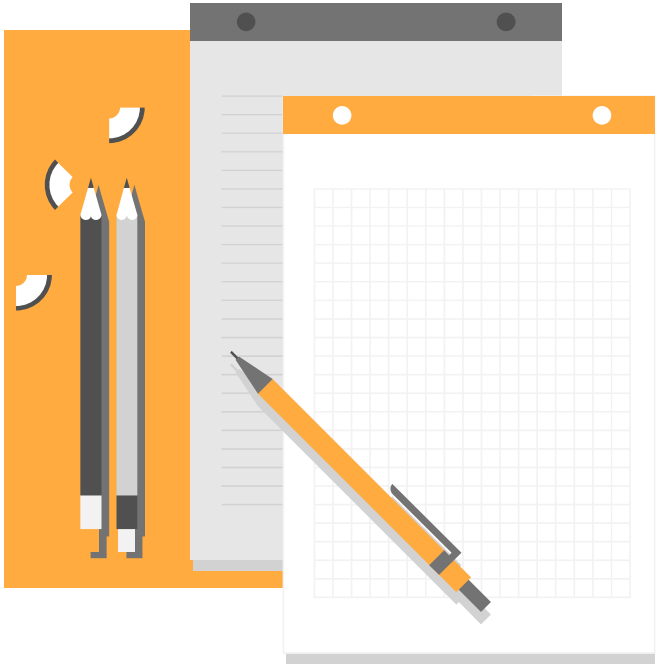
DATA EXPLORATION

- After reviewing the Airbnb NYC 2019 dataset we found that it contains 48895 Rows and 16 Column.
- The given Airbnb NYC 2019 dataset contains booking information Customer Details, Host details, Neighborhood details, Room details like type, price, reviews and availability.
- Also gives the location details with latitude and longitude, counts of Listings.
- All the data types are in the required format only, and no need to convert or change the datatype.
- There are total 4 columns containing the null values and We'll decide what to do with these null value columns depending on the column's importance
- To perform EDA we need to find the columns which are important for our analysis and then we'll only consider those columns for our analysis and ignore the remaining columns,
- On Broder scale we are having two types of Data one is categorical and other is Continuous Numerical data.

```
bnb.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
```

DATA PREPARATION



In Data Preparation we will be exploring following:

- Data Cleaning: In this part we will check for Null Values and decide how to handle for each column.
- Outliers: In this part will try to Look into the different column and see if there is outliners present if so, will handle the same.
- Correlation: Will See the if there is some potential correlations between the numerical data.

Data Cleaning

- As we can see there are some Null Values in **host_name**, **name** , **last_review** and **reviews_per_month**

- **host_name & name** : Missing Values are 21 & 16 both are less than 0.5% so will simply drop the rows corresponding to the missing values in those column.

- **last_review** and **reviews_per_month** : we found Null Values are for those properties which doesn't have any reviews. so will replace all the Null values of review_per_month with zero ('0') and drop the **last_review** as column has very high Null Values and negligible relevance with our problem set for EDA.

	Total	Percentage
last_review	10052	20.558339
reviews_per_month	10052	20.558339
host_name	21	0.042949
name	16	0.032723
id	0	0.000000
host_id	0	0.000000
neighbourhood_group	0	0.000000
neighbourhood	0	0.000000
latitude	0	0.000000
longitude	0	0.000000
room_type	0	0.000000
price	0	0.000000
minimum_nights	0	0.000000
number_of_reviews	0	0.000000
calculated_host_listings_count	0	0.000000
availability_365	0	0.000000

Handling Outliers

```
df_nyc.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

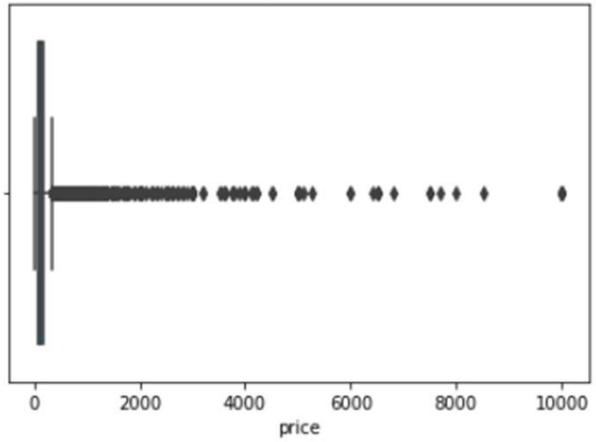
using `.describe()` to get a summary statistics of the numeric data, We Observe that there are some outliers for **price and minimum_nights**. Other columns such as **number_of_reviews and calculated_host_listings_count** are skewed toward right.

Handling Outliers

To confirm the Outlier in the **price** and **minimum_nights** Column we use Boxplot Chart to visualize the Outliers

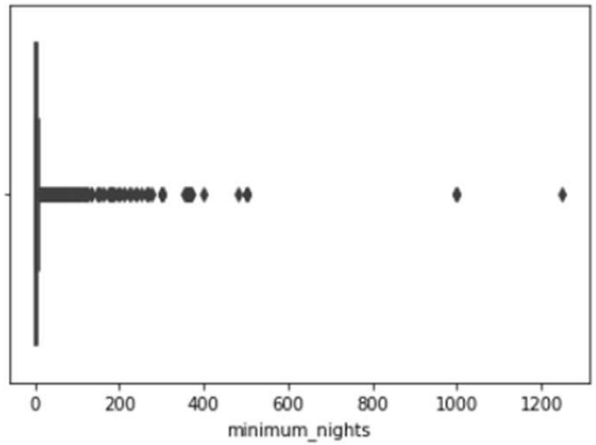
```
sns.boxplot(x=df['price'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fda01f46510>
```




```
[ ]  sns.boxplot(x=df['minimum_nights'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fda00111990>
```



Handling Outliers

Outliers can be easily seen in both column in Boxplot Chart.
To handle these outliners,

- Firstly, we inserted two column new Column with Z-Score of the corresponding values of the Respective column.
- Then Removed all those entries having Z-Score More than or equal to 3.
- As purpose solved, we drop the Column with Z-Score.

```
[ ]  df['z_price'] = np.abs(stats.zscore(df['price']))
     df['z_min_nights'] = np.abs(stats.zscore(df['minimum_nights']))
```

```
[ ]  # remove z scroe that are greater than 3

     df = df[(df['z_price'] < 3)]
     df = df[(df['z_min_nights'] < 3)]
```

```
[ ]  # Droping 'z_price' and 'z_min_nights'
     df.drop(['z_price','z_min_nights'],axis=1,inplace=True)
```

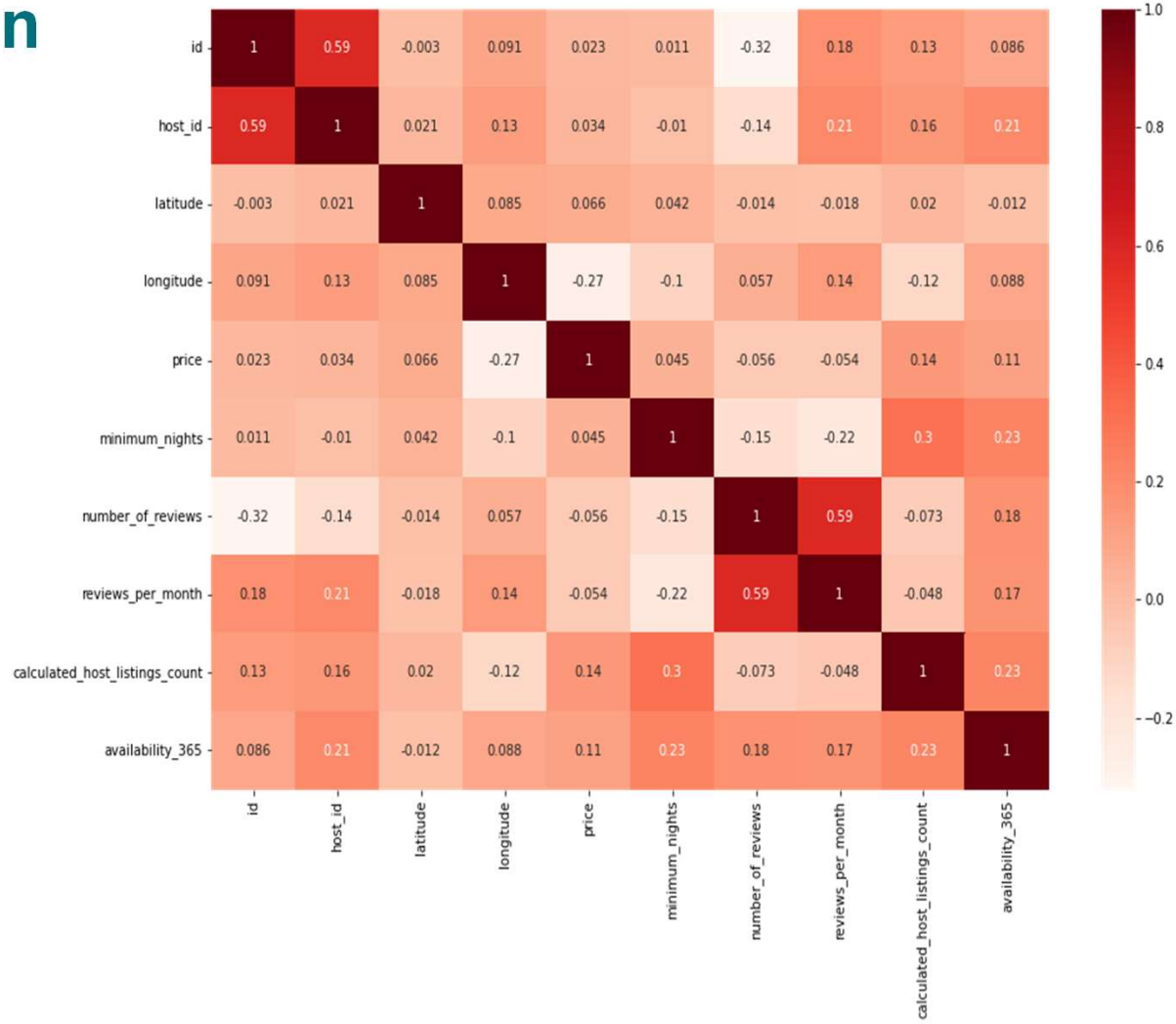
Understanding Correlation

To Understand if there is any correlation between the Numerical Data, we try to Plot Heat Map.

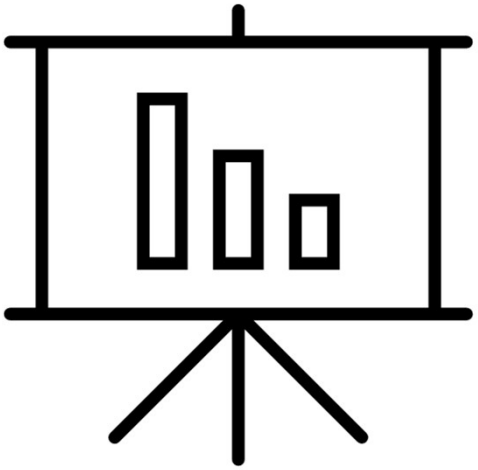
Clearly Only two correlation between

- Id and host_id
- number_of _reviews & reviews_per_month

That two are not having much analytical relevance so we assume among Numerical datas there is no potential correlation.



Exploratory Analysis



In Exploratory Analysis we will try to address the following Questions

- 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?

What can we learn about different hosts and areas?

With this Question we can :

- Come up first with the use of longitude and latitude available in our dataset, we can visualize the Different Neighborhood areas.
- Analyze the Top Hosts of the Area
- Explore the Booking Distribution in different neighborhood group.
- Analyze the top neighborhood of the areas.

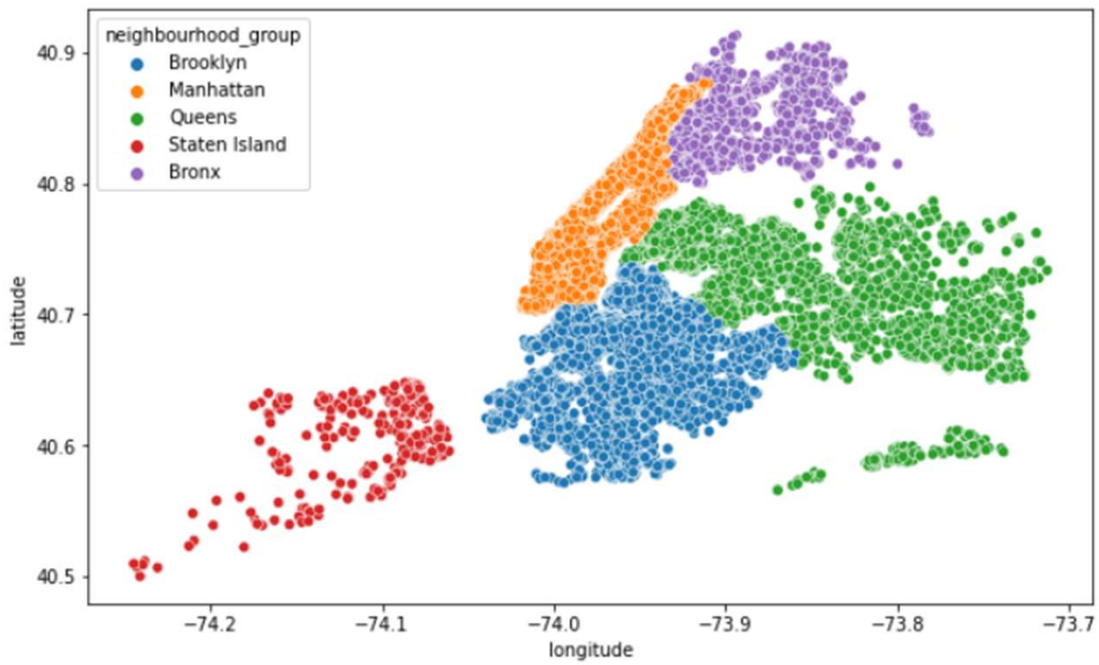
What can we learn about different hosts and areas?

- **First come first with the use of co-ordinates i.e longitude and latitude, we can visualize the Different Neighborhood areas.**

So here what we get:

- Each points in coordinates denotes booking to corresponding host of the area.
- Manhattan seems to be smaller by area but also very high number of booking.
- With scattered plot in Staten Island, we can think of either the listed host is very less, or booking is very less or both.
- Queen and Brooklyn seems good participation of Host listing.

```
plt.figure(figsize=(10,6))
sns.scatterplot(df.longitude,df.latitude,hue=df.neighbourhood_group)
plt.ioff()
```

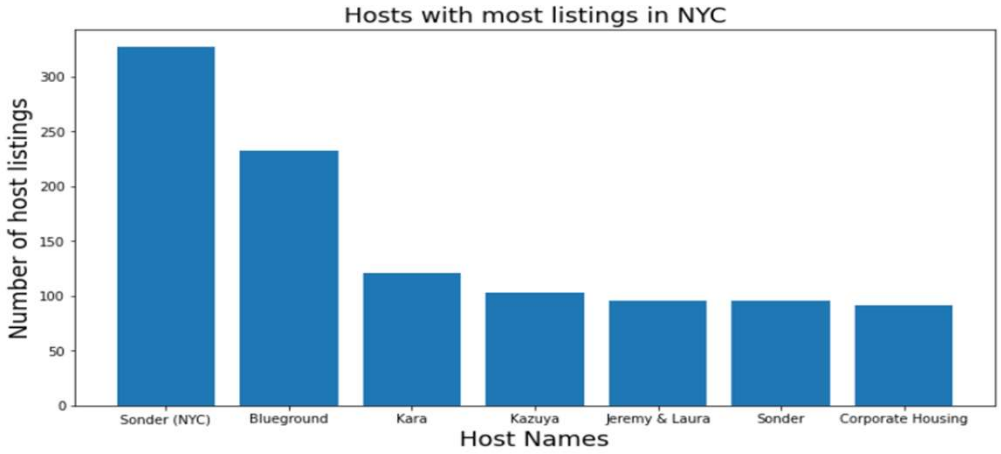


What can we learn about different hosts and areas?

❖ **Analyze the Top Hosts of the Area**

<u>So here what we get:</u>

- As Expected, most of the high listing counts hosts are in Manhattan some are in Brooklyn and Queens too.
- Sonder(NYC) from Manhattan has the highest Number of Listing.
- Based on our Last Analysis Outcomes we can think of Manhattan to be highest contributor among all neighborhood group, our hypothesis can be wrong or could be right we will find out in coming analysis in Neighborhood group.



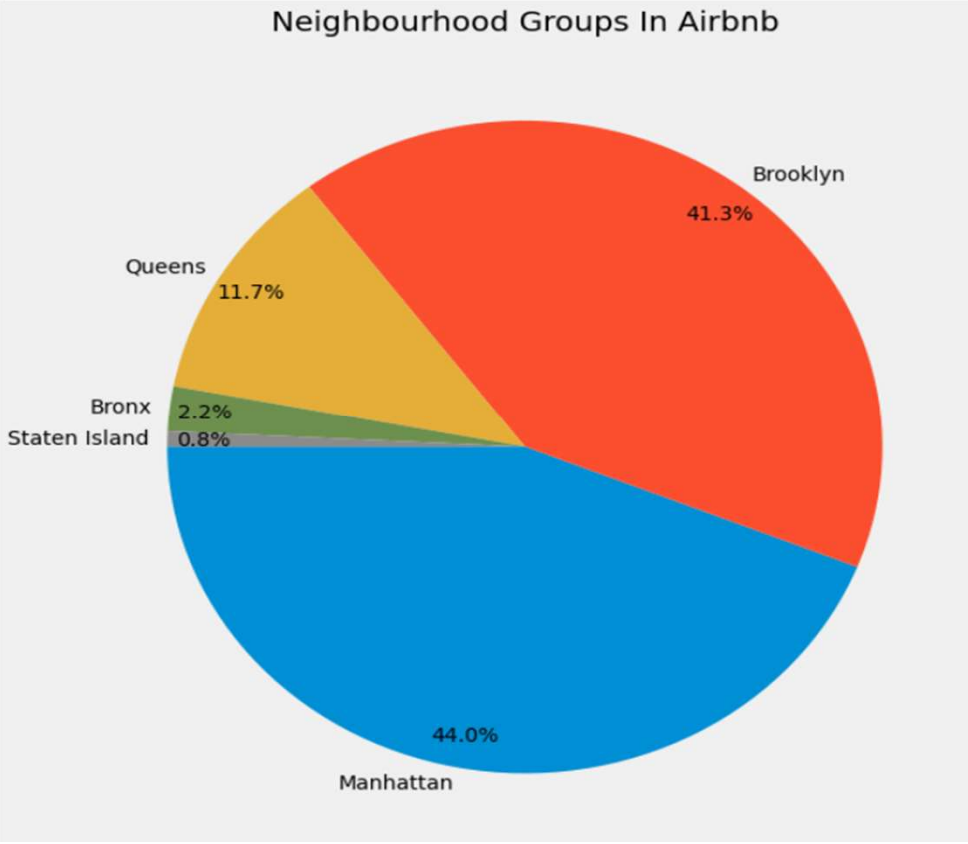
host_name	neighbourhood_group	calculated_host_listings_count
Sonder (NYC)	Manhattan	327
Blueground	Brooklyn	232
Blueground	Manhattan	232
Kara	Manhattan	121
Kazuya	Manhattan	103
Kazuya	Queens	103
Kazuya	Brooklyn	103
Jeremy & Laura	Manhattan	96
Sonder	Manhattan	96
Corporate Housing	Manhattan	91

What can we learn about different hosts and areas?

❖ **Explore the Booking Distribution in different neighborhood group.**

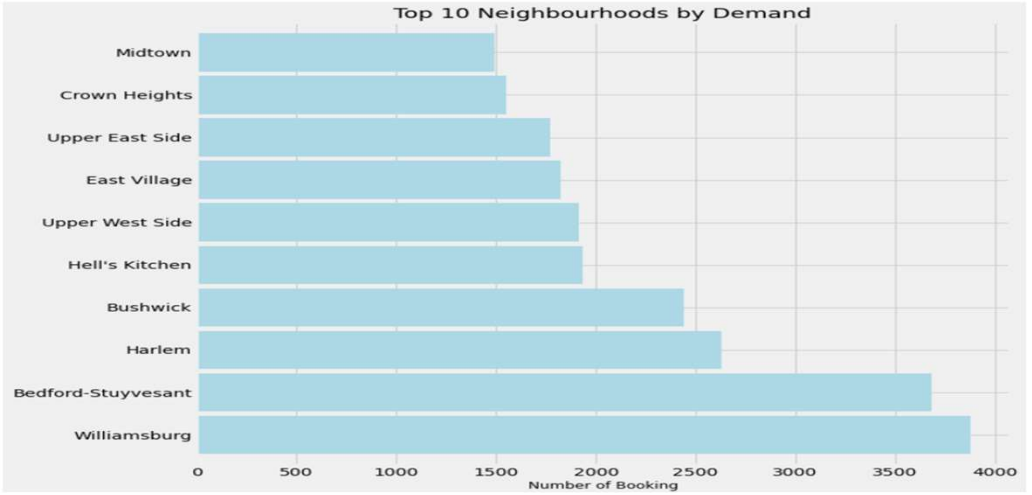
<u>So here what we get:</u>

- As per our Hypothesis we can conclude here that with 44% Manhattan has the highest no of bookings compare to all other neighborhood group followed by Brooklyn with 41.3%
- As expected, Staten Island with just 0.8% of total booking are at bottom.
- There is big difference between Queens and Brooklyn which was not as per our expectation.
- Also top two which is Manhattan and Brooklyn contribute the 85.3% of total Bookings.



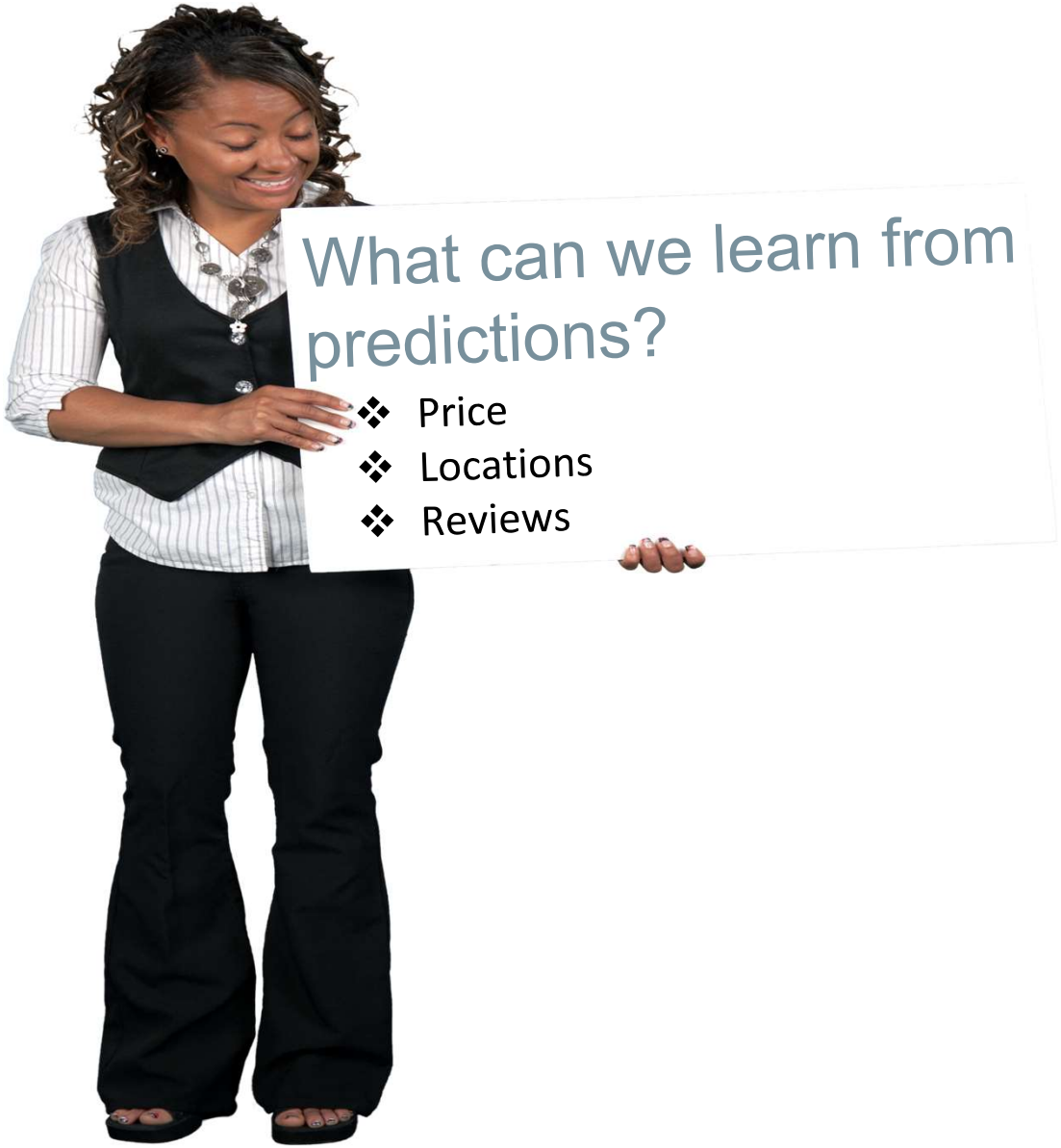
What can we learn about different hosts and areas?

neighbourhood	neighbourhood_group	Total Bookings
Williamsburg	Brooklyn	3878
Bedford-Stuyvesant	Brooklyn	3680
Harlem	Manhattan	2632
Bushwick	Brooklyn	2444
Hell's Kitchen	Manhattan	1933
Upper West Side	Manhattan	1915
East Village	Manhattan	1826
Upper East Side	Manhattan	1769
Crown Heights	Brooklyn	1552
Midtown	Manhattan	1494

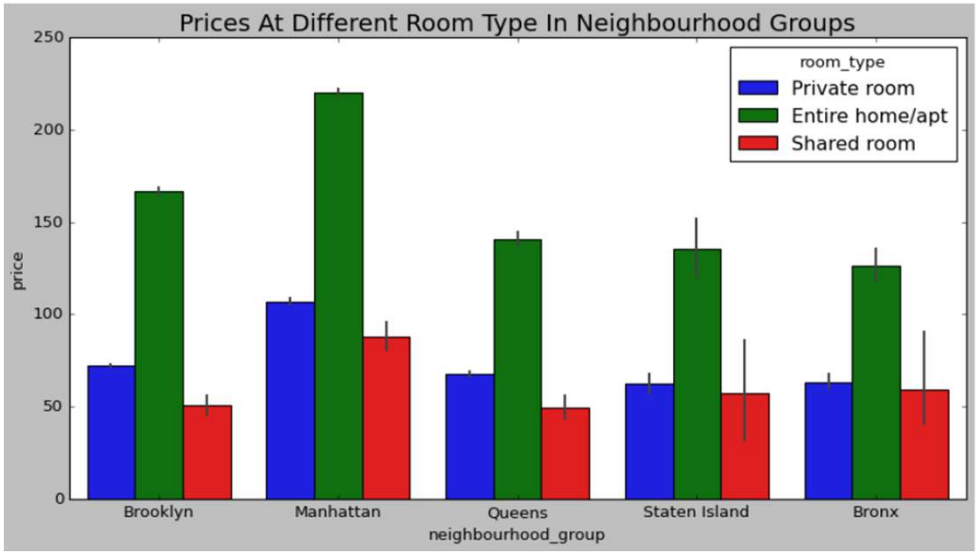


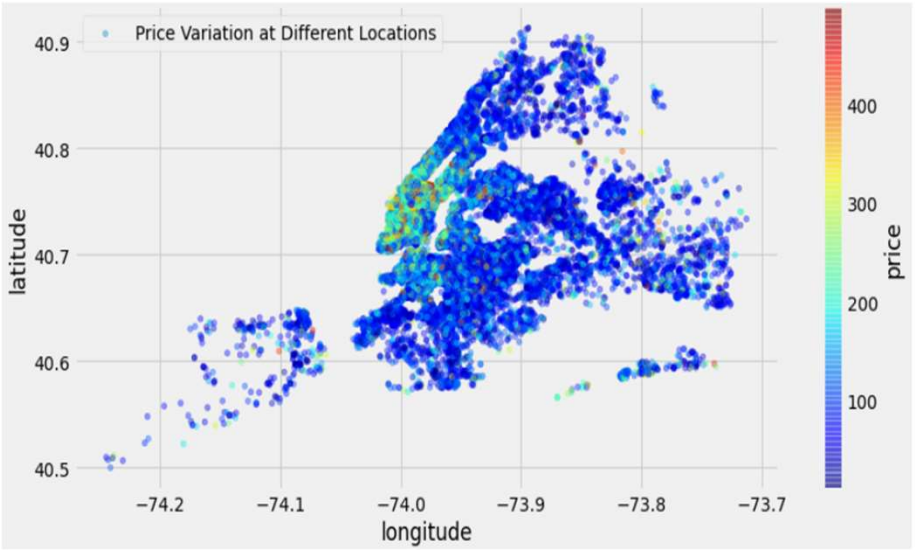
❖ **Explore the Booking Distribution in different neighborhood group.**

- Williamsburg with 3878 total number booking at the top Followed by Bedford-Stuyvesant from Brooklyn.
- Also we can see Top ten neighborhood is either from Brooklyn or Manhattan whish is obvious as together both constitute more than 85% of total bookings.


What can we learn from predictions?
Price
Locations
Reviews

What can we learn from predictions?



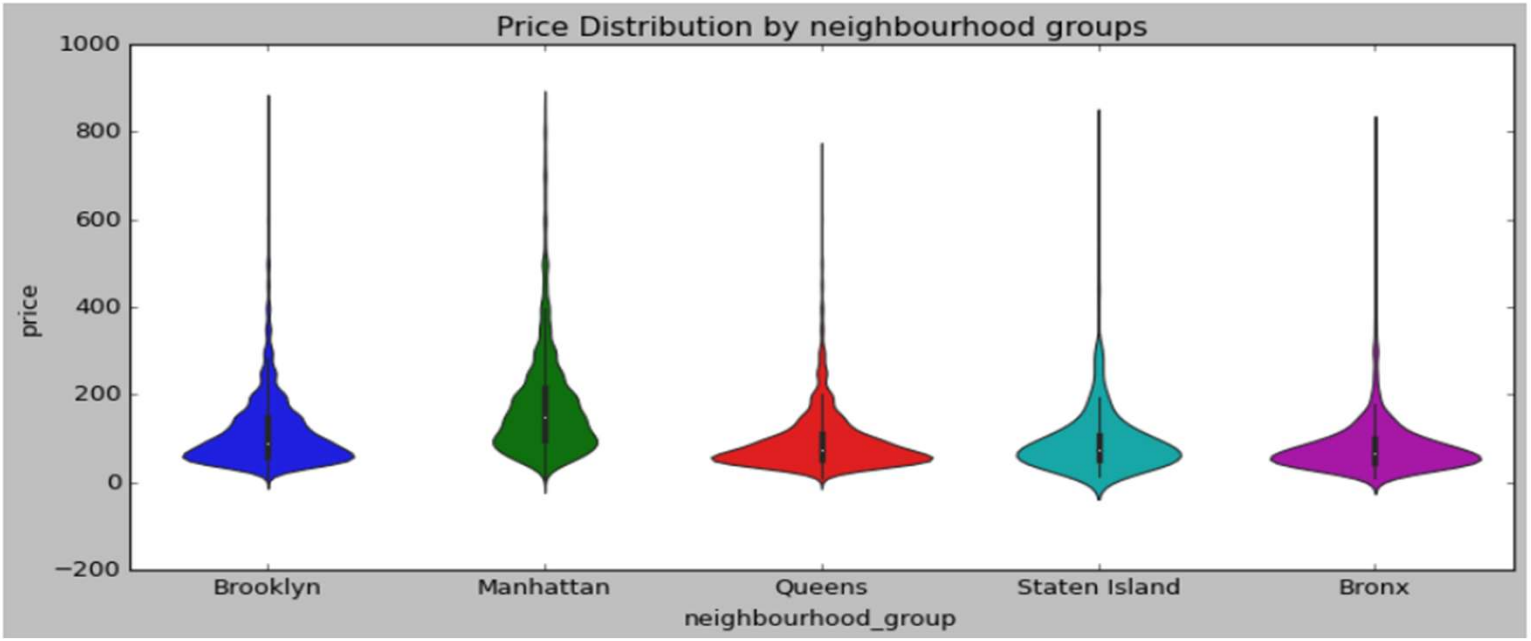


❖ **Understanding Price Distribution among different Location.**

- Scattered Plot shows the price at different location with the help of color bar we can see most of the expensive residence are in Manhattan.
- we can predict in Manhattan number of listed Entire home/Apt should be higher as we can witness from bar graph the price of Entire Home/Apt is much higher than other mode of residence in all neighborhood groups.
- Price gap between Entire home/Apt and Private or Shared rooms in Brooklyn is much higher this may be because of higher demand of Entire Home/Apt.

What can we learn from predictions?

```
plt.figure(figsize=(12,5))
ax = sns.violinplot(x="neighbourhood_group", y="price", data=df).set_title('Price Distribution by neighbourhood groups')
plt.show()
```



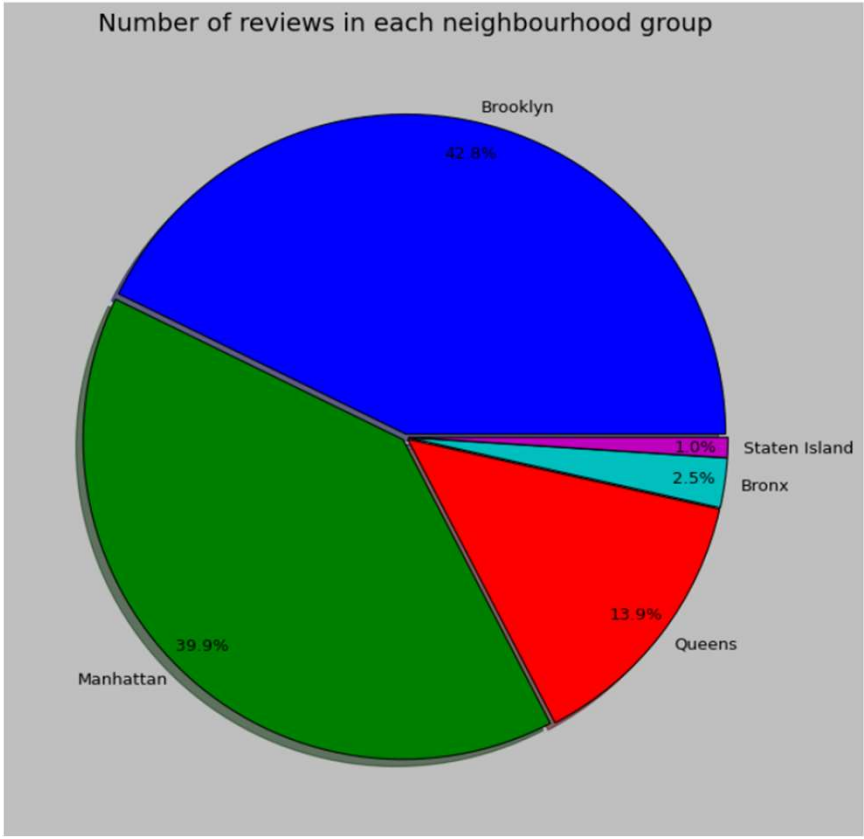
❖ **Understanding Price Distribution among different Location.**

- Using Beautiful Violin Plot we can understand and predict the price distribution in Manhattan is even compared to others neighborhood group.
- Except Manhattan in all neighborhood group price are skewed toward bottom indicating most of the bookings in these group are for shared rooms or private rooms.

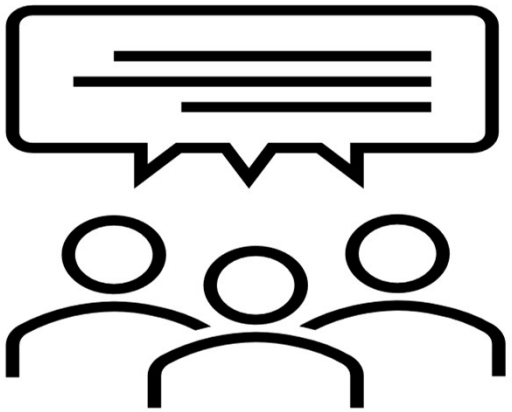
What can we learn from predictions?

- **Analyzing the total Reviews in different neighborhood group.**

- Surprisingly with 42.8% Brooklyn has the highest no of reviews with followed by Manhattan 39.9%
- Manhattan his slightly lower reviews 39.9%, compare to the percentage of total booking which is 44%.
- All Other group has higher percentage of reviews correspond to their total bookings for e.g. Queens has 13.9% reviews out of total however total booking in Queens is 11.7%, Similar for Brooklyn, Bronx and Staten Island.



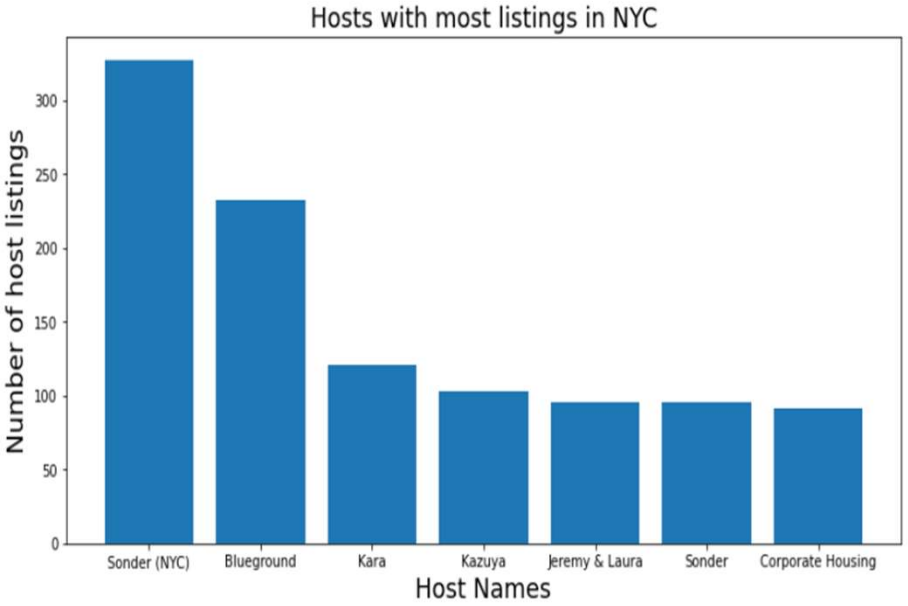
Which hosts are the busiest and why?

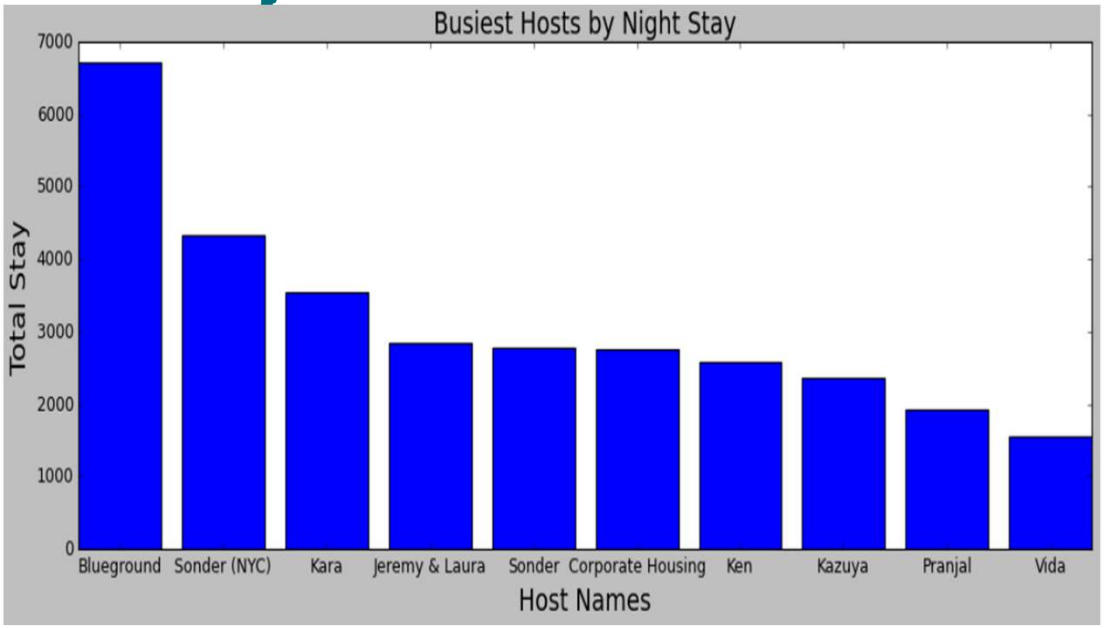


Busiest hosts would be one who have maximum number of reviews as people are booking frequently at those host's listings or One having highest total no night stay.

- Busiest Host by Total Night Stay of all customer.
- Busiest Host by No of Reviews.
- Busiest Host by total no of Bookings (Already analyze in Host analysis)

Which hosts are the busiest and why?

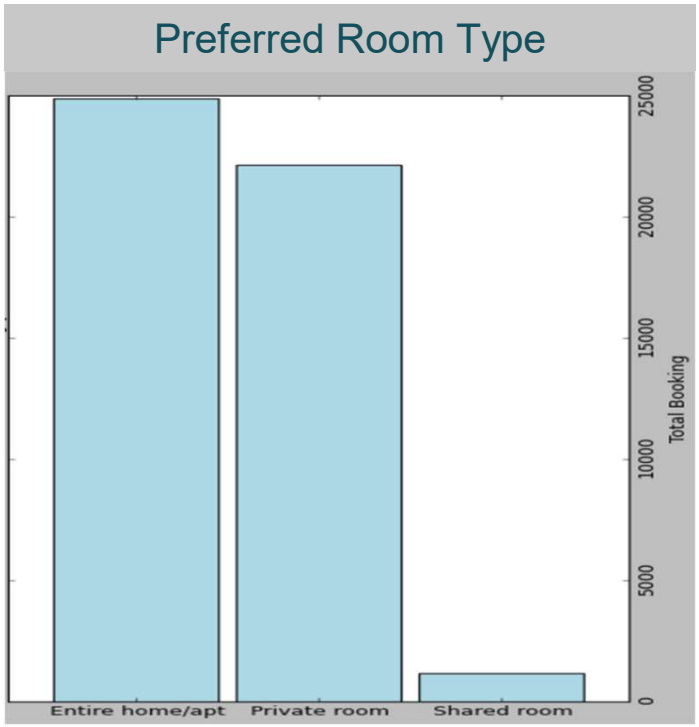




- **Busiest Host by Total Night Stays**.
- Considering total night stayed at different hosts regardless of the listed location we see Total night stayed by all customer for a Blueground is highest with 6720 nights in total followed by Sonder(NYC) with 4337 nights.
- From top host bar chart in our earlier analysis of hosts we see Sonder(NYC) at top followed by Blueground which means however number of bookings at Sonder(NYC) is higher, but customer stayed for shorter duration.

Which hosts are the busiest and why?

host_name	host_id	room_type	neighbourhood_group	minimum_nights
Blueground	107434423	Entire home/apt	Manhattan	6720
Sonder (NYC)	219517861	Entire home/apt	Manhattan	4337
Kara	30283594	Entire home/apt	Manhattan	3537
Jeremy & Laura	16098958	Entire home/apt	Manhattan	2850
Sonder	12243051	Entire home/apt	Manhattan	2784
Corporate Housing	61391963	Entire home/apt	Manhattan	2760
Ken	22541573	Entire home/apt	Manhattan	2580
Kazuya	137358866	Private room	Queens	2370
Pranjal	200380610	Entire home/apt	Manhattan	1920
Vida	7503643	Entire home/apt	Brooklyn	1560

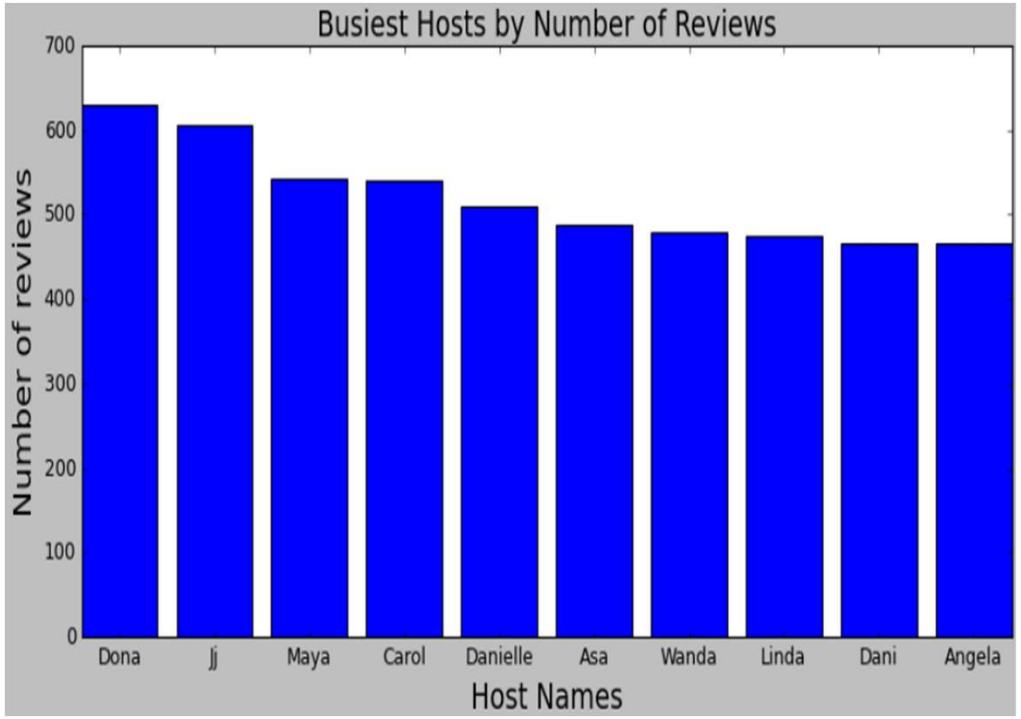


❖ **Busiest Host by Total Night Stays**.

- From Table we can see Busiest Host mostly are those listed Entire Home/Apt.
- However, Bar chart shows private rooms are also Preferred, **will see in our next analysis based on reviews**.

Which hosts are the busiest and why?

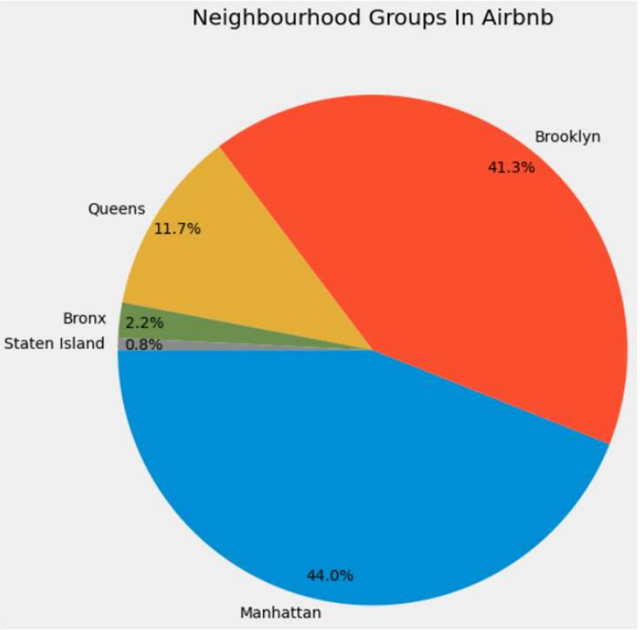
host_name	host_id	room_type	neighbourhood_group	number_of_reviews
Dona	47621202	Private room	Queens	629
Jj	4734398	Private room	Manhattan	607
Maya	37312959	Private room	Queens	543
Carol	2369681	Private room	Manhattan	540
Danielle	26432133	Private room	Queens	510
Asa	12949460	Entire home/apt	Brooklyn	488
Wanda	792159	Private room	Brooklyn	480
Linda	2680820	Private room	Queens	474
Dani	42273	Entire home/apt	Brooklyn	467
Angela	23591164	Private room	Queens	466

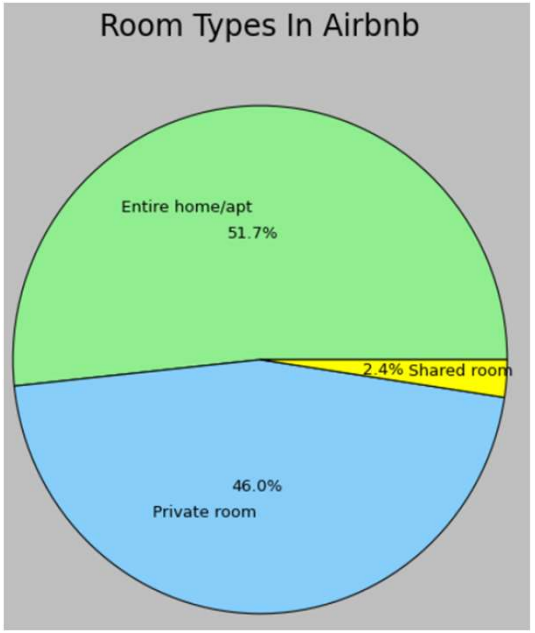


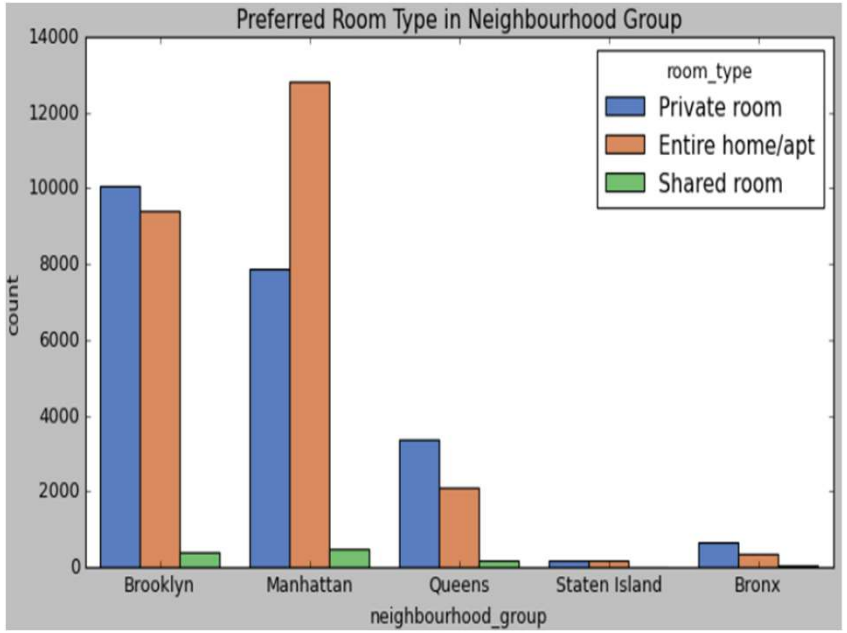
- **Busiest Host by number of Reviews**.

  - Dona has the most reviews followed by Jj and Maya.
  - So, in this analysis we can see no of reviews are higher in Private room's Hosts, hence we can conclude **Busiest Host are Mostly Private rooms or Entire home/Apt listed**.

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

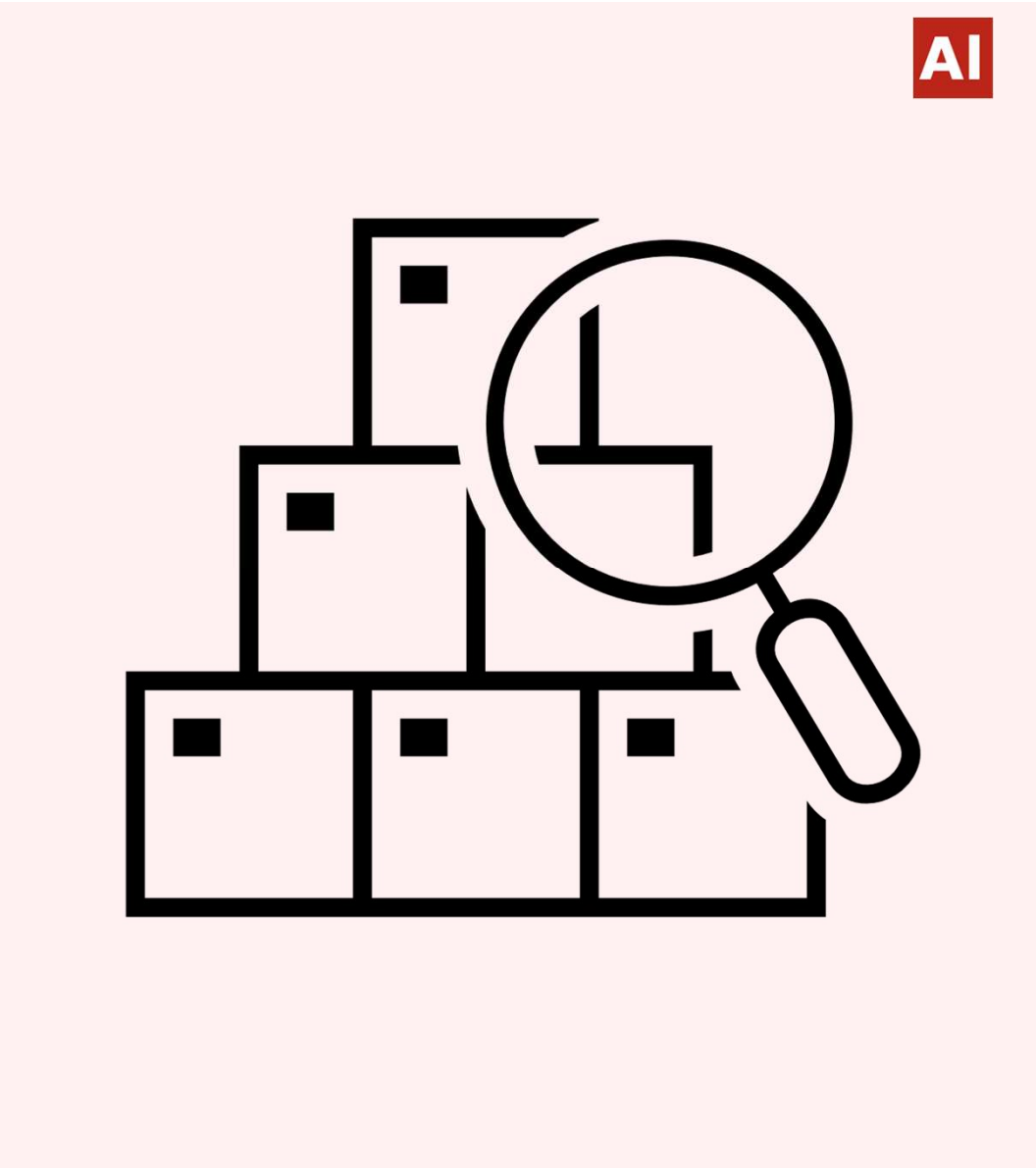






- Answer to the Question difference in traffic is obvious and can be seen in first fig that Brooklyn and Manhattan with 85.3% got most of the traffic.
- In second pie chart we analyze that with 97.7% of the bookings are in Entire room/apt or Private room and in Bar chart we can see most of the listings for Entire room/Apt and Private room are in Brooklyn and Manhattan only.
- Reason for the Difference in traffic is probably Lower listing this is due to lower demand in that area.

Key Findings & Conclusion



Key Findings & Conclusion

Key Findings

- Sonder(NYC) from Manhattan has the highest Number of Listing, followed by Blueground and Kara.
- With 44% Manhattan has the highest no of bookings compare to all other neighborhood group followed by Brooklyn with 41.3%
- Top Two neighborhood Group, Manhattan and Brooklyn contribute the 85.3% of total Bookings.
- Williamsburg with 3878 total number booking at the top Followed by Bedford-Stuyvesant from Brooklyn.
- Manhattan his slightly lower reviews 39.9%, compare to the percentage of total booking which is 44%.
- Total night stayed by all customer for a Blueground is highest with 6720 nights in total followed by Sonder(NYC) with 4337 nights.

Conclusions:

- We conclude that Manhattan is the Top neighborhood group by number of listings and highest rental prices 7 out of 10 top Host are from Manhattan followed by Brooklyn.
- One of the Probable reason for most preferred Neighbor Group is that Manhattan is a world-famous for its museums, stores, parks and theatres - and its substantial number of tourists thus attract Entire Home/Apt as favorite stay options and also stayed longer, as demand is high prices are much higher in this borough.
- At 2nd Brooklyn having significant number of listings and more affordable prices if compared to Manhattan.
- Rest 3 neighborhood groups namely Queens, Bronx and Staten island are observing very less listing options available, especially on Staten Island. Considering that those are residential areas, it is possible that many guests choose these locations to save up money or perhaps to visit family and friends who live in this area.

Thanks