**EDA Project on Airbnb NYC Dataset**

**By:-**

**Kartik Kumar**

**Submitted to:-**

**Almabetter**

# **Abstract:**

This study examined the relationship between various parameters of the AIRBNB dataset such as host id, hostname, neighbourhood group, neighbourhood, room type, price number of reviews, availability. An exploratory data analysis using field data points collected from the Airbnb listings in the metropolitan area of New York city reveals intriguing findings. The analysis helps us in understanding the most preferred hosts and neighbourhood groups by guests, the density of properties across the various neighbourhood, the number of room types belonging to each neighbourhood group, expensive neighbourhood groups, busiest hosts, preference of room types by guests, price of various room types. This analysis helps draw insights from the data and can be utilised for security, business decisions, understanding of customers and providers, behaviour and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

**Content:**

Airbnb is an online marketplace for arranging or offering lodging, primarily home stays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events; it acts as a broker, receiving commissions from each booking. The company is based in San Francisco, California, United States.The company was conceived after its founders put an air mattress in their living room, effectively turning their apartment into a bed and breakfast, to offset the prohibitive cost of rent in San Francisco; Airbnb is a shortened version of its original name, AirBedandBreakfast.com 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 entire 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.

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

**Problem Statement:**

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?

# Neighborhood Group vs. Availability Room.

* Properties in every neighborhood group.
* Map of neighborhood group.
* Price distribution in every types of room.

# **Approach:-**

# 1. Let us first check our dataset's and understand it.

# 2. Later we will check for any missing data in the data given. Does it hamper our analysis?

# 3. We would check the type of data and divide it for our analysis.

# 4**.** We checked where there any outlier or unethical data in it if so, we would filer such data for specific analysis.

# Number of quantitative variables:10 Number of qualitative variables:6

# Attributes: id, name, host\_id, host\_name, neighborhood-group, neighborhood, latitude, longitude, room type, price, minimum nights, number\_of\_reviews, last\_review, reviews\_per\_month, calculated\_host\_listings\_count, availability\_365

1. ID**:** There is a unique ID number for every entry in the dataset, with this unique ID information from the data can be easily extracted and identified
2. NAME**:** Every neighbourhood group has different hotels or renting rooms owned by the host which is termed as a name in the data frame.
3. HOST ID**:**

SameHosts may have properties in different neighbourhood groups so a

unique ID for the host is given as Host ID.

1. HOSTNAME**:**

Hosts who have listed their properties on Airbnb have a name which is termed the Hostname in the data frame

1. NEIGHBOURHOOD GROUP**:**

The name of groups of different hosts who have listed their property on Airbnb is termed a neighbourhood group.

1. NEIGHBOURHOOD**:**

Different localities of New York City are known as a neighbourhood in the data frame.

1. LATITUDE & LONGITUDE:

Latitude and longitude can be utilized to identify specific locations, which can also help identify landmarks.

1. ROOM TYPE:

Different types of rooms are available which are categorized as a private room, entire home/apartment, shared room

1. PRICE**:**

Every property listed on Airbnb has a rental price for owing over some time.

1. MINIMUM NIGHTS**:**

This data gives us information about the period of stay by guests in the hotels or renting houses

1. NUMBER OF REVIEWS**:**

Contains information on the Count of reviews given by particular guests staying at rooms

1. REVIEWS PER MONTH**:**

The count of reviews per month by every guest is stored in this column.

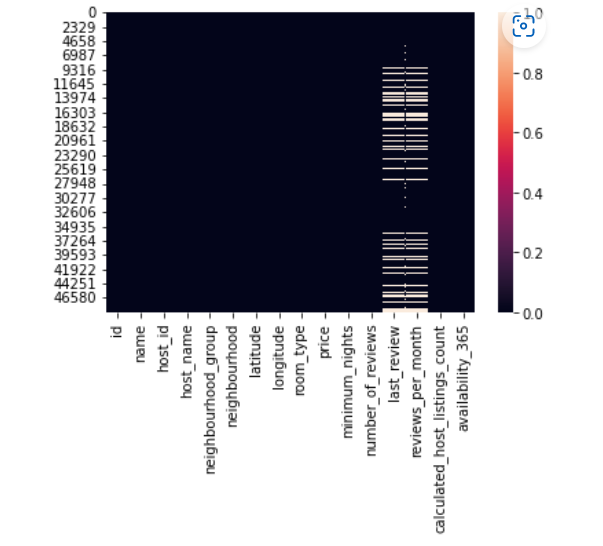
1. CALCULATED HOST LISTING COUNT:

Every host owns different properties across different neighbourhood groups and the count of this property of every host is listed.

1. AVAILABILITY 365:

This data helps us in knowing the number of days the hotel or renting a place is available in a financial year.

**Checking data for any null values**

****

Since our dataset's contain several missing values preprocessing must be done. Missing values will either be deleted or replaced with the column mean or nan. based on how important the attribute is. Also, with respect to preprocessing the datatype of certain attributes like last\_review must be changed to make processing easier. Our main goal is to analyze and find interesting patterns between the variables in our dataset’s. Visualization is an important aspect of finding patterns. Hence several visualization techniques like bar graph, pie chart, Violin chart, correlation, etc. will be plotted to gain insights. We then plan to predict certain variables such as price by using predictive models. Several models will be explored and models with the best accuracy will be selected. We also got that

Few columns like name, host name, last review had many missing values and then we replaced it with “missing”. Importance for analysis, hence they were deleted.

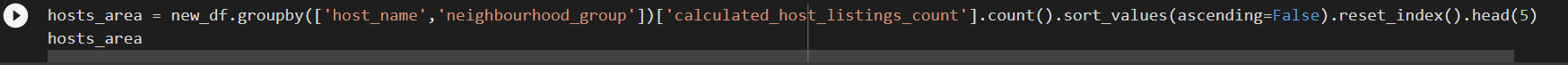
Reviews per month column had lot of missing rows but is important for analysis, hence missing values were replaced with the mean of that column.

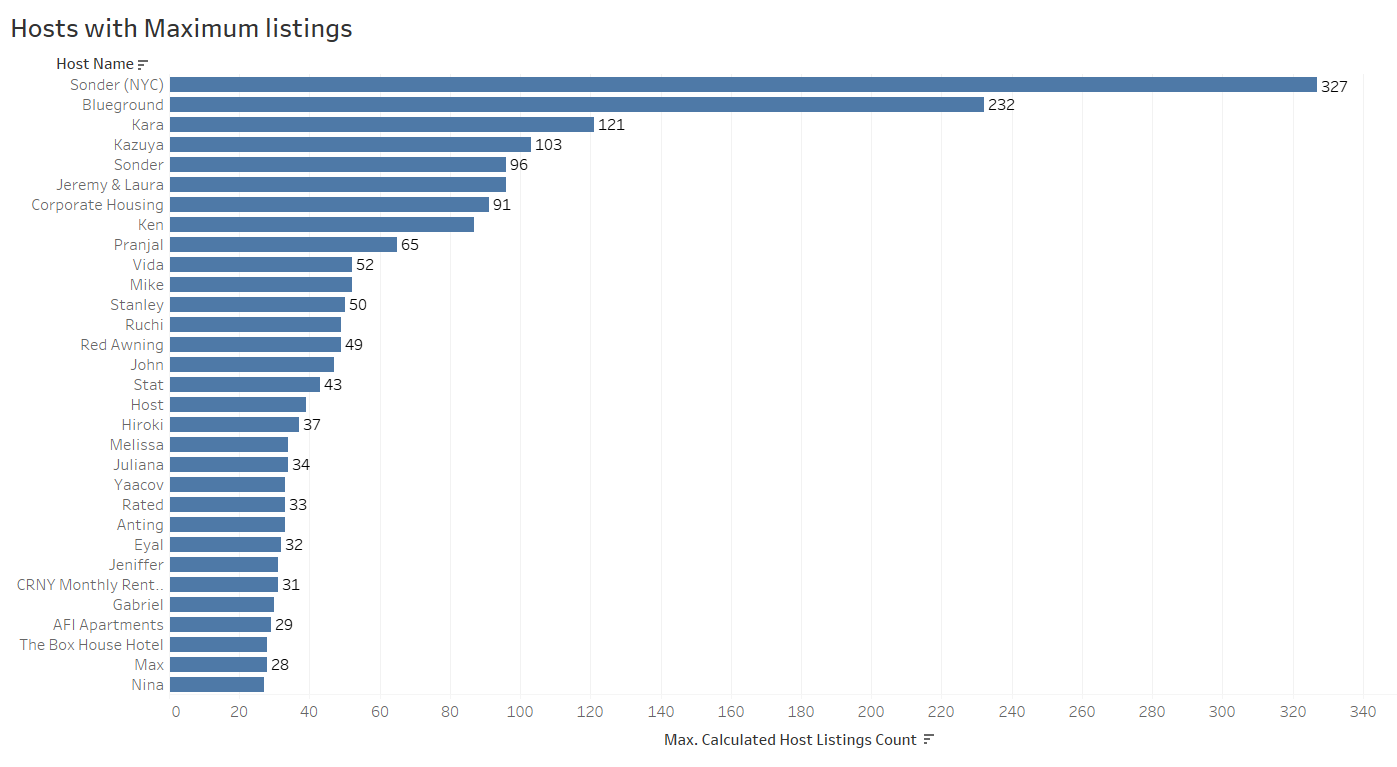
**EXPLORATORY DATA ANALYSIS**

Now we know that we are ready for an exploration of our data, we can make a rule that we are going to be working from left to right. The reason some may prefer to do this is due to its set approach - some datasets have a substantial number of attributes; plus, this way we will remember to explore each column individually to make sure we learn as much as we can about our dataset's.

# **Observation-1:-** What can we learn about different hosts and areas?

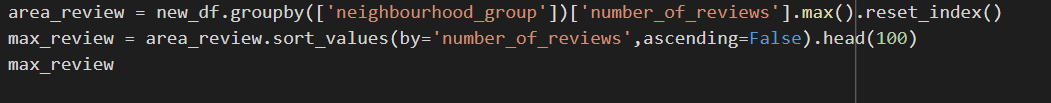
As shown in the adjacent bar chart, we can see there is a good distribution among the top 6 hosts. The host named Sonder (NYC) has highest number of listings of 327 in Manhattan neighbourhood group, The host named Blueground has second highest listings of 232 in Manhattan neighbourhood group.

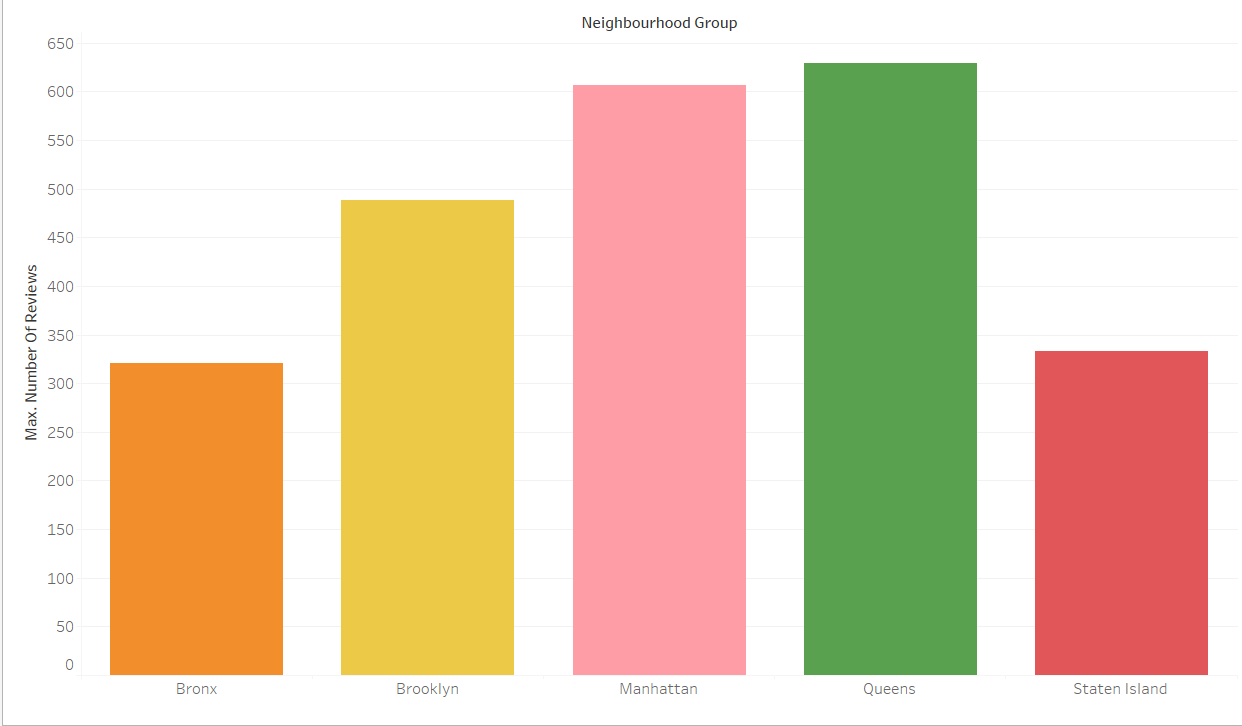




**Observation 2:-**What can we learn from predictions?

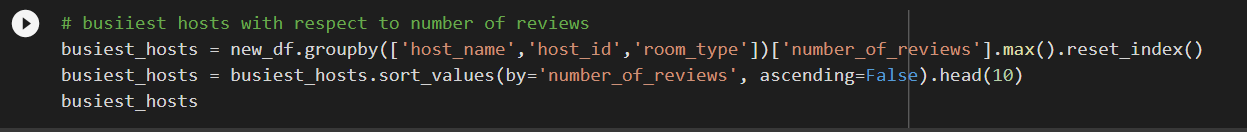
* The number of reviews feature in the dataset represent the customers who have given the reviews to a particular property they have stayed in
* Looking at the bar chart, Queens has maximum number of reviews.
* Manhattan has second highest number of reviews.
* Bronx and Staten island has least number of reviews.

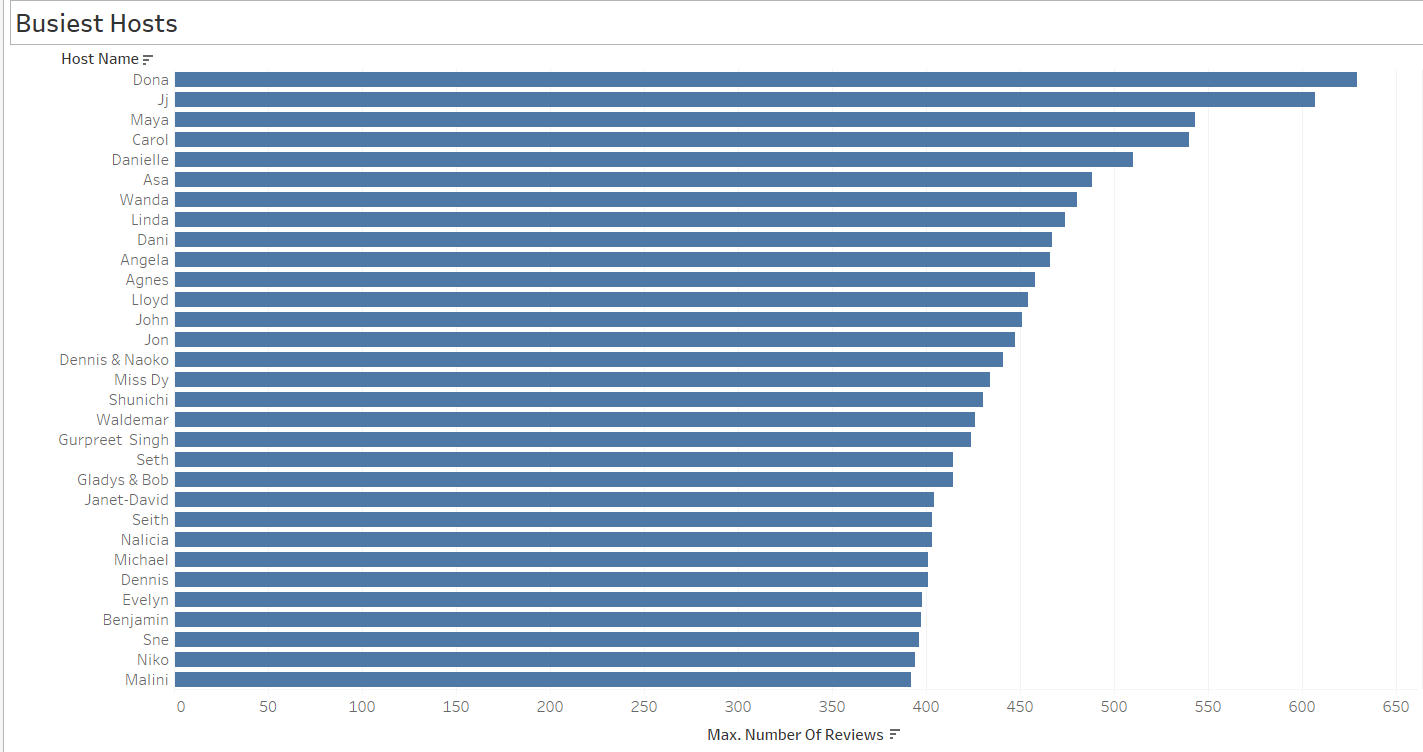




**Observation 3:-**Which hosts are the busiest and why?

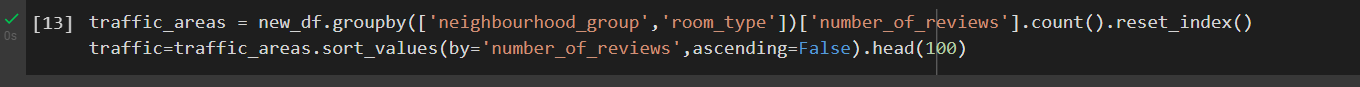
* The adjacent bar plot shows the top 30 hosts with to number of reviews.
* Among them Dona has the highest number of reviews and we can assume on the basis of number of reviews that Dona is popular and is the busiest host .
* The top hosts have listed private room, entire home/apartment.

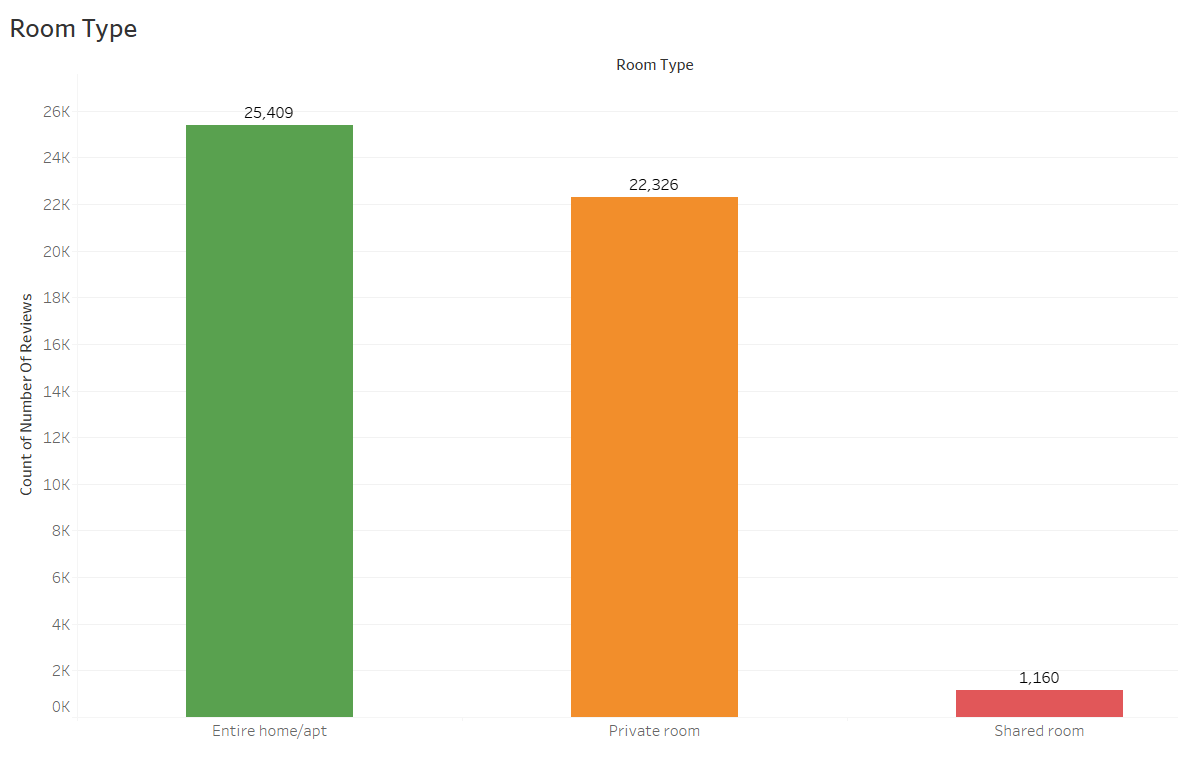




**Observation 4:-** Is there any noticeable difference of traffic among different areas and what could be the reason for it?

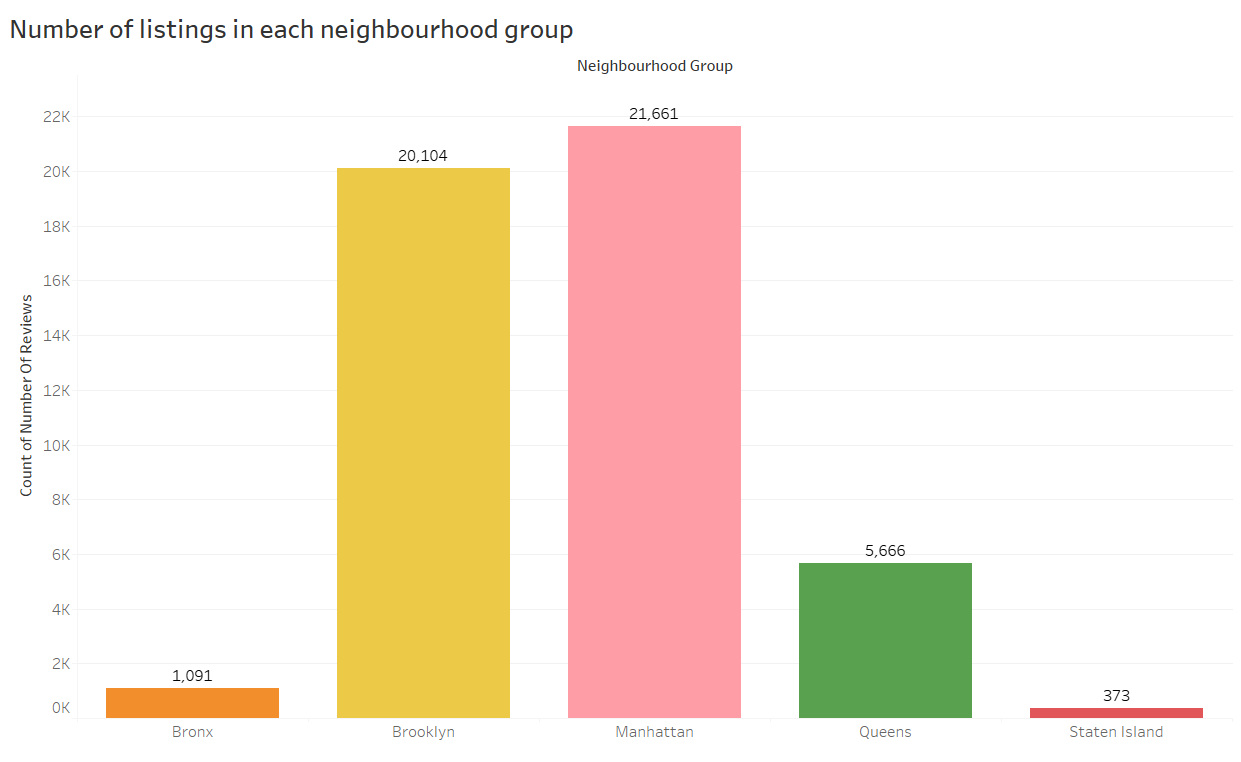
Most people prefer Entire home/apt in Manhattan or Brooklyn and the other prefer private rooms in Manhattan or Brooklyn so the traffic for both these areas is much high in comparison to Bronx and Staten Island. The entire home/apt and private rooms have a very demand with respect to minimum nights or number of reviews.



****

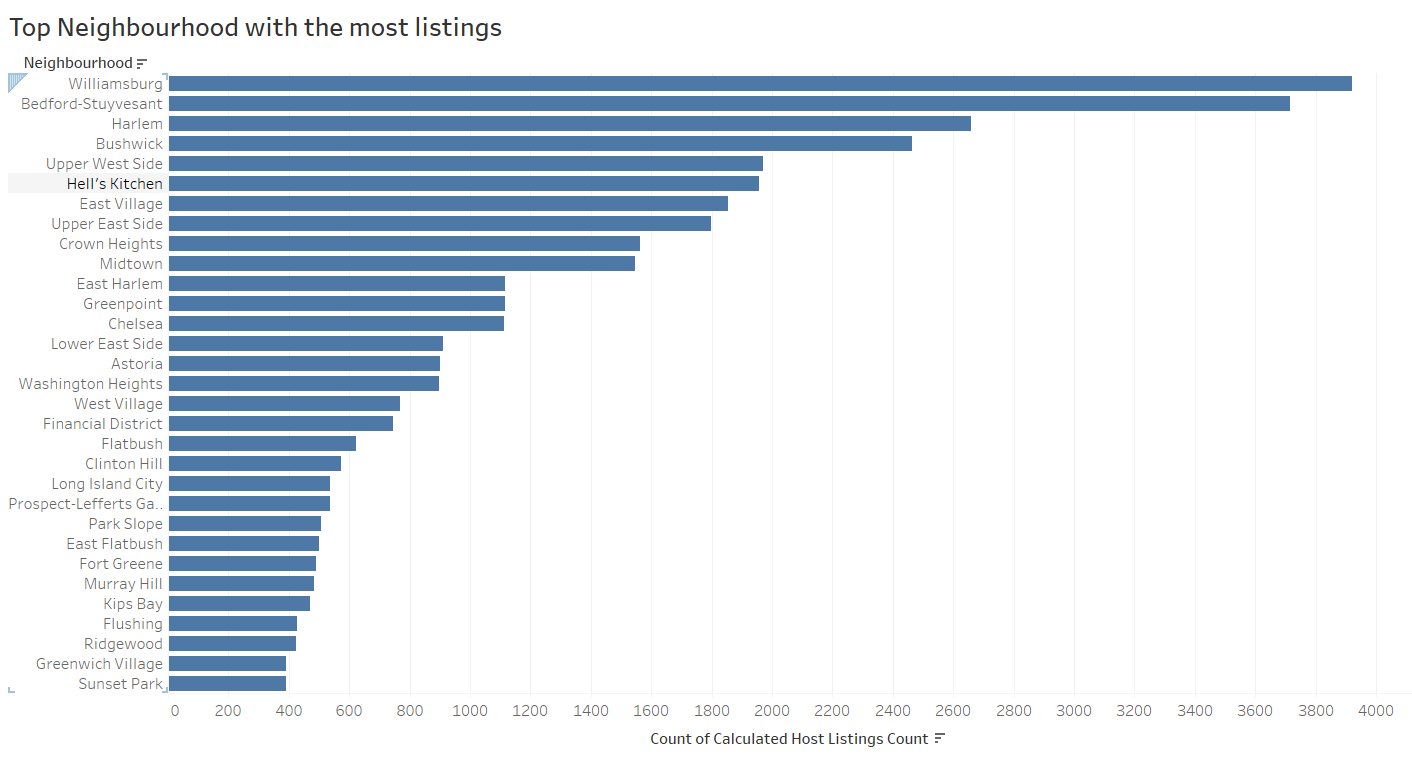
**Observation-5:-**Which Neighbourhood Group has most number of listings?

* I have done this analysis using Tableau public and It is observed that the Manhattan has highest number of listings of 21,661 which is 44.3% of total listings done on Airbnb.
* Brooklyn has second highest number of listings of 20,104 which is 41.13% of the total listings.
* Queen comes in third place with 5,666 listings whereas Bronx and Staten Island have least number of listings.



**Observation-6**:- Which Neighbourhood has most number of listings?

* From the various Neighbourhoods top neighbourhood are listed in the chart.
* Williamsburg has the highest number of listing which is around 3900 and it is located in Manhattan.

****

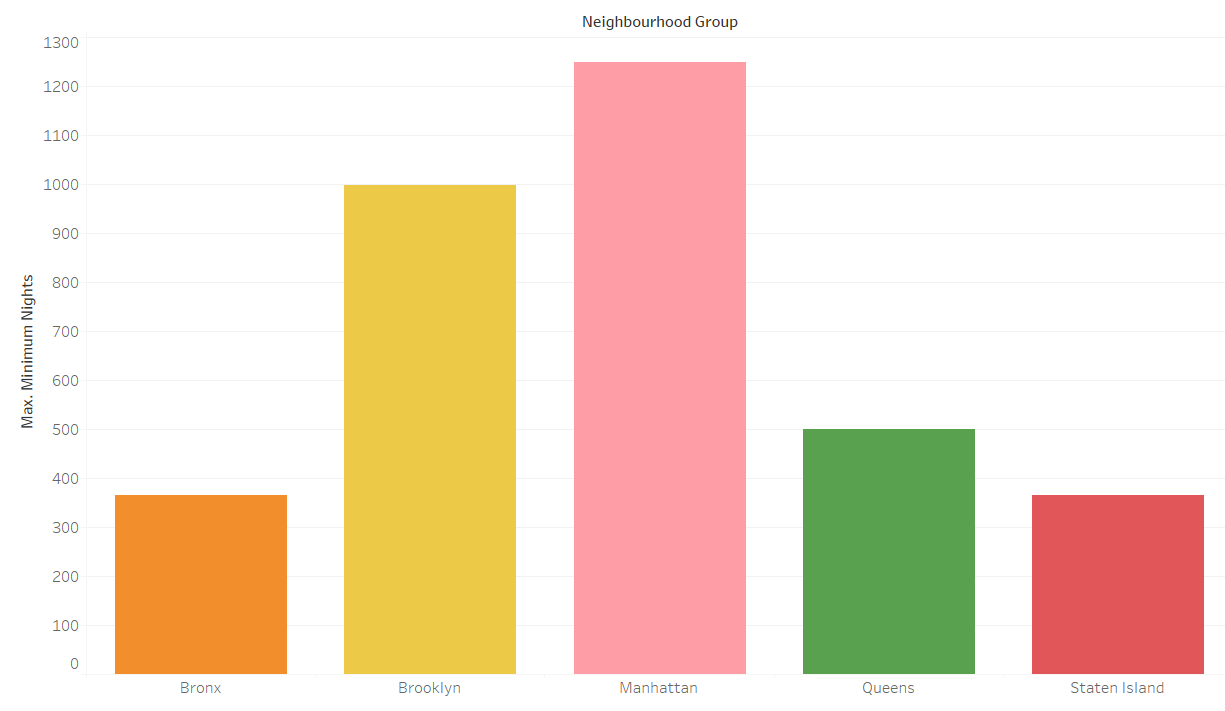
**Observation-6:-**Which among the all neighbourhood group has highest availability?

* We can infer that the availability for Manhattan is less we can assume that the tourist prebook Entire homes and private rooms so that they do not have to face any issues of availability
* The availability for Brooklyn is much higher in comparison to Manhattan.
* Bronx has the lowest availability among all and Staten Island has the highest availability among all.
* Queen has also less availability so we can assume that it also has its prebooking.



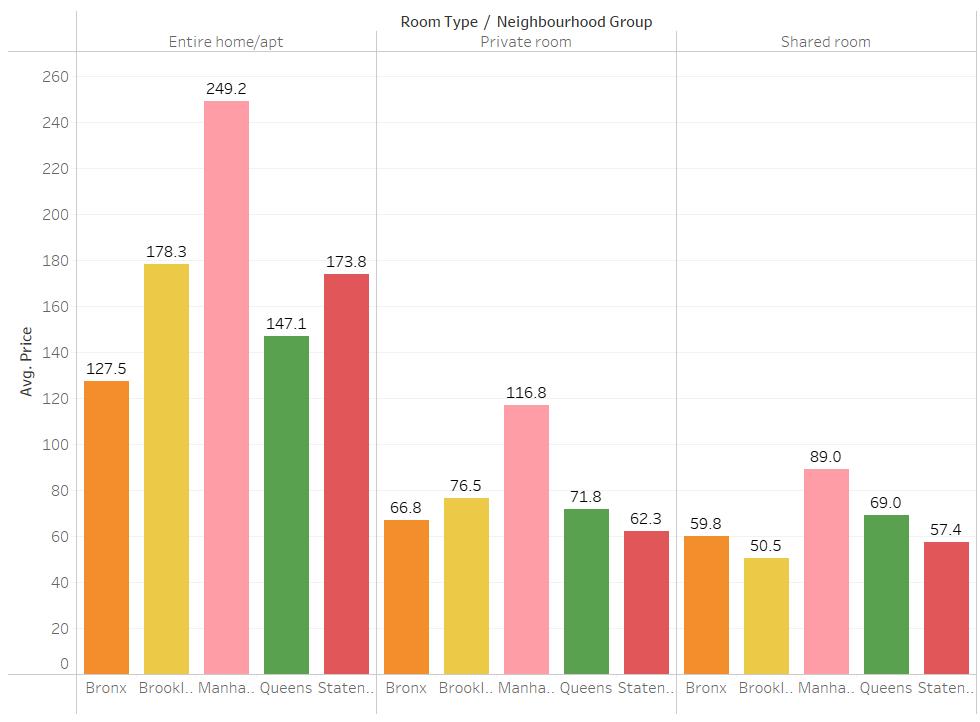
**Observation-7:-**In which neighbourhood groups maximum number of minimum nights is spent?

* The maximum number of minimum nights spent is in Manhattan and then Brooklyn.
* Manhattan has much more number of customers overall and then Brooklyn among all.
* The maximum number of minimum nights tells us the overall demand for different neighbourhood group



**Observation-8:-**What are the prices for different room types in for each neighbourhood group?

* We can easily view that the prices for the Entire home/apt is very much high in comparison to the other two room types.
* For Entire home/apt Manhattan has the highest average price following by Brooklyn and Staten Island.
* For private room the prices are lesser in comparision to entire home/apt and also the prices for Manhattan is highest for Private room aswell.



Private rooms on average are priced from 60-120 dollars per night on an average depending upon the neighborhood group it is located.

Most of the shared rooms have the price range between 50-90 dollars per night depending upon the neighborhood groups.

Most of the Entire home/apt are priced between 130to 250 dollars.

**Observation-9**:-What can we from latitudes and longitude from our dataset?

* From the latitude and longitude of different neighbourhood groups we have analysed that Manhattan , Brooklyn , Queens are at not so far from each other and that might be the reason that these three neighbourhood groups have been on highlight during our analysis .
* So I can analyse that if a person is going to either of these places can easily visit other two aswell.



**Problems Faced:-**

* Doing the capstone project alone was hectic task as all the work has to be done by me alone.
* Understanding the data and how to draw conclusions was major task to be done on the dataset

**Conclusion:**

* Manhattan has the 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 neighbourhoods groups by people.
* Queens has significantly less host listings in Manhattan. So we should take enough steps to encourage Host listings in Queens as there are decent demand in neighbourhood of Queens.
* The maximum demand is for Entire home/apartment and Private rooms.
* People are most interested in cheaper rentals.
* The top 10 neighbourhood with most listing are located either in Manhattan or Brooklyn with Harlem and Williamsburg presenting leading numbers in each borough , respectively.

# **References:-**

1.<https://www.almabetter.com/>(notes)

2. https://stackoverflow.com/questions/214741/wha t-is-a-stackoverflowerror

3. <https://startuptalky.com/airbnb-success-story/>

**Remarks:-**

**These write up is a documentation work done by Kartik Kumar**