**Airbnb Dataset Exploratory Data Analysis**

**Gaurang Bagga, Sushant Tripathi, Prerna Dave, J. Sai Naveen**

**Data Science Trainees @Almabetter**

**1. About Airbnb:**

Airbnb is an online marketplace connecting travellers with local hosts. On one side, the platform enables people to list their available space and earn extra income in the form of rent. On the other, Airbnb enables travellers to book unique homestays from local hosts, saving them money and giving them a chance to interact with locals. Catering to the on-demand travel industry, Airbnb is present in over 190 countries across the world.

**2. How it works?**

* Hosts list out their property details on Airbnb along with other factors like pricing, amenities provided etc.
* Airbnb sends a professional photographer (if available) to the property location in order to take high quality photographs.
* Travelers search for a property in the city where they wish to stay and browse available options according to price, amenities etc.
* Booking is made through Airbnb where traveller pays the amount mentioned by host and some additional money as transaction charges.
* Host approves the booking. Traveller stays there and finally Airbnb pays the amount to the host after deducting their commission.

**3. Overview of Our Airbnb Dataset:**

Our Airbnb dataset consists mainly the NYC’s five boroughs data which are Manhattan, Brooklyn, Queens, Bronx and Staten Island. The shape of our Airbnb dataset has initially 48895 records and 16 features in it.

The dataset is a mixture of Categorical and Numerical features. Each record in a dataset signifies a particular Airbnb listing and the features associated to it like its id, host, price, reviews etc. By looking at the head of a dataset we can get an overall idea of its columns and their description.

**4. Description of columns:**

* **id** - The id of the listing.
* **name** - The name of the listing.
* **host\_id** - This is to signify the id of host who are associated to the respective listing.
* **host\_name** - The name of the host.
* **neighbourhood\_group** - This column is signifying the five boroughs of NYC.
* **neighbourhood** - The neighbourhood of the respective neighbourhood groups.
* **Latitude** & **longitude** - The geographical co-ordinates of respective listings.
* **room\_types** - Listing space type.
* **price** - Price of listing in US dollars.
* **minimum\_nights** - This is the amount of nights minimum.
* **availability\_365** - The number of days when listing is available for booking within a year.
* **calculated\_host\_listings\_count** - The amount of listing per host.
* **number\_of\_reviews** - Total number of reviews.
* **reviews\_per\_month** - Average of reviews per month.
* **last\_review** - Date of the last review recorded.

**5. Steps Involved in our exploratory data analysis of Airbnb NYC dataset:**

**5.1 Getting the overview of our dataset**

The first and foremost thing for the analysis of any data is to get the idea and overview of how our data looks like, what are its features, what is its shape, and determining the data types of the columns. Initially we have set the name of our dataset to ‘df’.

**Shape** - The shape of our dataset was - (48895, 16)

**Structure** - To get the basic structure of our dataset we have simply called head function to print the first five rows of our dataset like ‘df.head()’. By doing this we will be able to visualize the structure of the dataset and many times it gives a lot of information about the presence of null values, categorical columns, numerical columns etc.

**Information of columns, their data types and presence of null values** - We will get the information about all the features running the code ‘df.info()’. It will give summarize data types and presence of any null count in features.

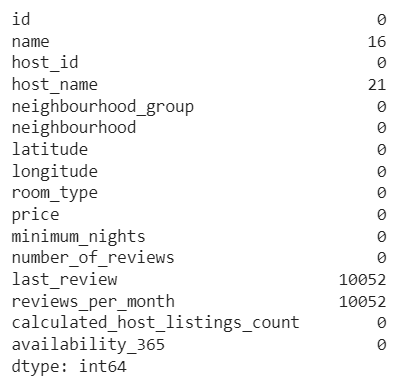


**5.2 Handling the null values**

From the isnull().any() method we have come to know about whether our data set consists of any null values or not in the form of Boolean.



Now to get the information about the number of null values present in our data set we can use isnull().sum() method.



Now from above two figures we can understand that ‘last\_review’ and ‘reviews\_per\_months’ columns are having the greater number of null values. Hence, I have removed these two columns from my dataset.

We can also see that ‘name’ and ‘host\_name’ columns are having the lesser number of null

values. Hence, I have replaced the null values with ‘Not Available’.

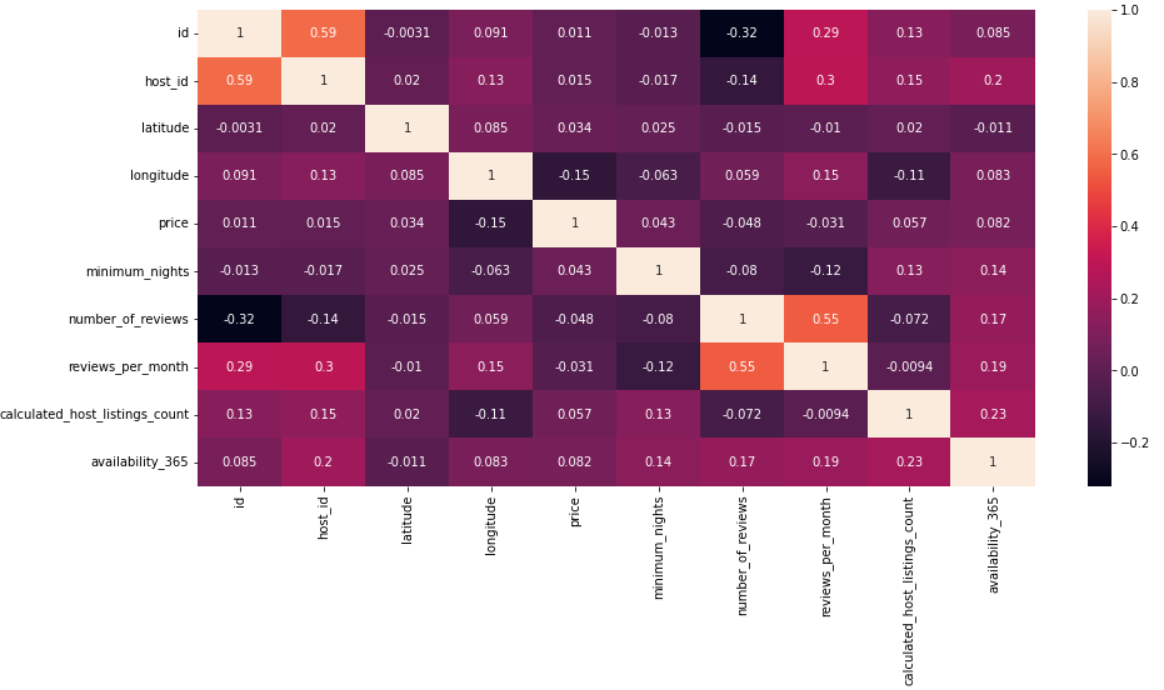
**5.3 Techniques used to draw information about our dataset**

Once the dataset set is prepared according to our need, we will start to dug out the information from it. We have used following techniques to draw information about our dataset:

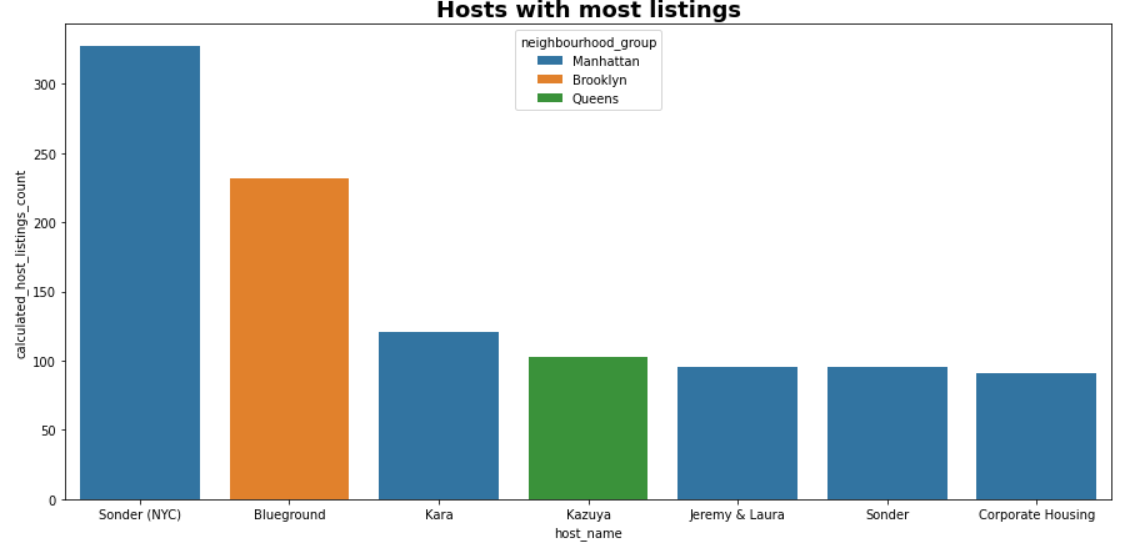
* Correlation Heatmap
* Univariate Analysis
* Bivariate Analysis
* **Correlation Heatmap** - A correlation heatmap is a heatmap that shows a 2D correlation matrix between two discrete dimensions, using coloured cells to represent data from usually a monochromatic scale The colour of the cell is proportional to the number of measurements that match the dimensional value. This makes correlation heatmaps ideal for data analysis since it makes patterns easily readable and highlights the differences and variation in the same data.
* **Univariate Analysis -** Univariate analysis is a basic kind of analysis technique for statistical data. Here the data contains just one variable and does not have to deal with the relationship of a cause and effect. It is basically the simplest form to analyze data. Uni means one and this means that the data has only one kind of variable. The major reason for univariate analysis is to use the data to describe. The analysis will take data, summaries it, and then find some pattern in the data.
* **Bivariate Analysis** - Bivariate analysis means the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values. It usually involves the variables X and Y. Univariate analysis is the analysis of one (“uni”) variable.

**6. Data visualization by exploring the dataset:**

**6.1 Verifying the correlation of different features in the data**

****

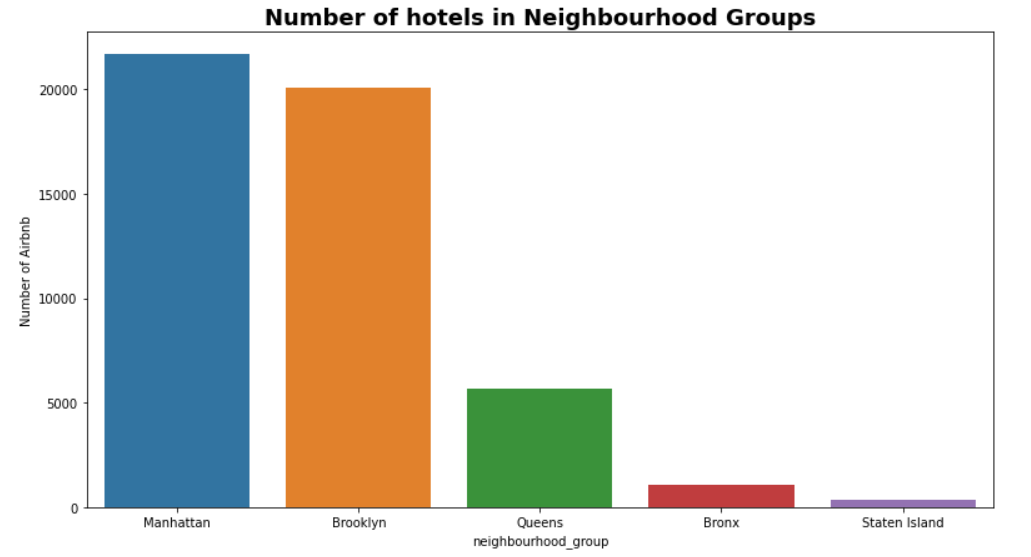
**6.2 Hosts with most listings**



**Observation:**

From the above bar plot, we can conclude that the name of the host with the most listings is Sonder (NYC) who has 327 listings in Manhattan neighbourhood group.

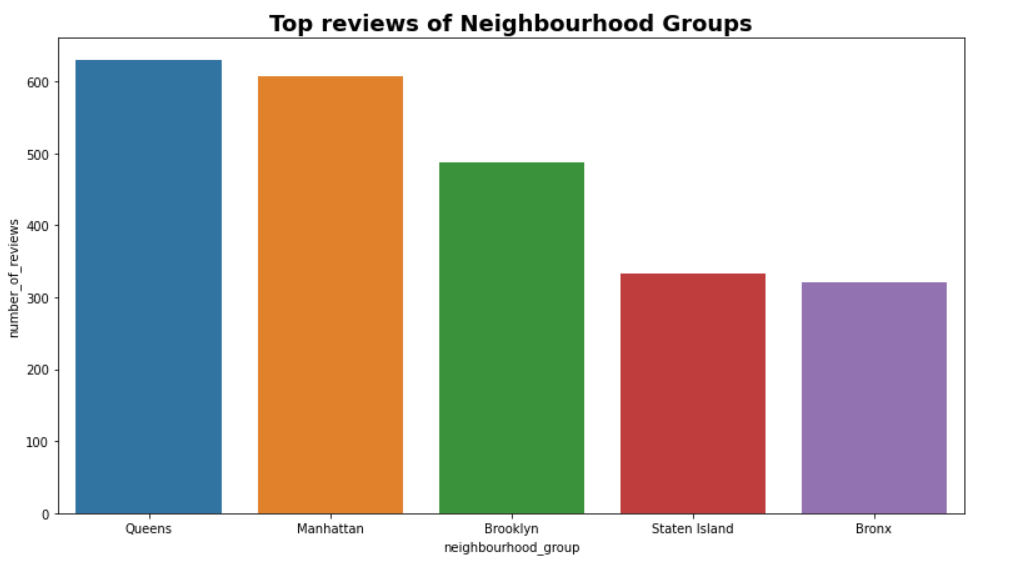
**6.3 Number of hotels in Neighbourhood Groups**

****

**Observation:**

From the above bar plot, we can conclude that the name of the neighbourhood group with most number of listings is Manhattan who has 21661 hotels and the name of the neighbourhood group with least number of listings is Staten Island who has 373 hotels.

**6.4 Top reviews of Neighbourhood Groups**

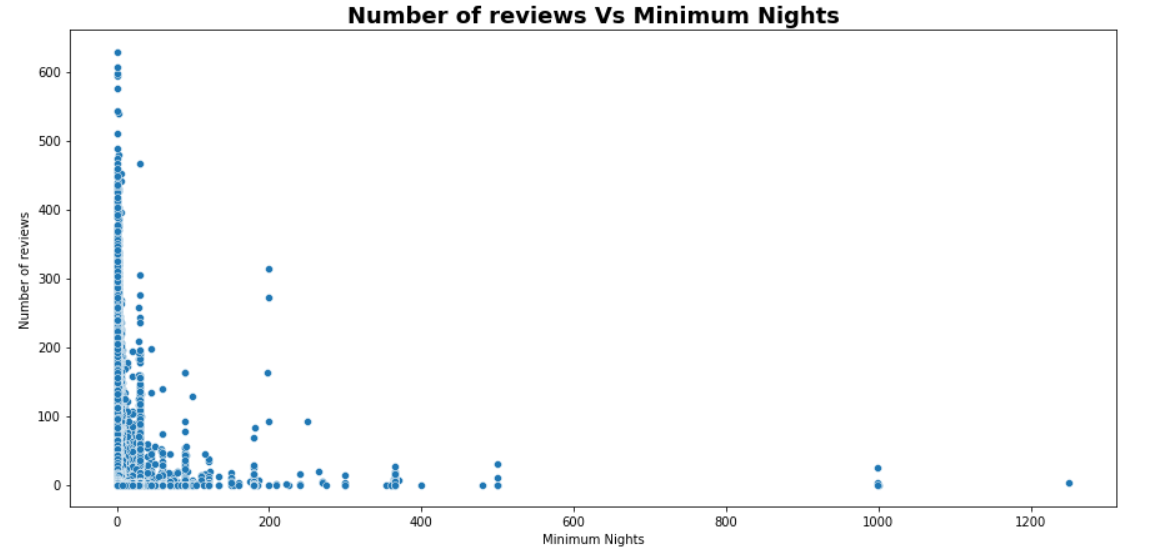


**Observation:**

From the above result, we can conclude that the name of the neighbourhood group with the most number of reviews is Queens who has 629 reviews and the neighbourhood group with the least

number of reviews is Bronx who has 321 reviews.

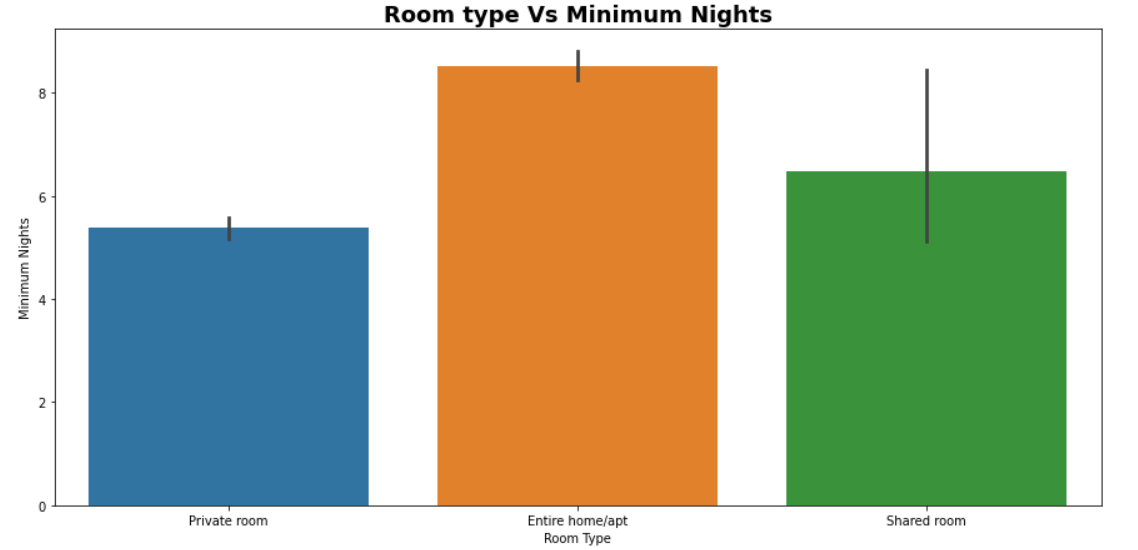
**6.5 Number of reviews Vs Minimum Nights**

****

**Observation:**

From the above plot, we can conclude that the customers used to give reviews to the rooms where they come for short duration like Private rooms and Shared rooms. As the number of night stay increases, customers don’t use to give reviews to them.

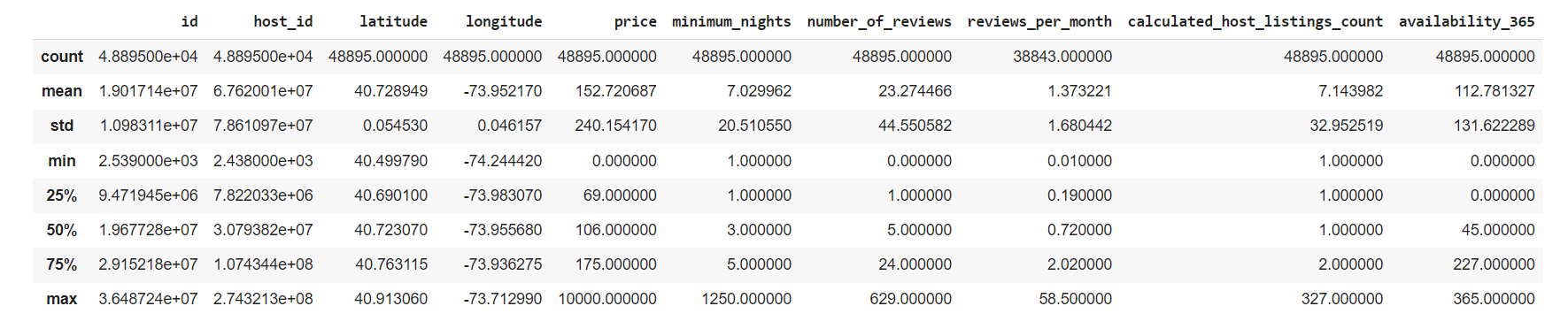
**6.6 Room type Vs Minimum Nights**



**Observation:**

From the above plot, we can conclude that for minimum night stay customers used to take Private rooms whereas for longer night stay customers used to take Entire home/apt.

**6.7 Checking the descriptive statistics of dataset**

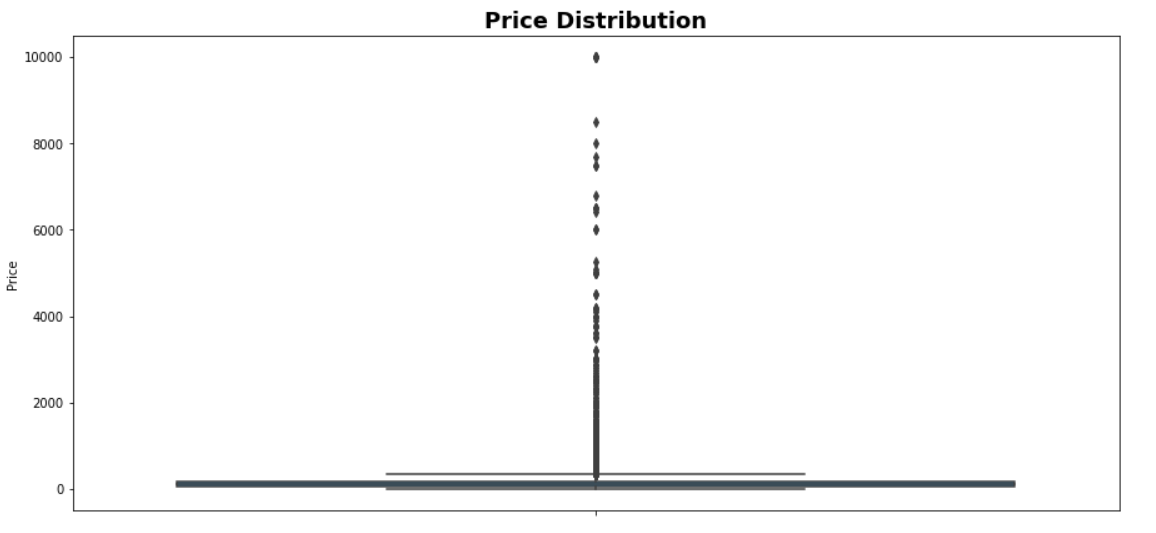
****

**Observation:**

The above summary statistics shows us that the average price is 152.72, the average minimum night’s stay is 7.03 nights, and the average number of reviews is 23.27 per listing. We also learn that a host has an average of 7.14 places listed and availability averages is 112.78 vacant days per year.

**Note:** Most importantly, the min price is showing as 0 and the max price as 10,000. Something isn't right with the data, so need to look into this issue and check for outliers.

**6.8 box plot before the outliers are removed**



**Observation:**

As expected, there are outliers present in the data. In order to fix this, we used the quantile-based flooring and capping. I

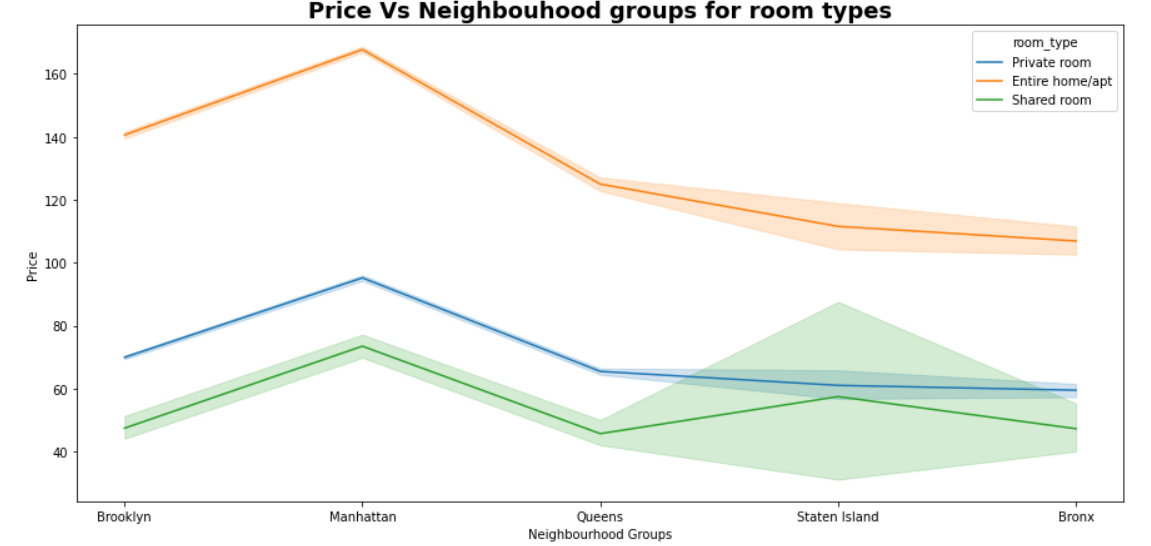
**6.9 box plot after the outliers are removed**



**Observation:**

As expected, the price has is reduced.

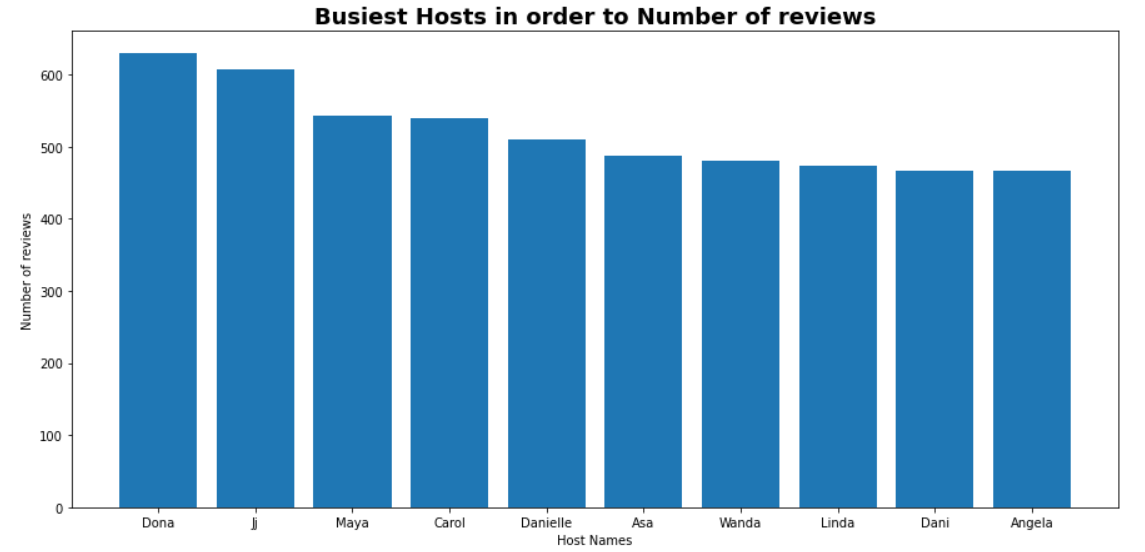
**6.10 Price Vs Neighbourhood groups for room types**



**Observation:**

By looking at the above line plot, we can see that Manhattan has a higher price range and is the most expensive one. Brooklyn has the second-highest rental prices, while the Bronx appears as the most affordable one.

**6.11 Busiest Hosts in order to Number of reviews**

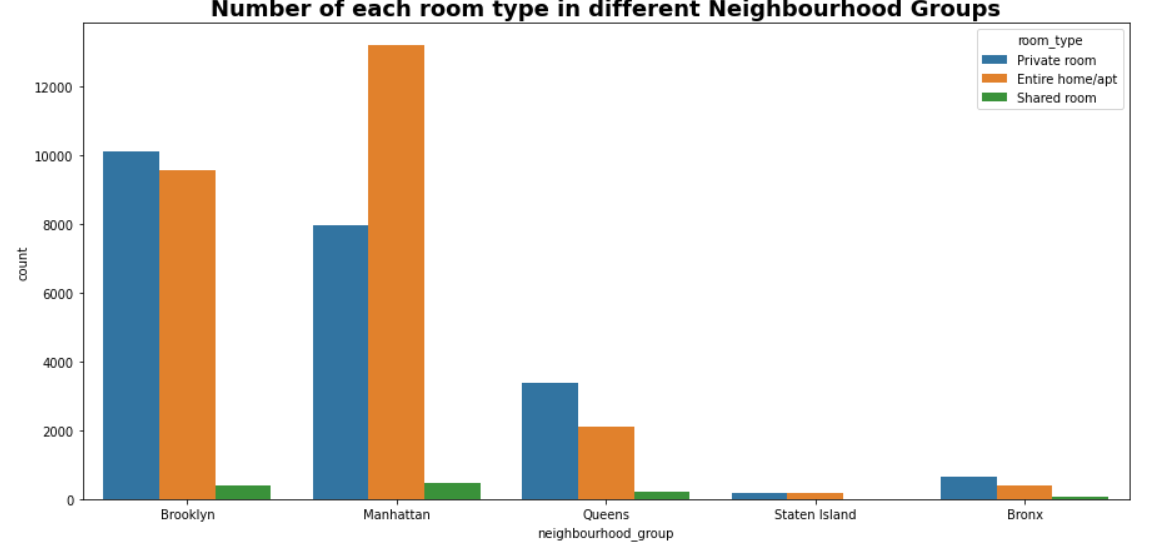


**Observation:**

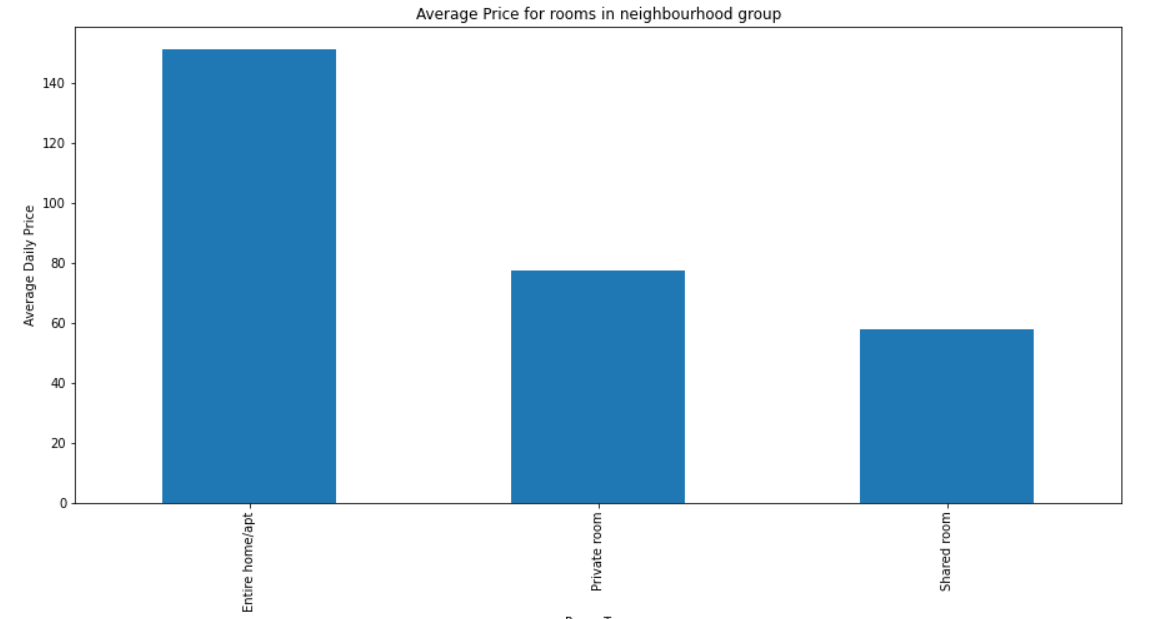
From the above result, we can conclude that the Name of the busiest host is Dona with 629 number of reviews.

**Note:** The busiest host is the person who got the most number of reviews i.e., people are frequently booking rooms at these hosts.

**6.12 Number of each room type in different Neighbourhood Groups**



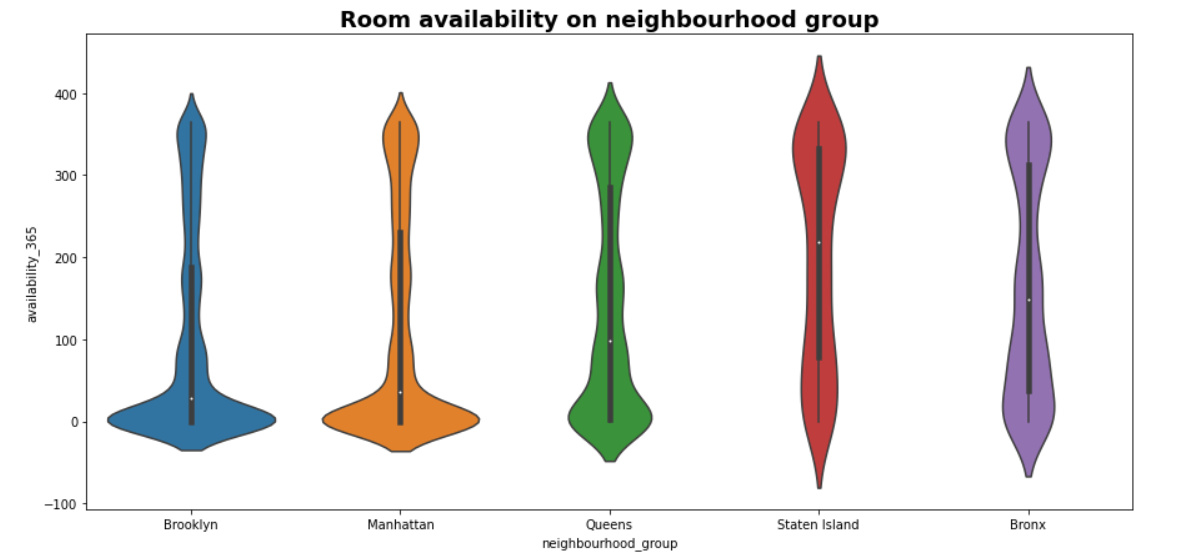
**6.13 Average Price for rooms in neighbourhood group**



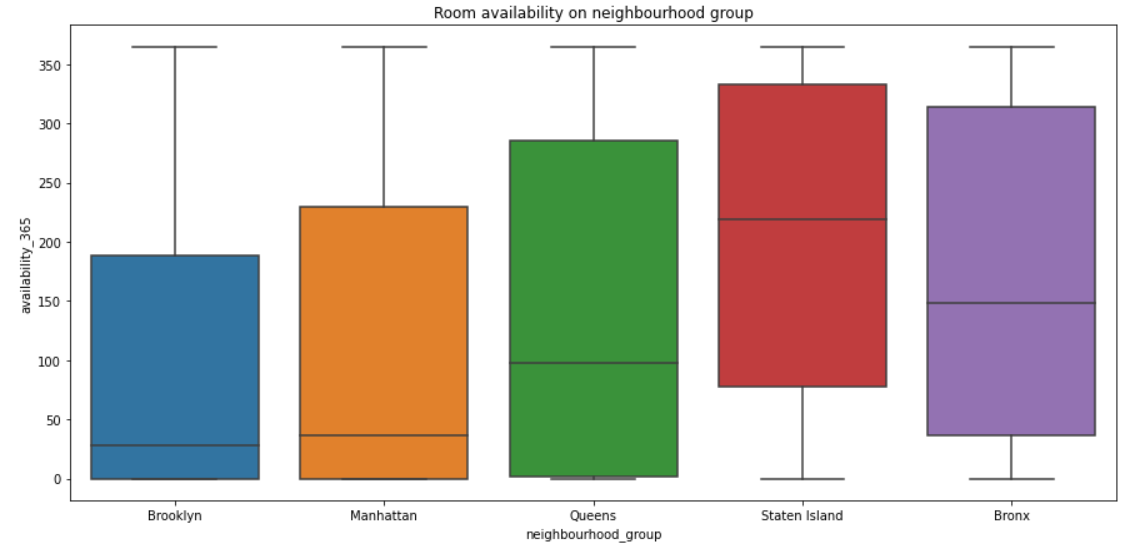
**Observation:**

From the above result, we can observe that the Average daily price for Entire home/apt is 151 USD, Private room is 78 USD and Shared room is 58 USD.

**6.14 Room availability on neighbourhood group with violin-plot**



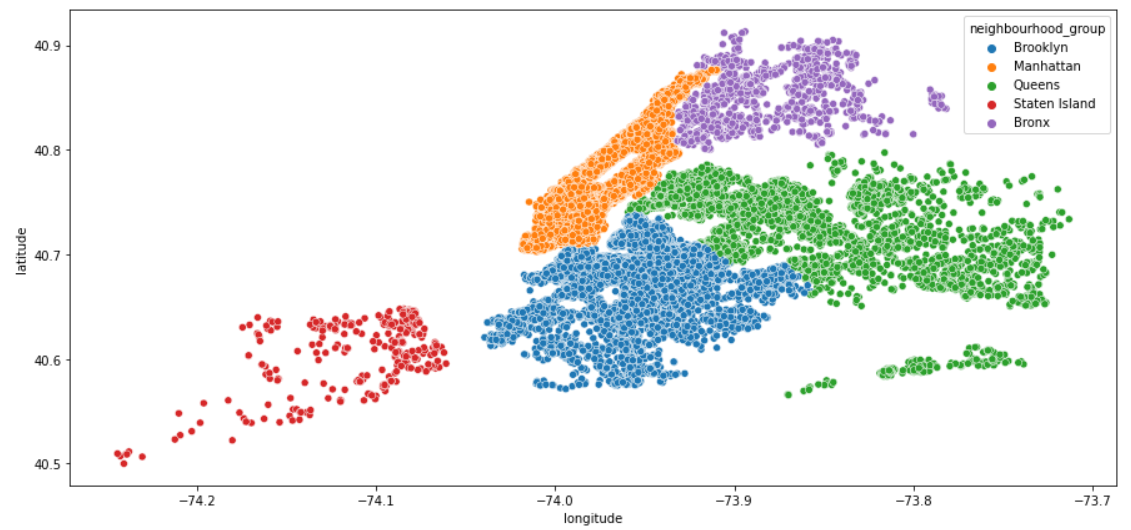
**6.15 Room availability on neighbourhood group with box-plot**

****

**Observation:**

From the above result, the neighbourhood group Manhattan and Brooklyn is having the less average of room availability and Staten Island is having the high average of room availability.

**6.16 Visualizing the traffic among different neighbourhood group**



**Observation:**

From the above two plots, we can conclude that the **Manhattan** and **Brooklyn** neighbourhood groups are having the most traffic because the room availability is throughout the year with all price ranges i.e., from expensive ones to the nominal ones. **Staten Island**neighbourhood group is having the least traffic because of less number of hotels.

**7. Graphs and Plots used for visualization:**

Following are the ways though which we have conducted this analysis:

* **Box Plot** – It is used to identify outliers in a particular feature and get the 5-point summary of the selected feature which is min, max, median, 1st and 3rd quartile.
* **Count Plot** - It is similar as bar plot. We used it to count number of appearances of a particular value of a feature.
* **Pie Plot** - It is used to show the distribution percentage of data points of a categorical column.
* **Scatter Plot** - It is used to show the relation between features like location and price, location and reviews, rooms and price etc.
* **Line Plot** - It is used to organize the data in a simple way and is very easy to interpret.
* **Bar Plot** - It is used to compare the items between different groups over time.
* **Violin Plot** - It is used to visualize the distribution of numerical data. It also depicts the summary statistics and the density of each variable.
* **Heatmap Plot** - It is used to show relationships between two variables, one plotted on each axis.

**8. Findings from the Airbnb Data Set Analysis:**

* Manhattan is the most focused place in New York since the competition among the hosts and number of hotels is higher here. However, Staten Island is the least focused place in New York since the competition among the hosts and number of hotels is lower here.
* Queens got the greater number of reviews which shows that the facilities provided by the hosts are good and customers are satisfied with them. However, Bronx got the smaller number of reviews which shows that the hosts need to improve their facilities to satisfy the customers.
* From share percentage of reviews for each neighbourhood group we can say that Queens and Manhattan have majority of reviews which implies that customers love staying in these neighbourhoods. Also, from these neighbourhoods almost half of the total revenue is generated for the Airbnb company.
* Customers pay highest amount in Brooklyn, Queens and Manhattan that is 10000 USD and lowest amount is 10 USD since 0 cannot be the genuine price because no one will provide room for free.
* Manhattan has a higher price range of rooms and is the most expensive neighbourhood group. Brooklyn has the second highest rental prices, while the Bronx appears as the most affordable one.
* Jamaica turns out to be a most popular neighbourhood since customers staying here are more satisfied.
* As Entire home/apt are generally preferred by a guest who would have booked a listing for long stay and are costlier than Private rooms and shared rooms. That's why most of the guest would go for other room types options i.e., Private room which are less costly than Entire home/apt.
* The share of Entire home/apt and Private rooms in New York city is almost 98% which shows that the customers want some privacy and that is why they are opting for these rooms.
* Shared rooms are for customers who don't have any privacy concerns or they don't have enough amount to pay for Entire home or Private rooms because the price of these rooms is comparably higher than the shared rooms.
* The most number of rooms is present in Williamsburg neighborhood located in Brooklyn which concludes that customers are more likely to stay in this neighborhood.
* Manhattan and Brooklyn are having the less average of room availability which indicates that customers mostly used to stay in these neighborhood groups only. However, Staten Island is having the high average of room availability which indicates that the customers are often comes to this neighbourhood group.

**9. Conclusion:**

That's it! We reached the end of our exercise. Starting from loading the raw data to get it cleaned and free from any non-uniformity and null values and getting the desired data for analysis was a challenge. In this process we have eliminated the features that are not required, removed outliers to get an unbiased result.

We have also used different visualization techniques, graphs and plots to get a clearer picture of what is preferred most by the guest and what they are looking for in Airbnb.

We got to know that people are preferring the locations that are the most popular and have numerous attractions the most. Rest of the other important insights of Airbnb dataset in mentioned in point 8.

**References-**

1. Almabetter Class Notes
2. GeeksforGeeks
3. Analytics Vidhya
4. GitHub