**Airbnb Bookings Analysis**

**Sunil R M, Rohan Solanke, Rishikesh Dubey,**

**Meghana RS, Prabhujeet Kaur**

**Data science trainees,**

**Alma Better, Bangalore**

**Abstract:**

Airbnb is an American home rental platform based in San Francisco that lets people list, find, and rent short-term lodging in 65,000 cities and more than 191 countries across the globe.

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.

**1. Problem Statement**

Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Explore and analyze the data to discover the key factors responsible for bookings.

**2. Introduction**

Airbnb is an online platform where you can find homestays/rentals for vacation and other tourism activities. They have listings in more than 220 countries and regions.

People search the hotels according to their needs depending upon their nearest localities, tourist places, reviews, rating, and room types.

We have a dataset of observation around 49000. Here we have customer reviews, the neighborhood of the hotels, and prices.

We will follow below roadmap before we deep dive into the solutions.

* + Loading the data
  + Cleaning the data
  + Exploratory data analysis
  + Conclusion

* **Loading the data**

We created a directorial path for the Airbnb dataset, using the Pandas read function we read it. It has a shape (48895, 16) which means it has 48895-row labels and 16 features or column labels.

After reviewing the data we can see that we have information on multiple hotels of different **locations, prices, reviews, etc.**

* **Cleaning the data**

We can see that there are multiple data which doesn’t contain any value, the data which doesn’t contain any value is called null value.

Our dataset has multiple null values which we have to replace so it doesn’t affect our accuracy. We are replacing them using fill na method so that we can get a better result.

## **3. Some Keywords**

There are some keywords we will be using throughout this discussion.

1. **IRQ(Inter-Quartile range):** The IQR describes the **middle 50% of values when ordered from lowest to highest**. To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.
2. **Outliers**: Outliers are defined as abnormal values in a dataset that don’t go with the regular distribution and have the potential to significantly distort any regression model. Therefore, outliers must be carefully handled in order to get the right insight from the data.
3. **Kurtosis:** It is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.
4. **Skewness:** It is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.
5. **Matplotlib, Seaborn, Plotly:** We used the above-mentioned three powerful libraries for visualization.

**4. Steps involved:**

**4.1 Exploratory Data Analysis**

Following are the observations using Exploratory Data Analysis and visualization.

**4.1.1 Hosts with maximum number of entries: Plotly line plot:**

We took the host\_name feature to find the top host with maximum number of entries in the data set by finding the value count of each host. Then we visualized it using Plotly line plot.

According to the analysis, Micheal has the maximum value count with 417 entries/rows in data set followed by David, Sonder, John, Alex so on. This visualization gives the top 10 busiest hosts in the provided Airbnb bookings data set.

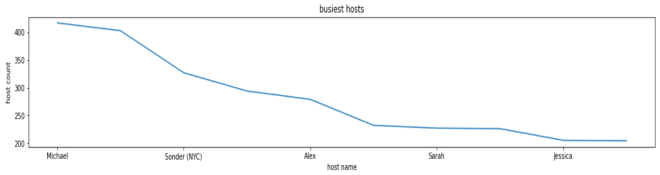


Fig 1: busiest host

**4.1.2 Calculated host listings count vs host name – Plotly bar plot**

We have taken the host name and calculated host listings count to analyze the relation between the two said features by using Plots bar plot. The visualization shows that the calculated host listings count feature basically tells the number of times that particular host has used Airbnb bookings in the data set.

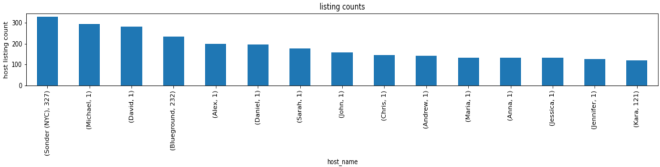
If calculated host listings count is 6 then that particular host has exactly 6 rows in the dataset. In the bar plot below, host Sonder has listings count of 327 which means he has 327 rows in the data set, it represents total number of listings made by a specific host. 

Fig 2: listing counts

**4.1.3 Different types of room available vs the number of rooms -- Plotly bar plot:**

There are three different types of rooms available for opting while renting a room under Airbnb bookings according to the data set. This can be confirmed by finding the unique contents of room type feature.

The below bar plot provides the visualization of three types of rooms along with the number of rooms named-

Entire home/apartment, shared room and private room. There are around 25000 of

Entire home/apartments entries, around 22000 of private rooms and around 1000s of shared rooms entries in the data set.

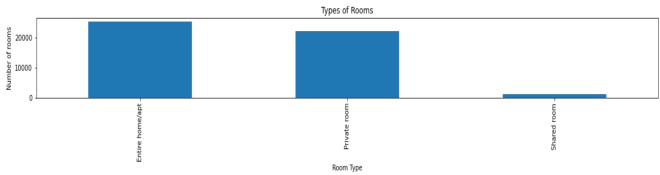


Fig 3: types of rooms

**4.1.4 average of review per month vs room type—Plotly line plot:**

We grouped by room type and took mean as aggregation operations on review per month feature, we calculated review per month mean for each room type, and then we visualized that using Plotline line plot.

The average review distribution for entire home/apartments is around 1.3, for private room and shared room is around 1.4. This means that there is a vast variation in the reviews given per month for each room type.

Your room reviews will affect its chances of being booked. Rooms with 3 stars or lower will not be encouraged by the people to book the room. Negative room reviews combined with a poor rating will hurt your rooms rank, but great room reviews and high ratings will help increase your bookings of rooms.

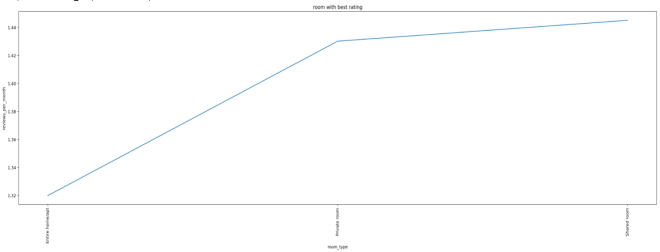


Fig 4: room type with best rating

**4.1.5 Correlation Between features Seaborn Heatmap:**

A heat map (or heatmap) is a [data visualization](https://en.wikipedia.org/wiki/Data_visualization) technique that shows the magnitude of a phenomenon as color in two dimensions.

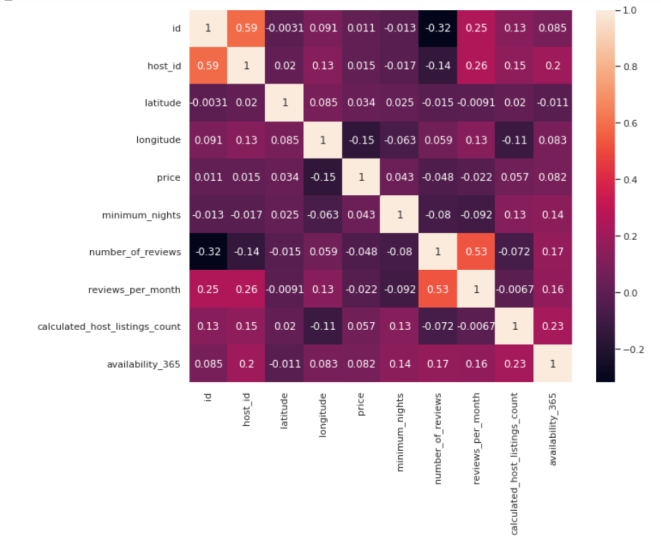


Fig 5: Correlation matrix

We used Airbnb data. corr () method to find out the correlation between features. There are three two types of correlation. There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation.

A **positive correlation** is a relationship between two variables in which both

variables move in the same direction. Therefore, when one variable increases as the other variable increases or one variable decreases while the other decreases.

A **negative correlation** is a relationship between two variables in which an increase in one variable is associated with a decrease in the other.

A **zero correlation** exists when there is no relationship between two variables.

By plotting seaborn heatmap correlation we got to know that there features positively correlated with each other, among which reviews per month and number of reviews are highly correlated. There is 53% of chance that number of reviews increases by reviews per month of Airbnb bookings data.

host\_id is correlated to reveiws\_per\_month & availability\_365. There is a correlation between calculated\_listings\_count, minimum\_nights and availability\_365.

**4.1.6Average price of top 10 most reviewed listings in NYC:**

We used number of reviews feature to find the average price of top 10 most reviewed listings in Airbnb bookings data set.

From the below table output, we can observe that the top 10 most reviewed listings on Airbnb bookings for NYC has average price of 65 dollars. Most of the listings has average price below 50 dollars. 9 out of 10 listings are private rooms type. The top reviewed listing has 629 reviews.

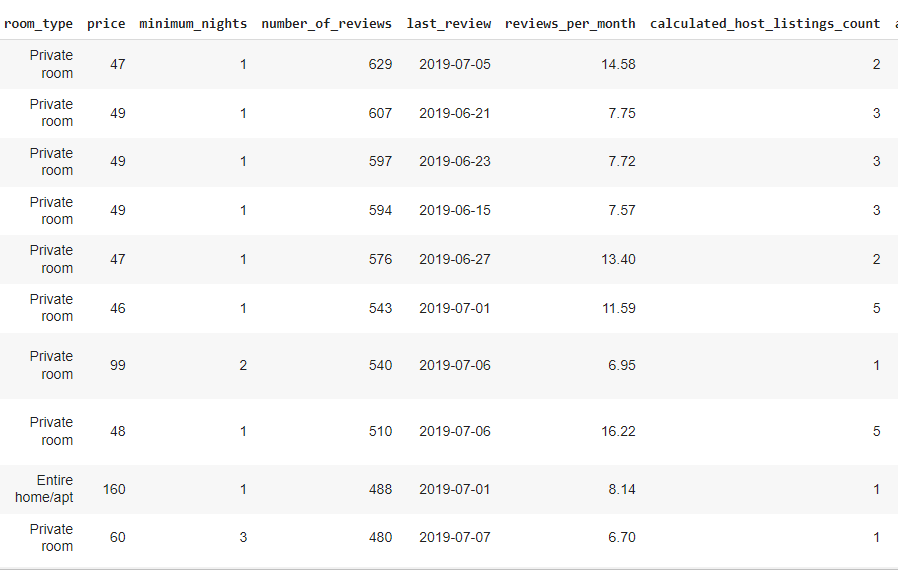


Table 1: top 10 most reviewed listings in NYC

**4.1.7 Price column-distribution and nature.**

The distribution curve of price column provides the distribution and nature of price column over the entire dataset.

We have used seaborn distplot to plot this distribution curve.

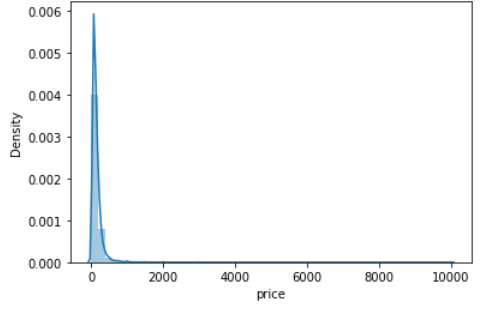


Fig 6: distribution plot for price

The distribution has a positively skewed tail at the extreme as we can see in the above distribution plot.

The **skewness** is found to be 19.118939 and **kurtosis** to be 585.672879. We can observe thattheskewness value being greater than 1 and kurtosis is high as 585, it indicates the presence of good amount of **outliers**. We have to look into how to handle these outliers.

Let’s see the boxplot of this **price** column to check the presence of outliers.

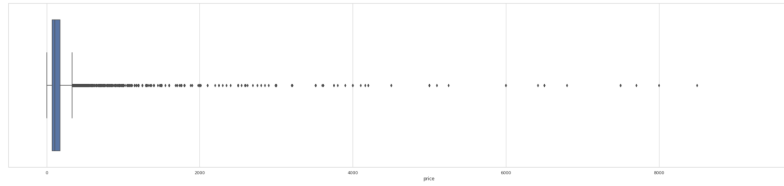


Fig 7: box plot for price column

**4.1.7.1 Dealing with outliers—How to deal with such anomalies in data?**

The common **data pre-processing steps**like missing values treatment is done but dealing with the outliers is very important part in EDA.

An **outlier** is a data point that lies outside the overall pattern in a distribution. These outlier value may come from an accidental response that was recorded correctly or from a data that is entered wrongly which leads to an error.

There can be low range or high range outliers and it can be found out using IQR (Inter-Quartile range) approach. A data point is an outlier if it is more than 1.5. IQR above the **third quartile** or below the **first quartile**. Low outliers are below **Q1–1.5\*IQR** and high range outliers are above **Q3+1.5\*IQR.** In this Airbnb dataset, I have used the**IQR approach** to handle the outliers as it had performed the best in removing almost all the extreme values present.

The below combination figure of distribution plot and box plot gives the comparison of price column before and after trimmings the outliers using IQR based filtering.

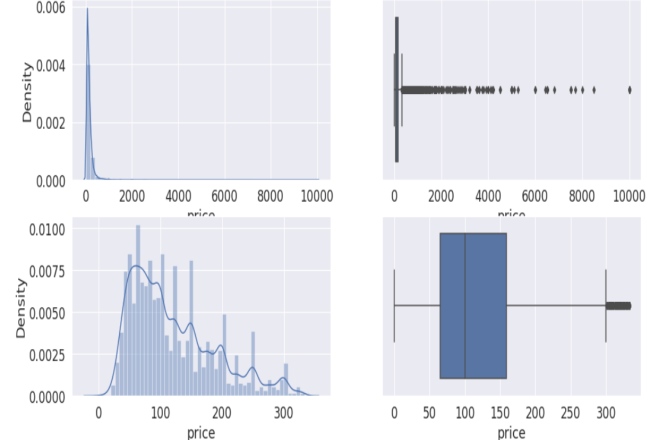


Fig 8: distplot and bar plot with and without outliers

The below box plot provides the variation of price with respect to the different type of rooms available.

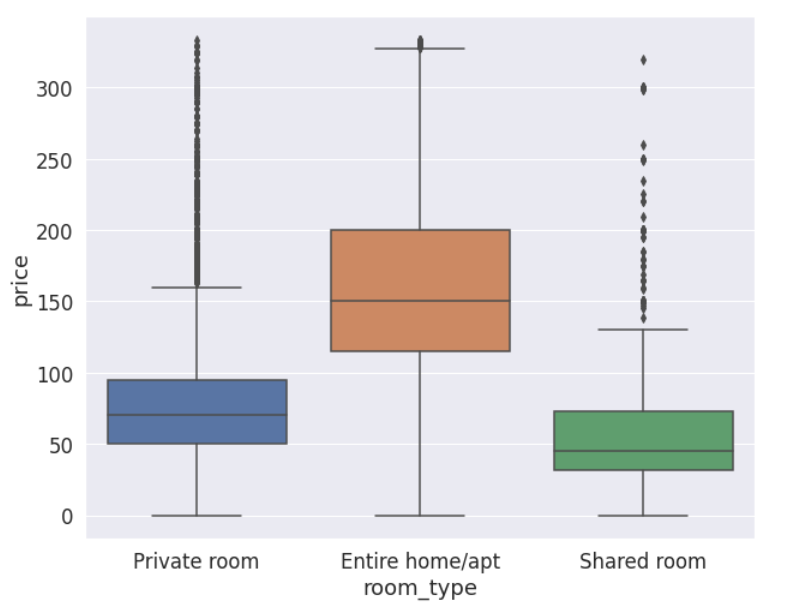


Fig 9: price vs room types boxplot

**4.1.8 Year wise types of rooms getting a last review**

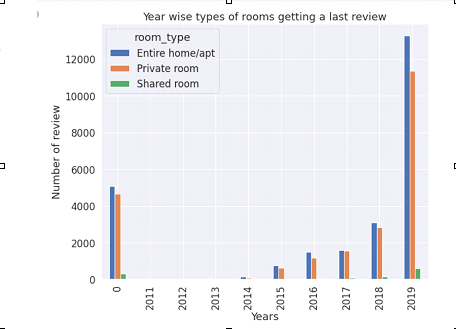


Fig 10: Year wise types of rooms getting a last review

The above bar graph shows the types of room getting a last review in year wise. As per the graph from 2011-2014 very less number, 2015-2018 review’s are slightly increase. In 2019 we can see more number reviews. Year 0 is missing years in Dataset. The x-axis shows the Year and the y-axis show number of last review.

In this graph group by types of room (Entire home/apt, Private room, Shared room), Entire home/apt is getting more review’s in every year and Shared room is very less review’s compare to other.

In 2019 compare with the other year’s all types room getting more last review’s.

Entire home/apt - 13k(13000)

Private room’s - 11k(11000)

Shared room’s - 200

**4.1.9 How many Neighbourhood\_group are available All days**

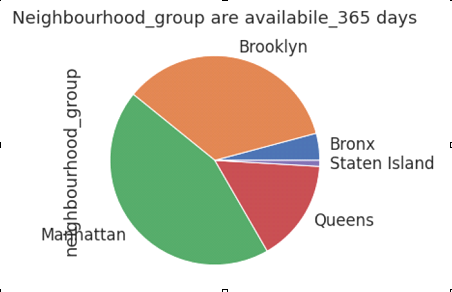


Fig 11: Neighbourhood\_group are available All days

The above Pie chart shows the how many of Neighbourhood\_group are available in all days of year. The Neighbourhood\_group is categorical column. Highest number Neighbourhood\_gruop is Manhattan (572) and lowest is Staten Island (12).

**4.1.10 How many Neighbourhood\_group are available All days**

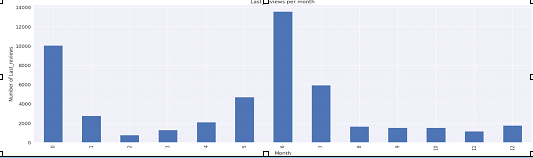


Fig 11: Neighbourhood\_group are available All days

In the above Bar graph shows the last reviews of every month in all year’s. The month 6(June) is the getting most last reviews. X-axis is shows the month and Y-axis is show’s the number of review’s.

In this graph we can observe month 6 (June) is busiest month in all year’s because more reviews are in June month. Most of busiest month’s are 5(May), 6(June) and 7(July). In bar graph we easy to know the which month is getting last review’s.

**4.1.11 Number of hotels in neighbourhood**

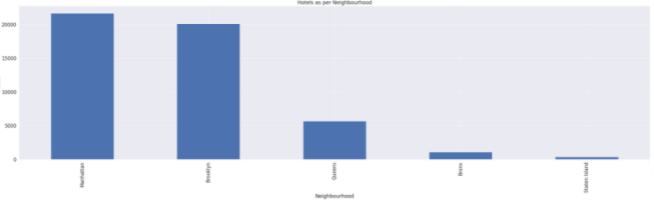


Fig 12: Number of hotels in neighbourhood

The above bar plot indicates the number of hotels according to a particular neighborhood. As per the graph, Manhattan has the highest number of hotels and Staten Island has the lowest number of hotels.

We have used bar graphs because they are **used to compare things between different groups or to track changes over time**. However, when trying to compare between different groups, bar graphs are best when the changes are larger.

From the above graph we can see that:

Manhattan has 21661 hotels

Brooklyn has 20104 hotels

Queens has 5666 hotels

Bronx has 1091 hotels

Staten Island has 373 hotels.

**4.1.12 Number of private room in all neighborhood**

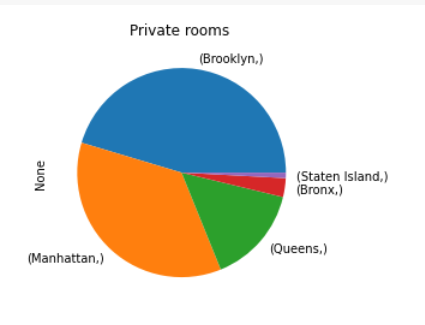


Fig 13: Number of private room in all neighborhood

The above pie chart displays private rooms in the neighborhood. The graph confirms that Brooklyn has the maximum number of hotels (10132) that have only private rooms.

This chart displays the whole relationship in our data on the basis of a private room, each slice represents one neighborhood.

The total number of hotels that have private rooms is 22,326. This data helps customers to determine which neighborhood they can choose when they wish to go for a private room.

**4.1.13 Minimum nights**

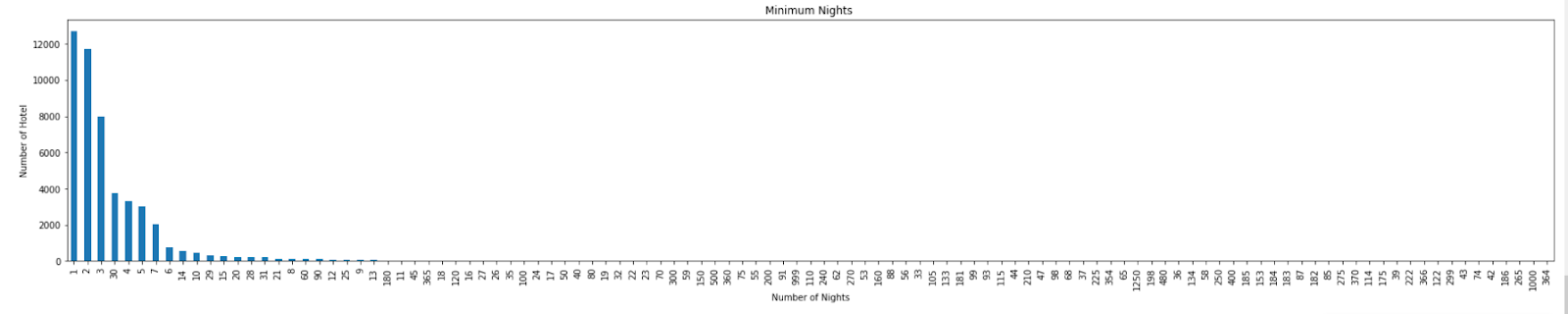


Fig 14: Minimum nights

In Airbnb **the minimum number of nights that a guest can book a short-term vacation rental**. It's determined entirely by the host, and it can be adjusted to correlate with yearly, monthly, and weekly trends.

The above bar graph shows the minimum nights available for all the hotels. As per the graph maximum hotels(12720) have 1 as minimum nights. In the above graph, the x-axis shows the number of nights and the y-axis shows the number of hotels.

**4.1.14 Average price of Airbnb based on Number of reviews**

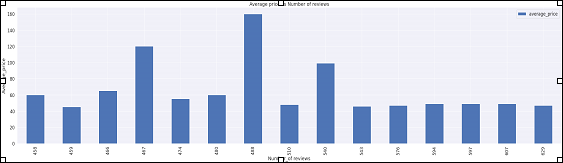


Fig 15: Average price of Airbnb based on Number of reviews

The above bar graph shows the average price according to the number of reviews available in the Airbnb data. As per the graph the highest average price of 166 have 488 total numbers of reviews. In the above graph, the x-axis shows the number of reviews and the y-axis shows the average price.

**4.1.15 Average reviews of Airbnb in different neighbourhood\_group**

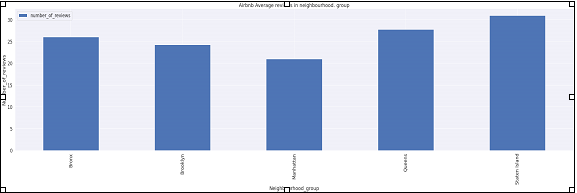


Fig 16: Average reviews of Airbnb in different neighbourhood\_group

The above bar graph indicates the average number of reviews as per different neighborhood group available in the Airbnb data. As per the graph above the highest number of reviews comes under the Staten Island neighborhood group which having the 30.97 of average reviews in the airbnb data and after that the Queens neighborhood group comes. In the above graph, the x-axis shows the different neighborhood group and the y-axis shows the average number of reviews here.

**4.1.16 Average reviews of AirBnb in different neighbourhood**

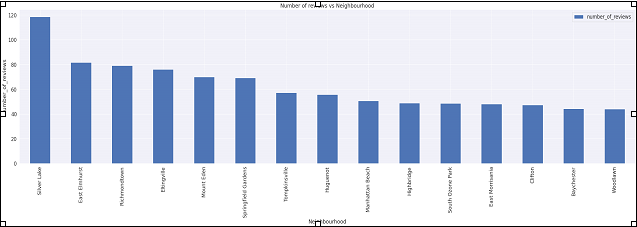
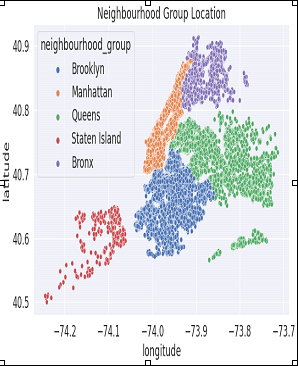
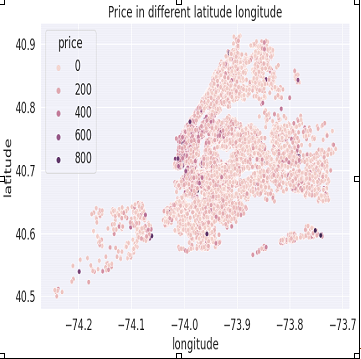


Fig 17: Average reviews of AirBnb in different neighbourhood

The above bar graph indicates here the average number of reviews vs the Neighborhood available in the airbnb data. The above graph the Silver Lake has the maximum number of reviews among all the neighborhoods. Here the x-axis indicates the Neighborhood and the y-axis indicates the average number of reviews.

**4.1.17 Price according to Neighborhood locations**

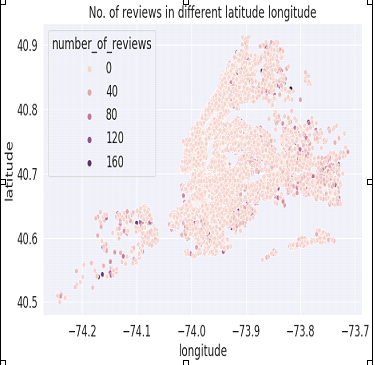


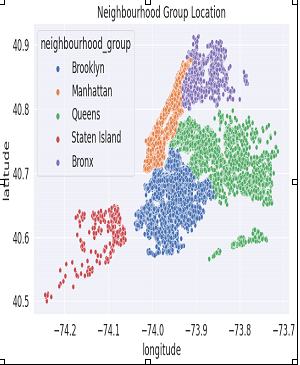


This scatter plot helps us to find out the neighborhood group in the map that is represented by the number of dots based on the latitude and longitude given to us in the dataset.

The other plot indicates the price in different longitude and latitude in the dots. So by comparing both plots the price of 800 is very less in number among the neighborhood group. The price of 200 and 400 is evenly distributed among the neighborhood group.

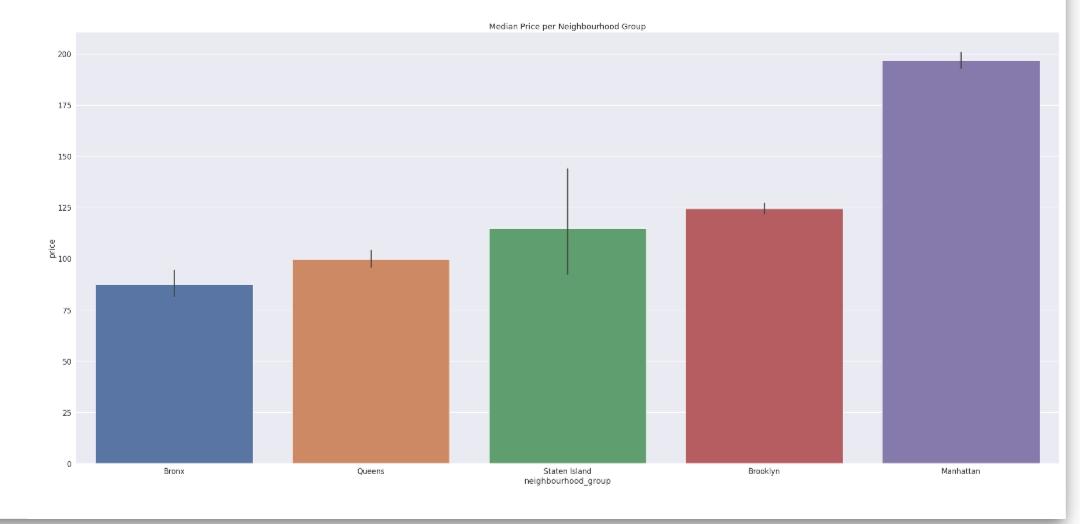
**4.1.18 Number of reviews according to neighborhood locations**





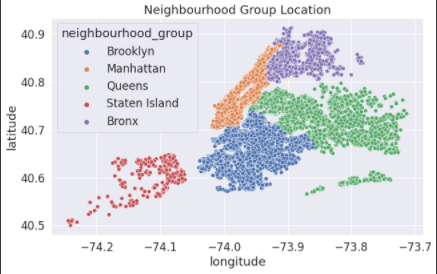
The first plot is same which is already mention above. Now the other plot shows the number of reviews among the different neighborhood group. Most of the average number of reviews are fallen between 40 and 80 among the neighborhood group.

**4.1.19 Median price per neighbourhood group**



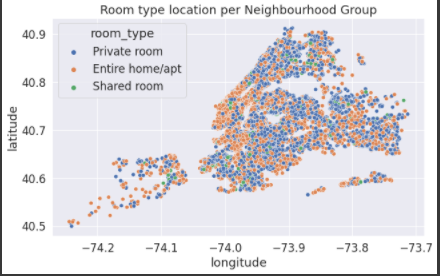
Median price is the middle point for real estate prices. It is not the same as the average price. The median price is the price in the very middle of a data set, with exactly half of the houses priced for less and half priced for more. We have created the bar plot of median price per Neighbourhood group from this we can see what are the median prices distributed among various neighbourhoods with the help of prices and neighbourhood\_group column. A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A bar graph shows comparisons among discrete categories. One axis of the chart shows the specific categories being compared, and the other axis represents a measured value.In our case x-axis is the neighbourhood group and y-axis is the price. Through study of this bar graph we can understand What is the median price for different neighbourhood groups.

**4.1.20 Neighbourhood group location**



A scatter plot is a diagram where each value in the data set is represented by a dot. We have plotted a scatter plot based on the latitude and longitude given to us in the dataset.We have distributed longitude among x-axis and latitude among y-axis and we have given separate colours for different neighbourhood groups. We have set hue for neighbourhood groups, hue will produce data points with different colors. Hue can be used to group multiple data variables and show the dependency of the passed data values to be plotted. Through study of this scatterplot we can understand how the neighbourhood groups are for different locations.

**4.1.21 Room type location per Neighbourhood Group**



We have plotted a scatter plot based on the latitude and longitude given to us in the dataset.We have distributed longitude among x-axis and latitude among y-axis and we have given separate colours for different types of rooms such as Private room, Entire home/Apartment or shared room. We have set hue for room type, hue will produce data points with different colors. Hue can be used to group multiple data variables and show the dependency of the passed data values to be plotted. Through study of this scatterplot we can understand how the room types are for different neighbourhood groups.

**5. Conclusions**

From the entire above analysis we can conclude that,

• Sonder is the busiest host on Airbnb. Though Michael did maximum entry after that Sonder, when we check the host listing count Sonder comes first before Michael.

• Shared rooms are less available compared to other room types but when we check the numbers of reviews we find out those private and shared rooms are preferred more compared to the entire room. So it means we have to increase the availability of room type other than entire must be increased to gain profit.

• The number of features positively correlated with each other. The correlation between the number of reviews and reviews per month is high. There is a correlation between calculated listing counts, minimum nights and availability 365 also.

• As skewness and kurtosis are high means a good amount of outliers are present in the price features. So to create uniformity in the data, remove the outliers by a boxplot.

• The price of private and shared room types is less compared to the entire room type. That’s why most of the visitors want to book private and shared rooms’ type.

• The Topmost reviewed room is the private type which has a price below 50$. It means users prefer cheap rooms.

• Manhattan has the highest number of hotels which have availability 365 then Brooklyn comes, as Manhattan is famous for museums, stores, parks and theatres that’s why more hotels are available.

• If we see closely the private room type is highest in Brooklyn then Manhattan comes. So as we discuss before private room type is high in demand and also this place near to Manhattan so visitors can visit both in cheap prices.

• Maximum hotels consider one minimum night which is good for visitors to stay accordingly.

• Most of the last reviews came in June month which means we can increase the price of Airbnb as most visitors visit this month.

• Staten Island has the maximum number of reviews than Queens that’s why most of the rooms are not available all 365 days compared to Brooklyn and Manhattan.

• The silver lake in Staten Island has more reviews than other places in NYC.

• Most of the prices are less in Staten Island and Queens because the number of reviews was more in these places than in other boroughs. It means we can increase the price to increase profit.

However, features related to accommodation, room type, and neighborhoods in Manhattan, Queens, and Brooklyn play an important role in determining future price of the listings. These features would be crucial for AirBnb to predict revenue.