|  |
| --- |
|  |
| **VISUALIZE NEWYORK CITY AIRBNB DATA** |
| **TANIA SULTANA**  **931031**  s87431@bht-berlin.de  &  **AHMED DIDER RAHAT**  **916146**  s40183@bht-berlin.de |
| Submitted on: 10th August 2022 |

Table of Contents

[**1.** **Introduction** 2](#_Toc111067629)

[**2.** **The data set** 3](#_Toc111067630)

[**3.** **Visualizing the data of Airbnb data sets** 4](#_Toc111067631)

[**3.1** **Basic statistics of the dataset:** 4](#_Toc111067632)

[**3.2** **Univariate Analysis:** 6](#_Toc111067633)

[**3.3** **Bivariate analysis:** 15](#_Toc111067634)

[**4.** **Final Discussion** 32](#_Toc111067635)

[**5.** **References** 33](#_Toc111067636)

Index

# **Introduction**

Data visualization is the process to develop an understanding of row data. By using this, we can get insights that are not visible in the first place. It is one of the core processes of generating information from data. In our project, we visualize the Airbnb data for New York City. We try to make some graphical representations of the features and also demonstrate the relationship between features.

Airbnb stands for Air Bed and Breakfast. According to business model navigator “Aribnb is an American company that operates an online marketplace and hospitality service for people to lease or rent short-term lodging including holiday cottages, apartments, homestays, hostel beds, or hotel rooms, to participate in or facilitate experiences related to tourism such as walking tours, and to make reservations at restaurants. The company does not own any real estate or conduct tours; it is a broker which receives percentage service fees in conjunction with every booking. Like all hospitality services, Airbnb is an example of collaborative consumption and sharing. The company has over 4 million lodging listings in 65,000 cities and 191 countries and has facilitated over 260 million check-ins.”

# **The data set**

The data set is collected from [kaggle.com](https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data). According to Kaggle, “Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019. This data file includes all needed information to find out more about hosts, geographical availability, and necessary metrics to make predictions and draw conclusions”.

**Features of the Dataset:**

1. **id:** The id assigned to each airbnb to identify them uniquely.
2. **name:** The name assigned to each airbnb.
3. **host\_id:** The id assigned to each host to identify them uniquely.
4. **host\_name:** The name assigned to each host.
5. **neighbourhood\_group:** The 5 boroughs that New York City is divided into: Manhattan, Queens, Brooklyn, Staten Island and Bronx.
6. **neighbourhood:** The neighborhood where the airbnb is located within the boroughs.
7. **latitude:** The latitude of the location where the airbnb is situated.
8. **longitude:** The longitude of the location where the airbnb is situated.
9. **room\_type:** The type of airbnb which is divided into two: Entire home/Apartment, Private room and Shared Room.
10. **price:** The rent of the airbnb per night.
11. **minimum\_nights:** The minimum number of nights the airbnb can be rented for.
12. **number\_of\_reviews:** Total number of reviews posted by customers.
13. **last\_review:** Date of the last review posted by a customer.
14. **reviews\_per\_month:** Monthly total of reviews posted by customers.
15. **calculated\_host\_listings\_count:** Number of total listings by a host.
16. **availability\_365:** Yearly number of days the airbnb is available for rent.

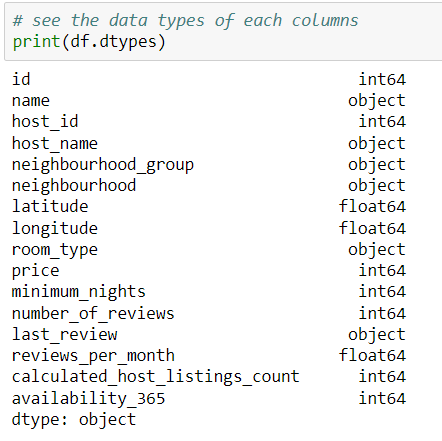
# **Visualizing the data of Airbnb data sets**

* 1. **Basic statistics of the dataset:** In this portion we load the data set and see some basic statistics of different features.
     1. **Data load in Pandas:** At first we load the data from csv file to pandas data-frame.

****

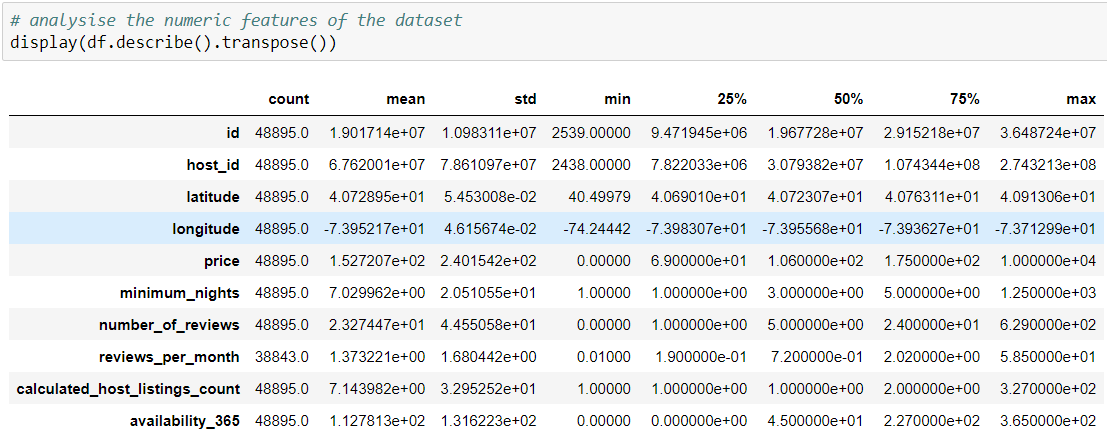
**Figure 01- data loading**

1. **Observe the data type:** In this section we analyze the data type of different features.

****

**Figure 02- get the data type for each columns**

1. **Statistical Analysis of Numerical Data:** Here, we have analyzed the statistical properties (count, mean, standard deviation, min, max, and percentile) of the numeric data.

****

**Figure 03- Analysis Numeric features**

|  |  |
| --- | --- |
| 1. **Null Value Analysis:** At this stage, we have counted exactly how many null values are in each column. At the end of the count, we see that the highest number of null values are in the review-related features. There are ten thousand and fifty-two (10052) null values ​​in each of this feature. These features have the highest number of null values ​​as customers do not give reviews. | **Figure 04- Null value analysis** |

1. **Univariate Analysis:** In this portion we analyze the univariate features of the dataset. For these type of analysis, at first we get the numbers and then plot them in different figures.
   * + 1. **Count neighborhood group:** We calculated the number of Airbnb in each group. After getting the count we plot the data into a bar-plot.

****

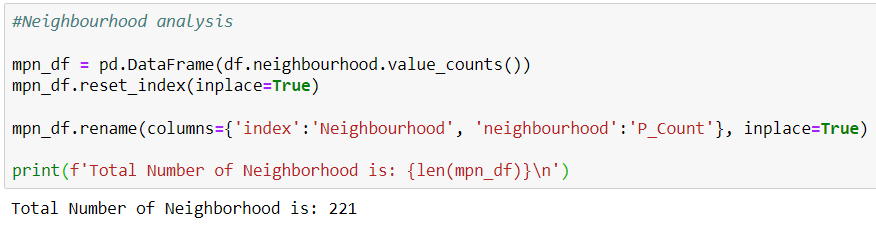
**Figure 05- Code and output of neighborhood group data**

|  |  |
| --- | --- |
| **Analysis:** In the bar chart, we can clearly see that Manhattan has the highest number of Airbnb with the percentage of 44.3%. Staten Island has the lowest number of Airbnb, and the count is only 373 (0.6%). The Bronx has about 2.2% of Airbnb which is 2nd lowest in NYC. And lastly, 41.1% and 11.6% of Airbnb are situated Brooklyn and Queens respectively. |  |

**Figure 06- Analysis of the number of Airbnb within different neighborhood group**

1. **Analyze the counts of Airbnb in different neighborhood:** Here, we analyze the volume of Airbnb in each neighborhood. For performing the analysis, we did some interrelated steps.

**Step 1:** We get the number of the unique neighborhood in NYC and we got 221 as our output.

****

**Figure 07- Code and output of unique neighborhood counting**

**Step 2:** Then we start investigating how much percentage of Airbnb situated in the top-nth neighborhood. So, we sort them based on their Airbnb counts and calculate their percentage. Later we plot a diagram that contains the percentage of total Airbnbs on the X-axis and the number of neighborhoods on the Y-axis. From the graph, we can conclude that the top 20 neighborhoods hold almost 60% of total Airbnbs. Lastly, we saw that 65.87% of total Airbnb situated on only top 20 neighborhood.

|  |  |
| --- | --- |
|  |  |

**Figure 08- Code and Visualization of Top nth Neighborhood and their percentage**

**Step 3:** Then we only focus on the top 20 neighborhoods. In this step, we show them in a table and make a bar plot with their percentages. **From the count and graph,** we make some decisions. Firstly, Williamsburg and Bedford-Stuyvesant have the maximum number of Airbnbs whose volumes are respectively 8.0% and 7.6%. Secondly, Clinton Hill has a minimum ratio of 1.2% in this top-20 list. Finally, all the other neighborhoods have roughly 5% to 1.5% Airbnb in their location.

|  |  |
| --- | --- |
|  |  |

**Figure 09- Implementation and analysis of top 20 neighborhoods based on their Airbnb counts**

**Step 4:** Here, we draw the top 20 neighborhoods on the NYC map. We also set a color scale to see the volume of the Airbnb count. **From the map,** we extracted some pieces of information. Firstly, all the 20 neighborhoods are situated in pretty much the same geographical region. Secondly, the top 2 neighborhoods named Williamsburg and Bedford-Stuyvesant are next to each other. Lastly, compared with the whole area of the city, most of the (65.87%) Airbnbs are densely located in a small area (only 20 neighborhoods out of 221).

|  |
| --- |
|  |

**Figure 10- Extract information from NYC map**

1. **Count of Apartment type:** There are only three types of Airbnb in the dataset. So we fetched the count of each group and plot them in a pie-chart.

****

**Figure 11- Implementation of Airbnb type with counts**

|  |  |
| --- | --- |
| **Analysis:** The pie chart shows that the number of “entire homes/apartments” is higher than the other two types, and the percentage is about 52.0%. The number of “shared room” has minimum counts with 2.4% of apartments. The “private room” type Airbnb has 45.7% of the apartments. |  |

**Figure 12- Pie-chart of apartment type**

1. **Counts of minimum night stay feature:** To analyze this figure we used a box-plot. And implementation, graph, and interpretations are given bellow:

|  |  |
| --- | --- |
| **Figure 13- Implementation of minimum night stay feature** | |
| **Analysis:** The minimum and maximum night’s stay in the dataset are 1 and 1250 respectively. The numbers between 1-100 are more densely populated in the graph, which means a huge percentage of the apartment has a minimum night stay value in this range. There also have some values from 100-300 (approximately). The rest of the numbers are some sort of outliers. |  |

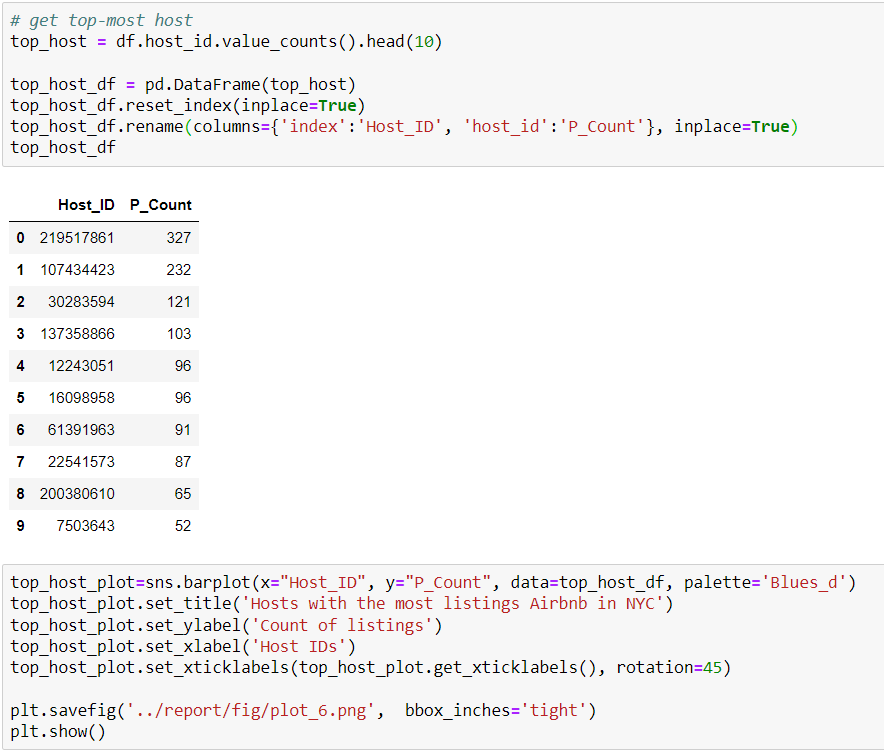
**Figure 14- Analysis of minimum night stay feature**

1. **Analyze the calculated host listing feature:** This feature represents the number of apartments listed by a host. We fetched the data and created a box plot of this feature.

|  |  |
| --- | --- |
| **Figure 15- Implement host listing feature** | |
| **Analysis:** According to the figure, we can see that a single host listed 327 apartments, which is pretty high compared with other hosts. Most of the hosts listed 1-50 apartments as the graph has much dense population is this section. |  | |

**Figure 16- Analysis of host listing in the data set**

1. **Get top most listed host:** In this part, we have calculated which hosts have been listed with the most apartments. And we've listed the top ten hosts (host IDs) that have been rented more frequently.

****

**Figure 17- Code implementation of most top host**

|  |  |
| --- | --- |
| **Analysis:** Looking at the drawn bar chart, it can be seen that this host id 219517861 has been used the most, that is, more than 300 times. Host 7503643 has been used a little more than about fifty times, and this host has been used the least number of times in the top 10 hosts. |  |

**Figure 18- Bar chat and insight of host id feature**

1. **Analysis of the 'number of reviews' feature:** To analyze this column, we fill all the empty cells with zero. Then we plot all the data into a box plot.

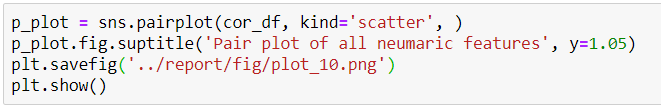
|  |  |
| --- | --- |
| **Figure 19- Code implementation number of reviews** | |
| **Analysis:** It is clear from the figure that the most number of reviews of the dataset is distributed into 100-400. Also, some apartments have more than 400 reviews. |  |

**Figure 20- Box-plot of ‘number of reviews’ and the Insight of the graph**

1. **Analysis the price of Airbnb:** To analyze the price, we draw two different type plot i.e. the box-plot and scatter plot.

|  |
| --- |
| **Figure 21- Implementation and Box-plot of feature price** |
| **Figure 22- Implementation and Scatter-plot of feature price** |
| **Analysis:** To analyze the box-plot and scatter-plot we can decide that, most of rent of the apartment are in between 100-1000. Also there are some costly apartment are visible in the graph. |

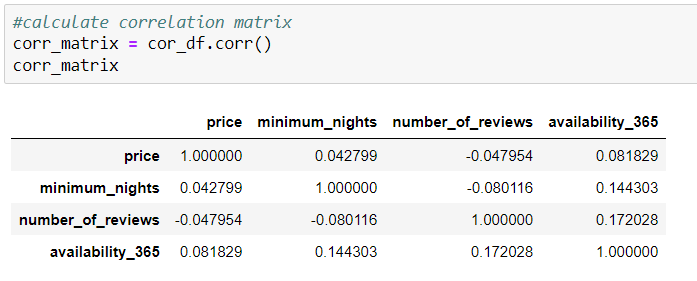
1. **Bivariate analysis:** In this portion we do bivariate features analysis.
2. **Drawing Pair plot:** To analyze the numeric features we make a new dataframe. Then we plot the data into a pair plot.

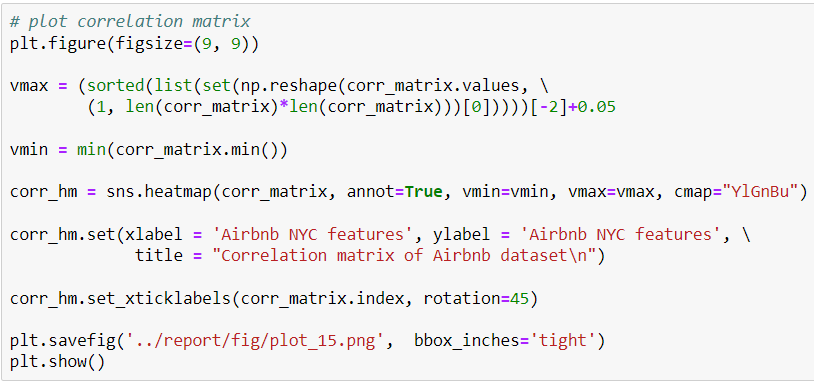


**Figure 23- Implementation of pair plot**

|  |
| --- |
| **Figure 24- Pair plot of numeric feature** |
| **Analysis:** From the above graph, we can see a correlation plot of the numerical variables of our data set. A Correlation Plot represents the strength of linear relation among the numeric variables. We assumed at least some linear relationship between the numeric variables initially (At least the price and the popularity of the housing). Here, the number\_of\_reviews variable is the way to measure the popularity of any Airbnb. Though there is no direct linear relationship between the variables, we observed some lightweight relationships for some of them.  **Firstly,** the relationship between price and the number of reviews. There is an exponential relationship between them where the number of reviews increases as the price decrease, which means the costly Airbnbs are less acceptable in NYC.  **Secondly,** we notice that a similar kind of relationship exists in minimum nights and number of reviews. The Airbnbs that have lower minimum night stay are most popular in NYC.  **Lastly,** we also observed a similar trend in the relationship between price and minimum night stay, with some slight exceptions. Here, most of the expensive Airbnbs have a high value of minimum night stay. |

1. **Correlation matrix of the numeric data:** In these section we calculate the correlation matrix and also draw them in a heat-map.





**Figure 25- Implementation of correlation matrix**

|  |
| --- |
| **Figure 26- Heat map of correlation matrix** |
| **Analysis:** From the plotted correlation figure, we clearly state that the correlation values are too small to make a linear relationship. The bluish-colored values represent a positive correlation, and the yellowish-colored values represent negative correlations. The highest positive correlation value is 0.17, between the number of reviews and availability\_365. The highest negative correlation between the two features is -0.08 (minimum night stay and number of reviews). |

1. **Price VS Neighborhood group:** In this portion, we have calculated the mean and median prices for each neighborhood group in NYC. We converted this stat into a bar chart.

|  |
| --- |
| **Figure 27- Implementation of the relationship between price and neighborhood group** |
| **Figure 28- Bar plot of the relationship between price and neighborhood group** |
| **Analysis:** The bar chart shows that both the mean and median prices of the Manhattan group's Airbnb are higher than the other groups. The mean price of this group is around 200, and the median price is approximately 150. The lowest cost Airbnb is available in the Bronx. The other three groups that have Airbnbs have an almost similar price. In all the cases, the mean price is much higher than the median price because of some extreme values. |

1. **Price VS Neighborhood:** At this stage, we try to pick the top-10 neighborhoods based on their price. For calculating the price, we calculate the mean and median value. As a result, we got two different sets with some common neighborhoods.

|  |
| --- |
|  |
|  |
|  |
| **Figure 29- Implementation of relationship between price and neighborhood group** |
| **Figure 30- Top 10 neighborhoods based on mean price** |
| **Figure 31- Top 10 neighborhoods based on median price** |
| **From the above graph, we made some interpretations.**  Firstly, it is clear that using both parameters Fort Wadsworth, Woodrow, and Tribeca are the most expensive neighborhoods.  Secondly, only five out of ten neighborhoods are common in the two price group. Including the above-mentioned three neighborhoods, Flatiron District and NoHo are also common in both sets.  Thirdly, we further investigate why there are two difference sets using mean and median. |
|  |
| **Figure 32- Investigation of the two different set of neighborhoods** |
| **Figure 33- Bar of the std. deviation of different neighborhoods** |
| **Analysis:** Using this figure, we could answer the question for two different sets of neighborhoods. Here it is clear that the mean set is different from the median only because of the high variance in the price. Using this figure, we could answer the question for two different sets of neighborhoods. Here it is clear that the mean set is different from the median only because of the high variance in the price. We can get much variety price apartments in the neighborhoods like Sea Gate, Riverdale, Prince's Bay, Battery Park City, and Randall Manor. |

1. **Price VS Airbnb type:** In this section, we have calculated the mean and median prices for all three types of Airbnb. Then, we plotted the stat using a bar chart.

|  |  |
| --- | --- |
| **Figure 28- Implementation of the relationship between price and Airbnb type** | |
| **Figure 29- Plot of the relationship between price and Airbnb type** | |
| **Analysis:** The bar chart shows that both prices (mean and median) for the entire home/apartment type are the highest. The lowest price is the shared room. Though both types of parameters private-room are in the middle position, the values are much lower than the entire home/apartment. |

1. **Room type VS Neighborhood group:** At this stage, we have counted how many Airbnb there are in each neighborhood group. Then, we plotted the data from the list in the form of graphs.

|  |
| --- |
|  |
|  |
| **Figure 30- Implementation of the relationship between neighborhood group and Airbnb type** |
| **Figure 31- Bar plot of the relationship between neighborhood group and Airbnb type** |
| **Analysis:** From the graph, we can get some assumption. Firstly, largest number of entire home/apt, and shared apartment are situated in Manhattan. Secondly, Brooklyn has maximum number of private apartment. Lastly, Staten Island has less Airbnb in all three types. |

1. **Get top reviewed neighborhoods:** Here, we reveal the percentage of reviews in different neighborhood groups. As we don't have any information about the meaning of the review (positive and negative review), we could not analyze the sentiment of the reviews. From the given figure, we only calculate the percentage of the reviews. If we calculate the percentage of reviews based on the review number, it would be highly biased on the count of Airbnbs in those neighborhoods. As a result, we see Brooklyn and Manhattan hold most of the reviews. So, we decided to calculate the review per Airbnb.

|  |
| --- |
|  |
|  |
| **Figure 32- Implementation of Neighborhood group VS number of review** |
| **Figure 33- Pie-chart of Neighborhood group VS number of review** |
| **Analysis:** Here, we reveal the percentage of reviews in different neighborhood groups. As we don't have any information about the meaning of the review, we could not analyze the sentiment of the reviews. From the given figure, we only calculate the percentage of the reviews. If we calculate the percentage of reviews based on the review number, it would be highly biased on the count of Airbnbs in those neighborhoods. As a result, we see Brooklyn and Manhattan hold most of the reviews. So, we decided to calculate the review per Airbnb. |
| **Figure 34- Implementation of Neighborhood group VS review per Airbnb** |
| **Figure 35- Bar plot of Neighborhood group VS review per Airbnb** |
| **Analysis:** According to this analysis, The Airbnbs on Staten Island has the maximum average number of reviews, Although it contains only 1% of the total number of reviews. Queens has the second highest number of reviews per Airbnbs (27.7 reviews per Airbnb). More than 80% of the overall reviews are listed on the neighborhood groups Brooklyn and Manhattan, but their average counts are the lowest among entire neighborhoods. |

1. **Get top reviewed Airbnb types:** Here, we reveal the percentage of reviews in different neighborhood groups. As we don't have any information about the meaning of the review (positive and negative review), we could not analyze the sentiment of the reviews. From the given figure, we only calculate the percentage of the reviews. If we calculate the percentage of reviews based on the review number, it would be highly biased on the count of Airbnbs in those neighborhoods. As a result, we see Brooklyn and Manhattan hold most of the reviews. So, we decided to calculate the review per Airbnb.

|  |
| --- |
|  |
|  |
| **Figure 36- Implementation of Airbnb type VS number of review** |
| **Figure 37- Pie-chart of Airbnb type VS number of review** |
|  |
| **Figure 38- Implementation of Airbnb type VS avg. Airbnb reviews** |
| **Figure 38- Bar plot of Airbnb type VS number of review (per Airbnb)** |
| **Analysis:** According to the analysis, we interpret some pieces of information.  **Firstly,** the entire home/apt type has almost half of the total reviews. The average number of reviews per Airbnb is 22.84 in this category.  **Secondly,** There are 47% of total reviews are given to this category. On average 24.11 reviews are given to each Airbnb of this category.  **Thirdly,** As shared apartments are most uncommon in NYC, only 2% of total reviews are owned by this type. |

# **Final Discussion**

**In this section we will summarize our finding according to the dataset and our analysis.**

1. Almost all the data set is null free except two review related columns.
2. In NYC the Airbnb are situated on all 5 neighborhood group and 221 neighbor.
3. There is not linear relationship among the numeric feature of the dataset.
4. Renting entire apt/home is most popular type as more than 50% of the data situated on this category also it is most expensive one.
5. Least costly Airbnb type is shared room also it is least popular.
6. Most Expensive Airbnb(s) are situated in Manhattan and also become most popular (based on Airbnb count) among other group.
7. Airbnb with lowest number of minimum night are most popular in NYC.

# **References**

1. <https://blog.datawrapper.de/beautifulcolors/>
2. <https://businessmodelnavigator.com/case-firm?id=4>
3. <https://www.schemecolor.com/flat-orange-blue-green-pie-chart.php>
4. <https://writing.wisc.edu/handbook/assignments/planresearchpaper/>
5. <https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data>