Capstone Project

Airbnb Data Analysis

Exploratory Data Analysis (EDA)

By : Nasim Aalam

1. Introduction :

1.1. What is Airbnb:

Airbnb is an online marketplace since 2008, which connects people who want to rent their homes with people who are looking for accommodations in a particular location. It covers more than 81,000 cities and 191 countries worldwide. The company ,which is based in San Francisco, California, does not own any of the property listings, but it receives commissions from each booking like a broker. The name “Airbnb” comes from “air mattress Bed and Breakfast.” The Airbnb logo is called the Bélo, which is a short version for saying ‘Belong Anywhere’. Airbnb hosts list many different kinds of properties such as private rooms, apartments, shared rooms, houseboats, entire houses, etc.

1.2. Airbnb Dataset:

This dataset describes the listing activity and metrics in NYC for 2019. It includes all the necessary information in order to find out more about hosts, prices, geographical availability, and necessary information to make predictions and draw conclusions for NYC. The explanation of the variables in our data, which consists of 16 columns and 48,895 rows, will be made in the next part. The data used in this assignment is called New York City Airbnb Open Data which is downloaded from AlmaBetter. This public dataset is a part of Airbnb, and the original source can be found on this website.

1.3. Objectives:

In this project, we will perform an exploratory data analysis(EDA) in order to investigate each of the variables and also come up with a conclusion for the relationship between variables. The main purpose is to identify which variables affect the price mostly. In addition to these, we will explore which neighborhood groups and room types are the most popular ones among the guests, and which hosts are the most preferred ones. The processes during the eda can be listed as below:

- ➔ Data Cleaning
- ➔ Data Preprocessing
- ➔ Data Manipulation
- ➔ Data Visualization
- ➔ Exploring the information

1.4 Questions that can be answered by this analysis.

- ➔ Number of Neighbourhood in each Neighbourhood Groups
- ➔ Distribution of rooms according to latitude and longitude and room density in their regions
- ➔ Room types and their percentage in each neighbourhood groups
- ➔ Relationship between room type and Price
- ➔ Price and Review Relationship
- ➔ Minimum, Maximum and Average Price according to the neighbourhood group
- ➔ Most popular neighbourhood group
- ➔ The Most and Least Expensive Neighborhoods

➔ The Most and Least Available Neighborhoods
➔ Availability of Room Types According to the Neighborhood Groups
➔ Minimum Nights and Neighborhood Relationship
➔ The Most Popular Hosts in NYC Airbnb
➔ Variation in price according to the last review year
➔ The Average Number of Reviews in Each Neighboorhood
➔ Most common words used in name column
➔ Price Group Analysis of Neighborhood Groups

2. Data Explanation:

2.1 Used Libraries:

I have used several packages during the analysis of the historical data of Airbnb in NYC in order to make data manipulation and visualization. The list of packages used in this EDA can be seen below:

numpy ==========> To perform mathematical operations
pandas =========> For dataframe
matplotlib =====> For Visualization of Data
seaborn ========> For Visualization of Data
plotly.express ===> For Visualization of Data

plotnine ========> For Visualization of Data
warnings ========> For ignorance of any kind of unnecessary warnings
dataprep ========> Describing and understanding the data
wordcloud ========> for preparing worldcolud
PIL ==============> opening, manipulating different image file formats
re ===============> Manipulate all kinds of text and data

This dataset contains 16 features/variables about Airbnb listings within New York City. Below are the features with their descriptions:

1.id: Listing ID (numeric variable)
2.name: Listing Title (categorical variable)
3.host_id: ID of Host (numeric variable)
4.host_name: Name of Host (categorical Variable)
5.neighbourhood_group: Neighbourhood group that contains listing (categorical variable)
6.neighbourhood: Neighbourhood group that contains listing (categorical variable)
7.latitude: Latitude of listing (numeric variable)
8.longitude: Longitude of listing (numeric variable)
9.room_type: Type of the offered property (categorical variable)
10.price: Price per night in USD (numeric variable)
11.minimum_nights: Minimum number of nights required to book listing (numeric variable)
12.number_of_reviews: Total number of reviews that listing has (numeric variable)
13.last_review: Last rent date of the listing (date variable)
14.reviews_per_month: Total number of reviews divided by the number of months that the listing is active (numeric variable)
15.calculated_host_listings_count: Amount of listing per host (numeric variable)
16.availability_365: Number of days per year the listing is active (numeric variable)

2.2. Dataset Insights:

Dataset Insights

1. last_review has 10052 (20.56%) missing values -- **Missing**
2. reviews_per_month has 10052 (20.56%) missing values --- **Missing**
3. host_id is skewed -- **Skewed**
4. longitude is skewed -- **Skewed**
5. price is skewed -- **Skewed**
6. minimum_nights is skewed -- **Skewed**
7. number_of_reviews is skewed -- **Skewed**
8. calculated_host_listings_count is skewed -- **Skewed**
9. availability_365 is -- **Skewed**
10. name has a high cardinality: 47905 distinct values **High Cardinality**
11. host_name has a high cardinality: 11452 distinct values **High Cardinality**
12. neighbourhood has a high cardinality: 221 distinct values **High Cardinality**
13. last_review has a high cardinality: 1764 distinct values **High Cardinality**
14. last_review has constant length 10 **Constant Length**
15. longitude has 48895 (100.0%) negatives **Negatives**
16. number_of_reviews has 10052 (20.56%) zeros **Zeros**
17. availability_365 has 17533 (35.86%) zeros **Zeros**

Dataset Statistics

- Number of Variables 16
- Number of Rows 48895
- Missing Cells 20141
- Missing Cells (%) 2.6%
- Duplicate Rows 0
- Duplicate Rows (%) 0.0%
- Total Size in Memory 23.5 MB
- Average Row Size in Memory 504.1 B
- Variable Types
  - Numerical: 10
  - Categorical: 6

```python
from dataprep.eda import create_report
create_report(df)
```

DataPrep Report | Overview | Variables ≡ | Interactions | Correlations | Missing Values

Dataset Statistics

Number of Variables	16
Number of Rows	48895
Missing Cells	20141
Missing Cells (%)	2.6%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	23.5 MB
Average Row Size in Memory	504.1 B
Variable Types	Numerical: 10 Categorical: 6

Dataset Insights

`last_review` has 10052 (20.56%) missing values	Missing
`reviews_per_month` has 10052 (20.56%) missing values	Missing
`host_id` is skewed	Skewed
`longitude` is skewed	Skewed
`price` is skewed	Skewed
`minimum_nights` is skewed	Skewed
`number_of_reviews` is skewed	Skewed
`reviews_per_month` is skewed	Skewed
`calculated_host_listings_count` is skewed	Skewed
`availability_365` is skewed	Skewed

1 2

```
from dataprep.eda import create_report
create_report(df)
```

DataPrep Report | Overview | Variables ≡ | Interactions | Correlations | Missing Values

Overview

Dataset Statistics

Number of Variables	16
Number of Rows	48895
Missing Cells	20141
Missing Cells (%)	2.6%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	23.5 MB
Average Row Size in Memory	504.1 B
Variable Types	Numerical: 10 Categorical: 6

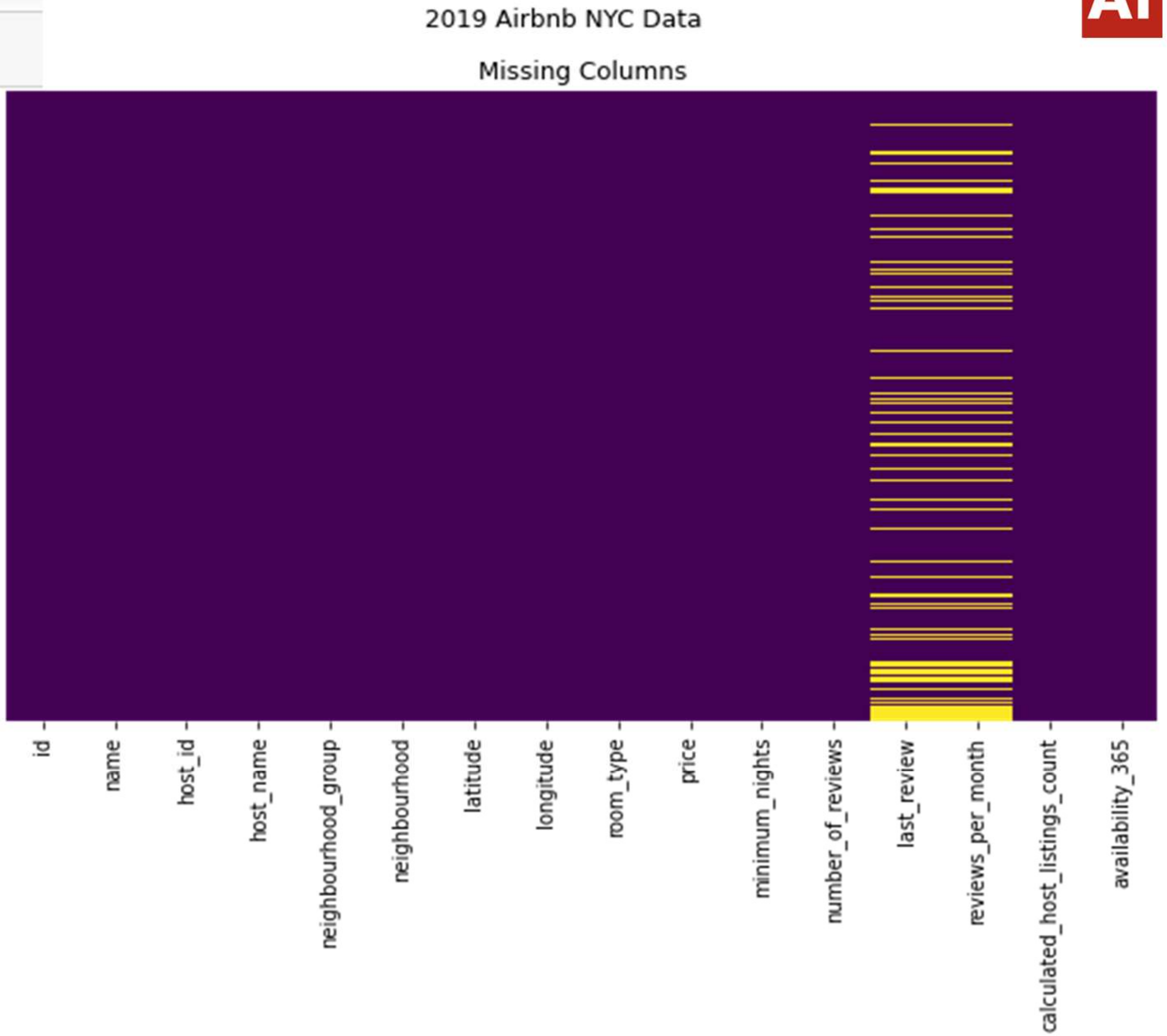
Dataset Insights

`name` has a high cardinality: 47905 distinct values	High Cardinality
`host_name` has a high cardinality: 11452 distinct values	High Cardinality
`neighbourhood` has a high cardinality: 221 distinct values	High Cardinality
`last_review` has a high cardinality: 1764 distinct values	High Cardinality
`last_review` has constant length 10	Constant Length
`longitude` has 48895 (100.0%) negatives	Negatives
`number_of_reviews` has 10052 (20.56%) zeros	Zeros
`availability_365` has 17533 (35.86%) zeros	Zeros

1 2

```
In [7]: df.isna().sum()

Out[7]: id                                    0
        name                                 16
        host_id                               0
        host_name                            21
        neighbourhood_group                   0
        neighbourhood                         0
        latitude                              0
        longitude                             0
        room_type                             0
        price                                 0
        minimum_nights                        0
        number_of_reviews                     0
        last_review                       10052
        reviews_per_month                 10052
        calculated_host_listings_count        0
        availability_365                      0
        dtype: int64
```



2.3. Data cleaning process (Handling Missing Values):

Null Values:

1. ➔ We can observe that name has 16 null values and host_name contains 21 null values, which is very few in comparison to the entire dataset, so I changed null values with **"not_present."**
2. ➔ last_review and reviews_per_month have 10052 (20.56%) in each, which is huge. It can affect the analysis, so I dropped them.

2.4 Unique Values:

➔ neighbourhood" contains **221** unique values.

➔ "neighbourhood_group" contains **5** unique values. There are five neighbourhood groups.

➔ "room_type" conatins **3** unique values. There are three types of rooms.

```
id                                0
name                              0
host_id                           0
host_name                         0
neighbourhood_group               0
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
calculated_host_listings_count    0
availability_365                  0
dtype: int64
```

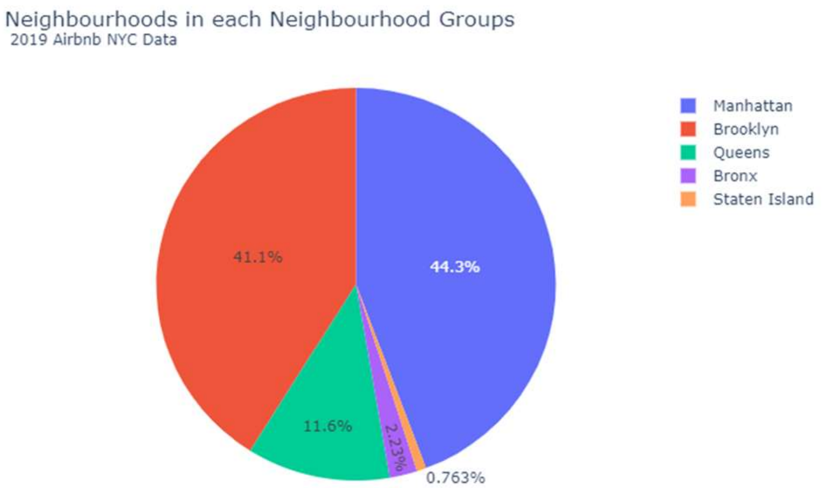
3. Exploratory Data Analysis:

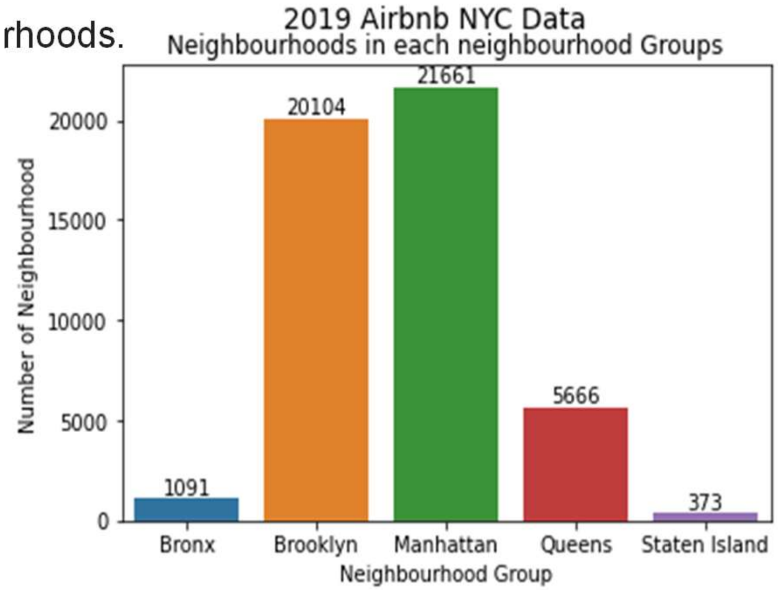
3.1 Number of neighbourhood in each neighbourhood_group:

➔ Manhattan has maximum number of neighbourhood which is followed by Brooklyn.

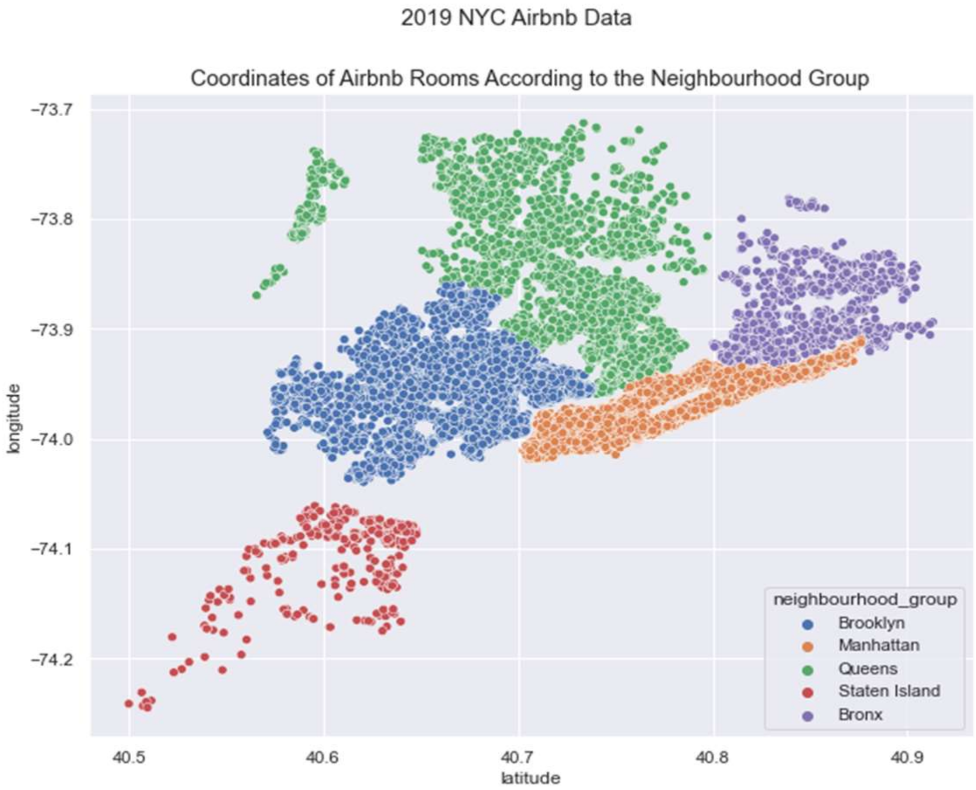
➔ Numbers of neighbourhoods in Manhattan is 21.661K (21661) which is 44.3%.

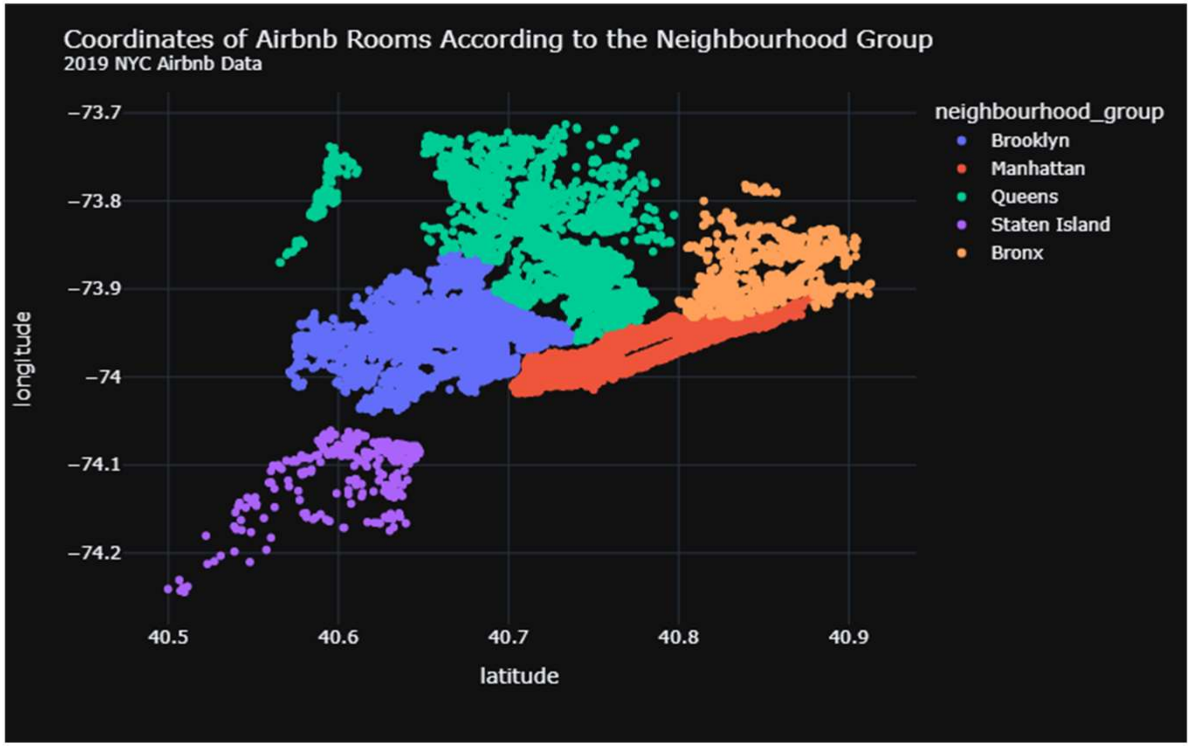
➔ Bronx and Staten Island has least Numbers of neighbourhoods.





3.2 Coordinates of Neighborhood Groups:

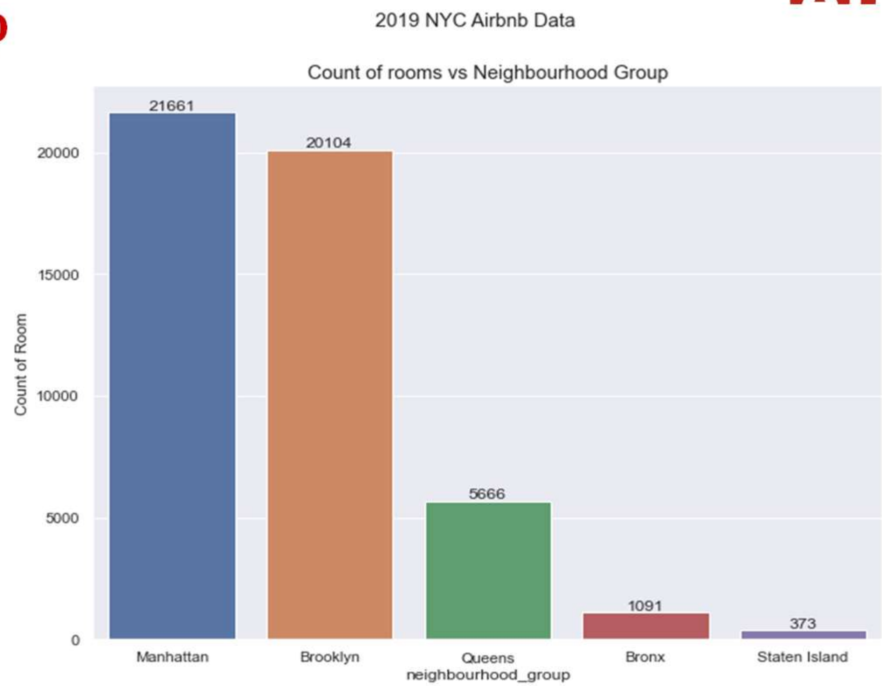




- ➔ Bronx and Staten Island have less room than the others.
- ➔ The room densities of Brooklyn and Manhattan are distributed balanced in their regions.

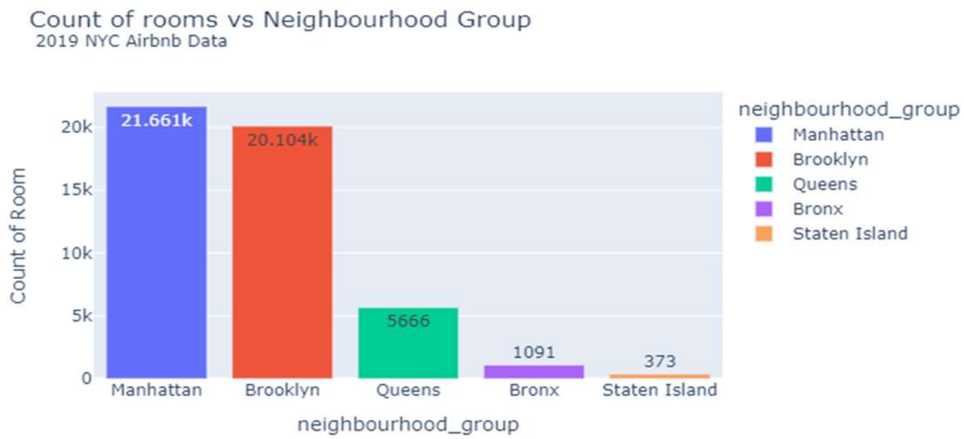
3.3 Count Of Rooms in every Neighbourhood Group

	neighbourhood_group	Count of Room
0	Staten Island	373
1	Bronx	1091
2	Queens	5666
3	Brooklyn	20104
4	Manhattan	21661



➔ We can see that Manhattan and Brooklyn have 21.66K(21661) and 20.104K(20104) rooms in compare to others.

➔ We can observe from above scatter plots and bar plots that region(spread) of Manhattan and Brooklyn have less area than other neighbourhood groups but number of rooms are maximum. It means that density of rooms is higher in Manhattan and Brooklyn.

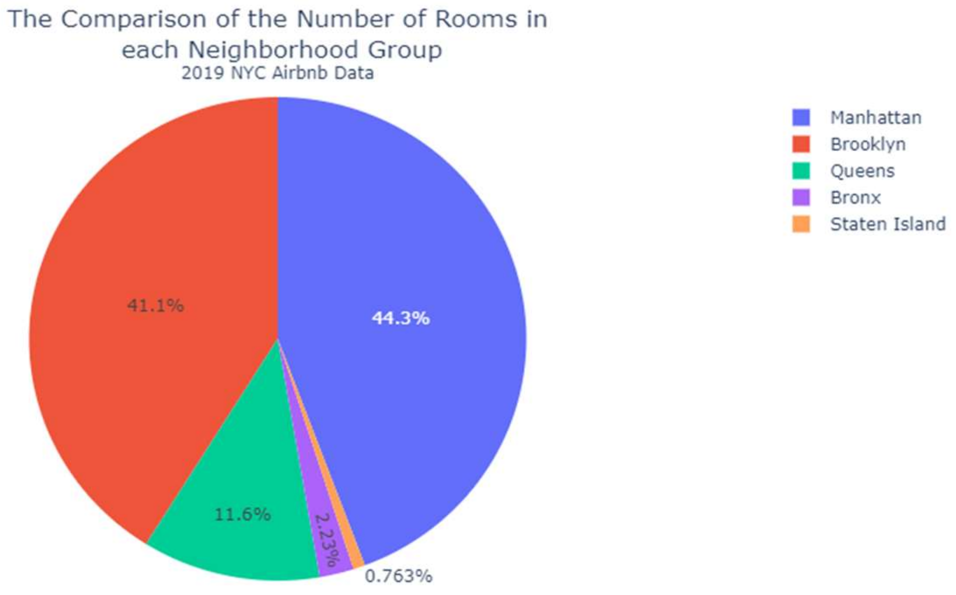


3.4 Percentage of Rooms in Each Neighborhood Group

➔ There are almost 50000 rooms in our data set.
➔ As we want to find the number of rooms and compare with each other, first we draw a pie chart and then we summarize in the table to provide clear difference.

➔ The results illustrate that the rooms in Manhattan and Brooklyn constitute the huge majority, i.e., the sum of these two percentage is equal to 85.42%.

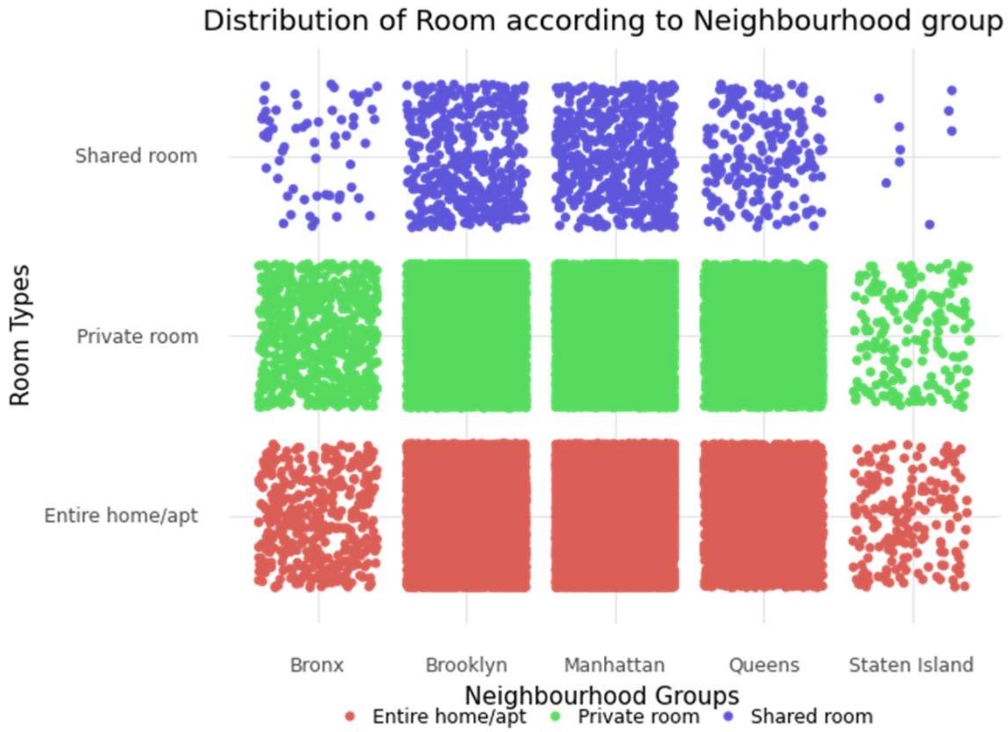
	neighbourhood_group	Count of Room	Percentage of Rooms
0	Bronx	1091	2.23%
1	Brooklyn	20104	41.12%
2	Manhattan	21661	44.3%
3	Queens	5666	11.59%
4	Staten Island	373	0.76%



Now, let's check for the distribution of room types across all neighbourhood groups of NYC!:

➔ We can infer there's is very less shared room throughout NYC as compared to private and Entire home/apt.

	neighbourhood_group	room_type	Number of rooms	Percentage of Rooms
0	Bronx	Entire home/apt	379	0.78%
1	Bronx	Private room	652	1.33%
2	Bronx	Shared room	60	0.12%
3	Brooklyn	Entire home/apt	9559	19.55%
4	Brooklyn	Private room	10132	20.72%
5	Brooklyn	Shared room	413	0.84%
6	Manhattan	Entire home/apt	13199	26.99%
7	Manhattan	Private room	7982	16.32%
8	Manhattan	Shared room	480	0.98%
9	Queens	Entire home/apt	2096	4.29%
10	Queens	Private room	3372	6.9%
11	Queens	Shared room	198	0.4%
12	Staten Island	Entire home/apt	176	0.36%
13	Staten Island	Private room	188	0.38%
14	Staten Island	Shared room	9	0.02%



➔ Distribution of all types rooms in Staten Island and Bronx is very less then others neighbourhood groups.

3.6 Relationship between room type and Price:

➔ Before starting my analysis, I checked for the outlier points in this dataset and I take the quantile 1 and 3 as references.

```
qtl1 = airbnb_df.price.quantile(0.25)
qtl3 = airbnb_df.price.quantile(0.75)
iqr = qtl3 - qtl1

lower = qtl1 - iqr * 1.5
upper = qtl3 + iqr * 1.5

lower,upper
```

```
(-90.0, 334.0)
```

```
from IPython.display import display, HTML
display(HTML("<style>.container { width:100% !important; }</style>"))
outlier = airbnb_df[(airbnb_df['price'] <= lower) | (airbnb_df['price'] >upper)][['neighbourhood_group','neighbourhood','price']]
outlier = outlier.sort_values(by = 'price', ascending = False)
outlier.head(10)
```

➔ ***Top 10 Outlier prices:***

When I analyzed the lower and upper bound of the non-outliers data, the lower bound was obtain as minus 90. In the given data set, as I considered the price of the airbnb room, there is no negative price.

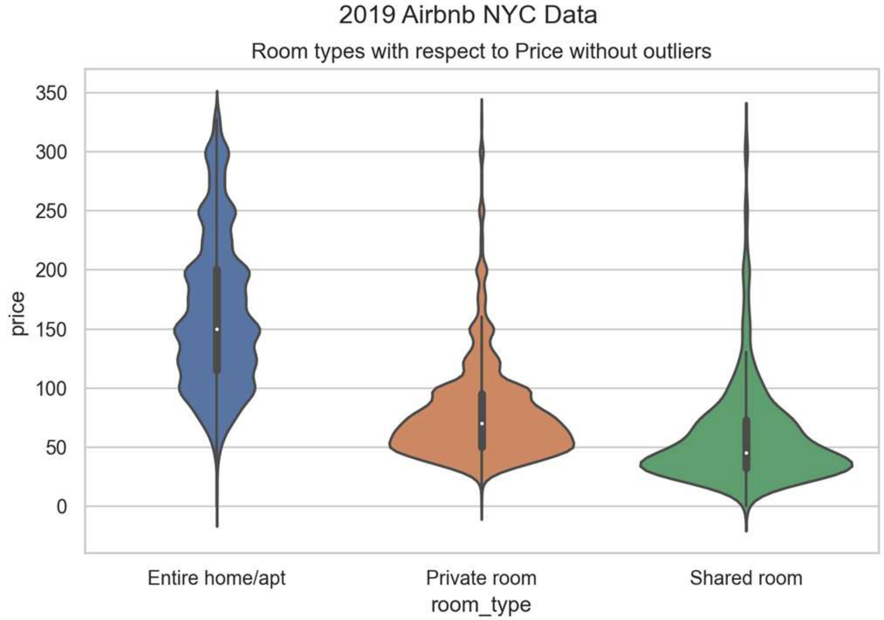
	neighbourhood_group	neighbourhood	price
9151	Queens	Astoria	10000
17692	Brooklyn	Greenpoint	10000
29238	Manhattan	Upper West Side	10000
6530	Manhattan	East Harlem	9999
12342	Manhattan	Lower East Side	9999
40433	Manhattan	Lower East Side	9999
30268	Manhattan	Tribeca	8500
4377	Brooklyn	Clinton Hill	8000
29662	Manhattan	Upper East Side	7703
42523	Manhattan	Battery Park City	7500

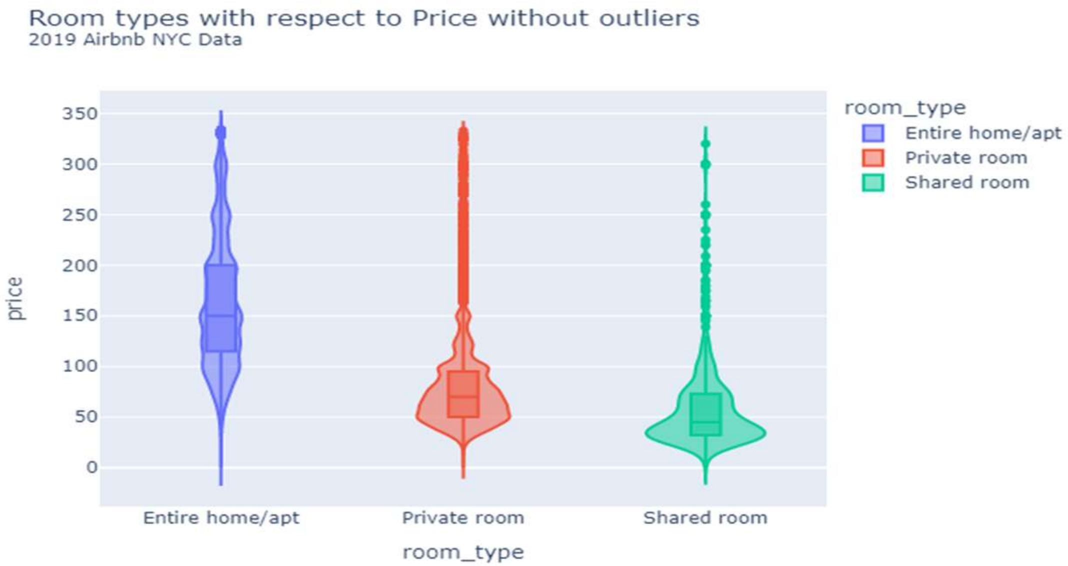
➔ For this reason, I only consider the upper bound. The upper bound address the 334. This means that, if the price value is greater than 334, it becomes an outlier values. In this data set, there are 2972 outliers and the top ten with the highest price is listed as sideward.

	neighbourhood_group	room_type	neighbourhood	price
2158	Manhattan	Entire home/apt	Greenwich Village	334
38688	Manhattan	Entire home/apt	Theater District	334
38683	Manhattan	Entire home/apt	Chelsea	334
38782	Manhattan	Entire home/apt	West Village	334
43351	Manhattan	Entire home/apt	Midtown	334
13798	Brooklyn	Entire home/apt	Williamsburg	333
19641	Manhattan	Private room	Lower East Side	333
28062	Manhattan	Entire home/apt	Upper West Side	333
32976	Manhattan	Entire home/apt	Midtown	333
45303	Manhattan	Entire home/apt	Chelsea	333

Violin Plot:

RoomType vs Price:





From above violin plot we can see that:

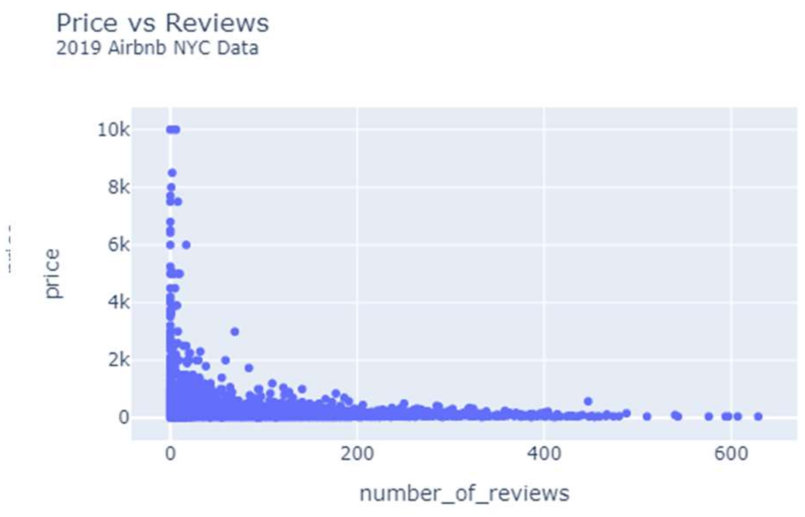
➔ Average cost of stay in Entire home/apt is near about 153USD

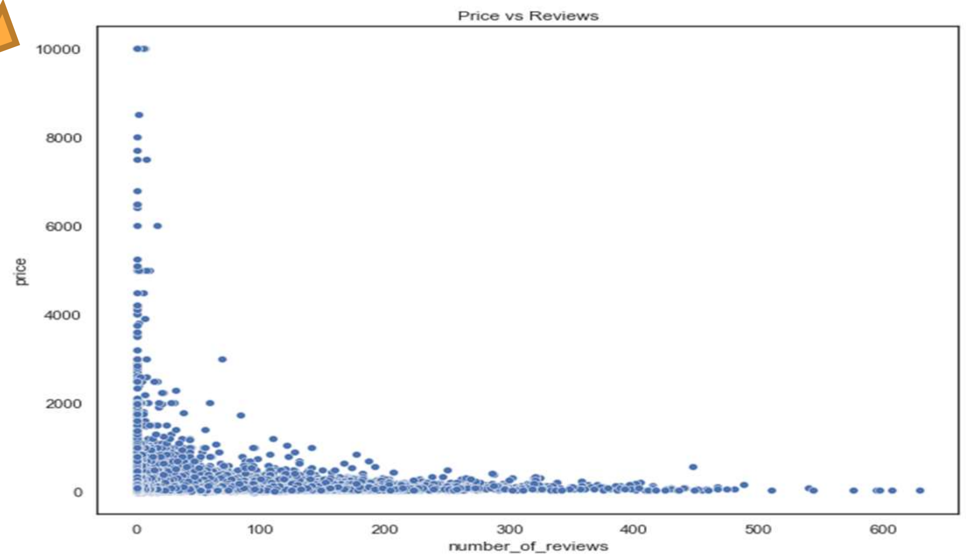
➔ Average cost of Private room type is near about 60 USD

➔ Average cost of Shared room type is very less which is 50 USD

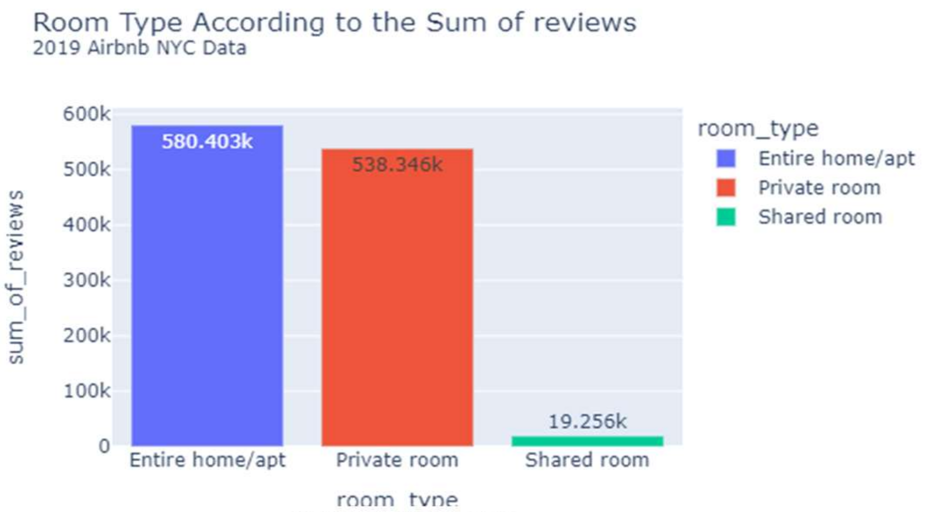
3.7. Price According to the reviews:

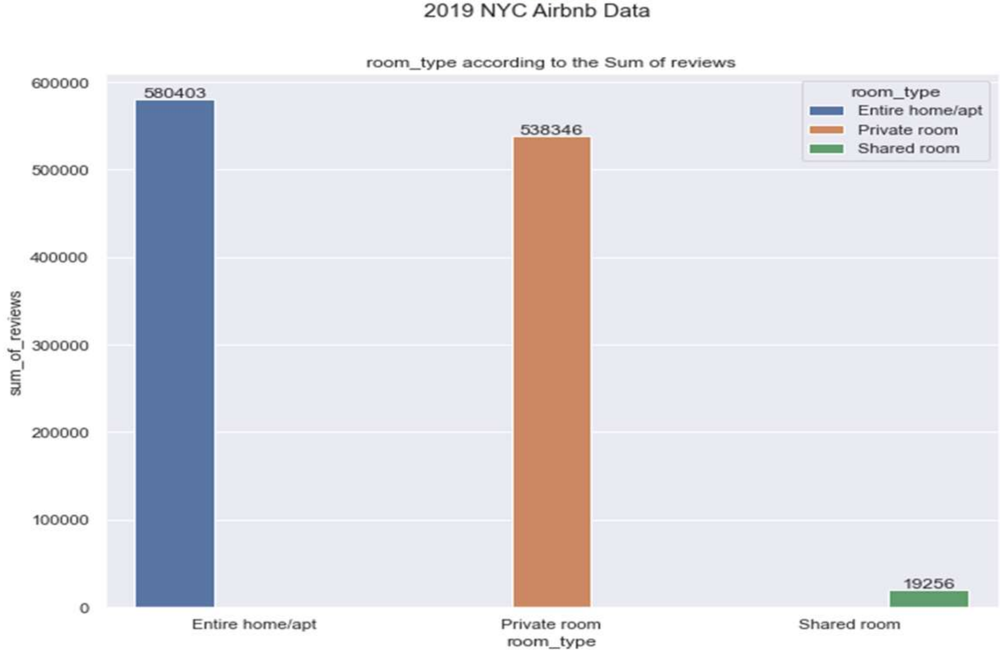
➔ We can observe from sideward scatterplots ,the first choice of customers is low price because spread of the points is very dense in lower region.





3.8. Visualization of Room Type according to the reviews to know the choice of customers for room type:





Conclusion:

➔ Number of reviews in Entire home/apt is 580.403K (580403)

➔ Number of reviews in Private room is 538.346K(538346)

➔ We can observe why the number of Entire home/apt and Private room is maximum because first preference of customers is Entire home/apt and second is Private room.

3.9 Minimum, Maximum and Average Price:

Conclusion:

➔ Average price in Manhattan is 145 USD

➔ Average price in Brooklyn is 106 USD

➔ We can observe from above that: Rooms in Manhattan are very costly.

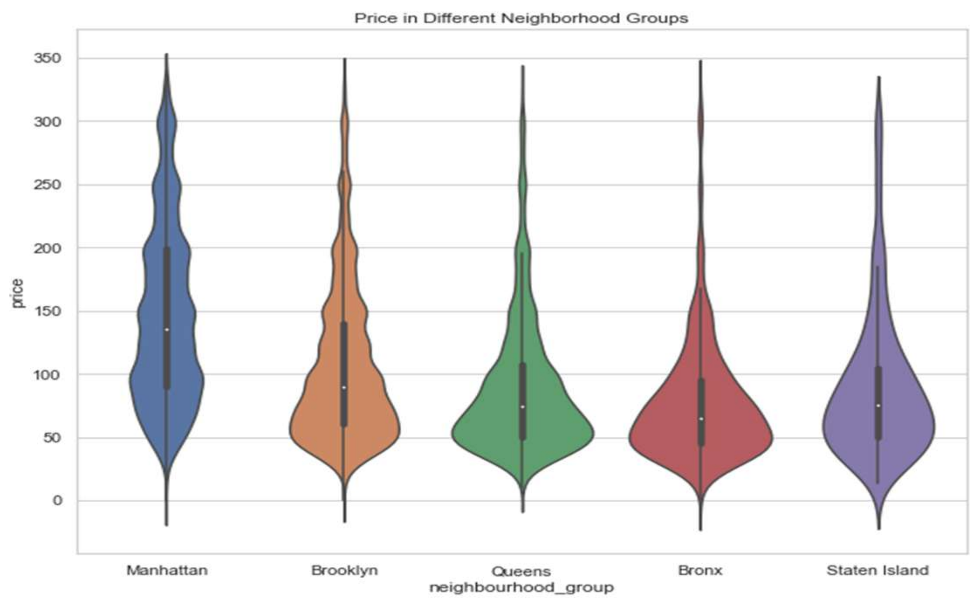
➔Cost of Rooms in Brooklyn are less than Manhattan

➔ Average price in Bronx is very low which is near about 77 USD

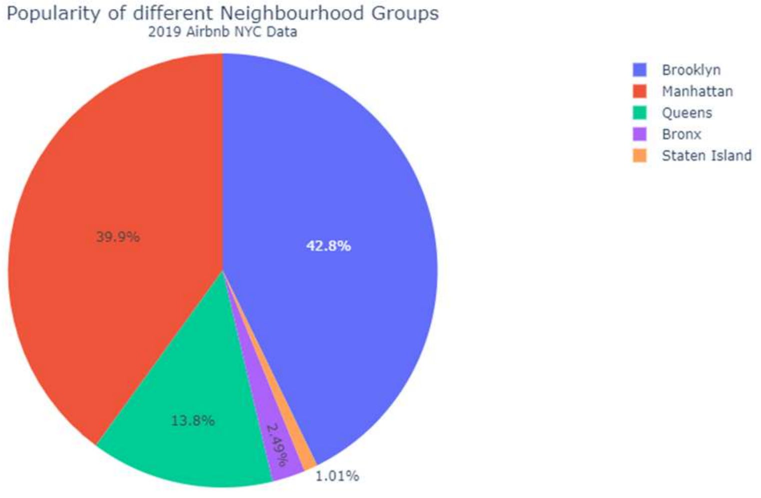
➔ Bronx , Queens and Staten Island are very cheap neighbourhood_group.

	neighbourhood_group	price		
		mean	min	max
0	Bronx	77.365421	0	325
1	Brooklyn	105.699614	0	333
2	Manhattan	145.952835	0	334
3	Queens	88.904437	10	325
4	Staten Island	89.235616	13	300

Price in Different Neighborhood Groups



3.10 Most popular neighbourhood_group :



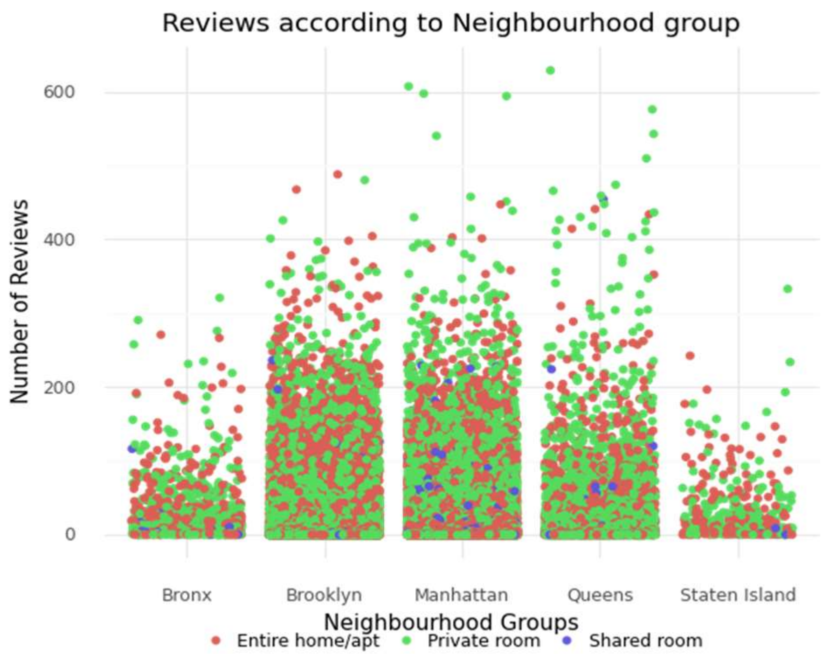
Conclusion:

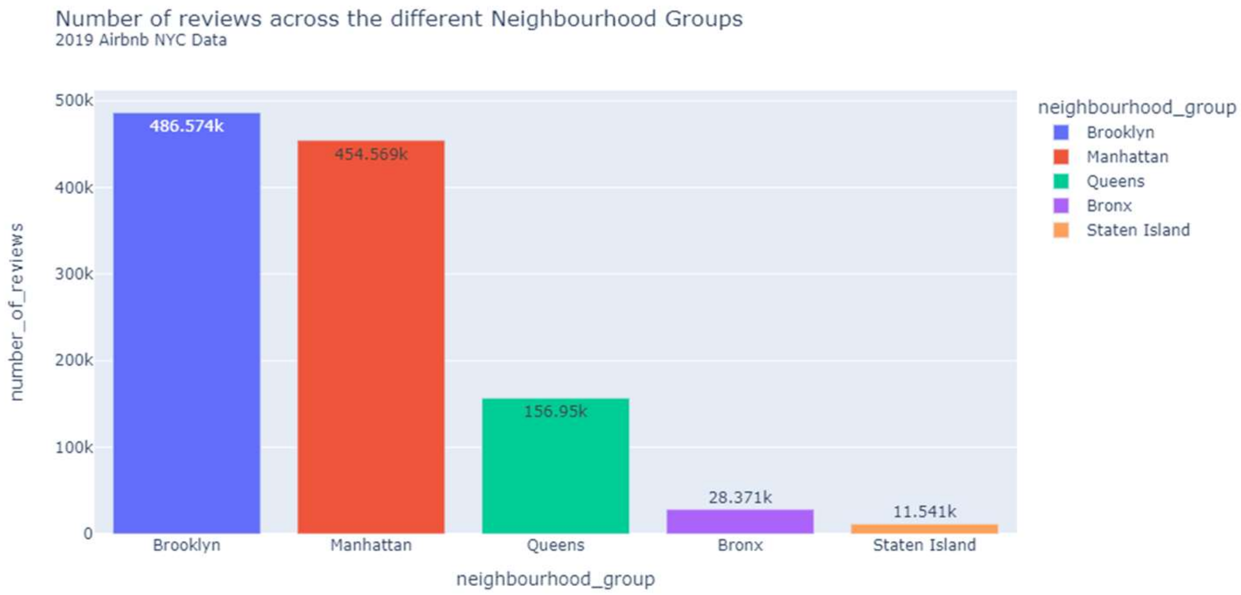
➔ From above plot we can observe that in each neighbourhood groups , Entire home/apt and Private room have maximum numbers of reviews.

➔ Brooklyn and Manhattan have maximum numbers of reviews which are followed by the Queens.

➔ It seems that the first choice of customers is Entire home/apt and Private room.

➔ Brooklyn and Manhattan are most popular neighbourhood group

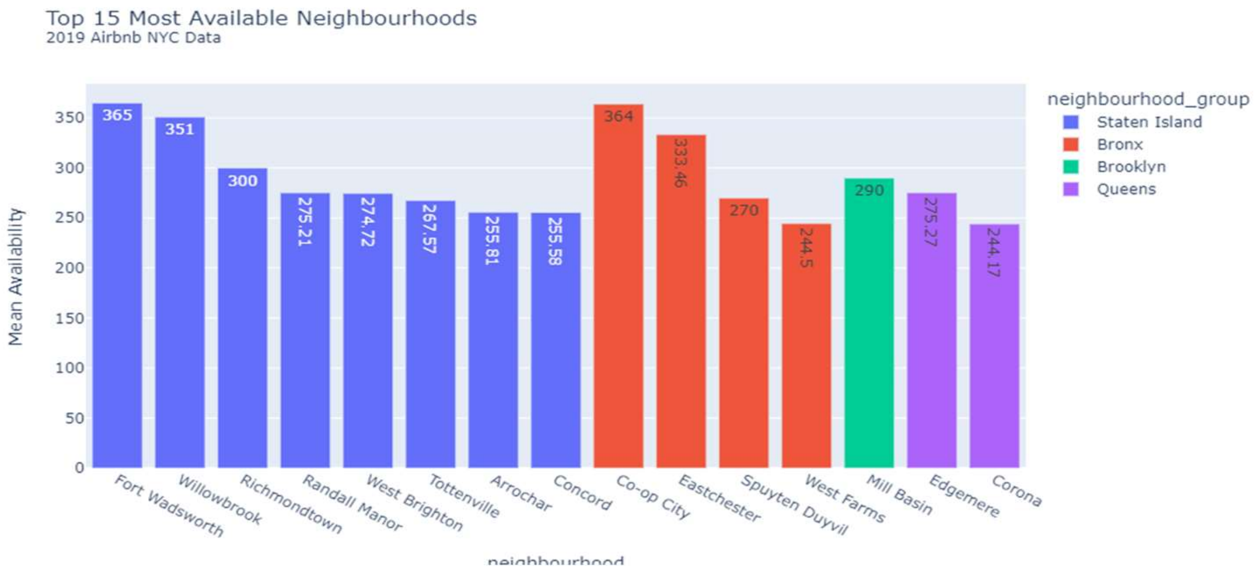




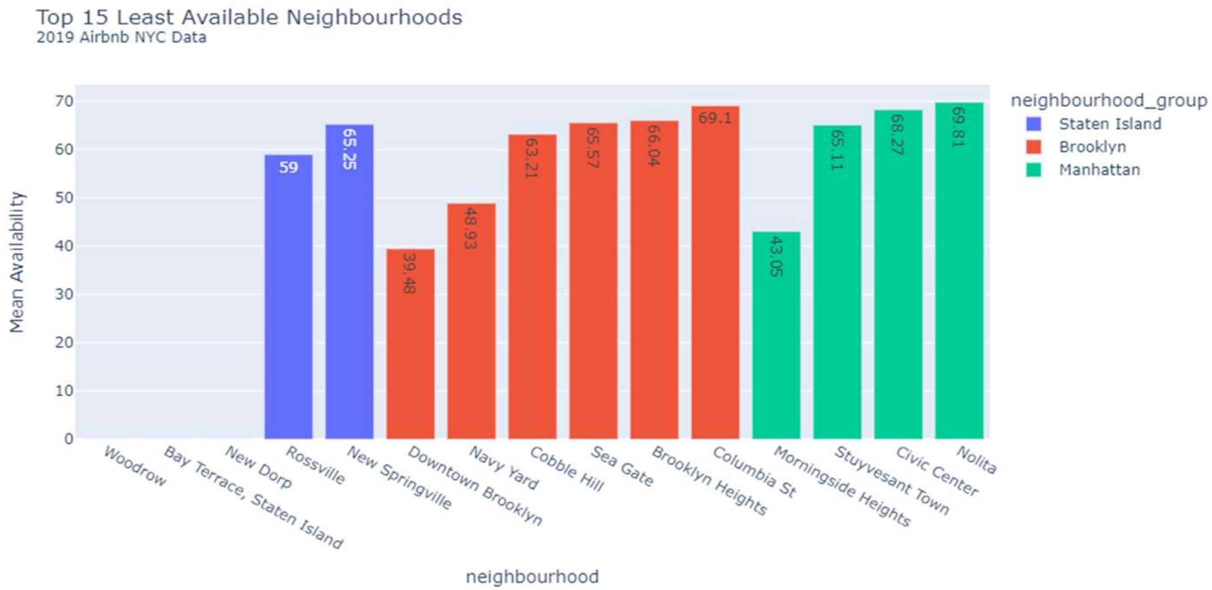
3.11. The Most and Least Expensive Neighborhoods

Conclusion:
➔By using average availability of the rooms, the graph shows that Staten Island is the most available neighborhood group in the top 15.

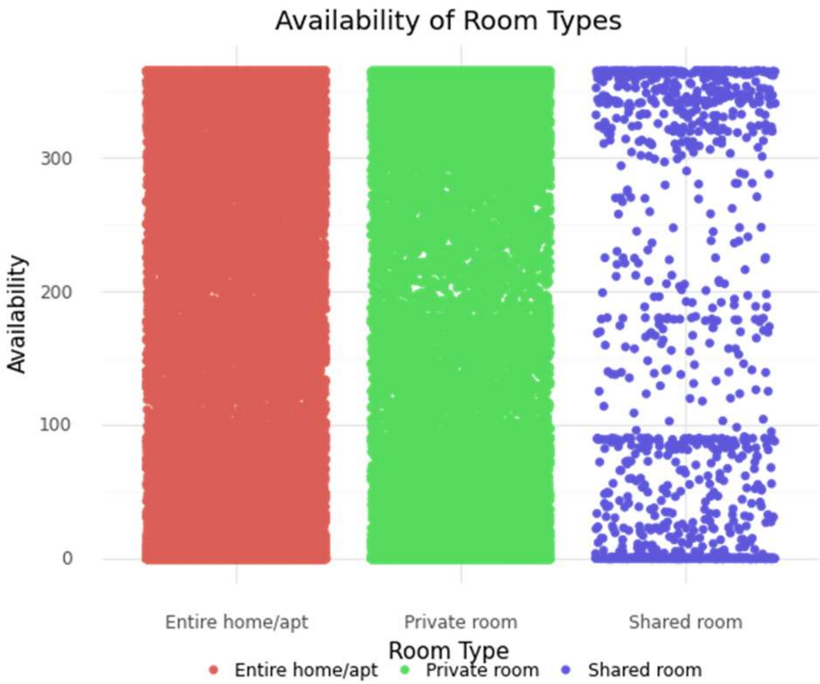
➔Manhattan, on the other hand, does not have any neighborhood in the top 15.

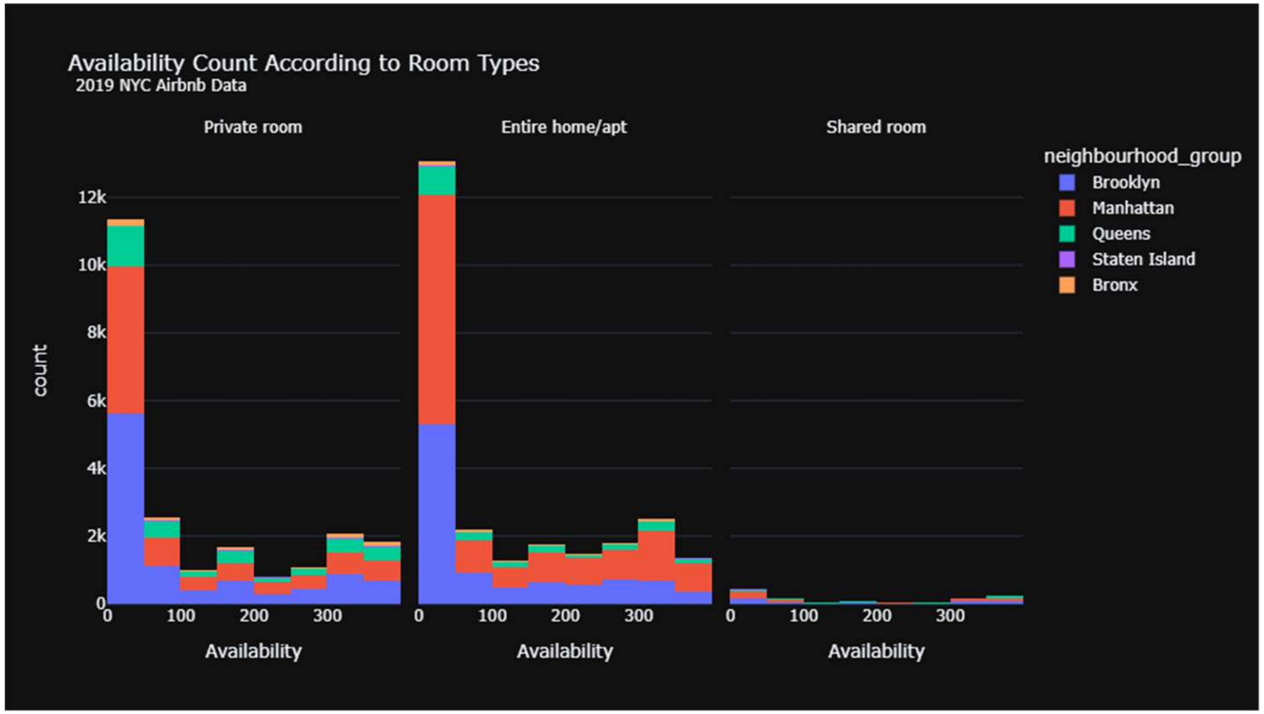


Conclusion:
➔ There are many neighborhoods in the data set with zero availability.
➔ Woodrow(Staten Island) , (Bay Terrace, Staten Island(Staten Island)) and New Dorp (Staten Island) do not have availability.



3.13. Availability of Room Types According to the Neighborhood Groups:





➔ Entire home/apt and private room have homogeneous distribution of availability, while the shared room accumulates on the edge of the intervals.

➔ **To make more analysis, we also plot histogram**.

➔ In this histogram, we want to analyze the availability of room types according to the neighborhood groups. It can be said that entire home/apt and private room can be reached every day in a year, whereas, shared room is not always accessible.

3.14. The most and leasavailability of Rooms at every neighbourhood groups:

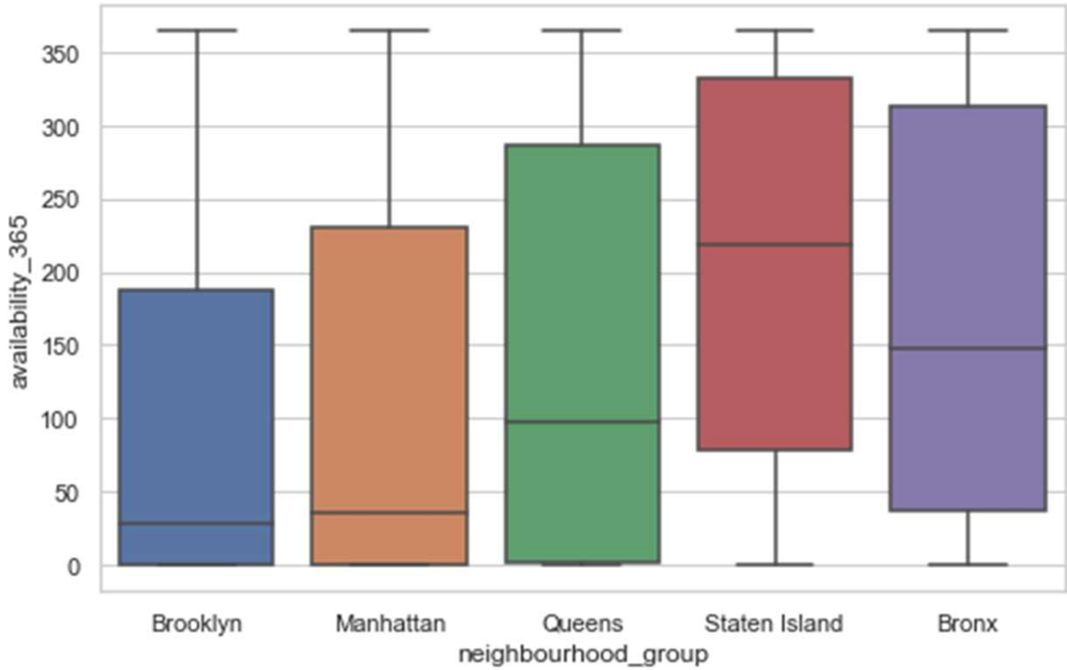
Conclusion

➔ Staten Island has most has most availability of room all over the year.

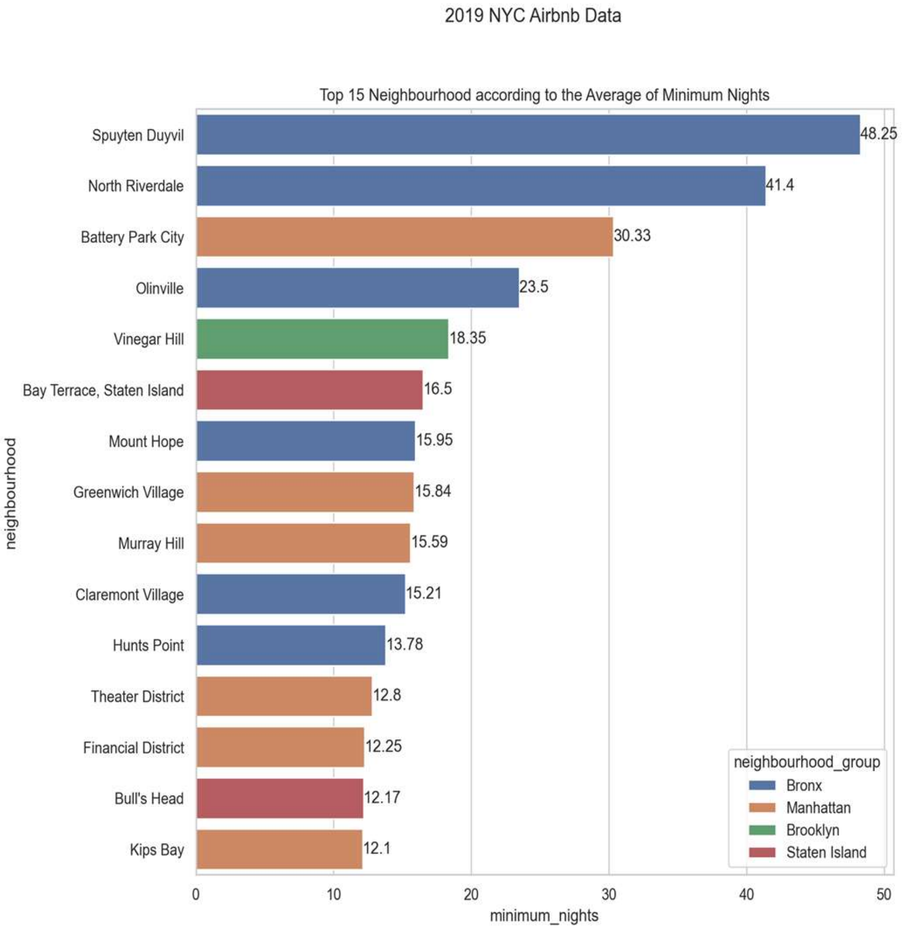
➔ Brooklyn has the least availability of room all over the year.

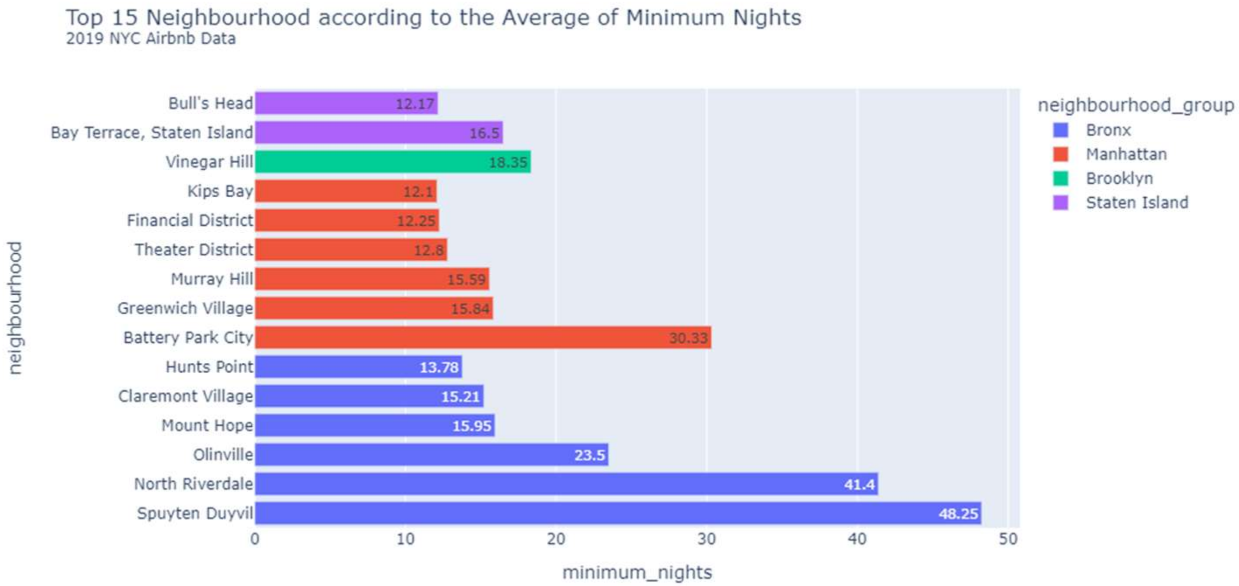
➔ Looking at the sideward categorical box plot we can infer that the listings in Staten Island seems to be more available throughout the year to more than 300 days.

➔ On an average, these listings are available to around 210 days every year followed by Bronx where every listings are available for 150 on an average every year.



3.15 Minimum Nights and Neighborhood Relationship:



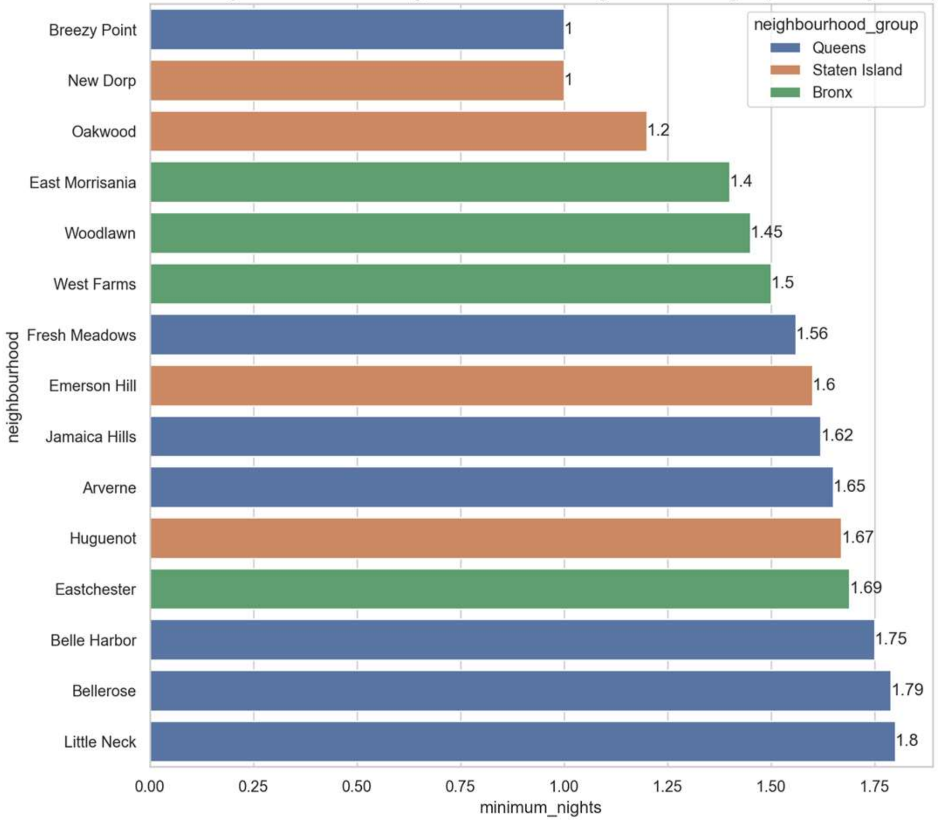


When we search information about NYC to make more clear:

➔ We realize that the most of the landmarks are located in Manhattan. Thus, we expect the more staying in this place. To check the assumption, we order the neighborhoods according to their Average of minimum nights. The above plot shows that neighborhoods in Manhattan are not included daily hosting.

2019 NYC Airbnb Data

Least Neighbourhood according to the sum of Average Minimum Nights(Descending Order)



Least 15 Neighbourhood according to the Average of Minimum Nights(Descending Order)
2019 NYC Airbnb Data

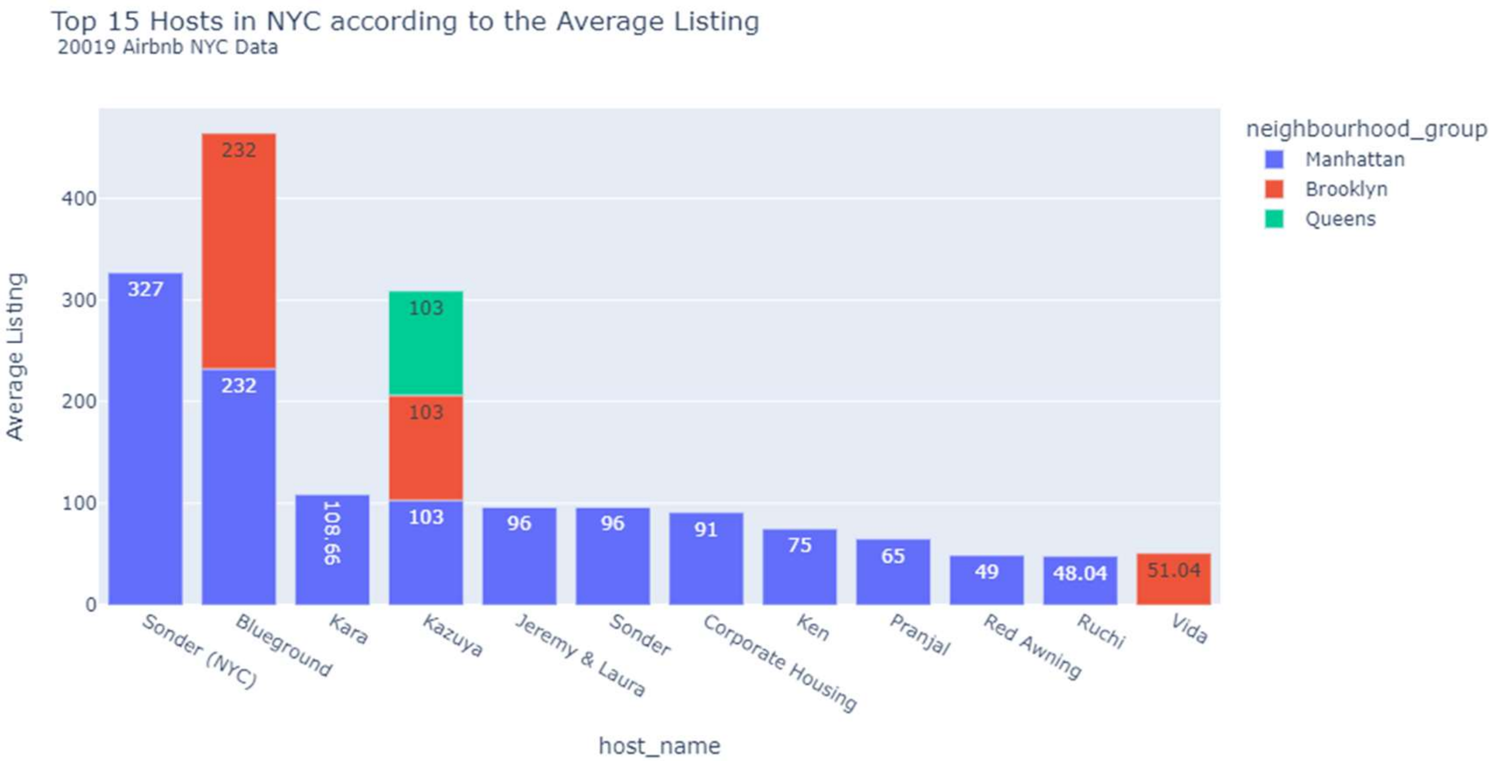
When I search information about NYC to make more clear:

➔ I realize that the most of the landmarks are located in Manhattan. Thus, I expect the more staying in this place. To check the assumption, I order the neighborhoods according to their Average of minimum nights. The above plot shows that neighborhoods in Manhattan are not included daily hosting.

3.16 The Most Popular Hosts in NYC Airbnb:

➔ Like we find the most popular neighborhoods, we can also determine the most popular host in NYC according to listing counts.

	neighbourhood_group	host_name	Average Listing
0	Manhattan	Sonder (NYC)	327.000000
1	Brooklyn	Blueground	232.000000
2	Manhattan	Blueground	232.000000
3	Manhattan	Kara	108.659259
4	Queens	Kazuya	103.000000
5	Brooklyn	Kazuya	103.000000
6	Manhattan	Kazuya	103.000000
7	Manhattan	Jeremy & Laura	96.000000
8	Manhattan	Sonder	96.000000
9	Manhattan	Corporate Housing	91.000000
10	Manhattan	Ken	75.000000
11	Manhattan	Pranjal	65.000000
12	Brooklyn	Vida	51.037736
13	Manhattan	Red Awning	49.000000
14	Manhattan	Ruchi	48.040000



➔ In top 15 listings , most of the neighbourhoods are from Manhattan.

➔ Blueground has equal average listings (232) through the Manhattan and Brooklyn

3.17 Variation in price according to the last review year:

Conclusion:

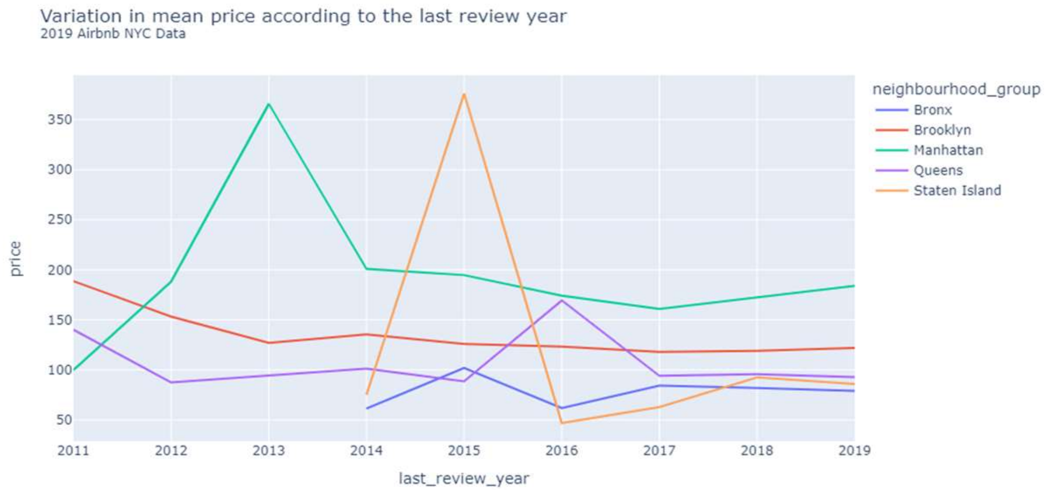
Null Values are –10052

From above line plot , hist plot and bar plot we can observe that:

➔ All over the year in 2011 average price was lowest in Manhattan.

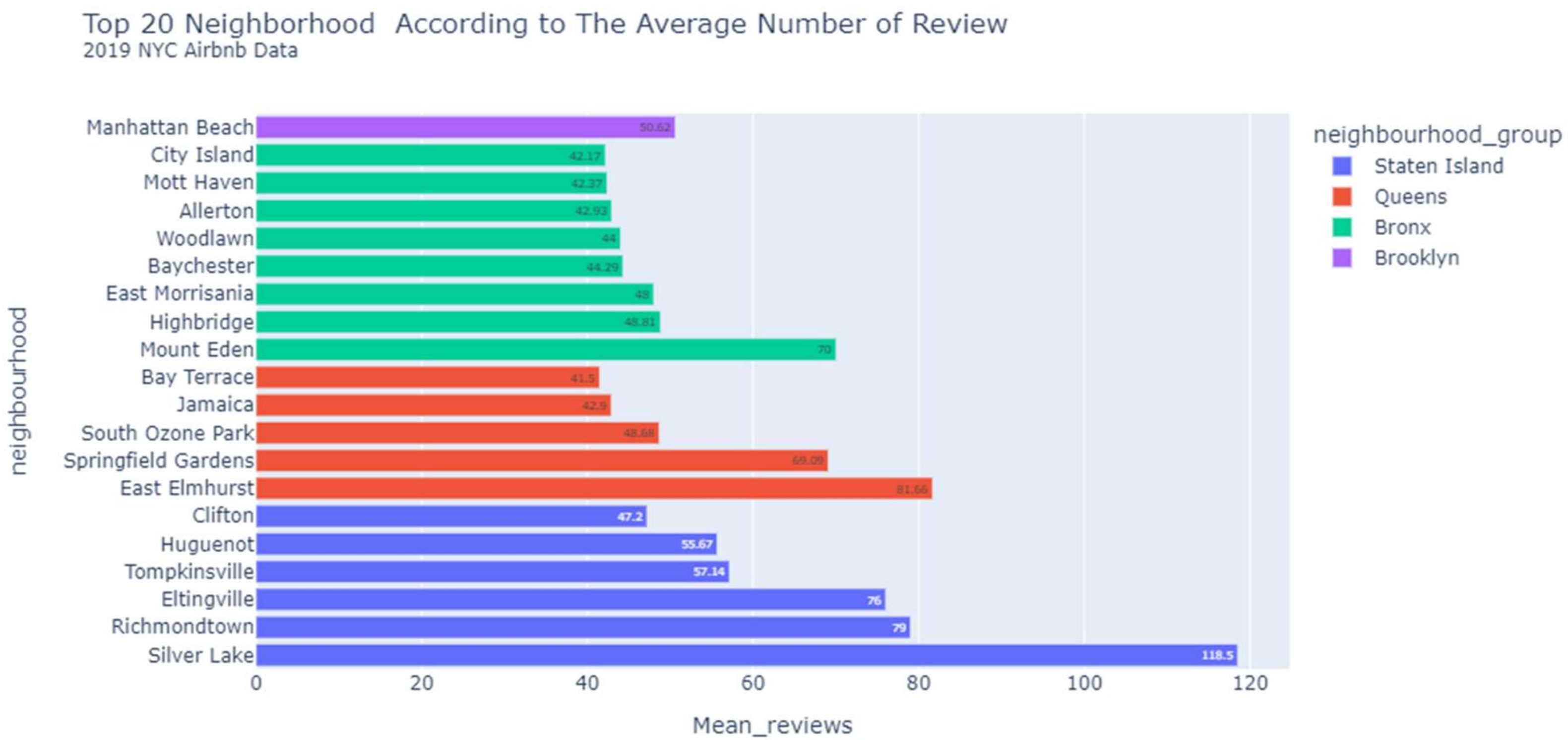
➔ In 2013 Manhattan had highest price.

➔ In 2015 Queens had very highest price.



3.18 The Average Number of Reviews to the Each Neighboorhood:

The another analysis can be made by using number of reviews:



- ➔ The results show that Bronx and Staten Island take the most of the reviews.
- ➔ On the other hand, there is no neighborhood from Manhattan in the top 20.

3.19 WordCloud

➔ Like the numerical values, Airbnb data includes verbal information such as name. By using this information, we can obtain the most used words in “name” column which describes the room features.

➔ With the WordCloud library, I tried to make the wordcloud process reproducible. After geting the data frame of the frequencies, it shows the wordcloud plot.

Conclusion:

➔ As we can see in the plot,"private","room","heart","nyc" and "apartment" words are the most common words in the name column. It means that most of the customers of Airbnb looks for the private rooms, so that these listings have these words in their names.

➔ As we can see from the plot that “brooklyn” and “manhattan” words are common in the name of the listings. We can infer that Brooklyn and Manhattan would have more listing than the others.

➔ bedroom comes with different words like "one bedroom" ,"private bedroom","bedroom apt" and "bedroom apartment". These are the most common words in the name column.

3.20 Price Group Analysis of Neighborhood Groups

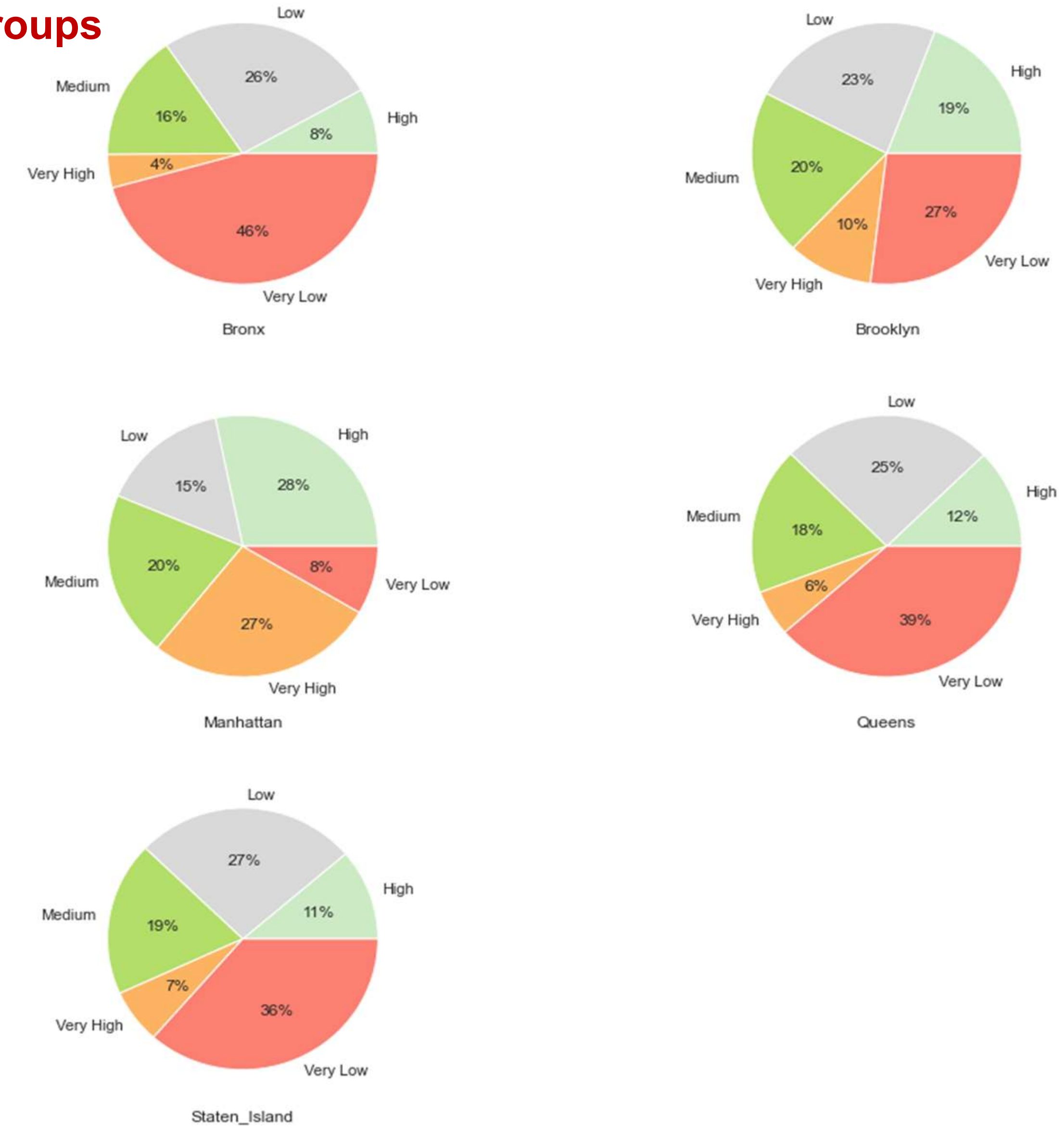
➔ By using quantile function, I divide price interval into five. Then, we define values in this intervals as "Very low", "Low", "Medium", "High" , and "Very high". Then by using these categorical values, I prepare pie chart and for each neighborhood group.

```
quant = list(df.price.quantile([0.2,0.4,0.6,0.8,1]))
quant
```

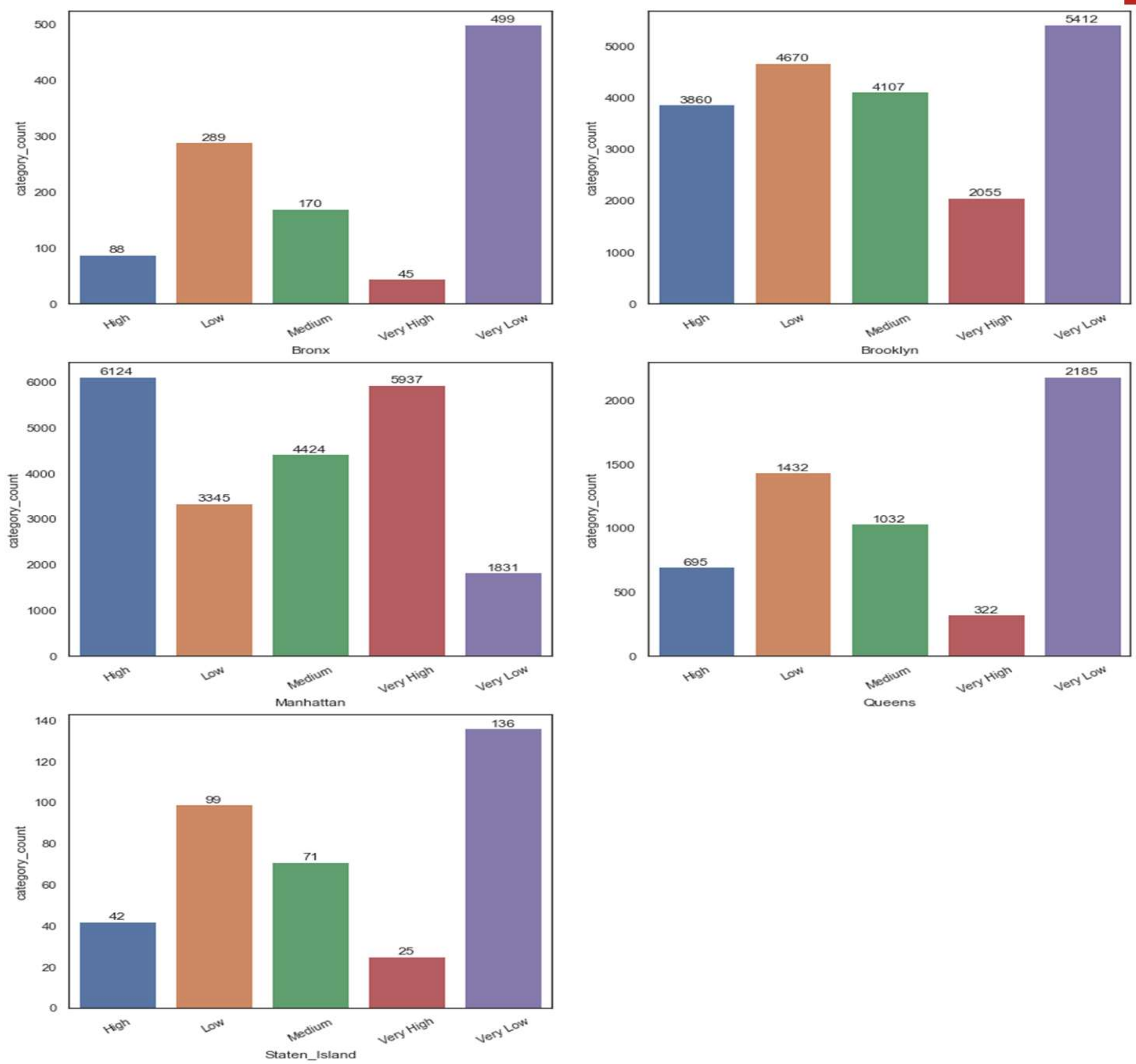
```
[60.0, 90.0, 130.0, 200.0, 10000.0]
```

```
def group_price(x):
    try:
        if 0 <= x <= quant[0]:
            return 'Very Low'
        elif quant[0] < x <= quant[1]:
            return 'Low'
        elif quant[1]< x <= quant[2]:
            return 'Medium'
        elif quant[2] < x <= quant[3]:
            return 'High'
        else:
            return 'Very High'
    except:
        return np.nan
```

```
df['price_group'] = df['price'].apply(group_price)
```



3.20 Price Group Analysis of Neighborhood Groups



We summarize the results as follow:

➔ The most of the rooms in Bronx, Queens and Staten Island has very low price.

- Bronx -- 46% (499)
- Queens -- 39% (2185)
- Staten Island -- 36% (136)

➔ The rooms with very low, low and medium prices in the Brooklyn are almost distributed equal percentage.

➔ The very high price group in Manhattan has higher percentage than the other price groups.

- Manhattan -- 27%

4. Final Conclusion:

➔ In this study, we address the explanatory analysis of the Airbnb data with several key features such as price, neighborhood, neighborhood group, room type, number of reviews, etc. By using these data

➔ We obtain price and neighborhood relationship, i.e., Manhattan is the most expensive Airbnb region when we compare the other neighborhood groups. On the other hand, the least expensive region is Bronx.

➔ Another analysis is conducted by using room type. The results show that the entire home/apt type is more preferable and the others are private room and shared room, respectively.

➔ In terms of listings Manhattan is on the top.

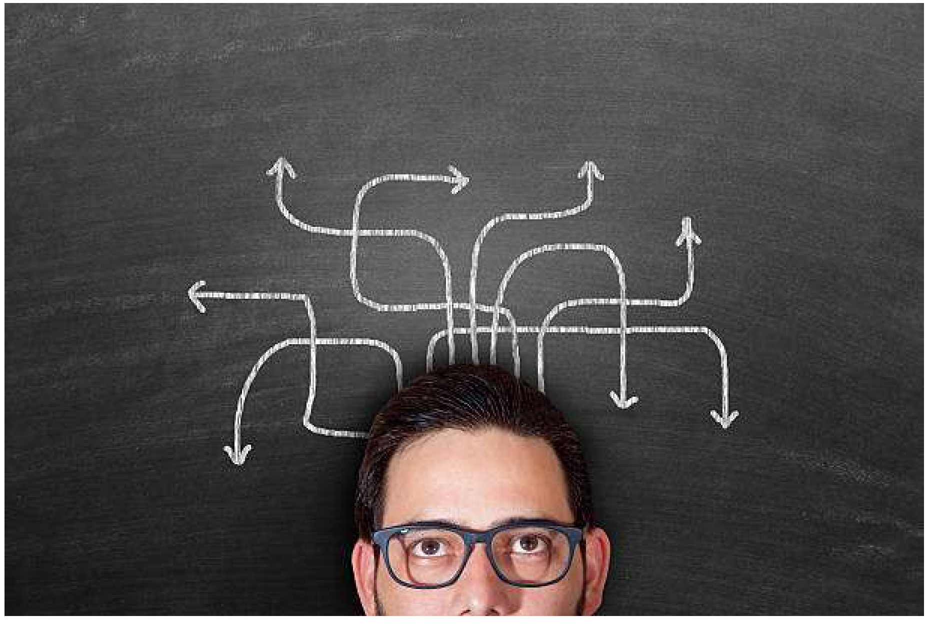
➔ In terms of reviews Staten Island is on the top.

➔ In terms of availability Staten Island has most has most availability of room all over the year

➔ To make a different analysis instead of numerical analysis, we use WordCloud which makes text mining.

➔ Number of reviews are also investigated to find which neighborhoods take the most review according to the neighborhood group.

5. Problem Faced:



- ➔ Collecting meaningful data
- ➔ Selecting the right analytics tool
- ➔ Cleaning the dataset
- ➔ Combining and analyzing
- ➔ Data visualization

Things that We have learned:

- ➔ Unable a business to take raw data and uncover patterns to extract valuable insights**.**
- ➔ Improve customer experience
- ➔ Work with complex datasets


Thank you
Adapt it with your needs and it will capture all the audience attention.