Capstone Project

EDA ON AIRBNB

Airbnb, as in “Air Bed and Breakfast,” is **a service that lets property owners rent out their spaces to travelers looking for a place to stay**. Travelers can rent a space for multiple people to share, a shared space with private rooms, or the entire property for themselves

❑ About Given Data Set:

- We are given a data set from Airbnb which holds values related to the tourists looking for renting a space in New York City.
- Not only hotels but Entire home/apt , Private Rooms or even Shared rooms.
- We here will try to get the most out of this data so that it can prove useful for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

❑ Data Summary:

➢ **This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.**

Understanding Data:

id :indicates the unique id of an entry in data set
name : indicates names of listings area in New York City
host_id :indicates unique id for a host in New York City
host_name : shows the name of the host
neighbourhood_group: indicates names of the cities where these listings present
neighbourhood: displays the names of different neighbouring cities for each neighbourhood_group
latitude: displays latitude values of a listing.
longitude: displays longitude values of a listing.
room_type: indicates the type of room available.
price: indicates the price.
minimum_nights: indicates minimum number of nights spent.
number_of_reviews: indicates total number of reviews for a particular listing.

❑ Data Summary (cont)

reviews_per_month :indicates reviews per month for a particular.
calculated_host_listings_count : shows the count of each host for a certain listing
availability_365 : indicates the number of days a particular listing is available in a year.

These were all the columns given in the Airbnb data set…

Lets now analyze them and try to gather some useful intel !!

❑ Work Flow :

AI

➢So we will divide our work flow into following 3 steps:

EDA will be divided into following 3 analysis.

1) **Univariate analysis:** Univariate analysis is the simplest of the three analyses where the data you are analyzing is only one variable.
2) **Bivariate analysis:** Bivariate analysis is where you are comparing two variables to study their relationships.
3) **Multivariate anlysis:** Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables.

❑ Data Cleaning and Manipulation:

AI

➢ We already have collected and understood our data so lets now directly jump into Data Cleaning and Manipulation:

Upon basic analysis we found a total of 4 columns with Null or missing values:

- **name**
- **host_name**
- **last_review**
- **reviews_per_month**

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              48895 non-null  int64
 1   name                            48879 non-null  object
 2   host_id                         48895 non-null  int64
 3   host_name                       48874 non-null  object
 4   neighbourhood_group             48895 non-null  object
 5   neighbourhood                   48895 non-null  object
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object
 9   price                           48895 non-null  int64
 10  minimum_nights                  48895 non-null  int64
 11  number_of_reviews               48895 non-null  int64
 12  last_review                     38843 non-null  object
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64
 15  availability_365                48895 non-null  int64
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

From this part we can conclude that there are a lot of NULL values in **last_reviews** and **reviews_per_month** column .... there are few in **name** and **host_name** column , **id** is should be unique for every element.

```
[ ] # Lets check the total number of null values in each column:
    df.isnull().sum()
```

```
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
```

Total Missing or Null values in each column

Here, we can see the exact number of null values from each column

❑ Data Cleaning and Manipulation:

AI

Checking for duplicates:

Lets check for any duplicate data just in case !

```
[ ] identical_data = df.duplicated()
    print(identical_data.sum())
```

0

None Found

Since there were a lot of null values i.e ≈10052 in 'last_review' column so deleting it was a fair step as we also couldn't have gather much of information from this column:

```
[ ] # Deleting:
    del df['last_review']
```

For 'name' and 'host_name' columns we replacing Null values with "Unknown" and "Host_Unknown"

```
# Replacing:

df['name'].fillna('Unknown',inplace=True)
df['host_name'].fillna('Host_Unknown',inplace=True)
```

Now, about outliers:

Presence of a lot of outliers was detected in 'price' column via box-plot, even the mean was not visible....

Replacing 'reviews_per_month' column's missing values with 0:

```
[ ] # Replacing with 0:
    df['reviews_per_month'] = df['reviews_per_month'].fillna(0)
```

❑ Data Cleaning and Manipulation:

So we tried removing outliers using Quantile approach:

```
[ ]  # Defining 0.1 and 98 percentile values from 'price' column:

     min_threshold , max_threshold = df.price.quantile([0.001 , 0.980])
     min_threshold , max_threshold
```

```
(18.0, 550.0)
```

In this cell of code, we have assigned the lowest and highest limits/threshold and later just selected the values between these thresholds:

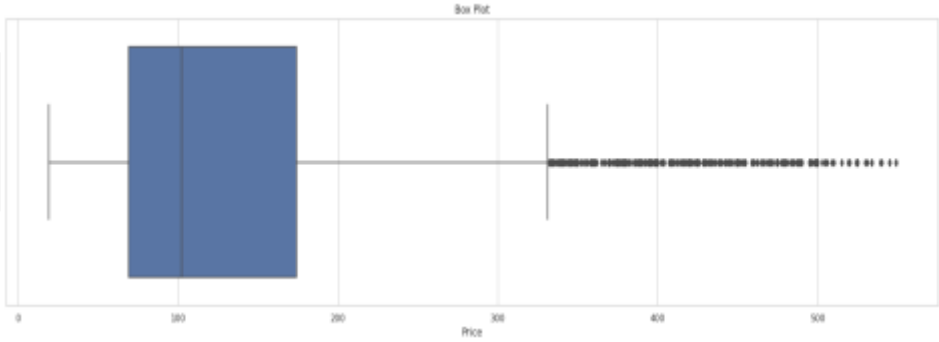
i.e values between 0.1 and 98 percentile were selected from the price column and a new Data Frame was created with these custom 'price' column values

```
[ ]  #Creating a new data frame:

     df2 = df[(df.price>min_threshold)&(df.price<max_threshold)]
     df2.shape
```

```
(47859, 15)
```

Box-Plot now:



Clearly a lot of outliers can still be seen here but now the mean, min, max and other quartile values were visible and that was quite enough for us to inspect the data further and perform EDA on it.

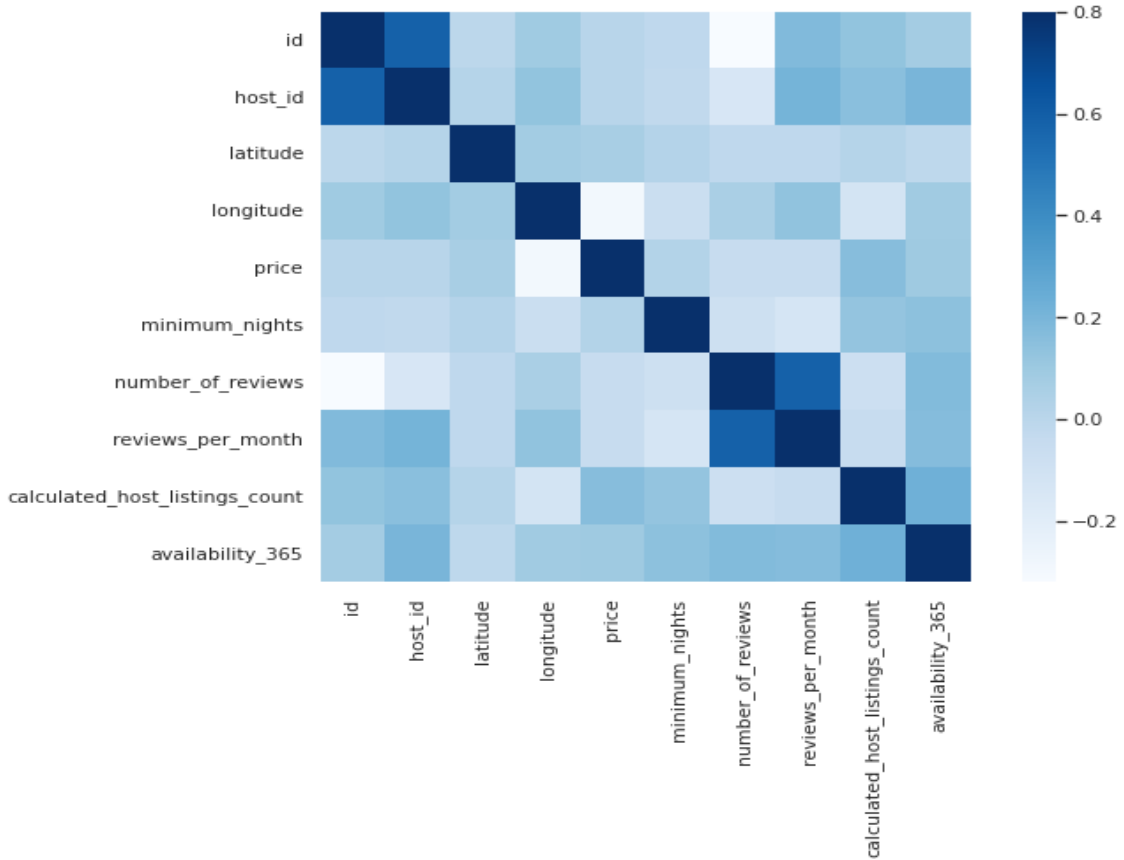
❑ Exploratory Data Analysis: :

Data Correlation with each other:

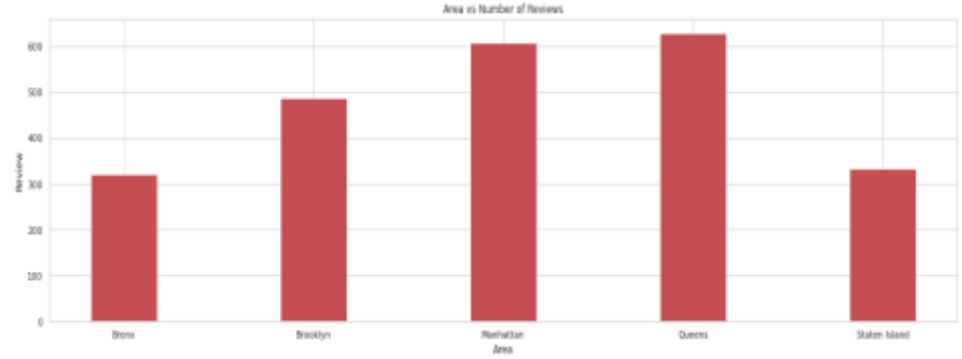
➤ From the heatmap we can see the correlation between different columns that can affect a Airbnb listing.

➤ Also....
There's correlation among host_id to reveiws_per_month & availability_365. And there's noticiable correlation between min_nights to no_of_listings_count & availability_365.

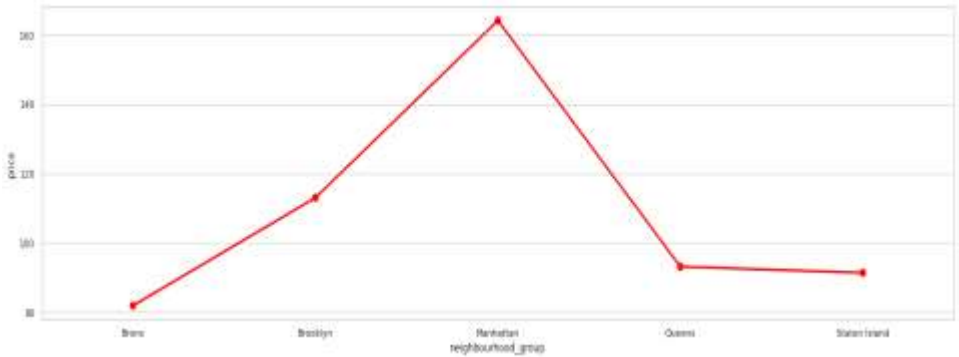
➤ Price also shows some correlation with availability_365 & host_listings_count.
no_of_reviews and reviews_per_month gives almost the same information.



Exploratory Data Analysis: (Bi-Variate Analysis)



Number of Total Reviews per Neighbourhood Group



AVERAGE PRICE OF DIFFERENT NEIGHBOURHOOD GROUPS

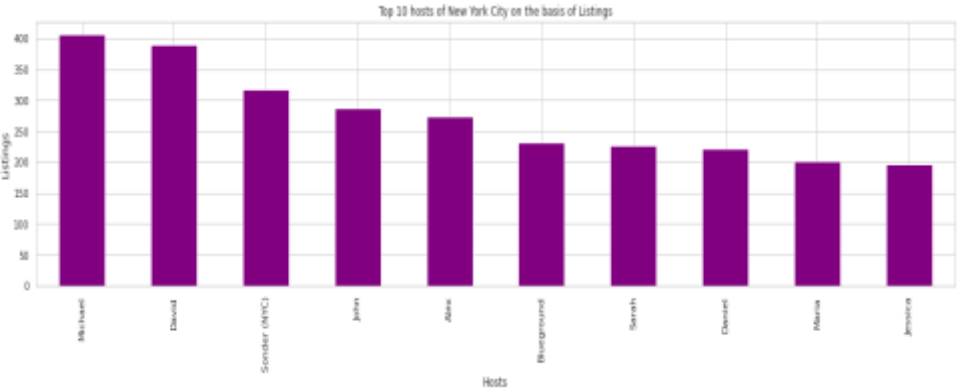
Conclusions:

- **People like staying in Queens the most as they have the highest total reviews ≈630**
- **Manhattan is the 2nd most reviewed city with a total review score of ≈608**
- **Brooklyn is at 3rd with ≈489 Total reviews.**
- **Then Staten island with ≈334 Total reviews.**
- **And Bronx is the least reviewed city in the entire New York.**

Conclusions:

- **Cost of living in Manhattan is the highest i.e ≈165**
- **Then in Brooklyn which is ≈114**
- **Queens comes at third place with ≈94 cost of living**
- **Then Staten island with ≈92**
- **And it is the cheapest in Bronx which is ≈82**

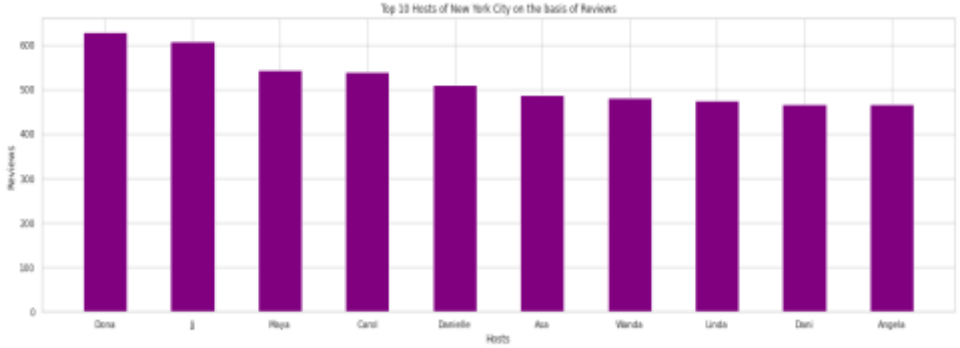
Exploratory Data Analysis: (Bi-Variate Analysis)



TOP 10 Hosts on the basis of Listings

Conclusions:

- **Michael holds the most number of listings properties ≈408 in entire New York City follower by David , Sonder (NYC) and others.**



TOP 10 Hosts on the basis of Reviews

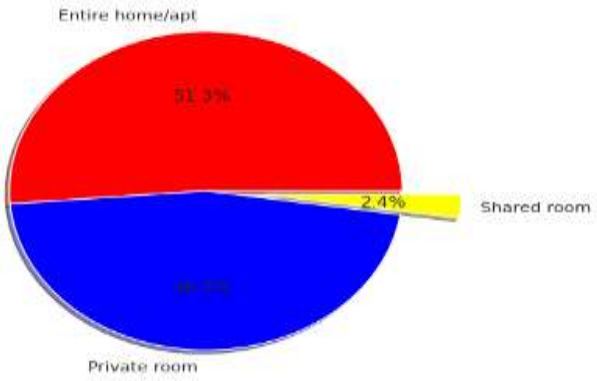
Conclusions:

- **However even after holding the most listings Dona is the most liked host of entire NYC with a total reviews of ≈630**

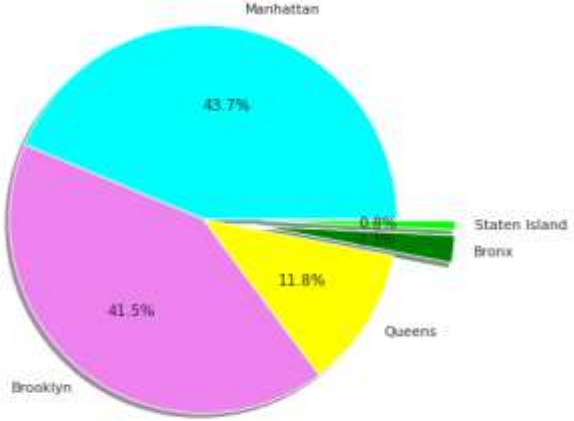
***Also*, we found out that a single host can have multiple listings over every other Neighbourhood Group with different host id's:**

```
# Using both the test conditions:
df2.loc[(df2['neighbourhood_group']=='Manhattan') & (df2['host_name']=='Michael')]
```

Exploratory Data Analysis: (Bi-Variate Analysis)



Total percentage of each room type in NYC



Total percentage of each neighbourhood group in NYC

Conclusions:

Total Percentage of **each room type** in entire NYC is:

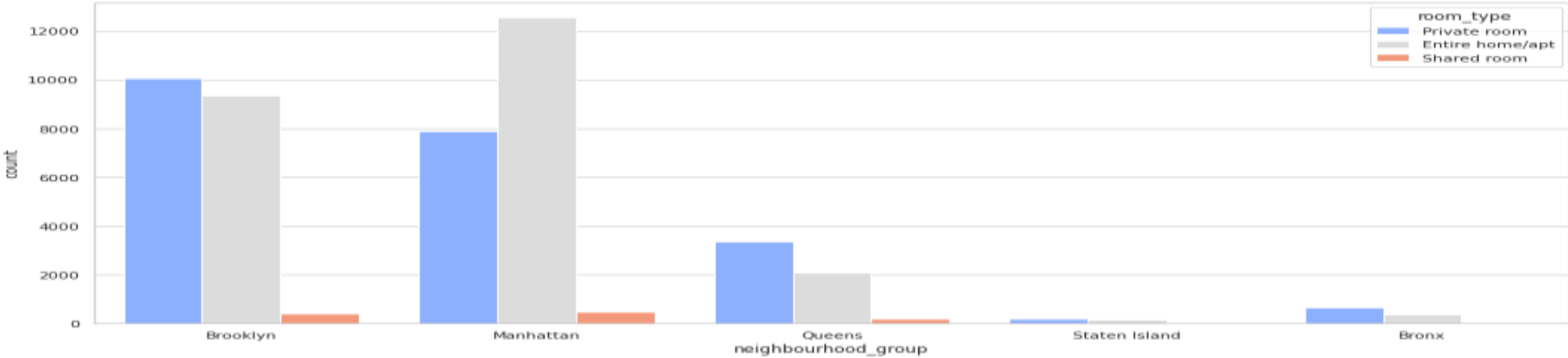
- Entire home/apt = 51.3%
- Private room = 46.3%
- Shared room = 2.4%

Conclusions:

Total Percentage of **area under each neighbourhood** group in NYC is:

- Manhattan = 43.7%
- Brooklyn = 41.5%
- Queens = 11.8%
- Bronx = 2.3%
- Staten island = 0.8%

Exploratory Data Analysis: (Multi-Variate Analysis)



Total listings of each room type in each neighbourhood group in NYC

Conclusions:

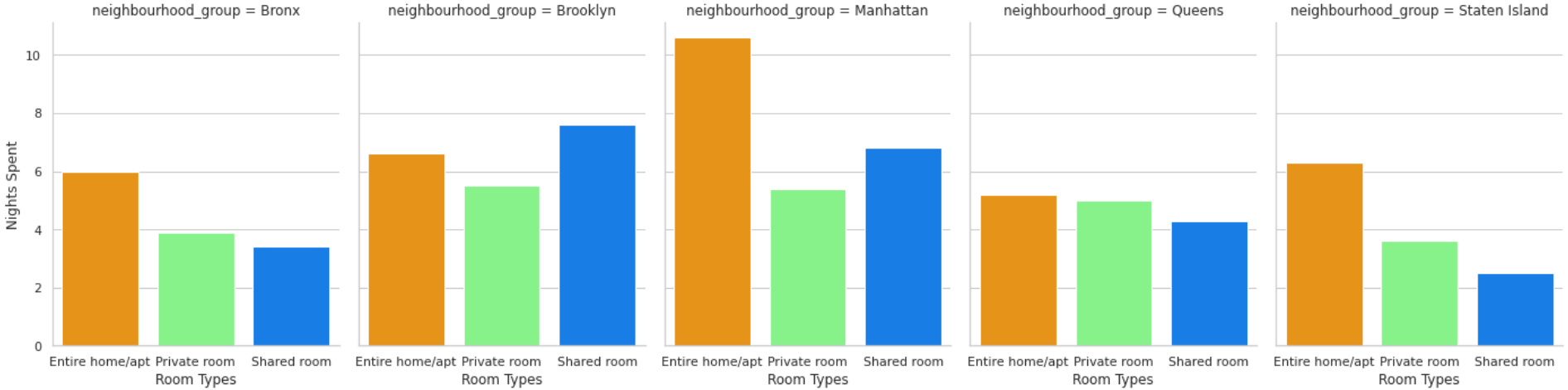
Total number of **room_type** listings in **each neighbourhood_group**:

Entire home/apt:
- ➤Manhattan ≈13000
- ➤Brooklyn ≈9500
- ➤Queens ≈2100
- ➤Bronx ≈400
- ➤Staten island ≈100

Private room:
- ➤Brooklyn ≈10000
- ➤Manhattan ≈8000
- ➤Queens ≈3500
- ➤Bronx ≈800
- ➤Staten island ≈200

Shared room:
- ➤Manhattan ≈500
- ➤Brooklyn ≈300
- ➤Queens ≈200
- ➤Bronx and Staten island with almost negligible shared room type listings

Exploratory Data Analysis: (Multi-Variate Analysis)



Average minimum nights spent in each room_type in each neighbourhood_group

Conclusions:

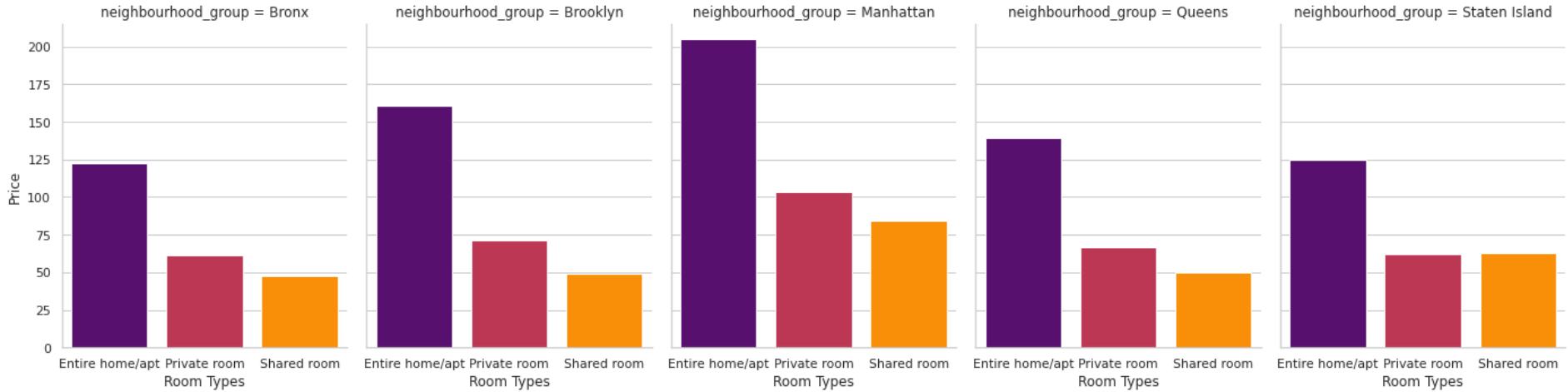
Average Minimum Nights spent in room_type in **each neighbourhood_group**:

Entire home/apt:
- Manhattan ≈11
- Brooklyn ≈7
- Staten island ≈7
- Bronx ≈7
- Queens≈6

Private room:
- Brooklyn ≈6
- Manhattan ≈6
- Queens ≈6
- Bronx ≈4
- Staten island ≈4

Shared room:
- Brooklyn ≈8
- Manhattan ≈7
- Queens ≈5
- Bronx ≈4
- Staten island ≈3

Exploratory Data Analysis: (Multi-Variate Analysis)



Average price of each room_type in each neighbourhood_group

Conclusions:

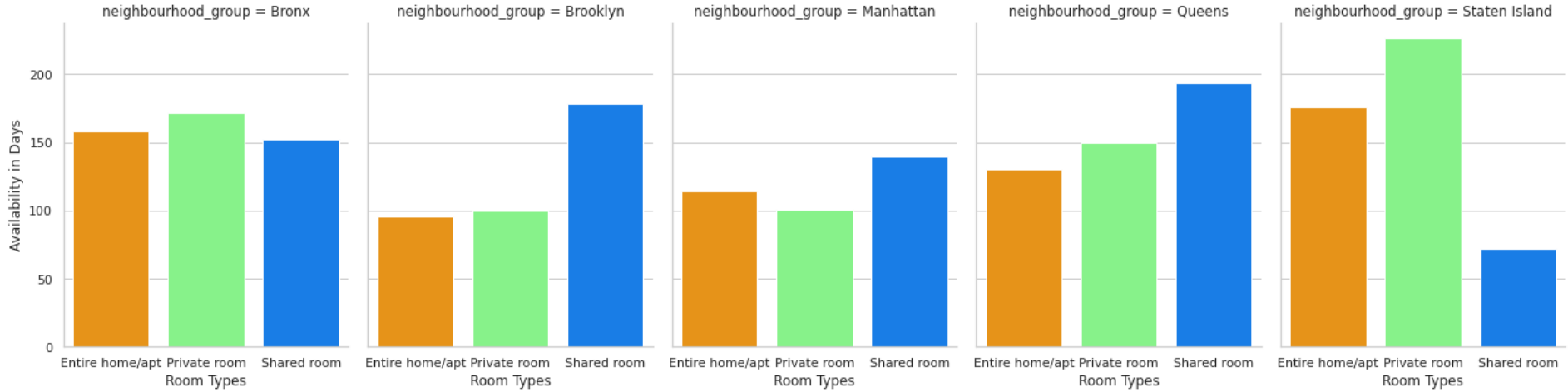
Average price of each room_type in **each neighbourhood_group**:

Entire home/apt:
- Manhattan ≈206
- Brooklyn ≈161
- Queens≈140
- Staten island ≈125
- Bronx≈123

Private room:
- Manhattan ≈104
- Brooklyn ≈72
- Queens ≈68
- Staten island ≈63
- Bronx ≈62

Shared room:
- Manhattan ≈85
- Staten island ≈64
- Queens ≈51
- Brooklyn ≈50
- Bronx ≈48

Exploratory Data Analysis: (Multi-Variate Analysis)



Average availability of each room_type in each neighbourhood_group

Conclusions:

Average availability of each room_type in **each neighbourhood_group**:

Entire home/apt:
- Staten island ≈176 days
- Bronx ≈159 days
- Queens≈131 days
- Manhattan≈114 days
- Brooklyn≈96 days

Private room:
- Staten island ≈227 days
- Bronx ≈172 days
- Queens≈150 days
- Manhattan≈101 days
- Brooklyn≈100 days

Shared room:
- Queens ≈194 days
- Brooklyn ≈179 days
- Bronx ≈153 days
- Manhattan ≈140 days
- Staten island ≈72 days

❑ Exploratory Data Analysis:

List of TOP 10 name of listings in entire NYC as per total reviews:

	name	number_of_reviews
32040	Private Bedroom in Manhattan	666
35742	Room near JFK Queen Bed	629
5513	Beautiful Bedroom in Manhattan	617
20433	Great Bedroom in Manhattan	607
35388	Room Near JFK Twin Beds	576
39694	Steps away from Laguardia airport	543
27090	Manhattan Lux Loft.Like.Love.Lots.Look !	540
14896	Cozy Room Family Home LGA Airport NO CLEANING FEE	510
33322	Private brownstone studio Brooklyn	488
25266	Loft Suite @ The Box House Hotel	481

Private Bedroom in Manhattan is the most famous listing property as per total reviews **≈667** followed by **Room near JFK Queen Bed** with **≈630** total reviews and others...

List of TOP 10 name of listings in entire NYC as per reviews per month:

	name	reviews_per_month
18151	Enjoy great views of the City in our Deluxe Room!	58.5
20560	Great Room in the heart of Times Square!	28.0
21943	Home away from home	21.1
25365	Lou's Palace-So much for so little	20.9
22975	JFK Comfort.5 Mins from JFK Private Bedroom & ...	19.8
22967	JFK 2 Comfort 5 Mins from JFK Private Bedroom	17.8
22969	JFK 3 Comfort 5 Mins from JFK Private Bedroom	16.8
32421	Private Room	16.8
14896	Cozy Room Family Home LGA Airport NO CLEANING FEE	16.2
21304	Harlem Gem	16.1

On the basis of monthly reviews **Enjoy great views of the City in our Deluxe Room!** is at the top with a total of **≈59** reviews per month.

❑ Exploratory Data Analysis:

AI

TOP 10 Neighbourhood on the basis of Total number of reviews

	neighbourhood	number_of_reviews
13	Bedford-Stuyvesant	109481
213	Williamsburg	85031
93	Harlem	75709
28	Bushwick	52329
94	Hell's Kitchen	49948
64	East Village	44067
61	East Harlem	36356
51	Crown Heights	36325
201	Upper West Side	35587
200	Upper East Side	31316

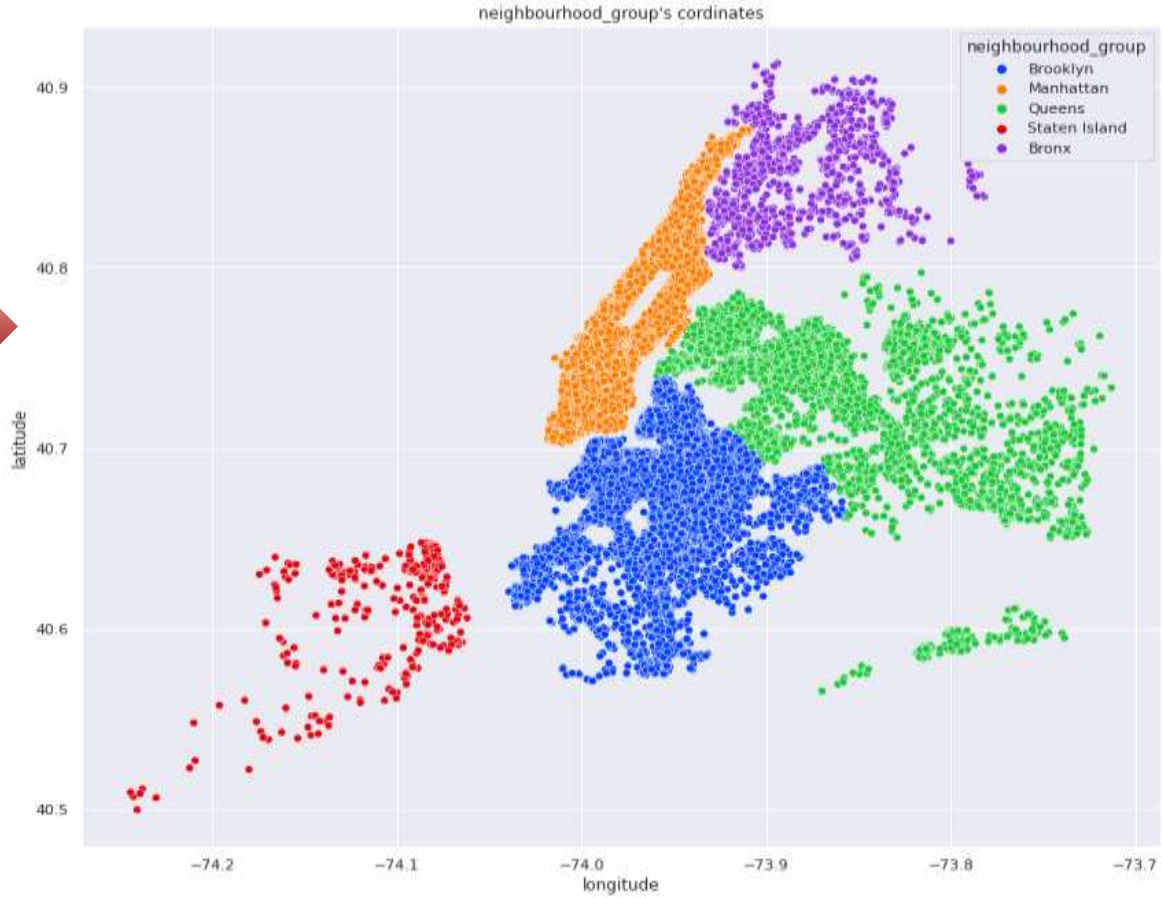
Top neighbourhood is **Bedford-Stuyvesant** with total reviews of ≈109482

Williamsburg is at 2nd with a total of ≈85032 reviews followed by **Harlem** and others

❑ Exploratory Data Analysis:

Latitude & Longitude values of each listing in each neighbourhood group in entire NYC

Latitude & Longitude values of each listing property of each neighbourhood group is shown here via scatterplot ....



❑ Overall Conclusion:

We were given a csv file containing a data set from Airbnb, we started our Analysis by:

- **First having a brief overlook at given Data Set ,then we fixed Null & NaN values.**

- **After that we tried dealing with the Outliers using Quantile approach.**
Which means to assign a minimum and a maximum threshold and then only catching the values in between them.

- **Created and overviewed Box Plot to get a basic idea of *mean, max, min and quartiles values*.**

- **Concluded the relationship between the Hosts and Neighbourhood Areas in NYC !**

- **Performed some Bi-Variate Analysis among Neighbourhood Groups , Prices and Reviews and plotted some Bar Graphs for the same.**
(Bi-Variate Analysis: comparing two different variables to study their relationship)

- **Figured out Top Airbnb Hosts on the basis of Listings and Reviews both and showed them using Bar Graphs.**

❑ Overall Conclusion: (cont)

- Saw the Total Percentage listings of each Room type and each Neighbourhood Groups in NYC using pie plot.

- We then went on performing some Multi-Variate Analysis and using CountPlot we figured out : (Multi-Variate Analysis: is quite similar to Bi Variate Analysis but here we compare more than 2 variables at once)

  - Average Nights spent in each Room Type for all Neighbourhood Cities from Longest to Shortest.
  - Average Price of each Room Type for all Neighbourhood Cities in NYC from Highest to Lowest.
  - Average Availability of each Room Type for all Neighbourhood Cities in NYC in a year.

- Sorted out TOP 10 listings Names according to the Reviews of people.

- Sorted out TOP 10 Neighbourhood Cities according to the Reviews of people.

- And finally played around with Latitude & Longitude a bit and represented them using Scatterplot.

❑ Challenges:

- The first challenge was to make the data set ready for EDA by fixing NULL & NaN values , which we did by replacing values in few columns and dropping/deleting the others.

- Next challange was to remove the outliers without doing so, the data could never be analyzed.
We removed the outliers by Quantile approach which makes the data set just ready for exploration...

- Using **Interquartile Range(IQR) Formula (***Interquartile range = Upper Quartile – Lower Quartile***)** for removing outliers was a better approach but Quantile method was just fine for our basic operations.

❑ THANKS !