NEW YORK AIRBNB VISUAL AND ANALYTICAL PRESENTATION

Submitted by-
Abhishek De
Rahul Roy

Agenda

- Overview
- Objective
- Importing necessary libraries and the dataset
- Descriptive Statistics for numeric columns
- Finding Null Values
- Missing Value Treatment
- Host Analysis
- Analysis of the Neighborhood_group column
- Price analysis of property rents

Objective

- Improve our shared understanding about the market conditions.
- Improve shared understanding about our customers.
- Provide recommendations to various departments to be prepared for business expansion.

Overview

- Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.
- The analytics and visualizations are performed in Python programming language using packages like NumPy, Pandas, Matplotlib and Seaborn.
- The different leaders at Airbnb want to understand some important insights based on various attributes in the dataset so as to increase the revenue.

Importing necessary libraries and the dataset

In [2]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [3]:
```
df=pd.read_csv('AB_NYC_2019.csv')
df.head()
```

Out[3]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_revie
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	2
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

In [4]:
```
# Checking the amount of rows and columns

df.shape
```

Out[4]: (48895, 16)

Looking at the dataset we find that the data is contained in 48,895 rows and 16 columns. df.head() function is used to visualise any pandas dataframe's first 5 rows.

Descriptive Statistics for numeric columns

```
# Descriptive Statistics

df.describe()
```

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.

From the above we can observe the Mean, Median and Maximum values of numeric columns.

Finding Null Values

```
# Checking for the null values

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
```

```
# Null Value Percentage

round(100*(df.isnull().sum()/len(df)),2)
```

```
id                                  0.00
name                                0.03
host_id                             0.00
host_name                           0.04
neighbourhood_group                 0.00
neighbourhood                       0.00
latitude                            0.00
longitude                           0.00
room_type                           0.00
price                               0.00
minimum_nights                      0.00
number_of_reviews                   0.00
last_review                        20.56
reviews_per_month                  20.56
calculated_host_listings_count      0.00
availability_365                    0.00
dtype: float64
```

df.isnull().sum() is the function which displays the sum of total null or NaN entries in the dataframe. We also depicted percentages of null values.

Missing Value Treatment

```
# Dropping the columns that could be insignificant or unethical for the future use

df.drop(['id','host_name','last_review'], axis=1, inplace=True)
```

```
df.shape
```

```
(48895, 13)
```

We dropped the columns which are having missing values and are considered insignificant for your analysis

```
# Repalcing the missing values of Reviews_per_month with 0

df.fillna({'review_per_month':0}, inplace=True)
```

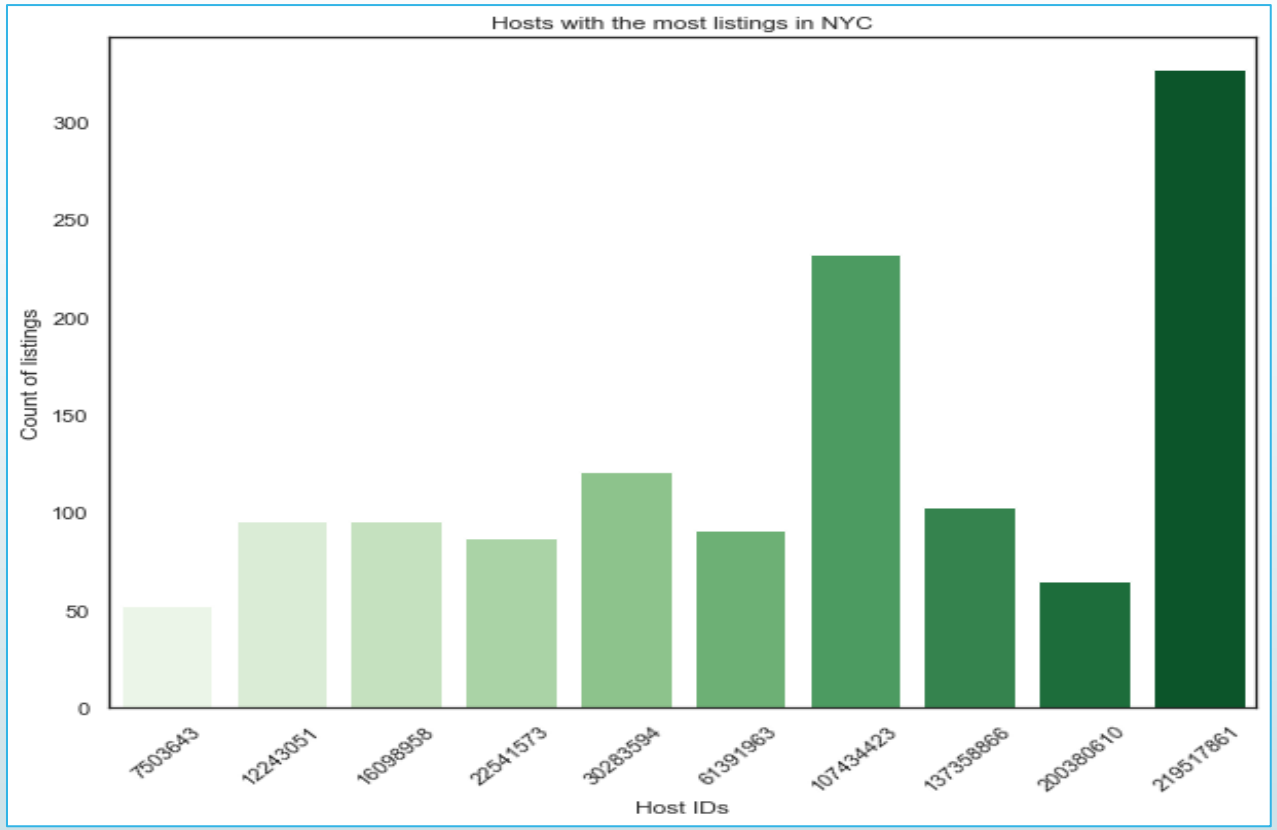
As the reviews are customer-dependent, it's insignificant to replace the missing values with the mean or median values that's why replacing them with 0

Host Analysis

```
sns.set(rc={'figure.figsize':(10,8)})
sns.set_style('white')

viz_1=sns.barplot(x="Host_ID", y="P_Count", data=top_host_df,
                  palette='Greens')
viz_1.set_title('Hosts with the most listings in NYC')
viz_1.set_ylabel('Count of listings')
viz_1.set_xlabel('Host IDs')
viz_1.set_xticklabels(viz_1.get_xticklabels(), rotation=45)

plt.show()
```



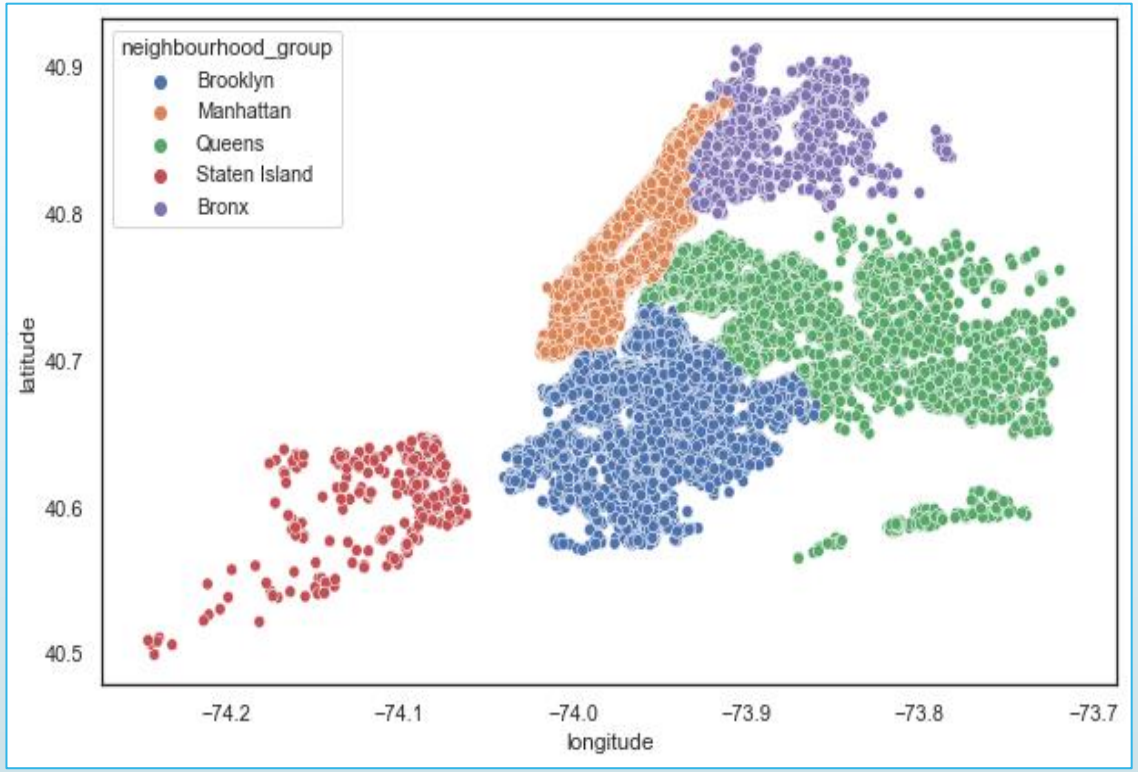
With the help of Seaborn Barplot, we can visualize the Hosts with the most number of Airbnb property listings in New York City.

Analysis of the Neighborhood_group column

```
# Analysis of the Neighbourhood_group column

df.neighbourhood_group.value_counts()
```

```
Manhattan        21661
Brooklyn         20104
Queens            5666
Bronx             1091
Staten Island      373
Name: neighbourhood_group, dtype: int64
```

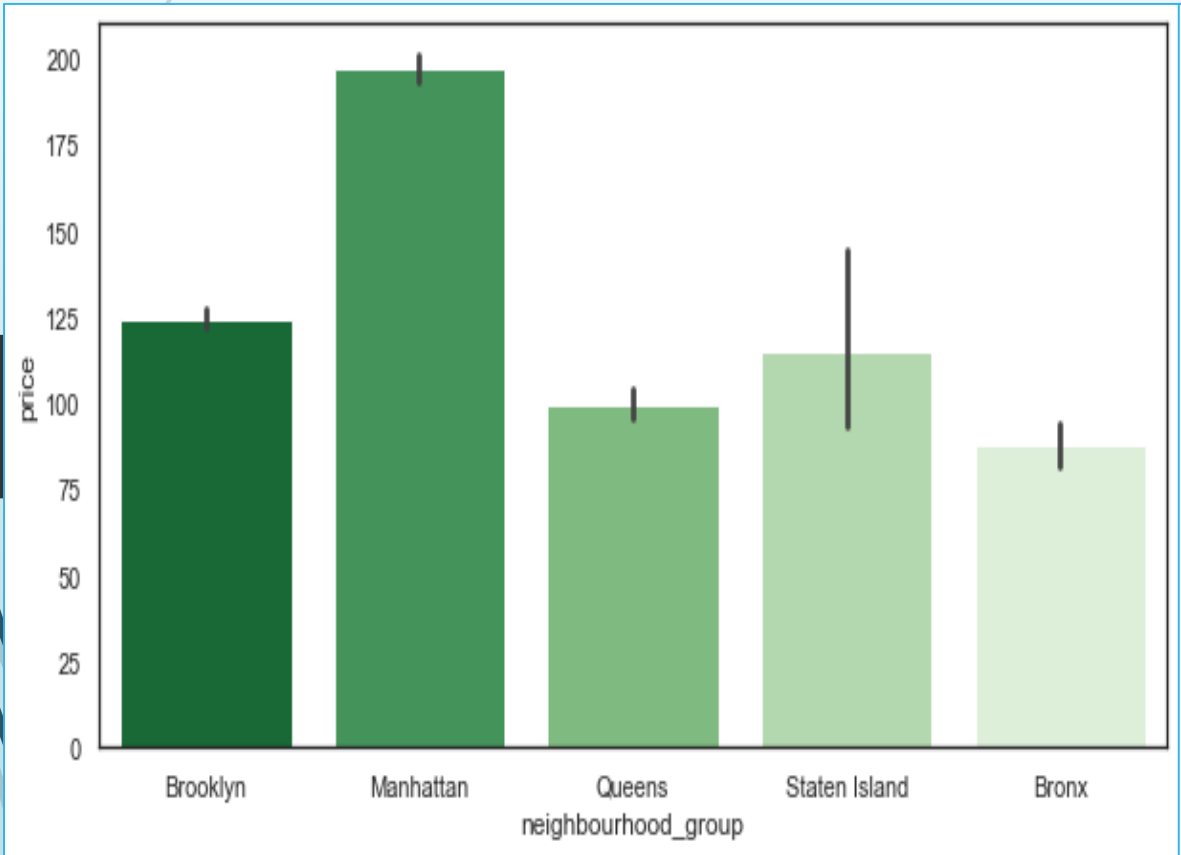


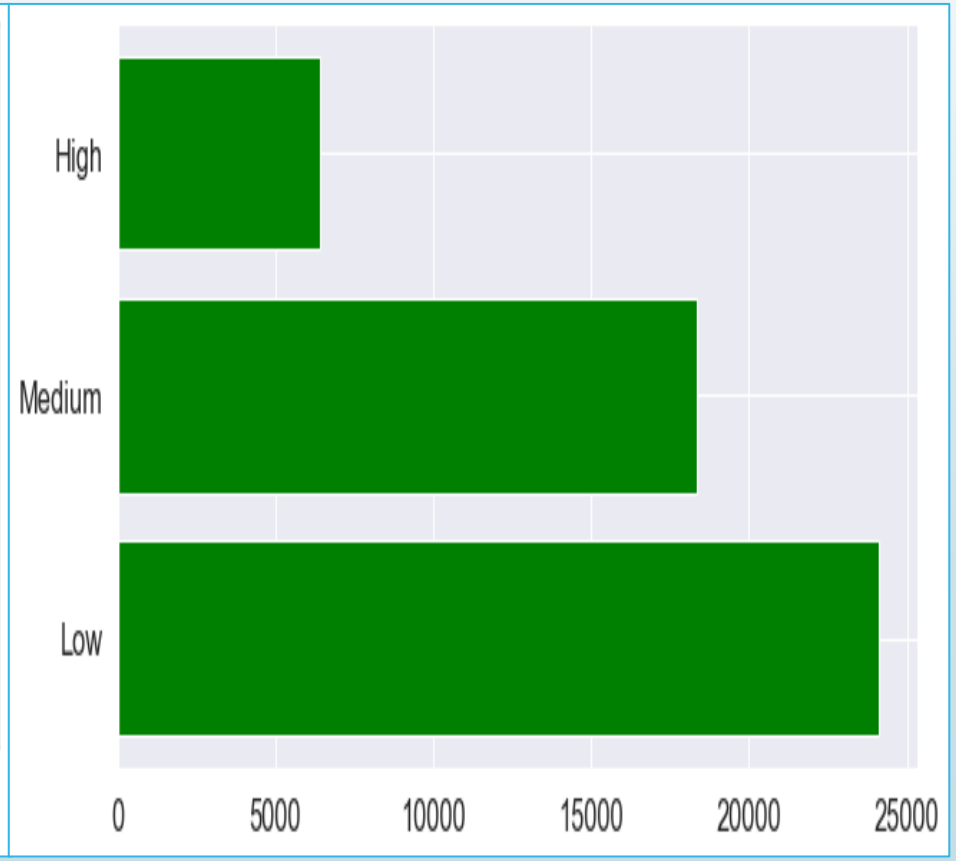
As evident from the count plot, Manhattan has the highest number of Airbnb holdings in New York (21,661) followed by Brooklyn (20,104), Queens (5,666), Bronx (1,091) and Staten Island (373). Visualising the neighbourhood groups using libraries like seaborn and matplotlib are very handy and can help in building great dashboards and report

Price analysis of property rents

```
# Does price affects the neighbourhood

plt.figure(figsize=(10,5))
sns.barplot(x='neighbourhood_group', y="price", data=df, palette='Greens_r')
plt.show()
```





THANK YOU