Methodology Document – Steps performed for EDA and Viz

1. Import required libraries and dataset –

```
In [1]:
1 import numpy as np
2 import pandas as pd

In [2]:
1 df=pd.read_csv("AB_NYC_2019.csv")

In [3]:
1 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	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

2. Checking datatypes of fields in the dataset –

```
In [6]:
1 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
```

3. Checking for nulls in the dataset –

```
In [7]:  1 df.isnull().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
```

4. Cleaning all the null by imputing values. I've decided to impute last_review and reviews_per_month as well despite they have huge nulls. As per dataset it may cause issues if we remove records with Nulls in reviews as customer may choose not to review a listing.

```
In [8]:  1 df['host_name'] = df['host_name'].fillna(df['host_name'].mode()[0])
         2 df['name']=df['name'].fillna(df['name'].mode()[0])
         3 df['last_review']=df['last_review'].fillna(df['last_review'].mode()[0])
```

```
In [10]:  1 df['reviews_per_month']=df['reviews_per_month'].fillna(df['reviews_per_month'].mode()[0])
```

```
In [11]:  1 df.isnull().sum()
```

```
Out[11]: 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
         last_review                         0
         reviews_per_month                   0
         calculated_host_listings_count      0
         availability_365                    0
         dtype: int64
```

5. Checking top 5 Neighborhood and Neighborhood groups including graphs

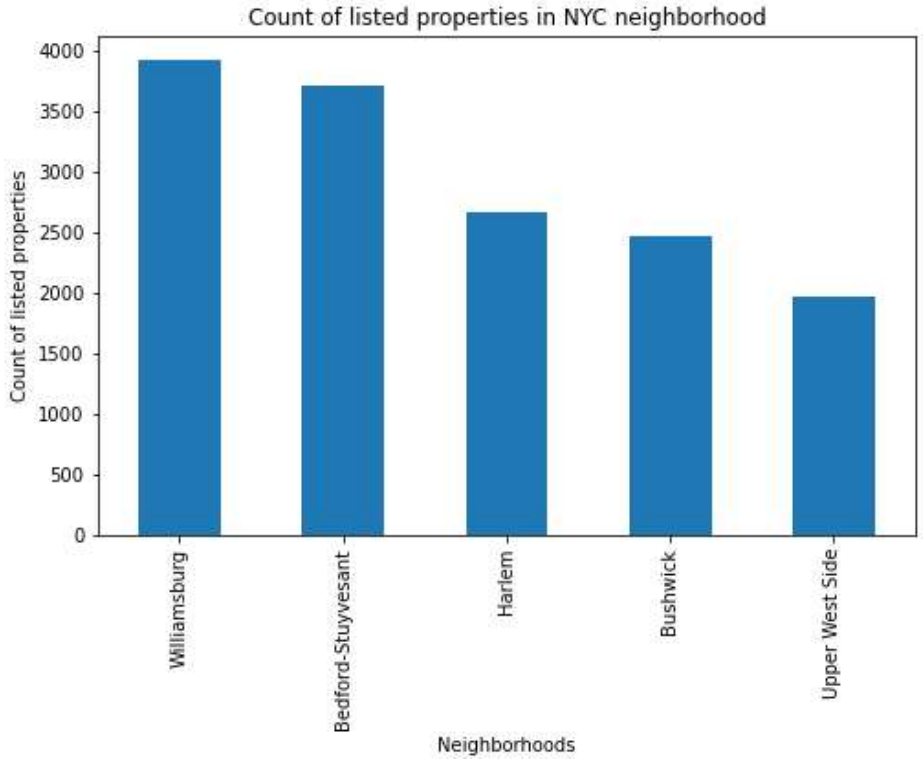
In [16]:
```
#checking the to 5 neighborhood where the properties are listed most.
top_5_neighbors = df.neighbourhood.value_counts().head(5)
print(top_5_neighbors)

```

```
Williamsburg          3920
Bedford-Stuyvesant    3714
Harlem                2658
Bushwick              2465
Upper West Side       1971
Name: neighbourhood, dtype: int64
```

In [14]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

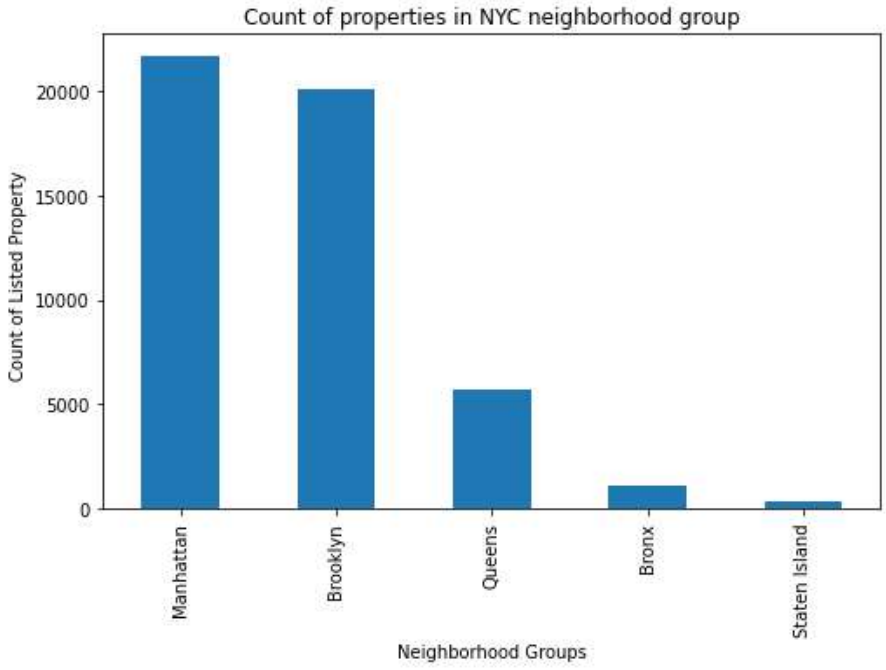
In [17]:
```
#plotting
plt.figure(figsize=(8,5))
top_5_neighbors.plot.bar()
plt.xlabel('Neighborhoods')
plt.ylabel('Count of listed properties')
plt.title('Count of listed properties in NYC neighborhood')
plt.show()
```



In [19]:
```
#checking the to 5 neighborhood groups where the properties are listed most.
top5_neighborhood_group = df.neighbourhood_group.value_counts()
print(top5_neighborhood_group)
```

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

In [20]:
```
plt.figure(figsize=(8,5))
top5_neighborhood_group.plot.bar()
plt.xlabel('Neighborhood Groups')
plt.ylabel('Count of Listed Property ')
plt.title('Count of properties in NYC neighborhood group')
plt.show()
```



please note the above two charts have not been shown in any of the presentation pdf's.

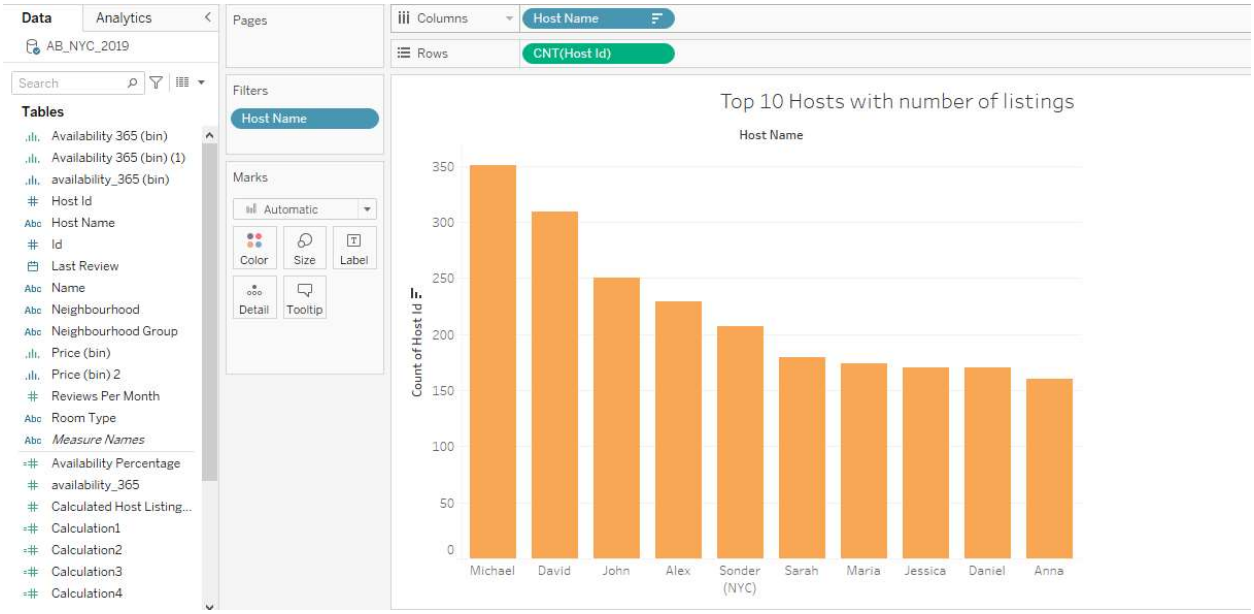
6. Exported CSV file attached used for tableau visualization –

AB_NYC_2019.csv

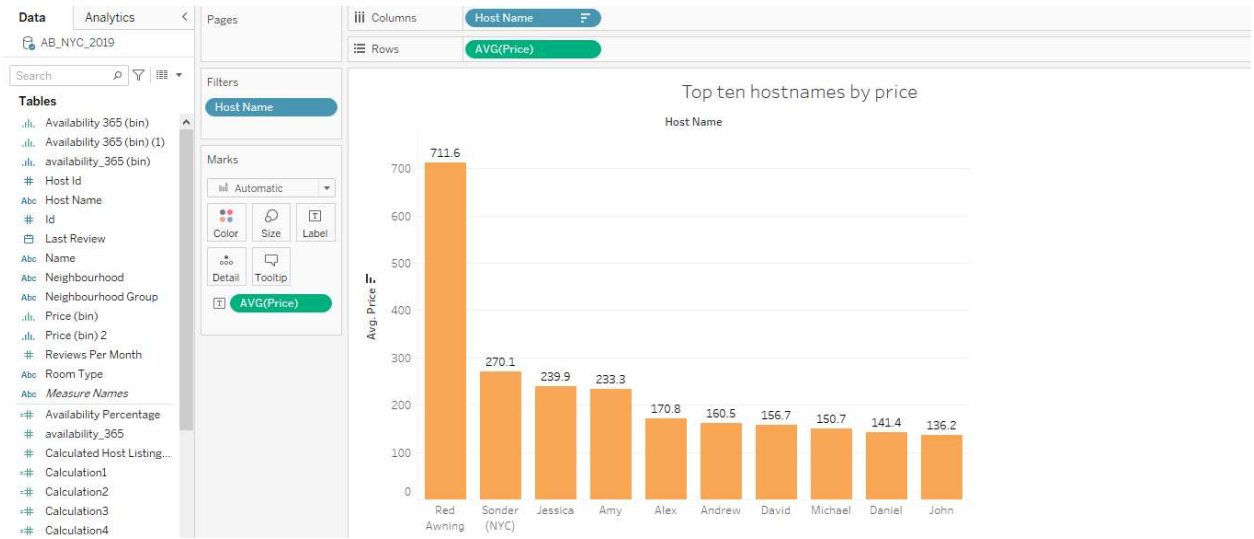
7. For presentation 1, which is to be shown to the Data Analysis Managers and Lead Data Analyst we Visualized all aspects with respect to price, minimum nights spent, hosts w.r.t multiple parameters, location.

For presentation 2, which is to be shown to the Head of Acquisitions and Operations, NYC and Head of User Experience, NYC we visualized core areas like location, pricing, customer reviews, room types and analysis for preference.

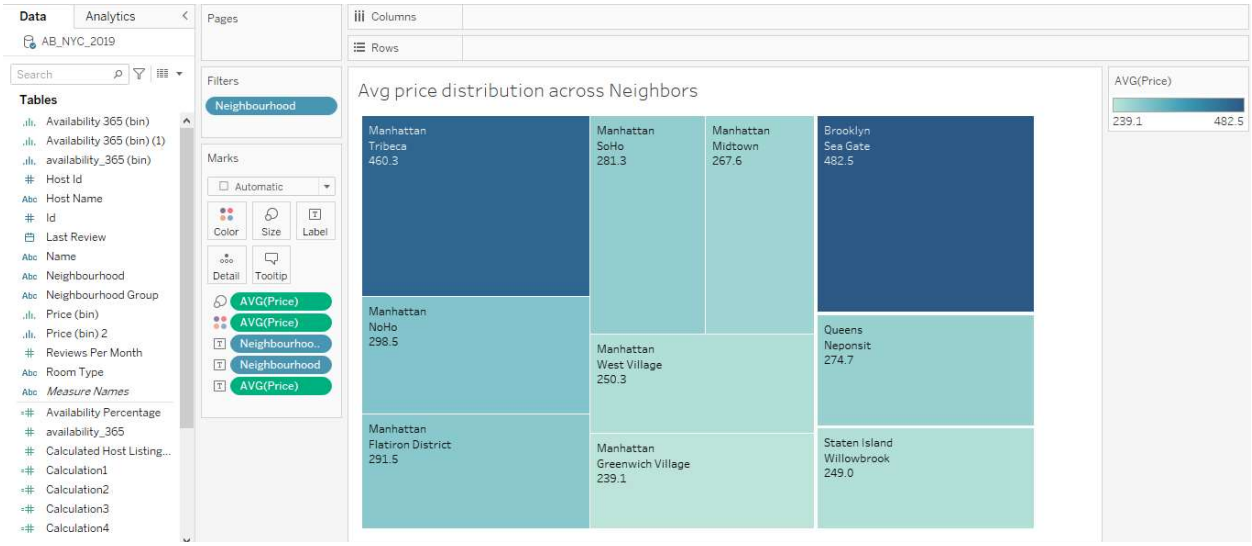
8. From Presentation 1 the first chart shows the top 10 hosts with the number of listings they have in NYC and across its neighborhood. Below is the tableau snip,



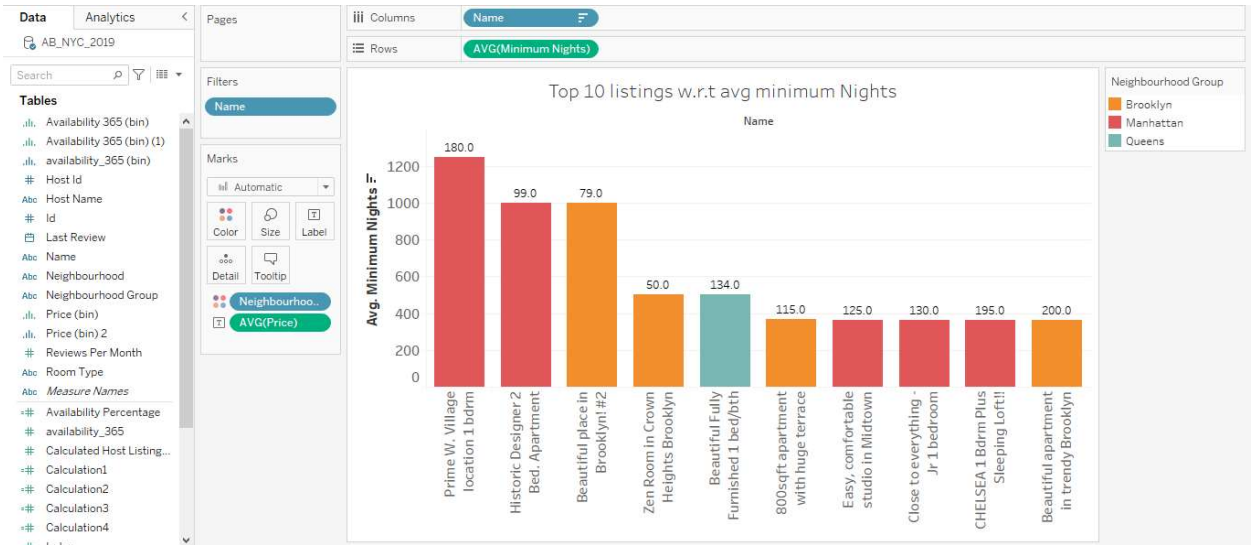
9. Second chart shows the top 10 hosts with highest average price for their listings in NYC and across its neighborhood. Below is the tableau snip,



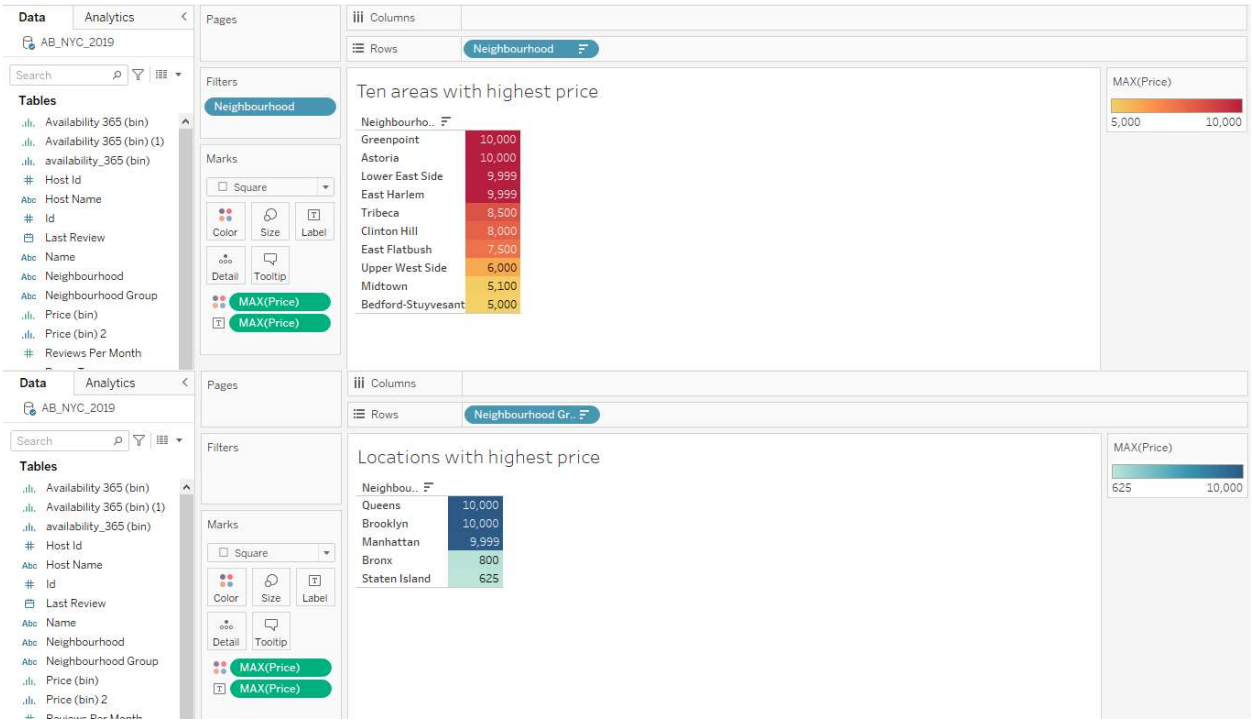
10. Below chart is showing a heatmap based on average price of listings across NYC neighborhood. This gives a glimpse on the areas where the listings are economic and places where the density of listings are huge. Below is the snip,



11. Below chart shows top 10 listings based on average nights spend over all the listings within all locations. This shows how much customers prefer this property and also the facilities given in these properties are much preferred. This also shows the average price at the tooltip to give an idea for the money spent.

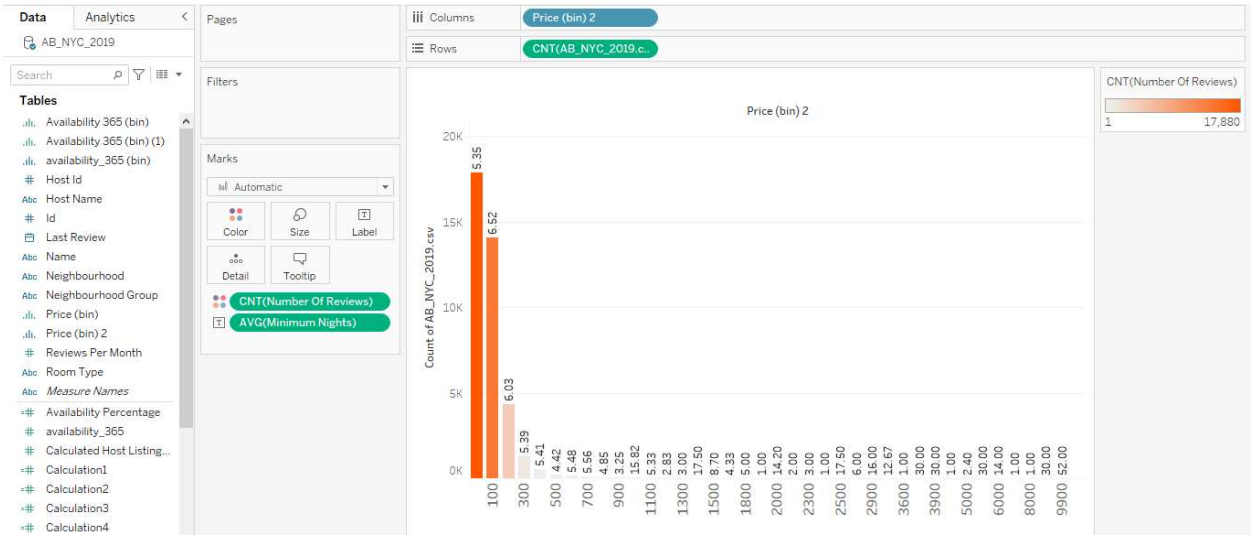


12. Below two charts shows the pricing of listings in a sorted fashion for both neighborhood and neighborhood groups within the dataset. Bronx and Staten are the cheapest neighborhood groups among all. Below is the snip,

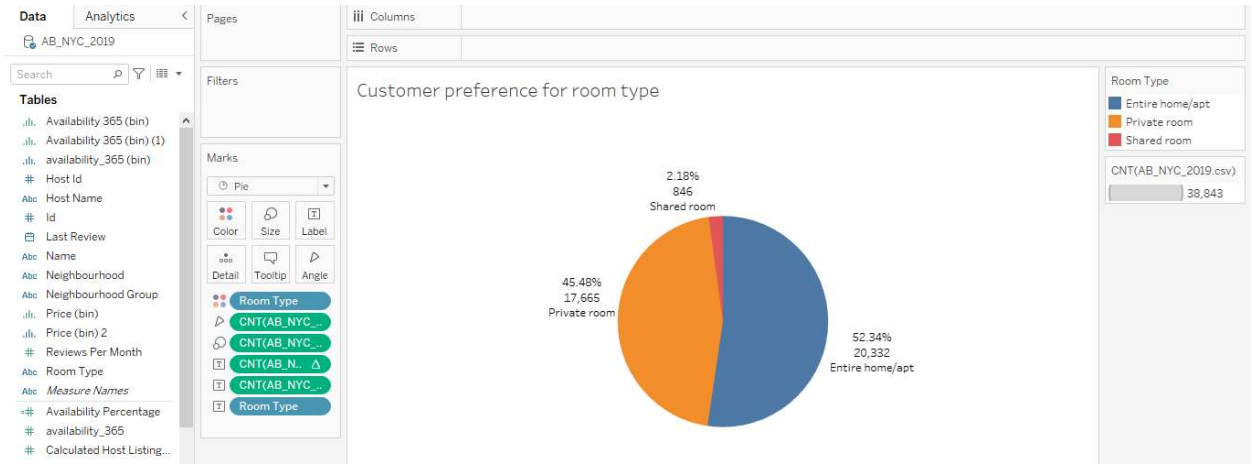


13. We get some conclusive points from the top charts shown above,
    a. Premium properties and rooms should be acquired at Staten Island and Bronx as rates are reasonable and tourism should be promoted in this two locations.
    b. Additional incentives to hosts who have properties at prime locations and are offering a mediocre cost per night for a room
    c. Manhattan and Brooklyn has the highest number of minimum nights available which should be capitalized by reducing the prices of the Private Room and Entire Home/Apt. as it will bring in more customers which ultimately will lead to more revenue generation

14. Below chart shows the reviews distribution based on price. Price in this chart has been divided in bins of 100. This chart shows the reviews has been given to the listings whose per night charge is within 100$ are maximum.



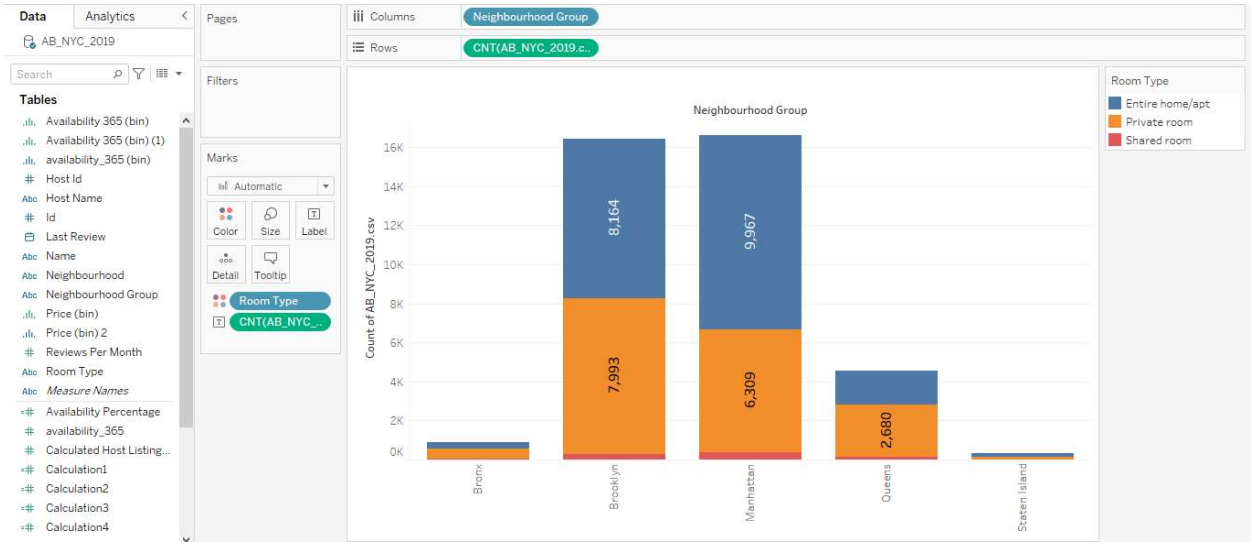
15. Below chart shows a pie chart distribution for the room type preference by the customers. Entire home/apt and Private rooms have more preference than any other type of rooms if available. Below is the snip,



16. Next we deep dive to reviews analysis as it helps the operations to improve if they were lacking somewhere. Also it helps to give recommendations to other new/existing customers. Below chart shows reviews distribution based on area/neighborhood. Below is the snip,

Although this turns out to be challenging if we could've modified it based on positive and negative kind of reviews else that would've given better perspective on this chart.

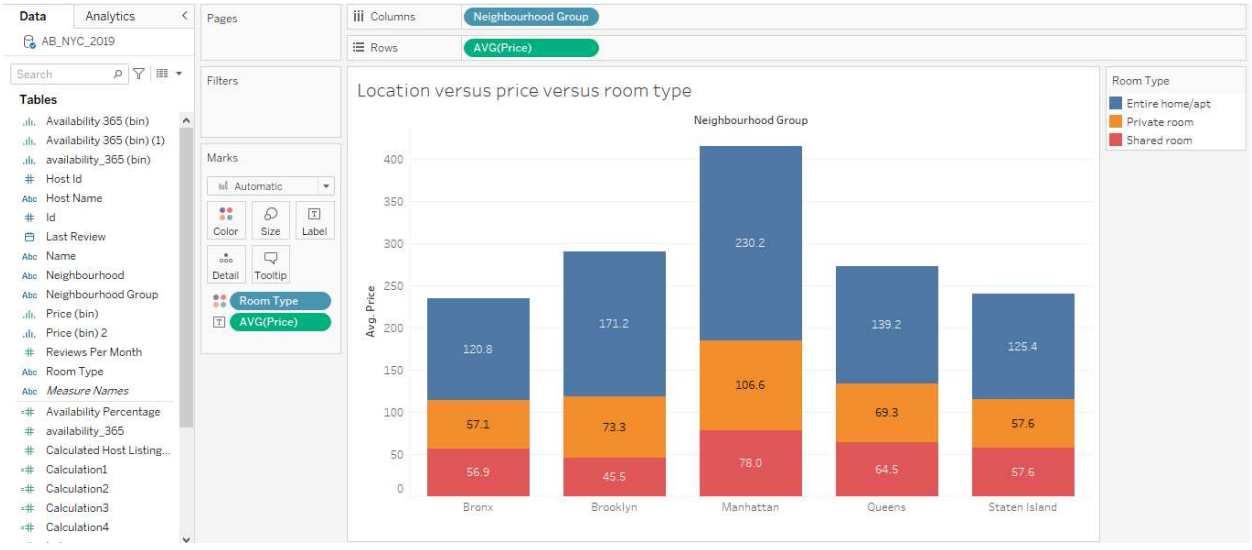
17. Below we can see which room type has been booked by the customers in the neighborhood groups. Since we already knew that Bronx and Staten Island has smallest booking and the cheapest we can focus more on other areas. The below graph shows that for Manhattan and Brooklyn, affluent people travel to this place, and they prefer Entire Home/Apt. Below is the snip,



missed to add percentage calculations.

18. Here, we have plotted average price with Sum (Number of Reviews) along with the Host Name.
    The below graph shows that the host with highest review is not the one whose property also has the highest price.
    In fact, hosts with highest Number of Reviews are more affordable in terms of price, and customers have given them a good review due to the value for money the property is offering in terms of facilities and room service. Below the snip,

19. In this chart we have shown a distribution among Location, Price and room type to give an understanding on the requirement for price adjustments if needed post Covid to attract more customers where average prices are more preferable to which room types and in which area. Below the snip,



20. Lastly we have given recommendations based out of drawn questions hypothetically from the given dataset.

- Efforts needed on advertising and expanding tourism in locations like Staten Island and Bronx.
- Cost cutting of the Private Rooms in Manhattan and Brooklyn so more people would visit and this in turn will normalize the cost reduction for the rooms.
- Encourage the top 10 hosts to open the listings at various other locations instead of focusing only on Manhattan and Brooklyn.
- Properties should be acquired at Staten Island and Bronx as rates are reasonable and tourism should be promoted in these two locations. But the land and properties availability can't be deduced from this dataset.
- Manhattan and Brooklyn tops the list of minimum nights available which should be capitalized by reducing the prices of the Private Room and Entire Home/Apt. as it will bring in more customers which will lead to more revenue generation.

----------*THANK YOU*----------