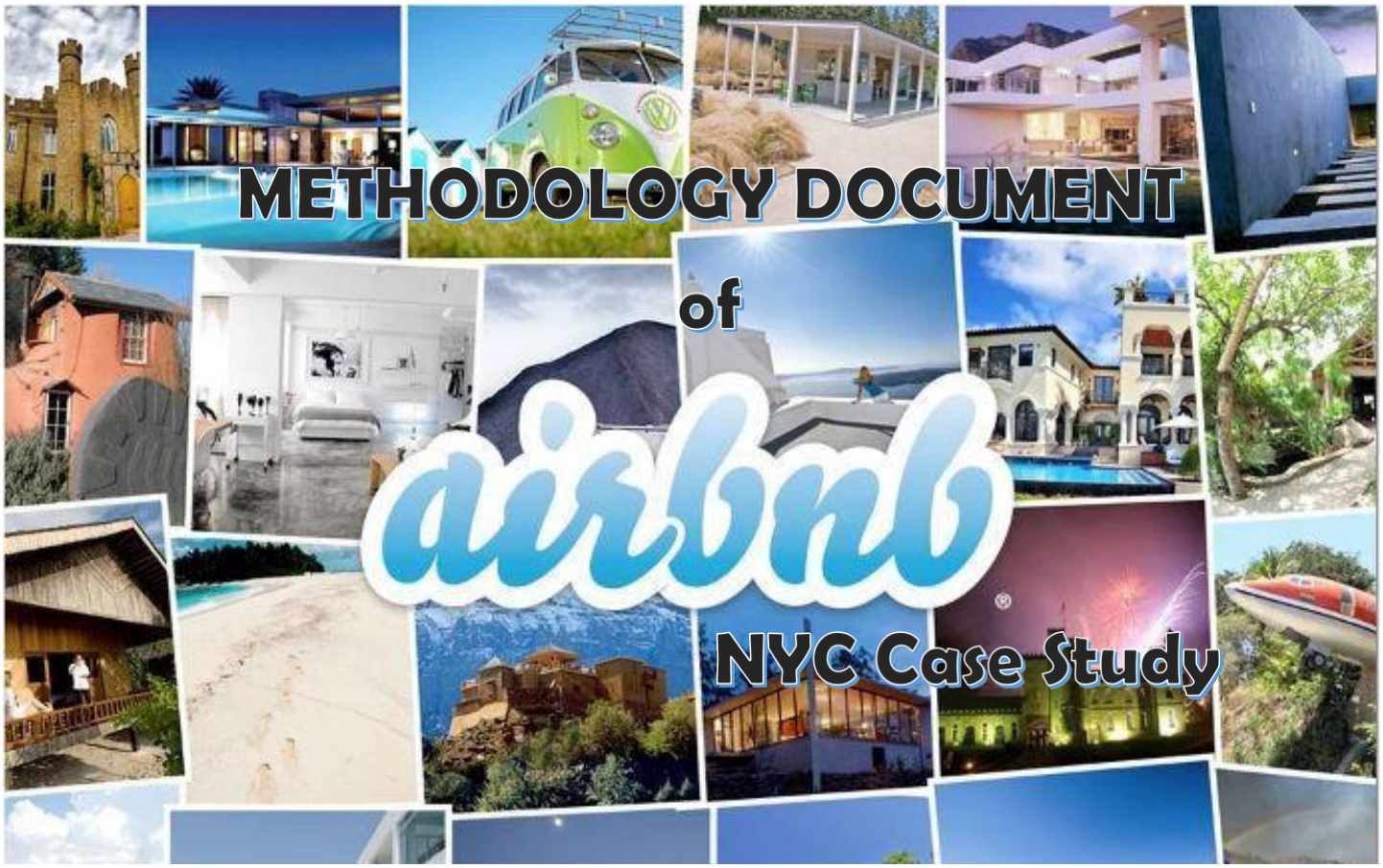

BY: -

1) SAMIKSHA BANSOD
2) DIPESH PIMPALE
3) NADHIYA C

Steps for Analyzing AIRBNB NYC Dataset:

Step1: Problem Understanding

Airbnb, like many other businesses, has been negatively impacted by the pandemic. As governments worldwide imposed lockdown measures and travel restrictions, the demand for short-term rentals and home-sharing services plummeted. Consequently, Airbnb saw a major decline in revenue as the number of bookings and travel activity dropped drastically.

However, as vaccination campaigns roll out and countries ease restrictions, the demand for travel is beginning to pick up. People are eager to make up for a lost time, and many are planning to travel as soon as they are able to do so.

To achieve this goal, Airbnb may be taking several measures to strengthen its platform, improve its services, and attract more customers.

Airbnb wants to make sure that it is fully prepared for this change.

Step 2: End Objective

To prepare for the next best steps that Airbnb needs to take as a business, we've been asked to analyze a dataset consisting of various Airbnb listings in New York. Based on this analysis, we need to give two presentations to the following groups.

Presentation - I

Data Analysis Managers: These people manage the data analysts directly for processes and their technical expertise is basic.

Lead Data Analyst: The lead data analyst looks after the entire team of data and business analysts and is technically sound.

Presentation - II

Head of Acquisitions and Operations, NYC: This head looks after all the property and host acquisitions and operations. Acquisition of the best properties, price negotiation, and negotiating the services the properties offer falls under the purview of this role.

Head of User Experience, NYC: The head of user experience looks after the customer preferences and also handles the properties listed on the website and the Airbnb app. Basically, the head of user experience tries to optimize the order of property listing in certain neighborhoods and citiesin order to get every property the optimal amount of traction.

Step 2: Business Understanding

The first step in our visualization project was to gain a clear understanding of the business problem and the objectives of the project.

Airbnb is an online marketplace that connects homeowners with people looking for accommodations in specific locales. Airbnb offers a relatively stress-free way for homeowners to earn some income from their property, while guests often find Airbnb accommodations to be cheaper, more characterful, and homier than hotels. Finally, Airbnb makes the bulk of its revenue by charging a service fee for each booking.

Step 3: Data Understanding in Python

After understanding the business problem and problem statement, we moved on to data understanding. This involved collecting and exploring the dataset to gain a better understanding of the variables, data types, and distributions. We imported the data in Python for more clarity of the given data.

```
# importing all the required libraries
import numpy as np
import pandas as pd
```

```
# importing csv files in jupyter using pandas
#AB NYC Data
df=pd.read_csv("AB_NYC_2019.csv")
df.head()
```

	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	

Figure 1. Importing libraries and data in python

```
# total number of rows and columns
df.shape
```

```
(48895, 16)
```

```
df.info(all)
```

```
<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
```

Column	Description
id	listing ID
name	name of the listing
host_id	host ID
host_name	name of the host
neighbourhood_group	location
neighbourhood	area
latitude	latitude coordinates
longitude	longitude coordinates
room_type	listing space type
price	
minimum_nights	amount of nights minimum
number_of_reviews	number of reviews
last_review	latest review
reviews_per_month	number of reviews per month
calculated_host_listings_count	amount of listing per host
availability_365	number of days when listing is available for booking

Figure 2. Description of dataset

As per the given dataset, there are 48895 rows and 16 columns while their names and description were given.

Step 4: Data wrangling in Python

During this phase, we first analyzed the values missing in the given dataset. It was found that more than 20% of missing data in the “last_review” and “reviews_per_month” columns. Then we identified that we have 16 places and 21 hostnames that are missing but they are having their IDs given in the dataset. So, maybe by mistake, there are null values in the hostname column.

But “last_review” and “reviews_per_month” carry “NaN” values on purpose, meaning they were not missing at random, but these hosted sites/places have not received any reviews from the customers. Hence, these places would be least preferred by future customers and would also be facing bad business.

Finally, we just imputed the missing values of “reviews_per_month” with a 0.

We noted that the last_review column had dates missing in it, so we decided to drop the 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
```

```
df.isnull().mean()*100
```

```
id                                0.000000
name                              0.032723
host_id                           0.000000
host_name                         0.042949
neighbourhood_group               0.000000
neighbourhood                     0.000000
latitude                          0.000000
longitude                         0.000000
room_type                         0.000000
price                             0.000000
minimum_nights                    0.000000
number_of_reviews                 0.000000
last_review                      20.558339
reviews_per_month                20.558339
calculated_host_listings_count    0.000000
availability_365                  0.000000
dtype: float64
```

Figure 3. Checking nulls in given data

```
# imputing the null values
df["reviews_per_month"].fillna(0,inplace=True)

# dropping "last_review" column
df=df.drop("last_review",axis=1)
```

```
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
reviews_per_month                 0
calculated_host_listings_count    0
availability_365                  0
dtype: int64
```

Figure 4. Imputing null values and dropping column

Then we categorized the columns into categorical, continuous, location, and time variables.

```
cat_vals=["neighbourhood_group","neighbourhood","room_type"]
cont_vals=["price","minimum_nights","number_of_reviews","reviews_per_month","calculated_host_listings_count","availability_365"]
loc_vals=["latitude","longitude"]
time_val=["last_review"]
```

Figure 5. Categorization of Columns

After this categorization, we identified outliers in the continuous variable columns i.e., “price”, “minimum_nights”, “no_of_reviews”, “reviews_per_month”, and "calculated_host_listings_count" columns.

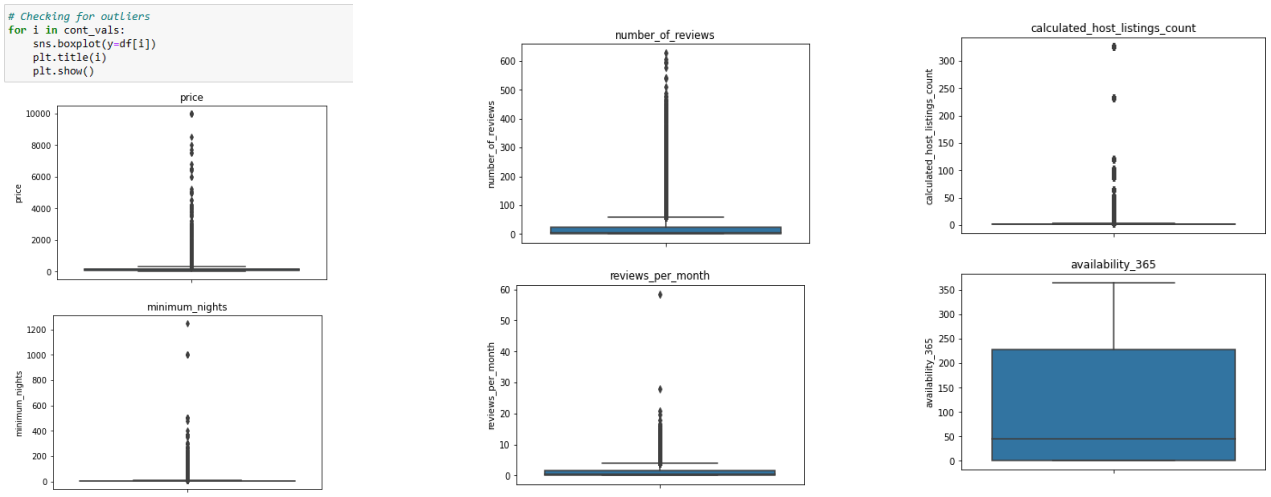

Figure 6. Checking for outliers

We treated these outliers with the “flooring” and “capping” methods.

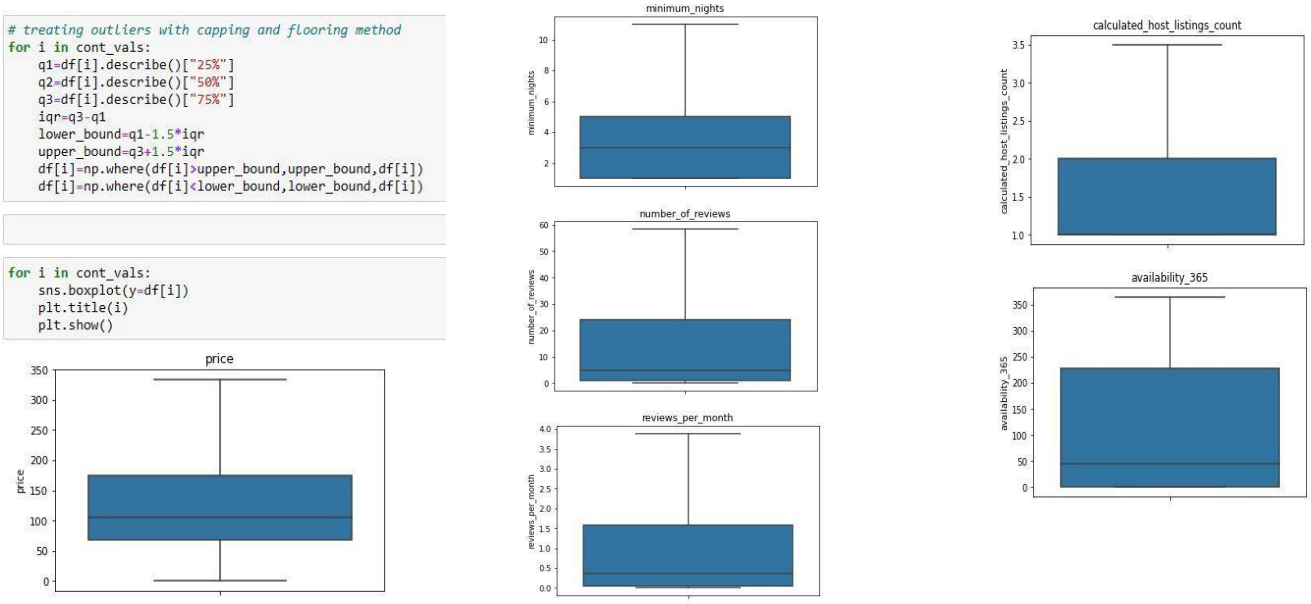

Figure 7. Treating outliers and columns after treating outliers in python

Step 5: Visualization in Tableau

Once we had a good understanding of the data, we moved on to the visualization phase. During this phase, we used Tableau to create 10+ visualizations that helped us gain insights into pricing trends and factors affecting pricing in Airbnb. To create these visualizations, we used advanced Tableau functions like calculated field, bins, and dual-axis charts. We also carefully selected appropriate chart types, color palettes, and labelling to make the visualizations intuitive, informative, and visually appealing. Throughout this phase, we paid close attention to the needs of our target audience and designed the visualizations to meet their needs.

1. **<u>Relationship between Neighborhood Groups, Listings and Prices –</u>**

   We explored the neighborhood groups relationship w.r.t. their average pricing and listing counts. Price and the customer visits of that neighborhood group are the important factors to know about which neighborhood groups we company needs to focus on in order to generate revenue.

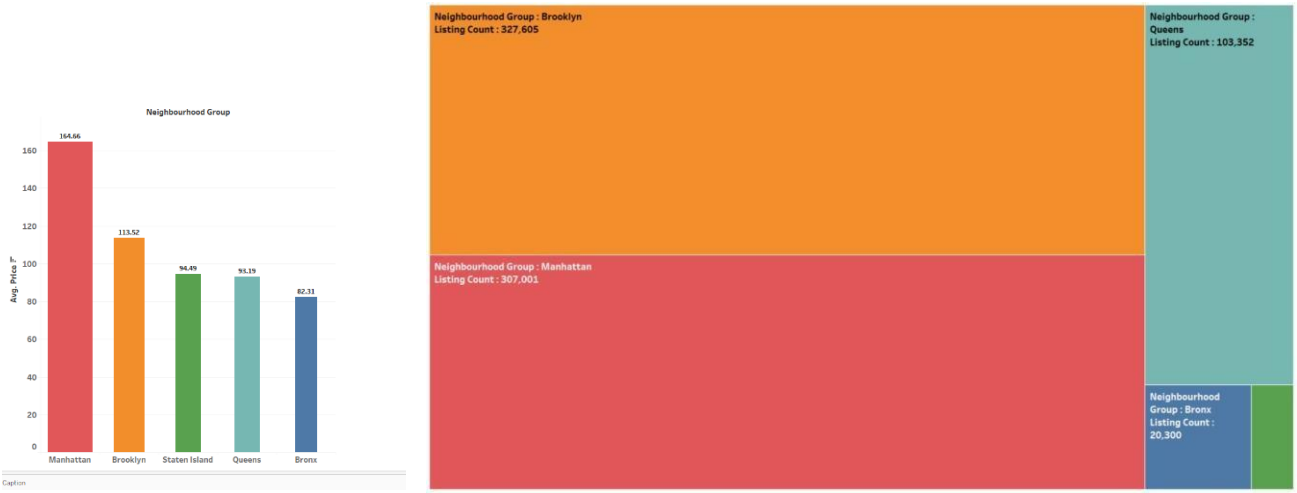

Figure 8. Neighborhood Groups w.r.t. pricing and listing counts

Inference: -

- We got to know, Manhattan and Brooklyn are expensive neighborhoods and so as their listing counts (customer visits).
- But, Staten Island (green colored) have high pricing than Queens and Bronx and the lowest listing of travelers.
- It states that may be due to high pricing not many visitors are eager to go there. Maybe pricing changes or discounts will attract customers to Staten Island.

2. Impact of Room Type and Neighborhood Group on Airbnb Prices and Reviews –

Most preferred Neighborhood Group and Room Type based on the Number of reviews. The customers reviews are the most important point to take into consideration at the time of analysis as it can be very helpful to know more about particular type. two different parameters that were taken for comparison: Number of reviews & pricing. The number of reviews a customer
gives for a particular listing directly implies the likability of the listing.

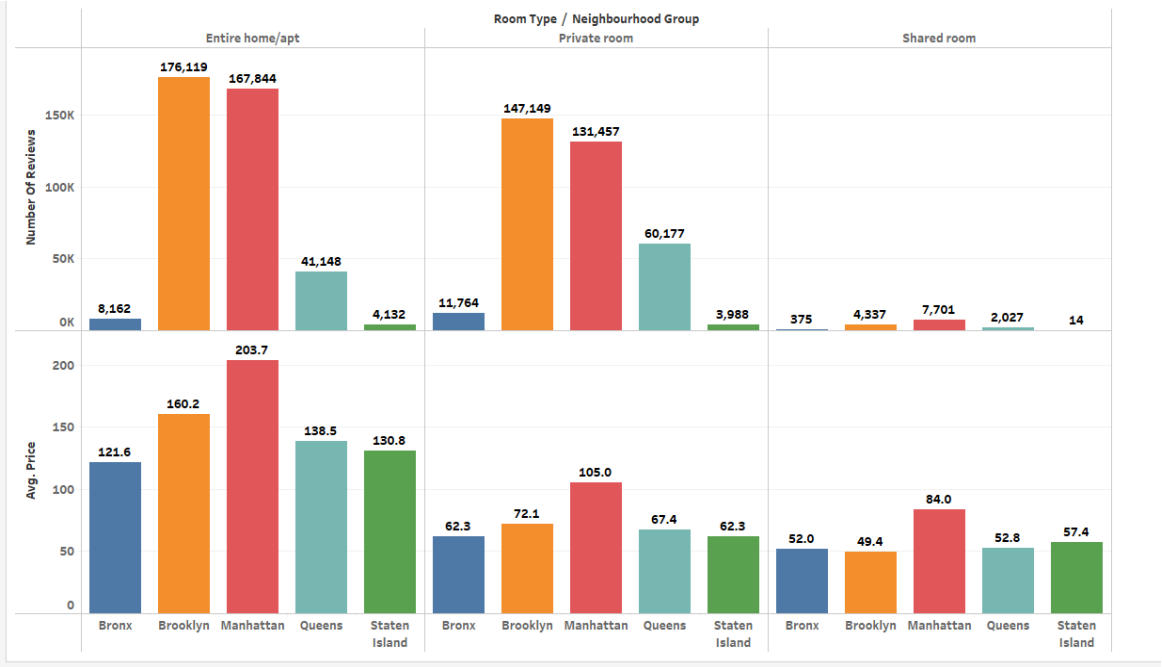

Figure 9. Reviews and pricing for Room type and their neighborhood groups

Inference: -

- Brooklyn and Manhattan have the greatest number of reviews for Entire homes/Apt and for Private rooms too. As per reviews, their pricing is high as well, but visitors prefer that.
- On the other hand, Queens Bronx and Staten Island have fewer visitors respectively but their pricing is far higher for all types of rooms.
- Very few visitors rent Shared rooms in every Neighborhood group.
- Staten Island is having the lowest number of visitors still the price for even shared rooms is higher than in Queens and Bronx.
- Promoting Shared rooms or omitting them or even lowering their prices can make a great difference.

3. Neighborhood-wise Occupancy rate and Availability of rooms –

Checking the occupancy rate and availability of rooms can tell us whether our findings match earlier questions or not. We created a new calculated field to find out the Occupancy rate of each neighborhood. This graph is further sorted according to the Average Availability of rooms.
Occupancy rate and Availability are inversely proportional to each other.
It is calculated as follows-

Figure 10. Calculation for Occupancy Rate

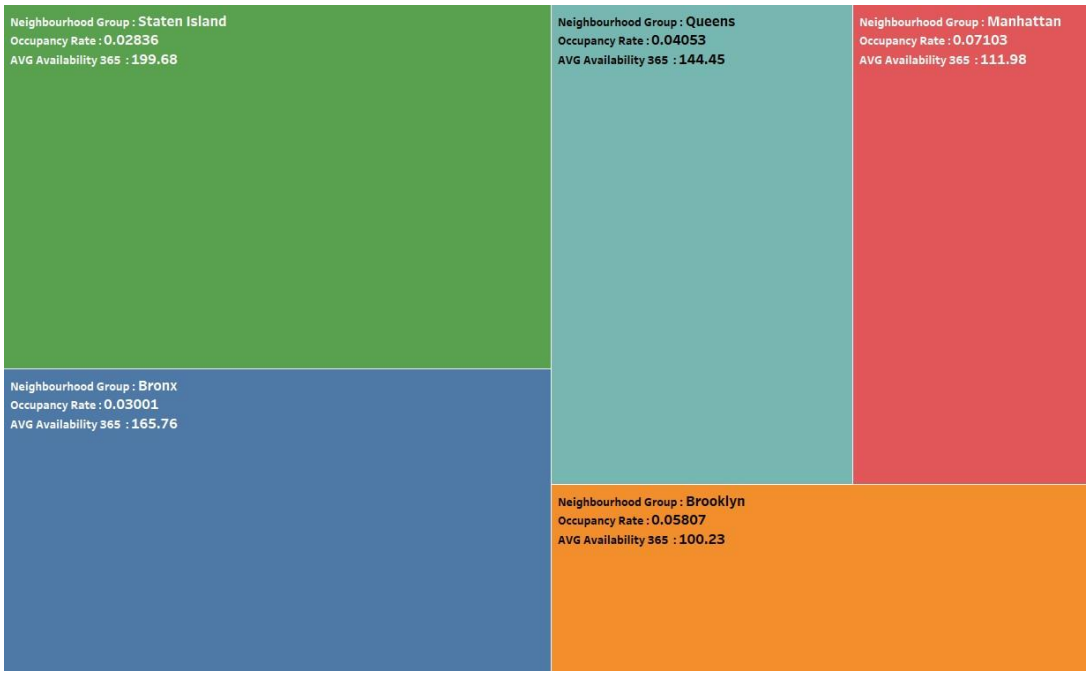

Figure 11. Neighborhood-wise Occupancy rate and Availability of rooms

Inference: -

- As per the graph, the Average Availability of Staten Island, Bronx and Queens are more than Brooklyn and Manhattan whereas their Occupancy rate is vice versa of Availability.
- Staten Island has fewer visitors hence the available Rooms will be higher and is the same for Bronx and Queens, whereas, Brooklyn and Manhattan have more visitors hence the available rooms will be less

4. Minimum Nights preferred by customers –

We wanted to observe the customer booking pattern and demand of property based on the minimum number of stay nights. This was chosen to understand for what type of stay customers use Airbnb; short-stay or long-stay. Here, we took into account the volume of booking and the neighborhood- wise volume of booking. The parameters taken into account were: -
CNT(Id), Minimum Nights (This was binned, with a bin size of 2 for easier visualization) & Neighborhood Group.

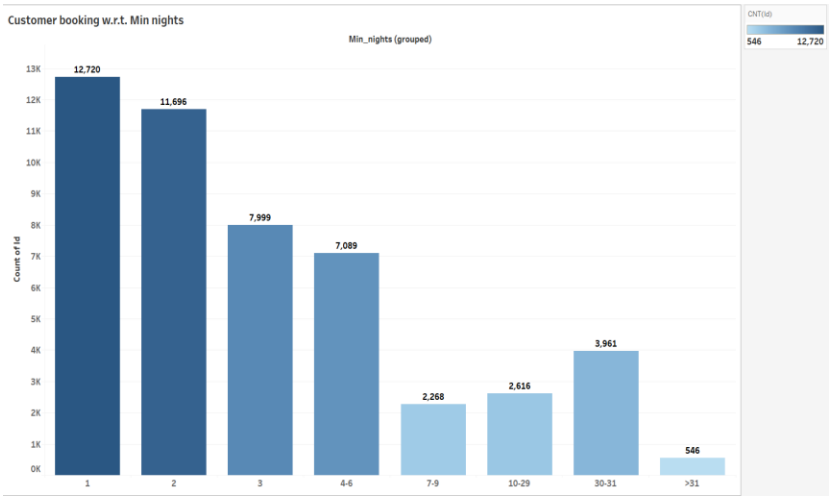

Figure 12. Customer booking w.r.t minimum Nights

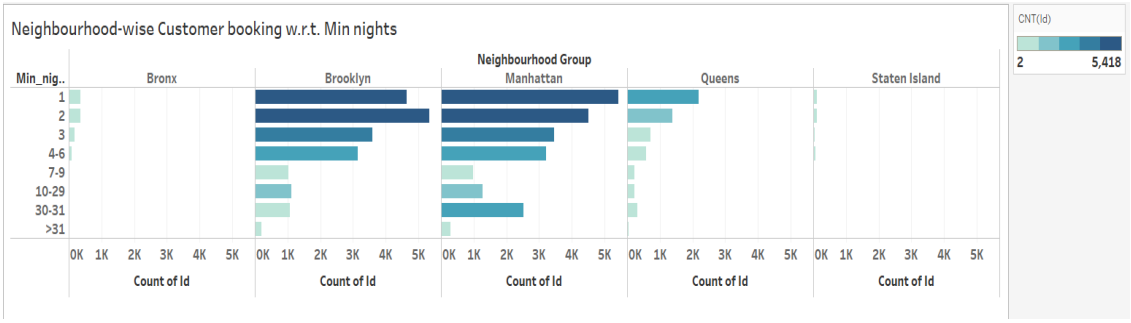

Figure 13. Neighborhood-wise Customer booking w.r.t minimum Nights

Inference: -

- The listings with Minimum nights 1-6 have the greatest number of bookings.
- We can see a prominent spike in 30 days; this would be because customers would rent out on a monthly basis. After 30 days, we can also see small spikes at 60 & 90 days, this can be explained by the monthly rent-taking trend.
- Manhattan & Brooklyn have a higher number of 30-day bookings compared to the others. The reason could be either tourist booking long stays

5. Pricing ranges preferred by the customers –

After the analysis, we need to find out which pricing range is best suitable for the customers. As the visitor's main thing to remember is about their budget/spent for the whole trip.
The parameter taken for the analysis is: - Pricing frequency with number of reviews

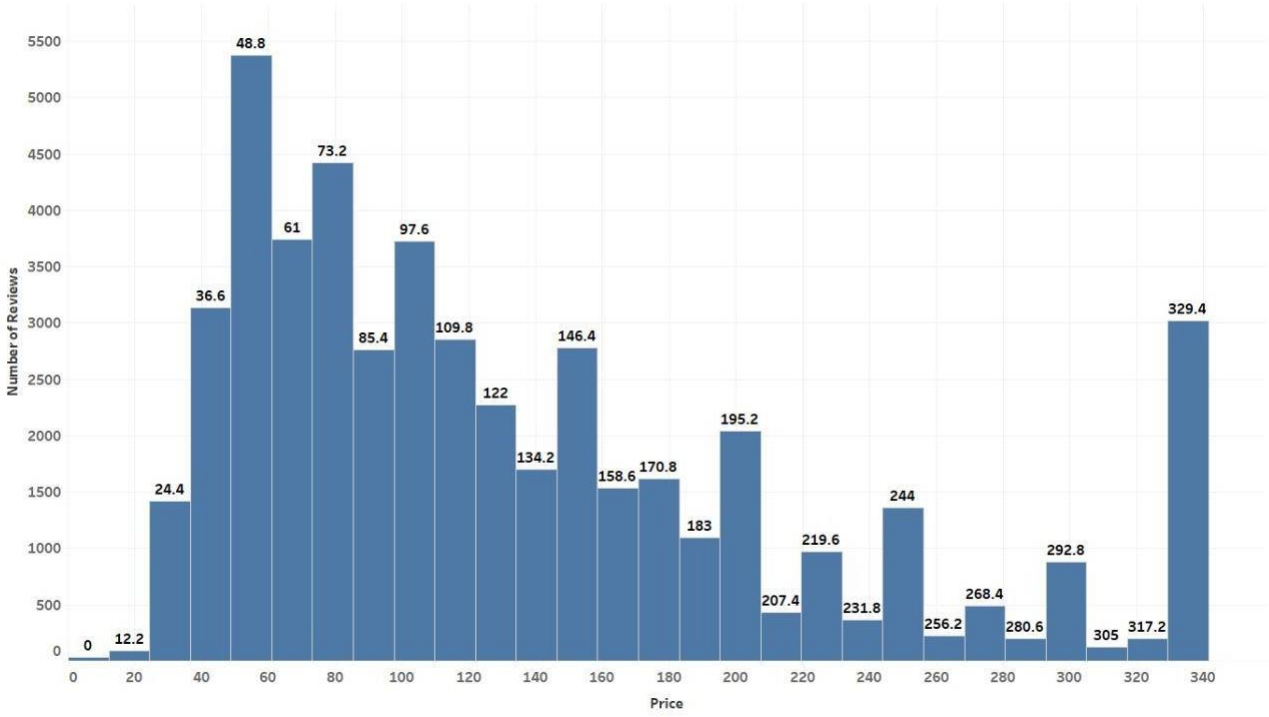

Figure 14. Price range w.r.t Number of Reviews

Inference: -

- Most of the customers prefer $48 to $150 pricing for their trip as most number of people have paid.
- As we did the capping for outliers the Number of Reviews is shown high for $330, but it might be because of the treatment of outlier, as $330 is the only showing large number of visitors in higher range.

6. Room Type and Neighborhood Group Effects on Average Price and Minimum Nights –

Pricing of room type according to the minimum nights can tell us if the pricing is proper according to the minimum nights given by the neighborhood groups. It will help to attract customer if the pricing of minimal stays is according to their preference.
For this question, first, we saw the percentage of visitors preferring which room type.

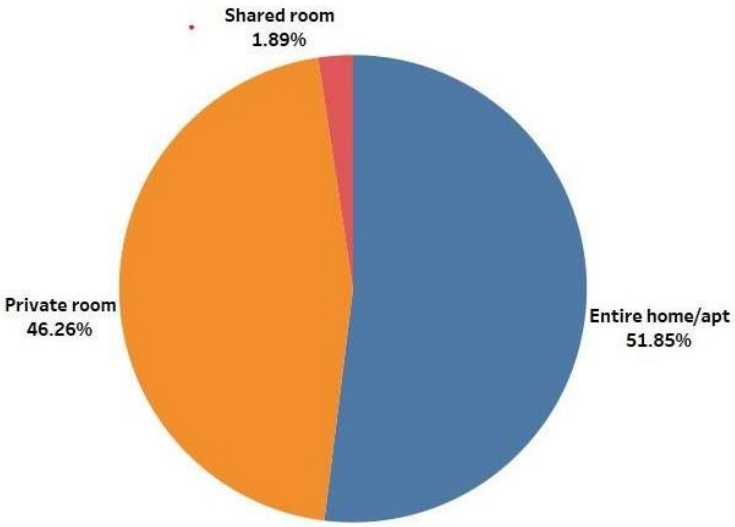

Figure 15. Percentage of visitors for each Room Type

Then, we prepared a Min. night for each Room type and their pricing graph where the parameters were: -
Average of minimum nights, average Price, room type, and neighborhood groups.

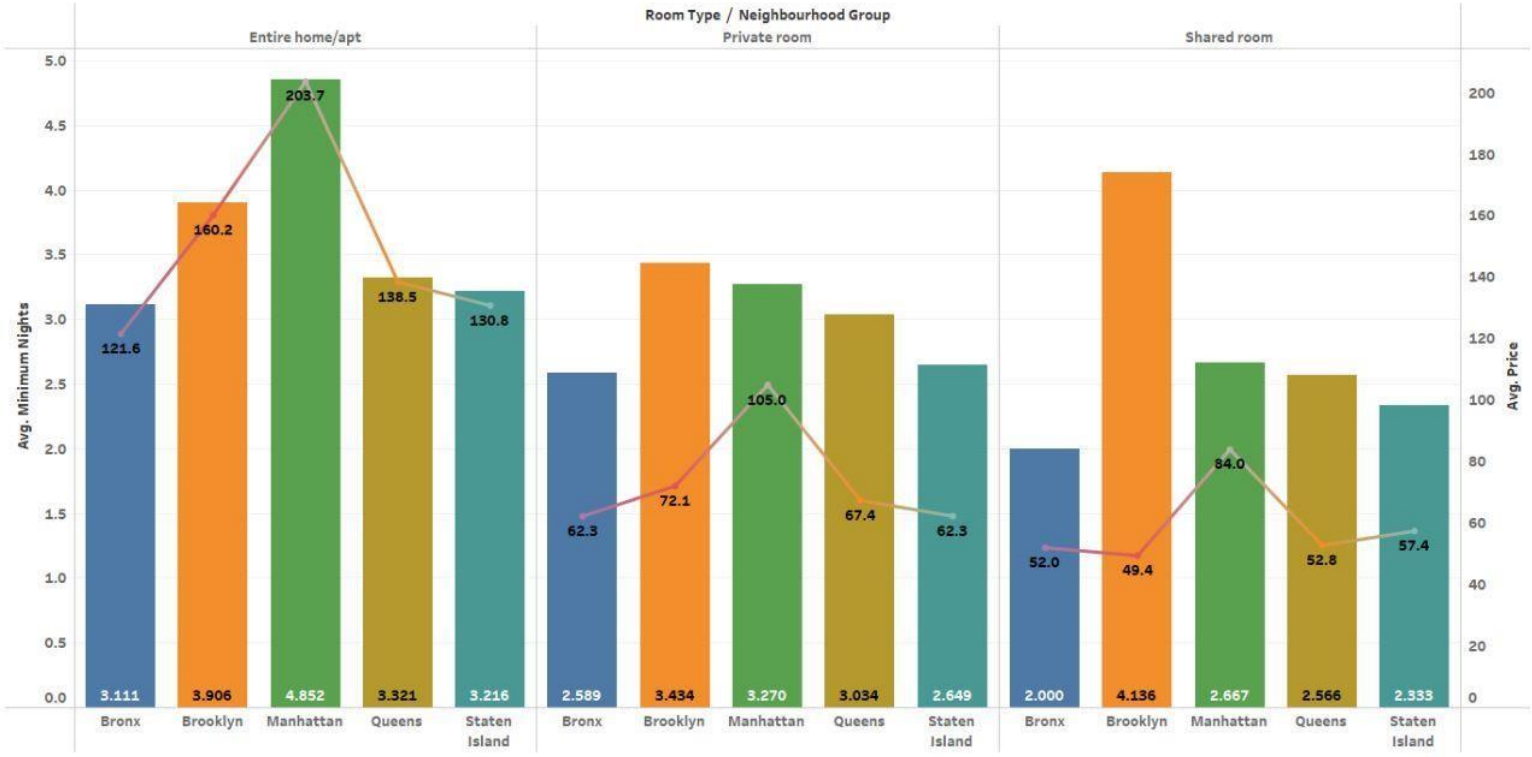

Figure 16. Avg Min. nights and Avg Price w.r.t each Room type and Neighborhood

Inference: -

- In Manhattan, Entire home/ Apt are most likely to be taken for an average of 5 minimal nights whereas Private rooms and Shared rooms are taken mostly in Brooklyn for 3-5 minimal nights.
- The average pricing for the Entire home/Apt is much higher than other room types.
- For Private rooms in Manhattan average minimum nights are almost 3.3 but pricing for that room is much high. If minimum nights for Manhattan are given more or the pricing decreased a bit, it will be very preferable for the customers/visitors.
- For Shared rooms in Brooklyn average minimum night's stay is of 4 nights but pricing is a little low. Shared rooms are not taken much so having high price for them will only repulse visitors' attention.

7. <u>Top Ten Neighborhoods w.r.t various Parameters –</u>

We analyzed which are the top ten Neighborhoods according to the Pricing, Number of reviews, and Availability.

With the help of various parameters, we can see their impact on the neighborhoods and accordingly take relevant measures to generate more revenue. These are the Top 10 locations :-

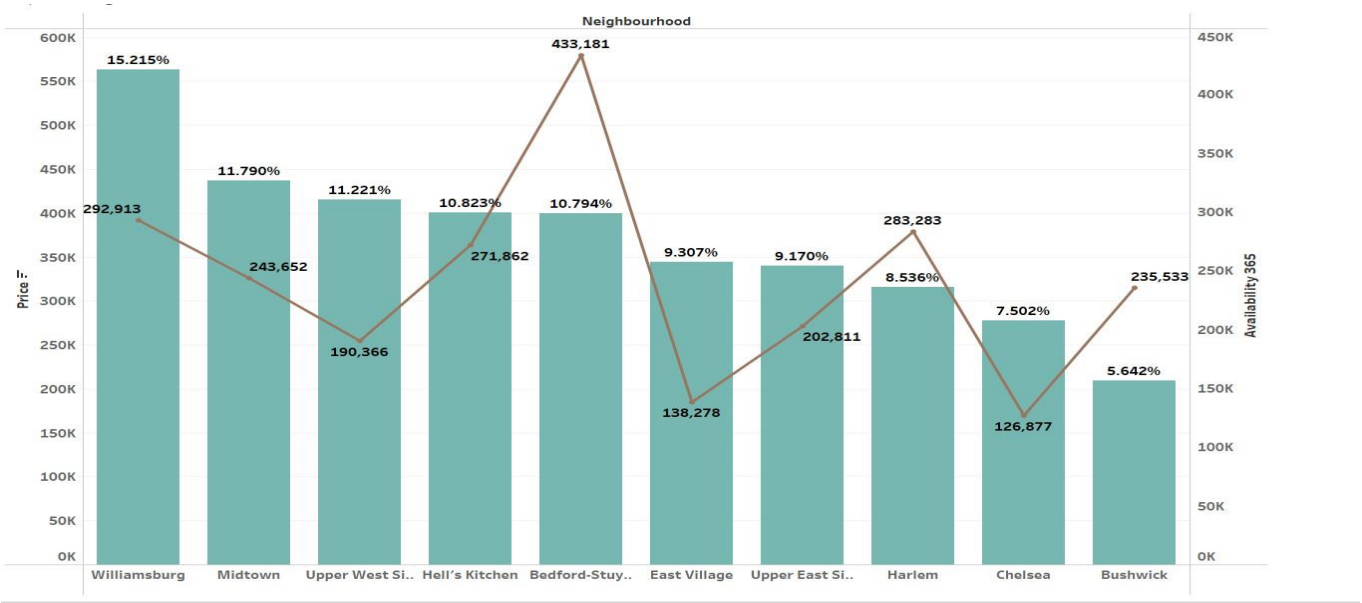

Figure 17. Top 10 Neighborhoods w.r.t Price and Availability

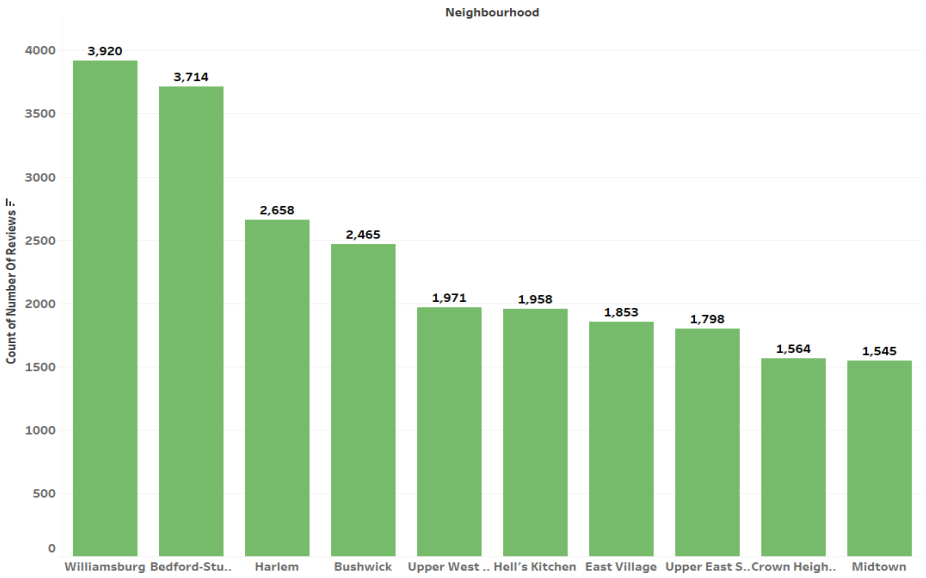

Figure 18. Top 10 Neighborhoods w.r.t No. of Reviews

Inference: -

- From these graphs, we infer that Williamsburg is the neighborhood which is costly and available, but still the count of reviews of that place is also higher than others.
- Bedford is the second highest famous neighborhood among the customers having the highest availability as its Pricing is respectively low.
- Bushwick is another developing and evolving neighborhood with less percentage of pricing and a greater number of visitors comparatively. It is an industrial area having imaginary and awesome street art which attracts people as it is affordable.

8. Top hosts according to various parameters –

Hosts are something quite important for any trip. So, we decided to analyze top hosts and their names according to different parameters and look into them to infer something. To see the top hosts, we used the parameters: -

1) Calculated host listing count
2) Number of Reviews.

Sum of calculated host listing count gives the sum of listings for that particular host and sum of number of Reviews will give total sum of reviews given to that host.

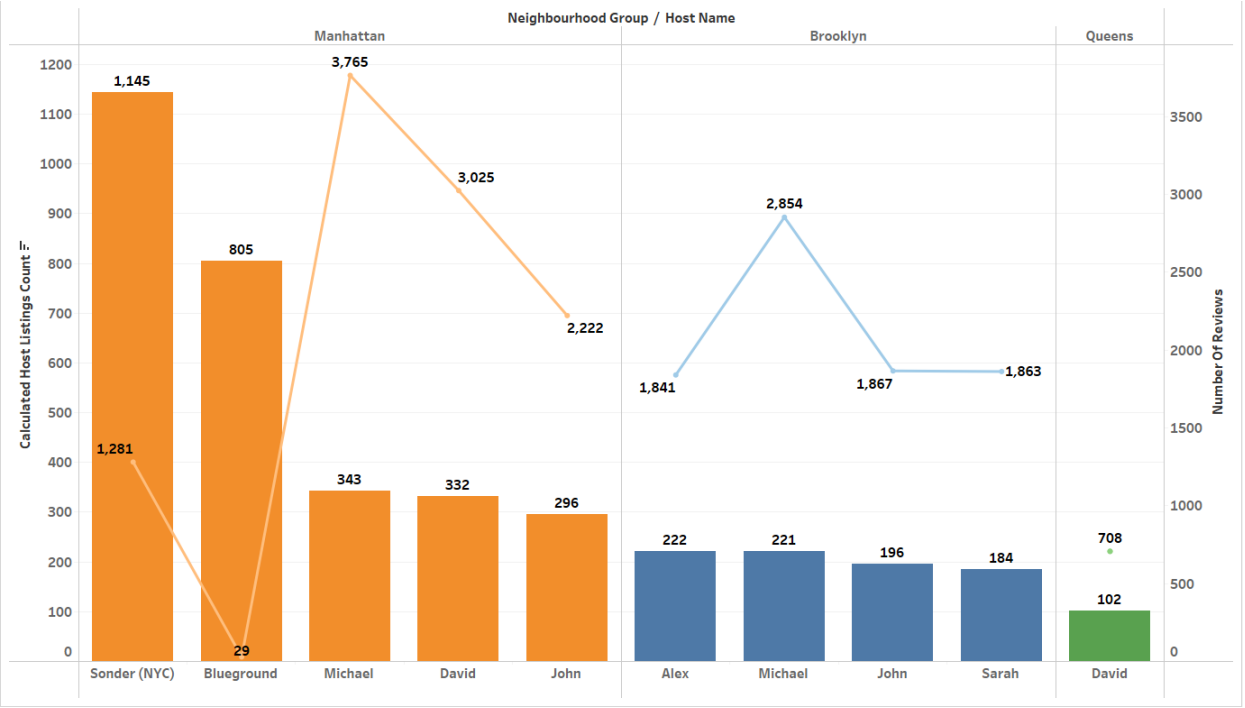

Figure 19. Top 10 Hosts w.r.t Reviews and Listings

Inference: -

- Sonder and Blue ground are the top 2 hosts who have the greatest count of listings. However, the Review given to them by the customers/visitors is way less than the other hosts in different neighborhood groups.
- Michael, David, John, Alex and Sarah are the top 5 hosts according to the reviews given to them by the visitors. To generate more revenue these 5 hosts, need to have more listings for all the neighbourhood groups.

To look if these 5 hosts generate revenue for Airbnb or not, we plotted another graph where wetook the top ten hosts according to the number of reviews and the price they generate.

The parameters used were: - Sum of the Number of Reviews and Total Price for each host inn sideby side graphs.

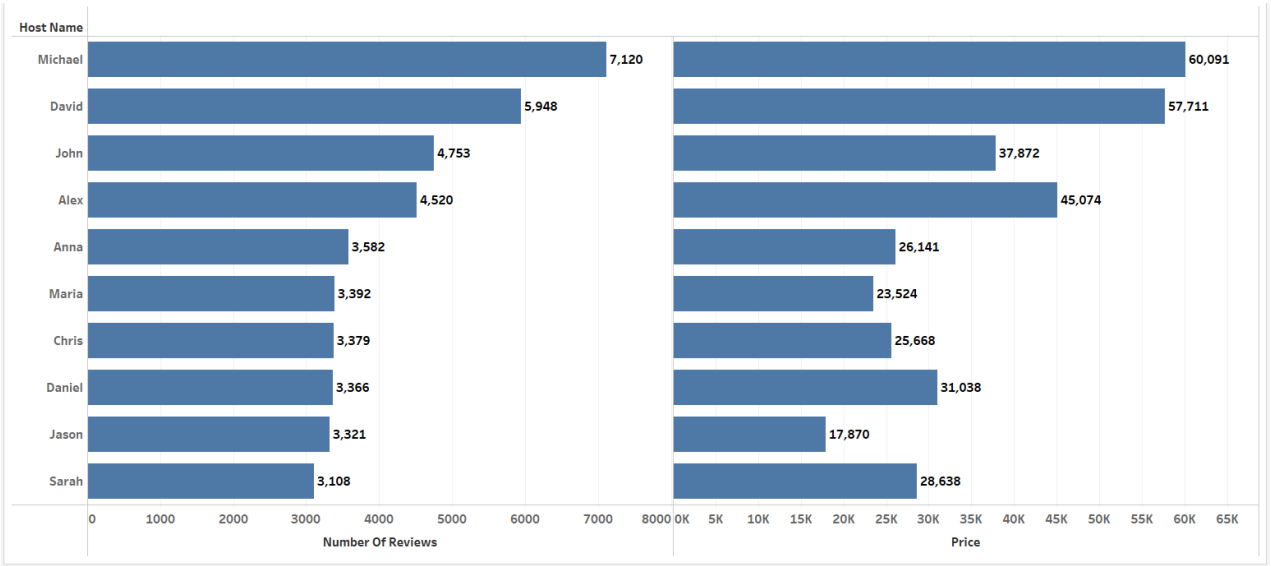

Figure 20. Top Hosts w.r.t Sum of No. of Reviews and their total Pricing

Inference: -

- Michael, David, Alex, John, and Daniel are the Top 5 hosts that seem to have received the highest number of reviews for their listed sites and have also sites listed with a high price range.
- These hosts need to acquire more in every neighborhood group so as to have more visitors attended by them.

9. <u>Understanding Price variation w.r.t Geography –</u>

We had earlier explored the price variation with respect to location. We now deep dive to understand how it varies across different areas/geographies.

We wanted to understand if geography played a part in rising prices. For this, we plotted a geographical map to understand the price density and variation

To further correlate our findings, we took the top 10 neighborhoods with maximum average price.

We used the findings in this to confirm our observation obtained from the geographical map.

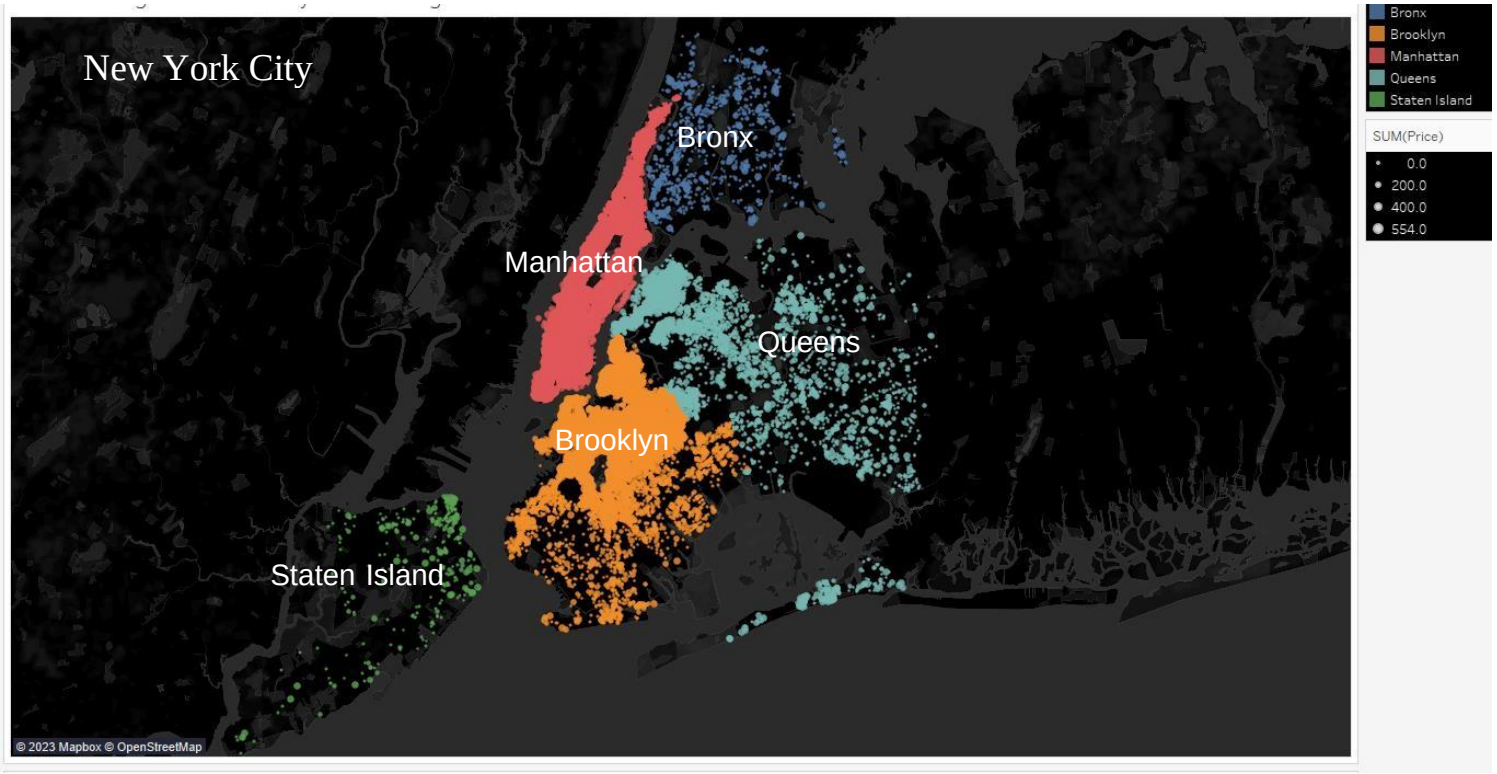

Figure 21. Neighborhood groups and Pricing density

Inference: -

- The map displays the price variation, which appears to be distributed uniformly in the inland areas.
- We see spike in prices in coastal cities, owing to better view from stays and easy ferry reachability.

- When we zoomed in, we also observed higher pricing near colleges or important monuments/landmarks.
- Increasing acquisitions and new properties in coastal regions can increase customer bookings.

After these visualizations, there were made two presentations which showed our visualizations about the Airbnb Problem Statement, to present in front of the Managers and Heads.

Two presentations prepared were:

1) Presentation-I: - Technical PPT for Data Analysis Managers and Lead Data Analyst.

2) Presentation-II: - Non-technical PPT for Head of Acquisitions and Operations, NYC and Head of User Experience, NYC.

RECOMMENDATIONS:

- It states that may be due to high pricing not many visitors are eager to go there. Maybe pricing changes or discounts will attract customers to Staten Island and Bronx.
- Staten Island is having the lowest number of visitors still the price for even shared rooms there is higher than in Queens and Bronx. Promoting Shared rooms or omitting them or even lowering their prices can make a great difference.
- More number of hosts & listings with monthly rental duration (30-60-90) can be acquired. We see a good potential in the 30-day rental window. Manhattan & Brooklyn have higher number of 30-day bookings compared to the others, these areas can be further targeted.
- Also, weekly or bi-weekly rentals can also be acquired as these can be used customers stranded in NYC for quarantine purposes.
- On average Entire home/apt types are preferred more by the customers followed by Private rooms and then Shared Rooms. Mostly because they are also available for a higher number of minimum night's stay window booking as compared to Private and Shared rooms.
- New acquisitions can be explored to acquire 'private rooms' in Manhattan and Brooklyn and 'entire homes' in Bronx and Queens.
- New acquisitions and expansion can be done in the price range of $40 - $150 as it satisfies both parameters of volume of customer traffic and customer satisfaction.
- If minimum nights for Manhattan are given more or the pricing decreased a bit, it will be very preferable for the customers/visitors. Shared rooms are not taken much so having high price for them will only repulse visitors' attention. Instead, dropping the prices and increasing promotion will help a lot.
- The price variation, which appears to be distributed uniformly in the inland areas. There are spike in prices in coastal cities, owing to better view from stays and easy ferry reachability. Increasing acquisitions and new properties in coastal regions can increase customer bookings.
- We can confirm that the greatest parameter for any customer to prefer a property and provide a review is having a maximum or minimum night stay window booking and their probability of good hosts to attend the customer to some extent.