<u>AIRBNB Case Study Data Methodology</u>

(Shital Prashant Jadhav and Inder Mukhopadhyay)

<u>Data Wrangling</u>:

In the case study, we utilized Jupiter Notebook for conducting initial data analysis and Tableau for data analysis and visualization.

Initial Analysis using Jupiter Notebook: Data Set Used: AB_NYC_2019.csv.

This dataset contains 48,895 rows and 16 columns.

In [1]:
```
# Import the necessary libraries
import warnings
warnings.filterwarnings("ignore")
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

In [2]:
```
# Data conversion and Understanding
airbnb_data = pd.read_csv("AB_NYC_2019.csv")
airbnb_data.head(5)
```

Out[2]:

	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 [3]:
```
# Check the rows and columns of the dataset
airbnb_data.shape
```

Out[3]: (48895, 16)

In [4]:
```
# To see Non-Null counts and data types
airbnb_data.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
```

last_review is of object Data type. datetime64 is a better Data type for this column.

In [5]:
```python
airbnb_data.last_review = pd.to_datetime(airbnb_data.last_review)
airbnb_data.last_review
```

Out[5]:
```
0        2018-10-19
1        2019-05-21
2               NaT
3        2019-05-07
4        2018-11-19
            ...
48890           NaT
48891           NaT
48892           NaT
48893           NaT
48894           NaT
Name: last_review, Length: 48895, dtype: datetime64[ns]
```

In [6]:
```python
# Percentage of missing values
round((airbnb_data.isnull().sum()/len(airbnb_data))*100,2)
```

Out[6]:
```
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
```

- There are a small proportion of null values which would not affect my analysis so let them stay as it is.
- Two columns (last_review , reviews_per_month) has around 20.56% missing values.
- We need to see if the values are Missing completely at random(MCAR) or Missing not at random(MNAR).
- There is no dropping or imputation of columns as we are just analyzing the dataset and not making a model.Also most of the features are important for our analysis.

In [7]:
```python
# Now reviews per month contains more missing values which should be replaced with 0 respectively
airbnb_data.fillna({'reviews_per_month':0},inplace=True)
```

In [8]:
```python
airbnb_data.reviews_per_month.isnull().sum()
```

Out[8]: 0

Missing values Analysis

In [9]:
```python
# Selecting the data with missing values for 'last_review' feature
airbnb_data_1 = airbnb_data.loc[airbnb_data.last_review.isnull(),:]
airbnb_data_1
```

Out[9]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	nu
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
19	7750	Huge 2 BR Upper East Cental Park	17985	Sing	Manhattan	East Harlem	40.79685	-73.94872	Entire home/apt	190	7	
26	8700	Magnifique Suite au N de Manhattan - vue Cloitres	26394	Claude & Sophie	Manhattan	Inwood	40.86754	-73.92639	Private room	80	4	
36	11452	Clean and Quiet in Brooklyn	7355	Vt	Brooklyn	Bedford-Stuyvesant	40.68876	-73.94312	Private room	35	60	
38	11943	Country space in the city	45445	Harriet	Brooklyn	Flatbush	40.63702	-73.96327	Private room	150	1	
...	...	...	...	...	...	...	...	...	...	...	...	

In [11]:
```python
# Count of 'neighbourhood_group'
airbnb_data.groupby('neighbourhood_group').neighbourhood_group.count()
```

Out[11]:
```
neighbourhood_group
Bronx             1091
Brooklyn         20104
Manhattan        21661
Queens            5666
Staten Island      373
Name: neighbourhood_group, dtype: int64
```

Missing values Analysis ('neighbourhood_group' feature)

```
In [10]: # Count of 'neighbourhood_group' with missing values
         airbnb_data_1.groupby('neighbourhood_group').neighbourhood_group.count()
```

```
Out[10]: neighbourhood_group
         Bronx              215
         Brooklyn          3657
         Manhattan         5029
         Queens            1092
         Staten Island       59
         Name: neighbourhood_group, dtype: int64
```

```
In [11]: # Count of 'neighbourhood_group'
         airbnb_data.groupby('neighbourhood_group').neighbourhood_group.count()
```

```
Out[11]: neighbourhood_group
         Bronx             1091
         Brooklyn         20104
         Manhattan        21661
         Queens            5666
         Staten Island      373
         Name: neighbourhood_group, dtype: int64
```

```
In [12]: (airbnb_data_1.groupby('neighbourhood_group').neighbourhood_group.count()/airbnb_data.groupby('neighbourhood_group').neighbourho
```

```
Out[12]: neighbourhood_group
         Bronx            19.706691
         Brooklyn         18.190410
         Manhattan        23.216841
         Queens           19.272856
         Staten Island    15.817694
         Name: neighbourhood_group, dtype: float64
```

```
In [24]: ((airbnb_data_1.groupby('neighbourhood_group').neighbourhood_group.count()/airbnb_data.groupby('neighbourhood_group').neighbourh
```

```
Out[24]: 19.240898461107257
```

Each neighbourhood_group has about 19 % missing values in 'last_review' feature.

Missing values Analysis ('room_type' feature)

```
In [26]: # Count of 'room_type' with missing values
         airbnb_data_2 = (airbnb_data_1.groupby('room_type').room_type.count()/airbnb_data.groupby('room_type').room_type.count())*100
         airbnb_data_2
```

```
Out[26]: room_type
         Entire home/apt    19.981109
         Private room       20.877004
         Shared room        27.068966
         Name: room_type, dtype: float64
```

'Shared room' has the highest missing value percentage (27 %) for 'last_review' feature while to other room types has only about 20 %

Missing values Analysis ('price' feature)

```
In [27]: print("Mean when last_review missing  = ", airbnb_data[airbnb_data['last_review'].isnull()].price.mean())
         print("Median when last_review missing  = ",airbnb_data[airbnb_data['last_review'].isnull()].price.median())

         print("Mean when last_review not missing   = ",airbnb_data[airbnb_data['last_review'].notnull()].price.mean())
         print("Median when last_review not missing  = ", airbnb_data[airbnb_data['last_review'].notnull()].price.median())
```

```
Mean when last_review missing  =  192.9190210903303
Median when last_review missing  =  120.0
Mean when last_review not missing   =  142.317946605566
Median when last_review not missing  =  101.0
```

INFERENCES:

- The pricing is higher when 'last_review' feature is missing .
- reviews are less likely to be given for shared rooms.
- When the prices are high reviews are less likely to be given.
- The above analysis seems to show that the missing values here are not MCAR (missing completely at random)

```
In [28]: # Now to check the unique values of other columns'
         airbnb_data.room_type.unique()
```

```
Out[28]: array(['Private room', 'Entire home/apt', 'Shared room'], dtype=object)
```

```
In [31]: len(airbnb_data.room_type.unique())
```

```
Out[31]: 3
```

```
In [29]: airbnb_data.neighbourhood_group.unique()
```

```
Out[29]: array(['Brooklyn', 'Manhattan', 'Queens', 'Staten Island', 'Bronx'],
               dtype=object)
```

```
In [32]: len(airbnb_data.neighbourhood_group.unique())
```

```
Out[32]: 5
```

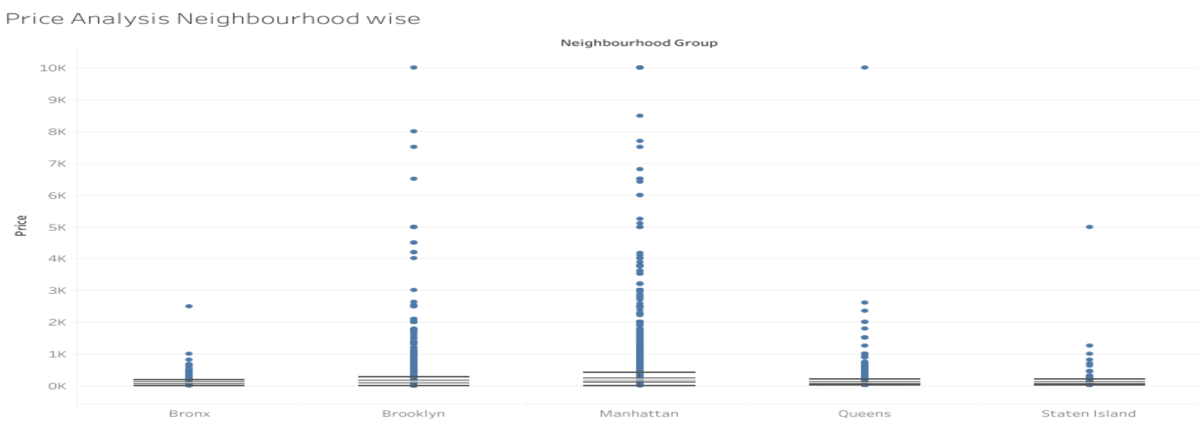
```
In [33]: len(airbnb_data.neighbourhood.unique())
```

```
Out[33]: 221
```

```
In [34]: airbnb_data.to_csv('AB_NYC_2019_processed.csv')
```

➢ Checked data type of variables. last _review Object Data type is converted to datetime64 Data type.

- Checked the Null Values in our dataset. Columns like name, host_name, last_review and review_per_month have null values. Columns with smaller portion of Null values would not affect my analysis so we let them stay as it is.
- Missing values in reviews_per_month column imputed with 0.
- Two columns (last_review, reviews_per_month has more than around 20.56% missing values. Missing value in last_review column is not MCAR. These columns not dropped or imputed as we are just analysing the dataset and not making a model.
- Checked the Duplicate row in our dataset and no duplicate data was found.
- Price was highly positively skewed so median was very close the lower quartile with some outliers as seen in the boxplot below.



- Created a grouped field for Minimum_Nights assuming null values belonged to the category.

Describe Field

Minimum night Bin

Role:	Discrete Dimension
Type:	Calculated Field
Contains NULL:	No
Locale:	
Sort flags:	Case-sensitive
Column width:	5
Status:	Valid

Formula

```
IF [Minimum Nights] = 1 THEN "1"
ELSEIF [Minimum Nights] = 2 THEN "2"
ELSEIF 4<=[Minimum Nights] AND [Minimum Nights]<=5 THEN "4-5"
ELSEIF 6<=[Minimum Nights] AND [Minimum Nights]<=7 THEN "6-7"
ELSEIF 8<=[Minimum Nights] AND [Minimum Nights]<=29 THEN "8-29"
ELSEIF 30<=[Minimum Nights] AND [Minimum Nights]<=31 THEN "30-31"
ELSE ">31" END
```

<u>Presentation 1</u>:

Objective:

- To conduct a thorough analysis of New York Airbnb Dataset.
- Ask effective questions that can lead to data insights.
- To process data, analyze and share findings by data visualization and statistical techniques.

EDA:

To understand some important insights we have explored the following questions:

- How are the Airbnb listings spread out in NYC and most contributing neighbourhood?
- What type of rooms do customers prefer?
- What could be the ideal number of minimum nights to increase customer bookings?

Based on customer review:

- Most preferred neighbourhood.
- Most preferred room type.
- Who are the Hosts who have the highest listings w.r.t. Neighbourhood?

Methodology:

- The data was analyzed through univariate and bivariate analysis.
- The analysis and visualizations were done using Tableau considering various parameters.
- The main parameters that have been taken into account for analysis are :

  1. Bookings based on Neighbourhood Groups
  2. Bookings based on Room type
  3. Number of reviews
  4. Minimum number of nights
- Inferences have been made keeping in mind the above parameters.

Explanation for EDA:

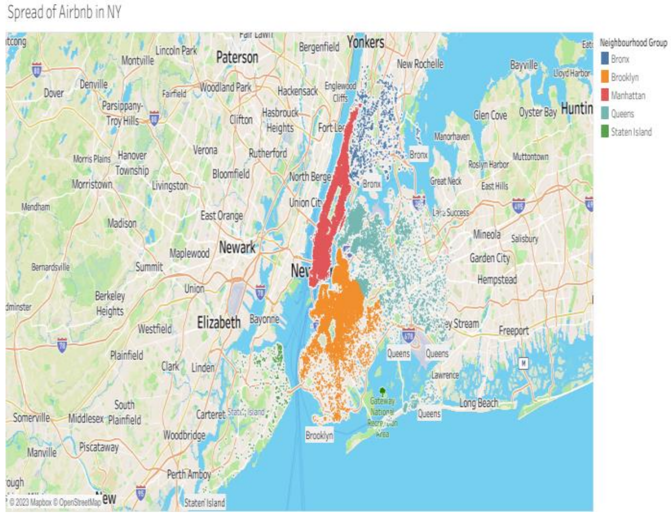
<u>How are the Airbnb listings spread out in NYC and most contributing neighbourhood</u>?

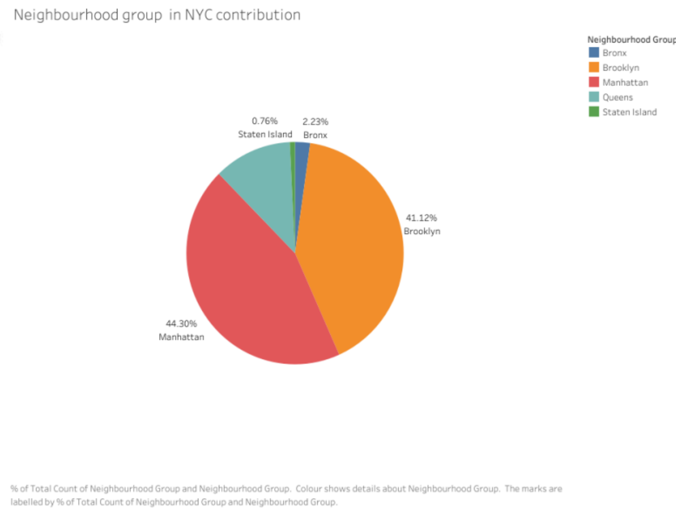
To understand the spread of listings in the NYC areas and the concentration of listings in each neighbourhood group, two visualizations were utilized: a geographical plot and a pie chart.

Two plots were used to explore this question:

Geographical plot: This was created using the parameters latitude, longitude, neighbourhoods, and neighbourhood group as parameters. This plot provided a visual representation of the areas under consideration, allowing for a better understanding of the geographic distribution of the listings across different neighborhoods in NYC.

Pie Chart: On the other hand, the pie chart was employed to examine the contribution of each neighbourhood group to the total count of listings. It made use of the parameters neighbourhood group and the percentage of the total count of neighbourhood groups.





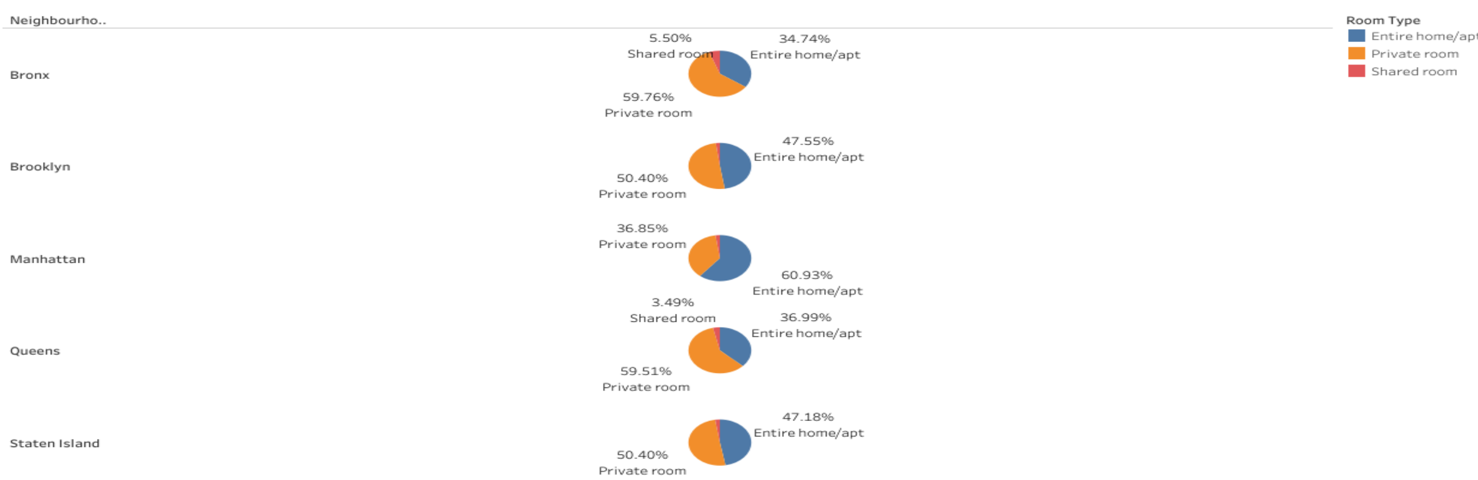
Inferences:

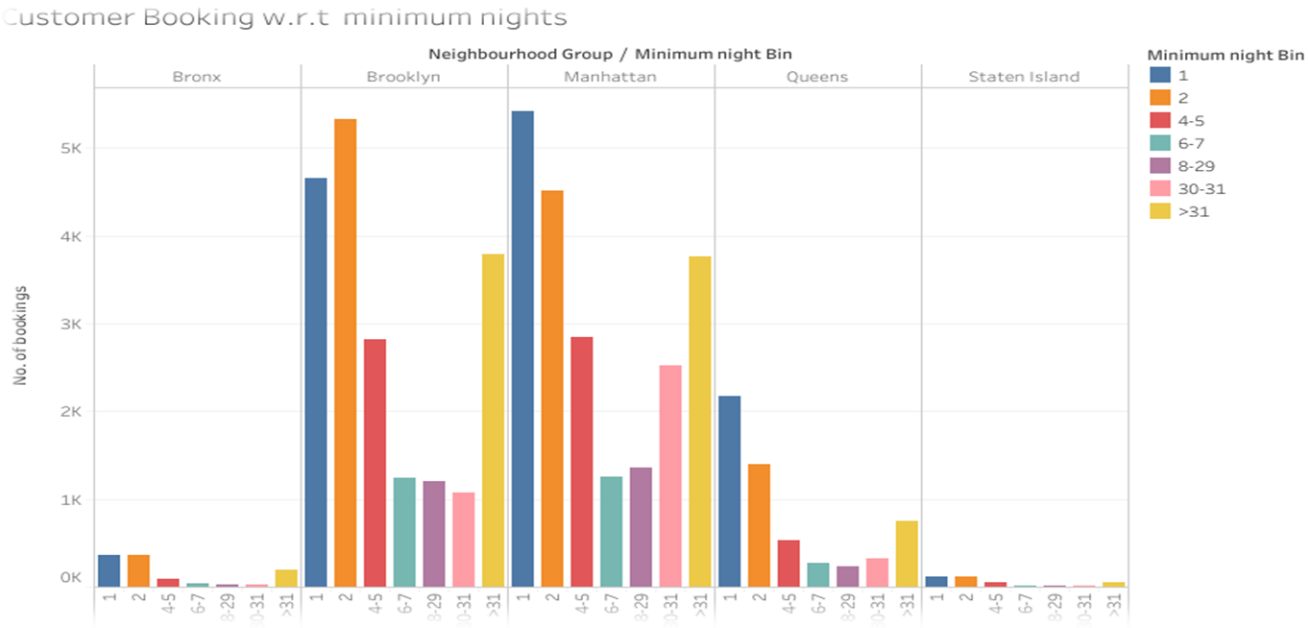
- Manhattan and Brooklyn are the prime hubs for Airbnb listings in New York City, with significant presence in both neighbourhood groups.
- Listings are maximum in Manhattan (44%) & Brooklyn (41%) neighbourhood group.
- Staten Island has the smallest proportion of listings, representing only around 1% of the total.

<u>What type of rooms do customers prefer?</u>

This question was addressed to understand the space needs of the customer and their preference. This has been explored using pie charts and side by side Bar graph.

- The first chart broke down the customer preference according to the neighbourhood group.
- The second chart showed the overall preference of the customer across NYC.





Inferences:

- The majority of bookings on Airbnb are made for listings with a minimum stay of 1-5 nights, indicating their popularity among guests.
- Notably, there is a significant increase in bookings for 30-day stays, which can be attributed to customers renting accommodations on a monthly basis.
- In terms of specific boroughs, Manhattan and Brooklyn stand out with a higher number of 30-day bookings compared to other areas. The reason could be either tourists booking long stays or mid-level employees who opt for budget bookings due company visits.

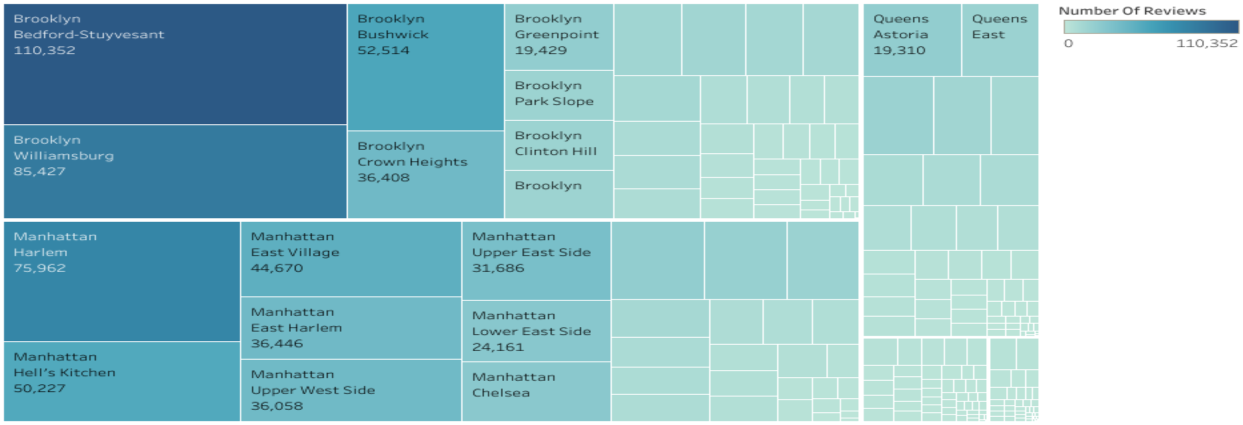
Based on customer review:

Most preferred neighbourhood & Most preferred room type:

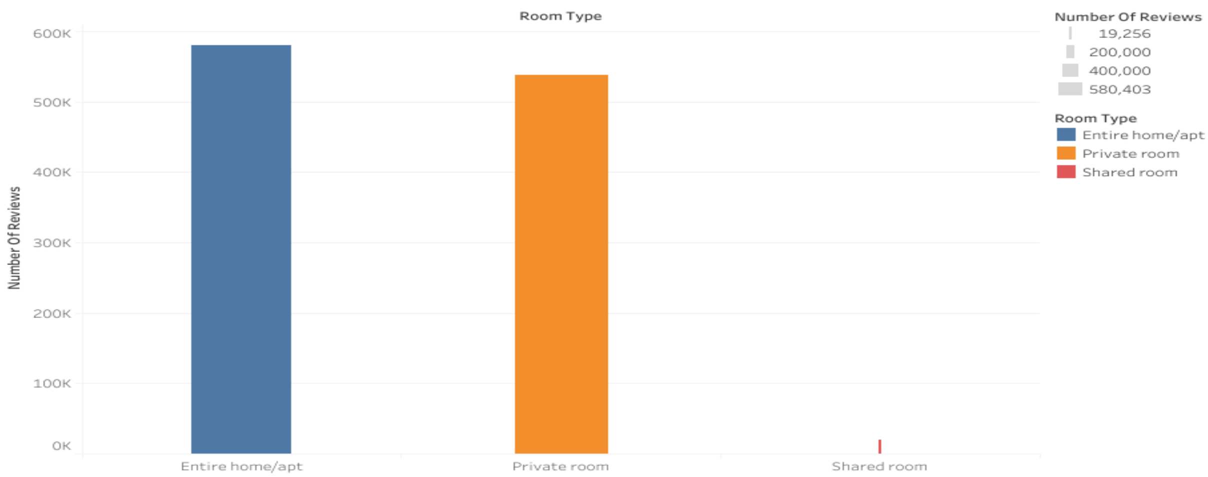
To further analyze the impact of customer reviews on listings in NYC, two parameters were considered: room type and neighbourhood. The number of reviews obtained for a particular listing is a direct indicator of its likability and can influence future bookings.

The parameters taken for analysis are: Room type; Neighbourhood, Neighbourhood group, SUM (Number of reviews).

Popular Neighbourhoods



Total Reviews w.r.t Room type



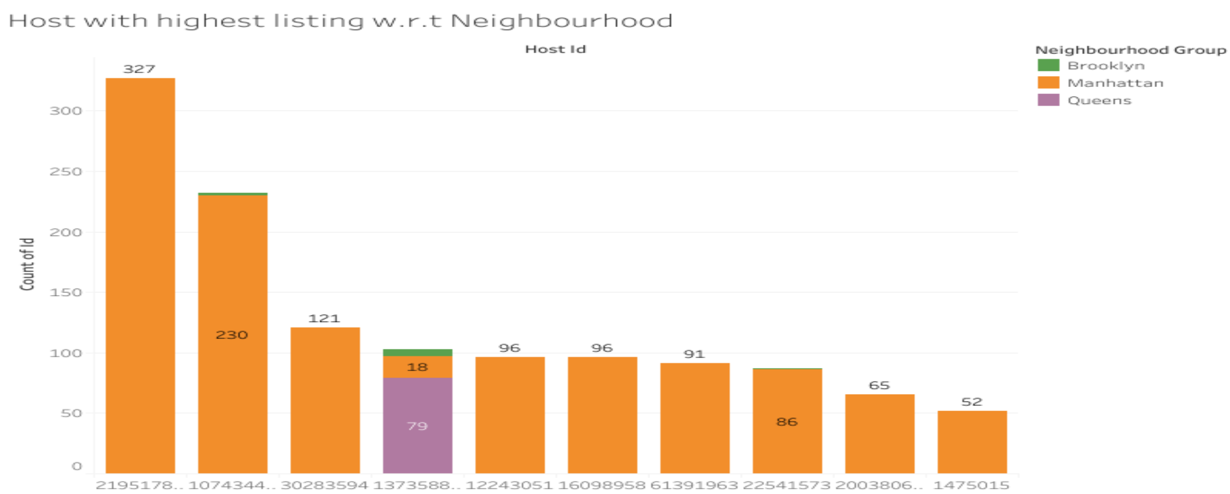
Sum of Number Of Reviews for each Room Type. Colour shows details about Room Type. Size shows sum of Number Of Reviews.

Inferences:

- Customers exploring the neighbourhoods of Brooklyn and Manhattan are inclined to offer their feedback.
- Encouraging customers who visit the vibrant neighbourhoods of Brooklyn and Manhattan to share their feedback is essential in order to facilitate continuous improvement and share valuable insights with other listings.
- Based on the maximum number of reviews received, it can be inferred that customers tend to favor the 'Entire home/apt' and 'Private rooms' options over 'Shared rooms'.

<u>Who are the Hosts who have the highest listings w.r.t Neighbourhood?</u>

The analysis focused on identifying the maximum number of listings held by individual hosts and their distribution across different areas. We have taken the Host ID in the x-axis with the CNT (Id) in the y-axis.The top 10 hosts were filtered based on their number of listings, and the graph was color-coded by neighbourhood group. This provided a concise overview of host investments and expansions in specific areas.



Inferences:

- An interesting observation is the presence of a single host managing multiple listings, particularly in the Manhattan area. This trend can be attributed to the fact that Manhattan attracts a significant number of tourists.
- Overall, the strategic decision of experienced hosts to focus on the Manhattan area stems from the high volume of tourists and financial enthusiasts it attracts, making it a profitable and sought-after location for short-term rentals.

Presentation 2:

Objective:

- To gain insights into customer preferences and enhance their experience when using Airbnb listings.
- To will analyse how various parameters influence pricing in Airbnb listings.
- To provide specific and actionable suggestions to enhance the quality of new acquisitions and elevate the overall customer experience in Airbnb listings.

EDA:

To understand some important insights we have explored the following questions:

1. Customer preference for neighbourhood & room type
2. Property demand based on minimum nights offered
3. Price range preferred by customers
4. Understanding Price variation w.r.t Room Type & Neighbourhood
5. Understanding Price variation w.r.t Geography
6. Top reviewed properties

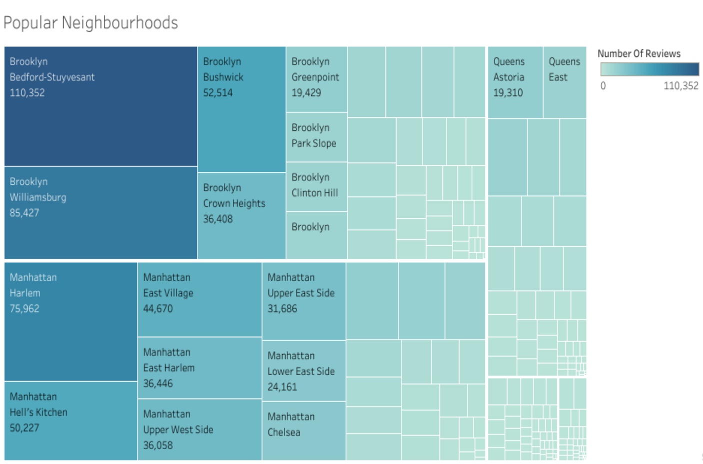
Methodology:

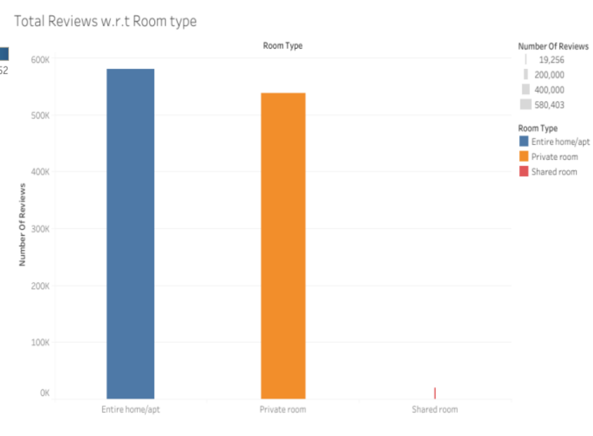
- The analysis and visualizations were done using Tableau considering various parameters. The analysis was done keeping in mind the business side of the project.
- The first half of the presentation focused on customer preference. The second half compared various parameters of customer preference with respect to price.
- The following parameters were considered :
  1. Customer experience: Neighbourhood, Room type & minimum nights offered.
  2. Price variation: Volume of customer booking, Room type, Neighbourhood, Number of reviews & Geography.
- The first half of the presentation focused on customer preference.

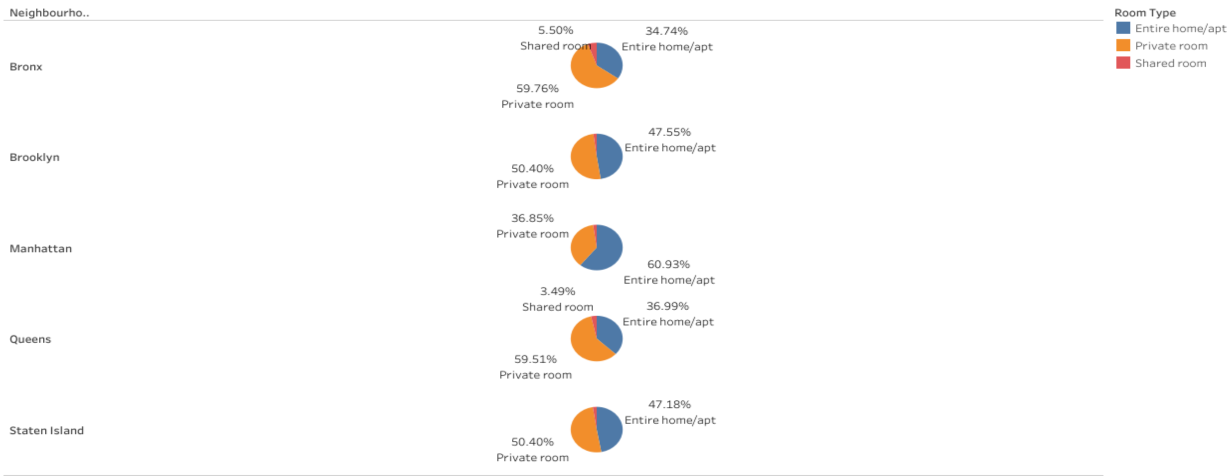
Explanation for EDA:

1. <u>Customer preference for neighbourhood & room type</u> :

We have explore the customer preference w.r.t volume and experience. The customer review parameter was chosen, as it is one of the most important factors to boost future bookings and listings. The number of reviews a customer gives for a particular listing directly implies the likability of the listing. The two different parameters were taken for comparison: neighbourhood & room type. The parameters taken for analysis are: Room type; Neighbourhood group, SUM (Number of reviews)







Inferences:

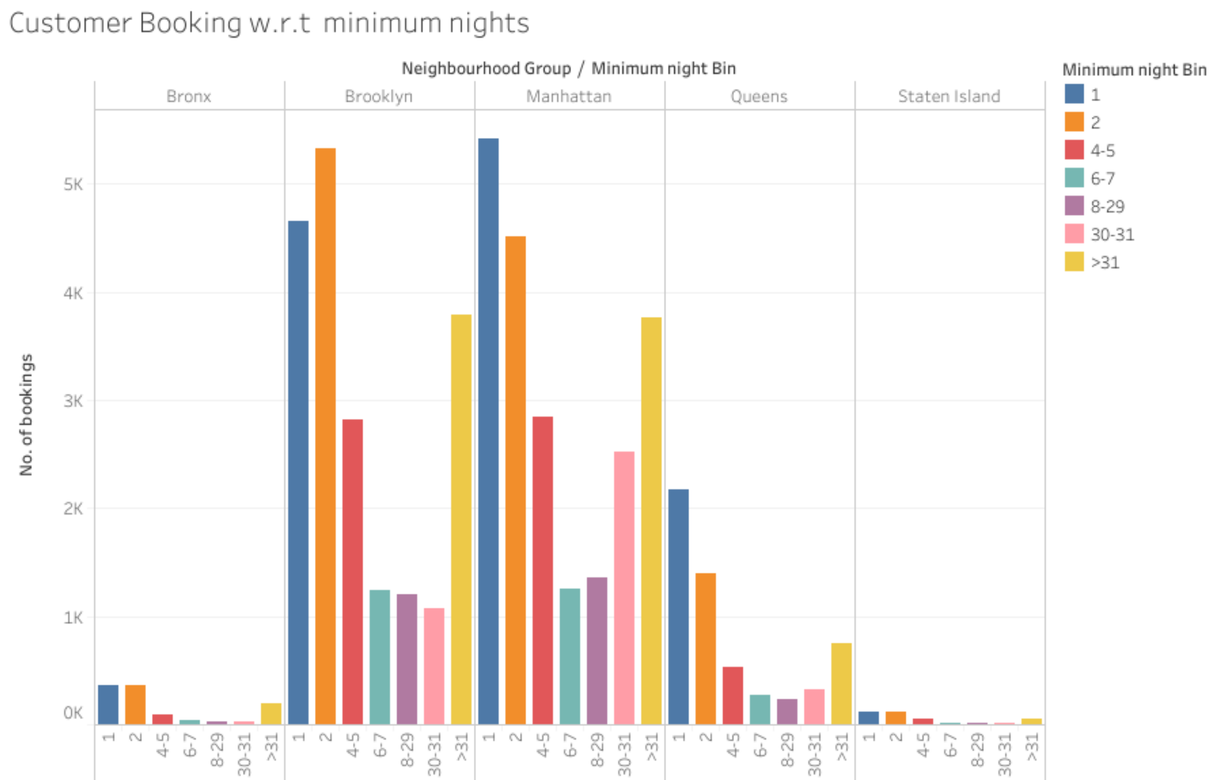
- Manhattan and Brooklyn stand out with the highest number of reviews in their listings, indicating a higher volume of bookings in these neighbourhoods. Means higher level of customer satisfaction in these areas.
- The analysis reveals that customers have a clear preference for private rooms or entire homes, as opposed to shared rooms.

Recommendation:

- Airbnb should promote shared rooms with targeted discounts to boost bookings.
- Consider acquiring private rooms in Manhattan and Brooklyn, and entire homes in Bronx and Queens to meet customer preferences and increase offerings.

2. <u>Property demand based on minimum nights offered:</u>

- 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 neighbourhood-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) & Neighbourhood Group.



Inference:

- Listings with 1-5 night minimum stays receive the highest number of bookings. A significant increase in 30-day bookings suggests a preference for monthly rentals. Additional spikes at 60 and 90 days further support the trend of longer-term stays driven by monthly rent considerations.

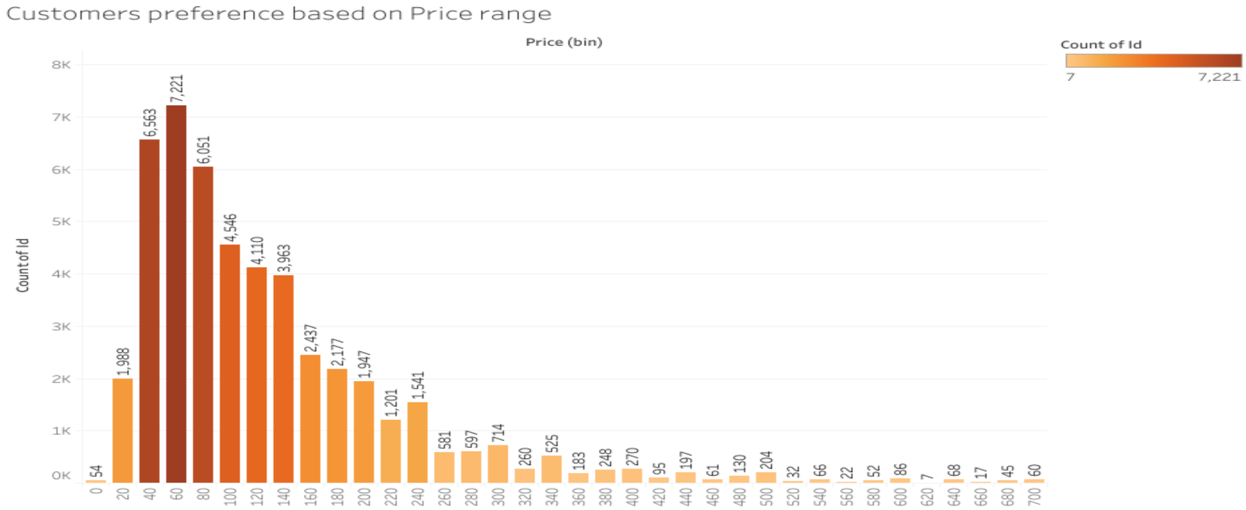
Recommendation:

- Expanding the inventory of hosts and listings offering monthly rentals (30-60-90 days) presents a significant opportunity. The popularity of 30-day bookings in Manhattan and Brooklyn indicates a potential target market in these areas.
- Furthermore, considering the demand for quarantine purposes, acquiring listings for weekly or bi-weekly rentals can cater to customers in need of temporary accommodations.

3. <u>Price range preferred by customers</u>:

For any business to operate it has to have a fair understanding of the customer-buying pattern. So we have tried to understand the most preferred price range for customers. Using this we can try to improve the listings in the price range preferred by the customer.

We have considered the volume of booking in a particular price range. For easy visualization, we have binned the Price with a bin size of 20. Also owing to the enormous value range, we have observed the variation until $700. As there was very little data beyond this, we decided to filter it.



4. **<u>Understanding Price variation w.r.t Room Type & Neighbourhood:</u>**

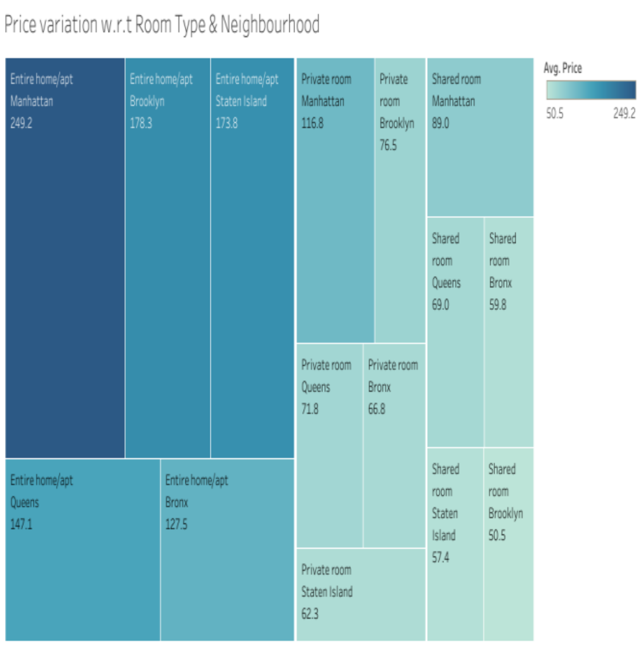
Now that we have obtained the optimum price range for listings, let us explore which neighbourhoods and room types fit in this category.

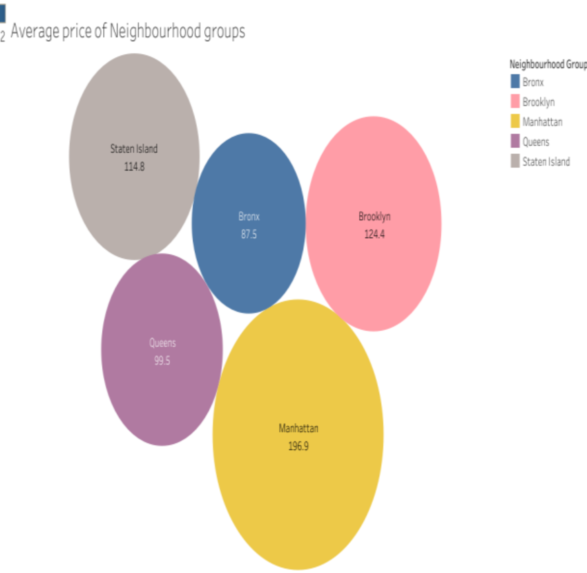
We have created two graphs to explore this question:

Tree map and Bubble chart:

We wanted to understand the average price distribution in the 5 boroughs of NYC. The tree map and Bubbles chart were created with Avg(Price) for ‘size’ and ‘color’. Highlight table and Bubbles.

As the comparison table in tree map containing the room type and neighborhood mainly consisted of numbers. we decided to go ahead with highlight table to display the highest and lowest values. Similarly in Bubble chart we decided to go ahead with highlight bubbles to display the highest and lowest values w.r.t 5 boroughs of NYC.





Inference:

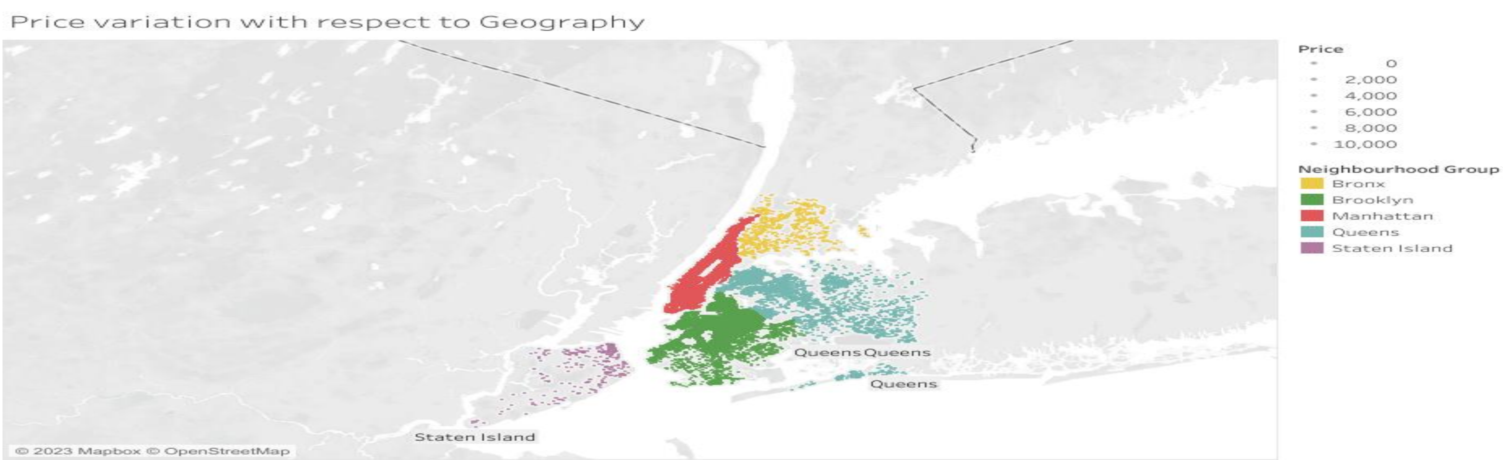
- Manhattan is the most expensive at $250, much higher than the overall average.
- Brooklyn Offers Affordable Shared Rooms.

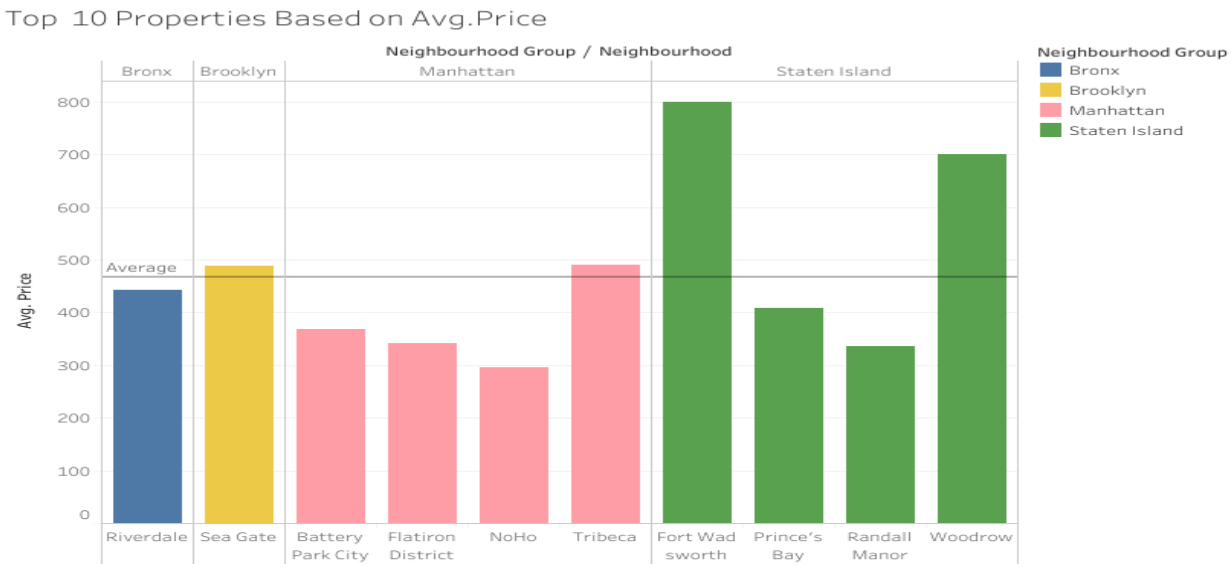
Recommendation:

- ‘Private rooms' in Manhattan and Brooklyn, as well as ‘Entire homes' in Bronx and Queens, fall within the favourable price range($40- $200).
- Additionally, explore Brooklyn for expansion, as it offers an average price of $124 and a less saturated market compared to Manhattan.

5. **Understanding Price variation w.r.t Geography:**

We had earlier explore the price variation with respect to location. We now deep dive to understand how it varies across difference areas/geographies. - We wanted to understand if the 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 finding; we took the top 10 neighbourhood with maximum average price. We used the findings in this to confirm our observation obtained from the geographical map.





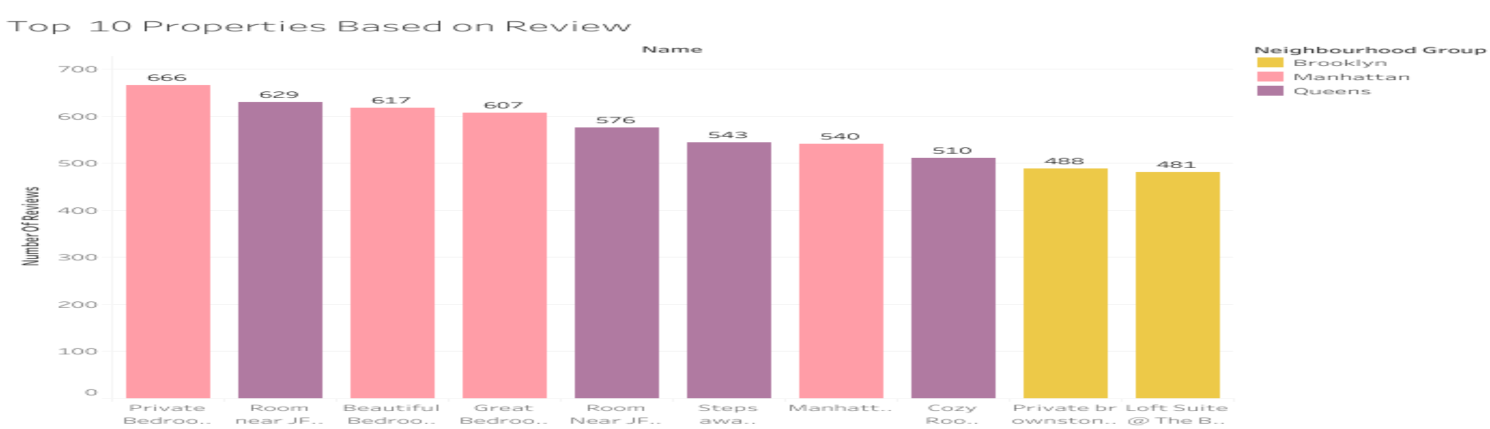
Inference:

- 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.
- The bar graph confirms our inference, as we observe that the top 10 neighbourhoods according to price are those that are situated near the sea or are next to important institutions/companies/landmarks.

Recommendation:

- Increasing acquisitions and new properties in coastal regions can increase customer bookings.

6. Top 10 Reviewed Properties:



Inferences:

- Among the boroughs of New York City, Manhattan, Brooklyn, and Queens stand out as properties with high ratings and reviews.

- Despite its steep price, the "Private Bedroom in Manhattan" has received the highest number of reviews, making it the most popular and favoured property in all of NYC.

Recommendations Consolidated:

- Promotion of shared rooms with targeted discounts to attract more bookings.
- Emphasize acquiring listings with a monthly rental duration (30-60-90 days). There is a potential market for 30-day rentals, especially in Manhattan and Brooklyn.
- Weekly or bi-weekly rentals can also be acquired, targeting customers who require temporary accommodation for quarantine purposes or extended stays in NYC.
- New acquisitions and expansion can be done in the price range of $40 - $200 to cater to a broader customer base and increase volume.
- Explore acquiring more ‘Private rooms’ in Manhattan and Brooklyn and ‘Entire homes’ in Bronx and Queens.
- Prioritize the expansion of property listings in Brooklyn due to its higher number of 30-day bookings and an average price of $124.
- Increase acquisitions and focus on new properties in coastal regions to attract customers seeking beachfront or waterfront accommodations, capitalizing on the appeal of these locations.