Storytelling Case Study: Airbnb, NYC

By : Anish Gautam
Sabita Rana
Sanket Badadal

Objective:

- Airbnb is an online platform using which people can rent their unused accommodations.
- During the covid time, Airbnb incurred a huge loss in revenue.
- To Conduct a thorough analysis of New York Airbnb Dataset.
- People have now started travelling again and Airbnb is aiming to bring up the business again and e ready to provide services to customers.
- Process, analyze and share findings by data visualization and statistical techniques

Data Preparation

- Cleaned data to remove any missing values and duplicates.
- Once data is cleaned, EDA is done and new features are created.
- Then Meaningful insights are derived using various analytical methods.

Importing libraries and reading the data

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```python
inp0 = pd.read_csv('AB_NYC_2019.csv')
inp0.head(5)
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights
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

Creating features

2.1 categorizing the "availability_365" column into 5 categories

```python
def availability_365_categories_function(row):
    """
    Categorizes the "minimum_nights" column into 5 categories
    """
    if row <= 1:
        return 'very Low'
    elif row <= 100:
        return 'Low'
    elif row <= 200 :
        return 'Medium'
    elif (row <= 300):
        return 'High'
    else:
        return 'very High'
```

2.2 categorizing the "minimum_nights" column into 5 categories

```python
def minimum_night_categories_function(row):
    """
    Categorizes the "minimum_nights" column into 5 categories
    """
    if row <= 1:
        return 'very Low'
    elif row <= 3:
        return 'Low'
    elif row <= 5 :
        return 'Medium'
    elif (row <= 7):
        return 'High'
    else:
        return 'very High'
```

2.3 categorizing the "number_of_reviews" column into 5 categories

```python
def number_of_reviews_categories_function(row):
    """
    Categorizes the "number_of_reviews" column into 5 categories
    """
    if row <= 1:
        return 'very Low'
    elif row <= 5:
        return 'Low'
    elif row <= 10 :
        return 'Medium'
    elif (row <= 30):
        return 'High'
    else:
        return 'very High'
```

Missing values

```
# Percentage of missing values
round((inp0.isnull().sum()/len(inp0))*100,2)
```

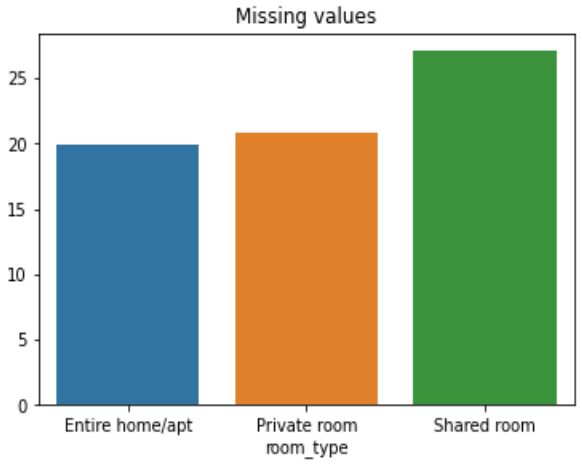
```
id                                  0.00
name                                0.03
host_id                             0.00
host_name                           0.04
neighbourhood_group                 0.00
neighbourhood                       0.00
latitude                            0.00
longitude                           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
availability_365_categories         0.00
minimum_night_categories            0.00
number_of_reviews_categories        0.00
price_categories                    0.00
dtype: float64
```

- Two columns (last_review , reviews_per_month) has around 20.56% missing values. name and host_name has 0.3% and 0.4 % missing values

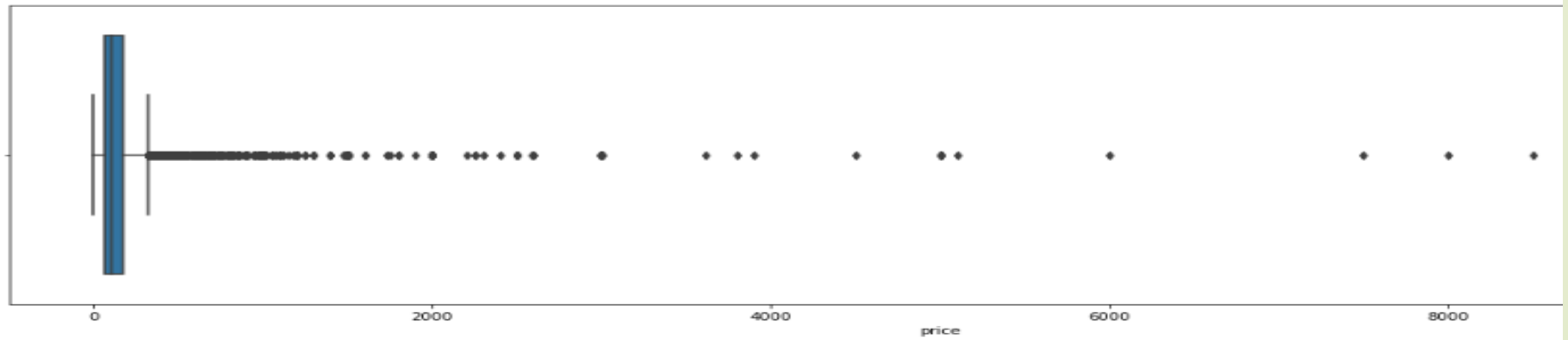
- We need to see if the values are, MCAR: It stands for Missing completely at random.

The reason behind the missing value is not dependent on any other features or if it is MNAR: It stands for Missing not at random. There is a specific reason behind the missing value.

- 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.



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



- 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)

Analysis

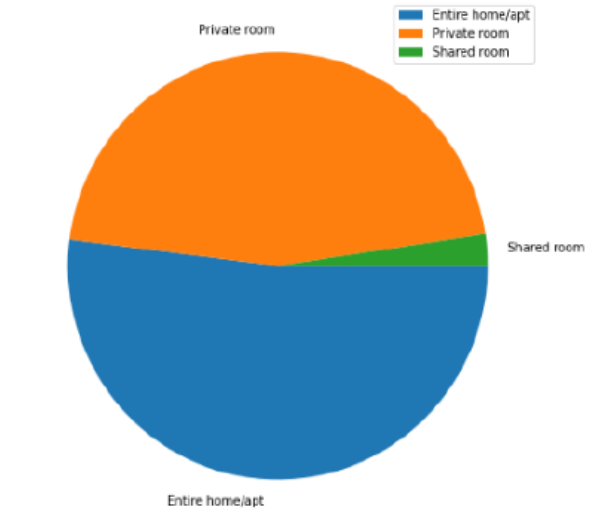
```
inp0.room_type.value_counts()
```

```
Entire home/apt    25409
Private room       22326
Shared room         1160
Name: room_type, dtype: int64
```

```
inp0.room_type.value_counts(normalize=True)*100
```

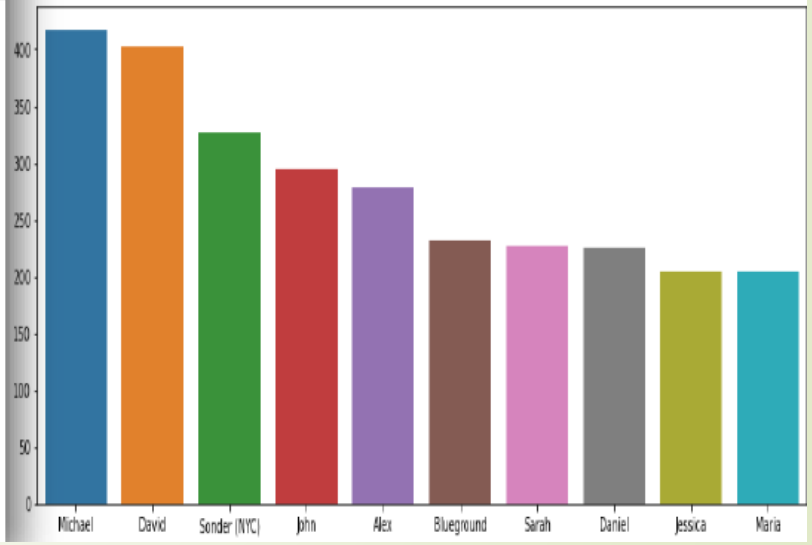
```
Entire home/apt    51.966459
Private room       45.661111
Shared room         2.372431
Name: room_type, dtype: float64
```

```
plt.figure(figsize=(8,8))
plt.pie(x = inp0.room_type.value_counts(normalize= True) * 100,labels = inp0.room_type.value_counts(normalize= True).
plt.legend()
plt.show()
```



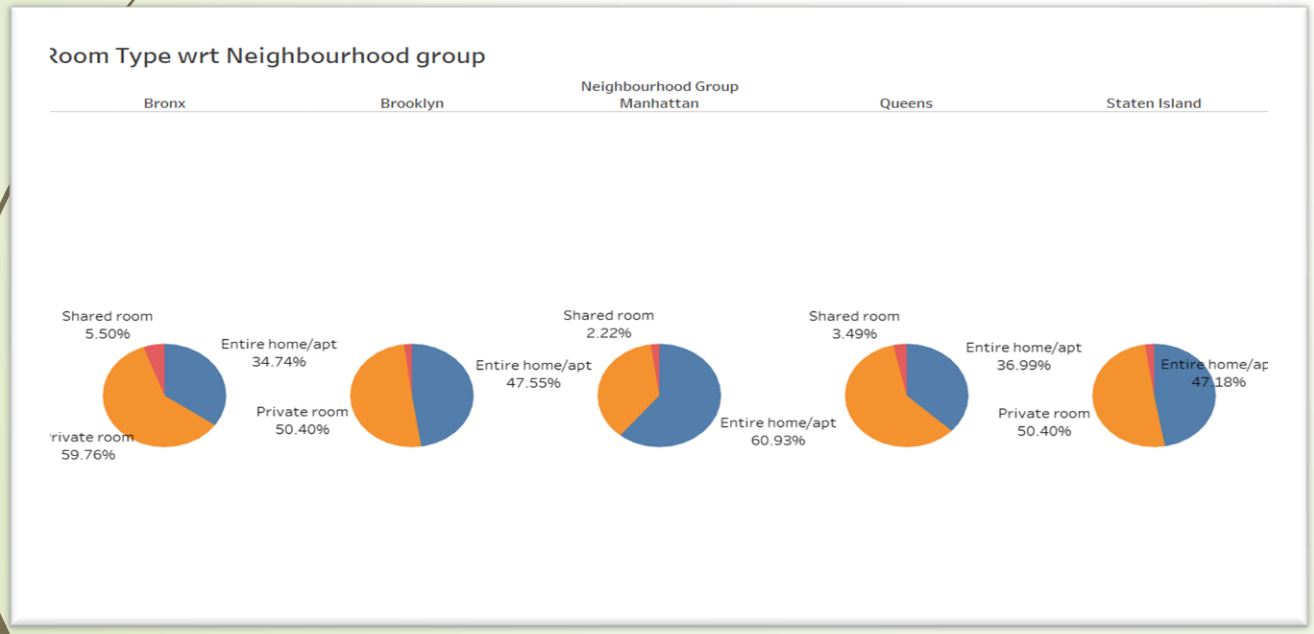

```
inp0.host_name.value_counts()
```

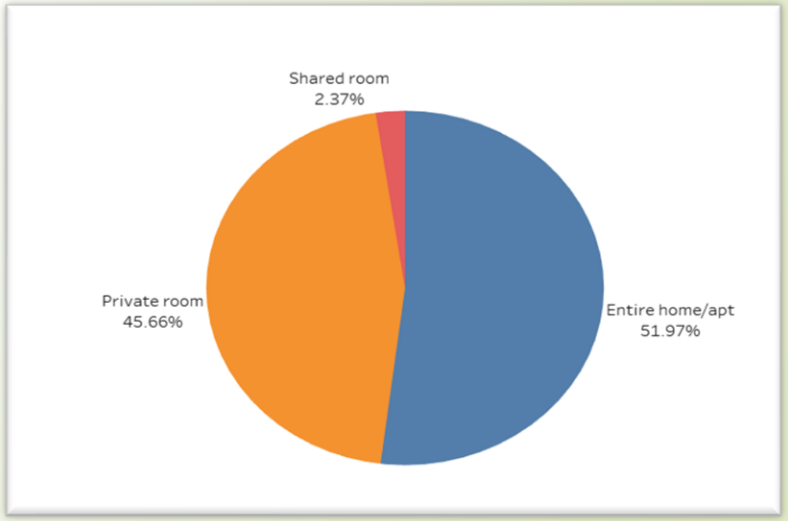
```
Michael              417
David                403
Sonder (NYC)         327
John                 294
Alex                 279
                    ...
Rhonycs                1
Brandy-Courtney        1
Shanthony              1
Aurore And Jamila      1
Ilgar & Aysel          1
Name: host_name, Length: 11452, dtype: int64
```

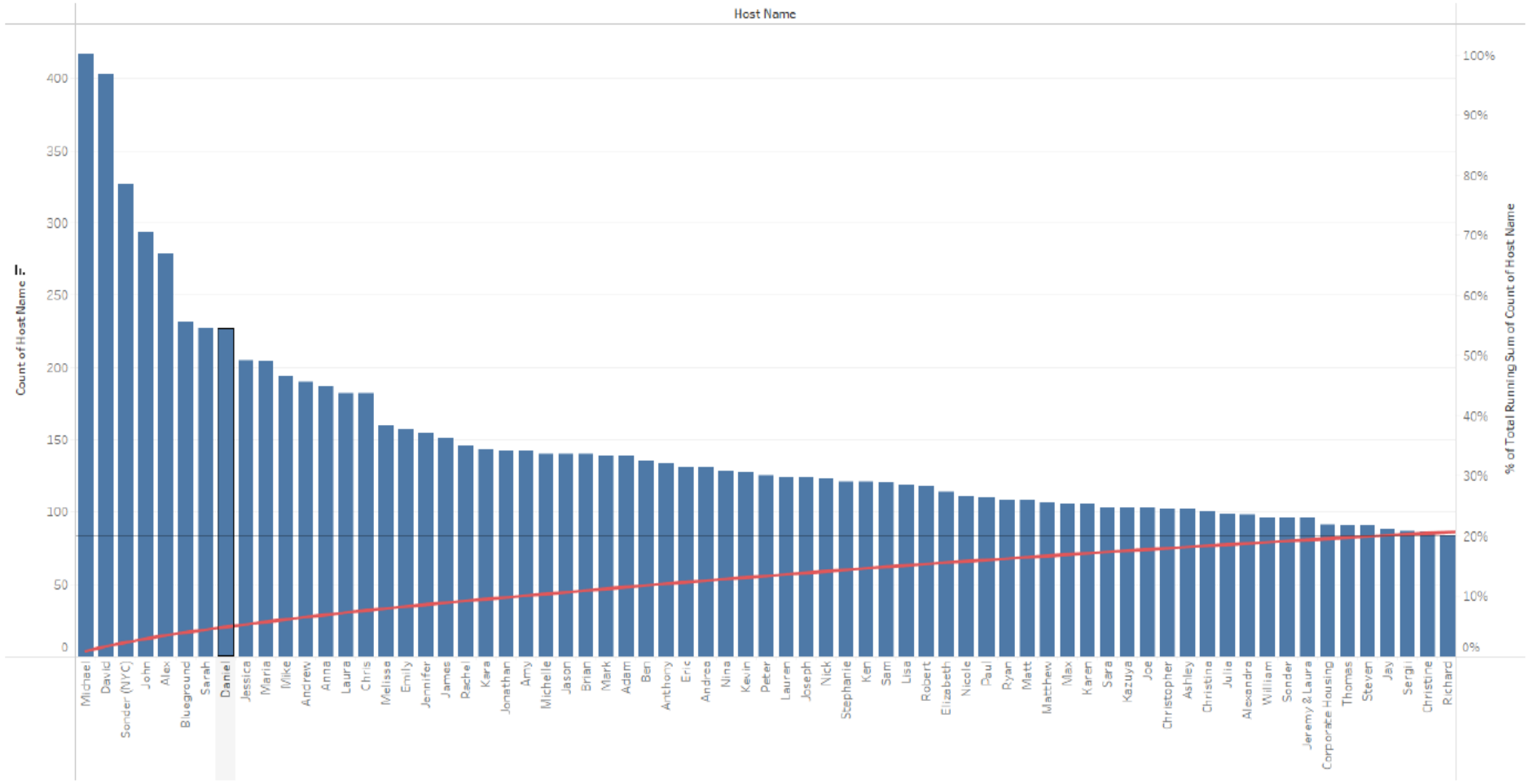


Room type with respect to Neighbourhood group

- There are three types of rooms - Entire home/Apartment, Private room & shared room.
- Overall, customers appear to prefer private rooms (45%) or entire homes (52%) in comparison to shared rooms (2.4%).
- Airbnb can concentrate on promoting shared rooms with discounts to increase bookings and also acquire more private listings.
- Queens & Bronx contribute 60% each to private rooms, more than the combined ratio of 45% Whereas, Manhattan has a higher contribution in entire home (61%), compared to the combined ratio of 52%.



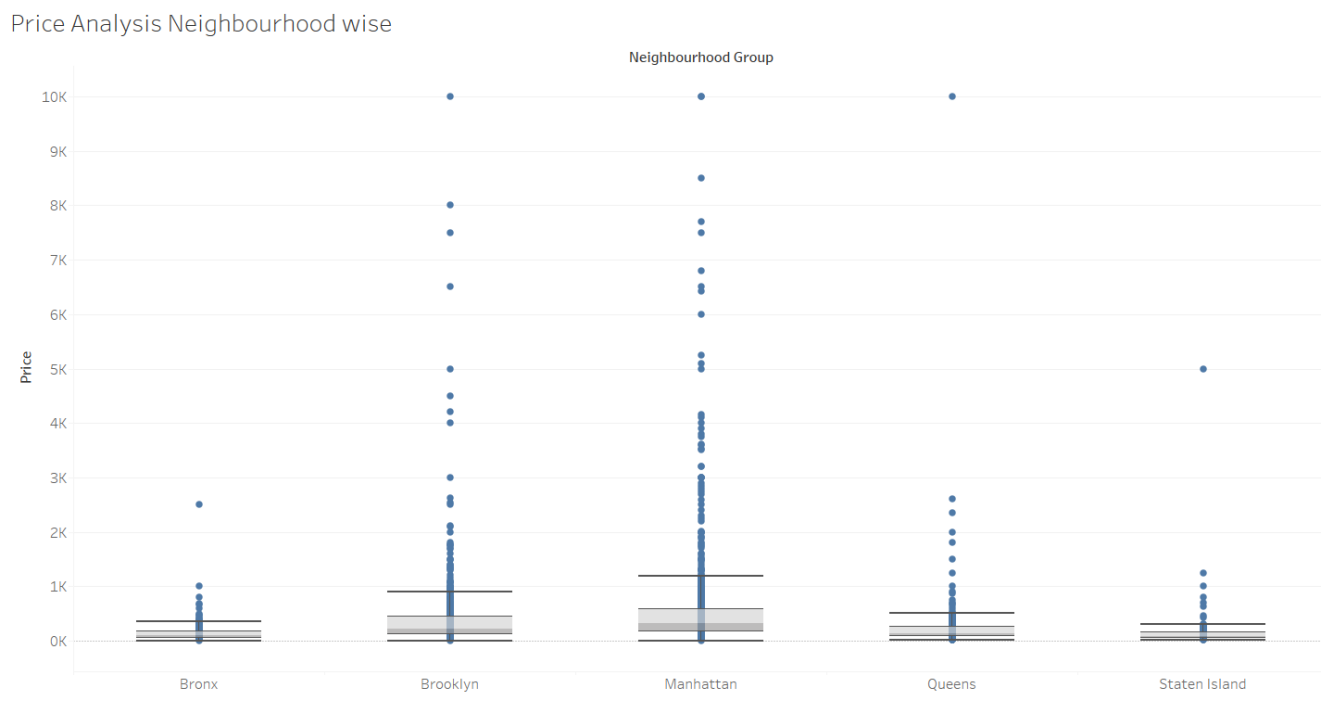




- The top 60 hosts only make up 20% of the total host count!

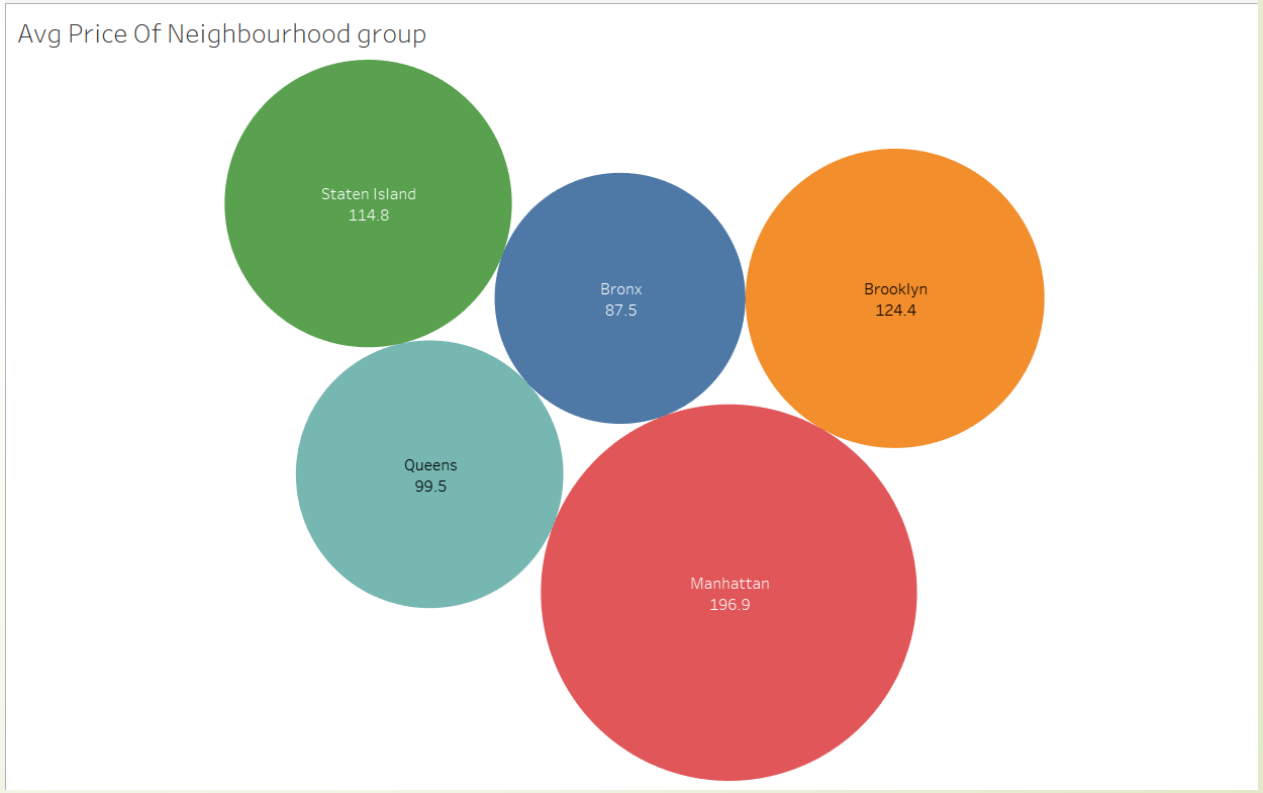
Price Analysis Neighbourhood wise

- Most of the outliers in Price column are for Brooklyn and Manhattan.
- Also, Manhattan has the highest range of prices for the listings.
- Bronx is the cheapest of them all.
- We can see the median price of all neighbourhood groups lying between $ 80 to $ 300.
- Price was highly positively skewed so median was very close the lower quartile with some outliers as seen in the boxplot below.



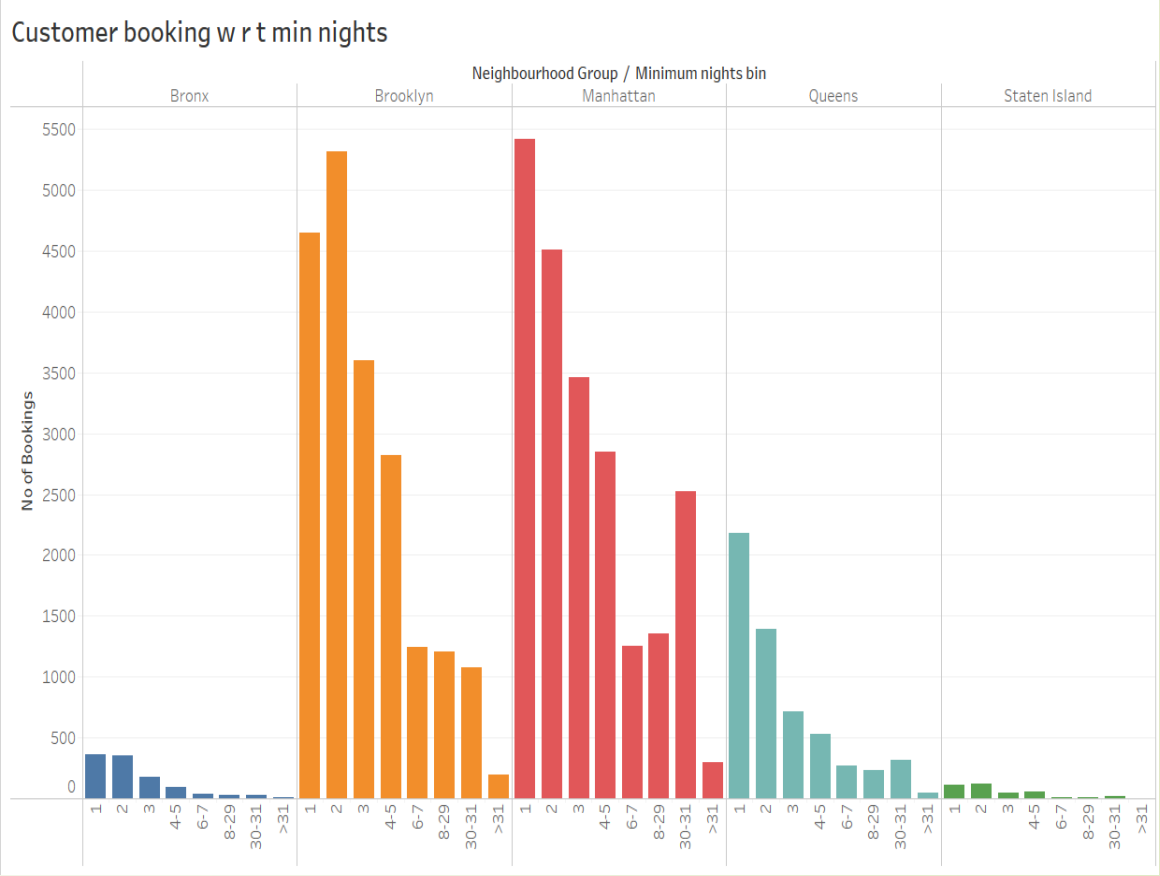
Average price of Neighbourhood groups

- The average price of listed properties in Manhattan is around 196.9, which is highest among all neighbourhoods.
- Average price for Brooklyn is second highest i.e. 124.4.
- Bronx appears to be an affordable neighbourhood as the average price is almost half than Manhattan's average price.



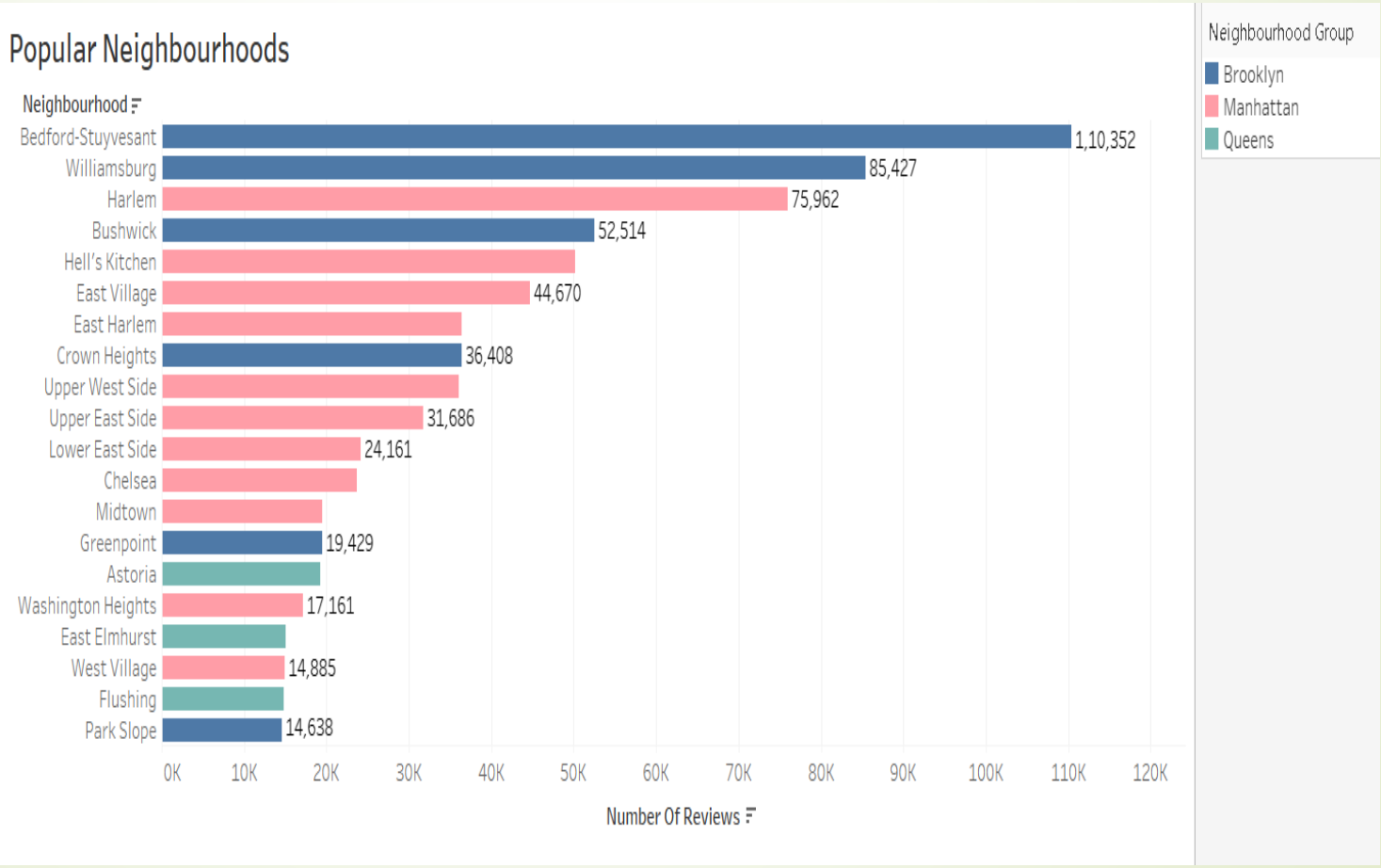
Customer Booking with respect to minimum nights

- The listings with Minimum nights 1-5 have the most 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, this can also be explained by the monthly rent taking trend.
- Manhattan &Queens have higher number of 30 day bookings compared to the others. The reason could be either tourists booking long stays or mid-level employees who opt for budget bookings due company visits



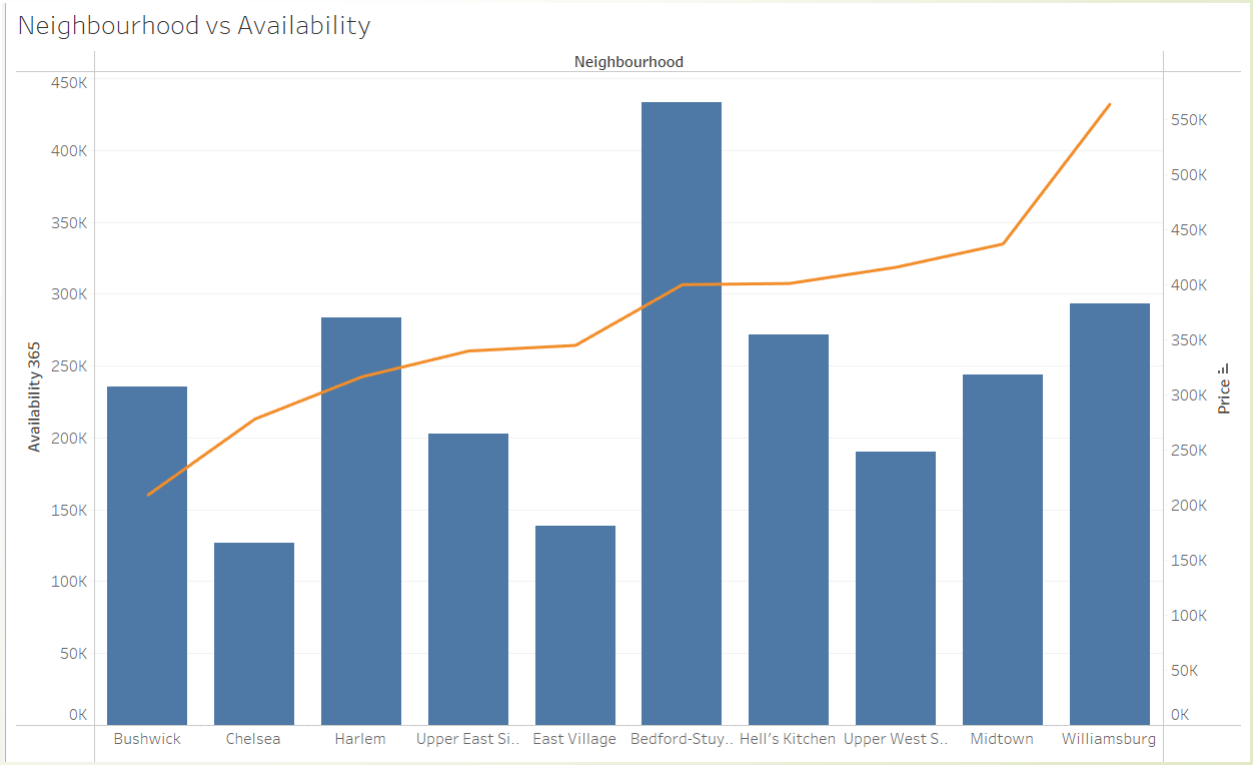
Popular Neighborhoods

- We see that Bedford-Stuyvesant from Brooklyn is the highest popular with 1,10,352 no of reviews in total followed by Williamsburg.
- Harlem from Manhattan got the highest no of reviews followed by Hell's kitchen.
- The higher number of customer reviews imply higher satisfaction in these localities.



Neighbourhood vs Availability

- Availability of Bedford is highest and its price is on the lower side. It is a good choice for customers.
- After Bedford, Harlem follows the same trend.
- Chelsea's availability low but it is costly.
- On the other hand, William's price is high and has average availability.



Thank You