Storytelling Case Study: Airbnb, NYC

By : Sanchit Kumar Behera, Potnuru Koushik, Nikitha R.

Objective:

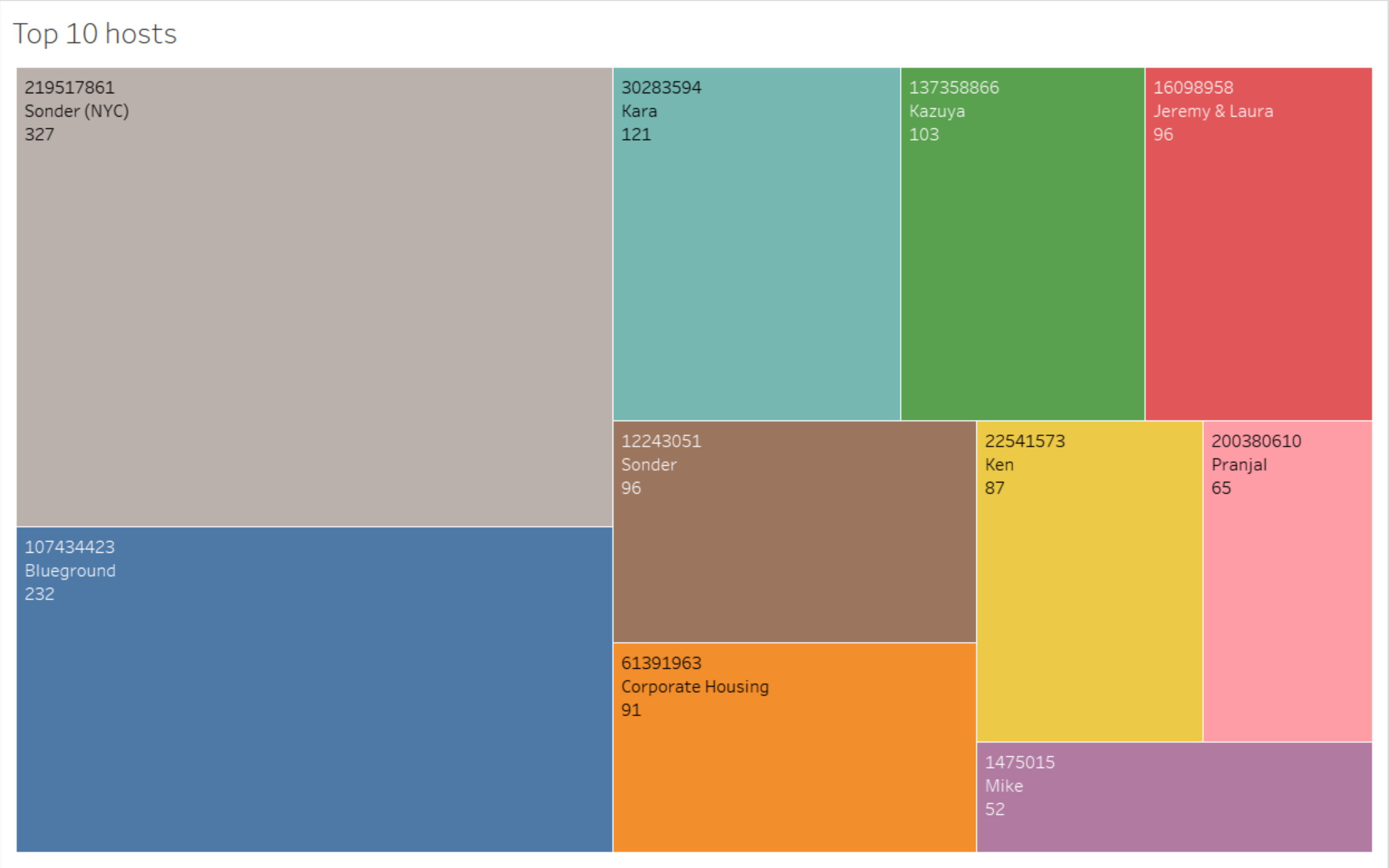
- Improve business strategies and estimate customer preferences to revive the business in the post-COVID period.
- Understand the critical pre-COVID period insights from the Airbnb NYC business.
- Make recommendations to various departments on how to prepare for post-pandemic changes.

Background

- Airbnb’s revenue has been significantly reduced in recent months as a result of COVID-19.
- People have begun to travel more now that the restrictions are lifted.
- Airbnb wants to make sure that it is fully prepared for this range.

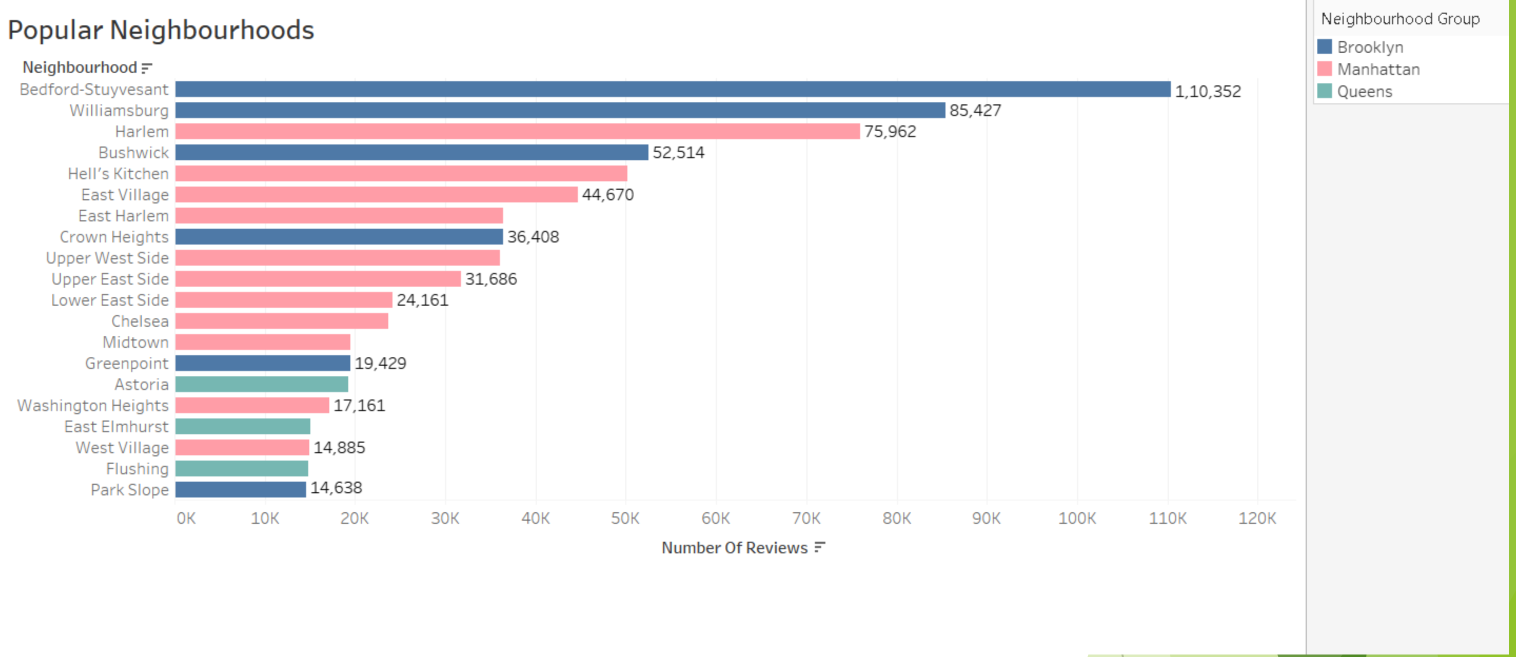
Top 10 Hosts

- The host with the ID 219517861, named Sonder, holds the record for the highest number of bookings, with a total of 327.
- Following Sonder, the second most popular host is Blue Ground.
- In addition to Sonder and Blue Ground, other hosts such as Kara, Ken, Pranjal, Jeremy, and Mike also rank among the top 10 hosts.



Popular Neighborhoods w.r.t Number of reviews

- Bedford-Stuyvesant, located in Brooklyn, holds the highest popularity with a total of 110,352 reviews, making it the most reviewed neighborhood. It is closely followed by Williamsburg.
- Among the neighborhoods in Manhattan, Harlem has received the highest number of reviews, indicating its popularity among customers. Hell's Kitchen follows closely behind in terms of review count.
- The larger number of customer reviews in these localities suggests a higher level of satisfaction among visitors and guests.

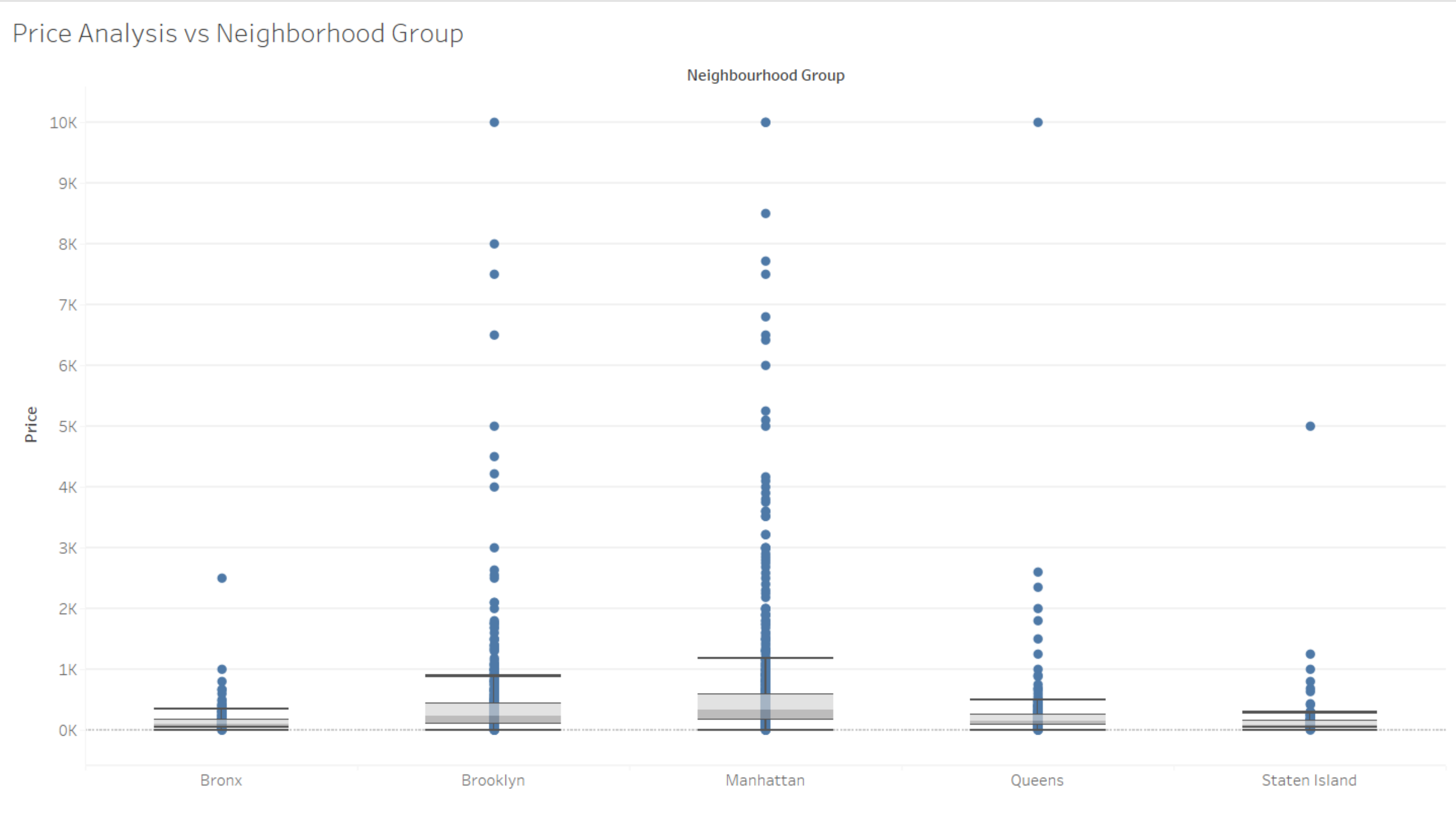


Neighbourhood vs Availability

- Bedford is highly available and offers affordable prices, making it a favorable choice for customers.
- Similar to Bedford, Harlem also exhibits high availability and relatively lower prices, making it another good option for customers.
- Chelsea, on the other hand, has limited availability but comes with higher costs.
- In contrast, Williams has higher prices and average availability, indicating it may be more suitable for customers who prioritize budget over immediate availability.

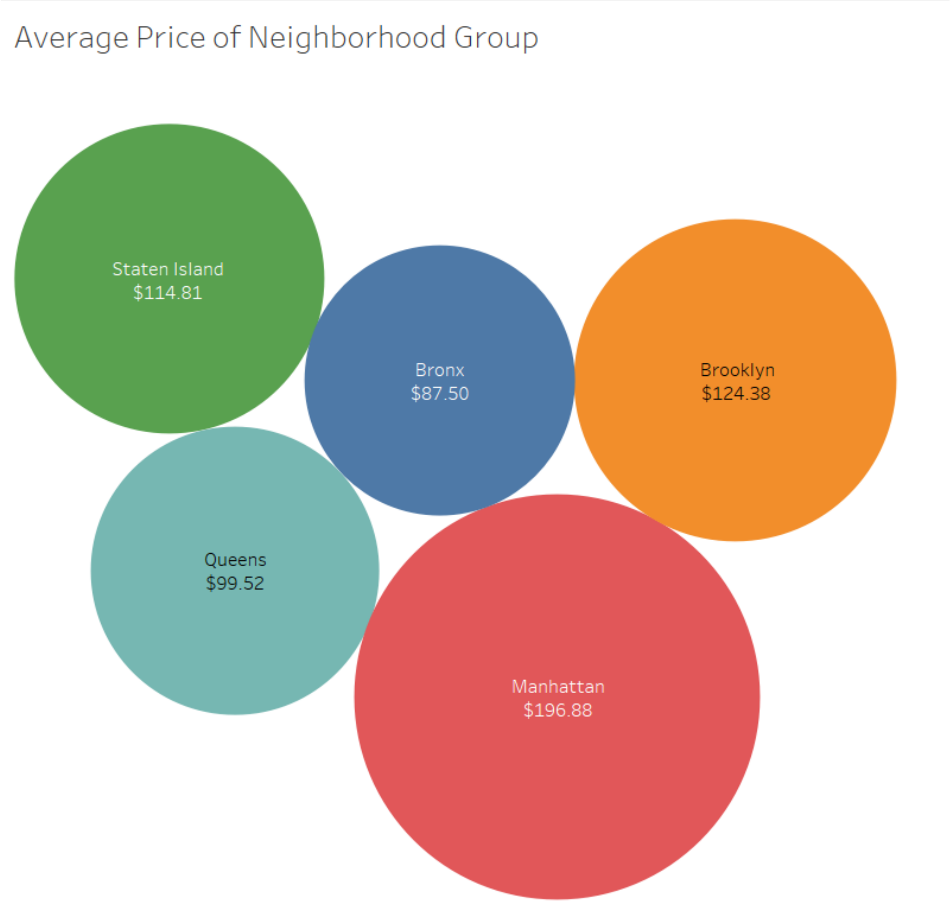
Price Analysis Neighbourhood wise

- Most of the outliers in Price column are for Brooklyn and Manhattan.
- 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 neighborhood groups lying between $80 to $300.



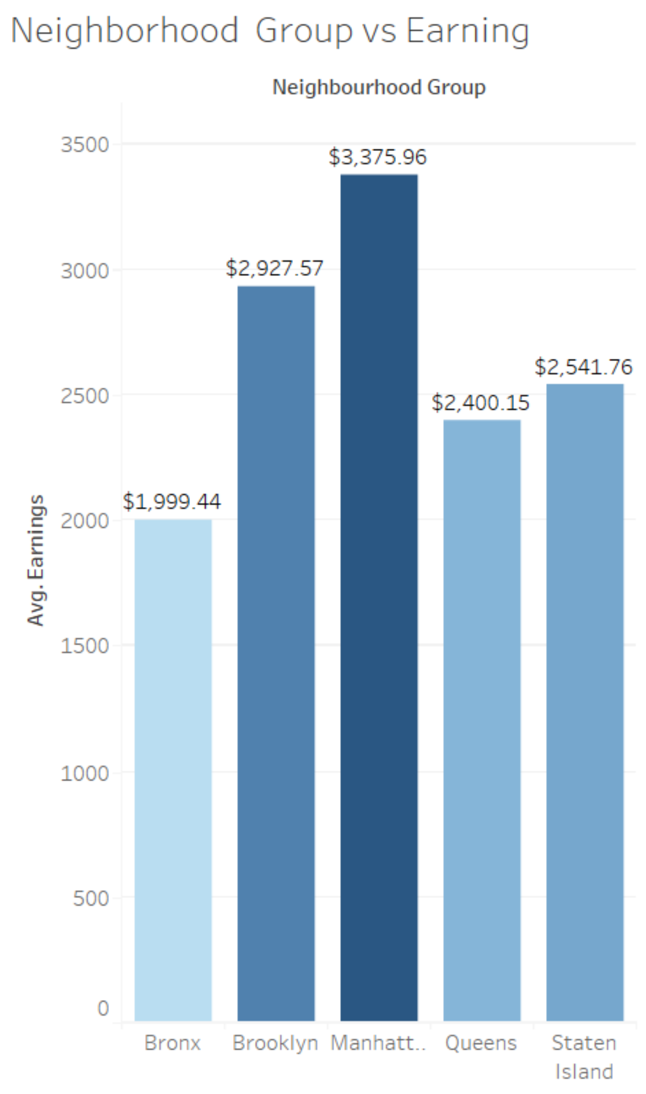
Average Price of Neighborhood Group

- 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 seems to be an affordable neighbourhood as compared to others as the average price is <50% than Manhattan's average price.



Neighborhood Group and Earning

- Manhattan has the highest average earning at $3.3k
- While Bronx is at the lower end at approximately $2k.
- The top 3 revenue generating areas are Manhattan, Brooklyn and Staten Island.



Conclusion

- The host with the ID 219517861, named Sonder, holds the record for the highest number of bookings, with a total of 327.
- Bedford is highly available and offers affordable prices, making it a favorable choice for customers.
- Bronx seems to be an affordable neighbourhood as compared to others as the average price is <50% than Manhattan’s average price.
- The top 3 revenue generating areas are Manhattan, Brooklyn and Staten Island.

Appendix : Methodology

AIRBNB Case Study

Methodology Document PPT 1:

In the case study we have used Jupiter notebook to perform initial analysis of the data and Tableau for data analysis and visualization.

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

Number of Rows: 48895

Number of Columns: 16

```
# 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
```

```
# Data conversion and Understanding
airbnb = pd.read_csv("AB_NYC_2019.csv")
airbnb.head(5)
```

	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	

```
# Check the rows and columns of the dataset
airbnb.shape
```

```
(48895, 16)
```

- The dataset contains 48895 rows and 16 columns
- Now we have to check whether there are any missing values in the dataset

```
# Calculating the missing values in the dataset
airbnb.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
```

```
# Now we have the missing values, there are certain columns that are not efficient to the dataset
airbnb.drop(['id','name','last_review'], axis = 1, inplace = True)
```

```
# View whether the columns are dropped
airbnb.head(5)
```

	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	

Methodology- Contd.

Step 2: Data Wrangling:

- Checked the Duplicate rows in our dataset and no duplicate data was found.
- Checked the Null Values in our dataset. Columns like name, host-name, last review and review-per-month have null values.
- We've dropped the column name as missing values are less and dropping it won't have significant impact on analysis.
- Checked the formatting in our dataset.
- Identified and review outliers.

Data Analysis and Visualizations using Tableau:

We have used tableau to visualize the data for the assignment. Below are the detailed steps used for each visualization.

1) **Top 10 Host:**
   - We identified the top 10 Host Ids, Host Name with count of Host Ids using the tree map.

2) **Preferred Room type with respect to Neighbourhood group:**
   - We created a pie chart for understanding the percentage of room type preferred w r t neighbourhood group
   - We added Room Type to the colours Marks card to highlight the different Room Type in different colours and count of Host Id to the size.

3) **For Variance of price with Neighbourhood Groups:**
- We used a box and whisker's plot with Neighbourhood Groups in Columns and Price in Rows.
- We changed the Price from a Sum Measure to the median measure.

4) **Average price of Neighbourhood groups:**
   • We created a bubble chart with Neighbourhood Groups in Columns and Price column in Rows.
   •We added the Neighbourhood Groups to the colors Marks card to highlight the different neighbourhood Groups in different colors. Also Put Avg price in Label.

5) **Customer Booking w r t minimum nights:**

- The bins were used to display the distribution of minimum nights based on the number of ids booked for each neighbourhood group.

6) **Popular Neighborhoods:**
- We took neighbourhood in rows and sum of reviews in column and took neighbourhood groups in colour.
- We used filter to show Top 20 neighbours as per the sum of reviews.

7) **Neighbourhood vs Availability:**
- We created a dual axis chart using bar chart for availability 365 and line chart for price for top 10 neighbourhood group sorted by price.

8) **Price range preferred by Customers:**
- We have taken pricing preference based on volume of bookings done in a price range and no of Ids to create a bar chart. We have created bin for Price column with interval of $20.

9) **Understanding Price variation w.r.t Room Type & Neighbourhood:**
- We created Highlights Table chat by taking Room Type in rows & Neighbourhood Group in column.
- We took the average price in colour Marks card to highlight the different Room Type in different colours.

10) **Price variation w r t Geography:**
- We used Geo location chart to plot neighbourhood, neighbourhood Group in map to show case the variation of prices across.

11) **Popular Neighborhoods:**
- We took neighborhood in rows and sum of reviews in column and took neighborhood groups in color.
- We used filter to show Top 20 neighbors as per the sum of reviews.

12) **Tools used:**
- Data cleaning and preparation: Jupyter notebook – Python
- Visualization and analysis: Tableau
- Data Storytelling: Microsoft PPT