# Capstone Project-1

## Airbnb Booking Analysis

by-

Kartik Kumar

(Self)

# POINT OF DISCUSSIONS

- About the Dataset
- Problem Statements
- Feature description
- Data Exploration
- Data Cleaning
- Hosts and neighbourhood groups
- Price Distribution across neighbourhoods
- Popular neighbourhood by reviews
- Preferred Room type
- Conclusion

# About the dataset-Airbnb NYC 2019

- Airbnb, Inc is an American company that operates as online marketplace for lodging , primarily home stays for vacation rentals, and tourism activities.
- Based in San Francisco , California , the platform is accessible via website and mobile app.
- The dataset that we would be analyzing consists the booking information on Airbnb from 2008 till 2019

# PROBLEM STATEMENTS:

With the help of explanatory data analysis techniques,'I will try to answer the following problem statements:

- What can we learn about different hosts and areas?
- What can we learn from predictions (prices,reviews,etc.)
- Which hosts are busiest and why?
- Which room type is preferred in most popular neighbourhood?

# FEATURE DESCRIPTION:

- The feature in the dataset can be described as follows:

1. id - This is the identity number of property listed by a particular host.
2. name - It stands for the name of the property listed by the host.
3. host_id - It is the identity number of the hosts who have registered on Airbnb website.
4. host_name – These are the names of the hosts who have listed their properties.
5. neighbourhood_group – These are the names of the neighbourhood groups present in NYC.
6. neighbourhood - These are the names of the neighbourhood present in the neighbourhood groups in NYC.
7. latitude - These represent the coordinates of latitude of the property listed.
8. longitude – These represent the coordinates of longitude of the property listed.
9. room_type - This represent the various types of room listed by host.

10. price - This is the rent of the property listed in USD.
11. minimum_nights - This represents the minimum number of nights customers rented the property.
12. number_of_reviews - This represents the number of customers reviewed the property.
13. last_review - This represent the date when the property was last reviewed.
14. review_per_month - It is the count of reviews per month which the property received.
15. calculated_host_listings_count - It is the number of listings done by a particular host.
16. availabilty_365 - This represent the number of days the property is available among 365 days

# DATA EXPLORATION

## Checking The Head of The Dataset

```
[6]  df = pd.read_csv('/content/drive/MyDrive/AirBnB Bookings Analysis/Airbnb NYC 2019.csv')
```

```
df.head(2)
```

| | id | name | host_id | host_name | neighbourhood_group | neighbourhood | latitude | longitude | room_type | price | minimum_nights | number_of_reviews | calculated_hos |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2539 | Clean & quiet apt home by the park | 2787 | John | Brooklyn | Kensington | 40.64749 | -73.97237 | Private room | 149 | 1 | 9 | |
| **1** | 2595 | Skylit Midtown Castle | 2845 | Jennifer | Manhattan | Midtown | 40.75362 | -73.98377 | Entire home/apt | 225 | 1 | 45 | |

- The dataset consist of 48895 observations (rows) and 16 features (columns)

```
[22] # Checking the shape of the datset
     print(f'The shape of the Airbnb Dataset is {df.shape}')
```

```
The shape of the Airbnb Dataset is (48895, 16)
```

- Checking out the 16 features:

```
# checking the list of features names
print(f'The names of features present in the dataset are: ')
list(df.columns)
```

```
The names of features present in the dataset are:
['id',
 'name',
 'host_id',
 'host_name',
 'neighbourhood_group',
 'neighbourhood',
 'latitude',
 'longitude',
 'room_type',
 'price',
 'minimum_nights',
 'number_of_reviews',
 'last_review',
 'reviews_per_month',
 'calculated_host_listings_count',
 'availability_365']
```

- Checking for the categorical and non-categorical columns in the dataset:

```
[27] #Checking the categorical columns
     cat_cols = df.select_dtypes(include = 'object').columns
     print(f'The following are the categorical features in the dataset {(list(cat_cols))}')
```

```
The following are the categorical features in the dataset ['name', 'host_name', 'neighbourhood_group', 'neighbourhood', 'room_type', 'last_review']
```

```
# Checking the numeric/non categorical columns
num_cols = df.select_dtypes(exclude = 'object').columns
print('The following are the non categorical features in the dataset:')
list(num_cols)
```

```
The following are the non categorical features in the dataset:
['id',
 'host_id',
 'latitude',
 'longitude',
 'price',
 'minimum_nights',
 'number_of_reviews',
 'reviews_per_month',
 'calculated_host_listings_count',
 'availability_365']
```

- Checking for Null values:

The columns like number_of_reviews and reviews_per_month have largest number of null values and these two columns are not necessary for my data exploration and analysis.

The columns like name and host_name contains fewer number of null values

```
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
```

# DATA CLEANING

Fixing the null values:
I have dropped the unnecessary columns like number_of_reviews and reviews_per_month.

Null values after the cleaning the data are as shown

```
[ ]  # our first motto is to deal with missing values
     #last_review-10052
     #reviews_per_month-10052
```

```
[33] df.drop('reviews_per_month',axis=1,inplace=True)
     df.drop('last_review',axis=1,inplace=True)
```

```
[35] # Filling missing values
     df['name'].fillna('Absent', inplace = True)
     df['host_name'].fillna('Absent', inplace =  True)
```

```
df.isnull().sum()
```

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

# CREATING A NEW DATA FRAME

I have created a new dataframe consisting of only columns that I think are required for my analysis

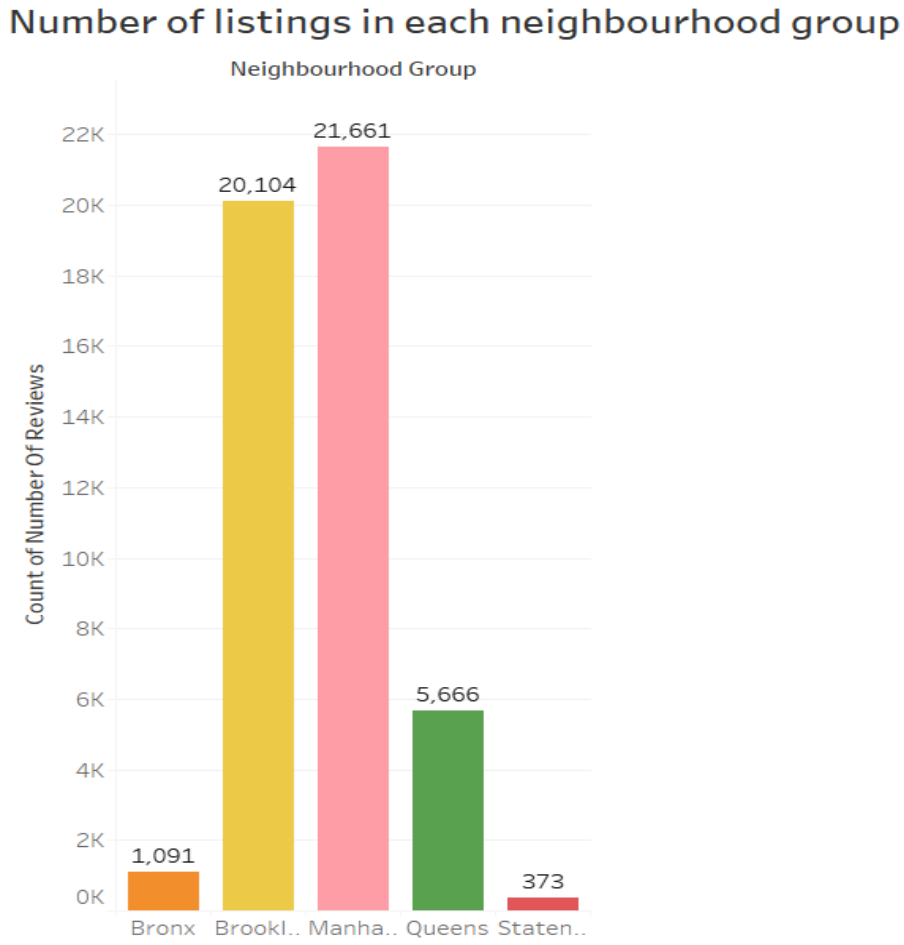
```
[7] new_df =df[['id','name','host_id','host_name','neighbourhood_group','neighbourhood','room_type','price','minimum_nights','number_of_reviews',
```

```
[ ] new_df
```

| | id | name | host_id | host_name | neighbourhood_group | neighbourhood | room_type | price | minimum_nights | number_of_reviews |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2539 | Clean & quiet apt home by the park | 2787 | John | Brooklyn | Kensington | Private room | 149 | 1 | 9 |
| **1** | 2595 | Skylit Midtown Castle | 2845 | Jennifer | Manhattan | Midtown | Entire home/apt | 225 | 1 | 45 |
| **2** | 3647 | THE VILLAGE OF HARLEM....NEW YORK ! | 4632 | Elisabeth | Manhattan | Harlem | Private room | 150 | 3 | 0 |
| **3** | 3831 | Cozy Entire Floor of Brownstone | 4869 | LisaRoxanne | Brooklyn | Clinton Hill | Entire home/apt | 89 | 1 | 270 |
| **4** | 5022 | Entire Apt: Spacious Studio/Loft by central park | 7192 | Laura | Manhattan | East Harlem | Entire home/apt | 80 | 10 | 9 |

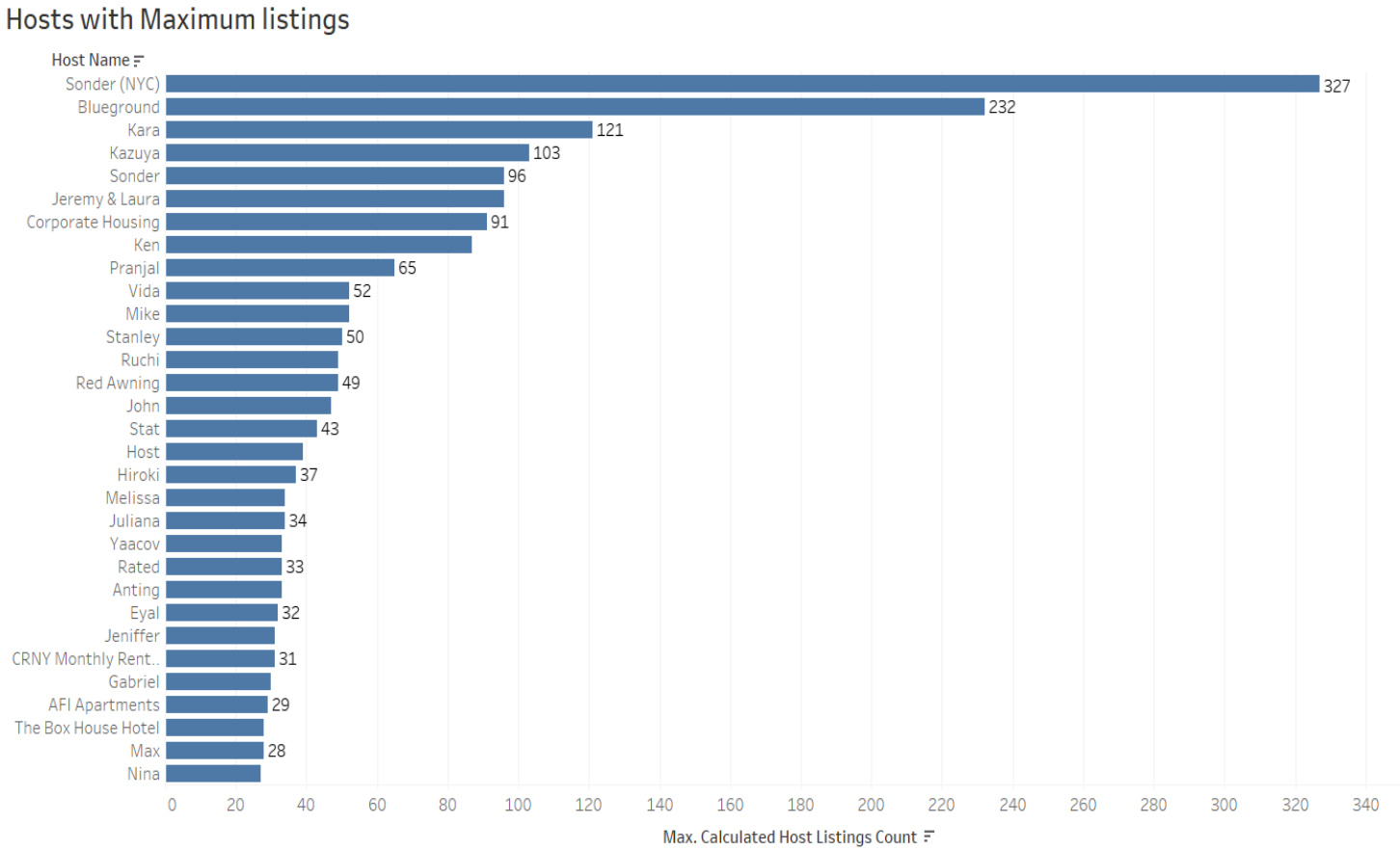
# Number of listings across neighbourhood groups

- It is observed that the Manhattan has highest number of listings of 21,661 which is 44.3% of total listings done on Airbnb.
- Brooklyn has second highest number of listings of 20,104 which is 41.13% of the total listings.
- Queen comes in third place with 5,666 listings whereas Bronx and Staten Island have least number of listings.



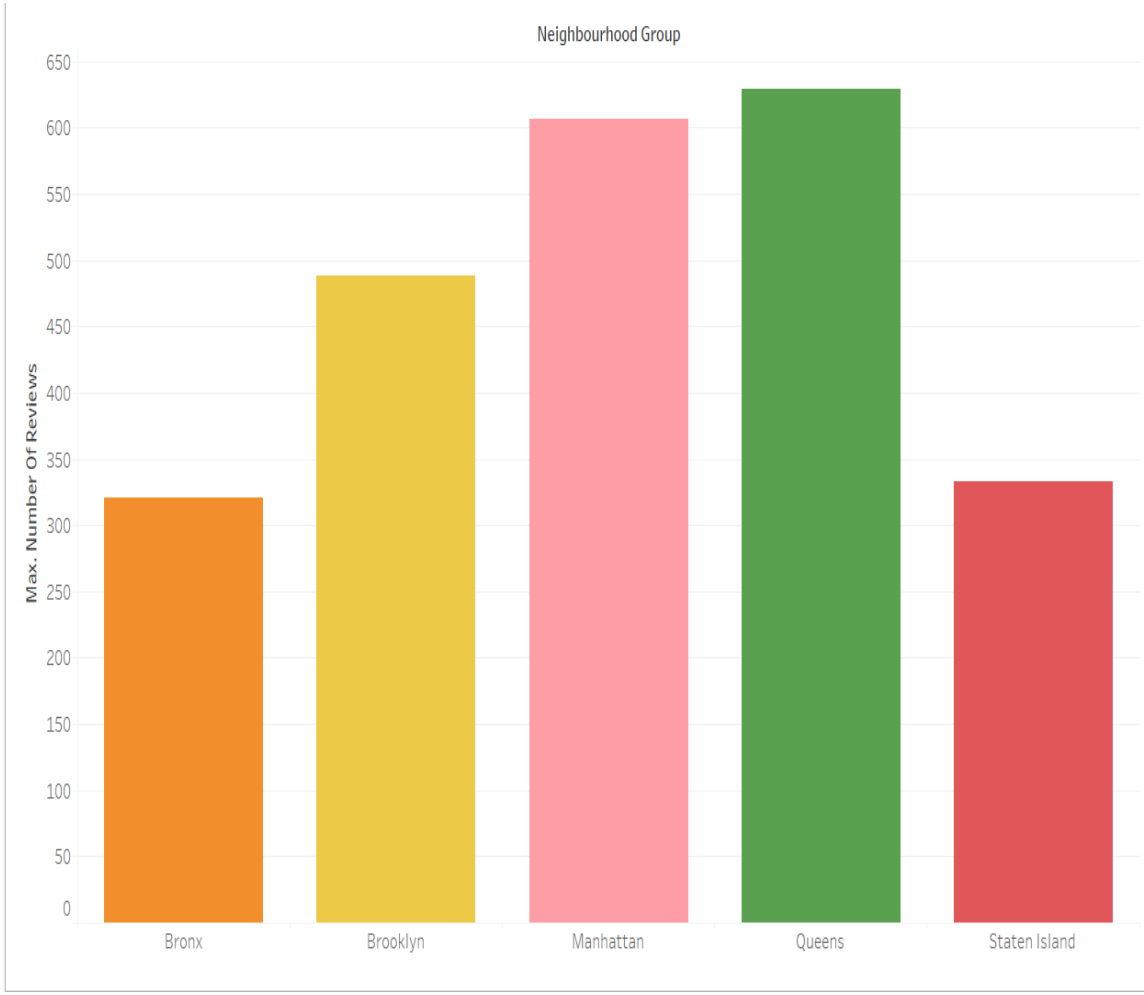
# Hosts with maximum listings

- As shown in the adjacent bar chart,we can see there is a good distribution among the top 6 hosts.
- The host named Sonder(NYC) has highest number of listings of 327 in Manhattan neighbourhood group.
- The host named Blueground has second highest listings of 232 in Manhattan neighbourhood group.
- The host Blueground also has 232 listings in Brooklyn



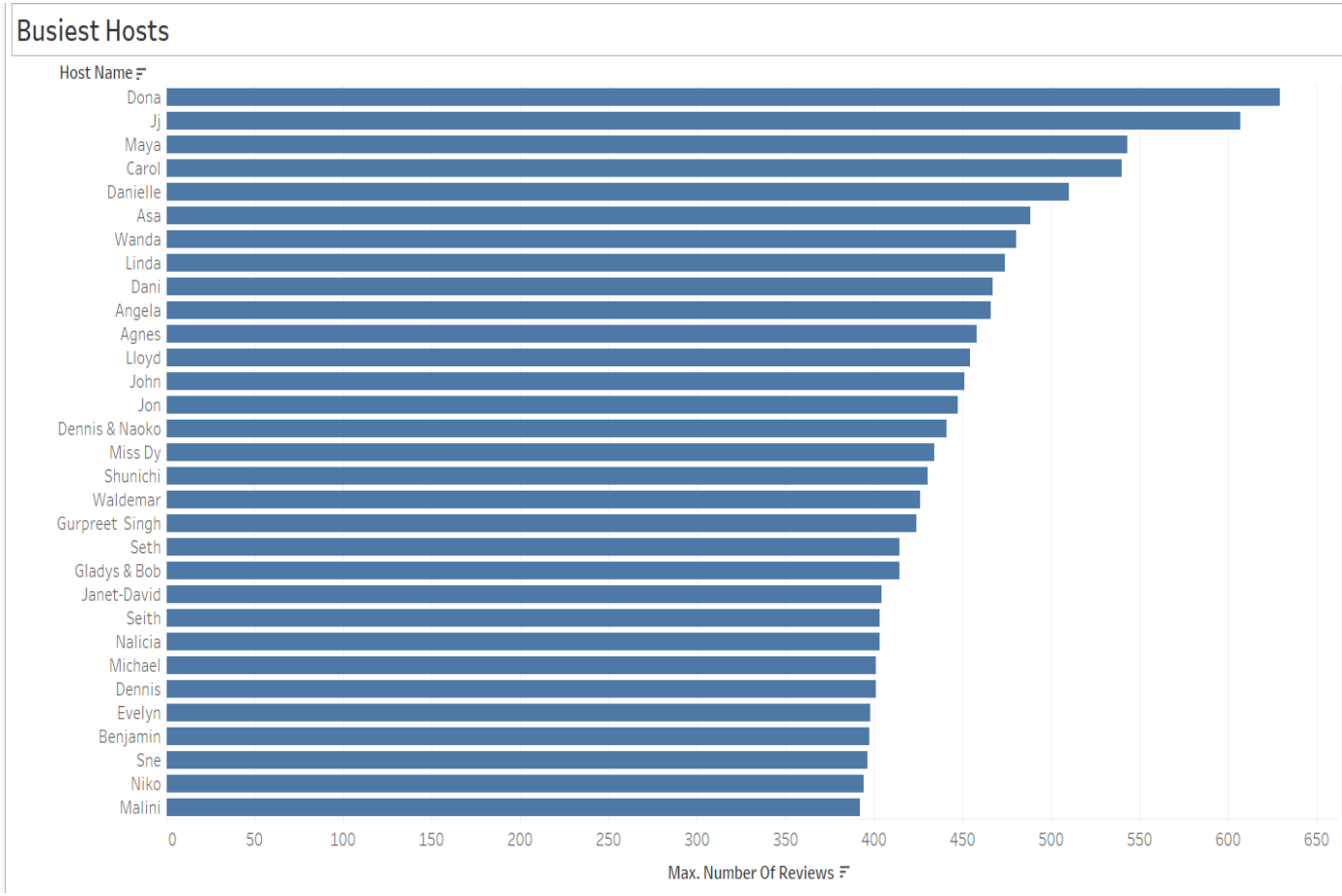
# Areas with maximum reviews

- The number of reviews feature in the dataset represent the customers who have given the reviews to a particular property they have stayed in
- Looking at the bar chart, Queens has 26.45% of total reviews which is maximum share.
- Manhattan has second highest number of reviews constituting 25.53%.
- Bronx has 13.50% of total reviews.



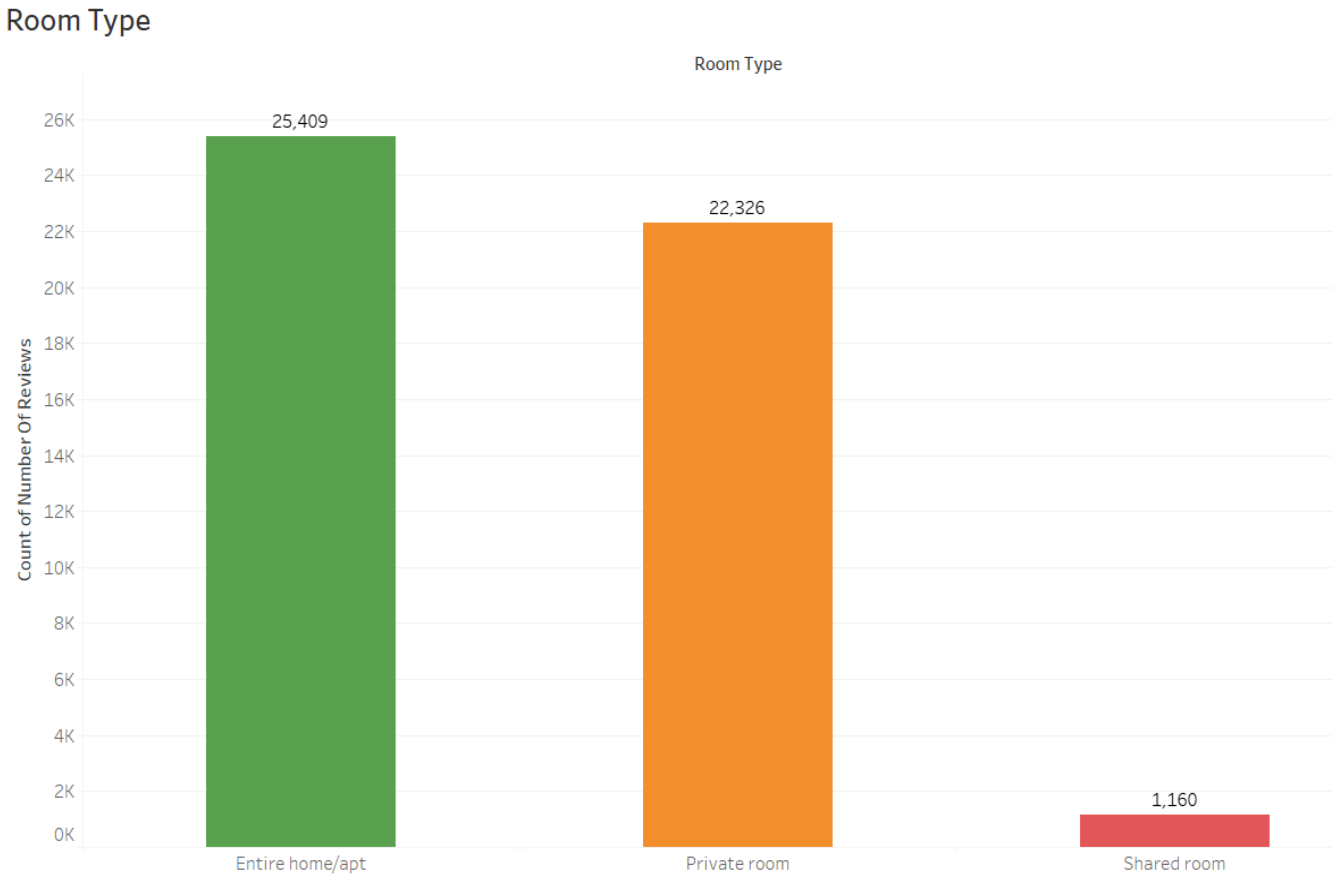
# BUSIEST HOSTS

- The adjacent bar plot shows the top 30 hosts with to number of reviews.
- Among them Dona has the highest number of reviews and we can assume on the basis of number of reviews that Dona is popular and is the busiest host .
- The top hosts have listed private room, entire home/apartment.



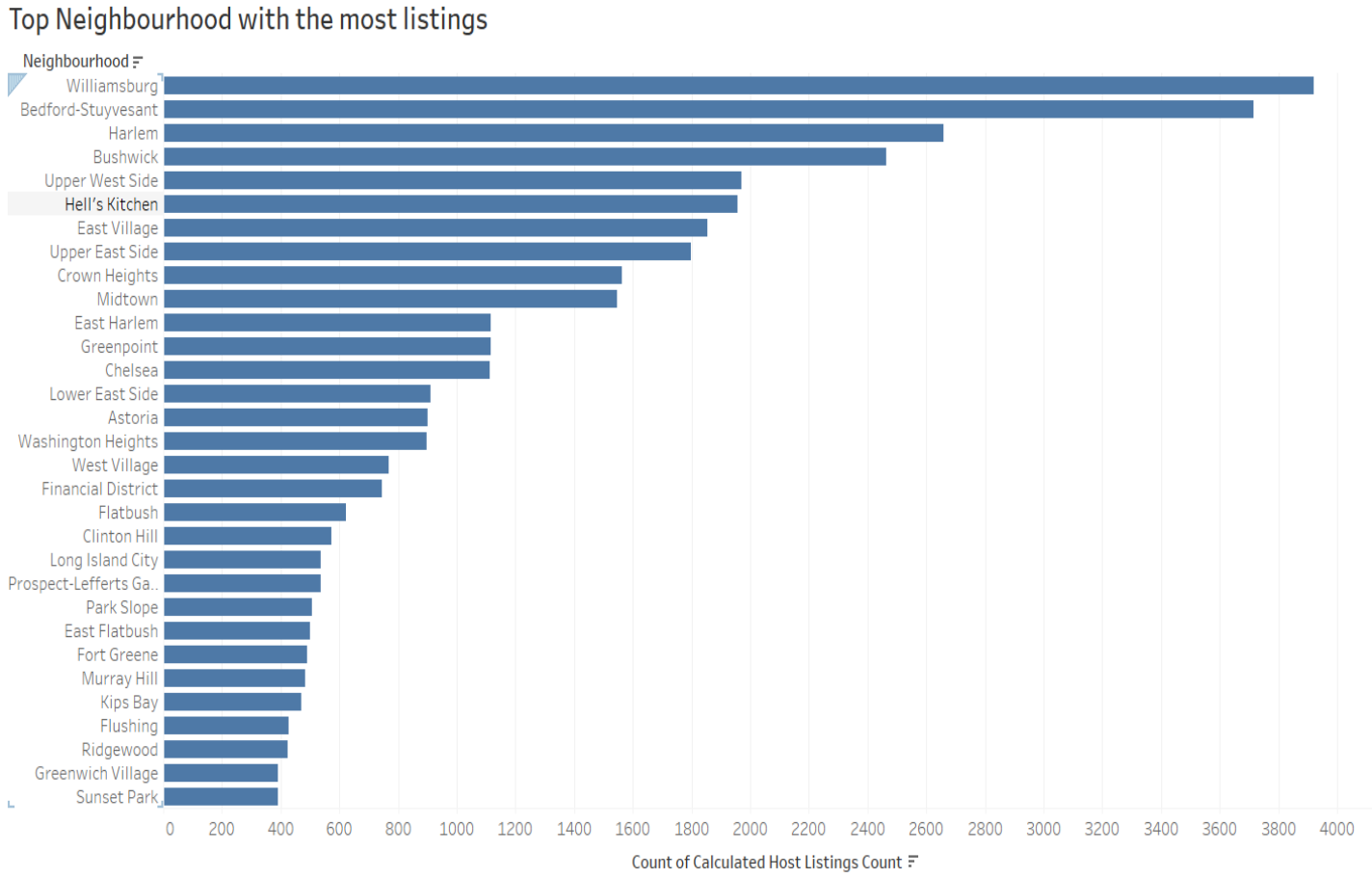
# Most Preferred Room Types

- Looking at the adjacent histogram, we can say that there are 3 room types listed in the entire dataset namely , Private room , Entire home/apt , Shared room.
- Among this types the most preferred room type is Entire home/apt as well as private room.
- Shared room is least preferred by people.



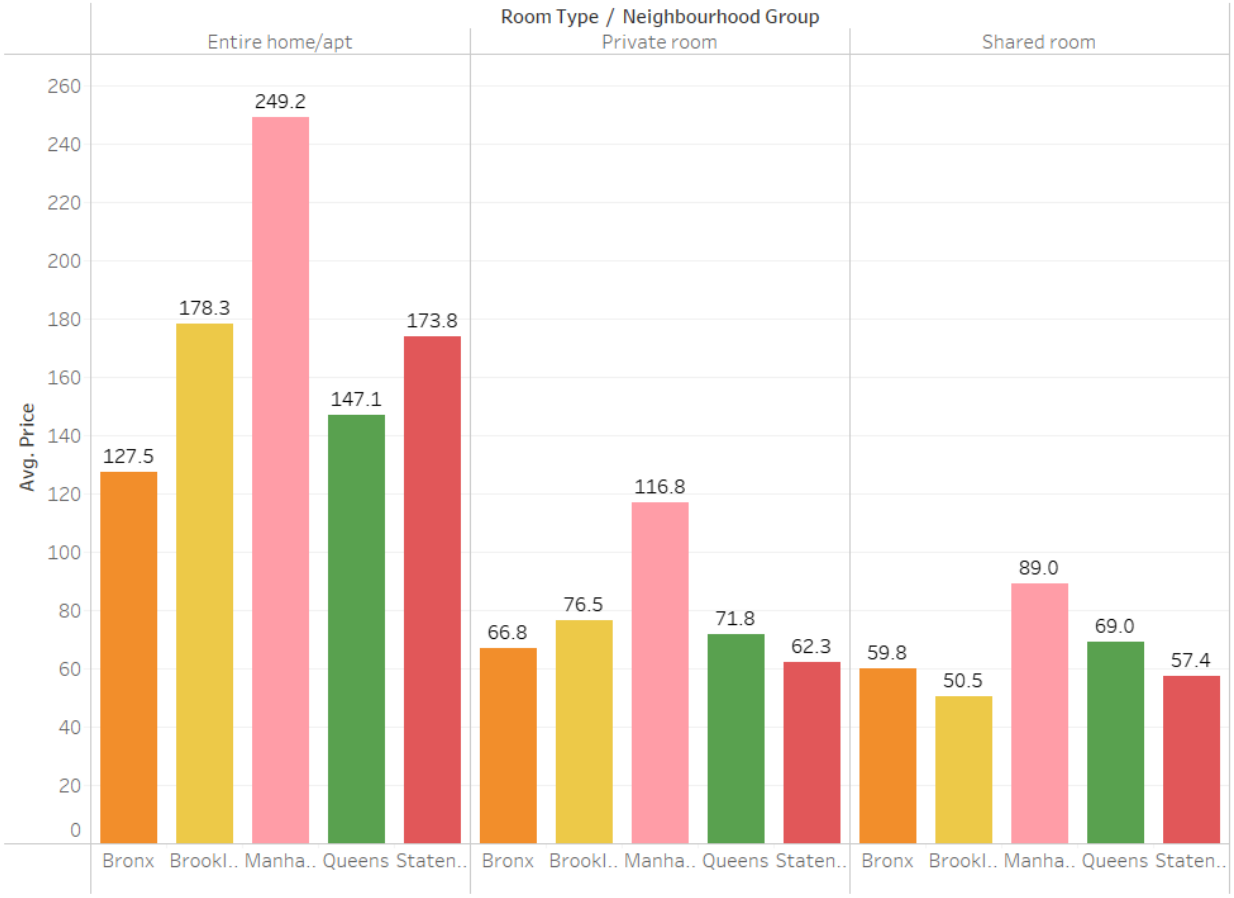
# Top 10 Neighbourhoods with Most Listings

- From the various Neighbourhoods top neighbourhood are listed in the chart.
- Williamsburg has the highest number of listing which is around 3900 and it is located in Manhattan.



# Prices for different room types

- We can easily view that the prices for the Entire home/apt is very much high in comparison to the other two room types.
- For Entire home/apt Manhattan has the highest  average price following by Brooklyn and Staten Island.
- For private room the prices are lesser in comparision to entire home/apt and also the prices for Manhattan is highest for Private room aswell.

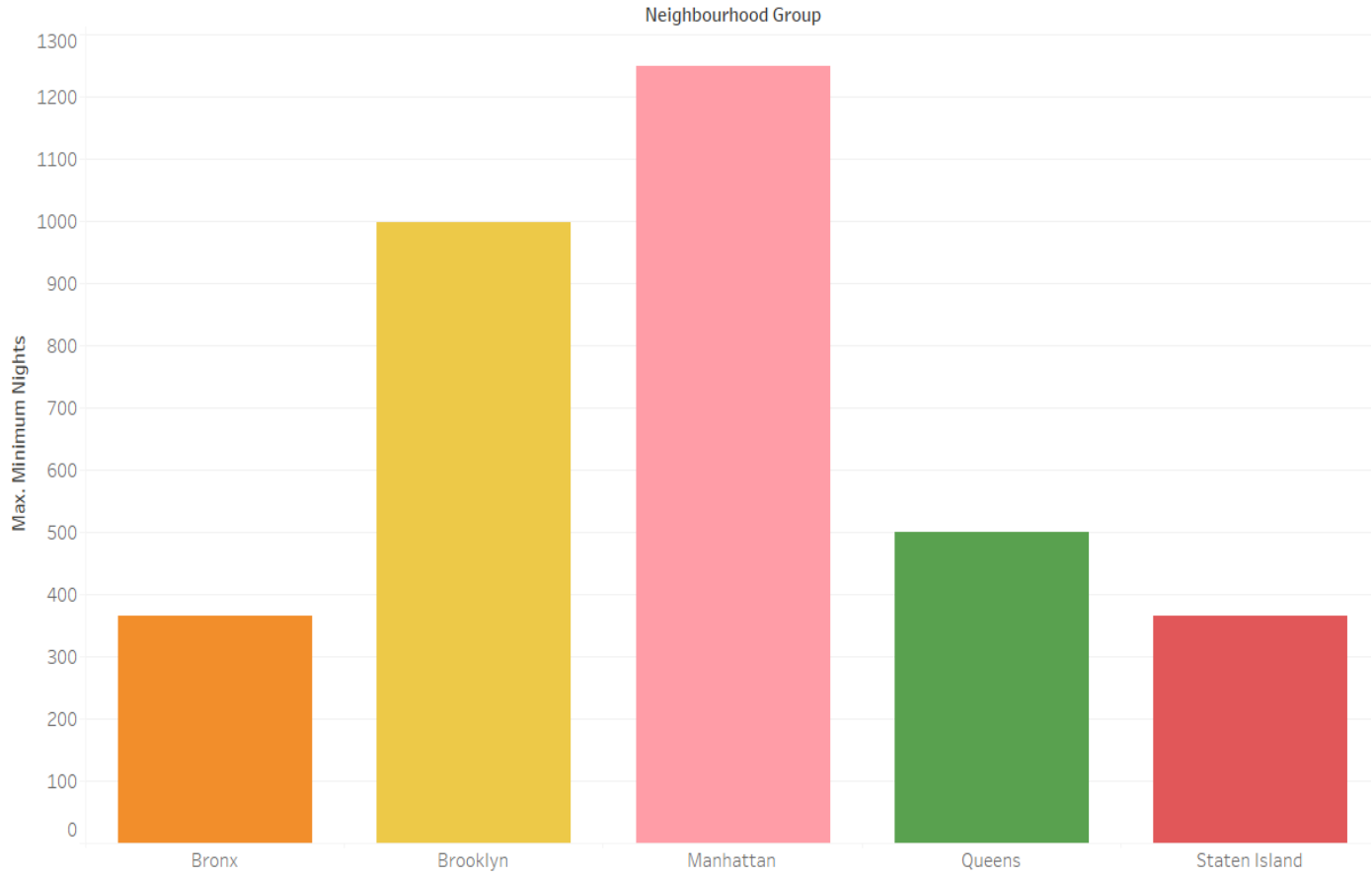


# Availaibility for different neighbourhood groups

- We can infer that the availability for Manhattan is less we can assume that the tourist prebook Entire homes and private rooms so that they do not have to face any issues of availability
- The availability for Brooklyn is much higher in comparison to Manhattan.
- Bronx has the lowest availability among all and Staten Island has the highest availability among all.
- Queen has also less availability so we can assume that it also has its prebooking.

# Minimum nights for different neighbourhood group

- The maximum number of minimum nights spent is in Manhattan and then Brooklyn.
- Manhattan has much more number of customers overall and then Brooklyn among all.
- The maximum number of minimum nights tells us the overall demand for different neighbourhood group



# Latitude and Longitude

- From the latitude and longitude of different neighbourhood groups we have analysed that Manhattan , Brooklyn , Queens are at not so far from each other and that might be the reason that these three neighbourhood groups have been on highlight during our analysis .

- So I can analyse that if a person is going to either of these places can easily visit other two aswell.

# Conclusion:

- Manhattan has the most number of listings,followed by Brooklyn and Queens, Staten Island has least number of listings.
- Manhattan and Brooklyn make up for 87% of listings available in NYC.
- Brooklyn and Manhattan are most liked neighbourhoods groups by people.
- Queens has significantly less host listings in Manhattan. So we should take enough steps to encourage Host listings in Queens as there are decent demand in neighbourhood of Queens.
- The maximum demand is for Entire home/apartment and Private rooms.
- People are most interested in cheaper rentals.
- The top 10 neighbourhood with most listing are located either in Manhattan or Brooklyn with Harlem and Williamsburg presenting leading numbers in each borough , respectively.

- Manhattan is the top neighbourhood group in the terms of listings as well as highest price range . It was assumed that Brooklyn might have most number of listings as it is a quite popular place.

- Given that Manhattan is world-famous for its museums, stores , parks, and theaters , also its substantial number of tourists throughout the year, it makes perfect sense that prices are much higher in the neighbourhood group.

- Brooklyn comes in second with the significant number of listings and cheaper prices as compared to the Manhattan , with most listings located in Williamsburg and Bedford Stuyvesant-two neighbourhoods strategically close to Manhattan-tourists get the chance to enjoy both boroughs equally while spending less.


***With Regards***

***Kartik Kumar***