METHODOLOGY

Step 1: **Exploratory Data Analysis:**

- Checked the Null values in the dataset. Found some columns with the null values i.e., **name**, **host_name**, **last_review**, and **review_per_month**.

```
In [4]: df.isnull().sum().sort_values(ascending = False)*100/len(df)

Out[4]: last_review                     20.558339
        reviews_per_month               20.558339
        host_name                        0.042949
        name                             0.032723
        id                               0.000000
        host_id                          0.000000
        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
        calculated_host_listings_count   0.000000
        availability_365                 0.000000
        dtype: float64
```

- Same was rectified by removing the rows with null values from **name**, **host_name,** and by replacing null values by "**0**" in **review_per_month**.
- Checked the outliers in the dataset.
- Created a calculated field of number of reviews per listing

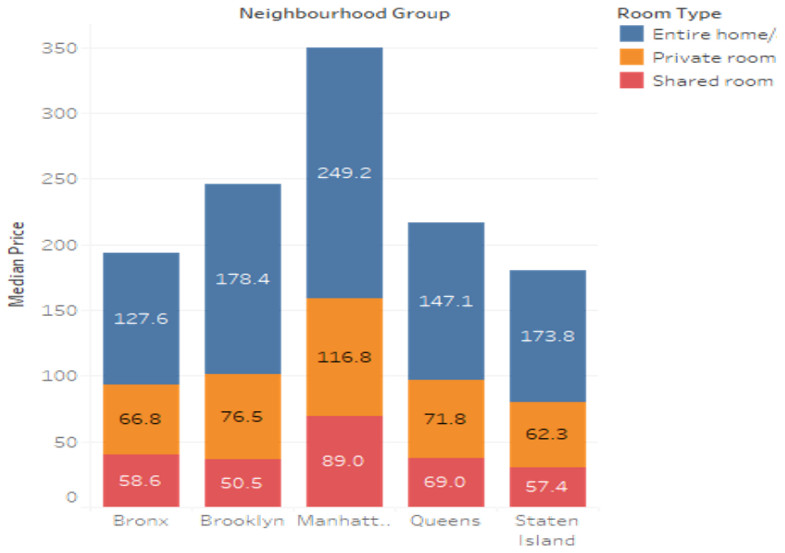
- For better understanding of **reviews_month_range** column, it was divided into bins ['**Low**', '**Medium**', '**High**'] and for **min_nights_range** as follows.

```
# Creating bins for the minimum_nights column :

min_nights_range = ['Less than a week','1-2 Weeks','2-4 Weeks','1-3 Months','3 Months - 1 Year','More than a Year']
min_nights_range
```

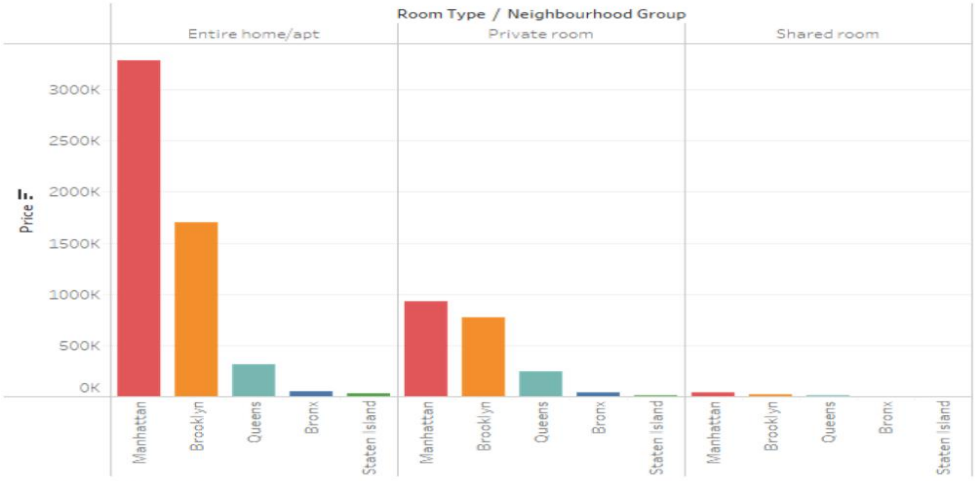
Step 2: Data Analysis

- Tried to analyse the data using different columns on the basis of their **price**, **availability_365**, **minimum_nights** and the reviews of the customers we received.
- Checked neighbourhood grouped wise distribution of price and room type.



Median of Price for each Neighbourhood Group. Colour shows details about Room Type. The marks are labelled by average of Price.

- Room type v/s Neighbourhood



Sum of Price for each Neighbourhood Group broken down by Room Type. Colour shows details about Neighbourhood Group.

Room Type,Neighbourhood Group,count of Calculated Host Listings Countandaverage of Number Of Reviews. Colour shows average of Number Of Reviews. Size shows count of Calculated Host Listings Count. The marks are labelled by Room Type,Neighbourhood Group,count of Calculated Host Listings Countandaverage of Number Of Reviews.

Avg. Number Of Reviews

1.56 33.28

- Through these we can clearly infer:
  - We saw that people like to visit the centre of New York from where they can see the beauty of the city.
  - Number of listings of shared rooms are limited but their average price is placed less and availability is high.
  - Number of reviews and reviews per month are more at less price than the higher price as there is less chance of people going for a high price room.
  - Manhattan and Brooklyn are very costly neighnourhood_groups.
  - People show interest in the host Blueground, and spend more nights here.
  - Minimum number of nights to stay reduces with increase in price.
  - Focus on prime locations like Manhattan and Brooklyn where people show interest.

Step 3: Presentation

- A presentation has been given to the Management and Data-Analyst team adhering to the best practices and pyramid principle.
- Recommendations on the basis of thorough analysis has been included in the presentations for the respective departments.