Methodology Document for Storytelling Case Study: Airbnb, NYC

Step -1

Using Jupiter notebook to perform the initial analysis of the data

Info

```
airbnb.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
```

Missing values

```
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
```

Unique Values

```
airbnb.nunique()

id                                48895
name                              47896
host_id                           37457
host_name                         11452
neighbourhood_group                   5
neighbourhood                       221
latitude                          19048
longitude                         14718
room_type                             3
price                               674
minimum_nights                      109
number_of_reviews                   394
last_review                        1764
reviews_per_month                   937
calculated_host_listings_count       47
availability_365                    366
dtype: int64
```

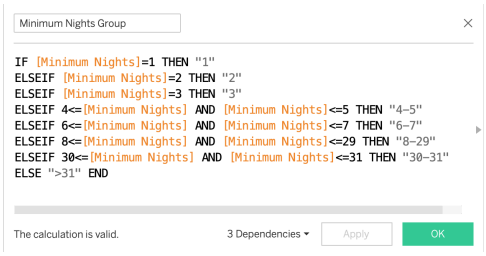
- We checked for duplicate rows in the dataset and found no duplicate data.
- We also identified null values in certain columns such as name, host_name, last_review, and reviews_per_month.
- However, we decided to leave the dataset unchanged as the missing values do not significantly impact our analysis."

Step -2

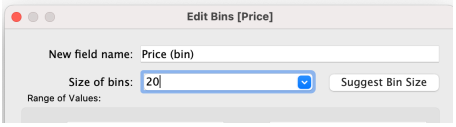
We used Tableau to create required columns for our data analysis. The following are the detailed steps involved

Adding new columns to the data source

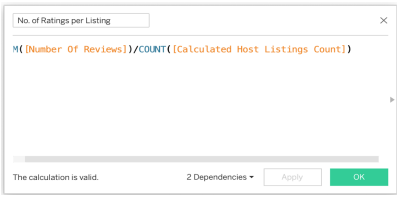
1. Grouping the Minimum nights



2. Binning the Price column



3. Creating calculating field

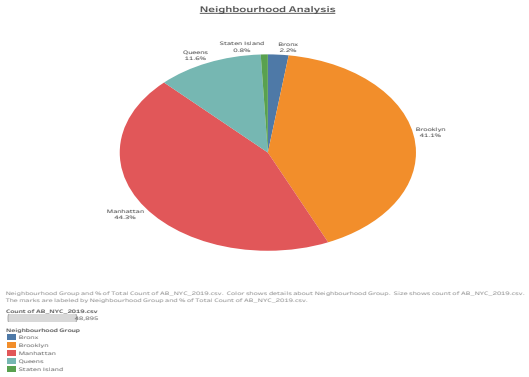


Step – 3

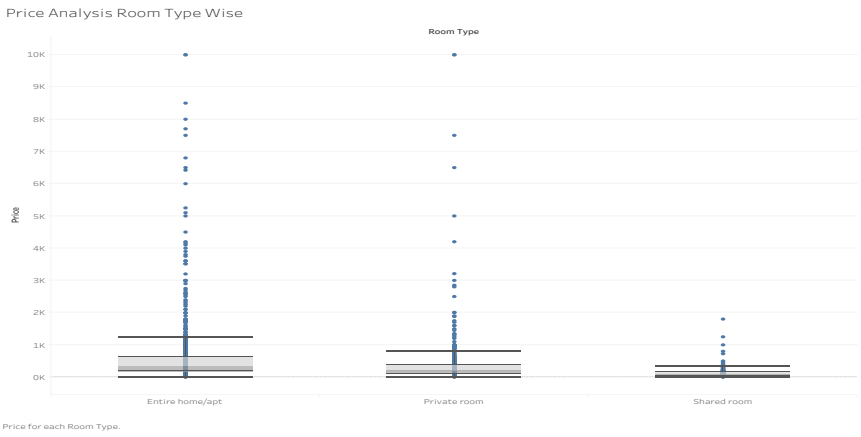
We used Tableau to create visualizations for our data analysis. The following are the detailed steps we used for each visualization

Visualizations:

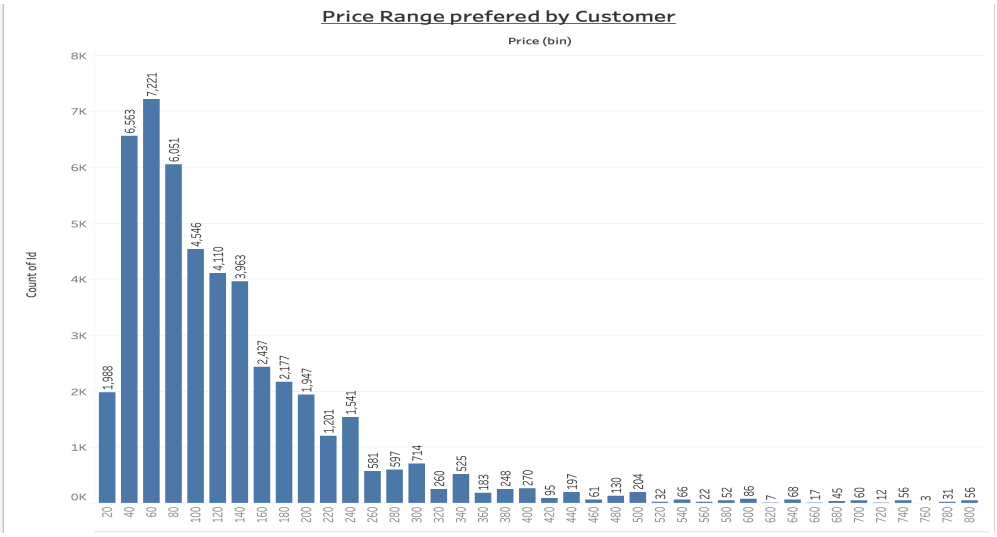
1. Analysed through PIE chart that Manhattan has the highest % in the Neighbourhood group



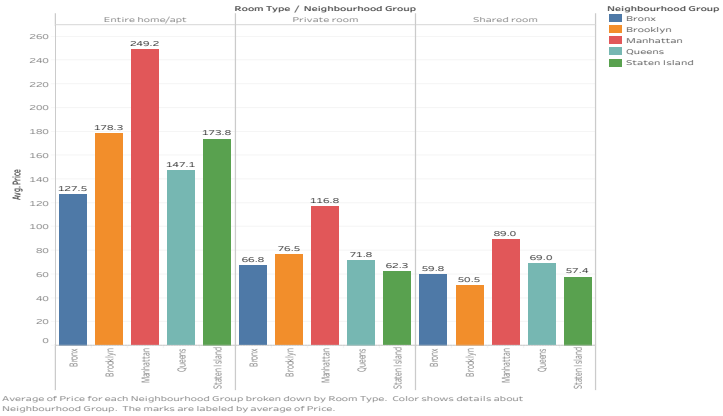
2. Utilized Box plot to check the outliers in the Price w.r.t to the room type



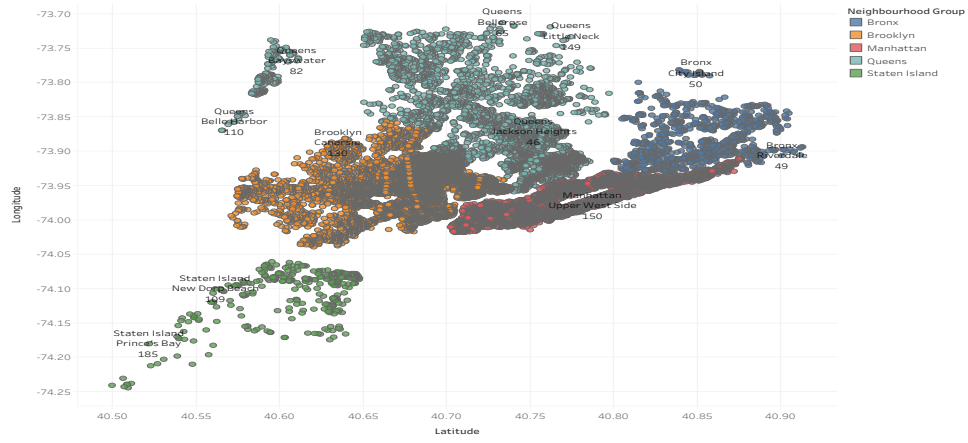
3. Analysed the customer preferred price range through Bar chart



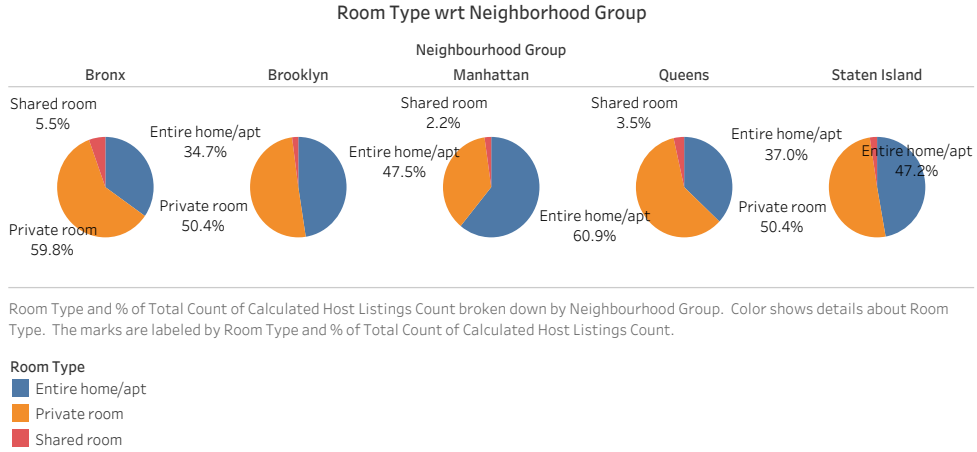
4. Analysed the two categorical variables – Room type & Neighbourhood group with average price through Bar chart by adding Neighbourhood group to the colour marks card to highlight the different Neighbourhood group in different colours.



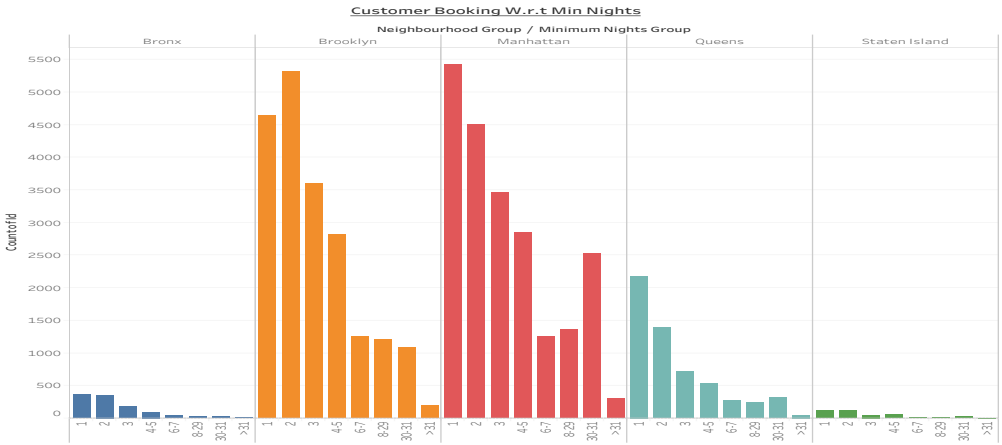
5. Analysed the Listings of the neighbourhood group through Geographical locations in the Maps and also indicated the labels through selecting the mark labels at several locations



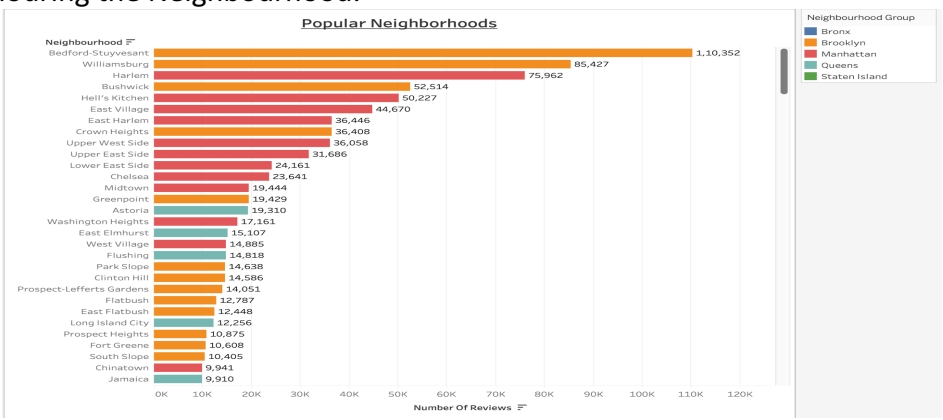
6. Analysing Room type portion w.r.t Neighbourhood Group from the calculated Host listing count and represented in the Pie Chart



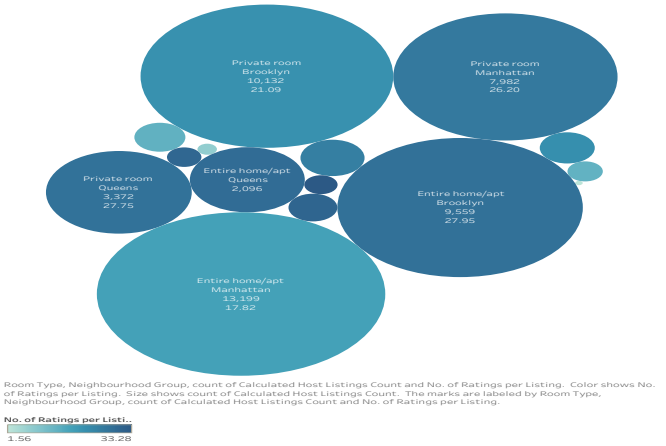
7. Analysing the customer booking w.r.t minimum nights by considering the Neighbourhood group and Minimum nights group in the columns and count of IDs in the rows and highlighted the different Neighbourhood groups in different colours to visualize the data.



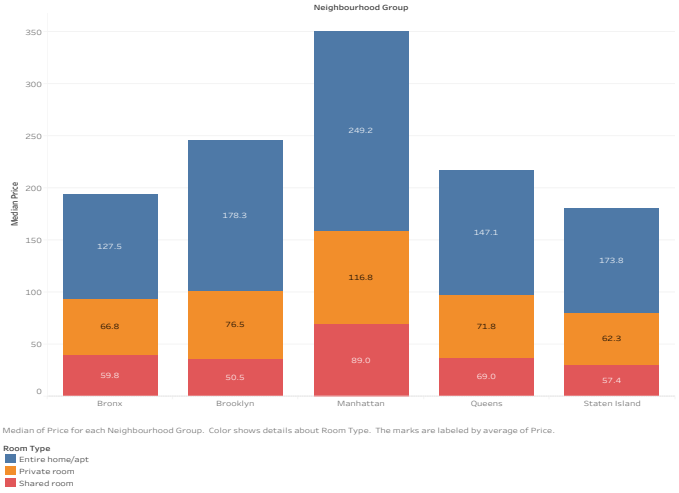
8. Analysing the categorical – Neighbourhood and continuous variable – No. of reviews in the Bar chart with labelling the sum of no. of reviews and also differentiating colouring the Neighbourhood.



9. In bubble chart sizing the Calculated Host Listings count and labelling the Room type, Neighbourhood group, Calculated Host Listings count, no. of ratings per listing and also colouring the No. of ratings per listing.



10. Analysing the categorical variable – Neighbourhood group with continuous variable – median price by colouring the room type



Step – 4 Presentation

- We prepared the presentation adhering to best practices, including the Pyramid Principle and the Rule of Three.
- We also included recommendations for the respective departments.