Data Methodology

Step 1: Storyboarding

- Went through the data to get familiarized with it and noted down important fields
- Made a mind map of the various slides of the presentation
- Made a rough template based on this mind map

Step 2: Data Wrangling

- Explored all the columns in the dataset by importing it to python notebook
- Checked for the Missing values and found out columns name, host_name, last_review and reviews_per_month had missing values.

```
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
```

- Missing values are treated using Python .Attached below is the Python Notebook along with a few snapshots.

Airbnb_CaseStudy_B
y_Madhumitha_And_I

Missing values are present in the name, host_name, last_reviews and reviews_per_month columns. In the above exploration part we can see that if the number_of_reviews is 0 then it does not make sense to have last_review and reviews_per_month and are marked as NaN. Hence the missing values in the data is following a pattern and will be treated accordingly.

Let us check if the assumption made above holds true.

```
#checking the assumption -> 0 reviews will have missing values in last_review and reviews_per_month columns.

assumption_test = data.loc[(data.last_review.isnull()) & (data.reviews_per_month.isnull())][['number_of_reviews',  'last
assumption_test.head()
```

	number_of_reviews	last_review	reviews_per_month
2	0	NaN	NaN
19	0	NaN	NaN
26	0	NaN	NaN
36	0	NaN	NaN
38	0	NaN	NaN

The exact amount of null values present in both the columns. It proves that the assumption made was clear. We will substitute 0 for the missing values present in reviews_per_month column.

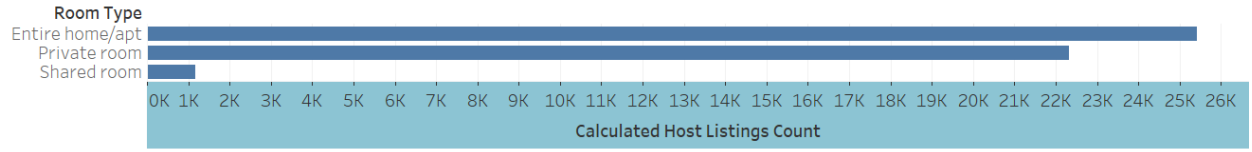
As for the last_review column we know that it is a datetime object of the pandas and substituting 0 won't make sense here. We will have to leave the null values of last_reviews as it is for now.

```
#filling the missing values in reviews_per_month with 0.
data.reviews_per_month.fillna(0, inplace=True)
```

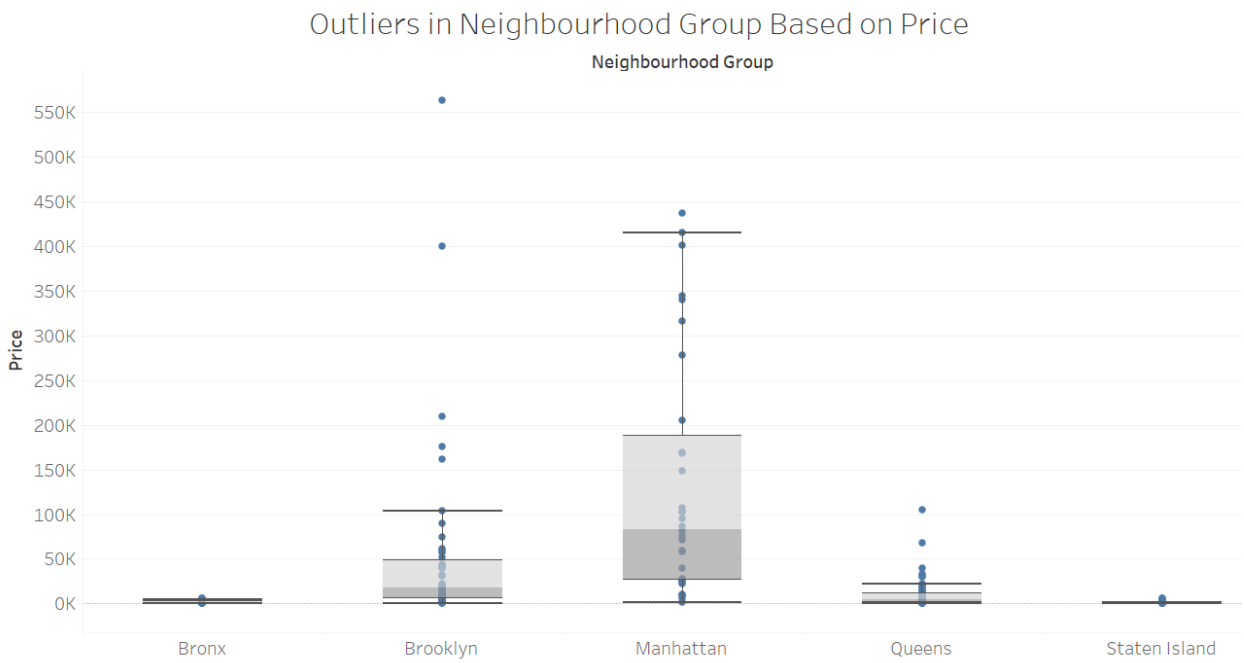
That been done, we will also replace the null values present in the host name and name columns with NA.

```
data.loc[data["name"].isnull(),'name']=data["name"].apply(lambda x:"NA")
data.loc[data["host_name"].isnull(),'host_name']=data["host_name"].apply(lambda x:"NA")
```

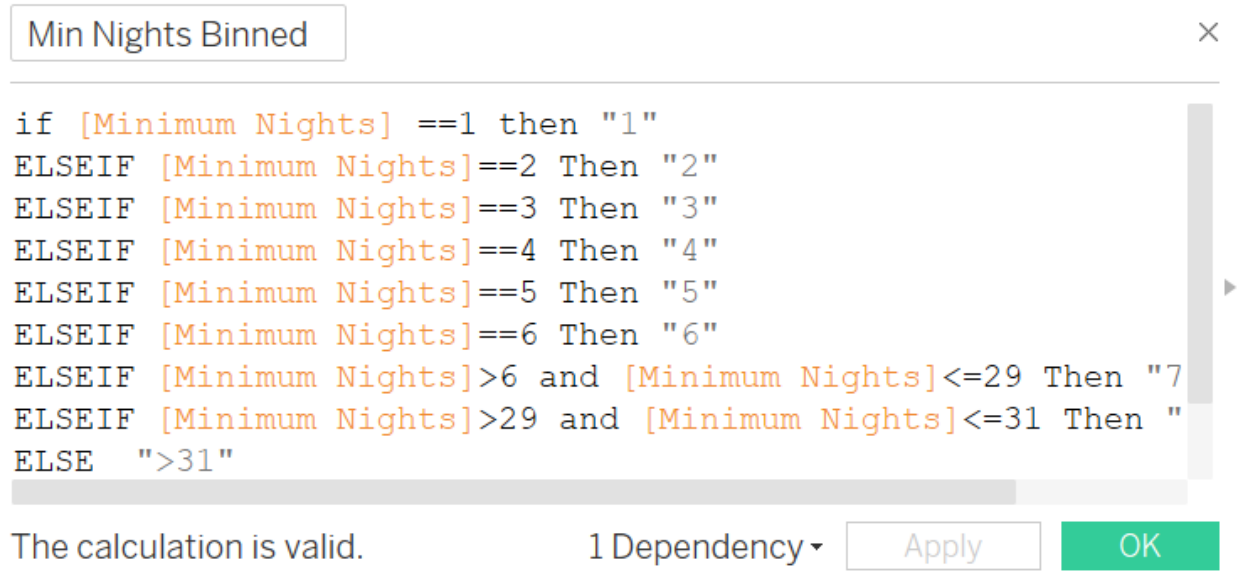
- Host Listings count is maximum for entire apartment and private room and is very small for shared room as seen below.



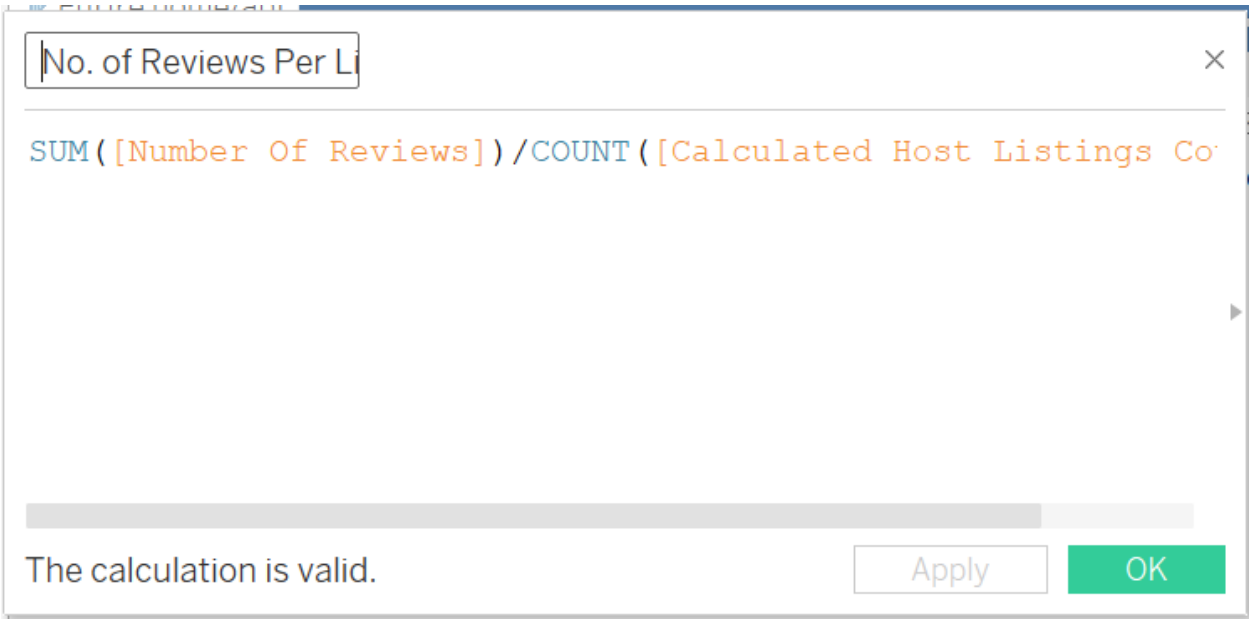
- Checked if any outliers are present w.r.t. price



- Created a grouped field for Minimum Number of Days assuming null values belonged to the category.

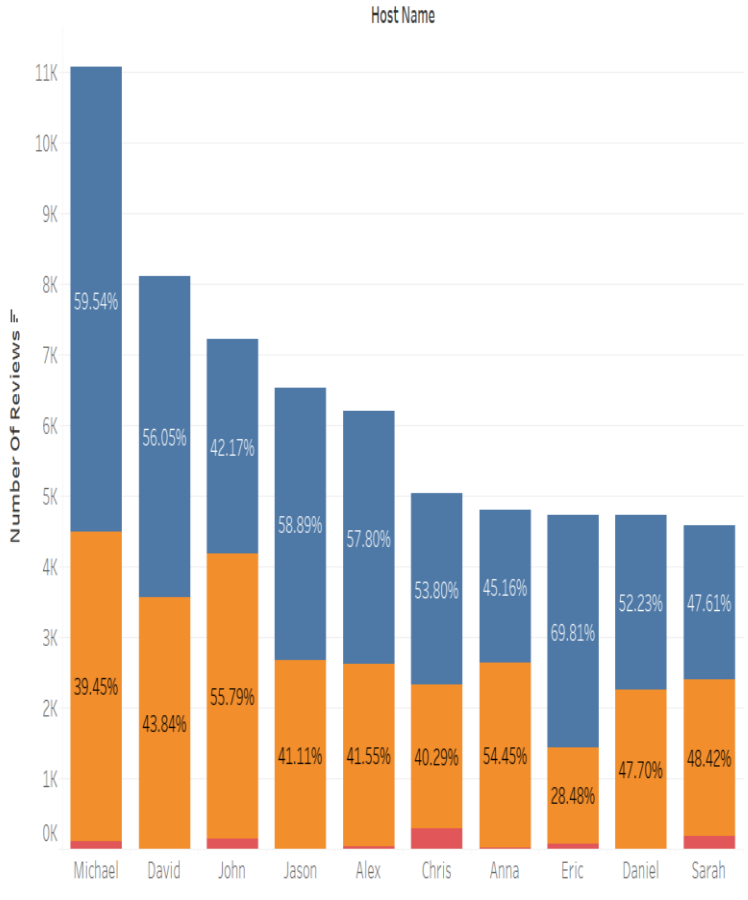


- Created a calculated field of number of reviews per listing


No. of Reviews Per Li
SUM([Number Of Reviews])/COUNT([Calculated Host Listings Co
The calculation is valid.
Apply
OK

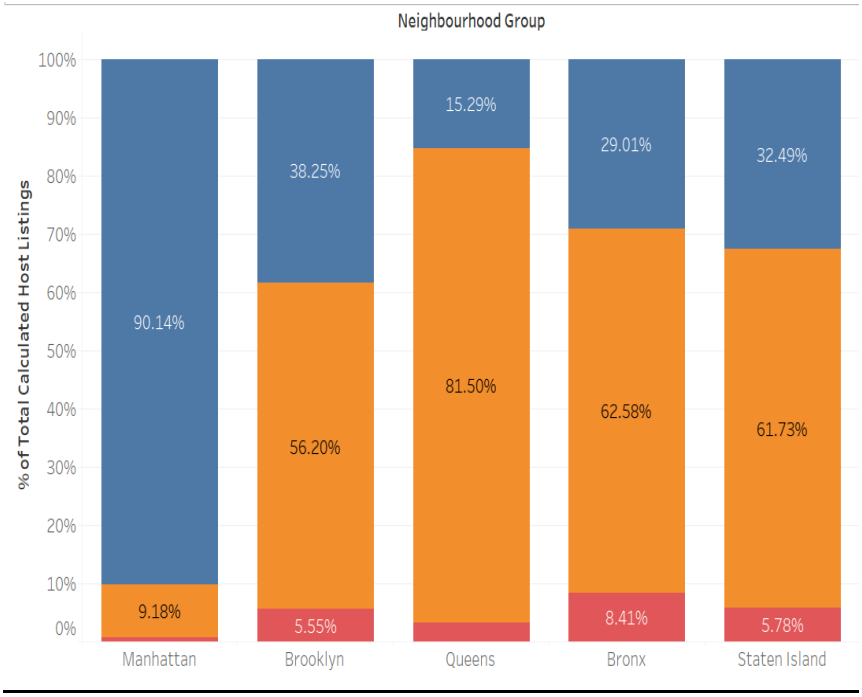
Step 3:Data Analysis

Top 10 Hosts by Reviews



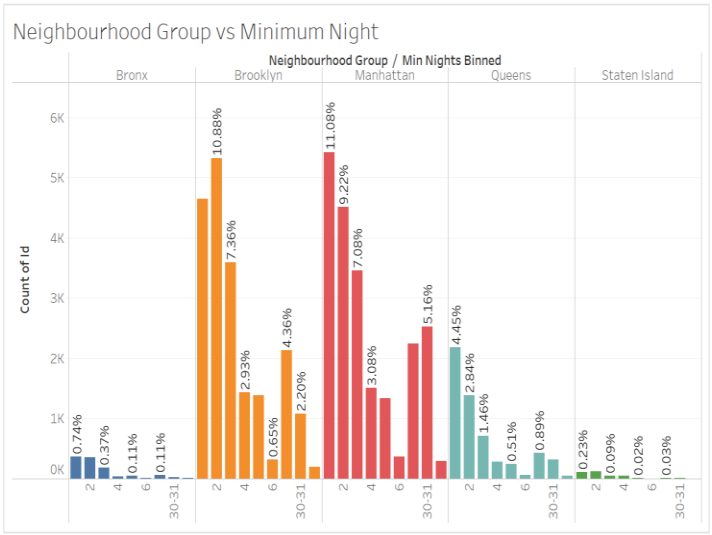
- Shared rooms accounts the least number of reviews of the total listed properties
- More than 50% of the hosts prefer renting out the entire home/apartment
- Private room & Entire home/apartment seems to be popular(more than 90% reviews)

Customer Preferences of Properties in NYC Areas



- The properties in Manhattan are the most expensive than any other area.
- Manhattan has the highest contribution of ‘Entire home/apt’ compared to the overall contribution of ‘Entire home/apt’.
- Queens has a higher contribution of ‘Private room’ compared to the overall contribution of ‘Private room’.

Preferences of Neighbourhood Group w.r.t Minimum Nights



- 20% of the bookings are made either in Manhattan or Brooklyn.
- As Manhattan & Brooklyn are expensive areas, majority of the people prefer staying below 3 nights
- Staten Island is still a developing place while Bronx is the poorest borough, only 10% of the people prefer renting out.

Tableau Workbook attached: 
Tableau_Case_Study_Final.twbx

Step 4:Presentation

- Made the presentation adhering to best practices and pyramid principle.
- Here **Data Analysis Managers** & **Lead Data Analyst** are our audience
- Added recommendations for the respective departments