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
- Listed out all the attributes for which we need to plot graphs and type of the charts

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 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.

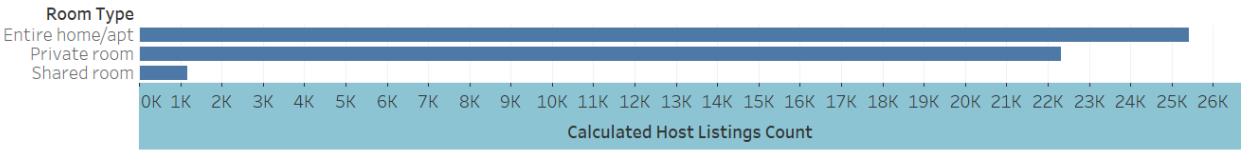
As for the last_review column we know that it is a datetime object of the pandas and substituting 0 won't make sense here. We will have to leave the null values of last_reviews as it is for now.

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

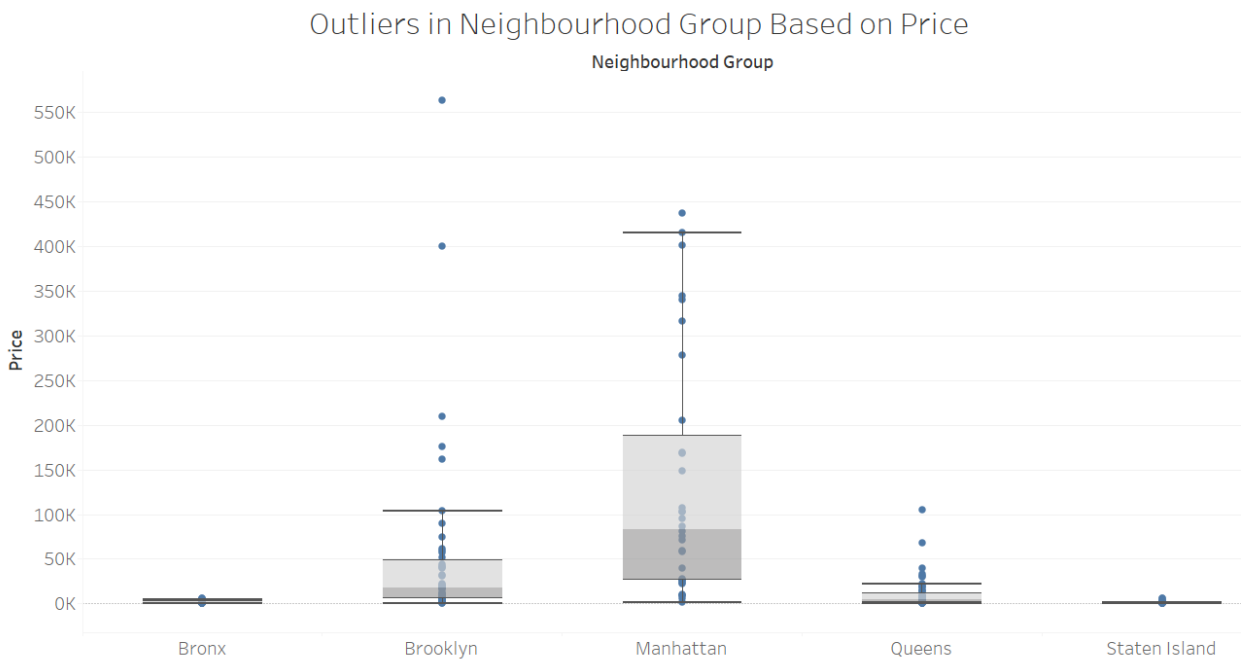
That been done, we will also replace the null values present in the host name and name columns with NA.

```
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")
```

- Exported the above changes to a csv file & used it in Tableau for further Visualization.
- 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



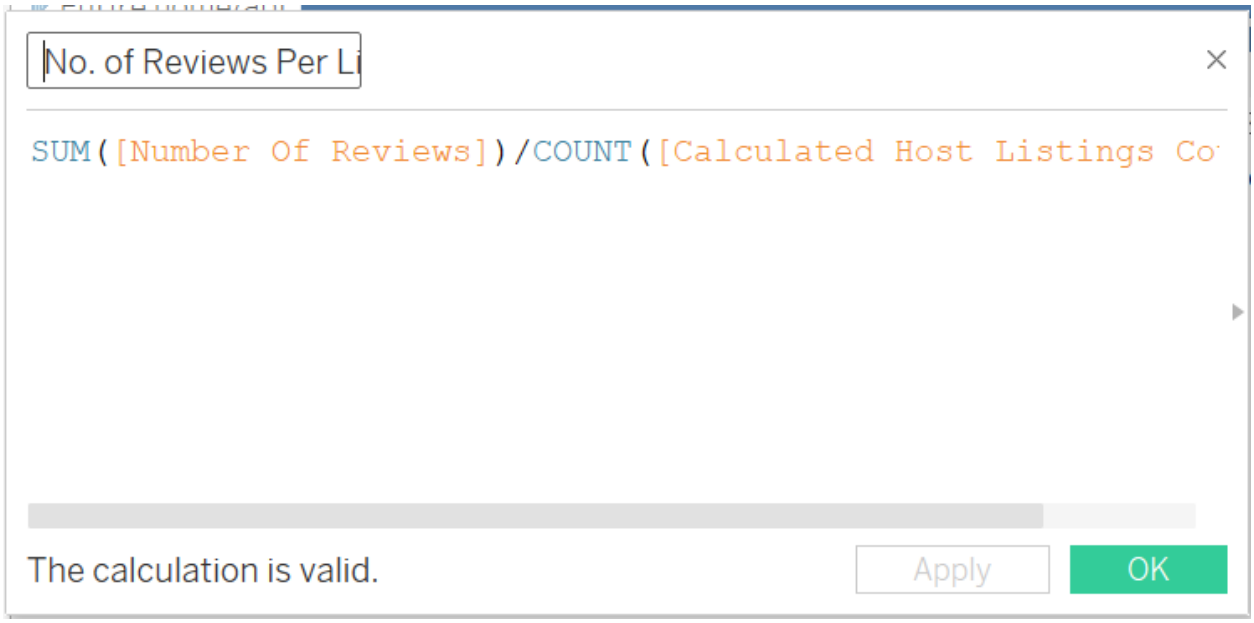
- Created a grouped field for Minimum Number of Days assuming null values belonged to the category.

Min Nights Binned

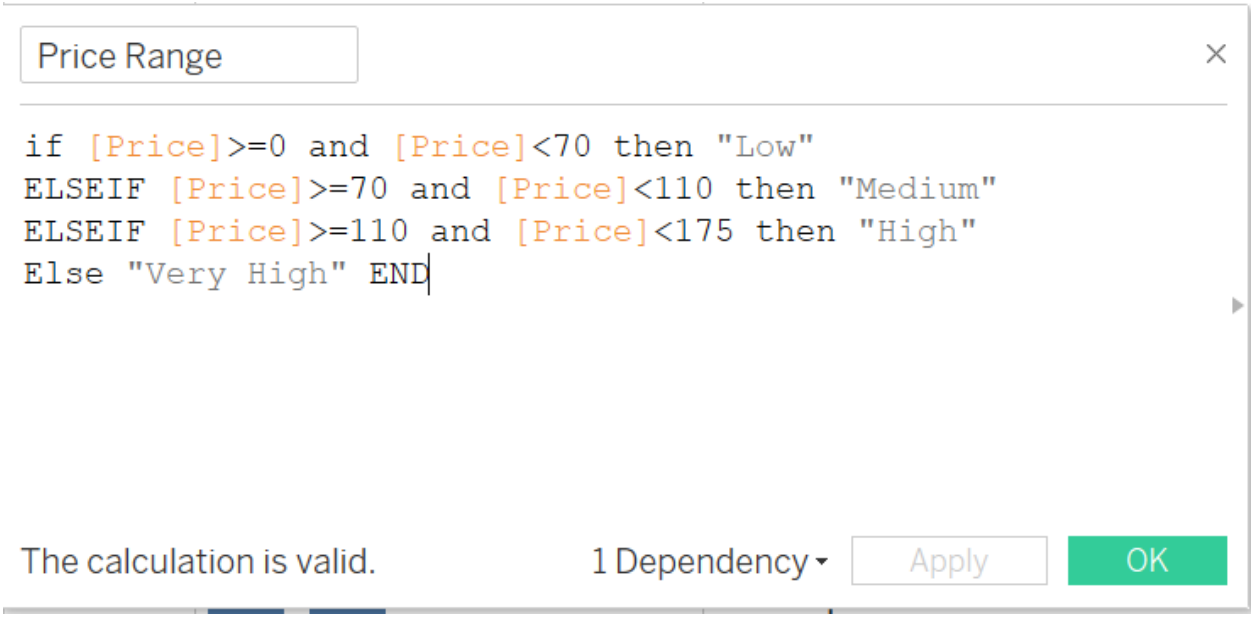
```
if [Minimum Nights] ==1 then "1"
ELSEIF [Minimum Nights]==2 Then "2"
ELSEIF [Minimum Nights]==3 Then "3"
ELSEIF [Minimum Nights]==4 Then "4"
ELSEIF [Minimum Nights]==5 Then "5"
ELSEIF [Minimum Nights]==6 Then "6"
ELSEIF [Minimum Nights]>6 and [Minimum Nights]<=29 Then "7
ELSEIF [Minimum Nights]>29 and [Minimum Nights]<=31 Then "
ELSE  ">31"
```

The calculation is valid. 1 Dependency Apply OK

- Created a calculated field of number of reviews per listing

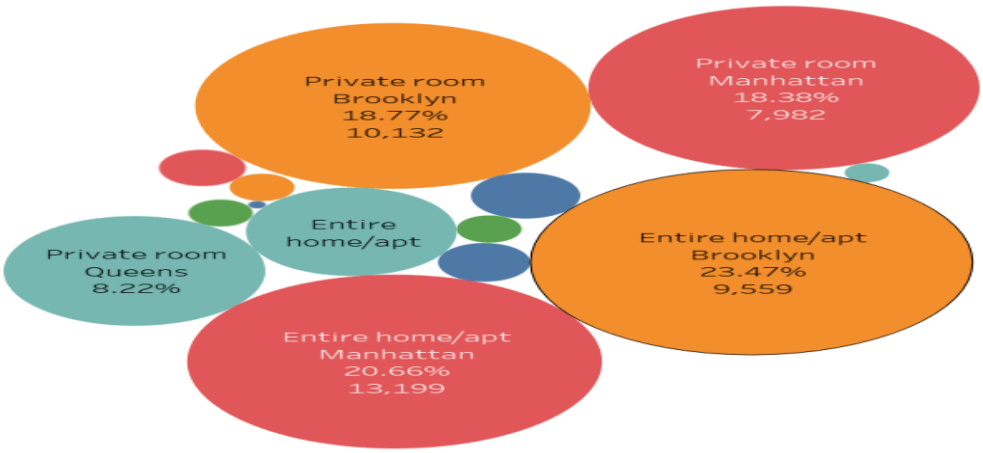


- Created a Price Range (Low, Medium, High, Very High) for the price column using calculated field



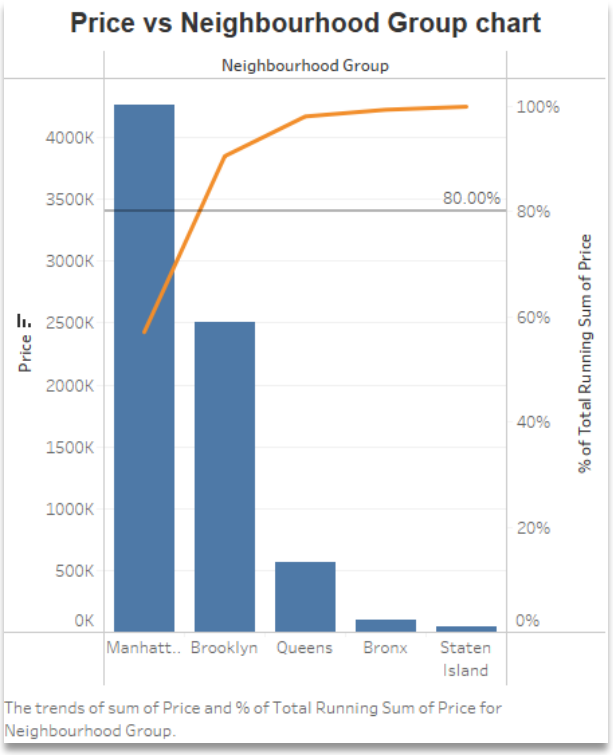
Step 3:Data Analysis

Neighborhood Wise Business Distribution



- **Private Rooms & Entire home/apt** should be targeted in Brooklyn & Manhattan as they seem to be popular.
- **Shared Rooms** should be targeted as the rates already cheap and a reasonable approach will get more customers.
- The properties in Manhattan & Brooklyn are the most expensive than any other area.

Insights on the price based on Neighbourhood Group



- Focus on increasing the listings in Queens ,Bronx, Staten Island to increase their revenue.
- Manhattan contributes to 90% of revenue
- As Manhattan leads the list and is quite popular, hosts can reduce the price to attract customers.

Popular Accommodations across NYC

Neighbourhood contributing to higher income

Neighbou..	Neighbourhood	Fixed LOD	Price	Percent of Income Contribution
Brooklyn	Bedford-Stuyves..	2,500,600	399,917	11.82%
	Bushwick	2,500,600	209,033	6.18%
	Williamsburg	2,500,600	563,707	16.66%
Manhattan	Hell's Kitchen	4,264,527	400,987	6.95%
	Midtown	4,264,527	436,801	7.57%
	Upper West Side	4,264,527	415,720	7.21%
Queens	Astoria	563,867	105,469	13.83%
	Long Island City	563,867	68,449	8.97%
Staten Island	Randall Manor	42,825	6,384	11.02%
	St. George	42,825	5,671	9.79%

- The following are popular places in each of the Neighborhood group and contributes to higher income and hence target these places more for revenue generation.

- Brooklyn –Williamsburg
- Manhattan-Midtown
- Queens-Astoria
- Staten Island-Randall Manor

Enclosed is the packaged workbook:

Tableau_Case_Study_Final.twbx

Step 4:Presentation

- Made the presentation adhering to best practices and pyramid principle.
- Here **Head of Acquisitions and Operations** & **Head of User Experience** are our audience
- Added recommendations for the respective departments