AIRBNB CASE STUDY - METHODOLOGY

In the case study we have used Python to perform initial analysis and data cleaning and then exported back the data as csv file. The further analysis and data visualization was done using MS-Excel and Tableau.

- DATA SOURCING

```
In [1]: import warnings
        warnings.filterwarnings('ignore')
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt,seaborn as sns
        %matplotlib inline
```

```
In [2]: data= pd.read_csv('C:\\Users\\Admin\\Downloads\\AB_NYC_2019.csv')
        data.head()
```

Out[2]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_revie
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	2
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

- DATA CLEANING

```
In [3]: data.shape
Out[3]: (48895, 16)
```

```
In [4]: #Checking NULL values
        data.isnull().sum()
Out[4]: 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
```

Removing null values

```
data= data[~data.name.isnull()]
```

```
data= data[~data.host_name.isnull()]
```

```
data.drop('last_review',inplace=True,axis=1)
```

```
In [9]: data.room_type.value_counts()
```

```
Out[9]: Entire home/apt    25393
        Private room       22306
        Shared room         1159
        Name: room_type, dtype: int64
```

```
In [13]: data.neighbourhood.value_counts()
```

```
Out[13]: Williamsburg          3917
         Bedford-Stuyvesant    3713
         Harlem                2655
         Bushwick              2462
         Upper West Side       1969
                               ...
         Fort Wadsworth           1
         Richmondtown             1
         New Dorp                 1
         Rossville                1
         Willowbrook              1
         Name: neighbourhood, Length: 221, dtype: int64
```

```
In [14]: data.neighbourhood_group.value_counts()
```

```
Out[14]: Manhattan        21643
         Brooklyn         20089
         Queens            5664
         Bronx             1089
         Staten Island      373
         Name: neighbourhood_group, dtype: int64
```

Checking for wrong values

In [28]:
```
data.latitude.describe()
```

Out[28]:
```
count    48858.000000
mean        40.728941
std          0.054528
min         40.499790
25%         40.690090
50%         40.723070
75%         40.763107
max         40.913060
Name: latitude, dtype: float64
```

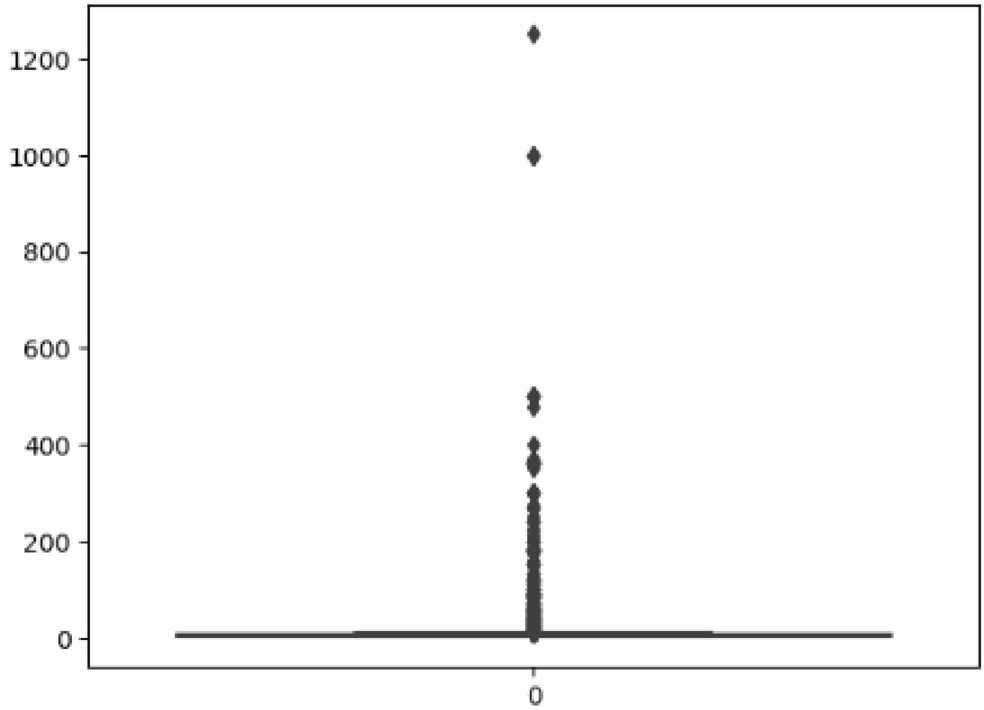
In [29]:
```
data.longitude.describe()
```

Out[29]:
```
count    48858.000000
mean       -73.952170
std          0.046159
min        -74.244420
25%        -73.983070
50%        -73.955680
75%        -73.936280
max        -73.712990
Name: longitude, dtype: float64
```

In [31]:
```
data.minimum_nights.describe()
```

Out[31]:
```
count    48858.000000
mean         7.012444
std         20.019757
min          1.000000
25%          1.000000
50%          3.000000
75%          5.000000
max       1250.000000
Name: minimum_nights, dtype: float64
```

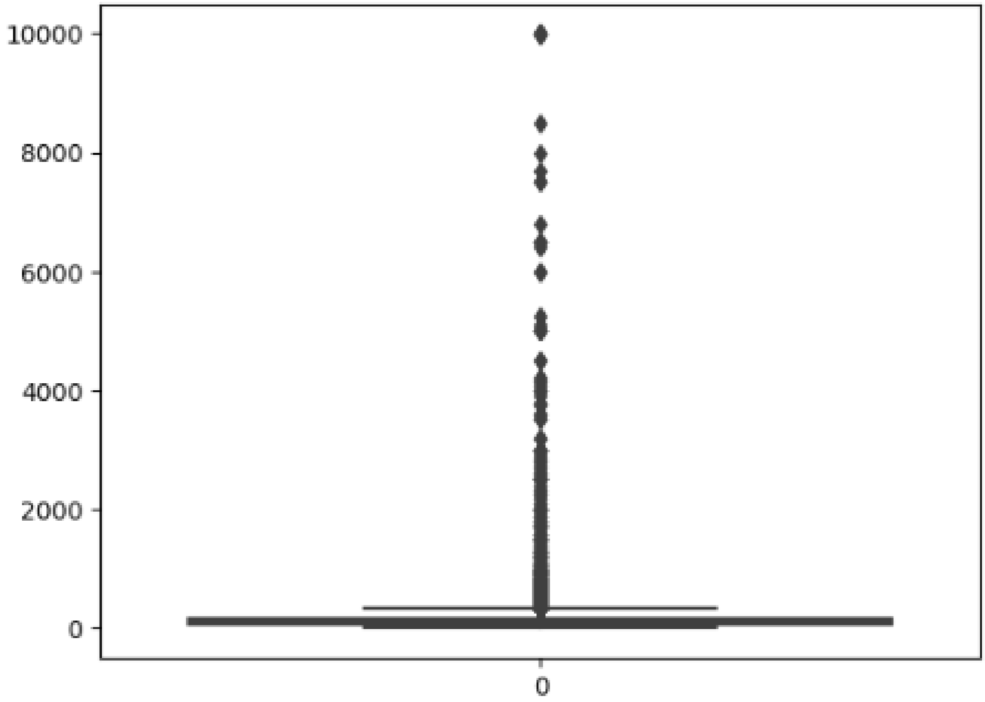
In [25]:
```
sns.boxplot(data=data.minimum_nights)
plt.show()
```



In [30]:
```
data.price.describe()
```

Out[30]:
```
count    48858.000000
mean       152.740309
std        240.232386
min          0.000000
25%         69.000000
50%        106.000000
75%        175.000000
max      10000.000000
Name: price, dtype: float64
```

In [29]:
```
sns.boxplot(data=data.price)
plt.show()
```



In [15]:
```
data.number_of_reviews.describe()
```

Out[15]:
```
count    48858.000000
mean        23.273098
std         44.549898
min          0.000000
25%          1.000000
50%          5.000000
75%         24.000000
max        629.000000
Name: number_of_reviews, dtype: float64
```

In [17]:
```
data.calculated_host_listings_count.describe()
```

Out[17]:
```
count    48858.000000
mean         7.148369
std         32.964600
min          1.000000
25%          1.000000
50%          1.000000
75%          2.000000
max        327.000000
Name: calculated_host_listings_count, dtype: float64
```

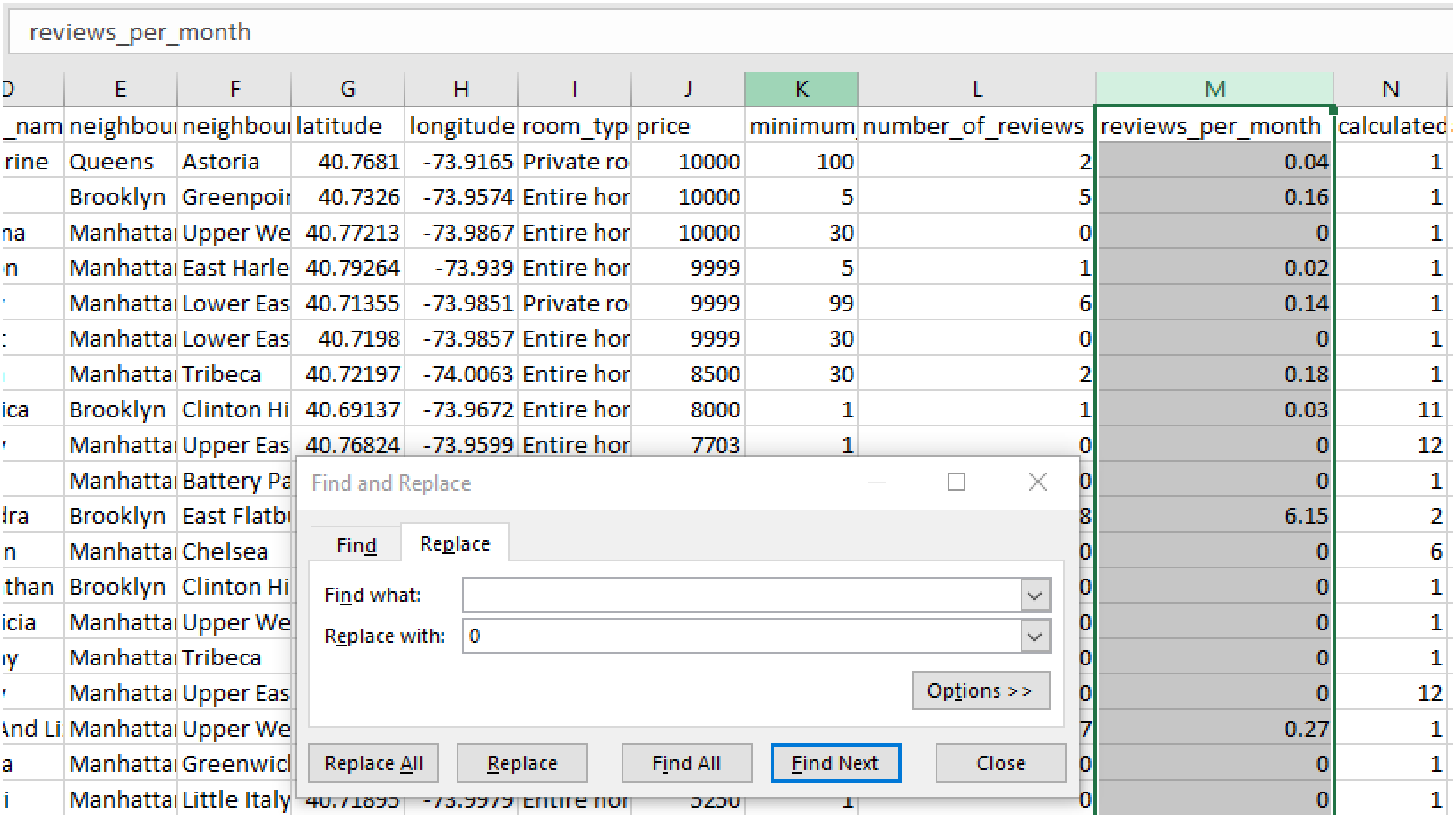
```
In [18]: data.availability_365.describe()

Out[18]: count    48858.000000
         mean       112.801425
         std        131.610962
         min          0.000000
         25%          0.000000
         50%         45.000000
         75%        227.000000
         max        365.000000
         Name: availability_365, dtype: float64
```

No wrong values or major outliers were found

```
In [28]: # Exporting back the new edited file
         data.to_csv(r'C:\Users\Admin\Downloads\airbnb.csv',index=False, header=True)
```

After exporting the data back to csv, some data manipulations were done in MS-Excel such as replacing null values with 0 in “reviews_per_month “column.



reviews_per_month

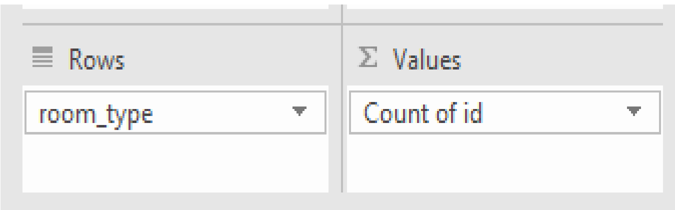
D	E	F	G	H	I	J	K	L	M	N
_nam	neighbou	neighbou	latitude	longitude	room_typ	price	minimum	number_of_reviews	reviews_per_month	calculated
rine	Queens	Astoria	40.7681	-73.9165	Private ro	10000	100	2	0.04	1
	Brooklyn	Greenpoi	40.7326	-73.9574	Entire hor	10000	5	5	0.16	1
na	Manhatta	Upper We	40.77213	-73.9867	Entire hor	10000	30	0	0	1
n	Manhatta	East Harle	40.79264	-73.939	Entire hor	9999	5	1	0.02	1
'	Manhatta	Lower Eas	40.71355	-73.9851	Private ro	9999	99	6	0.14	1
:	Manhatta	Lower Eas	40.7198	-73.9857	Entire hor	9999	30	0	0	1
	Manhatta	Tribeca	40.72197	-74.0063	Entire hor	8500	30	2	0.18	1
ica	Brooklyn	Clinton Hi	40.69137	-73.9672	Entire hor	8000	1	1	0.03	11
'	Manhatta	Upper Eas	40.76824	-73.9599	Entire hor	7703	1	0	0	12
	Manhatta	Battery Pa						0	0	1
dra	Brooklyn	East Flatb						8	6.15	2
n	Manhatta	Chelsea						0	0	6
than	Brooklyn	Clinton Hi						0	0	1
icia	Manhatta	Upper We						0	0	1
ly	Manhatta	Tribeca						0	0	1
'	Manhatta	Upper Eas						0	0	12
And Li	Manhatta	Upper We						7	0.27	1
a	Manhatta	Greenwicl						0	0	1
i	Manhatta	Little Italy	40.71895	-73.9979	Entire hor	5250	1	0	0	1

- DATA VISUALIZATION

1. Room types and their percentage share.

a) A pivot table was created in excel and "room_type" was selected in rows and count of "id" was selected as values.

b) Using this pivot table, a pie chart was created to depict the room types and their shares.

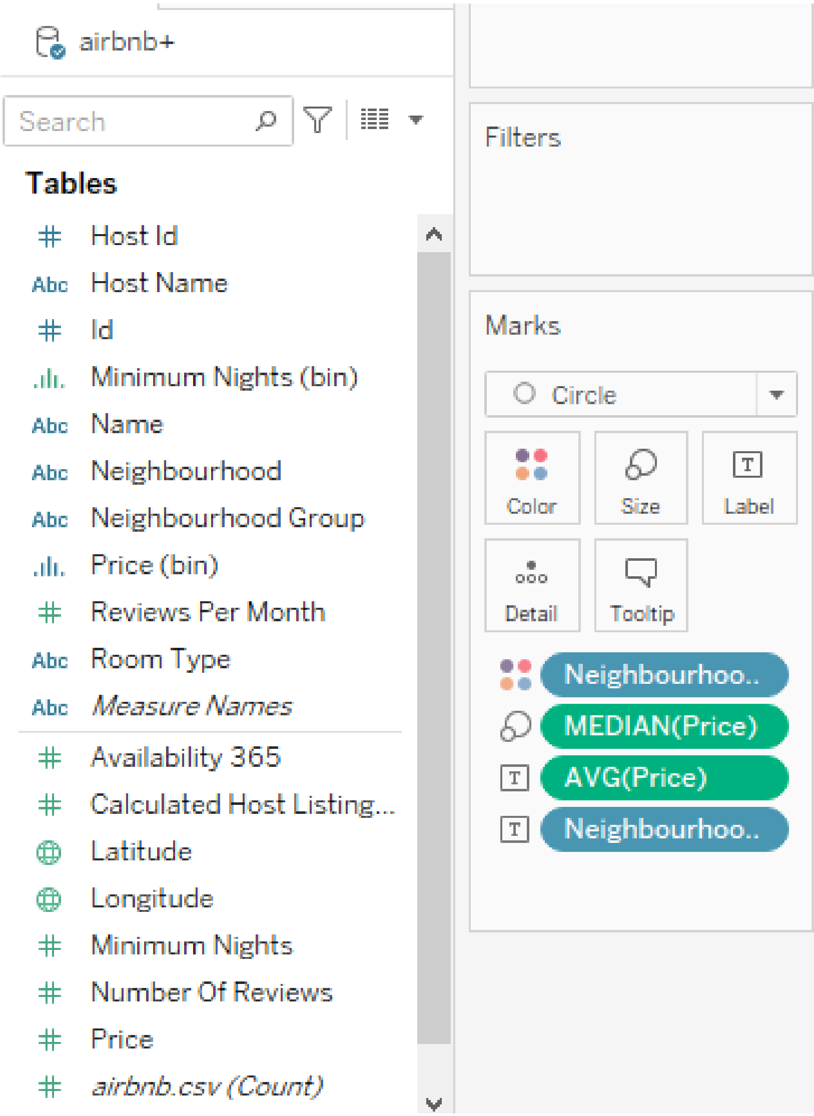


Row Labels	Count of id
Entire home/apt	25393
Private room	22306
Shared room	1159
Grand Total	**48858**

2. Price distribution with respect to neighbourhood group.

a) Using Tableau, a bubble chart was created to visualize the average price of neighbourhood group.

b) The selections made for this chart is attached below.

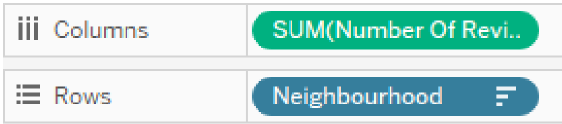


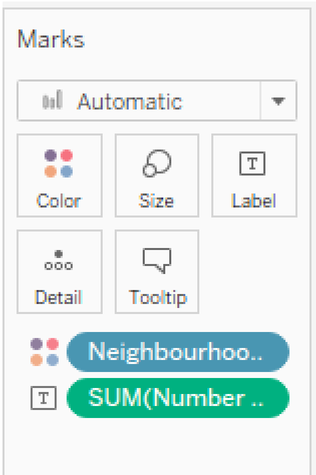
3. Top Neighbourhoods.

a) Using Tableau, a bar chart was created that shows the top neighbourhoods and also which neighbourhood so they belong to.

b) Neighbourhood and sum of reviews was put in row section and column section respectively.

c) Neighbourhood Group was attached to color and total reviews to the label.

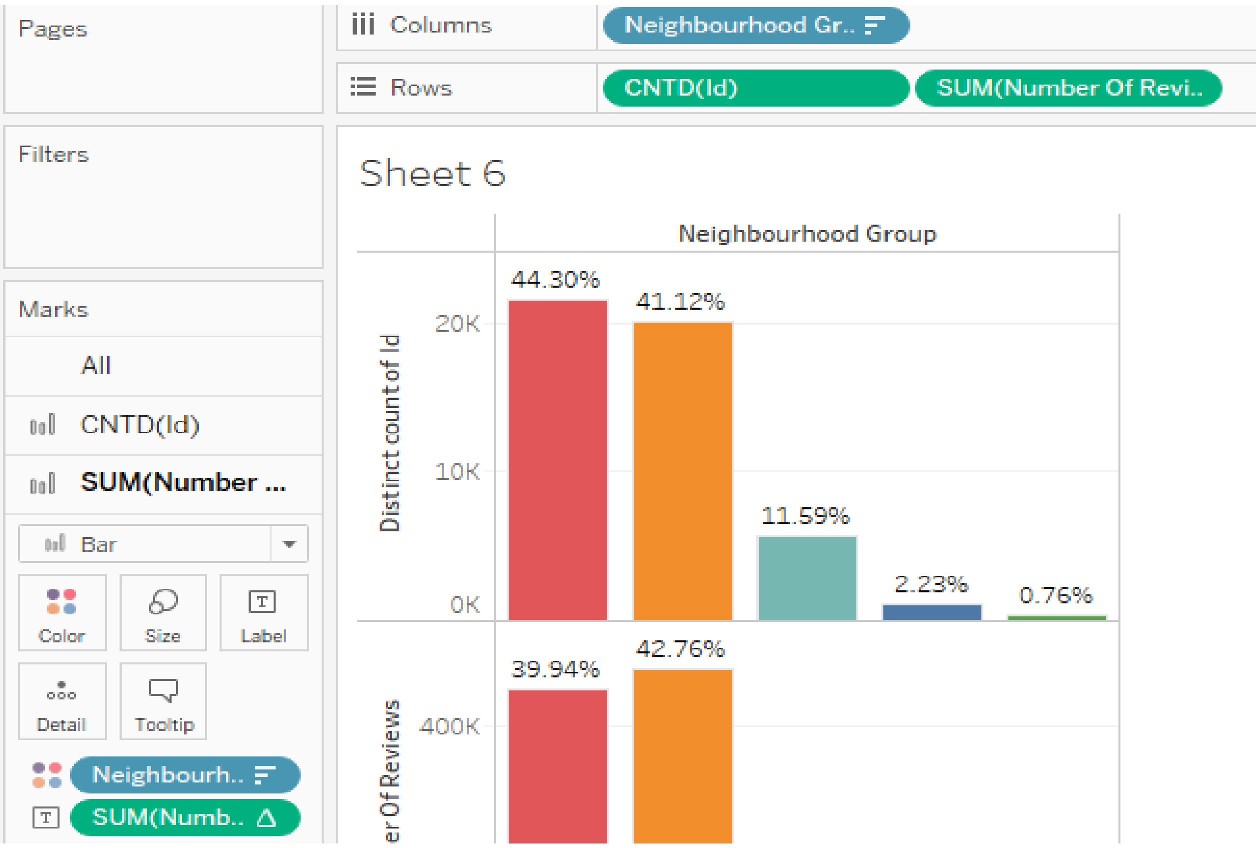




4. Neighbourhood group vs Total reviews & Total share.

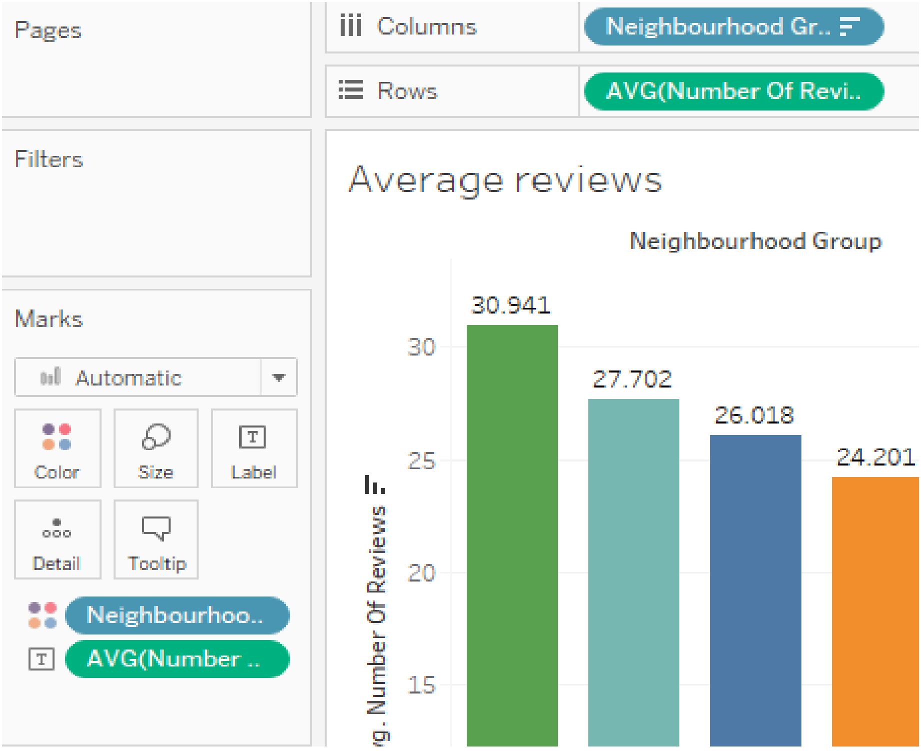
a) Using Tableau, a dual axis chart was prepared to compare the total bookings and total share of properties for the different Neighbourhood groups.

b) The selections made are shown below.



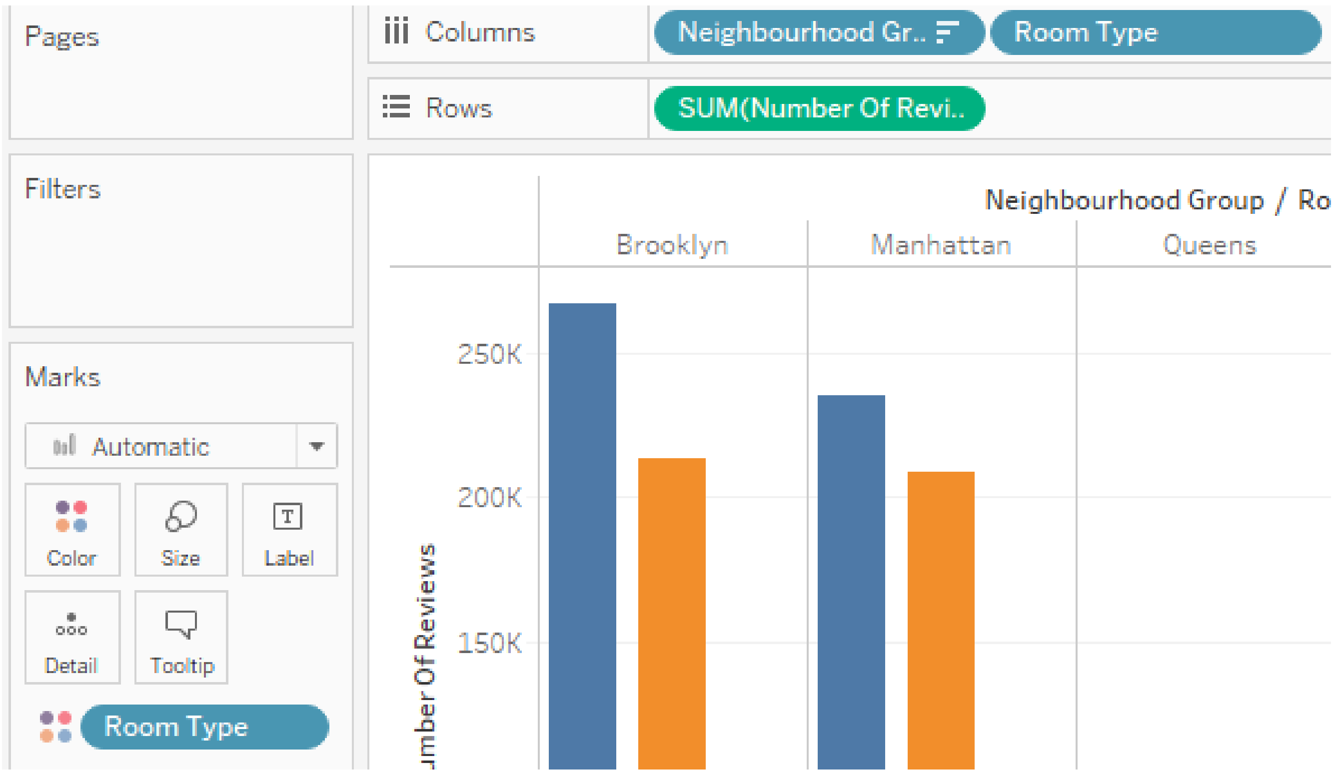
5. Average Reviews of Neighbourhood Group.

a) Using Tableau, a bar chart was created that shows the average ratings of the neighbourhood groups.
b) Average number of ratings and Neighbourhood Group was put in row section and column section respectively.
c) The selections made are attached below.



6. Total Reviews with respect to neighbourhood group and room type.

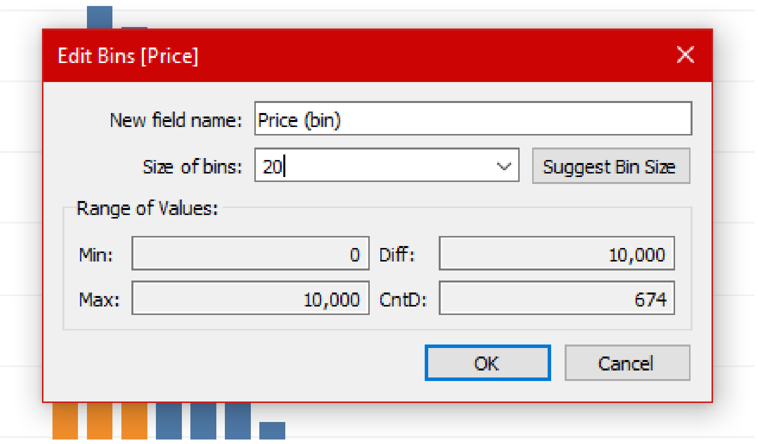
a) The different variables attached to different attributes are attached below.

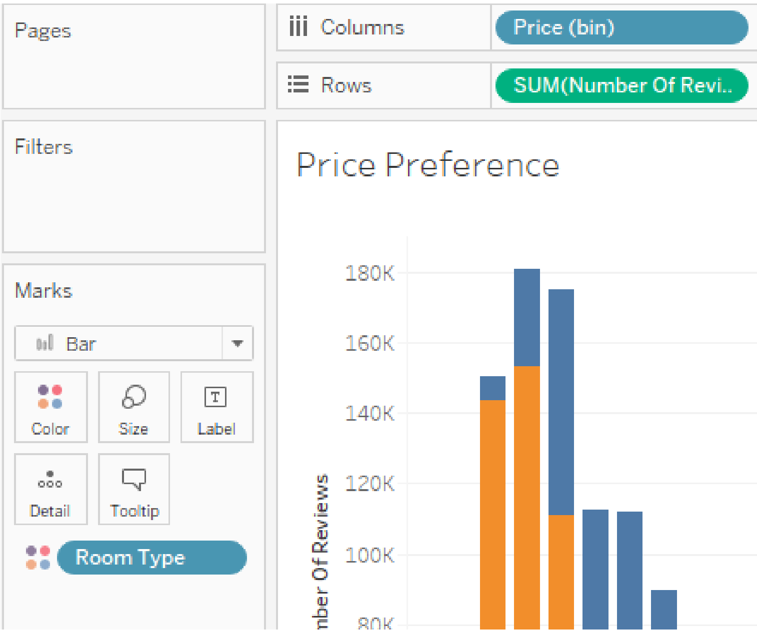


7. Price preference as per Number of Reviews.

a) To plot a bar graph depicting number of reviews according to price ranges a bin was created for variable “price” in tableau.

b) The Price bin was then attached to column and sum of reviews to the rows. Moreover, room type was attached to colours to depict the room type dimension also.

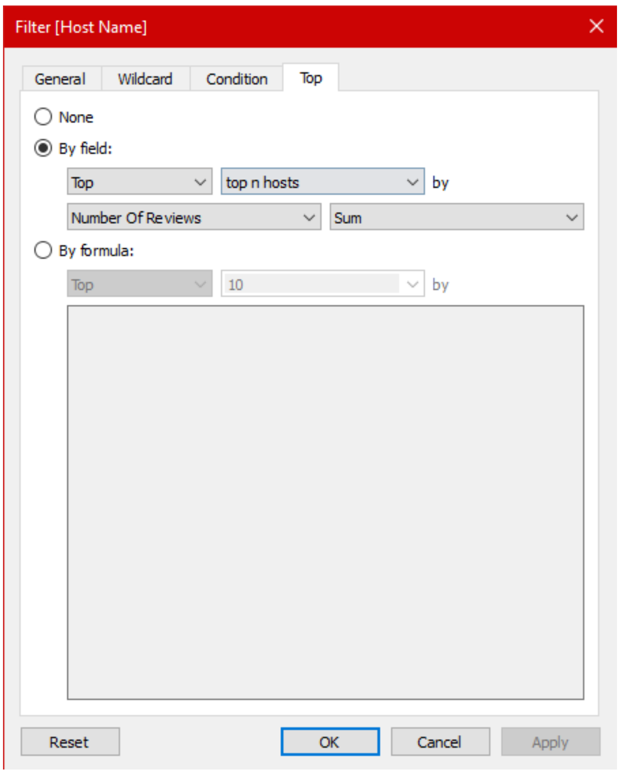


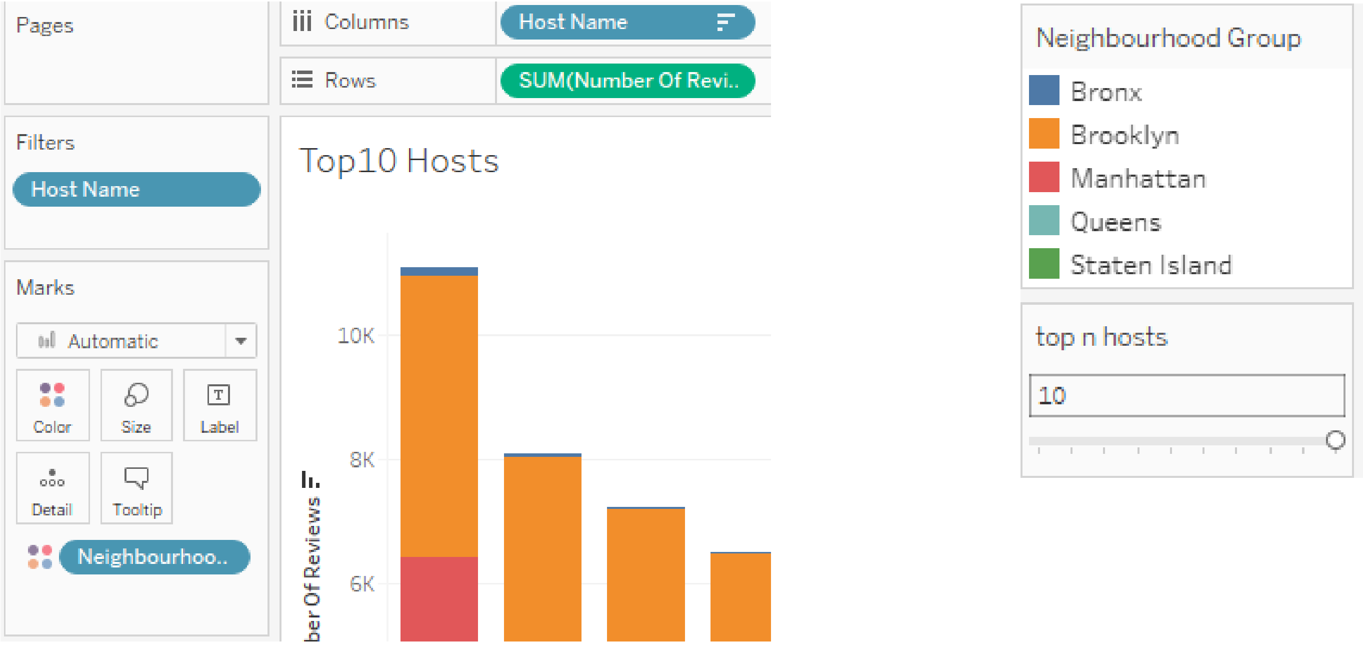


8. Top 10 Hosts

a) To generate the graph for Top N Hosts, a filter was created to show top n hosts.

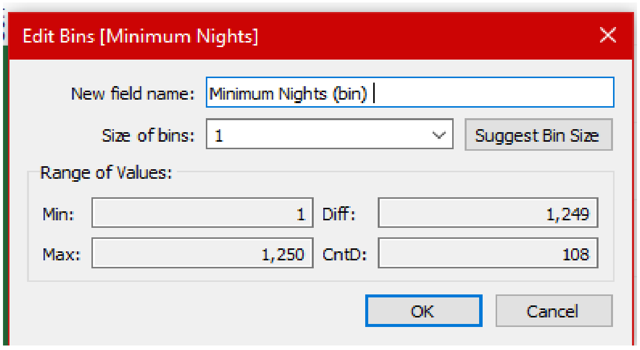
b) And the other selections there were made are attached below as snaps.

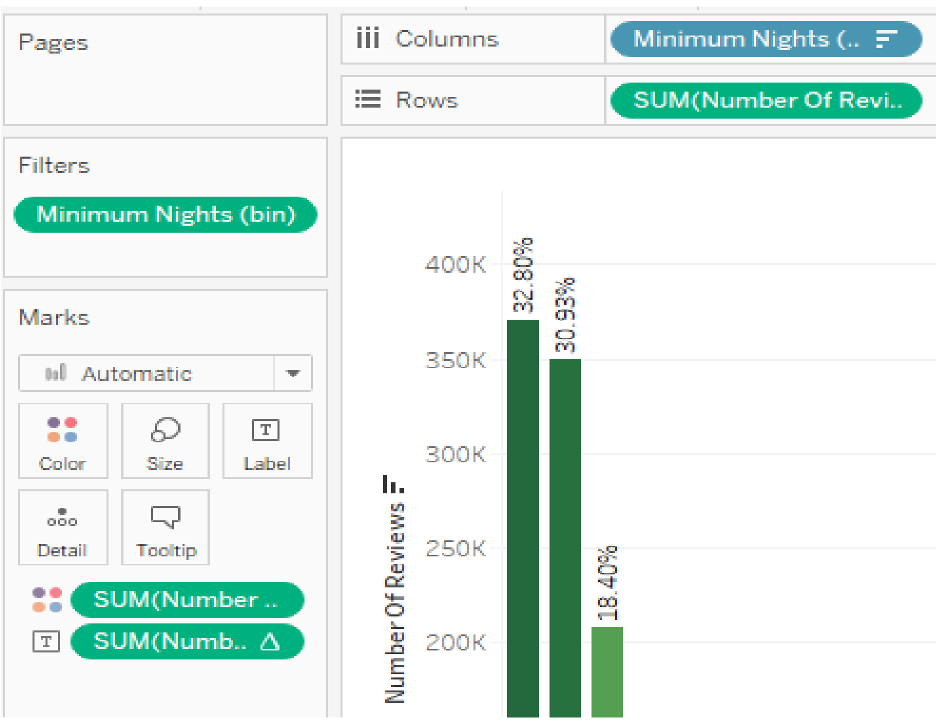




9. Most preferred properties as per minimum nights criteria.

- **a)** To plot a bar graph depicting number of reviews according to minimum night ranges a bin was created for variable “minimum_nights” in tableau.
- **b)** The Minimum Nights bin was then attached to column and sum of reviews to the rows.





10. Heatmap showing price variation wrt to neighbourhood group and room type.