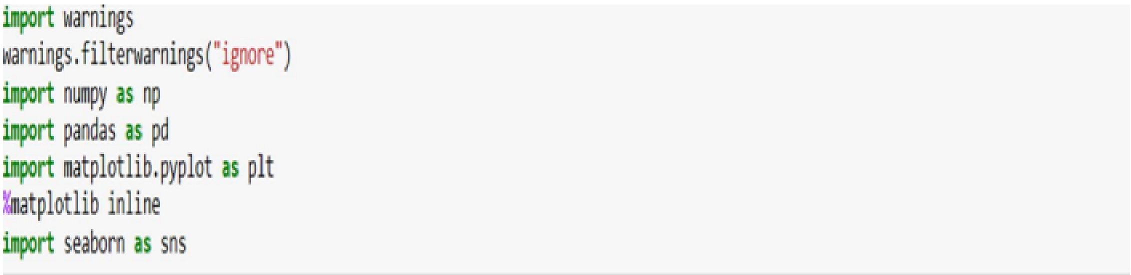
Methodology Document - AIRBNB Case Study

In the case study we have used Jupiter notebook to perform initial analysis of the data and Tableau for data analysis and visualization.

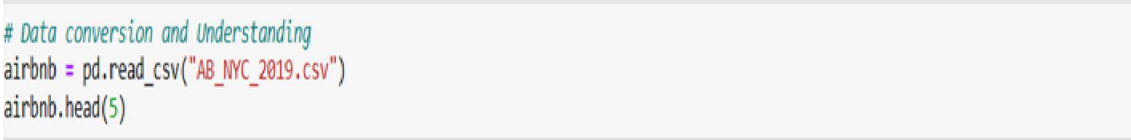
Initial Analysis using Jupiter Notebook: Data Set Used: AB_NYC_2019.csv

Number of Rows: 48895

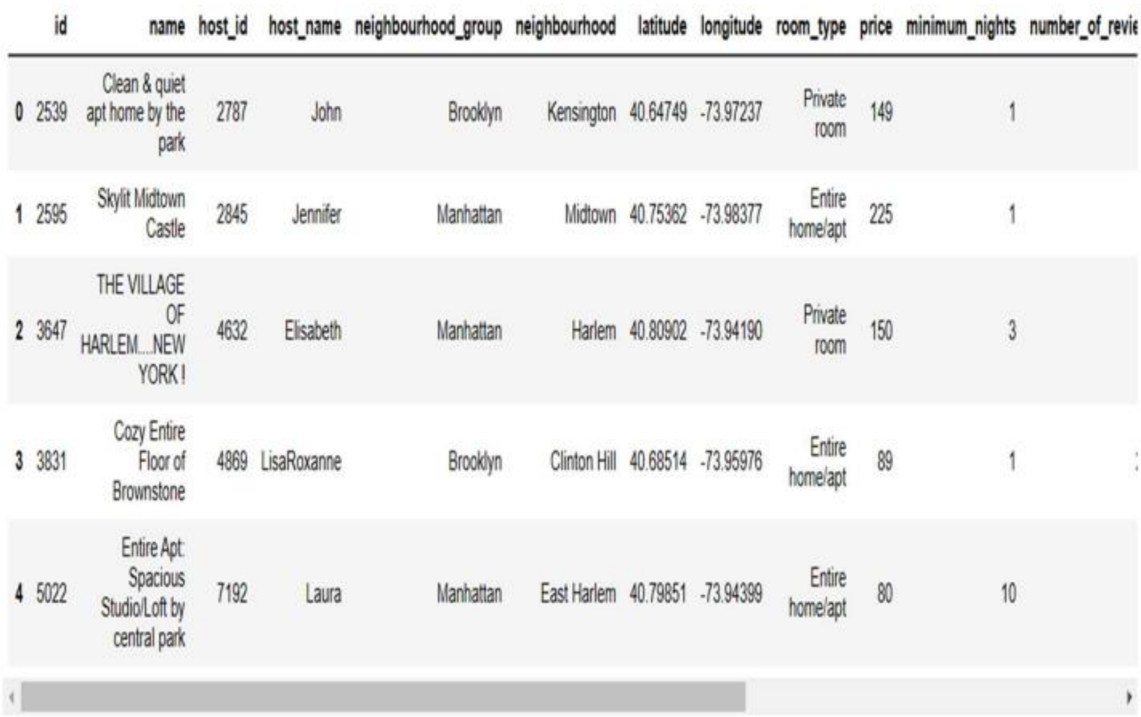
Number of Columns: 16



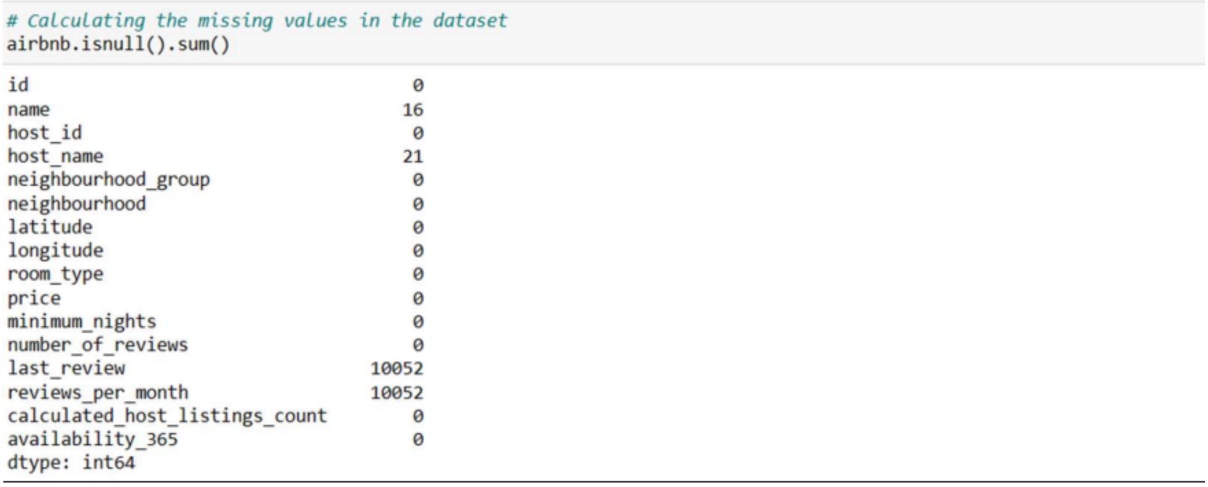
```
import warnings
warnings.filterwarnings("ignore")
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```



```
# Data conversion and Understanding
airbnb = pd.read_csv("AB_NYC_2019.csv")
airbnb.head(5)
```

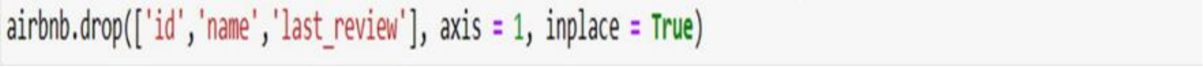


	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_revie
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-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	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

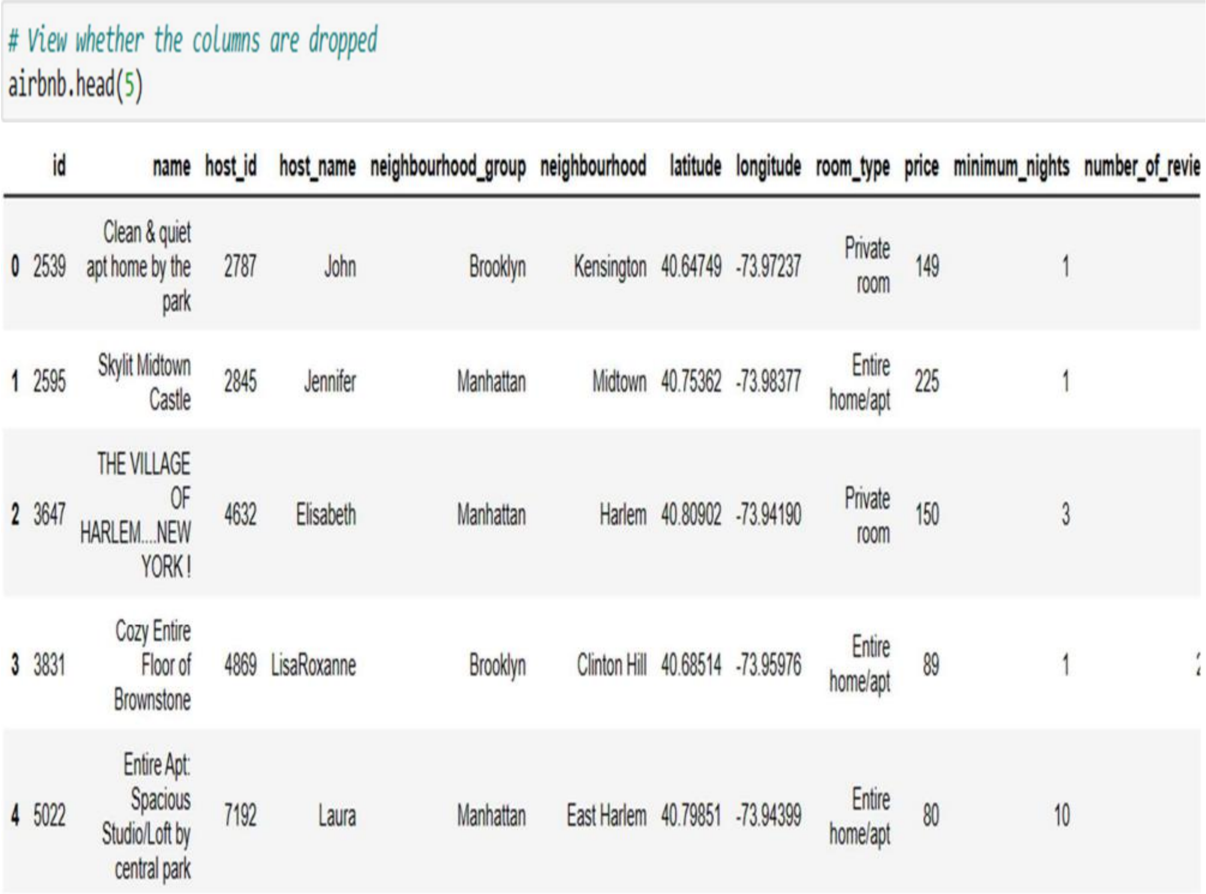


```
# Calculating the missing values in the dataset
airbnb.isnull().sum()
```

```
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
```



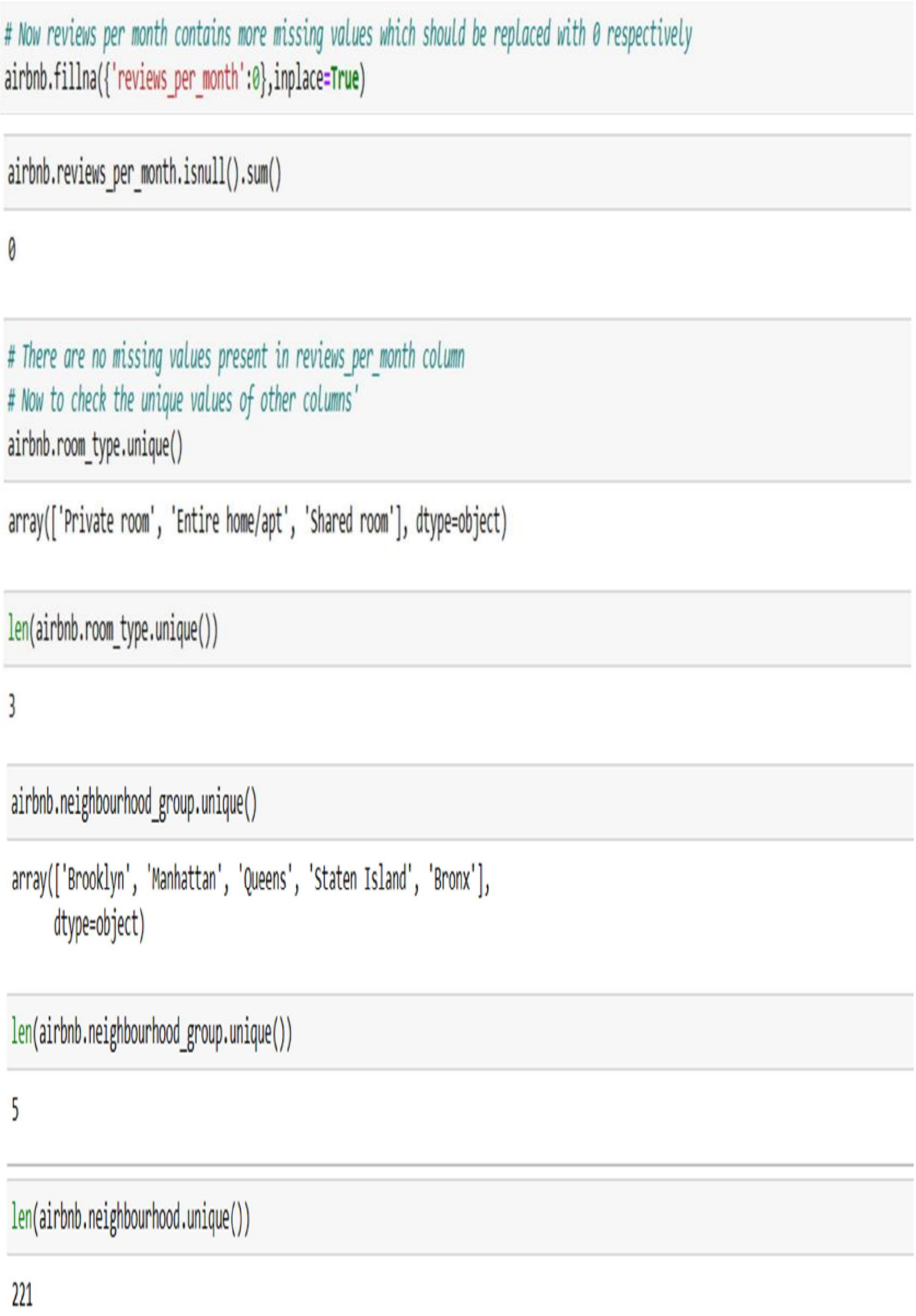
```
airbnb.drop(['id','name','last_review'], axis = 1, inplace = True)
```



```
# View whether the columns are dropped
airbnb.head(5)
```

	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	

- We removed the columns like Id, Name, Last Review which was not giving much information.



```
# Now reviews per month contains more missing values which should be replaced with 0 respectively
airbnb.fillna({'reviews_per_month':0},inplace=True)
```

```
airbnb.reviews_per_month.isnull().sum()
```

```
0
```

```
# There are no missing values present in reviews_per_month column
# Now to check the unique values of other columns'
airbnb.room_type.unique()
```

```
array(['Private room', 'Entire home/apt', 'Shared room'], dtype=object)
```

```
len(airbnb.room_type.unique())
```

```
3
```

```
airbnb.neighbourhood_group.unique()
```

```
array(['Brooklyn', 'Manhattan', 'Queens', 'Staten Island', 'Bronx'],
      dtype=object)
```

```
len(airbnb.neighbourhood_group.unique())
```

```
5
```

```
len(airbnb.neighbourhood.unique())
```

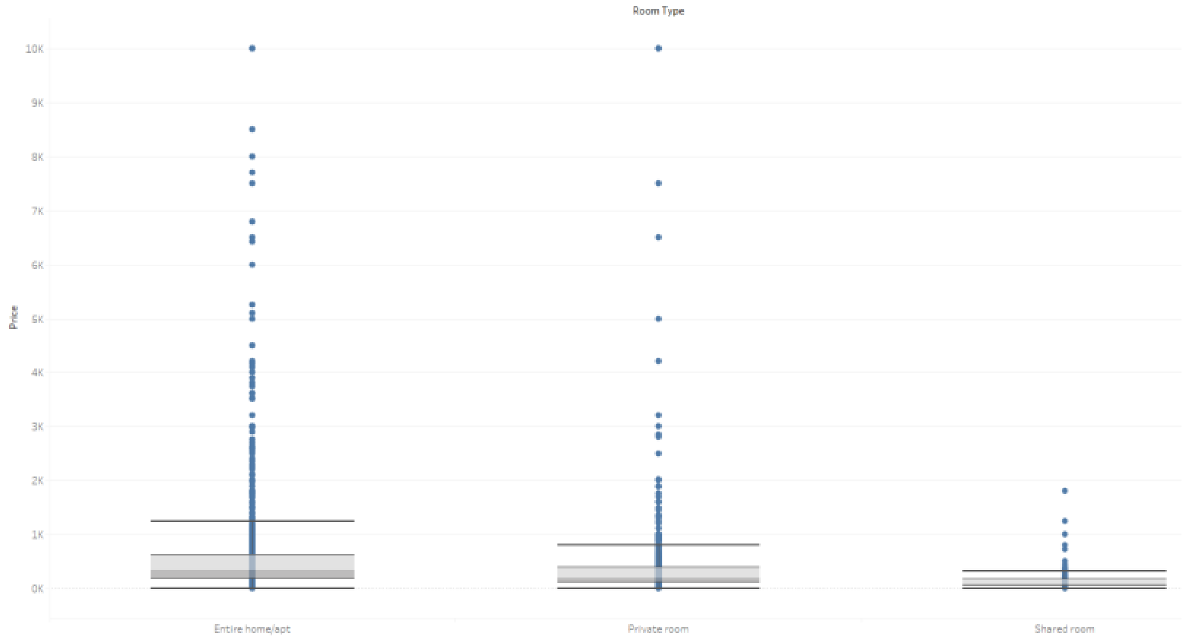
```
221
```

Data Wrangling:

- Did univariate analysis using Tableau on the fields to see their distributions, the unique values in a field, the missing values and to check for outliers if any.

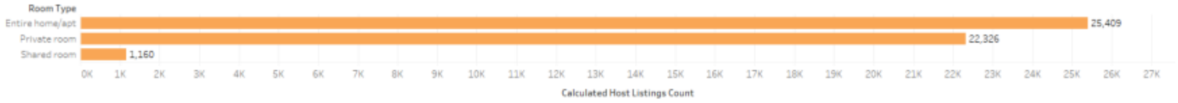
- There was a small proportion of null values which would not affect my analysis so let them stay as it is.

- Price was highly positively skewed so median was very close the lower quartile with some outliers as seen in the boxplot below:



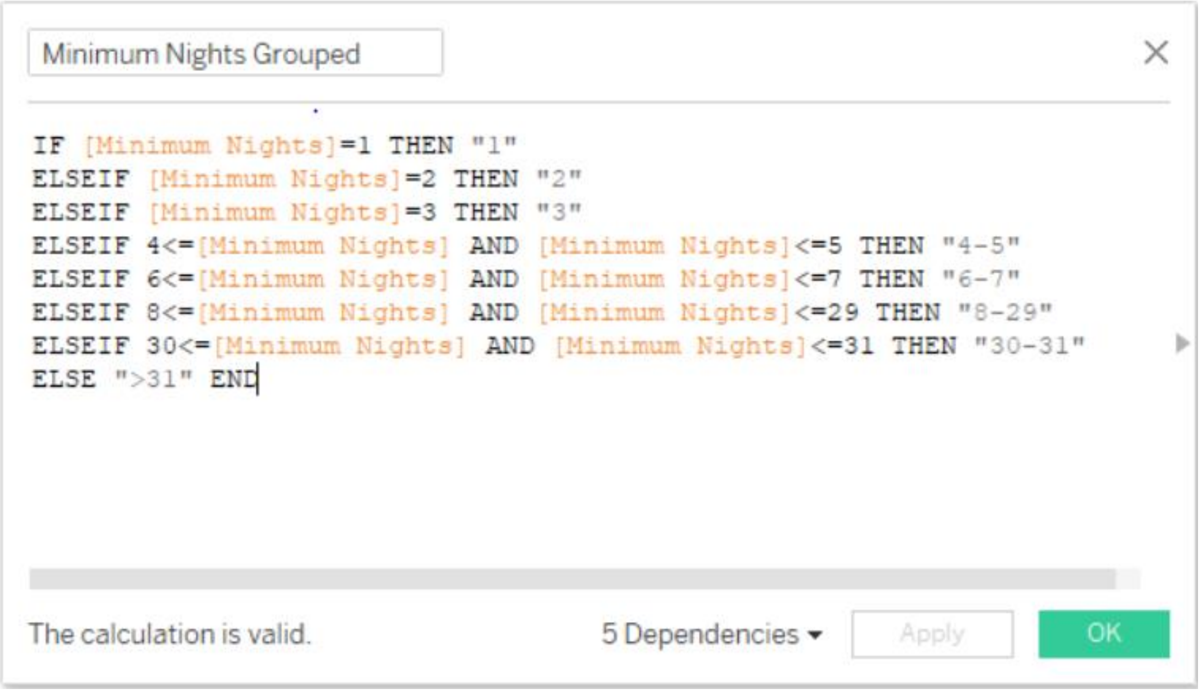
- Since price has outliers, used median instead of mean as the measure for price.

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



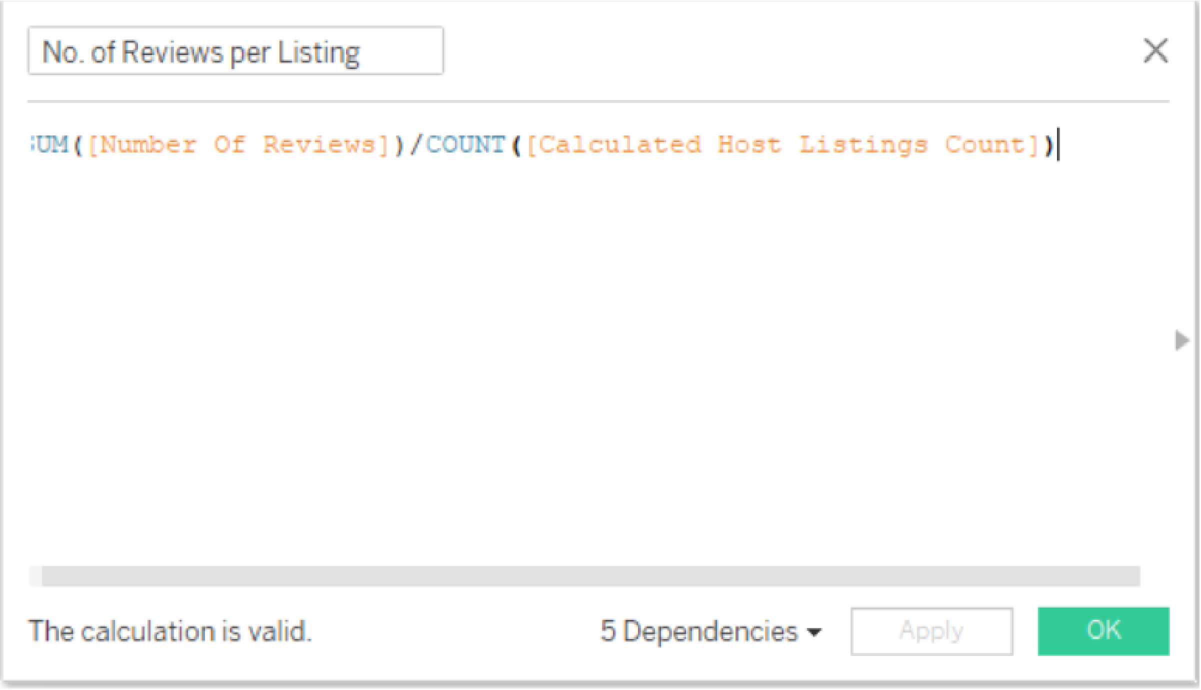
Count of Calculated Host Listings Count for each Room Type. The marks are labeled by count of Calculated Host Listings Count.

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



```
IF [Minimum Nights]=1 THEN "1"
ELSEIF [Minimum Nights]=2 THEN "2"
ELSEIF [Minimum Nights]=3 THEN "3"
ELSEIF 4<=[Minimum Nights] AND [Minimum Nights]<=5 THEN "4-5"
ELSEIF 6<=[Minimum Nights] AND [Minimum Nights]<=7 THEN "6-7"
ELSEIF 8<=[Minimum Nights] AND [Minimum Nights]<=29 THEN "8-29"
ELSEIF 30<=[Minimum Nights] AND [Minimum Nights]<=31 THEN "30-31"
ELSE ">31" END
```

- Created a calculated field of number of reviews per listing.



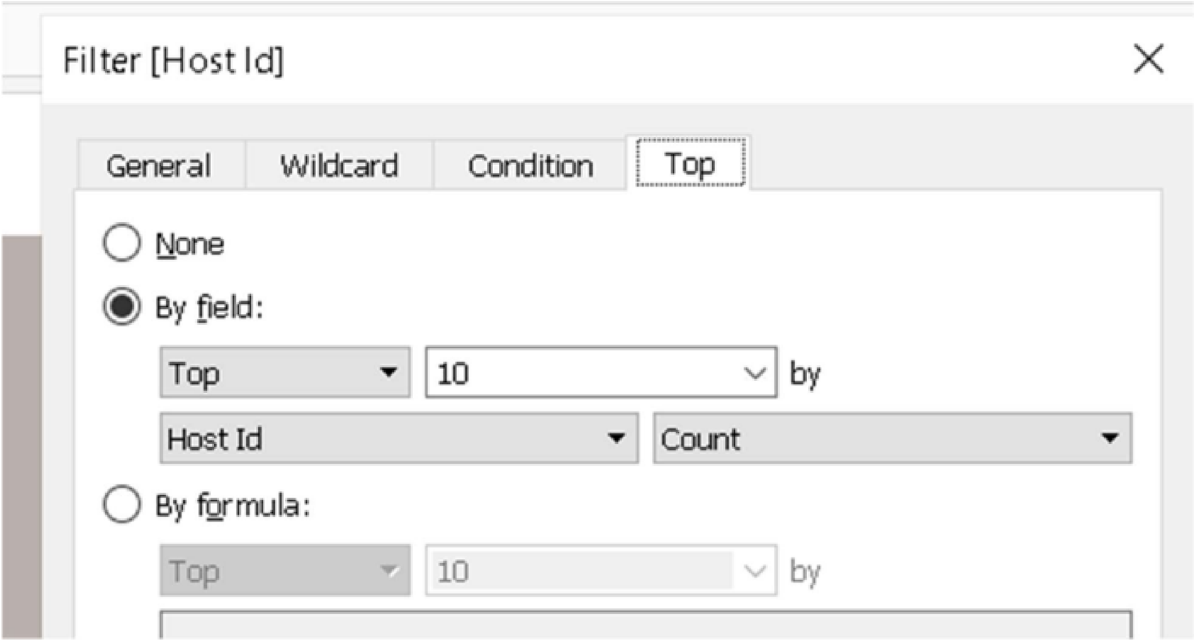
```
:UM([Number Of Reviews])/COUNT([Calculated Host Listings Count])
```

Data Analysis ppt1:

We have used tableau and python to visualize the data. Below are the steps used for the visualisation :-

1. **Top 10 Host:**

- ➢ We identified the top 10 Host Ids, Host Name with count of Host Ids using the tree map.



Filter [Host Id]

○ None

● By field:

Top ▾ 10 ⌄ by

Host Id ▾ Count ▾

○ By formula:

Top ▾ 10 ⌄ by

2. **Average prefer price by people:**

- ➢ We created a bubble chart with Neighbourhood Groups in Columns and Price column in Rows.
- ➢ We added the Neighbourhood Groups to the colours Marks card to highlight the different neighbourhood Groups in different colours. Also Put Avg price in Label.

3. **Types of Properties by Customer Preferences:**

- ➢ We created a pie chart for understanding the percentage of room type preferred w r t neighbourhood group.
- ➢ We added Room Type to the colours Marks card to highlight the different Room Type in different colours and count of Host Id to the size.

4. Most Popular Localities and Properties in New York:

- ➢ We took neighbourhood in rows and sum of reviews in column and took neighbourhood groups in colour.
- ➢ We used filter to show Top 10 neighbours as per the sum of reviews.

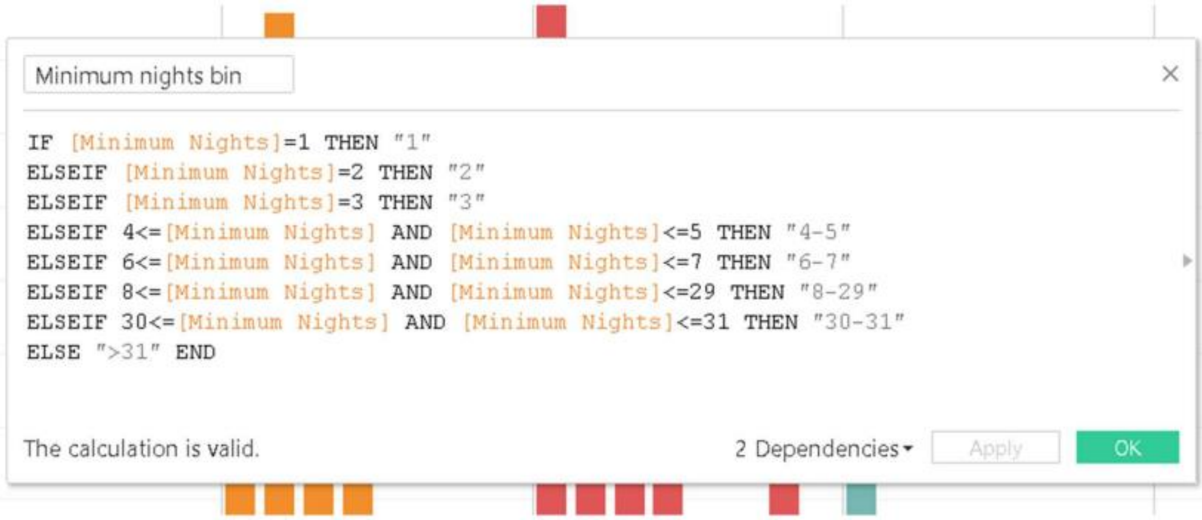
Data Analysis ppt2:

1. **Room type with respect to Neighbourhood group:**

   - We created a pie chart for understanding the percentage of room type preferred w r t neighbourhood group • We added Room Type to the colours Marks card to highlight the different Room Type in different colours and count of Host Id to the size.

2. **Customer Booking with respect to minimum nights:**

   - We created the bin for Minimum nights as shown below:



Minimum nights bin

```
IF [Minimum Nights]=1 THEN "1"
ELSEIF [Minimum Nights]=2 THEN "2"
ELSEIF [Minimum Nights]=3 THEN "3"
ELSEIF 4<=[Minimum Nights] AND [Minimum Nights]<=5 THEN "4-5"
ELSEIF 6<=[Minimum Nights] AND [Minimum Nights]<=7 THEN "6-7"
ELSEIF 8<=[Minimum Nights] AND [Minimum Nights]<=29 THEN "8-29"
ELSEIF 30<=[Minimum Nights] AND [Minimum Nights]<=31 THEN "30-31"
ELSE ">31" END
```

The calculation is valid.

3. **Host Listings and cheap negotiations on availability:**

   - We created a dual axis chart using bar chart for availability 365 and line chart for price for top 10 neighbourhood group sorted by price.

4. **Price range preferred by Customers:**

   - We have taken pricing preference based on volume of bookings done in a price range and no of Ids to create a bar chart. We have created bin for Price column with interval of $20.

5. **Neighborhood variation with respect to Geography:**

   - We used Geo location chart to plot neighbourhood, neighbourhood Group in map to show case the variation of prices across.