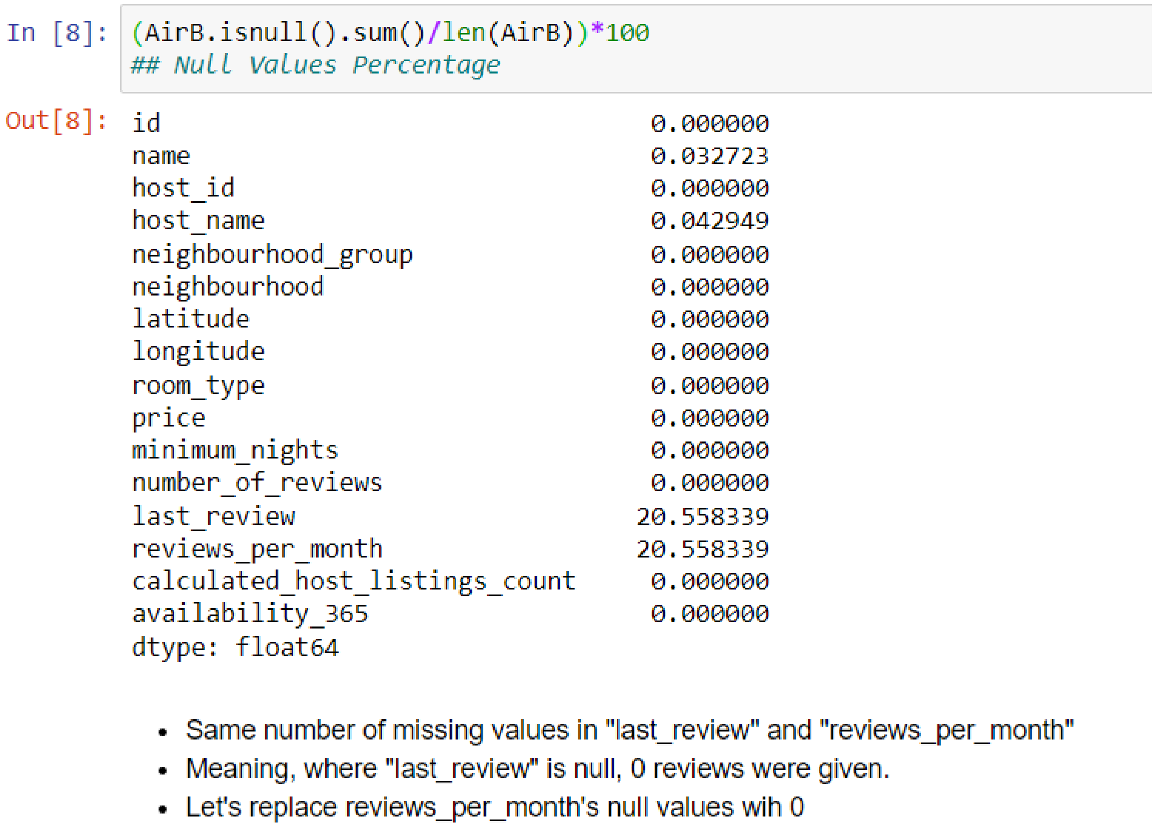
Data Methodology 1:

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.

- Loaded the provided dataset into pandas and tried to understand the variables present.
- Analysed each attribute and checked the data type of each column.’
- Then calculated the null values in each column:



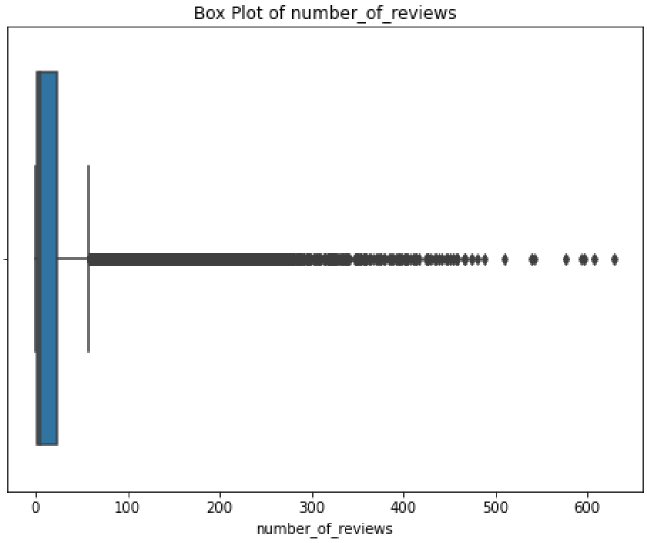
```
In [8]: (AirB.isnull().sum()/len(AirB))*100
        ## Null Values Percentage
```

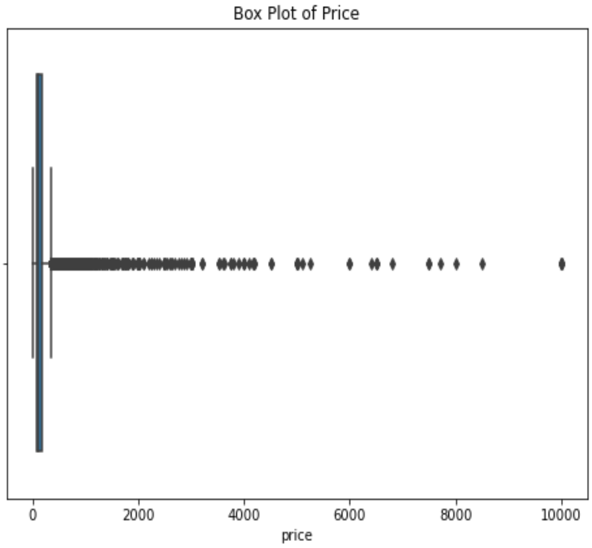
```
Out[8]: id                                 0.000000
        name                               0.032723
        host_id                            0.000000
        host_name                          0.042949
        neighbourhood_group                0.000000
        neighbourhood                      0.000000
        latitude                           0.000000
        longitude                          0.000000
        room_type                          0.000000
        price                              0.000000
        minimum_nights                     0.000000
        number_of_reviews                  0.000000
        last_review                       20.558339
        reviews_per_month                 20.558339
        calculated_host_listings_count     0.000000
        availability_365                   0.000000
        dtype: float64
```

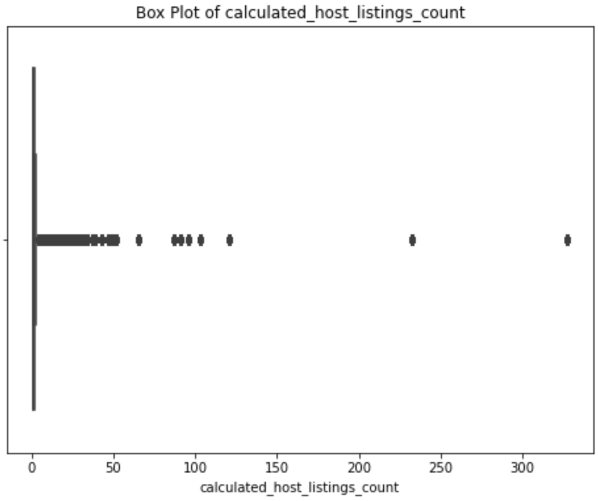
- Same number of missing values in "last_review" and "reviews_per_month"
- Meaning, where "last_review" is null, 0 reviews were given.
- Let's replace reviews_per_month's null values wih 0

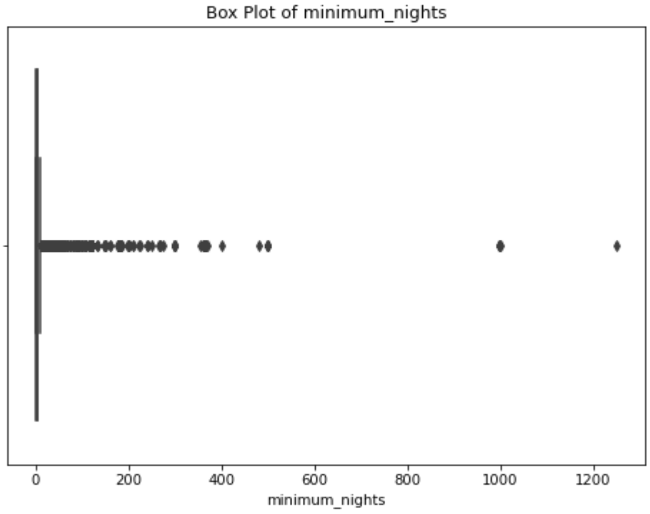
Thus, replaced the ‘reviews_per_month’ column with ‘0’.

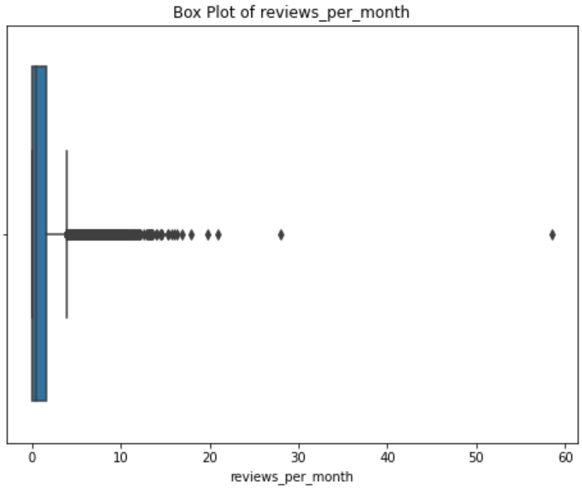
- Checked the spread of the numerical variables using Box Plot:

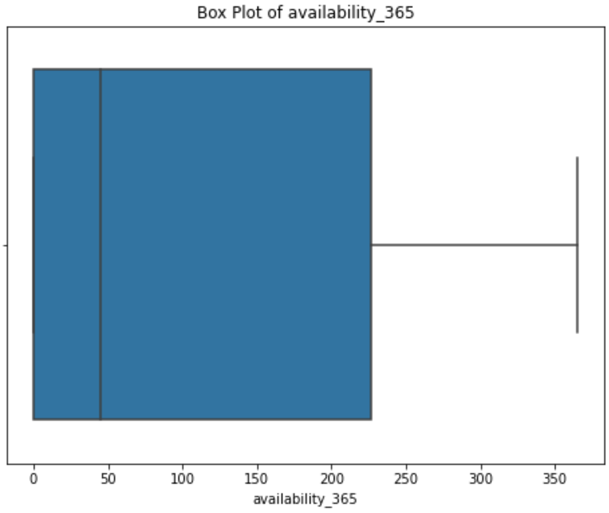






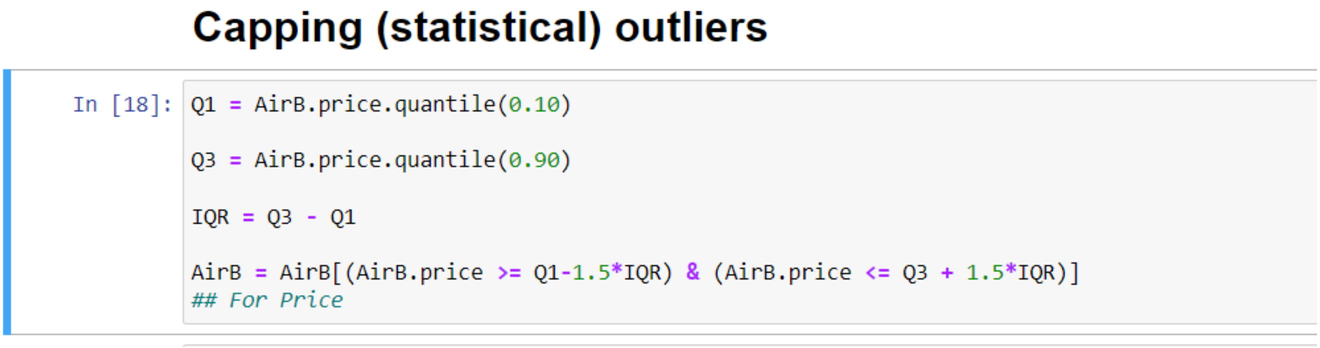






We observe in the above given box plots, that there are a lot of outliers. These can massively skew the data.

- Thus, we got rid of a few outliers using the IQR approach:

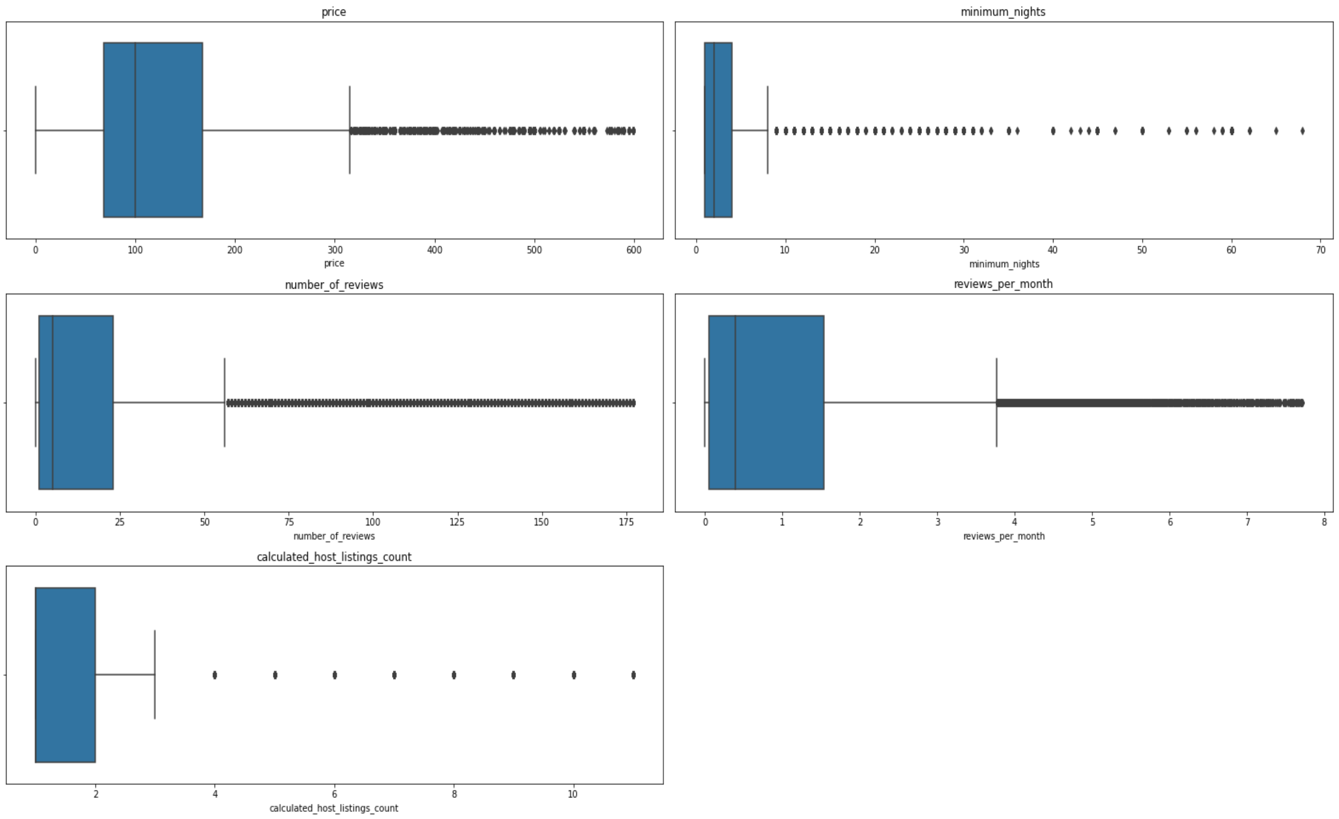


Capping (statistical) outliers

```
In [18]: Q1 = AirB.price.quantile(0.10)

         Q3 = AirB.price.quantile(0.90)

         IQR = Q3 - Q1

         AirB = AirB[(AirB.price >= Q1-1.5*IQR) & (AirB.price <= Q3 + 1.5*IQR)]
         ## For Price
```

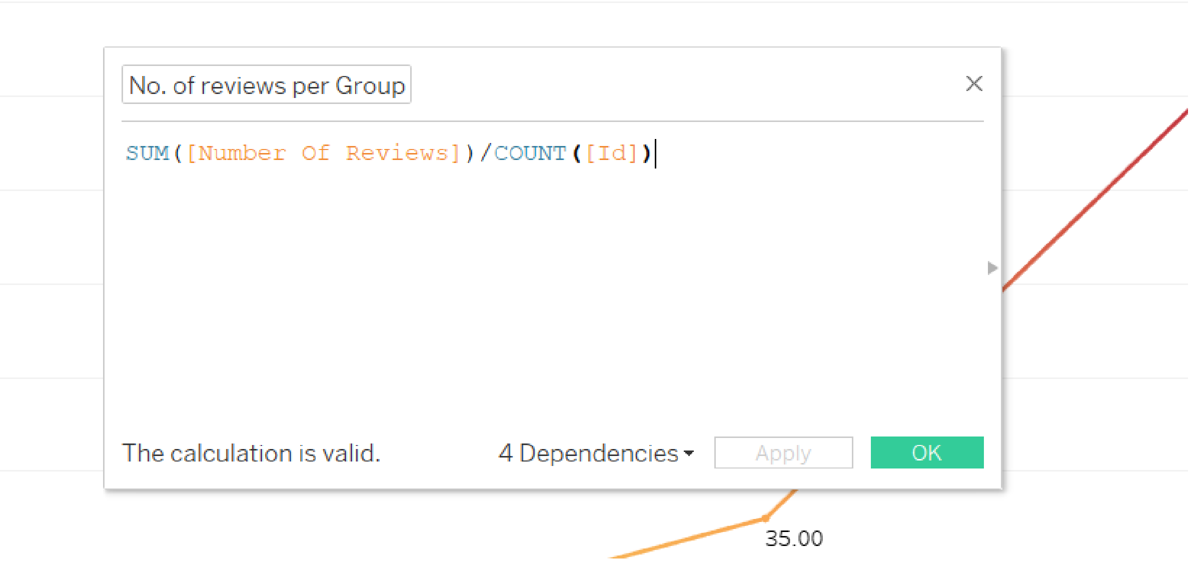
This method was done for all numerical columns (except “availability_365” since it did not have significant outliers).

- Now, boxplot were plotted again to see the difference:



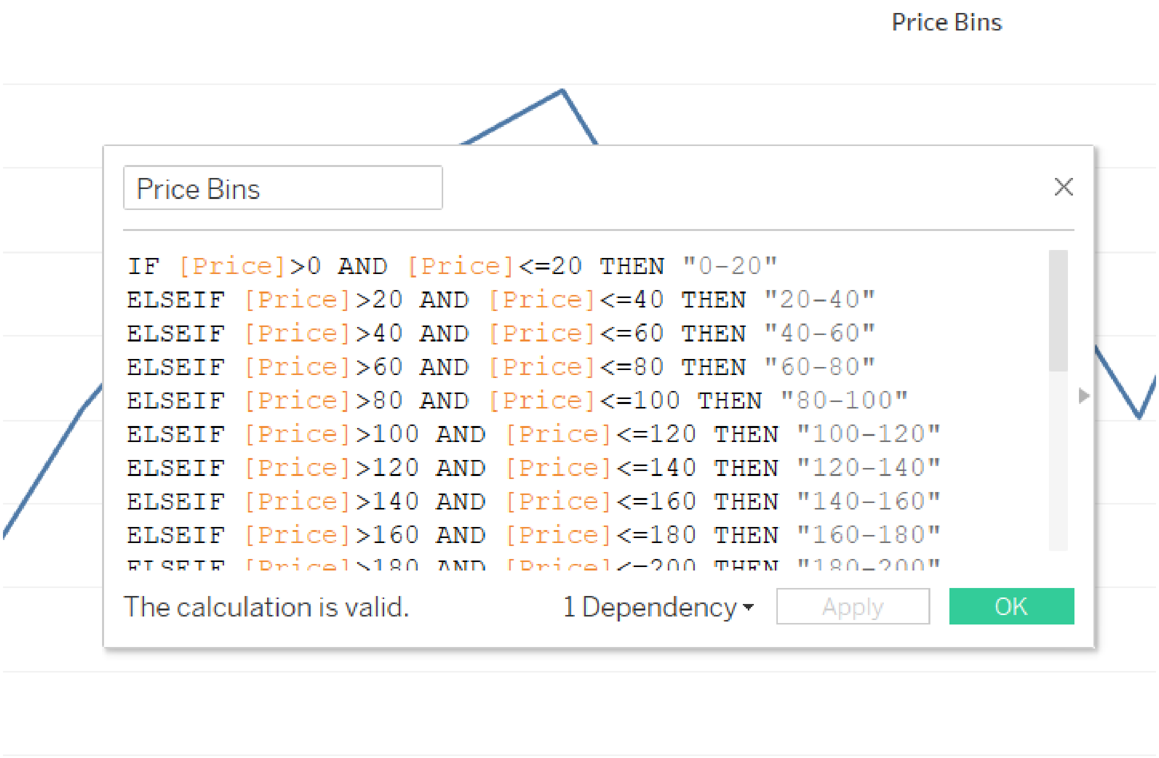
We see a significant difference in the Box Plots now.

- Loaded the data in Tableau for visualization.
- Created a calculated field for **Average number of reviews**:



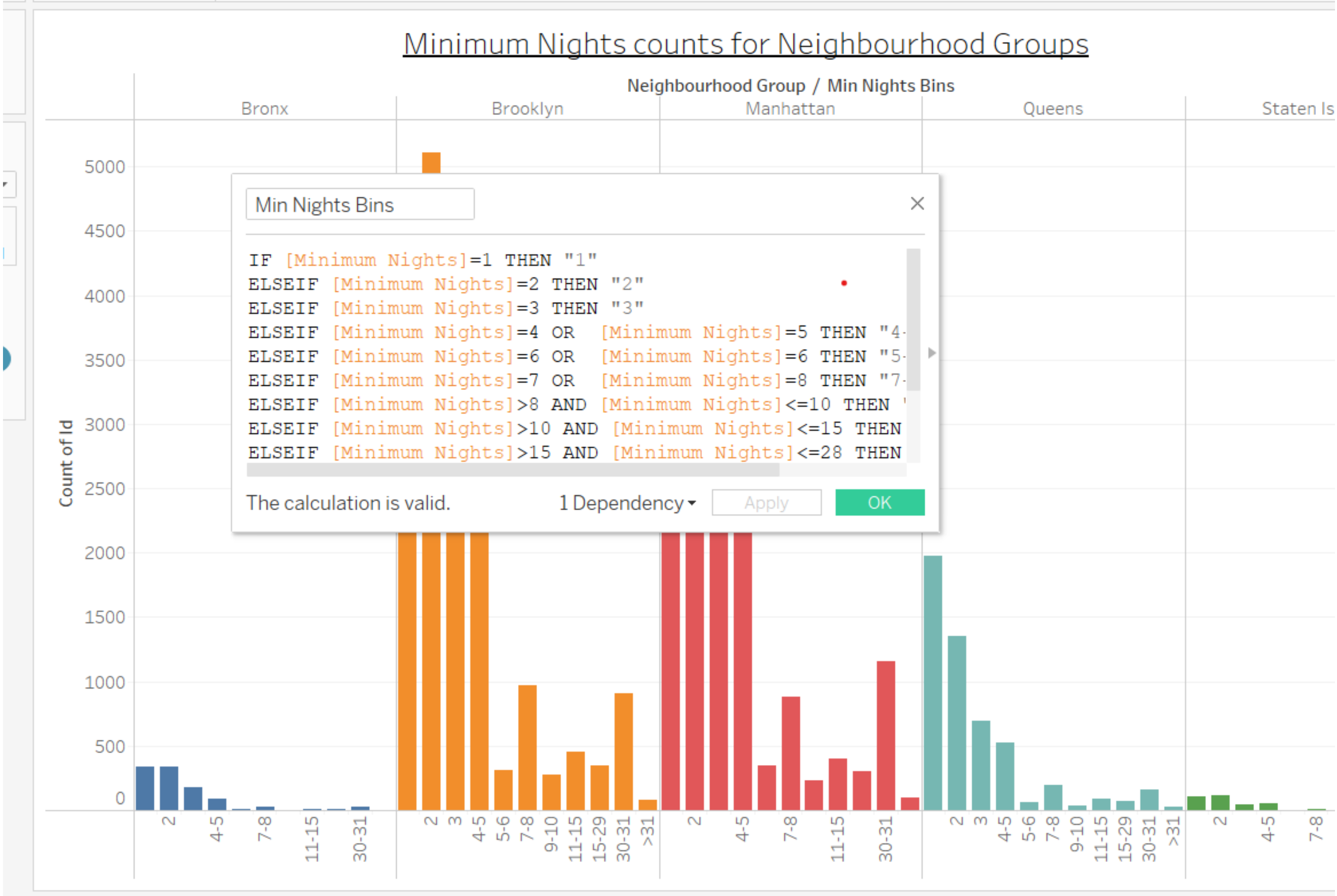
```
SUM([Number Of Reviews])/COUNT([Id])
```

- Created bins for price range:



```
IF [Price]>0 AND [Price]<=20 THEN "0-20"
ELSEIF [Price]>20 AND [Price]<=40 THEN "20-40"
ELSEIF [Price]>40 AND [Price]<=60 THEN "40-60"
ELSEIF [Price]>60 AND [Price]<=80 THEN "60-80"
ELSEIF [Price]>80 AND [Price]<=100 THEN "80-100"
ELSEIF [Price]>100 AND [Price]<=120 THEN "100-120"
ELSEIF [Price]>120 AND [Price]<=140 THEN "120-140"
ELSEIF [Price]>140 AND [Price]<=160 THEN "140-160"
ELSEIF [Price]>160 AND [Price]<=180 THEN "160-180"
ELSEIF [Price]>180 AND [Price]<=200 THEN "180-200"
```

- Created Bins for Minimum number of nights offered:

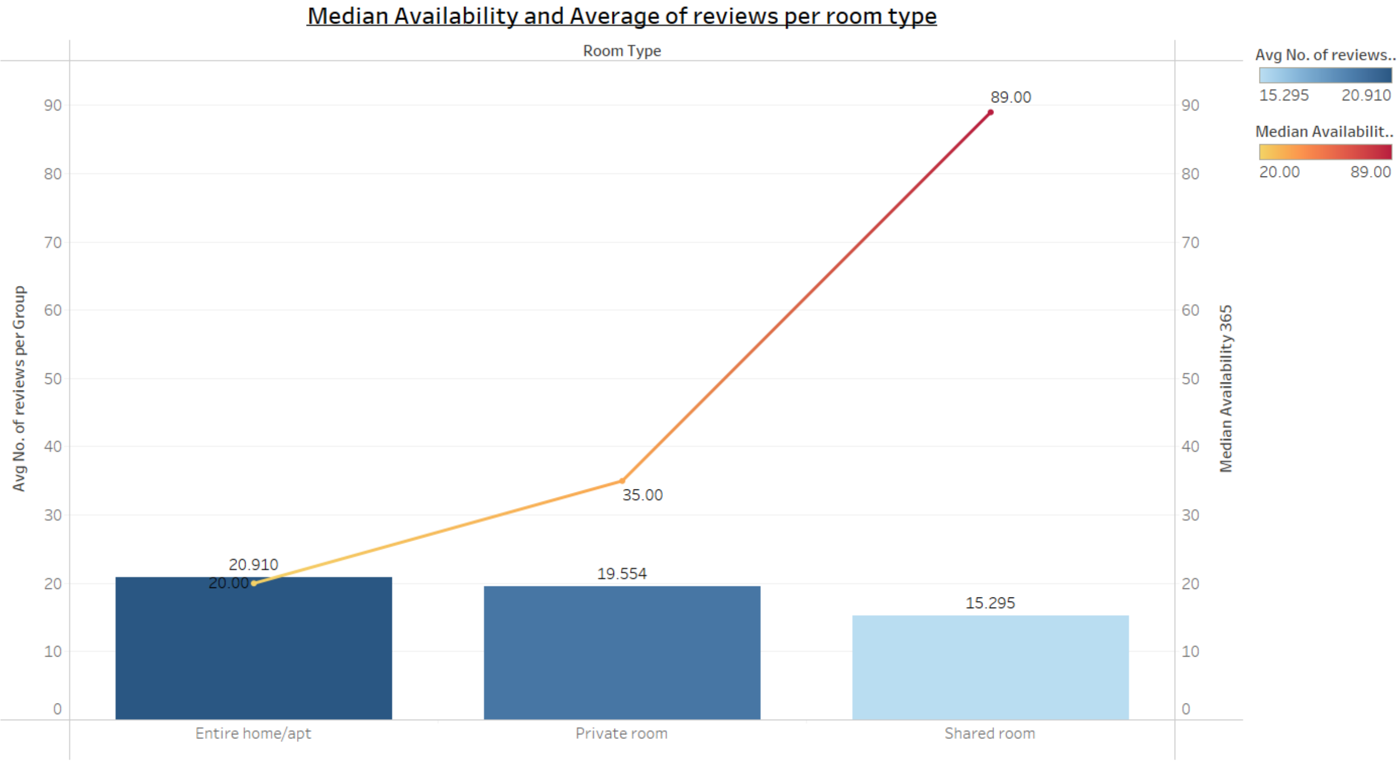


```
IF [Minimum Nights]=1 THEN "1"
ELSEIF [Minimum Nights]=2 THEN "2"
ELSEIF [Minimum Nights]=3 THEN "3"
ELSEIF [Minimum Nights]=4 OR  [Minimum Nights]=5 THEN "4-
ELSEIF [Minimum Nights]=6 OR  [Minimum Nights]=6 THEN "5-
ELSEIF [Minimum Nights]=7 OR  [Minimum Nights]=8 THEN "7-
ELSEIF [Minimum Nights]>8 AND [Minimum Nights]<=10 THEN '
ELSEIF [Minimum Nights]>10 AND [Minimum Nights]<=15 THEN
ELSEIF [Minimum Nights]>15 AND [Minimum Nights]<=28 THEN
```

The calculation is valid.

Step 3: Data Analysis.

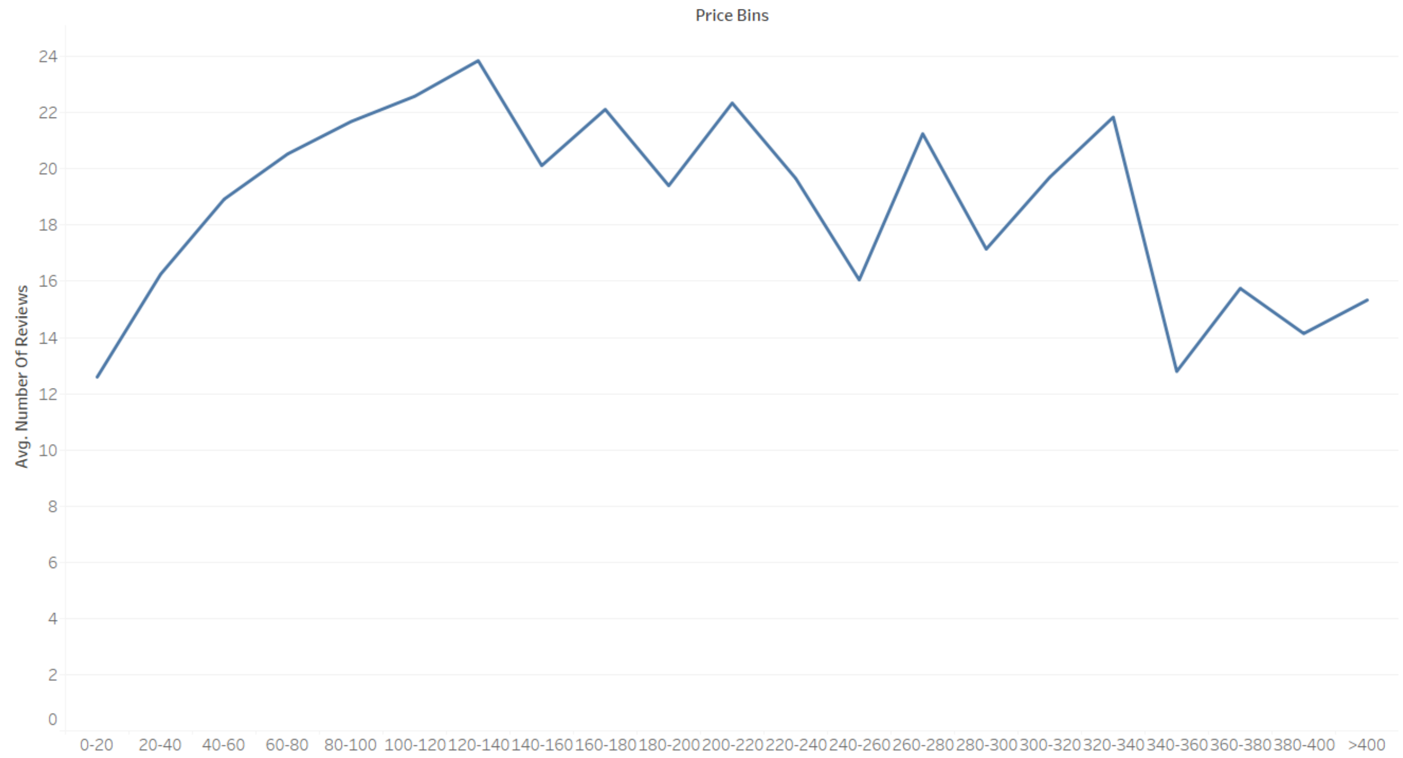
- Compared the average number of reviews (popularity measure) with the median of number of available days in a year for different room types:



The trends of Avg No. of reviews per Group and median of Availability 365 for Room Type. For pane Median of Availability 365: Color shows median of Availability 365. The marks are labeled by median of Availability 365. For pane Avg No. of reviews per Group: Color shows Avg No. of reviews per Group. The marks are labeled by Avg No. of reviews per Group.

- Checked the trend of average of reviews w.r.t increase in the price range:

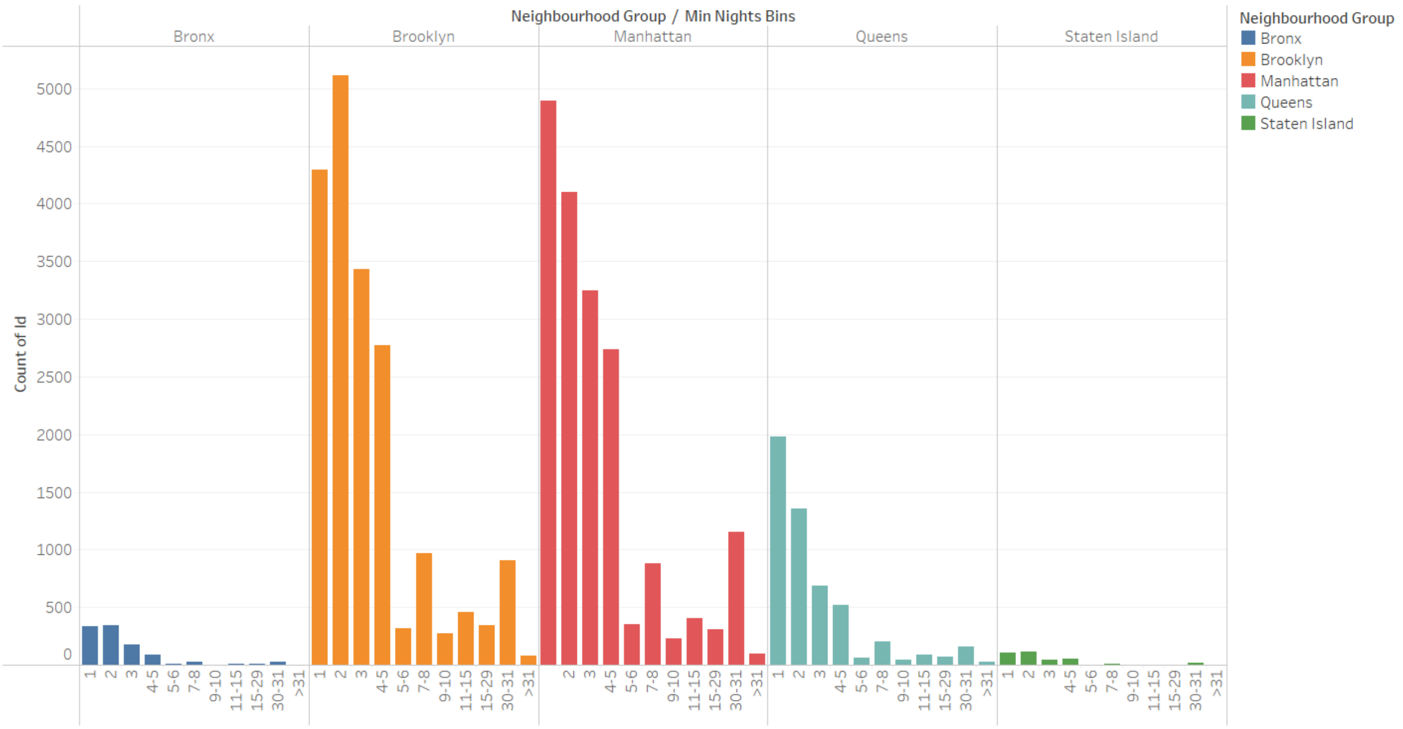
Avg No. of Reviews for Price Range



The trend of average of Number Of Reviews for Price Bins.

- Compared the trend for number of room bookings w.r.t number of minimum nights stay offered in each neighbourhood:

Minimum Nights counts for Neighbourhood Groups



Count of Id for each Min Nights Bins broken down by Neighbourhood Group. Color shows details about Neighbourhood Group.

Step 4: Presentation.

- Made the presentation using the above given insights and visualization using the pyramid principle and keeping the best business practices in mind.