CAPSTONE PROJECT

Exploratory data analysis
on
Airbnb Bookings Analysis

Points for Discussion

- Data summary
- Data Cleaning (Checking Null Values)
- Exploratory Data Analysis
- Conclusion

Data summary

Problem statement: Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Explore and analyze the data to discover the key factors responsible for bookings.

Data set name : Airbnb Nyc 2019.csv (This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.)

Data set shape : The Dataset contains 48895 rows and 16 columns.Out of 16 columns, we have 6 categorical column and rest numerical column.

```
df.dtypes
```

```
id                                  int64
name                               object
host_id                             int64
host_name                          object
neighbourhood_group                object
neighbourhood                      object
latitude                          float64
longitude                         float64
room_type                          object
price                               int64
minimum_nights                      int64
number_of_reviews                   int64
last_review                        object
reviews_per_month                 float64
calculated_host_listings_count      int64
availability_365                    int64
dtype: object
```

Data Cleaning

- Null Values before cleaning

```
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
```

- After cleaning

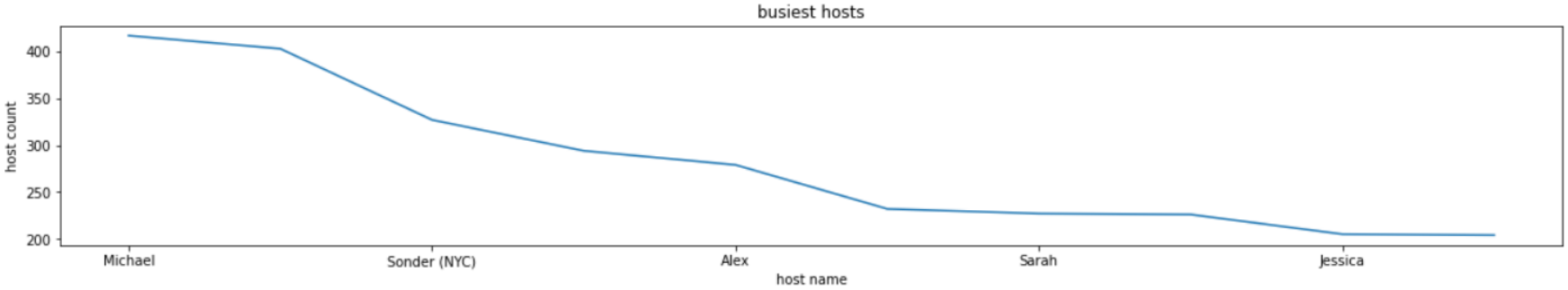
```
id                                  0
name                                0
host_id                             0
host_name                           0
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
last_review                         0
reviews_per_month                   0
calculated_host_listings_count      0
availability_365                    0
dtype: int64
```

Cont

- Missing values can be either replaced or that respective column is dropped.
- Dropping of column only takes place when null values are more than 50%.
- Null values can be replaced with mean,median and mode.

Hosts with maximum number of entries

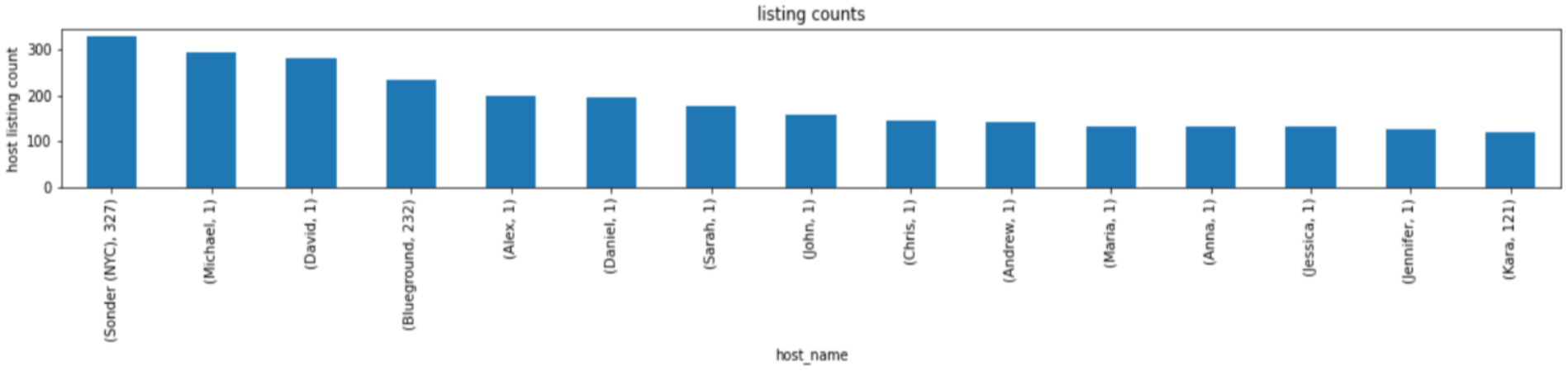
Top host with maximum number of entries in Airbnb Bookings the data set.



The Airbnb Bookings data set have different host entries along with their names and related data. The line plot visualization gives the top 10 busiest hosts in the Airbnb bookings data set. According to the analysis, host Micheal has the maximum value count with 417 entries/rows in data set followed by David, Sonder, John, Alex so on.

Calculated host listings count vs host name

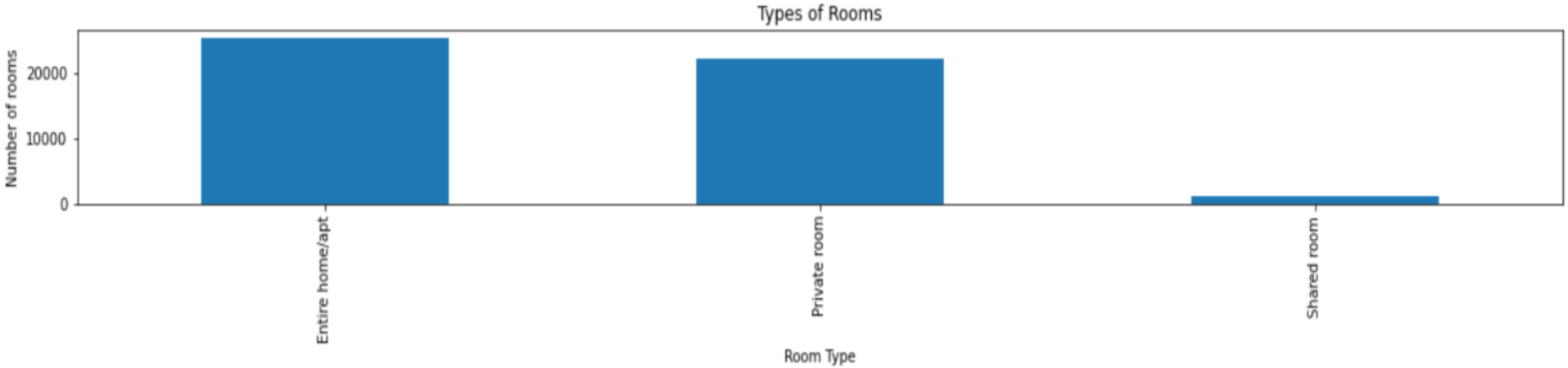
Analysing the relationship between the different host names and their respective calculated host listings count.



The bar plot visualization shows that the calculated host listings count feature basically tells the number of times that particular host has used Airbnb bookings in the data set. Host Sonder has listings count of 327 which means he has 327 rows in the data set, it represents total number of listings made by a specific host.

Different types of room available vs the number of rooms

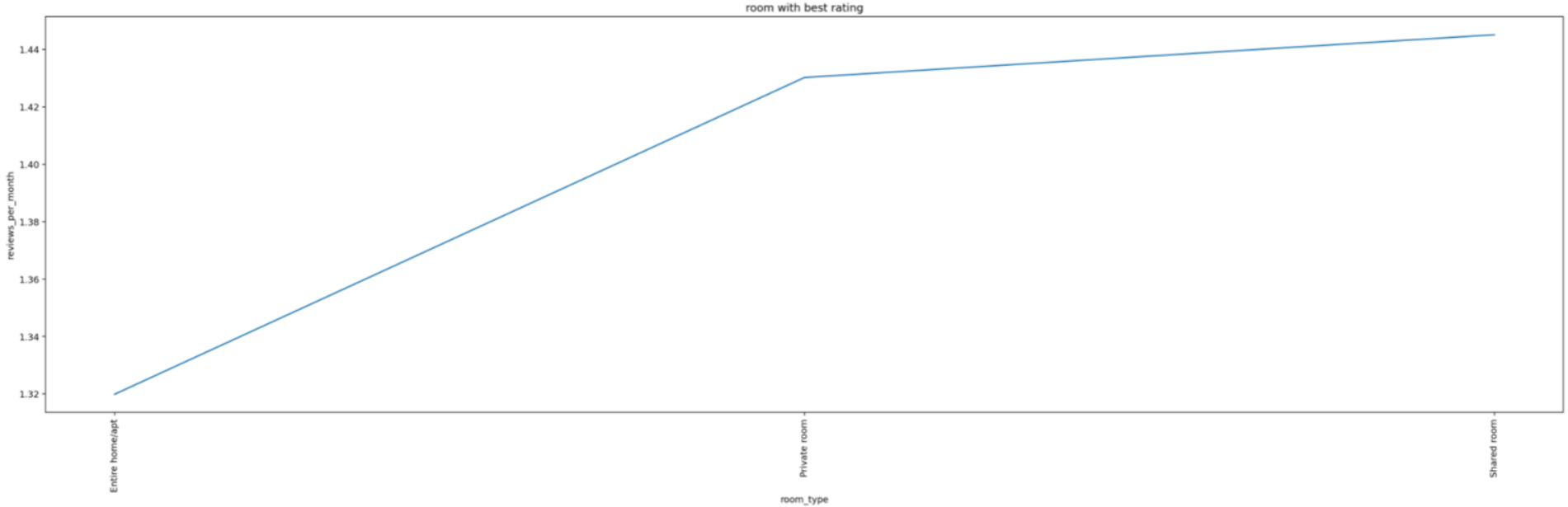
Three different types of rooms available while renting a room, enlisted under Airbnb Bookings data set.



The Plotly bar plot provides the visualization of three types of rooms along with the number of rooms named- Entire room/apartments, private rooms and shared rooms.There are around 25000 of entire home/apartments entries, around 22000 of private rooms and around 1000s of shared rooms entries in the data set.

Average of review per month vs room type

Reviews for rooms will affects its chances of being booked.



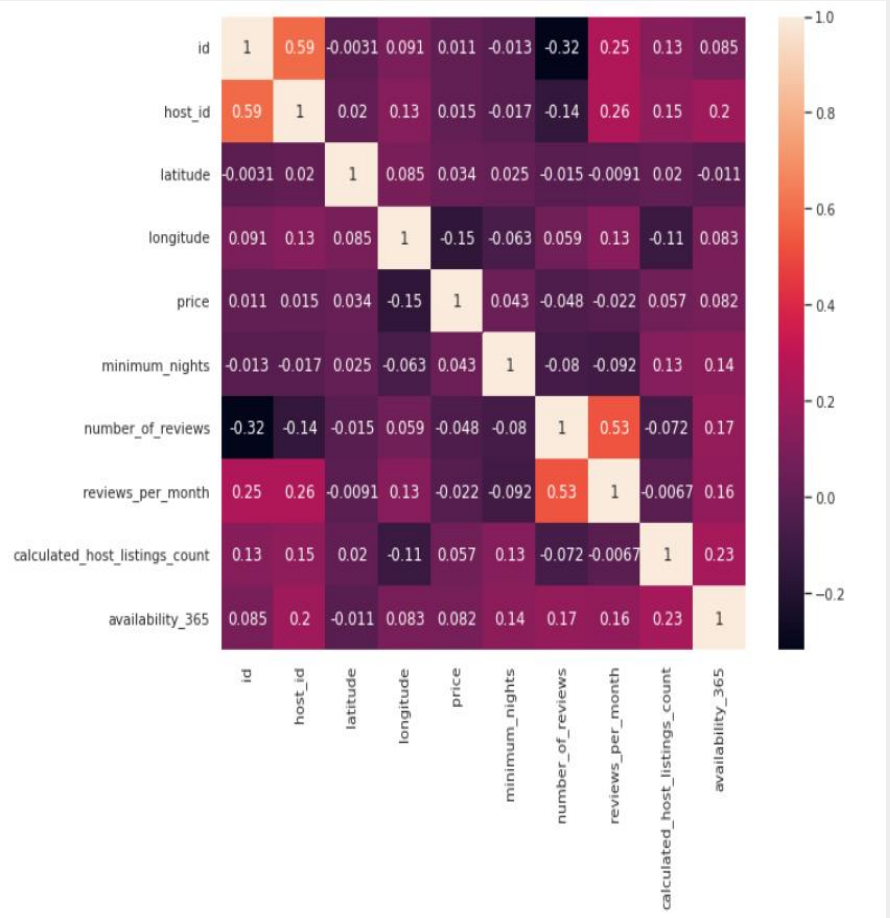
The line plot visualization provides the average review distribution for entire home/apartments is around 1.3, for private room and shared room is around 1.4. This means that there is a vast variation in the reviews given per month for each room type.

Heatmap Correlation between features

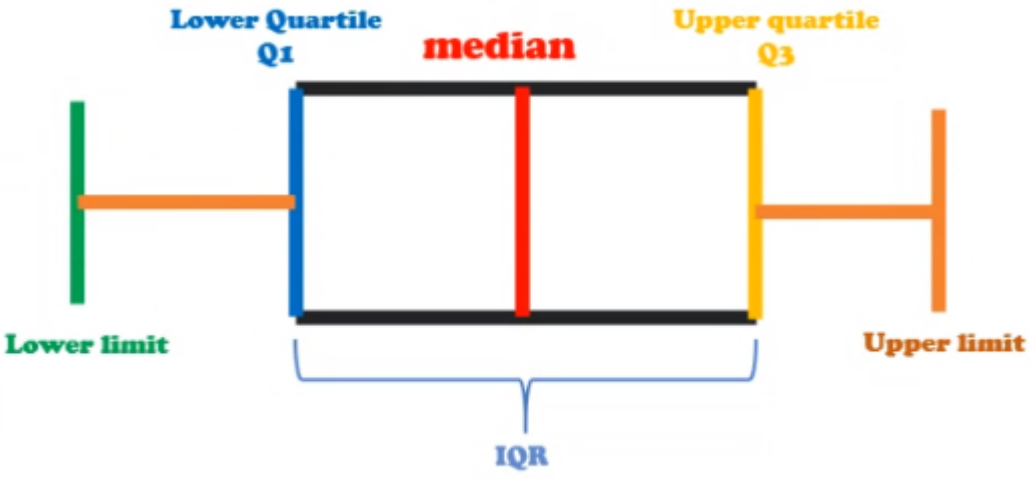
This is a heatmap chart that represents the relation between one feature and another one.

By plotting seaborn heatmap correlation we got to know that there features positively correlated with each other, among which reviews per month and number of reviews are highly correlated. There is 53% of chance that number of reviews increases by reviews per month of Airbnb bookings data.

Host_id is correlated to reveiws_per_month & availability_365. There is a correlation between calculated_listings_count, minimum_nights and availability_365.



Handling outliers - Box Plot



Q1 is 25 % percentile

Q3 is 75 percentile

We get (IQR) Inter Quartile Range by Q3 –Q1

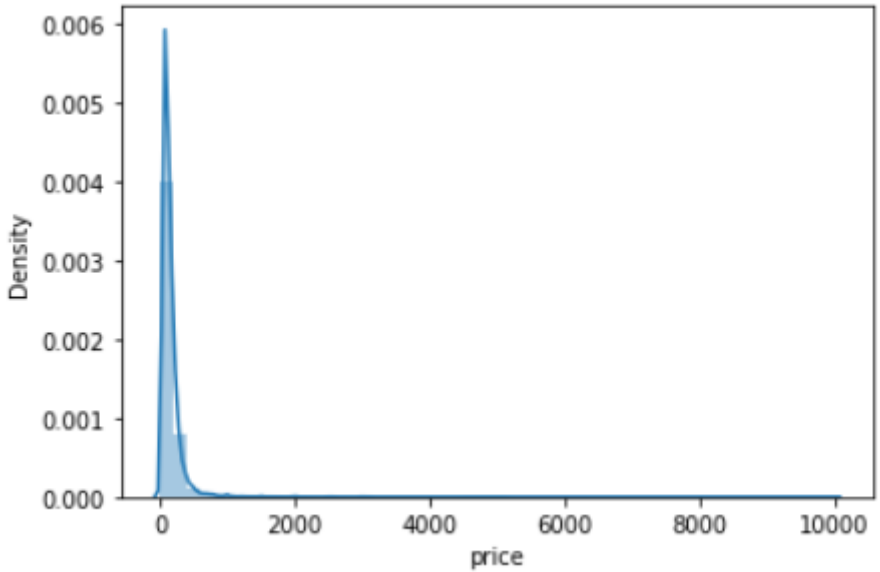
Lower limit = Q1 – 1.5 x IQR

Upper limit = Q3 + 1.5 x IQR

A box plot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles.

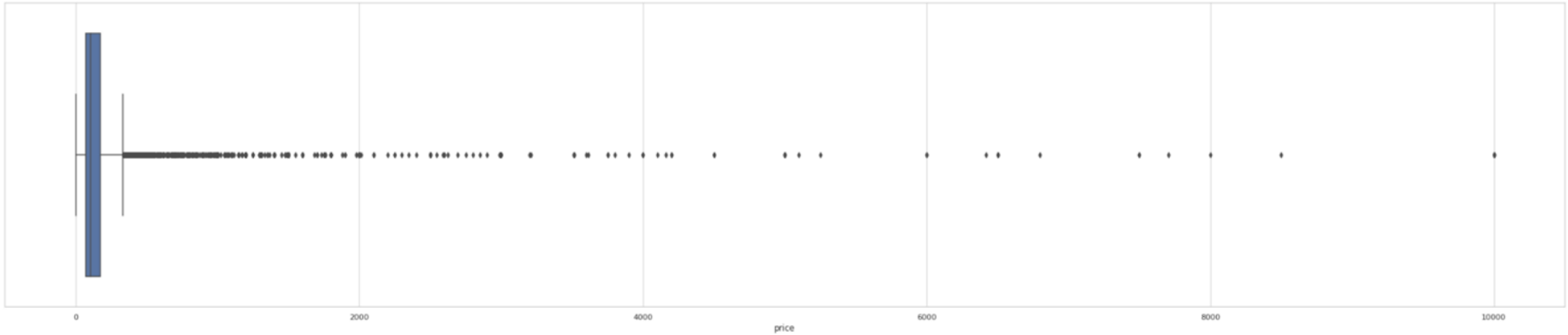
Distribution and nature of price feature

The seaborn distplot will provide the distribution curve to study the nature of price feature in data set.



The **skewness** is found to be 19.118939 and **kurtosis** to be 585.672879. We can observe that the skewness value being greater than 1 and kurtosis is high as 585, it indicates the presence of good amount of **outliers**.

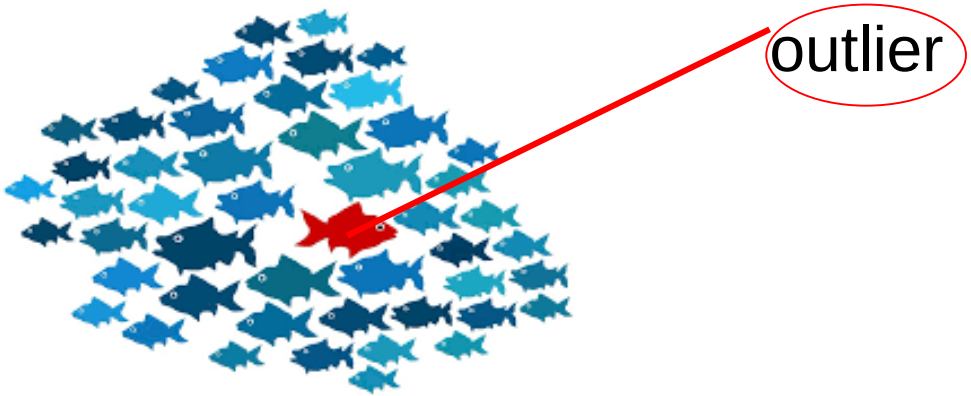
Presence of outliers in price feature: Box plot



An **outlier** is a data point that lies outside the overall pattern in a distribution.
The box plot visualization also confirms the presence of outliers in price feature.
These outliers has to be handled in a way that it does not affect the analysis in any way.

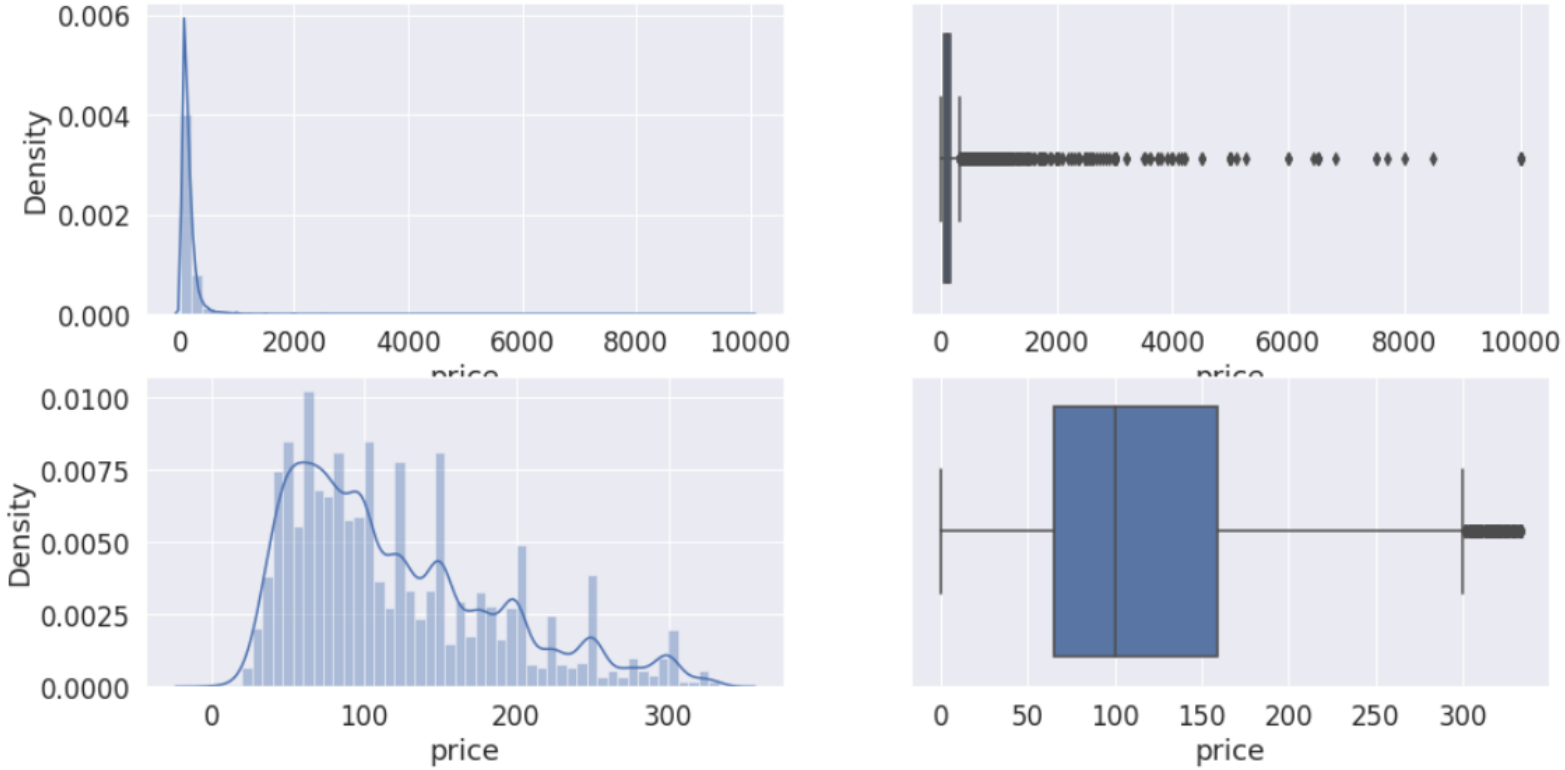
Handling outliers in price column

- Inter-Quartile Range(IQR) approach can be used to handle the high range and low range outliers.
- These outlier value may come from an accidental response that was recorded correctly or from a data that is entered wrongly which leads to an error.
- Low outliers are below Q1**–**1.5*IQR.
- High range outliers are above Q3+1.5*IQR**.**

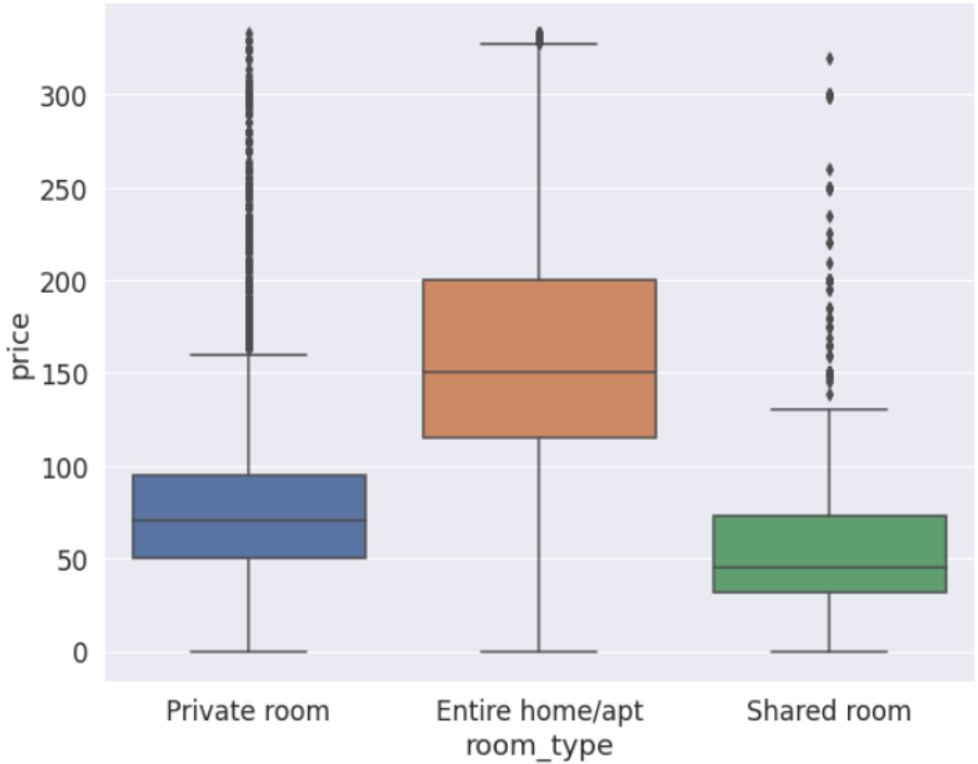


Distribution curve and box plot for price column

The combination of distribution curve and box plot gives the visualization of price column with and without outliers being handled using IQR filtering method.



Price vs room type: Box plot



This combination of box plots provides the variation of price with respect to the different type of rooms available. This visualization shows that the mean price of entire room/apartments is around 150$, mean price of private is around 70$ and that of shared room is 40$.

Average price of top 10 most reviewed listings in NYC:

the top 10 most reviewed listings on Airbnb bookings for NYC has average price of 65 dollars. Most of the listings has average price below 50 dollars. 9 out of 10 listings are private rooms type. The top reviewed listing has 629 reviews.

room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count
Private room	47	1	629	2019-07-05	14.58	2
Private room	49	1	607	2019-06-21	7.75	3
Private room	49	1	597	2019-06-23	7.72	3
Private room	49	1	594	2019-06-15	7.57	3
Private room	47	1	576	2019-06-27	13.40	2
Private room	46	1	543	2019-07-01	11.59	5
Private room	99	2	540	2019-07-06	6.95	1
Private room	48	1	510	2019-07-06	16.22	5
Entire home/apt	160	1	488	2019-07-01	8.14	1
Private room	60	3	480	2019-07-07	6.70	1

Neighbourhood group

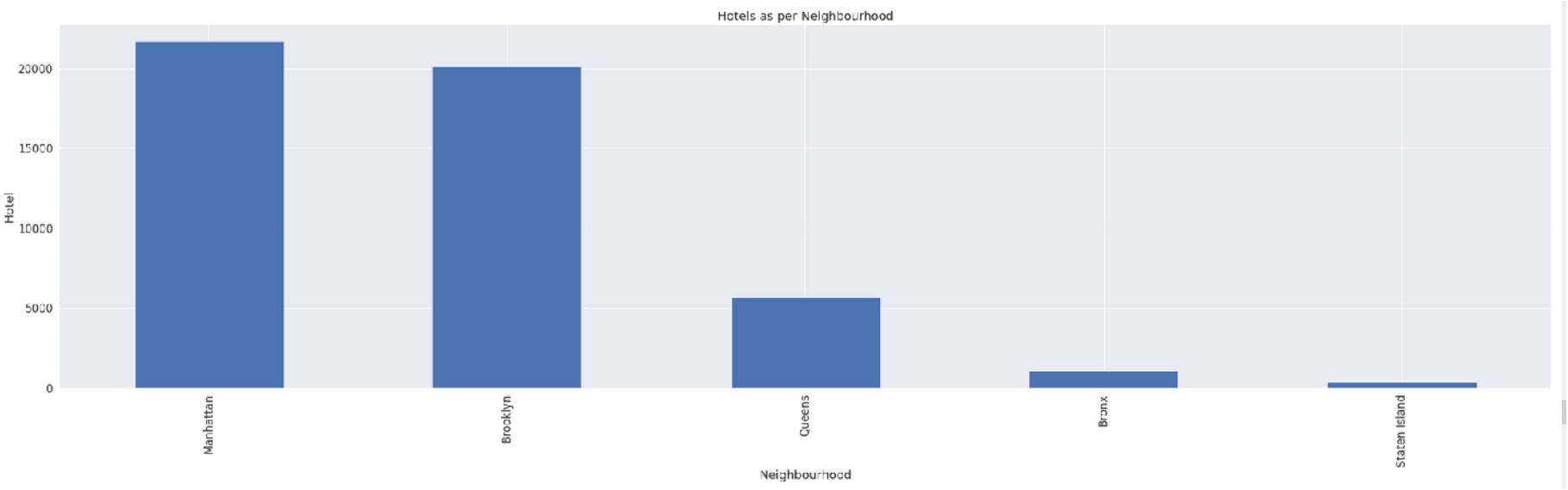
1) Brooklyn

2) Manhattan'

3) Queens

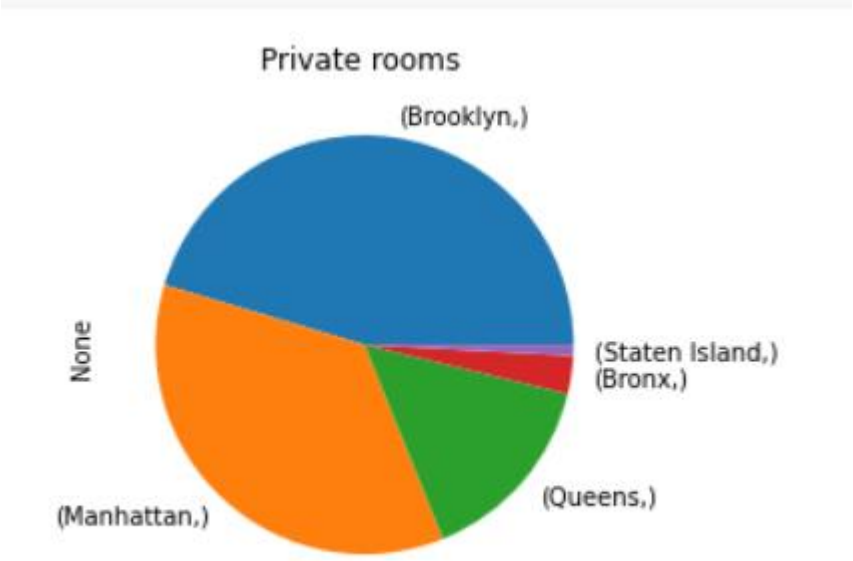
4) Staten Island

5) Bronx



The above bar plot indicates the number of hotels according to a particular neighborhood. As per the graph, Manhattan has the highest number of hotels and Staten Island has the lowest number of hotels.

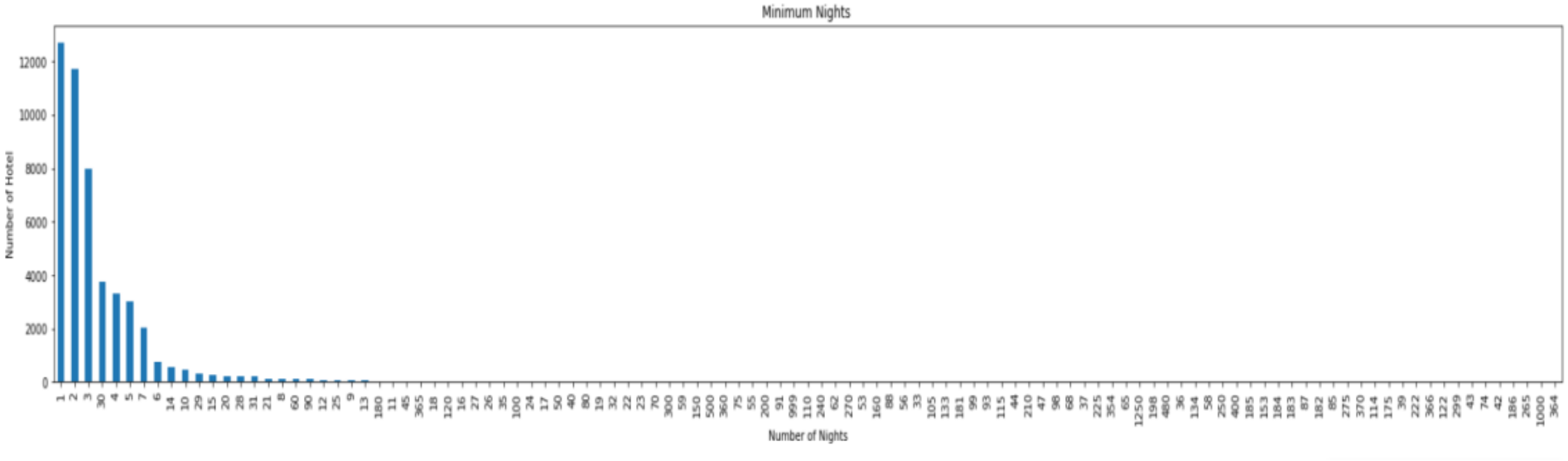
Private room in all neighbourhood



The pie chart displays private rooms in the neighborhood. The graph confirms that Brooklyn has the maximum number of hotels (10132) that have only private rooms.

The total number of hotels that have private rooms is 22,326.

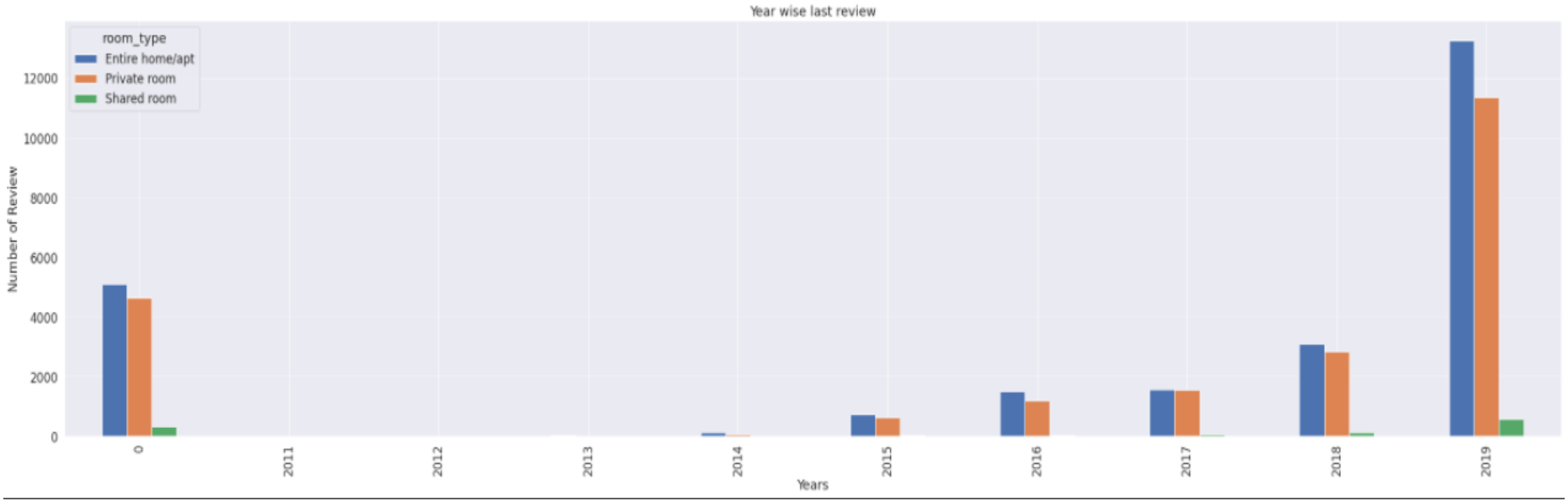
Minimum nights



The above bar graph shows the minimum nights available for all the hotels.,In the above graph. the x-axis shows the number of nights and the y-axis show number of hotels.

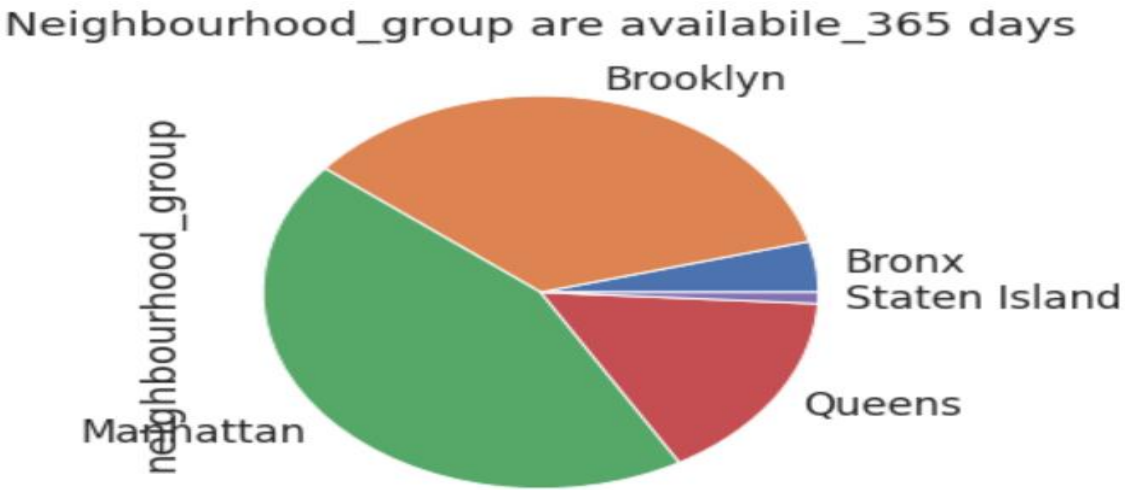
As per the graph maximum hotels(12720) have 1 as minimum nights.

Year wise types of rooms getting a last review



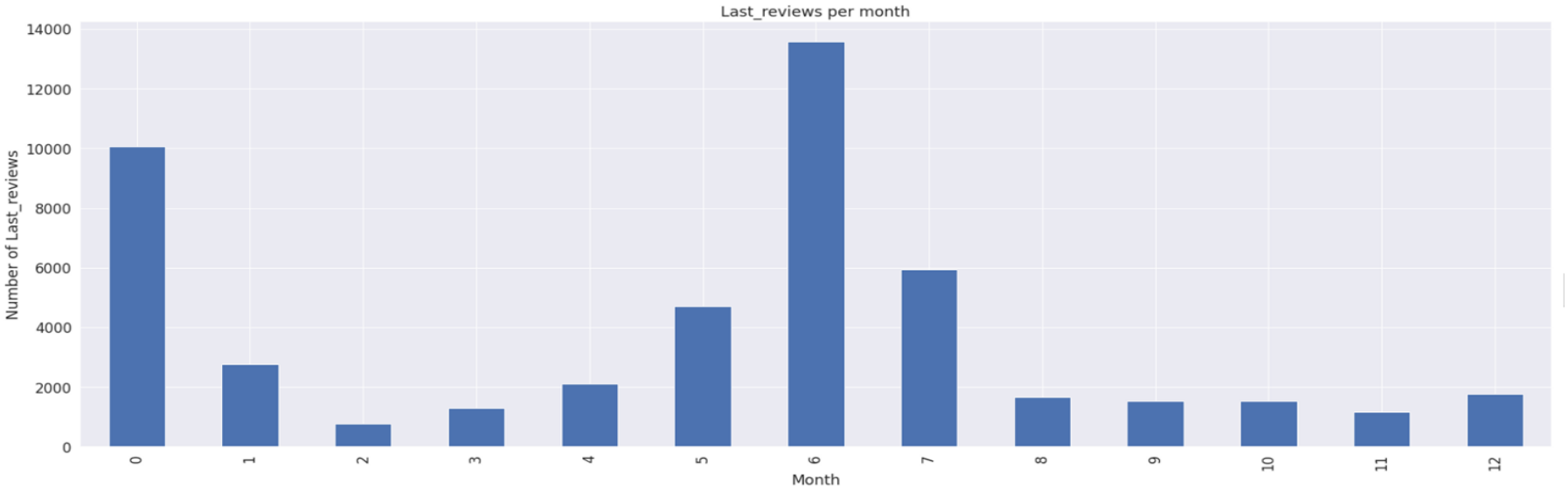
The above bar graph shows the types of room getting a last review in year wise. As per the graph from 2011-2014 very less number, 2015-2018 rooms are slightly increase. In 2019 we can see more number reviews. Year 0 represents the missing year's in Dataset.

How many Neighbourhood group are available All days



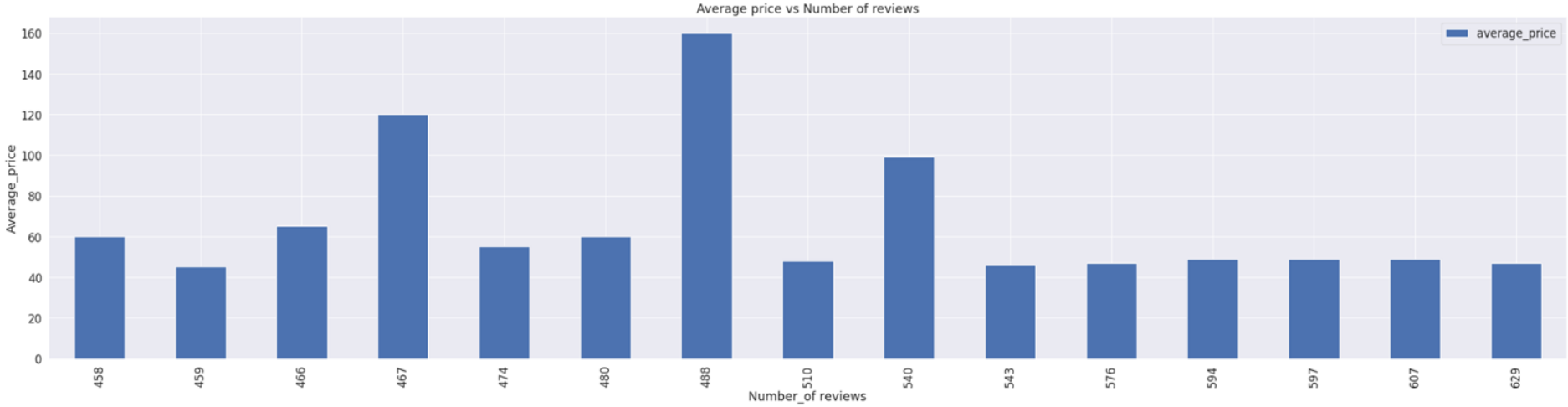
The above Pie chart shows the how many of Neighbourhood_group are available in all days of year. The Neighbourhood_group is categorical column. Highest number Neighbourhood_gruop is Manhattan (572) and lowest is Staten Island (12).

How many last reviews of every month



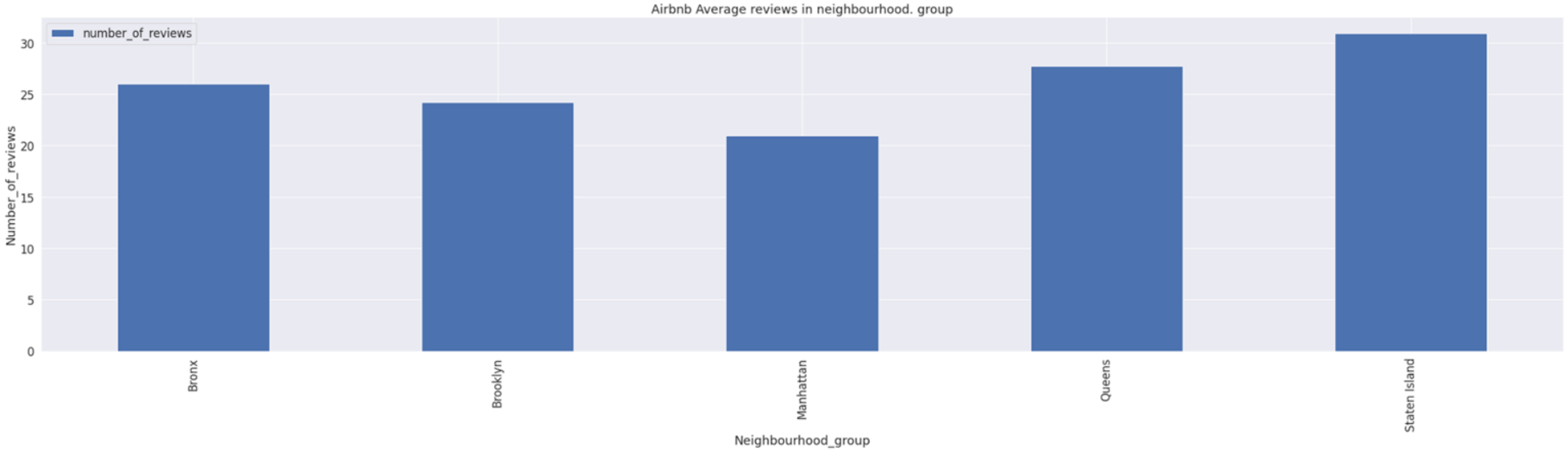
In the above Bar graph shows the last reviews of every month in all year's. The month 6(June) is getting the most last reviews in all year.

What is the highest average price corresponding to total number of reviews?



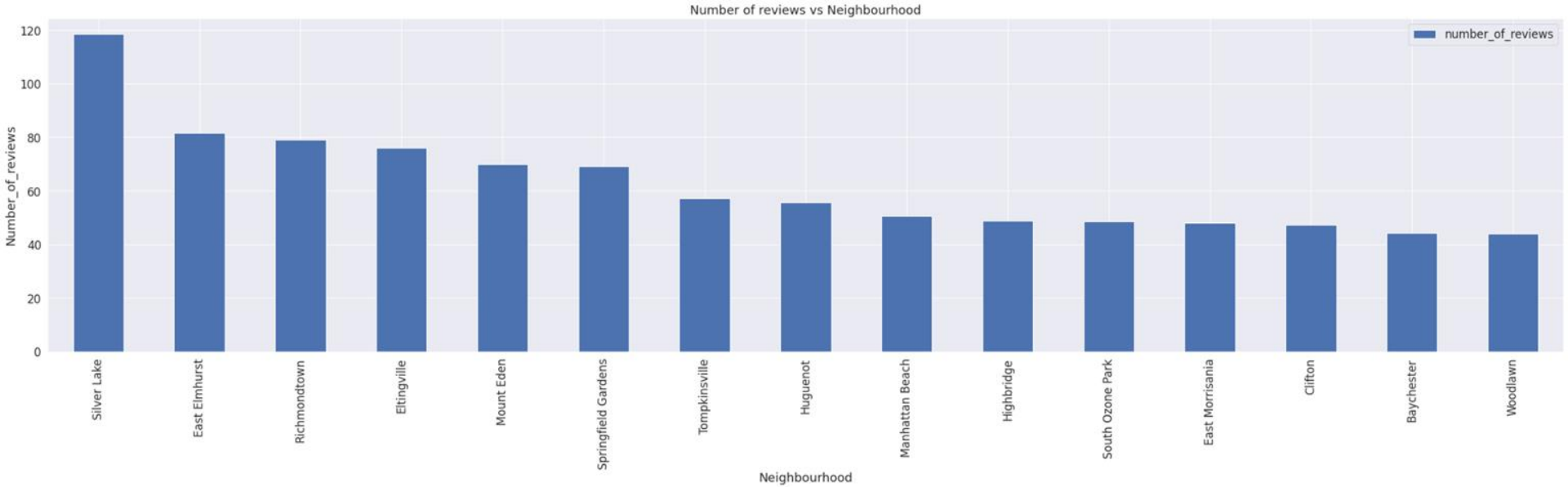
The above bar graph shows the average price according to the number of reviews available in the Airbnb data. As per the graph the highest average price of 166 have 488 total numbers of reviews.

How much the average number of reviews among neighborhood group?



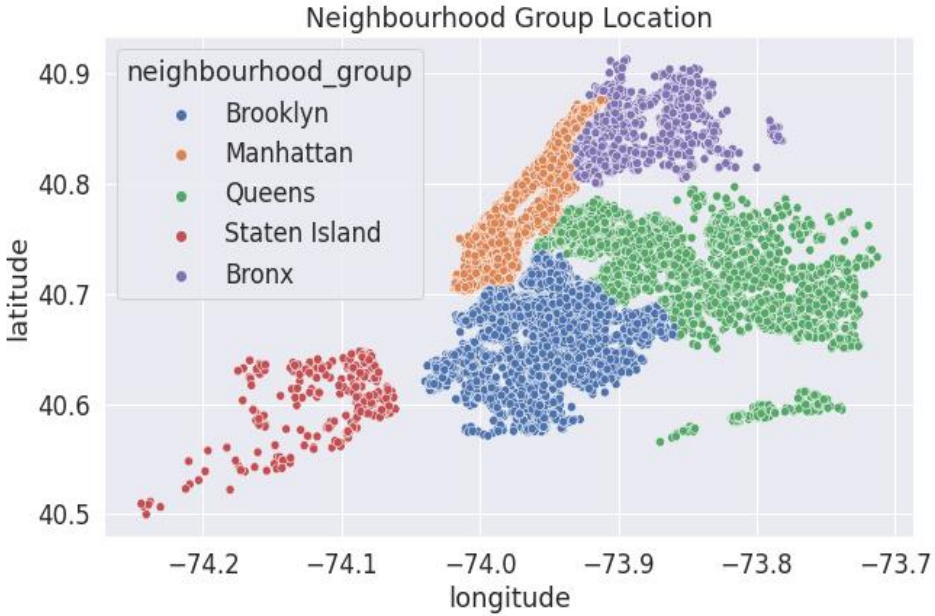
The above bar graph indicates the average number of reviews as per different neighborhood group available in the Airbnb data. As per the graph above the highest number of reviews comes under the Staten Island neighborhood group which is having the 30.97 of average reviews in the airbnb data.

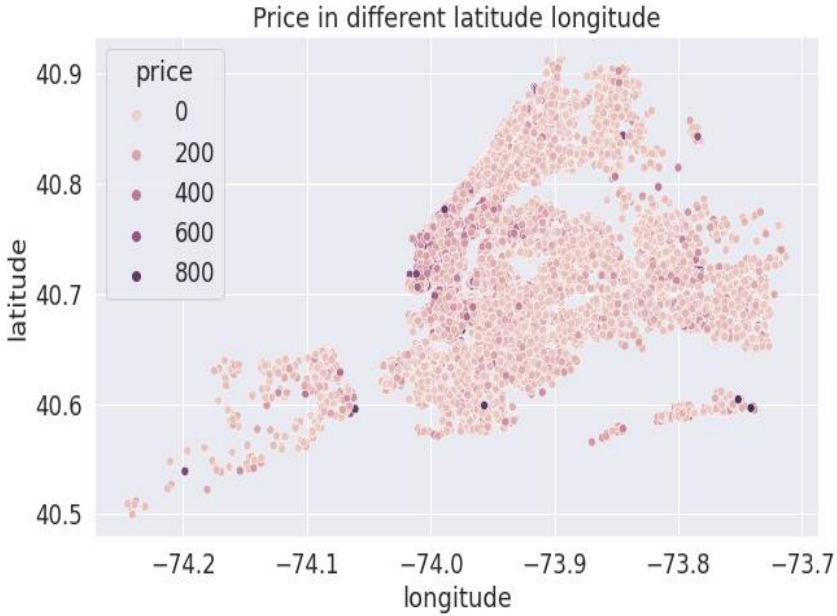
How much the average number of reviews comes in the neighborhood?



The above bar graph indicates here the average number of reviews vs the Neighborhood available in the airbnb data. In the above graph the Silver Lake has the maximum number of reviews among all the neighborhoods.

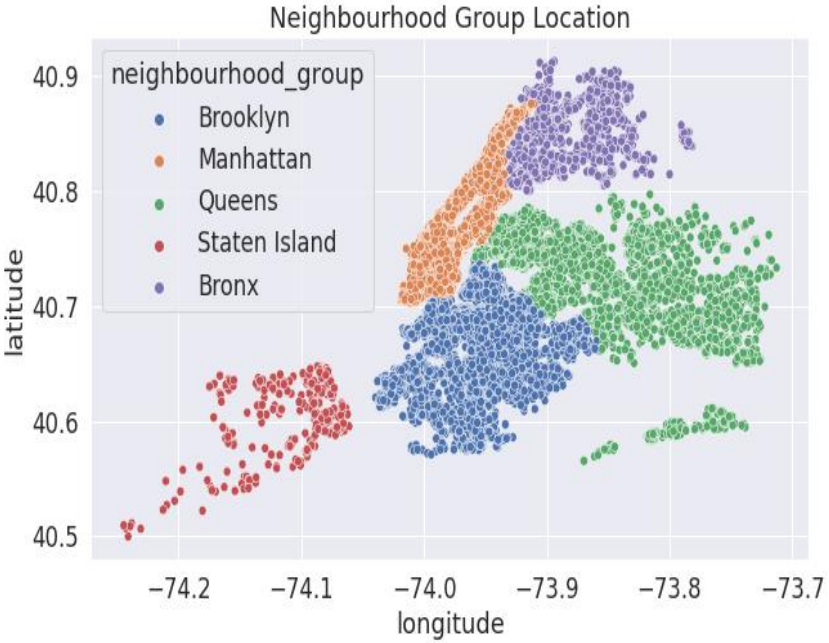
Price distribution corresponding to neighborhood group

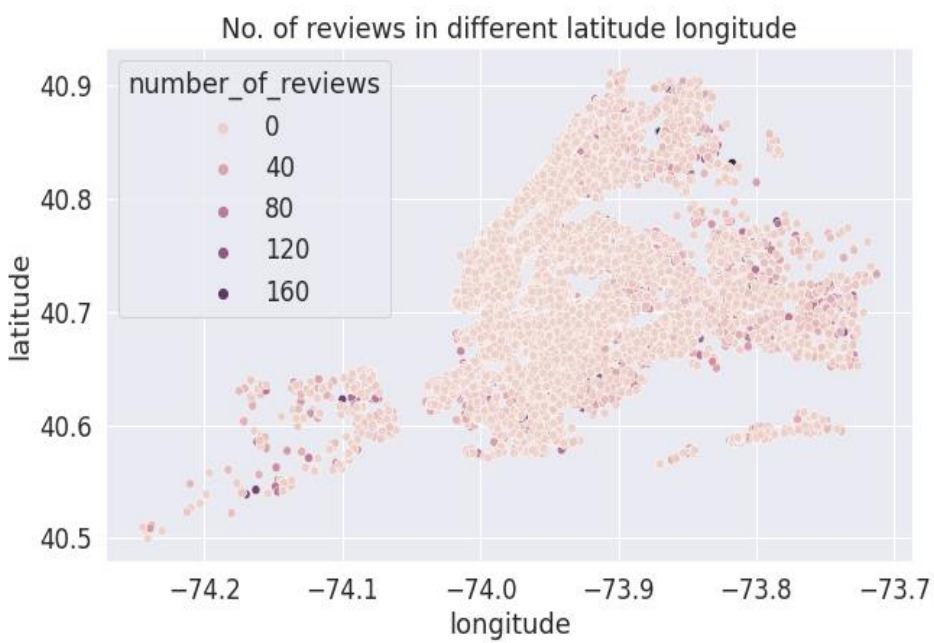




This scatter plot helps us to find out the neighborhood group in the map that is represented by the number of dots based on the latitude and longitude given to us in the dataset. The other plot indicates the price in different longitude and latitude in the dots. So by comparing both plots the price of 800 is very less in number among the neighborhood group.

Number of reviews distribution corresponding to neighborhood group





The first plot is same which is already mentioned above. Now the other plot shows the number of reviews among the different neighborhood group. Most of the average number of reviews are fallen between 40 and 80 among the neighborhood group.

Conclusion

- The Airbnb dataset gave us a great source to very well understand New York's rental landscape and bookings. New York has proven to be one of the Airbnb's fastest growing cities with around 49K listings in the last 9 years.
- Through this exploratory data analysis and data visualization, we obtained very interesting insights into the Airbnb rental market. This Airbnb dataset for year 2019 looks like a variety of columns that allowed us to do deep data exploration on each important features. We analyzed neighbourhood groups and neighbourhood listings, the important areas and their popularity with one another. We obtained the distribution and nature of the price variations with respect to different features. We handled the outliers present in the price column by using box plot and IQR method in order to have better analysis of the observations.
- We also emphasized on main findings like room types and their preference by the customers, top average listing counts and the top reviewed hosts.
- We took an opportunity to make good use of latitude and longitude columns to create geographical scatter plots to find the neighbourhood groups in the map, to find the price variations in different latitude and longitudes and the number of reviews among the different neighbourhood groups.