Capstone Project

Airbnb Bookings Analysis

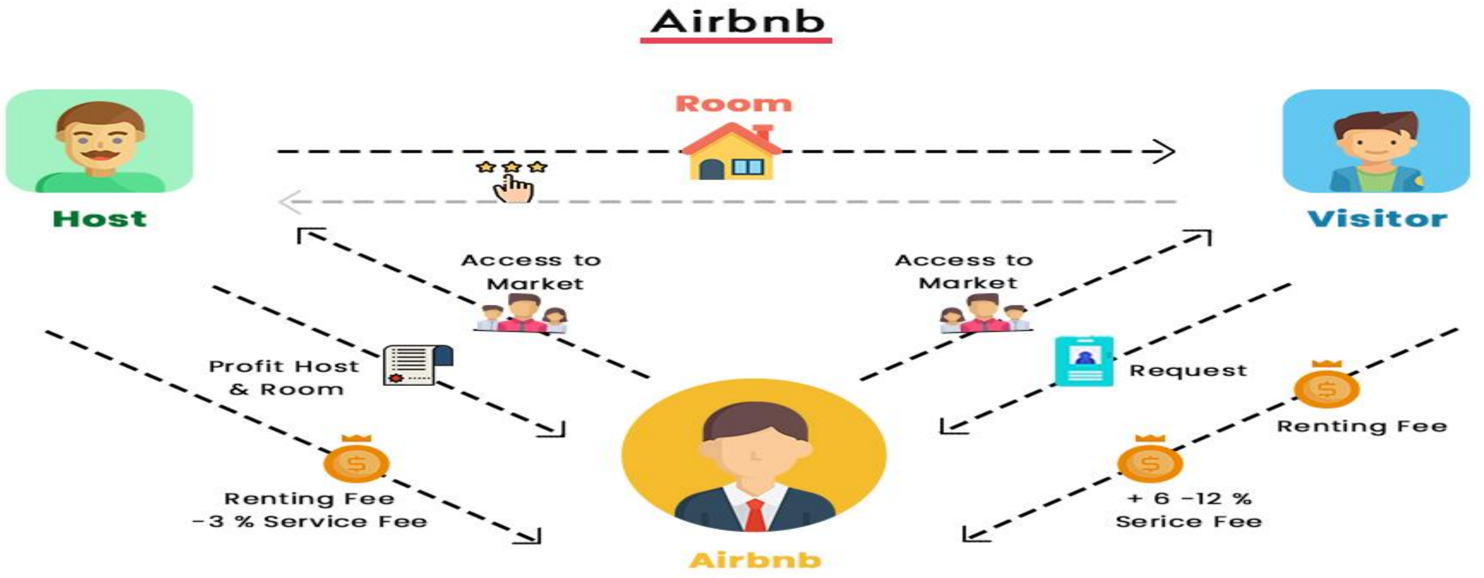
By-Prabhujeet Kaur

Points for Discussion

- Defining Problem Statement
- Data Pipeline
- Data summary
- Data Cleaning (Checking Null Values)
- EDA
- Conclusion

Defining Problem Statement

How the Airbnb works?



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 Pipeline

- Data processing-1: In this first part we have removed unnecessary features, if required. Since there were few columns with some null values.
- Data processing-2: In this part, we manually go through each features selected from part 1, and encoded the categorical data and numerical data separately for better analysis.
- EDA: In this part, we do exploratory data analysis (EDA) on the data selected in part 2 to see the number of trend.
- Conclusion: Finally , In this last part, we concluded some points by the EDA.

Data Summary

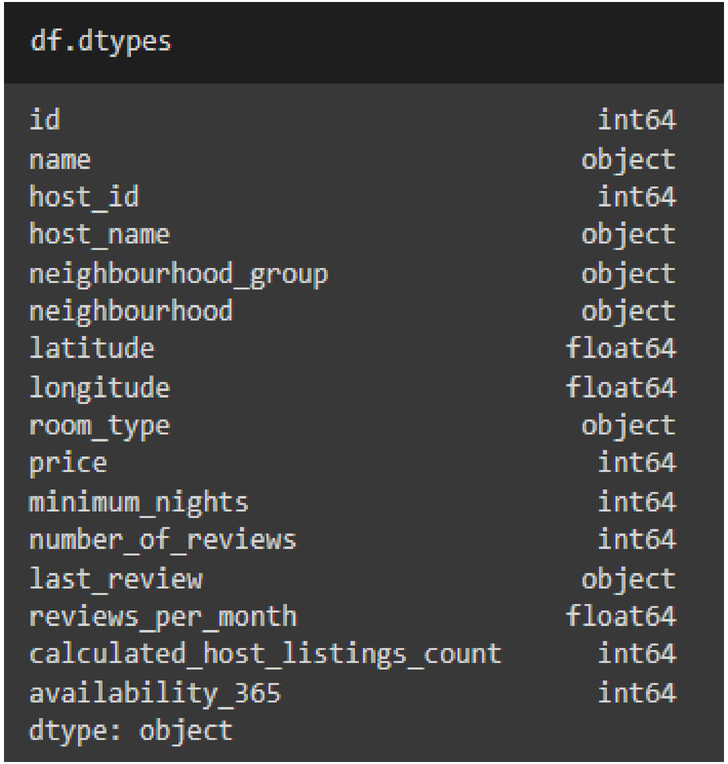
Data set:

- id: listing ID
- name: name of the listing
- host_id: host ID
- host_name: name of the host
- neighbourhood_group: location
- neighbourhood: area
- latitude: latitude coordinates
- longitude: longitude coordinates
- room_type: listing space type
- price: price in dollars
- minimum_nights: amount of nights minimum
- number_of_reviews: number of reviews
- last_review: latest review
- reviews_per_month: number of reviews per month
- calculated_host_listings_count: amount of listing per host
- availability_365: number of days when listing is available for booking

Data Summary

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 numerical data.)

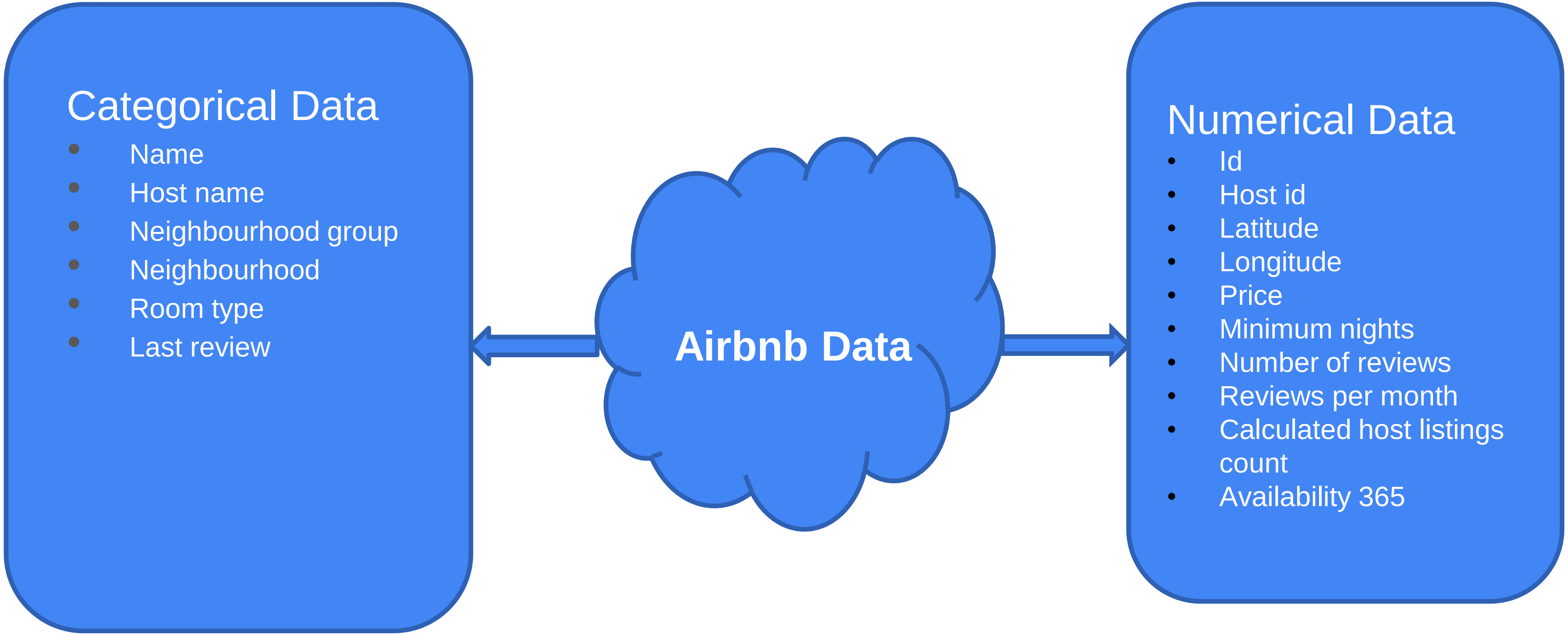
Data set shape : The Dataset contains 48895 rows and 16 columns.Out of 16 columns, we have 6 categorical column and rest numerical column.



```
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 Summary



Airbnb Data

Categorical Data

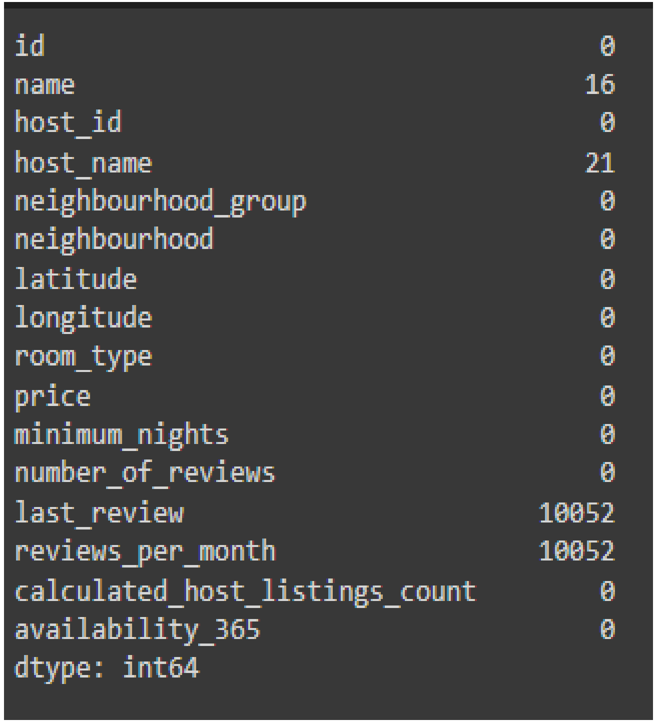
- Name
- Host name
- Neighbourhood group
- Neighbourhood
- Room type
- Last review

Numerical Data

- Id
- Host id
- Latitude
- Longitude
- Price
- Minimum nights
- Number of reviews
- Reviews per month
- Calculated host listings count
- Availability 365

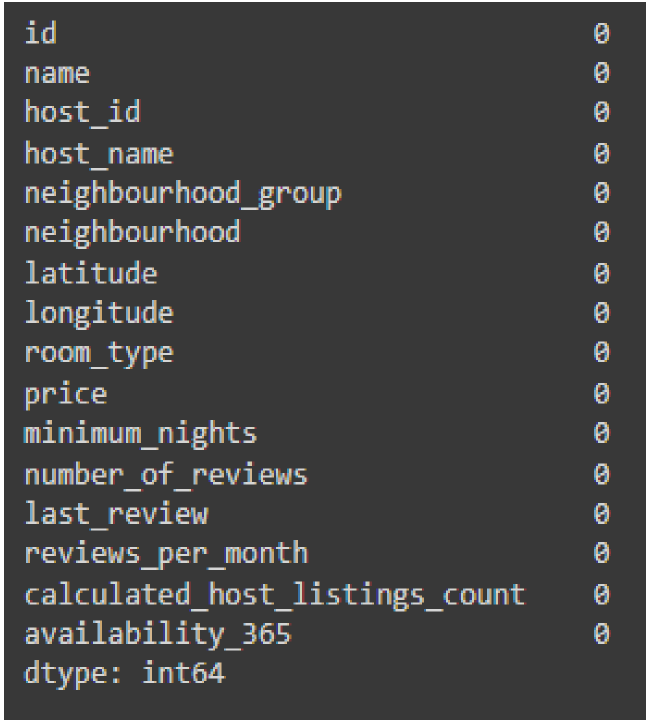
Data Cleaning

- 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
```

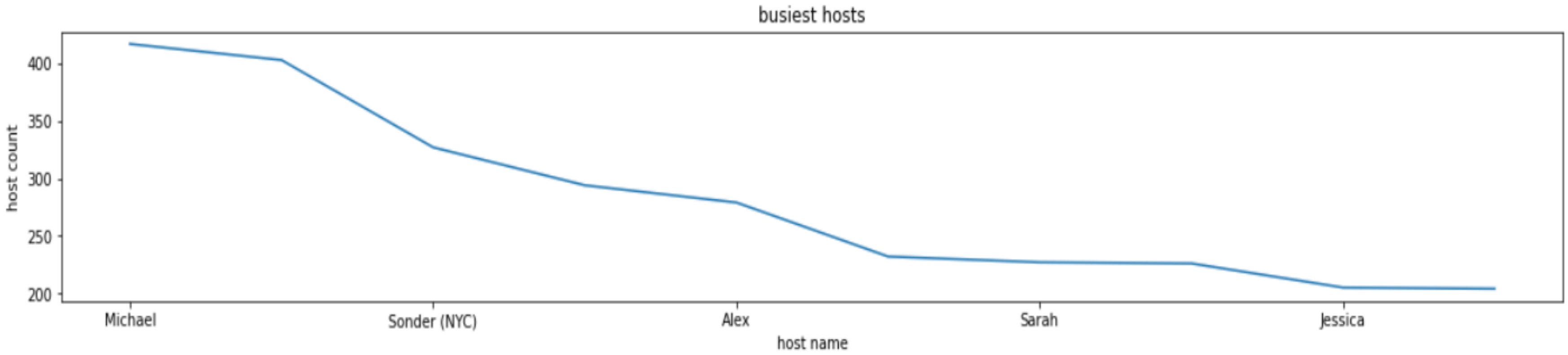
Data Cleaning

- 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.

EDA(Exploratory Data Analysis)

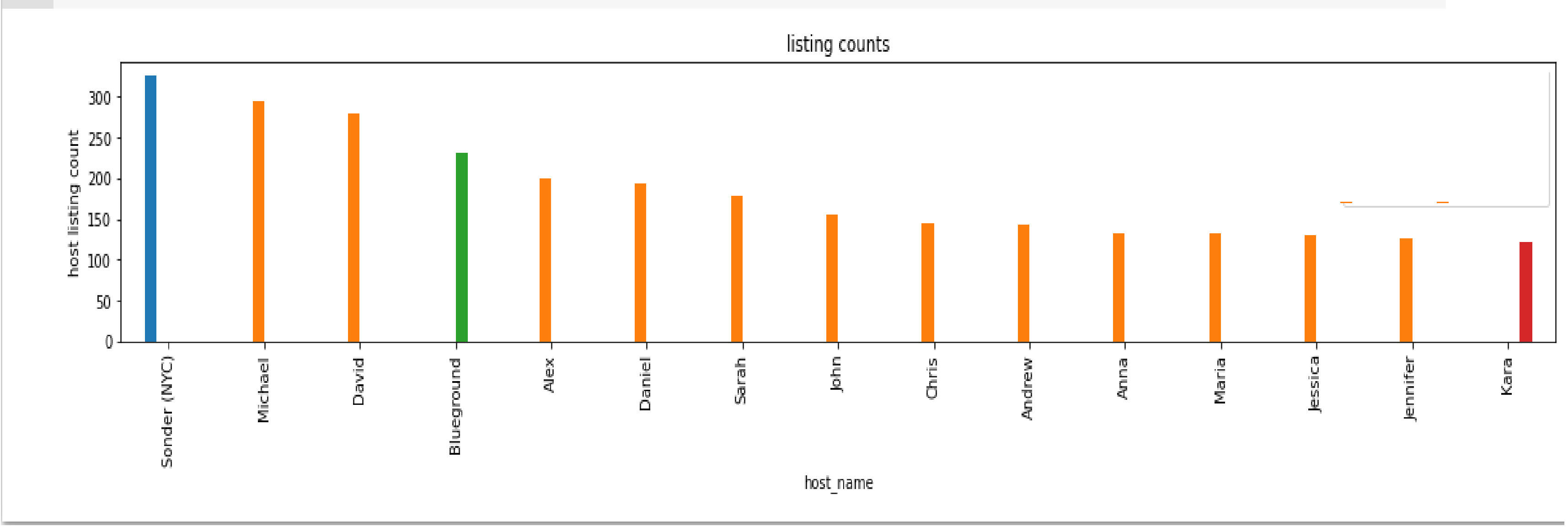
- What can we learn about different hosts and areas?
- What can we learn from predictions? (ex: locations, prices, reviews, type of rooms etc)
- Which hosts are the busiest and why?
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?
- How long ago the last review was left for the host?
- Price related to the review per month?
- Which features are highly correlated with among selves
- Average price of top 10 most reviewed
- Number of nights per number of hotel variation

Hosts with maximum number of entries



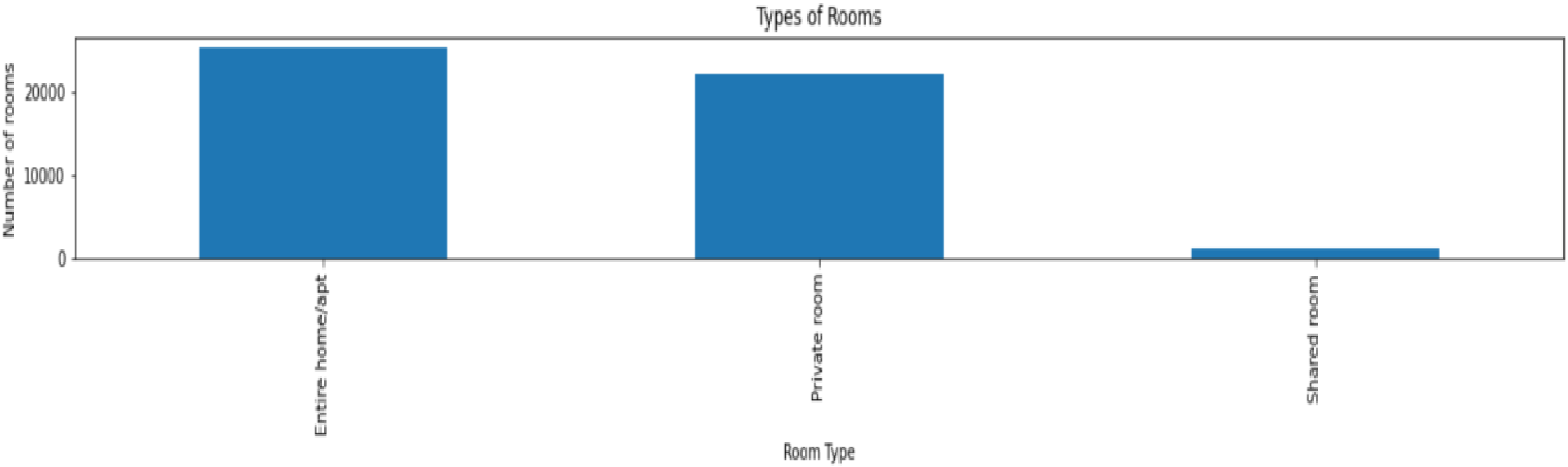
The Airbnb Bookings data set have different host entries along with their names and related data. The line plot visualization gives the top 10 busiest hosts in the Airbnb bookings data set. According to the analysis, host Micheal has the maximum value count with 417 entries/rows in data set followed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

Host listings count vs Host name



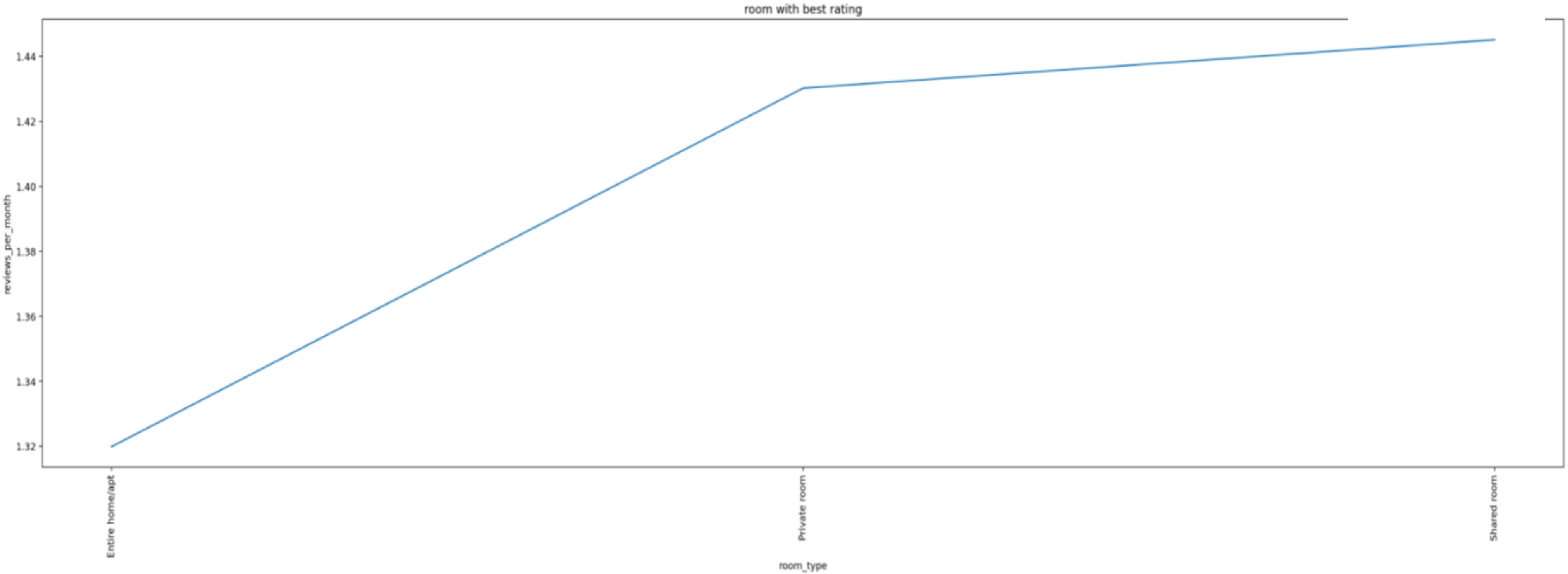
Now here we make the bar plot that shows the calculated host listings count feature. Basically tells the number of listing done by particular host used Airbnb bookings in the data set. Host Sonder has host listings count of 327, it represents total number of host listings count made by a specific host.

Different types of room available vs the number of rooms



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



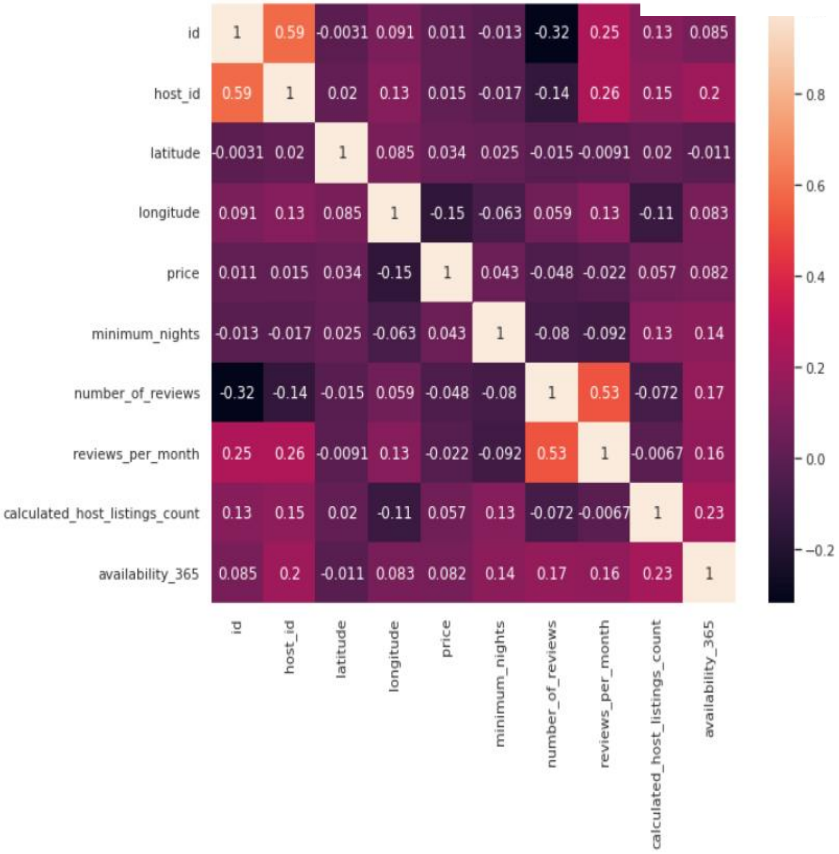
The line plot visualization provides the average review distribution for entire home/apartments is around 1.3, for private room and shared room is around 1.4. This means that there is a vast variation in the reviews given per month for each room type.

Heatmap Correlation between features

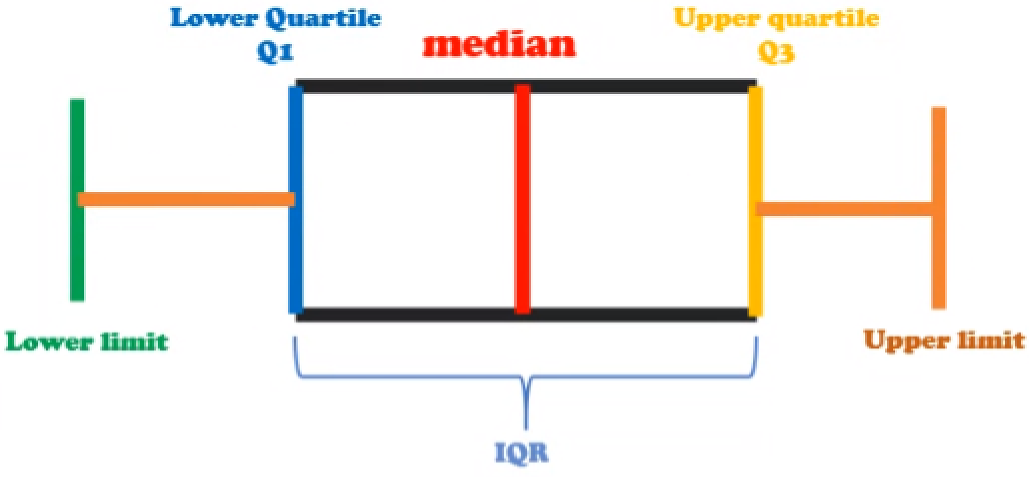
This is a heatmap chart that represents the relation between one feature and another one.

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 reviews_per_month & availability_365. There is a correlation between calculated_listings_count, minimum_nights and availability_365.



Handling outliers - Box Plot



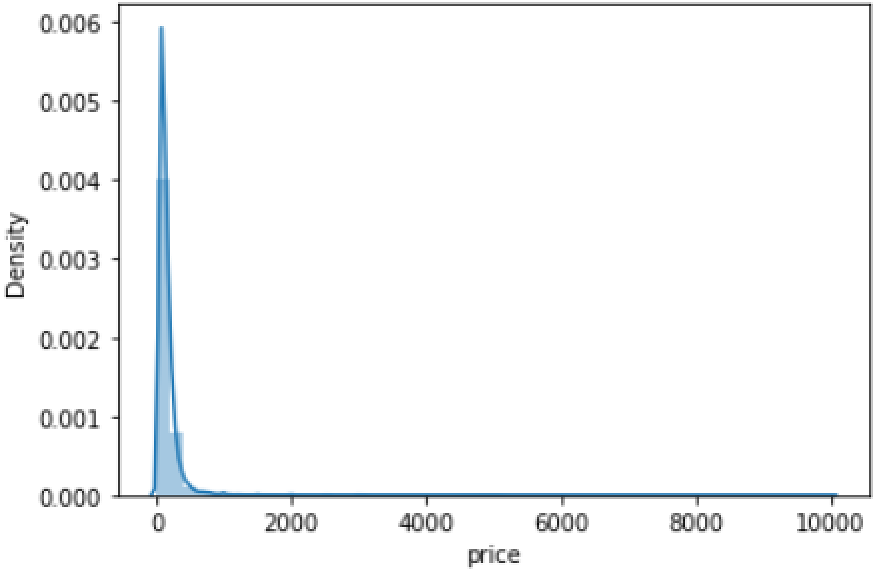
Q1 is 25 percentile.
Q3 is 75 percentile.

We get (IQR) Inter Quartile Range by Q3-Q1
Lower limit = Q1 – (1.5 x IQR)
Upper limit = Q1 + (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.

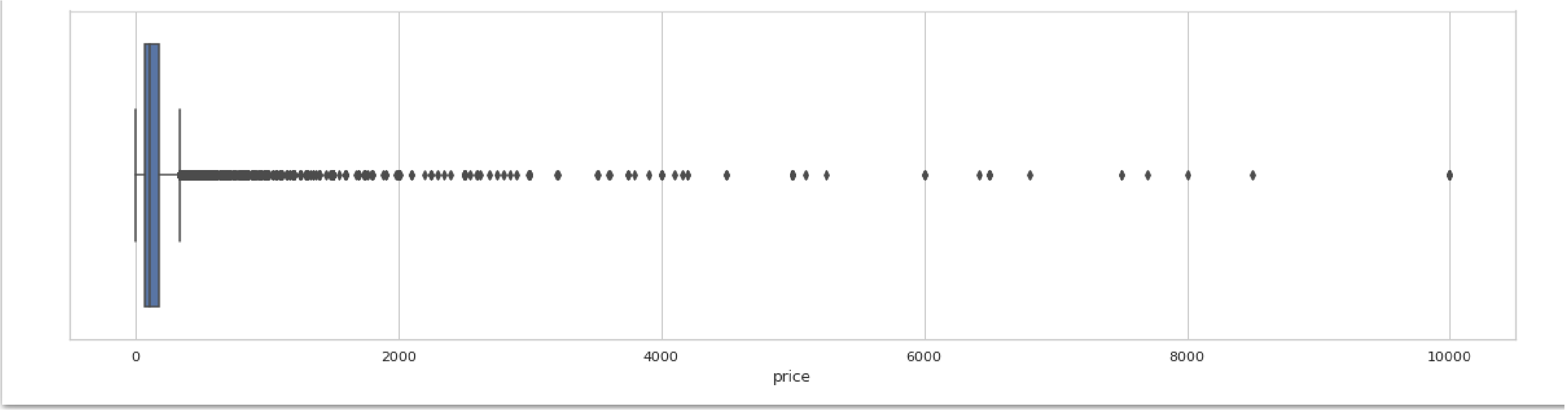
Distribution and nature of price feature

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**.

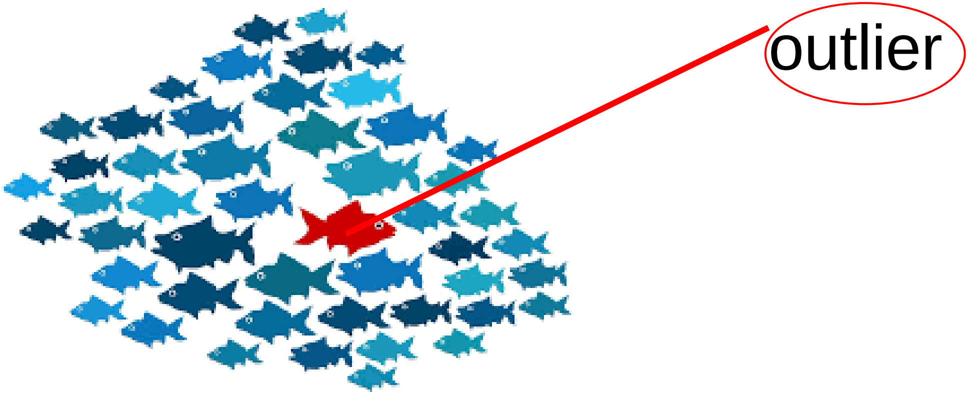
Presence of outliers in price feature: Box plot



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.

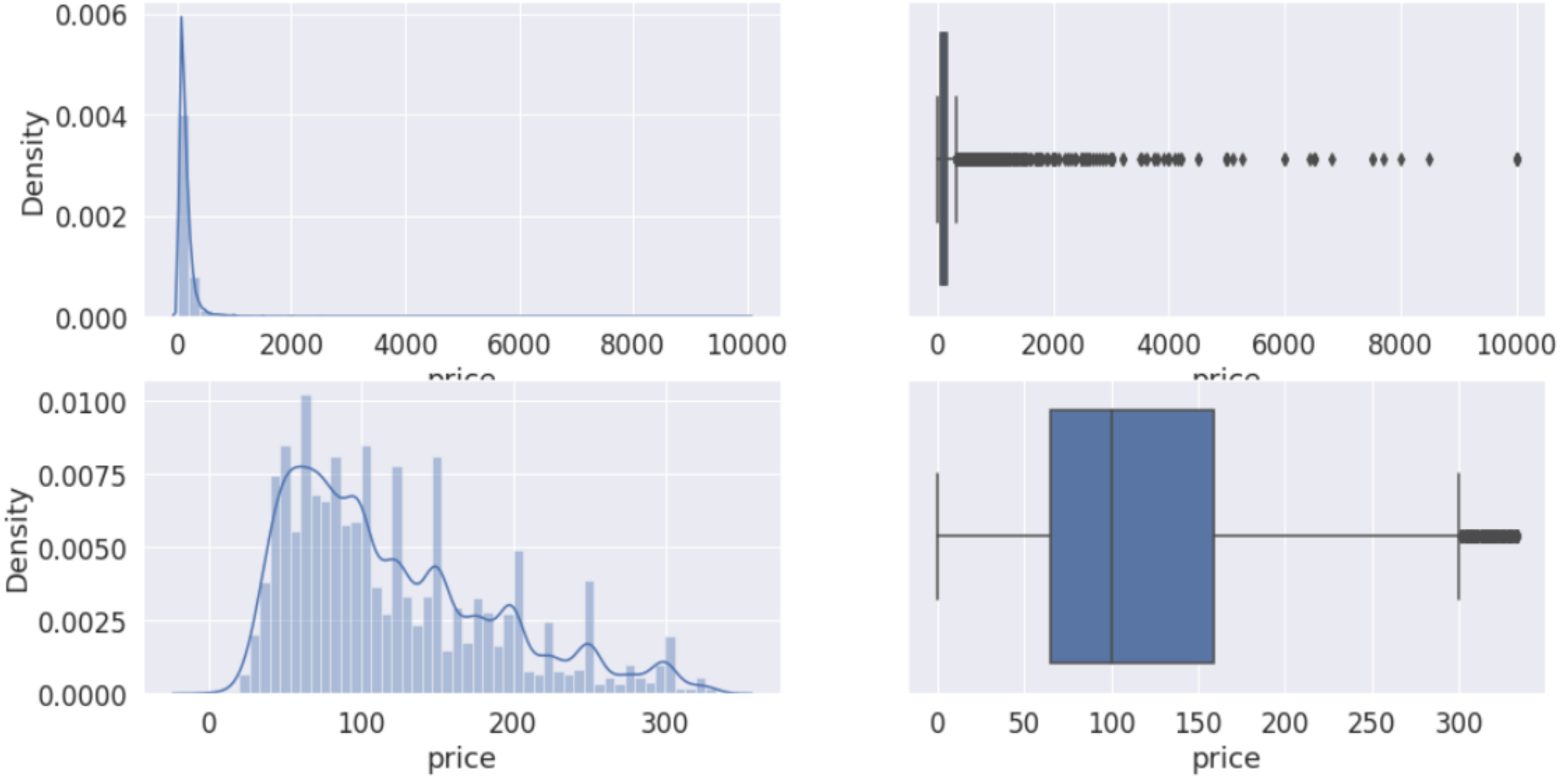
Handling outliers in price column

- 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**.**

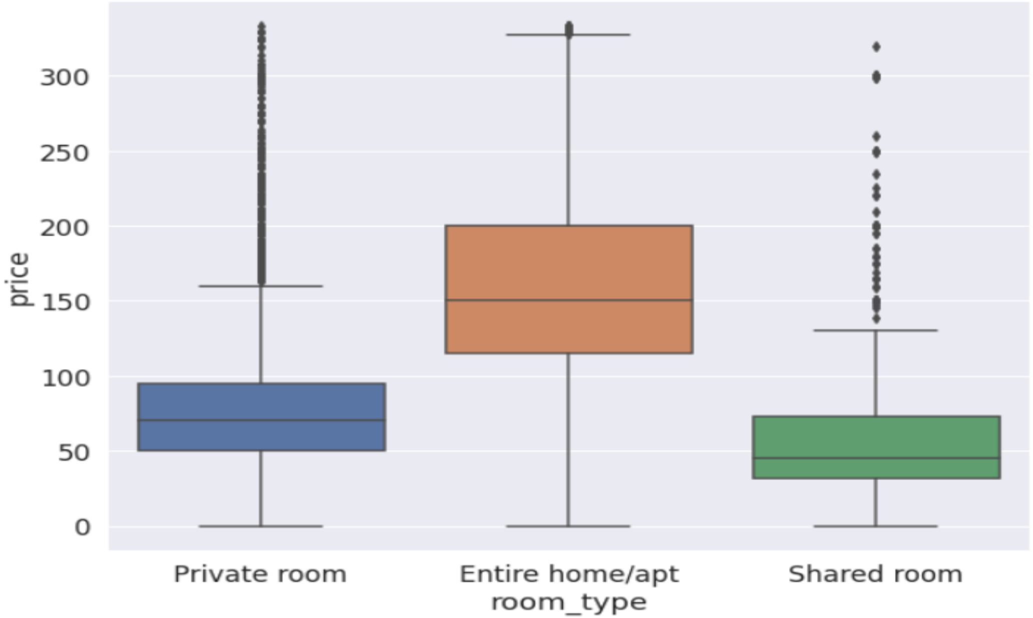


Distribution curve and box plot for price column

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

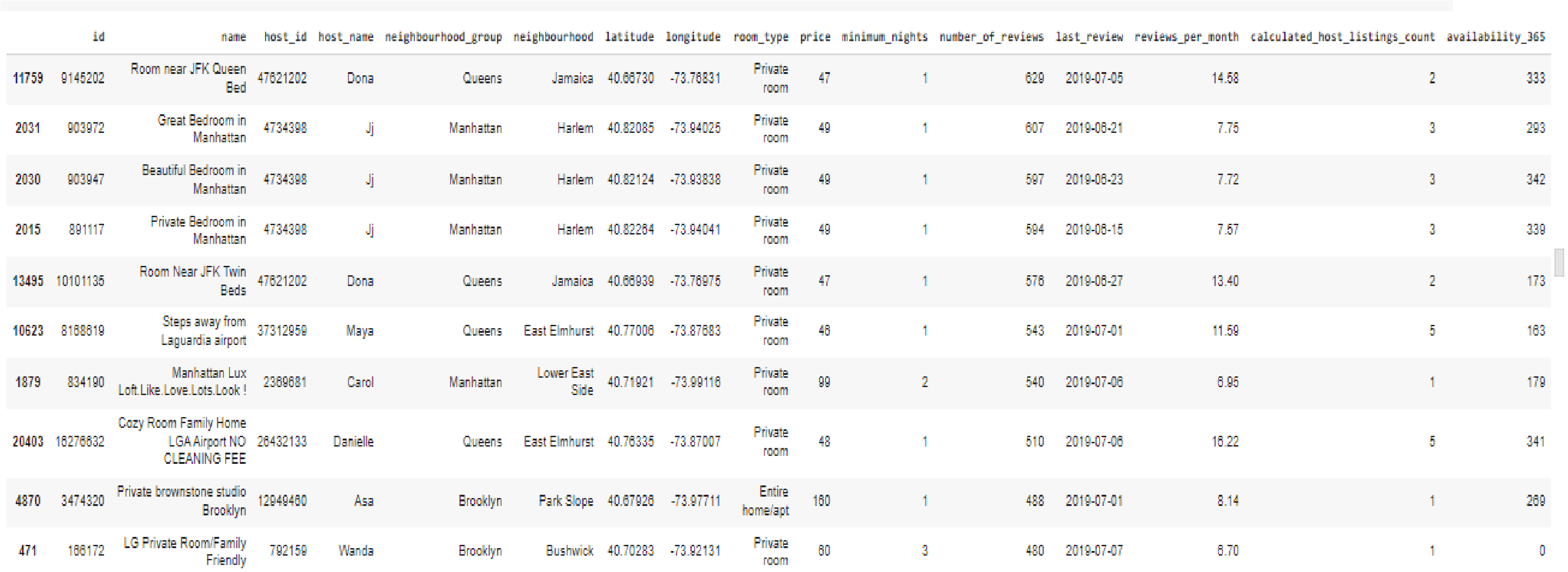


Price vs room type: Box plot



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$.

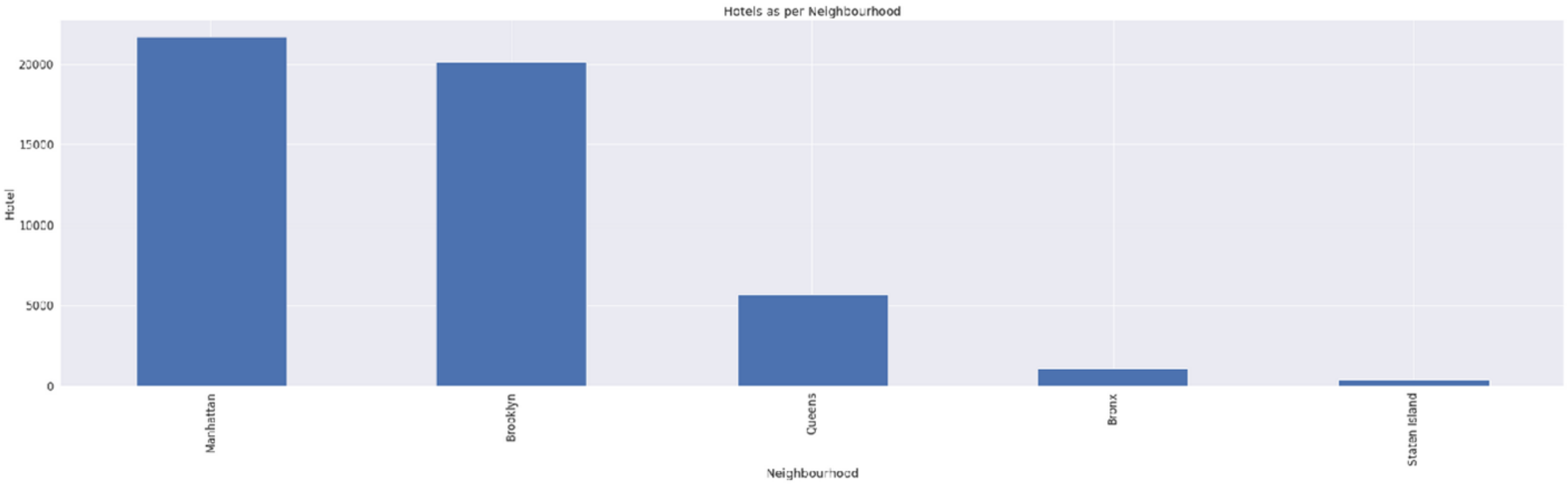
Average price of top 10 most reviewed



	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
11759	9145202	Room near JFK Queen Bed	47621202	Dona	Queens	Jamaica	40.66730	-73.76831	Private room	47	1	629	2019-07-05	14.58	2	333
2031	903972	Great Bedroom in Manhattan	4734398	Jj	Manhattan	Harlem	40.82085	-73.94025	Private room	49	1	607	2019-06-21	7.75	3	293
2030	903947	Beautiful Bedroom in Manhattan	4734398	Jj	Manhattan	Harlem	40.82124	-73.93838	Private room	49	1	597	2019-06-23	7.72	3	342
2015	891117	Private Bedroom in Manhattan	4734398	Jj	Manhattan	Harlem	40.82264	-73.94041	Private room	49	1	594	2019-06-15	7.57	3	339
13495	10101135	Room Near JFK Twin Beds	47621202	Dona	Queens	Jamaica	40.66939	-73.76975	Private room	47	1	576	2019-06-27	13.40	2	173
10623	8168619	Steps away from Laguardia airport	37312959	Maya	Queens	East Elmhurst	40.77006	-73.87683	Private room	46	1	543	2019-07-01	11.59	5	163
1879	834190	Manhattan Lux Loft.Like.Love.Lots.Look !	2369681	Carol	Manhattan	Lower East Side	40.71921	-73.99116	Private room	99	2	540	2019-07-06	6.95	1	179
20403	16276632	Cozy Room Family Home LGA Airport NO CLEANING FEE	26432133	Danielle	Queens	East Elmhurst	40.76335	-73.87007	Private room	48	1	510	2019-07-06	16.22	5	341
4870	3474320	Private brownstone studio Brooklyn	12949460	Asa	Brooklyn	Park Slope	40.67926	-73.97711	Entire home/apt	160	1	488	2019-07-01	8.14	1	269
471	166172	LG Private Room/Family Friendly	792159	Wanda	Brooklyn	Bushwick	40.70283	-73.92131	Private room	60	3	480	2019-07-07	6.70	1	0

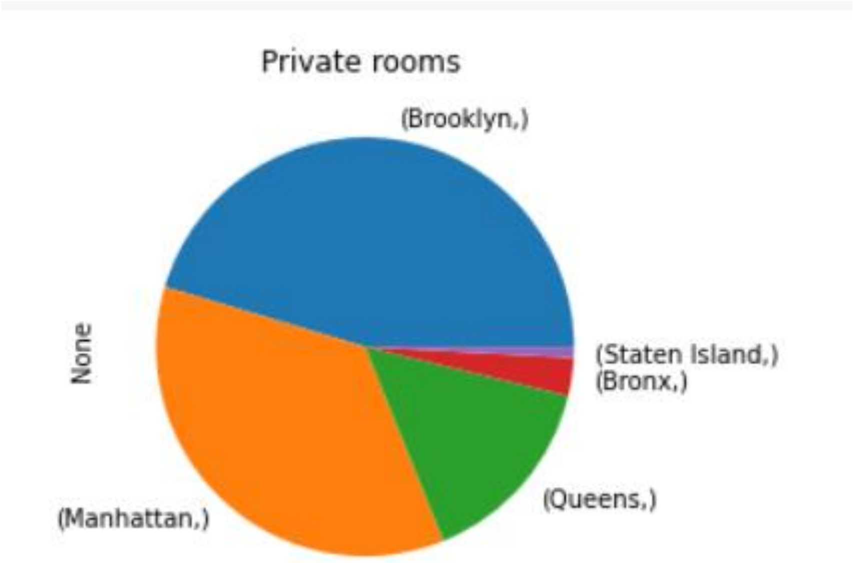
The top 10 most reviewed listings on Airbnb bookings for NYC has average price of 65 dollars. Most of the listings has average price below 50 dollars. 9 out of 10 listings are private rooms type. The top reviewed listing has 629 reviews.

Hotel near to Neighbourhood group



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.

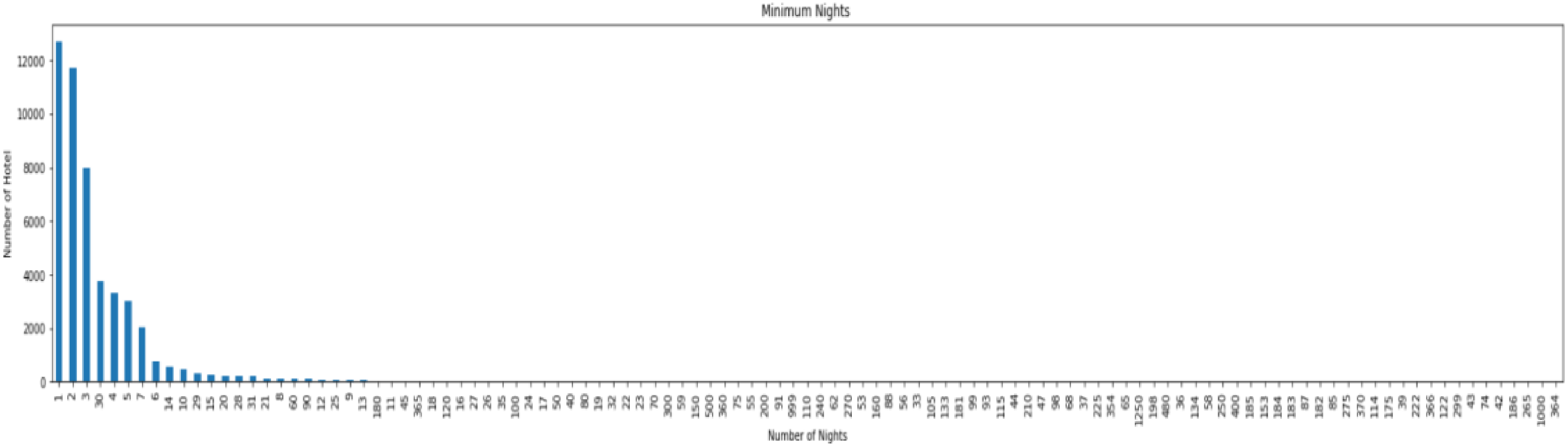
Private room in all neighbourhood



- 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.

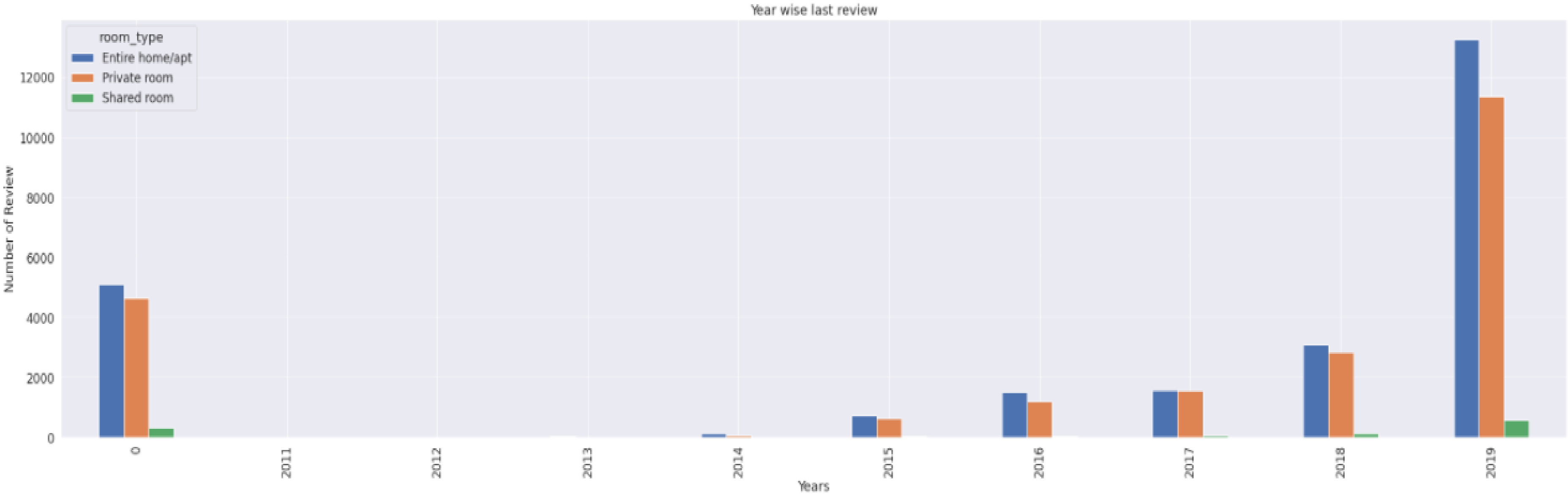
- The total number of hotels that have private rooms is 22,326.

Minimum nights



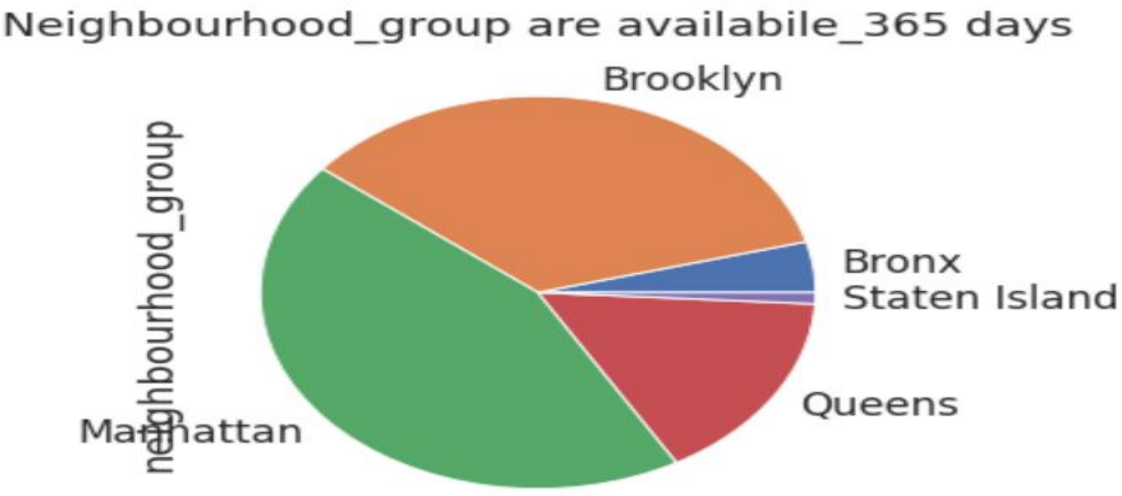
The above bar graph shows the minimum nights available for all the hotels.,In the above graph. The x-axis shows the number of nights and the y-axis show number of hotels. As per the graph maximum hotels(12720) have 1 as minimum nights.

Year wise types of rooms getting a last review



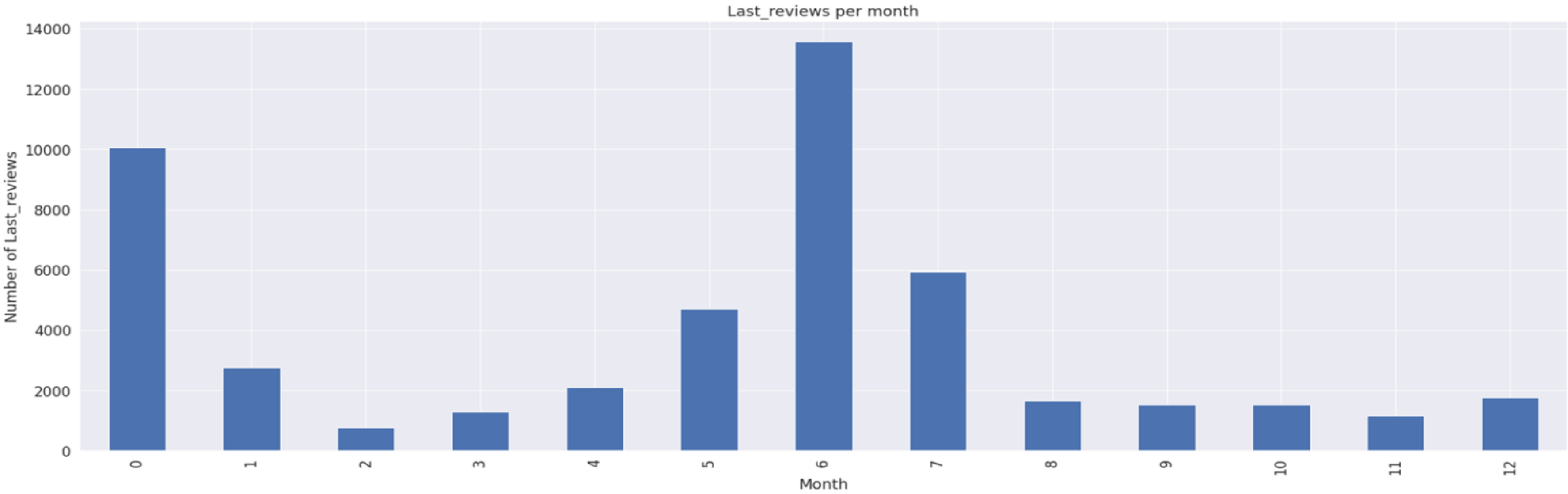
The above bar graph shows the types of room getting a last review in year wise. As per the graph from 2011-2014 very less number, 2015-2018 rooms are slightly increase. In 2019 we can see more number reviews. Year 0 represents the missing year's in Dataset.

Neighbourhood group are available all days



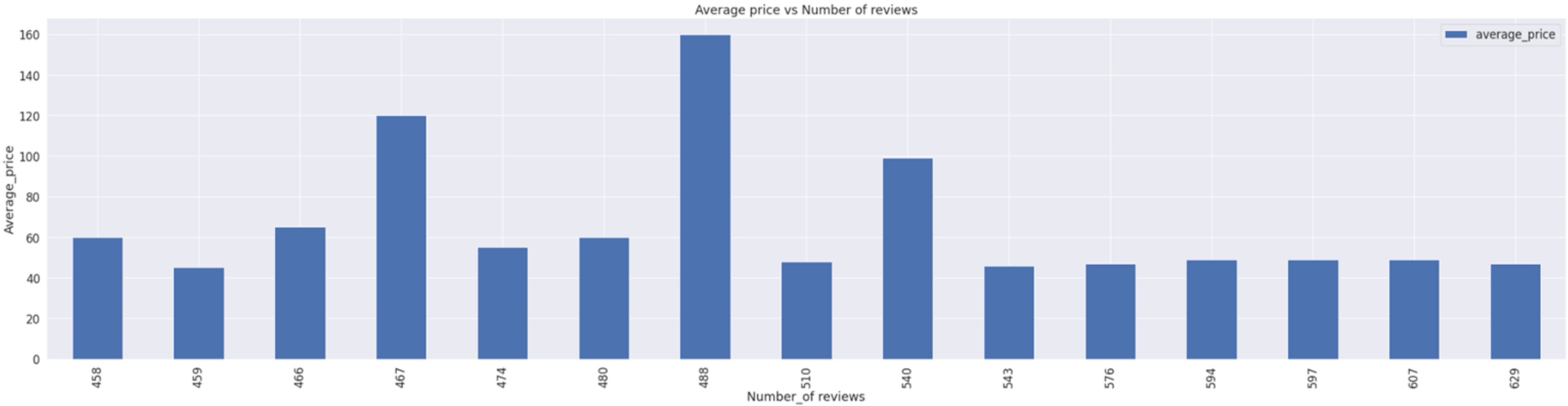
The above Pie chart shows the how many of neighbourhood group are available in all days of year. Highest number Neighbourhood group is Manhattan (572) and lowest is Staten Island (12).

Last reviews trend in particular months



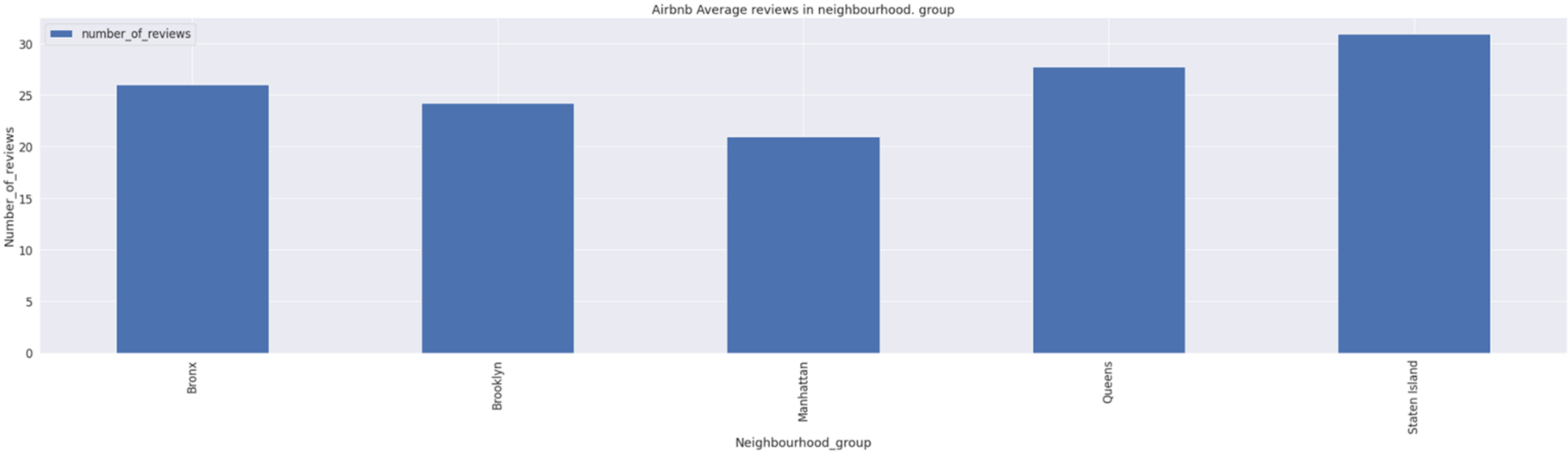
In the above Bar graph shows the last reviews of every month in all year’s. The month 6(June) is getting the most last reviews in all year.

Highest average price corresponding to total number reviews



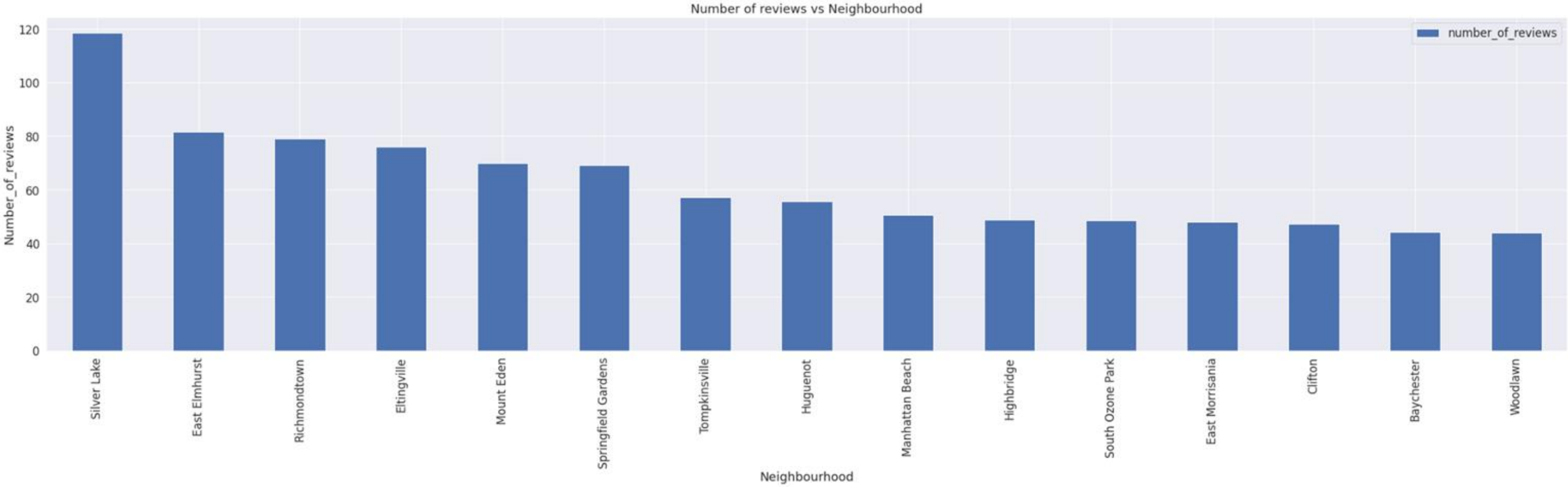
The above bar graph shows the average price according to the number of reviews available in the Airbnb data. As per the graph the highest average price of 166 have 488 total numbers of reviews.

Average number of reviews among neighborhood group



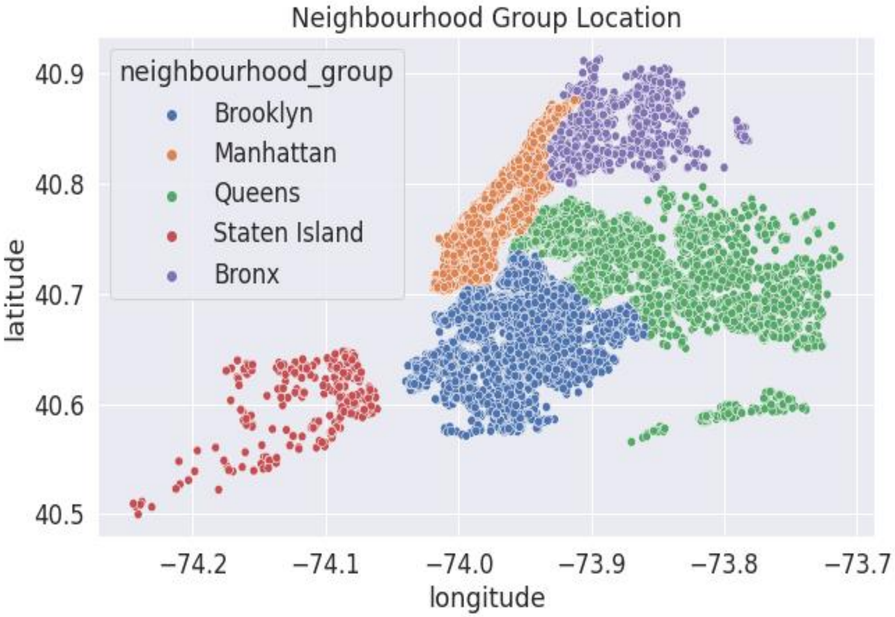
The above bar graph indicates the average number of reviews as per different neighborhood group available in the Airbnb data. As per the graph above the highest number of reviews comes under the Staten Island neighborhood group which is having the 30.97 of average reviews in the airbnb data.

Average number of reviews comes in the neighborhood

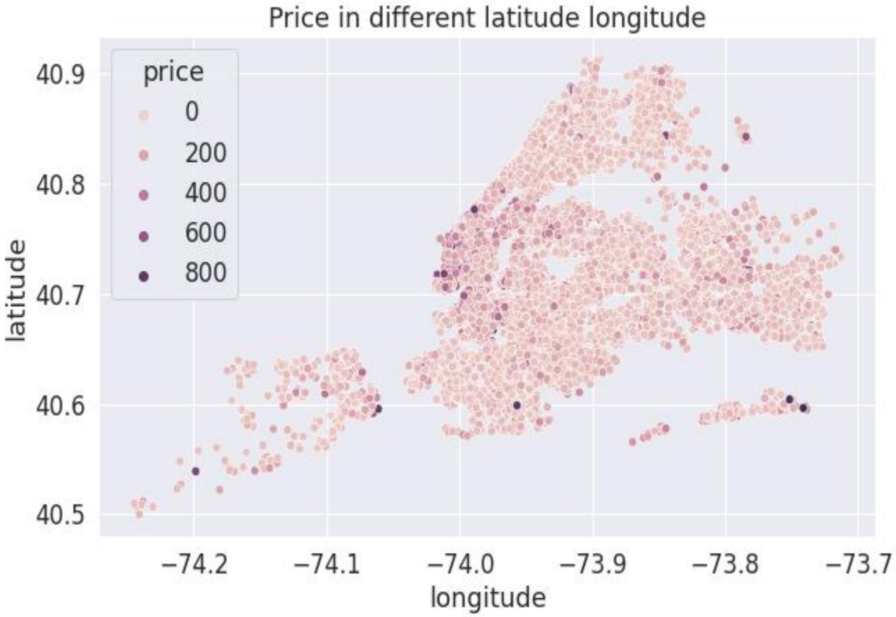


The above bar graph indicates here the average number of reviews vs the Neighborhood available in the airbnb data. In the above graph the Silver Lake has the maximum number of reviews among all the neighborhoods.

Price distribution corresponding to neighborhood group



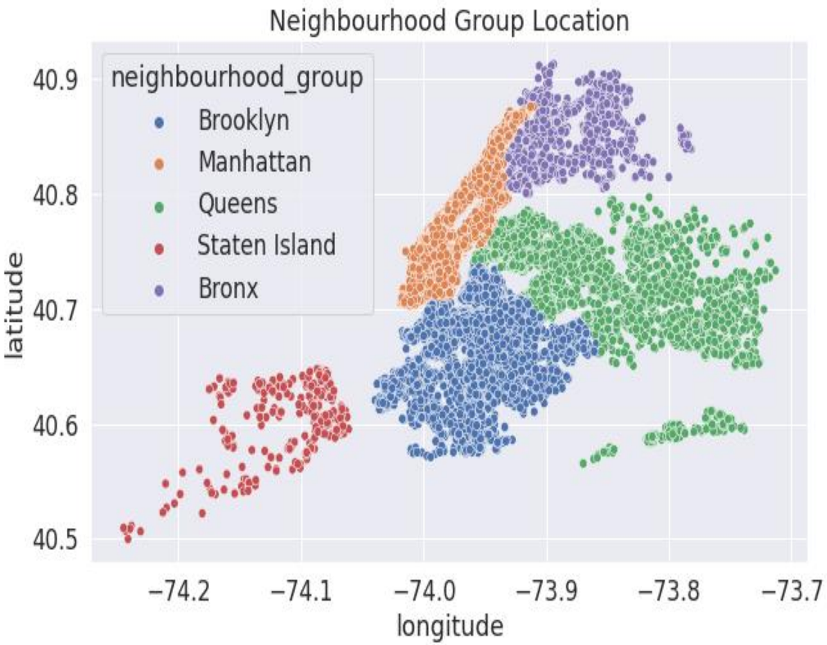
Neighbourhood Group Location

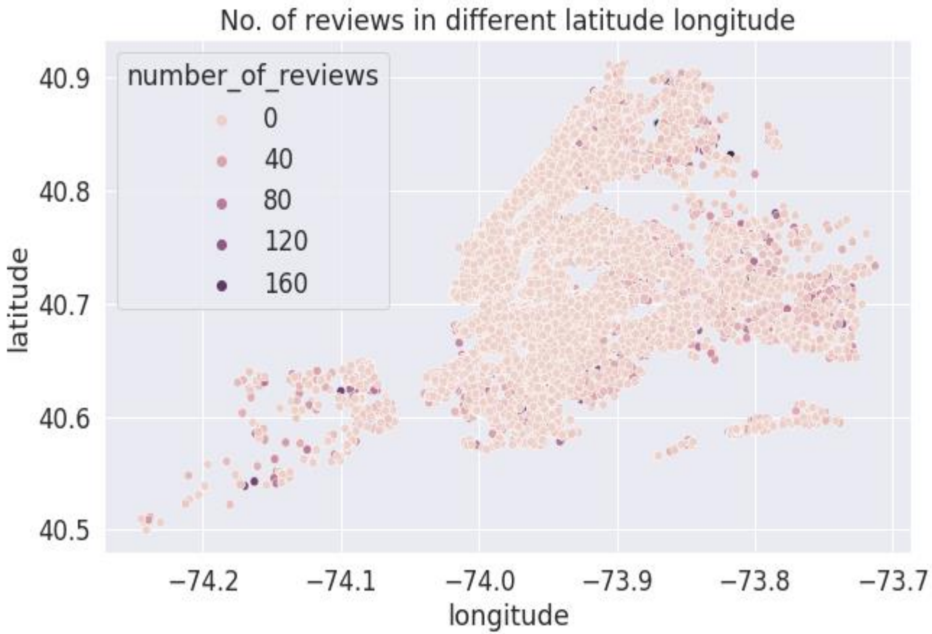


Price in different latitude longitude

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

So, Airbnb dataset is a rich in data but not on features. From the entire above analysis we can conclude that,

- Sonder is the busiest host on Airbnb. Though Michael did maximum entry after that Sonder, when we check the host listing count Sonder comes first before Michael.
- Shared rooms are less available compared to other room types but when we check the number of reviews we find out those private and shared rooms are preferred more compared to the entire room. So it means we have to increase the availability of room type other than entire must be increased to gain profit.
- The number of features positively correlated with each other. The correlation between the number of reviews and reviews per month is high. There is a correlation between calculated listing counts, minimum nights and availability 365 also.
- As skewness and kurtosis are high means a good amount of outliers are present in the price features. So to create uniformity in the data, remove the outliers by a boxplot.
- The price of private and shared room types is less compared to the entire room type. That’s why most of the visitors want to book private and shared rooms type.

Conclusion..

- The Topmost reviewed room is the private type which has a price below 50$. It means users prefer cheap rooms.
- Manhattan has the highest number of hotels which have availability 365 then Brooklyn comes, as Manhattan is famous for museums, stores, parks and theatres that's why more hotels are available.
- If we see closely the private room type is highest in Brooklyn then Manhattan comes. So as we discuss before private room type is high in demand and also this place near to Manhattan so visitors can visit both in cheap prices.
- Maximum hotels consider one minimum night which is good for visitors to stay accordingly.
- Most of the last reviews came in June month which means we can increase the price of Airbnb as most visitors visit this month.
- Staten Island has the maximum number of reviews than Queens that's why most of the rooms are not available all 365 days compared to Brooklyn and Manhattan.
- The Silver Lake in Staten island has more reviews than other places in NYC.
- Most of the prices are less in Staten Island and Queens because the number of reviews was more in these places than in other boroughs. It means we can increase the price to increase profit.

Thank You!!