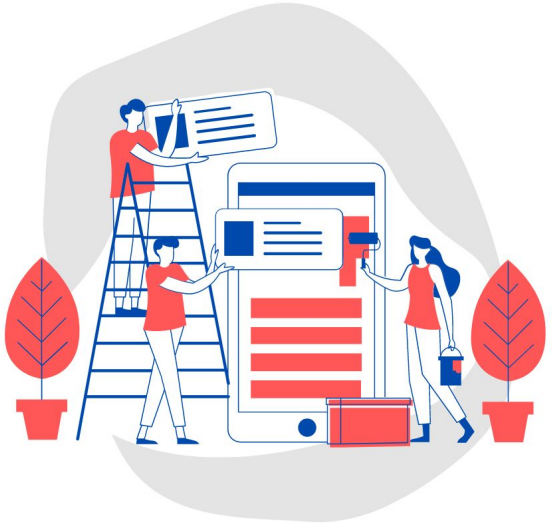
Capstone Project-1

Airbnb Bookings Analysis

Individual Project by:
Anish Johnson

Contents:

1. Introduction.
2. Objective of the analysis.
3. Preview of the provided Dataset.
4. Exploratory Data Analysis.
5. Inferences and conclusions.



Introduction:

What is Airbnb?

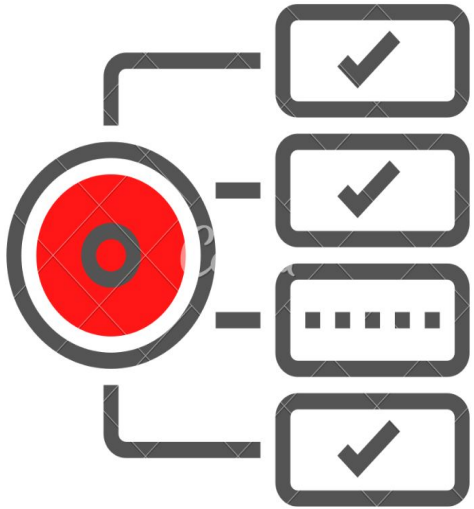
Airbnb is essentially an online marketplace that involves the renting of property to travellers. Airbnb does not own any of the properties. It simply provides a platform from which people can rent out their properties or spare rooms to guests. Prices are set by the property owners and money is collected via the Airbnb app.

How does it operate?

Objective:

We have been provided with a dataset containing around 49,000 observations in it regarding the Airbnb Bookings.

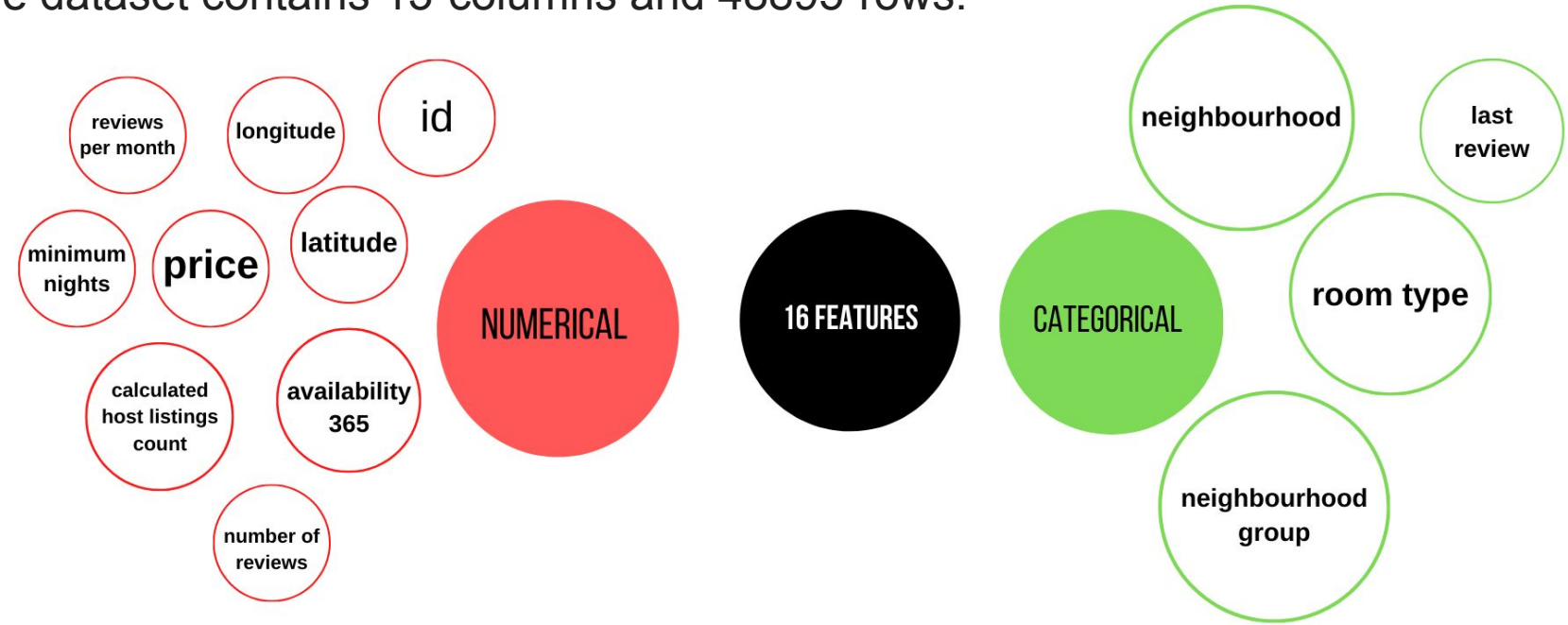
Our objective is to explore and analyse the data to discover key understandings and provide helpful conclusions about the dataset through Exploratory Data Analysis, which then could be used in decision making and model building for further predictions.



Preview of the data:

After dropping the unwanted columns ['name', 'host_id', 'host_name']

The dataset contains 13-columns and 48895-rows.



It's always better to handle the null values before starting with further analysis in order to get best results.

- Check for null values showed that the columns **last_review** and **reviews_per_month** contain many null values that need to be removed.

- To do so we utilize the dropna() function in python to remove rows containing the null values from the data.

price	0
minimum_nights	0
number_of_reviews	0
last_review	10052
reviews_per_month	10052
calculated_host_listings_count	0
availability_365	0

Before removing null values

price	0
minimum_nights	0
number_of_reviews	0
last_review	0
reviews_per_month	0
calculated_host_listings_count	0
availability_365	0

After removing null values

Now let's get some more details from our data.

	Price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count
75%	170.000000	4.000000	33.000000	2.020000	2.000000
max	10000.000000	1250.000000	629.000000	58.500000	327.000000

Summary using the .describe() function in python.

- A brief summary for the data showed the huge difference between the 75th percentile and the max values of these features which indicates the presence of outliers in the data.

- Also the minimum **price** is given as zero, which is impossible unless Airbnb plans to provide rooms free of cost. To resolve this problem let's replace these values.

count	48895
mean	152
std	240
min	0
25%	69
50%	106
75%	175
max	10000

- To replace the price values which are zero lets swap it with the average of the minimum prices paid for 1 night multiplied by the total nights stayed, i.e
- **[price = avg(min_prices)x(total_nights_stayed)]**.

- We assume that the minimum prices to be paid for 1 night ranges between (0-100) and get an average of all the prices less than 100$.
- There are total 19 rows in which the price 0 has to be replaced with the `min_avg_price` that we found to be 65$.

- As we can see the minimum price has changed to 10$ after replacing with `min_avg_price`.

mean	142.448472
std	197.363353
min	10.000000
25%	69.000000

Exploratory Data Analysis:

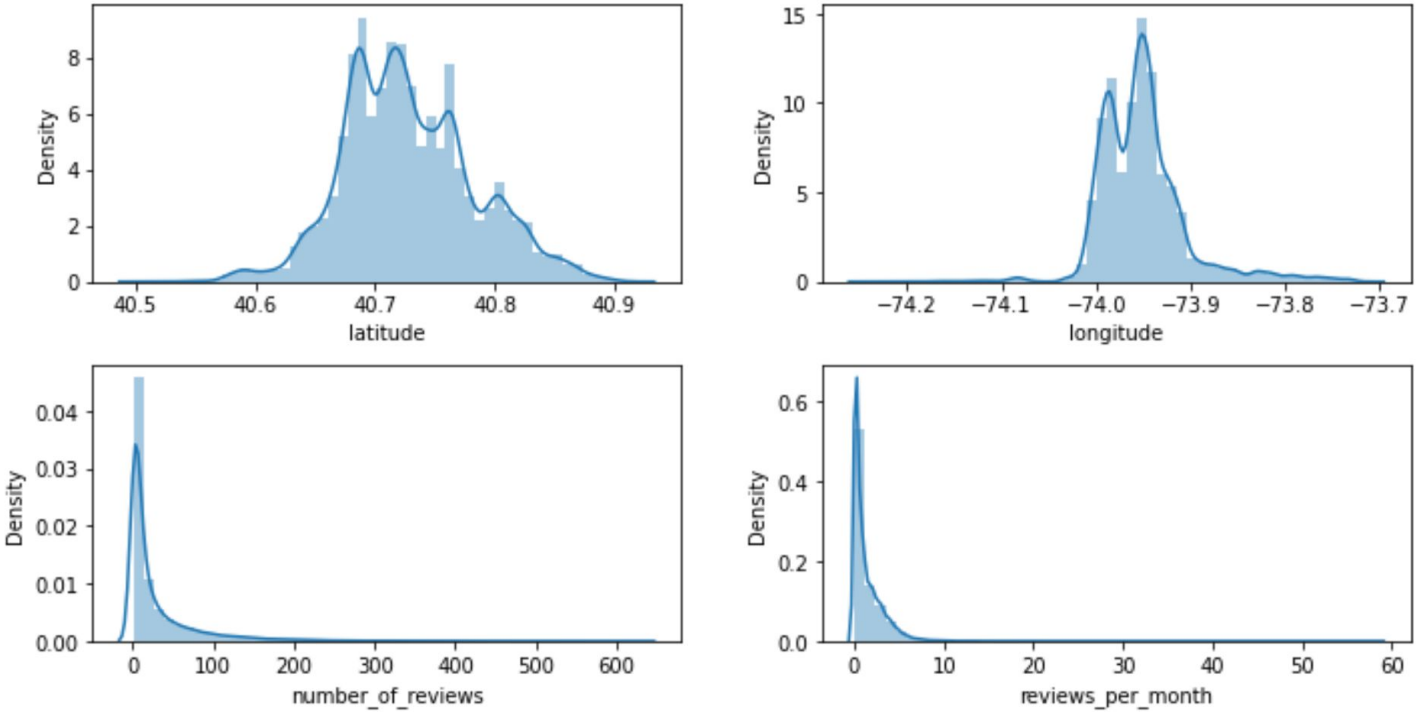
What is Exploratory Data Analysis?

Simply defined, EDA is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

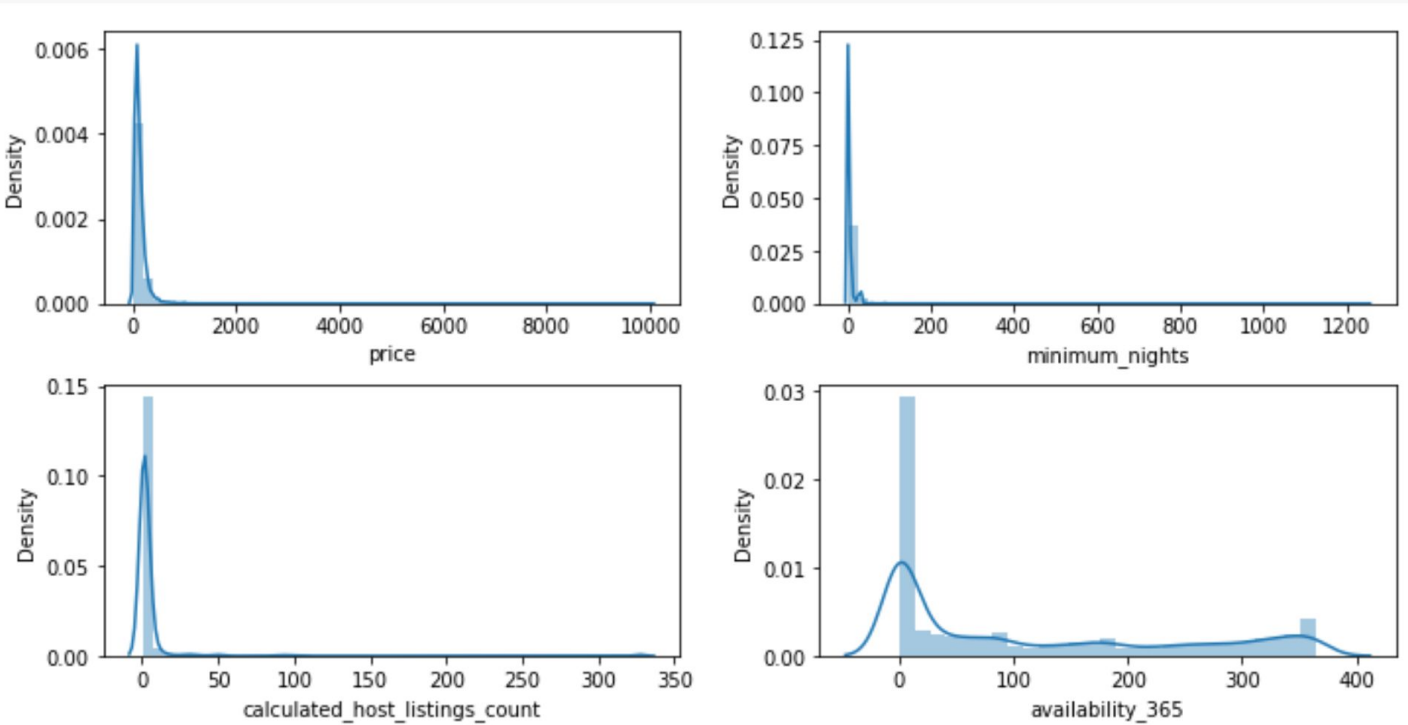
Why is EDA important?

Explore Data
Helps to discover trends,
Visualize the data,
Understand the features, etc.

Distplots: Numerical Features



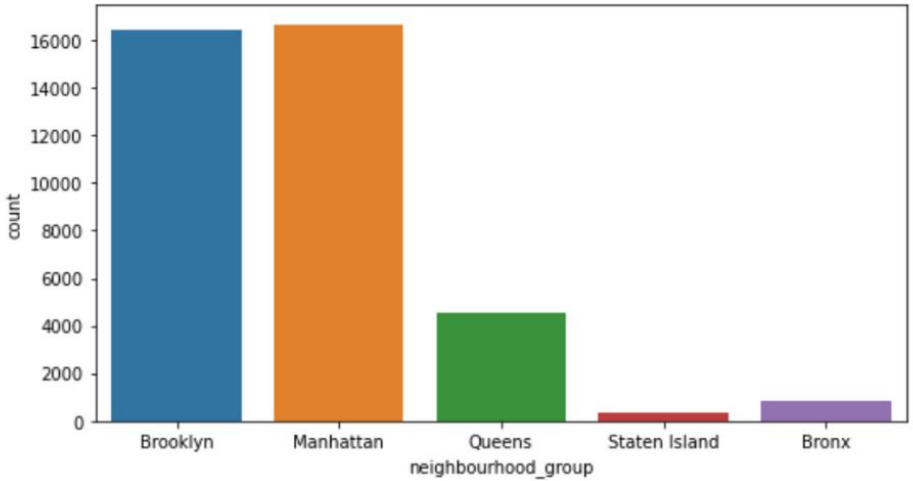
Longitude and Latitude seem normally distributed whereas severe positive skewness is present in number_of_reviews and revies_per_month.

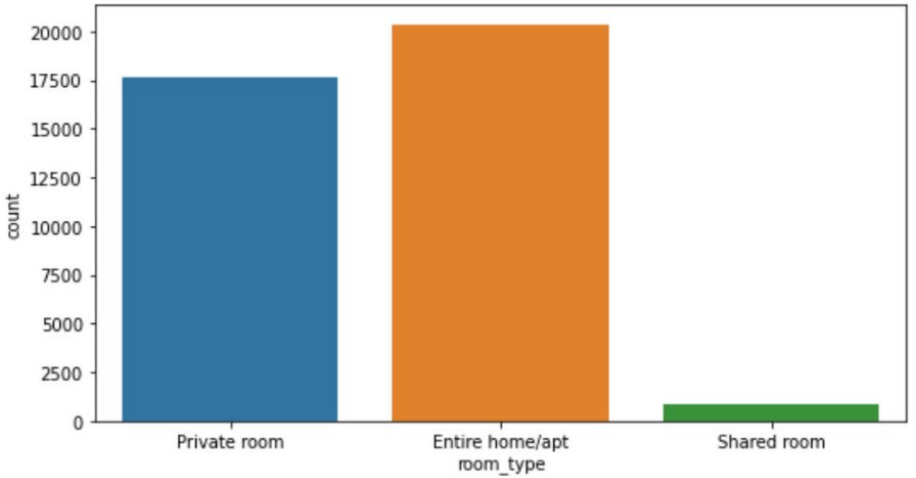


Severe positive skewness is present in all the above features.

This can be treated using log transformation on the these features.

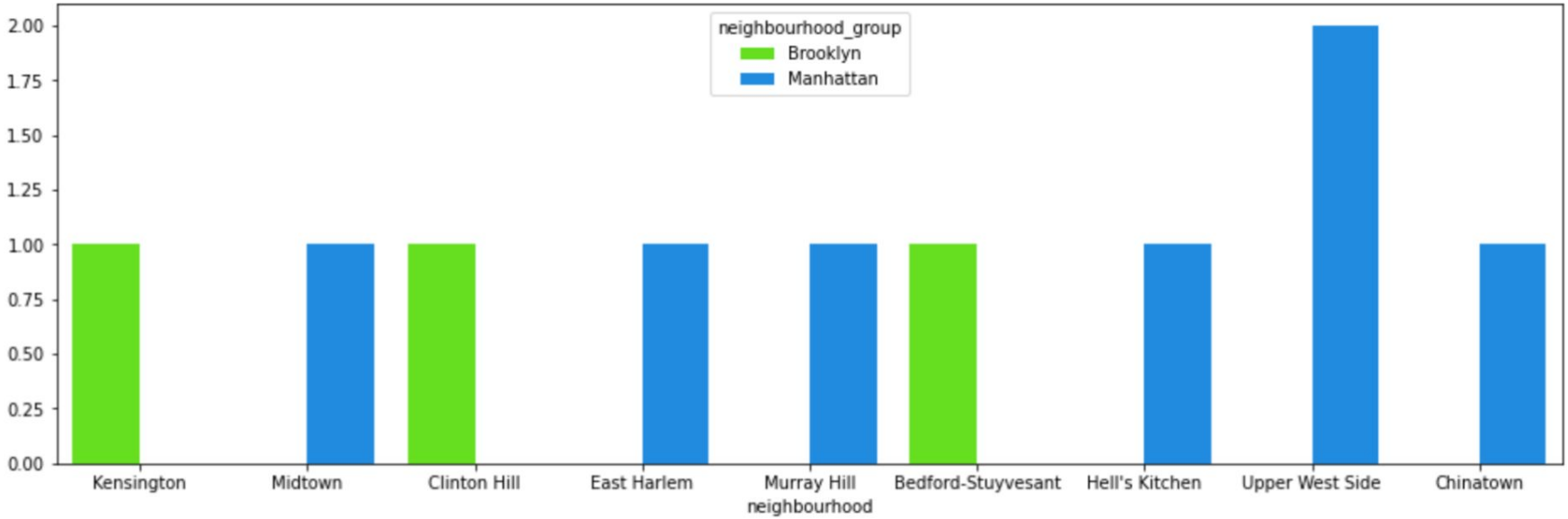
Barplots: Categorical Features





Most preferred neighbourhood_group are Manhattan and Brooklyn followed by Queens whereas Staten Island and Bronx are least preferred.

From the three types of rooms provided Private rooms and Entire home/apt are most booked from all of these locations.



As we can see the top 10 neighbourhoods in are data belong to Manhattan or Brooklyn, making these places most popular.

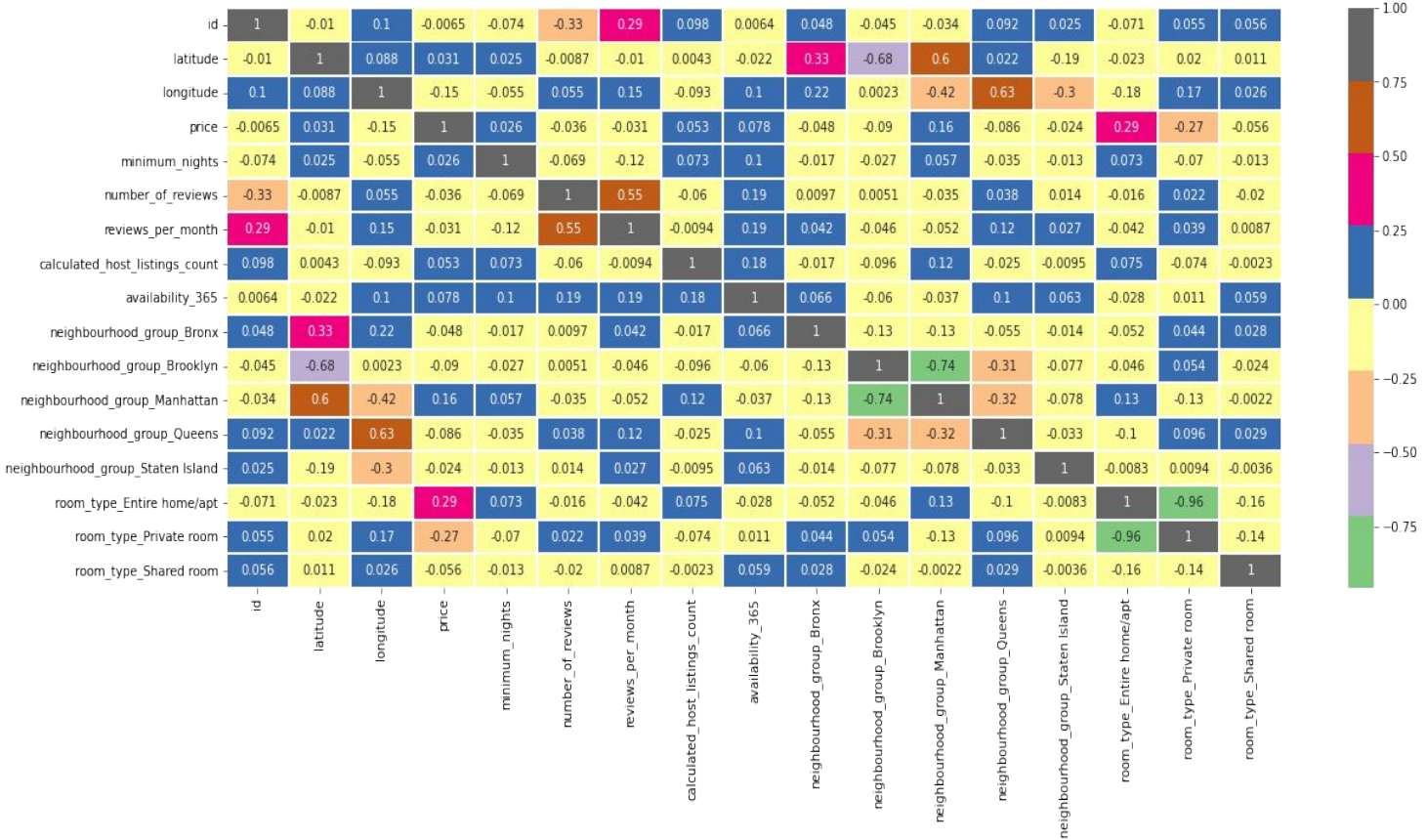
Correlation Heatmap:

It's difficult to understand the heatmap in a first look, so let's break it down.

If the color of the box moves towards grey the features are positively correlated, if moving towards green then negatively correlated.

While building a model positive correlation is preferred whereas the variables with negative correlation are dropped.

This helps a lot to understand the data and selecting the required features.

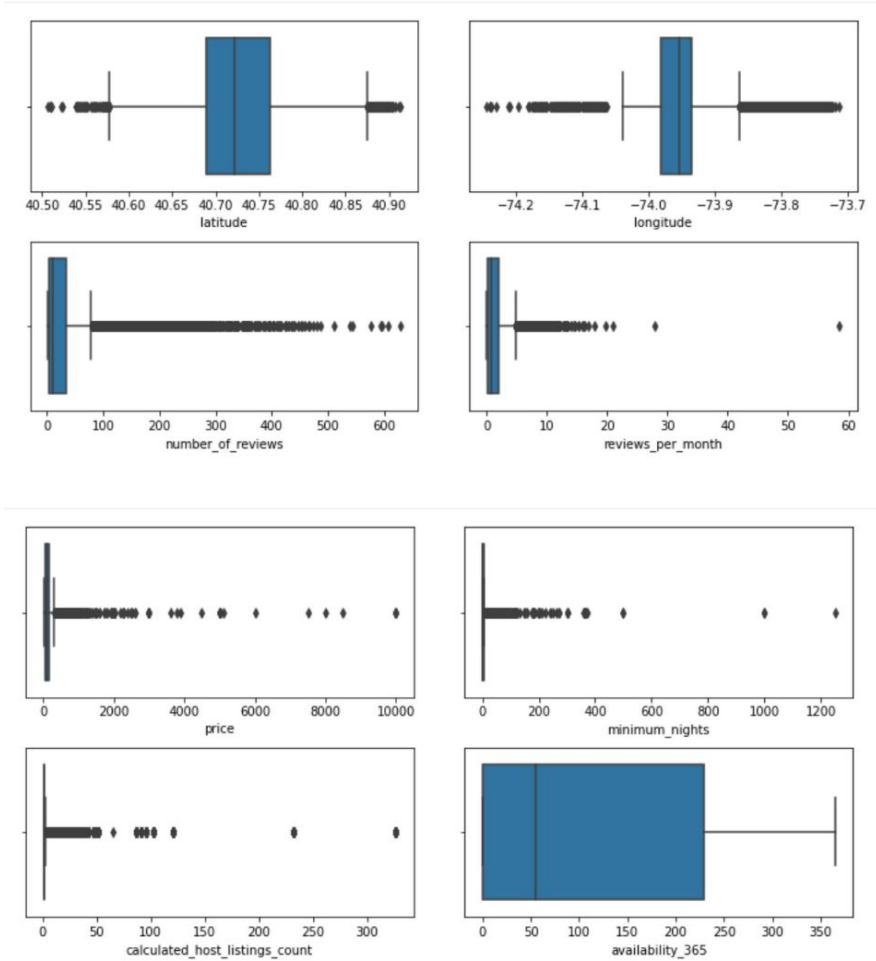


Outliers:

As we had seen earlier few features from our data had outliers, this boxplot proves it.

There are many outliers present in these features that need to be handled.

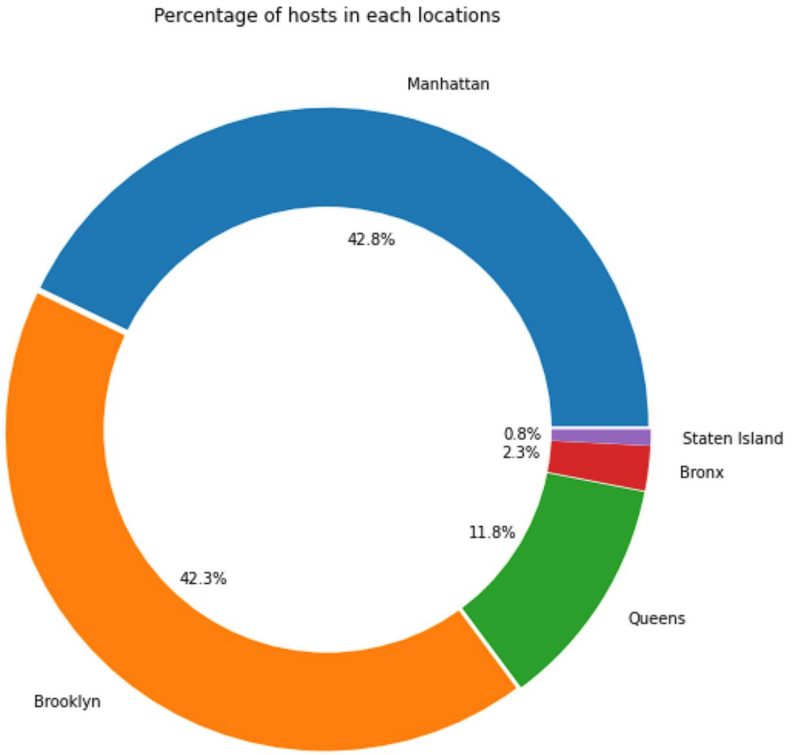
These can be treated either using the IQR or the Z-score.

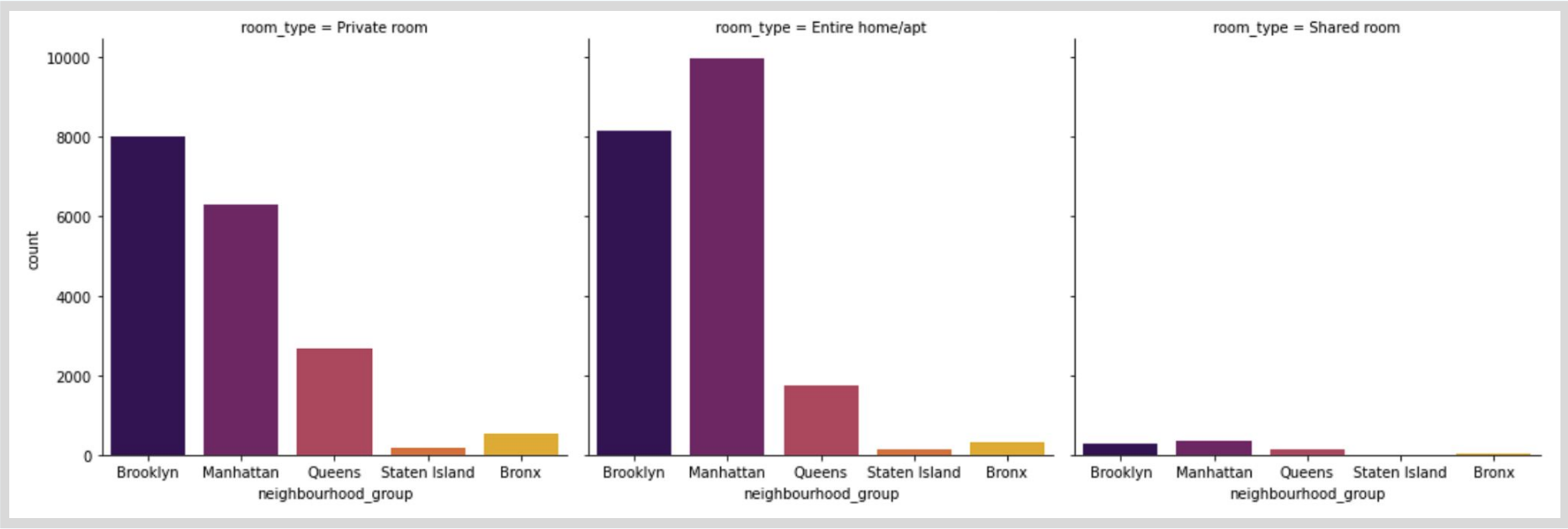


What can we learn about different hosts and areas?

Following points can be understood from the pie chart:

- Manhattan and Brooklyn are home to 85.1% of the hosts followed by Queens with 11.8%.
- Whereas Bronx and Staten Islands are occupied by only 3.1% of the hosts.
- This makes clear that most of the hosts prefer Manhattan or Brooklyn for their business.
- This also shows that people prefer either Manhattan or Brooklyn for their stays.



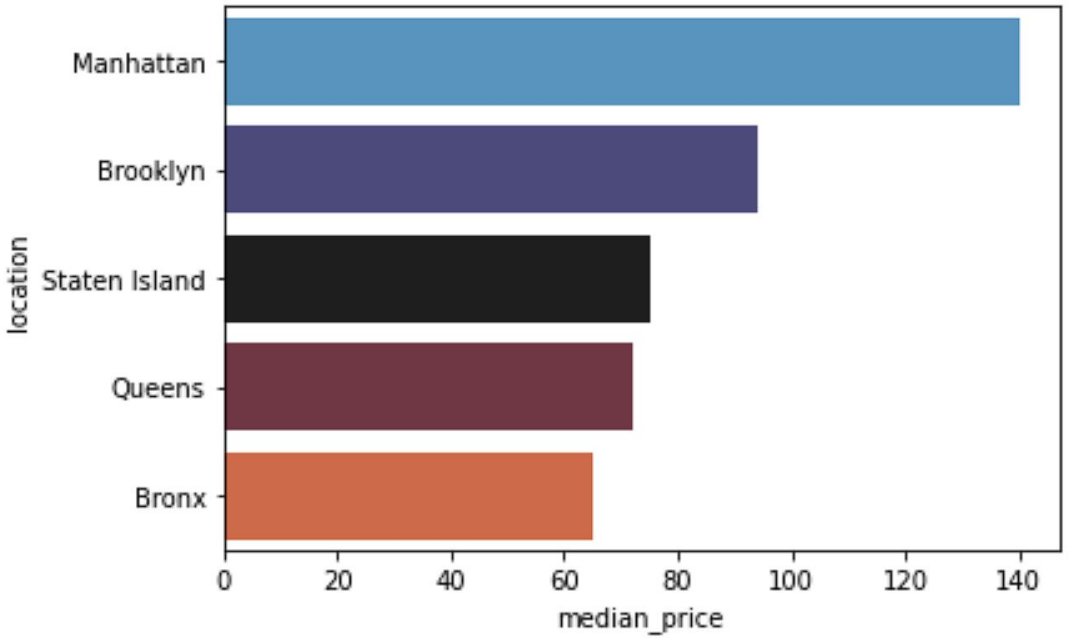


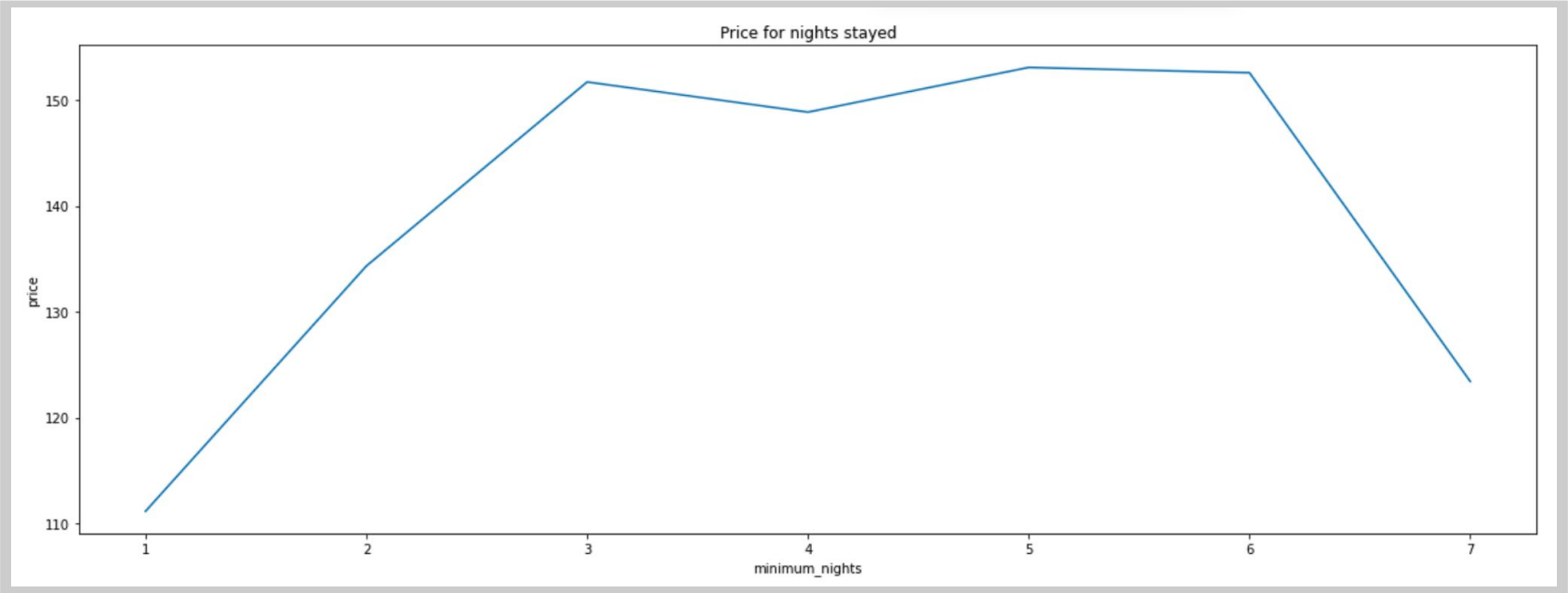
- From the three room_types provided by the hosts, Private rooms and Entire home/apt are more preferred and Shared rooms are least preferred.

- Even here Manhattan and Brooklyn are the major players followed by Queens.

What can we learn from predictions?

- Since the data is skewed we will choose median in order to calculate the price difference according to locations, as mean may give bias results.
- As shown in the barplot Manhattan and Brooklyn are more expensive than the other three locations.
- This indicates more demand for the top two locations.

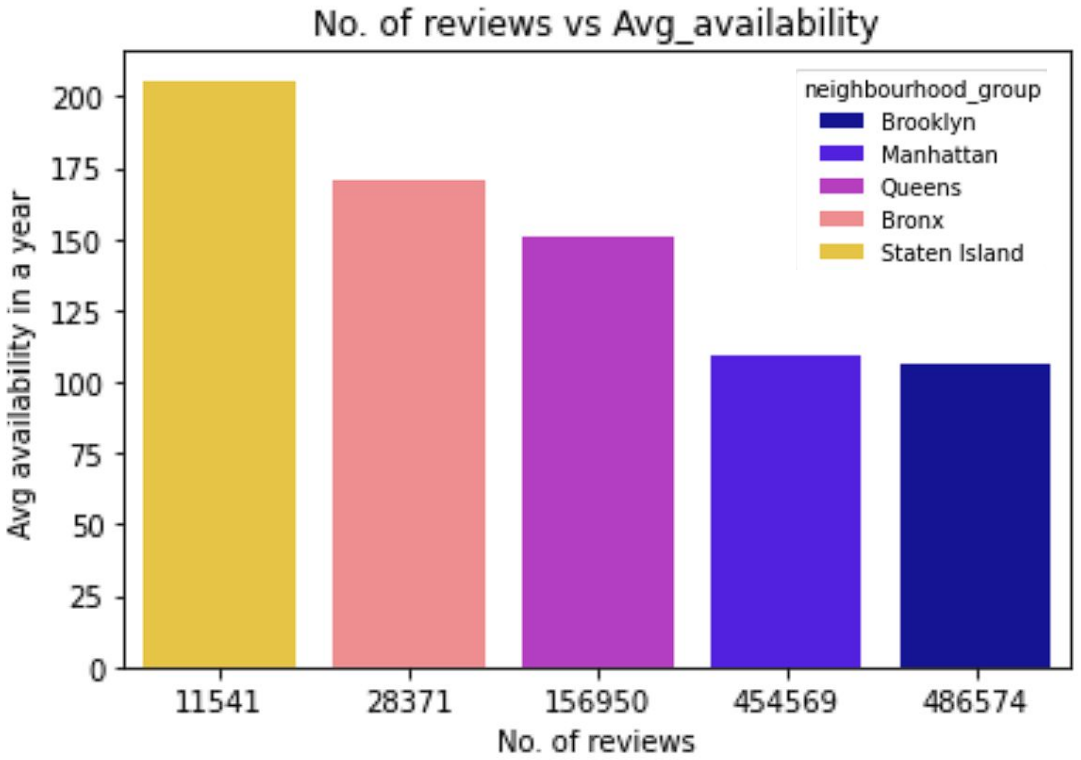




It can be observed that the price is higher when the number of nights stayed is less than a week and gradually decreases as the number of nights stayed increases. Which implies that one would have to pay comparatively higher price when booking a room for less than a week.

Which hosts are the busiest and why?

- We have already seen that Manhattan and Brooklyn are the top two locations booked by people.
- This plot shows how number of review affects the availability of that particular host throughout the year.
- As the number of review increases the average availability in a year decreases.
- Which implies that hosts in Manhattan and Brooklyn i.e the locations where hosts receive maximum reviews are most busiest through the year.

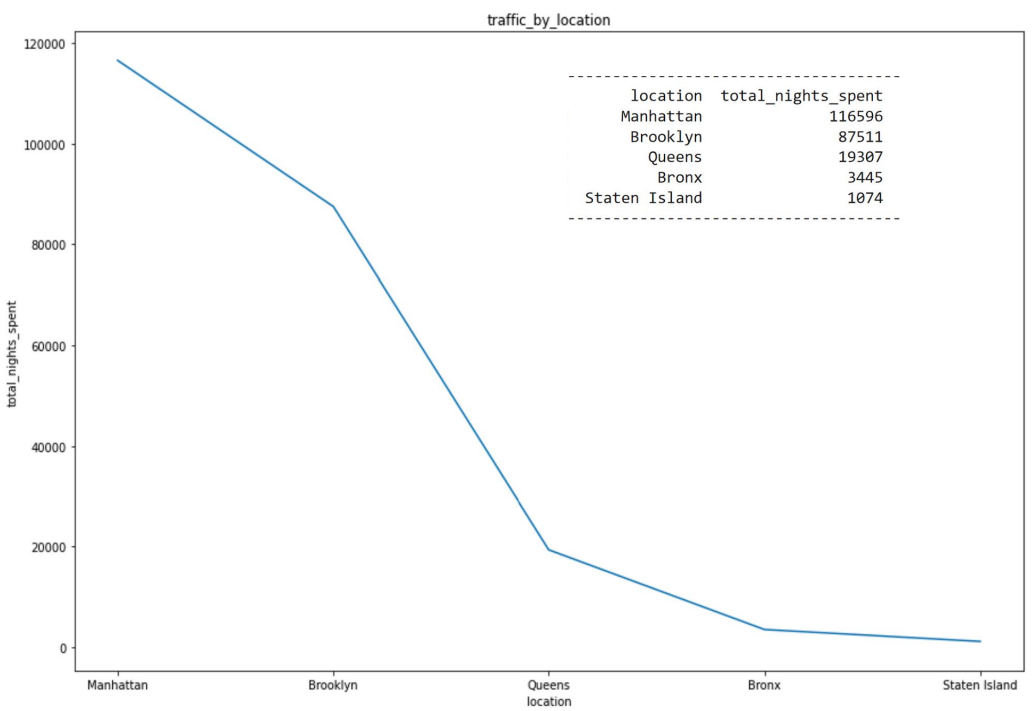


Is there any noticeable difference of traffic among different areas and what could be the reason for it?

- To understand the traffic among different areas we can look at the given plot:
- As shown total_nights_spend are comparatively higher in Manhattan and Brooklyn and it drops drastically once it reaches Queens.

Which Implies:

- Traffic in Manhattan and Brooklyn are higher, which could be due to their demand among the customers.
- Traffic is very low in locations like Queens, Bronx and Staten Island, which might be due to fewer number of hosts present in these areas and also the lack of services that are available in these locations like the preferred room_type.



Conclusions and Inferences:

- Drop unwanted columns ['name', 'host_id', 'host_name'] as these either contain high cardinality or reveal sensitive information about the host.
- Null values are present in **last_review** and **reviews_per_month** which can be dropped.
- Price values are 0 for 19 of the rows which have been replaced with the 65$ that we computed assuming that min avg price paid for one night was in the range 0-100 and multiplying it with total nights stayed.
- Distplots for the numerical features show severe skewness in the data which can be treated with log transformation on the respective features.
- Barplots for the categorical fetures revealed Manhattan and Brooklyn as the most preferred locations with Private rooms and Entire home/apt as most sought out room_types. Also the top 10 neighbourhoods belong to these two locations making them the most popular destination among the users.

- Multicollinearity among the variables can be determined using the heatmap, which indicates high correlation between:
  1. number_of_reviews and reviews_per_month
  2. latitude/longitude and neighbourhood_groups

  If going ahead with linear regression multicollinearity has to be eliminated.

- Most of the numeric features contain outliers which can be observed using the boxplot and can be removed using the IQR method or using the Z-score.

- By grouping different hosts and areas it is observed that 85.1% of the hosts are present in Manhattan and Brooklyn followed by Queens with 11.8% and the rest of locations having 3.1% of stake. Also business is great for people having Private rooms and Entire home/apt in Manhattan ,Brooklyn and Queens while those preferring shared rooms could go with Bronx and Staten Island.

- Manhattan is the most expensive destination when compared with median prices for all the locations.

- Prices are higher when the number of nights stayed is less than a week and gradually decreases as the number of nights stayed increases. Which implies that one would have to pay comparatively higher price when booking a room for less than a week.

- To understand who are the busiest hosts we compare the total number of reviews received by the particular host with their availability throughout the year. As a result it is observed that as the number of reviews go higher the availability decreases indicating that the busiest hosts are the people receiving the most reviews.

- As expected Manhattan and Brooklyn are the locations with high user traffic with a drastic difference with rest of the locations. Major reasons for this are the number of hosts present in these locations and having the most preferred room_types. Reviews obtained by these locations contribute towards the traffic difference, as people tend to book stays with higher number of reviews.

- For further analysis this data can be used in predicting the future price of a particular location. Also the insight from the EDA can be used to understand which locations might be more profitable to start business and what sells the most in which region.

Challenges Faced:

- Huge amount of data had to be dealt with keeping in mind not to loose anything of value.
- Data contains many outliers which affect the decision making.

Thank you.