Exploratory Data Analysis on Airbnb Bookings

Ganesh Bayas

Data Science Trainee

AlmaBetter, Bengaluru

CONTENTS

1. Introduction

2. Problem Statement

3. Dataset Analysis

4. Plot Analysis

5. Limitation

6. Scope of Improvement

7. Conclusion

Introduction:

Airbnb is an American Company since 2007, it is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales.

The dataset from Airbnb based on NY. NY is amongst the most expensive places to live in USA. We would like to perform an in-depth analysis on one of the most densely populated cities of world. Our dataset is feature rich containing, location with co-ordinates, prices, host name, room types, availability throughout season.

With these features we've done exploratory data analysis and tried to extract information like most expensive places to live in NY, is location really varies with occupancy rate, what type of room people tends to choose most, is there any particular season for tourists or locale when we can follow a surge in prices or occupancy rate of properties etc.

Problem Statement:

Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

We need to explore and analyze the data to discover key understandings (not limited to these) such as:

- What can we learn about different hosts and areas?
- What can we learn from predictions? (ex: locations, prices, reviews)
- 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?

Dataset Analysis:

The dataset contains 48895 observations with 16 features. This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions. Let us look through our features,

- Id: a unique id identifying an Airbnb listing or property
- name: name representing the accommodation
- host_id: a unique id identifying an Airbnb host
- neighbourhood_group: a group of area
- neighborhood: area falls under neighbourhood_group
- latitude: coordinate of listing
- longitude: coordinate of listing
- room_type: type to categorize listing rooms
- price: price of listing
- minimum_nights: the minimum nights required to stay in a single visit
- number_of_reviews: total count of reviews given by visitors
- last_review: date of last review given

- reviews_per_month: rate of reviews given per month
- calculated_host_listings_count: total no of listing registered under the host
- availability_365: the number of days for which a host is available in a year.

latitude and longitude have represented a co-ordinate, neighbourhood_group, neighborhood and room_type are columns of categorical type. last_review is a column of date type; we will convert it as required.

The distribution of numerical columns are as follows,

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

Fig 1. Statistical Distribution of Numerical Features

Other 3 important columns are,

- neighbourhood_group: It contains 5 unique hoods which are Manhattan, Brooklyn, Queens, Bronx & Staten Island.
- neighbourhood: It contains 211 unique neighborhoods.
- room_type: It contains 3 unique room types which are Entire home/apt, Private room, Shared room

The distribution of our numerical columns has positive skewness.

Out of all columns, 4 columns containing null values which are name, host_name (looks like listing name and host_name doesn't really matter to us for now) and last_reviews, reviews_per_month (obviously, if a listing has never received a review, it’s possible and valid). So, those null values have been replaced with 0 during our analysis.

Plot Analysis:

Fig 2. Wordcloud Image

Above word-cloud Image Shows most frequently used words in the dataset.

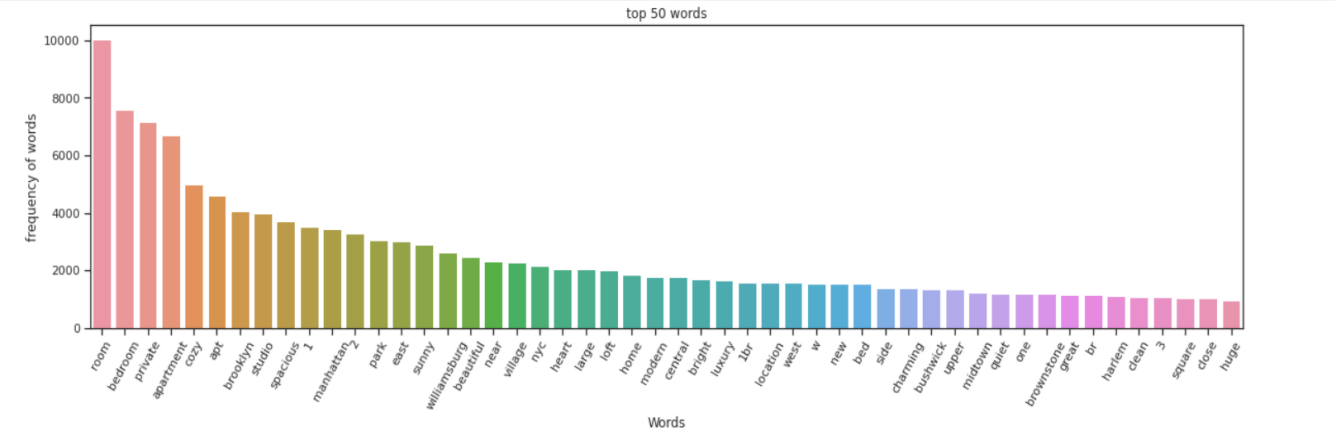

Fig 3.Top 50 words

The above observation shows us the Top 50 words and there frequency. These words can be helpful in model building point of view.

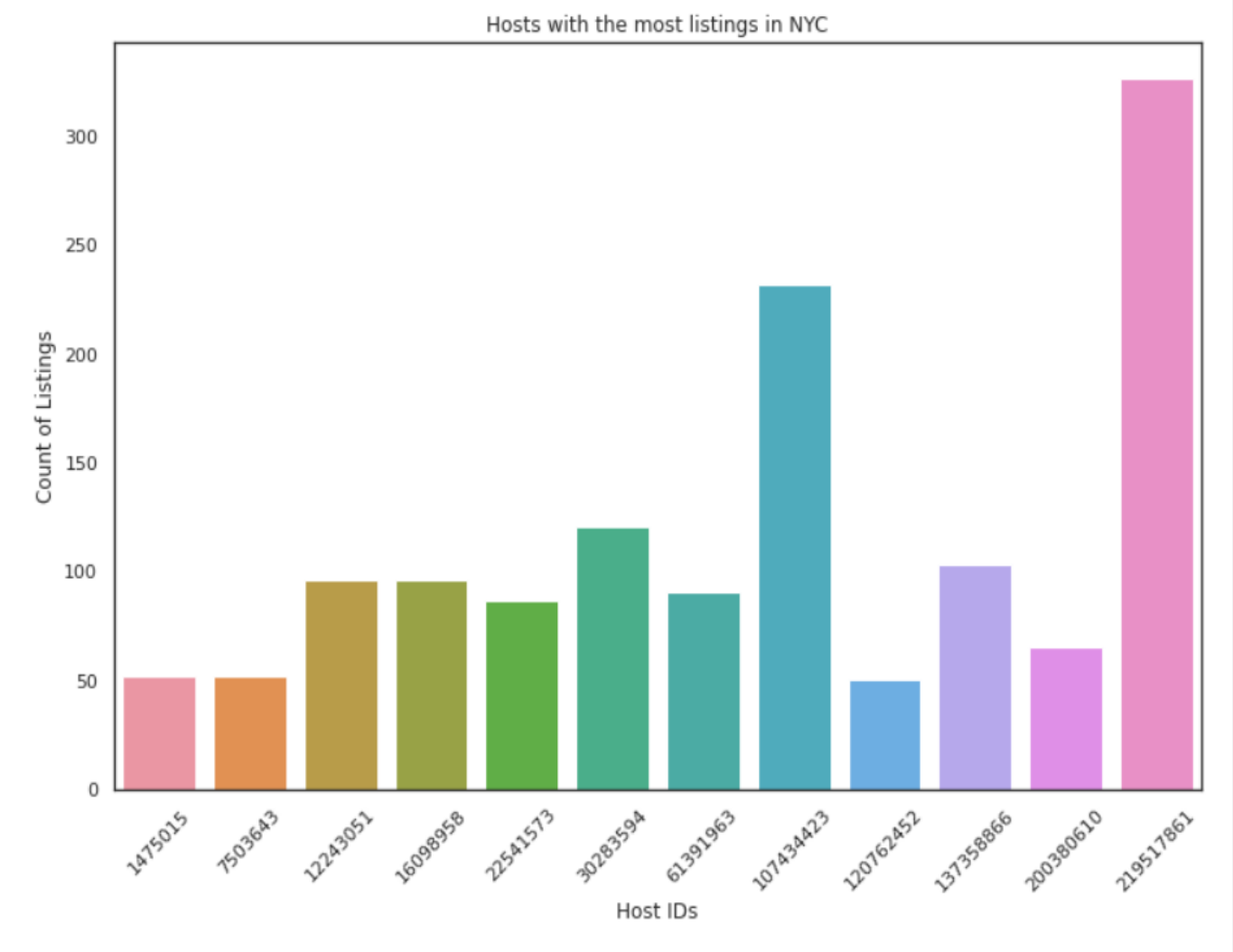

Fig 4.Host listings Chart

The above bar chart shows Hosts with the most listings in NYC it also shows distinct host ids with number of listings.

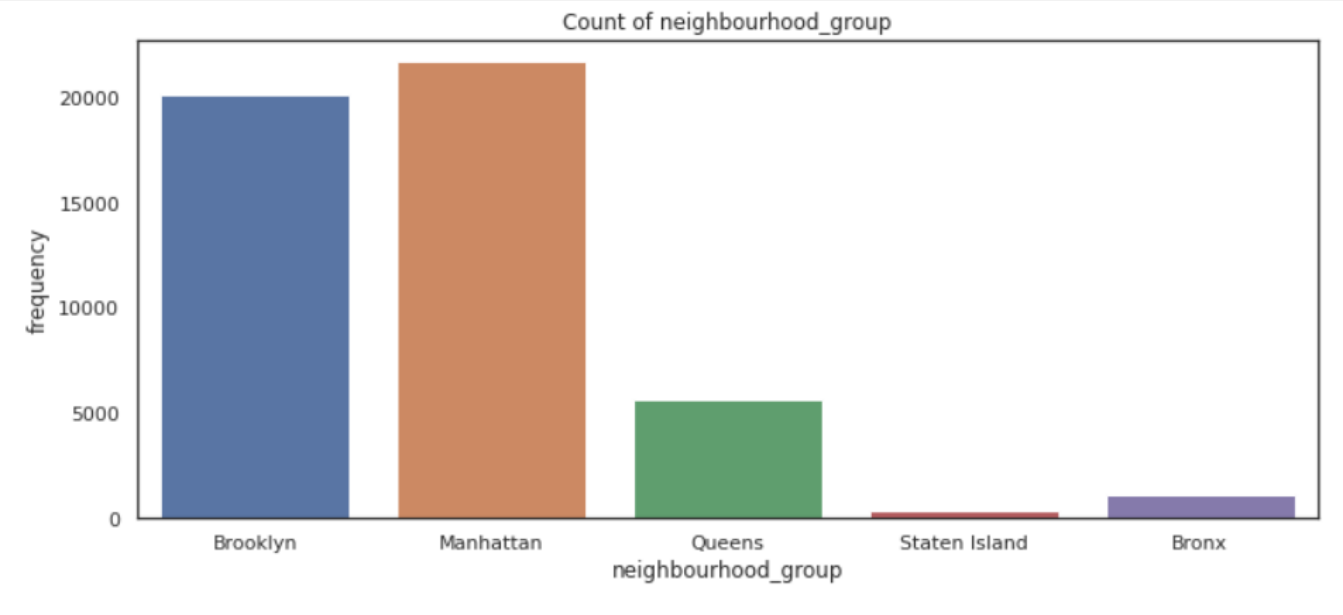

Fig 5. Bar chart of neighborhood data

The above chart shows neighborhood groups booking frequency the highest bookings are in Brooklyn and manhattan.

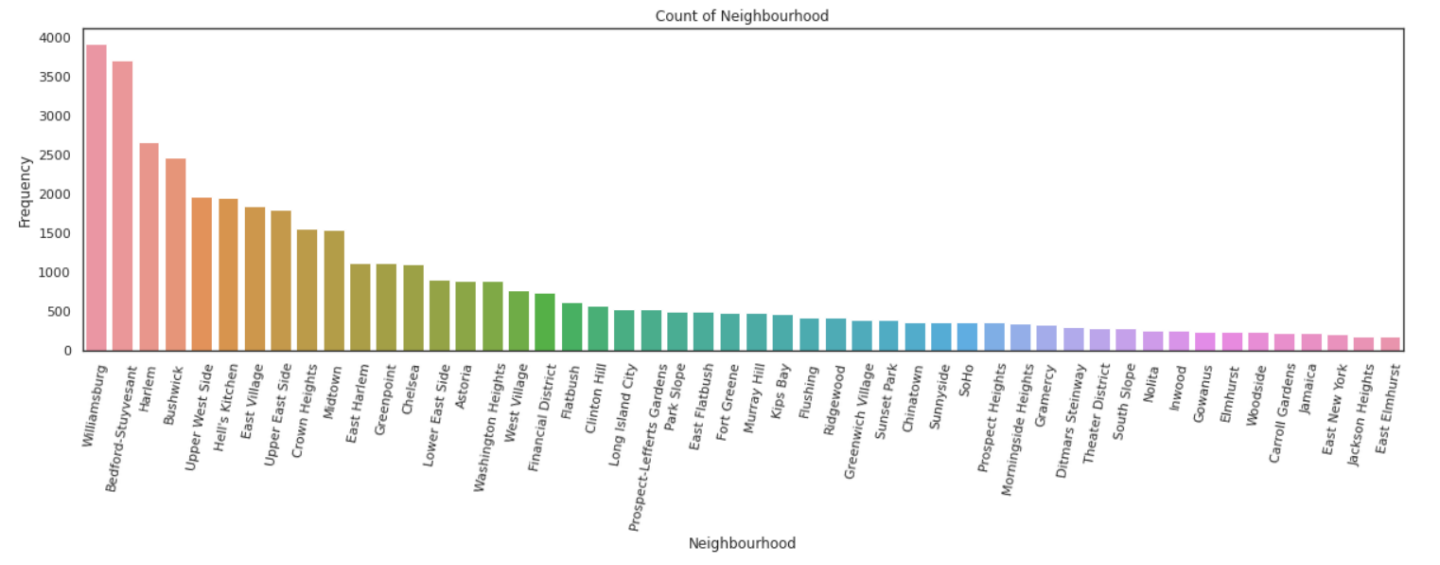

Fig 6. Neighborhood bar chart

The above chart shows us neighborhood towns with the highest and lowest demand from customers across the platform.

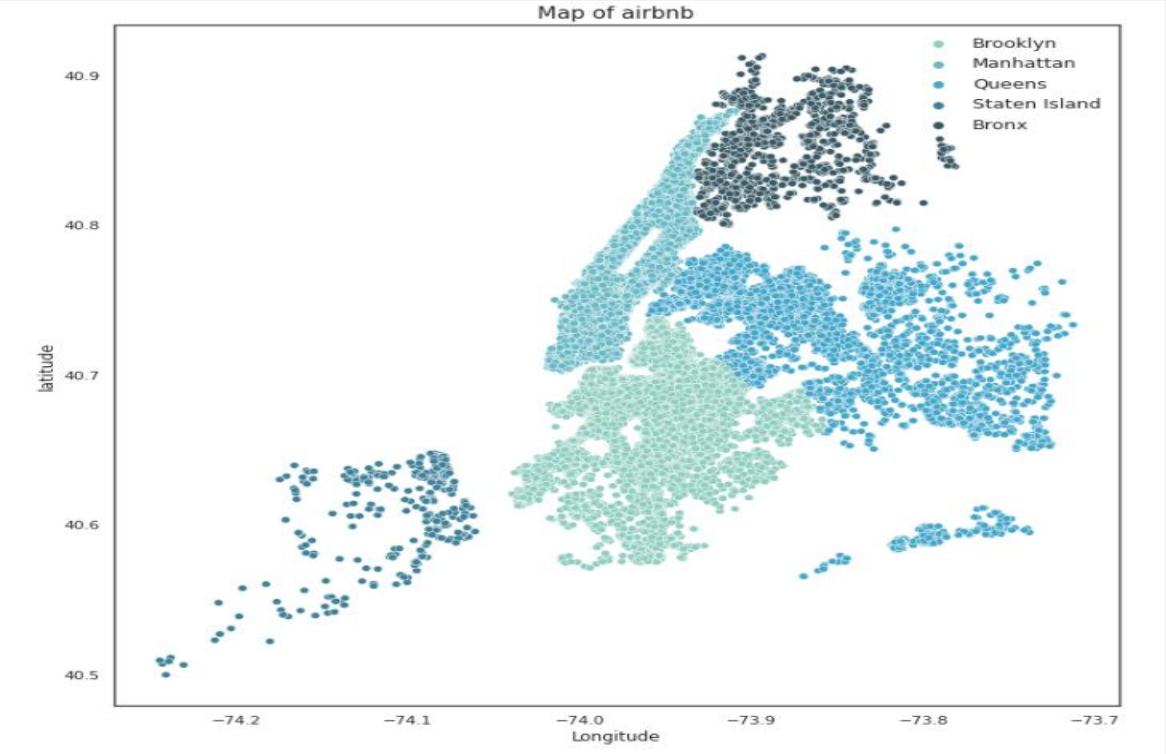

Fig 7. Scatterplot of airbnb listings

The above latitude and longitude visualizes us that Brooklyn and Manhattan are the most dense with hotels and apartments followed by queens island.

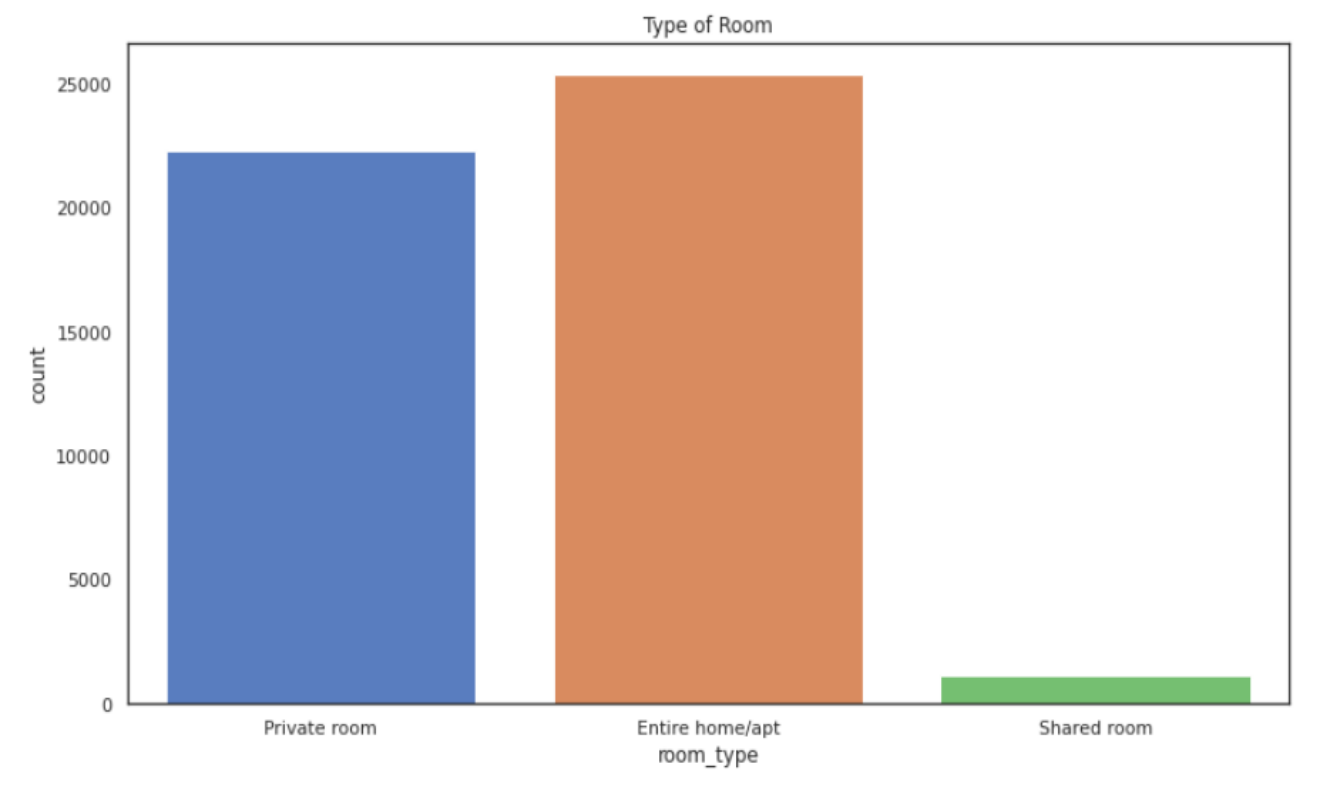

Fig 8.Room type chart.

The above chart shows there are three types of rooms

Namely:

1.Private room

2.Entire home/apt room_type

3.Shared room.

4.People mostly preferred to take whole apartment on rent followed by Private room.

5.very few people preferred to have shared rooms.

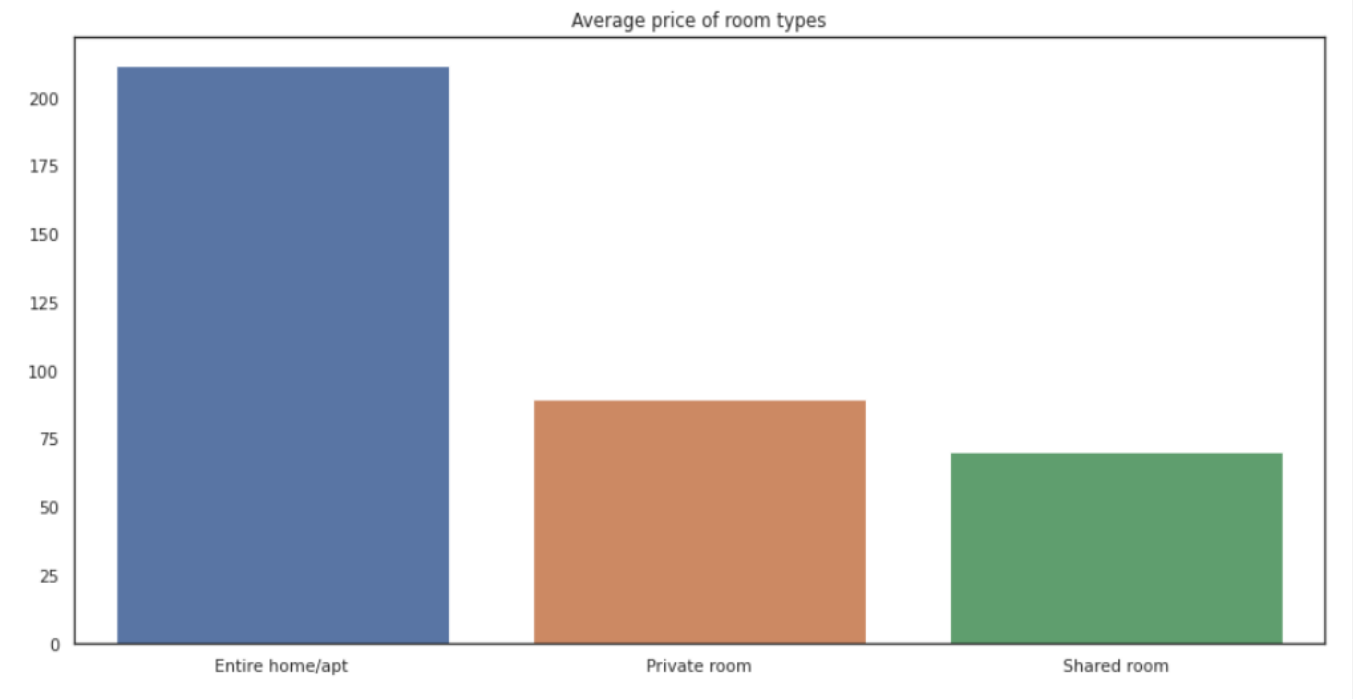

Fig 9. Room pricing chart

In above chart Entire home/apt has more than 50% avg cost private room and shared room avg cost is not having price difference morethan 20%.

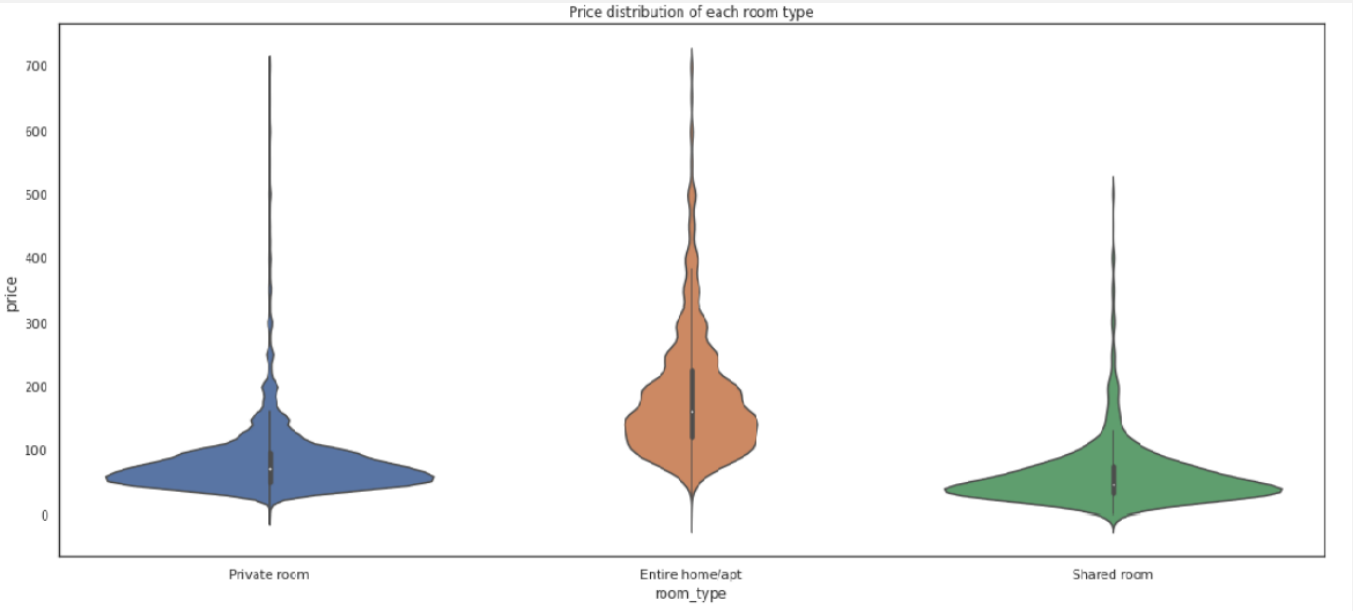

Fig 10. Proportional distribution of room prices.

we have considered to divide the whole price range into three categories

1.cheep (price range below or equal to 80$)

2.Affordable(for price range 80 to 500$)

3.Expensive(for price range more then 500$) so, it look like people have more intrest in having "affordable" rooms/apartments rathre then having cheep and expensive rooms.

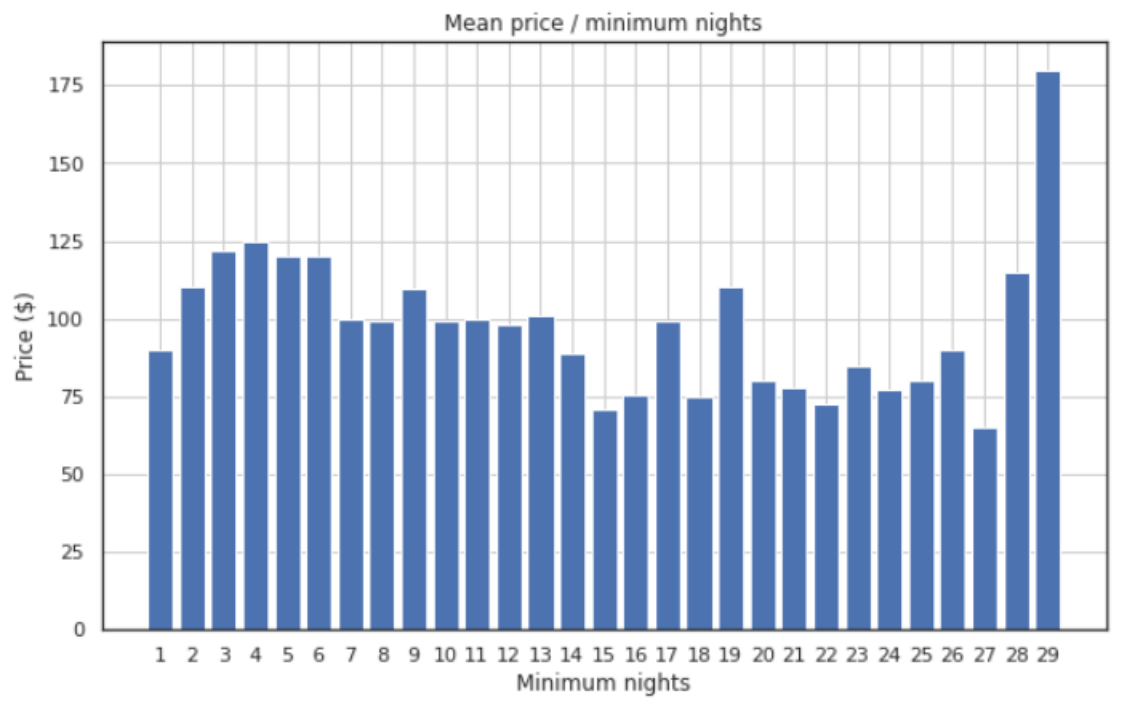

Fig 11. Mean price chart

1.It's generally cheaper to stay in rooms between 14 and 28 nights.

2.Usually, the minimum required nights to stay in a room is around 2.

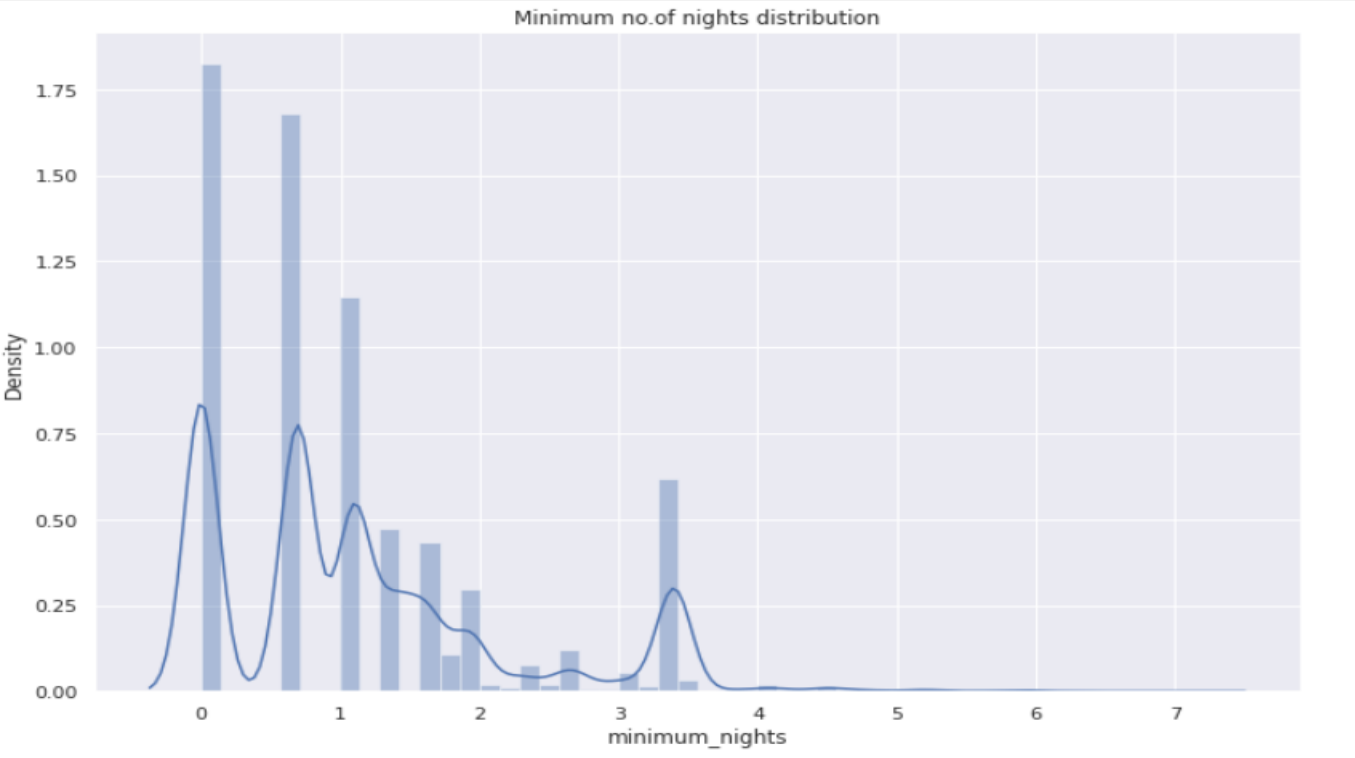

Fig 12. Min No Night Stay chart

This plots shows that majority of room booking is one for 1 to 4 days. Box-Cox transformed plot strictly shows that the majority of booking lies between 0 to 3 days. We have set the lambda parameter not equal to zero so it by definition of box-cox transform selected the best value of lambda.

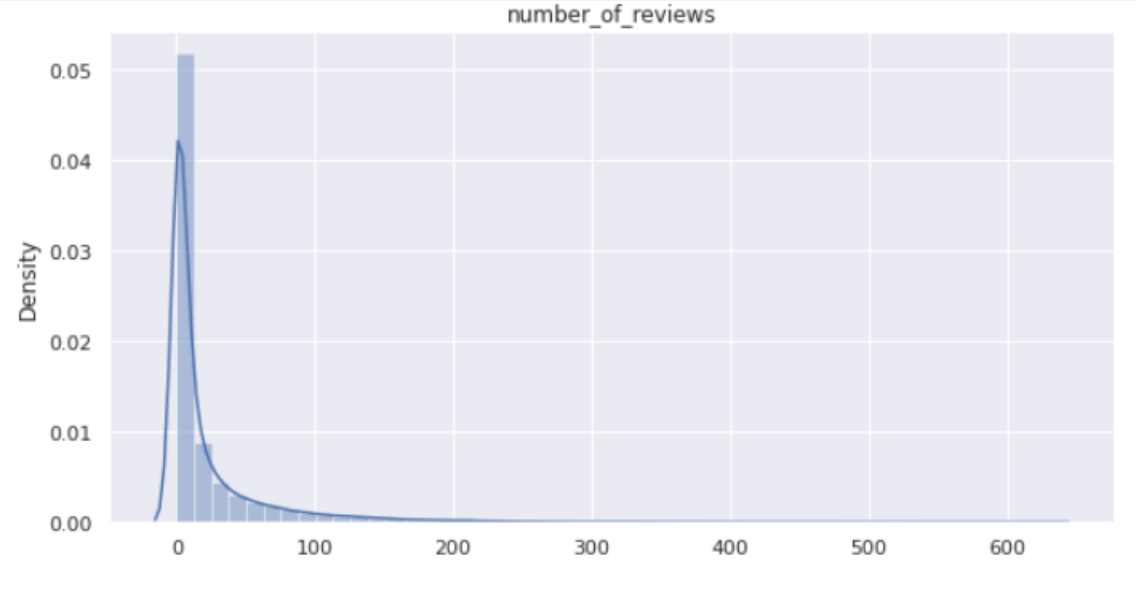

Fig 13.Number of reviews analysis

Number of reviews are highly dense from 0 to 100 reviews. We can say that most of the rooms are not rated and those which are frequently occupied only those are rated. maximum 629 times the particular room is rated. Average rating is around 23.

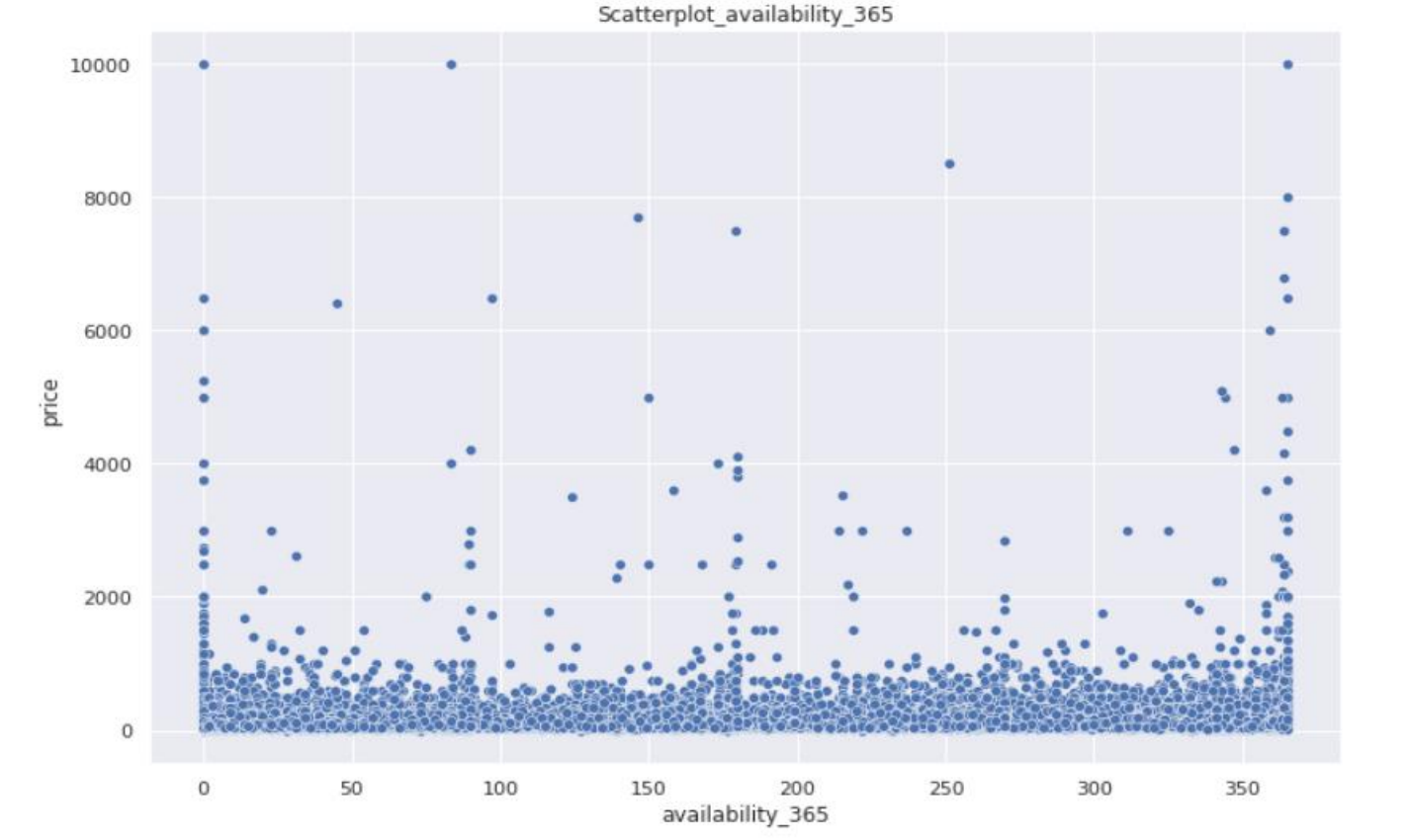

Fig 14.Availability around the year plot

From above plot we can see that most of the available rooms are in the proce range of 0 to 2000. Very few are available for price above 2000$,this is quite obivious that there are very few peoples who prefer to have expensive rooms.

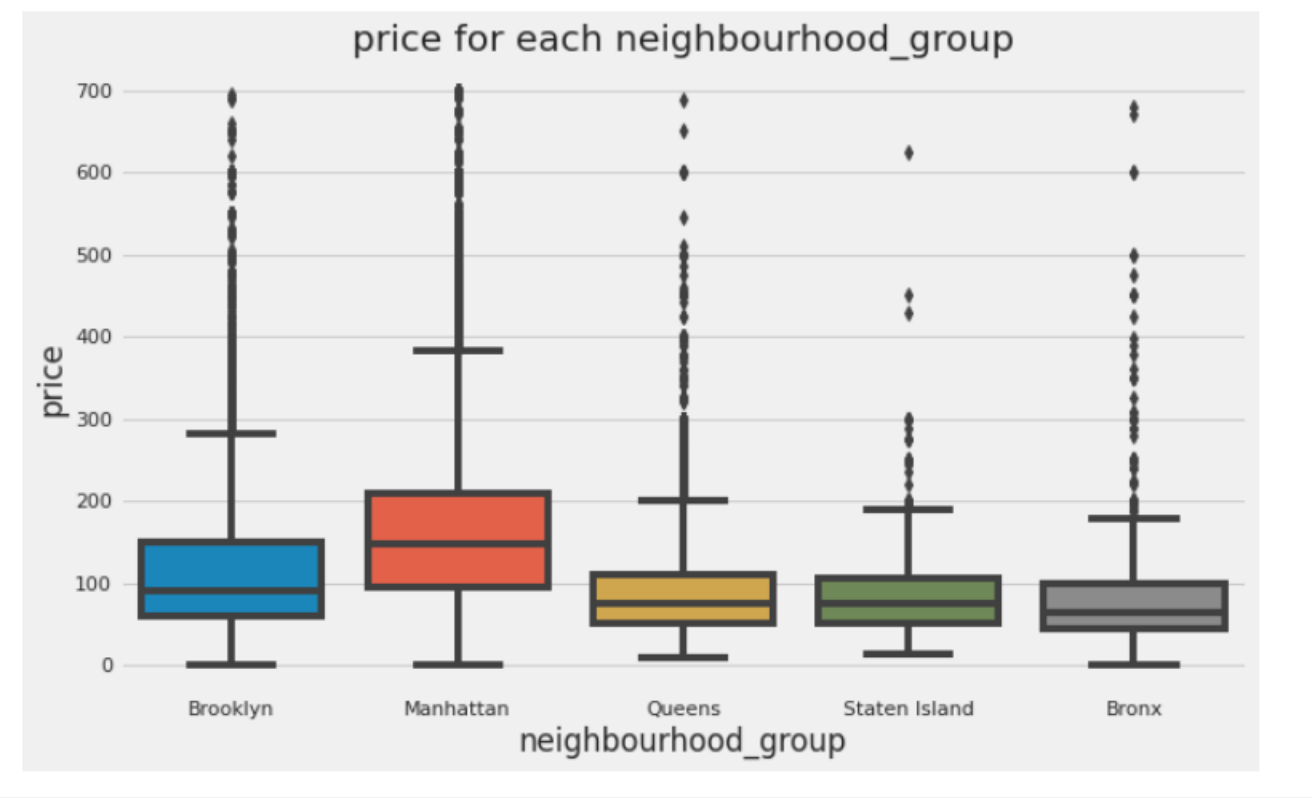

Fig 15.Neighborhood wise price

We can see that Manhattan is the most expensive destination immediately followed by Brooklyn. Queens, Staten island and Bronx, are having price range less as compared to other two. we know no of reviews is directly proportional to no of guests.

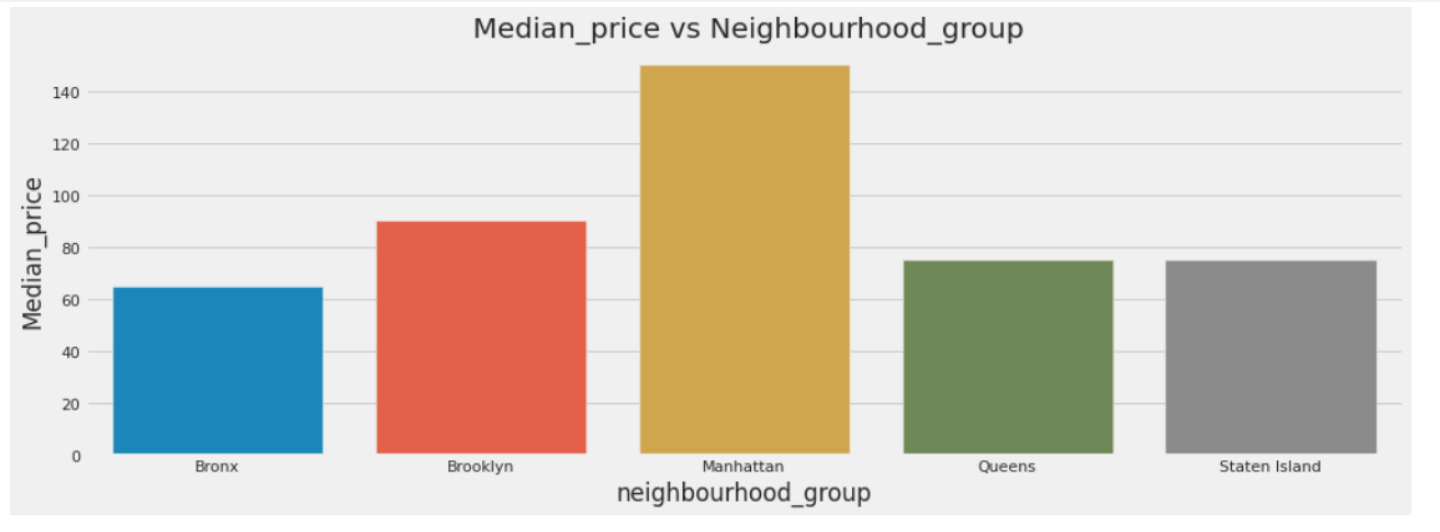

Fig 16. Median Price vs Neighborhood group

Above chart shows neighborhood wise median prices as we can see Manhattan is the most expensive neighborhood group followed by Brooklyn then Queens, staten islands,Bronx.

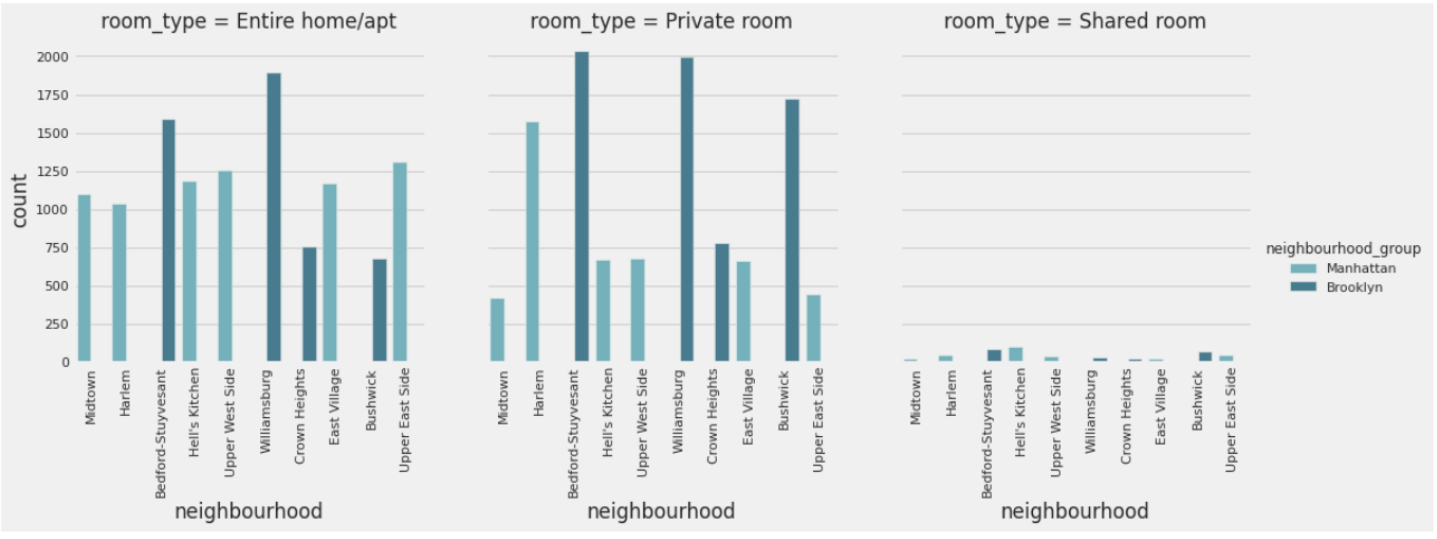

Fig 17.Neighborhood analysis

let's break down what we can see from this plot. First, we see that our plot consists of 3 subplots - this is the power of using a catplot; with an interesting output, we can easily proceed to compare the distributions between the interesting attributes.

The x and y axes remain exactly the same for each subplot, the observations on the X axis we want to count and the Y axis represents the number of observations. However, there are 2 more important elements: shade and column; these 2 distinguish subplots. After we specify the column and specify the hue, we are able to observe and compare our x and y axes between the specified columns and color coded. So, what do we learn from this? The observation that certainly contrasts the most is that Airbnb's "shared room" offering is barely available among the 10 most populous neighborhoods.

Then we can see that for these 10 boroughs, only 2 boroughs are represented: Manhattan and Brooklyn; this was somewhat expected as Manhattan and Brooklyn are one of the most visited destinations and therefore would have the most deals available. We can also observe that Bedford-Stuyvesant and Williamsburg are the most popular for Manhattan and Harlem for Brooklyn.

Limitation:

Although the dataset is very feature-rich and shares less correlation and contains enough sample to perform regression on price prediction, the correlation with target price is also low. So this will result in high MSE. Also, the features that the dataset provides in terms of the modern world are of very poor quality when making real estate valuation decisions. Since the features are positively skewed, we need to treat them before prediction.

To have a better analysis of real estate quality, it would be interesting to have sentiment analysis with real estate valuations.

User ratings of hosts are not available, it would be better to rank our hosts based on user satisfaction and ratings. In these cases, further analysis can also be done to see how guests typically rate the price or room type, or how the rating determines the property's valuation. A property with a low rating usually lowers their price.

The exact number of guests is also missing; it is only assumed that guests by column: last_review. A new host may never have been rated, that doesn't mean no guest has ever stayed there.

Scope of Improvement:

As the dataset has few qualifying attributes for property valuation, other features can be added such as bedrooms, bathrooms, property age (may be one of the most important), applicable tax rate, distance to nearest airport, hospital or schools.

Based on the rating, the hosts can be ranked and ranked, the highest rated hosts can be given a special discount or offer according to the marketing strategy.

Time series analysis can be performed to forecast occupancy rates based on the tourist season.

Conclusion:

From the entire analysis, it can be concluded that,

1) NYC shared rooms tend to be clustered in the city center, perhaps because there are more travelers who want to visit the most famous cities.

2) Comparing the price/popularity variables suggests that people who travel and use Airbnb tend to prefer listings that are cheaper.

3) Room availability per year is highest in Queens Island and lowest in Brooklyn.

4) The best place nearby is Williamsburg and Bedford.

5) Because people liked to stay in need, the maximum number of nights in Entire Apartment and Brooklyn is the second most focused place of people. Also, because the average cost in Brooklyn is $80, which is less than in Manhattan. Airbnb may increase the number of an entire apartment in Brooklyn.

6) There are two types of user rooms: Professionals, which are remote areas, each with a high number of rooms; and Amateurs, of which there are usually only a few. Although amateurs can make money like a business, their volume is clearly lower than that of professionals.

7) Expert positions are located in the city center. The way rooms are announced differs between professionals and amateurs. The former uses more objective terms to describe the room, while the latter uses more subjective ones.

8) Have a room "close" to things that affect popularity (it might be a good idea to include those words in the room name).

THE END