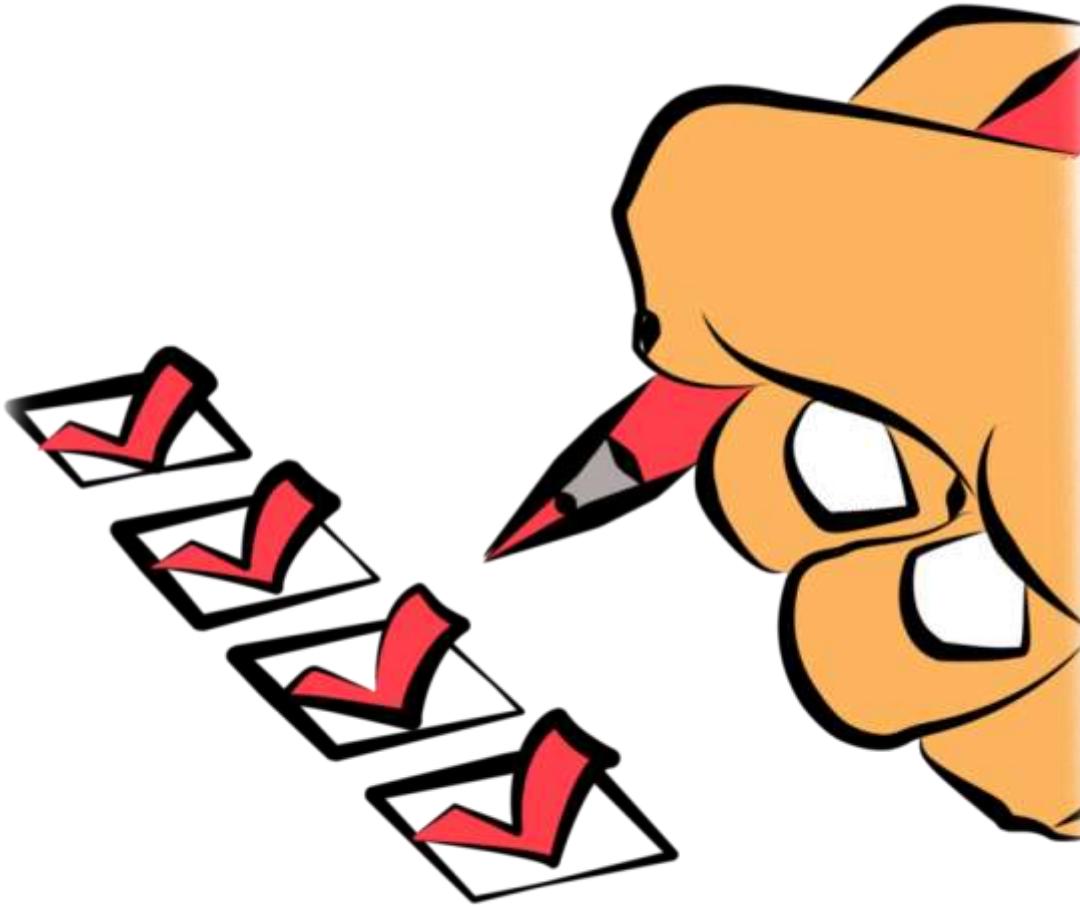
Capstone Project
EDA on Airbnb Booking Analysis

Team - Data Manipulators

Abhishek Dhasmana
Nishant kalanta (Leader)
Ramkrishan Shahi
Dhrishanda Mehdi
Himanshu Awasthi

Contents:



Introduction:

I have a place to rent out.

Airbnb

I need a place to stay for my vacation.

Introduction:

- Airbnb (ABNB) is an online marketplace that connects people who wish to rent out their homes with people who are looking for lodging in a given location.
- According to the most recent data, Airbnb has over 6 million listings, covering over 100,000 cities and towns in over 220 countries worldwide.
- Since 2008, guests and hosts have used Airbnb to broaden their travel options and provide a more distinctive and personalized way to tour the world.
- Today, Airbnb is a one-of-a-kind service that is used and recognized worldwide.
- A crucial component for the firm is data analysis of the millions of listings made accessible through Airbnb.

Problem Statement:

- Major firms like Airbnb constantly generate data; ever day, millions of bits of data are produced.
- Therefore, managing this data is the real challenge, and managing data is essential for any business to be on the right track, understand what customers want, and determine whether they are receiving it or not.
- By working with the data or performing data analysis, all of these issues are clarified or addressed.
- The descriptive analysis can be used for security, business decisions, marketing initiatives, implementing cutting-edge add-on services, understanding customer and vendor (host) behavior on the platform, and more.

Dataset Analysis:

The given Airbnb dataset contains 48895 observations with 16 features. This dataset contains all the required information to perform Exploratory Data Analysis. We will Understand and conclude based on the given data set.

id = Unique ID assigned to the entry.

name = This is a column containing the name provided by each host for customer reference.

host id and host name = Many hosts serve many objects. This host id and host name contains this records.

neighborhood and neighborhood group = These columns contain information about the city and area of properties offered by Airbnb New York.

Longitude and Latitude = Contains the longitude and latitude of the property location.

Room type = Private room / entire home / shared room.

price = An important column that contains price values for all these properties.

minimum nights = This gives you information about the minimum number of nights a host offers for a particular accommodation.

number of reviews and reviews month = It includes the number of reviews and ratings per month for these accommodations, as well as information about the host's hospitality.

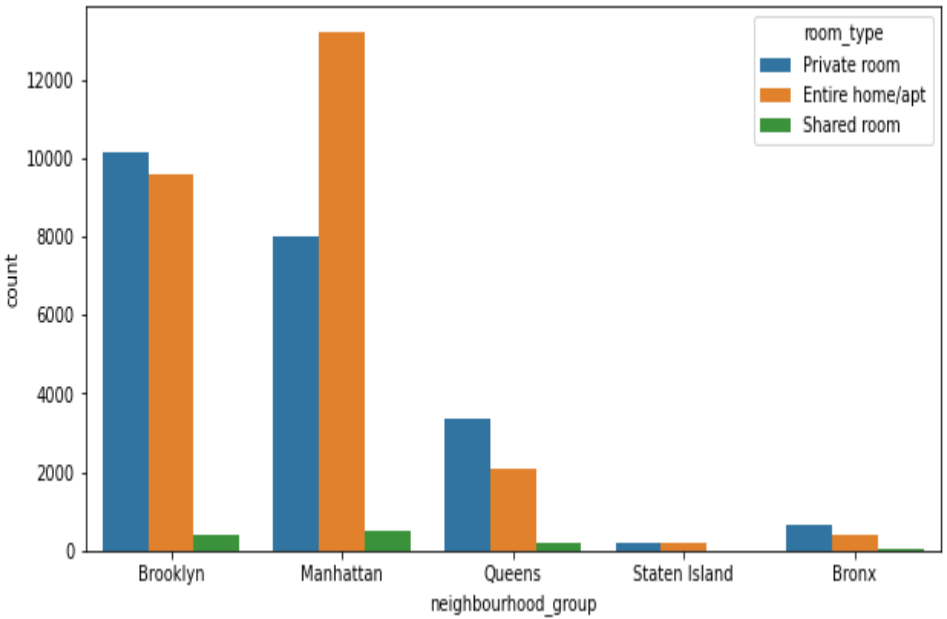
availability 365 = Provides information about offer availability

Description:

	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

As shown above figure, Statistical description of Dataframe was returned.

Plot between neighbourhood_group and room_type (count)

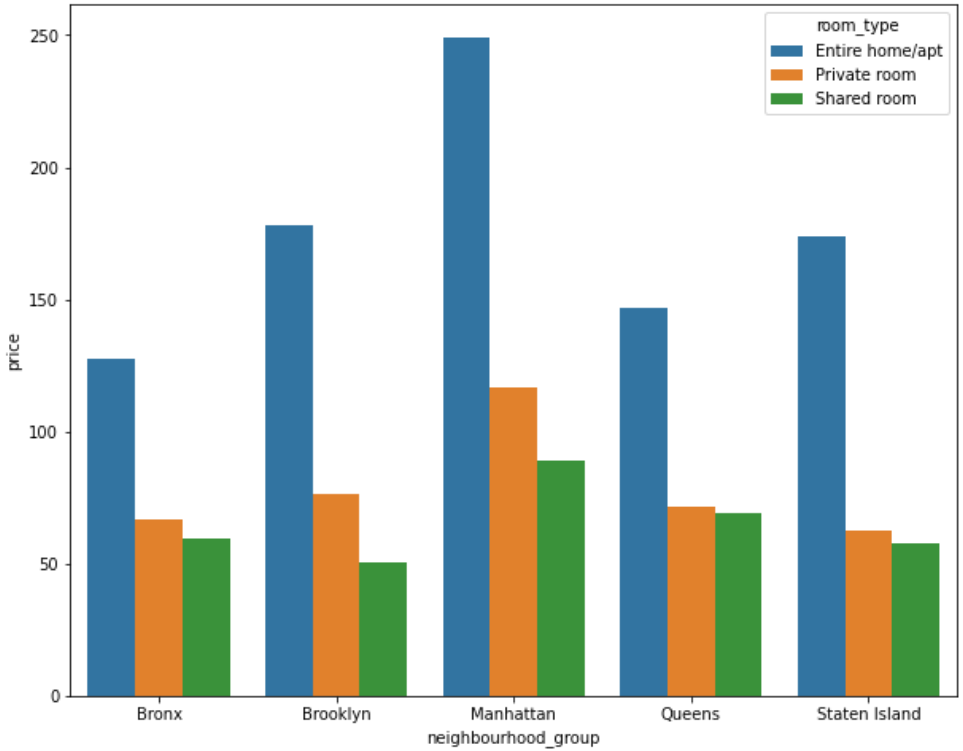


we can understand that the stay of customers are distributed amoung 5 Neighbourhood groups namely Brooklyn, Manhattan, Queens, Staten Island, Bronx.

room_type has three categories. We can also observe the number of accommodations of each room type in these five cities.

we can also understand that Brooklyn and Manhattan provides highest number of Room type for their customers.

Plot between neighborhood_group and mean price:

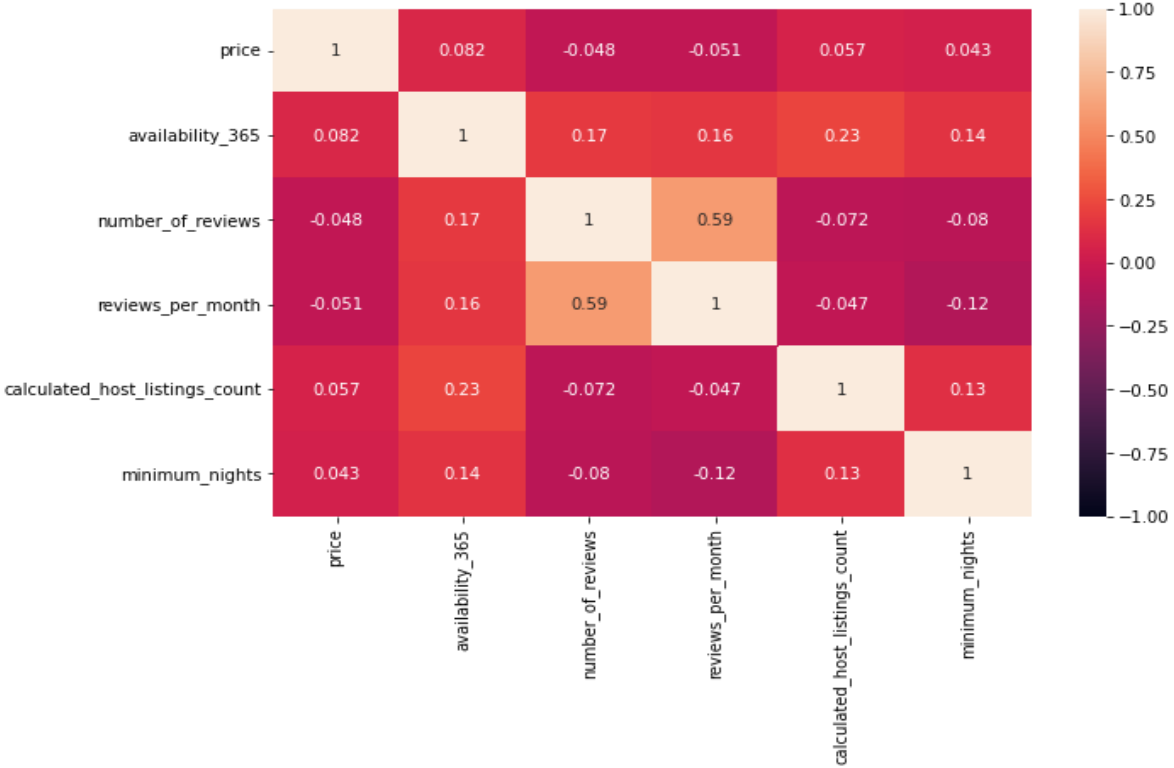


we can conclude that:-

1- The price of entire home/apt are very high as compared to the prices of private room and shared room which are somewhat comparable.

2- Among all the neighborhood the most expensive one is Manhattan.

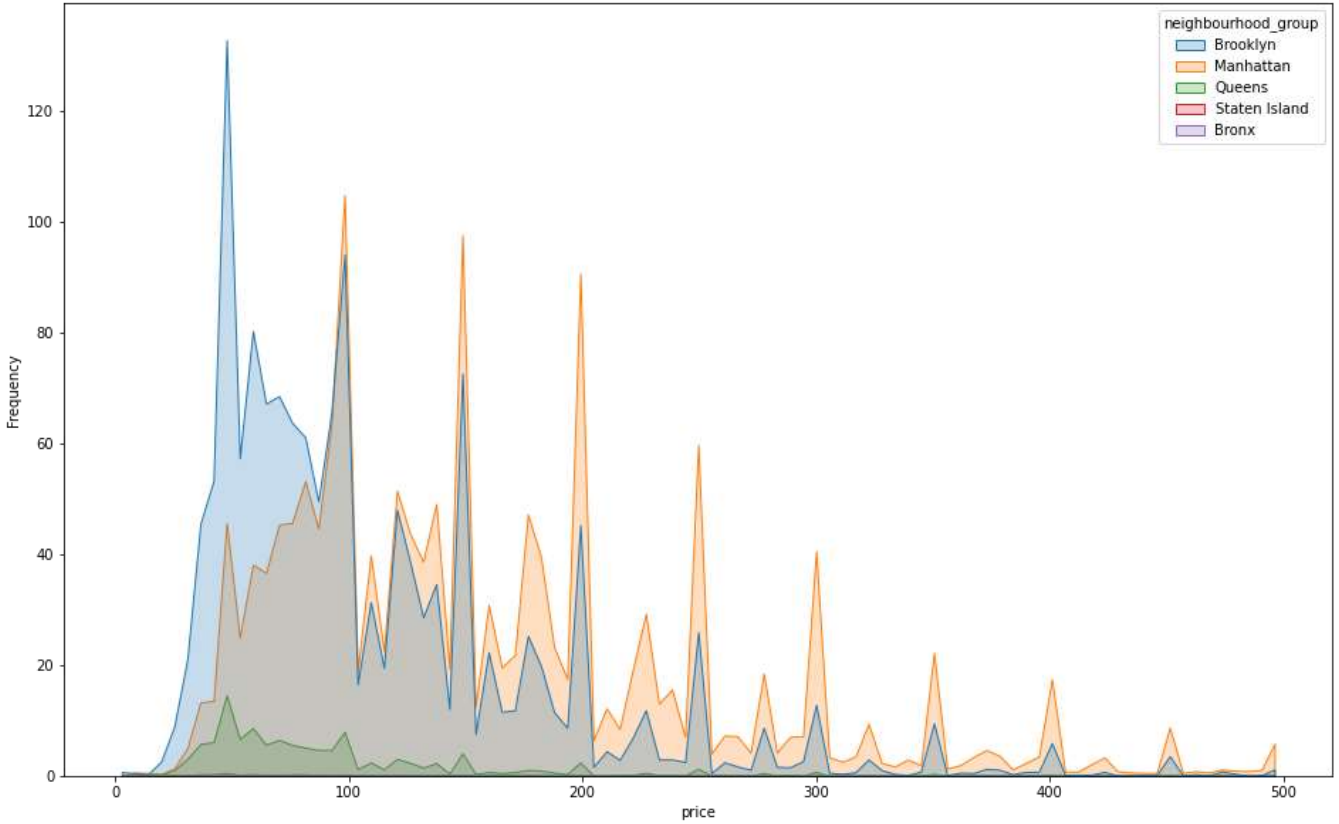
To display a more generalized view of numeric values:



1- number of reviews is highly correlated with reviews per month

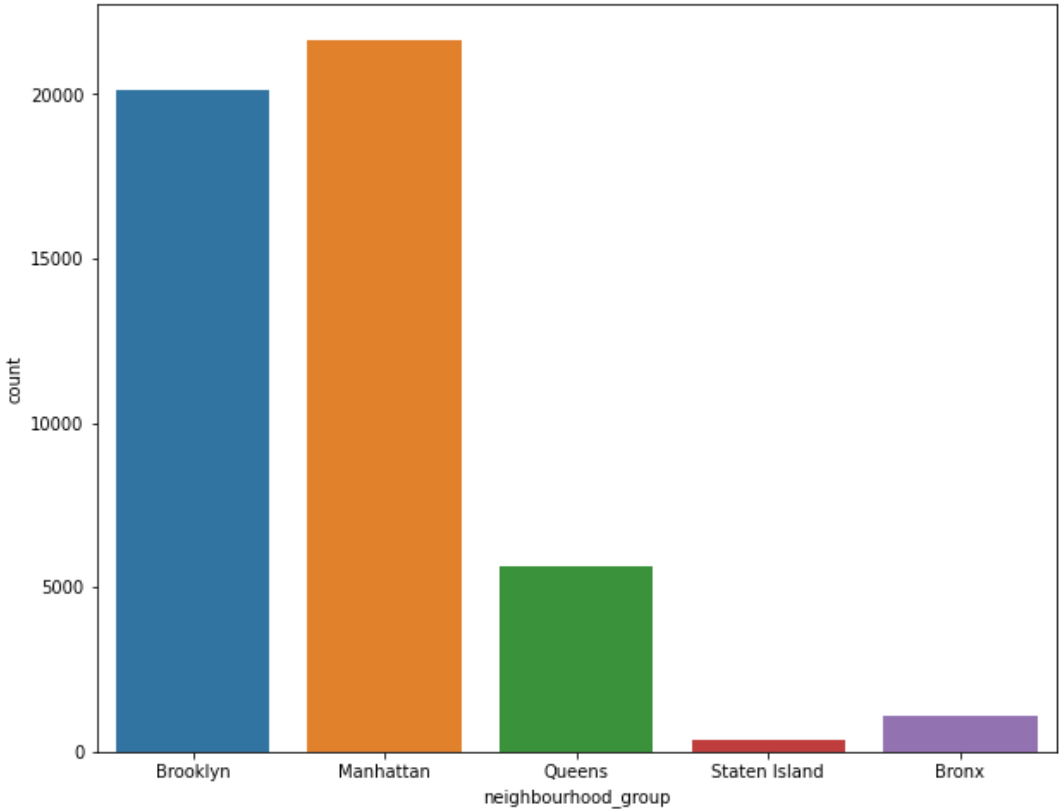
2- price is negatively correlated with reviews but correlated with availability_365, minimum_nights, calculated_host_listing_count.

Plot between mean price and its frequency:



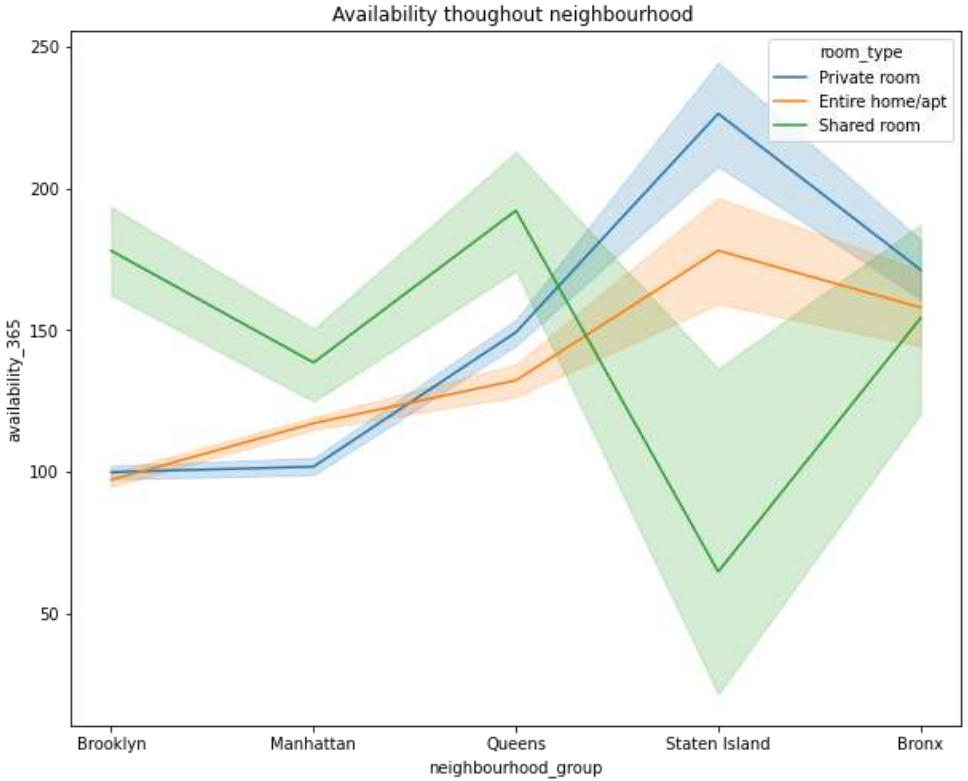
Brooklyn and Manhattan are most expensive Neighbourhood_groups.

Plot to show the count of each neighborhood:



We can say
that Brooklyn and Manhattan are the most crowded place in NewYork and most of the guests wants to visit these two as compared to other neighbourhood.

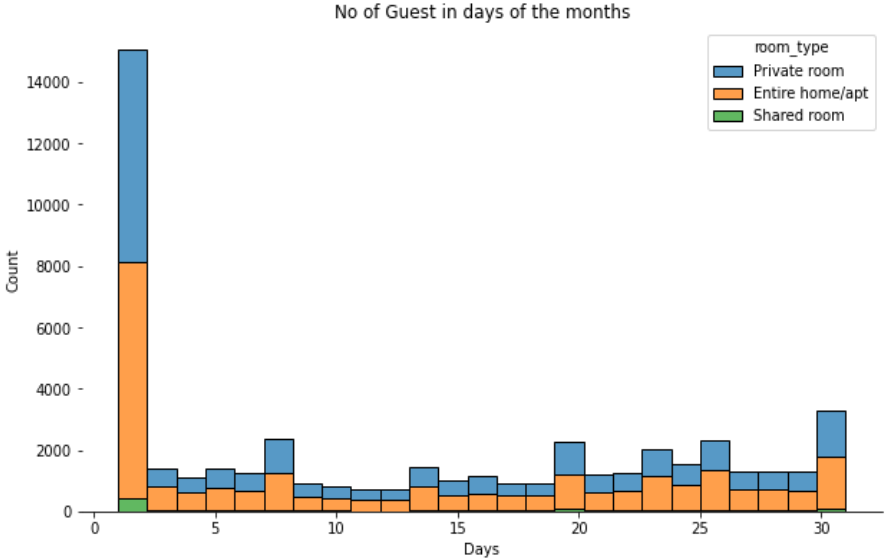
Let's see the availability of different room types of each neighborhood_group:

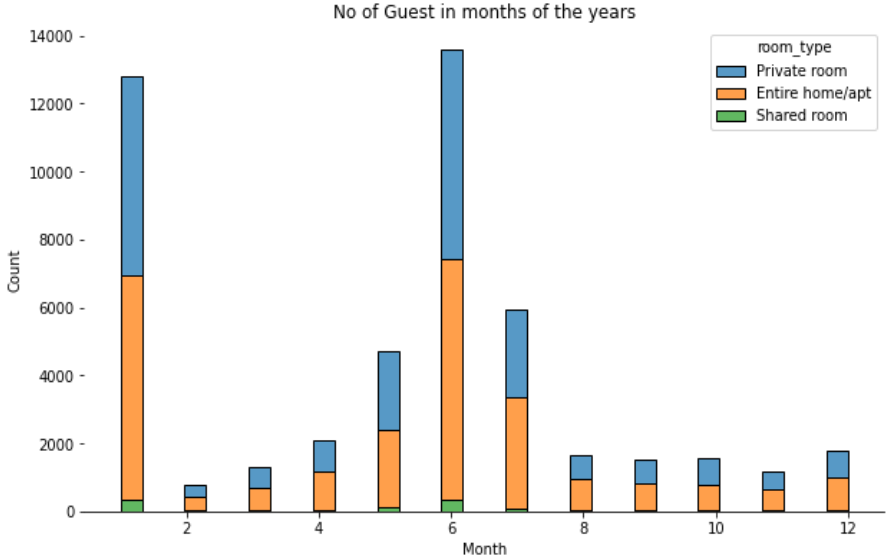


Here we can say that, Private room and Entire home is available at most in STATEN ISLAND but the availability of Shared rooms is lowest as compared to other neighbourhood groups.

Also, The availability of Private rooms and Entire home is almost similar in Brooklyn and Manhattan.

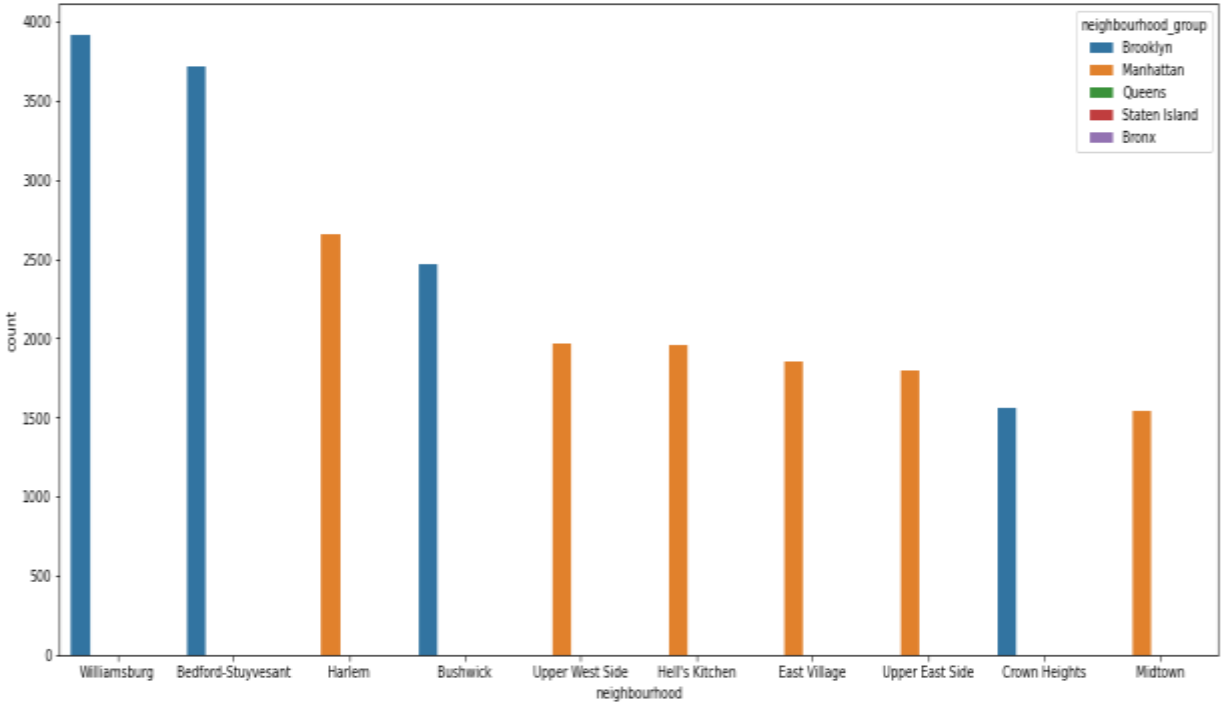
Now we will analyze the data with the help of reviews we have received:





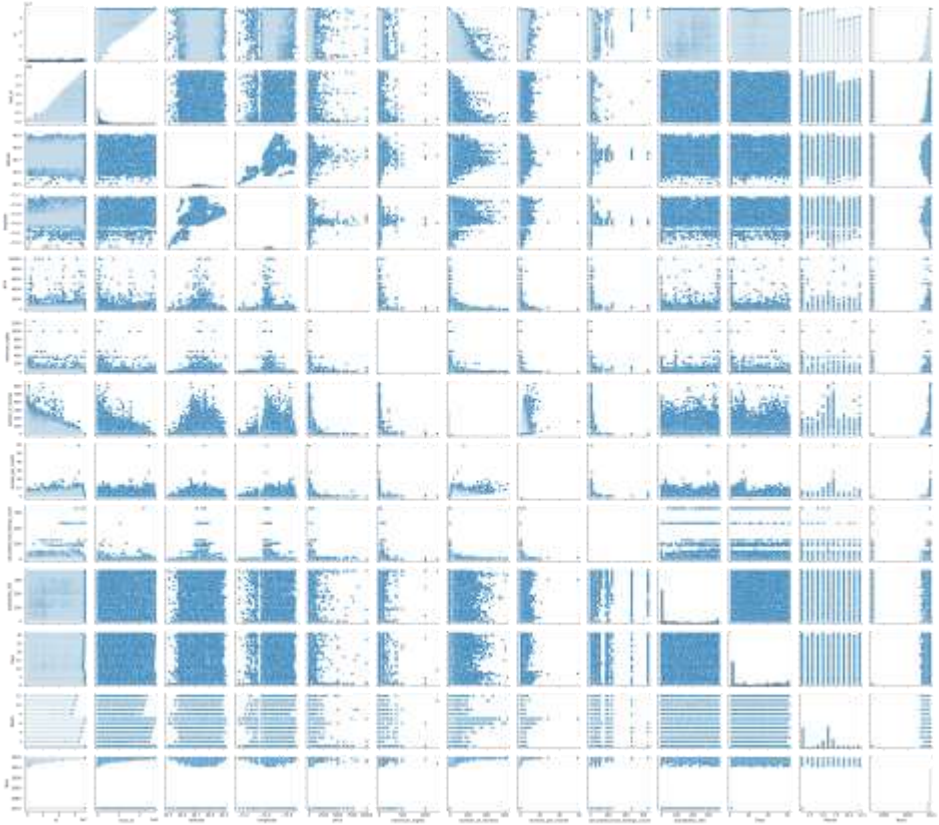
Before making conclusions we must keep in mind that these two predictions are based on the data user have provided but many users did not provide the reviews so we are restricted to those who did.

Count plot of Neighborhood:



Williamsburg and Bedford-stuyvesant are most listed neighborhood

Pairplot:



In these graphs, I am experimenting how much variable we have and their relationship with other variables pretty much like correlation matrix but in well visualized way.

```
sns.pairplot(data=df)
```

Let's find out which neighbourhood_group is expensive for Customers:

	Location	Maximum price
1	Brooklyn	10000
2	Manhattan	10000
3	Queens	10000
4	Staten Island	5000
5	Bronx	2500

Clearly Manhattan, Brooklyn and Queens have the maximum rent price.

Let's check which host is highly suited for customers based on number_of_review:

	Host_id	Host_name	Number_of_reviews
21316	37312959	Maya	2273
1052	344035	Brooklyn& Breakfast -Len-	2205
18636	26432133	Danielle	2017
20884	35524316	Yasu & Akiko	1971
21934	40176101	Brady	1818

We found that Maya, Brooklyn & Breakfast-len-, Danielle, Yasu & Akiko and Brady are the most reviewed host_name.

End of our Exploratory Data Analysis

Limitation:

- 1- Although the dataset is rich in quantity but it lacks quality as there is very less correlation between features.
- 2- The dataset provides very poor characteristics for the modern world when determining property valuations.
- 3- Highest number of guests in Manhattan along with minimum reviews can cause analysis to not have a mutual conclusion.
- 4-Host user reviews are not available. Better to rate hosts based on user satisfaction and reviews. In order to have best analysis.
- 5- The exact number of guests is also unknown. I'm just assuming the guest for column last_review. New hosts may not be rated, but this does not mean guests have never stayed there

Scope of Improvement:

- Based on the limitations lets define the Scope of Improvement.
- We can add features like number of days stayed, customer service, tax rate, amenities (like gym, hospital, airport, taxi service, etc.), etc. to have better correlation within dataset.
- So that the analysis becomes more interesting.
- Time can also play crucial role when it comes to EDA as it can be used to relate with no. of days stayed to predict the busiest time of the month.
- We are doing all this EDA to provide better service for our (Airbnb) customers so getting info which can help us doing that is beneficial.
- We must provide such service so that maximum number of customers wants to review the neighborhood.
- As in modern days, there are very few people who cares about longitude and latitude instead we can provide amenities (as mentioned earlier) so that it would be more easy for customers to choose the best neighborhood.

Conclusion:

- The stay of customers are distributed among 5 neighbourhood_groups namely Brooklyn, Manhattan, Queens, Staten Island, Bronx. Most crowded places are Manhattan, Brooklyn, Queens in the same order. Number of Private_room and entire home is almost equal in every neighborhood group in Manhattan.

- The price of an entire home/apt are very high as compared to the prices of private room and shared room which are somewhat comparable. Among all the neighborhood the most expensive one is Manhattan and it is located in the East Village area of New York City.

- Brooklyn and Manhattan are most expensive Neighborhood groups.

- Brooklyn and Manhattan are the two most popular places to visit in New York City. Most of the guests wants to visit these two as compared to other neighborhood. The availability of Private rooms and Entire home is almost similar in Brooklyn and Manhattan. However, the availability of Shared rooms is lowest as compared to other neighborhood groups.

- Throughout the year maximum number of guests/customer visit neighborhoods on January and June.

- we also plot Pairplot to see the relation between variable among different variables.

- Manhattan, Brooklyn and Queens have the most reviewed host_name. Maya, Brooklyn & Breakfast-len-, Danielle, Yasu & Akiko and Brady are the most popular host.

Thank You