Exploratory Data Analysis on Airbnb Bookings

By

Sunil Kumar Bodugu

CONTENTS

1. Introduction

2. Problem Statement

3. Dataset Analysis

4. Plot Analysis

5. Scope of Improvement

6. 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, and 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, which neighborhood group has most number of rooms and the reasons for that, what are the average price of various room types in different neighborhoods 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:

A. What can we learn about different hosts and areas?
B. What can we learn from predictions? (ex: locations, prices, reviews)
C. Which hosts are the busiest and why?
D. Is there any particular factor which makes some of the hosts more profitable when compared to the rest?

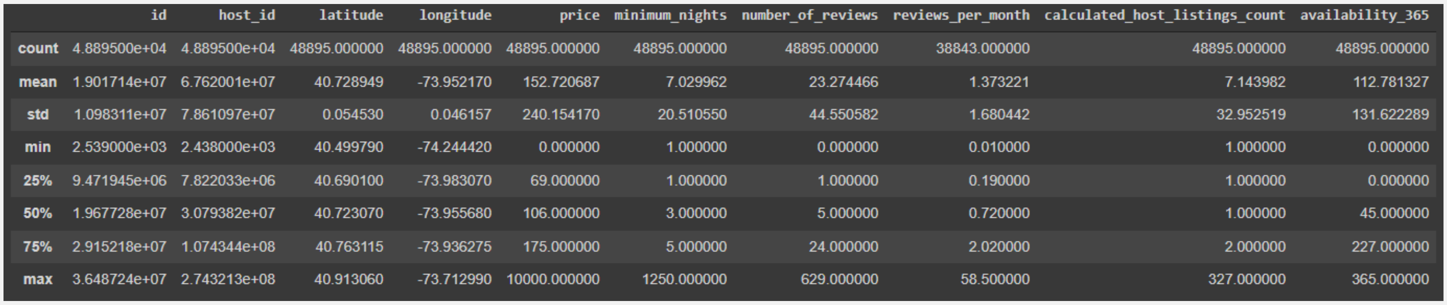
Dataset Analysis:

The dataset contains 48895 observations with 16 features. This data file includes all needed information to find out more about hosts, geographical availability, and necessary metrics to 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

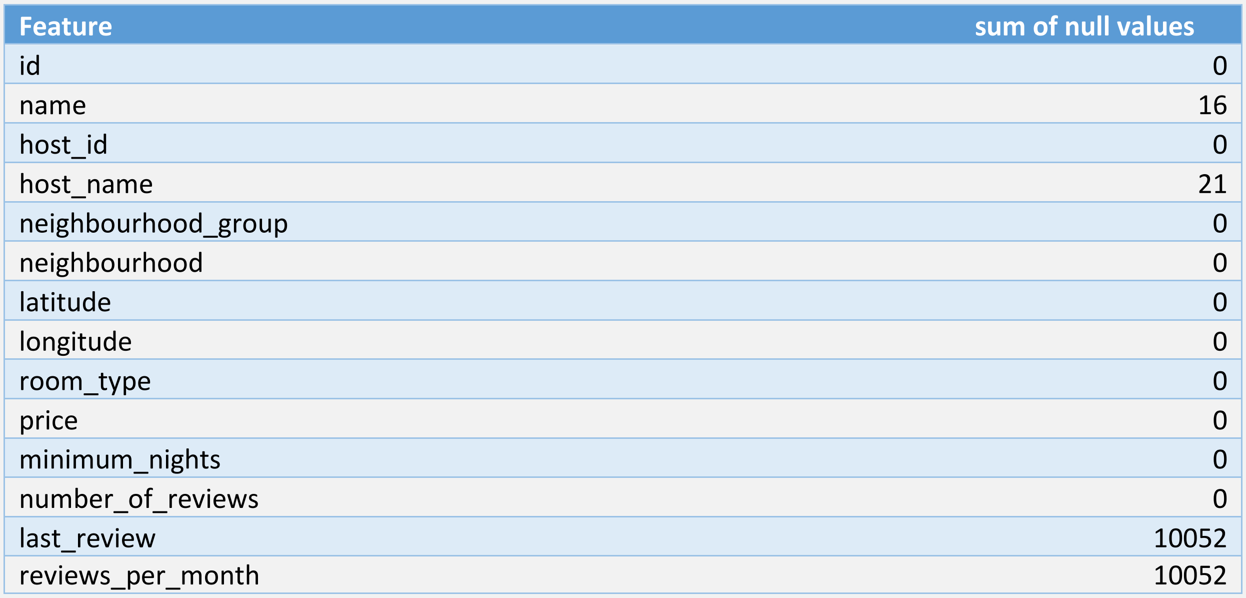
Other 3 important columns are,

1. neighbourhood_group: It contains 5 unique hoods which are Manhattan, Brooklyn, Queens, and Bronx & Staten Island.
2. Neighbourhood: It contains 211 unique neighborhoods.

3. room_type: It contains 3 unique room types which are Entire home/apt, Private room, Shared room

Out of all columns, four columns containing null values which are name,

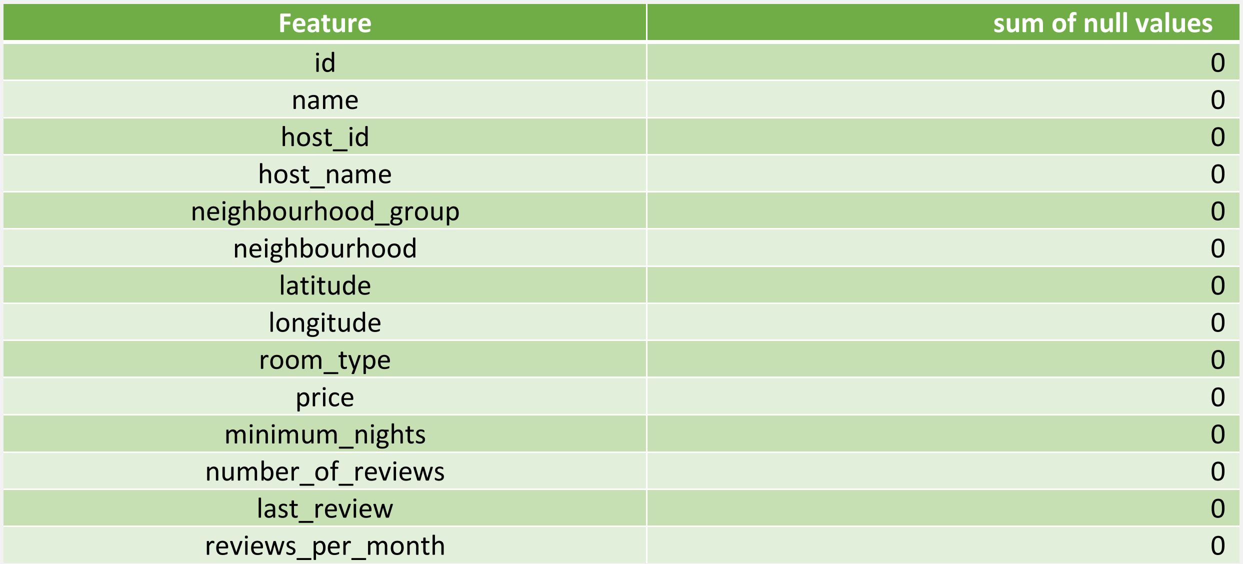
- Name column having total 16 null values.
- Host_name column having 21 null values.
- Last_review and reviews_per_month are having 10052 null values
- We will look at the columns and decide what we can do with them



Feature	sum of null values
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

Fig 2. Sum of total null values in all features.

As we are not going to use columns named Name and last_review, we are going to drop those columns and for the host_name column we are going to fill it with room type column, null values in review_per_month column are replaced with zero values.



Feature	sum of null values
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

Fig 2. Sum of total null values in all features after data cleansing

We haven't missed value anymore.

Plot Analysis:

Review is the one of the important criteria with online activity these days. This gives a lot of insights to a particular place for tourist and they can swing mood when it comes to online booking, so our first objective is to find out top 5 properties name with maximum number of Reviews.

property id	hotel name	max_reviews
11759	Room near JFK Queen Bed	629
2031	Great Bedroom in Manhattan	607
2030	Beautiful Bedroom in Manhattan	597
2015	Private Bedroom in Manhattan	594
13495	Room Near JFK Twin Beds	576

Table 1. Top 5 properties with max reviews.

Bar chart of good reviews on each neighbourhood groups.

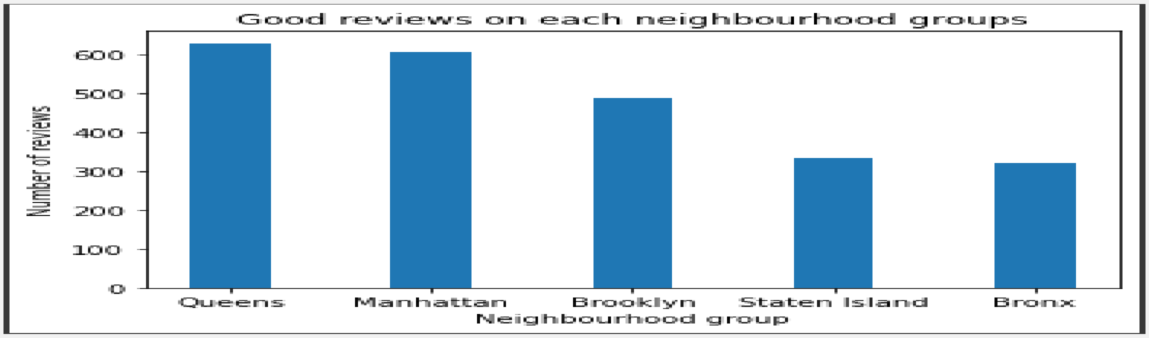


Fig 3. Good reviews on each neighbourhood groups

From the above plot we can conclude Queens and Manhattan neighbourhood group have more good review, one the reasons behind getting more number of reviews is providing feedback from to the customers. One thing we are not able to analyse the nature of reviews whether it is positive or negative.

Next, we will look for the the maximum and average price of all room type

Room type	Average price	Maximum price
Entire home/apt	211.79	10000
Private room	89.78	10000
Shared room	70.13	1800

Table 2.Prices of all room types.

Let’s find out which hosts are busiest and will plot the bar graph for the top 7 busiest host. The below analysis is done after considering the no of reviews they got and price.

These hosts are busy because they are charging minimum price, having good reviews and providing good services to the customers. So customer will prefer for these hosts.

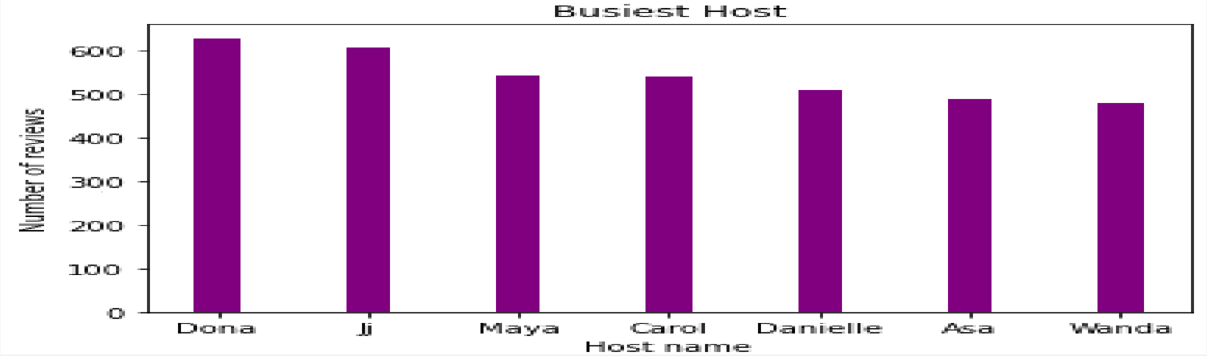


Fig 4.Top 7 busiest hosts

Below plot describes democratic view of properties listed also it provides a clear view of the city area.

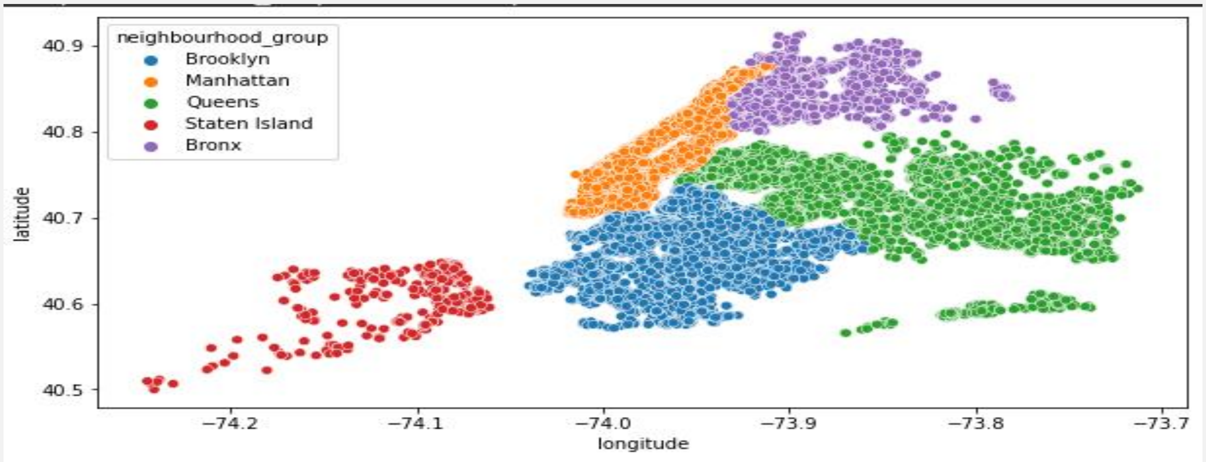
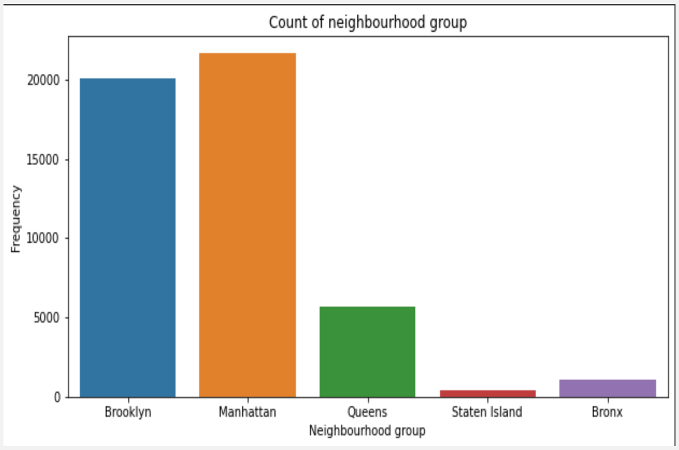


Fig 5. Location of Neighborhood Groups

Next, we will look for the distributions of properties throughout the area.



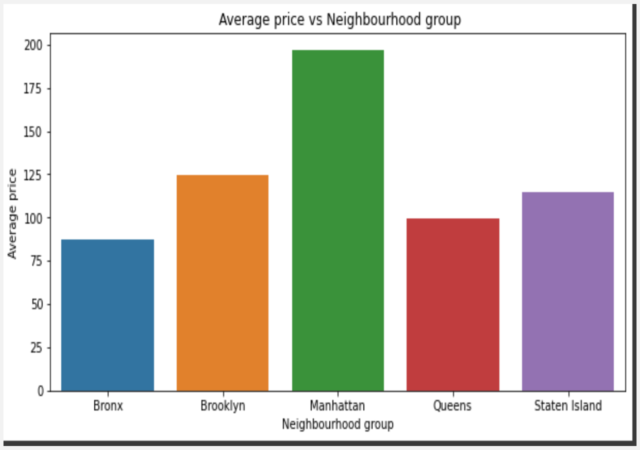


Fig 5. Count of Properties in Neighborhood Groups

Fig 6. Average price in Neighborhood Groups

From above graph it can be understood that Brooklyn and Manhattan stand within the most urban and active area, in terms of listing count and pricing both. We further analyzed the room types occupied by a neighborhood group. Shared rooms are the cheapest and also has lowest count in every neighborhood, whilst Manhattan has the greatest number of Entire home/apt category, but Brooklyn has the greatest number of Private room category. Entire home category rooms also maintained a higher price range in almost every hood with an average price of $211.79 next to Private Room which shares an average price of $89.78, pretty large margin!

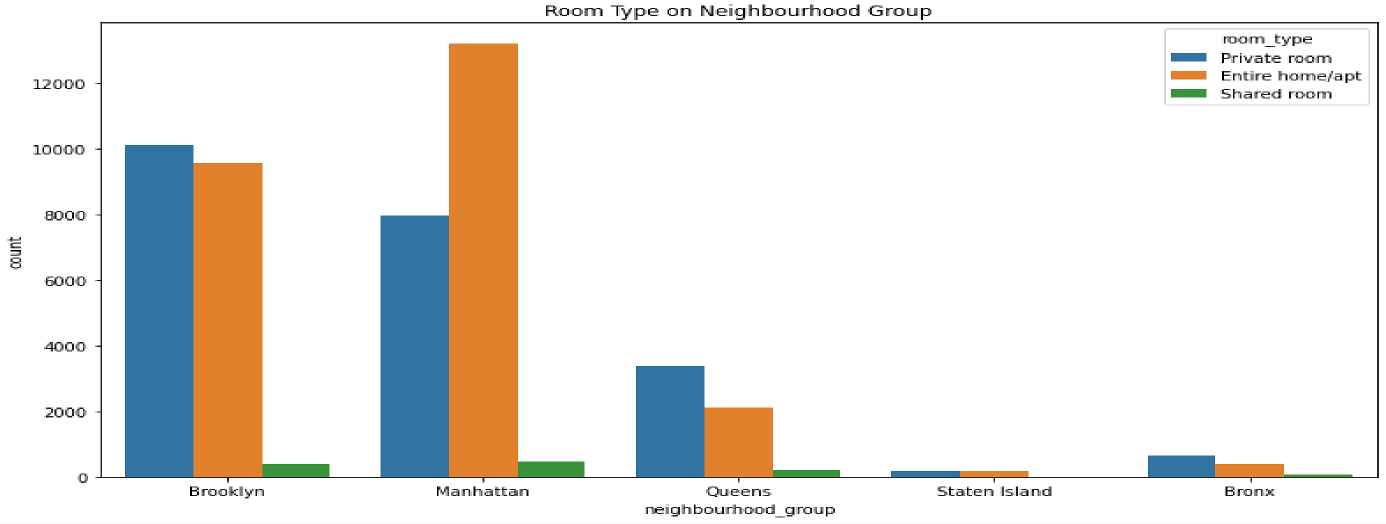


Fig 7. Count of Room types in Neighborhood group

There are 221 unique neighborhoods. Below plot describes the distribution of properties over top 10 various neighborhoods. The objective of doing this analysis is to give proper insight to hosts who want to set up new properties, as this analysis will easily find out the neighbourhood where least properties are listed.

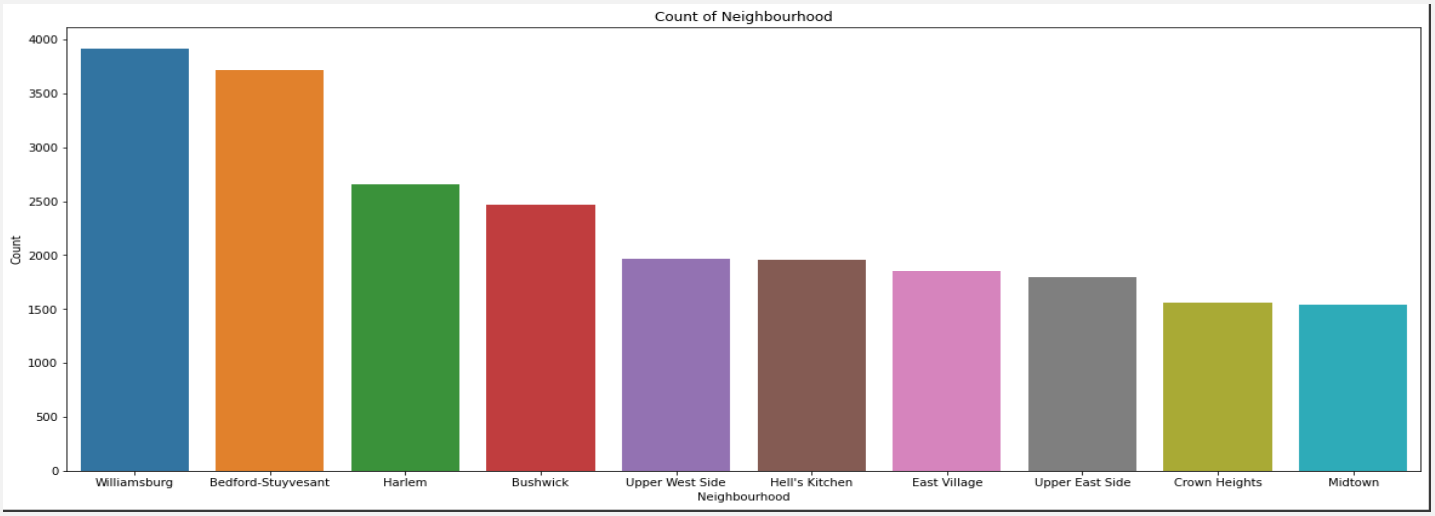
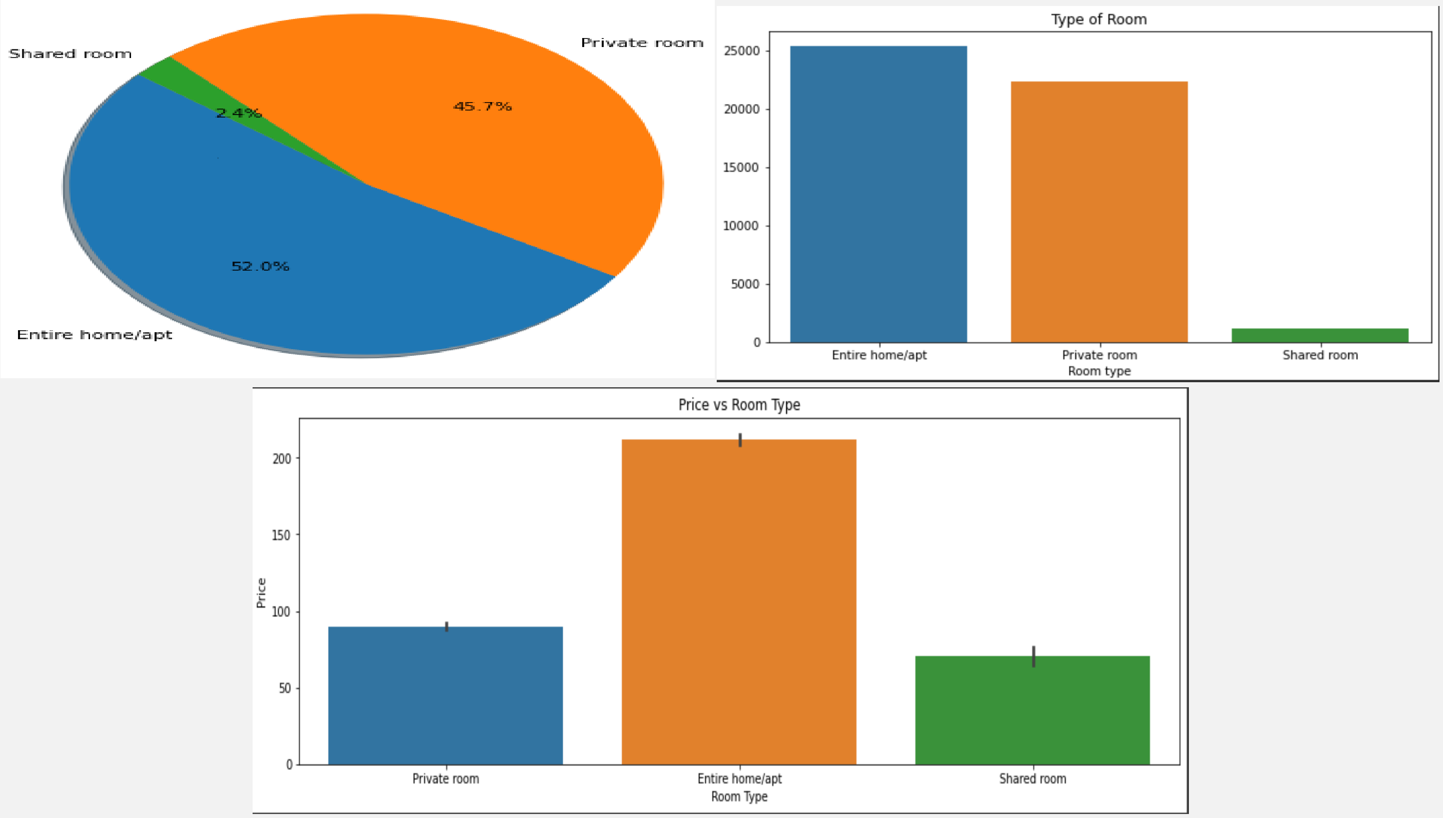


Fig 7. Count of Properties in top 10 Neighborhoods.

Entire home/apt has more than 50% proportion in New York City and it too has highest average price also. Shared room are the cheapest, but only has 2.4% proportion. No wonder New York life is of high standard



.Fig 8: Proportion, Count and Average price of Room Types

Now to get even more understanding of how total revenue is varying in neighbourhood for various room types lets analyze the dataset, for that we need to find maximum and minimum prices for each room type among the 221 unique neighborhoods.

Room type	Maximum revenue details		Minimum revenue details	
	Neighbourhood	Total Revenue($)	Neighbourhood	Total Revenue($)
Entire home/apt	Williamsburg	389724	New Dorp	57
Private room	Williamsburg	171265	Graniteville	20
Shared room	Hell's Kitchen	9488	Randall Manor	13

Table 3.Maximum and Minimum prices for various room types in Neighborhood’s.

So from the above outputs, the neighbourhood 'Williamsburg' has generated highest revenue for Entire home/apt and Private Room types. The reason behind this may be the presence of tourist places near this neighborhood’s.

Now divide the properties based on their booking price, for simplified analysis I consider the properties whose price is less than 100$ as 'cheap', if the booking price is in between 100$ and 500$ then it is considered as '3 Star Hotel' and for properties whose booking price greater than 500$ I can considered it as '5 Star Hotel'.

To get clear view of which type of properties were mostly booked by the customer, we used minimum_nights column. The below bar graph shows the visualization of above analysis.

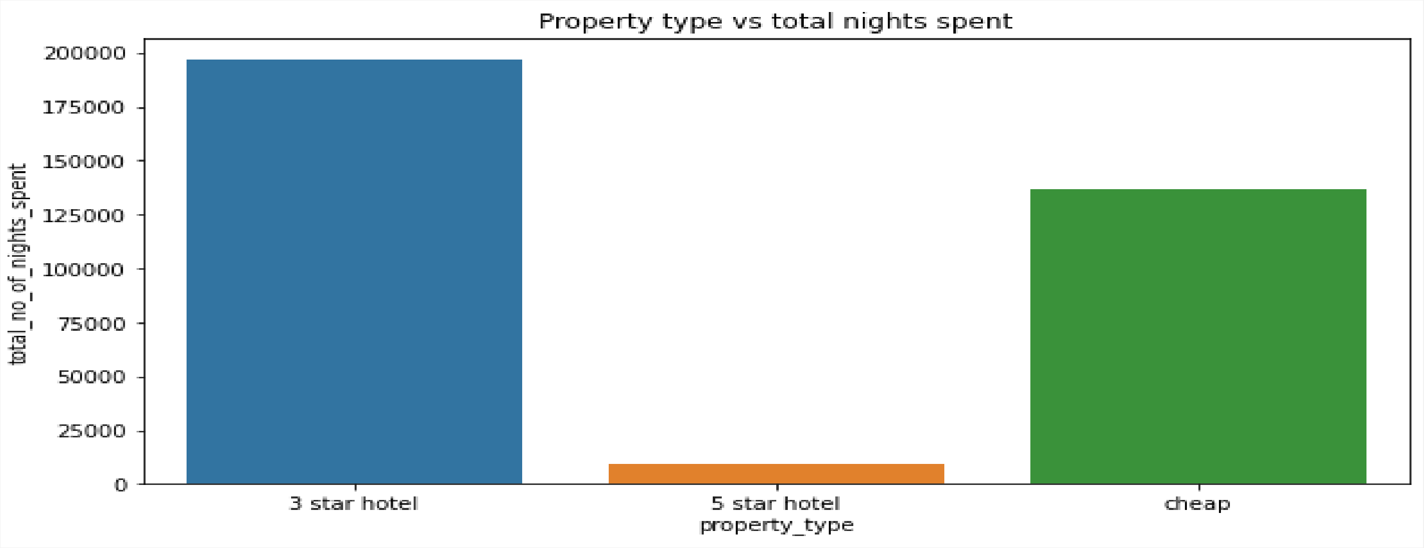


Fig 9: Property type vs Total nights spent

'3 star hotels' has most demand in the NYC area followed by cheap type properties and '5 Star hotels' are booked by very few people.

Now it's time look into some statistical data for price column which is grouped by room_type column.

room type	count	mean	std	min	25%	50%	75%	max
Entire home/apt	25409	211.794246	284.041611	0	120	160	229	10000
Private room	22326	89.780973	160.205262	0	50	70	95	10000
Shared room	1160	70.127586	101.725252	0	33	45	75	1800

Table 4. Statistical data of price column.

The key observations from the above table is that the majority of properties price is below 100$ for room types Private and shared room. Whereas the price is below 300$ for room type Entire home/apt.

The below scatter plot describes the price variation of properties in different neighborhood groups.

For properties in Staten Island and Bronx the booking prices lies below 2000$.

For properties in Queens the booking prices lies below 2500$

For properties in Brooklyn and Manhattan the maximum booking prices lies in the range of 4000$.

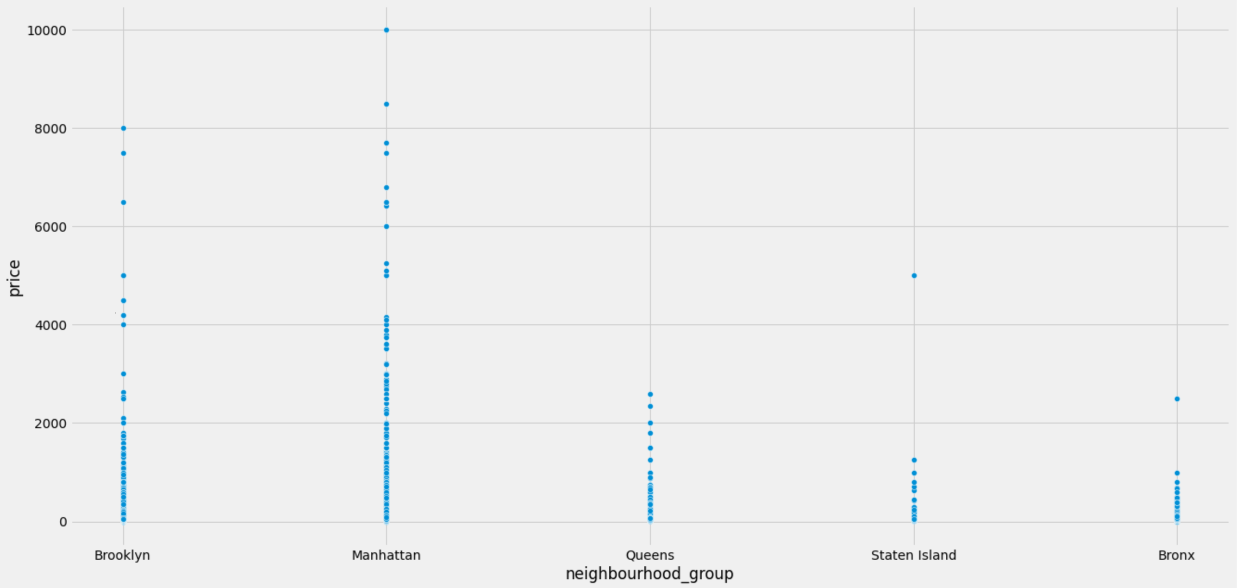


Fig 10: Scatter plot of prices in various neighborhood groups.

From the Airbnb data set we know that some of the hosts own more than fifty properties, so my objective is to find at least one factor which makes them to expand their properties when compared to rest.

To find out the relation I had chosen minimum nights spent columns for better understanding. The maximum no of properties are owned by the host with host id as 219517861 and the total no of properties owned by him are 327.

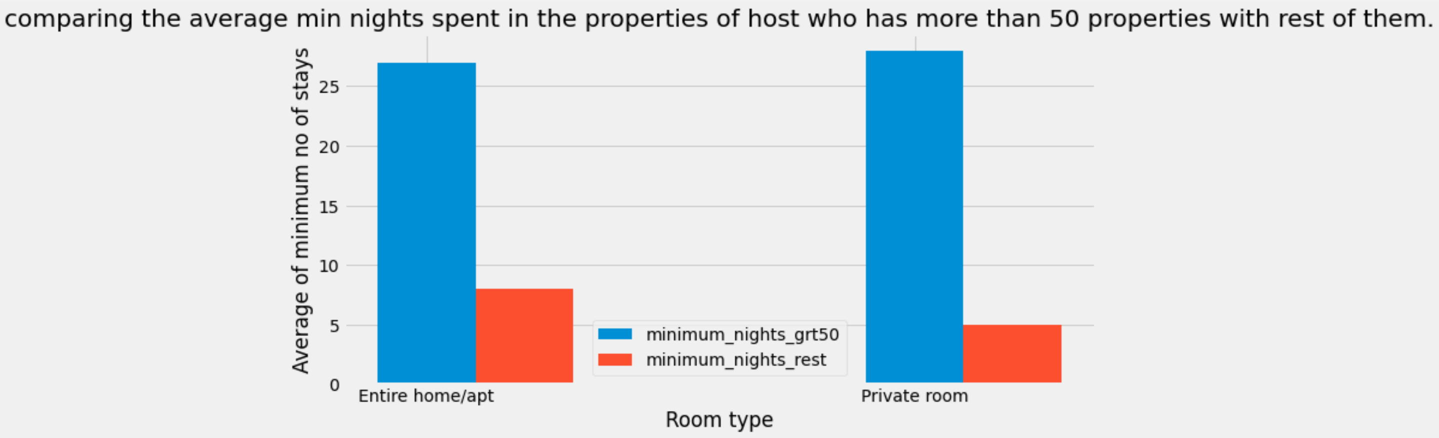

comparing the average min nights spent in the properties of host who has more than 50 properties with rest of them.

Fig 11: comparing the average min nights spent in the properties of host who has more than 50 properties with rest of them.

When we examine the above plot we can notice a considerable change in minimum nights spent in both the room types owned by the hosts who has more than 50 properties and rest of them. The ratio of minimum nights spent in host's property who has properties more than 50 to rest of them is approximately 3 for Entire home/apt and more than 5 for Private room type.

There may be n number of factors which influences the ownership of multiple properties, but one cannot succeed in managing them without satisfying his customer. They might providing some complimentary facilities to their customer's for the expansion of their business.

Scope of Improvement:

As dataset has few qualifying attributes to value a property, more features can be added like bedroom, bathroom, property age (it might be one of the most important one), tax rate applicable, distance to nearest airport, hospital or schools.

In order to have a better analysis regarding the quality of the properties, it would be interesting if we had an analysis of sentiments with property valuations.

User ratings of hosts aren't available, it would've been better to rank our hosts based on user satisfaction and ratings

In presence of ratings, hosts can be classified and ranked, special discount or offer can be given to highest rated hosts following marketing strategy.

Time series analysis can be done to make prediction on occupancy rate based on tourist season.

Conclusion:

That's it! We reached the end of our exercise.

Starting with loading the data so far we have done EDA, null values treatment, encoding of categorical columns, feature selection and data visualization. In this simple yet power full way we had done the EDA on Airbnb dataset, certainly this is not the end rather this this the start we can say as per business requirement changes we need to find the insights in that direction and justify the business problems. There can be n-number of questions and n-number of dimension to explore the dataset and find the insight from them, this there is no limit unless the business constrain is solved.

THE END