Exploratory data analysis

on

AIRBNB

Team Data Manipulators ↓

Nishant Kalanta (Team leader)

Abhishek Dhasmana

Dhrisanda Medhi

Himanshu Awasthi

Ramakrishnan Shahi

Data Science Trainee, Almabetter,

Bengaluru, Karnataka

Contents:

Introduction:

Airbnb (ABNB) is an online marketplace that connects people who want to rent a home with people who are looking for accommodation in a specific area. The company has come a long way since 2007, when its co-founders first came up with the idea of having paying guests sleep on air mattresses in their living rooms. According to Airbnb's latest data, Airbnb has over 6 million listings, covering over 100,000 of his cities and towns and over 220 countries around the world.

Since 2008, guests and hosts have turned to Airbnb to expand their travel choices and offer a more unique and personal way to experience the world.

Today, Airbnb is a unique service, used and recognized around the world. Data analysis of the millions of listings offered through Airbnb is a key factor for the company. These millions of lists generate a lot of data.

Data is analyzed and used for security, business decisions, understanding customer and vendor (host) behavior and performance on the platform, driving marketing initiatives, implementing innovative add-on services, and more.

Problem Statement:

Big firms like Airbnb generate data constantly, every day the number of data generated is roughly in millions. So the real challenge is to handle this data and handling data is vital for any firm to be on the right track, to know what customer want and are they getting what they asked for?

All these things are cleared or answered by handling the data or doing DATA ANALYSIS.

This analyzed data can be used for security, making business decisions, understanding customer and vendor (host) behavior and performance on the platform, driving marketing initiatives, implementing innovative add-on services, and more.

Data should be explored and analyzed to discover key concepts such as (but not limited to):

1- **What can I know about the various hosts and zones?**
2- **What can we learn from forecasting? (e.g. location, price, reviews)**
3- **Which host is the busiest and why?**
4- **Are there any noticeable differences in traffic between the different regions? Why?**

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 make predictions and conclusions based on the given data set.

Features:

'id' = Unique ID assigned to the entry.

'name' = This is a column have 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.

'neighbourhood' and 'neighbourhood_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 on hospitality.

'availabilty_365' = Provides information about offer availability.

Description:

Below is the numerical description of our dataframe.

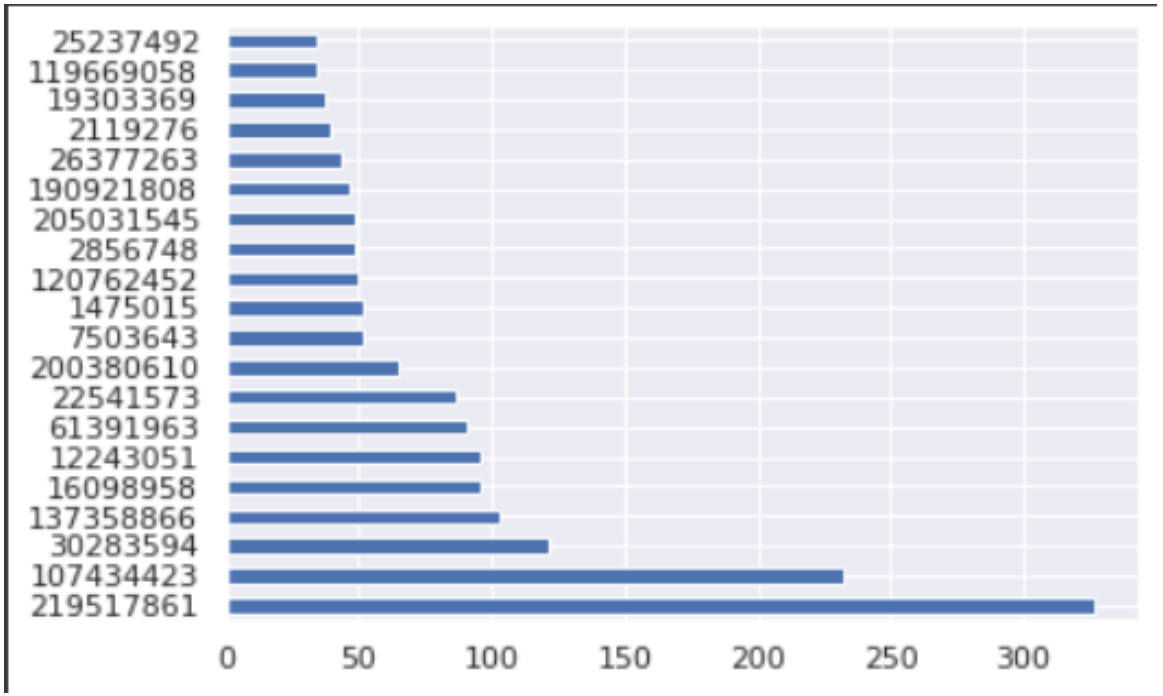
	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	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

(Figure 1)

As shown above figure, Statistical description of dataframe was returned

Plot analysis:

Plot analysis for host_id.



(Figure 2)

We can observe that the highest times transaction done by a customer is 327.

These 20 hosts (as shown in above figure 2 on y axis) are busiest host.

Analysis on host_name.

	host_name	neighbourhood	shared
451	Max	Lower East Side	23
622	Sergii	Bushwick	20
461	Melissa	Bedford-Stuyvesant	18
47	Anchor	Bedford-Stuyvesant	17
225	Güney	Hell's Kitchen	11
79	Baboucarr	Sunnyside	10
6	Abraham	East Harlem	9
198	Erik	Hell's Kitchen	9

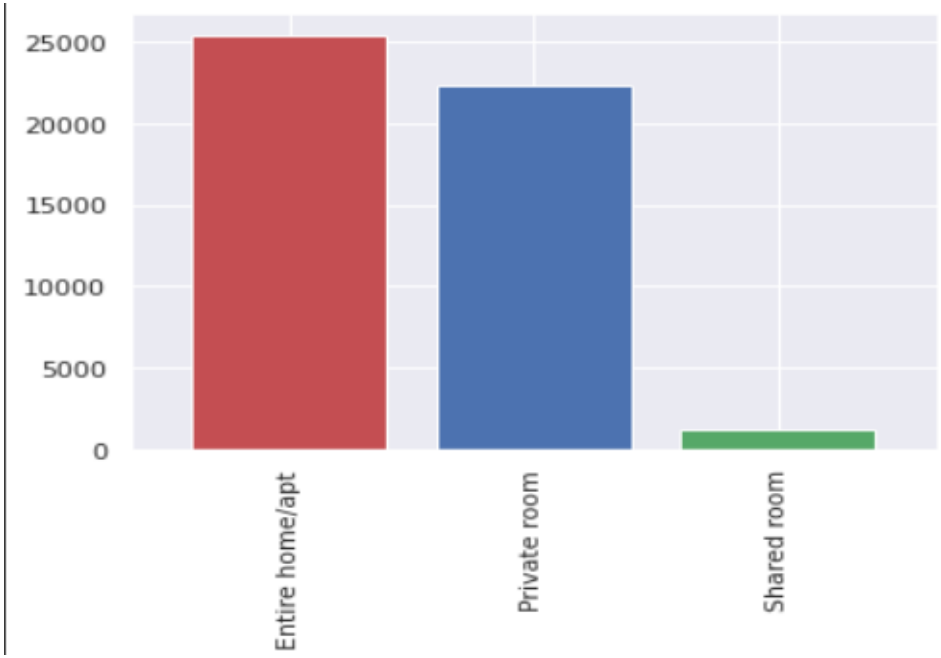
Figure in left, right and below shows the name of busiest host with their respective room type.

	host_name	neighbourhood	private
6603	John	Hell's Kitchen	47
7391	Kazuya	Woodside	35
4356	Eyal	Upper West Side	32
4689	Gabriel	Hell's Kitchen	31
10345	Nina	Bedford-Stuyvesant	28
14042	Zach	Fort Greene	26
14015	Yuval	Ridgewood	23
9540	Michael	Williamsburg	22

	host_name	neighbourhood	Entire home/apt
15502	Sonder (NYC)	Financial District	210
16961	Vida	Greenpoint	52
15497	Sonder	Financial District	51
15505	Sonder (NYC)	Murray Hill	50
13841	Red Awning	Midtown	49
15625	Stanley	Murray Hill	49
8734	Kara	Hell's Kitchen	41
2188	Blueground	Chelsea	37

(Figure 3(a, b, c))

Analysis on customers preference on room_type.

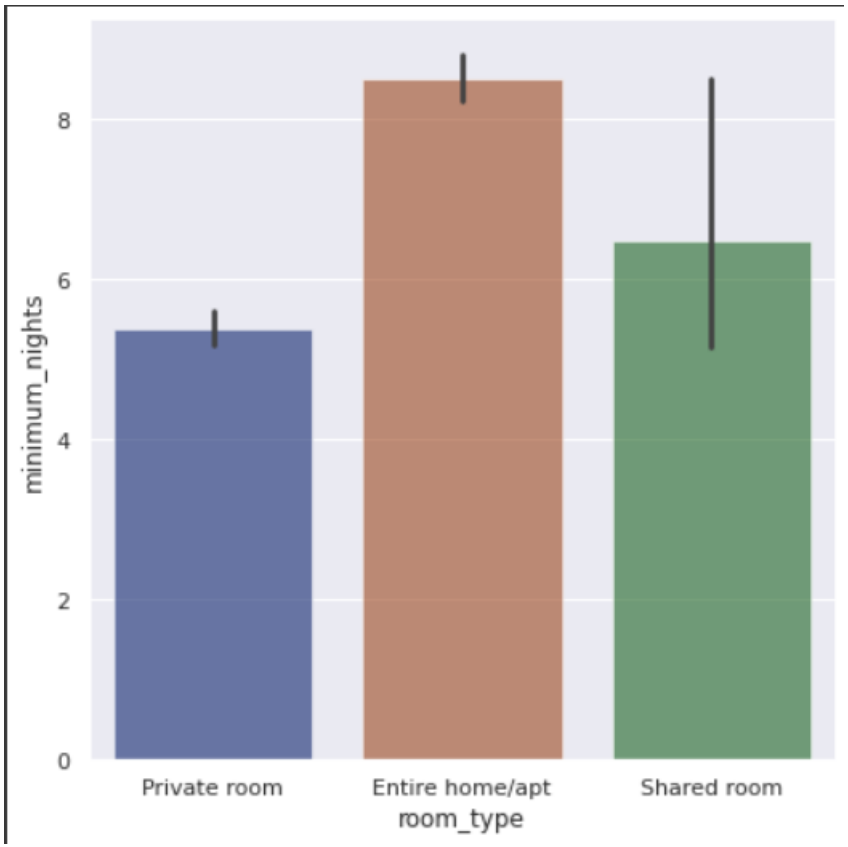


(Figure 4)

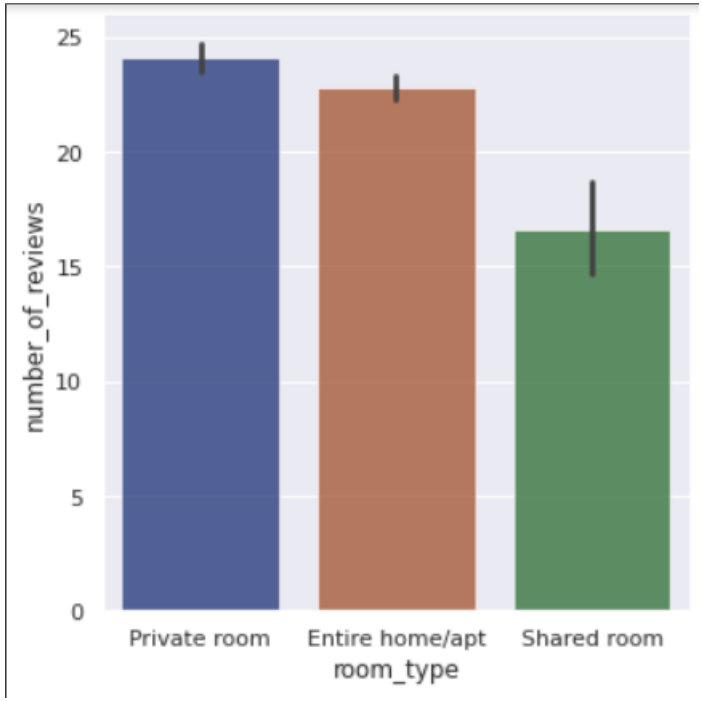
As we can see that most of the customers prefer entire home/apt.

The reason could be anything but from the numbers we can conclude that those who book entire home must have been staying with their family and hence we can provide services according to the basic need of a family.

Analysis on minimun_nights and room_types:



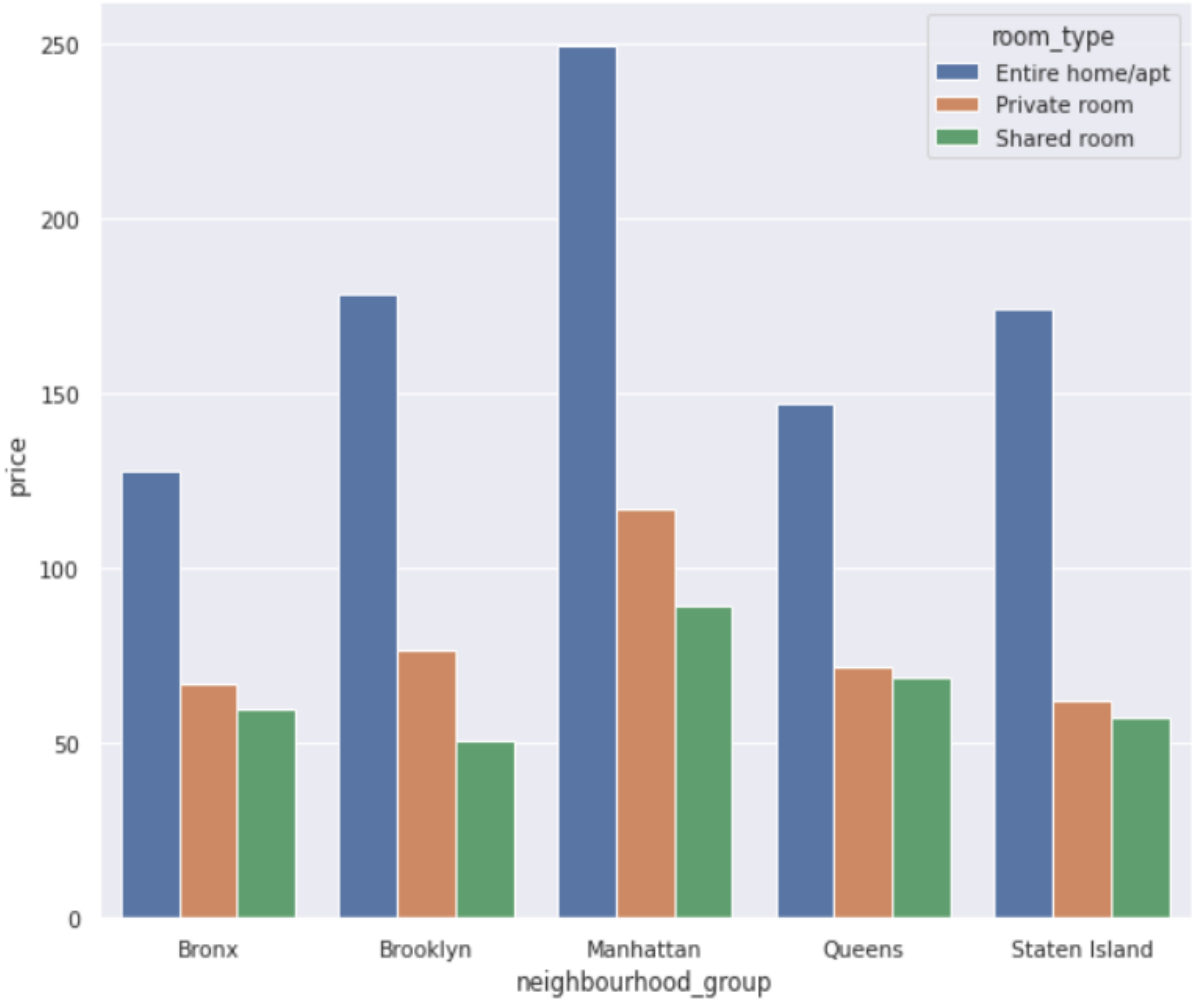
(Figure 5)



(Figure 6)

From figure 5 we can see that minimum nights spend on Private rooms and that maximum nights are spend on entire home. As we stated in *page 9*, that most of the entire home was booked by families this could also say that they wanted to spend more time to visit New York. From figure 6 we can conclude that number of reviews are comparable in Private room and entire home but least in shared room type.

Analysis on Neighbourbhood_groups and price with respect to room type:



(Figure 7)

In figure 7 we can see the plot between neighbourhood_group and price with respect to room type.

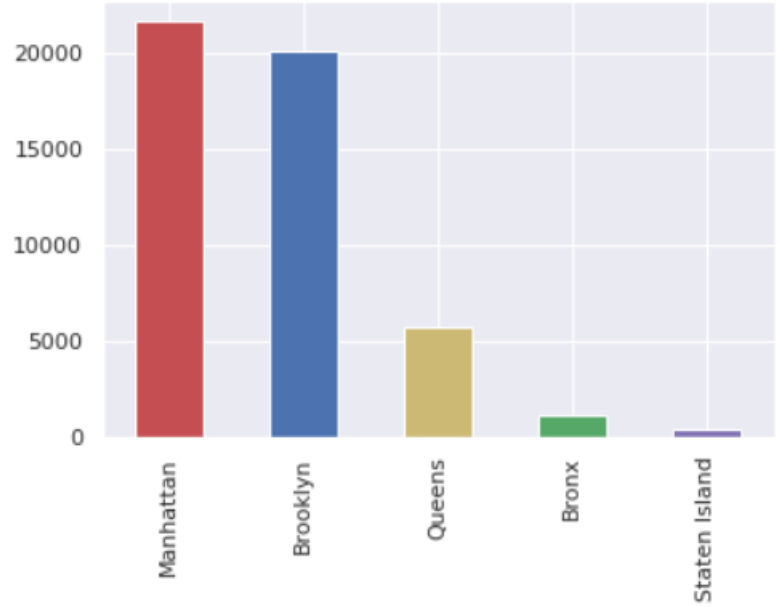
We figured: -

1 – Most expensive neighbourhood_group is Manhattan.

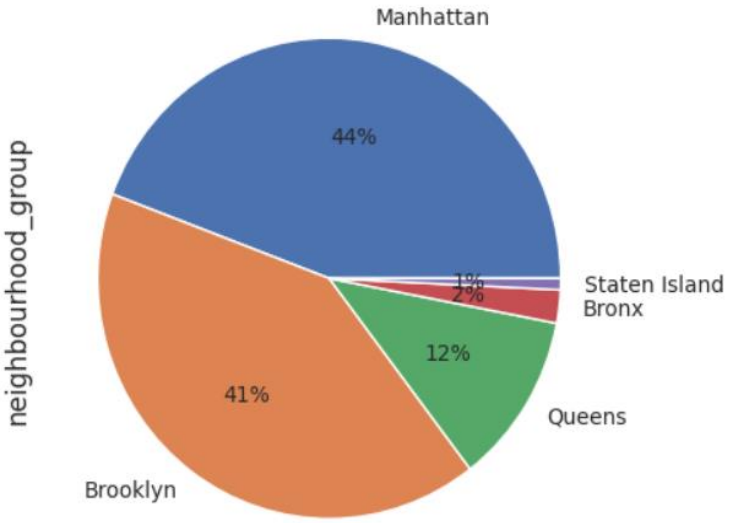
2 – In every neighbourhood_group the expensive room type is Entire home, and the least expensive is shared room with price comparable to private room type.

3 – We can also say that Bronx is having low average price of every room type.

Analysis on the most crowded neighbourhood_group.



(Figure 8-a)

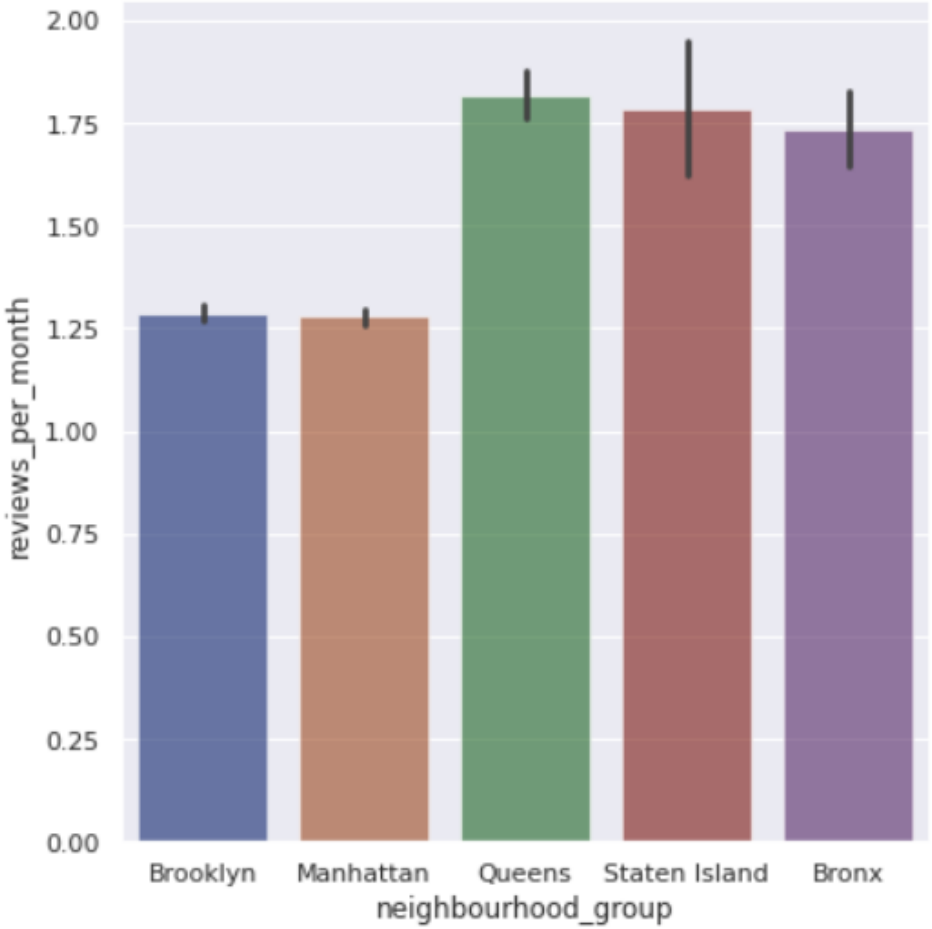


(Figure 8-b)

In *(Figure 8-a)* we can see that Manhattan and Brooklyn are the busiest neighbourhood_group and Staten Island is the least priority of the customers.

In *(Figure 8-b)* we can see that Manhattan is the city where most popular location with 44% of entire dataset. The least happened in Staten Island only 1%. Brooklyn consisted of 41% of transactions with 12% Queens and 2 % in Bronx.

Analysis on neighbourhood_group and reviews_per_month:

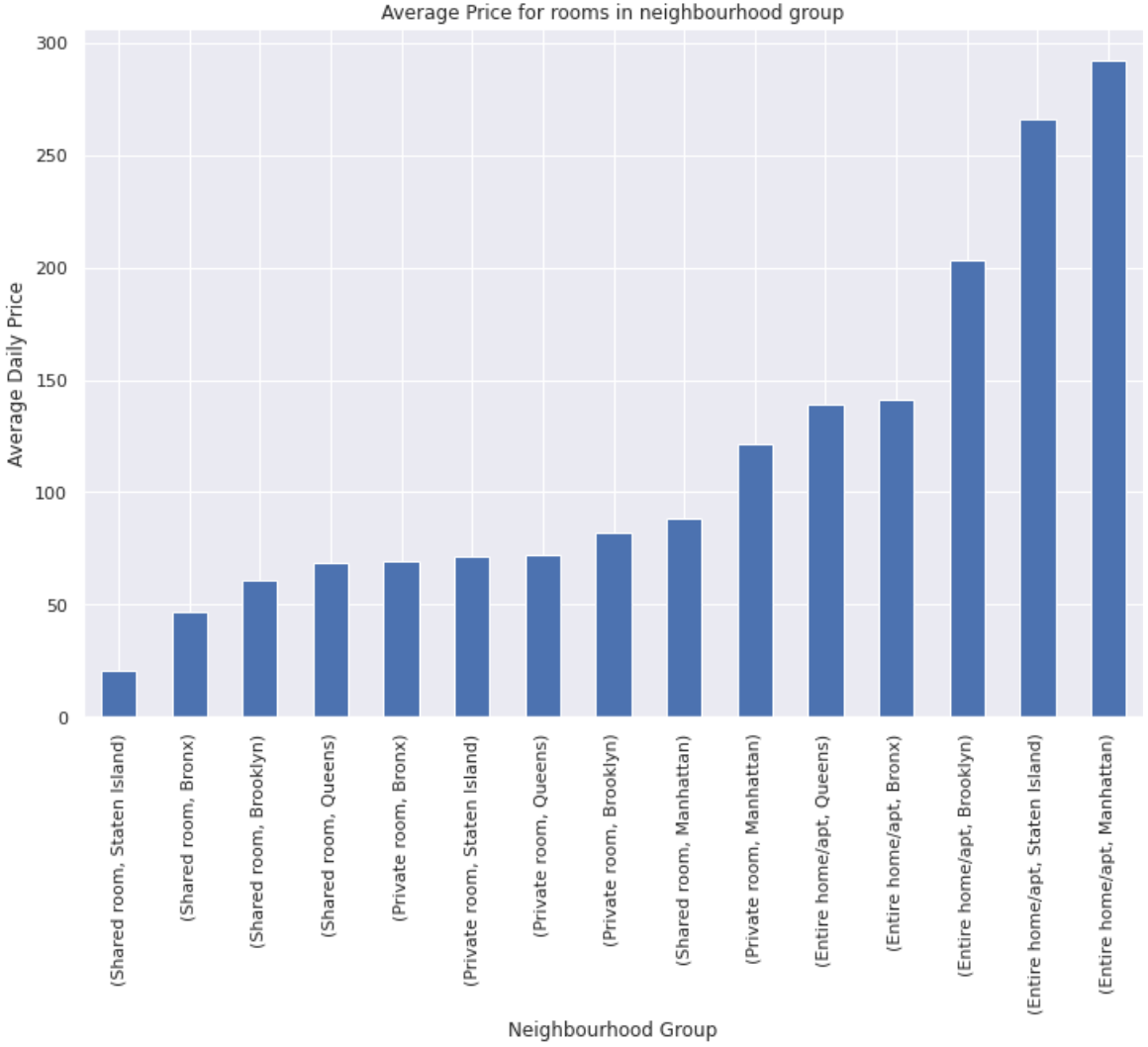


(Figure 9)

From figure 9 we can say that the reviews for the Queens and Staten Island are the most and Manhattan is the least.

Also, besides what could be the reason that least visiting places have maximum reviews?

Analysis on average daily price of different neighbourhood_group:

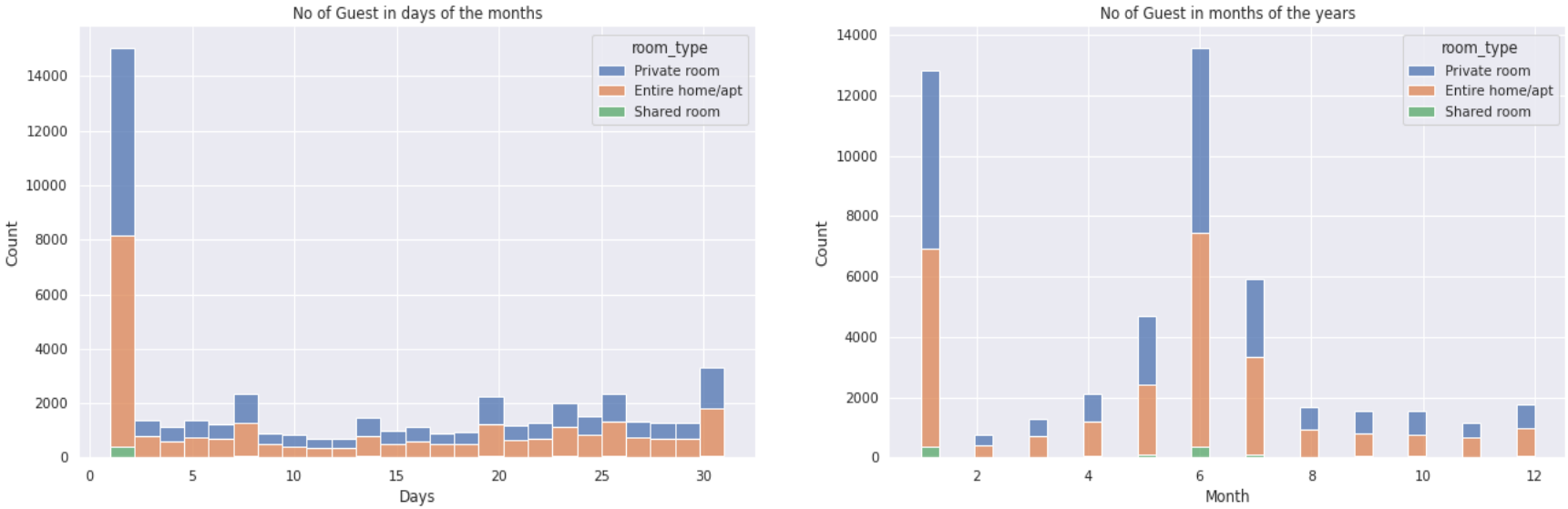


(Figure 10)

Here below is the table which shows the average daily price of different neighbourhood_group according to room types.

Entire home/apt	
Queens	139.036260
Bronx	141.541176
Brooklyn	202.895245
Staten Island	266.205128
Manhattan	291.784595
Private room	
Bronx	69.025862
Staten Island	71.394366
Queens	72.454958
Brooklyn	81.713284
Manhattan	121.434183
Shared room	
Staten Island	21.000000
Bronx	46.711111
Brooklyn	60.921212
Queens	68.459459
Manhattan	88.462898

Analysis on number of customers:



(Figure 11-a and b)

From combination of above two figures, we can conclude that the busiest day of the month is the start i.e. 1st and then the end of the month i.e. 31st. in the middle of the month the number of customers is quite low and comparable.

Also, the busiest month of the year is January and June.

After looking at different plots we can surely confirm that Entire home is the most preferable room type for the customers.

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:

What can we learn about different hosts and areas?

Predications = Host name Sonder (NYC) from Financial District have the most accommodation (Home and Apartments) i.e. 210 for all other private (47) and shared rooms(23)

- **The Staten Island is having the most available rooms throughout the year and Brooklyn is having the least available**

What can we learn from predictions? (Ex: locations, prices, reviews, etc)

Locations with respect to price availability

Predications = The room available in different locations are:

1. **Manhattan = 21661**
2. **Brooklyn = 20104**
3. **Queens = 5666**
4. **Bronx = 1091**
5. **Staten Island = 373**

- **The group with high number of neighbourhood_group (Manhattan) are having more costly room's and less room are having less price except Staten Island**

- **The Staten Island is having the most available rooms throughout the year and Brooklyn is having the least available**

Price with respect to availability throughout the year

- **As we can see that the most of the price is below the range of 2000 and the highest price for the hotel is 10000 for 3 hotels and among which one available all the time and one is not available all the time for some reason and one is available at some time and not available some time.**

Reviews with respect to room type and neighborhood.

- **All the room's get almost same type of reviews just having minor difference**
- **The reviews for the Queens and Staten Island are the most and Manhattan is the least**

Which hosts are the busiest and why?

Count of listing by top 10 hosts is almost 2.5% (1270 listings) of the whole dataset.

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

More customer preferred Manhattan location for night stay than Brooklyn • 63.2% customer spend night in Entire home and 1.6% spend night in Shared room.