Airbnb Booking Analysis by EDA

(Abhishek Mishra, Arunesh Mishra, Vineeta Singh, Tushar Gupta, Kurva Mallesh)
Data science trainees,
Alma Better, Bangalore

Abstract:

Airbnb is an online marketplace that connects people who want to rent out their homes with people looking for accommodations in that locale. We were provided the Airbnb dataset which has large number of observations in New York City. Analysis was done on this data set based on different parameters.

Our EDA can make us understand about the famous place to live, the famous host, the affordable place to live based on different parameters.

Keywords: EDA, Airbnb

1.Problem Statement

Airbnb is an online marketplace that connects hosts and guests. Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one-of-a-kind services that is used and recognized by the whole world. 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 analysed and used for security, business decisions, understanding of customers' and providers' (hosts) behaviour and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

The main objective is to do Exploratory Data Analysis (EDA) on the dataset which was provided, to conclude the facts about the certain features which are mentioned.

The Airbnb dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

Id: It is the unique ID given to Listings available in New York

Name: It is the description/name of the listings mentioned by hosts

Host id and Host name: This column gives information about the host

Neighbourhood and Neighbourhood group: These columns hold the information about the city and areas of the properties which contains the listings of Airbnb NYC.

Longitude and Latitude: It holds the information about the Longitude and latitude of the listings available in New York

Room type: It displays the room type of the property (either private room / entire home / shared room)

Price: It's an important column which holds the price value of all those properties.

Minimum nights: It gives us information about the minimum number of nights that is offered by hosts for particular listings.

Number of reviews and reviews per month: It gives information about the number of reviews and reviews per month for those properties and hosts hospitality.

Availability 365: Through this column, we can conclude for how many days the properties are available.

2. Steps involved:

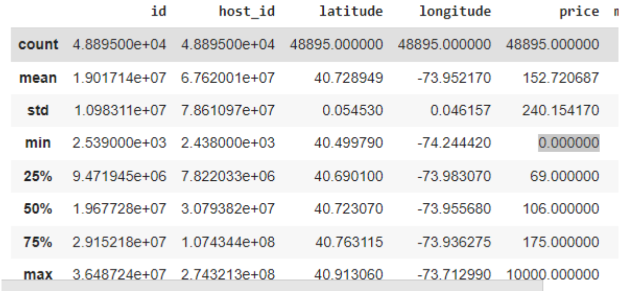
- **Data Wrangling:**

  After loading the dataset, we performed this method by cleaning, organizing, and transforming raw data into the desired format which makes us to understand the data clearly. This process helped us to tackle the unwanted data, to produce accurate results, to make better decision.

- **Null Value Treatment:**

  The data also had null values. To preserve all the information, we imputed or dropped the rows and columns containing null values while conducting exploratory analysis that made use of these features. id', 'last_review' attributes are dropped from the initial analysis part 'reviews_per_month' attribute has null values which are then replaced with 0. Because we can assume that there should be no host without the reviews on Airbnb data. we have treated the null values by filling with zeros in order to produce more accurate results.

- **Data Cleaning**
- Here in the below table, we can easily identify the presence of outliers for column 'Price'. There we can see that the min price is 0$ and maximum price is 10000$.
- After doing some background research it is admissible that Airbnb customers mostly searched for rent between 20$ to 800$.
- Removal of outliers if possible will give new ways to look into the data. Two graphical techniques for identifying outliers are scatter plots and box plots initially.



	id	host_id	latitude	longitude	price
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000

- **Exploratory Data Analysis (EDA):**

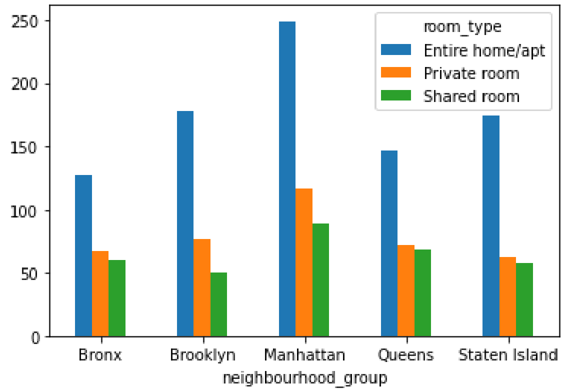
  After the data wrangling step we performed EDA by comparing different parameters which are involved in the dataset. EDA help us to find the different relations among the parameters. It involves the visualization of the data by comparing the different parameters to find out the best among all.

3. EDA

In the process of understanding the data we found that price, neighbourhood group, room type, reviews and host are most important parameters in Airbnb. So, we performed analysis based on neighbourhood group, room type, price and host.

Which area is most expensive?

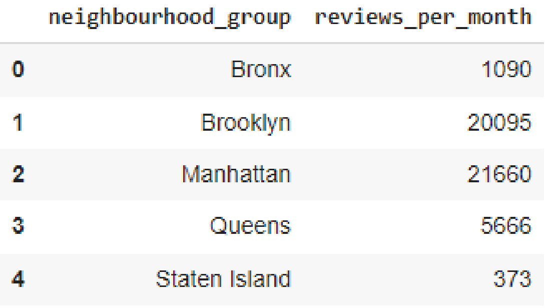
Here we have analysed the neighbourhood_group and the room type against the mean of the prices.



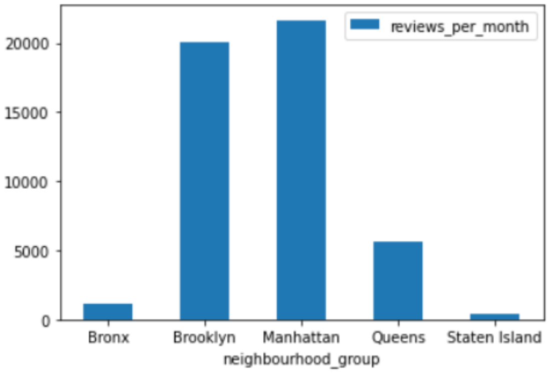
From the above bar chart Manhattan is the clear winner when it comes to high rents. Airbnb customers prefer again entire home/apt to be booked in Manhattan. After doing research again as Manhattan considered as the financial capital for US then guests visit will also be considered high. It can also be said that as Manhattan being the expensive state then could be possible that up and down of rent may lead to proportionate change in bookings.

Which is the busiest neighbourhood group?

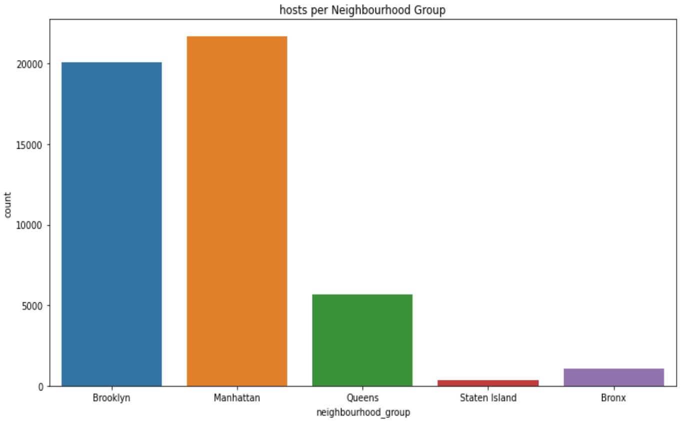
Here we have analysed the demand of Airbnb customers for location and their reviews

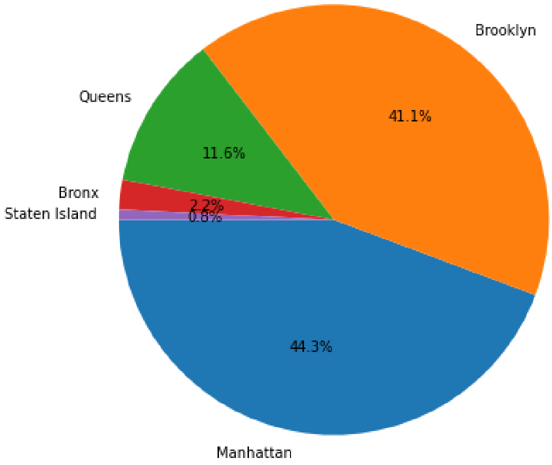


	neighbourhood_group	reviews_per_month
0	Bronx	1090
1	Brooklyn	20095
2	Manhattan	21660
3	Queens	5666
4	Staten Island	373



After the analysis of attributes 'neighbourhood_group' containing information about different states of US and 'reviews_per_month' have information about the customers ratings given to particular host present in a particular state. We can clearly state from the above table that Manhattan and Brooklyn are the busiest neighbourhood groups because of the reviews given by customers.

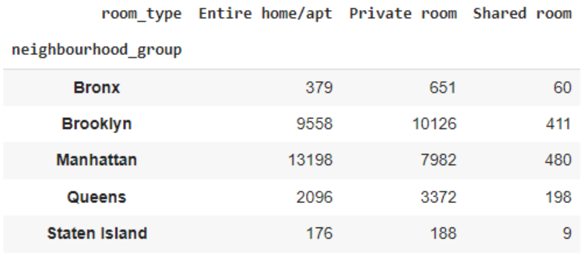




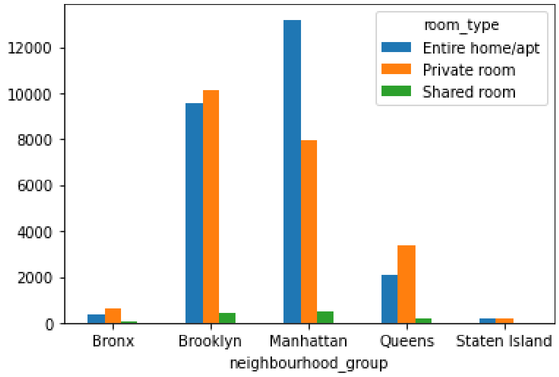
From the above shown bar chart and pie chart it is very clear that the highest 21642 (approx. 44%) hosts available in Manhattan and second highest 20080 (approx. 41%) hosts available in Brooklyn. It can be clearly understandable that majority of the hosts are from these two areas alone. We have already seen that the Manhattan is more expensive neighbourhood group.

Type of rooms preferred by hosts in different neighbourhood group

Here we have analysed the data of the number of available room types (such as entire room/apt, private room, shared room) in each neighbourhood group.



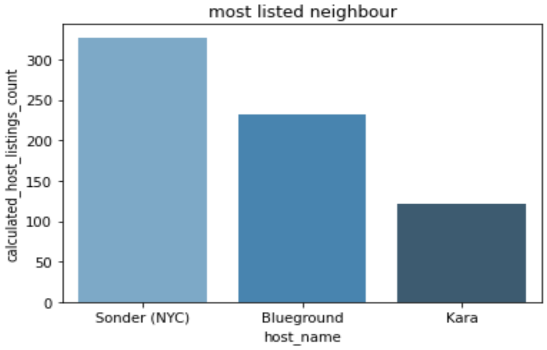
room_type	Entire home/apt	Private room	Shared room
neighbourhood_group			
Bronx	379	651	60
Brooklyn	9558	10126	411
Manhattan	13198	7982	480
Queens	2096	3372	198
Staten Island	176	188	9



The above plot clearly says that shared rooms are least populated. that most of the people select privacy and safety over the less price. So, it is evident that most of the people preferred to stay in the entire home/apt and private rooms.

Top Busiest hosts insights:

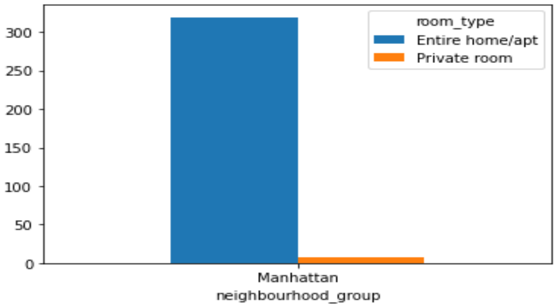
We made analysis between Neighbourhood group who has most listed host in country. After data analysis we found that Sondar is the most demand side for the booking. when we search, we found that Sondar is a hotel chain running in Manhattan.



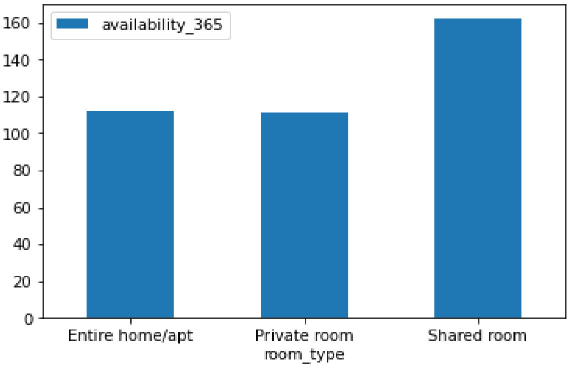
After more exploratory data analysis, we have found out more interesting facts about some of the hotel chains or location, about the availability of hotels or hosts.

- When we deep dive inside the information about hosts. We concluded the below graph.
- After doing some google research, we found that Sonder is a hotel chain company which used Airbnb for listing its property. it's hotel is based on Manhattan.

The below plot tells the shared room have less no of booking because the people of U.S preferred the privacy over the price. That is why Manhattan and Sondar hotel chain either did not have any shared room type option or very less rooms available

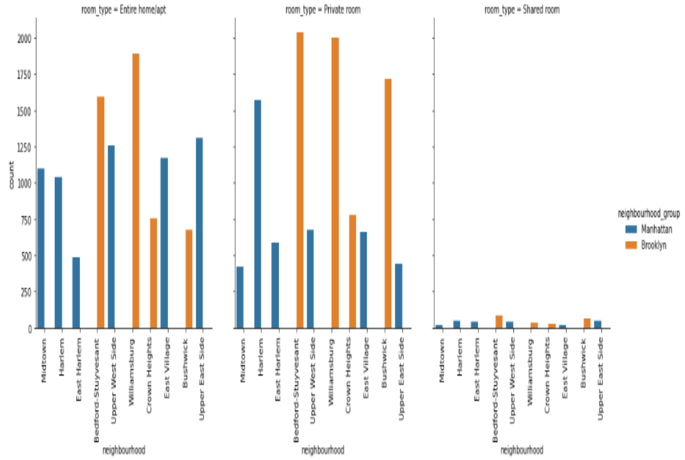


Room type and Availability All Year



This graph above shows the availability of different types of rooms. Clearly shared rooms have high availability in the coming 365 days as they are least preferable.

Plot for top ten neighbourhood showing the preferred room type



From the above graph for three different types of rooms, what we observed is that the shared room type has less no of interest. we can see from the top 10 neighbourhood only Manhattan, Brooklyn are the most travel destination therefore would have the most listing availability. We can also observe Bedford-Stuyvesent and Williamsburg are most populate from Manhattan and Harlem from Brooklyn.

Conclusion:

- Most Airbnb is located in Manhattan and Brooklyn due to capital income state. Manhattan is most expensive

- Correlation graph shows that in the U.S prefer privacy then the price so the entire home is the most preferable room type followed by private rooms.

- Correlations between price and review per show that paying extra always be less satisfying so Airbnb with more reviews be affordable than average, but not a strong negative correlation there.

- The overall satisfaction level was affected primarily by the type of the room i.e., private room, shared room or entire home

References-

1. Google Search
2. GeeksforGeeks
3. Stack Overflow