**AIRBNB EXPLORING DATA ANALYSIS**

**SAI VENKATA RAVI TEJA ANDE,**

**AMOGHA K**

**SUNIL KUMAR BODUGU**

**Data science trainees,**

**AlmaBetter**

Abstract:

Airbnb stand for Air, Bread and Breakfast. It is an online marketplace that connects people who want to rent out their homes/spare rooms with people who are looking for accommodations in specific locales. It has millions of listings, which generate lots of data. We are analyzing this data for making business decisions, for looking best room type, affordable, best place.

We have our dataset from Airbnb based on NY and it has 48,895 rows and 16 columns. NY is the most expensive city in the USA. From data, we retrieved the basic information to understand what data is saying and we started looking out for null values, duplicate values to get better understand of the data and we found 4 column values are missing. After that we cleaned the data entirely and we started the Exploratory data analysis (EDA) to get insight from the data. The EDA helps customer to make choice decision which room to take according to their price, availability etc.

While doing exploratory data analysis we had faced many difficult things such as code error on python code, how to deal with data etc. and we came out from all of that by taking help teammates and from google, other website related to data science. It was good hands-on experience for me and learns lots things in this project.

**1.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:

1.What can we learn about different hosts and areas?

2.What can we learn from predictions? (ex: locations, prices, reviews)

3.Which hosts are the busiest and why?

4.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.

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,

a.Name column having total 16 null values.

b.Host\_name column having 21 null values.

c. Last\_review and reviews\_per\_month are having 10052 null values

d.We will look at the columns and decide what we can do with them

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 |

|  |  |
| --- | --- |
| 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 |

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.

Bar chart of good reviews on each neighbourhood groups.

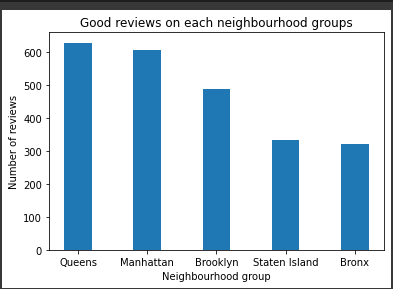
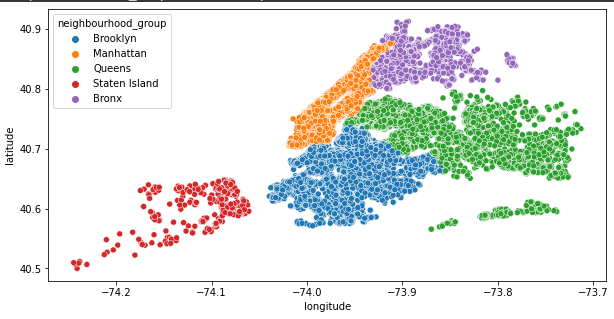


Fig 3. Good reviews on each neighbourhood groups

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 |

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.

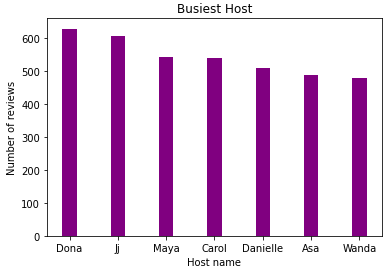
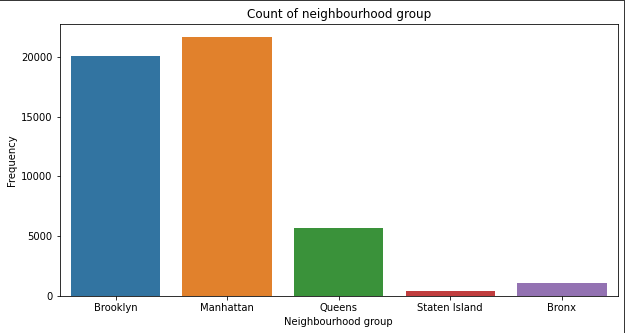
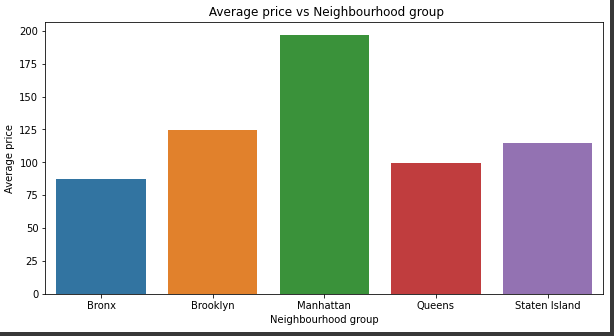
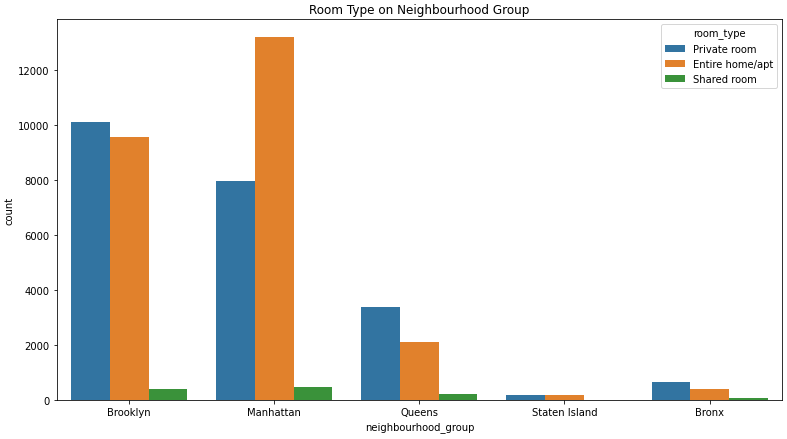
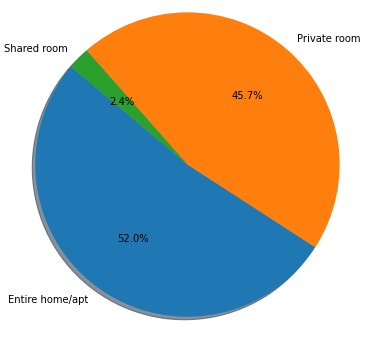


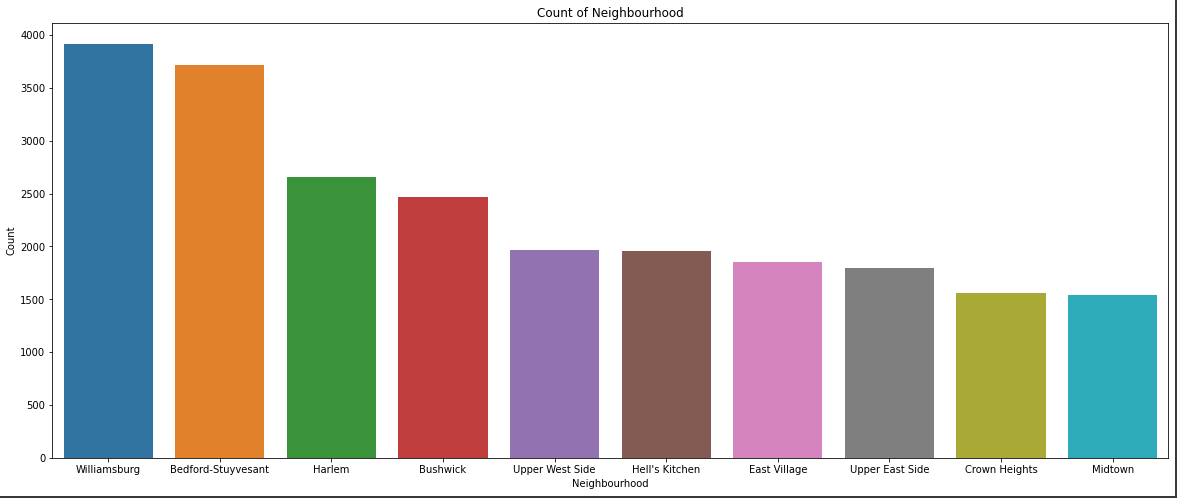
Fig 4.Top 7 busiest hosts

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

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



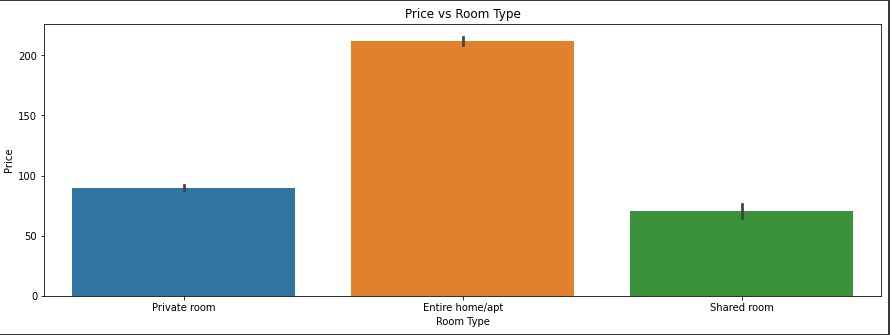
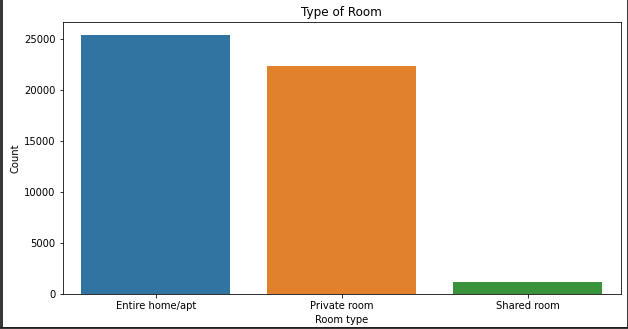


From above graph it can be understood that Brooklyn and Manhattan stand within the most urban and active area, in terms of listing count and pricing both. We further analyzed the room types occupied by a neighborhood group. Shared rooms are the cheapest and also has lowest count in every neighborhood, 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!

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.

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

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 |  |
|  | Neighbourhood | Total Revenue($) |
| Entire home/apt | Williamsburg | 389724 |
| Private room | Williamsburg | 171265 |
| Shared room | Hell's Kitchen | 9488 |

|  |  |  |
| --- | --- | --- |
| Room type | Minimum revenue details |  |
|  | Neighbourhood | Total Revenue($) |
| Entire home/apt | New Dorp | 57 |
| Private room | Graniteville | 20 |
| Shared room | Randall Manor | 13 |

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 consider 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.



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.

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 provide some complimentary facilities to their customers for the expansion of their business.

8. Conclusion:

Finally, we are done with the analysis and some insight from the data. By the analysis we can say that Manhattan has most listed neighborhood and has overall good reviews compared to others. The customer will choose either Entire home or Private room for stay and revenue from Entire room is highest from all other room types.

Shared room are more affordable in all the neighborhoods but customer will not prefer to stay in that. Queen and Manhattan have good number of visitors. If customer is staying for 1-2 days, then they might be travelers.

References-

1. GeeksforGeeks
2. Towards data science
3. GitHub
4. Stack overflow
5. Google search ad YouTube