**Airbnb Booking Analysis**

**Raneev K , Radha Gharima**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

Airbnb is an online marketplace connecting travelers with local hosts. On one side, the platform enables people to list their available space and earn extra income in the form of rent. On the other, Airbnb enables travelers to book unique homestays from local hosts, saving them money and giving them a chance to interact with locals. Catering to the on-demand travel industry, Airbnb is present in over 190 countries across the world.

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 service 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 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.

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

**Problem Statement:**

We have a dataset of Airbnb having almost 16 columns and 48k entries,we are going to analyse this dataset to find solution for some business problems,they are,

What can we learn about different hosts and areas?

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

Which hosts are the busiest and why?

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

**Exploring Dataset :**

Lets take a look at given dataset and understand different columns.

**Id**:

This is the unique identity number of each room/apartment which is an “int(64)”datatype having 48895 entries and has no null values.

**Name**:

This is the name of room/apartment ,which is “object” datatype and having 16 null values.

**Host\_id**:

This is the unique identity number of each host which is an “int(64)”datatype having 48895 entries and has no null values.

**Host\_name**:

This is the name of host ,which is “object” datatype and having 21 null values.

**Neighbourhood\_group**:

This column represents a broad area where the particular room/apt is located. There are five neighbourhood\_group present in our dataset they are 'Brooklyn', 'Manhattan', 'Queens', 'Staten Island', 'Bronx'.This is an “object” datatype having no null values.

**Neighbourhood**:

This is the local place within each neighbourhood\_group,where the particular room/apt is located. This is an “object” datatype having no null values.

**Latitude**:

This columns gives idea about the particular geographical location where the room/apt is located. This is an “float(64)” datatype having no null values.

**Longitude**:

This columns gives idea about the particular geographical location where the room/apt is located. This is an “float(64)” datatype having no null values.

**Room\_type**:

This column gives idea about particular room type.There are three room types available in this given dataset they are “private rooms”, “shared rooms”, “entire home/apt”. which is an “object” type dataset having no null values.

**Price**:

This column gives price data for each rooms, price varies accordings to different locations and different room types.Which is an int(64) datatype having no null values. As per dataset Maximum price for the room is 10000$ and mean price is 152.7 $. This is an “int(64)” datatype having no null values.

**Minimum\_nights**:

Minimum number of nights people stayed in each room/apt.As per dataset maximum value is 1250 and minimum value is 1. on an average, people stay 7 days in a room. This is an “int(64)” datatype having no null values.

**Number\_of\_reviews**:

Total number of reviews for each rooms/apt. The maximum value is 629 and the mean value is 23. This is an “int(64)” datatype having no null values.

**Last\_review**:

The date at which particular room/apt got last review,which is given as an “object” datatype,it should be in datetime format but for our current project this column has no significance so we can left it as it is.This column also has 10052 null values.

**Reviews\_per\_month**:

This column is all about how many reviews are getting per month for each room/apt.Which is a “float(64)” type dataset having 10052 null values.

**Calculated\_host\_listings\_count**:

This column represent how many listings are there corresponding to each host\_id. This is an “int(64)” datatype having no null values. The maximum number of host listing is 327 and the average value is 7.

**Availiability\_365**:

This column gives data about availability of rooms per year. This is an “int(64)” datatype having no null values.

**Data Cleaning**

In this process we explore the number of rows & columns, ranges of values etc. Then Handle missing, incorrect and invalid data and Perform any additional steps (parsing dates, creating additional columns, merging multiple dataset etc.)

There we have total 4 columns with null values but "last\_review" has no significance for this current EDA project.The missing values present in "reviews\_per\_month" can be replaced by 0.

Now lets take a look at "availability\_365" column, there we can see eventhough availability is zero, there are some corresponding values available in "reviews\_per\_month" column.So that means that airbnb reviewed some rooms having no guests yet.We can drop those values.

Null values

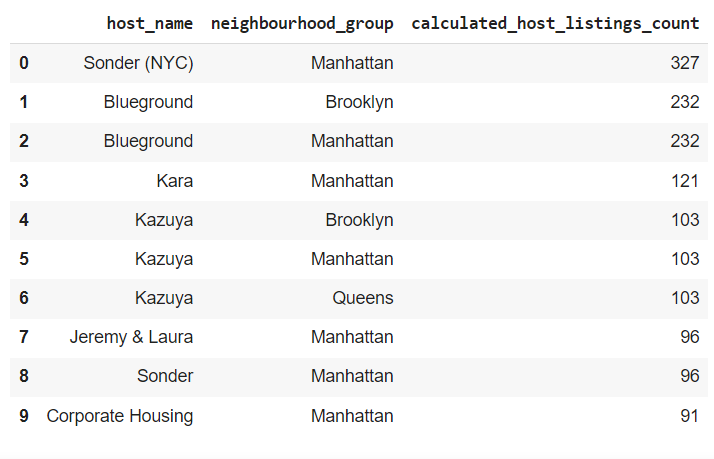

Figure 1- Null values

**Exploratory Analysis and Visualization**

Exploring the data by analyzing its statistics and visualizing the values of features and correlations between different features. Explaining the process and the results.Now that we are ready for an exploration of our data, we can start working from left to right. The reason some may prefer to do this is due to its set approach - some datasets have a big number of attributes, plus this way we will remember to explore each column individually to make sure we learn as much as we can about our dataset.

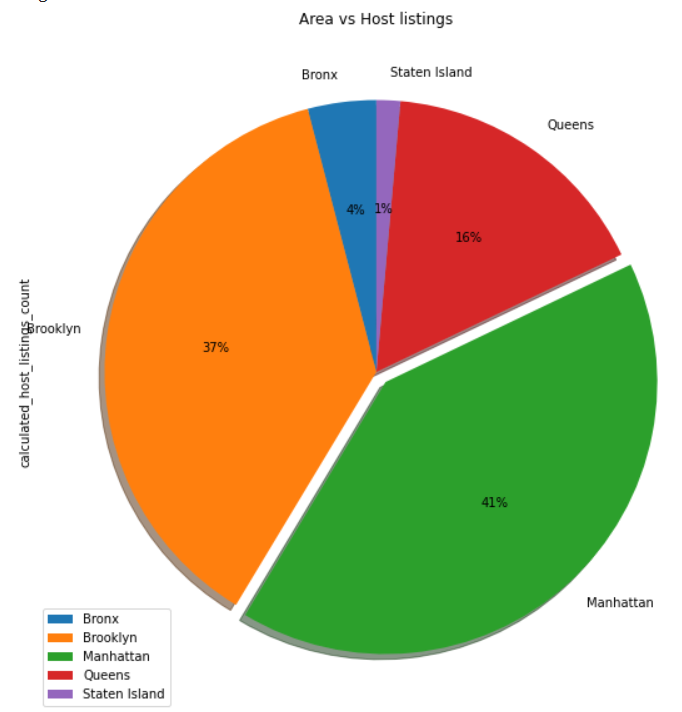
**Objectives and Approach**

* **What can we learn about different hosts and areas?**



From the top 10 observations accordings to highest “calculated\_host\_listings\_count”, we can found that 7 results are from -Manhattan area,2 from Brooklyn and 1 from Queens.So it is clear that airbnb is a popular businesss model in manhattan.The host who has most host\_listings are sonder(NYC).

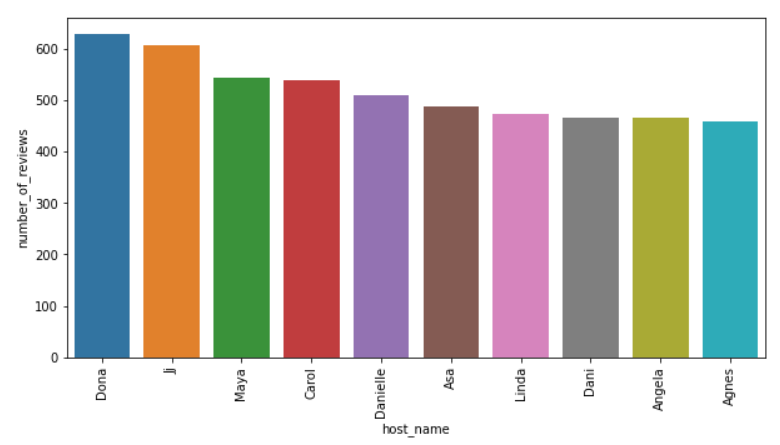
Now lets find out distribution of Airbnb rooms in each areas by plotting a pie diagram



Manhattan has maximum host listings which is 41% of the entire listings,then brooklyn has 37%,Queens has 16%,Bronx has 4% and at last staten island has only 1% listings.

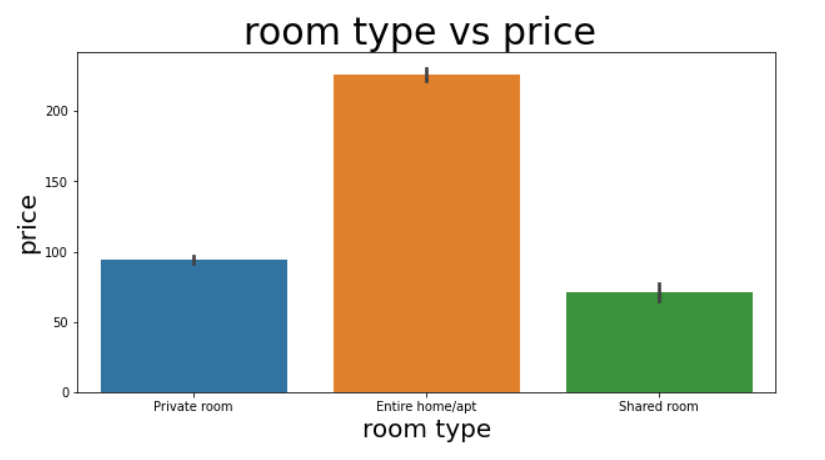
* **Which hosts are the busiest and why?**

Lets find out who is the busiest host.For this we are going to do an analysis on “number of reviews and “host name”.Those hosts who has got maximum number of reviews obviously will be the busiest one.



* From top 10 list of busiest hosts we can see that 8 results are from private room type.
* The name of the busiesthost is Dona from queesarea having 629 review.
* **What can we learn from predictions? (ex: locations, prices, reviews, etc)**

Now lets do an analysis on room type and price ,we can find which room type has high price and which one is cheap.

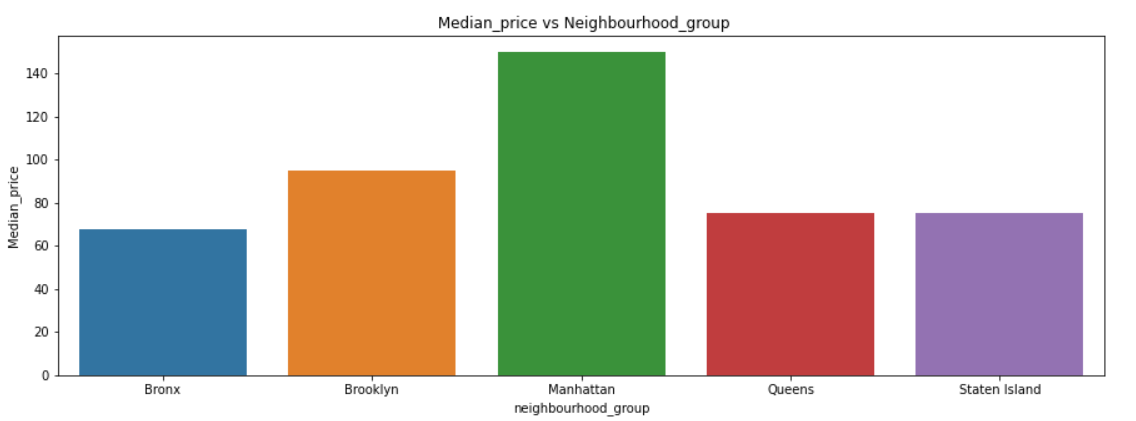


Shared rooms are the cheapest which has an average price of 71$.

Private rooms are little costlier than shared rooms which has an average price of 94$.Entire home/apt are the most expensive out of these three,the average price is 226$.

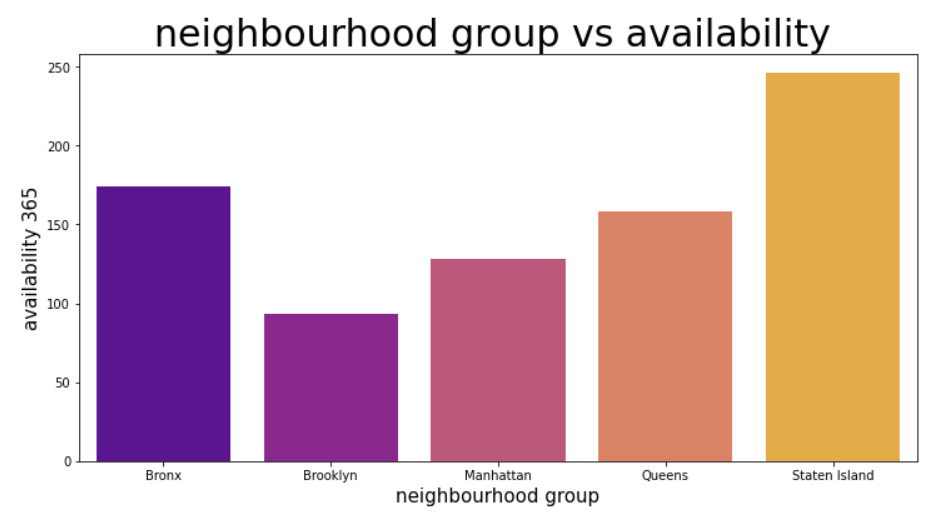
Even though shared rooms are much cheaper as compared to private rooms most reviews are for private rooms which means that most number of people prefer private rooms over other room types.Probably couples or small group of peoples who need more privacy and prefer budget friendly stay will go for private rooms,so we can say that most of the airbnb customers are from this category.

Now lets do an analysis on neighbourhood\_group and median\_price so that we can find out which area has highest price.



'Manhattan' is the area having highest median price,followed by Brooklyn.So we can say that most costly hosts are situated in "manhattan".

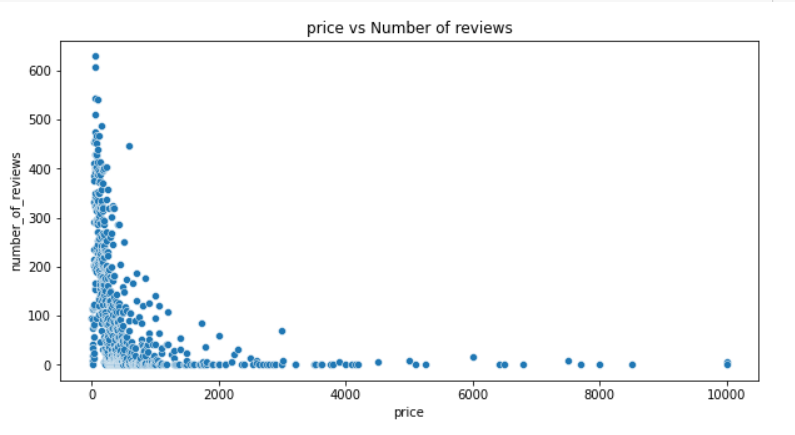
Lets take a look at availability column and find areas having high availability



**Some Observations From graph of “Neighbourhood group vs availability”**

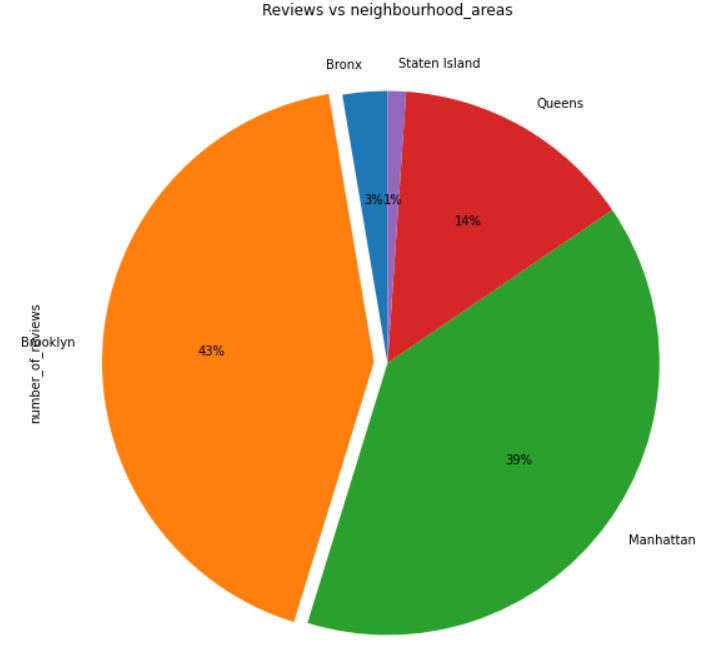
1. Room availability is high in "Staten Island" ,the average value shows that rooms are available for 246 days in each year.
2. Room availability is very low in "Brooklyn", as per available data rooms are available only for 93 days in each year.
3. From the previous observations we found that maximum host listings are in "manhattan area" but room availability is very low as compared to other areas that means number of people visiting "manhattan" is more than available rooms. Which creates high demands for rooms and thats the reason for high price.

For business development it is really important to understand customer’s choice.We can plot a scatterplot on “price” and “number\_of\_review” so that we can find out at what price range reviews are more.



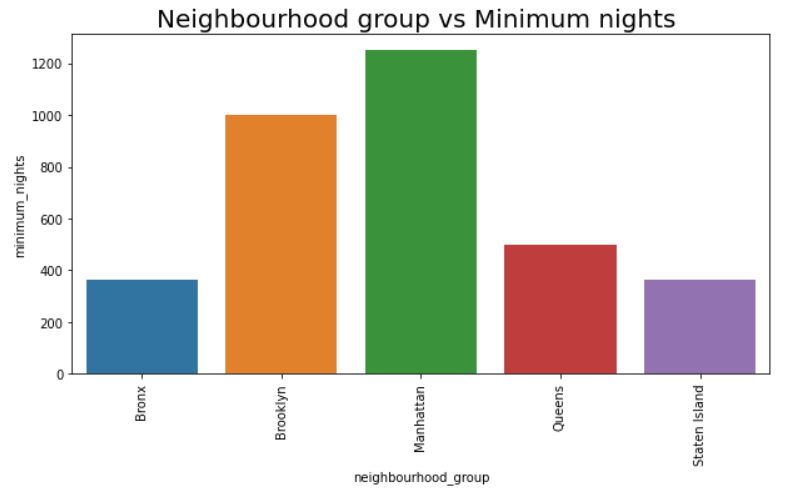
* Number of reviews are more at low price and reviews decreasing when price increases.
* From the scatterplot it is clear that most number of people prefer budget friendly rooms.
* **Is there any noticeable difference of traffic among different areas and what could be the reason for it?**

From the previous observations we seen that maximum host listings are in "manhattan area" and "brooklyn" but room availability is very low in both these areas that means number of people visiting these areas are high. To strengthen this assumption lets do some more analysis based on number of reviews in different neighbourhood\_areas.



From the diagram it is clear that both "manhattan" and "brooklyn" has got most number of reviews which strengthen our assumption that these areas are having high traffic.

Lets do some analysis using “minimum\_night” column



From minimum nights data,it is clear that people like to spend more days in "manhattan" and "brooklyn".

From these last three observations based on review,minimum nights and availability, we can say in “Manhattan” and “Brooklyn” has more traffic.

**Conclusions:**

1. From the top 10 observations accordings to highest calculated\_host\_listings\_count, we can found that 7 results are from “Manhattan” area,2 from “Brooklyn” and 1 from “Queens”.So it is clear that airbnb is a popular businesss model in “Manhattan”.The host who has most host\_listings are “sonder(NYC)”.
2. From top 10 list of busiest hosts we can see that 8 results are from private room type.
3. The name of the busiest host is “Dona” from “queens” area having 629 reviews.
4. Even though shared rooms are much cheaper as compared to private rooms most reviews are for private rooms which means that most number of people prefer private rooms over other room types.probably couples or small group of peoples who need more privacy and prefer budget friendly stay will go for private rooms,so we can say that most of the airbnb customers are from this category.
5. 'Manhattan' is the area having highest median price,followed by “Brooklyn”.So we can say that most costly hosts are situated in "manhattan".
6. Room availability is high in "Staten Island" ,the average value shows that rooms are available for 246 days in each year.
7. Room availability is very low in "Brooklyn", as per available data rooms are available only for 93 days in each year.
8. From the previous observations we found that maximum host listings are in "manhattan area" but room availability is very low as compared to other areas that means number of people visiting "manhattan" is more than available rooms. Which creates high demands for rooms and thats the reason for high price.
9. From the scatterplot it is clear that most number of people prefer budget friendly rooms.
10. by analysing available datas of "number\_of\_reviews","minimum nights" we found "manhattan" and "brooklyn" are the high traffic areas.