**Team Capstone Project - AirBnb**

**Biswajit Paul, Shreya Pattanayak**

Data science trainees - AlmaBetter, Bangalore

**Abstract:**

AirBnb operates on online marketplace for primarily [homestays](https://en.wikipedia.org/wiki/Homestay) for [vacation rentals](https://en.wikipedia.org/wiki/Vacation_rental), and [tourism](https://en.wikipedia.org/wiki/Tourism) activities. It is the most popular rental stay due to its affordability and collaboration with the countries and city all over the world.

After collecting the relevant data of Airbnb hotel listings and performing different statistical analysis and tests, we can have an insight of the factors that impact the most and least on the overall satisfaction level of customers.

***Keywords: Data, Pandas, Matplotlib, Seaborn, Data Visualization, Exploratory data analysis etc.***

**What we have to find out**

* Predict whether a new AirBnb user will effectively book a destination or not.
* Predict which country a new AirBnb user's first booking destination will be.

**Introduction**

### The Air platforms adjust their prices using a specific algorithm which is real time and dynamic known as **“Surge Pricing”** or **“Dynamic Pricing”**. It gives output a multiplier which is adjusted along with the base fare, locality and reviews.

### Our goal here is to build a predictive model, which could help customers in finding the best option among many.

## **Stepwise process :**

## **Data importing/mounting**

airbnb\_df = pd.read\_csv("/content/drive/MyDrive/EDA Capstone project/"+ "Airbnb NYC 2019.csv")

or [Airbnb NYC 2019.csv](file:/C:/Users/Zindaram/Downloads/Airbnb%20NYC%202019.csv)

## **Data profiling and cleaning**

In this we clean the data in useful manner to find out solution.

*Range Index: 48895 entries, 0 to 48894*

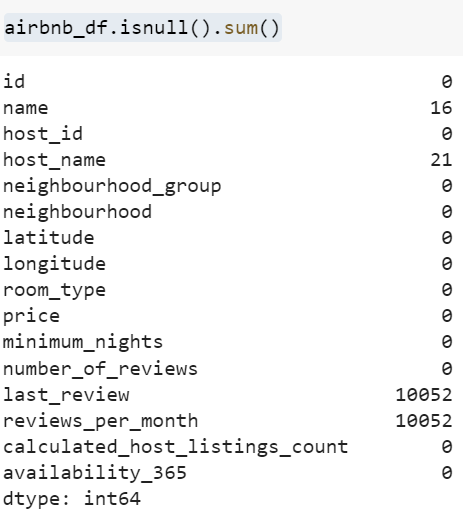
*Data columns (total 16 columns):*

*Data types: float64(3), int64(7), object(6)*

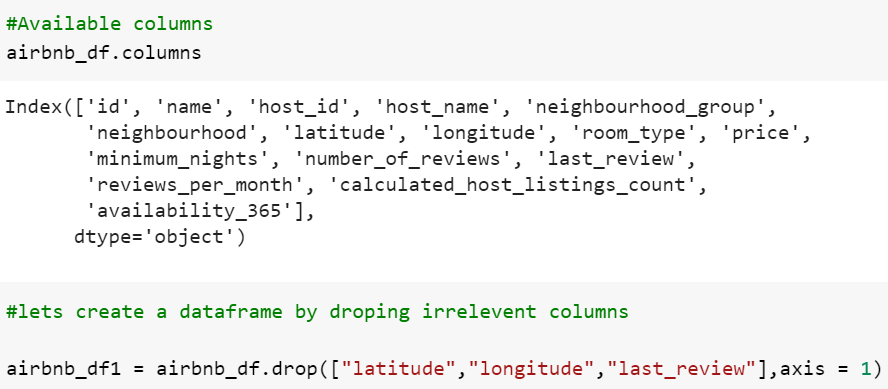
*memory usage: 6.0+ MB*

**airbnb\_df.describe()** – using this in Pandas describe() is used to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values.

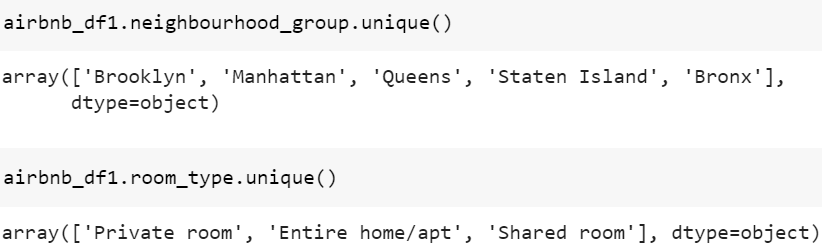
**airbnb\_df.isnull().sum() -** using this in Pandas isnull() is used to finding the missing or null value which are not useful for us.



In python, we can find out all columns, so that we can take only those columns which are useful for us otherwise dops the columns form table like this.

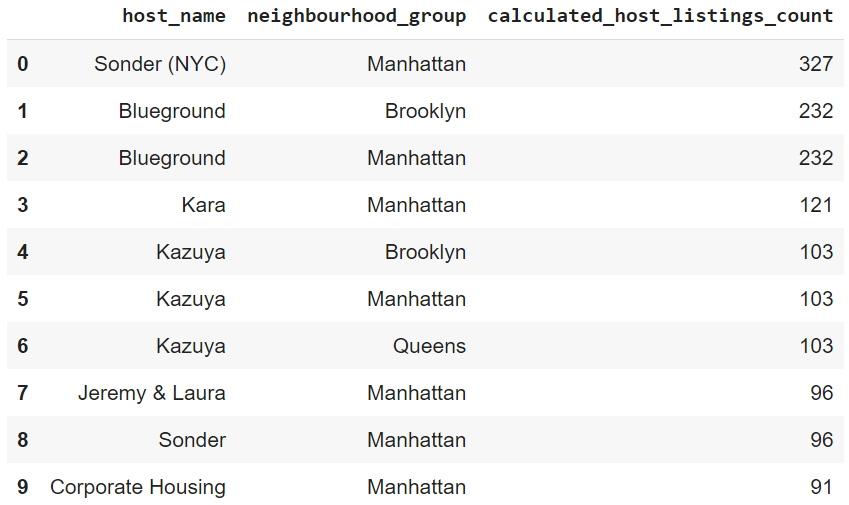


In our data many things are repeated many time, so we can find use Unique technique identifies the unique values of a Pandas Series like in neighbourhood\_group and room\_type.

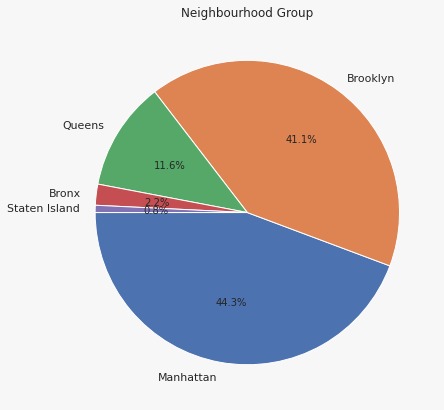


After EDA process, we do the question asked in our project. We use different of methods like group by, maximum, drop etc to solve the query.

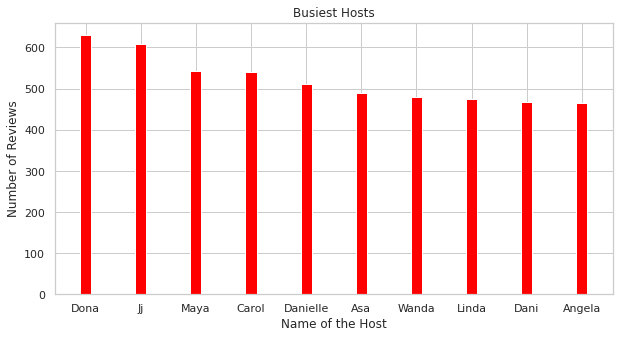
* In our first question, we find out top 10 host name using neighbourhood\_group and visualization graph below:



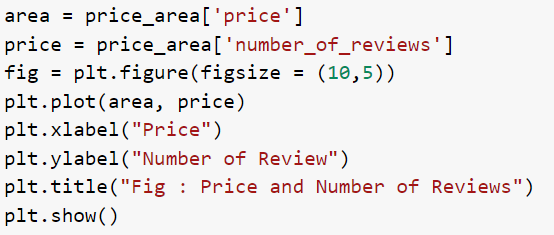
Now, graphical visualization as below:

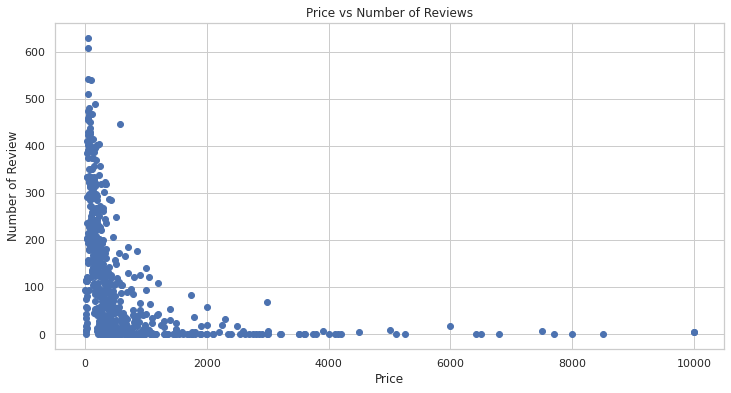


* Next, we find out the busiest Host using neighbourhood\_group, room\_type & numbers\_of\_reviews.

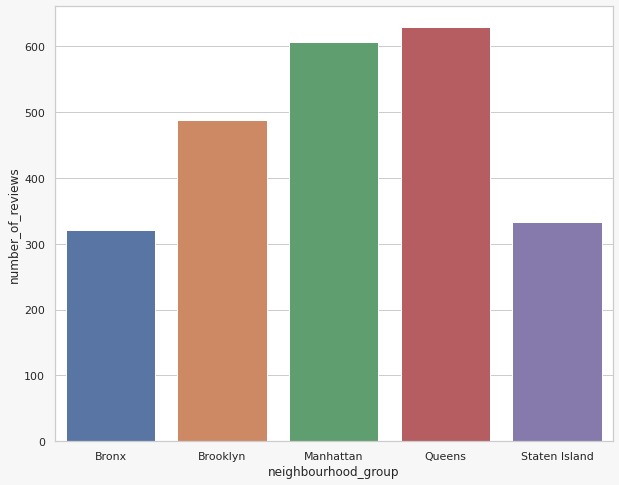


* After that for affordable room type, we have to check the prelationship of price and no of reviews.





* Now, we check out the availability of room in different neighbourhood\_group as shown in graph.



Using data visualisation the relationship between Area and Host listing, which makes data more understandable to all.

For different types of bar, pie chart etc. we use library as mention below:

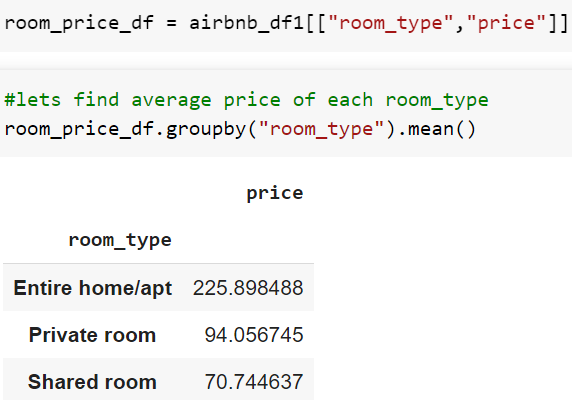
**%matplotlib inline**

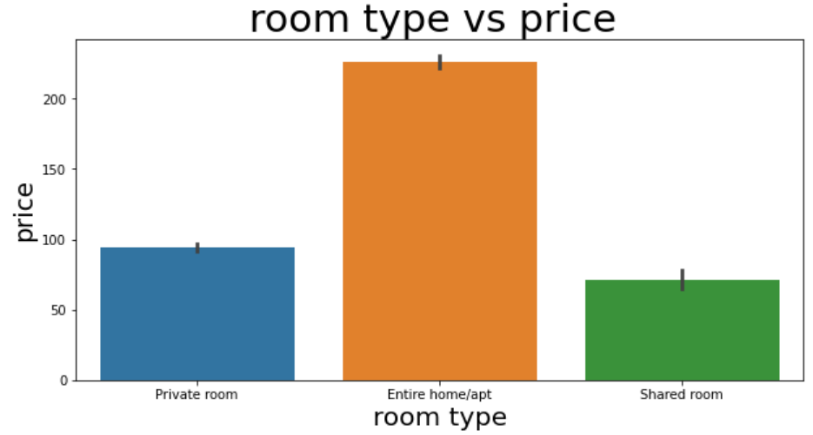
**import matplotlib.pyplot as plt**

**import seaborn as sns**

These help us to visualize the data in graphics.

Now, we are going to find out the price and room\_type relationships using coding and graph.





**Conclusion:**

Through this exploratory data analysis and

visualization project, we gained several interesting insights into the Airbnb rental market. Below we will summarise the answers to the questions that we wished to answer at the beginning of the project:

* Firstly we did EDA on data and extract the useful data and drop the extra data as some operation mentioned in slides.
* Extracted top 10 Host from data and makes the pie cart of extracted data.
* Make bar chart of 10 busiest Host as shown above.
* After that shows the relationship of price and review using line chart.
* Represented room availability in different neighbourhood group in bar graph.
* Last we find out the price in different room type.
* Finally, we see people prefer the rooms which comes under their budget.
* After analysing of data, we see more traffic in which having more booking.

**References-**

1. Guidance of Alma Better Instructors & Moderators
2. Google
3. Discussion with my team members