**NYC Airbnb Data Analysis**

**SARANYA.N**

**DATA SCIENCE TRAINEE**

**ALMABETTER , BANGALORE**

**ABSTRCT:**

Airbnb is an American company that facilitates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. It basically connecting travelers with local hosts who want to rent out their homes with people who are looking for accommodations in that locality. In the world of rising new technology and innovation, Travel industry is advancing with the role of Data Science and Analytics. Data analysis can help them to understand their business in a quiet different manner and helps to improve the quality of the service by identifying the weak areas of the business. This study demonstrates the how different analysis help out to make better business decisions and help analyze customer trends and satisfaction, which can lead to new and better products and services.

**PROBLEM STATEMENT**

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.

Explore and analyze the data to discover key understandings (not limited to these) such as :

* 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

**The dataset includes three main tables:**

#### Listings -Detailed listings data about hosts, Airbnb houses and price. The attributes used in the analysis are id (listing ID), name (name of the listing),  host ID,  host name(name of the host),  room type, and price

#### reviews - Detailed reviews given by the guests. Key attributes include  number of reviews, latest review, and  number of reviews per month

#### location - Provides details about Airbnb location in New York city. Four attributes including neighbor hood group , neighbour hood  (area), longitude, and latitude .

**INTRODUCTION**

The goal of this data exploration is to understand the significance related to the Airbnb dataset, we got to know different hosts having different kinds of properties for rent. Prices of rooms depending upon location and facilities provided and with a minimum number of nights of stay required, also check various aspects related to price, room type, availability 365, reviews per month etc. Understand the popularity of Airbnb based on locations, Analyze the reason about variation based on locations by prices (which area is expensive), reviews (which area is the best), and type of rooms ,Realize the situation of the busyness of the host .

**Exploratory Data Analysis**-

For Exploratory Data Analysis it is important to import libraries because these are necessary for data exploration.

**Import NumPy** -NumPy can be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.

**Import pandas**- Pandas is a Python library for data analysis. Pandas is built on top of two core Python libraries—matplotlib for data visualization and NumPy for mathematical operations.

**Import seaborn**- Seaborn is an open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis. Seaborn works easily with data frames and the Pandas library. The graphs created can also be customized easily.

**Import matplotlib.pyplot** - Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. %matplotlib inline -sets the backend of matplotlib to the 'inline' backend, with this backend, the output of plotting commands is displayed inline within frontends.

**Handling Null values**- Data Cleaning is the process of finding and correcting the inaccurate data that are present in the dataset. Our dataset contains a large number of null values which might tend to disturb our understanding hence we dropped them at the beginning to get better insights. Replacing null values in the column reviews\_per\_month with 0 in the dataset. Name and hostname are not the main aspects of our analysis, that's why replaced them with some characters. Dropping the column the last review having a large number of data points contains null values.

A **bar plot** is a plot that presents categorical data with rectangular bars with lengths proportional to the values that they represent. A bar plot shows comparisons among discrete categories. One axis of the plot shows the specific categories being compared, and the other axis represents a measured value.  Plotting neighbourhood\_groups and total listings gives insights on the data in the particular neighbourhood\_groups values or which is most preferred in the neighbourhood\_gropus.

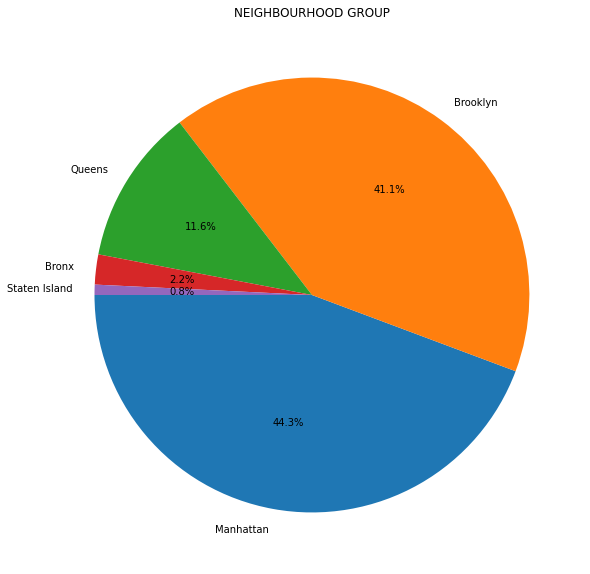
A **pie chart** is a type of graph that represents the data in the circular graph. The slices of pie show the relative size of the data, and it is a type of pictorial representation of data A pie chart requires a list of categorical variables and numerical variables. Here, the term “pie” represents the whole, and the “slices” represent the parts of the whole.

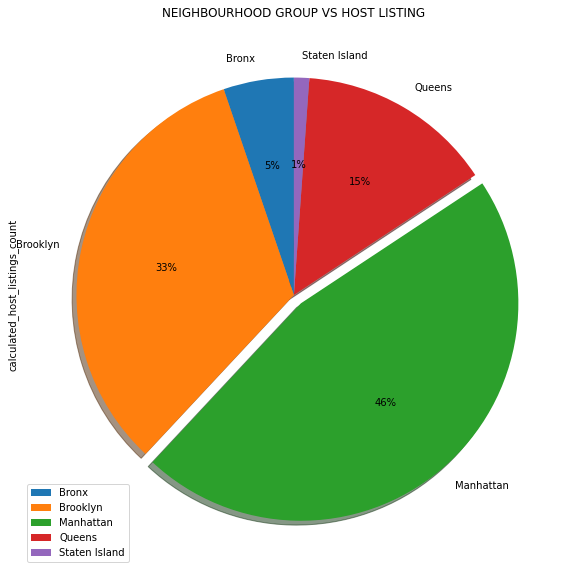
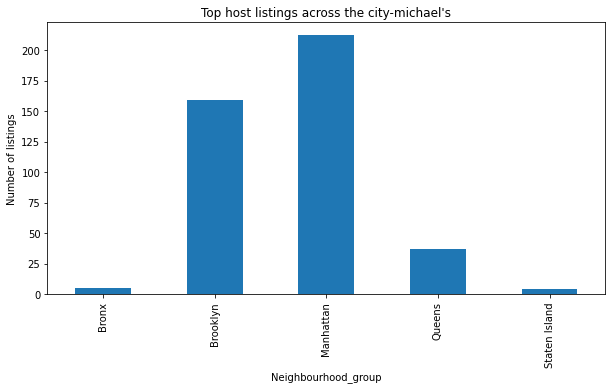
Here also included scattered chart for analyzing the datas.

Graphs are helps to understand datas more efficiently

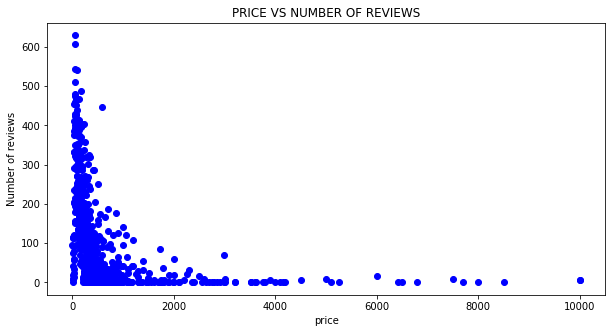
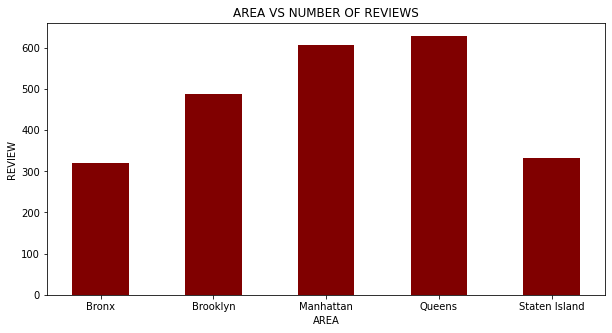
**Analysis**

First Based on the neighbourhood locations, the pie chart shows that 44.3% of Airbnb housings are located in Manhattan, and 41.12% of Airbnb housings are located in Brooklyn. Almost 90% of Airbnb housings are located near by the most famous city hotspots

Also plotting different graphs helps to lead relevant observations , while comparing host name and no of properties , michael’s is the top host soby plotting a bar graph with number of listings vs neighbourhood group we get most of the michaels properties are in manhattan , so we can understand that most of them are visiting near this neighbourhood.



**What can we learn from locations , prices, reviews, etc...**

Airbnb users rate their stay on the basis of location, price, cleanliness, living experience and a host of other parameters. Here I used the data to get the chart about variation of price and numbers of reviews by locations. It is really interesting because price and numbers of reviews didn’t have a large positive correlation. Highly expensive neighbourhoods will not tend to have more numbers of reviews because travelers will decide more cheaper Airbnb to live in and take the transportation to famous city hotspots**** 

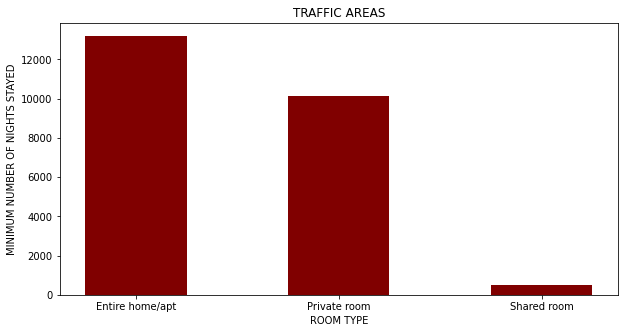
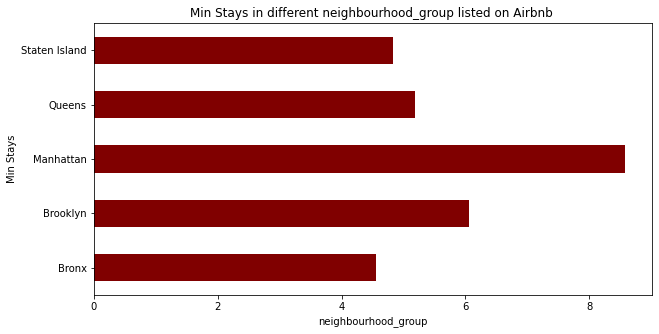
# **THE BUSIEST**

Based on the charts we have above. Let’s explore the busyness of the host that are there in NYC. In line with different charts about numbers of Airbnb by locations in NYC, variation of price by location, and variation of numbers of review by location, we understand Brooklyn and Manhattan provide the most houses and rooms. Therefore, we can imagine that top 10 busy host are from the area of Brooklyn and Manhattan.

Plotting neighbourhood\_groups and availability\_365 gives insights on the data in the particular neighbourhood\_groups values or which is most preferred in the neighbourhood\_gropus. staten island seems to be more available throughout the year it is available more than 300 days in a year.so which is less busiest .and also we get Brooklyn and manhattan are the most busiest.also while checking number of host vs number of review we also can find the busiest host ‘

**Different types of rooms in New York City**

Now let’s explore the types of rooms that are there in New York City. Below is a chart that shows distribution of types of rooms by different locations (Bronx, Brooklyn, Manhattan, Queens, and Staten Island) .According to the chart, it shows that Brooklyn and Manhattan provide the most houses and rooms, representing a high demand there. Also, because of the high housing prices, New York City in Airbnb gives users more choices for private rooms to reduce the burden on life. from this graph we can analys that most people prefer entire home/apt and then private room which are in manhattan ,queens ,Brooklyn from this observations people mostly living in manhattan on an average of more than 8 nights followed by guests who stayed in brooklyn where average stay is 6 nights . so we figure out that most people prefer to saty in manhattan

. **** 

Airbnb NYC dataset contains variables like price, room\_type, neighborhood\_group, neighborhood, host\_id, availability\_365 etc. The dataset values will help in the EDA visualization by giving an idea about price in a particular neighborhood, neighborhood groups, room types, 365 availability of rooms, and preferred room types in a particular area, criteria to select a room, reviews, minimum stay by customers etc. plots like scatter plots ,bar graphs , pie graphs will provide insights on the data dataset

# **Conclusion:**

# 1. The people who prefer to stay in Entire home or Apartment they are going to stay bit longer in that particular Neighbourhood only.

# 2. The people who prefer to stay in Private room they won't stay longer as compared to Home or Apartment..

# 3. Most visitors don't prefer shared rooms, they tend to visit private room or entire home.

# 4. Most people prefer to pay less price.

# 5. If there are more number of Reviews for particular Neighbourhood group that means that place is a tourist place.

# 6. If people are not staying more then one night means they are travellers.

# 7.Manhattan and Brooklyn are the two distinguished, expensive & posh areas of NYC

So Manhattan and Brooklyn are the most preferred neighborhoods.

**Reference:**

* [pandas - Python Data Analysis Library (pydata.org)](https://pandas.pydata.org/)
* [NumPy](https://numpy.org/)
* [Matplotlib — Visualization with Python](https://matplotlib.org/)
* [seaborn: statistical data visualization — seaborn 0.11.2 documentation (pydata.org)](https://seaborn.pydata.org/)
* [Airbnb - Wikipedia](https://en.wikipedia.org/wiki/Airbnb?msclkid=289e1ea6a81111ec8b49fbebf9ed1197)