**AIRBNB BOOKINGS ANALYSIS - EDA**

**CAPSTONE PROJECT**

**Pretesh Agarwal**

**Data science trainees,**

**AlmaBetter**

**Abstract:**

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 document aims to better understand what factors are considered when an individual chooses to book an Airbnb and what features contribute most to their experience. With this understanding, it is the goal to be able to effectively market Airbnb to potential customers to increase the number of bookings and generate more revenue. We focus our analysis on Airbnb listings in New York City through data sets that detail a variety of topics, including listingspecific attributes, host qualities, and user experience. We utilize a variety of data analysis techniques to identify the most important factors that affect a listing’s booking rate throughout the year, as well as, take a deep dive to determine which specific attributes affect a review score rating and how the price of an Airbnb, regardless of the borough, compares to hotels within the area.

***Keywords: Python, Pandas, Matplotlib***

**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.

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

**2. Introduction**

**1) What is Airbnb?**

Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Based in San Francisco, California, the platform is accessible via website and mobile app. Airbnb does not own any of the listed properties; instead, it profits by receiving commission from each booking. The company was founded in 2008 by Brian Chesky, Nathan Blecharczykand Joe Gebbia. Airbnb is a shortened version of its original name, AirBedandBreakfast.com.

The idea behind Airbnb is simple: Find a way for local people to make some extra money renting out their spare home or room to people visiting the area. Hosts using this platform get to advertise their rentals to millions of people worldwide, with the reassurance that a big company will handle payments and offer support when needed. And for guests, Airbnb can offer a homey place to stay that has more character, perhaps even with a kitchen to avoid dining out, often at a lower price than what hotels charge.

## 2) How Does Airbnb Make Money**?**

Airbnb’s [business model](https://www.investopedia.com/terms/b/businessmodel.asp) is quite profitable. The company, like Uber, Lyft, and others, has capitalized on the [sharing economy](https://www.investopedia.com/terms/s/sharing-economy.asp), essentially [making money](https://www.investopedia.com/articles/investing/112414/how-airbnb-makes-money.asp) renting out property that it doesn’t own.

Every time a reservation is made, Airbnb takes a cut. When you click on a property, you'll find to the right of the page a breakdown of the fees you'll be charged if you go ahead and book. One of these fees is a service fee, which covers the cost of running the platform and providing support; this basically makes up the bulk of Airbnb's [revenue](https://www.investopedia.com/terms/r/revenue.asp).

1. **Key Objectives**

* 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?

1. **Data Preparation & Data Wrangling**

● Dropping unnecessary data:

as "last\_review" and "reviews\_per\_month" have more than 10,000 null values, it affects the outcomes of Data analysis; So, we are removing these columns and also as we are not doing any analysis specifically on latitude and longitude, we're also removing these variables as well.

● Verifying Data quality:

We have gone through whole data and checked null values and reviewed any missing data or wrong data. And prepared whole data ready for exploratory data analysis.

● Basic data exploration:

Using describe() and info () and size functions of pandas. Gone through a basic exploration of data before entering into EDA.

1. **Data Description**

● **id:** unique reference number for each different hotel.

● **name:** name of different hotels of various neighborhood groups.

● **host\_id:** unique reference id of each individual host.

● **host\_name:** name of host hosting different hotels.

● **neighbourhood\_group:** aggregate group of neighborhood cities of some particular regions.

● **neighbourhood:** cities present in NYC.

● **latitude:** latitude is a geographic coordinate that specifies the north–south position of a point on the Earth's surface. Latitude is an angle which ranges from 0° at the Equator to 90° at the poles.

● **longitude:** Longitude is a geographic coordinate that specifies the east–west position of a point on the Earth's surface, or the surface of a celestial body.

● **room\_type:** Different room types available for booking, which contains Private room, Entire home/apt, Shared room.

● **price:** price per each night stay of different room types at various hotels.

● **minimum\_nights:** minimum nights booked in particular hotel.

● **number\_of\_reviews:** count of reviews got for each hotel.

● **last\_review:** date of last review got by a customer to a particular hotel.

● **reviews\_per\_month:** count of reviews getting per month of a particular hotel.

● **calculated\_host\_listings\_count:** It represents total number of listings made by a specific host. In some cases, the properties are same but some of the other features differ like(room\_type).

● **availability\_365:** number of available days for booking in a year

1. **Exploratory Data**

**Analysis**

### Understanding, Wrangling and Cleaning Data.

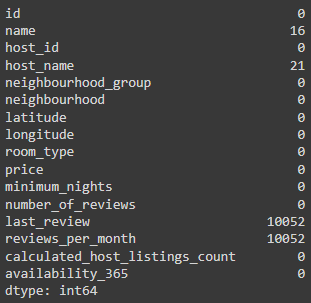
##### Presenting the code and methods for acquiring the data. Loading the data into appropriate format for analysis.

After looking at the head of the data-set we already were able to notice some NaN values, therefore need to examine missing values further before continuing with analysis.

Looking to find out first what columns have null values.

Using ‘sum’ function will show us how many nulls are found in each column in dataset.

##### airbnb.isnull().sum()



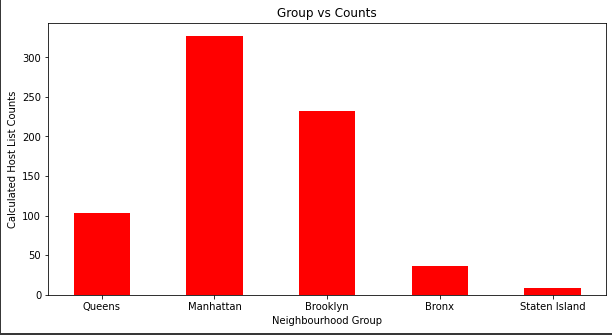
In our case, missing data that is observed does not need too much special treatment. Looking into the nature of our dataset we can state further things: columns "name" and "host\_name" are irrelevant and insignificant to our data analysis, columns "last\_review" and "review\_per\_month" need very simple handling. To elaborate, "last\_review" is date; if there were no reviews for the listing - date simply will not exist. In our case, this column is irrelevant and insignificant therefore appending those values is not needed. For "review\_per\_month" column we can simply append it with 0.0 for missing values; we can see that in "number\_of\_review" that column will have a 0, therefore following this logic with 0 total reviews there will be 0.0 rate of reviews per month. Therefore, let's proceed with removing columns that are not important and handling of missing data.

### Exploring and Visualizing Data

1. **Top Hosts and their listings count:**

For analyzing this we used group by function and took 'host\_name', 'neighbourhood\_group', and "calculated\_host\_listings\_count" and calculated the top 10 hosts.

* **Key findings**
* Manhattan neighbourhood group hosts are out-performing in listings.
* 7 out of top 10 hosts are from the ''Manhattan'' neighbourhood group.
* 2 out of top 10 hosts are from the ''Brooklyn" neighbourhood group.
* 1 out of top 10 hosts are from the ''Queens" neighbourhood group.



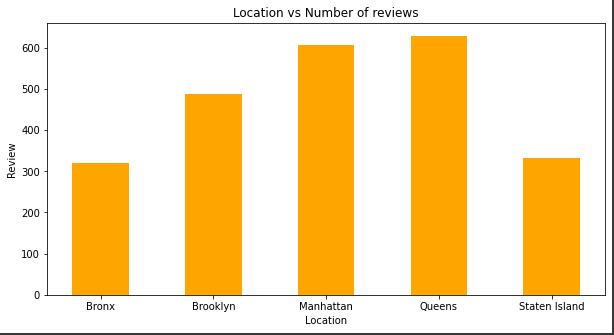
1. **What can we learn from predictions?**

**(ex: locations, prices, reviews, etc)**

For this question, we approached with 3 different corelations:

* **“neighbourhood\_group” and “number\_of\_reviews”.**

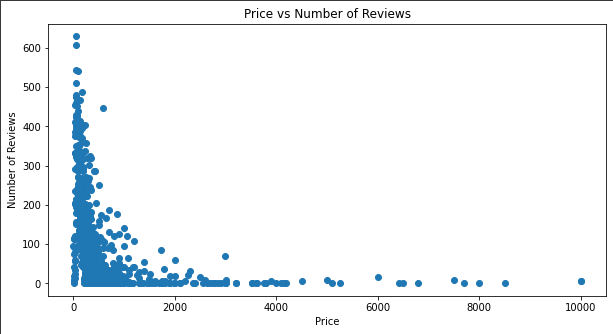
For analyzing this we used group by function and took “neighbourhood\_group” and “number\_of\_reviews” and got the following result.



After this analysis we can say that people who lived in Queens, Manhattan and Brooklyn have given most number of reviews.

* **“Price” and “number\_of\_reviews”.**

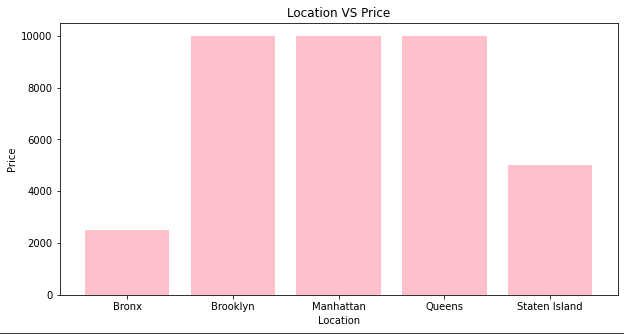
For analyzing this we used group by function and took “price” and “number\_of\_reviews” and got the following result.



After this analysis we can say that most people prefer to stay in place where price is less.

* **“neighbourhood\_group” and “Price”**.

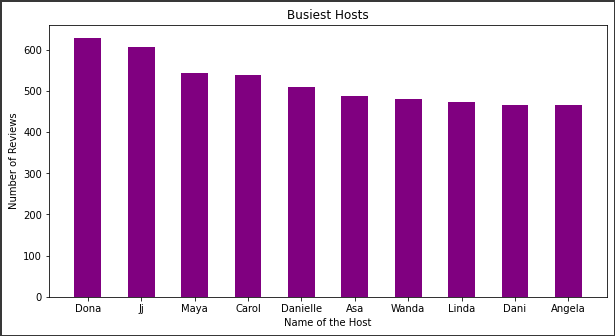
For analyzing this we used group by function and took “neighbourhood\_group” and “price” and got the following result.



After this analysis we can say that maximum prices for Brooklyn, Manhattan and Queens are same 10000.

1. **Finding Busiest Host:**

Next, we were interested in finding the busiest hosts by considering “number\_of\_reviews”. We took, “host\_name”, “host\_id”, “room\_type”, “number\_of\_reviews”, and got the following result.



* **Key findings:**

Top 5 busiest hosts are:

- Dona

- Ji

- Maya

- Carol

- Danielle

Because these hosts listed room

type as Entire home and Private

room which is preferred by most

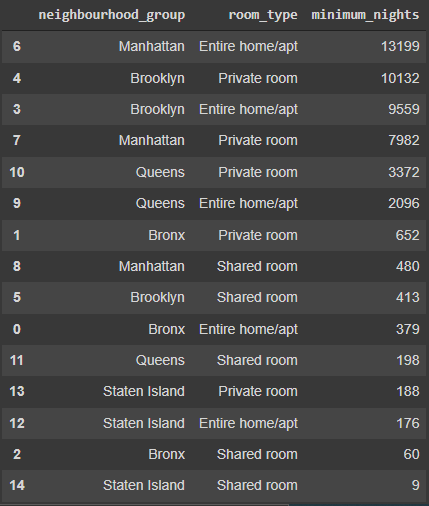
number of people.

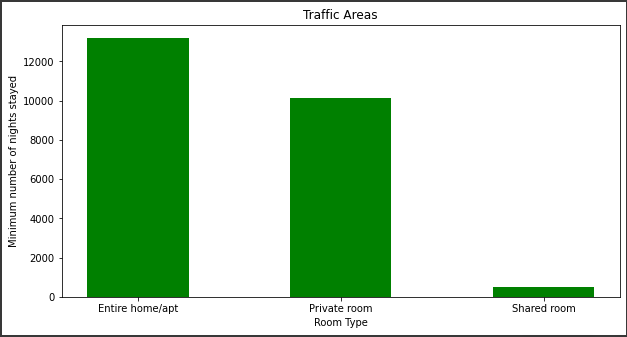
1. **Is there any noticeable difference of**

**traffic among different areas and what could be the reason for it?**

Our main motive through this step was to find traffic among different types of rooms at different neighborhood groups.

For this we took, 'neighbourhood\_group', 'room\_type', 'minimum\_nights' to analyze this question.





* **Key findings:**
* In Manhattan, people are preferring "Entire Home/apt".
* But, in Brooklyn, Queens and Bronx people are preferring private rooms.
* In Staten Island, people are having equal preference over all three types of rooms.

1. **Conclusion**

Airbnb dataset-2019 appeared to be a very rich data-set with a variety of columns that allowed us to do deep data exploration on each significant column presented.

* First, we have found hosts that take good advantage of the Airbnb platform and provide the most listings; we found that our top host has 327 listings. After that, we proceeded with analyzing, boroughs and neighbourhood listing densities and what areas were more popular than another.
* From the entire analysis on Airbnb bookings analysis, our assumptions before analysis went totally different after getting results from the analysis. The whole EDA process gave very fascinating results and insights that will be helpful for business development and expansion, budget allocations and focusing on things people prefer.

1. **References**
2. GeeksforGeeks
3. Kaggle
4. Wikipedia