**Airbnb booking analysis**

**Praful M. Gedam,**

**Data Science Enthusiast**

**AlmaBetter, Bangalore**

**Abstract:**

Abstract overview of the notebook 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 analysed and used for security, business decisions, understanding of customers' and providers' (hosts) behaviour and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

1. **Problem Statement**.

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

1. **Introduction**

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. On the other hand, this platform enables host to list their available space and earn extra income in the form of rent and it also enables travelers to book unique homestays from local hosts, saving them money and giving them a chance to interact with locals.

### .

1. **Airbnb Booking Dataset Insight**

This dataset has around 49,000 observations in it with 16 columns and it is a mix of categorical and numeric values It contains different hosts, the neighborhood group the properties are located in and the type of property customers most wish for. Exploring them will definitely help us have a very good understanding of the booking trends.

**Column Information**

* name = Description about the listings.
* host\_id = unique id for each listed host.
* host\_name = Hostname for the listings.
* neighbourhood\_group = Location
* neighborhood = Area
* latitude = Latitude coordinates
* longitude = Longitude coordinates
* room\_type = Listing space types
* price = Price in dollars
* minimum\_nights = minimum nights required to stay
* number\_of\_reviews = No. of reviews written for the listing
* last\_review = Last reviewed date for the listing
* reviews\_per\_month = Total review per month for the listings
* calculated\_host\_listings\_count = Total no of listing against the host id
* availability\_365 = Available days of a listing in a year.

1. **Steps involved**

* **Setting up the notebook**

# The notebook is set up in Google Collaboratory platform. The Google drive containing the dataset is mounted in the notebook and it is loaded as a pandas dataframe. The necessary libraries such as NumPy, pandas (for working on the dataframe), seaborn and matplotlib (for visualization) are imported

* **Data Exploration**

During the preparation of the dataset for EDA, a basic statistical summarization was done. The information regarding the data types of each column was explored.

This was followed by a detailed exploratory data analysis. The distribution of bookings with respect to different categories of features was visualized. Relationship between variables was observed.

## **Null value handling**

The dataset contains a good number of null values. The columns name, host name & reviews contained large number of null values. So, we tried to preserve as many rows as possible by replacing null values with suitable values.

1. **Exploratory Data Analysis**

Throughout the analysis, we tried to answer questions that help us understand the factors determining the booking trends.

1. **What is the price of each room type?**

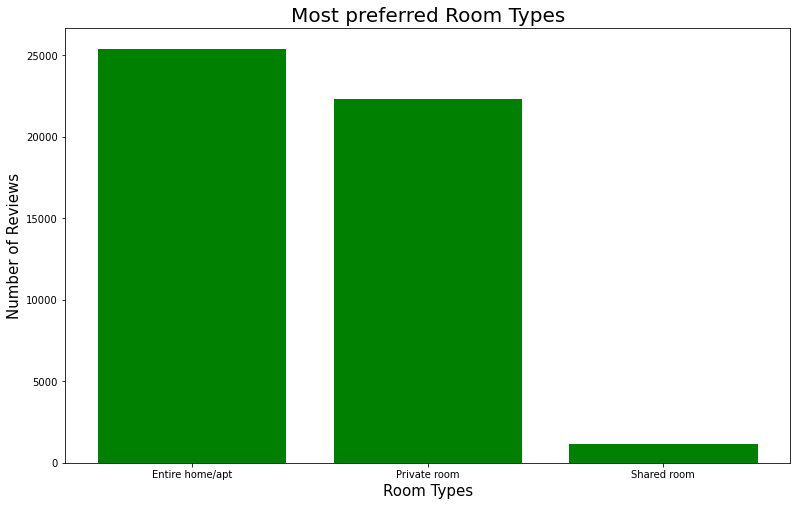
The mean price of the entire room/apartment type was found to be $211; the mean price of a private room was $89 followed by a shared room which was $72.

1. **Which was the most wished/booked room type?**

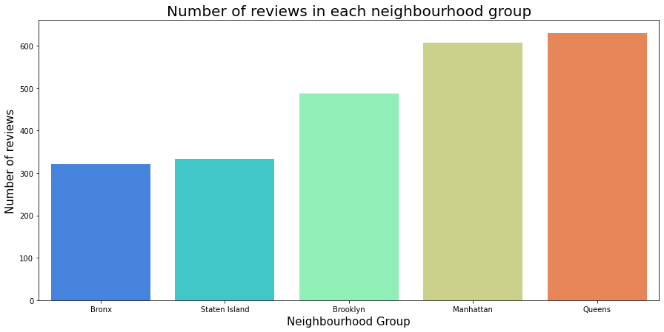
Entire home/apt is the most wished/booked, followed by private room and shared room

So, we can say that, when people go on vacation with their loved ones they prefer staying in an Entire home/Apartment followed by a private room, who wish for some privacy and people on a budget prefer shared rooms.

Fig1.most preferred room type



1. **Price analysis of each room type in all neighborhood groups.**

All the groups have Entire home/apt as the most expensive commodity. Manhattan and Brooklyn agree with the general price trend completely, while for Bronx and Queens, the price for shared rooms and private rooms are almost equal and we see a trend reversal in the price of shared rooms and private rooms in Staten Island.

We can also see that Manhattan is the most expensive for all room types by a considerable margin.

1. **number of hosts in each neighborhood groups?**

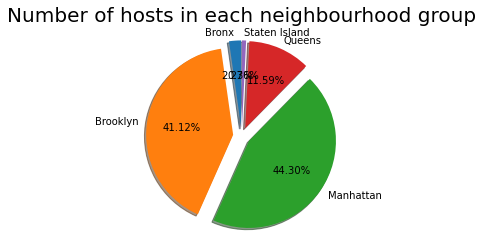
****Manhattan has 21661 host which is the highest of all, followed by Brooklyn which has 20104 listings, Queens having 5666, Bronx having 1091 and at last Staten Island having 373.

Fig2.host in each neighborhood groups

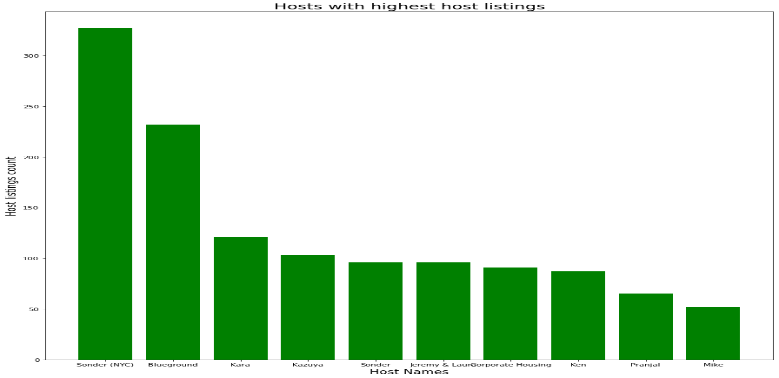
* **Latitude & longitude of Airbnb**

The information we got from the graph above is red color dots are the rooms in higher price. Also, we can see that the Manhattan region has the more expensive type.



**Highest host listing**

Sonder (NYC) is the most successful host andhas about 327 listings which is 30.61% of the total listings and these are present only in Manhattan with just private rooms and apartments. They deal with the most expensive aspect of the business, hence possess more money, which in turn increases their number of listings.



1. Which is the most expensive neighborhood in each neighborhood groups? 80% of the hosts are in Manhattan and Brooklyn
2. Sonder (NYC) and blue ground has the greatest number of host listings
3. Most of the people likes Queens and Manhattan
4. Even with moderate number of hosts in Queens, it generated the highest revenue out of all the neighborhood group
5. Most of the guests prefer cheaper price
6. Dona and Ji are the busiest hosts
7. About 90% of the guests prefer entire home/apt or private room types.

* The most expensive neighborhood in Brooklyn is Sea Gate.
* The most expensive neighborhood in Manhattan is Tribeca.
* The most expensive neighborhood in Queens is Neponset.
* The most expensive neighborhood in Staten Island is Fort Wadsworth.
* The most expensive neighborhood in Bronx is Riverdale.

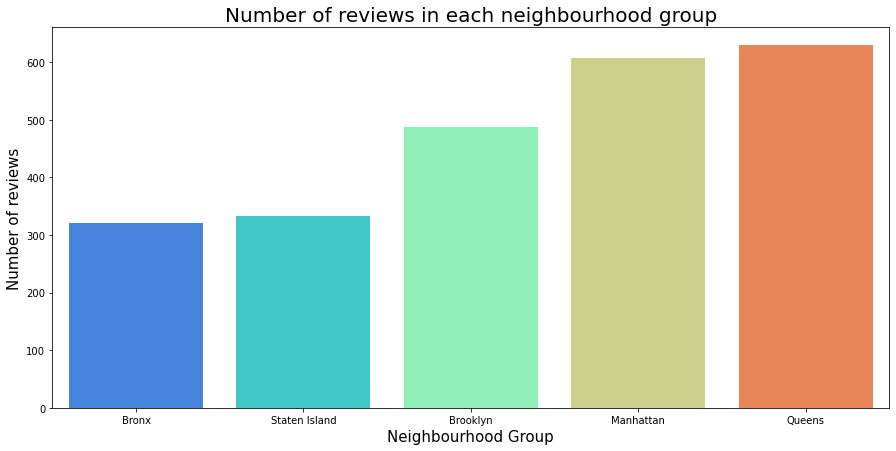


Fig. neighborhood group

**Conclusion**

Finally able to answer some really important questions about the bookings analysis using this dataset. By studying the trends of the past bookings, we can take steps to satisfy the customer demands and understand where to focus the efforts the most to make the most positive results.

* 80% of the hosts are in Manhattan and Brooklyn
* Sonder (NYC) and blue ground has the greatest number of host listings
* Most of the people likes Queens and Manhattan
* Even with moderate number of hosts in Queens, it generated the highest revenue out of all the neighborhood group
* Most of the guests prefer cheaper price
* Dona and Ji are the busiest hosts
* About 90% of the guests prefer entire home/apt or private room types.

**References**

1. [*https://towardsdatascience.com/*](https://towardsdatascience.com/)
2. [*https://stackoverflow.com/*](https://stackoverflow.com/)
3. *https://medium.com/*