**Airbnb Bookings Analysis**

**Kanika Bhardwaj**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

This San-Francisco based startup offers you someone’s home as a place to stay instead of a hotel. You might be thinking of another unicorn in town as to OYO Hotels which has kind of a relatable business model but **Airbnb** allows you to be host for anyone anywhere with rooms/beds available in your personal space. OYO Rooms and Airbnb are by no means similar to each other, in fact, they are almost as opposite as the sky and sea. So, having much said let’s just deep dive into our actuals on why are we basically here?

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

**2. Introduction**

A host on Airbnb holds multiple properties in a neighborhood group(boroughs of NYC) with different host-ids but a host with a particular property/listing in a particular neighbourhood of a neighbourhood group holds a same host-id(also not mandatory as there are exceptions where few hosts have different id’s for each listing/property in a neighbourhood). Also the data tells, there might be cases where a particular host has co-hosted someone else’s property/listing in a neighbourhood on Airbnb

Data analytics is simply the analysis of various data, i.e., cleaning the data, converting it into an understandable form, and then modeling the data to extract useful information for business or organizational purposes. It is mainly used for business decision making. Many libraries are available for doing the analysis. For example, NumPy, Pandas, Seaborn, Matplotlib, pyplot, etc.

* NumPy: NumPy is a library written in Python, used for numerical analysis in Python. It stores the data in the form of nd-arrays (n-dimensional arrays).
* Pandas: Pandas is mainly used for converting data into tabular form and hence, makes the data more structured and easily to read.
* Matplotlib: Matplotlib is a data visualisation and graphical plotting package for Python and its numerical extension NumPy that runs on all platforms.
* Seaborn: Seaborn is a Python data visualisation package based on matplotlib that is tightly connected with pandas data structures. The core component of Seaborn is visualisation, which aids in data exploration and comprehension.
* Pyplot: matplotlib. Pyplot is a collection of command style functions that make Matplotlib work like MATLAB. Each Pyplot function makes some change to a figure.

Data visualization will help the data analysis to make it more understandable and interactive by plotting or displaying the data in pictorial form. Pandas, a Python open-source package that deals with three different data structures: series, data frames, and panels, solves that need of analyzing and visualization of data.

Data analysis with Python makes the task easier because the Python programming language has many advantages over other programming languages. It is characterized by being a high-level programming language (the codes are in human-readable form) and can be easily understood and used by any programmer or user. Many libraries and functions for statistical and numerical analysis are available in Python.

This paper contains all the basic terms and functions that a beginner absolutely needs to know what data analysis are.

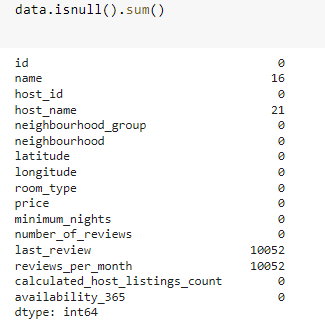
1. DATA ANALYSIS

### **Data requirements:** Data are the most important unit in any study. Data must be provided as inputs to the analysis based on the analysis’ requirements. The term “experimental unit” refers to the type of organization that would be used to gather data (e.g., a person or population of people). It is possible to identify and obtain specific population variables (such as height, weight, age, and salary). It doesn’t matter whether the data is numerical or categorical.

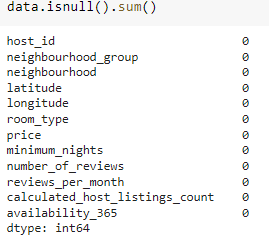
### **Data Collecting:** The collecting of data is simply known as Data Collecting. Data is gathered from a variety of sources, including relational databases, cloud databases, and other sources, depending on the study’ needs. Field sensors, such as traffic cameras, satellites, monitoring systems, and so on, can also be used as data sources.

### **Data processing:** Data that are collected must be processed or organized for analysis. For instance, these may involve arranging data into rows and columns in a table format (known as structured data) for further analysis, often through the use of spreadsheet or statistical software like SQL.

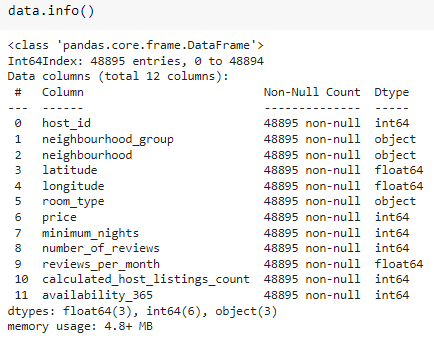
### **Data cleaning:** The method of cleaning data after it has been processed and organized is known as data cleaning. It scans for data inconsistencies, duplicates, and errors, and then removes them. The data cleaning process includes tasks such as record matching, identifying data inaccuracy, data sort, outlier data identification, textual data spell checker, and data quality maintenance. As a consequence, it keeps us from having unexpected outcomes and assists us in delivering high-quality data, which is essential for a successful outcome.



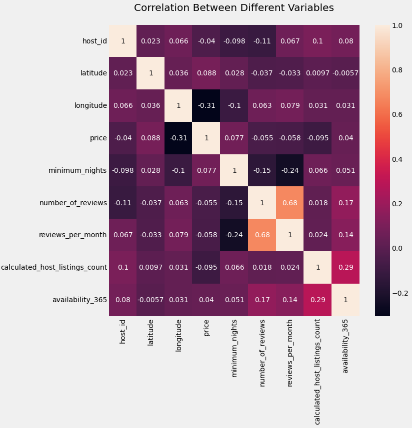
Here we are having some missing values, so we have to remove these missing values. After removing missing values we get



### **Exploratory data analysis:** Once the datasets are cleaned and free of error, it can then be analyzed. A variety of techniques can be applied such as exploratory data analysis- understanding the messages contained within the obtained data and descriptive statistics finding average, median, etc. Data visualization is also a technique used, in which the data is represented in a graphical format in order to obtain additional insights, regarding the information within the data.



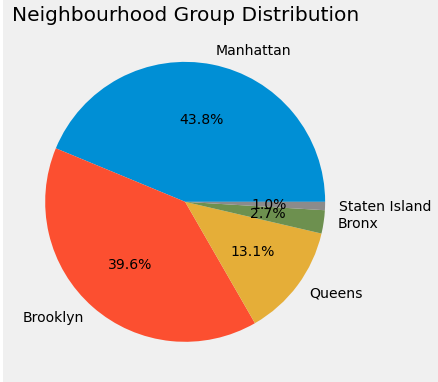
### **Modeling and algorithms**: Mathematical formulas or models (known as algorithms), may be applied to the data in order to identify relationships among the variables; for example, using correlation or causation.

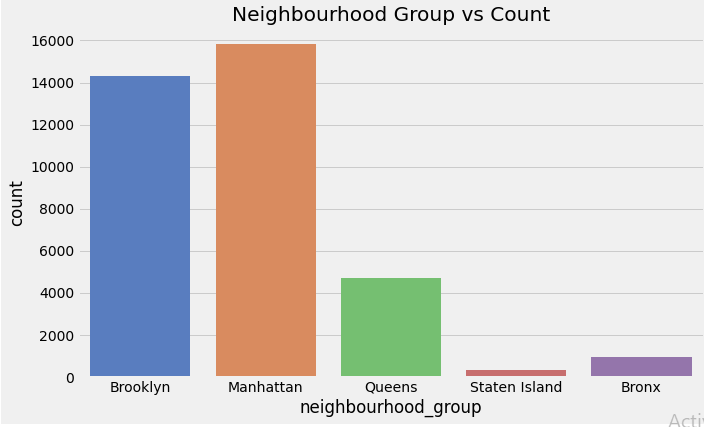


### **Data product:** A data product is a computer application that takes data inputs and generates outputs, feeding them back into the environment. It may be based on a model or algorithm.

# **3. Univariate Analysis on Airbnb data:**

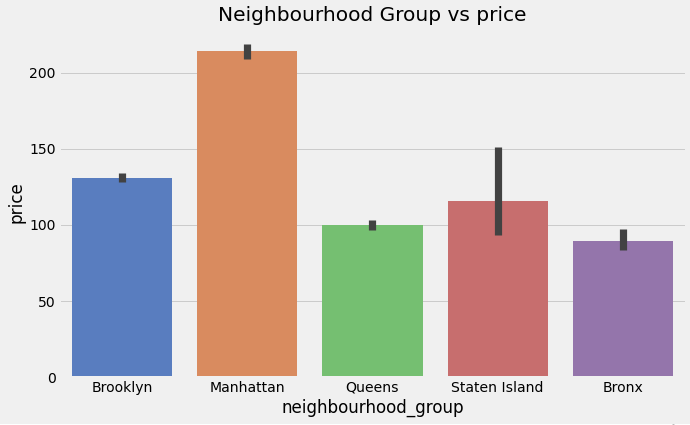
Univariate analysis is **the simplest form of analyzing data**. Uni means one, so in other words the data has only one variable. Univariate data requires analyzing each variable separately. Data is gathered for the purpose of answering a question, or more specifically, a research question.



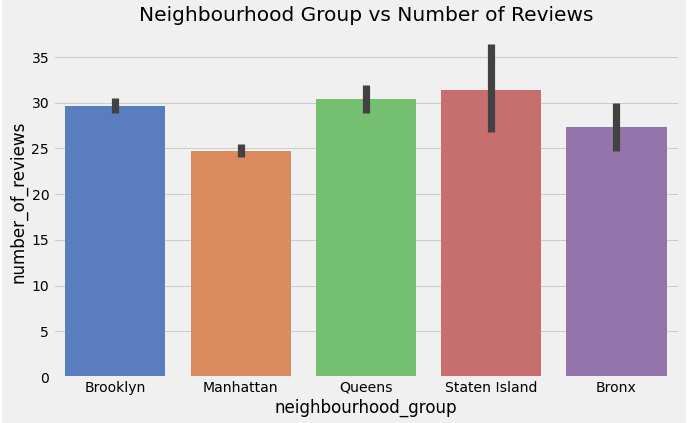


# **4. Bi-variate analysis on Airbnb**

Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables, for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association.



# **Reviews Prediction**



**5. Conclusion:**

Through this exploratory data analysis and visualization, we gained several interesting insights into the Airbnb rental market. This Airbnb dataset for 2019 year appeared to be a very rich dataset with a variety of columns that allowed us to do deep data exploration on each significant column presented. After that, we proceeded with analyzing boroughs and neighborhood listing densities and what areas were more popular than another, their price variations, their availability as per room types. Also we emphasized on key findings like room types and their preferred stays by guests, the top reviewed hosts and their listings.

Next, we put good use of latitude and longitude columns to create a geographical heatmap color-coded by the price of listings.

I have used [Seaborn](https://seaborn.pydata.org/tutorial.html) and [Matplotlib](https://matplotlib.org/tutorials/introductory/pyplot.html#sphx-glr-tutorials-introductory-pyplot-py) for creating all the visualizations. This is just a glimpse of eda on the Airbnb dataset and there’s no any prediction involved.

**References-**

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya