**Abstract:**

**Airbnb is one kind of service that connects the guest and host to share their property. Basically, property owner wants to utilize their property in the right direction so that for host it can be an income source and for guest, it can be a destination for a stay. The problem for Airbnb is there are tons of data generated through the hosts and guests as well. So, to find the right direction according to marketing and find the correct business-driven solution we did the data analysis on 49,000 observations. Through which we identify many solutions related to pricing, location, and room types that we mentioned in this project.**

**Introduction:**

**What is Exploratory Data Analysis?**

To simplify, exploratory data analysis (EDA), also known as Data Exploration is a step in the Data Analysis process, where several techniques are used to better understand the dataset being used.

**Some of the techniques are:**

* Extracting important variables and leaving behind useless variables
* Identifying outliers, missing values, and human error
* Understanding the data, maximizing our insight on a dataset and minimizing potential error that may occur later in the process

By conducting EDA, we can turn an almost useable or unusable dataset into a useable dataset.

**Main components of Exploratory Data Analysis:**

1. Acquire and loading data
2. Cleaning dataset
3. Exploring and Visualizing Data

**1. Acquire and loading data**

For this project, we are using google collab notebook.

To get the data, we are using Airbnb data of new york that was shared by almabetter. Before we are able to load the data into our collab notebook, first we need to import various external libraries/modules that needed for visualization and analysis.

**A. Load python libraries**

* **Pandas** and **NumPy** library used for data analysis
* **Matplotlib** and **Seaborn** library used for data visualization

import pandas as pd  
import numpy as np  
import matplotlib. pyplot as plt  
%matplotlib inline  
import seaborn as sns

**B. Load dataset**

* To load the dataset, we use panda’s library and function to read the CSV file of New York city Airbnb 2019 dataset, convert it to the Data Frame and check the top 5 index data.

**C. Understanding data**

* After we load the dataset, we need to understand the dataset by using various techniques. First, we need to look for information on how big is our dataset. By using **shape** attributes, we get to know our data size from a number of rows which consist of listing index, and the number of columns with the content of every feature related to the index.
* Then we check all the data type of every column if it already matches our requirement. For instance, we need a numerical data type (integer and float) on the longitude and latitude, for listing names we need to make sure the data is using string/object data type.

**We see that Airbnb have 3 room type. Based on the information on the Airbnb website, the definition of each room type is:**

* **Private room**

Guests have exclusive access to the bedroom/sleeping area of the listing. Other parts area such as the living room, kitchen, and bathroom are likely open either to the host even to other guests.

* **Entire home/apt**

Guests have the whole place for themselves. It usually includes a bedroom, bathroom, and kitchen.

* **Shared Room**

Guest sleep in a bedroom or a common area that could be shared with others.

# 2. Cleaning dataset

The next step is cleaning up the data, oftentimes the data we load have various faults, such as typo, missing value, incomplete data, etc. By doing cleaning up, the data quality will have better quality to be used for further analysis.

* **Checking column with missing values**
* **Replacing all the missing values**

# 3. Exploring and visualizing data

After we clean up the data, the next step is exploring the data by visualizing and analysing the values of the features, explaining the process and the results.

For our case, we will look up a various listing category consisting of each biggest value, visualize the listing distribution using a map, looking for selling value from their listing name, and finding the average price of the most popular listing.

1. **Answering and visualizing following questions in collab notebook and ppt.**

* Number of active hosts per location.
* 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?
* Finding Relation between neighborhood group and availability of rooms.
* Find top 10 hosts with most listings.
* Find total no. of nights spend per location.
* Total no. of nights spends per room types.
* Top 10 highest listing neighborhood.
* What is the average preferred price by customers according to the location?
* Find the total count of each room type.
* Which is the most preferred place by hosts to do their business.

**Conclusion**

**During this EDA project, the following concepts were learnt:**

* Basic inspection of the raw data.
* Handing the duplicate, error and NaN values present in the dataset, i.e., cleaning the data.
* Using different Python functions and libraries to clean and manipulate data.
* Data wrangling to come up with different insights on the data.
* Designing multiple visualizations to summarize the information in the dataset and successfully communicate the results and trends to the reader.
* Different Python libraries used to complete this EDA:
* Pandas
* NumPy
* Matplotlib. Pyplot
* Seaborn

**Reference:**

* [W3Schools Online Web Tutorials](https://www.w3schools.com/)
* [GeeksforGeeks | A computer science portal for geeks](https://www.geeksforgeeks.org/)
* [YouTube](https://www.youtube.com/)
* [GitHub: Where the world builds software · GitHub](https://github.com/)