Airbnb new user booking prediction

1. Introduction :

Airbnb is a lodging company that connects host and guest depend on lodging type, date, location, and price through the website and mobile app. The host provides hospitability to the guest in order to gain a higher rate and advertise their rental. For this company to develop, first it is important to provide a great impression with their first experience. Airbnb wants to predict where new user likely to book their first travel experience. Depend on their age, gender, and personal preference, each customer will have a difference in choosing their first destination. Through accurate data analysis, the company will be able to provide a shorter time to find a destination and better-personalized contents for the first users

2. Objectives :

Through Kaggle competition, Airbnb has provided all the data that is necessary to predict the destination country. data is given in 6 different files. There is two background information about destination countries and the ages and gender of the new users. Three files are train\_users, test\_users, and sessions that are provided in a CSV file. The train user is a dataset that contains Airbnb users includes their destination. The next dataset is tested users has the same format as train users.csv without a destination. This is the file that Airbnb wants to have a prediction on. The session file is supplementary data that contains activities such as the weather they click to see the lodge or add to a personal list.

3. Outlines

The country destination is the most important column that has to be considered. About 58% of the user have selected NDT which means there was no booking. Since the priority of this data, an analyst is to identify where is the first destination

The data should be separated between NDF and the rest of destinations. Depend on the age, gender and devices used and all different column may have a bias in choosing destinations. Each of the columns will be diagnosed using multiple linear regression models to see the relationship with a destination. Among those variables, the variables that have the least effect on the destination will be omitted. Through omitting variables, the most related variable will be considered most in prediction.

1. Deliveries :

For each variable that has a strong effect on the destination will be considered as a bar graph or scatter plot to reveal the relationship with a destination. For example, if there is gender bias in choosing destinations, the bar graph will be drawn with x variable gender and y variable destination with a different color for each country. Using the multiple linear regression model, and its scatter plot, the destination will be visualized to identify which variable should be considered for a new user to select their destinations