**Capstone Project 1: Exploratory Data Analysis**

Among the variables, gender and age is most significant in determining the destination. In most of the countries, females are more likely to use Airbnb while in other and Canada and Germany has a more male or equal rate of gender. For age, in us, there is a constant rate of people in each age range. While in other countries like Canada, Australia and Portugal have no age range above 70.

Some other variables that have a minor difference within variables are first device type, language, affiliate provider and second elapsed.

Some other variables that have minor differences are first device type, language, affiliate provider and second elapsed. For the first device type, In the US, iPhone is more used than iPad while Canada has an equal rate of usage and in France, people tend to use more iPhone than iPad. Language is also one of the significant variables. The customers who speak French have a higher rate to choose the destination in France than other language speakers. For affiliate provider, in the US, the customer tends to use Facebook than bing while bing is chosen over Facebook in Europe.

The last variable that matter with the destination is second elapsed. Some customer took more than 1750000s while in Australia, Canada, and other countries took less than 250000s.

In order to find the correlation between variables, the categorical variables were converted to dummy variables. The correlation color table is plotted for each of the variables. However, with the graph, it is hard to tell the exact correlation between variables. Therefore, the correlation is calculated using Kendal method. The result came out that most of the variables have a correlation close to 0 which means that they are slightly correlated or likely to be uncorrelated. According to the table, the variables, signup\_method, affiliate provider, first\_affiliate\_traced, signup app, have less than 0.01 which means these variables are not or least correlated to the country destination.

Among the variables, gender has the highest positive correlation with value 0.04. The age, affiliate\_channel, and first browser have correlation value larger than 0.01 and negative correlation with the result. Other variables that have larger than 0.01 but positive are signup flow, language and first device type used.

Since the variables are categorical and have low correlation with dependent variables, the variable should be analyzed with multivariate logistic regression. Even though the correlation came out low in univariate analysis, it could be significant in the presence of other variables in the multivariate model. Furthermore, multicollinearity can be tested to identify. The multicollinearity makes the precision of estimation therefore, it should be identified in order to find the precise regression model between variables.