Airbnb new user booking prediction

1. Introduction :

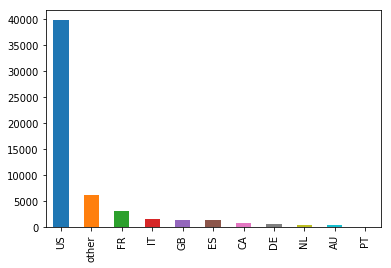
Airbnb is a lodging company that connects host and guest depend on lodging type, date, location, and price through the website and mobile app. The host provides hospitability to the guest in order to gain a higher rate and advertise their rental. For this company to develop, first it is important to 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. Datasets:

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.

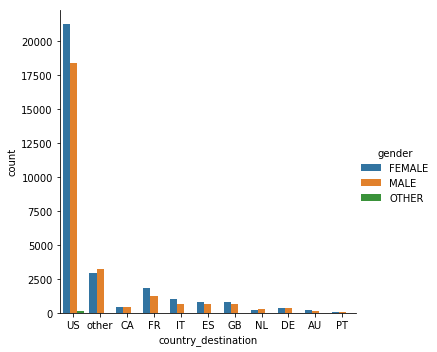
Since there are lot of missing data, first data was cleaned. In the variable age, since it was able to get average from other data, the empty spot for age was filled with mean. However, most of the data were categorical, it was hard to find the mean. Therefore, those NaN data were dropped using dropna function. Also, for the age, since the age below 18is restricted to use Airbnb and age approximately over 95 would not use the Airbnb, therefore the data with age below 18 and over 95 were also dropped . Then train and test data were merged together in order to see how variables in train is related to the country destination in test data. The empty data in the session is also replaced as NaN value and dropped. This session data is also merged into test and train data by same column ‘user id’ in order to see how variables in session related to country destination.

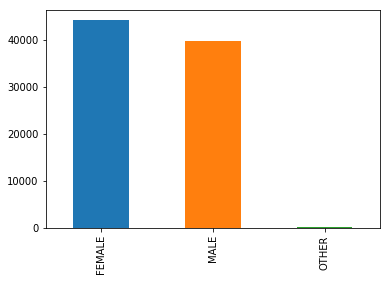
The variable secs\_elapsed is grouped and summed by same user\_id to check how each id took and choose their destinations.



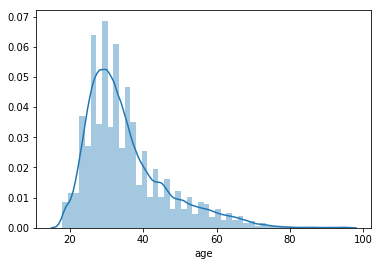
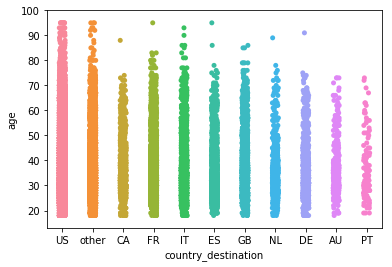
3. exploratory analysis:

After wrangling the data, each variable were plotted as a bar graph or scatter plot to reveal the relationship with a destination. But first, target variable was plotted. In the graph, most of the users selected US as there first destination.

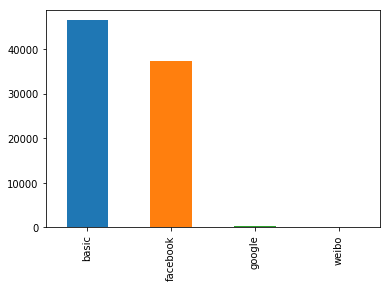
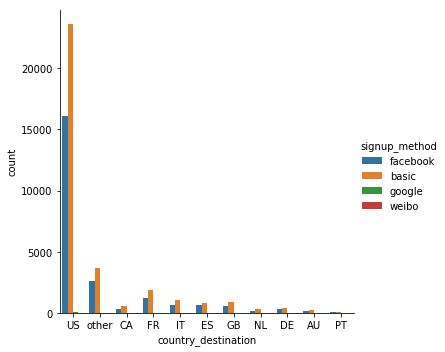




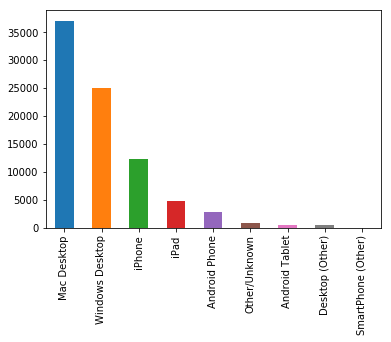
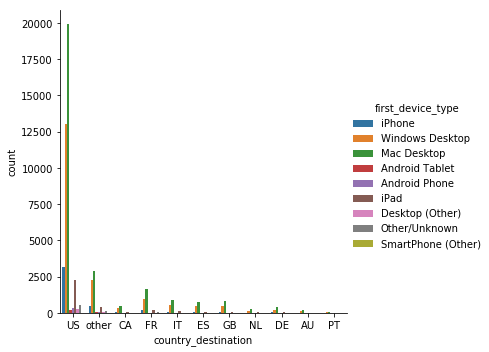
By counting only the gender variable, there was more female user than the male. However, by comparing the gender for each country, there were more female booker or equal amount in gender in other, Canada and Denmark. Since there is difference in gender in country selection, the gender may considered to be significant variable that effect the country destination.

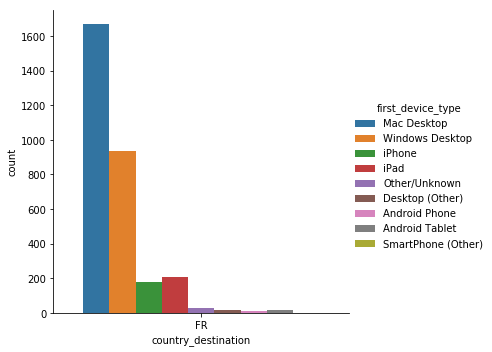
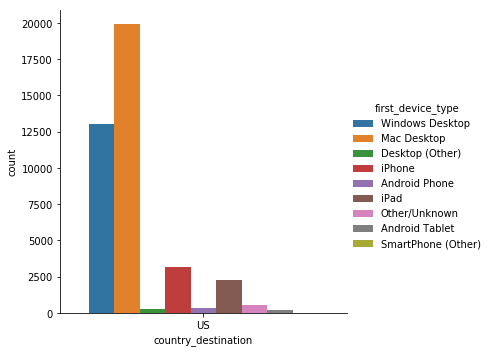


The age was plotted as bar graph. Since the age below 18 and over 95 is dropped, ther eis no data exist. The graph showed that age between 20 and 40 has highest amount of users. For the age in each country, it seems there are large amount of user until 60s. By comparing this plot with country selection plot, the age range tend to drop as country selection is dropped. However, while GB is more selected than ES in country destination, the ES has wider age range than GB. Therefore, the age could be considered as import variable that effect country destination.

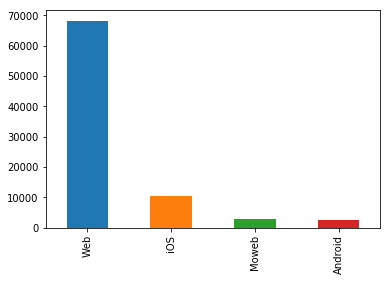
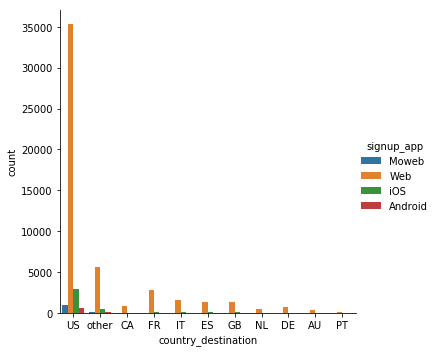
 

For Signup method, The basic and facebook was selected for most of the users as their signup method. Since the pattern matches in all the countries, the signup method variable was considered to be not related to the country destination.

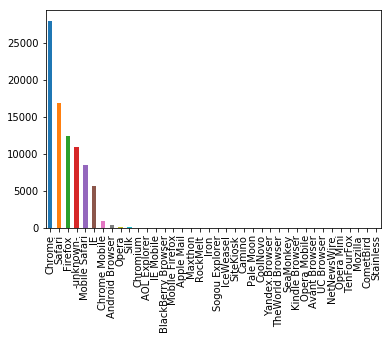
 

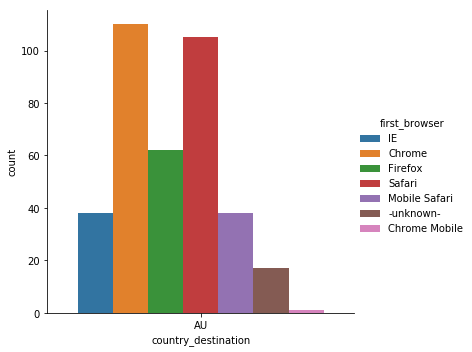


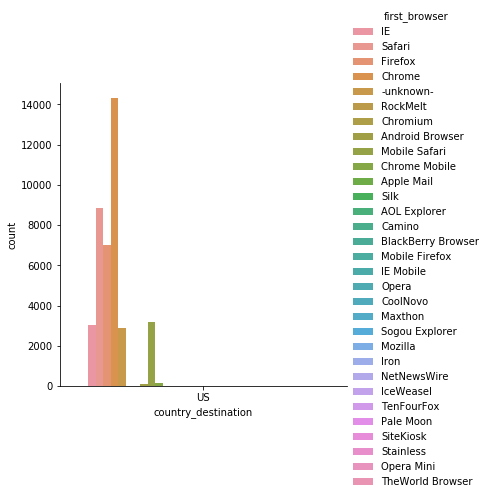
For the variable first device type, the mac and window desktop were used most. When comparing the countries, in all of the countries the mac desktop were used most then the window desktop were used. However, in US, while iphone were 3rd most used device, in France, the ipad were used more than iphoen. Eventhough it is small difference, is could be considered as important variable for the country destination.

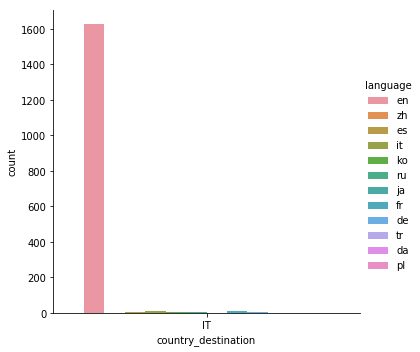
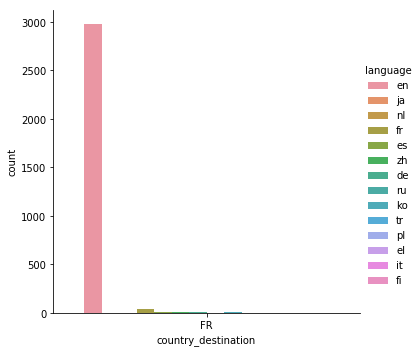
 

In signup app variable, all the countries revealed same pattern. The web was used most then ios was selected next. Since there was no significant pattern in each country, the signup app was considered to be not significant variable for predicting country destination.

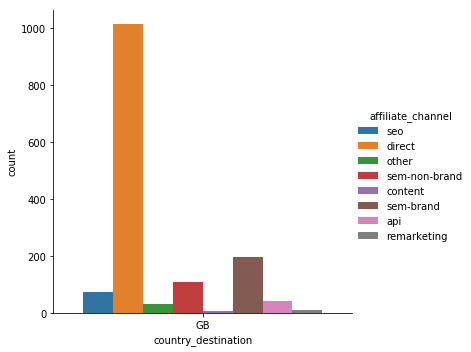
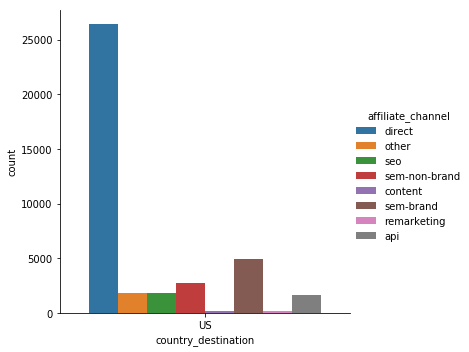
For the first brower, Among the lot of the browser, chrome , safari and firefox were used repectively. It showed same pattern in each country. However, while there were a lot of the browser in US, about 31 kind of browser, there were only 7 browser were used in the Australia. Even though the pattern for top 6 broswers are same, the minor browsers might effect in selecting country destination. Therefore the browser was considered to be important variable



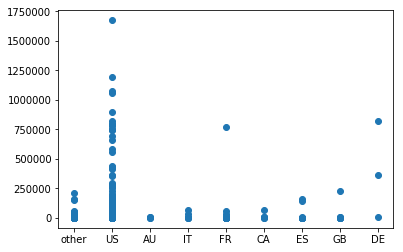


For language, in all the country, English was most used language. However, while it was hard to tell what language is used next to English in US, the French is used more than other languages for user who selected France as their destination. Which is obvious result, therefore, the language was considered to be important variable to predict country destination.

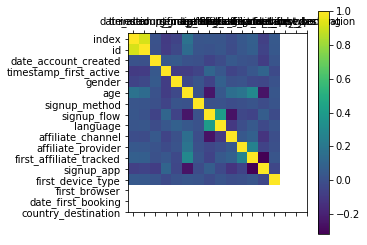
For afilliate channel, the most used channel were same in most countries but there was also minor difference in US and GB. There was more API user than the other channel in GB. This is minor difference, but this could result in country destination, therefore this variable was also considered to be important variable.

****

The variable affiliate provider and date created were plotted but there was no significant difference in each countries. Direct, google were selected most respectively. Also the date created were start to increase in 2012-01 in most of the country and gradually increased until 2014-07. Therefore, these two variables were considered to be not important as other variable.

while US, France and Italia were selected as country destination , For the variable second\_elapsed the user selected Denmark took more seconds than the user selected France and Italia. Therefore, this second elapsed variable was considered to be important variable for deciding country destination.

Since it is hard to identify which of the variable has effect on country destination, the correlation color graph was plotted to identify the variable has relationship with country destination. However, it was hard to tell the color of the plot and identify which of the variable has correlation with country destination.



Therefore, the correlation table were plotted using .corr function. However, all of the variable has 0 which means that they are slightly correlated or likely to be uncorrelated. Since the variables are categorical and have low correlation with dependent variables, the variable will 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.