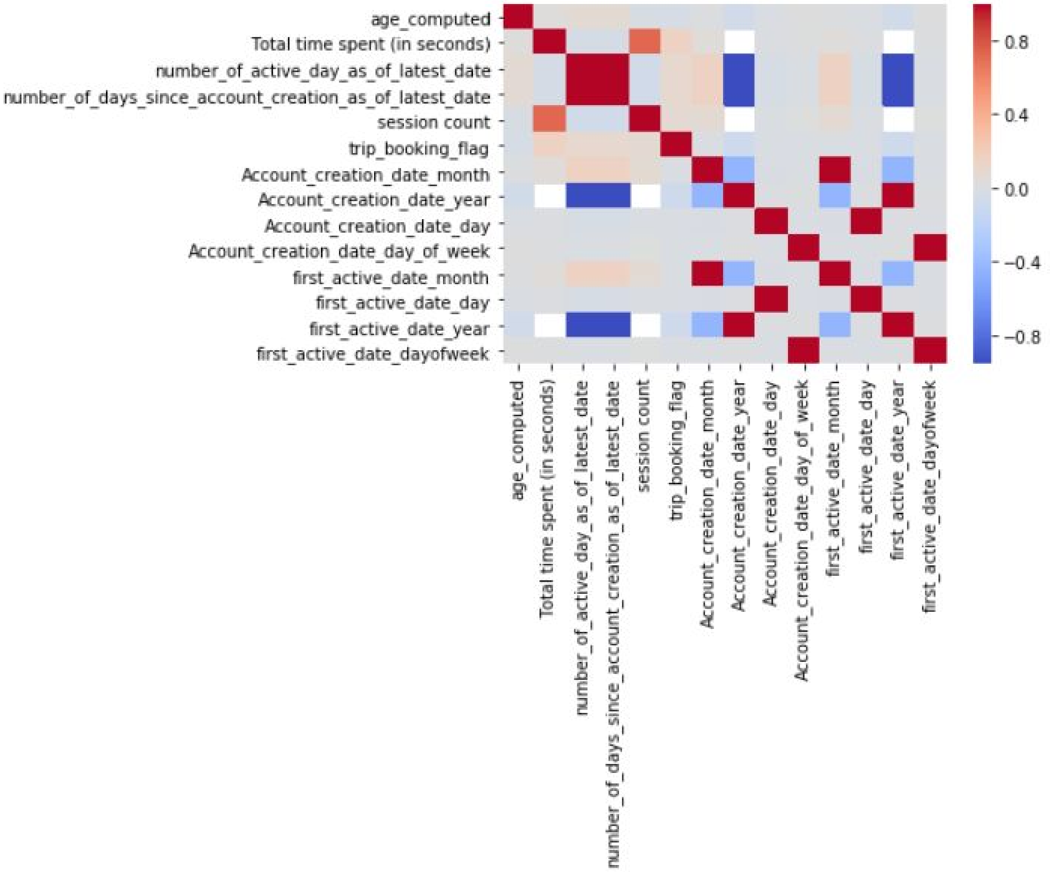
Airbnb Project Exploratory Analysis Report

- ## Is there any strong correlation between the label (trip booking flag) and the features?

By looking at the heatmap below, the feature that is most correlated to the trip_booking_flag is the total time spent (in seconds).



- ## Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?

By looking at the data, below are the three question/hypothesis I have regarding the relationship between the variables.:

- Bookers has stayed with the airbnb website longer than the non bookers (based on the first date active)

- ❖ Bookers has more activities on the airbnb website than non bookers (based on the session count)
- ❖ Age and gender could be important factors to determine which country users booked

- **What are the most appropriate tests to use to analyse these relationships?**

For each of the hypothesis above, I have come up with the test below:

- ❖ Hypothesis 1: Bookers has stayed with the airbnb website longer than the non bookers (based on the first date active)

  Ho: There's no difference between bookers and non-bookers on their number of days since they got active on the website.
  H1: Bookers stay longer with the website than non-bookers.

  For hypothesis 1, I have used z test to compare the average number of days since the first active date between the bookers and non bookers

  The conclusion is the high z test statistic indicates that the p value is very low and there's signification difference between bookers and non-bookers in number of active days.

- ❖ Hypothesis 2: Bookers has more activities on the airbnb website than non bookers (based on the session count)

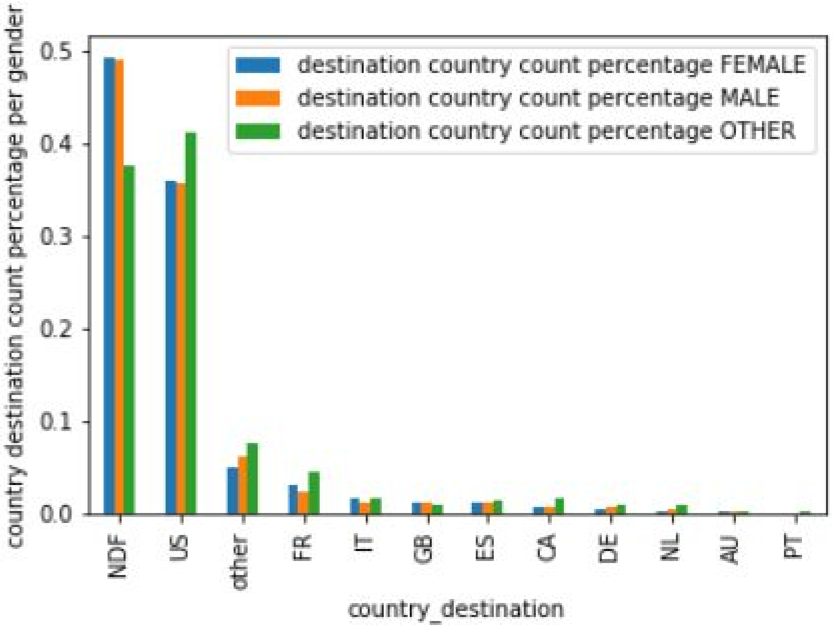
  For hypothesis 2, I have also used z test to compare the average session count between bookers and non-bookers.

  Ho: There's no difference between bookers and non-bookers on their web activities.
  H1: Bookers have more web activities on airbnb than non-bookers on average.

  The conclusion is bookers are on average more active than non-bookers on the airbnb website.

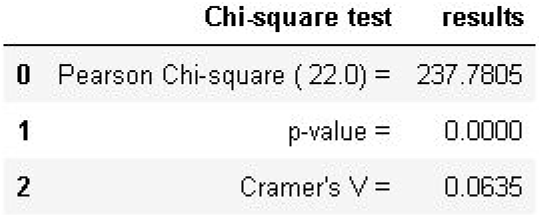
- ❖ Hypothesis 3: Age and gender could be important factors to predict which country users booked

For hypothesis 3, I have used data visualization, chi square test and One Way Anova test to determine the interaction between the label and featureslike user age and user gender.



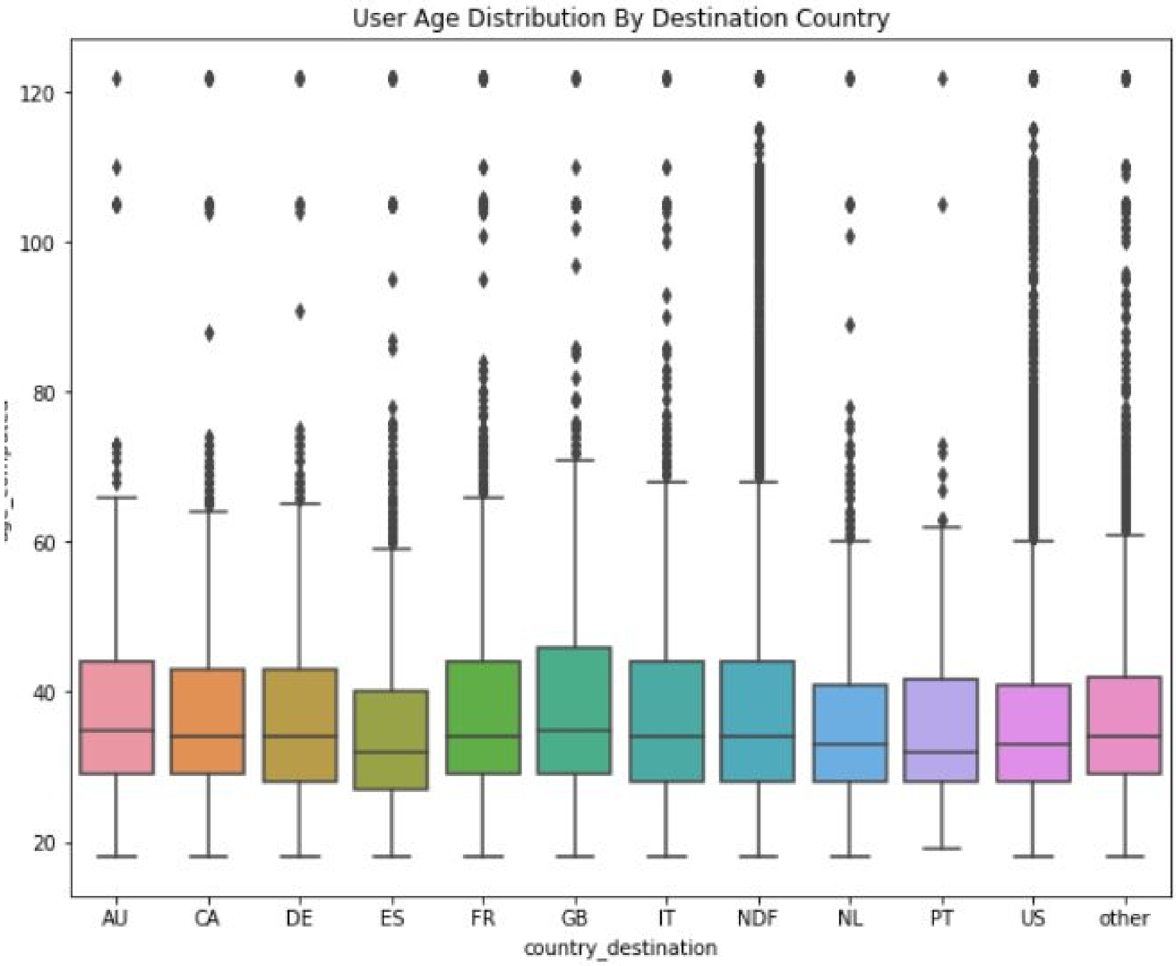
As you can see the bar plot above:

1. Other gender is more likely to book a trip in US, other destinations and France compared with users identified as Male and Female gender.
2. Users identified as male and female gender are less likely to book a trip.



	Chi-square test	results
0	Pearson Chi-square (22.0) =	237.7805
1	p-value =	0.0000
2	Cramer's V =	0.0635

By doing the Chi-square test, we can reach the conclusion that there's a significant relationship between gender and booking destination. But the relationship is not that strong given Cramer's V is only 0.0635.



By looking at the box plot above, you will find people going to countries like Spain, Portugal and Netherland tend to be relatively younger and people going to UK tend to be relatively older.

By looking at the test result from the one-way Anova test, there's a significant difference in age between users going to different destination countries.