Capstone Project 1: Milestone Report

# Problem statement and Objective

Client Name: Airbnb

In which country will a new guest book their first travel experience?

The solution to this problem will help Airbnb understand the following:

* Better forecast demand
* Share customized content with client
* Decrease the average time to first booking

# Data Description:

* Data from Kaggle competition started 4 years ago. It contains
  + Test data
  + Train data
  + Other data which might be needed
* The link to the competition is: <https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data>

In this challenge, Kaggle has given a list of users along with their demographics, web session records, and some summary statistics.

All the users in this dataset are from the USA.

There are 12 possible outcomes of the destination country: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found), and 'other'. Please note that 'NDF' is different from 'other' because 'other' means there was a booking, but it is to a country not included in the list, while 'NDF' means there wasn't a booking.

* The training and test sets are split by dates. In the test set, we need predict all the new users with first activities after 7/1/2014
* The sessions dataset, the data only dates back to 1/1/2014, while the users dataset dates back to 2010.

The list of datasets is as follows:

1. train\_users.csv - the training set of users
2. test\_users.csv - the test set of users
3. id: user id
4. date\_account\_created: the date of account creation
5. timestamp\_first\_active: timestamp of the first activity, note that it can be earlier than date\_account\_created or date\_first\_booking because a user can search before signing up
6. date\_first\_booking: date of first booking
7. gender
8. age
9. signup\_method
10. signup\_flow: the page a user came to signup up from
11. language: international language preference
12. affiliate\_channel: what kind of paid marketing
13. affiliate\_provider: where the marketing is e.g. google, craigslist, other
14. first\_affiliate\_tracked: whats the first marketing the user interacted with before the signing up
15. signup\_app
16. first\_device\_type
17. first\_browser
18. country\_destination: this is the target variable to predict
19. sessions.csv - web sessions log for users
20. user\_id: to be joined with the column 'id' in users table
21. action
22. action\_type
23. action\_detail
24. device\_type
25. secs\_elapsed
26. countries.csv - summary statistics of destination countries in this dataset and their locations
27. age\_gender\_bkts.csv - summary statistics of users' age group, gender, country of destination
    1. This dataset contains number of Male/Female in different age buckets who travelled to these destination countries in 2015 **[Own interpretation]**
28. sample\_submission.csv - correct format for submitting your predictions

Further, the following data processing steps were done to create the dataset for EDA and modeling

1. Train and Test datasets were loaded with the NaN replacing all the missing values.
2. All the date time variables were converted to the right date time format using Pandas library
3. Merge the ‘users’ dataset with ‘countries’ data to get the language of the destination country.
4. The following additional variables were created:
   1. Ind\_same\_lang: It has a value = 1 if the user’s international language preference is same as the language of the destination country
   2. days\_since\_first\_book: This variable contains difference in days between the first day of activity and day of booking
   3. age\_bucket: This variable is created so that it can be used with ‘country\_destination’ and ‘gender’ to pull population from the dataset ‘age\_gender\_bkts’

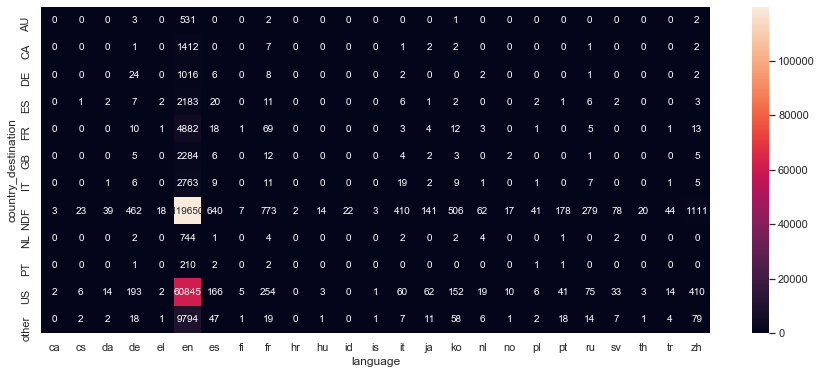
# Exploratory Data Analysis

The following section gives you a breakdown between the country of destination and predictor categorical variables.

1. **Country of destination and gender:** 
   1. US is the destination country with highest number of users who are evenly distributed between Male, Female and Unknown categories.
   2. In other countries also, users are evenly distributed between Male and Female



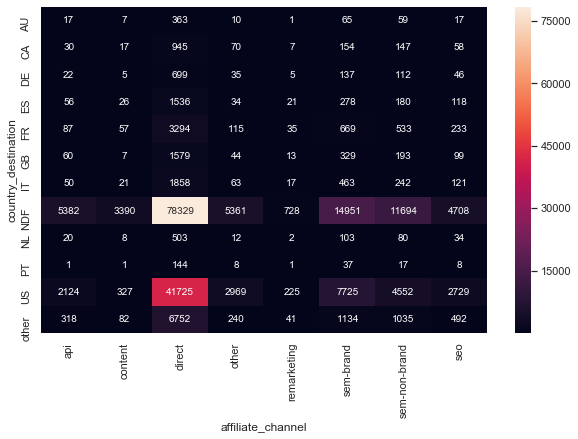
1. **Country of destination and international language of preference:** 
   1. Again, most users have ‘English’ as their language of preference



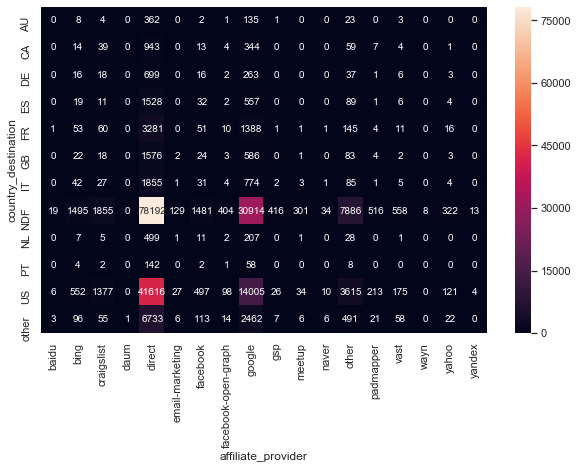
1. **Country of destination and signup method:**
   1. Most signups take place through Airbnb’s own website and facebook



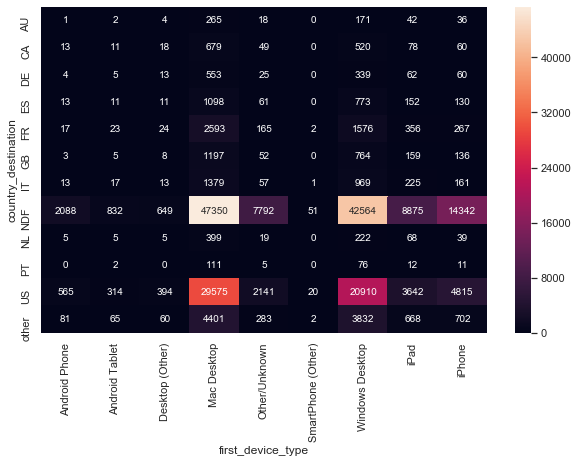
1. **Country of destination and affiliate channel:**

****

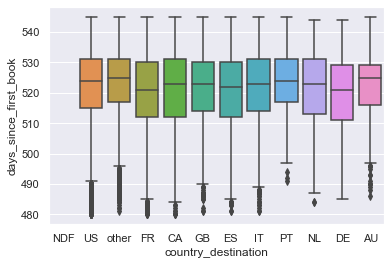
1. **Country of destination and affiliate provider:**

****

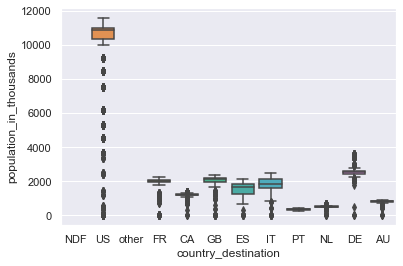
1. **Country of destination and first device type:**
   1. Most users are concentrated in the US, and they are either Mac Desktop or Windows Desktop users

****

1. **Country of destination and days since first booking:**
   1. The median number of days since first booking is more or less similar for different countries with similar variability as well.



1. **Country of destination and population:**
   1. This population could be the total number of people who travelled to these destination countries in 2015. US was the most popular travel destination in 2015. Since user base is from US, it shows most users travelled within US which makes sense as well

****

# Statistical Data Analysis

1. Chi-Square test between target variable and independent variables to measure association between them: Since the target variable is a categorical variable and most of the predictor variables are also categorical variables, Chi-square test has been used to measure association.

The data also satisfies the assumptions of chi-square test:

* 1. The levels (groups) of the variables being tested are mutually exclusive
  2. Each participant contributes to only one cell within the Chi-square table
  3. The groups being tested must be independent
  4. The value of expected cells should be greater than 5 for at least 20% of the cells

**Results**: There’s an association between the country of destination and the following predictor variables based on chi-square testing:

1. Gender
2. International language of preference
3. Sign up method used
4. First device type
5. First browser
6. ANOVA can also be used to test association between target variable and independent variables. Since the target variable is a categorical variable and a few predictor variables are continuous variables, ANOVA is an appropriate test to measure association in such cases. The results are as follows:
   1. **Association between country of destination and # of days since first booking:** Since the F-statistic= 38.9 and the p-value= 1.9e-77 which is indicates that there is a significant effect of # of days since the first booking on country of destination but this test may not be very reliable as it may violate some of the assumptions of ANOVA
   2. **Association between country of destination and population: [To be updated]**
   3. **Association between country of destination and seconds per session: [To be updated]**

# Approach **(To be updated)**

This is a supervised learning algorithm because we have labeled training data

This is also a classification problem because we have a fixed number of categories we need to predict as part of the target variable

The variable to be predicted is destination country.

The predictor variables are be as follows:

User’s current country, age, gender, month, distance between their country of residence and destination country, user’s native language and language of the destination country

1. References
2. <https://pythonfordatascience.org/chi-square-test-of-independence-python/>
3. <https://pythonfordatascience.org/anova-python/>