## horizontal line



Airbnb New User Prediction

# Overview

Instead of waking to overlooked "Do not disturb" signs, [Airbnb](https://www.airbnb.com/) travelers find themselves rising with the birds in a whimsical treehouse, having their morning coffee on the deck of a houseboat, or cooking a shared regional breakfast with their hosts.

New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. By accurately predicting where a new user will book their first travel experience, Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand.

# Problem Statement

The problem is to predict to which country the new user of AirBnb will book his/her first travel experience.

# Data Source and Specification

The dataset is taken from the kaggle competition page — <https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data> .

We are given a list of users along with their demographics, web session records, and some summary statistics. You are asked to predict which country a new user's first booking destination will be. 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 is to a country not included in the list, while 'NDF' means there wasn't a booking.

# Solution Process

## Exploratory Data Analysis

## Data Cleaning

## Featurization and Encoding of Data

## Building the Model

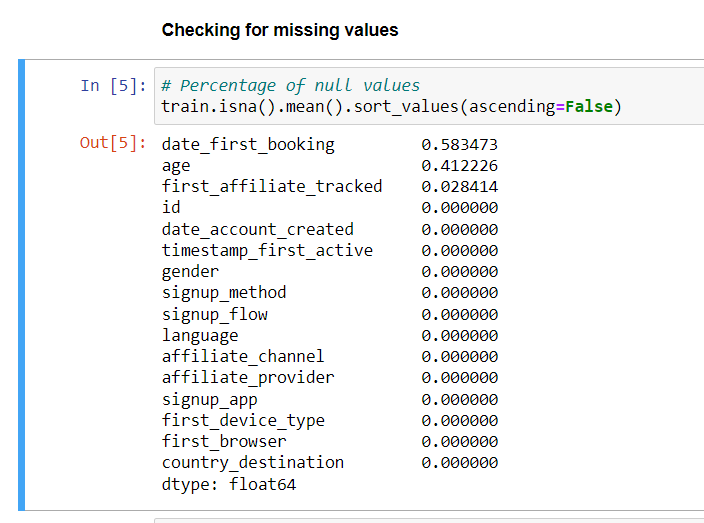
## Model Evaluation & Iteration

# Exploratory Data Analysis :

For exploratory data analysis(EDA) we decided to divide the task into two teams after analyzing raw data . We were divided into two teams, one to perform univariate exploration and one to perform multivariate exploration so that we could understand all the features and also the relation between features.

## Raw Data Analysis:

### Missing data analysis:

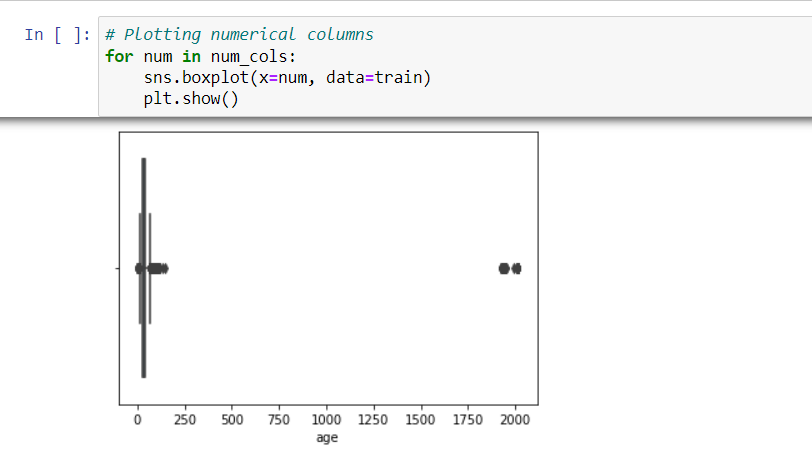


From the train dataset, 58.34 % of the **date first booking,** 41.22% of the **age,** and0.0322% of the **first affiliate tracked columns were missing.**

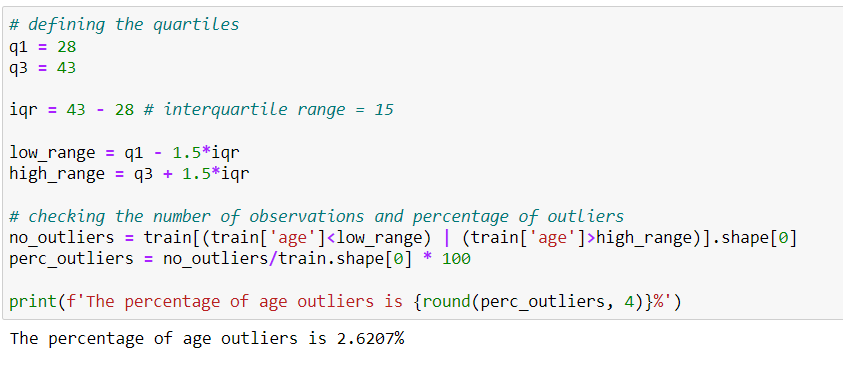
## Univariate analysis:

We did univariate analysis to check for outliers and to better understand the data we got two main conclusions;

### AGE FEATURE:

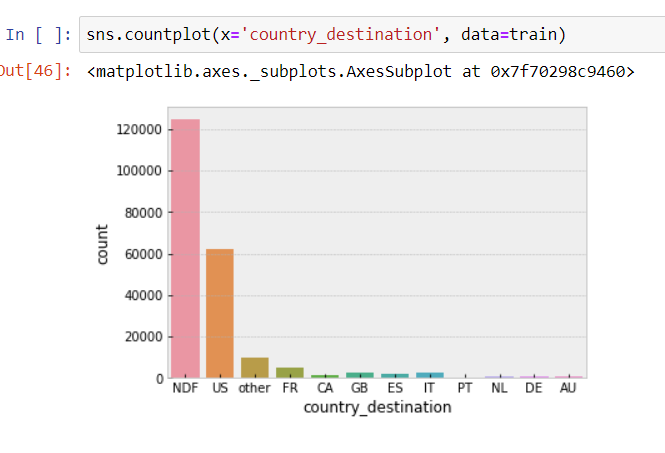


This plot showed that the age column has outliers.



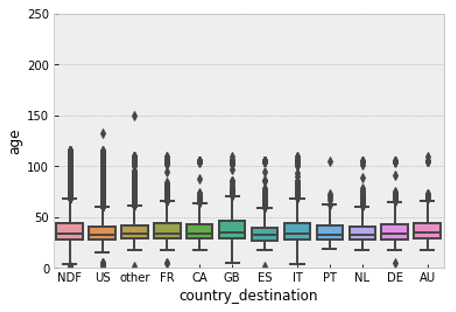
The percentage of outliers came to **2.670%**. We decided that outliers can be filled by taking the median of values.

### Country destination:

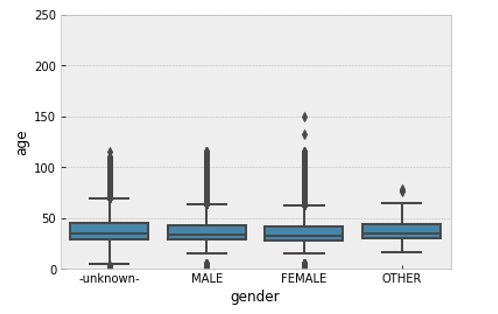
This graph showed that most people do not book a flight and most people travel from the US to the US, that is they use domestic flights.

## Bi-variate analysis:

We then performed bivariate analysis, an analysis of any concurrent relationship between two attributes or a variable. So to see concurrent relationships between attributes in our dataset we used bivariate analysis.

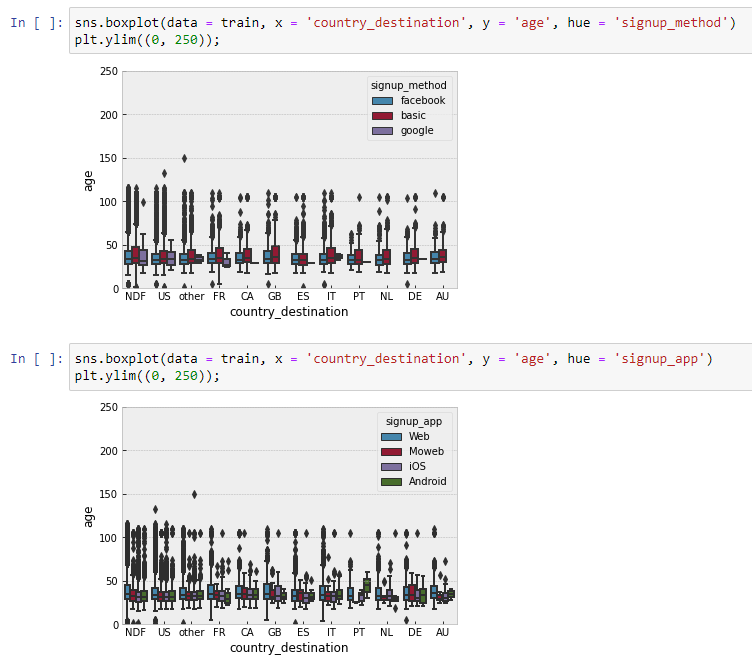


This graph shows us that younger travellers predominantly travel to the US, ES, IT and CA while older travellers turn to travel to GB, IT, AU, DE and NDF.



Based on the figure above we saw that female travellers are the youngest and some of the unknown travellers are the younger.

## Multivariate Analysis:



Based on the above two plots we were able to determine that younger travellers use Facebook as a signup method while the older generations used the basic way of signing in to travel. The younger the traveller the more technology is used. On the basis of the second graph, we can see that the younger travellers used Moweb, Android, and iOS as a signup.

# Data Cleaning:

After the analysis of the data the next step was to clean the dataset to remove the unfavourable variables for our model to perform better. The cleaning of the dataset was divided into steps.

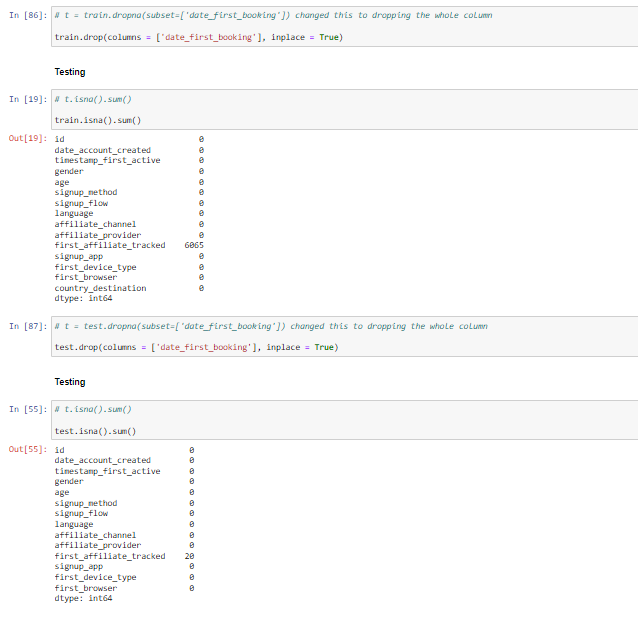
#### Step-1;

Filling null values and replacing outliers with the median.

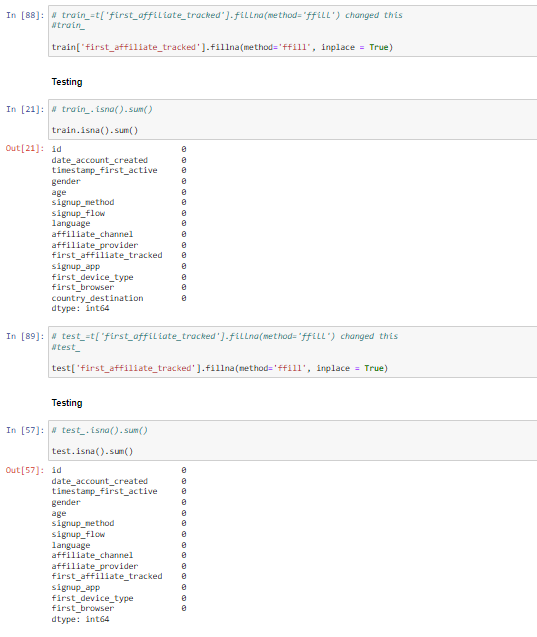


#### Step-2;

Dropping the data\_first\_booking as the null values represent the users who did not travel.

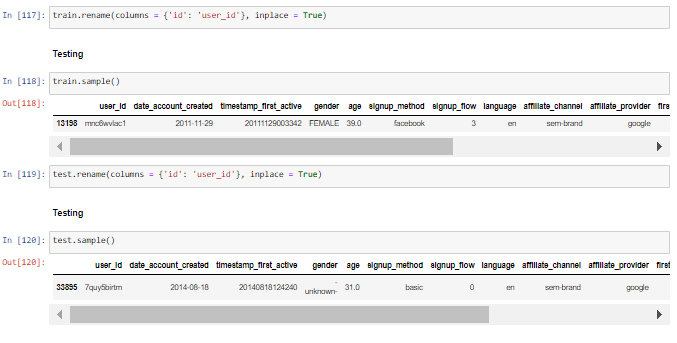


#### Step-3;

Filling missing first\_affiliate\_tracked with forward fill.

#### Step-4;

Renaming id to user-id.



# Feature engineering and encoding :

After the cleaning of data the next step is feature engineering. Feature engineering is a very important step in machine learning. It consists of designing artificial features into an algorithm.

Encoding is a technique of converting categorical values into numerical values.

# Building the Model :

Now we have got our dataset cleaned and we have selected our features to implement the model . We went ahead and tested with various models and merged various datasets to get the best accuracy.

### Baseline Model :

We first built a baseline model using three different classifiers, KNN classifier, RandomForestClassifier and GradientBoostingClassifier. The best performing out of the three was GradientBoostingClassifier with a 62.9% ndcg score. The least performing was KNN classifier with 58.67% score. Thus we decided to tune the GradientBoostingClassifier.

We then tried another classifier that is RandomizedSearchCV and it got a ndcg score of 63%.

Since our model still performs slightly above 62.9% we decided to apply more feature engineering to beat the baseline model score.

### Best Model:

For our best model we **merged our train and sessions** dataset and went on to perform feature engineering and encoding on the merged dataset. We performed **one hot encoding** on the independent categorical variables and **label encoding** on the target variable. We **dropped redundant values** by the use of vectorization. Then we merged this dataset with our train data frame.

We used **GradientBoostingClassifier** on the final dataset and we got an accuracy of ***87.26%.***

To further try to improve this model, we performed feature importance and dropped the 20th least important features from the dataset. This however gave a rather inconsiderable change in ndcg score.

# Conclusion:

Working with our best model obtained, we were able to predict the top 5 most probable destinations for each user.

**Future work:**

1. More features could have been extracted from the session data as well as train data, also other data sets like countries and age\_gender\_bkts can be used to build a new data set for training the model.
2. More hyper parameter tuning on different parameters of the model can be done.

# Team Members:

1. Volker Felvic Katche Tachin
2. Harsh Pant
3. Peter Baba
4. Chris Macharia
5. Thandazi Mnisi
6. Tanu Deep
7. Ameh Lawal
8. Sharon Prempeh
9. Gift Wilfred
10. Rahul TG
11. Isaac Baah
12. Elenpe Oluwakemi Deborah
13. Abdulrauf Taofeek Bolaji