Rounak Banik
IIT Roorkee
rounakbanik@gmail.com
+91 84398 60325

Airbnb Final Report

29th September 2017

OVERVIEW

The problem this capstone project aims at solving is predicting where a newly registered Airbnb user will book his/her first travel experience. By analyzing a variety of data such as user informations, country information and user browsing session records, the project aims at coming up with a model that could accurately detect where a user would book his/her first experience from an array of 34000+ cities across 190+ countries.

THE CLIENT

The Client in question is Airbnb, an online marketplace and hospitality service enabling people to lease or rent short-term lodging including vacation rentals, apartment rentals, homestays, hostels beds, or hotel rooms,

Since its inception in 2008, Airbnb has expanded into more than 190 countries and 34,000 cities across the planet. It has disrupted the hotel industry and has become the premier choice of booking for many travelers across the globe with people booking a wide variety of spaces from tree houses to houseboats to apartments.

When a user first signs up for Airbnb, it is in the user's and Airbnb's best interests to show homes and places in a city and country that the user intends on visiting. Since it is not possible to explicitly ask the user for this data and expect the user to update his/her preference (a user may just be browsing, s/he may change preference from time to time and may not update that in his/her profile, etc.), it would make sense to actually build a predictive model around the user's browsing data and give predictions based on it.

This would serve the following purposes:

1. Share more personalized content with community
2. Decrease the average time of booking
3. Better forecast demand

In other words, based on the analysis done and the model built, Airbnb will be able to provide a much better user experience by offering a news feed tailored to each user's preferences for their next destination.

This, in turn, would decrease the time the user took to book their first host as they will spend considerably less time searching for homes and more time comparing appropriate options.

Finally, it would also help Airbnb which cities and countries are in hot demand at a particular time of the year. This would help them in suggesting more competitive pricing to hosts in these regions. Additionally, it would also help the client focus their marketing efforts in a region where demand is high, leading to better sales.

THE DATA

The data is already available to us in the form of a Kaggle competition hosted by Airbnb in 2015. Therefore, there isn't a need for additional data mining or web scraping. The data will however have to be cleaned and wrangled before any analysis is performed on it.

The data provided by Airbnb is in the form of CSV files and are listed below:

1. **train_users.csv**: The training set of users
2. **test_users.csv**: The test set of users. Contains user information such as gender, age, language, signup and device information
3. **sessions.csv**: Web sessions log for users. Contains time, type and details of various user actions.
4. **countries.csv**: Summary statistics of destination countries in this dataset and their locations
5. **age_gender_bkts.csv**: Summary statistics of users' age group, gender, country of destination
6. **sample_submission.csv**: Correct format for submitting predictions

DATA WRANGLING

Overview

This section describes the various data cleaning and data wrangling methods applied on the Airbnb datasets to make it more suitable for further analysis. The following sections are divided based on the datasets provided.

]Age, Gender and Statistics

1. The age bucket was converted into a mean age feature. This was done to treat age as a numerical feature. Furthermore, the numerical nature of the data will make it easier to perform one hot encoding and label encoding on this feature later.
2. The year feature was dropped as it had only one value, 2015. Therefore, it was giving us no extra information and could be safely dropped.

Countries

The dataset is clean and extremely small. No wrangling or cleaning techniques were used on this dataset.

Sessions

1. All the unknown fields were converted into NaN to give it more semantic meaning.
2. The missing second values were interpolated using the Pandas Series Interpolate function. This was done as this treatment did not significantly alter the summary statistics of seconds elapsed.

Training Users

1. All unknown values were converted into NaN to give it more semantic meaning.
2. Samples for which age was greater than 120 was converted into NaN as these clearly represented polluted data.

INFERENTIAL STATISTICS

Overview

This section presents the results of the inferential statistical methods applied on two hypothesis tests namely:

1. Gender Preferences for Airbnb Bookings
2. Relationship between Signup Method and Device Type.

Gender Preferences for Airbnb Bookings

This test was performed to test if there is a relationship between the gender of a user and the Airbnb Country Destination. In other words, does your gender influence the country you will travel to and book an Airbnb in? To perform this analysis, we only took into consideration users

that identified themselves as either male or female. Users with no destination or a destination to a country not listed as a class were not considered either.

Since we were comparing two categorical variables, of which one was multivariate, the ideal statistical tool to be used was the Chi Square Test for Significance. The data available to us was pivoted into a form usable by Scipy's Chi Square Contingency Method.

	AU	CA	DE	ES	FR	GB	IT	NL	PT	US
gender										
FEMALE	207	455	358	853	1962	881	1091	254	78	22694
MALE	188	477	416	677	1335	682	699	278	69	19457

1. There is a significant relationship between gender and country destination.
2. The P-Value obtained was 5.8×10^{-21}

Device and Signup Preferences

This test was performed to check if there is a relationship between the type of device and the signup method. In other words, were you more likely to signup through Facebook if you were using a mobile? To perform this analysis, we only took into consideration the basic and the Facebook signup methods as they made up the bulk of the signups. Also, we considered two types of devices: Computer and Mobile. iOS, Android and the Mobile Web Browser were all clubbed into the same category.

Since we were dealing with two binary categorical variables, we had the choice between two statistical tests: The Chi Square Test for Significance and the Two Sample Significance Test. We applied both these tests and compared the results to arrive at our conclusion.

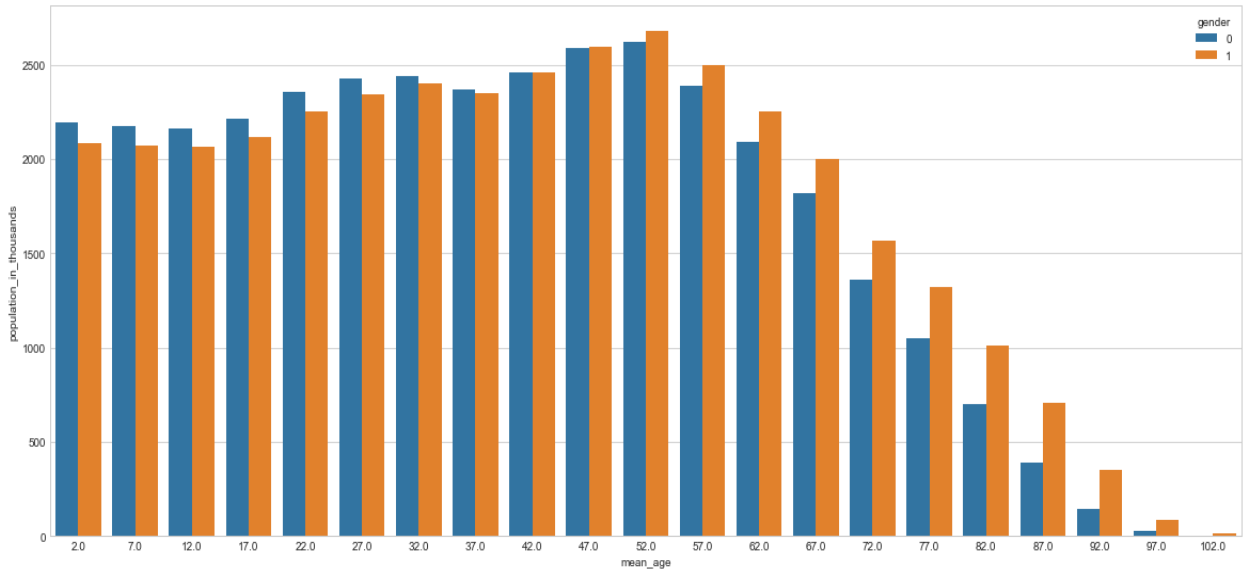
	Basic	Facebook	Total
Computer	131237	51480	182717
Mobile	21660	8528	30188
Total	152897	60008	212905

1. There is no relationship between device type and signup method. The two variables are independent of each other.
2. The result obtained from both the two sample significance test and the chi square test were the same. The Chi Square Test was performed without the correction term.
3. The P-Value obtained in both tests was 0.78.

EXPLORATORY DATA VISUALIZATION AND ANALYSIS

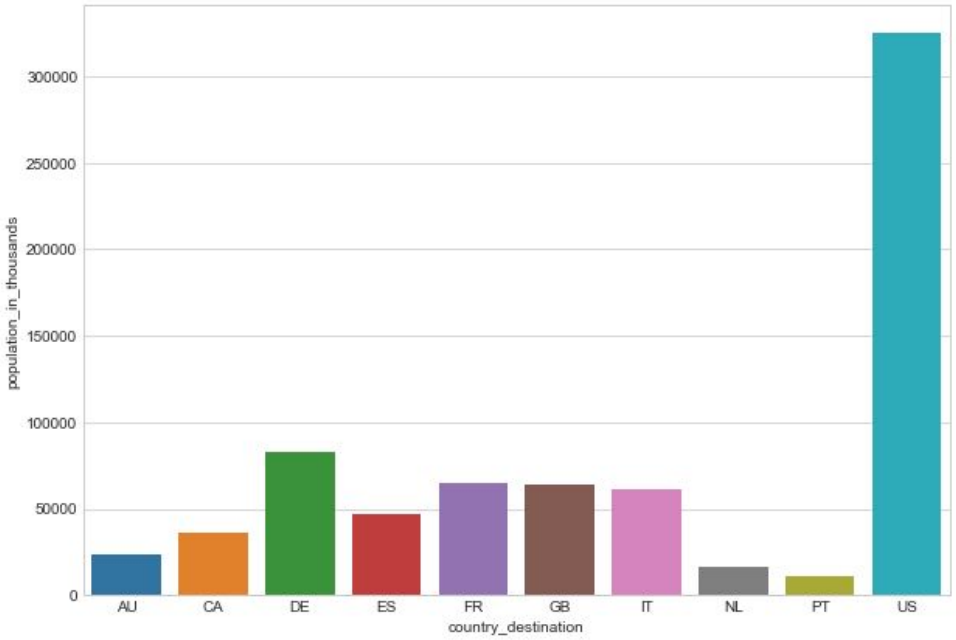
In this section, the various insights produced through descriptive statistics and data visualisation is presented.

Country Statistics



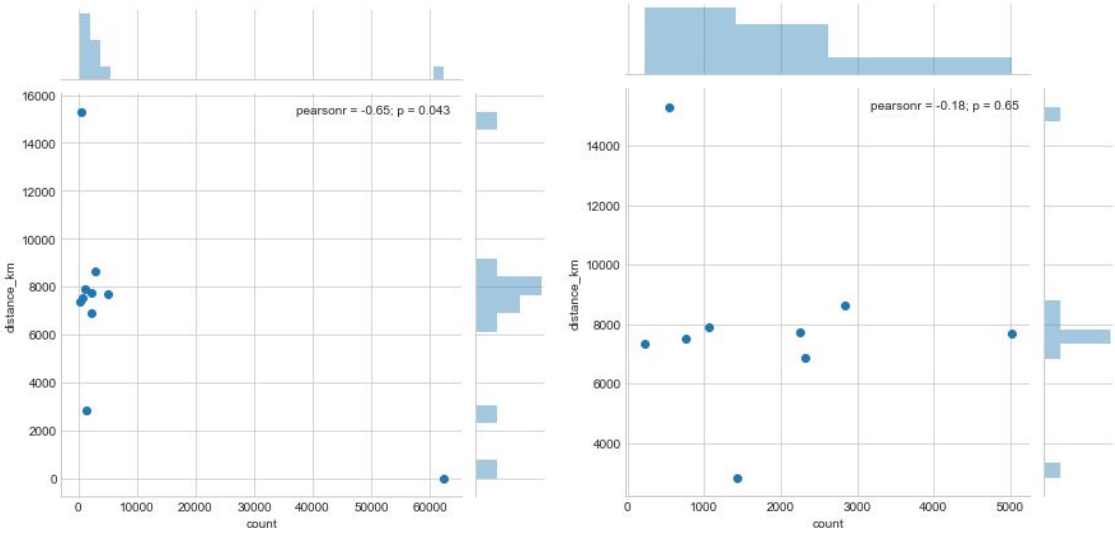
1. The countries that are represented in this dataset largely consist of an **aging population**. The largest groups are people with **mean ages 47 years and 52 years**.
2. The distribution resembles a skewed bell curve. The middle aged people occupy the largest share of the population, closely followed by the youth and finally, the old.
3. The population counts of young and middle aged people are fairly comparable. But as we transition towards old age (age > 57 years), the population count for every successive bucket decreases steadily.
4. One very interesting thing to note is that the sex ratio is skewed towards men for younger age groups but as the mean age increases, the ratio skews more towards women. **Women indeed live longer than men.**

Next, let us try and graph the population count in each country.



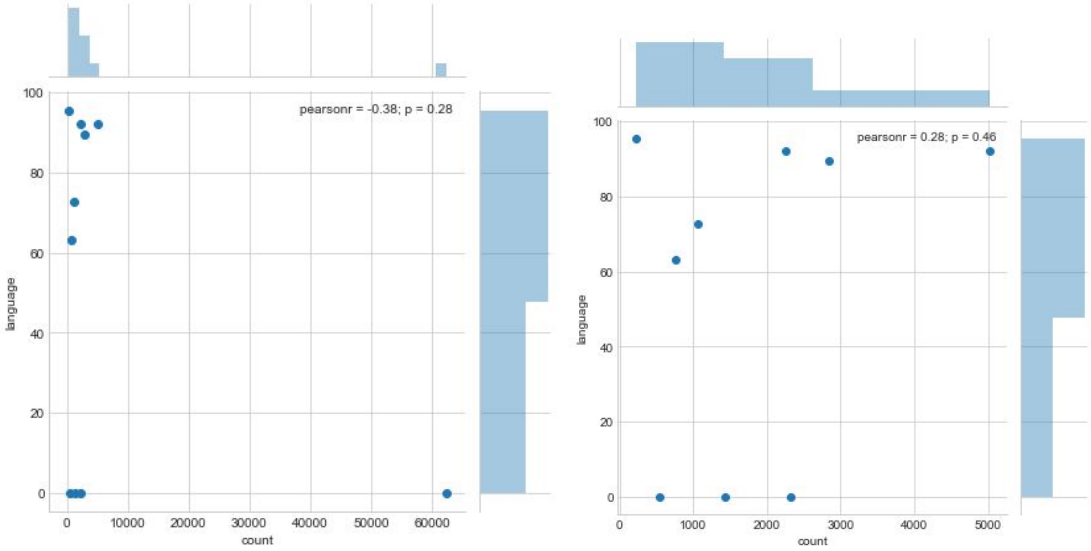
The United States of America is clearly the most populated nation amongst the destination countries with a population of over 300 million. All the other countries in the list have a population less than 100 million.

Distance of Countries



There is a **strong negative correlation of -0.65.** People overwhelmingly prefer booking in the United States than any other country in the world.
However, when taking only international countries into consideration (i.e except United States), the **correlation is much weaker at -0.18.**

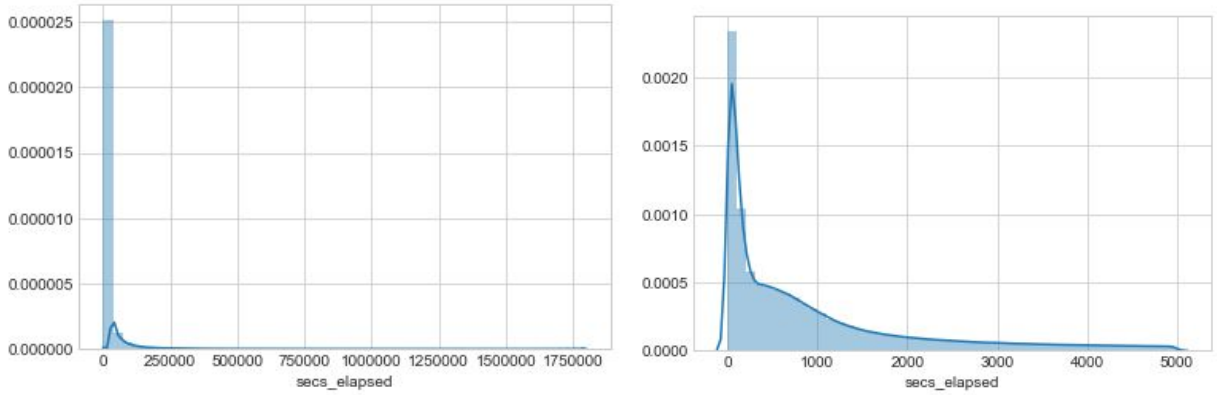
Language Levenshtein Distance



There seems to be a **medium negative correlation of -0.38** between the two quantities. But when the United States is taken out of the equation, there is actually a **positive correlation between language distance and booking frequency.**
At first glance, this may suggest that people prefer countries with different language (and therefore culture) while travelling abroad. Another way of looking at it would be that Europe is an extremely popular travel destination for Americans and they tend to prefer it to other English Speaking countries like Canada and Australia. So this may have nothing to do with language difference as it may have to do with destination preference.

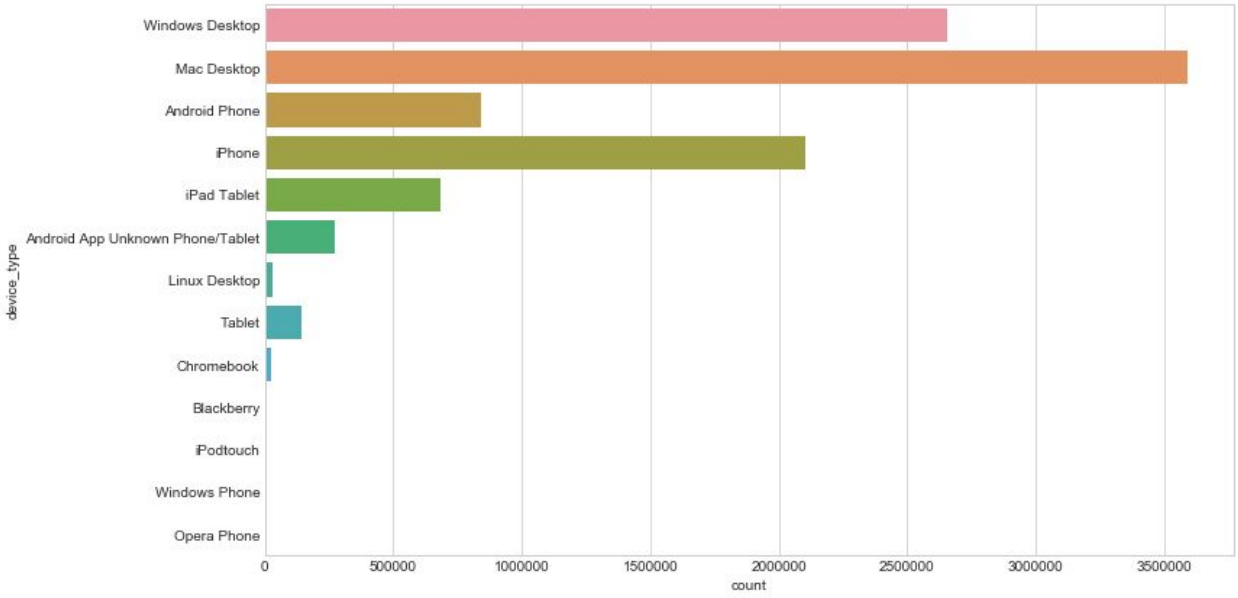
Session Statistics

Seconds Elapsed Distribution



We can see that most the number of sessions greater than 1000 seconds decreases almost exponentially. It is fair to assume that most sessions were less than 1000 seconds long. **Almost 47% of all sessions were less than 1000 seconds long**. This strongly suggests a decreasing exponential distribution of seconds elapsed on each session. In other words, as the number of seconds increases, the number of instances of sessions requiring that much time exponentially decreases.

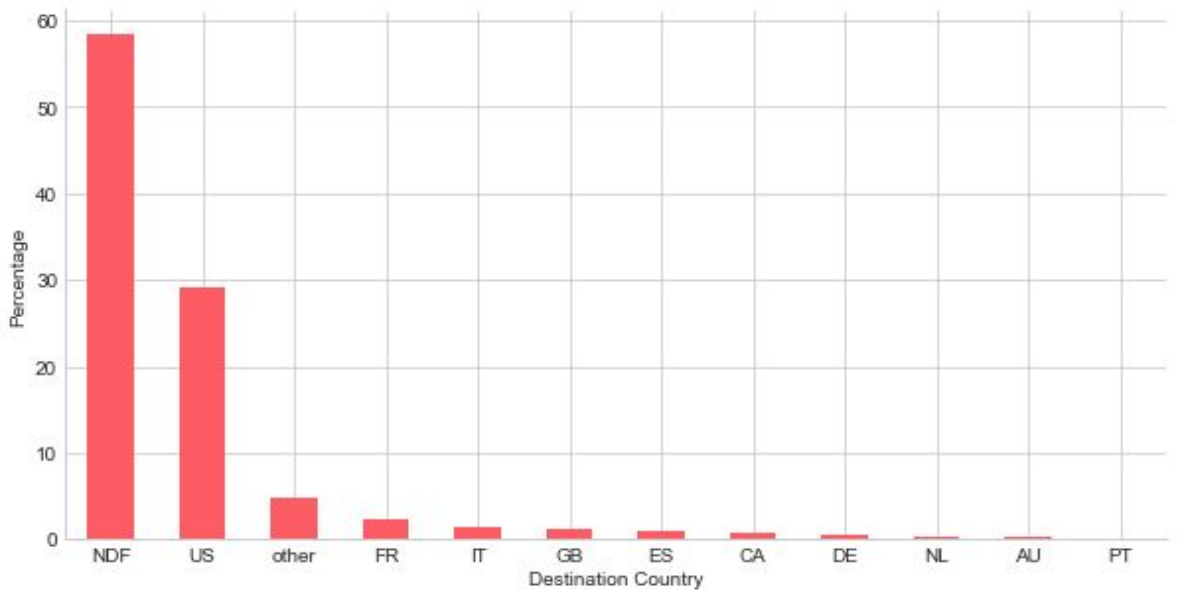
Device Preferences during Sessions



The Mac Desktop is the most popular device among Airbnb Users, followed by the Windows Desktop. An interesting insight is that **Apple Products are extremely popular** with Airbnb Users. The iPhone, iPad and the Mac all appear in the list of top 5 most popular devices.

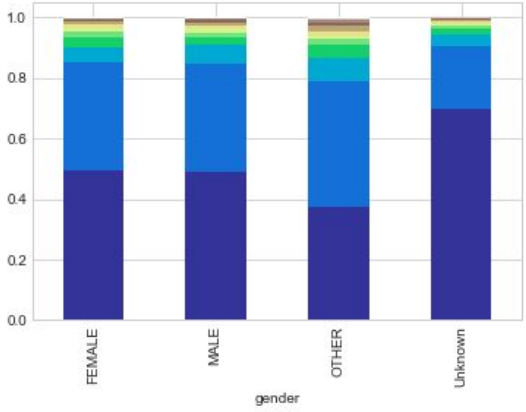
Training Dataset Statistics

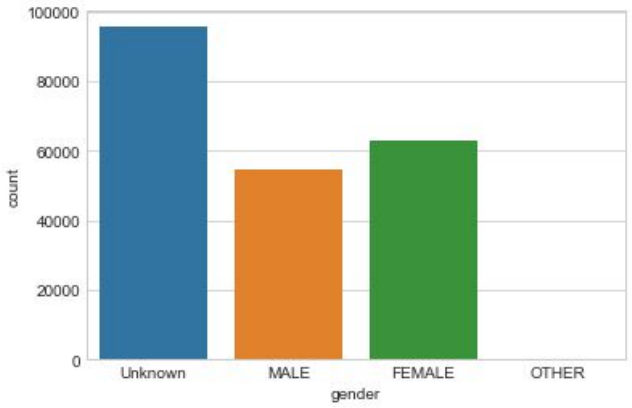
Distribution of all Airbnb Bookings



As can be seen above, **close to 60% of users have never booked an Airbnb**. Among the users that have, they have overwhelmingly chosen **United States** as their first destination. When training our machine learning model, it is of interest to us to separate the bookers from the non bookers. Subsequent classification amongst bookers would yield a high accuracy as we could use the imbalance of classes to our favor.

Gender



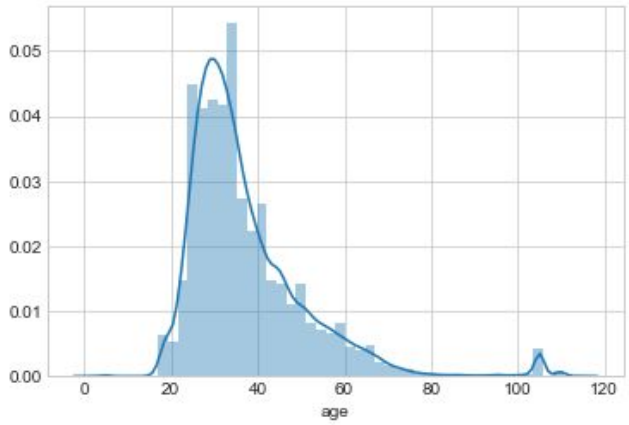


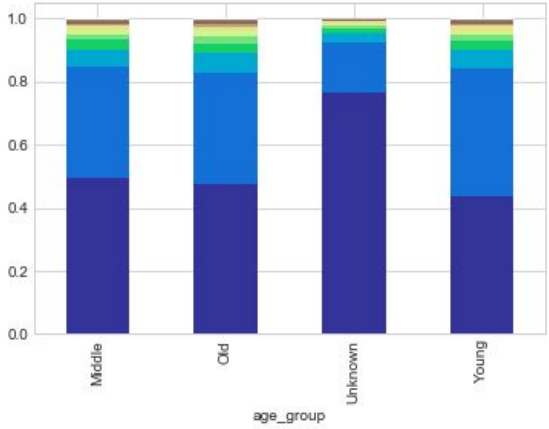
We see that the users whose gender is unknown form the majority. Out of the users whose gender is known, **there are more females than males**. This can suggest two things:

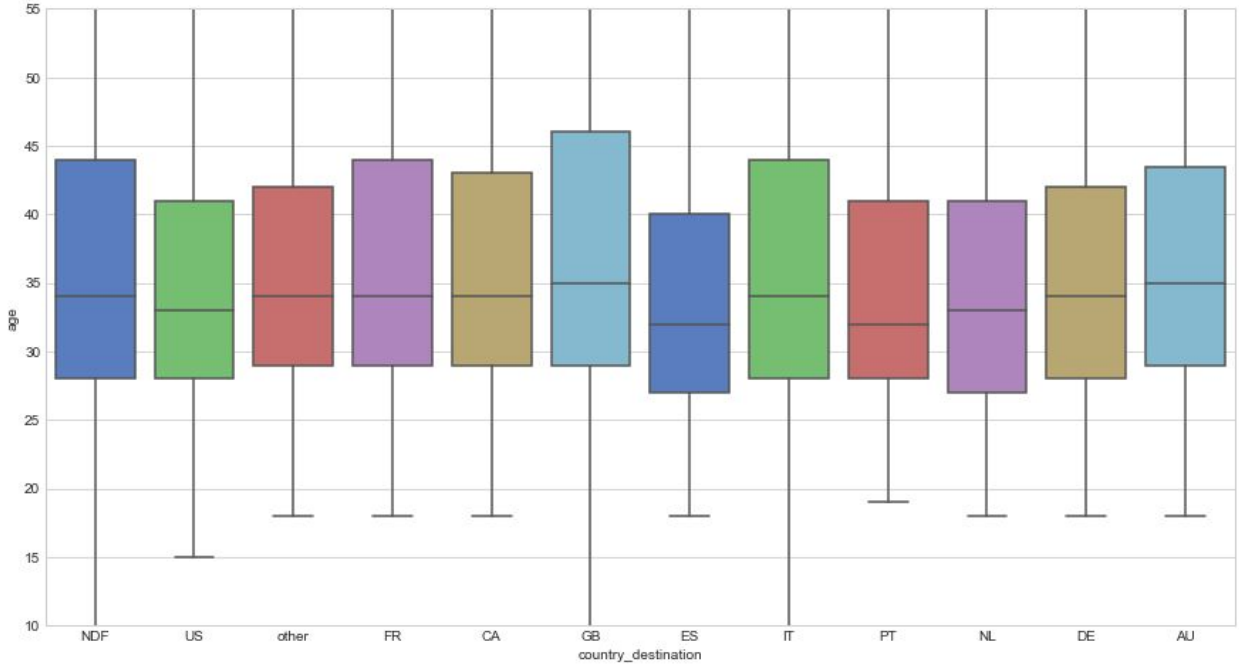
1. There are more female Airbnb Users than male
2. Women are more likely to disclose their gender than men.

One very interesting point of note is that **people who haven't marked their gender are less likely to book an Airbnb.** Also, people who have marked themselves as 'other' are more likely than any other group to make a booking.

Age



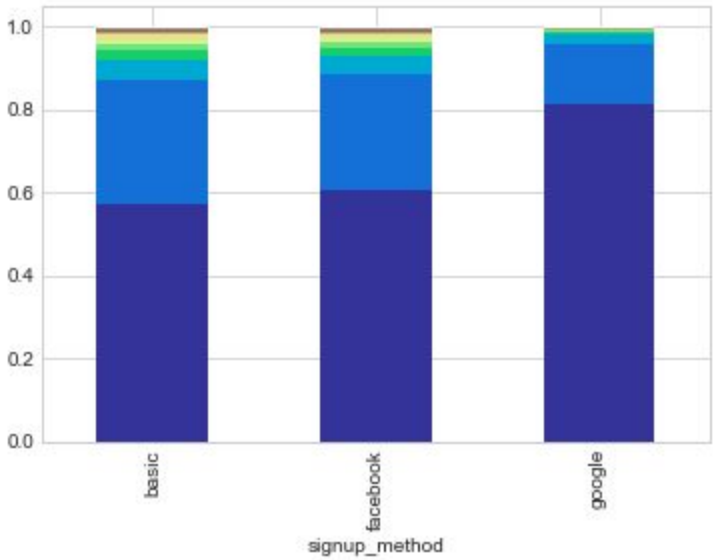




From the boxplot above, we find that the distribution is more or less the same for every country. **Great Britain** has the highest average age of travellers and **Spain** is more popular amongst younger travellers.

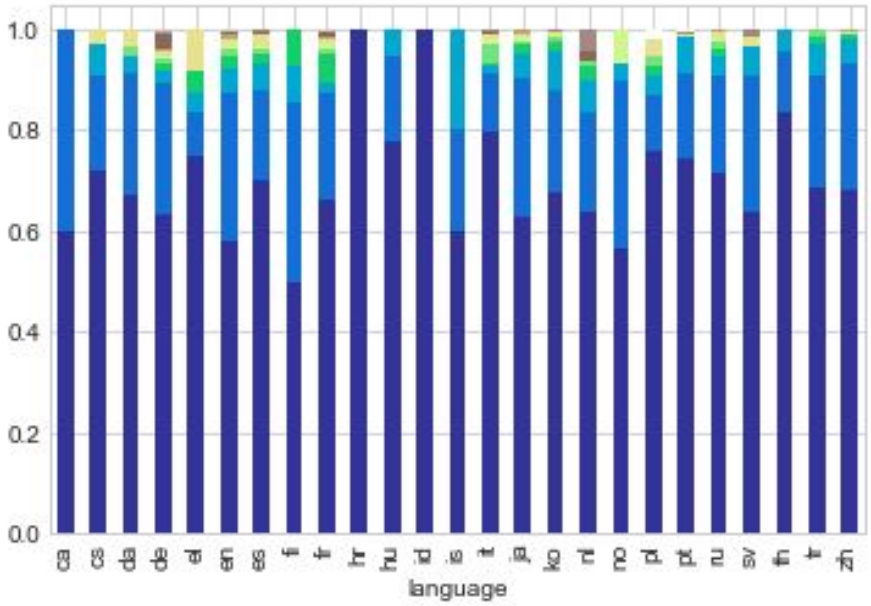
We also discover that people who have not disclosed their ages are least likely to book an Airbnb. Out of the users whose age we know, **Middle Aged People** are most likely to book an Airbnb although it must be noted that there isn't a very significant difference amongst the three groups.

Signup Method



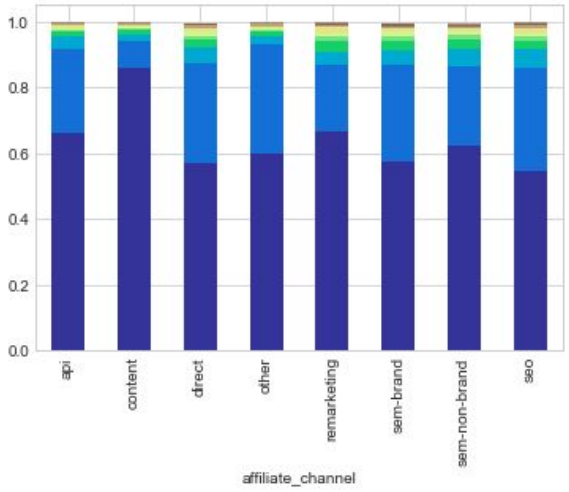
Basic and Facebook are the most popular methods of Signup. People who use the **Basic** Method are the most likely to book an Airbnb whereas people signing up using **Google** are the least.

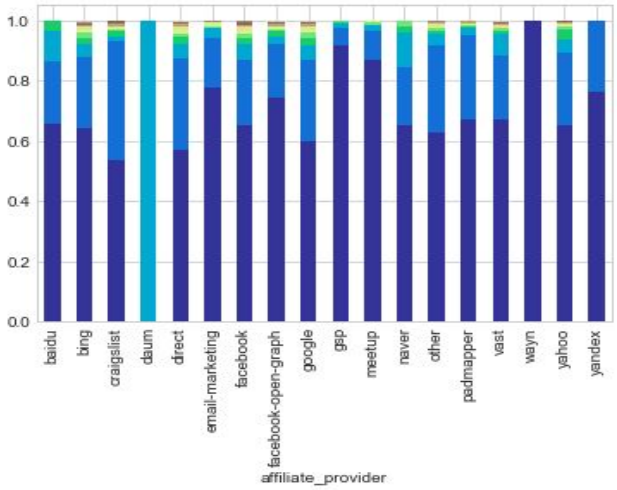
Languages

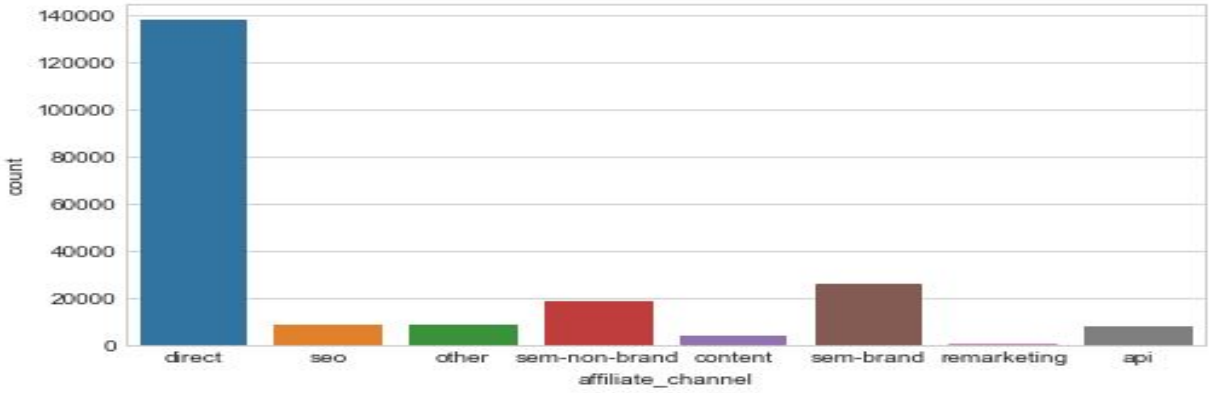


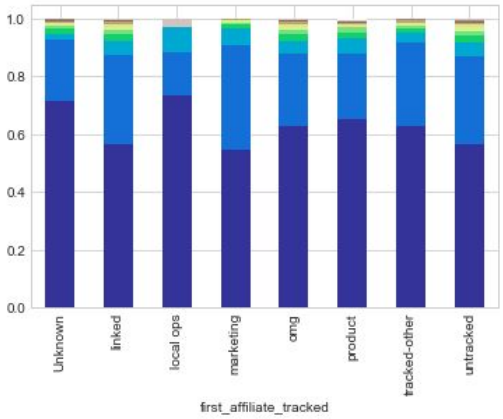
We see that people who speak **Croatian** and **Indonesian** made almost no bookings. People who spoke **Finnish** made the most bookings amongst all languages. The large number of languages is also surprising considering that Americans usually converse and interact with their apps primarily in English.

Affiliate Channels and Providers





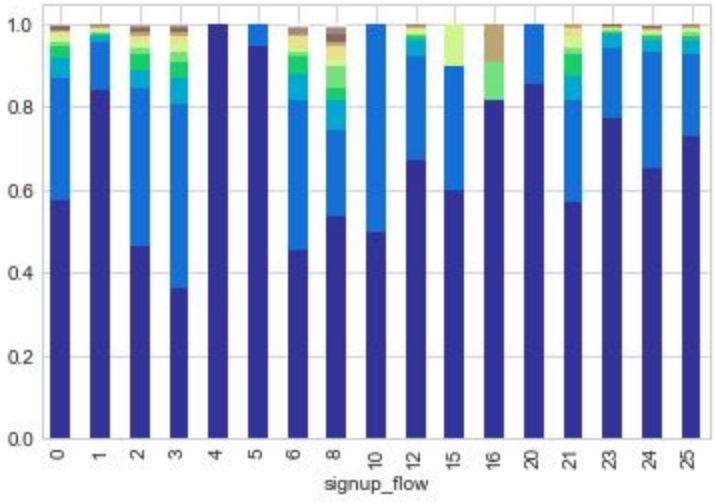




The following observations can be made regarding affiliates:

1. The **Direct Channel** has the most number of conversions to bookings whereas the **Content Channel** has the least.
2. **Direct and Google** are the most popular affiliate providers.
3. **Wayn** has the least percentage of conversions whereas **Daum** has the most. However, we must take this with a pinch of salt as it might be the case that the **number of sample points of these categories are extremely few in number** (as the count plot suggests).
4. Apart from the above, **Google and Craigslist** have a good percentage of conversions.
5. People with **Marketing affiliates** were most likely to book. People whose first affiliate was tracked as **Local Ops or was Unknown** were least likely.

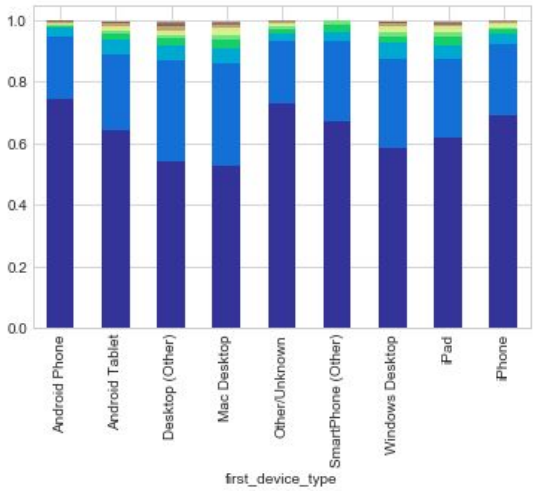
Signup Flow

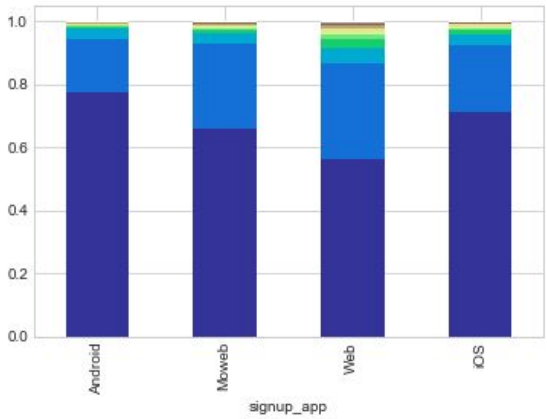


The Signup Flow stacked bar chart raises very interesting observations: people with signup flow 3 are most likely to book an Airbnb. Conversely, people with signup flows 4 and 5 are least likely to convert.

Type of Device, Browser and App Used

As there are too many browsers, we will ignore it for the time being and try to reduce the categories in a later step.

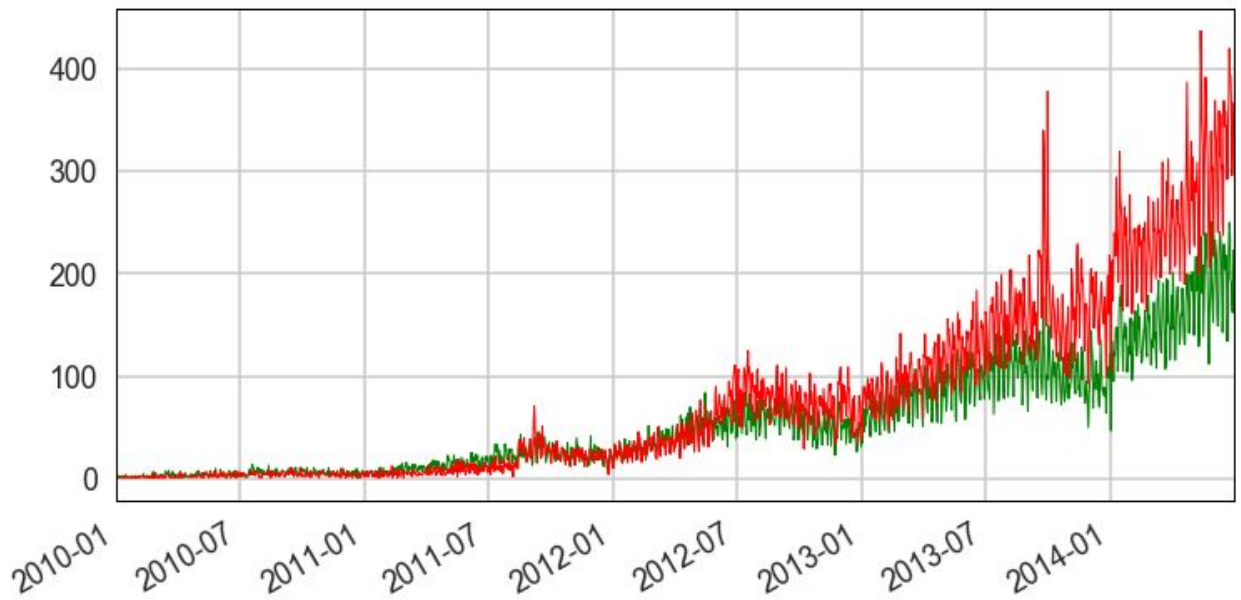




Users using the **Web App** are most likely to book an Airbnb whereas Android Users are least likely to do so. People with an **Android Phone or whose devices were unknown** bought fewer Airbnbs. **People on Desktops** (Mac or otherwise) bought more.

This strongly suggests that **users on their desktop** will be more likely to book an Airbnb and Apple Users are more prone to buying on the website whereas Android Users are the least.

Dates



The number of non booking users have increased more than the number of booking users as time passed by. This can be attributed to the fact that more people are using the platform just for exploration.

Another reason might be that since the user was on the platform for a longer time, s/he was more likely to go ahead and book a space. This directly implies that the earlier the user had created an account on Airbnb, **the more likelier s/he was to actually make a booking.**

FEATURE ENGINEERING

With these insights in our pocket, we now proceed to extract and build features from the data that has been provided to us.

Session Data

First, we will take a look at the session data that gives us information about user activity on the Airbnb Website and App. From what we've learnt we'll construct the following features:

1. **Number of Sessions:** The total number of sessions registered by the user.
2. **Number of types of Sessions:** The distinct types of activities logged by a particular user.
3. **Total Seconds:** The total amount of time spent by the user on Airbnb
4. **Average Seconds:** The average amount of time spent in each session by the user.
5. **Short Sessions:** The number of sessions which were less than 5 minutes long.
6. **Long Sessions:** The number of sessions which were more than 2000 seconds (or 33 minutes and 20 seconds) long.
7. **Number of Devices:** The number of devices used by the user to use the website.

The initial hunch was that a greater number of devices would imply the user is a traveller, thus making him/her more likely to book an Airbnb. The other intuition was that the longer time the user spent on the platform, the more serious they were about booking an Airbnb.

Training Data

The Training dataset contained the bulk of the feature engineering performed to come up with the final training dataset. All the date features were removed as they were not of too much use to us considering that the test dataset begins somewhere during mid 2014. So all our analysis regarding users dating back to 2010 is moot.

From the insights gained in the exploratory data analysis section, the number of categories were reduced for each variable in such a way so as to increase disparity. Continuous variables such as

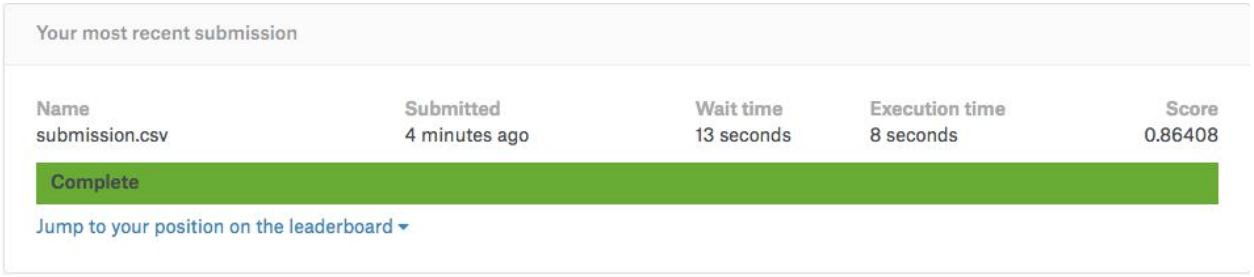
age were binned into groups. Finally, one hot encoding was performed on these categorical variables to arrive at the final training dataset.

MACHINE LEARNING

The next step was to build a classifier to train the data on and then test its performance against the test data. With all the feature engineering already done in the previous step, applying machine learning was a fairly concise step. The Gradient Boosting Classifier was used as the model of choice for this problem.

The training and test data was split in the ratio 3:1. The accuracy achieved was a little over 63.5%

Finally, the submission dataframe was constructed with the top 5 destinations for each users based on the value of their prediction probabilities. This dataset was uploaded to Kaggle and the score achieved was 0.86408.



This score can of course be improved through hyperparameter tuning and more advanced feature engineering but the improvement is extremely minimal (considering that the highest score was 0.88).

CONCLUSION

This report highlighted the processes of data wrangling, inferential statistics, data visualization, feature engineering and predictive modelling performed on the Airbnb Dataset. All the results and insights gained as part of these processes were also highlighted. With these insights, a Gradient Boosting Classifier was built and a score of 0.86408 was achieved on the Kaggle Public Leaderboard.

The code associated with this report is available at: https://github.com/rounakbanik/airbnb