Mehreteab Kidane
Springboard

Milestone Report: Airbnb New User Bookings

Problem Statement

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. Airbnb is the client of this project, the goal is to predict in which country a new user will make his or her first booking so to better forecast demand and therefore increase revenue.

Dataset

The data set is already available to in the form of Kaggle competition by Airbnb in 2015.
https://www.kaggle.com/rounakbanik/airbnb-new-user-bookings/data
The data provided by Airbnb is in the form CSV files and are listed below.

```
train_users.csv - the training set of users
test_users.csv - the test set of users. Contains User information such as gender,
                  age, language, signup and device information
sessions.csv - web sessions log for users. Contains time, type and details of
               various user actions
countries.csv - summary statistics of destination countries in this dataset and
                their locations
age_gender_bkts.csv - summary statistics of users' age group, gender, country of
                      destination
sample_submission.csv - correct format for submitting your predictions
```

Significance of the problem and my approaches

The approach to solving this problem is subject to change as I progress with career track and learn new concepts and approaches. Here are my approaches –

1. Data collecting: downloading all the data provided by Airbnb to local.
2. Data Wrangling: data cleaning, seek mistakes in data, look for peculiar behavior, fix missing data.
3. Data exploration: use of classification, inferential statistics and data visualization to find interesting trends and identify significant features in the data set.
4. Data Analysis: data manipulation and modeling.
5. Complete and submit final deliverables

Deliverables:

The deliverables will be the codes and visualization techniques on GitHub in the form of Jupyter Notebooks, and a slide desk. This will include a report and I intend to write a documentation explaining the code and the results.

Data Wrangling

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 technique was used on this dataset.

Sessions

1. All the unknown fields were converted into NaN to give it more semantic meaning.
2. The missing second value were interpolated using the panda Series interpolate function. This was done as this treatment didn't significantly alter the summary statistics of second elapsed.

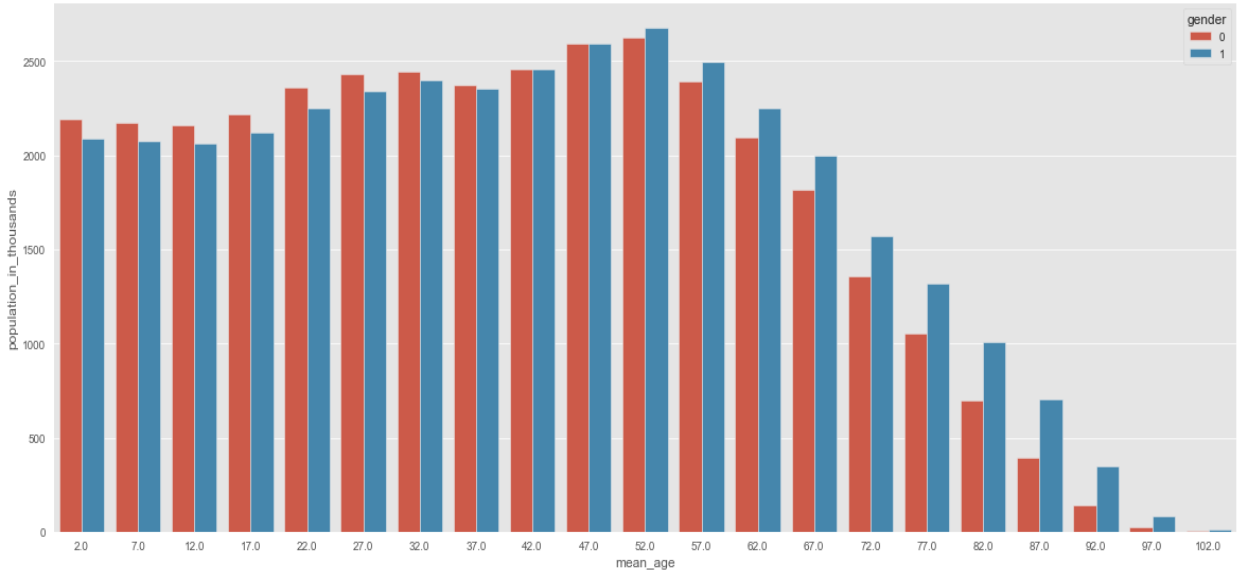
Training Users

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

Exploratory Data Analysis

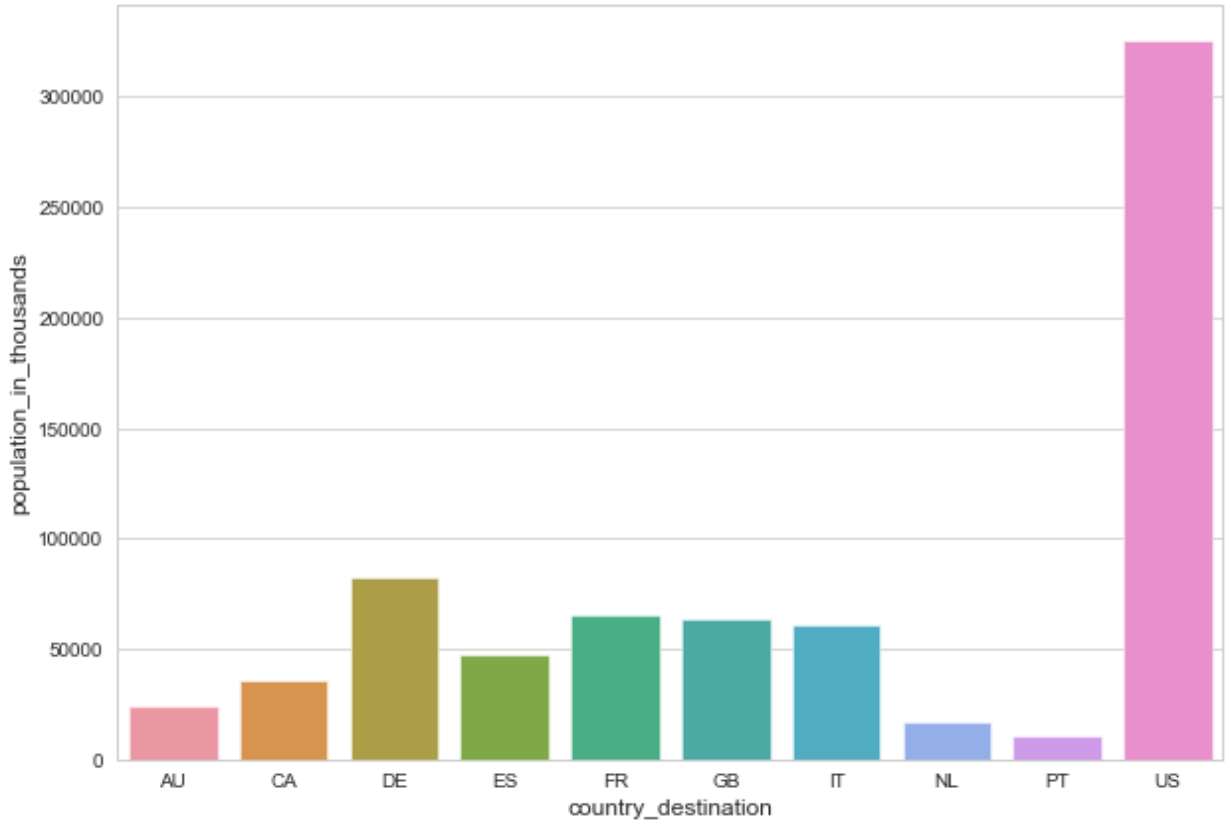
In this section, the various insights produced through descriptive statistics and data visualization is presented.

Country Statistics.



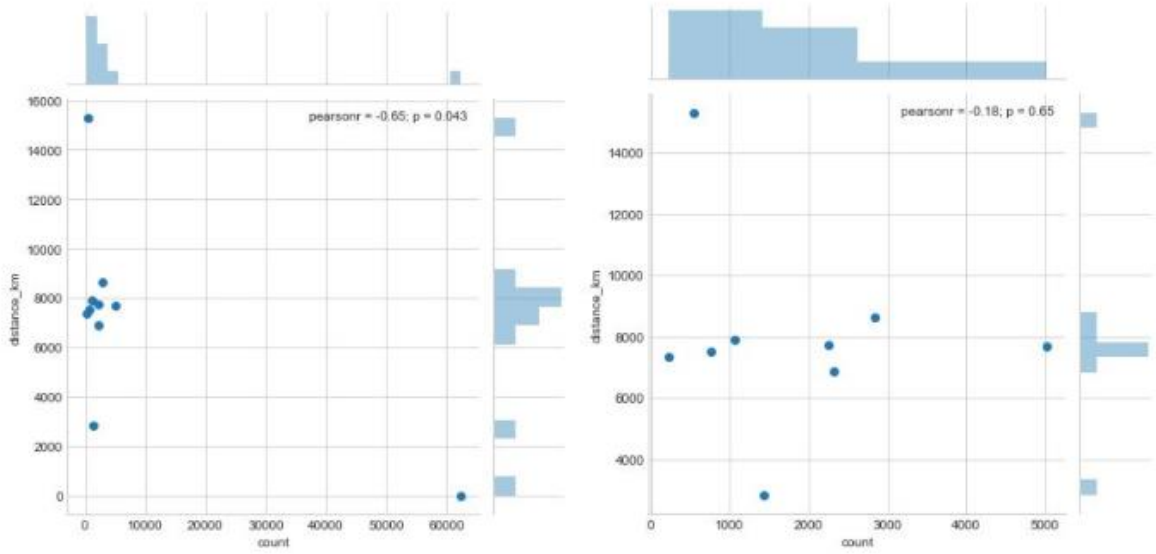
1. The countries that are represented in this dataset largely consist of an aging population. The largest groups are people with **mean ages 47 years and 52 years.**
2. The distribution resembles a skewed bell curve. The middle aged people occupy the largest share of the population, closely followed 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 steady.
4. One very interesting thing to note is that the sex ratio is skewed toward men for younger age groups but as the mean age increase, the radio skews more towards woman.
   Woman indeed lives longer than men.

Next, let's try 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

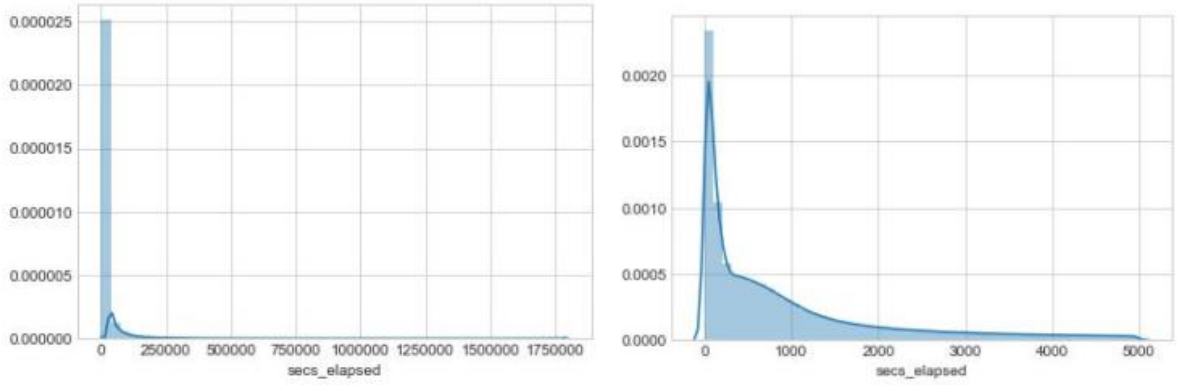


There is a strong negative correlation of -0.65. People overwhelmingly prefer booking in the United States than any other country in the world.

However, when taking only international countries into consideration (i.e except United States), the correlation is much weaker at -0.18.

Session Statistics

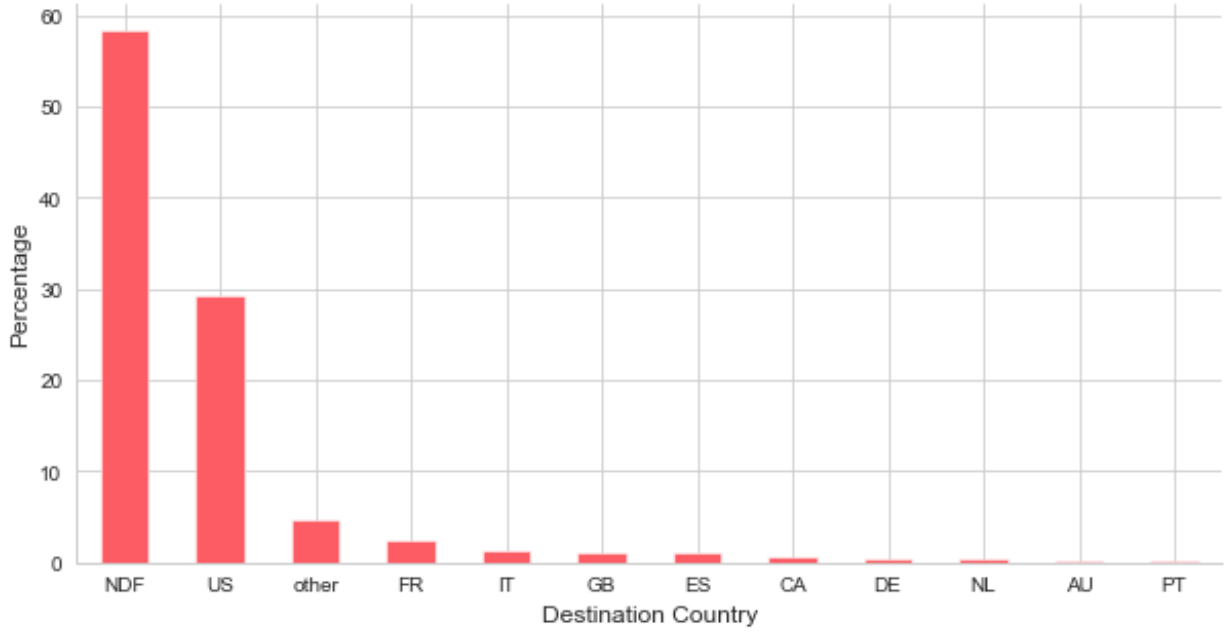
Seconds Elapsed Distribution



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.

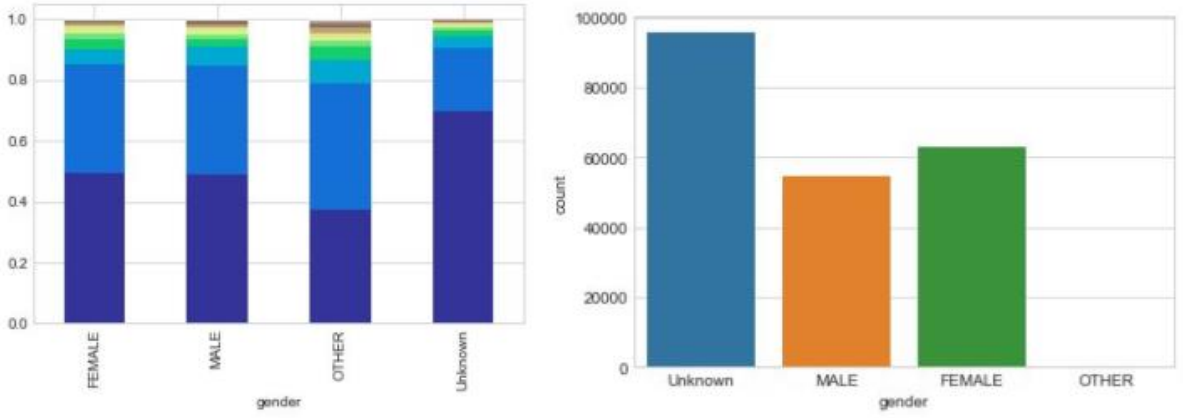
Training Dataset Statistics

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

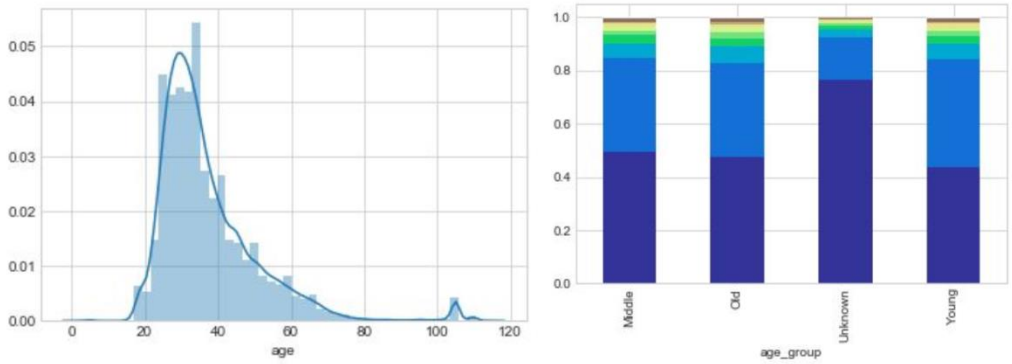


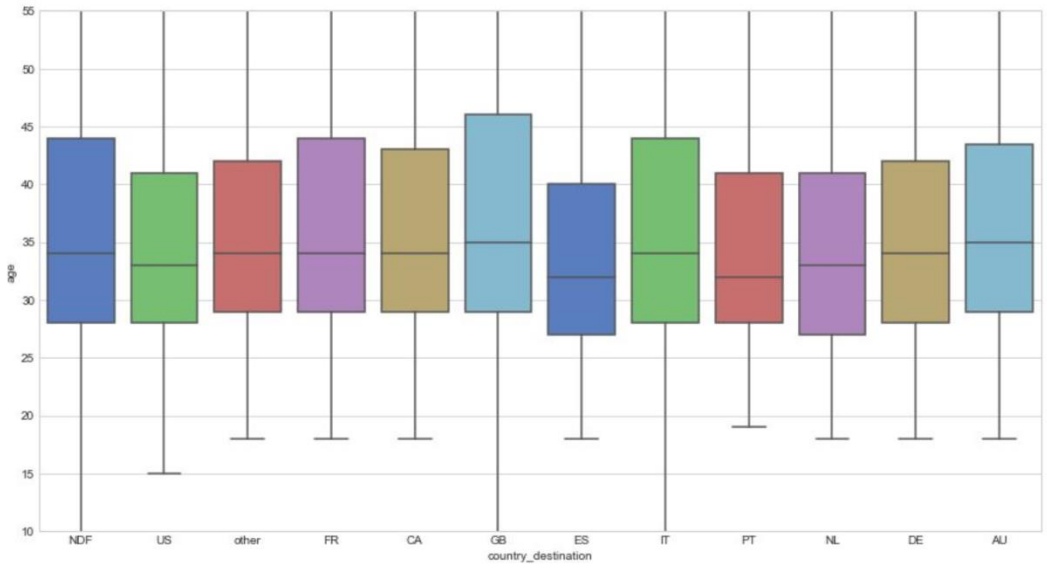
We see that users whose gender is unknown form the majority. Out of the users whose gender is known, there are more females than males. This can suggest two things:

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 one 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 booking.

Age





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

We also discover the 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 amongst the three groups.

Signup Method

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

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. For now, we can make the following observations about User Devices:

- 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.

Inferential Statistics

Gender Preference for Airbnb booking

This test was performed to test if there is a relationship between the gender of a uset and the Airbnb Country destination. In other words, does your gender influence the county you will travel to and an Airbnb in? To perform this analysis, I 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 are comparing two categorical variables, of which one was multivariate, the ideal statistical tool to be used was Chi Square Test for Significance. The data available to us was pivoted into a form usable by Scipy Chi Square Contingence Method.

Out[71]:

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

Results:

1. There is a significant relationship between gender and country destination
2. The p-value obtained was $5.8 * 10^{-28}$

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 likely to sign up through Facebook if you were using a phone? To perform this analysis I only took into consideration the basic and the Facebook signup methods as they made up the bulk of signups. Also I considered two types of devices; computers 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, I had the choice between two statistical tests. The Chi Square Test for Significance and the two sample significance Test. I applied both test and compared the results to reach at my conclusion.

Out[84]:

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

Results:

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 samples 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

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 session 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 minute log
6. Long Sessions: The number of session which were more than 2000 seconds(or 33 minutes and 20 seconds) long
7. Number of Devices: The number of devices used by user to use the website

The initial hunch was that a greater number of devices would imply the user is a traveler, 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 data features were removed as they were not of too much used to us considering that the test dataset begins somewhere during mid 2014. So all our analysis regarding users dating back to 2010 is ignored.

From the insights gained in the EDA section, the numbers of categories were reduced to each 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.

1. The timestamp, data of first booking and date of account created features were dropped as these information provided no value in the test user dataset.
2. People with signup flow 3 were identified with a feature due to the disproportionate number of them booking Airbnb.
3. The language feature was reduced to a binary variable to identify if the language was English or not.
4. Affiliate feature were binned based on the number of samples available from each type.
5. All the minor browsers were clubbed into a single category of other browsers
6. The devices were classified based on Desktop, Laptop and Mobile phones.
7. One hot encoding was performed on al resulting categorical variables.

```
In [111]: def feature_engineering(df):
              df = session_features(df)
              df = df.drop('age', axis=1)
              df = browsers(df)
              df =devices(df)
              df =affiliate_tracked(df)
              df = affiliate_provider(df)
              df = affiliate_channel(df)
              df = languages(df)
              df['is_3'] = df['signup_flow'].apply(lambda x: 1 if x==3 else 0)
              df = first_booking(df)
              df = df.drop('timestamp_first_active', axis=1)
              df = account_created(df)
              df = df.set_index('id')
              df = pd.get_dummies(df, prefix='is')
              return df
```

Machine Learning

The next step is to build a classifier to train our 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 should be fairly concise.

Choosing a Model

Three different classifiers were tested for their performance namely:

1. Logistic Regression
2. Random Forest Classifier
3. Gradient Boosting Classifier

```
In [141]: classifiers = [RandomForestClassifier(verbose=1), LogisticRegression(verbose=1),
                        GradientBoostingClassifier(verbose=True)]

          for classifier in classifiers:
              classifier.fit(train_X, train_y)
              print("Score: " + str(classifier.score(test_X, test_y)))
```

The Logistic Regression Classifier performed the worst with an accuracy of less than 60% whereas the **Gradient Boosting Classifier** performed the best with accuracy close to 64%.

Hyper Parameter tuning

Hyper parameter tuning was performed on the Gradient Classifier using 3-Fold Grid Search Cross Validation. The following were tuned.

1. N Estimators : values of 100 and 200
2. Max Depth; Values of 3 and 5
3. Max Features: Logarithmic and Square Root

```
In [142]: parameters = {
              'n_estimators': [100,200],
              'max_features': ['auto', 'log2'],
              'max_depth': [3,5]
          }
```

```
Out[144]: GridSearchCV(cv=None, error_score='raise',
                estimator=GradientBoostingClassifier(criterion='friedman_mse', init=None,
                       learning_rate=0.1, loss='deviance', max_depth=3,
                       max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_sampl...      subsample=1.0, tol=0.0001, validation_fraction=0.1,
                       verbose=0, warm_start=False),
                fit_params=None, iid=True, n_jobs=1,
                param_grid={'learning_rate': [100, 200], 'max_features': ['auto', 'log2'], 'max_depth': [3, 5]},
                pre_dispatch='2*n_jobs', refit=True, return_train_score=True,
                scoring=None, verbose=100)
```

Conclusion

This notebook demonstrated all the major steps that take place in performing data analysis and predictive modeling in a typical data science problem. The data was wrangled and cleaned. Some inferential statistics to deduce relationships between features. Extensive EDA was performed to gain insights on the data and these insights were used to extract and engineer new features.

Finally, several classifiers were considered and the **Gradient Boosting Classifier** was selected and its parameters were tuned.