

Airbnb New User Bookings

Predicting User Destinations from the Airbnb Dataset

SPRINGBOARD

JULY, 2019

Prepared by:

Mehreteab Kidane

Mentor:

Kenneth Gil-Pasquel

The problem

Company

Airbnb is an online marketplace and hospitality service that enables people to lease or rent short-term lodging including vacation rentals, apartments, homestays, hostels, or hotel rooms,

Goals

Share more personalized content with community

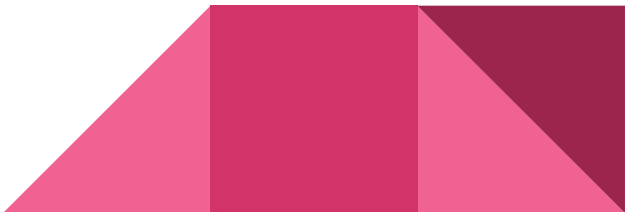
Decrease the average time of booking

Better forecast demand

Problem statement

Given data on the user and their sessions, predict which country the user will book his/her first Airbnb in.

Data Wrangling

1. The data provided to us by Airbnb was already relatively clean and required very little wrangling.
 2. The missing values of seconds elapsed in the sessions dataset was interpolated using Pandas.
 3. People with ages > 125 were given an age of 'Unknown'.
 4. All missing and unknown values were initially converted to NaN to give it more semantic meaning.
- 

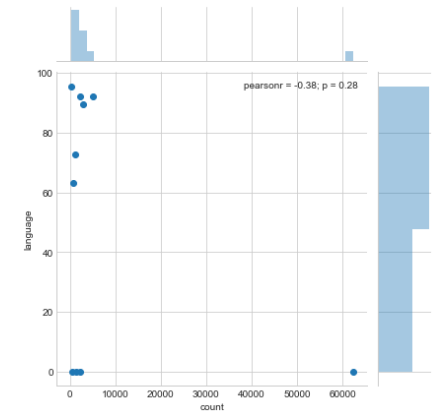
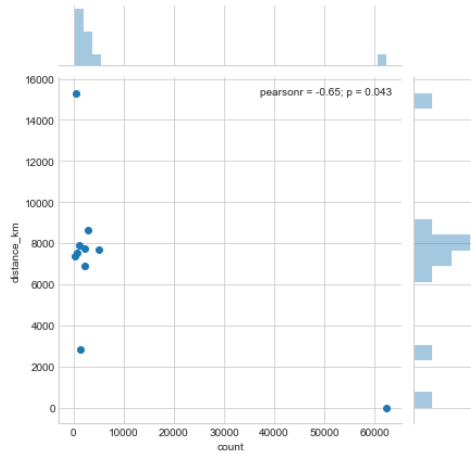
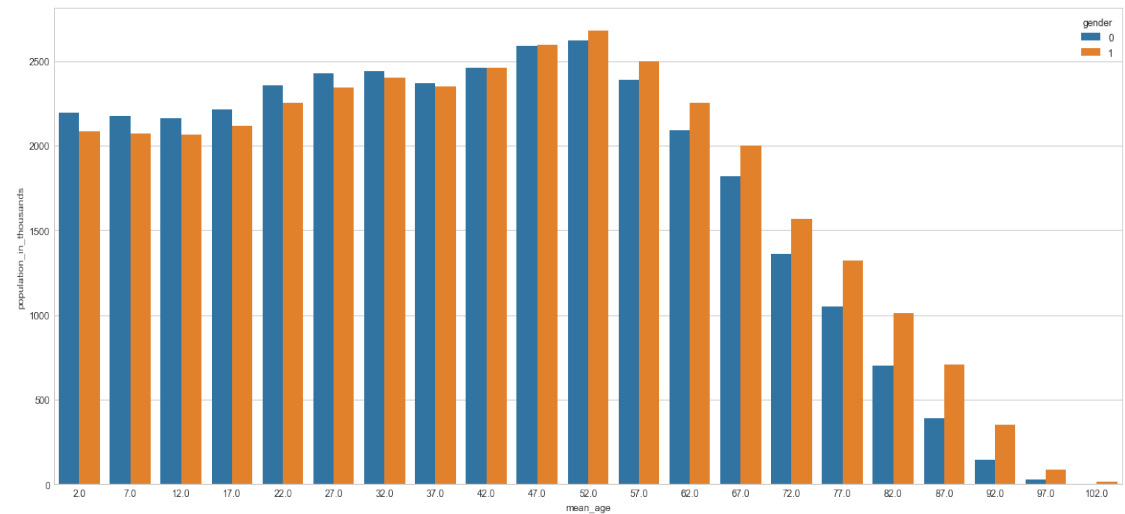
Inferential Statistics

The Chi Square Test for Independence and the Two Sample Significance Test were used to gain more insight of the data via inferential statistics. The following results were obtained:

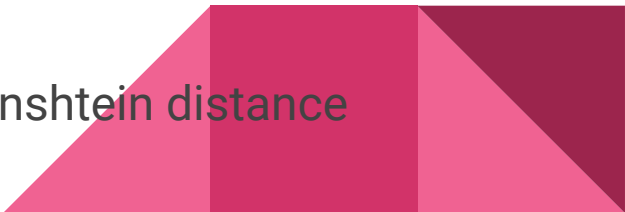
1. There is a gender based preference for countries. In other words, the gender of a person influences the destination of choice.
2. There is no relationship between the device used and the method of signup. The two quantities were found to be independent of each other.



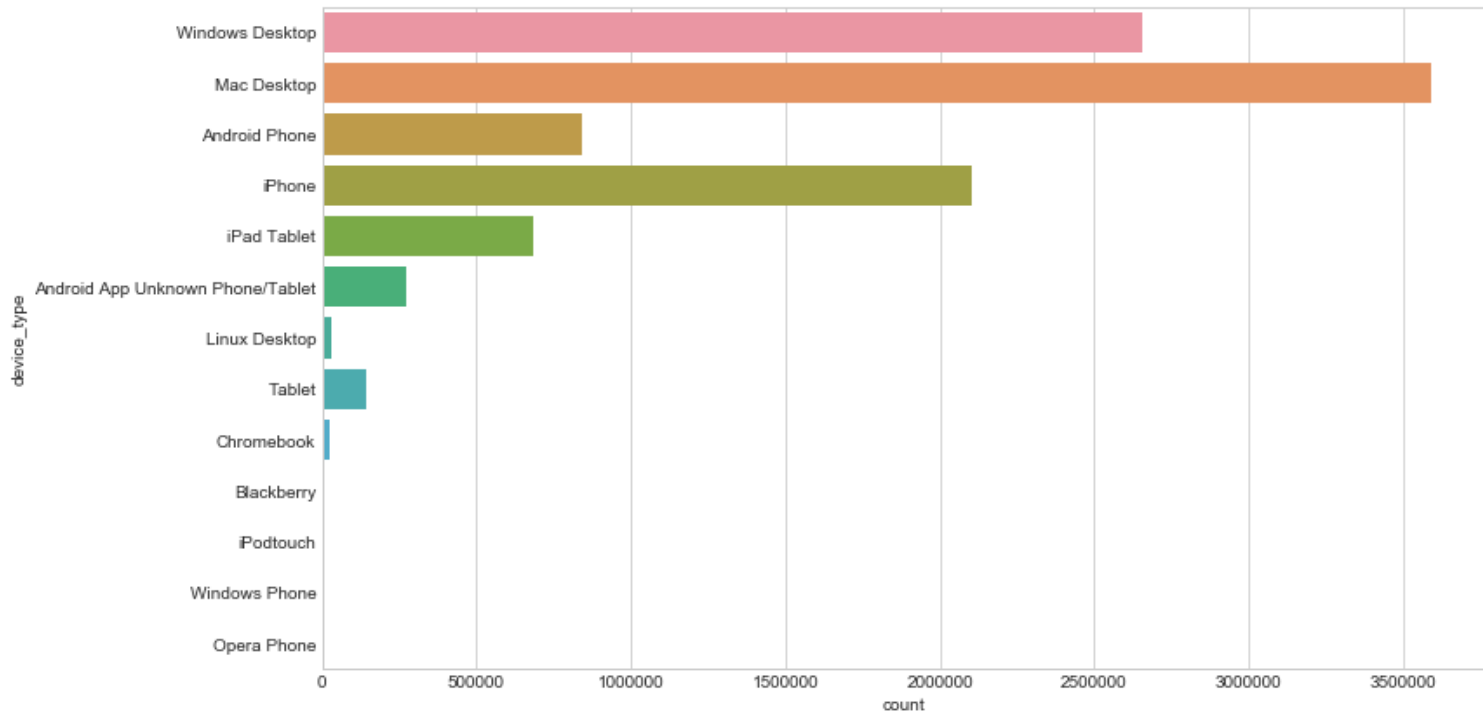
Data Visualization and Analysis



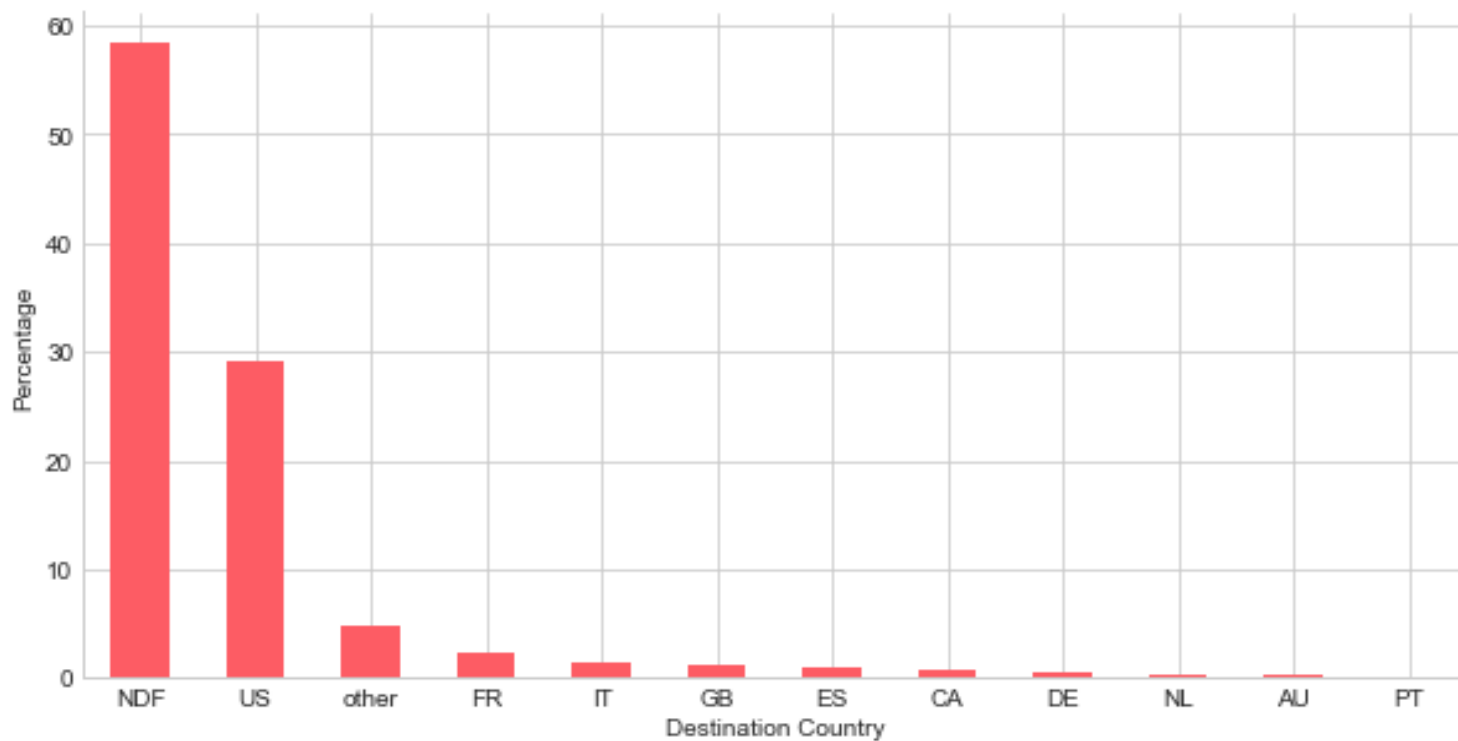
Country Statistics

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. 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.
 3. There is a negative correlation between country preference and country distance.
 4. The correlation is more complicated for language levenshtein distance (Negative including US but positive excluding it).
- 

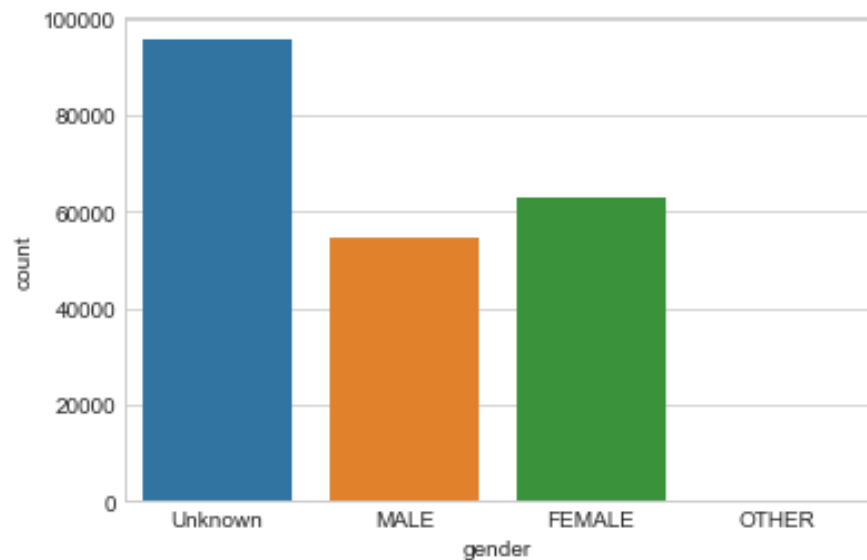
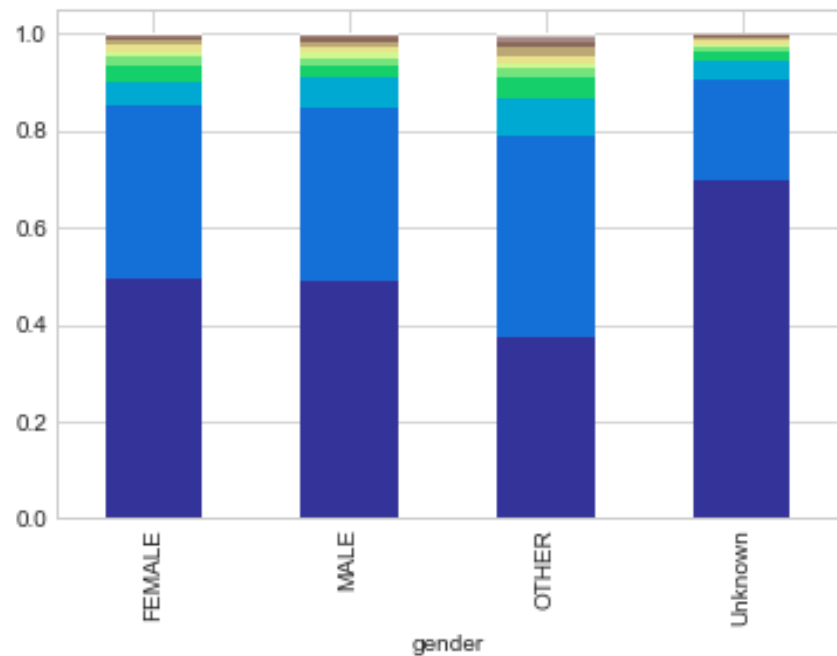
Device Usage during Sessions



Destination Popularities



Gender

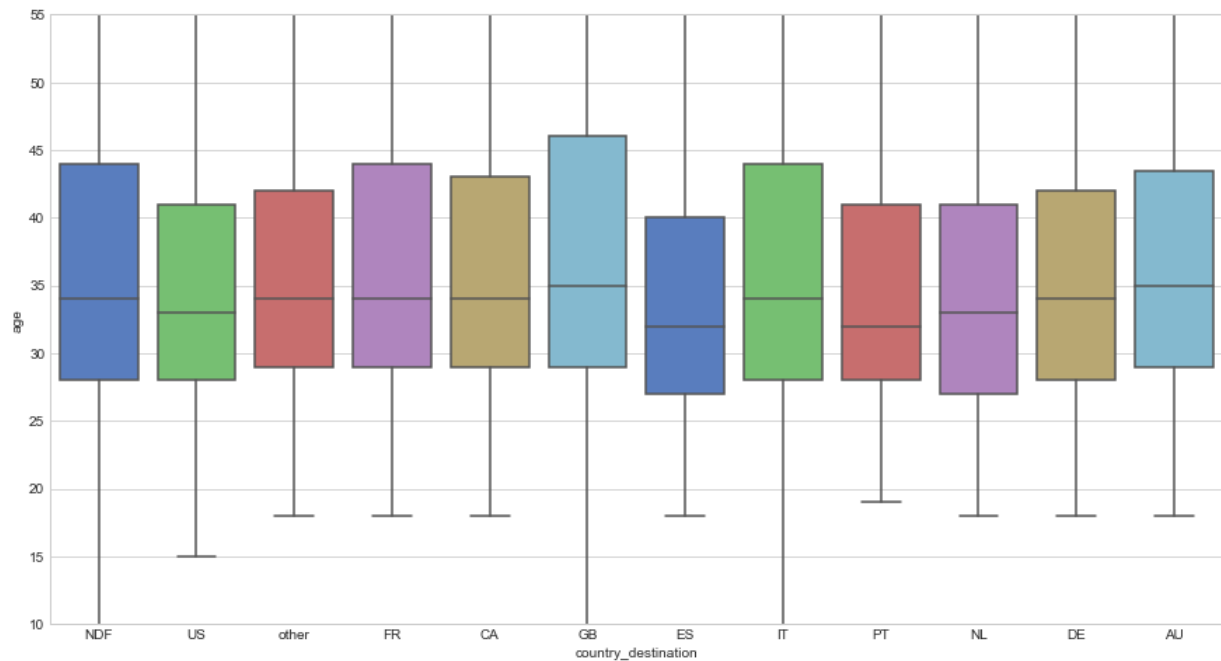
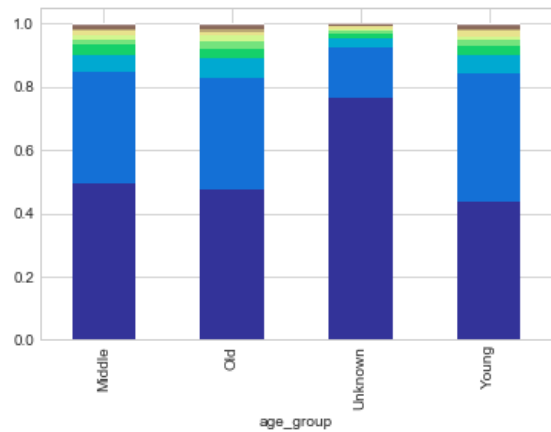


Gender Insights

1. People who haven't marked their gender are less likely to book an Airbnb.
2. People who have marked themselves as 'other' are more likely than any other group to make a booking.
3. There are more identified female users than male users.



Age

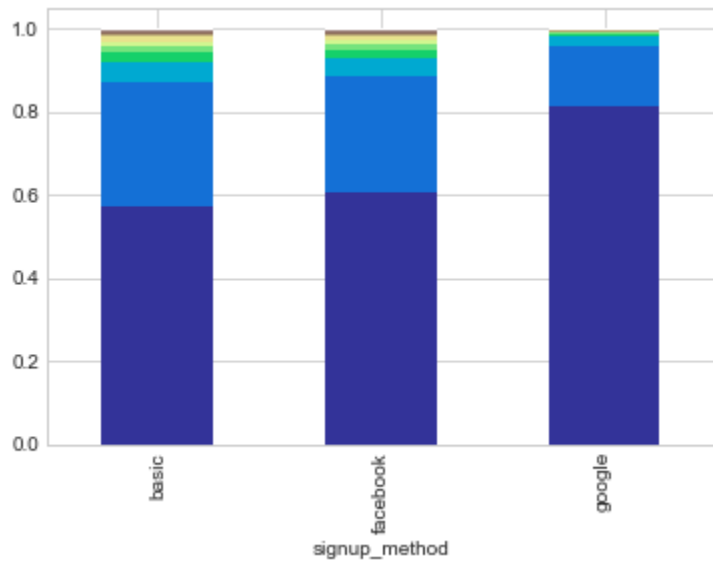


Age Insights

1. Great Britain has the highest average age of travellers.
2. Spain is more popular amongst younger travellers.
3. People who have not disclosed their ages are least likely to book an Airbnb.
4. 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.

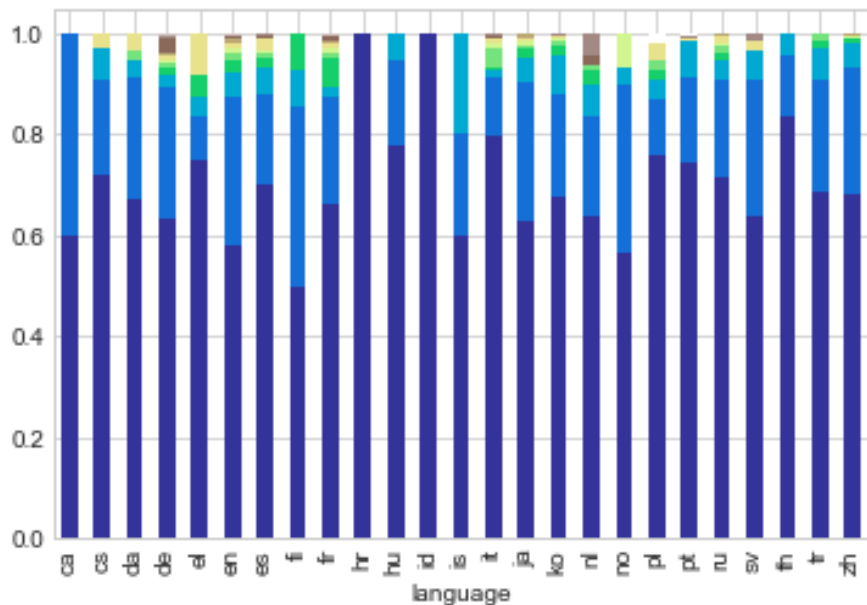


Signup Method



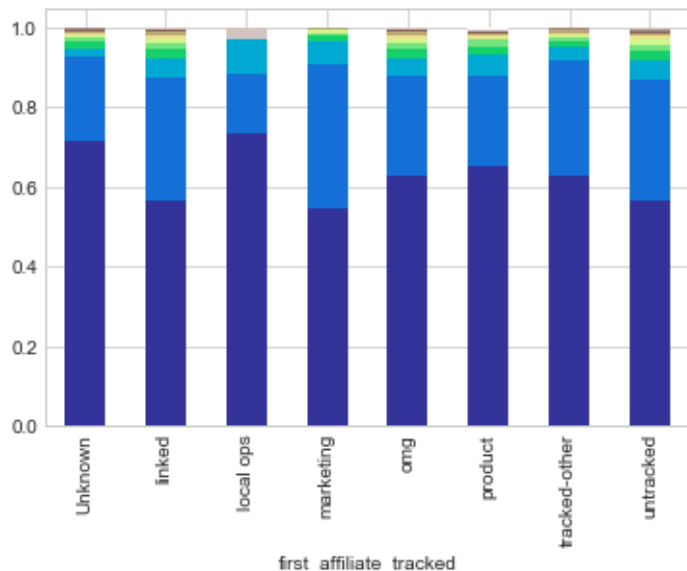
1. Basic and Facebook are the most popular methods of Signup.
2. People who use the Basic Method are the most likely to book an Airbnb.
3. People signing up using Google are the least.

Languages



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

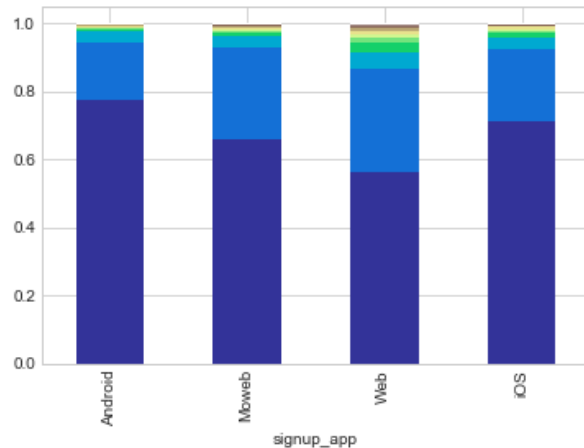
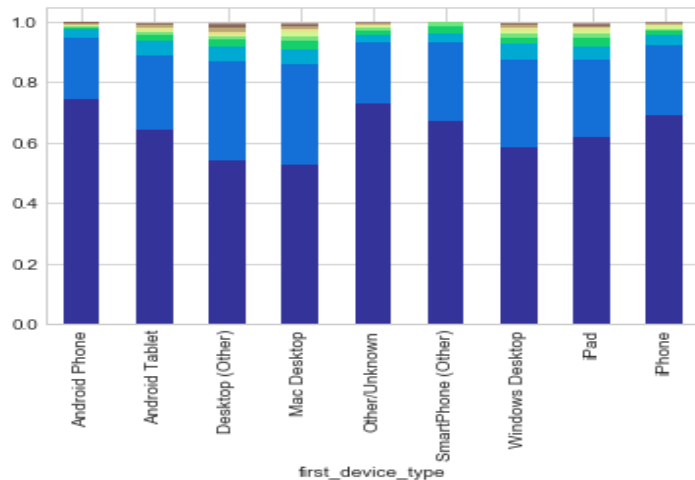
Affiliate Channels



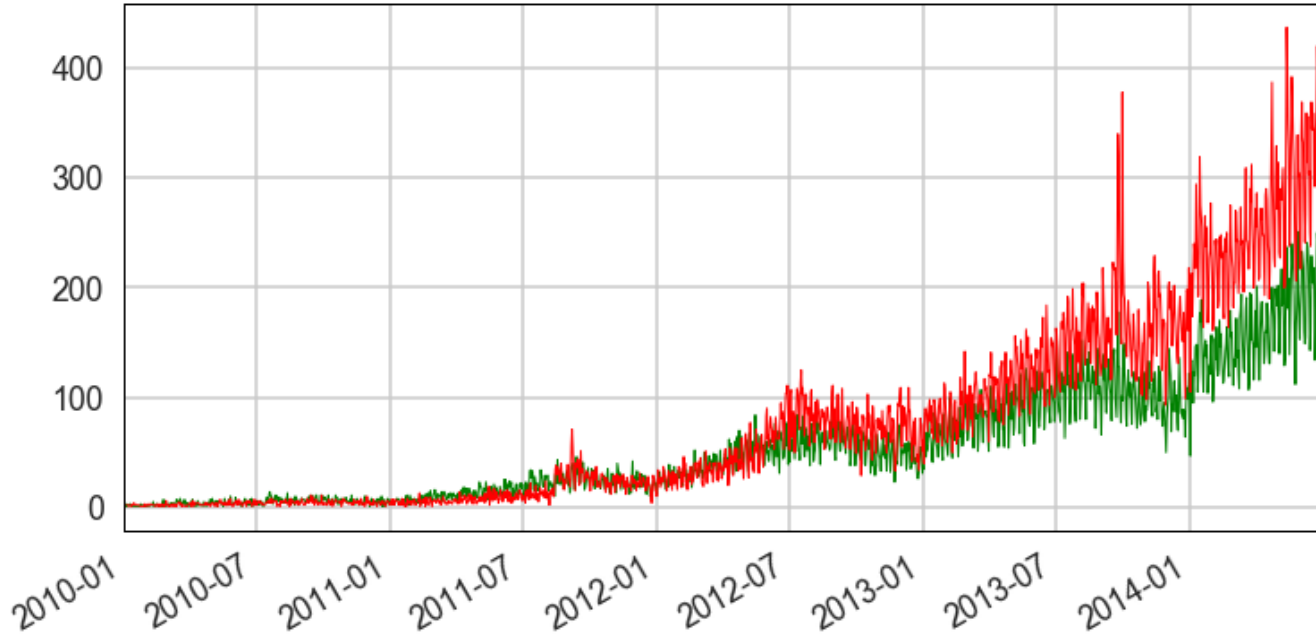
1. The Direct Channel has the most number of conversions to bookings whereas the Content Channel has the least.
2. Direct and Google are the most popular affiliate providers.
3. Wayn has the least percentage of conversions whereas Daum has the most.
4. Apart from the above, Google and Craigslist have 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.

Devices

1. Users using the Web App are most likely to book an Airbnb whereas Android Users are least likely to do so.
2. People with an Android Phone or whose devices were unknown bought fewer Airbnbs. People on Desktops (Mac or otherwise) bought more.
3. 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.



Trends of users booking Airbnbs

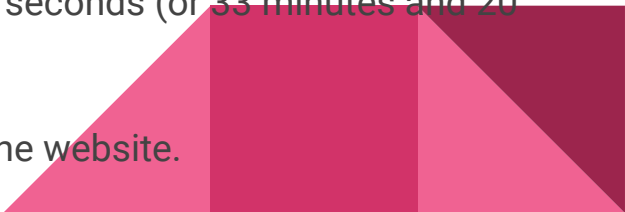




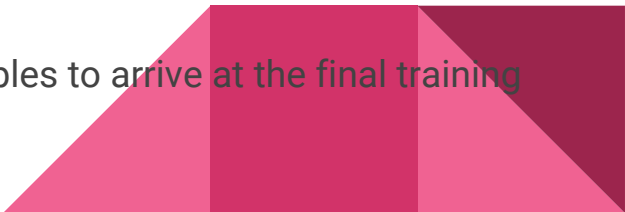
Feature Engineering and ML

Session Data

The following features were constructed from the session data:

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.
- 

Training Data

1. The Training dataset contained the bulk of the feature engineering performed to come up with the final training dataset.
 2. 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.
 3. 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.
 4. Continuous variables such as age were binned into groups.
 5. Finally, one hot encoding was performed on these categorical variables to arrive at the final training dataset.
- 

Model Choice

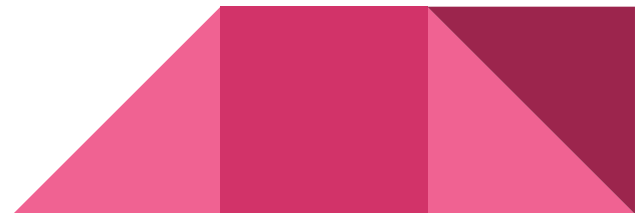
1. 3 Models were tested for their accuracy: Logistic Regression, Random Forest Classifier and Gradient Boosting Classifier.
2. The Gradient Boosting classifier had the highest accuracy of close to 64% whereas Logistic Regression performed the worst with an accuracy less than 60%.



Hyperparameter Tuning

The Gradient Boosting Classifier was tuned for the following parameters:

Feature	Range	Best Value
n_estimators	100,200	200
max_depth	3,5	3
max_features	auto, log2	auto



Conclusions

To summarize the following steps were performed to arrive at the final result:

1. Data Wrangling
2. Inferential Statistics
3. Exploratory Data Analysis
4. Feature Engineering
5. Machine Learning

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

