Airbnb Milestone Report

Vikram Lucky
Dr.vlucky@gmail.com

26 Dec 2018

PROBLEM STATEMENT

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 information, country information, user browsing session records etc. The project aims at coming up with a machine learning model that could accurately predict in which country a user would book his/her first travel experience in. So, that more personalized content can be recommended, and overall booking process can be optimized.

THE CLIENT

The client in question is Airbnb: A world's biggest accommodation-sharing site, enabling people to short-term lodging including vacation rentals, apartment rentals, homestays, hostels etc, usually more affordable alternative to standard hotel bookings. Guests get good value accommodation at a huge range of price points, from a few dollars a night to hundreds and often in prime locations where a normal hotel would cost infinitely more. There is more than a financial impetus though: many guests like living like a local and getting restaurant recommendations from people really in-the-know, while many hosts just enjoy meeting new people and showing off their home.

THE DATA

The data is obtained from Kaggle in the form of competition hosted by Airbnb in 2015. The data is composed of 5 .csv files (including 1 test file for competition submission), and it needs to be cleaned and wrangled before we analysis it.

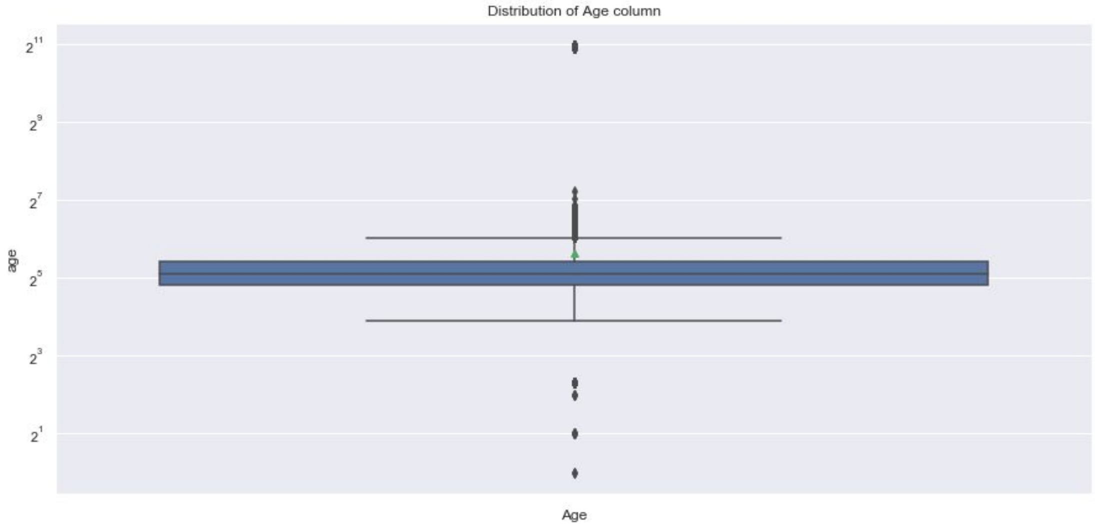
DATA WRANGLING

Training data: Training dataset contains 213,451 rows and 16 columns.
Column name and information about data it contains:

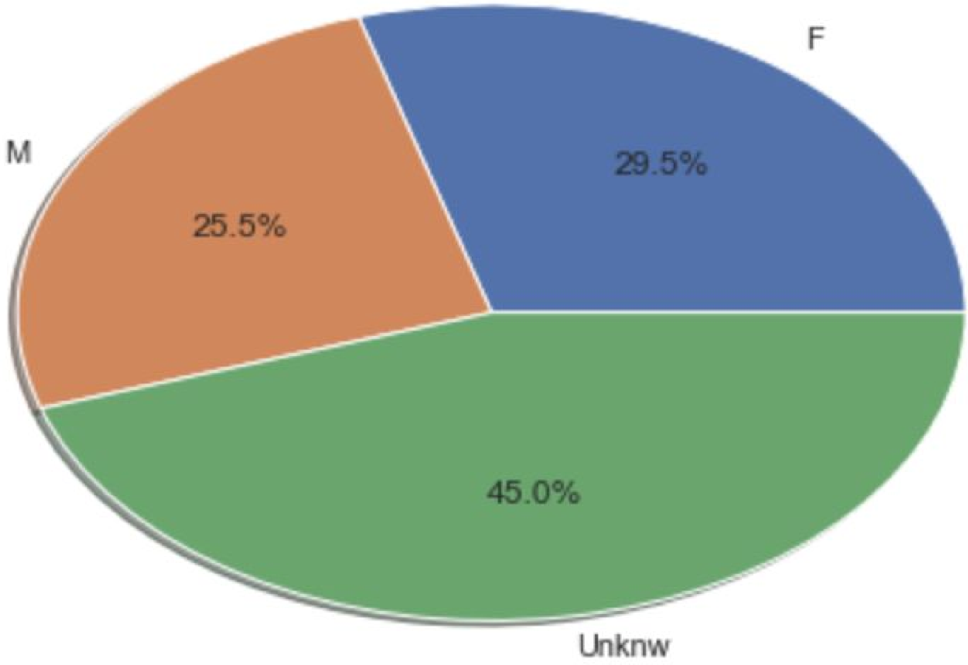
1. Id: user id, unique for each user.
2. date_account_created: The date of account creation.
3. timestamp_first_active: Timestamp of the first activity, it can be earlier than date_account_created or date_first_booking as user can search before signing up.
4. date_first_booking: Date of first booking on Airbnb
5. gender
6. age
7. signup_method: How user signed up for example: using basic method, facebook or google
8. Signup_flow: The page a user came to signup from
9. Language: International language preference
10. affiliate_channel: Kind of paid marketing
11. affiliate_provider: Where the marketing
12. first_affiliate_tracked: What is the first marketing the user interacted with
13. singup_app
14. first_device_type
15. first_browser
16. country_destination: Target variable

Most of the columns in training dataset contained missing and unknown values.

- Columns such as date_account_created and timestamp_first_active were not in their correct dtypes. These columns were converted to datetime object.



- All the unknown and irregular fields *mostly in age column shown in above boxplot and gender columns* were converted into np.nan to give it more semantic meaning, later columns with large number of nan values >58% such as *Date first booking* were completely dropped, and the rest were imputed with sensible data relevant to the column.



Sessions data: sessions dataset contains 1053324 rows and 6 columns.

- This file contains information about web sessions log for users
- This file contains many np.nan values, but as this is not our training dataset, we left this out for now until we get to Machine learning part, if required we might extract some information from this dataset.

Age gender bucket: This dataset has 420 rows and 5 columns

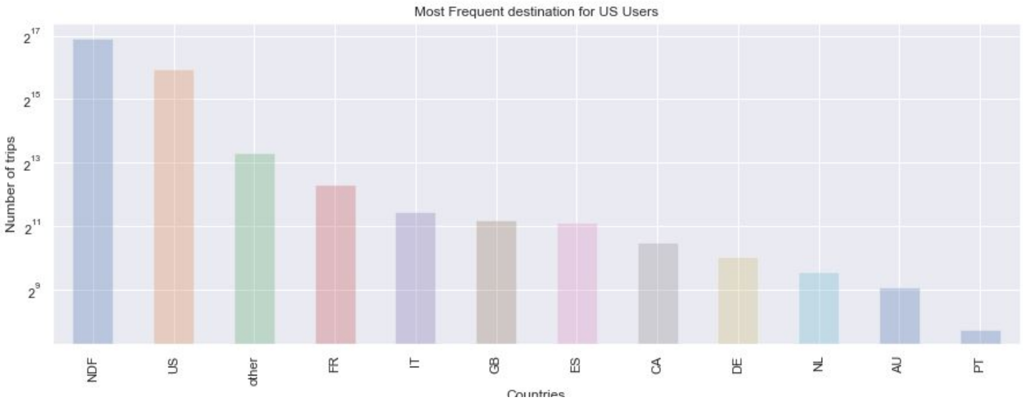
- This dataset contains summary statistics of users age group, gender and country destination
- Contains no missing values

Countries: This dataset has 10 rows and 7 columns

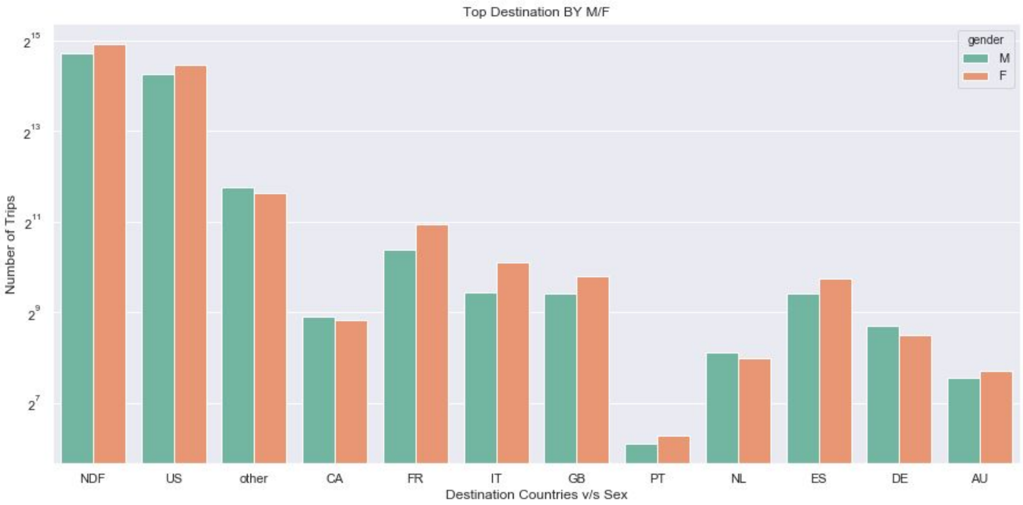
- This dataset contains summary statistics of destination countries
- Contains no missing values

EXPLORATORY DATA ANALYSIS AND VISUALIZATION

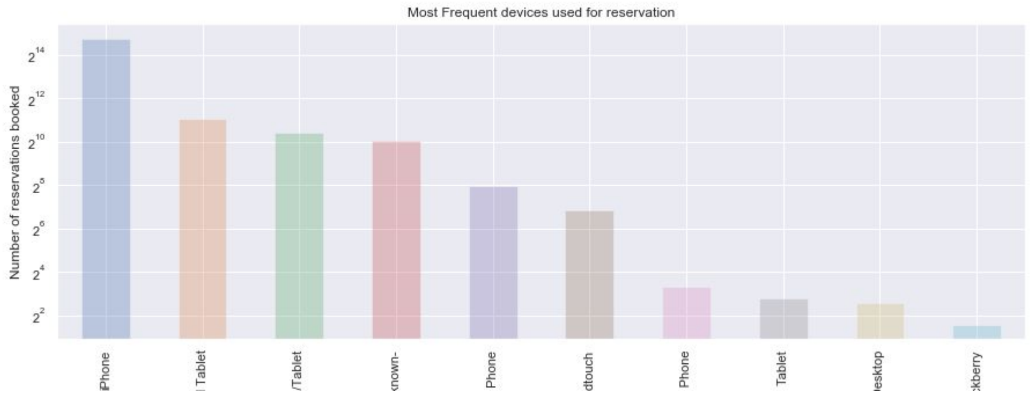
Some insights into our data:



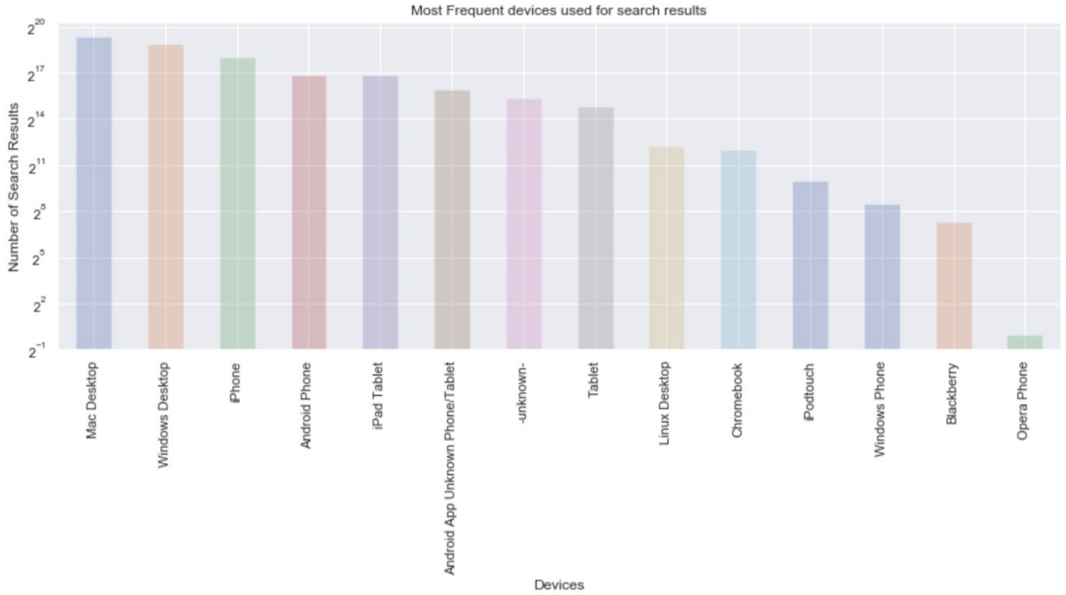
As shown in the Bar chart above, most of the user ended their search in NDF (mean no trips were booked), followed by US, and other (which represents trips were booked by not in our 12 category). Almost 60% of the users did not book their trips.



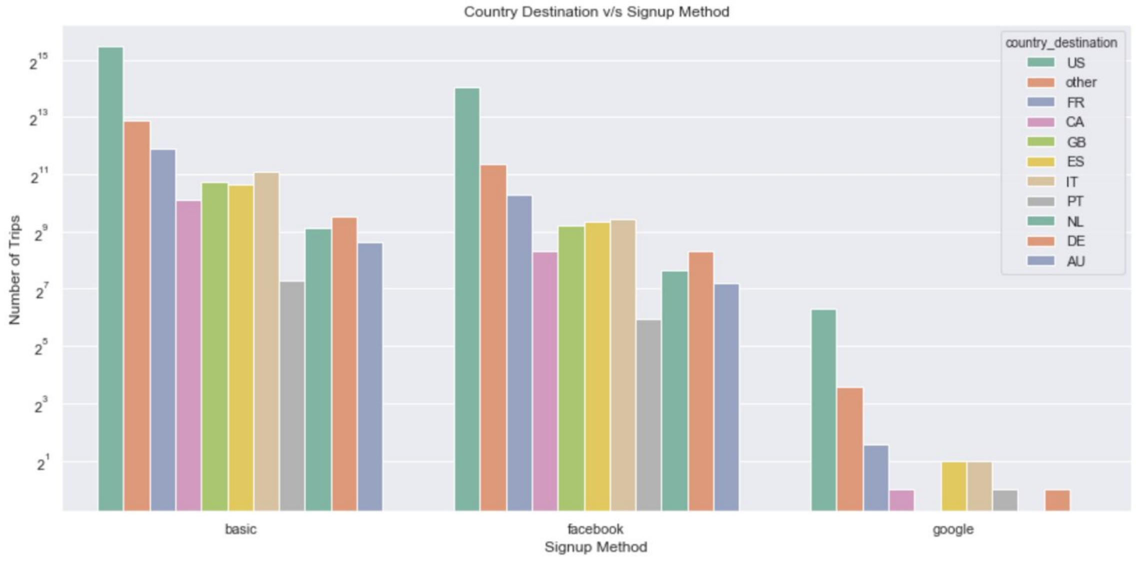
Graph above represents trip booked by each gender.



Above Bar chart represents number of most frequent devices used for **Reservation**, We clearly see Apple products come on top of the list for reservation, surprisingly Mac Desktop neither Windows Desktop made to the top of the list. But how search results. Below we will see what devices are used for search results (to browse destinations.)



Looks like users prefer to search and browse destination on bigger screens, As shown in the above Bar graph Mac Desktop and Windows Desktop are on top of the list followed by Iphone and Android phone. Another interesting fact according to above to bar graph, Android users who browse and search using their phones are less like to lock the deals, whereas IOS user are more likely to book their trips.



Above Chart represents the connection between Signup method used and destinations. Looks People who signed up using basic method and using Facebook, have slightly higher chances of booking a trip with Airbnb compare to users who signed up using Google.

CONCLUSION

This report highlighted the initial business problem, introduced us to the Dataset, showed us some data wrangling techniques and some visualizations to find insights in data. With these insights it is now possible to jump into feature engineering and machine learning.