***Milestone report***

***BUSINESS REQUEST/GOALS:***

* Airbnb has a wide range of options to choose from 34000+ cities around 190 countries for lodging, primarily homestay or tourism experience.
* Predicting where the new user will book their first travel experience has a great value.
* Having such insights or information can help Airbnb share more personalised content with the community, decrease the average time for first booking, understand how a user engages with the service, what factors would encourage them to engage more deeply and better forecast demand and many more.

***WHO CARES ABOUT THIS ?***

* Airbnb is keen on knowing where the new user will book their first travel experience.
* As a new user getting a personalised treatment is of great value.

***DATA COLLECTION AND WRANGLING:***

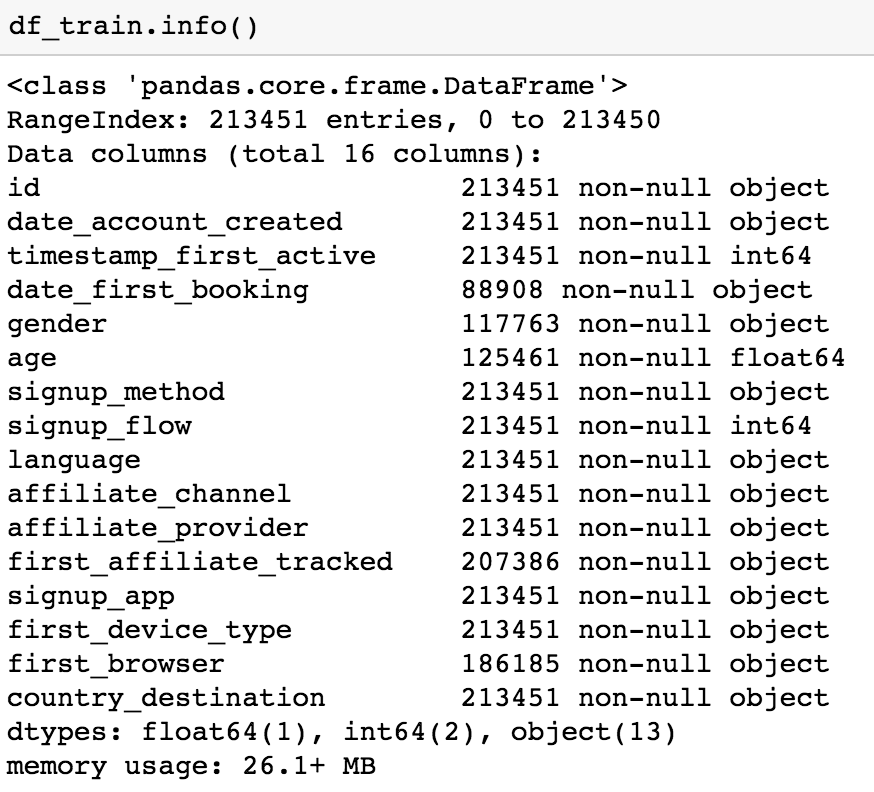
* The data is collected from Kaggle. Ref [Data](https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data)
* Data mainly comprises demographics information, web session records of the user and some summary statistics.
* Most of the data is clean.
* 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' are possible destination countries(target variable)
* Timings are transformed to datetime formats.
* Missing values are transformed to np.NAN.
* Some outliers were observed, like in user age which were replaced by mean age.

***EXPLORATORY DATA ANALYSIS SUMMARY***

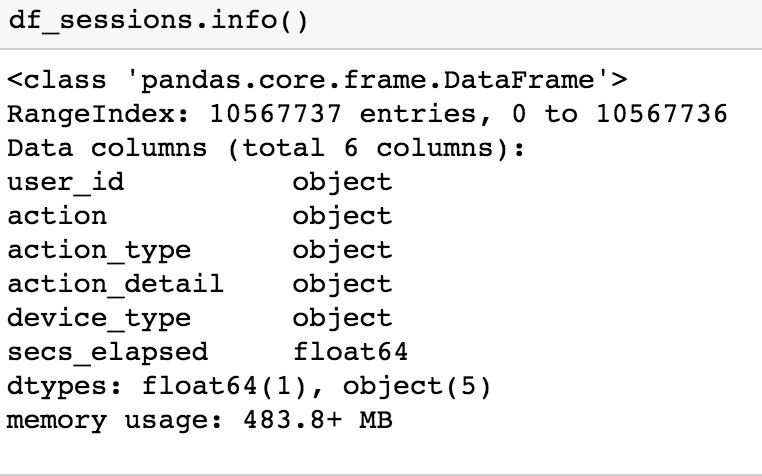
Ref script: [EDA](https://github.com/Anandpatil412/DSC/blob/master/CapstoneProject2/DataCleaning/DataCleaning.ipynb)

* We majorly have the following datasets.
  + Train,sessions and countries.
* Quick glance of data

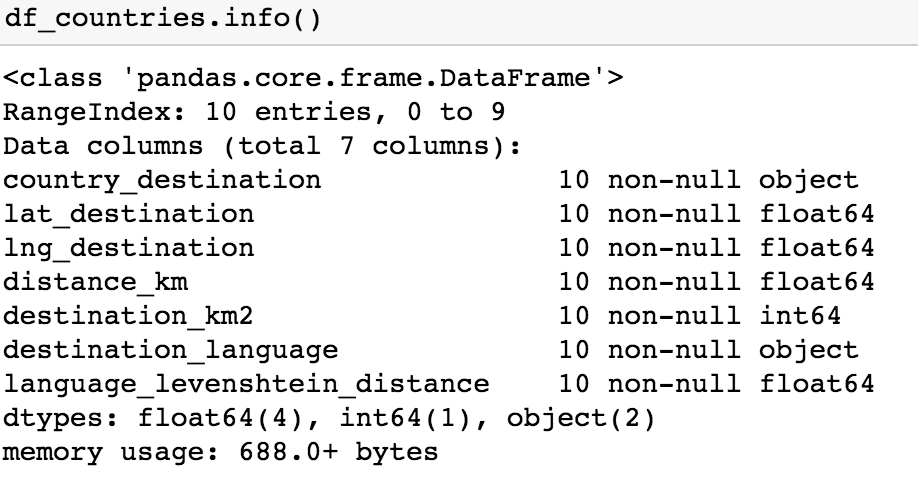
**Train data**



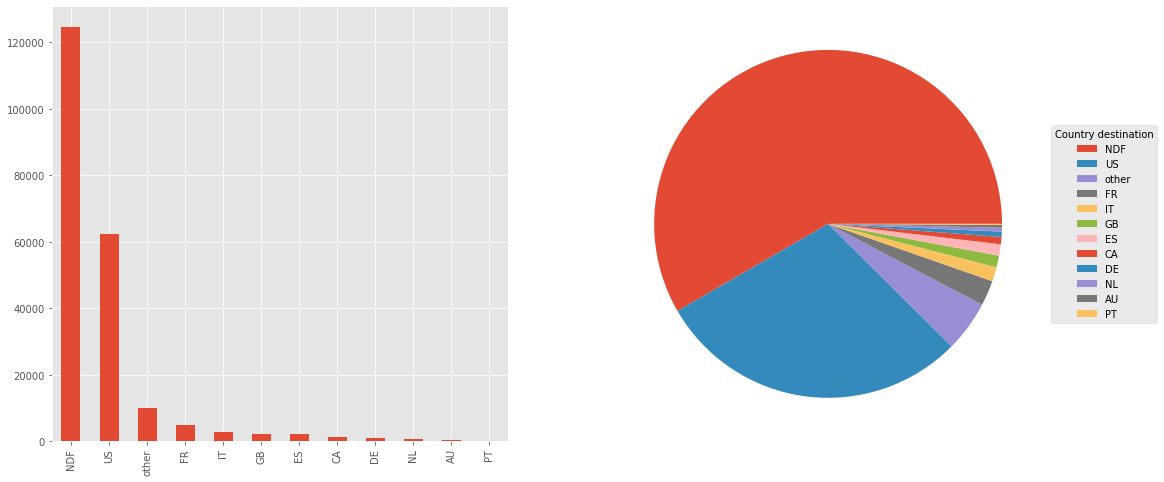
**Session data**



**Countries data**



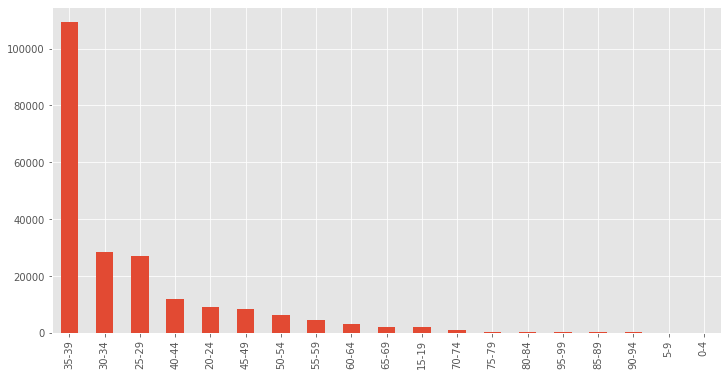
* ***Distribution of destination countries:***



Most of the users land up doing no bookings.

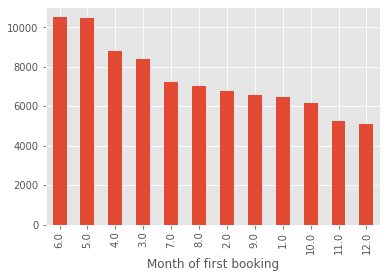
US is the destination country for most of the users, could be because all user data are from people of US which also implies that most users do bookings within the country.

* ***Age group with max bookings:***



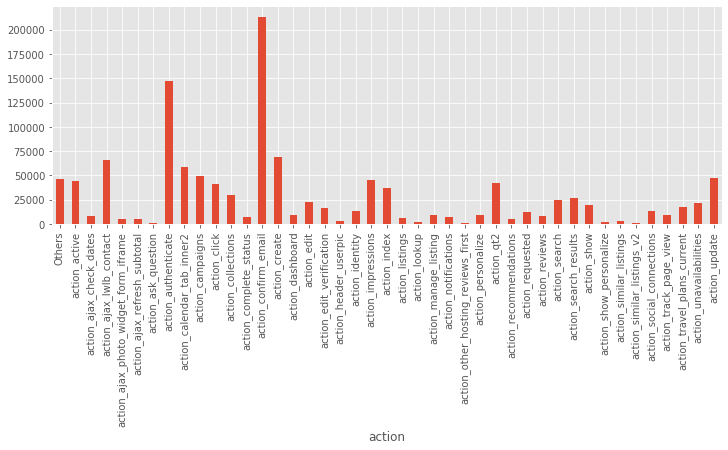
Most users belong to the age bucket 35-39. Also,there is a lot of variance in booking count as age bucket varies.

* ***Age group with max bookings:***



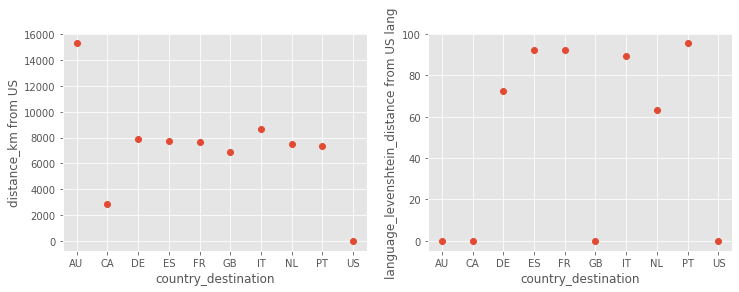
Mid year(ie May, June) seems to have relatively higher first time bookings.

* ***User session action having highest time elapsed:***

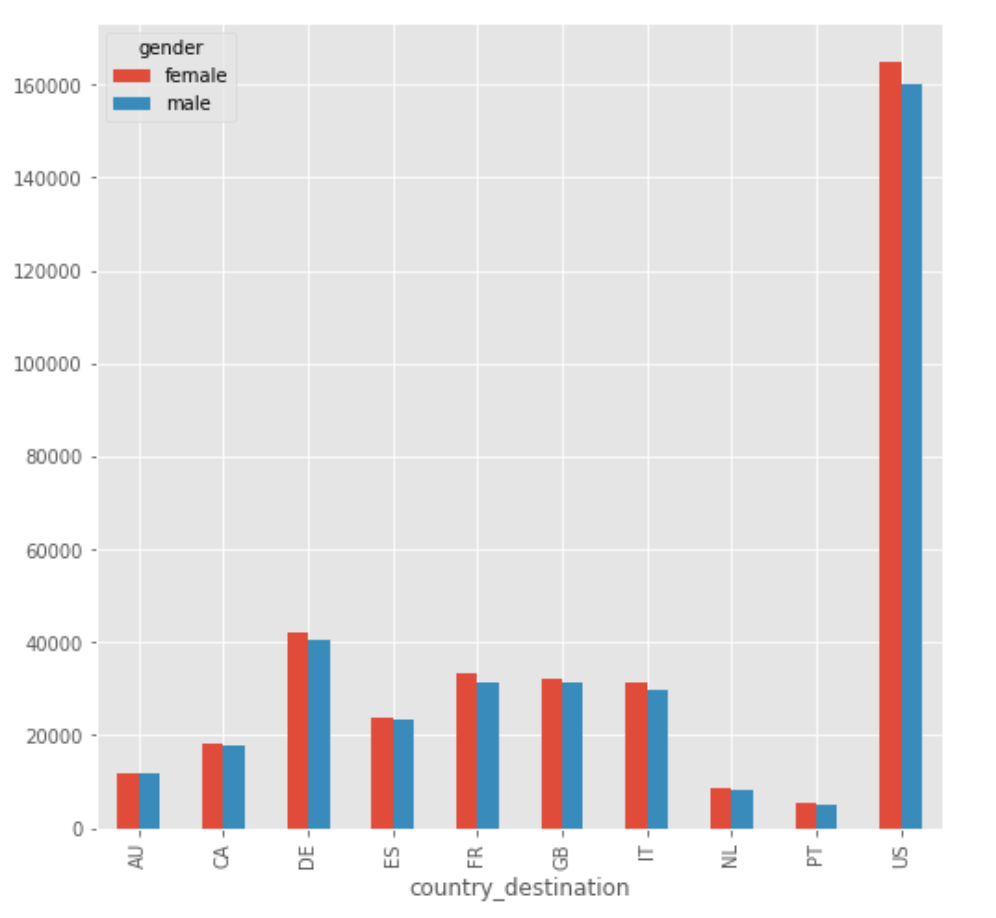
******

Action 'confirm\_email' and 'authenticate' has the highest mean secsElapsed in a user session.

* ***Language difference and km distance for a US user:***

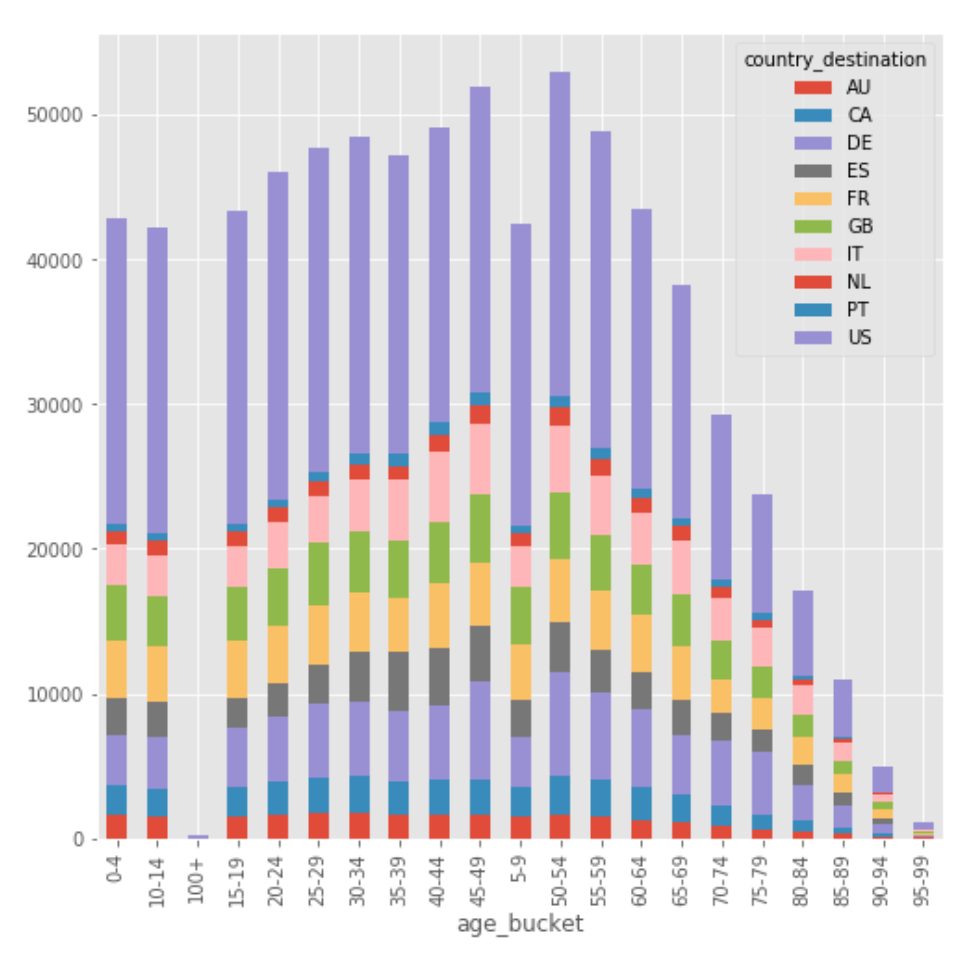
******From plot 1, AU looks farest from the US in km distance.ES, FR, PT have the highest language\_levenshtein\_distance i.e these languages have the highest difference score from US english.

* ***Demographic information of cities***

******

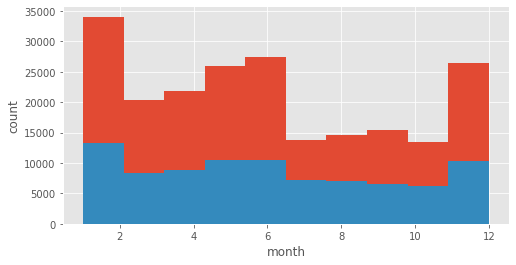
The US seems to have the highest population, also female population is higher compared to male for all destination countries.

* ***Age bucket wise distribution of destination country***

******

There is no significant variation in the segments with age buckets.

* ***Highest first bookings and accounts created***

******

From the plot, we see that the shapes of accounts created and first bookings are quite similar, December and January have the maximum count.

***DATA PREPROCESSING AND FEATURE ENGINEERING:***

As a part of data preprocessing and feature engineering following steps were performed.

* Datetime format transformations.
* Extracting important features from datetime like month were added as separate features.
* Less frequent categories considering a threshold were transformed to single categories like ‘Others’.
* Grouping and aggregations.
* Dropping redundant columns.
* Joining eg. Session data was joined with train data.
* Age to Age\_group transformation.
* Adding features like user language, age group preferences from the demographics information of the destination countries.