Raw Data Set:

5 raw datasets:

- sessions.csv - web sessions log for users
    - user_id: to be joined with the column 'id' in users table
    - action
    - action_type
    - action_detail
    - device_type
    - secs_elapsed
- train_users.csv - the training set of users
- test_users.csv - the test set of users
    - id: user id
    - date_account_created: the date of account creation
    - timestamp_first_active: timestamp of the first activity, note that it can be earlier than date_account_created or date_first_booking because a user can search before signing up
    - date_first_booking: date of first booking
    - gender
    - age
    - signup_method
    - signup_flow: the page a user came to signup up from
    - language: international language preference
    - affiliate_channel: what kind of paid marketing
    - affiliate_provider: where the marketing is e.g. google, craigslist, other
    - first_affiliate_tracked: whats the first marketing the user interacted with before the signing up
    - signup_app
    - first_device_type
    - first_browser
    - country_destination: this is the target variable you are to predict
- 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

Data Cleaning Step:

- Group count of session by user id as count of activity count and sum of “secs elapsed” by user id as total time spent with airbnb website
- Merge user training data with the session data by user_id
- 58% of users have not made a booking yet. So I created the trip_booking_flag field for the binary classification problem.
- In the user dataset, there are two timestamp fields: account creation date and first active timestamp. By leveraging these two timestamp fields, two numerical fields have been created: number of days since the account creation as of latest day and number of days since the first activity as of latest day..
- There are some users with unreasonable ages that are lower than 18 or higher than 122. For the users with age that is lower than 18, the age is computed as 18. For the users with age that is higher than 122, the age has been computed as 122.
- Since there are some null values in the age. To resolve the issue, I have cut the age into buckets based on quantiles and null values are categorized into the unknown bucket.
- Since there are some null values in total time spent with airbnb website and total activity count fields, I have cut the fields into bucket based on quantile and the null values are categorized into the unknown bucket.
- For “total time spent” and “activity count” fields, I have also filled the null value with zeros to keep the continuous variables.
- For the account creation date, four more fields have been created: year, month, day and day of week.
- For the first active date, four more fields have been created: year, month, day and day of week.

Final Cleaned Dataset:

Below is the screenshot of the fields in the cleaned dataset:

```
user_training_dataset_update_binary_classfication.columns
```

```
Index(['id', 'date_account_created', 'timestamp_first_active_cleaned',
       'gender', 'signup_method', 'signup_flow', 'language',
       'affiliate_channel', 'affiliate_provider', 'first_affiliate_tracked',
       'signup_app', 'first_device_type', 'first_browser', 'age_computed',
       'Total time spent (in seconds)',
       'number_of_active_day_as_of_latest_date',
       'number_of_days_since_account_creation_as_of_latest_date',
       'session count', 'trip_booking_flag', 'Account_creation_date_month',
       'Account_creation_date_year', 'Account_creation_date_day',
       'Account_creation_date_day_of_week', 'first_active_date_month',
       'first_active_date_day', 'first_active_date_year',
       'first_active_date_dayofweek', 'age_bucket',
       'Total time spent (in seconds)_fill_null_zero',
       'session count_fill_null_zero', 'Total time spent (in seconds)_bucket',
       'session count_bucket'],
      dtype='object')
```