# **AirBnb Data Set Description**

# **Where will a new guest book their first travel experience?**

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.

In this recruiting competition, Airbnb challenges you to predict in which country a new user will make his or her first booking.

In this challenge, you are given a list of users along with their demographics, web session records, and some summary statistics. You are asked to predict which country a new user's first booking destination will be. All the users in this dataset are from the USA.

There are 12 possible outcomes of the destination country: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found), and 'other'. Please note that 'NDF' is different from 'other' because 'other' means there was a booking, but is to a country not included in the list, while 'NDF' means there wasn't a booking.

The training and test sets are split by dates. In the test set, you will predict all the new users with first activities after **7/1/2014**. In the sessions dataset, the data only dates back to 1/1/2014, while the users dataset dates back to 2010.

**Source**

Kaggle competition from 25 November 2015 to 11 February 2016

<https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings>

AirBnb provided 5 data files.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Source file*** | ***Description*** | ***Shape*** | ***Additional Details*** |
| users.csv | Information about users  Data from 2009-03-19 to 2014-06-30  Accounts created from 2010-01-01 to 2014-06-30 | 213451 lines and  16 columns | 41% of age data is missing |
| sessions.csv | Web sessions log for users (train and test) | 10567737 lines and  6 columns | Session data is provided only for 35% of train users, but for 99% of test users |
| countries.csv | Summary statistics of destination countries in this dataset and their locations | 10 lines and  7 columns |  |
| age\_gender\_bkts.csv | Number of population of specified age in different country destinations (on year 2015) | 420 lines and  5 columns | age\_bucket ['100+' '95-99' '90-94' '85-89' '80-84' '75-79' '70-74' '65-69' '60-64'  '55-59' '50-54' '45-49' '40-44' '35-39' '30-34' '25-29' '20-24' '15-19'  '10-14' '5-9' '0-4'] |
| test\_users.csv | Data set for prediction.  Contains the same columns as users data set except for country destination (target variable)  Data from 2014-07-01 to 2014-09-30 | 62096 lines and  15 columns | 46% of age data is missed  date\_first\_booking is excluded and is always null (if it were not null, it could mean that booking was made and thereby excluded category ‘NDF’ ) |

**Variable Description**

Variables from users and test\_users datasets

|  |  |
| --- | --- |
| Id | Unique 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 | User`s gender |
| Age | 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 | Application used for signing up |
| first\_device\_type |  |
| first\_browser |  |
| country\_destination | this is the **target variable** for prediction |

Variables from sessions dataset

|  |  |
| --- | --- |
| user\_id | Unique user id, can be joined with the column 'id' in users table |
| Action | none |
| action\_type | Type of action: click, submit, booking request and etc |
| action\_detail | none |
| device\_type | Type of device used by user (Windows Desktop, iPad, etc) |
| secs\_elapsed | Number of seconds between actions were recorded. |

Variables from countries dataset

|  |  |
| --- | --- |
| country\_destination | Country destination ('US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU') |
| lat\_destination | Latitude |
| lng\_destination | Longitude |
| distance\_km | Distance to destination country |
| destination\_km2 | Area of destination country |
| destination\_language | Language of destination country |
| language\_levenshtein\_distance |  |

Variables from age\_gender\_bkts

|  |  |
| --- | --- |
| age\_bucket | Age group |
| country\_destination | Country destination |
| gender | Gender |
| population\_in\_thousands | Population in specified age group and gender |
| year | Year of observation |