**Link to the original datasets ->**

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

The original and cleaned datasets are too large to be attached to github.

**Train.users.csv**

Totally 213,452 rows with 16 columns, with substantial missing values in column Age.

In Age column, there are many values in thousands and some in single digits. Assume an age range of 15 to 100, and then assume the 4 digit years from 1924-1995 as birth years.

1. Replace all ages outside the valid range to NA
2. Replace all missing values to NA
3. Calculate the 4 digit years as birth year by subtracting given year from 2016 (commencement of competition in Kaggle)

**Sessions.data.csv**

Totally 10,567,737 rows with 6 columns, with multiple rows for a user-id.

User\_id

1. Replace missing or blanks as “Not given”

Action

1. Replace “-unknown-“ as “unknown” and
2. Replace blanks as “Not Given”
3. Append each Action with a suffix (“\_AC”) to differentiate between similarly named columns
4. Rename all Actions with Cumulative Percentage > 95% as MISC

Action\_type

1. Replace “-unknown-“ as “unknown” and
2. Replace blanks as “Not Given”
3. Append each Action with a suffix (“\_AT”) to differentiate between similarly named columns
4. Rename all Action\_type with Cumulative Percentage > 98% as MISC

Action\_detail

1. Replace “-unknown-“ as “unknown” and
2. Replace blanks as “Not Given”
3. Append each Action with a suffix (“\_AD”) to differentiate between similarly named columns
4. Rename all Action\_detail with Cumulative Percentage > 95% as MISC

Device\_type

1. Replace “-unknown-“ as “unknown” and
2. Replace blanks as “Not Given”
3. Append each Action with a suffix (“\_DT”) to differentiate between similarly named columns
4. Rename all Device\_type with Cumulative Percentage > 96% as MISC

Secs\_elapsed

1. Replace all NA or blanks to 0