Rounak Banik
IIT Roorkee
rounakbanik@gmail.com
+91 84398 60325

Data Wrangling

OVERVIEW

This document describes the various data cleaning and data wrangling methods applied on the Airbnb datasets to make it more suitable for further analysis. The following sections are divided based on the datasets provided.

AGE, GENDER AND STATISTICS

1. The age bucket was converted into a mean age feature. This was done to treat age as a numerical feature. Furthermore, the numerical nature of the data will make it easier to perform one hot encoding and label encoding on this feature later.
2. The year feature was dropped as it had only one value, 2015. Therefore, it was giving us no extra information and could be safely dropped.

COUNTRIES

The dataset is clean and extremely small. No wrangling or cleaning techniques were used on this dataset.

SESSIONS

1. All the unknown fields were converted into NaN to give it more semantic meaning.
2. The missing second values were interpolated using the Pandas Series Interpolate function. This was done as this treatment did not significantly alter the summary statistics of seconds elapsed.

TRAINING USERS

1. All unknown values were converted into NaN to give it more semantic meaning.
2. Samples for which age was greater than 120 was converted into NaN as these clearly represented polluted data.