Exploratory Data Analysis

Ahmed M. Hasan

Dataset used : FordGoBike

Summary

In this exploratory data analysis, we have used trip data collected from June 2017 through December 2019. Because of memory limitation, data is kept in three different files corresponding to three years 2017, 2018 and 2019.

- 2017 data file has 5,19,700 trips information. It has 13 features.
- 2018 data file has 18,63,721 trips information. It has 14 features.
- 2019 data file has 25,06,983 trip information. It has 15 features.

The column header descriptions are as follows:

- Trip Duration (seconds)
- Start Time and Date
- End Time and Date
- Start Station ID
- Start Station Name
- Start Station Latitude
- Start Station Longitude
- End Station ID
- End Station Name
- End Station Latitude
- End Station Longitude
- Bike ID
- User Type (Subscriber or Customer – "Subscriber" = Member or "Customer" = Casual)
- Bike share for all trip (present 2018 and 2019 data)
- rental access method (present in 2019 data)

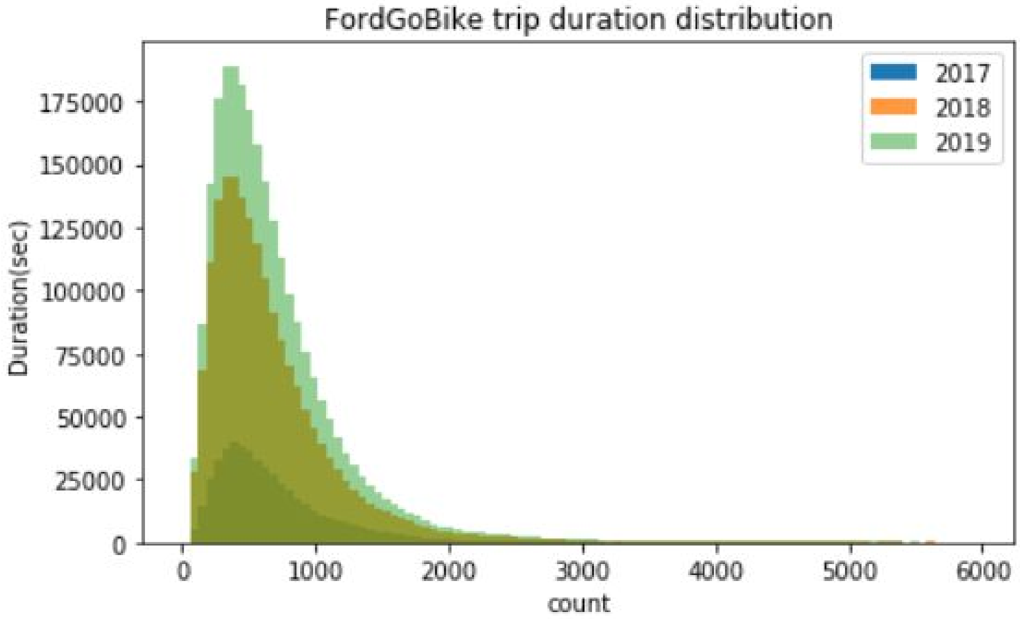
I added the following columns:

- Start Time Month
- Start Time Month Num
- Start Time Day of the Week
- Start Time Hour
- Distance (km)
- Distance (miles)

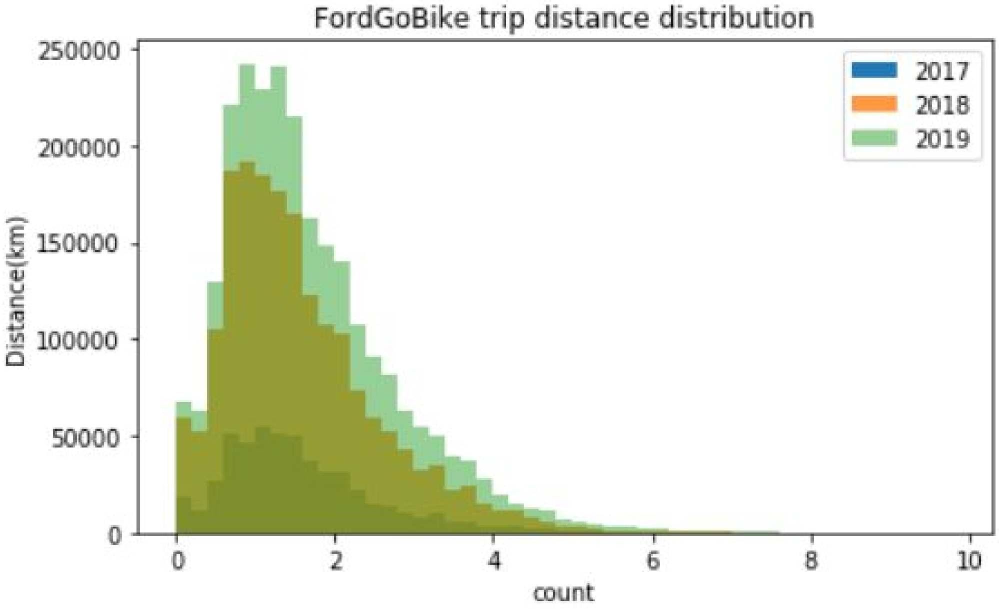
Interesting observations:

- Most of the trip data have short distance and short duration.
- User Type feature seems to be the most useful categorical feature to get more insight.
- Duration, distance, start time (day, month, hour) seem most useful features in explanatory analysis later.
- Trip data are anonymized. So all user personal information (e.g. age, gender etc) are removed in this data.

Trip Duration Distribution



Trip Distance Distribution



For more details and code, check the *exploratory_data_analysis.ipynb* notebook.