Analysis of Ford GoBike System Data

Cathy Moy
July 2019

Ford GoBike System Data: https://s3.amazonaws.com/fordgobike-data/index.html

Dataset

This data set, (2017-fordgobike-tripdata.csv) that I renamed in my exploratory data analysis to (2017-bike-data.csv) includes information about individual rides made in a bike-sharing system covering the greater San Francisco Bay area for the whole year of 2017. Each trip is anonymized and includes many variables such as trip duration in seconds, start time, end time, start and end station information, bike ID, and user information.

This dataset started with 519,700 entries but after some preliminary data wrangling, I will be working with 435,098 entries and there will be focus on four columns, two of which I created. I created new two columns, duration_min and member_age. I used the original duration_sec column to create the new column duration_min. I converted the entries of duration_sec that was in seconds to minutes for better read of data and saved that to a new column, duration_min. I also changed member_birth_year to member_age by calculating the age from year 2017, the year that this data was obtained. Then, I saved those calculations to a new column, member_age. The purpose of this change was for easier understanding of bike riders regarding their age and bike ride duration. The four columns I worked with for my data analysis were duration_min, member_age, member_gender, and user_type. My cleaned dataset has been saved to a new file as 'clean-2017-bike-data.csv'.

Main Findings

From the exploratory data analysis, there were some interesting findings about how gender, user type, and age group play an interaction effect to average trip duration for bike rides by the Ford GoBike system. I mainly discovered that for gender, females had longer bike rides on average than males. For user type, customers had longer bike rides on average than subscribers of the bike sharing system. Then for age groups that were grouped by a span of 20 years, age group of 0-20 had longer bike rides on average than both age group of 20-40 and 40-60. During my exploratory data analysis, I chose to account for the number of bike riders in each category and to group them by each other to look for connections and if average trip duration would be affected by the three variables at hand: age, gender, and user type.

Key Insights

From examining the dependence of bike riders' characteristics such as gender, age, and user type, I found some key insights to their expected average trip duration for bike rides with the Ford GoBike system. When comparing these variables against one another and to the other two, the same pattern followed in which bike riders who are females, customers, and those between ages 0-20 are more likely to have longer individual bike rides on average than others. This finding is supported with the averages found for each characteristic and grouped characteristic. Although, there was an uneven distribution of representation among the characteristics, the average trip duration showed similar correlations. From my data analysis, I can conclude that gender, user type, and age group may have some effect to the average trip duration for bike rides taken by the Ford GoBike system.

Resources

1. https://stackoverflow.com/questions/17578115/pass-percentiles-to-pandas-agg-function

2. https://stackoverflow.com/questions/53277718/pandas-dataframe-easier-syntax-to-drop-rows-by-condition-on-values

3. https://stackoverflow.com/questions/31583151/count-number-of-rows-when-row-contains-certain-text

4. https://stackoverflow.com/questions/34828701/mean-line-on-top-of-bar-plot-with-pandas-and-matplotlib/34829398

5. https://stackoverflow.com/questions/48978550/pandas-filtering-multiple-conditions

6. https://stackoverflow.com/questions/18992086/save-a-pandas-series-histogram-plot-to-file

7. https://stackoverflow.com/questions/17071871/select-rows-from-a-dataframe-based-on-values-in-a-column-in-pandas