Ford GoBike Bike Sharing Data Explanatory Analysis

Dataset

The data of Ford GoBike bike sharing service was selected for analysis. The original datasets contain the information about 2,290,554 bike rides with 16 columns. There a 2 time-relevant variables ('start_time', 'end_time'). The latitude and longitude coordinates of both start and end locations are provided. One numeric variable ('member_birth_year') and two categorical variables ('user_type', 'member_gender') about the users are also available.

The dataset can be found here: https://s3.amazonaws.com/baywheels-data/index.html

Summary of Findings

- Monthly ride count ranges between 150,000 and 200,000 from May 2018 to February 2019
- There is a significant drop of count in November and December, 2018, which may be due to the weather
- There is a big jump of ride count in March 2019. The count drops slightly in the following month, but still fairly higher than the counts in 2018
- The demand for the service is significantly higher during the week (Monday to Friday) than weekends
- The peak hours of the service are from 8 to 9 a.m. and 4 to 6 p.m.
- Most rides are within 4 kilometer in distance and 20 minutes in duration. There is a linear correlation between distance and duration per ride within this range
- The most frequent rides are about 0.5 to 1 kilometer in distance and 3 to 5 minutes in duration

- Duration per ride is relatively shorter between November 2018 and February 2019. However, the distance per ride during this period of time is not significantly shorter

- Distance per ride increases steadily over time

- Duration per ride is longer on weekends, but the distance per ride becomes shorter

- Rides by male users are more than twice as much than female users

- Female users generally spend more time per ride than male users

- ’20-29’, ’30-39’ and ’40-49’ are the top3 user groups, with the ’30-39’ group being the largest. These three groups account for almost 89% of the rides

- While the ’30-39’ group is the largest during the week, the ’20-29’ group uses the service more than other groups on weekends

- While the ’30-39’ group is the largest during the day (6 a.m. to 7 p.m.), the ’20-29’ group uses the service more than other groups from 8 p.m. to 2 a.m.

- Compared to other groups, duration per ride for the ’18-19’ group is more sparsely distributed

- The majority of the rides in the data is by the subscriber group, and the frequency is much higher than the customer group

- The customer group is younger than the subscriber group. Specifically, the proportion of the ’20-29’ age group is slightly higher for the customer group

- The proportion of female users in the customer group is slightly higher than the subscriber group

- The subscriber group uses the service more during the week while the customer group more on Saturdays

- Both overall duration per ride and overall distance per ride for the customer group are longer than that of the subscriber group

- A further investigation shows that while the overall duration per ride for the customer group is longer than the subscriber, it is only true for Friday and Sunday

- Similar pattern also found in distance: while the overall distance per ride for the customer group is longer than the subscriber group, it is only true for Friday and Sunday

Descriptive statistics are first given in the explanatory analysis. The findings are presented by features/variables. Only findings with significance are chosen for the explanatory analysis.

Sources

- For calculating distance between two coordinates: https://stackoverflow.com/questions/33029396/using-pandas-to-calculate-distance-between-coordinates-from-imported-csv

- Markdown specifications: https://daringfireball.net/projects/markdown/syntax