Paras Patel
Data Analytics – Udacity
Exploratory Data Analysis Report

In this project, Ford GoBike's 2018 system data was analyzed. Pythons data science and visualization libraries were used to explore the dataset's variables and understand the data's structure, oddities, patterns, and relationships.

First, a distance column was added to the dataset using Python's geopy library that uses a spherical model to calculate the shortest distance between 2 points. Although not a perfect representation of the true distance traveled, it's a good estimate. After creating this column, many of the starting and ending locations were observed to be the same, resulting in a distance of '0'. An attempt was made to determine a relative distance based on the duration of the trip. An average velocity was calculated and a new distance was calculated (Distance = Velocity x Duration). Unfortunately, there was a high mean difference (665 meters) between the locational distance calculated by geopy and the velocity based distance. The locational distance was used for subsequent analysis.

Second, null data was explored. Typically, null data analysis would be part of data wrangling, but in this case, null data was explored and visualized to determine potential issues with the Ford GoBike service. The data revealed a high concentration of null values for Bike ID's in the range of 4100-4400. Although the number of trips made is almost double compared to the beginning of the year, this does not provide enough supportive evidence that null values are correlated to the number of trips. Instead, it suggests potential product failure i.e. a bad manufacturing batch of these bikes, a possible software glitch, or issues with the geo tracking system.

<u>Univariate Relationships</u>

First, the distribution of the 'duration_sec' and 'distance' columns were examined. Binning revealed the data was skewed to the right, so a log transform was used on both the 'duration_sec' and 'distance columns. Accounting for outliers, the histograms revealed the mean duration traveled was roughly 800 s with most trips lasting 650 s. The mean distance traveled was approximately 1600 m with a majority of trips lasting anywhere between 700-1500 m.

A distribution of the start times revealed most of the trips were occurring at 8:00 AM and 5:00 PM, which coincides with typical starting and ending work times.

The top 3 starting and ending locations were the same: Townsend St. at 4th St (station 2), Harry Bridges Plaza, and Townsend St at 4th St.

Number of subscribers far exceeded number of customers using the service (1.58 million vs. 280,000), and only 162,000 trips were bike share for all trips out of the 1.8 million+ trips made in 2018.

The distribution of the user age was explored to determine which age ranges are most likely to use the product. A line plot was made to show most users who used the Ford Go-Bike service were young adults (late 20's to early 30's) with a steady drop off beginning at 33 years old.

Bivariate Relationships

Bivariate relationships between trip duration, trip type (Ford Go-Bikes Bike Share For All), membership, and user gender were examined. Hypothetically, Bike Share For All trips should level off at 60 min due to the fee for additional time over this time limit. This hypothesis was supported by the data as only 1.2% of bike share for all trips lasted over 60 min. Visualization by a clustered bar graph showed gender distribution among user types (Customer vs. Subscriber) favors males over females/other. Over 700,000 more trips were made by male subscribers vs female subscribers and over 70,000 trips more trips were made by male customers vs female customers.

Multivariate Relationship

This analysis focused on the top 10 starting locations (based on number of trips) and the length of the trips made from these locations at different time frames of the day (3 hour timeframes). The heat map was fairly distributed but a few insights can be made. For one, the longest rides occurred from 12AM - 3AM, but it's also important to note this group contained only 1,380 data points compared to 3PM-6PM time frames for instance, that contained over 80,000 data points. The shortest rides occurred in the 6AM-9AM time frame, most likely users using the service to travel short distances to work. Further investigation also revealed the longest rides were starting from the San Francisco Ferry Building.