Ford GoBike Exploration Report

By: Mehrnaz Siavoshi

About the Project

This project started by gathering data from the Ford GoBike website (now under Lyft) for all months of 2019. Data was aggregated and cleaned. A series of visualizations (univariate, bivariate, and multivariate) were created to explore the data.

Data Collection and Cleaning

Data was collected from the Ford GoBike Data Page (now part of Lyft services). Trip history data from all months of 2019 was downloaded. According to the data providers, each trip entry includes:

- 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)

First, the 12 datasets were combined into one master dataset. The data was very clean with no significant missing values. However, in order to streamline analysis, new columns for day and month were created.

The final set contained 2,506,983 observations with 22 features. The main feature of interest is the trip length, as well as specific time of the trip, including day of the week and time of day.

Exploratory Visualizations

Univariate Exploration

Univariate exploration involves looking at trends in one variable at a time. In this section, the following will be explored:

- {x} Trips by month.
- {x} Top 10 starting stations.
- {x} Starting station geographic density.
- {x} Trips by length.
- {x} Trips by day of the week.
- {x} Trips by time of day.
- {x} Bike ID.

When considering when trips were taken, there were some clear peaks in specific months. However, there was no easy explanation as to why, for example, there was a peak in March and October.

Trips by time of day followed a bimodal distribution with clear spikes during common commute times. This suggests that most of the usage of these bikes is for commuters.

Trips by day of the week had a constant distribution during the week with a dropoff on the weekend, also supporting this conclusion.

It was discovered that when looking at trip length, there were a lot of very long outliers. The vast majority of trips were under 1 hour long, so these trips were separated into a new dataset for visualization and later analysis.

Bivariate Exploration

Bivariate explorations look at whether two variables are correlated. In this section, the following will be explored:

- {x} The day of week and trip duration.
- {x} The day of the week and trip start time.
- {x} The trip duration and subscriber status.

Note: For all analyses that use trip time, the short_trip dataset will be used. This dataset only contains trips 60 minutes or shorter, which encompasses the vast majority of trips without including the effect of significant outliers.

In the previous section, we learned that that more trips occurred during the week than on weekends. However, here, we learned that longer trips tend to occur on the weekends. This suggests that commutes tend to be shorter than leisure bike rides.

Interestingly, we also determined that customers tend to have longer bike rides than subscribers. This correlates nicely with the longer average trip time on weekends as customers tend to not be primary commuters.

Multivariate Exploration

Multivariate exploration considers more than 2 variables to determine trends. In this section, the following will be explored:

- {x} Interaction between trip hour, day, month and length.
- {x} Trips by day of week and month.
- {x} Trips by hour and day of the week.
- {x} Trips by day of week for different subscriber statuses.
- {x} Interaction between trip hour and subscriber status.

In this section, the interaction of multiple variables was considered. In agreement with previous results, we see that subscribers (at this point assumed to be primarily commuters) have significantly more trips during the week than customers, who have a uniform distribution throughout the week.

Interestingly, there was no correlation between when a trip was taken and its length. This may be due to the effect of such a large number of very short trips (under 10 minutes) that seem to be uniformly distributed across all days and times.

The difference between the number of trips per hour for subscribers and customers is striking. While both have a bimodal shape with peaks occurring at commute times, the customer bar graph is much more uniform in shape.

Conclusion: Key Insights

Through the analysis, some interesting observations were made:

- There is increased usage in March, April, July, and October.
- The vast majority of trips are less than 10 minute long.
- Most of the bikes are used for commutes, as demonstrated by the peaks in usage around 8-9 AM and 5-6 PM.
- Customers, who likely are not using the bikes for commutes, typically have longer rides.
- The time of day when a bike is rented has no impact on how long the trip will be.
- However, any true effect of time of day may be hidden by the overwhelming number of trips less than 10 minutes in duration.
- In the customer use by time of day graph, we see that some customers may be using the bikes for a one-off commute, or to test the service for a commute before becoming a subscriber.

More complete analyses could have been done if more data was included. For example, it would have been interesting to look at the age or gender of the rider. However, as this data was not available, only information about usage could be considered.

Additionally, for future exploration, it would be interesting to compare these analyses over the years. Data from 2017, 2018, and part of 2020 is currently available, so a few different comparisons can be made.