Exploring Ford Go-Bike Data Read Me

Vikram Tharakan

In this project I chose to investigate the Ford Go-Bike data. This dataset includes information over the past 2 years on over 3.5 million user rides. After congregating all the data, I was able to being exploratory analysis and the results are shared below.

Looking at gender, we found that GoBike users are predominantly male, with 69.2% of all users registering as male, while only 22.6% registered as female. There was also a small percentage that identified as 'Other' and many users that did not put down their gender. Out of only the users that filled out their gender information, 74% were men, 24% were women, and 2% identified as other.

When broken down by membership type, we see that 84.4% of the users are subscribers while only 15.6% of the users are customers. The ratio of customers to subscribers is also much higher for females than it is for males.

Next, I looked at average distance traveled. The most common trip length is about 0.5 to 0.6 miles and this did not vary based off of gender. However, when broken down by membership, the most common ride for customers involved returning the bike to the same location, thus giving an effective total distance traveled of 0.0 miles.

Thus to get a more accurate reading the next logical step was looking at the average duration of a trip. Upon plotting the data, we see that the most common ride duration falls between about 5-10 minutes regardless of the gender or membership type for the user.

I then looked what the age breakdown of the user's looked like. I separated the users out into bins by decade (i.e. 20's, 30's, etc) and found that the most common users were in their 20's and 30's occupying 33.2% and 39.2% of the total rides in the data respectively. I also looked to see if age affected the average distance traveled, but there did not appear to be to much of a difference here.

After that I tried to analyze what time of the day was most popular to ride the Go-Bike. The resulting histogram of rides vs. time of day is bimodal, with the two peaks coming at either side of rush hour (around 9am and around 5-6pm).

I then tried to further look at this "rush hour" theory by looking at the start and end locations of these rides. By plotting a heat map of longitude vs latitude it can be seen that the most common spot for rides is near the financial district/Caltrain stop in San Francisco, which is a highly commuted to area in San Francisco.

Finally, I wanted to see how biking trends have changed over time. By binning the data by month over the last 2 years and then plotting a histogram of the data, we see that there has been a steady upward trend, with the most recent month (4/19) containing over 5 times the number of users in 7/17. I then broke the plot down further by analyzing this trend by age group. It was seen that those in their 20's, 30's, and 40's experienced the biggest increase in number of users, while the other age groups had relatively constant numbers over the past 2 years.

Most of the feedback I got on my plots involved making titles more clear and changing the y axis in histograms to relative values for a clearer message. I was also given feedback into interesting areas that I could explore that I had not explored before which I then tackled, such as breaking down the increase in users plot by age group. The feedback helped make my visuals clearer and introduced new questions into the analysis that I had not considered before.

The main resource used during this project was the Ford Go-Bike data repository. I also had to look up exactly how to unzip files programmatically, which I was able to find with a quick Google search.