Samuel Bendekovic - sbendeko

Andrew Wu – AndrewWu20

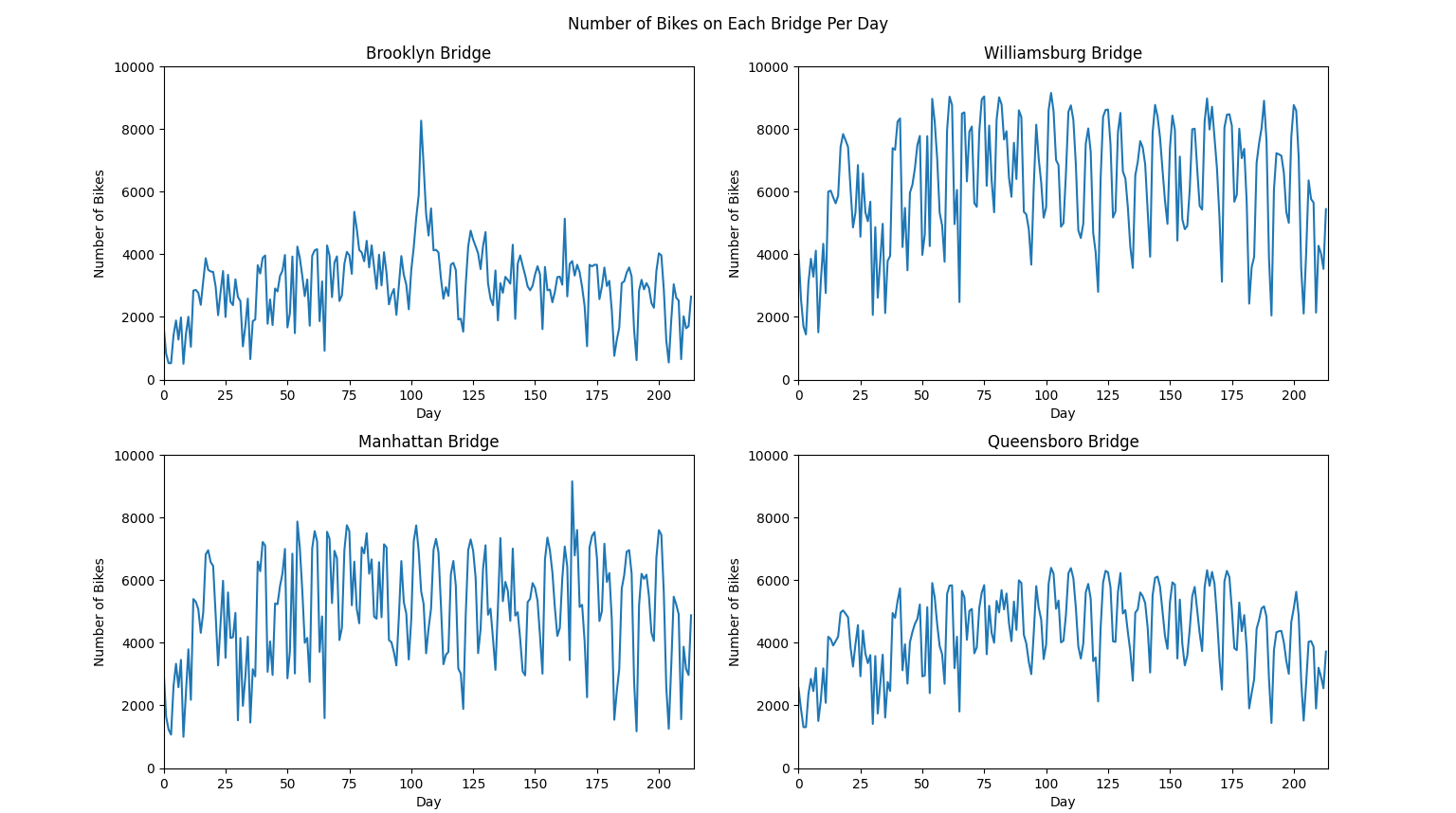
ECE 20875 Project Path 2

**Final Project Report**

We have chosen to solve Path 2, which is the bike traffic problem. That is, we are trying to analyze bike traffic across four bridges in New York City: Brooklyn Bridge, Manhattan Bridge, Queensboro Bridge, and Williamsburg Bridge, and see if it is possible to predict bike traffic based on weather (temp and precipitation) and day of the week. To solve these problems, we used deductive reasoning, linear regression, and confidence intervals to find relationships between number of riders and the factors we are analyzing.

**Problem 1:**

In the first problem, we are tasked with finding which 3 bridges to install sensors on to best describe the overall average number of bike riders. In order to find out which three bridges to pick, we wanted to find out the variance of each bridge and how close the average number of bike riders on each bridge was to the average of all the bridges. This is illustrated in Figure 1 below.

*Figure 1: Graphs Showing Number of Bike Riders on Each of the Four Bridges*

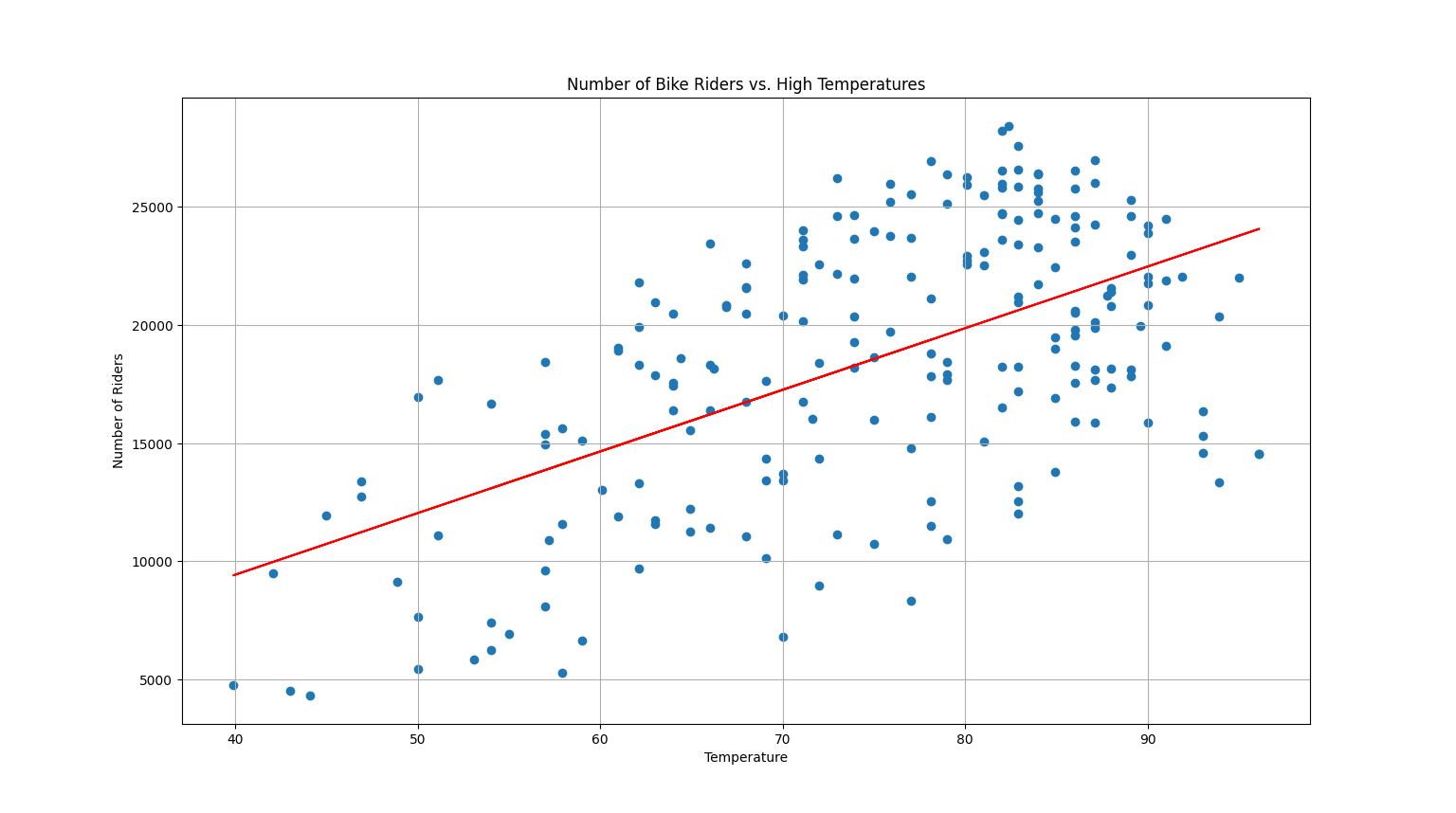
*Table 1: Mean, Standard Deviation, and Variance for Each Bridge Analyzed*

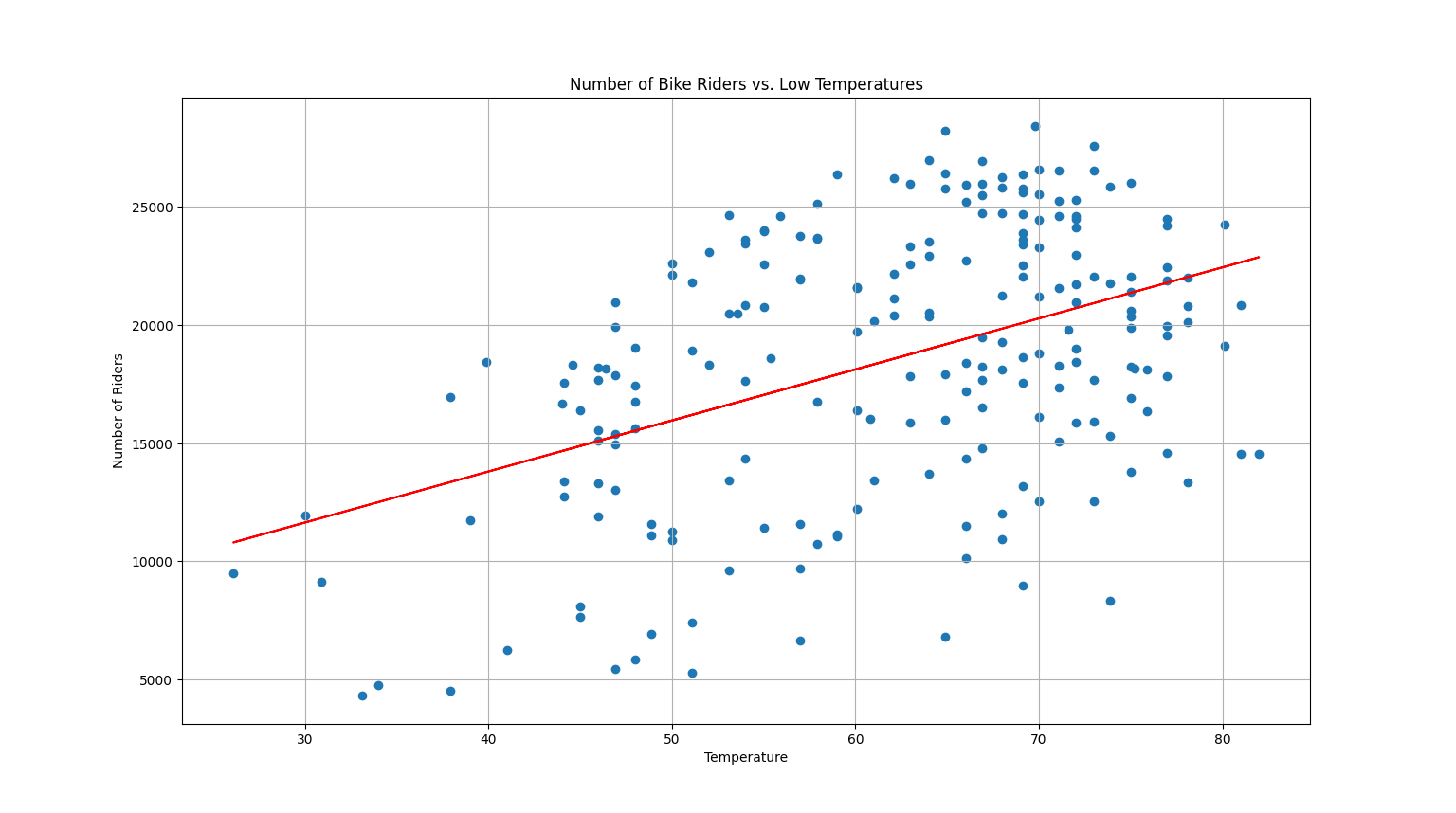
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Brooklyn Bridge | Manhattan Bridge | Queensboro Bridge | Williamsburg Bridge | Overall Bridge Traffic |
| Mean Number of Bikers | 3030.7 | 5052.2 | 4300.7 | 6160.9 | 4636.1 |
| Standard Deviation | 1131.4 | 1741.4 | 1258.0 | 1906.2 | X |
| Variance | 1,280,048 | 3,032,482.3 | 1,582,654.7 | 3,633,498.4 | X |

As shown in the table above, the Williamsburg Bridge has the largest average number of bikers, the Brooklyn Bridge has the lowest number of bikers, and the Queensboro and Manhattan Bridges have the closest averages to the overall average. However, in terms of variance, the Williamsburg Bridge also has the highest variance, followed closely by the Manhattan Bridge, with the Queensboro and Brooklyn Bridges with the lowest. Therefore, due to the similarities in average number of bike riders between the Manhattan and Queensboro Bridges and the high variance of the Manhattan Bridge, the Manhattan Bridge would be the optimal bridge to leave out of the sensor selection. The similarities in the means of the two bridges means that installing sensors on both of them would not contribute any impactful results to the study, since the results would be the same. Furthermore, since the mean number of riders of the Queensboro and Manhattan bridge are close to the total mean number of riders, including sensors on both bridges would exclude the inputs of the other two bridges which would be the outliers in this case. In conclusion, we suggest installing sensors on the Brooklyn, Williamsburg, and Queensboro Bridges, since these three bridges would accurately predict the overall bike traffic with as little variance as possible.

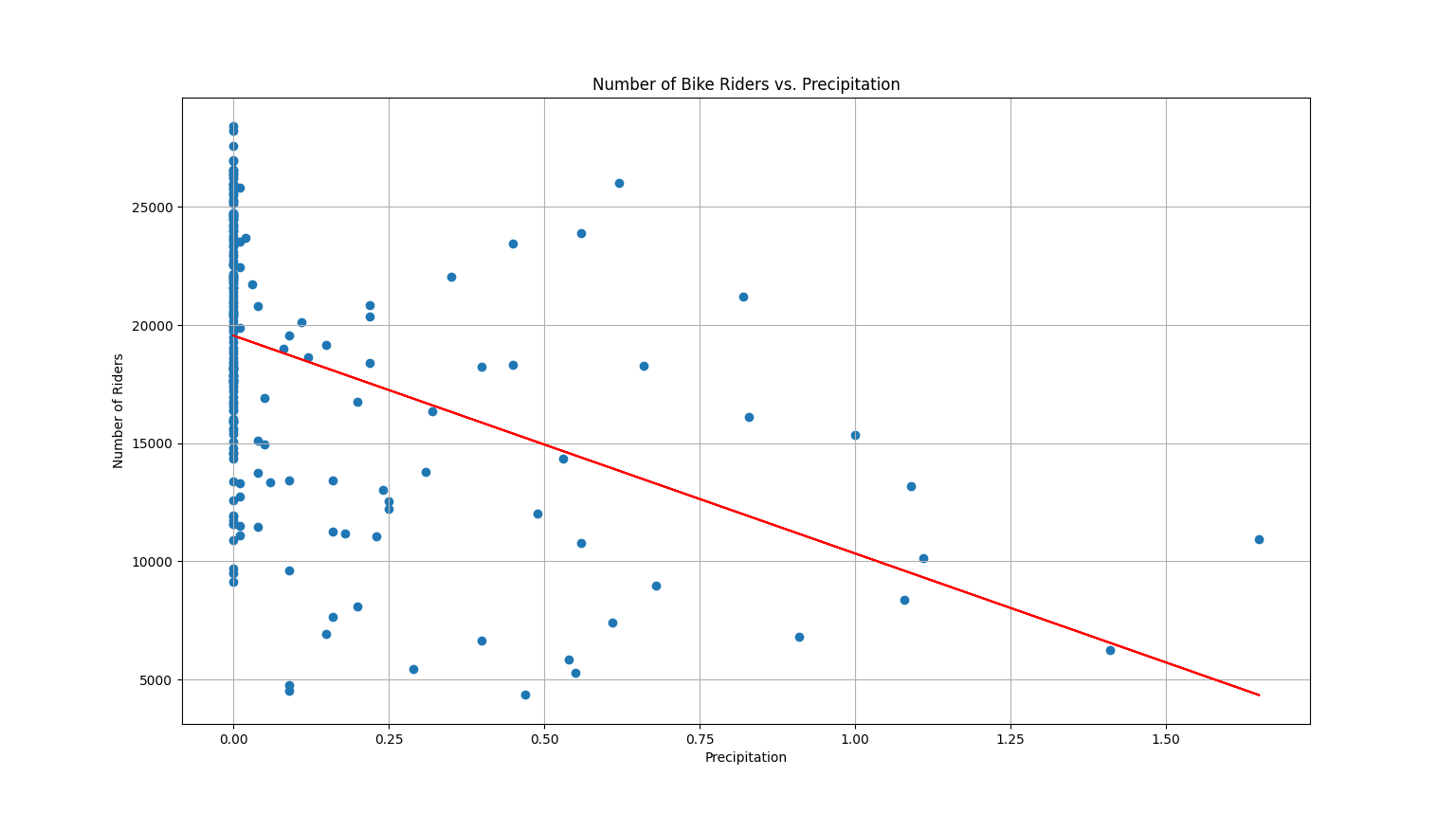
**Problem 2:**

For the second problem, we are asked to see if the use of the weather forecast for the next day, being temperature or precipitation, would be a good predictor for the number of bike riders on that day. To do so, we separated the data into those three factors, high temperature, low temperature, and precipitation, and plotted the three factors against the total number of riders for each day. The resulting data can be seen below in Table 2, and the resulting graphs can be seen below in Figures 2, 3, and 4.



*Figure 2: Graph Showing Number of Bike Riders vs. High Temperatures*

*Figure 3: Graph Showing Number of Bike Riders vs. Low Temperatures*



*Figure 4: Graph Showing Number of Bike Riders vs. Precipitation*

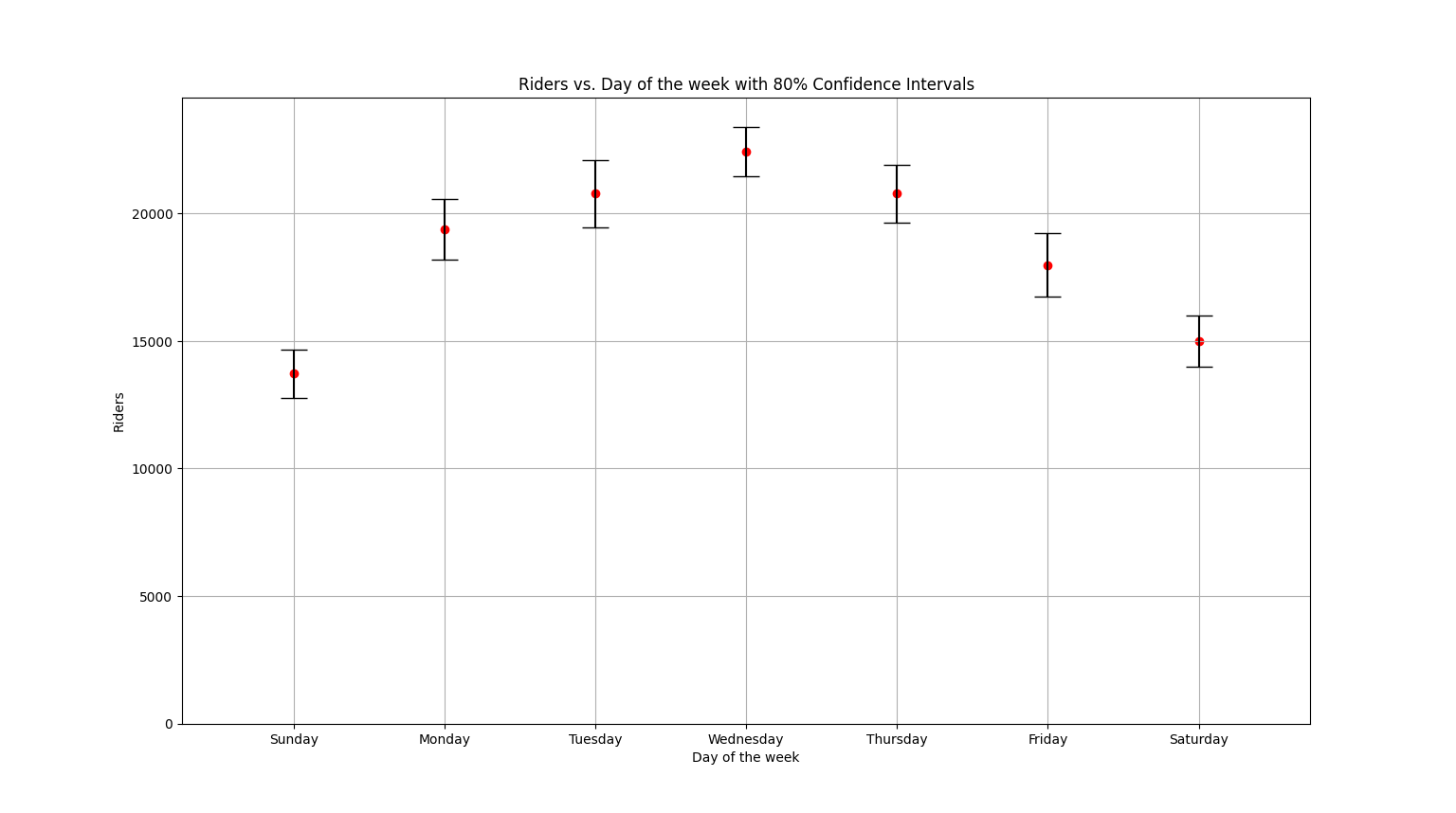
*Table 2: Linear Regression Equations for Each of the Weather Conditions*

|  |  |  |  |
| --- | --- | --- | --- |
|  | High Temperature | Low Temperature | Precipitation |
| M (Slope) | 260.973 | 216.028 | -9228.128 |
| B (Y-Intercept) | -1011.132 | 5156.734 | 19551.002 |
| Full Equation  (with rounded values) |  |  |  |
| r2 Value | .330 | .195 | .177 |

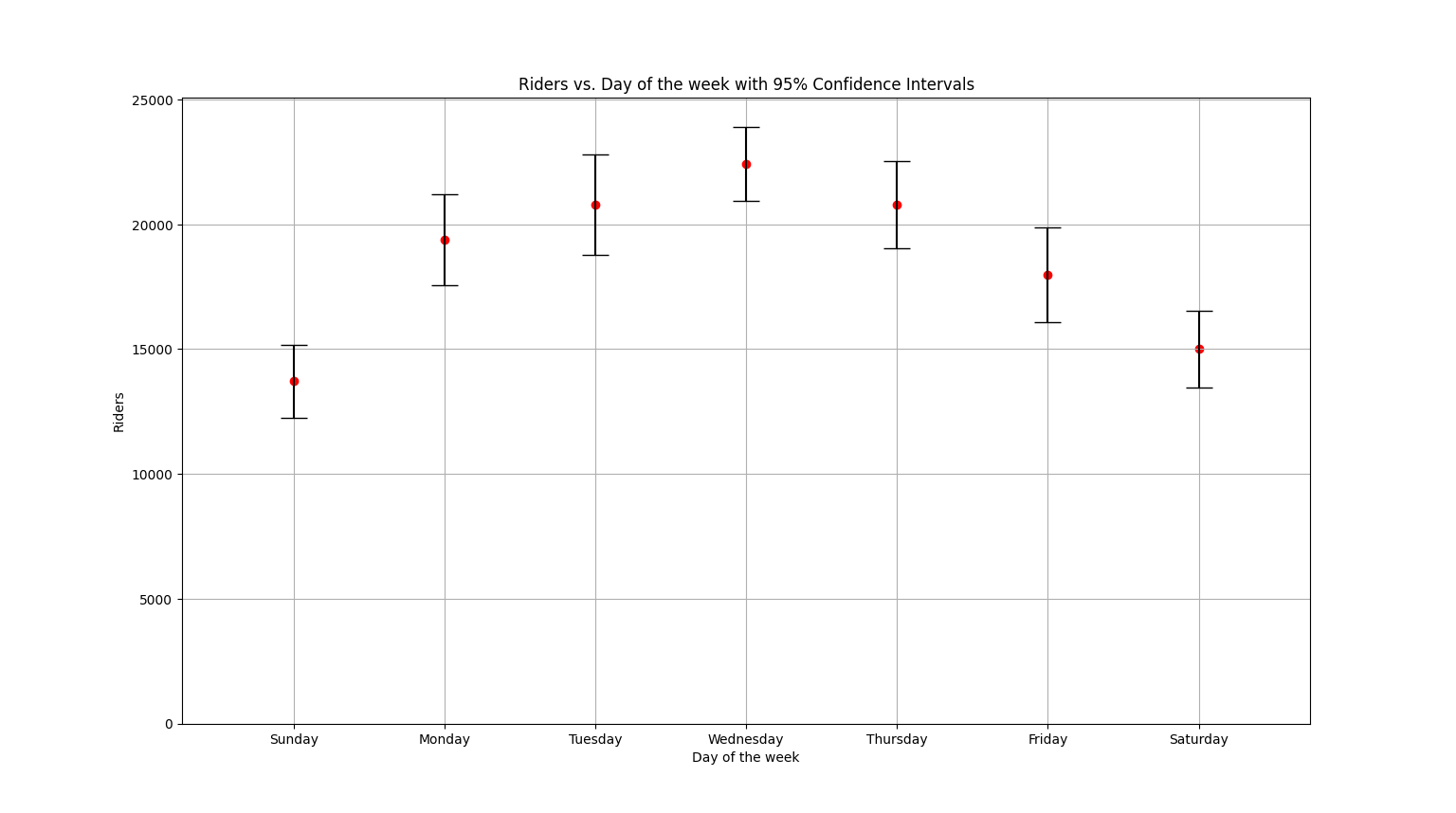
Looking at the figures and table above, the r2 values of each model is very low. The r2 value represents how well the model fits with the data, with one meaning that the model fits perfectly with the data and zero meaning the model does not fit the data at all. With all three r2 values being low, this means that the models generated are generally not good fits of their respective data. However, there is an obvious generally positive trend that can be seen easily with the High/Low Temperatures in Figures 3 and 4, thus these models cannot be completely disregarded. Unfortunately, the precipitation model does not have a good fit, since the 0-precipitation data is evenly spread across ~10,000 riders and there are not a lot of data points past 0.5-precipitation. In conclusion, the High/Low Temperature model predictions could be used with caution to predict the amount of bike riders on the next day, while the precipitation model cannot be.

**Problem 3:**

We are asked to determine if we can use the data given to us to predict what day it is based on the number of riders. To do this, we calculated the means number of riders for each day. We then calculated 80% and 95% confidence intervals around the means for each day to get a range of values each day could be. These can be seen in Figures 5 and 6.



*Figure 5: Graph Showing 80% Confidence Intervals for Each Day of the Week*



*Figure 6: Graph Showing 95% Confidence Intervals for Each Day of the Week*

Based on Figures 5 and 6, you can see that there is overlapping of confidence intervals of different days. This means that a certain number of riders could correlate to multiple days based off the confidence interval. An example would be if we are given ~20,000 riders, the days that have ~20,000 riders would be Monday, Tuesday, and Thursday with 80% confidence, or Monday, Tuesday, Thursday, and Friday with 95% confidence. However, there is an obvious trend that can be seen, with the beginning and end of the week having the least number of bike riders and the middle of the week having the greatest number of bike riders. For example, if we were to be given ~22,000 riders, at both 80% and 95% confidence intervals, the middle of the week (Tuesday, Wednesday, Thursday) could have that number of riders. Similarly, if we had ~14,500 riders, the beginning/end of the week (Sunday, Saturday) could have that number of riders. In conclusion, due to the amount of overlap among individual days, we cannot construct a model that can predict the day based on the number of riders; however, due to the trends in the data, we could hypothetically construct a model that could predict the time of the week it is based on the number of riders on a given day.