Timothy Hein and George Brechbill
heint and gbrechbi
HeinTimothy and BadMonkeyBoss
Path 1

We are working with a dataset that contains the numbers of bikers that crossed 4 specific bridges in New York City per day over a 214 day period, as well as the high and low temperatures experienced and the precipitation on each day. We are also given the total number of bikers across all four bridges and the date of each data point's collection. The bridge crossing samples are given in whole numbers, as well as the total, while the date is given as a mixture of the name of the month and the date within the month. The temperatures, both high and low, are given as numbers rounded to the nearest tenth, and the precipitation is given as a decimal below 0 rounded to the nearest hundredth, or as the letters T and S. We interpreted the T as an error on the part of the data entry, and replaced all the instances of T in the data with 0.01, and interpreted the S as there having been snow that day, replaced the S with a 0, and created a new column to the data which we labeled as 'Snow' and granted a value of 1 for days marked S and 0 for days marked with T or a number.

Analyses:

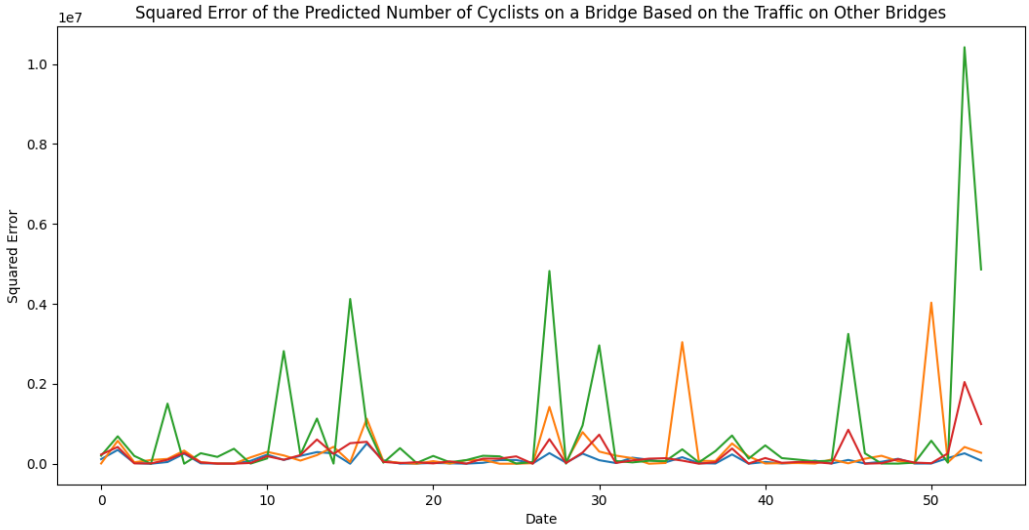
1. To decide which bridge should be excluded, we decided to perform a ridge regression to find which of the bridges is the most related to the other three, based on the idea that this bridge would then be the most well predicted by the data collected with the sensors on the other bridges. We will be given the mean squared error of how well the generated model fits the data from the other three bridges, and then select the model with the least MSE.

2. To check whether the number of bikers is related to the weather, we plan to once again generate a model using ridge regression to attempt to predict the number of bikers given the precipitation on any given day. Using this model, we can compare to the actual numbers of bikers on those days and calculate a correlation coefficient, which we can then compare to an alpha level of 95%. Under these conditions, we would set the null hypothesis as there being no relationship between precipitation and number of bikers, with the alternative hypothesis being that there is a relationship.

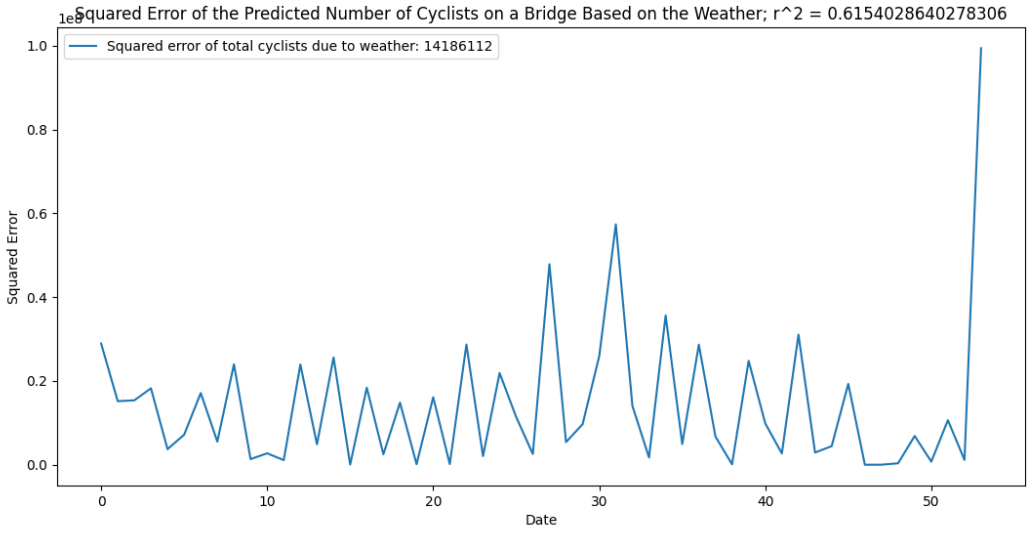
3. To predict the weather using the number of bikers out on the bridges, we will create a training set and a test set of data, and then train a model to predict the weather using ridge regression. Similarly to part 2, we will then be able to calculate a correlation coefficient and compare to the actual values of the data to see how effectively the model is able to predict the weather, comparing the calculated r value to a confidence level of 80%, which is lower than the 95% from part 2 since this is a predictive model and is likely less accurate than a test of relationship.

Results and Answers

1. After completing this regression, we concluded that the bridge that should not receive a sensor was the Queensboro Bridge (blue line) since it has the lowest squared error when tested in this manner, with a calculated value of 93,678. The other values were as follows: Manhattan Bridge (green) - 828,710; Brooklyn Bridge (orange) - 300,512; and Williamsburg Bridge (red) - 207,235.



2. Since the r^2 statistic was 0.615, the correlation coefficient for this relationship r would be 0.7845. This value is below the 95% confidence level, so even though it does suggest there may be some correlation, we will reject the hypothesis that precipitation and the number of bikers are somehow connected.



3. Since we got an r^2 value of 0.2478, the correlation coefficient for this data is r = 0.4978, which indicates a weak correlation. Therefore, at the 80% level of certainty we would

reject this model and conclude that the relative lack of correlation between the weather and the number of bikers would make using one as a predictor of the other a bad idea.