Han Tun Oo (ooh)
Kristopher Kunovski (kkunovsk)
Path #1

Bike Traffic Dataset

This data set contains the data on weather forecast (with the para and biking traffic on four bridges for 214 days in New York City. The weather forecast accounts for data on the daily high temperature, low temperature, and precipitation level. The traffic is recorded on the Brooklyn, Manhattan, Williamsburg, and Queensboro bridges, and we examine the data of each bridge to find the ones that will provide the best predictions. In addition, we observe, determine and analyze the correlation between the weather and total biking traffic to make projections.

Analyses

For the first question, we chose to use linear regression with respect to combinations of three bridges and the total bike traffic. Our expectations with this approach is to obtain the coefficients of each model, and we can use these coefficients to compare which combination of three bridges out of four should have sensors installed to achieve the best overall traffic predictions that best represent traffic data across all the four bridges.

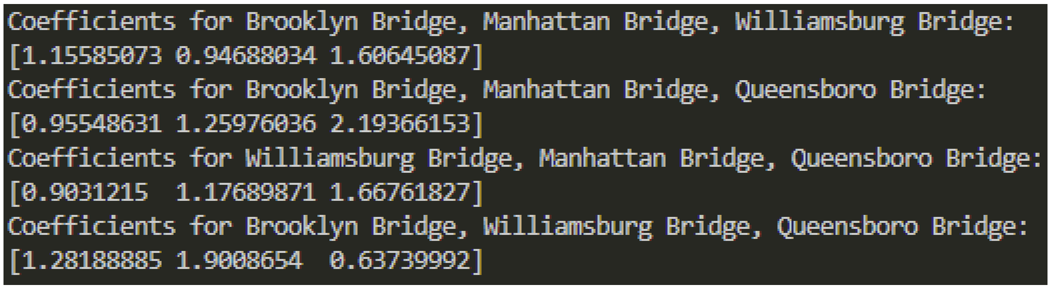
Furthermore, our second question uses a similar tactic of linear regression; however, we normalize all the data prior to fitting a model. This method allows us to explore and determine the overall correlation of the entire weather forecast (accounting for all the given weather parameters of high temperature, low temperature, and precipitation) with respect to the total bike traffic. We also output a coefficient of determination which tells us how predictable traffic is based on the normalized weather data.

Lastly, for the third question, we are merely observing how the total biker population of the bridges is affected by whether or not it rained or snow (whether the precipitation value is zero or nonzero). Because we were attempting to determine the correlation between population data and “binary” categorical constraints of whether or not it rained/snowed, we used logistic regression between the presence of rain/snow (zero or nonzero precipitation levels) and the total bike traffic. By looking at the logistic model score, we should be able to determine if rainy days can be predicted based on the bikers for that day.

Han Tun Oo (ooh)
Kristopher Kunovski (kkunovsk)
Path #1

Results

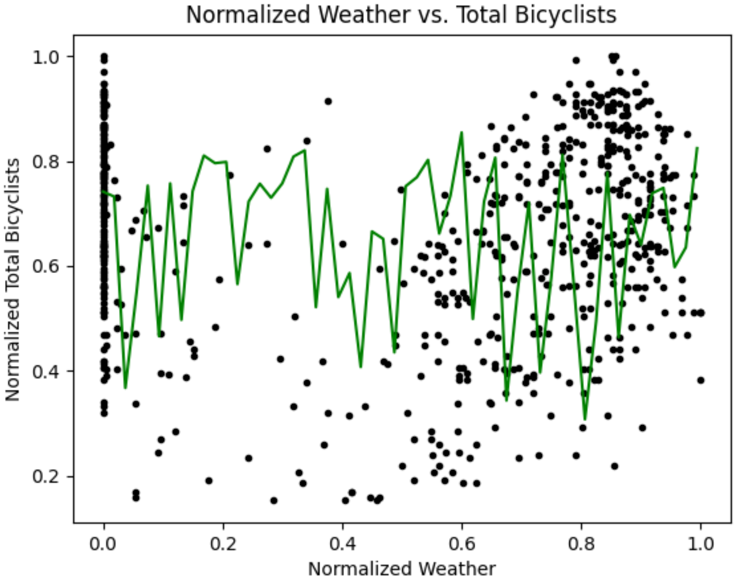
Question 1

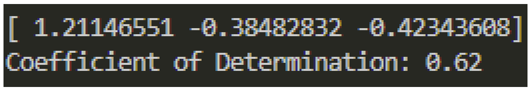


```
Coefficients for Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge:
[1.15585073 0.94688034 1.60645087]
Coefficients for Brooklyn Bridge, Manhattan Bridge, Queensboro Bridge:
[0.95548631 1.25976036 2.19366153]
Coefficients for Williamsburg Bridge, Manhattan Bridge, Queensboro Bridge:
[0.9031215  1.17689871 1.66761827]
Coefficients for Brooklyn Bridge, Williamsburg Bridge, Queensboro Bridge:
[1.28188885 1.9008654  0.63739992]
```

The figure above shows the model coefficients of the feature combinations for each bridge trio. The combination of Brooklyn, Manhattan, and Queensboro Bridges seems to have the highest feature weights across the board with Queensboro Bridge's feature weight value of 2.19 in the particular combination being the highest magnitude across all combinations of bridges and their respective feature weights. Based on these quantities, the sensors should be installed on the Brooklyn, Manhattan, and Queensboro Bridges for the best overall traffic prediction.

Question 2

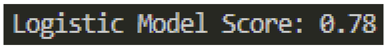




```
[ 1.21146551 -0.38482832 -0.42343608]
Coefficient of Determination: 0.62
```

The plot, normalized model coefficients, and coefficient of determination above sum up our question two results. These findings prove that the overall weather forecast explains about 62% of the bike traffic. In other words, the coefficient of determination proves that this model is a mediocre fit. However, with each individual weight of each feature weight of the model having some low magnitudes, the accuracy of the extrapolation of the correlated number of bikers on the day of the weather forecast will not be completely accurate. Therefore, the next day's weather forecast should somewhat be able to provide rough predictions on the estimated number of bicyclists across the four bridges for that day.

Question 3



```
Logistic Model Score: 0.78
```

For the third question, our results are the logistic model score that we achieved by performing logistic regression on the days it rained/snowed vs. the total traffic. With a score of 0.78, the bike traffic explains 78% of the days it's supposed to rain. Thus, this data can be used to reliably predict whether it is raining/snowing based on the number of bicyclists on the bridges.