Project ECE 20875: Python for Data Science
Spring 2022

1. Project team information

Mini-Project Spring 2022
ECE 20875
Aidan Caputi
GitHub Username: aidancaputi
Email: acaputi@purdue.edu
Path 1 – Bike Traffic

2. Descriptive Statistics

The dataset provides the date, day of week, temperature, precipitation, and traffic numbers on four different bridges throughout the months April through October.

Given our desired analysis, it will be crucial to analyze the traffic numbers (problem 1), and weather information (problem 2/3). The date-related information might not be as crucial in this case, but it will be important to check for trends there as well that may affect the analysis. Additionally, the precipitation data will need to be converted to "rain" or "no rain" for problem 3. We will define no rain as a precipitation of zero and rain as any precipitation above zero.

Below is a summary statistics table for the data given:

Data	Number of observations	Mean	Median	Maximum	Minimum	Standard Deviation
High Temperature	214	74.934	78.1	96.1	39.9	12.545
Low Temperature	214	61.972	64.9	82.0	26.1	11.671
Precipitation	214	0.1091	0.0	1.65	0.0	0.25996
Brooklyn Bridge	214	3030.7	3076.5	8264	504	1134.0
Manhattan Bridge	214	5052.2	5132.0	9152	997	1745.5
Queensboro Bridge	214	4300.7	4342.5	6392	1306	1261.0
Williamsburg Bridge	214	6160.9	6334.5	9148	1440	1910.6
Total for all Bridges	214	18544.5	19001.5	28437	4335	5702.1

Table 1: Descriptive statistics of all data

Below is a histogram for the total traffic on all the bridges combined:

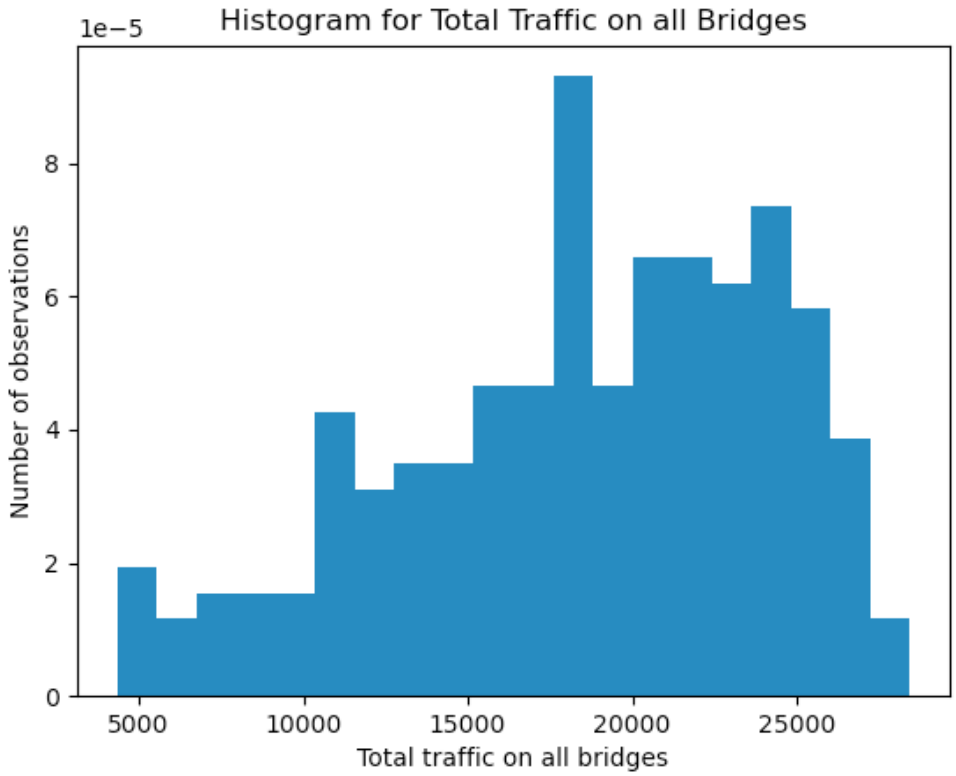

Figure 1: Histogram representation of total traffic data for all the bridges combined

The above histogram is skewed to the left. This is caused by the fact that the mean is lower than the median, which is due to more outliers on the left side pulling the average down while most of data points are towards the right. This is in line with what we would expect looking at this data, for there are weather variables that have the possibility of pulling this number down and when they do, it will be dramatic causing outliers.

Imagine this: there are four nice days of weather in a row with no rain, followed by one day with a torrential downpour all day. This rainy day will create a data point much lower than the other four nice days, pulling the average down despite the median remaining within the four data measurements from the rainless days.

3. Approach

To determine which bridge will not have a sensor installed, we will use a simple polynomial fit to the total bridge data and compare this polynomial with each individual bridge data. For this to be feasible, we will have to do so with all the data normalized so that we can compare the different magnitudes from each bridge.

For question 2, we are going to fit a Lasso Regression model using the data for high temperatures, low temperatures, and precipitation as our x-values and the total traffic on the bridges as our y-values. This model will then be tested and if it is accurate in predicting the total bicycle traffic, we will be able to use it as a predictor based on the next day's weather forecast.

For question 3, it is important to recognize that if it would be possible to predict whether it is raining or not based on the bike traffic, we will not be able to use a typical regression model that predicts a value based on an input. Instead, we will use a special regression type called Logistic Regression that is made for classification (aka. Binary data). In our case, it will be rain or no rain.

4. Analysis

DISCLAIMER: Where applicable, instead of splitting data into testing and training data, I have decided to use all the data given as training data and create my own test scenarios. This is because there are not many data points to begin with (only 214), so I would like to use as many as possible in training to create the best performing model possible. Additionally, I would like to make my own testing scenarios so that I can display results and predictions as clearly as possible (i.e., creating test cases that I believe exemplify the validity of these predictions as opposed to just showing performance data metrics).

Question 1) When observing the raw bridge traffic data, it is obvious that it would be best represented by a polynomial of the second degree due to its shape. In this case, we will use the days as the x-axis values so that we can accurately compare data from each bridge on the same day. This will allow us to see which individual bridges are following most similar trends to the overall traffic.

Below is each bridge's normalized data plotted on top of the second-degree polynomial that was fit to the normalized total bridge data. The coefficients of this polynomial are unimportant as we are not worried about predicting normalized traffic values based on the day, we just want to use a line of best fit to visually compare similarity between the datasets.

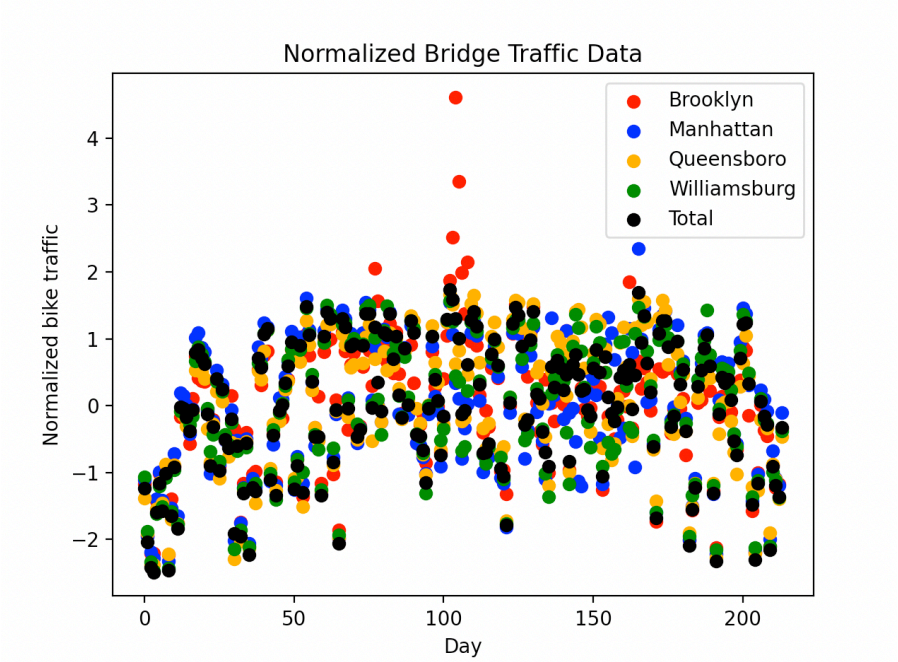

Figure 2: Each individual bridge traffic data plotted on top of total traffic data (all normalized)

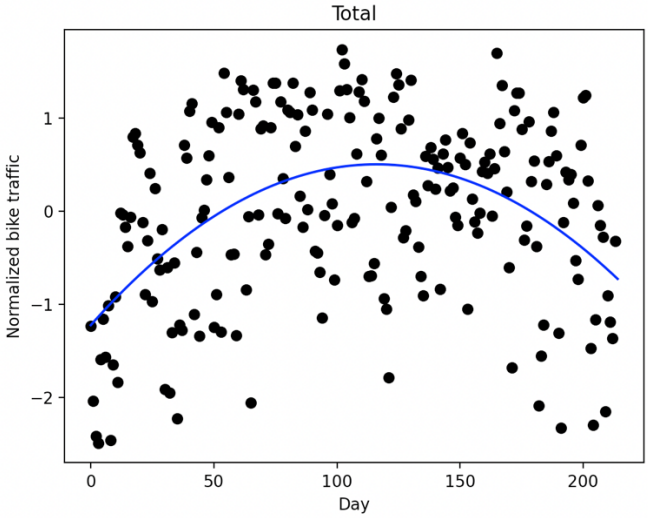

Figure 3: Quadratic curve fit on top of normalized total bridge traffic data

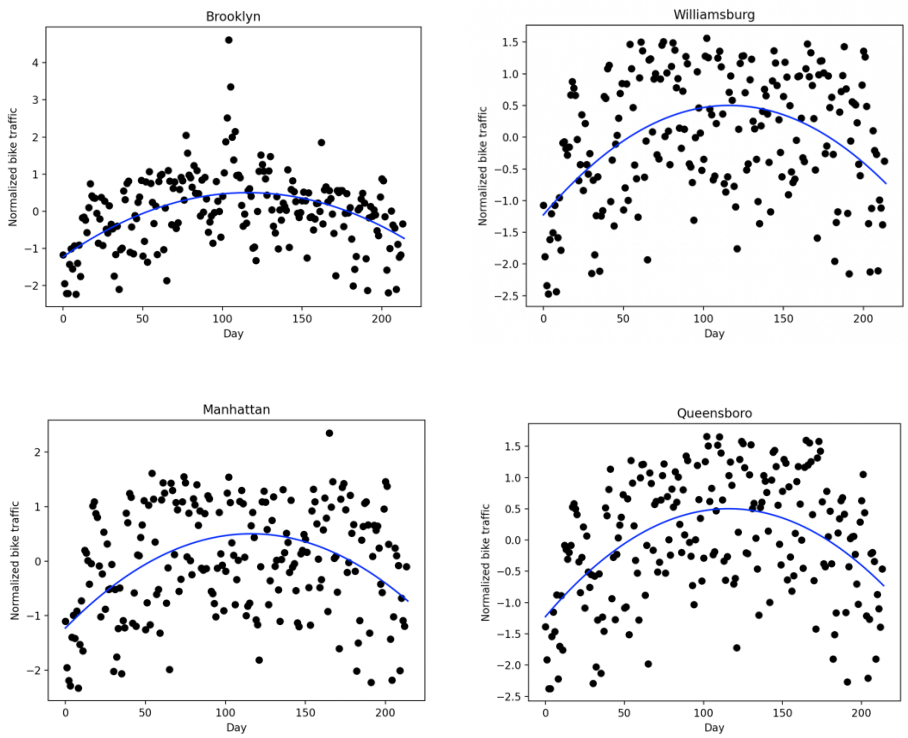

Figure 4: Each individual bridge traffic visualized on top of the curve built to fit the total normalized data

Using figures 3 and 4 above, one can see that the Brooklyn Bridge clearly differs in traffic trends more than any of the other three bridges when compared to the total bridge data. Specifically, it has a unique period of spikes in its traffic at around day 100. If this bridge was used, for example, there could be artificially high false predictions on overall traffic data due to these spikes. For this reason, **sensors should be installed on the Williamsburg, Manhattan, and Queensboro bridges** to most accurately represent the total bike traffic on all the bridges. These three bridges follow a similar trend to the overall data and would be the most accurate predictor of total traffic.

Question 2) Unlike the data from the first problem above, it will not be convenient to visually interpret our results when trying to use weather to predict bicycle traffic. Instead, we will test a model we have trained and observe if it predicted values that we would expect for our test situations that we create.

For this we will use the Lasso regression model to train and test a system for predicting the number of cyclists on a certain day using the weather information. Our feature matrix will consist of the high temperatures, low temperatures, and precipitation levels. Our dependent variable will be the total bike traffic across all bridges.

We use the sklearn python library to fit a Lasso regression model to our feature matrix (X) and output matrix (y), giving us the predictions below:

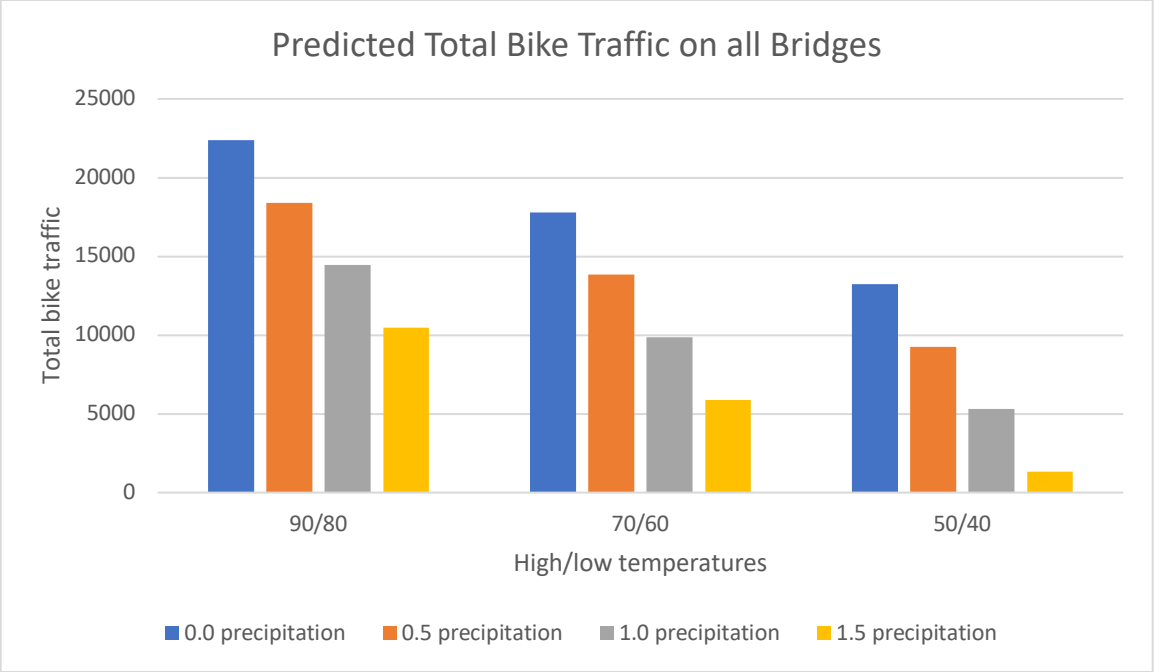

Figure 5: Predicted total bike traffic based on weather (using Lasso regression)

Figure 5 above appears to have the characteristics we would expect from a predictor in this case: the total bike traffic trends downward as the temperatures drop, and the traffic also trends downward within a temperature range as the precipitation increases.

So, now that we have a model with the trends we would hope to have, we can check its r-squared value to see if the model is a good fit for the data or not.

The r-squared produced by the model is around 0.499. This may not seem like a model that predicts well based on the r-squared value. However, this is not the case as human behavior is quite difficult to predict and most of the time human choices are modeled, the r-squared is below 0.5. It is important to consider outside factors that could affect the bike traffic such as:

- the previous weather
    - if it has been rainy and cold for days and then suddenly becomes moderate and partly sunny, more people will probably decide to bike that day since the day will now seem "good" in comparison to the previous days
    - on the contrary, if it has been warm and sunny and then suddenly the temperature drops to moderate and partly sunny, more people might decide not to bike that day since the day will now seem "bad" compared to the previous days
- day of the week
    - more commuters on weekdays
    - more leisure riders on weekends
- wind
    - independent of precipitation and temperature but will still affect bike traffic
- time of year
    - students riding to class during school year but not during summer

- holidays
    - less commuters and less leisure riders

and many, many more. One could add to this list for quite a long time, but that is not the point. As far as the weather data we are given is concerned, the model trained does quite well at predicting reasonable traffic values without access to all the other factors that we know could affect the data.

All in all, we can indeed use the weather for the next day to predict the total bike traffic on the bridges. The model may not be precisely accurate, but it will give a reasonable figure more often than not. Below is the model created:

$$\boldsymbol{total\ traffic} = \mathbf{391.01591073}(\boldsymbol{high\ temp}) - \mathbf{162.42271548}(\boldsymbol{low\ temp}) \\ - \mathbf{7936.18793292}(\boldsymbol{precipitation})$$

Question 3) In this third problem, we opt to use a logistic regression model to classify data points as rain (1) or no rain (0) based on the total traffic numbers from all the bridges.

To test our model and see if it makes reasonable predictions, we will test it on values ranging from zero to 30,000 (seeing as the maximum traffic recorded was around 28,000). We are then going to compare these models to some of the descriptive statistics for the rain traffic and no rain traffic individually.

Below are some of the statistics for the rain and no rain traffic data:

	Rain	**No Rain**
Mean	14320.9	20511.7
Median	13430.5	20952.0
Standard Deviation	5816.6	4415.7
Minimum	4335	9126
Maximum	25999	28437

Table 2: Descriptive statistics of total traffic data separated into rain and no rain

Below we have a visualization of the rain-separated data with the top points being the rain data and the bottom points being the non-rain data. As you can see, the data for when it is raining is generally lower than the data for when it is not raining, and it is also more spread out (hence the higher standard deviation). This could be accounted for by the other weather factors such as temperature that also affect the total traffic (ex. cyclists are more likely to be out when it is warm and rainy than cold and not rainy).

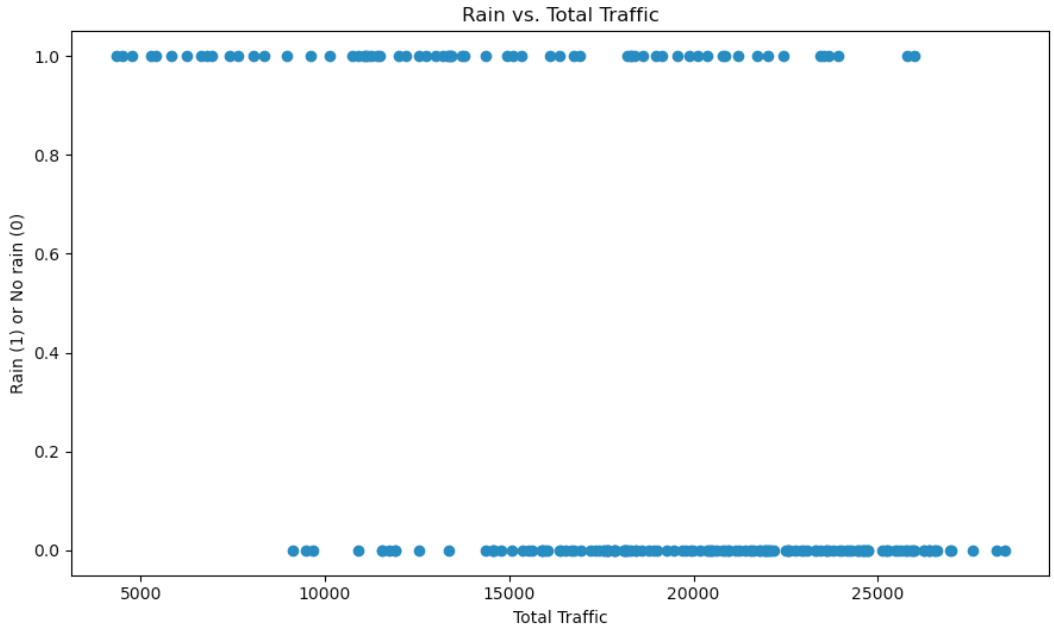

Figure 6: Raining data based on total traffic

We will now observe our model's predictions:

Total traffic	Prediction
30000	No rain
28000	No rain
26000	No rain
24000	No rain
22000	No rain
20000	No rain
18000	No rain
16000	Rain
14000	Rain
12000	Rain
10000	Rain
8000	Rain
6000	Rain
4000	Rain
2000	Rain
0	Rain

Table 3: Logistical Regression model predictions for if it is raining based on total bike traffic

Using tables 2 and 3, we can observe that the model switches between predicting rain and no rain somewhere between 16000 and 18000. Considering that the mean traffic on days without rain was around 20000 and the mean traffic on days with rain was around 14000, these predictions indeed do make sense as 16000 and 18000 are equidistant from 14000 and 20000, respectively.

In simpler words, the prediction decided the transition from rain to no rain was right in the middle of the means of the two, which would make sense for this scenario. We can now go back and run another test with values between 16000 and 18000 to find a more specific transition point.

Total traffic	Prediction
18000	No rain
17800	No rain
17600	Rain
17400	Rain
17200	Rain
17000	Rain
16800	Rain
16600	Rain
16400	Rain
16200	Rain
16000	Rain

Table 4: Logistical Regression model predictions for if it is raining based on total bike traffic (smaller range)

Using table 4 above, we can determine that the transition between raining and not raining predictions would be at a total traffic value around 17,700, with over being no rain and under being rain.