Name: Dorien Penebacker
Github Username: dpenebac
Purdue Username: dpenebac
Path 1
Instructor: Mahsa Ghasemi

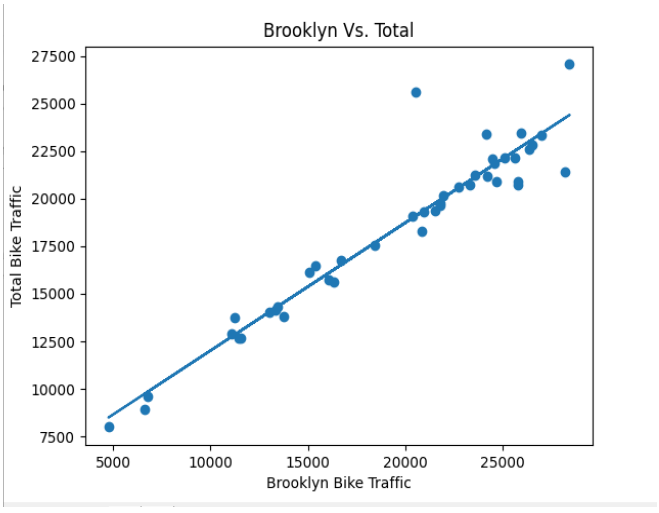
The data set I have been given is a csv file containing information about bike traffic in what I assume to be Europe, I honestly don't know, between April 1st and Oct 31st in 2016. There are quite a few metrics given in this csv file. Each is represented in their own format, but they each represent the data appropriately. This data is the temperature high/lows, precipitation levels, what day of the week it is, and total bike traffic per day. There is also the total bike traffic separated to each bridge and their sum is equal to the total amount of bike traffic column.

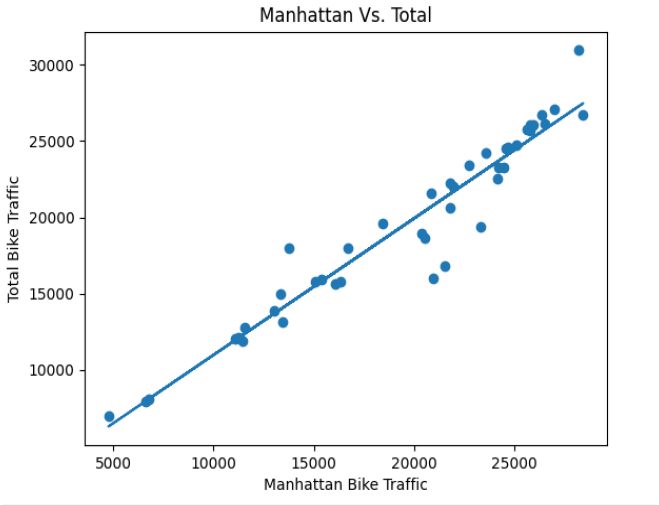
Problem 1

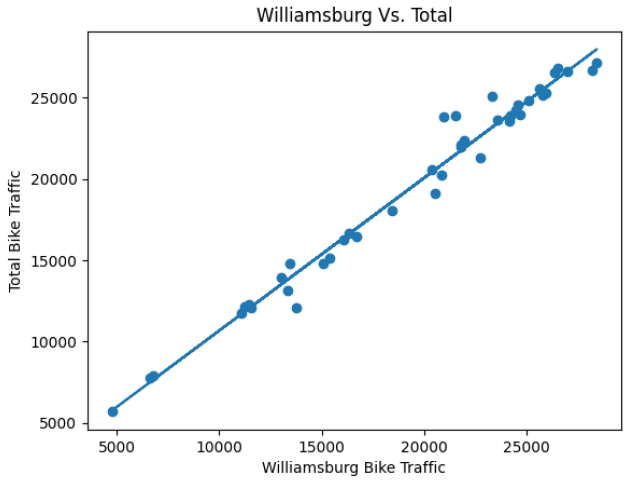
Prompt:
You want to install sensors on the bridges to estimate overall traffic across all the bridges. But you only have enough budget to install sensors on three of the four bridges. Which bridges should you install the sensors on to get the best prediction of overall traffic?

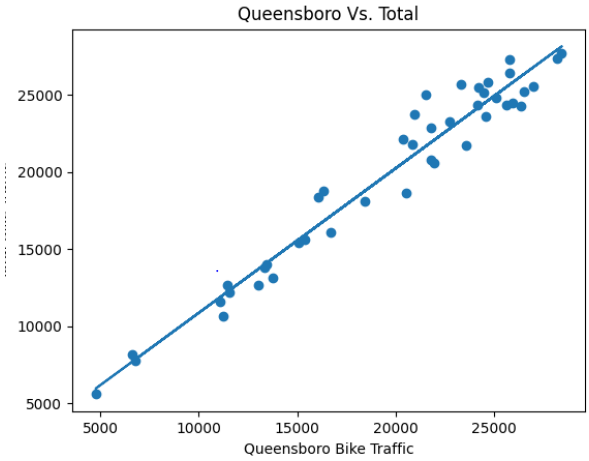
To determine which bridges we want to use to estimate overall traffic across each bridge, we need to choose three of the four bridges to use. To do this, we must determine which bridge is the least likely to estimate overall traffic across each bridge and remove that from the list of bridges we want to use.

To choose which bridge to remove from the list, I created four linear regression models where the amount of bike traffic in each bridge is the feature variable, and the total amount of bike traffic per day is the target variable.









Below is the best fit line information and the coefficient of determination (r2) for each bridge:

Brooklyn
y = 4729.05042791 * x + 18286.74269006
r2 = 0.8162226795024599

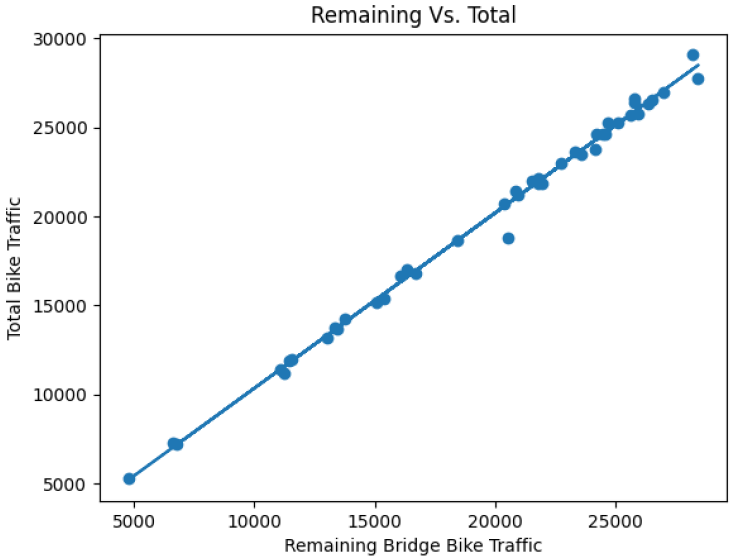
Manhattan
y = 5091.48969409 * x + 18286.74269006
r2 = 0.9273759337009827

Williamsburg
y = 5339.11091244 * x + 18286.74269006
r2 = 0.9765915927612067

Queensboro
y = 5273.14500177 * x + 18286.74269006
r2 = 0.9529271496363251

As we can see from the graphs along with the corresponding r2 scores for each bridge, we can interpret that Brooklyn, having the lowest r2 score of 0.81, estimates with the least accuracy, the amount of total bikers for each day over the entire data set.

Along with that interpretation, we can also see what happens when we remove Brooklyn from the set of features and perform another linear regularization, where the feature variables are the three remaining bridges, and the target variable is the total amount of bike traffic.



Remaining {'Queensboro', 'Williamsburg', 'Manhattan'}
y = 0.9826265854889314 * x + 536.876749631949
r2 = 0.9943639891703062

When performing a linear regularization using all the bridges, we should approximately get an r2 score of 1, where the data matches the linearization exactly. As we can see, by removing only Brooklyn from the list of features, the r2 score only decreases by .005, which means that the data still represents with about 99% certainty, the total bike traffic over all the bridges.

Therefore, we can choose Brooklyn to be the bridge that we do not choose when placing three of the four bridges to place cameras, and we get the best prediction of the overall traffic while only using data from three of the four bridges.

Name: Dorien Penebacker
Github Username: dpenebac
Purdue Username: dpenebac
Instructor: Mahsa Ghasemi

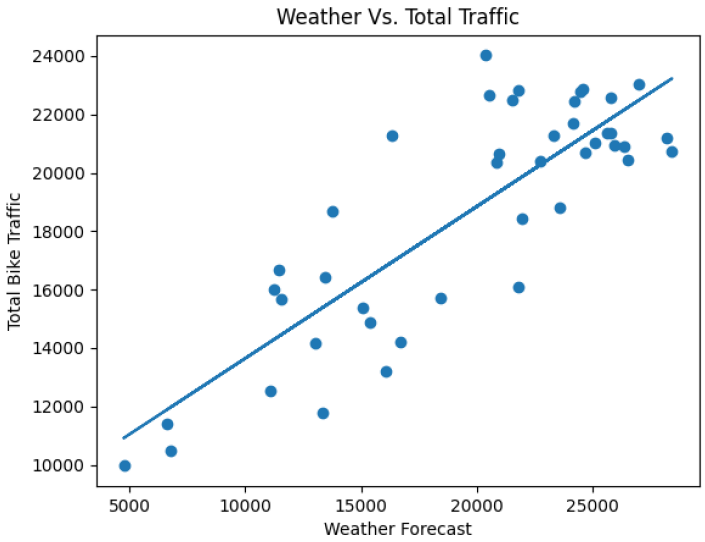
Problem 2

Prompt:
The city administration is cracking down on helmet laws, and wants to deploy police officers on days with high traffic to hand out citations. Can they use the next day's weather forecast to predict the number of bicyclists that day?

Similarly to problem 1, we can use a linear regression model to try and determine whether the weather can predict the number of bicyclists for the day.

To create this model, I used the High, Low, and Precipitation columns from the dataset and used those as the feature variables, and once again used the total amount of bikers per day as the target variable. For the precipitation, when I saw trace of rain, I simply set it to be 0 as trace amounts of rain is likely a slight drizzle and won't affect the overall biking traffic. With snow I simply converted it to be 10 times the amount of rain since snow can be more of a hindrance to bikers than rain normally is due to blocked roads etc.

Below is the graph and information:
High Coef: 4151.07764602
Low Coef: -1421.43547379
Precipitation Coef: -1984.52809033
R2, score 0.6350768663477959



Because the data is normalized we can make a couple of assumptions regarding the graph and the coefficients and how they weigh against each other. We see that both a higher amount of precipitation and a lower temperature causes there to be less bikers overall as they are both negative values. Along with a higher temperature causing there to be more bikers overall because it is a positive value.

To go back to the original question we mainly look at the r2 score value. With r2 being a lower score, we cannot assume that the weather specifically can absolutely predict the amount of

bicyclists total for each day. The main problem with this model is that we are assuming every biker changes their plans for the day based on the weather. However, some people choose to, or even have to, bike every day, along with those who only bike on weekends

However, by using the coefficients given, we can make a couple assumptions and since r2 is still greater than 0.5, we can use the model to get a “loose bound” assumption of the amount of bikers there are based on the weather. This means we can predict with about a 65% certainty the amount of bikers based on the weather forecast.

Name: Dorien Penebacker
Github Username: dpenebac
Purdue Username: dpenebac
Instructor: Mahsa Ghasemi

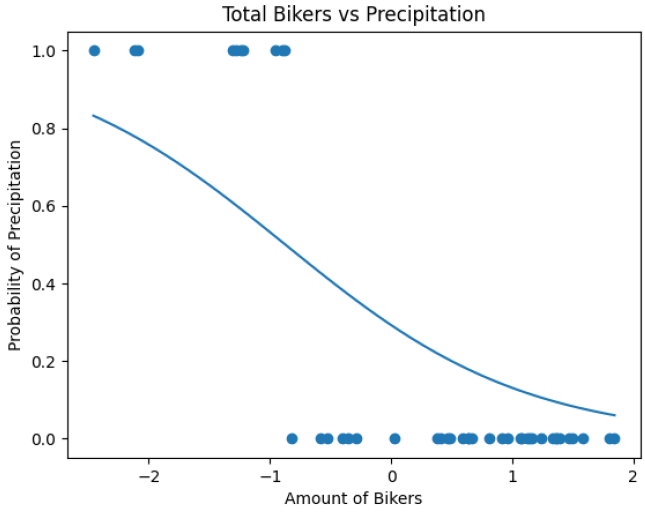
Problem 3

Prompt:
Can you use this data to predict whether it is raining based on the number of bicyclists on the bridges?

To predict whether it is raining or not based on the number of bicyclists, rather than using a linear regression, we can use a logistic regression instead. This logistic regression will tell us, depending on the amount of bikers, whether there is a high or low probability of rain.

The same idea is applied. However, now our feature variable is the total number of bicyclists on the bridge and our target variable is the amount of precipitation for each day. To perform logistic regression, I changed each value in the precipitation matrix to a 0 or a 1. If it was raining or snowing I set the value of that day to be a 1 and if it was raining or not raining, I set the value of that day to be 0.

Log Score 0.8372093023255814



I calculated the score using logreg.score(X_test, y_test) which computes the percentage with which the data can be accurately used, similarly to the r2 score used in linear regression.

What this graph is showing, is that, as the amount of bikers increase, the probability of any precipitation decreases. Note, the data for the amount of bikers is normalized and has yet to be normalized. We also can see vice versa, that as the probability of precipitation increases, the amount of Bikers decreases.

Based on the high logscore along with the graph, we can properly assume that based on the number of bicyclists, we can predict up to an 83% certainty that it is raining or not.