ECE 20875 Mini Project
Path 1: Bike Traffic

Authors: Agrim Bharat (bharata)
Kahaan Patel (pate1410)

Introduction:

For this project, we chose the first path. The first path asks us to analyze a dataset that tells us about the total number of bicyclists on four major bridges of New York. These bridges are the Brooklyn Bride, the Manhattan Bridge, the Williamsburg Bridge, and the Queensboro Bridge. The dataset also has the highest and lowest temperature (in degree Fahrenheit) for every single day of the 7 months it describes.

Data Overview:

The dataset as mentioned above gives us the day, date, precipitation, and highest and lowest temperatures for all seven months. However, there are certain discrepancies in the dataset. Certain erroneous measurements are listed as 0.47s and T. Both these entries were nullified to 0 so that our analysis would not be thrown of due to measurement errors. The datatype for the temperatures and traffic was integer, and we changed it to float so that we could have more complex mathematical computations.

Analysis:

1. ***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?***

For the first part, we need to find which three of the four bridges should have sensors installed on them to best predict the number of bicyclists on that day. For this, we want to know which of the four bridges contributes the least to the total number of bikers every day.

We do this by finding the probability of a randomly selected biker being from one of the four bridges. We do this by summing all the bikers on each bridge throughout the seven months and then divide it by the total number of bikers that were recorded over the four bridges throughout this time. The results for this will show us which of the bridges contributes the least number of bikers to the total traffic. We can plot this probability distribution as a pie chart to make it easy to comprehend.

Using this pie chart, we can find out which of the three bridges contribute the most amount of traffic to the total number of bikers. If we accurately measure the traffic on the three most significant bridges using sensors and estimate the traffic on the fourth bridge using a mathematical model, it will lead to the least amount of error in our prediction as its contribution to the traffic is the least amongst all the bridges. Ignoring any of the other bridges would lead to a large error as they contribute a much larger number of bikers to the total traffic.

2. *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?*

For the second part, the city wants to enforce helmet laws for bikers. For this they want to deploy police officers on days with high traffic to hand out citations. They want us to find whether we can predict the number of bicyclists on a particular day using the weather forecast for that day. The data from the weather forecast that has been provided to us is the highest and lowest temperature for that day along with the amount of precipitation.

For this, we first need to find how these three parameters individually affect the total number of bicyclists. This can be done by plotting three different scatterplots with the each of the three parameters on the x axis and the total number of bicyclists on the y axis. We can start by analyzing these plots to find whether there is any direct and obvious correlation between the total number of bicyclists and the weather parameters. If we find correlation of total traffic with more than one parameter, we can find the beta value by finding this MSE value. This way we can use regularization to set up the independent parameters as features and the total traffic as the target. This yields a mathematical model that can be further improved using training and testing. Thus, we get a good mathematical representation of how the traffic is affected by each of these parameters. This enables us to find a good prediction of the total traffic based on the significant features.

3. *Can you use this data to predict whether it is raining based on the number of bicyclists on the bridges?*

For the third problem, we need to find whether the amount of rain can be predicted using the total biker traffic on all the bridges. For this, we can start by plotting the precipitation on the dependent variable axis and total traffic on the independent axis. Then we can repeat the analysis of step 2 to find whether there is an acceptable mathematical model to predict rain. We can do this by using regularization. This time we only have one feature which is total number of bicyclists, and the target

value is the precipitation. We can then train and test this model to make it fit better, and to finally test whether we can make an accurate prediction of rain based on the traffic, we can calculate the r^2 score to check whether our model fits the dataset with an acceptable amount of error. The following Lambda solution was viable for a single feature , hence using the Linear Regression from sklearn to find the linear regression of the precipitation and total in correspondence

Result:

1. ***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?***

The analysis yields the pie chart shown below. As can be seen, the probability of a randomly selected bicyclist being on the Brooklyn Bridge is the lowest amongst all the four bridges. Therefore, we can conclude that we should install sensors on the Manhattan Bridge, the Williamsburg Bridge and the Queensboro Bridge. This is because these three bridges contribute the most (84%) to the overall bike traffic on these bridges. We do not install sensors on the Brooklyn Bridge as it only contributes to 16% of the overall traffic. We can use the data gathered by the sensors on the other three bridges and model the data for the fourth bridge to predict the total amount of traffic most accurately. As the contribution of the Brooklyn Bridge is the least, estimating the traffic on it and accurately measuring the traffic on the other bridges will lead to the smallest amount of error in our final prediction.

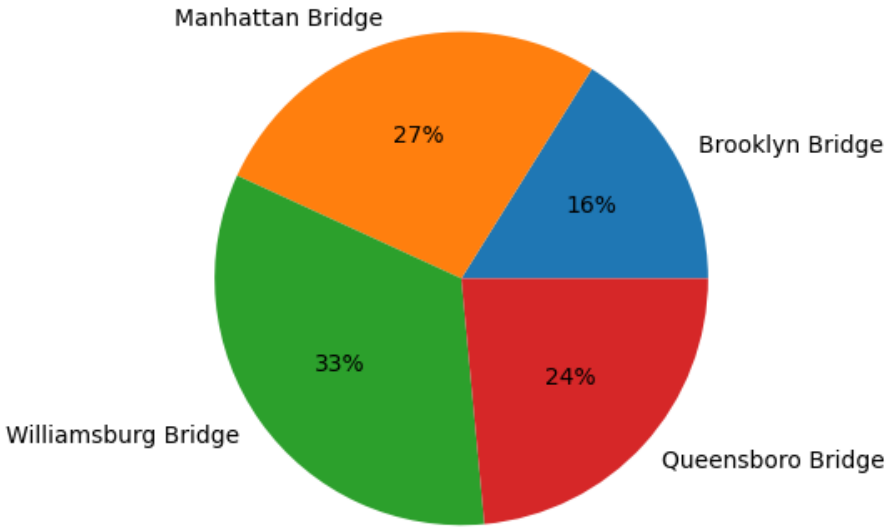

Figure 1 : The following is the pie chart with the probabilities of bikers on each bridge.

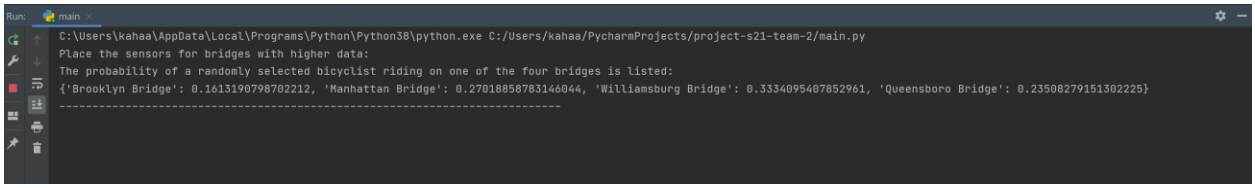

Figure 2 : The following is output screen visible on running the code.

2. ***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?***

Using the analysis described above, we plotted the three scatter plots shown below. Firstly, upon analyzing the scatter plot for total number of bicyclists against the highest temperature of the day, we can see a direct positive correlation between the two. As the temperature increases, the total number of bicyclists on that day also increases. We can explain this correlation by the fact that New York is a relatively cold state, so during the warmer months from April to October, people like to come out and bike, and as the days grow warmer, more and more people come out to ride a bike. At the bottom of the plot, we can see that when the highest temperature is between 40- and 50-degrees Fahrenheit, the total number of bikers is only around 5000, but on the warmest days this number starts touching 30000 bicyclists.

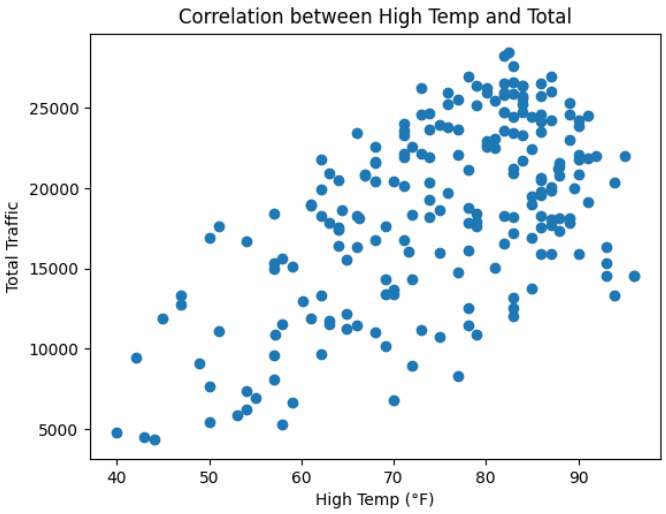

Figure 3 : Correlation between High Temperature and Total Traffic.

Secondly, the plot with total number of bicyclists against the lowest temperature also shows the same results as the first plot. Even here, we can see that as the days grow warmer, the total number of bicyclists increases. On cold winter days, the number of bikers is small but on days where the lowest temperatures are also relatively high, more people come out to ride bicycles.

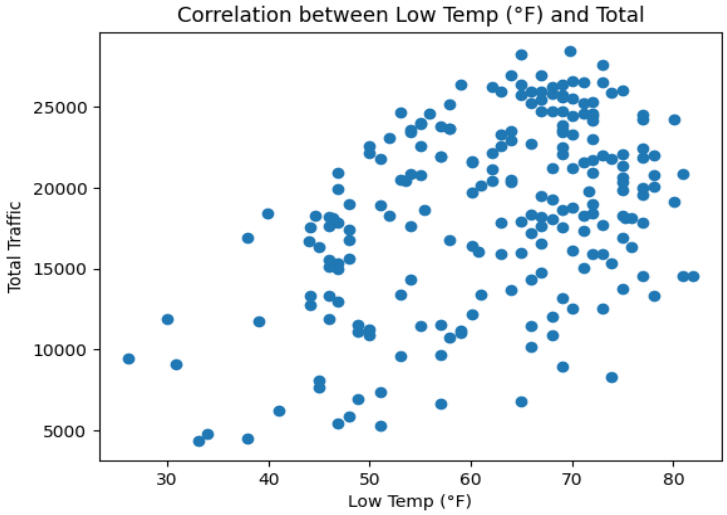

Figure 4 : Correlation between Low Temperature and Total Traffic.

The third scatterplot is plotted against the amount of precipitation and the total number of bicyclists that day. Upon analyzing this graph, we do not see any direct correlation, neither negative nor positive, as we did in the previous two scatterplots. But upon closer inspection, we can conclude that as the precipitation tends to 0, the number of bicyclists increases manyfold, and on days with more precipitation, the number of bicyclists show no correlation with the amount of precipitation. Using this information, we can conclude that most bicyclists only ride their bikes on days with little to no rainfall, and on days with rainfall, the number of bicyclists depends on precipitation. r factors.

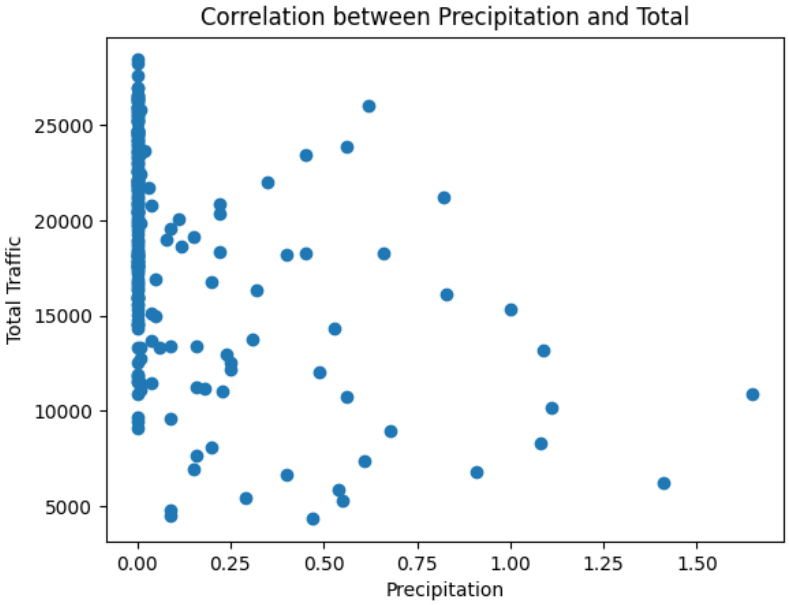

Figure 5 : Correlation between Precipitation and Total Traffic.

Using the results above we can conclude that all three parameters affect the number of bicyclists on the bridges. Now we can use regularization on the data and find multipliers for each of the three parameters to find a fitting mathematical model for the total traffic. We then use training and testing on this model to make it better, and this yields the following predictions for the total traffic on each day. To explain outliers, we find the MSE, which comes out to be 13975003.01. We can predict that the measured traffic is within $\pm\sqrt{MSE}$ i.e., 3738.31, of our predicted value.

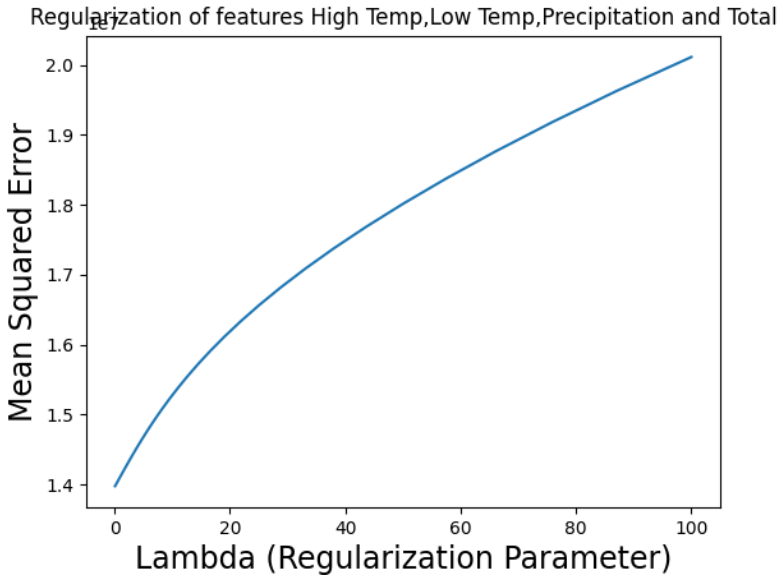

Figure 6 : Regularization of the features in correspondence to the target.

```
predicted traffic by normalizing data [19770.75797903 12761.90518943  9775.8195598   8617.53721473
 12255.89666546 12776.96093267 13360.65842689 11527.3229163
 10368.18549211 14049.00266124 16696.41901071 13622.73723638
 15753.3126899  16953.50350359 17696.61417796 18291.29008813
 20976.21614738 22718.48698645 17742.66773275 18346.85817927
 19452.03398823 20238.22601103 16445.50566063 18609.83760319
 17428.74017341 14186.0870081  16651.07716608 15009.10351595
 14467.89022555 17183.80430131 11419.1745426  15372.24651338
  8758.17027583 12393.50726587 14832.88470618  9355.92316287
 14890.76277519 16073.70003229 19115.44292135 16098.82037585
 20042.63919246 21229.68103495 14088.66387636 18210.19926179
 14496.5682642  18178.50615648 16353.383763   17873.49521622
 18531.60600451 20184.96789776 16684.39738403 17674.40916507
```

Figure 7 :Output of the predicted traffic bases on the features . The order is from left to right that corresponds to top to bottom in the excel.

3. *Can you use this data to predict whether it is raining based on the number of bicyclists on the bridges?*

Using the analysis explained above, we first start by plotting the precipitation on the dependent axis and the total traffic on the independent axis to see whether there is any obvious dependence of the precipitation on the number of bicyclists. Upon visually inspecting the graph shown below, we can see that there is no such obvious correlation between the two measurements, which is unlike the apparent correlations we found in part 2. We use regularization of parameters to try to find a mathematical model. Even after training and testing, , we see that the error between the true measured values and our predicted values is too large for the model to be of any practical significance. We then calculate the r^2 score that comes out to be 0.18, which shows that even after training and testing, we cannot predict

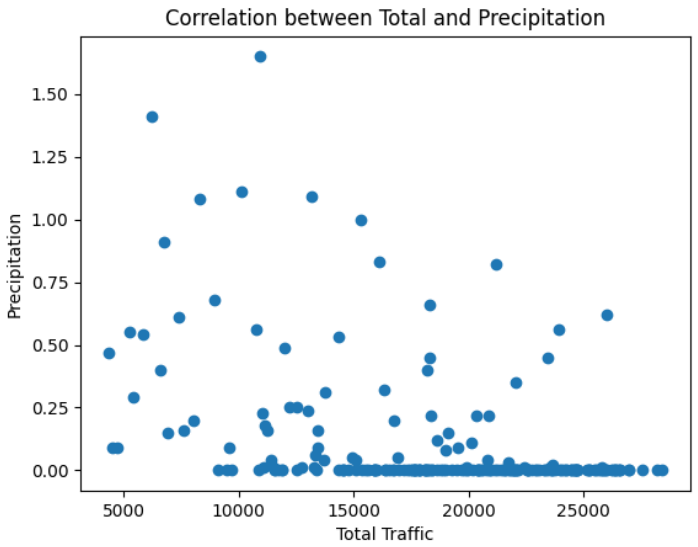

Figure 8 : Correlation between Precipitation and Total Traffic. Precipitation as the dependent

the rainfall based on the bicycle traffic with any significant accuracy. On doing linear regression we get the graph obtained in figure 11, that shows the training data and test data set , from the following results we can infer there isn't a accurate connection between the Total traffic and precipitation.

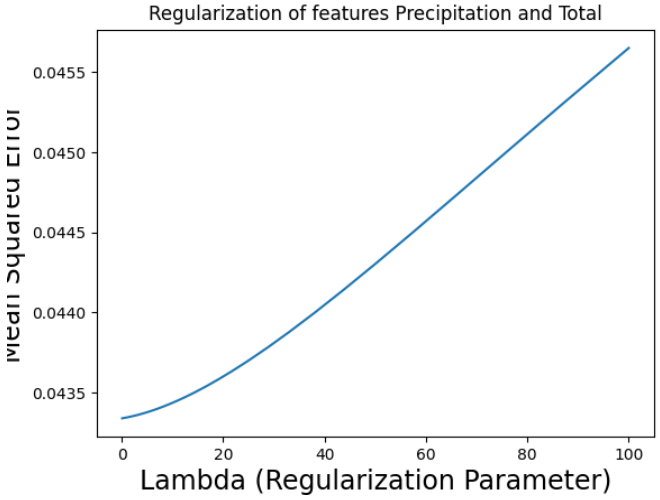

Figure 9: Regularization of Feature and Target

```
predicted precipitation by normalizing data [ 2.37633025e-01  3.24425136e-01  3.65459308e-01  3.73502992e-01
 2.76068176e-01  2.29627282e-01  2.73696807e-01  2.13976246e-01
 3.70183076e-01  2.82613155e-01  2.03200745e-01  3.02703395e-01
 1.06259173e-01  1.08289065e-01  1.22422425e-01  1.45092714e-01
 1.10850144e-01  1.77027629e-02  1.33773856e-02  2.71692686e-02
 3.59717910e-02  1.17357181e-01  2.01038056e-01  1.38168317e-01
 6.01976984e-02  2.09024827e-01  7.79545107e-02  1.59207104e-01
 1.72050439e-01  1.25040417e-01  3.11012672e-01  1.69413477e-01
 3.15357021e-01  2.45069638e-01  1.64196464e-01  3.44970679e-01
 2.36399913e-01  2.42242966e-01  2.74158910e-02  4.22132346e-02
-1.20247209e-02 -2.10169528e-02  2.24258503e-01  1.51675635e-01
```

Figure 10: Output of the predicted precipitation of the model based on testing and training.

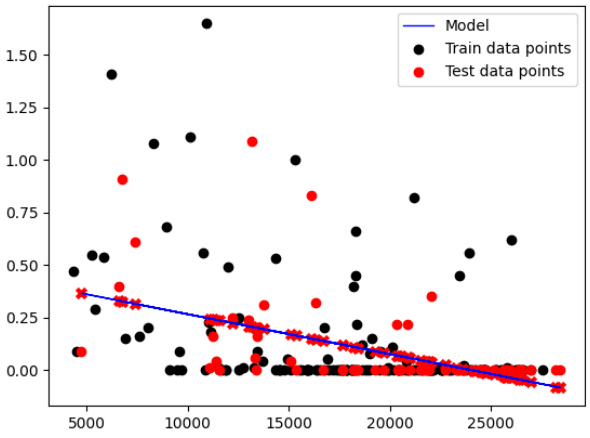

Figure 11: Training and testing the model