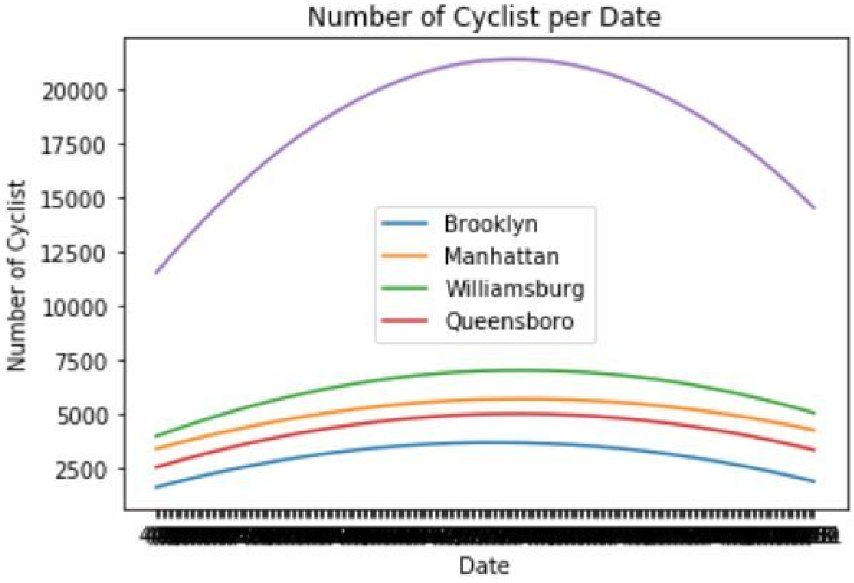
Santiago Guada
Marc Zabit
ECE 20875: Python for Data Science
Professor Milind Kulkarni
Final Project: Path 1: Bike Traffic in New York City

Problem 1

This group has been tasked with estimating the overall bike traffic on 4 bridges in New York City, the Manhattan Bridge, the Brooklyn Bridge, the Queensboro Bridge, and the Williamsburg Bridge. The problem is, there are only enough sensors to be installed on three of the four bridges.

Figure 1

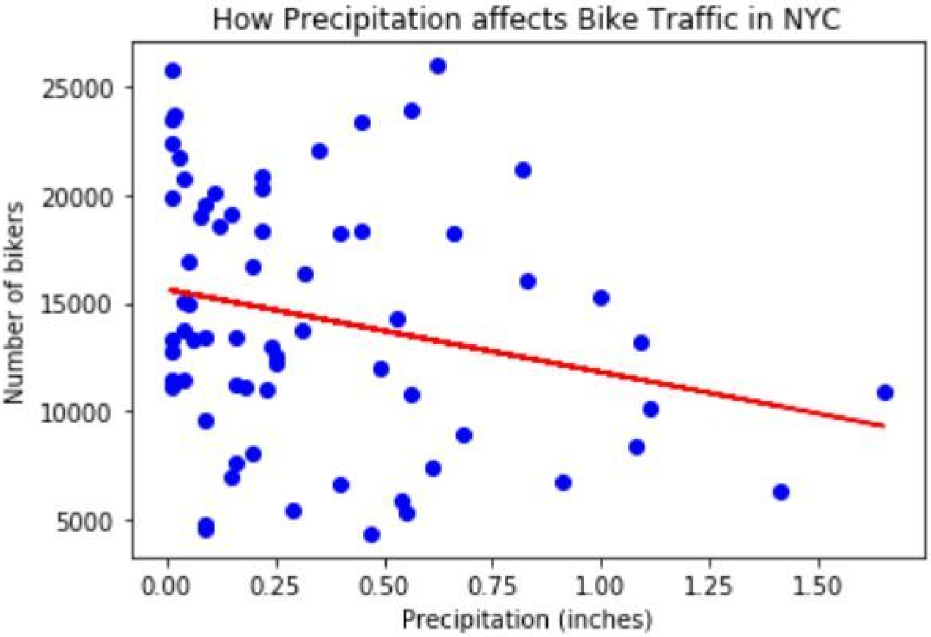


As can be seen in Figure 1, of the 4 bridges that could be monitored the Brooklyn Bridge has the least overall traffic. The job of this group is to predict overall bike traffic through the city so that the municipal government can deploy its resources appropriately to deal with the traffic. Since all of the bridges cannot be covered with sensors, the most important bridges that have the most traffic should be covered so that the most good can be done for the most number of people with the resources that are available at this time. Therefore, it would be the recommendation of this group that the Manhattan, Williamsburg, and Queensboro bridges receive the sensors as on every day measured these have been the top bridges in terms of traffic. On not one day has the Brooklyn Bridge ranked above any of the other bridges, so it should be the bridge that is left out. This way the best estimate can be observed of the bike traffic flow of the city. If any other combination of bridges was chosen, the estimate of the bike traffic would not be as accurate and fewer people would be helped by the city's policies.

Problem 2

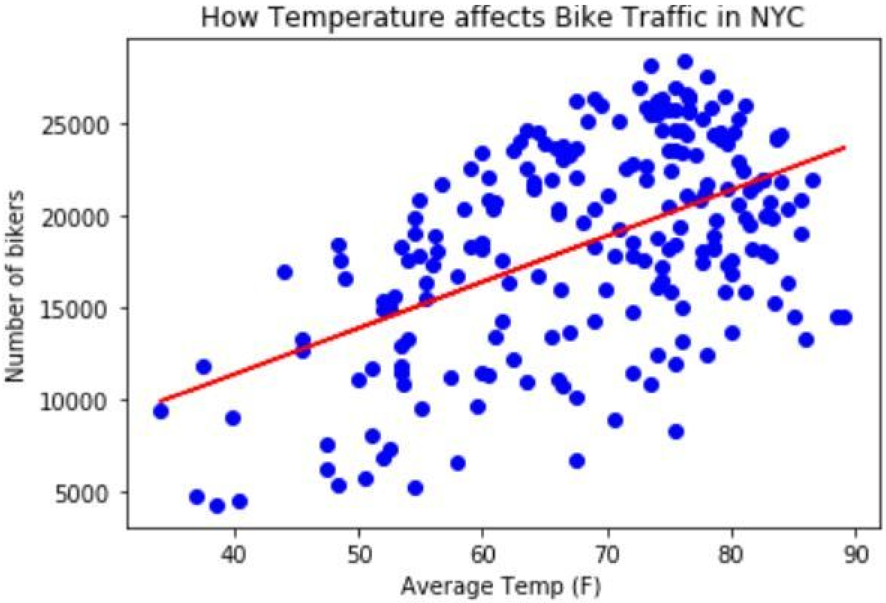
The next problem is about using the next day's weather forecast to predict bike traffic so that the city can decide how many police officers to deploy. This group took an approach to solving this problem by separating the weather forecast into two distinct parts. The distinction is between whether it is raining or not raining outside. If it is raining, than the number of bikers will most likely be determined by the amount of rain that is going to fall. There is a negative linear relationship between these two parameters that is shown in Figure 2.1. The equation for this linear relationship is y = (-3834)x + 15638, where x is the amount of precipitation in inches and y is the amount of bike traffic.

Figure 2.1



If there is no rain and the weather is clear, then the factor that will determine if people are to go biking is most likely the temperature. There is a positive linear relationship between these two factors as shown in Figure 2.2. The linear relationship described between temperature and bike traffic on clear days is described by the equation y = 251x + 1378, where x is the temperature in degrees Fahrenheit and y is the total number of bikers.

Figure 2.2



Problem 3

For this problem, we were required to predict whether it is raining in the city of New York based on the number of cyclists on the bridges. For this type of analysis, we have to make our explanatory variable the number of cyclists in the city and the response variable will be a two-available response of 1 or 0. The reason behind comes from the question being asked, we need to predict whether it is raining or not, so it is a yes or no response instead of how much rain there will be in inches. We will be using 1 for when the amount of precipitation exceeds 0.25, which we considered being a fair amount of rain to be considered for this data, and 0 for when it is less than 0.25.

We decided to predict our result with the Logistic Regression classifier since it demonstrates the final goal correctly, emphasizing our two outputs and the number of cyclists that day. We created a code to separate the data into raining or not raining, as explained above and then separated the data into training and test data. The idea for this is to use the training data to get the model and measure the effectiveness of it with the test data. Then we used sklearn LogisticRegression to get the model and the prediction of our model, which is where we encounter our problem. We used the predict command of the LogisticRegression to then plot the values and get the curve plot, but the command gave us an output of all 0, no matter the size we used. There is nothing wrong inherently, but the difference between the features of raining days and not raining days is too high. Looking at Figure 3.1, we can see that the not raining output (0) has a much greater weight than the raining output.

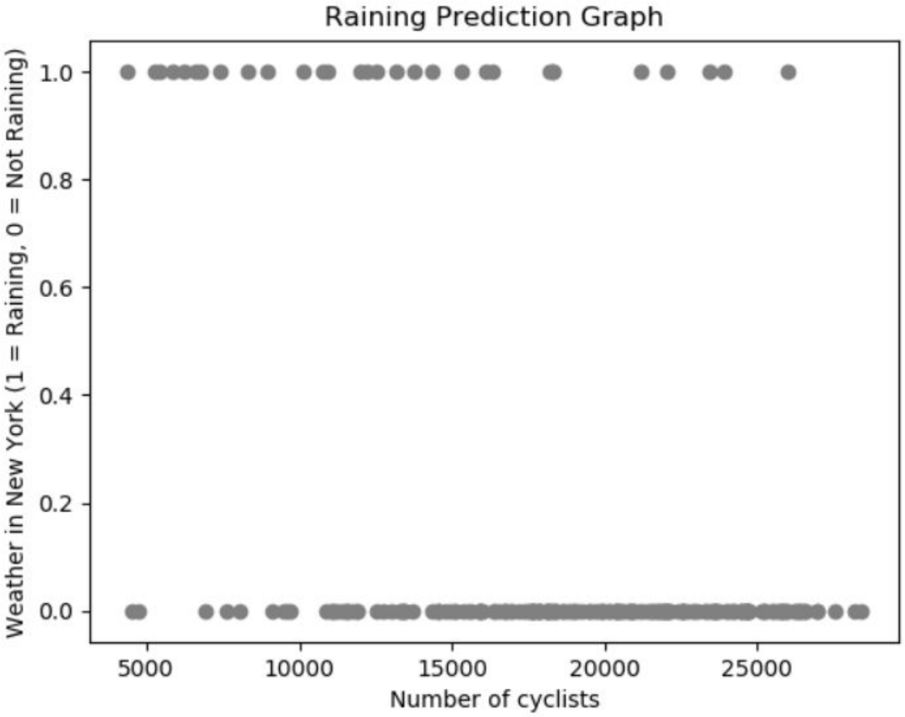
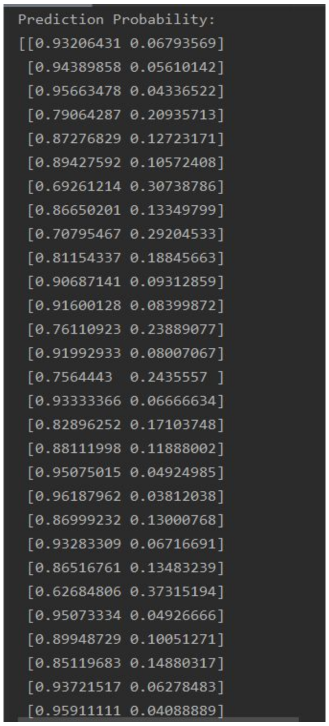


Figure 3.1

Also, we decided to print the predicted_proba, which is the predicted probability of where the point will be classified, and the whole data was favoring the first classifier (not Raining = 0), in Figure 3.2 we can see a small representation of this explanation.



```
Prediction Probability:
[[0.93206431 0.06793569]
 [0.94389858 0.05610142]
 [0.95663478 0.04336522]
 [0.79064287 0.20935713]
 [0.87276829 0.12723171]
 [0.89427592 0.10572408]
 [0.69261214 0.30738786]
 [0.86650201 0.13349799]
 [0.70795467 0.29204533]
 [0.81154337 0.18845663]
 [0.90687141 0.09312859]
 [0.91600128 0.08399872]
 [0.76110923 0.23889077]
 [0.91992933 0.08007067]
 [0.7564443  0.2435557 ]
 [0.93333366 0.06666634]
 [0.82896252 0.17103748]
 [0.88111998 0.11888002]
 [0.95075015 0.04924985]
 [0.96187962 0.03812038]
 [0.86999232 0.13000768]
 [0.93283309 0.06716691]
 [0.86516761 0.13483239]
 [0.62684806 0.37315194]
 [0.95073334 0.04926666]
 [0.89948729 0.10051271]
 [0.85119683 0.14880317]
 [0.93721517 0.06278483]
 [0.95911111 0.04088889]
```

Figure 3.2

Even though there is no predicting line, we can see from the graph in Figure 1 that when the total number of cyclists on the bridges is approximately 10,000 or less it can be considered a rainy day and vice versa with no rainy days.