Name: <Lin Gan, Yu Liang>

Purdue Username: <gan30, liang289>

Instructor: <Inouye>

Path: Path 1 Bike Traffic

Dataset: The dataset consists of several elements including weather, date, and the bike traffic for each of the NYC bridges for 214 days. From the dataset we could analyze the correlation between the weather and the total traffic, as well as the effect of each bridge to the total bike traffic.

Analyses:

Question 1.

To determine three sensors among four, we generate four models to observe their performance. We plot each bridge (e.g., Brooklyn Bridge) versus the total traffic to have a best fit model for each of the four bridges. If the model shape looks like a linear model, we expect to know the r^2 and the linear model $y = kx + b$ by using a polyfit function. However, if the model looks like nonlinear model, we use regression functions to calculate r^2 and the predicted model $y_d(x) = a_d x^d + a_{d-1} x^{d-1} + \ldots + a_1 x + b$. Finally, we choose three models that have the highest r^2 values.

Question 2.

For this question, we are asked to predict the number of bicyclists on the bridges by weather based on the dataset. What first comes up to my mind is there should be a positive linear relationship. Therefore, I take the average of every day’s highest temperature and lowest temperature as my whole day's temperature. I used polyfit for the data, which x value is the temperature and y value are the total bicyclists on the bridges and get the plot below. The model is y = 250.78245271x + 1377.71209967.

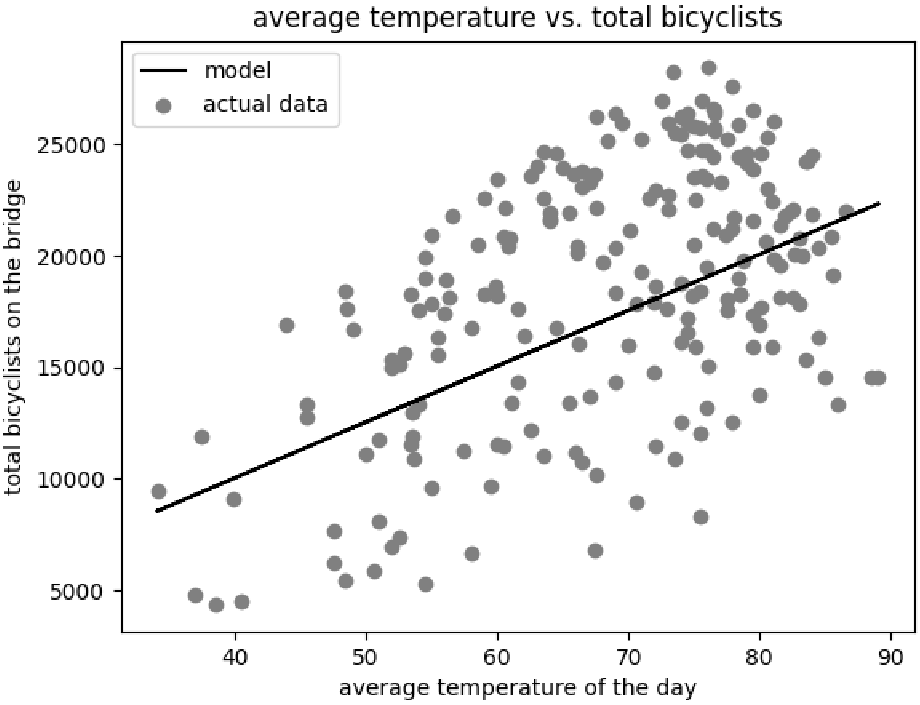

Figure 1. Average temperature versus total bikes

As you can see, the model does not predict the data well, the r score we get is 0.213310371737907, which is a weak correlation.

By looking into this plot, we found out that there are more people in the temperature section from 70-80 degrees Fahrenheit, when the temperature goes higher after reaching 80-degree Fahrenheit, the number of bicyclists goes down.

Consequently, my new plan is to separate the temperature into two sections, one with an increasing model (lower than 70-80 degrees Fahrenheit), another one with a decreasing model (higher than 70-80 degrees Fahrenheit). Also, I separate the average temperature back to the highest and the lowest temperature for each day. Thus, I'll get 2 (increasing & decreasing) * 2 (highest temperature & lowest temperature) = 4 models in total. For the temperatures that are used for separating increasing and decreasing sections, I use weighted averages to determine it. For the highest temperature's two models, I use each day's highest temperature times the total bicyclists on the bridge and add the products together. Then use the sum divided by the average of 214 days' total bicyclists then use the quotient divided by the total number of days which is 214, to get the weighted average for highest temperature in 214 days. Also do the same thing for lowest temperature. The highest and lowest temperature of the day will determine which two models are going to be used. Then for each day, use the average of the two numbers of bicyclists I have from the highest & lowest models as our prediction for the total number of bicyclists on bridges.

Question 3.

We use confidence intervals and clusters to predict whether it is raining or not based on the total traffic. Utilizing Z test to construct a 99% confidence interval to maximize the range of the total bikes included. After that, calculate the lower and upper bound of the interval in the form of $\mu \in (\underline{x} - Z_c * SE, \underline{x} + Z_c * SE)$. Where x is the average total bikes for raining days and no rainy days. Finally, use clusters to determine the centers of the data to verify the capability of our results. We expect to see two cluster centers based on the density and variance of the dots shown in the figure below (figure 1), and the range of intervals. The interval means if the number of total traffic is given in a certain range, we are possible to predict if it is raining or not.

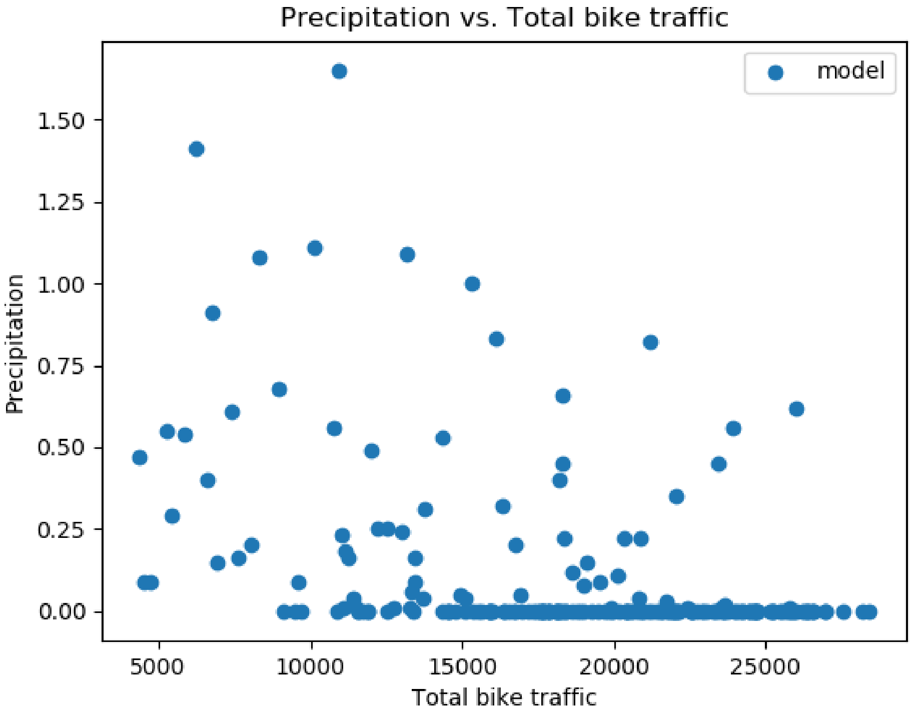

Figure 2. Scatter plot of precipitation and total traffic

Result:

Question 1.

Our results show that we obtain four linear models.

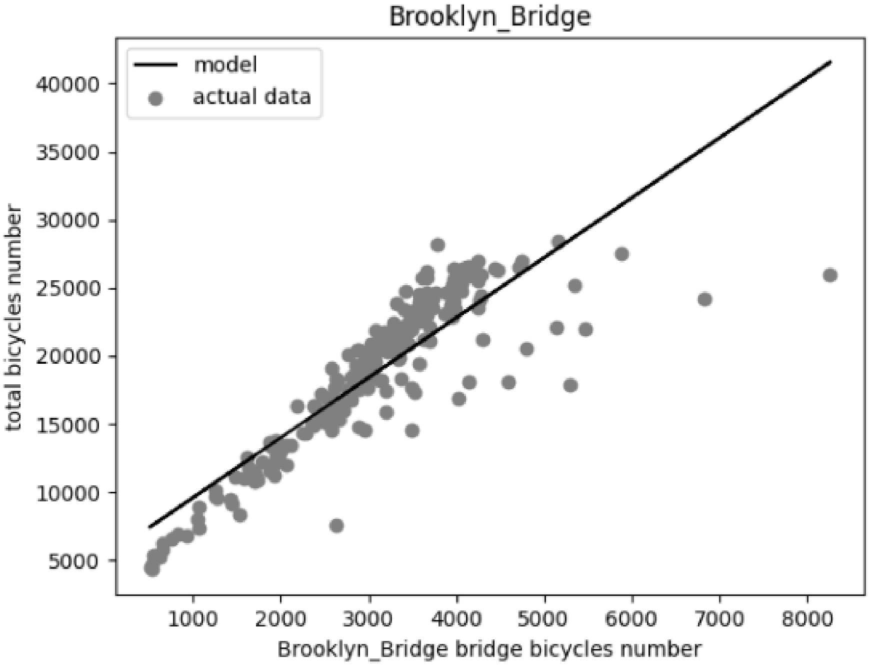

Figure 3. Bike traffic of Brooklyn Bridge versus total traffic

Our linear model can be defined as $y(x) = 4.397x + 5219.67$. The $r^2$for Brooklyn Bridge 0.7645972720914207.

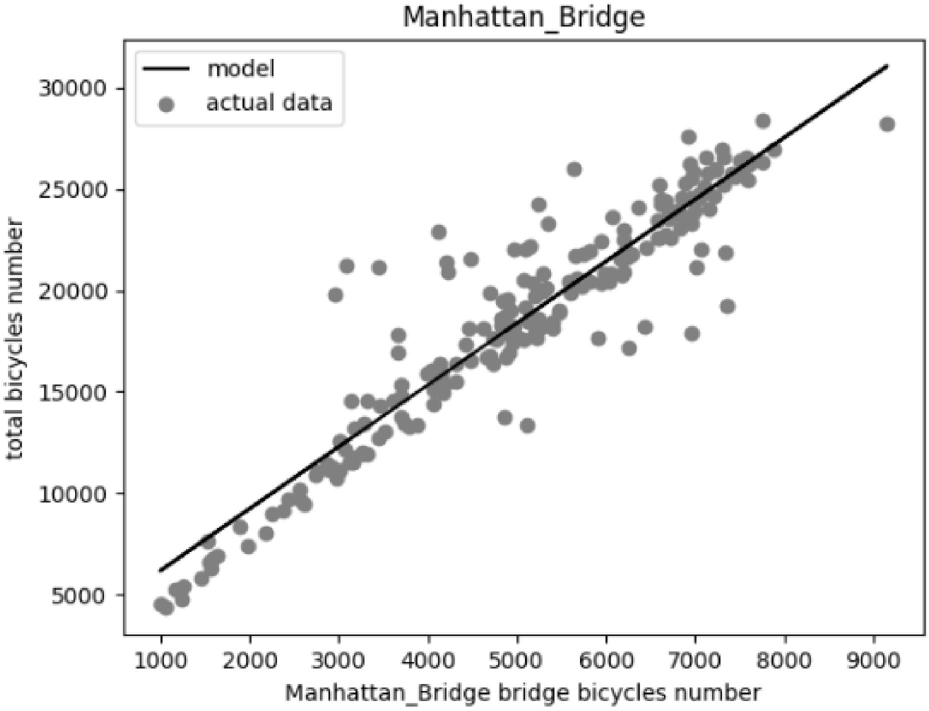

Figure 4. Bike traffic of Manhattan Bridge versus total traffic

Our linear model can be defined as $y(x) = 3.056x + 3105.06$. The $r^2$for Brooklyn Bridge 0.8751119334222758.

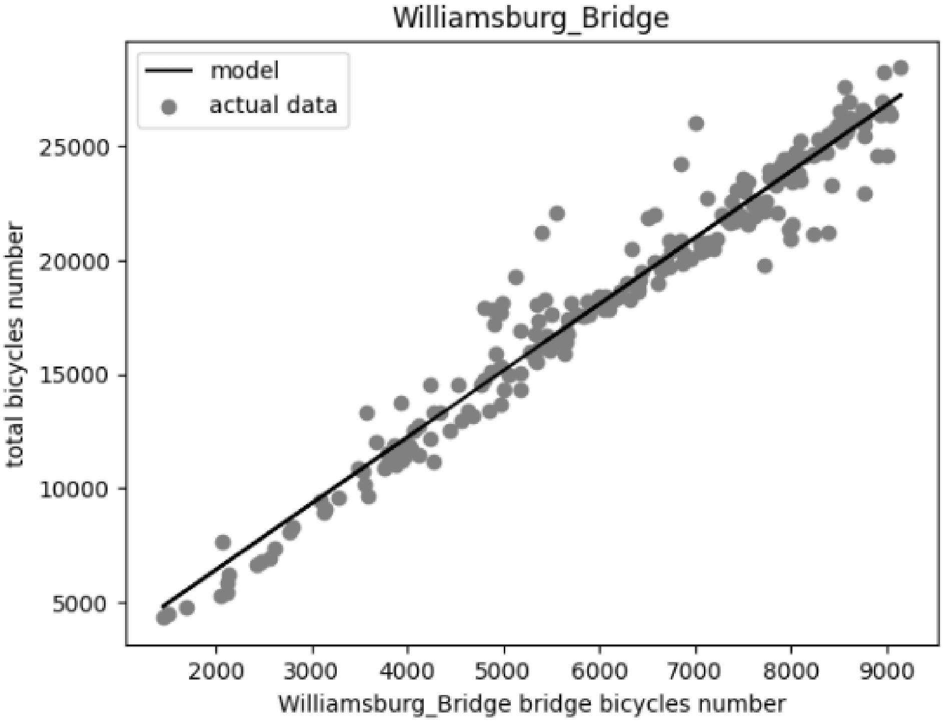

Figure 5. Bike traffic of Williamsburg Bridge versus total traffic

Our linear model can be defined as $y(x) = 2.91x + 616.17$. The r^2 for Brooklyn Bridge 0.9507989423628156.

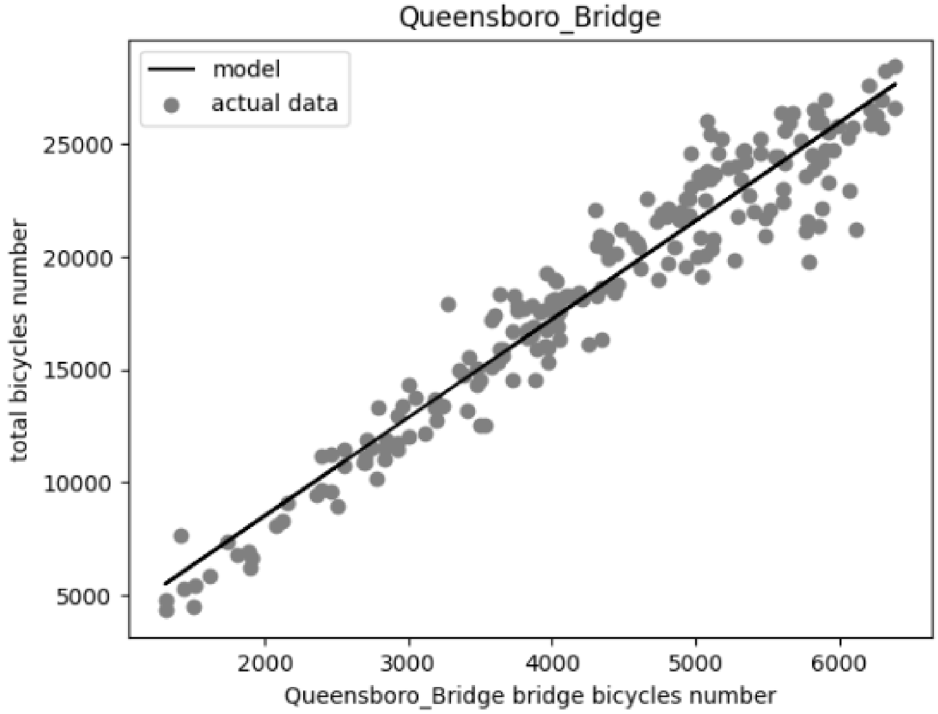

Figure 6. Bike traffic of Queensboro Bridge versus total traffic

Our linear model can be defined as $y(x) = 4.355x - 186.97$. The r^2 for Brooklyn Bridge 0.9277165921829049.

Based on our linear models, we should choose Manhattan Bridge, Williamsburg Bridge, and Queensboro Bridge that have the r^2 values of 0.875, 0.951, and 0.928. Since Brooklyn Bridge model only has r^2 value equals 0.765, we can conclude that the other three models could best represent the total traffic.

Question 2.

The weighted average for highest temperature in 214 days is: 77.13817423580012.

The weighted average for lowest temperature in 214 days is: 63.5516562807891.

The model for days having higher temperature than the weighted average for highest temperature is below. The r score for tempHabove 0.01957014816790492. The model is y = -136.65500918292733x + 32662.70839271426.

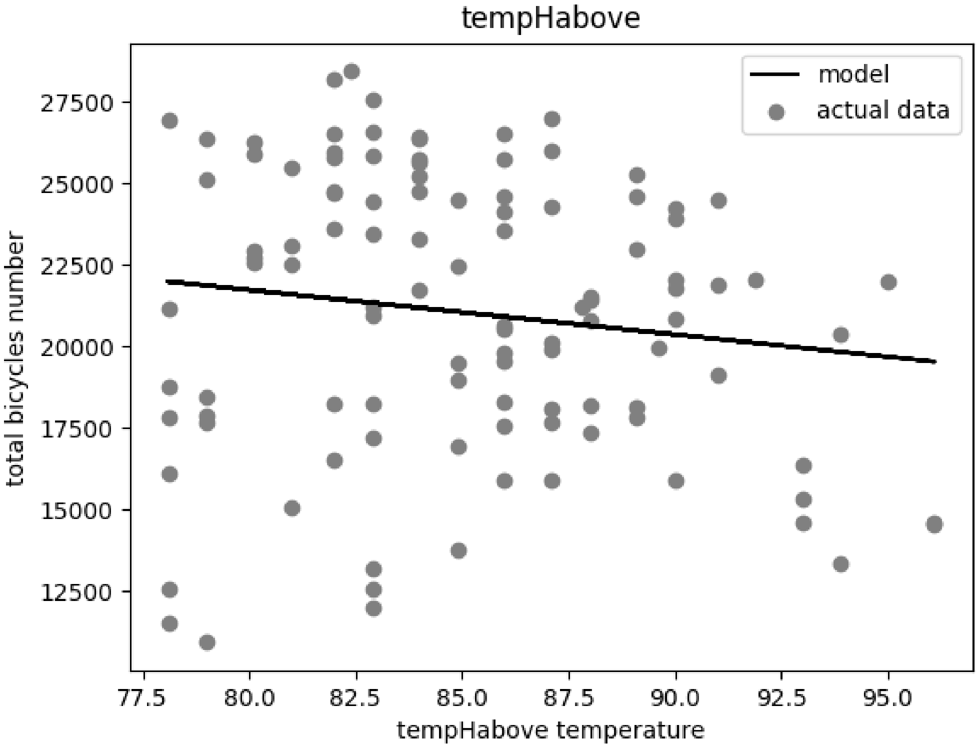

Figure 7. Days having higher temperature than the weighted average for highest temperature versus total bikes

The model for days having lower temperature than the weighted average for highest temperature is below. The r score for tempHbelow 0.38992170372738577. The model is y = 397.7422210463019x - 9667.58599536097.

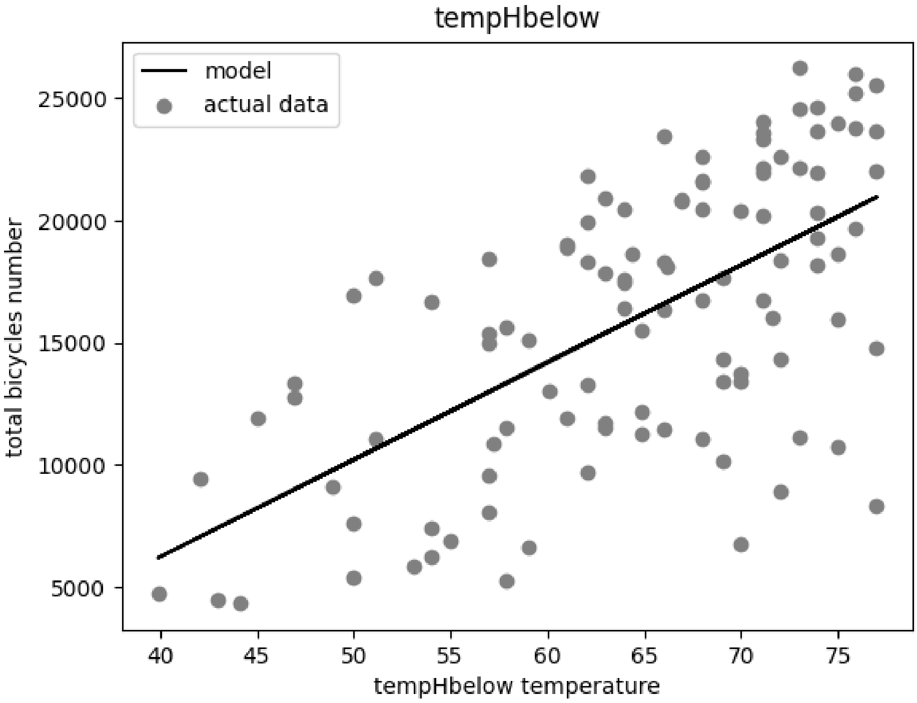

Figure 8. Days having lower temperature than the weighted average for <u>highest</u> temperature versus total bikes

The model for days having higher temperature than the weighted average for <u>lowest</u> temperature is below. The r score for tempLabove 0.005677088337388558. The model is y = -82.49703721392494x + 26216.184112810035.

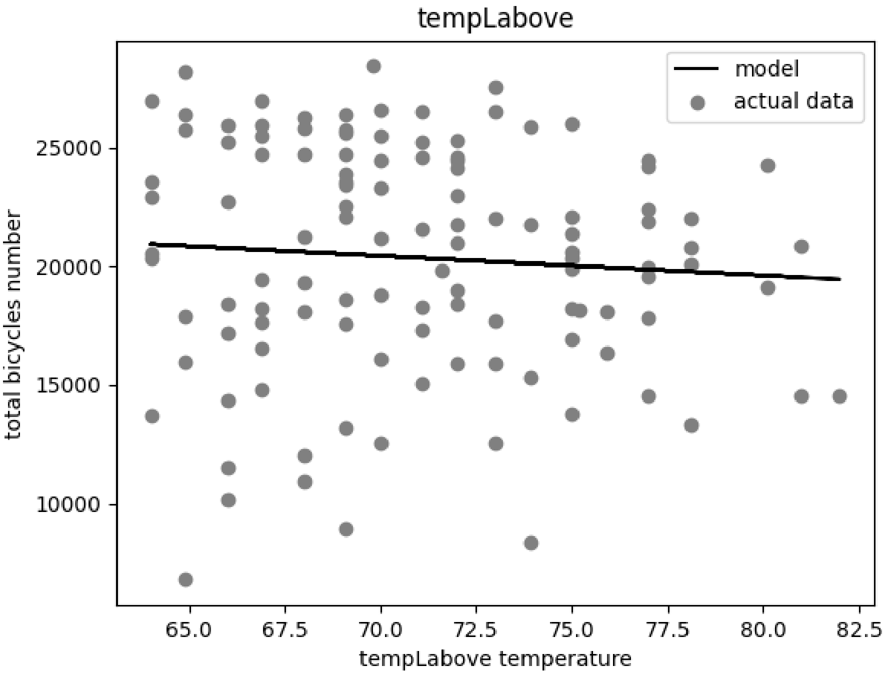

Figure 9. Days having higher temperature than the weighted average for lowest temperature versus total bikes

The model for days having lower temperature than the weighted average for lowest temperature is below. The r score for tempLbelow 0.2526869147586326. The model is y = 370.859149568954x - 2636.649624308875.

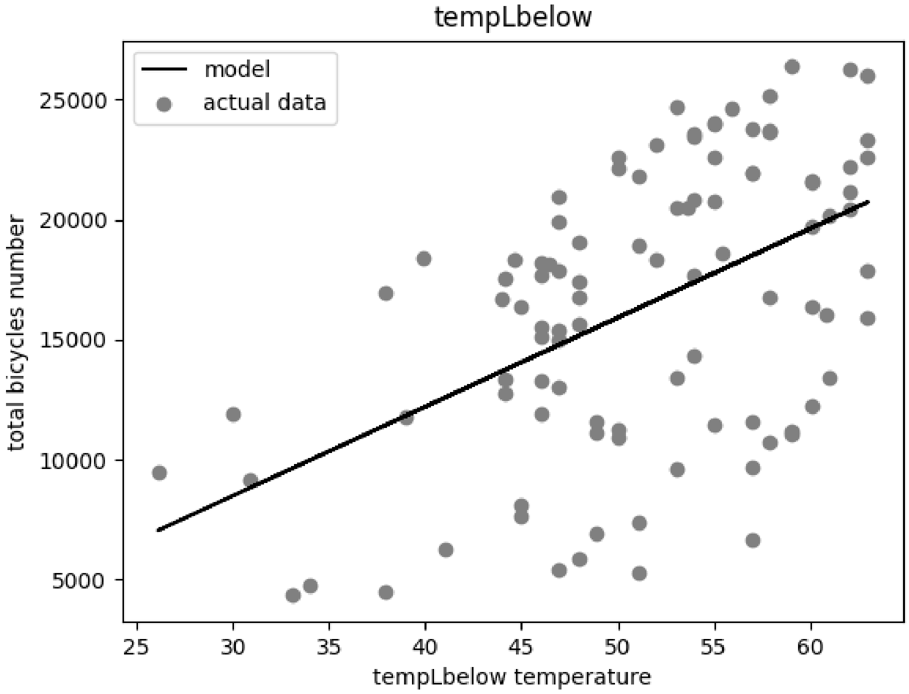

Figure 10. Days having lower temperature than the weighted average for lowest temperature versus total bikes

From 4 plots and r scores above, we can see that when temperature goes higher than the weighted average of highest/lowest temperature, the total number of bicyclists becomes even more unpredictable comparing to the method of using the average temperature of the days we used. When the temperature is less than the weighted average of highest/lowest temperature, the model for the total number of bicyclists becomes better compared to the previous method, but the r score is still not high enough, which means the correlation is still weak. Our conclusion for this question is we can hardly use the highest and lowest temperature of the day to predict the total bicyclists on the bridges.

Question 3.

Our result represents some correlation between the total traffic and the precipitation. Since we are going to decide whether it is raining instead of getting numerical value of the precipitation, we construct two lists: rain and does not rain. The number of days not raining 146 and raining is 68. Subsequently, we calculate the average of total bikes on rainy days and on no rainy days. The average Bike traffic for raining days is 14320.882352941177, and the average Bike traffic for no rainy days is 20511.712328767124.

Next, we plug in values to the equation $\mu \in (\bar{x} - Z_c * SE, \bar{x} + Z_c * SE)$ to determine the interval for rainy days and no rainy days. Result for constructing a 99% confidence interval of total traffic (number of bikes) when raining is [12490.465529413877,16151.299176468476]. Result for constructing a 99% confidence interval of total traffic (number of bikes) when not raining is [19567.152009146783,21456.272648387465]. From the interval we could see that the number of total bikes increases when it is not raining. Thus, we could also conclude that when the total number is relatively large, it is most likely that the day does not rain.

However, there are some numbers neither in the rainy interval, nor in the no rainy interval. Considering normal distribution and real-world cases, we use clusters to further analyze the data.

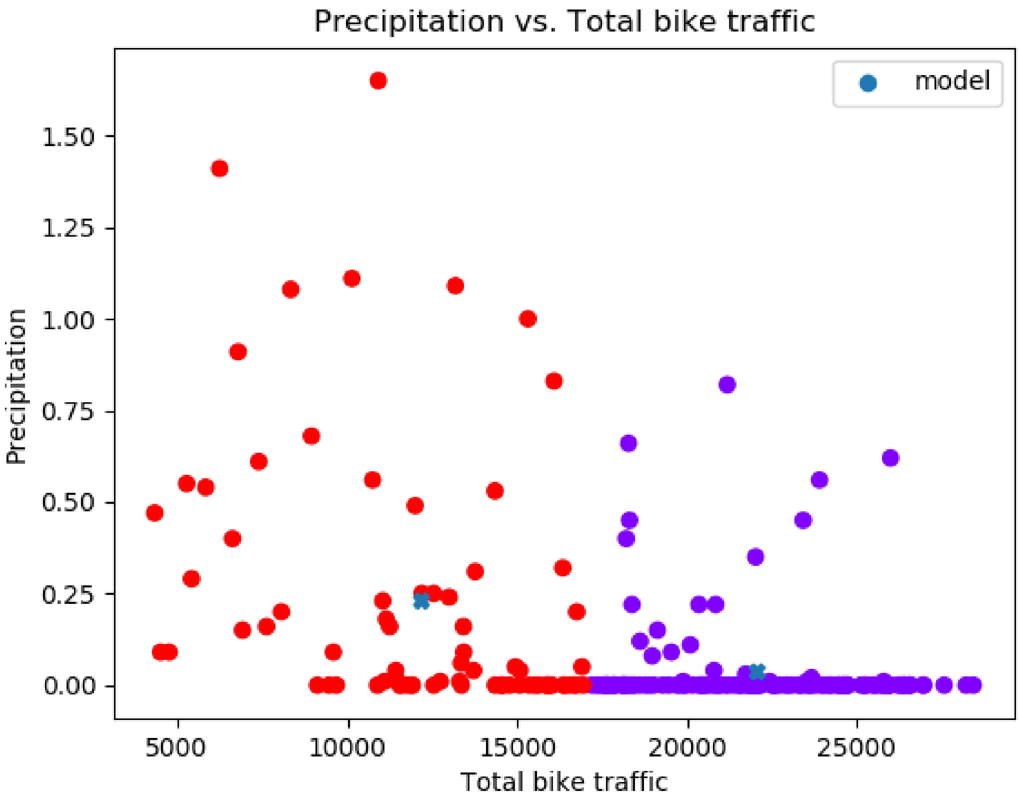

Figure 11. Clusters and centers of precipitation versus total traffic (Centers are blue cross marks)

The centers of the clusters are [0.040942029 22046.471] and [0.232763158 12185.7500]. The two confidence intervals present normal distribution in a relatively narrow range, and for sure there are some numbers outside the intervals. But this is reasonable. Under real world situations, like figure 11 shows, there are some outliers that are far away from the center of each cluster. The outliers will affect the confidence interval and we cannot simply use our model to predict those numbers. From the observation and calculation of centers, we see that the positions of both centers depend highly on the cluster density. On the other hand, the variance of the dataset is another reason for the huge gap between the two intervals. In conclusion, we can predict whether it is raining or not from the total traffic in the two 99% confidence intervals. For those numbers outside the range, using properties of normal distribution, we could possibly predict the precipitation from observing the distance between the number and the interval bound. For instance, if the total number of bikes 17000 is given, it is most likely that it is a rainy day since the number is close to the upper bound of the rainy interval [12490.465529413877,16151.299176468476]. The concept can also be interpreted by the cluster figure (figure 11) because the point with the smallest distance from the center is chosen to form a new cluster and update the center.