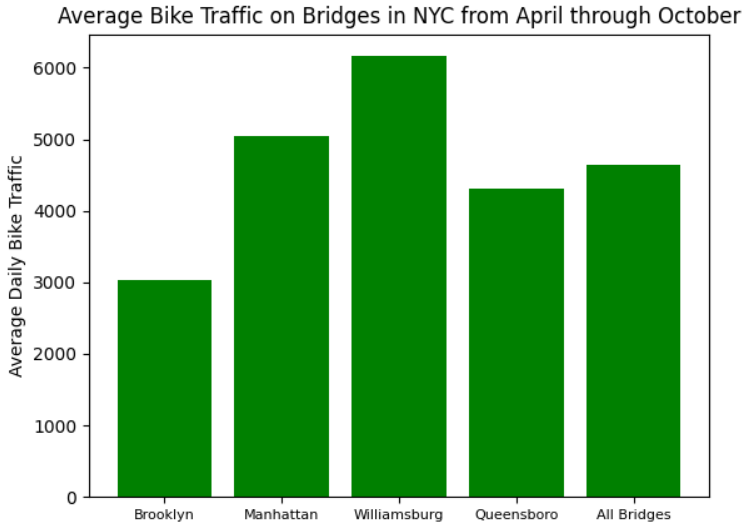
Name: Karthik Selvaraj
Purdue Username: kselvara
Path: 1

Mini Project ECE 20875

Dataset:

The dataset used in this mini-project was collected by the NYC Department of Transportation and describes the bike traffic on 4 different bridges in New York City. From April 1st to October 31st in 2016, the number of bikers using the Brooklyn, Manhattan, Williamsburg, and Queensboro bridges were counted to quantify bike traffic on each bridge. The data also includes the total bike traffic of the day, day of the week, the high and low temperatures in Fahrenheit, and the precipitation in inches. Further analysis of the dataset will include taking the average bike traffic on each bridge in the given time period. For example:



This figure displays the average bike traffic on each of the four bridges being analyzed from April 1st to October 31st. The figure also includes the average total bike traffic per bridge. The values displayed by each bar are 3031, 5052, 6161, 4301, and 4636 bikers, respectively.

Analysis Methods:

In question 1, we are tasked with determining which three of the four bridges should have sensors installed to monitor and accurately predict bike traffic in New York City. In order to determine which bridges should receive the sensors, I will create a similar figure to the one in the previous section to examine which three bridges have the highest average bike traffic. However, to account for the weather, I will make a line chart that contains the average bike traffic on each

bridge during each month of the year. This way, the data can be examined in a manner that takes seasonal effects into account. The three bridges with the highest average bike traffic will receive the sensors. This method was chosen, so that the bikes with the highest average traffic, or in other words the bridges that contribute most to bike traffic, will be chosen to have the sensors.

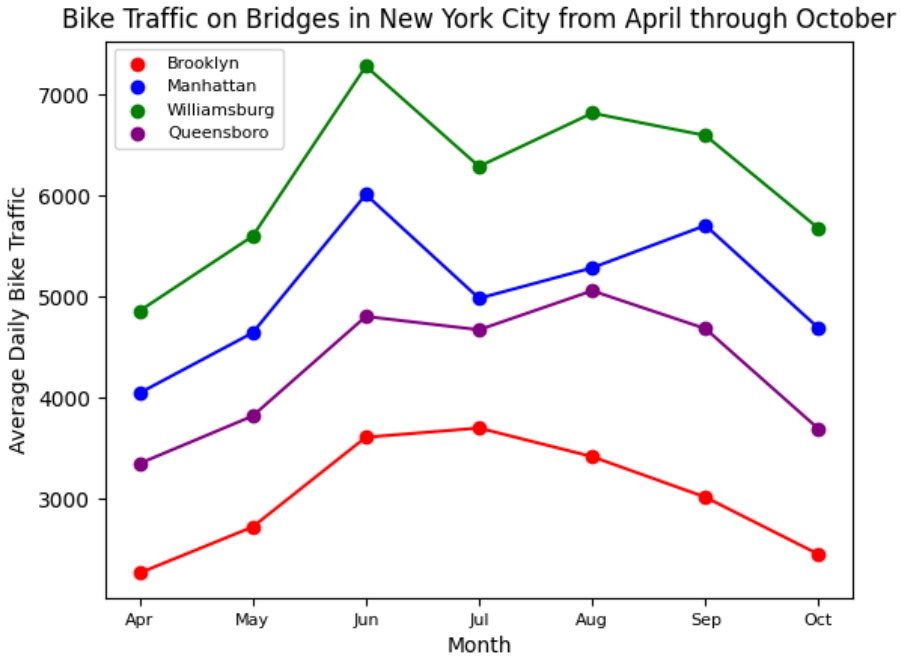
In question 2, we are tasked with determining whether or not police officers can utilize the weather forecast to predict bike traffic. To do this, I will use linear regression on the weather data given to us to create a model to predict bike traffic. This includes creating figures comparing bike traffic to precipitation, high temperature, and low temperature. This method was chosen because we will be able to develop a model that the police would use and then evaluate its effectiveness to answer whether or not they can predict the traffic. It was also chosen because a multiple variable linear regression model will yield a model that uses the precipitation level, high temperature, and low temperature to get a single number that represents the predicted total bike traffic. To determine whether or not there is a strong correlation between bike traffic and our weather data, I will calculate the r^2 value of the model. If the r^2 value is high, then the police officers can predict bike traffic with the weather forecast.

In question 3, we are tasked with determining whether or not the data can be used to determine which day of the week it is based on the given data. To do this, I would like to use a classification analysis method. This is because the problem requires us to *classify* the data into days of the week. Based on this problem, the best way to classify this data is to train a KNN model. This model is used because of its ability to classify data based on multiple features and into many groups. Prior to building the model, I will visualize features in the data and how they relate to the day of the week to see if there are any strong correlations. This will help determine whether or not the model will be successful. To evaluate the model, I will use the Accuracy, Precision, Recall, and F1 metrics of the KNN model to conclude whether or not the day of the week can be predicted based on bike traffic. I will also use these metrics to determine the best K value for the model.

Results:

<u>Question 1</u>

This is the graph comparing bike traffic between each bridge across the months:



As expected, the data follows seasonal trends across all bridges. Thus, separating the data by months helped determine how each bridge represents the total bike traffic throughout the year. As seen above, the Brooklyn Bridge consistently has the lowest average bike traffic throughout the year. This means that the Brooklyn Bridge contributes the least to the total average bike traffic.

It could be argued that the Williamsburg Bridge has too much traffic on it compared to the rest of the bridges. However, the Brooklyn Bridge is not affected by the seasonal changes as much as the other bridges, more specifically the Williamsburg Bridge. This is best seen in the jump from traffic in May to June; the Brooklyn Bridge does not see that much of an increase compared to the other three bridges. This shows that the Brooklyn Bridge does not represent the total average bike traffic as well as the other bridges. Therefore, the three sensors should be placed on the Manhattan, Williamsburg, and Queensboro bridges.

Question 2

For question 2, I used the following equation to determine the linear regression coefficients of all three listed weather features:

$$\beta = (X^T X)^{-1} X^T y$$

In the context of this problem, X is a N x 4 matrix, where N equals the number of given data points. Respectively, the first, second, and third columns represent high temperature, low

temperature, and precipitation. The fourth column is for the fixed offset of the model. The matrix y is an N x 1 matrix that contains bike traffic for each day.

Since most features operate on different scales, I normalized the data to ensure the mean and standard deviation for each feature are 0 and 1, respectively. Afterward, I calculated the coefficients for the following model, where T is predicted traffic, H is normalized high temperature, L is normalized low temperature, and P is normalized precipitation:

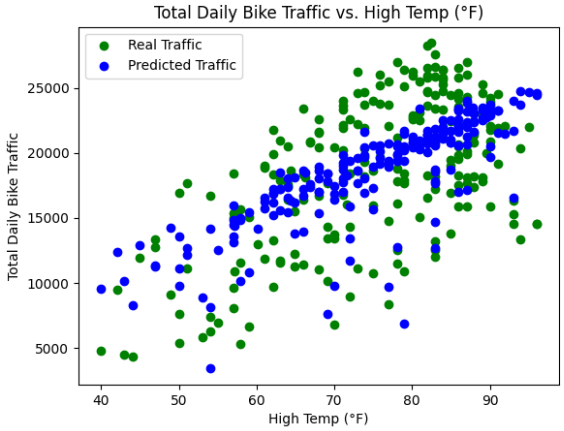
$$T = 4892.761H - 1889.936L - 2062.222P + 18544.533$$

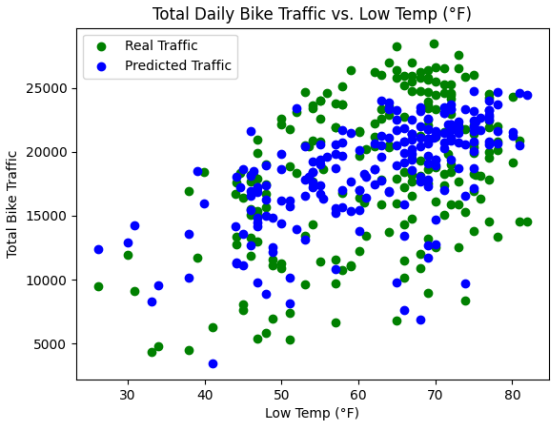
To test the validity of this model, the coefficient of determination was calculated by subtracting the mean squared error over the variance from one. The equation is as follows:

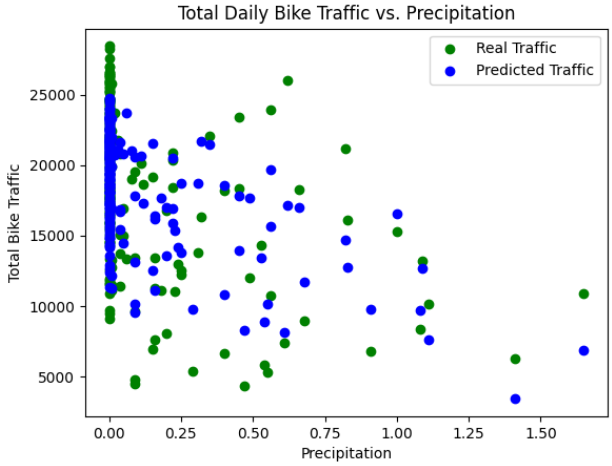
$$r^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} = 1 - \frac{MSE}{\sigma_Y^2}$$

- y_n: Measured value, $\hat{y}_n$: Predicted value
- $\bar{y}$: Mean measured value, σ_Y^2: Variance of measured value

Our results gave us a coefficient of determination of 0.499. In other words, about half of the variance can be explained by the model. This is not a very strong correlation, but there is a moderate correlation. This means police officers will be able to somewhat predict bike traffic based on the weather with a lot of variability in the prediction. The graphs below display the real and predicted bike traffic based on the data:



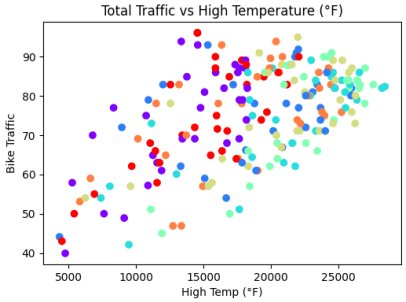


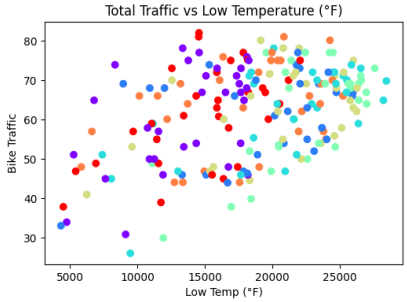


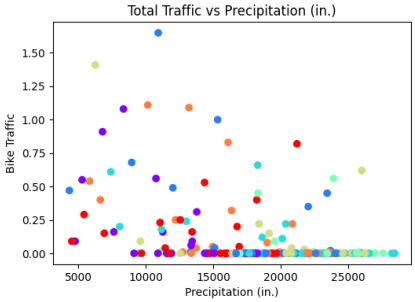
As seen above, the model is near the data points, indicating the weather is somewhat accurate when used to predict bike traffic.

Question 3

Prior to creating the classification model, I created graphs to visualize relationships in the data as they compare to the day of the week. This is the first set of graphs:





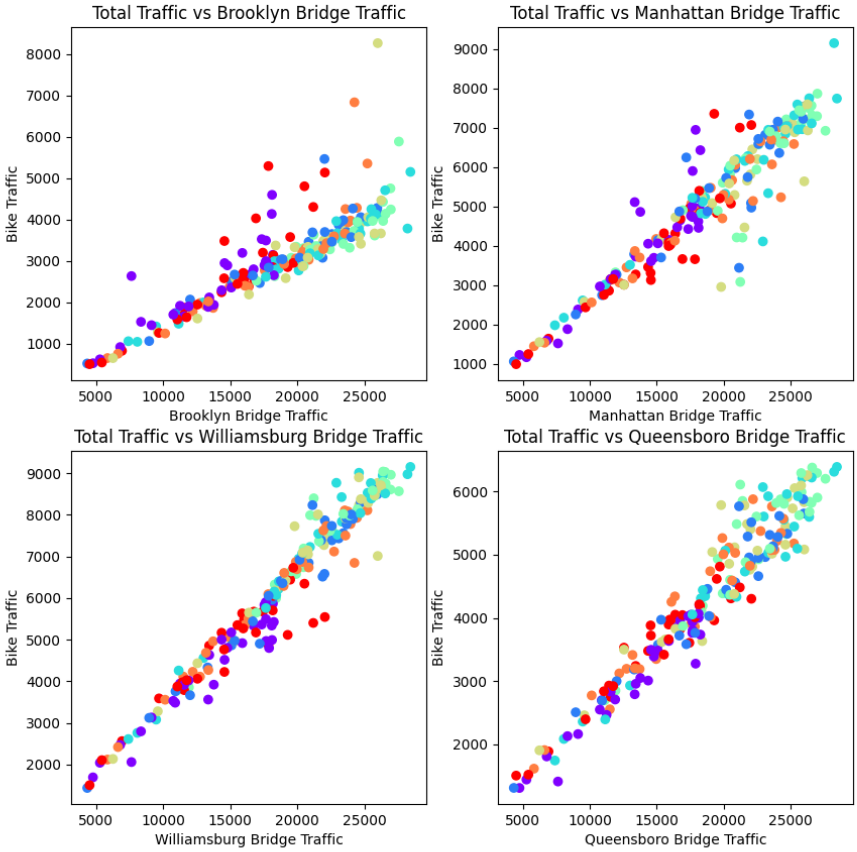


These scatter plots compare the high temperature, low temperature, and precipitation data to the total daily bike traffic. Each color represents a different day of the week, this is the key:

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday

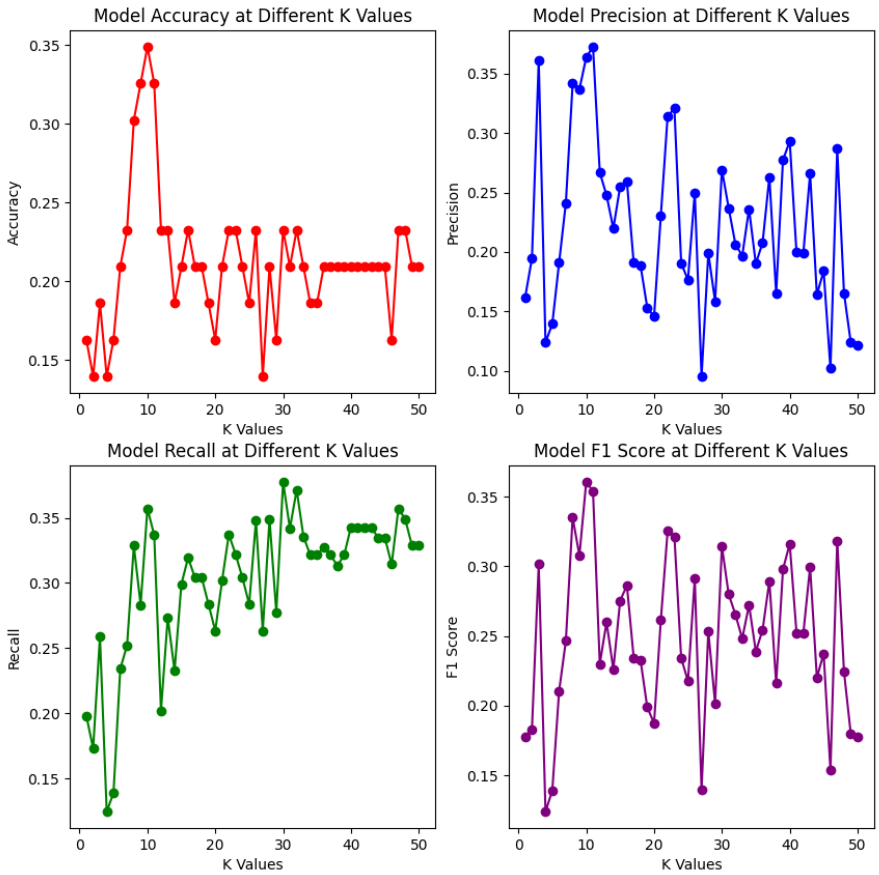
As seen, there is no strong visible correlation between these relationships and the days of the week. There are general trends in the data that can be seen, but there are no strong clusters of points from the same day of the week.

These are the next set of graphs:



These graphs visualize the relationships between total daily bike traffic data and the traffic on each bridge. In other words, they visualize how bike traffic on each bridge is distributed throughout the week on each bridge. These graphs use the same legend as the ones above. In this case, there are more clusters of data points seen, showing a stronger correlation between the given bike traffic data and the day of the week. Specifically, the days with higher traffic on all bridges typically are Tuesdays, Wednesdays, and Thursdays, or the middle of the working week. The rest of the days are not as strongly concentrated, but this does show that a classification approach could yield a model that predicts the day of the week based on this information.

Based on the question at hand and how the data appears, the model of choice here is KNN. To find the optimal K value, I have trained models with K values ranging from 1 to 50 and graphed the metrics used to evaluate a KNN model.



This figure shows the relationship between the various K values used to create these different models and the metrics used to evaluate the KNN model: Accuracy, Precision, Recall, and the F1 Score. The K values that result in the maximization of each metric are 10 (Accuracy), 11 (Precision), 30 (Recall), and 10 (F1). The F1 score is a combination of the precision and recall scores, so if K = 10 is the best K value for both the F1 score and Accuracy, then this is most likely the optimal K value for the model. In addition, it can be seen that at K = 10, the precision value is very close to that at K = 11, and there is still a high recall score at K = 10.

A K value of 10 was used to train the final model. The feature matrix used includes total bike traffic, bike traffic on each of the four bridges, the high and low temperatures, and the precipitation. Once the model is trained with a K value of 10, these are the following values for each metric:

Accuracy Score: **0.3488372093023256**
Precision Score: **0.3638528138528138**
Recall Score: **0.35714285714285715**
F1 Score: **0.36046661234292576**

It can be concluded that the model is not very good at predicting the day of the week, as the metrics are just over one third. This means the model is not very strong. **Even after removing the high and low temperature as well as the precipitation from the feature matrix, all the metrics stay the same (including the ideal K value of 10).** Therefore, it can be concluded that the day of the week can not be predicted based on bike traffic, or rather that the day of the week can not be *well* predicted.

This is most likely due to the similarity in bike traffic on certain days. For example, it was noted earlier that Tuesday, Wednesday, and Thursday have the highest bike traffic. The fact that all three of the days have similar bike traffic can make it difficult to properly categorize these data points. If there was more variability in the bike traffic on these days, then it would be much easier for the model to classify the data. Furthermore, the bike traffic on the weekends seemed to vary much more than on other days of the week. This also most likely made it more difficult for the model to classify these data points. Thus, it is reasonable to say that the model can not predict the day of the week based on bike traffic that well.