Team members: Junyoung Kim, Vikram Baid

Purdue Usernames: kim3722, vbaid

Path 2: Bike Traffic

Description of the Dataset

The data contains the number of bikers on four different bridges: Brooklyn, Manhattan, Williamsburg, Queensboro, along with the total each day from the 1st of April to 31st October in the year 2016. Additionally, the data also contains the precise day of the week, the high and low temperatures in Farenheit, and the precipitation in inches for each day.

Methods

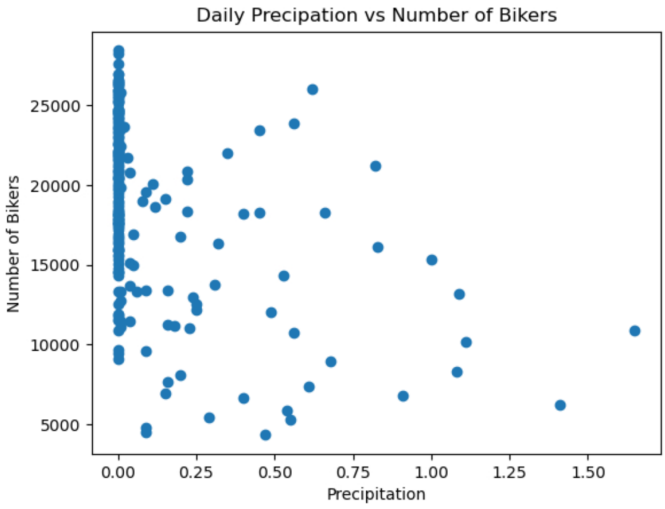
Question 1

To find the three bridges to install our sensors, we have decided to select the bridges with the largest overall average number of bikers. This is because we desire to analyze bridges with large sample sizes as that gives us more accurate and representative information about the overall biking traffic. The bridge with the smallest average also has a possibility of suffering from high variability.

Question 2

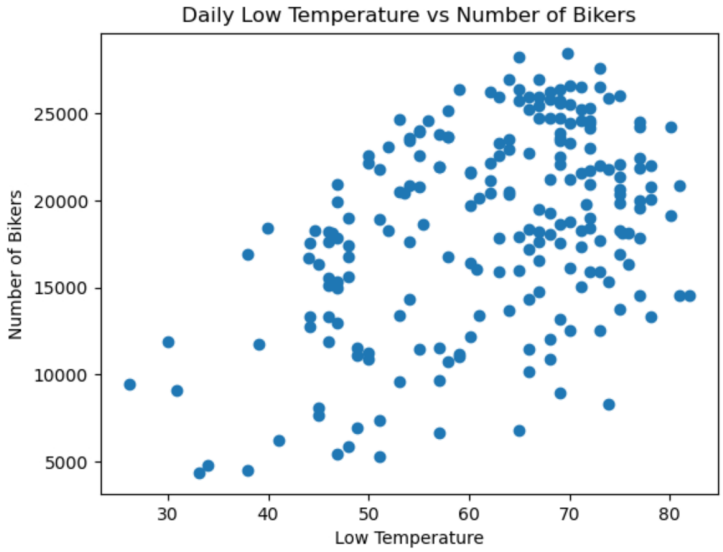
To determine how we can use the weather forecast to predict days of high traffic, we first decided to observe the scatterplots of the total number of bikers per day against three different measures of weather forecast - the precipitation, lowest temperature, and highest temperature - for each day.

The plot below shows the distribution of the number of bikers each day against the precipitation.



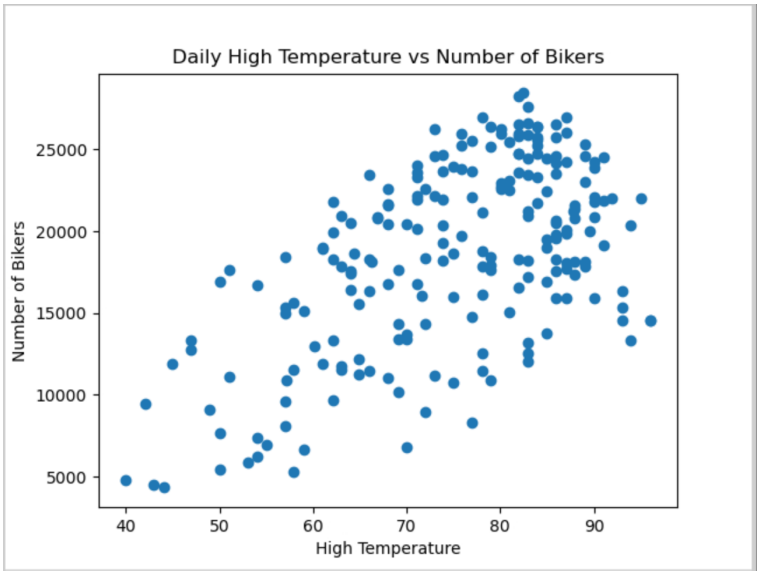
By observing the plot, we are unable to discern any noticeable pattern or relationship. There are also a lot of data points for days of low precipitation compared to days of higher precipitation. Any analysis using this data could potentially lead to biased conclusions as the data would not be representative since it lacks information for days with higher levels of precipitation.

The next plot shows the relationship between the number of bikers per day and the daily lowest temperatures



There appears to be a weak positive linear trend as there are a larger number of bikers for higher values of the low temperature. We can train a linear model to predict the number of bikers given the low temperature.

Finally we look at the scatterplot for the number of bikers vs the high temperature.



Similar to the plot for low temperatures, the data appears to follow a weakly positive linear trend. Thus we can combine these two measures to train a 2 variable linear model that takes as input the daily low and high temperature and predicts the total number of bikers on that day.

Question 3

In order to predict what day it is given the number of bikers on the different bridges, we decided to use a K-nearest neighbors algorithm. This is because we expect there to be approximately similar levels of traffic on each bridge on a given day each week - as most people have a fixed schedule to follow and thus use their bikes to commute on particular days each week. Given the number of people on each bridge on a given day, we believe that by finding the closest days with similar traffic, we may be able to classify which day it is.

Results and Analysis

Question 1

We found that the Manhattan, Queensboro, and Williamsburg bridges had averages of approximately 5052, 4300, and 6160 bikers, whereas the Brooklyn Bridge had an average of only 3031 bikers which was the lowest out of the others. Therefore we decided it would be best to place our sensors on the Manhattan, Queensboro, and Williamsburg bridges.

Question 2

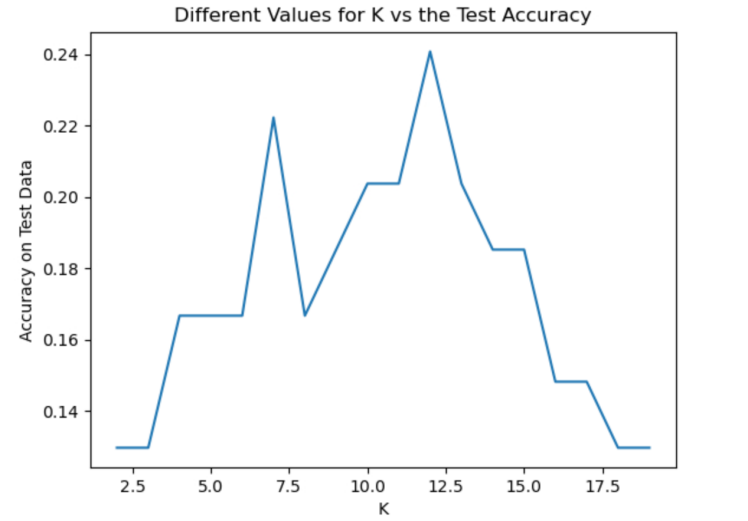
The model we obtained from linear regression was: $\widehat{y} = 483.6\,x_0 - 260.9\,x_1 - 1526.7$

- $\widehat{y}$: Predicted number of total bikers
- x_0 : Highest Temperature
- x_1 : Lowest Temperature

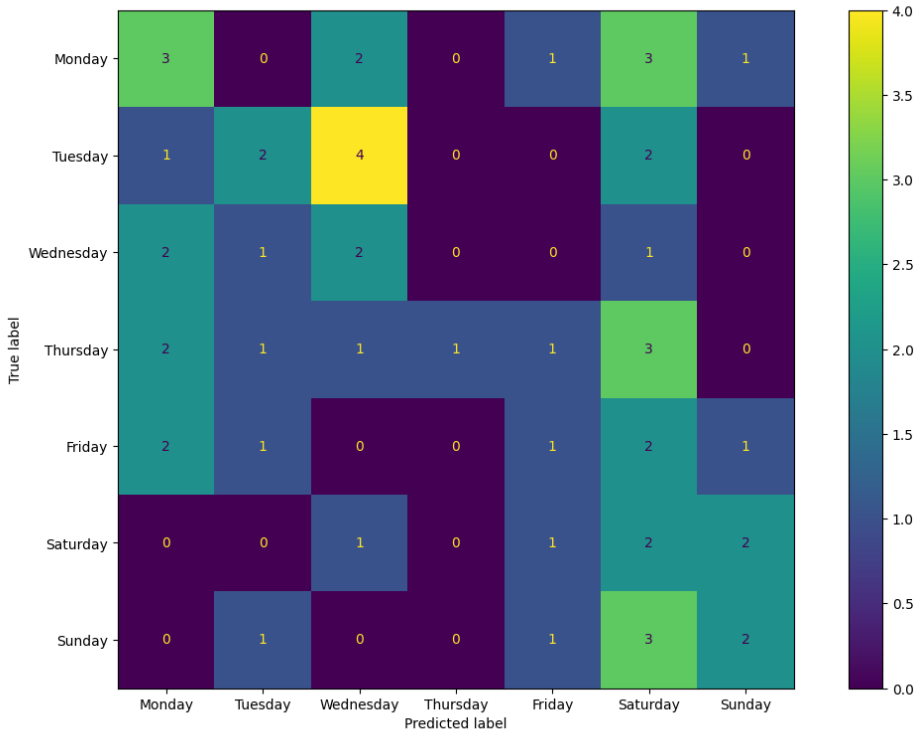
The R-squared value obtained was about 0.375 which indicates a relatively weak linear relationship. The predictions using this model may not be reliable and do not account for other weather conditions such as precipitation, and therefore should only be used as a very rough estimate for the number of bikers given the low and high temperatures.

Question 3

To train our kNN algorithm, we tried out different values for k from 2 to 20 and plotted each value of k against the model's accuracy on the test set.



We found that a k of 12 gave the best accuracy which was around 0.241. Although this was not a relatively high accuracy that we hoped for, the model did much better than random guessing which would produce an accuracy of about 1/7 or 0.143. To see how our algorithm misclassified different data, we obtained the following confusion matrix.



The confusion matrix reveals an important insight into why our algorithm did not perform very well. We noticed that generally, the weekdays got mistaken for each other, and so did the weekends. This is possibly because the traffic patterns are mimicking the patterns of people commuting to work on the weekdays. The relationship between bike traffic on each bridge and the exact day of the week may not be strong enough to precisely estimate day of the week using only information about the number of bikers on each bridge.