Names: Kaitlyn Yu, Ben Pekarek
Purdue Usernames: ksyu, bpekarek
GitHub Usernames: kaitsadilla, benpekarek
Path: 1

1. The Dataset

For this project, we selected the first dataset, the NYC 2016 Bike traffic data. In this set, we were provided the date, day of the week, high and low temperatures, precipitation metric, traffic over the Brooklyn, WIlliamsburg, Manhattan, and Queensboro bridge, in addition to the total bike traffic. When conducting our analysis, we filtered out the dates including snow and trace amounts of precipitation as there were not enough points to affect the data significantly, nor enough info as for what “trace amounts” entailed to quantitatively determine its relationship with regards to precipitation and traffic.

2. Analysis Methods

Question 1:
A multifaceted approach was taken to determine which three of the four bridges should have sensors installed to best predict total traffic. First, the mean traffic crossing each bridge was calculated to give a general sense of which bridges had the highest proportion of bike crossings. Then, the number of people crossing each individual bridge was plotted against the total number of people crossing all bridges to determine the nature of the relationships between the values. A linear relationship was observed. Next, 4 linear regressions were created with the feature being the number of people crossing each bridge and the target being the total number of people crossing all bridges. r^2 values were calculated for each regression to determine which bridge's crossing traffic correlated most with the total traffic observed around the city. It is believed that data that has a higher correlation value with total bridge traffic is the most useful for predicting future bridge traffic.

Problem 2:
To tackle this problem, a regression model was created. This was done by passing in training data: daily high temperatures, low temperatures, precipitation as features, all to be compared against total bike traffic across all 4 bridges. As conducted in the previous problem, r^2 was then calculated from the regression to determine the correlation between our factors with our output. To increase the accuracy of the model, we constructed a ridge regression, however setting our lambda equal to zero minimized error.

In addition, for analysis a trend line was produced to see the relationship of each factor as coefficients. From this line’s equation, we anticipate to see positive coefficients from the factors

which increase traffic, and negative coefficients for those which negatively impact bike traffic. Overall, with a higher R^2 value, we anticipate attributing that with a higher correlation between weather and bicyclists which would support whether weather forecasts can predict the number of bicyclists.

Problem 3:

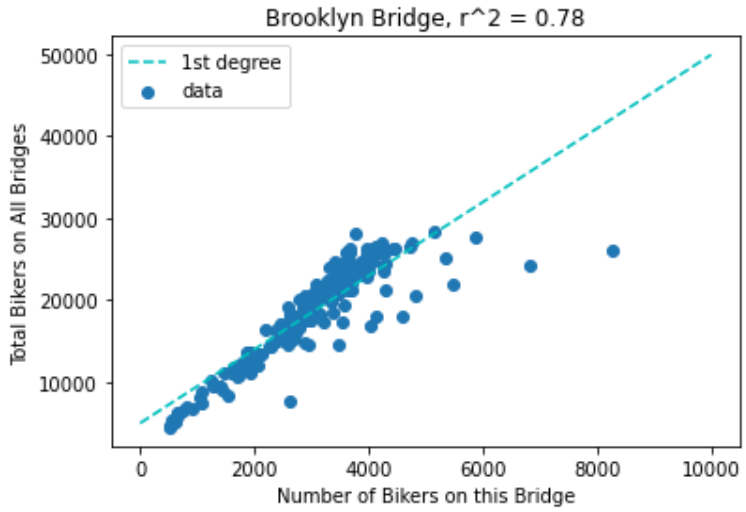
To solve this problem, we first had to filter the precipitation data to remove values that would not be able to be accurately quantified ('Trace amounts' and 'Snow'). Upon completing this, we removed dates with these irregular patterns resulting in a cleaner dataset. From there we ran a linear model with the number of bicyclists against the precipitation metric. Like in problem two, the r^2 was calculated to determine the relationship between the number of bicyclists and how much rain could possibly be there.

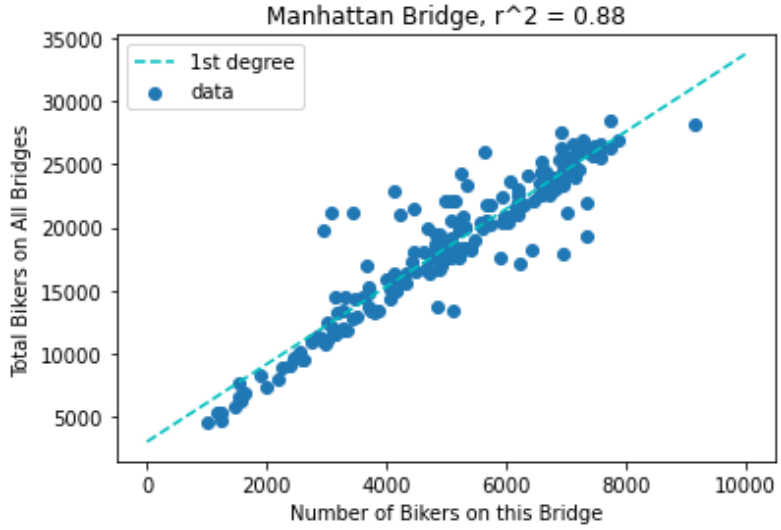
In addition to this, to answer the yes or no question of whether or not rain can be predicted, a knn classifier was used. This entailed filtering the precipitation data to a binary class, with 0 being no rain, and 1 being with rain. With this data, a model was trained and tested with 100 randomized splits into 75% training data and 25% testing data. From there, the best k value was determined by averaging the accuracy of various k-valued models across the 100 folds. Given a high accuracy in this model, we can classify whether it is raining or not based on the number of bicyclists outdoors.

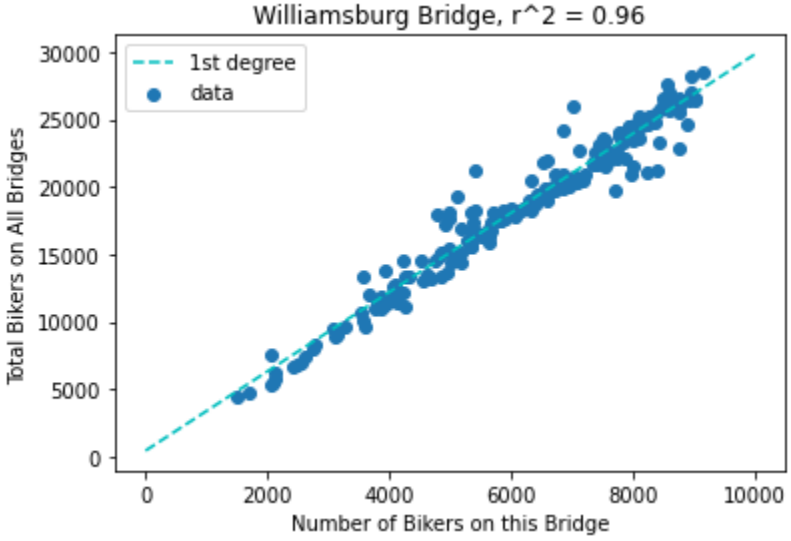
3. Results

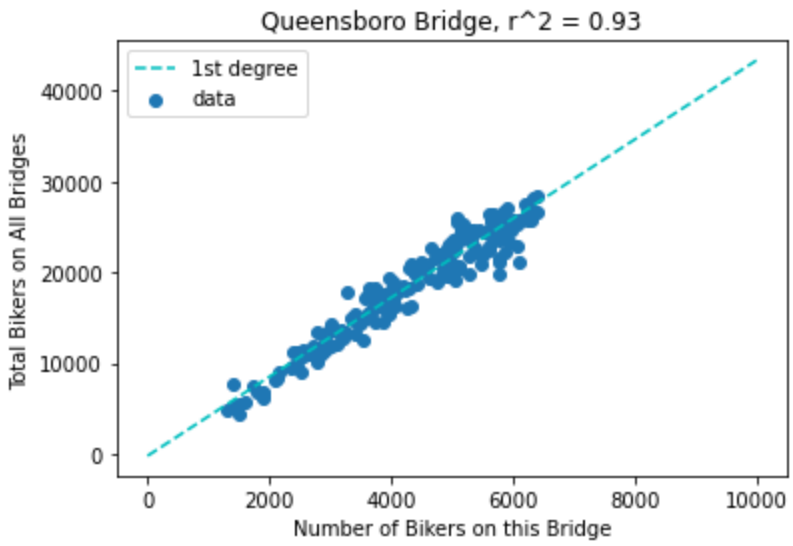
Question 1:
After running the regression models, in addition to calculating the means, from the process described above we were able to determine that the sensors should be placed on the Manhattan, Williamsburg, and Queensboro Bridges. We came to this conclusion, as upon running the linear regression of the traffic of each bridge against total traffic, the relationship between each bridge's traffic and total traffic produced an r^2 value of greater than .78, with the lowest r^2 value belonging to the Brooklyn Bridge. In comparison, Manhattan, Williamsburg, and Queensboro Bridges had values of 0.88, 0.96, and 0.93 respectively. Similarly, when comparing the average percent of total bike traffic over each bridge, Brooklyn Bridge had the lowest average total share of bike traffic.

Figure 1: Brooklyn, Manhattan, Williamsburg, Queensboro bridges


Brooklyn Bridge, r^2 = 0.78
1st degree
data
Total Bikers on All Bridges
Number of Bikers on this Bridge
50000
40000
30000
20000
10000
0
2000
4000
6000
8000
10000


Manhattan Bridge, r^2 = 0.88
1st degree
data
Total Bikers on All Bridges
Number of Bikers on this Bridge
35000
30000
25000
20000
15000
10000
5000
0
2000
4000
6000
8000
10000


Williamsburg Bridge, r^2 = 0.96
1st degree
data
Total Bikers on All Bridges
Number of Bikers on this Bridge
30000
25000
20000
15000
10000
5000
0
2000
4000
6000
8000
10000



From this we can conclude that not only does the Brooklyn Bridge have the lowest total traffic, but also with the high R^2 values, we can determine that total traffic is highly dependent on the traffic over the other Manhattan, Williamsburg, and Queensboro Bridges, and placing sensors on these three bridges would assist in predicting overall traffic.

Question 2:
After running the processes as described above for question 2, we were able to determine that there is a relationship between weather and total cyclists. This can be concluded from the regression model of precipitation compared against total bicyclists, as the average R^2 produced with cross validation was .46. While this value is not incredibly high, it is enough to show a low-to-medium correlation, this can also be verified logically with the derived equation below.

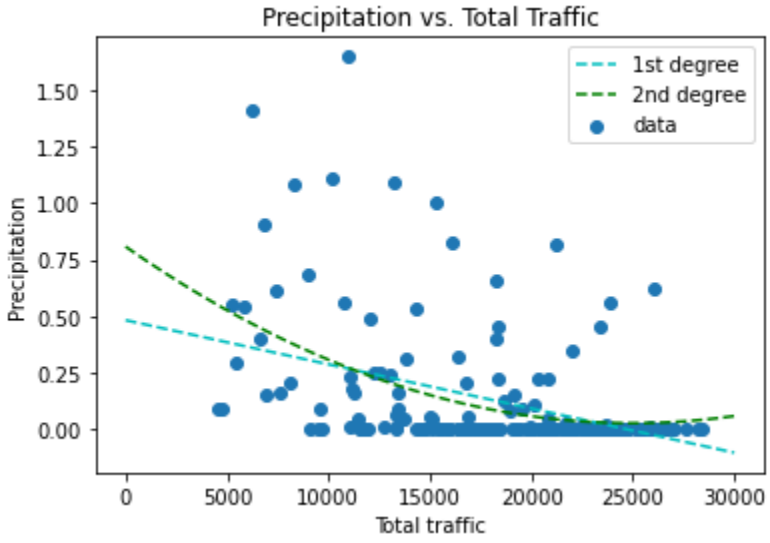
Total Bikers = 4776.05(High Temp) - 1730.56(Low Temp) -2146.32(Precipitation) + 18610.28

Overall, what this means is that officers can use the weather to predict which days there will be more traffic based on the rain, however it will not always be a definitive factor whether riders increase or decrease significantly. This model may possibly be improved by including additional features, such as the day of the week. Such a factor most likely influences the quantity of people riding bikes, as many use bikes to get to and from work, and would not be influenced by weather. However, such factors are not taken into consideration by the model.

Question 3:
Through running the regression model as described above for question 3, we were able to determine that you cannot use this data collected to predict the amount of rain from the number of bicyclists on the bridges. This was concluded as the regression with precipitation being the

dependent variable against total bikers produced an r^2 value of .16 signifying very low correlation and little to no statistical relationship.

Figure 2: Precipitation vs Total Traffic Data with Regression Overlayed



That being said, whether it is raining or not could be determined with a high degree of accuracy from the number of bicyclists on the bridges using a kNN classifier. In terms of the results of the kNN analysis, a k = 4 was found to produce the highest average accuracy after cross validation. This k value produced an accuracy of 77% which is fairly high as it means ~77% of the time, the model was able to sort whether it was raining or not based on how many people were biking.