Name: Christopher Chan

Purdue Username: chan328

Path: 1

Mini Project Report: Bike Traffic Across Bridges in New York City

Dataset:

For this project, a dataset consisting of the population of cyclists across four bridges in New York City was used. This includes the date, day of the week, the high and low temperatures for the day, the precipitation, the population on each of the four bridges and the total population across the four bridges combined. Data was collected across seven months from April until October.

Analysis of Approach:

For the first question, we had to answer what bridge we should install traffic sensors onto if we only had the budget to install sensors on three of the four bridges. To find out, I decided to find the average cyclist population across the four bridges from April to October and for individual months covered in the dataset. This allows for the analysis of the entire dataset at once as well as being able to see monthly trends to give us a better understanding of the data. In addition, the monthly data was added to see if factors such as the temperature or precipitation would contribute to changes in the average daily bike traffic.

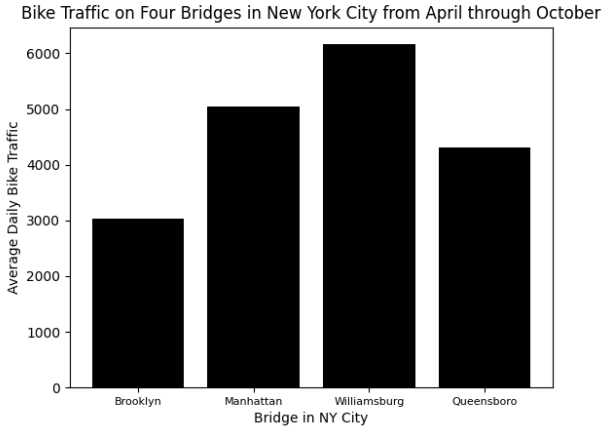
For the second question, we were asked to see if it was possible to predict the number of cyclists given data for the weather forecast the next day. This was to aid in the police's efforts to crack down on helmet laws and keep the citizens safe. To find out, I first had to see if there was even a correlation between the weather data and the population data. After normalizing the data, I first performed a linear regression on the three categories of weather data provided: the high temps of the day, low temps of the day and precipitation levels for the day. This helped create a model for me to use to predict bike traffic. I then used the coefficient of determination to see how well my regression fit the data and how accurate my predictions would be.

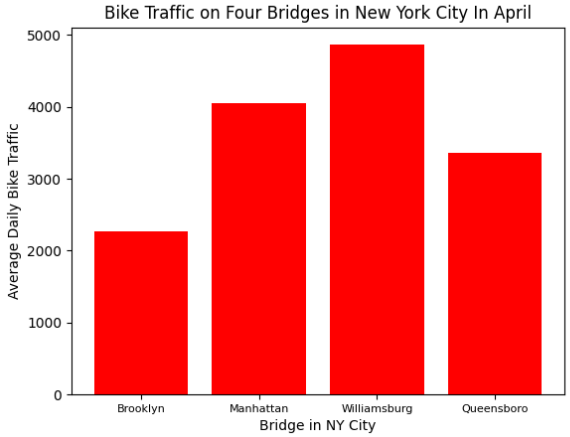
For the last question, we were asked if it was possible to predict what day of the week it was given the population data for cyclists on the bridges. I wanted to first visualize the data better, so I separated the data into categories based on the day of the week and graphed the average daily bike traffic for each bridge and for all the bridges combined. I then normalized the data and proceeded to perform a linear regression on the data. With the new model, I then attempted to predict the day of the week and used the coefficient

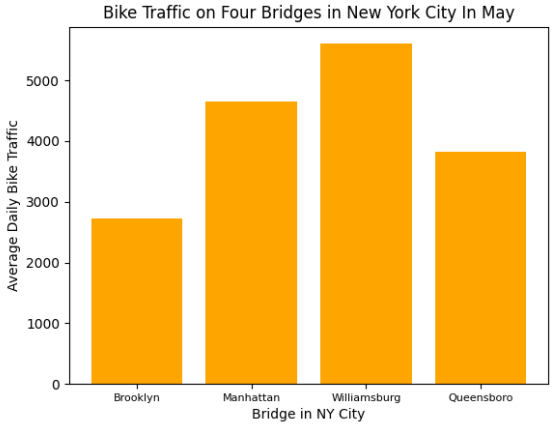
of determination to check how well the new model fit the data to see if my predictions would be accurate.

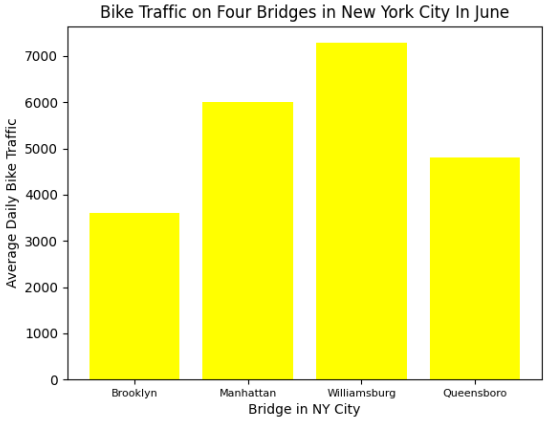
Analysis of Results:

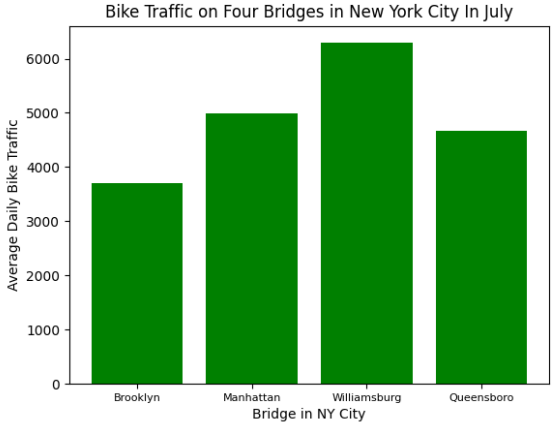
Here are the results of the first question in graphical form:

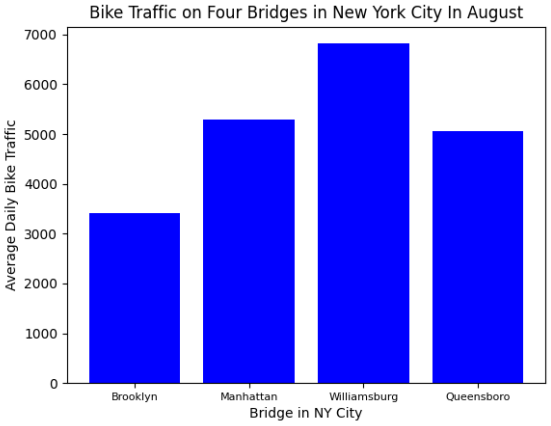


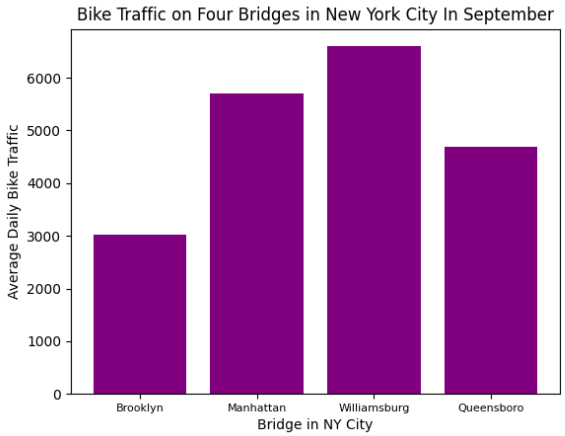


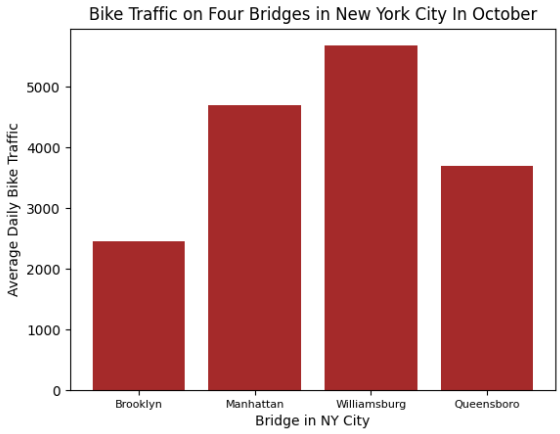










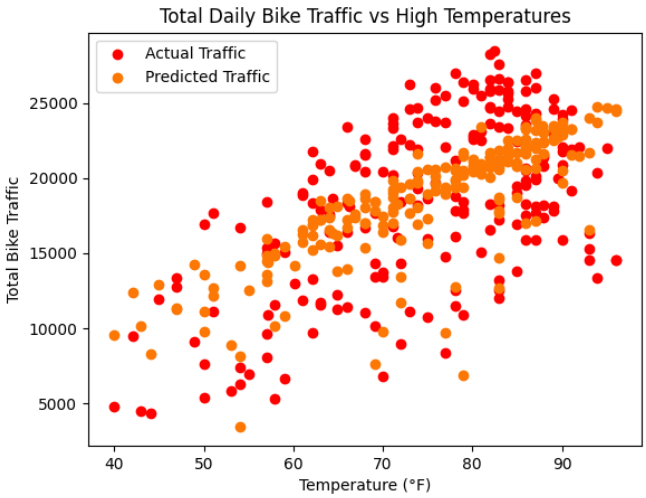


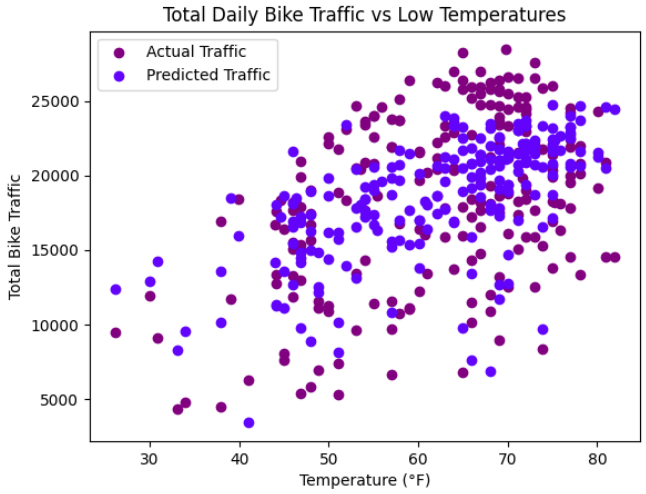
The graphs show the average daily bike traffic across the bridges with the first one being the total for the seven months and each graph after being data for a particular month. The result of the graphs can clearly be seen with Williamsburg Bridge having the highest average daily bike traffic for every month and in total. This is followed by the Manhattan Bridge, Queensboro Bridge and lastly, Brooklyn Bridge. All the graphs follow a similar pattern throughout. By looking at the graphed data, we can conclude that since Brooklyn Bridge receives the least amount of bike traffic on average, we can exclude it from our traffic sensor program and only need to install the sensors on the other three bridges.

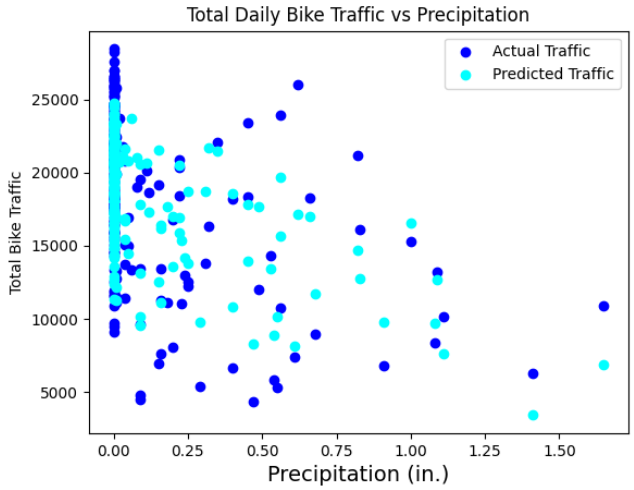
For the second question, I used the equation $\beta = (X^TX)^{-1} X^Ty$ to solve for beta and find the coefficients of my linear regression. In this equation, X is an N by 4 matrix where N is the number of data points present with one point per day. The four columns are made up of high temps, low temps, precipitation and a fourth column of 1s for the offset of the model. The y matrix is made of an N by 1 matrix that represents the bike traffic for a day. After doing the calculations, I was able to find the coefficients for the model I was going to use to predict the bike traffic with. The equation looked like this: T = 4892.76h - 1889.94l - 2062.22p + 18544.53 where T is for predicted traffic, h is for high temp, l is for low temp and p is for precipitation. I then did the calculations to find the coefficient of determination using this equation below:

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

The r squared value I got from the calculations was 0.499. The r squared value is normally between 0 and 1 with 1 being a perfect fit and 0 being no better than a horizontal line. Since our value is around halfway between them, we can conclude that about half of the variance of the data is explained using the model we found. I then graphed predicted data on top of the actual data below for each weather condition:

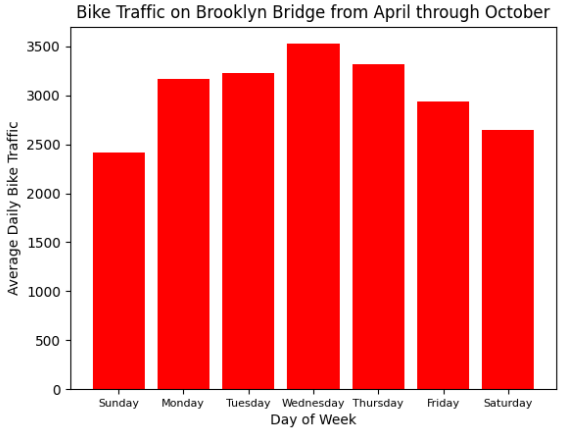

Total Daily Bike Traffic vs High Temperatures
Actual Traffic
Predicted Traffic
Total Bike Traffic
25000
20000
15000
10000
5000
40
50
60
70
80
90
Temperature (°F)

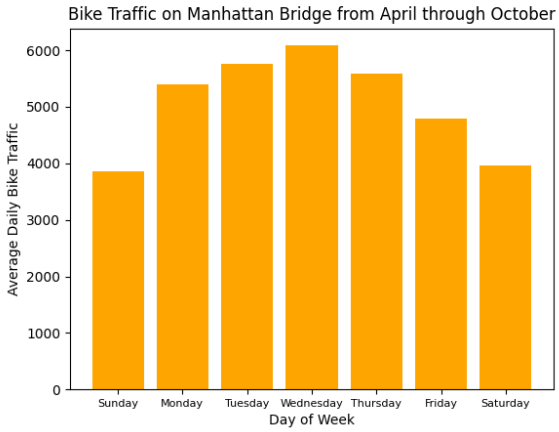

Total Daily Bike Traffic vs Low Temperatures
Actual Traffic
Predicted Traffic
Total Bike Traffic
25000
20000
15000
10000
5000
30
40
50
60
70
80
Temperature (°F)

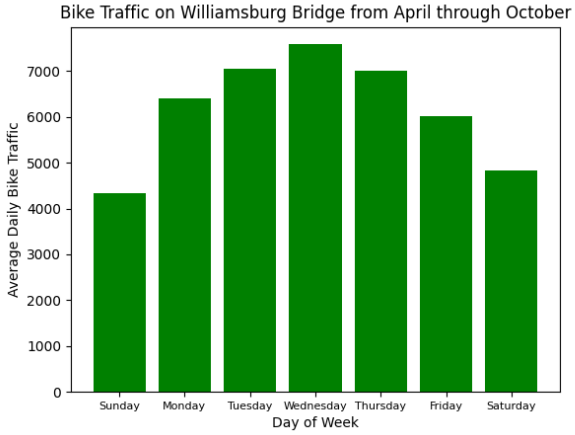

Total Daily Bike Traffic vs Precipitation
Actual Traffic
Predicted Traffic
Total Bike Traffic
25000
20000
15000
10000
5000
0.00
0.25
0.50
0.75
1.00
1.25
1.50
Precipitation (in.)

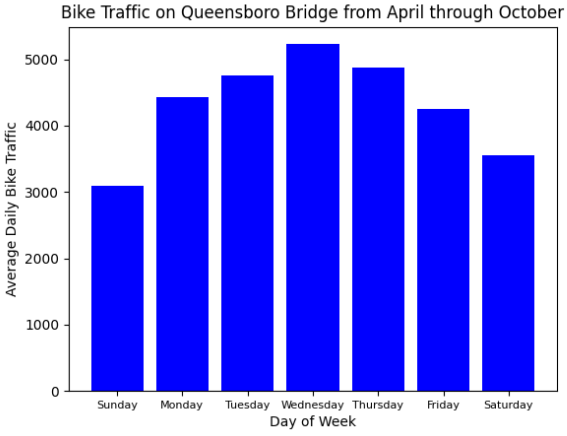
As shown in the graphs above, we can see that there is a decent amount of correlation between the points predicted by the model we found and the actual data we were given so with this, we can conclude that we can make a decent estimate as to how many people there will be on the bridges on a given day but we can not be completely exact on the number of cyclists and we can't be a hundred percent sure on our prediction either.

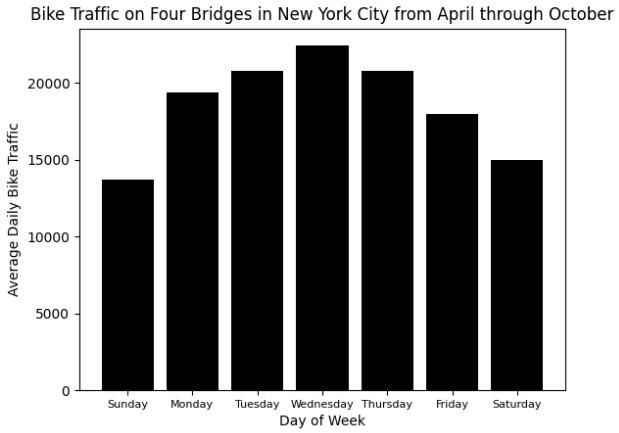
For the last question, I wanted to first reorganize the data a little bit and visualize the data categorized by the day of the week before starting to find the model to predict the day of the week. I decided to graph the data in terms of daily average number of cyclists in total and on each individual bridge. The graphs I created can be found below:









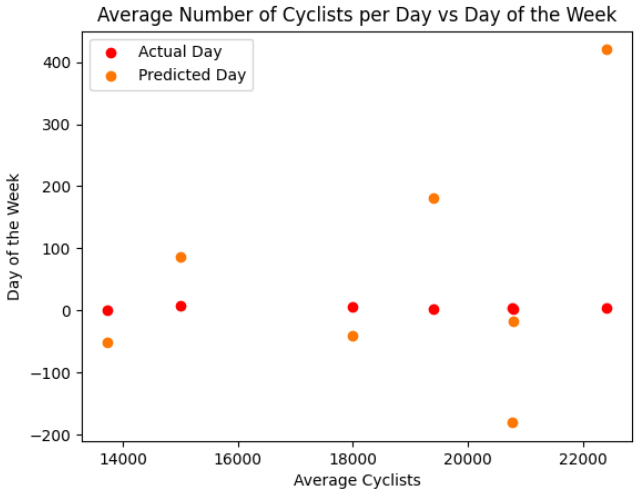


As you can see from the data in graphical form, they seem to follow the same general trend but I can already see issues in terms of being able to accurately predict the day based on the population. The large numbers isn't the issue here as that would probably help us in the long run if there was a large range of averages between the days of the week. However, the averages between the days of the week are in fact quite tightly grouped together towards the top of the spectrum, which doesn't bode well as it is much harder to predict the day of the week with such a small range of values. However, I chose to go ahead with the linear regression just to check if my doubts were correct or not. The linear regression was done using the same equation as mentioned in question 2 and it as done on the total population for all the bridges. This time, X was an N by 32, with N being the number of data points which is 7 as there are 7 days per week. The columns are made up of 31 days in a month, which is the upper limit of days in a month and the column of onesfor the offset of the model. And this is where my first calculation issue popped up. I realized that if I took a look at the dataset, you can clearly see that some months only had 30 days and not the upper limit of 31. I didn't want to withold data from my model so I decided to add dummy data to the months without 31 days by adding the average population value for the corresponding day of the week that was missing that last data point. The y value for the equation was an N by 1 matrix representing the 7 days of the week starting from 1 to 7 with N being 7 for the seven days of the week. After solving for beta, I found the coefficients for the model as shown here:

[75.75, -255.75, -694.25, 519.75, -230.94, 123.44, 204.56, 23.59, 13.82, 144.16, 120.25, -106.44, 108.06, -1.59, -15.69, 60.90, 8.41, -92.42, -14.69, 267, -72.5, 61.88, -41.37, -16.38, 81.64, -5.125, -16, -27.97, -57.81, -12.41, 87.38, 85.31]

This long list of coefficients turns into our model with the variables using the coeffiecients being the data for each day of the month and the result being our predicted day of the week based on that data. I was starting to have doubts about our model and the feasibility of being able to accurately predict the day of the week so instead of finding the coefficient of determination right away after predicting the new datapoints, I quickly printed out the predicted data and sure enough, the model was not accurate at all,

predicting days of the week that didn't exist as far as a 420th day of the week. I graphed the predicted data along with the actual data below:



With no doubt in my mind that our model was absolutely off in its predictions, I did the calculations to find the coefficient of determination and sure enough, I ended up with an r squared value of -8956.49, a value which was supposed to be between 0 and 1. This ultimately proves that with the data provided, it is not possible to to predict the day of the week purely based on the population of cyclists on the bridges. After further through, I have concluded that the huge range in population from day to day over such a long period of time is to blame for this as the seasons, the weather and other factors play big roles in the population, not just the day of the week.