ECE 2087-002 Mini Project

Partner Names: Cameron Woods, Legnar Tarrago

Purdue Usernames: woodscm, ltarrago

Chosen Path: Path 1

Data Set Analysis

This mini project aims to analyze the pedestrian bike traffic data across four bridges (Brooklyn, Manhattan, Queensboro, and Williamsburg) in New York City. The dataset, NYC_Bicycle_Counts_2016_Corrected.csv, was collected by the New York City Department of Transportation during a six-month span from April to October and contains records of the number of bicyclists that crossed into/out of the city on each respective bridge, the high and low temperatures, the day of the week/date, and the total number of bicyclists entering the city each day.

Problems to Solve

1. You want to install sensors on the bridges to estimate overall traffic across all the bridges. But you only have enough budget to install sensors on three of the four bridges. Which bridges should you install the sensors on to get the best prediction of overall traffic?
2. The city administration is cracking down on helmet laws and wants to deploy police officers on days with high traffic to hand out citations. Can they use the next day's weather forecast(low/high temperature and precipitation) to predict the total number of bicyclists that day?

3. Can you use this data to predict what day (Monday to Sunday) is today based on the number of bicyclists on the bridges?

Problem 1 Analysis

In problem 1, we were tasked with analytically choosing which three bridges with sensors installed would allow the New York Department of Transportation to get the best overall prediction of traffic on the bridges. Initially, we thought to plot the number of people using each bridge while removing outliers and the average of each bridge to find the three bridges with the highest average number of cyclists each day. But quickly realized this method is not a great predictor of traffic throughout time and a mathematical model would be more accurate. So then we decided to create a linear regression model using the traffic data of all the bridges. Choosing three bridges to install sensors on and having four total bridges gives us four possible unique combinations of three bridges which the sensors can be installed on. The unique combinations are as follows:

1. Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge
2. Brooklyn Bridge, Manhattan Bridge, Queensboro Bridge
3. Brooklyn Bridge, Williamsburg Bridge, Queensboro Bridge
4. Manhattan Bridge, Williamsburg Bridge, Queensboro Bridge

After understanding how the data should be organized, we then wanted to begin developing our linear regression model. We split our data and trained a set amount of the data and built the regression model using the training data. We then tested the model on the remaining data and evaluated the R-Squared Values of the 4 unique combinations of three bridges. This R-Squared

is an indicator of how accurate the model is based on the data tested. The combination of three bridges which has the highest R-Squared will be the three bridges the sensors should be installed on. It is important to note that the data which the model is tested on is the same on all cases so there is no variation in the testing procedure.

Problem 1 Results

Our initial analysis method using the average of the number of people on each bridge while removing outliers left us with the following diagram:

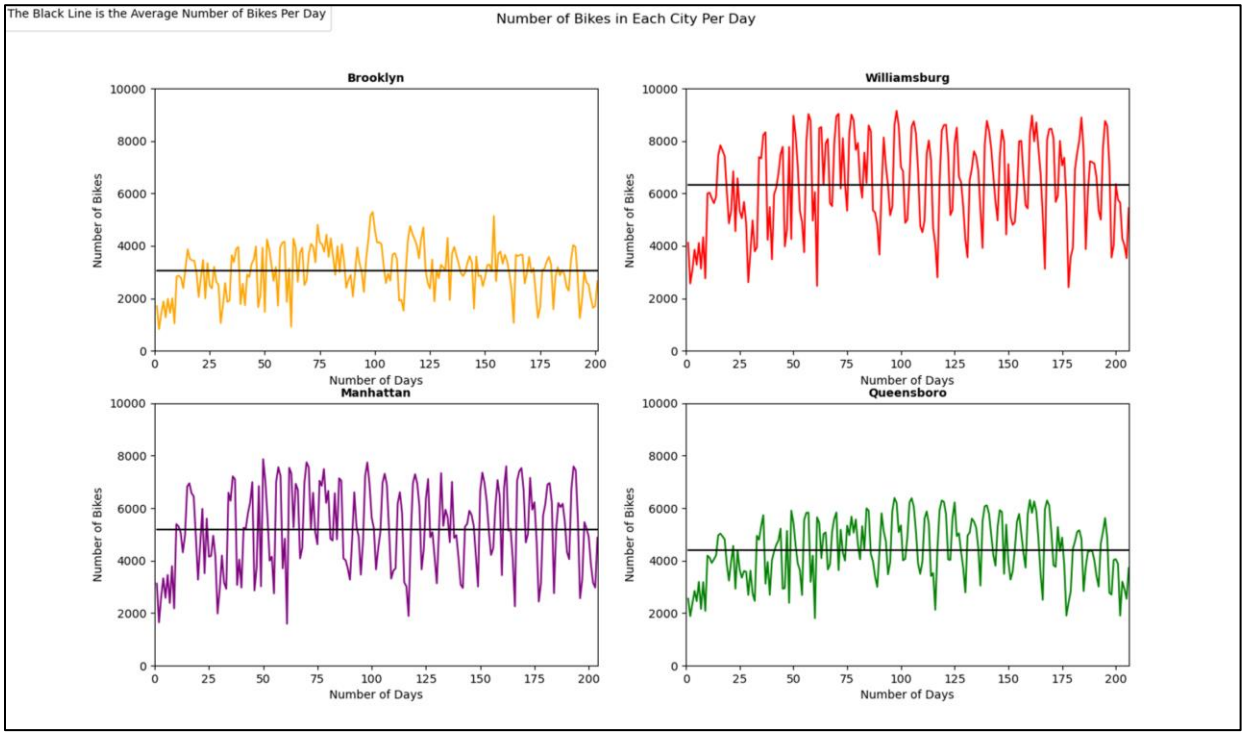


Figure 1: Naïve Approach

The above diagram would leave us to choose to install the sensors on the Williamsburg Bridge, the Manhattan Bridge, and the Queensboro Bridge. But we wanted a more precise way to

determine the three bridges to install the sensors on, so we calculated a linear regression and gathered the following results:

R-Sqaured Values				
	Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge	Brooklyn Bridge, Manhattan Bridge, Queensboro Bridge	Manhattan Bridge, Williamsburg Bridge, Queensboro Bridge	Brooklyn Bridge, Williamsburg Bridge, Queensboro Bridge
R-Sqaured Value	.996 or 99.6%	.988 or 98.8%	.946 or 94.6%	.982 or 98.2%

Table 1: Problem 1 R^2 Values

&

Total Bike Traffic Model Part 1

$$= 1.136(\text{Brooklyn Bridge}) + .9449(\text{Manhattan Bridge}) + 1.612(\text{Williamsburg Bridge}) + 380.413$$

Using this data and the understanding that a regression model is more accurate than choosing the three highest averages to predict the overall traffic in the city we can declare that installing the sensors on the Brooklyn Bridge, Manhattan Bridge, and Williamsburg Bridge will lead to best predictor of traffic within the city. The R-Squared value was .996, which was higher than the other combinations.

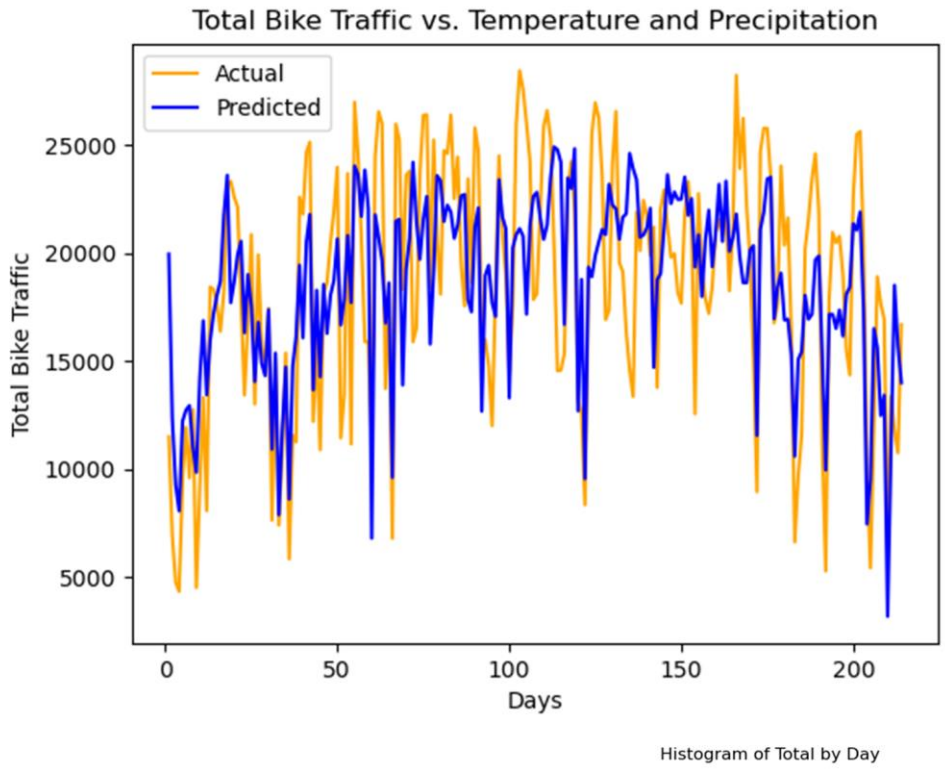
Problem 2 Analysis

In this part of the problem, we are tasked with determining if there is any correlation between the number of people riding bikes on the bridges and the weather conditions. We chose to make another linear regression model using minimum temperature, maximum temperature, and precipitation as features of out model and total traffic as the output value. Once we split our

data and trained our model, we once again tested the model using the remaining data and used the results to determine if there is a correlation between the number of people on the cycling on the bridges and the weather conditions.

Problem 2 Results

The model we developed, and the results of the R-Squared value is as follows:



$$Total\ Bridge\ Traffic\ Model\ 2$$
$$= 406.179(High\ Temp) - 170.783(Low\ Temp)$$
$$- 8034.91(Precipitation) - 422.826$$

&

$$R - Squared\ Value: .433\ or\ 43.33\%$$

Unfortunately, the model that was developed was not very accurate when tested and therefore we are not able to state that there is a clear correlation between weather conditions and the total number of people riding their bikes on the bridges. The R-Squared value was .433 meaning that the model does not produce accurate and reliable results. Our results would force us to recommend that the police not fully rely on the weather conditions to determine when the bike population is high and thus when they enforce the helmet law.

Problem 3 Analysis

In this section of the problem, our objective is to find any relationship between the amount of people using the bridges and the day of the week, in order to determine what day of the week it is. To try to relate the quantity of bicycles to the day of the week, we decided to create a Naive Bayes clustering. Each day was given a number in order to convert the string into a number that could be examined. A confusion matrix was developed after the model was built in order to assess its performance and see if it could be used to forecast the day of the week based on the volume of bicycle traffic.

Problem 3 Results

After applying the Naive Bayes algorithm with the sample, the following results appeared:The percentage of correctly predicted days of the week was 23%, with a confussion matrix

Predicted and Real Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Monday	0	0	4	0	1	0	1

Tuesday	0	0	4	0	1	0	1
Wednesday	0	0	4	0	0	0	0
Thursday	0	0	7	0	0	1	0
Friday	0	0	2	0	1	1	1
Saturday	0	0	1	0	1	1	3
Sunday	0	0	0	0	2	2	4

Table 2: Confusion Matrix of Naïve Bayes

As a result, the approach was unsuccessful, and given the volume of traffic on the roads, we cannot safely anticipate the day of the week.

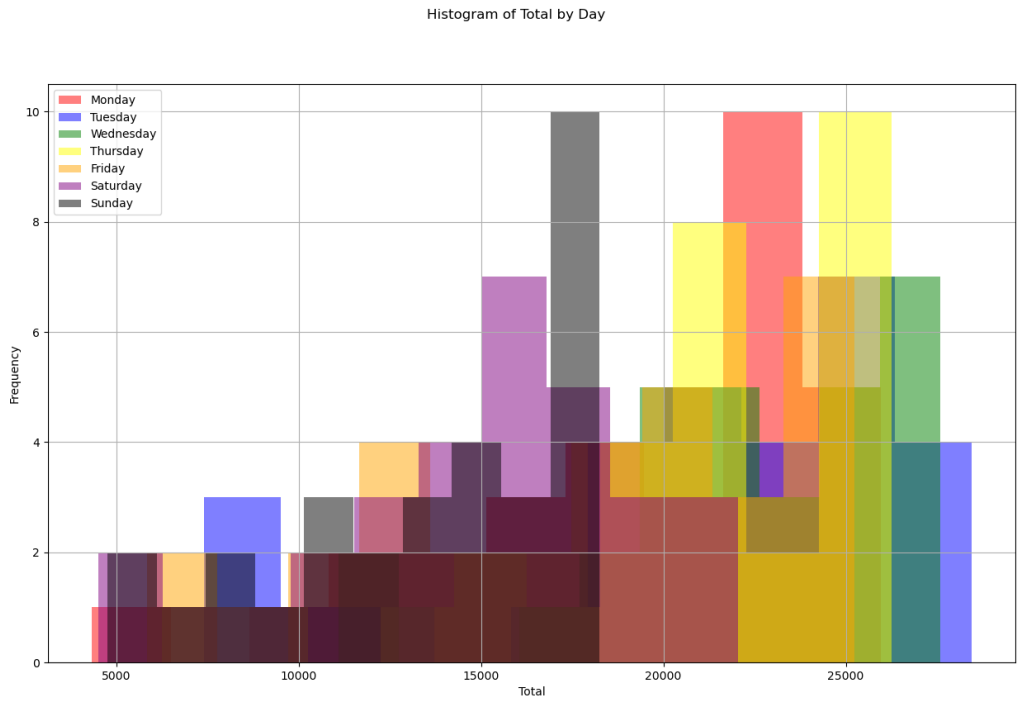


Figure 3: Histograms for each day of the week

From the picture we can see right away that the histograms overlap a lot meaning that the traffic each day is very similar among all the days. This makes the clustering algorithm to be unable to distinguish between different days. We can confirm this information by analyzing the Mean and Standard Deviation of each day.

Day	Mean	Standard Deviation
Monday	19394	5253
Tuesday	20782	5841
Wednesday	22422	4198
Thursday	20781	5033
Friday	17984	5387
Saturday	15000	4402
Sunday	13716	4140

Table 3: Statistics for Each Day of The Week

Because of the distribution of the traffic per day, predicting which day of the week only from the bike traffic would not result in an accurate prediction.