ECE 20875 Mini Project - Bike Traffic

Hai Lam Le and Harini Subramanian

Purdue University

ECE 20875

Instructor: Qiang Qiu

Date: December 6th, 2023

Author Note

First paragraph: Authors' Details

Second paragraph: Dataset Description

Third paragraph: Analysis Questions and Method

Fourth paragraph: Results and discussion

Authors' Details

Path taken: Path 2 - Bike Traffic

1. **Hai Lam Le**

   Section: 002

   Github Username: HaiLamLe

   Purdue Email: le171@purdue.edu

   GitHub repository: https://github.com/ECEDataScience/miniproject-f23-HaiLamLe

2. **Harini Subramanian**

   Section: 002

   Github Username: subra114

   Purdue Email: subra114@purdue.edu

Dataset Description

The New York City Department of Transportation provided a data set to solve a problem. The data tracked the daily bicycle traffic over four New York City bridges: Brooklyn, Manhattan, Williamsburg, and Queensboro. For each day, the data set included the number of bicyclists crossing each bridge along with the low and high temperatures that day, the precipitation that day, and the total bicyclists across all bridges.

Analysis Questions and Method

Question 1

Sensors could be installed on the bridges to estimate overall traffic across all four bridges. However, the budget only allows for sensors to be placed on three of the four bridges - Brooklyn, Manhattan, Williamsburg, and Queensboro. In order to determine which three bridges should have sensors installed in order to most accurately predict the total traffic across all four, we made use of percentage errors. We first created a list that has the names of the bridges, and a list named ‘errors’, which will contain all the percentage errors for the bridges, orders are determined by the bridge names.

After that, we take the average and of each data point by dividing the sum to the number of data points and append it to the list. By the end, we get the list ‘errors’ which contain 4 numbers which are the percentage errors of the bridges: 'Queensboro Bridge', 'Brooklyn Bridge', 'Manhattan Bridge', 'Williamsburg Bridge', respectively. We will not need to install a sensor at the bridge whose error is the highest for that is the most accurate and close to the true mean. “The purpose of a percent error calculation is to gauge how close a measured value is to a true value.” (Anne Marie Helmenstine, 2020)

Question 2

The city administration aims to enforce helmet laws by deploying police officers on high bicycle traffic days to issue citations. By considering factors such as the next day's weather forecast, which encompasses predicted low and high temperatures as well as precipitation, the city will be able to predict the days in which the number of bikers are on the high. Understanding the correlation between these weather variables and the total number of bicyclists across all bridges is necessary for an effective strategy. Therefore, the use of ridge regression is used, focusing on examining the relationship between total traffic levels and the temperature highs, temperature lows, and precipitation to see if there are any clear signs of correlation that would allow us to predict biker levels based on weather.

Question 3

We need to determine if the data set tracking daily bicycle traffic over the Brooklyn, Manhattan, Williamsburg, and Queensboro bridges can be utilized to predict the current day of the week (Monday through Sunday) based solely on the bicycle volumes recorded that day on each bridge. To do this, we use logic regression on the data under the column "Total" and "Day". "Logistic regression is a data analysis technique that uses mathematics to find the relationships between two data factors. It then uses this relationship to predict the value of one of those factors based on the other. The prediction usually has a finite number of outcomes, like yes or no." (*What is Logistic Regression?* 1978) Using Logic Regression will give us a number that describes the accuracy (between 0 and 1), which we can use to predict the accuracy of the data. Since randomness plays a crucial role in logistic regression, every time we run the program, we might get different numbers. We initially thought this would create an issue since accuracy in this case is important, however, we realized that the output number we get every time falls within an acceptable range, and most importantly, is always extremely low.

Results and Calculations

Question 1

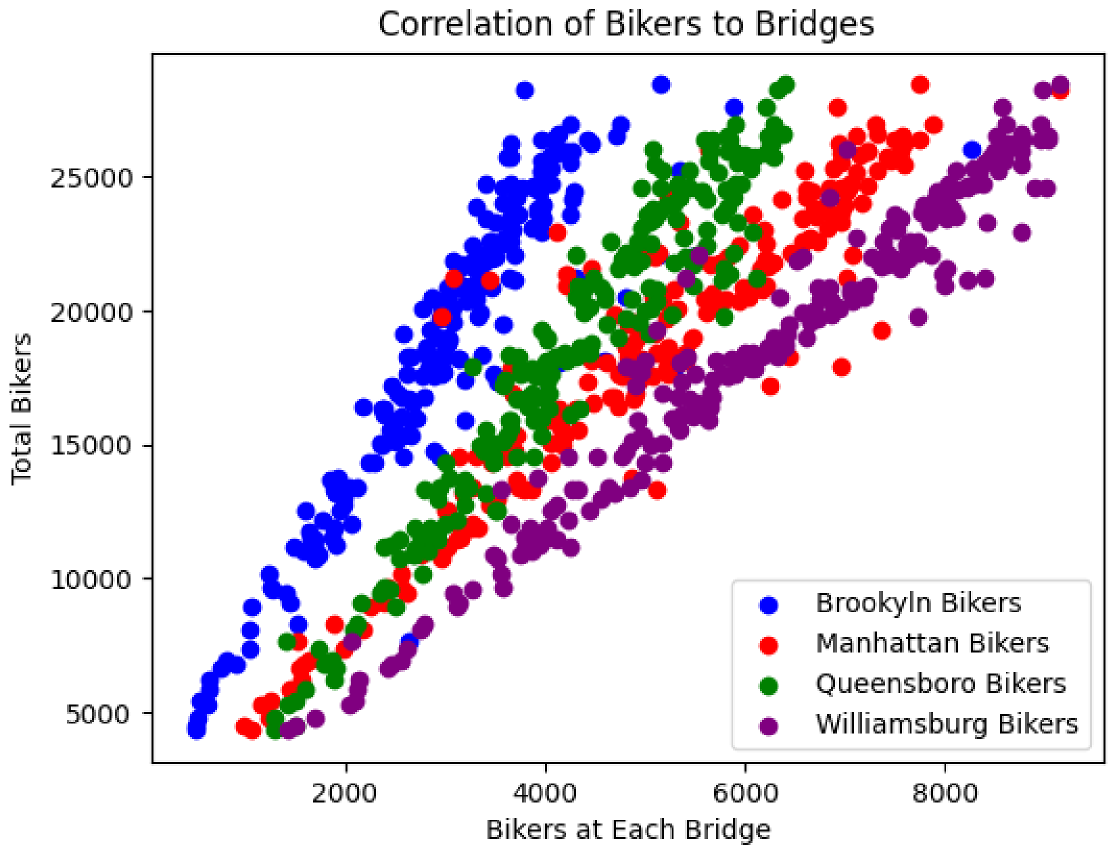

Figure 1.Traffic Plot of each bridge

As observed in the graph, there is a clear division between the traffic on the bridges, making it clear to decide which bridge has the least amount of traffic. Figure 1 shows the bridge with the least amount of traffic in comparison to the others is the Brooklyn Bridge.

We solve this problem using percentage errors. We calculated the percent error of each of the bridge by the basic formula:

$$Percent\ Error = \frac{Each\ data\ point - average}{average}\ 100\%$$

Since a percentage error calculation is used to determine how near a measured number is to an actual value, the higher the percentage error the closer and more accurate it is to the true mean. After performing the calculation in python, we got the percentage error for the 4 bridges are [0.09022294729532944, 0.3638890052857534, 0.12065588345970343, 0.333638163141111837], for ['Queensboro Bridge', 'Brooklyn Bridge', 'Manhattan Bridge', 'Williamsburg Bridge'], respectively. Looking at this data, we can effectively conclude that the percentage error for Brooklyn Bridge is the highest (approximately 36.39%), therefore, we will not have to install a sensor at this bridge.

Question 2

To determine if there is a correlation between the weather and the number of bikers ridge correlation was performed between the number of bikers and the temperature highs, the temperature lows, and the precipitation levels.

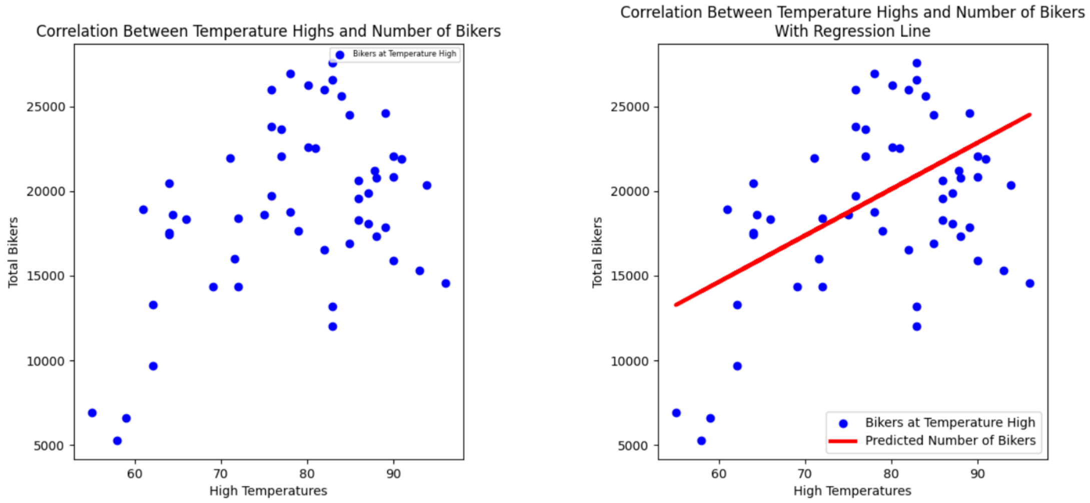

Figure 2.Correlation between Temperature Highs and Bikers

As shown is figure two, there is a positive correlation between the total number of bikers and the temperature highs. The data is typically above the curve between high temperatures between sixty-five and eighty-five degrees and is typically below the curve otherwise.

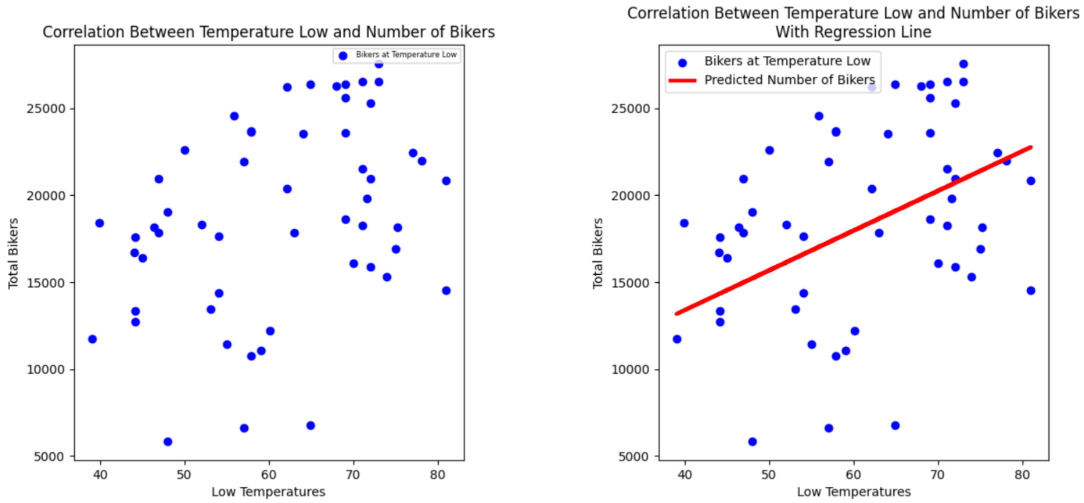

Figure 3.Correlation between Temperature Lows and Bikers

As shown in figure three, just as was shown in figure two, there is a positive correlation between the total number of bikers and the temperature lows. The data is typically above the curve between high temperatures between sixty-five and eighty-five degrees and is typically below the curve otherwise. There are about twenty-eight points above the curve and twenty-four points below the curve. Because of this, it is determined that there is no strong correlation between the number of bikers.

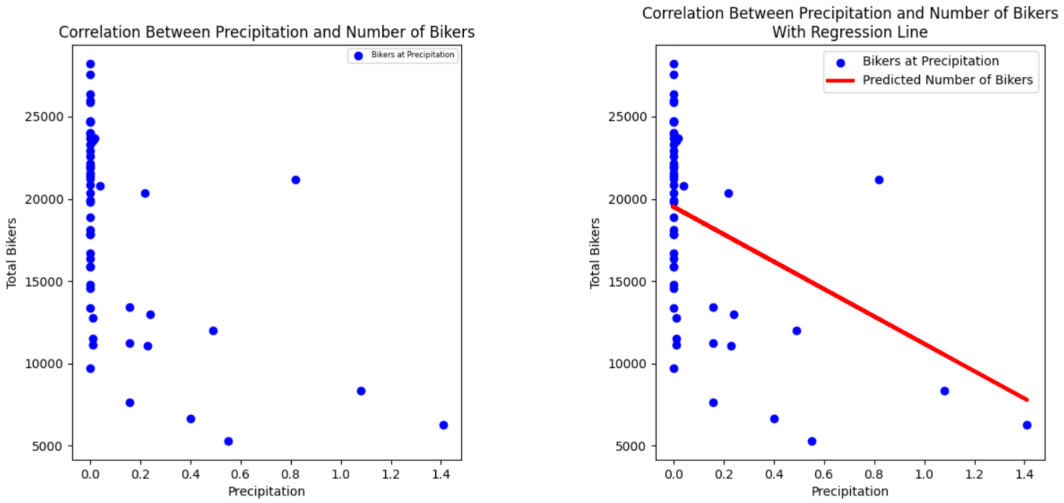

Figure 4.Correlation between Precipitation and Bikers

As shown in figure four, the regression line demonstrates a negative correlation between the number of bikers and the precipitation levels. However, when looking at the data points, most of the data points are when the precipitation levels are at zero, and have a varying number of bikers from about ten thousand and thirty thousand. This demonstrates that there is no strong correlation between the number of bikers and precipitation.

Based on the data from figures two, three, and four, we recommend that the police are deployed on days where there are no precipitation levels and and have high temperatures between sixty-five and eighty-five degrees.

Question 3

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Due to the inherent randomness in logistic regression models, repeated runs of the program can yield varying numerical outputs. Though initially concerned given the importance of high accuracy, further analysis revealed that the results consistently fall within a satisfactory range across iterations. Most significantly, the error rate generated remains extremely minimal irrespective of the precise figures. Here are some of the values that we got while running the program: 0.09302325581395349, 0.11627906976744186, 0.06976744186046512.

Due to the incredibly low accuracy score, it is safe to say that it would be very difficult to predict the day (Monday to Sunday) based on the number of cyclists on the bridges.

References

Anne Marie Helmenstine, Ph. D. (2020, November 2). *This is how to calculate percent error*. ThoughtCo. https://www.thoughtco.com/how-to-calculate-percent-error-609584

The University. (1978). *What is Logistic Regression?*. Amazon. https://aws.amazon.com/what-is/logistic-regression/#:~:text=Logistic%20regression%20is%20a%20data,outcomes%2C%20like%20yes%20or%20no.