ECE 20875-001

Prof. Pare

Seeun Kim (GitHub: kim3729)

Hyunsang Cho (GitHub: phenomenon928)

ECE 20875 Mini Project: Bike Traffic

1. Introduction

From NYC Bicycle Counts in 2016 CSV file, the information on bike traffic across four different bridges in New York City is given. The information from the data gives us the high and low temperatures, precipitation, and the number of bicyclists across each bridge for each day starting from April 1st to October 31st. Using the given information, we are going to install sensors on three of the four bridges to estimate overall traffic across all the bridges. The reason for selecting three bridges out of four is due to the lack of financial assets needed to assist all four bridges with sensors. To get the best prediction of overall traffic, the bridge that will be excluded from the data would be selected with reasonable reasons. Then, the information related to the remaining three bridges would be used for the next steps. The city administration is strictly enforcing the helmet laws and intends to dispatch police officers to impose fines on days when there is a lot of traffic. We use those three bridges to predict the total number of bicyclists on a day with its weather forecast, including low and high temperatures and precipitation. Lastly, we decide if we can use the data to predict what day it is from Monday to Sunday based on the number of bicyclists on the bridges.

2. Approach

Problem 1:

There are two different approaches that we use to select the bridge that would be excluded from installing the sensors. First, with the given data which indicates the number of traffic for each bridge, we created scatter plots from the values of the number of trafficking each day starting from April 1st. However, we identified several outliers in raw data (see Figure 1A), so we preprocessed the data using Interquartile Range (IQR) method to remove them. Then, we created new plots without outliers in order to create more precise and accurate plots (see Figure 1B). From the new scatter plots, we calculated the r-squared values in order to find out the bridge with the lowest r-squared value, as it represents how well the regression model explains observed data.

In order to support the first method, the other approach was using Mean Absolute Deviation (MAD) method. We computed the average total bike traffic on all four bridges. Then, we measured the deviations from total traffic on each bridge to the average of four. From those measurements, we concluded to exclude the bridge with the highest absolute deviation from the data.

Based on the results of the two different approaches, we exclude the bridge that has the lowest R-squared value and the largest absolute deviation from the average.

Problem 2:

To predict the day's total number of bicyclists, we plotted three graphs of total bike traffic over the average temperature: one with the precipitation data, one without the precipitation data, and two plots aforementioned in one graph for better

visualization. We made assumptions that precipitation at 0 be ‘no precipitation’ and above 0 be ‘precipitation’ and the average temperature of each day is to be used. With the data, we drew a linear regression line and trendline for each graph (see Figure 3). A linear regression line shows the size and direction of the change in Total Bike Traffic as a result of the change in Daily Average Temperature, and a trendline shows that Total Bike Traffic is increasing or decreasing at a steady rate.

The user is asked to input the weather condition: precipitation, highest temperature, and lowest temperature. The average temperature is calculated by averaging the highest and lowest temperature the user inputted. If the inputted precipitation value is 0, a ‘no precipitation’ graph and linear regression equation are used, and if the imputed precipitation value is above 0, a ‘precipitation’ graph and linear regression equation are used to predict the total bike traffic. Then, with the average temperature of the user inputted highest and lowest temperature value, we expect the total number of bicyclists on that day using the chosen linear regression line.

Problem 3:

To predict the day (Monday through Sunday), we calculated the average of the total number of bicyclists for each day and compared the input for the total number of bicyclists. Therefore, we found the day that has the closest number for the average. However, we need to check how accurate and precise the result of the predicted day is, because this can be affected by many different facts.

3. Conclusion / Analysis

Problem 1:

Using the two approaches, we chose the one bridge to be excluded and installed sensors on the rest bridges. As seen in Figure 1B, the r-squared value of the Brooklyn bridge is 0.0057, that of the Manhattan bridge is 0.0202, that of Queensboro is 0.0337, and that of Williamsburg is 0.0263. According to figure 2, the Brooklyn bridge has the largest deviation from the average, which is 0.992 million bike traffic. The absolute deviations of each bridge from the average are the following:

Brooklyn Bridge: 359299.75, Manhattan Bridge: 94291.25,

Williamsburg Bridge: 331540.25, and Queensboro Bridge: 66531.75

Therefore, the Brooklyn bridge has the lowest r-squared value and largest Mean Absolute Deviation. As a result, since both approaches show the Brooklyn bridge is the least favorable dataset, we concluded that it is the best choice to install sensors on the Manhattan, Queensboro, and Williamsburg bridges. (see Figure 5)

Problem 2:

We concluded with two different linear regression line equations for precipitation and no precipitation. The linear regression equation for the 'precipitation' graph is $y = 257.427x - 5250.524$ and that for the 'no precipitation' graph is $y = 136.486x + 7713.267$ where x is the average temperature in Fahrenheit and y is the total bike traffic within a 24 hours period. The third graph of Figure 3 shows the comparison of 'precipitation' and 'no precipitation' regression lines.

Problem 3:

As shown in Figure 4, we found the average number of bicyclists on each day: Monday through Sunday. The input from the user is compared with the calculated number of average. Then, the output is the predicted day that has the closest number of average with the input (see Figure 5). We conclude this is just “predicted day”, not one hundred percent correct. This means there is a large error possibility. The reason is because the average number of bicyclists can be changed depending on holidays, weather, etc.

4. Figures

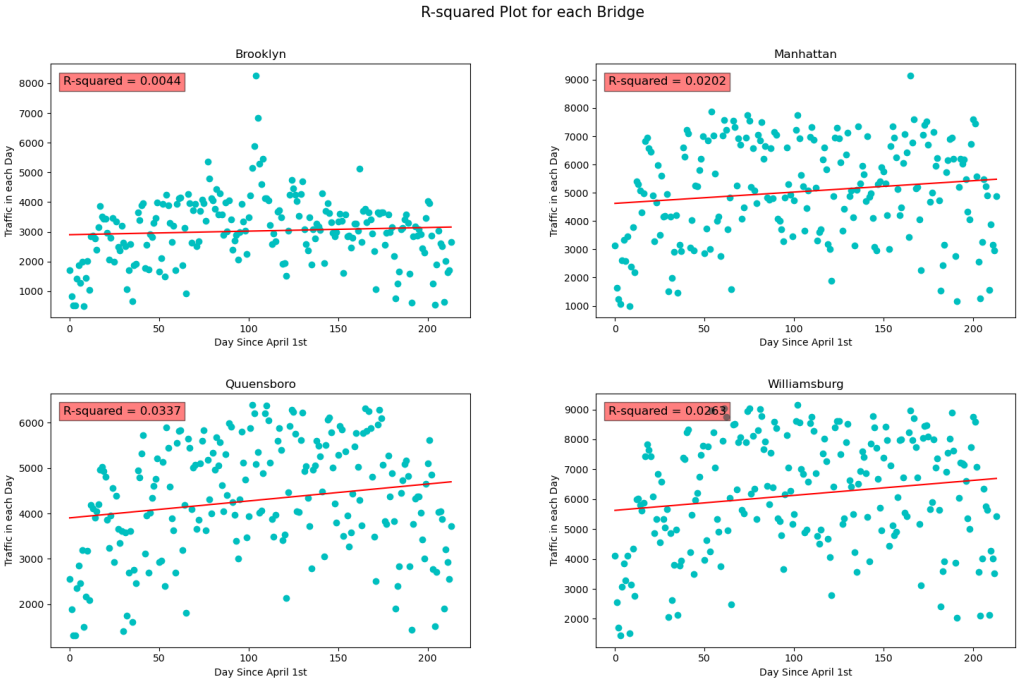

Figure 1A. Data before IQR Preprocessing

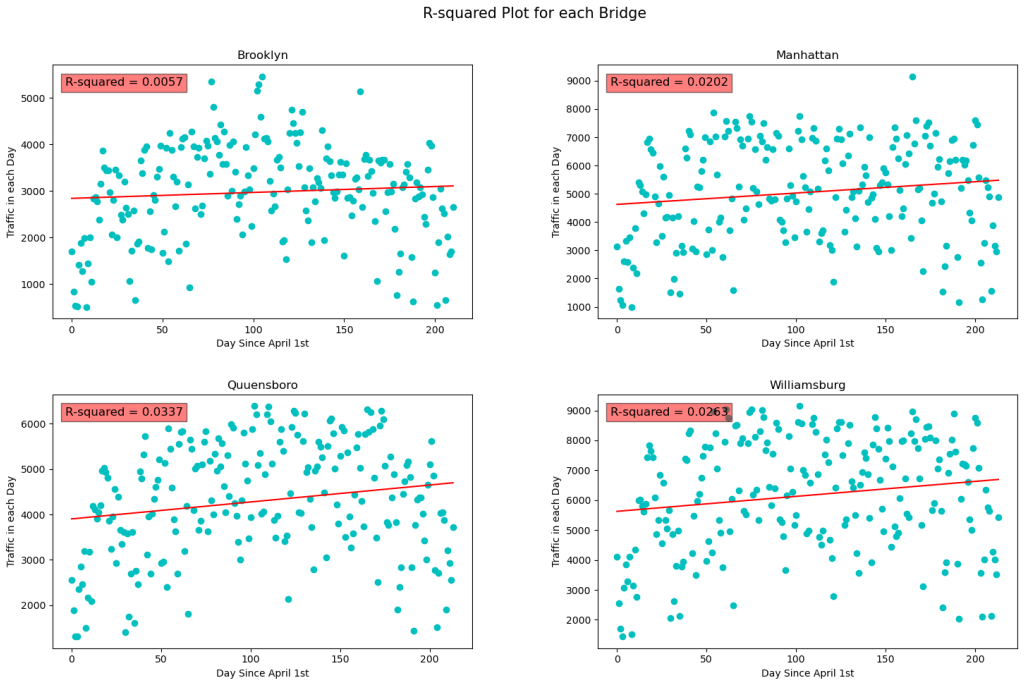

Figure 1B. Data after IQR Preprocessing

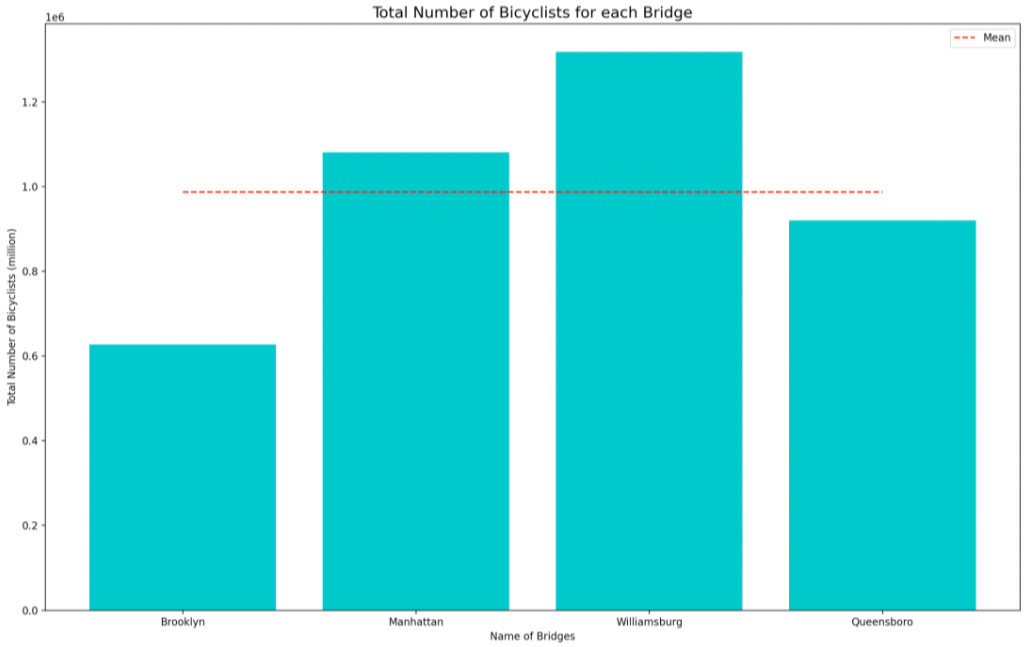

Figure 2. Total Number of Bicyclists for each Bridge compared with Mean

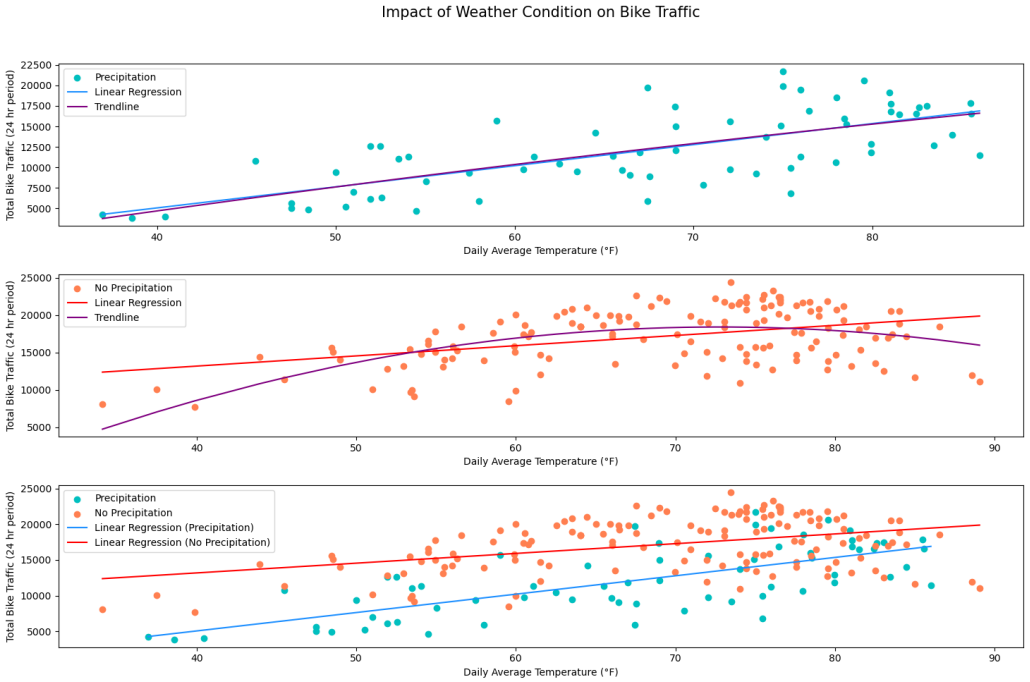

Figure 3. Impact of Weather Conditions on Bike Traffic

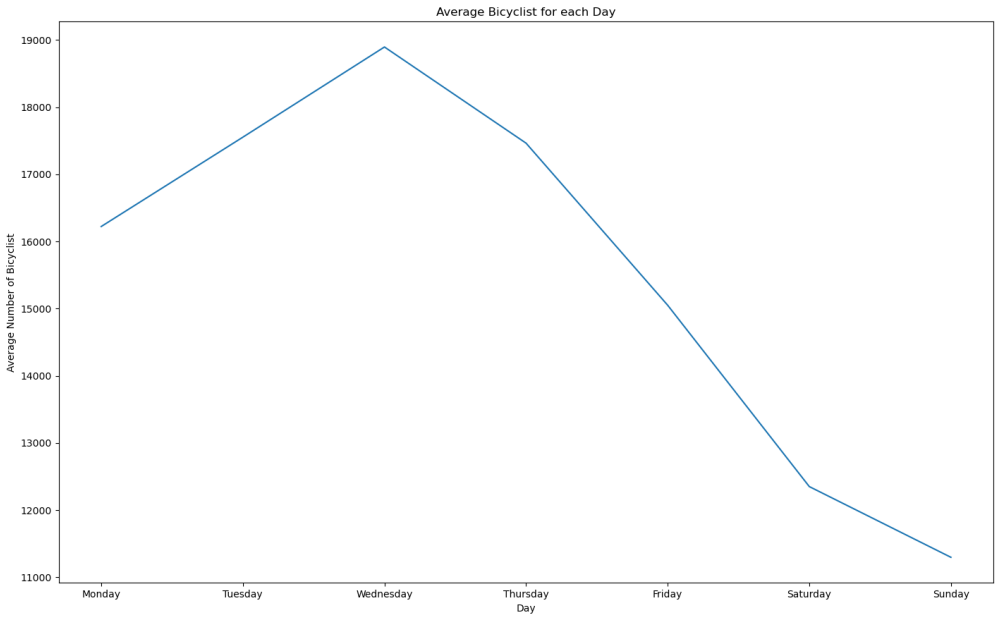

Figure 4. Average Bicyclist for each Day

```
Problem 1:
Selected Bridges: ['Manhattan Bridge', 'Williamsburg Bridge', 'Queensboro Bridge']

Problem 2:
Enter Precipitation: 0.5
Enter Highest Temperature (°F): 75
Enter Lowest Temperature (°F): 50
Predicted Traffic: 10838

Problem 3:
Enter the total number: 17500
Predicted Day: Thursday
```

Figure 5. Output