ECE 20875 Mini-project

Name: Rushil Shah
Email: shah660@purdue.edu
Github Username: RushilShaw
Name: Avigdor Roytman
Email: aroytman@purdue.edu
Github Username: av1roytman

Path #1

Dataset Description

The path one problem set gives us a dataset with a comma-separated-version(.csv) file named “NYC_Bicycle_Counts_2016_Corrected”. This file has the following columns: Date, Day, High Temp, Low Temp, Precipitation, Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge, Queensboro Bridge, Total. So, bases on the day, we could determine the weather and amount of cyclist on bridges based on different bridges in NYC. Using this data, we can solve the problems laid out in the path one problem set.

Analysis

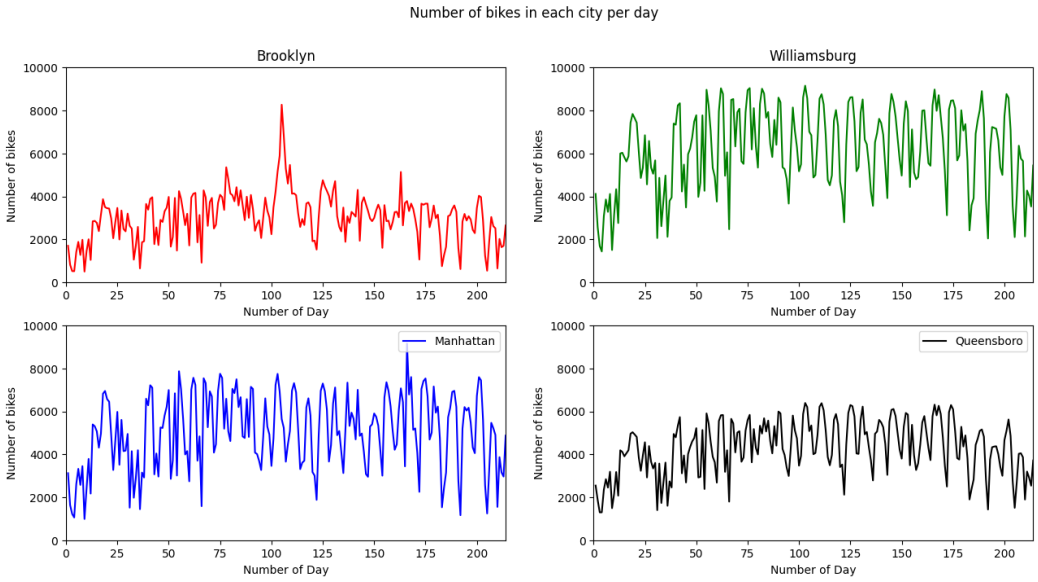
In problem one, we were asked to install sensors on bridges to estimate overall traffic across all the bridges. However, the budget only allows for three of the four bridges to be equipped with these sensors. We need to determine which bridges would need these sensors. To approach this problem, we believed that the best method to solve the problem is to look at the R-Squared of each of the datasets. This is because, if a one bridge has a particularly high R-Squared, then we know that it has a similar amount of people on the bridge regardless of temperature, date, or precipitation, and therefore wouldn't need sensors, as the other factors could give an accurate prediction of the amount of people on the bridge during that particular day. Another approach would be to look at the standard deviation of the bridges and select the bridge with the lowest standard deviation since the number of cyclists can be predicted without the use of sensors.

In problem two, we are tasked to use the next day's weather forecast to predict the total number of bicyclist that day. To solve this problem, we hypothesis that the best method to solve the problem is to look at the R-Squared of each of the datasets and see the correlation between temperature, precipitation and number of cyclists. If these values has a high R-Squared, then we can predict the amount of cyclist on the bridge from temperature and precipitation data, and can efficiently deploy the police force on high traffic days.

In problem three, we want to find the day of the week based on the number of cyclists on the bridge. The idea being that of people ride their bikes on bridges a different amount based on the day of the week, then given the number of cyclists, we should be able to determine the day of the week. To solve this problem, we formulated that the best solution was to look at the mean and standard deviation, if each day had a different mean and each day had a relatively low standard deviation. Then we should be able to accurately predict the day based on the number of cyclists.

Results

For the first problem, we were asked to install sensors on bridges to estimate overall traffic across all the bridges. However, the budget only allows for three of the four bridges to be equipped with these sensors. We needed to determine which bridges would need these sensors. We completed this by looking at the R-Squared of each of the datasets. The rationale behind this being if a one bridge has a particularly high R-Squared, then we know that it has a similar amount of people on the bridge regardless of temperature, date, or precipitation, and therefore wouldn't need sensors, as the other factors could give an accurate prediction of the amount of people on the bridge during that particular day. Below is a graph showing the number of bikers on a bridge on any given day.



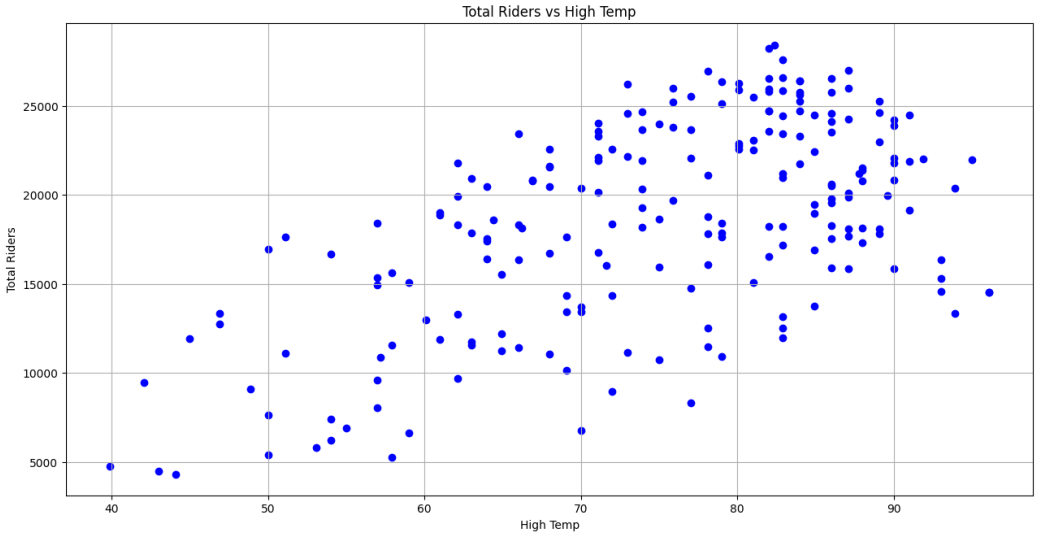
From these graphs we were NOT able to conclude that that the relationship between weather forecasting and number of bikers. So instead, we decided to pick the bridge with the lowest standard deviation. The logic behind this being that the bridge with the lowest standard deviation

is the most likely to not need sensors since the numbers of biker can be predicted without the need of sensors.

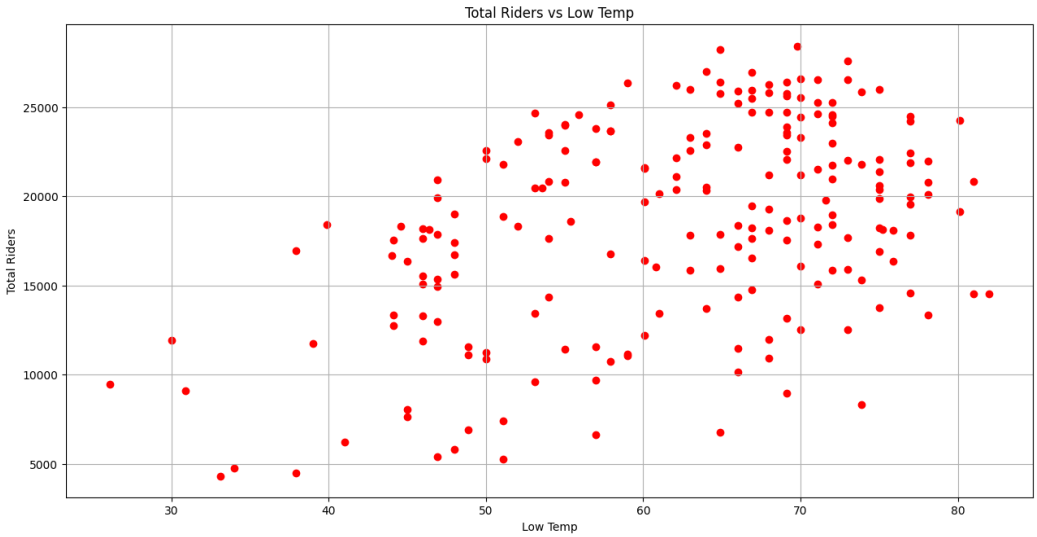
```
Standard Deviation of  of Brooklyn Bridge is 1131.392
Standard Deviation of  of Manhattan Bridge is 1741.402
Standard Deviation of  of Williamsburg Bridge is 1906.174
Standard Deviation of  of Queensboro Bridge is 1258.036
```

From this approach we found that the **Brooklyn Bridge** would be the bridge of choice to not be equipped with sensors.

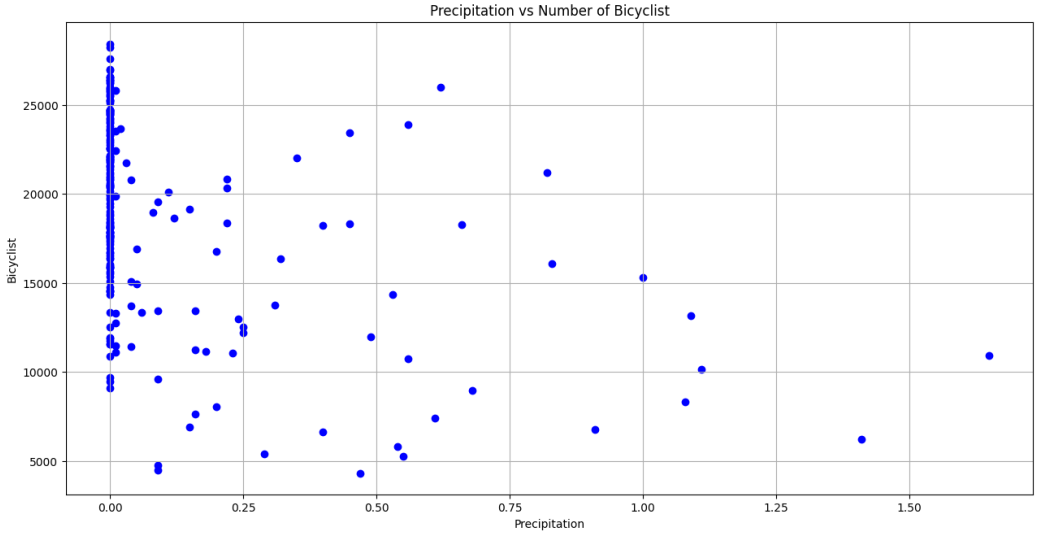
For the second problem, we are tasked to use the next day's weather forecast to predict the total number of bicyclist that day. To solve this problem, we took a look at the R-Squared of each of the datasets and saw the correlation between temperature, precipitation and number of cyclists. If these values have a high R-Squared, then we can predict the amount of cyclist on the bridge from temperature and precipitation data, and can efficiently deploy the police force on high traffic days. We prepared the following graphs to represent our data. This prediction model



This is a graph shows the relationship between the number of riders with high temperature.



This is a graph shows the relationship between the number of riders with low temperature.



This is a graph shows the relationship between the number of riders with precipitation.

Using the above data, we created a prediction model to predict the amount of cyclist on a bridge give the information of temperature and precipitation. The predication model had an R-squared of 0.374. Since this R-Squared is below a 0.4. We can say that this dataset has no correlation and that we cannot reliably use temperature and precipitation data to predict the amount of cyclist on a bridge.

For the third problem, we wanted to find the day of the week based on the number of cyclists on the bridge. The idea behind this being that of people ride their bikes on bridges a different amount based on the day of the week, so given the number of cyclists, we should be able to determine the day of the week. To solve this problem, we formulated that the best solution was to look at the mean and standard deviation, if each day had a different mean and each day had a relatively low standard deviation. Then we should be able to accurately predict the day based on the number of cyclists using the z-score values. The image below has each day of the week and the respective standard deviation and mean.

```
Days:   ['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday']
STDEV:  ['5167.76', '5743.62', '4127.46', '4948.69', '5300.31', '4330.53', '4073.50']
Mean:   ['19393.71', '20782.27', '22422.27', '20781.30', '17984.58', '15000.65', '13716.39']
Standard Dev of Means:  2974.70
```

Since the means, are close together and the standard deviations are high compared to that (the standard deviations are all higher than the standard deviation of the means) we cannot confidently the day of the week based on the number of bicyclists on a given day.