Team Name: Wayne Enterprises

Members: Adhil Akbar

Username: akbara@purdue.edu

Chosen Path: (1)

Bike Traffic Analysis

The Dataset:

The dataset contained in the file "NYC_Bicycle_Counts_2016_Corrected.csv" gives information on bike traffic across a few bridges in New York City. Within this file, data is recorded from April 1st to October 31. Between this timeframe, data has been collected for *Day*, *High Temp (°F)*, *Low Temp (°F)*, *Precipitation*. In addition to this, the number of bikes on *Brooklyn Bridge*, *Manhattan Bridge*, *Williamsburg Bridge*, *Queensboro Bridge* have also been collected. In addition, the *Total* number of bikes across all three bridges is also collected. Within this dataset, modifications must be done first to perform an analysis. In the *Precipitation* column, two areas of particular interest are 0.47 (S) and T. To accommodate this a decision was made to change 0.47 (S) with 0 as it seems to be an outlier. T has been changed to 0, because it is defined as an amount measuring less than 0.01.

The Analysis:

#Which bridges should you install the sensors on to get the best prediction of overall traffic?

To first answer this question, tests were conducted to better understand the data being used. The first step was to develop descriptive statistics and histograms for the traffic on each Bridge. This will allow us to better understand the spread of traffic amongst all the bridges. In addition to this, it also gives us information about the Sample Mean, Standard Error and Standard Score which may aid in determining which three bridges to use. However, this is simply a preliminary analysis. To finalize which bridges to use, four linear regression models will be analyzed, treating one of the bridges as the target variable and all others as explanatory variables. From this analysis, the target bridge which yields the lowest Mean Squared Error will be the bridge we decide not to use.

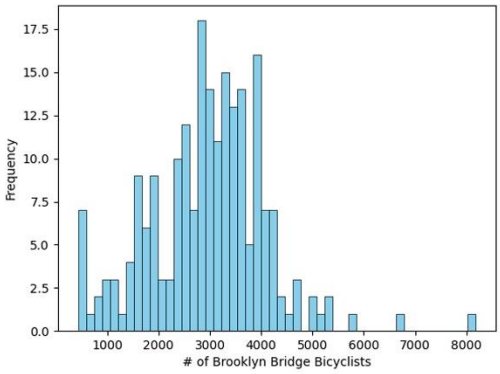
#Can they use the next day's weather forecast to predict the number of bicyclists that day?

To answer this question, linear regression models were performed like the one above. However, this time, multiple variables are needed to determine if a prediction can be made. Linear Regression is used when we want to predict the value of a variable based on the value of another. The variables we are given relating to weather consist of High Temp (F), Low Temp (F) and Precipitation. To determine whether we can predict based on the forecast, different independent variables will be tested. These variables will consist of High Temp (F), Low Temp (F), Avg/Median Temp (F), and Precipitation. These variables will be plotted against Total Traffic and Average Traffic. From all these plots, linear regression using the least squares method will be performed and R^2 Values will be calculated to determine if next day's weather forecast can predict the number of bicyclists. An R^2 value between 0 and 1 determines the strength of our ability to predict based on weather.

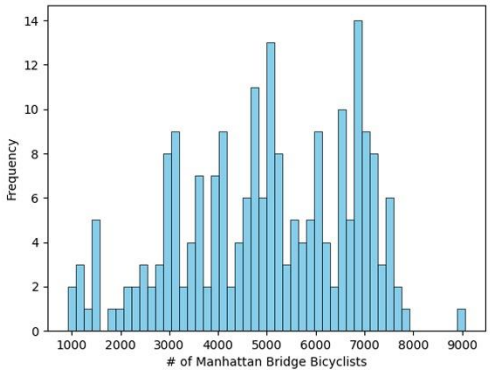
#Can you use this data to predict whether it is raining based on the number of bicyclists?

To answer this question, linear regression models were performed like the one above. However, this time an addition test is used. The additional test being used is Spearman's Correlation test. For this specific test, the two variables we are testing can be related in a nonlinear manner and may have non-Gaussian distribution. This is ideal for this scenario because we are unsure if there is a linear relationship or a Gaussian distribution. After performing both tests, the calculated R^2 values and Spearman's Correlation Coefficient will be found. An R^2 value between 0 and 1 determines the strength of our ability to predict based on rain. In addition, Spearman's Correlation Coefficient will yield a score between -1 and 1. A value of 1 means there is a perfect association, 0 means no association and -1 means a perfect negative association.

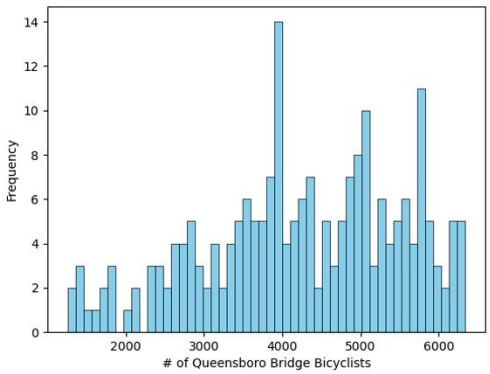
Q1 Results:



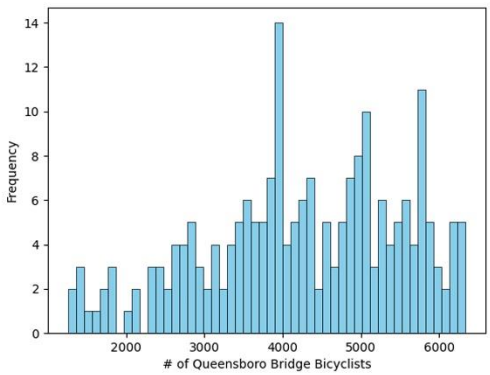

```
Data Values for Brooklyn Bridge:
Sample Size: 214
Sample Mean: 3030.700934579439
Standard Error: 77.5217083256169
Standard Score: 39.08519303847949
```



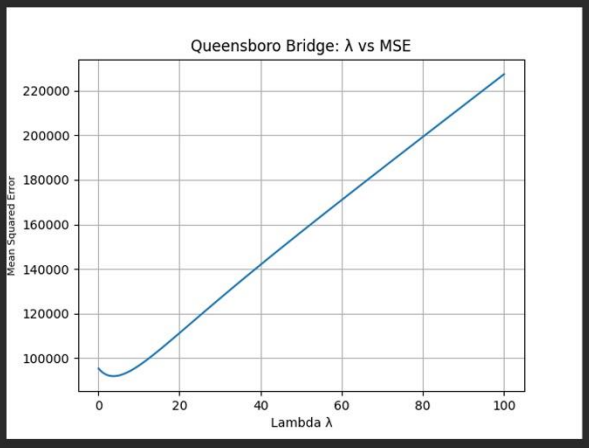

```
Data Values for Manhattan Bridge:
Sample Size: 214
Sample Mean: 5052.2336448598135
Standard Error: 119.31892601708162
Standard Score: 42.33597982717885
```



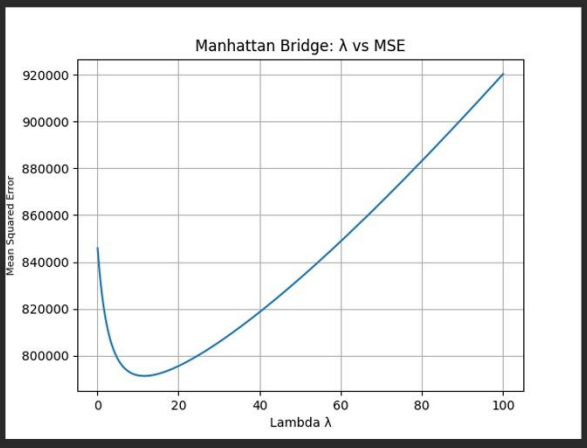

```
Data Values for Williamsburg Bridge:
Sample Size: 214
Sample Mean: 6160.873831775701
Standard Error: 130.6088738720642
Standard Score: 47.164665379549554
```



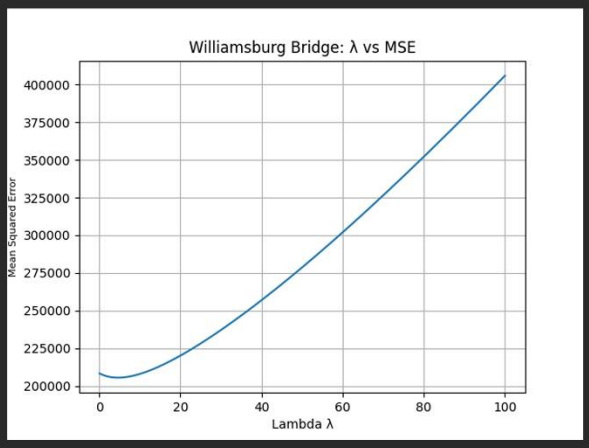

```
Data Values for Queensboro Bridge:
Sample Size: 214
Sample Mean: 4300.72429906542
Standard Error: 86.19920958149402
Standard Score: 49.8841499816789
```



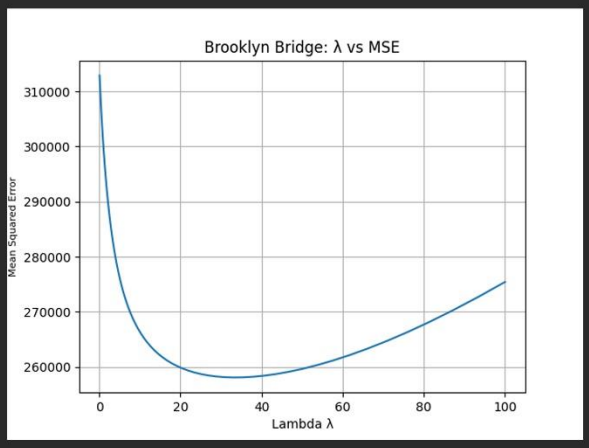

```
Best lambda Tested: 3.890451449942805
Yielded MSE Value: 91879.94913249096
Model Best Coefficients:
[[ 222.95898611  -56.04951768 1028.09150025]]
Model Best Intercept:
[4264.94375]
```




```
Best lambda Tested: 4.786300923226383
Yielded MSE Value: 205521.2246253878
Model Best Coefficients:
[[ -30.27893545  559.977064   1312.52152314]]
Model Best Intercept:
[6124.15]
```




```
Best lambda Tested: 11.748975549395292
Yielded MSE Value: 791306.759971486
Model Best Coefficients:
[[ 308.53905047 1033.48292446  147.28667656]]
Model Best Intercept:
[4980.025]
```



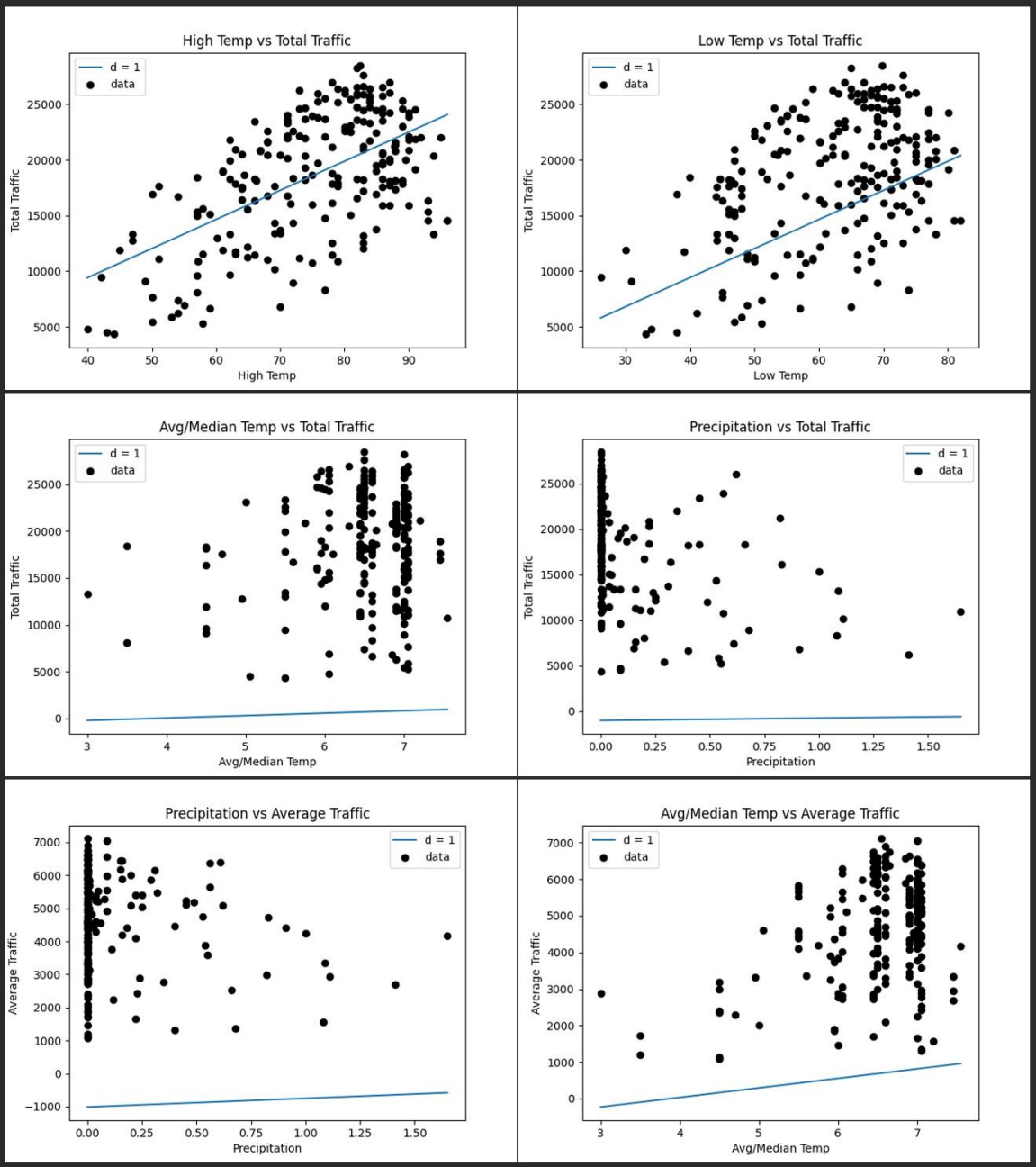

```
Best lambda Tested: 33.11311214825911
Yielded MSE Value: 258082.48067652673
Model Best Coefficients:
[[263.60859838 203.66150417 410.60862125]]
Model Best Intercept:
[3043.48125]
```

Q1 Conclusion:

#Which bridges should you install the sensors on to get the best prediction of overall traffic?

The Standard Score is the number of standard deviations a given data point lies above or below the mean. The mean is the average of all values in a group, added together, and then divided by the total number of items in the group. From the descriptive statistics that were calculated, it was found that amongst the four bridges, Queensboro and Williamsburg both had the highest Standard Score and were very close to each other. When the Standard Score is higher it indicates that the expected results are likely to be different from what is expected. Furthermore, the spread of all 4 bridges seemed to follow a normal distribution. Williamsburg had the highest average traffic indicating the popularity, lack of efficiency or convenience of its location. The Mean Squared Error tells us how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the "errors") and squaring them. After calculating the MSE values treating each bridge as the target variable, it was found that Queensboro yielded the lowest MSE amongst the four models and Williamsburg the highest MSE. From this we can concur that it is best to remove Queensboro from the analysis to reduce the costs of installing sensors. Manhattan, Brooklyn, and Williamsburg are the recommended bridges to use.

Q2 Results:



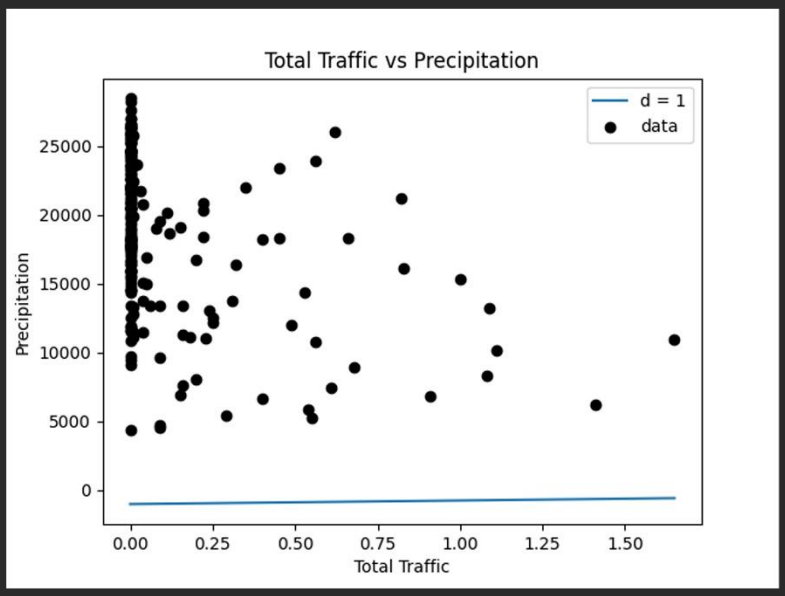

```
The R-Squared Value for High Temp vs Total Traffic is 0.3296814258860437
The R-Squared Value for Low Temp vs Total Traffic is 0.1954960163493372
The R-Squared Value for Avg/Median Temp vs Total Traffic is 0.015742783067257633
The R-Squared Value for Precipitation vs Total Traffic is 0.16098276881324405
The R-Squared Value for Precipitation vs Average Traffic is 0.029203600296799814
The R-Squared Value for Avg/Median Temp vs Average Traffic is 0.11363865056915033
```

Q2 Conclusion

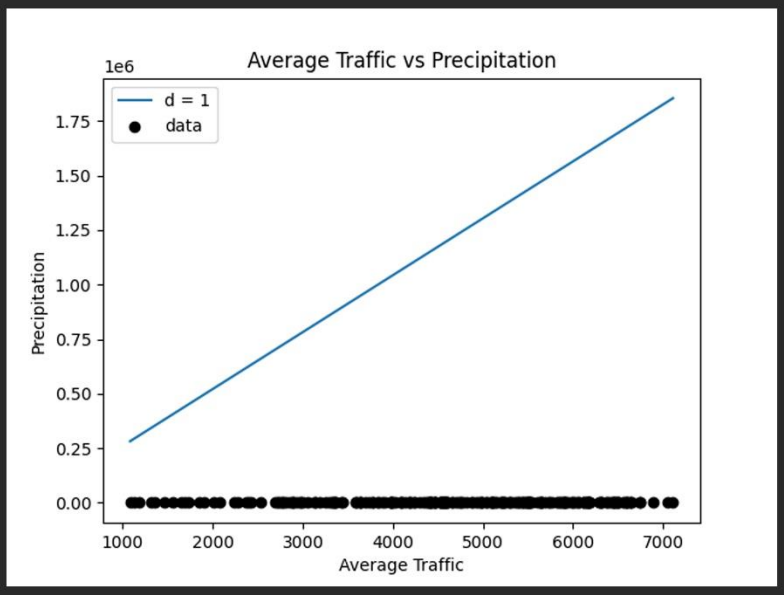
#Can they use the next day's weather forecast to predict the number of bicyclists that day?

For this question, many different relationships were looked at and linear regression using the least squares method was performed on each one to determine if one could use the next day's weather forecast to predict the number of bicyclists on that day. Visually looking at the graph one can see that almost all high levels of traffic are clustered in areas where there is no precipitation. The same can be said for average/median temp as well. However, this is not enough information to simply say a prediction can be accurately made. This doubt is clarified upon calculating the R^2 values for each relation. R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R^2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs. Amongst the many relationships that were analyzed, the highest R^2 value was 0.3296 and the lowest was 0.0157. This means that at best, 33% of the observed variance can be explained by the model's inputs. Based on this we can conclude that next day's weather forecast cannot be used to the predict the number of bicyclists that day.

Q3 Results:




```
The R-Squared Value for Total Traffic vs Precipitation is 0.16098276881324405
Spearman"s Correlation for Total Traffic vs Precipitation: -0.057
```




```
The R-Squared Value for Average Traffic vs Precipitation is 0.029203600296799814
Spearman"s Correlation for Average Traffic vs Precipitation: -0.057
```

Q3 Conclusion

#Can you use this data to predict whether it is raining based on the number of bicyclists?

For this question, Precipitation vs Total and Avg Traffic were analyzed, and linear regression using the least squares method was performed on each one to determine if one could use the next day's weather forecast to predict the number of bicyclists on that day. Visually looking at the graph one can see that almost all high levels of traffic are clustered in areas where there is no precipitation. Upon calculating the R^2 values for each relation we can develop an initial conclusion. R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R^2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs. Amongst the many relationships that were analyzed, the highest R^2 value was 0.029 and the lowest was 0.016. This means that at best, 1.6% of the observed variance can be explained by the model's inputs. To further solidify this conclusion, Spearman's Correlation coefficient is calculated. This measures the strength and direction of the association between two variables. A value of 1 means there is a perfect association, 0 means no association and -1 means a perfect negative association. From the analysis we see that for both variables, the coefficient is -0.057. This indicates that are almost no association between both variables. Based on these 2 tests we can conclude we cannot predict whether it is raining based on the number of bicyclists that.