ECE 20875 Mini Project Report

Jake Scherer (schere16) and Aastha Patel (pate1695), Section 001
Path 2: Bike Traffic
GitHub Link: https://rb.gy/07ikxw

Dataset Description

The dataset presented to us, nyc_bicycle_counts_2016.csv, contains bike traffic data from four New York City bridges in the 2016 year. These bridges include Brooklyn, Manhattan, Williamsburg, and Queensboro from April 1 to October 31. The traffic data includes the number of bikes that crossed the given bridge each day, along with the total number of bikers across all four bridges. Along with traffic, this dataset consists of the day of the week and the weather forecast, including the high temperature, the low temperature, and the precipitation for the day. Overall, the data presents bike traffic and the overall conditions of the day for New York City.

Analysis and Predictions

Question 1:

In question one, we are tasked with installing sensors on the three of the four bridges to best estimate overall traffic. The analysis we chose to use was a correlation matrix, which will show the degree of correlation between each bridge and the total bike traffic for the day. The value provided is the coefficient of determination (R^2), a value from 0 to 1 describing variance between x and y, with 1 being no variance. We expect this to show us which bridge least correlates to the other bridges as well as the overall traffic. Whichever bridge does not correlate as well to the other bridges will be the bridge we choose not to install a sensor on. We decided to choose the three bridges that have the least variance to each other and the total to install sensors on.

Question 2:

Question two focused on finding a correlation between the weather forecast and the total number of bicyclists a day, focusing on good days to hand out traffic citations. Initially, we created a correlation matrix between the forecast (high/low temperature and precipitation) and the total traffic each day. This matrix gives us the coefficients of determination for each relationship, and will demonstrate how related each element of the forecast is to the total bike traffic. We then created a polynomial fit regression model between forecast and total traffic with a training size of 80% and a testing set of 20% to see if there would be a strong correlation. To continue further, we ran the model with different degrees, varying from 1 to 4, and ran each degree a minimum of five times. We then took the mean squared error of each iteration and

created an average MSE for each degree. These analysis techniques should illustrate the type of correlation between forecast and total traffic if there is one, as well as how accurate that correlation is.

Question 3:

To determine if the day (Monday to Sunday) could be predicted based on the total number of bicyclists that day, we extracted the data into separate arrays by day. From this we can graph and fit those datasets to find patterns and correlation between the total bike traffic and the day of the week. We expect to find relatively consistent total traffic numbers by day with a linear fit line that is nearly horizontal.

Results

Question 1:

Based on the analysis done through a correlation matrix of the bridge/total traffic, we found that the Brooklyn Bridge correlated the least with the other bridges' traffic as well as the total traffic. For this reason, we decided to install sensors on the Manhattan, Queensboro, and Williamsburg bridges to best estimate the overall traffic.

	Brooklyn Bridge	Manhattan Bridge	Queensboro Bridge	Williamsburg Bridge	Total
Brooklyn Bridge	1.000000	0.751713	0.813207	0.792604	0.874413
Manhattan Bridge	0.751713	1.000000	0.838967	0.878377	0.935474
Queensboro Bridge	0.813207	0.838967	1.000000	0.965399	0.963180
Williamsburg Bridge	0.792604	0.878377	0.965399	1.000000	0.975089
Total	0.874413	0.935475	0.963180	0.975089	1.000000

Table I: Correlation Matrix of Bridge Traffic vs. Total Traffic

Question 2:

Based on our initial correlation matrix (Table II), we were unable to determine a strong correlation between the weather forecast and the total traffic. All of the correlation coefficients were below 0.6, which is not sufficient enough to discern a relationship. For our polynomial fit regression model, we printed out the mean squared error (MSE) for each regression model. The MSE values were consistently high for each iteration, and as the degrees increased, the MSE values continued to skyrocket. Since we were unable to solidify a model of fit for the forecast and total traffic data, we have determined that there is not a strong enough correlation between the two to predict total traffic based off of the weather forecast.

	High Temp	Low Temp	Precipitation	Total Traffic
High Temp	1.000000	0.917376	-0.052069	0.574179
Low Temp	0.917376	1.000000	0.040390	0.442149
Precipitation	-0.052069	0.040390	1.000000	-0.420711
Total Traffic	0.574179	0.442149	-0.420711	1.000000

Table II: Correlation Matrix of Weather vs. Total Traffic

	Trial Number					
Degree	1 MSE	2 MSE	3 MSE	4 MSE	5 MSE	Avg. MSE
1	19078104.8	14166830.5	17233866.6	18984425.3	21365881.6	18165821.8
2	17895110.0	14847977.3	13075620.3	15915521.1	13498531.6	15046552.1
3	13679644.3	95245418.1	15111296.9	13436526.4	17636891.1	31021955.4
4	56056018.7	263743523.1	181521151.9	60085517.7	28406674.8	117962577.2

Table III: Average Mean Squared Error Values for Polynomial Fit with Varying Degrees

Question 3:

After separating the bike traffic by day and running linear regression, we were able to find a coefficient of determination for each of the days of the week. This helped show how consistent the bike traffic was on a given day of the week over the course of 7 months. Initially, we examined the graphs to determine if there seemed to be any correlation to the naked eye. From this, we determined that they did not look immediately consistent and decided that linear regression would prove a better test. Based on the generated coefficients of determination, the highest being 0.2227, we determined that the total number of bicyclists on any given day of the week was not consistent over 7 months.

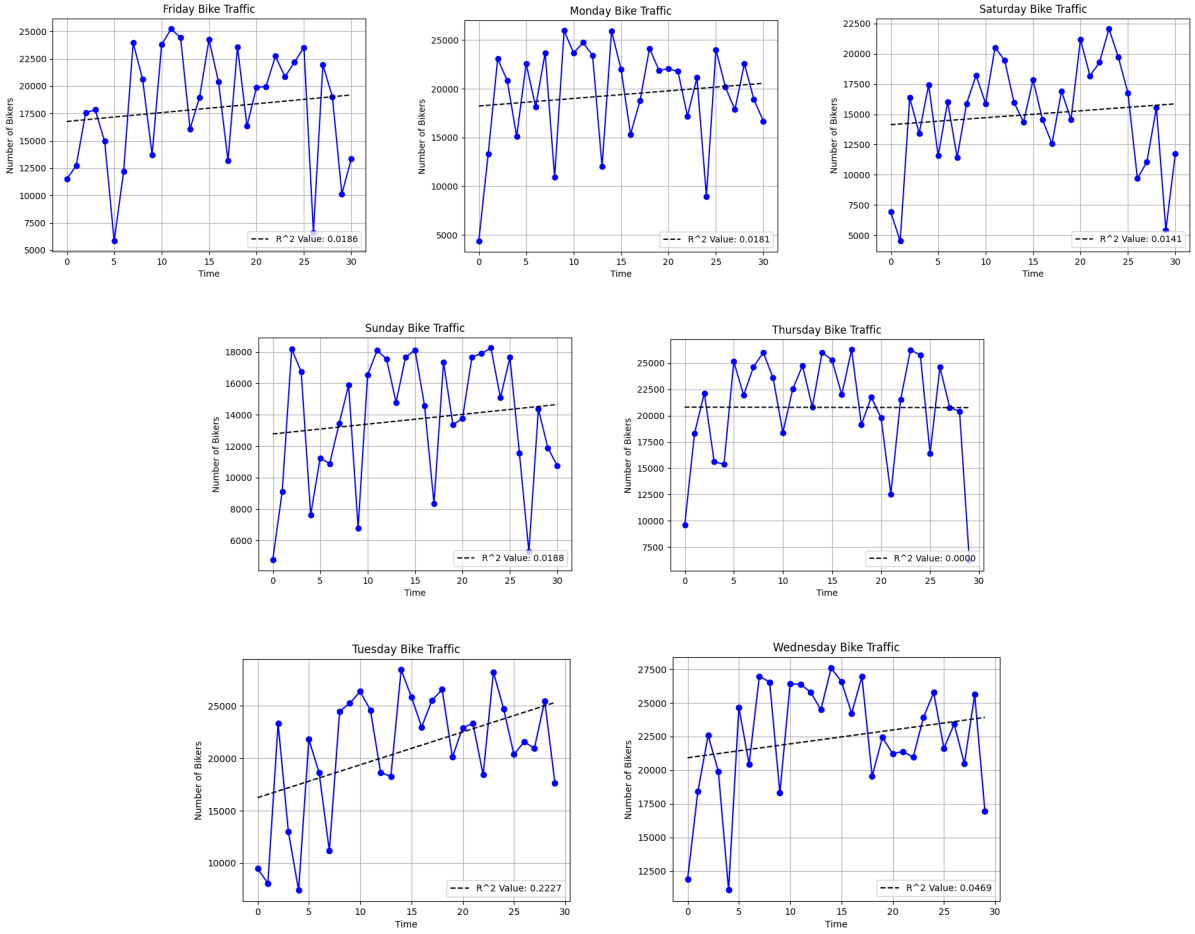

Figure I-VII: Linearly Fitted Bike Traffic (Apr-Oct) by Day

Day	Linear Regression R^2 Value
Friday	0.0186
Saturday	0.0141
Sunday	0.0188
Monday	0.0181
Tuesday	0.2227
Wednesday	0.0469
Thursday	0.0000

Table IV: Linear Regression R^2 Values for Daily Bike Traffic