Answering Analysis Questions on Bike Traffic Across a Number of Bridges in New York City

Analysis of Bike Traffic in New York City

Isha Ghodgaonkar & Sneha Mahapatra

Data Set

This report analyzes bike traffic across a number of bridges in New York City: Brooklyn, Queensboro, Williamsburg, and Manhattan Bridge. The data set is a daily recording of the number of bicycles that crossed into/out of the city of Manhattan using these four bridges from April 1st, 2016 to October 31st, 2016. There are 10 columns in the data set:

1. Date
   Info: Date of the recording
   Format: Month/Day
   Type: String

2. Day
   Info: Day of the Week.
   Format: Monday, Tuesday etc.
   Type: String

3. High Temp (°F)
   Info: Highest recorded temperature that specific day. Given in Fahrenheit.
   Type: Float

4. Low Temp (°F)
   Info: Lowest recorded temperature that specific day. Given in Fahrenheit.
   Type: Integer

5. Precipitation
   Info: Weather of the day. The T stands for trace precipitation (meaning little to no precipitation) while S stands for snow. Each float represents how many inches are present.
   Type: Float

6. Brooklyn Bridge
   Info: Number of Bicyclists on this bridge recoded that day.
   Type: Integer

7. Manhattan Bridge
   Info: Number of Bicyclists on this bridge recoded that day.
   Type: Integer

8. Williamsburg Bridge
   Info: Number of Bicyclists on this bridge recoded that day.
   Type: Integer

9. Queensboro Bridge
   Info: Number of Bicyclists on this bridge recoded that day.
   Type: Integer

10. Total
    Info: Total number of bicyclists that rode across all the bridges that day.
    Type: Integer

Analysis & Results

Question 1

Question

You want to install sensors on the bridges to estimate overall traffic across all the bridges. But you only have enough budget to install sensors on three of the four bridges. Which bridges should you install the sensors on to get the best prediction of overall traffic?

Analysis

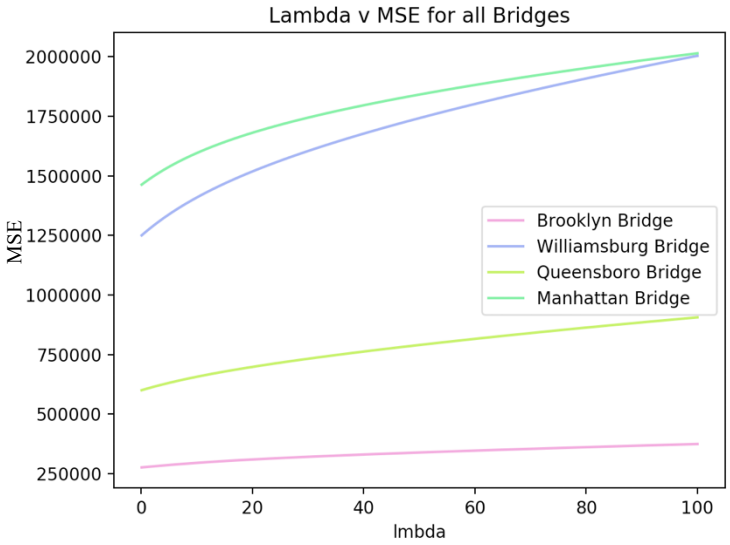
To choose the three bridge we decided to do apply two ideas. We will use Ridge Regression to find out which variable or variables impact the number of cyclists per day. Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity[1]. Since there are multiple independent variables such as Date, Day, Precipitation, High Temperature, Low Temperature, we have to see which variable or variables affects the data. In our case we only took into account three variables that could be greatly affecting the number of riders: Precipitation, High Temperature, and Low Temperature. Once we find the coefficients that greatly impact the data, we will then use them to plot each variable against the number of cyclists that was there due to that variable.

Afterwards, will use the poly fit method to see what degree of polynomial it follows. We will than determine which three bridges follow a similar pattern and exclude the one that does not. This way we can determine overall traffic by looking at the bridges that follow similar patterns instead of having one bridge that may disrupt the trend.

We will also use a test size of 10% and a train set of 90%.

Results

First we tested multiple lambda's after training the models. The graph to the right. Shows the lambda verses MSE (mean squared error). Down below, the first table shows what was the best lambda, mean squared error and r-squared value for each bridge. The second table is the calculated coefficients.



Characteristics	Brooklyn	Williamsburg	Queensboro	Williamsburg
Lambda	.1	.1	.1	.1
MSE*	276898.9222204	1250209.073406	600704.005115	1462806.60894
R^2**	.44789	.45334	.51412	.39249

*MSE has been rounded to the 7th decimal place. The whole number can be found after running file.
**R-Squared has been rounded to the ten-thousandths place. The whole number can be found after running file.

Coefficients Bridge	High Temperature F°	Low Temperature F°	Precipitation
Brooklyn	878.02866169	-237.24977689	-311.38789431
Williamsburg	1477.77064376	-602.75437551	-675.32662465
Queensboro	930.39613812	-225.07930222	-409.5696558
Manhattan	1334.74455384	-619.00505188	-593.39345853

After looking at the variables it is very clear that the higher temperature has a much bigger impact compared to Lower Temperature and Precipitation. We can directly use these values and not have to worry about High temperature > Lower Temperature because we normalized each column. This allows use to use direct comparisons.

After deciding that the single most important variable was High temperature, we plotted all the values and used poly fit to determine the degree they each follow. The equations below is the different degrees of fit equations. Below this are the graphs that show the poly fit and the scatter plot of the original data versus high temperature.

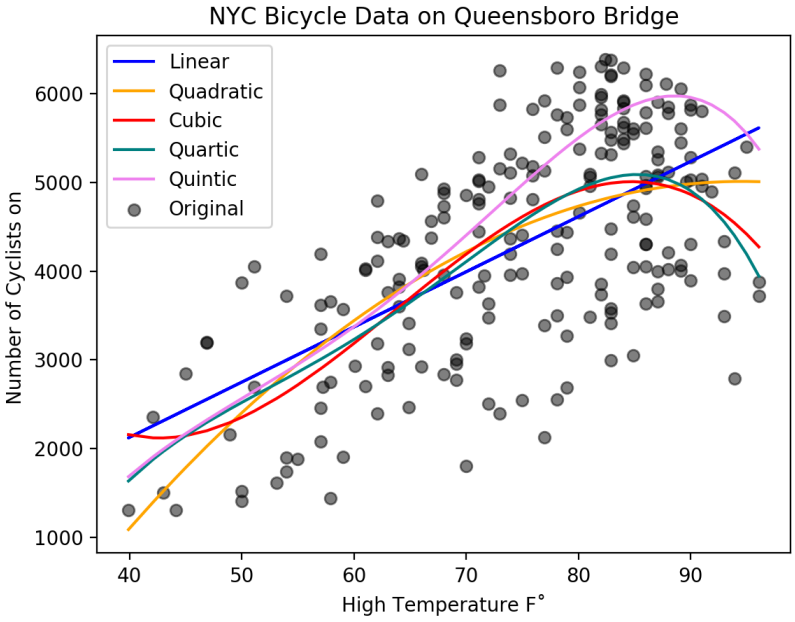
Manhattan
linear: 65.161497X + 169.445142
Quadratic: -2.721402X^2 + 451.030005X + -13037.961441
Cubic: -0.100979X^3 + 18.29858X^2 + -969.204216X + 17980.018893
Quartic: -0.001621X^4 + 0.348441X^3 + -27.457443X^2 + 1052.952591X + -14669.64821
Quintic: -6e-06X^5 + 0.000601X^4 + 0.046904X^3 + -7.372733X^2 + 397.31437X + -6288.871892

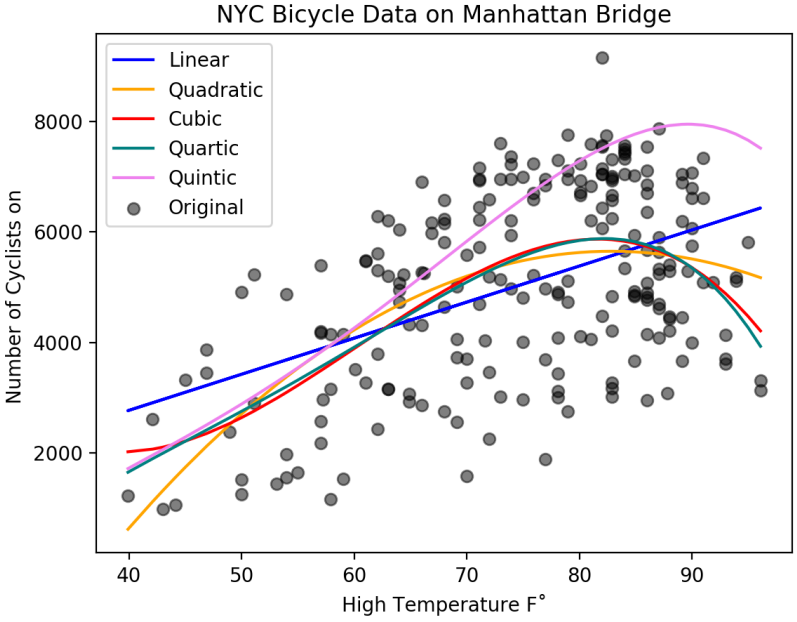
Brooklyn
linear: 53.068115X + -945.88634
Quadratic: -0.852169X^2 + 173.897389X + -5081.599208
Cubic: -0.05806X^3 + 11.233692X^2 + -642.694727X + 12752.809942
Quartic: -0.002669X^4 + 0.682164X^3 + -64.129428X^2 + 2687.927904X + -41023.296712
Quintic: -1.6e-05X^5 + 0.002954X^4 + -0.08109X^3 + -13.290854X^2 + 1028.371349X + -19809.810852

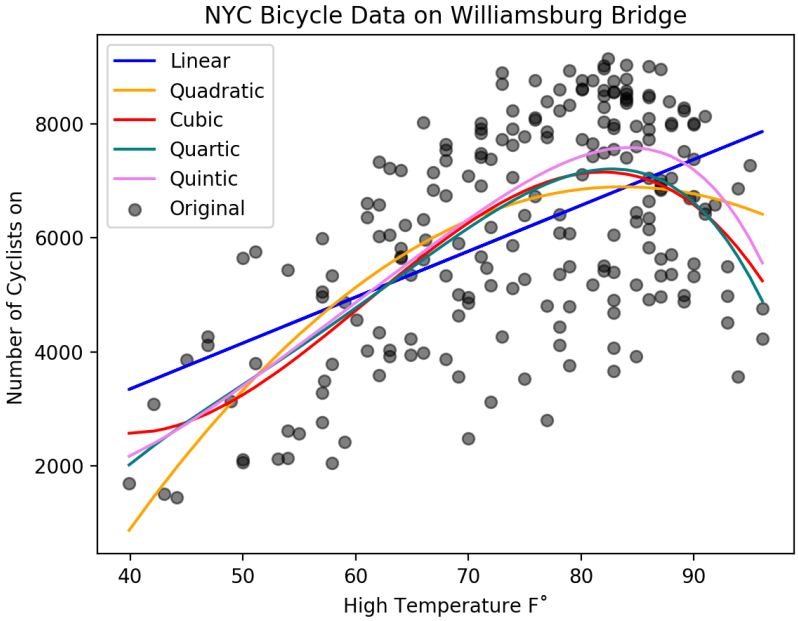
Queensboro
linear: 62.206203X + -360.613202
Quadratic: -1.313192X^2 + 248.404059X + -6733.74482
Cubic: -0.077205X^3 + 14.757987X^2 + -837.45974X + 16981.570194
Quartic: -0.002298X^4 + 0.56012X^3 + -50.128814X^2 + 2030.169024X + -29319.055067

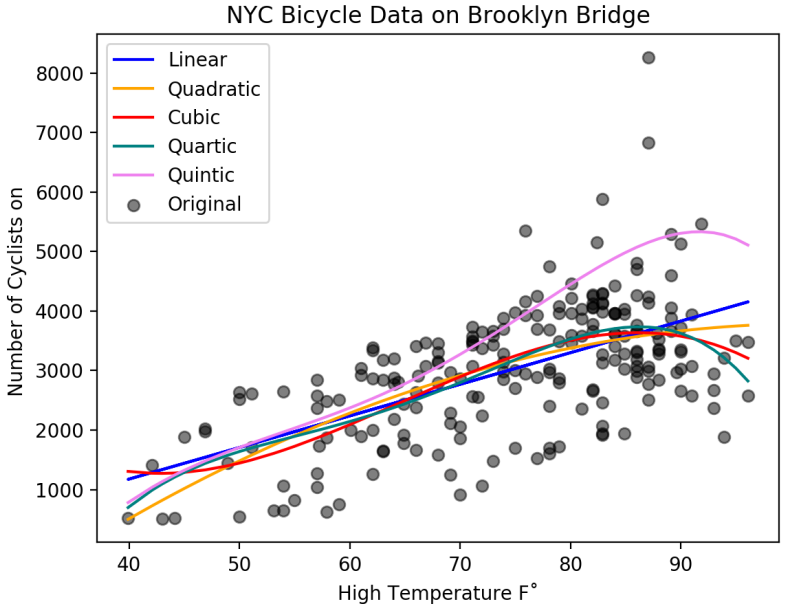
Quintic: -9e-06X^5 + 0.000871X^4 + 0.129977X^3 + -21.478034X^2 + 1094.90301X + -17363.902097

Williamsburg
linear: 80.537267X + 125.922874
Quadratic: -3.139675X^2 + 525.712816X + -15111.429629
Cubic: -0.122779X^3 + 22.418192X^2 + -1201.127716X + 22602.846528
Quartic: -0.002437X^4 + 0.553083X^3 + -46.392156X^2 + 1839.899578X + -26497.464812
Quintic: -4.2e-05X^5 + 0.012098X^4 + -1.419821X^3 + 85.018396X^2 + -2449.820257X + 28336.406751









Looking at the equations we can see that High Temperature Versus Bicyclists for Manhattan, Queensboro, and Williamsburg follow a quartic fit while Brooklyn follows a more linear trend(these equations have been highlighted, in blue, above, to clarify).This shows that Brooklyn deviates from the trend that the other bridges follow and therefore should be the bridge we exclude to put the sensor on as we want the overall trend in traffic. Since Brooklyn doesn't follow the other three bridges it will not give an overall accurate estimation of traffic.

In addition, the total ridership of Brooklyn Bridge is very low (648570) across the whole year, compared to the other bridges (Manhattan Bridge: 1081178, Queensboro Bridge: 920355, Williamsburg Bridge: 1318427). This corroborates our selection of Brooklyn Bridge as removing the bridge with the least traffic would have lesser impact on total traffic than removing a bridge with much higher traffic.

We also created many test cases to look at the predicted number of riders using the model created from our ridge regression. Looking at the r-squared value of the Brooklyn bridge it has the third lowest regression which further shows that the Brooklyn bridge would be ill-suited to use.

Test Cases	[Highest Temperature, Lowest Temperature, Precipitation range[0,1]	Predicted Number of Riders on Brooklyn	Predicted Number of Riders on Williamsburg	Predicted Number of Riders on Queensboro	Predicted Number of Riders on Manhattan
0	[-0.69999, -15, 0.4]	3561.0	7130.0	4735.0	6046.0
1	[106, 91.7, 0.4]	3901.0	7646.0	5278.0	6363.0
2	[50, 35, 0.4]	4064.0	7953.0	5443.0	6655.0
3	[61, 45, 0.4]	4027.0	7884.0	5406.0	6589.0
4	[84, 69, 0.4]	3944.0	7728.0	5323.0	6441.0
5	[83, 68, 0.4]	3947.0	7732.0	5325.0	6444.0
6	[40, 25.7, 0.4]	4117.0	8054.0	5495.0	6752.0
12	[78.1, 66, 0.01]	3920.0	7682.0	5298.0	6397.0
13	[43.0, 37.9, 0.09]	3884.0	7615.0	5261.0	6333.0
14	[63.0, 46.9, 0]	4019.0	7869.0	5398.0	6574.0

*Selected Test Cases to show how Brooklyn bridge differs due to different values for the features.

Question 2

Question

The city administration is cracking down on helmet laws, and wants to deploy police officers on days with high traffic to hand out citations. Can they use the next day's weather forecast to predict the number of bicyclists that day?

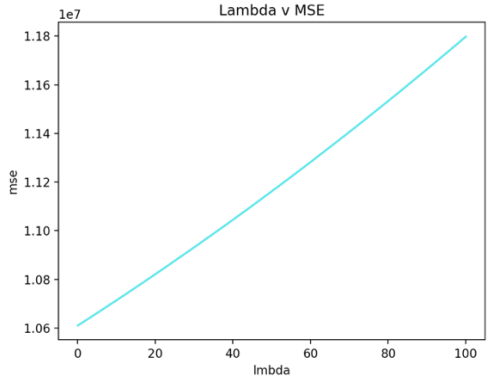
Analysis

To answer this question, we build a regression model that relates weather forecast (high temp, low temp, and precipitation) to number of bicyclists. The decision we need to make is whether to deploy police officers on that day at all, so we should consider total bicyclists rather than the bicyclists on any individual bridge.

After cleaning the dataset (converting all values to floats, translating T in precipitation to no rain, and filtering S out since there is only one instance and can be grouped together with high precipitation (rain) since its value is relatively high), we normalize the input data. We then run a Lasso regression model on the normalized data to learn the trend between the weather and total number of bicyclists.

Results

We split the data into 75% training data and 25% testing data. We try different values of lambda when running the model such that we can find the model with the lowest MSE on the test data. The graph of lambda values vs. MSE looks like this:



Our final model uses a lambda of 0.1, which yields an MSE of 10610896.1228186 and an R^2 of 0.632. The model's equation is Total Riders = 4639.858* High Temp -1702.073* Low temp - 1989.069 * Precipitation.

We can conclude that using the next day's weather forecast to predict the number of bicyclists is feasible. We have a model that reasonably correlates (with an R^2 of ~0.6) weather conditions to

total number of bicyclists, but more data would improve the quality of our model. The above results were obtained using a train-test split of 90-10.

We did attempt other train-test splits (70-30, 80-20), and other models (Ridge Regression). The Ridge regression model resulted in an R^2 of 0.473, which is slightly worse than Lasso Regression.

The table below shows the R^2 values given different models and different train-test splits.

Percentage data used (train, test) / Model	70, 30	80, 20	90, 10
Ridge Regression	0.423	0.449	0.473
Lasso Regression	0.542	0.596	0.632*

* The best model R^2 value out of all tested.

Question 3

Question

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

Analysis

We can approach this problem using either regression or classification. Using regression, we would build a model to predict the amount of rain given total number of bicyclists minus the ones from the irrelevant bridge (Brooklyn Bridge, as determined in the first problem), and determine the R^2 value of the model as compared to the ground truth values of precipitation. Using classification, we would discretize the precipitation values to be either Rain or No Rain and use logistic regression to predict either whether it is raining or not given a total number of bicyclists on a given day.

We tried both strategies (Regression including Lasso Regression, Ridge Regression, and Classification using Logistic Regression) to see whether any models would be good predictors of whether it is raining.

Results

The Table below shows the R^2 values of the models tested with the different training and testing splits:

Percentage data used (train, test) / Model	70, 30	80, 20	90, 10
Ridge Regression	0.1740127159429855	0.23743528091825916	*0.3015647810105274
Lasso Regression	-0.00237072728564	-0.0195017853394967	-0.00077622761002715

* The best model R^2 value out of all tested.

Neither of these models seem to be very good predictors, although with more data, Ridge regression could become a good predictor as the trend for R^2 seems to be increasing with more training data, culminating in the starred value of ~0.3.

We tried logistic regression by treating the precipitation as a binary label (rain or no rain) using the following logic: raining (1) if the precipitation level is greater than threshold, non-inclusive, else no rain (0). We ran logistic regression using three different thresholds (0, 0.2, 0.5) for generating the precipitation label. For this experiment we held the train-test split constant at 90-10 as previous experiments have shown that with this small amount of data, a train-test split of 90-10 generally yields the best models. Below are the results for each test case.

Threshold = 0

Condition	Precision	Recall
0 (no rain)	0.84	0.94
1 (rain)	0.67	0.40

Threshold = 0.2

Condition	Precision	Recall
0 (no rain)	0.9	1
1 (rain)	1	0.33

Threshold = 0.5

Condition	Precision	Recall
0 (no rain)	0.91	1
1 (rain)	0	0

From these results, we can conclude that the best model is a logistic regression model with a threshold of zero. This model was able to recall 94% of no rain instances with a high precision (84%), and 40% of raining instances with a precision of 67%. Increasing the threshold decreases model quality, as more and more instances get predicted as no rain, leading to an eventual 0% precision and 0% recall of raining instances with a threshold of 0.5. It is not possible to decrease the threshold below zero, so we can conclude that this is our best predictor.

Based on the results from treating the problem as a continuous regression problem (lasso and ridge regression results) and a classification problem (logistic regression), we can conclude that a classification model is a better model for the question and we can reasonably predict whether it is raining or not based on number of bicyclists, although the model seems too precise for cases of rain.