Name: Justin Chan
GitHub: justinachano04
ID: chan279
Path1

Introduction:

The dataset NYC_Bicycle_Counts_2016_Corrected.csv shows information about bicyclists across the 4 major bridges in New York City. In the dataset, there is data on: temperature, precipitation, # of bicyclists across the bridge on a given day, and total number of bikers across all 4 bridges on a given day. The data was collected from April 2016 to October 2016. The goal of this task is to help the city make better decisions to help better regulate bike traffic.

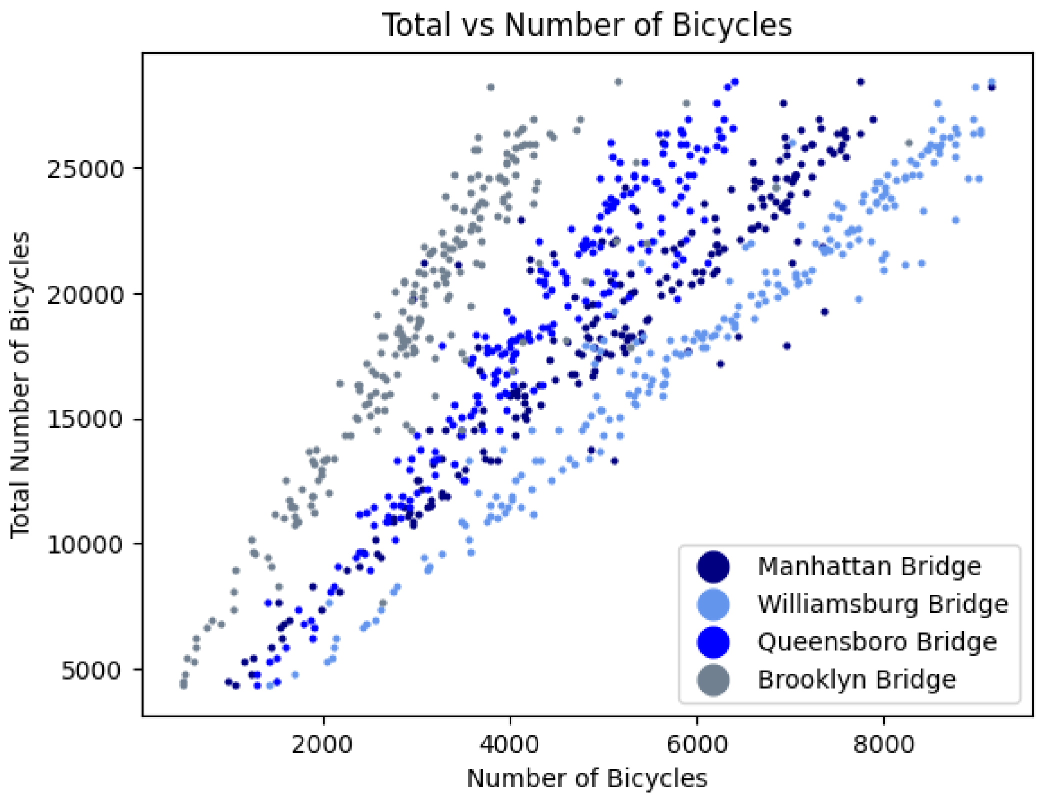
With the power of Python, we can parse and perform analysis on the given data to answer some of the key questions NYC is trying to answer in regards to bike traffic. Within Python, various libraries are used to further extend the functionalities of the scope of the project. Numpy is used to store arrays and perform matrix operations utilizing topics found in linear algebra while Pandas is used to parse through the data which simplifies the process of breaking down the large amount of data into manageable arrays.

Before conducting analysis, the data needs to be loaded and transformed to prevent any potential errors. For example, the data regarding the number of bicycles for each bridge has commas separating the digits which needs to be removed. On top of that, the precipitation data has certain elements marked with “S” or “T”. Since these are not values, it is important to set them to 0 to prevent errors down the line.

With Analysis 1, information on the bicycle data per bridge and total number of bicyclists is used to perform a polynomial regression to determine how each bridge impacts the total amount of bike traffic. Analysis 2 utilizes high temperature and the bicyclist data over the 4 bridges with a linear regression to better understand the correlation between weather and bike traffic. Analysis 3 utilizes mean squared errors and hypothesis testing to understand the correlation between precipitation and bike traffic.

Question 1

Question 1 utilizes polynomial regression to determine how much effect each bridge has on overall traffic across the 4 bridges. Utilizing the least squares regression equation, the coefficients of the polynomial regression can be determined to see how much weight each bridge has on the overall bridge traffic. Packages such as Pandas and Numpy were used to first parse through the data and then matplotlib was used to display the results. To better understand the weights, the data is first normalized to better understand weights.



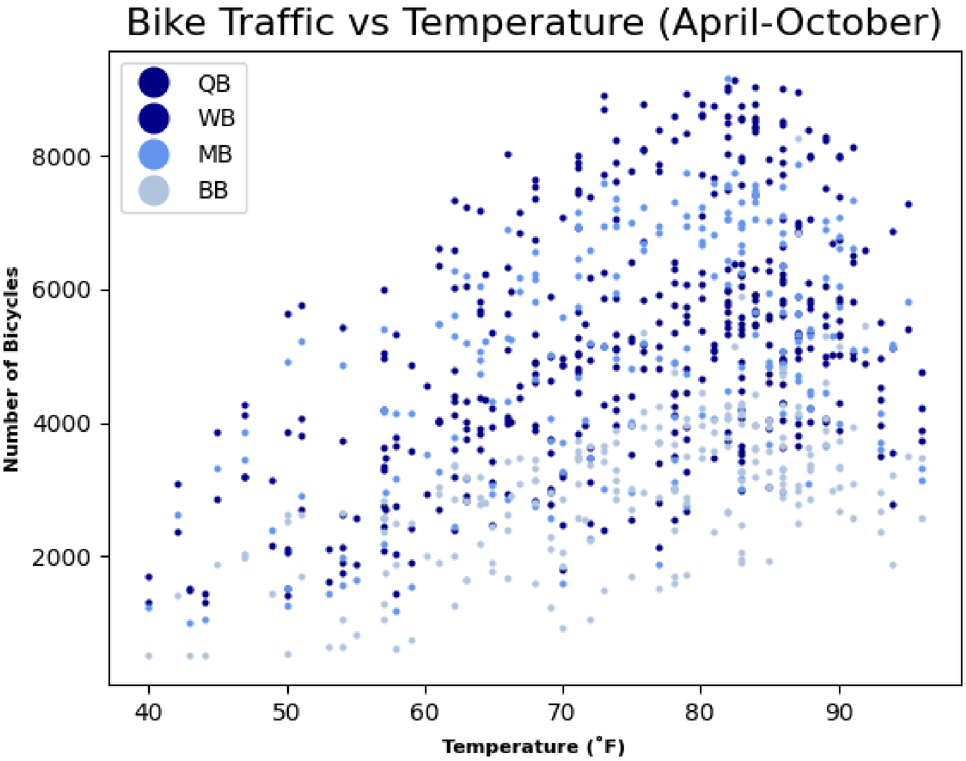
Just based on looking at this graph visually, it can be shown that the Brooklyn Bridge data points deviate from the other bridge data shown. However, further analysis must be done to better understand the data. Through calculating the weights of each bridge's numbers in reference to the total number of bicycles, the results are shown:

Williamsburg	Queensboro	Brooklyn	Manhattan Bridge
94341.036	65550.60	47323.91	78174.90

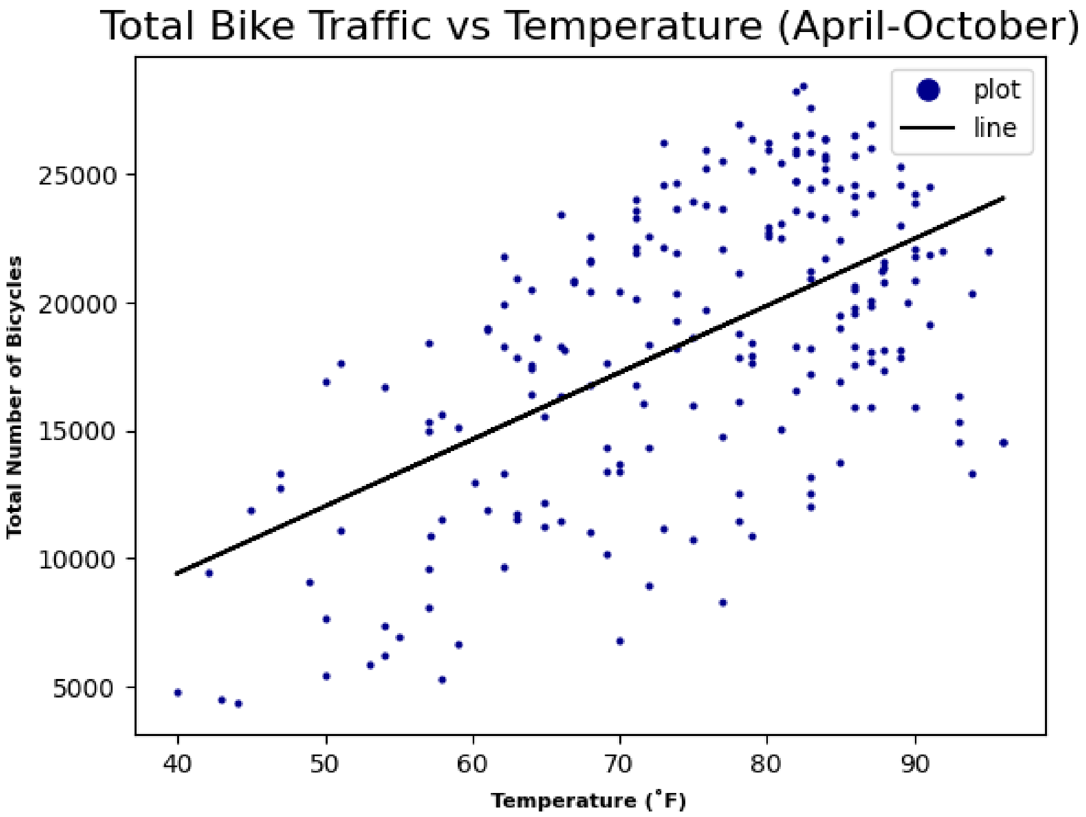
Based on the results shown, it can be further supported that the bicyclist data on Brooklyn Bridge has the least impact on the overall traffic across the bridges. With a weight of 47323.91, the Brooklyn Bridge has the lowest weight compared to the other bridges. Knowing this information allows the city to make the decision on installing the sensors on the Williamsburg, Queensboro, and Manhattan Bridge to get the best estimation for overall traffic.

Question 2

To best answer the question, a linear regression was taken to understand the correlation between high temperatures and bike traffic to make conclusions on the weather forecast in predicting bike traffic. The graph below shows the number of bicyclists across each bridge as a function of temperature.



With New York City being located on the Coast Northeast, the temperature has implications on the precipitation during a given day. In general, higher temperatures leads to less precipitation while lower temperatures are associated with precipitation. The reason this assumption can be made is because rain usually falls from a cooler body of air above having a cooling effect when hitting the ground. Additionally, New York City is not considered a tropical region where rain is more prevalent during warmer days therefore temperature is used for analysis.

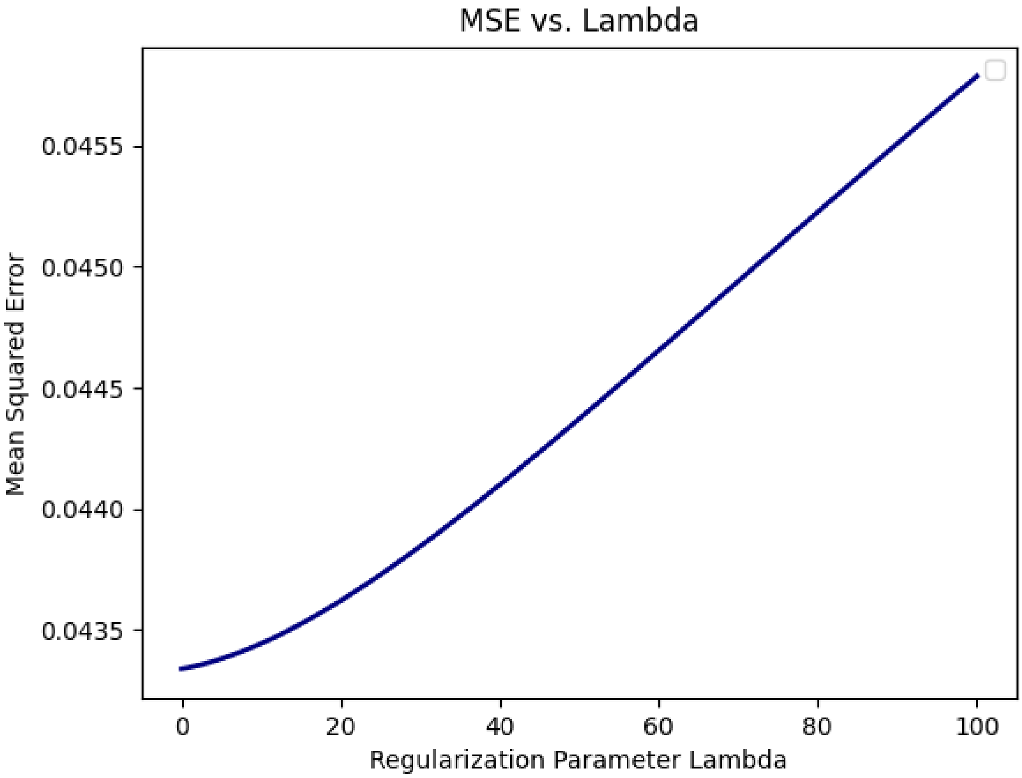


$$Best\ Fit\ Line\ Equation\colon\ y\ =\ 260,97x\ -\ 1011.131$$

Through analysis on bike traffic and temperature, a linear regression is shown above. However, the best fit line by itself is not enough to deduce correlation. Using concepts from the regression lecture, calculating the coefficient of determination is a great way to quantify the correlation. With the analysis a R^2 value of 0.57 is found. Understanding the bounds and meaning of R^2 is needed to determine the meaning of 0.57. A R^2 value of 0 means there is no correlation between the 2 variables while a R^2 value of 1 means there is direct correlation between the 2 variables. With a value of 0.57, it is safe to say there is a moderately positive correlation between the total number of bicyclists and the temperature. With higher temperatures, one for the most part says there are going to be more bicyclists outside. The city can deploy officers on the days with higher temperatures to hand out citations. However, the city should also explore further to see if there are better ways to measure high traffic to maximize the number of citations handed out for cyclists that aren’t wearing helmets.

Question 3

For this question, the question that needs to be answered is if there is a correlation between the number of bicyclists on the bridges and whether there is rain or not. In this case, a ridge regression alongside a hypothesis test is used to understand correlation. A hypothesis test is suitable in this case between the question is a yes or no question not requiring a quantified answer allowing us to use statistical evidence to support the claims. In this case, 75% of the data was allocated to training the dataset while 25% is used for testing the model. The model is trained using ridge regression over a range of lambda values which is used to determine the mean squared error.



Finding the lowest MSE is used in hypothesis testing. It allows us to do hypothesis testing by observing the lowest error which should be 0

Null Hypothesis: x = 0 meaning there is no correlation between total number of bicyclists and precipitation.
Alternative Hypothesis: x != 0 meaning there is correlation between the total number of bicyclists and precipitation.

Standard Error = 2.62e-06
Z-Score = -158908.06
P-Value = 0.99

Based on the calculations and implementation, the p-value is 0.99. At various significance levels of 0.01, 0.05, 0.1, the results are INSIGNIFICANT which means we cannot reject the **Null Hypothesis** in favor of the **Alternative Hypothesis** meaning precipitation does not depend on bike traffic. Through the process, the lecture on Hypothesis testing as well as the lecture on Regression were helpful in determining the various stats to draw the right conclusion.

$$SE_{a_m} = \frac{\sqrt{\frac{\sum_{n=1}^{N} (y_n - \hat{y}_n)^2}{N-2}}}{\sqrt{\sum_{n=1}^{N} (x_{n,m} - \bar{x}_m)^2}}$$

Equation 2: Standard Error for Regression