An Analysis on the Data of New York City Bicycle Count from 2016

(Path 1)

By: Patrick Li (li3299) and John Gentner (jgentne)

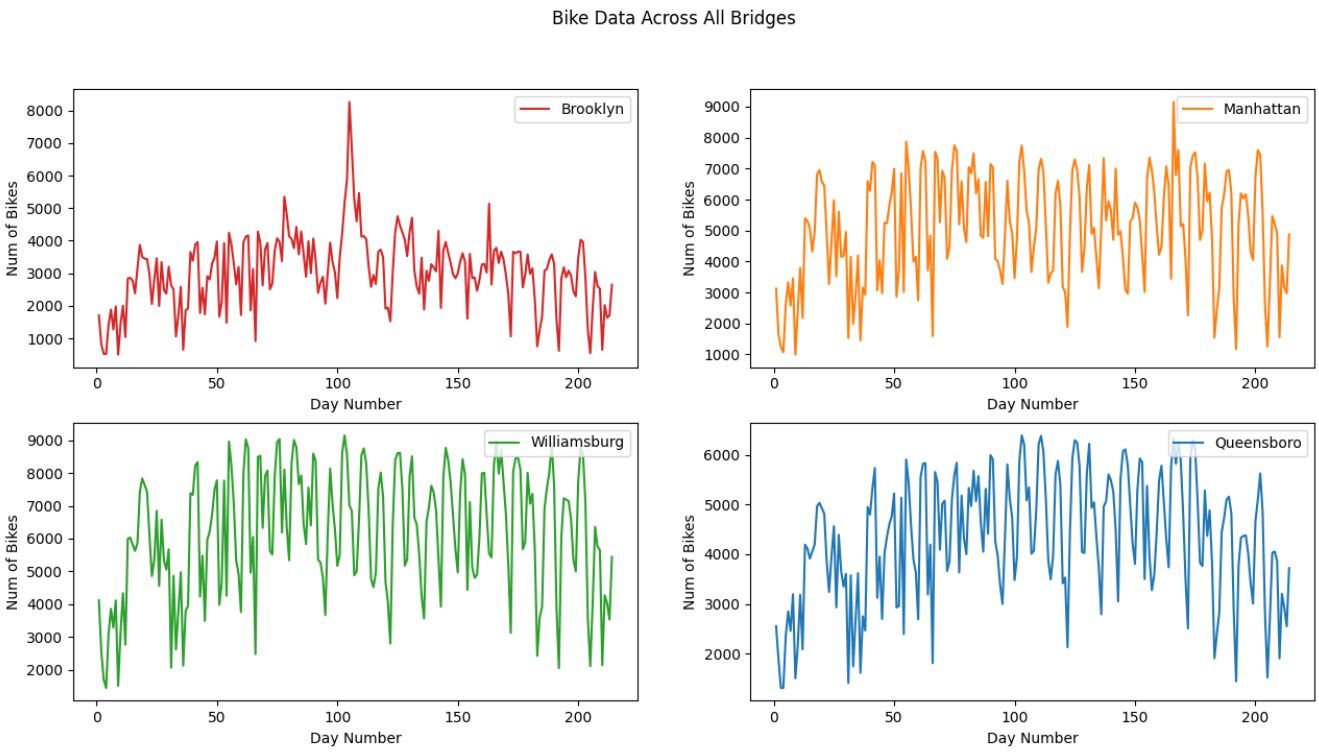
Description

This data set represents the bicycle information for New York City from April through October, 2016. The type of information that it provides are the high temperature, low temperature, the precipitation, and the bicycle counts that crossed brooklyn bridge, manhattan bridge, williamsburg bridge, and Queensboro Bridge as well as the total bicycle count for all of the bridges. We will use statistical tools and make conclusions based on our outputs and ensure our data will suffice the problems requested that must be attended to. In this path we chose, we used several modules to assist in our statistical calculations.

Analysis

Question 1

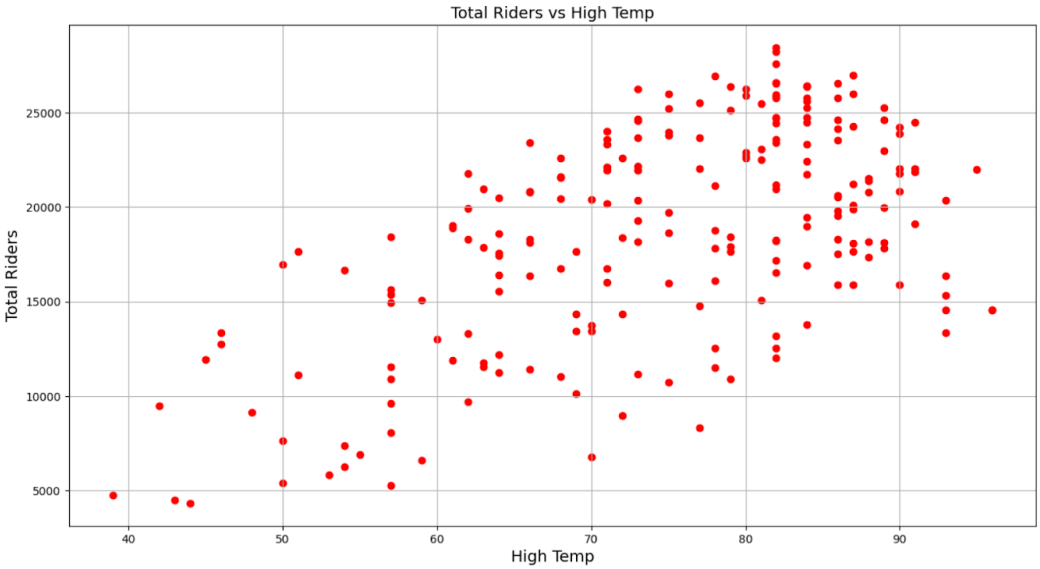
For the first problem, we were tasked with installing sensors across four bridges located in New York City. However, we were limited in that we could only place three sensors on the four bridges due to a limited budget. In order to determine which of the three bridges to put the sensors on, we did an analysis. We conducted a very simple analysis which was to create four different graphs showing the number of bikes used for a specific day at a specific bridge (Brooklyn, Manhattan,Williamsburg, and Queensboro). Here, we took the amount of bikes each day by each bridge and plotted our results. Here, we notice Manhattan, Queensboro, and Williamsburg are all very similar and all 3 tend to contain higher biker use. With this we can conclude that those will be the three bridges we place the sensors. The Brooklyn bridge we notice has a very high peak around 100 - 110 days into the year. The rest of the data happens to be the lowest of the four, year round.

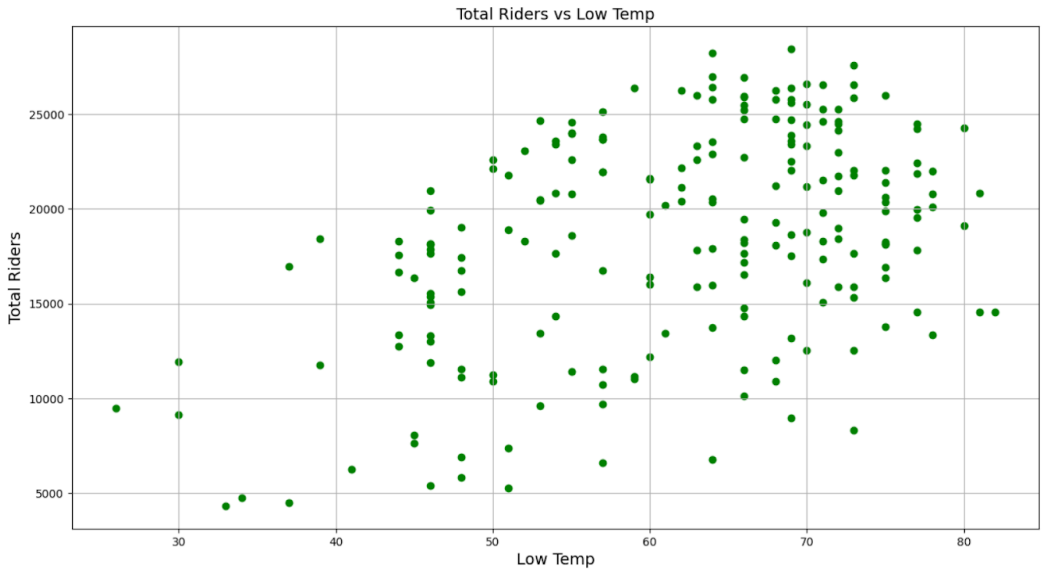



```
Total number of bicyclists for Brooklyn Bridge:  648570
Total number of bicyclists for manhattan Bridge:  1081178
Total number of bicyclists for Williamsburg Bridge:  1318427
Total number of bicyclists for Queensboro Bridge:  920355
```

Question 2

For the second problem, we were assigned the task of predicting the number of bicyclists using next day's weather forecast. In order to determine the number of bicycles for each day, we chose to use linear regression in order to see the correlation between the total number of bikers with high temps and total number of bikers and low temps. In this step, we will create a hypothesis test for the testing variables. Our null hypothesis will be that there is no correlation between bikers in New York and temperature in any given day. Our alternative hypothesis will be there is a correlation between the proportion of bikers in New York and temperature in any given day.





```
==============================================================
                 coef    std err          t      P>|t|
--------------------------------------------------------------
const      -1514.3090   1880.387     -0.805      0.422
High Temp    479.9600     61.736      7.774      0.000
Low Temp    -255.8632     66.064     -3.873      0.000
==============================================================
R-squared:                       0.374
Intercept:
 -1514.3089827597614
Coefficients:
 [ 479.96000673 -255.8632283 ]
```

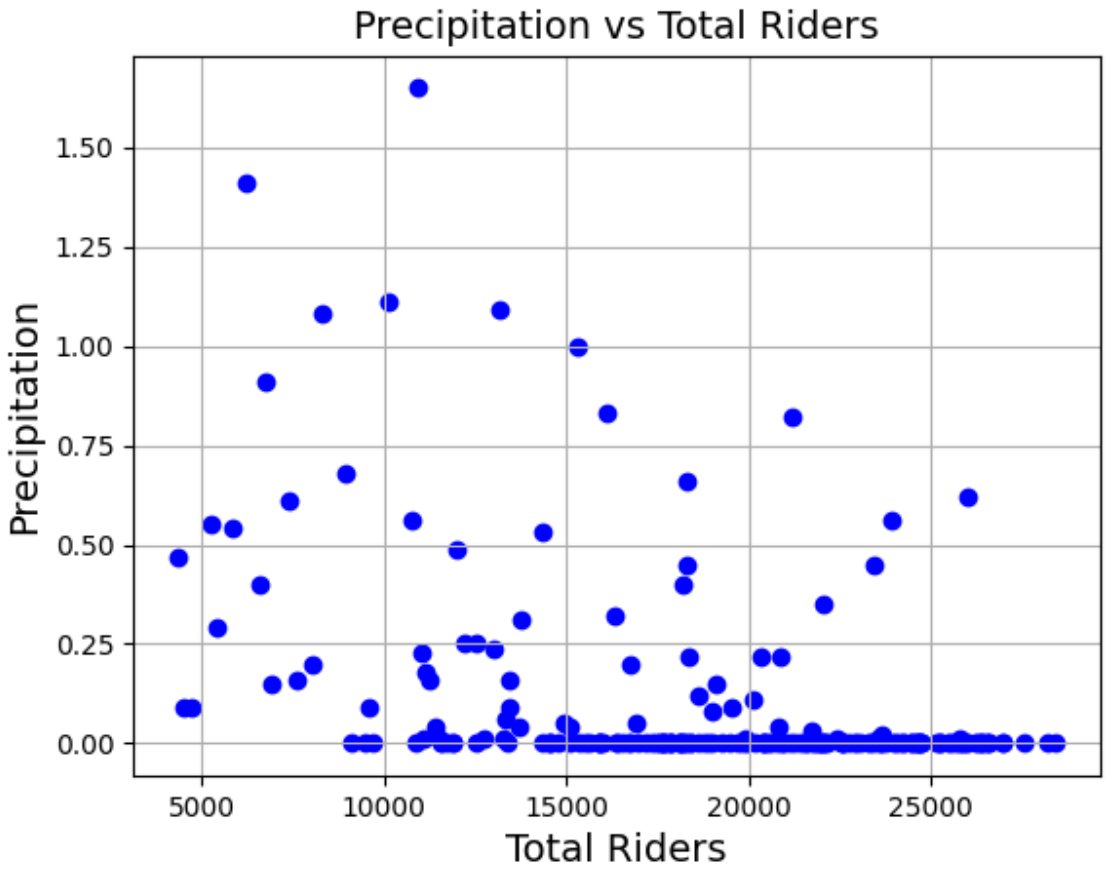
As we can see from the calculated data, the program outputs a p-value that is extremely large. 0.422 to be exact. By stark contrast, the accepted p-value in statistics is 0.05 in order to reject the null hypothesis. In this case, we fail to reject the null hypothesis. We have no evidence to suggest that a correlation between the bikers in New York and the temperature exists on any given day . The calculated equation for high temperature and low temperature is Y = 479.96X1 - 255.86X2 - 1514.31 where X1 is for the high temperature and X2 is for the low temperature. One possible reason why there would not be a correlation between the temperatures and the total number of bicyclists would be the fact that we used a linear regression on a set of data points that varies by an extreme amount.

Question 3

For the third problem, we were assigned the task of using the data to predict whether we could determine whether it was raining or not based on how many people were traveling across the bridges. In order to determine whether it will rain or not, we had to create a regression model using the data from the total number of bikes and the precipitation. One issue we came across on the data was that some of the data contain an S to indicate that it was snowing. Another issue was that the data also contained T which meant that there was no rain that day. In order to solve these issues, we had to clean the data by changing the T to a zero and removing the S. We then started plotting the total number of bikers' vs precipitation. After this we created a linear regression model and got the following figures from the linear regression.

	coef	std err
const	1.955e+04	384.511
x1	-9228.1282	1366.665

R-squared: 0.177



As we can see from the calculated data, the program outputs a large standard error and a very small r^2 value which means that there is a very small correlation between the total number of riders and the precipitation rate. As a result, we can only conclude that there is a very small correlation between the number of total riders and the amount of precipitation. One possible reason why there might be such a small amount of correlation between these two types of data is the fact that we used linear regression to find the correlation.