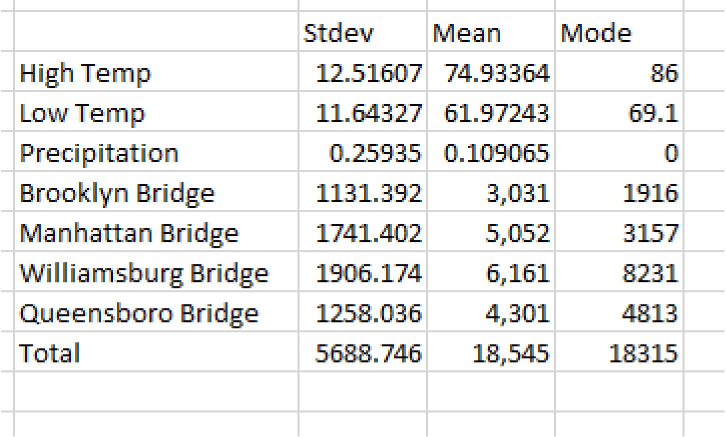
Project ECE20875: Python for Data Science Spring 2022

1. **Project team information** :
   Mini-Project Spring 2022
   ECE20875
   Devon Holloway – khdev00 – dkhollow@purdue.edu
   Path (data set) chosen: Path One Cyclist data.
2. **Descriptive Statistics** :

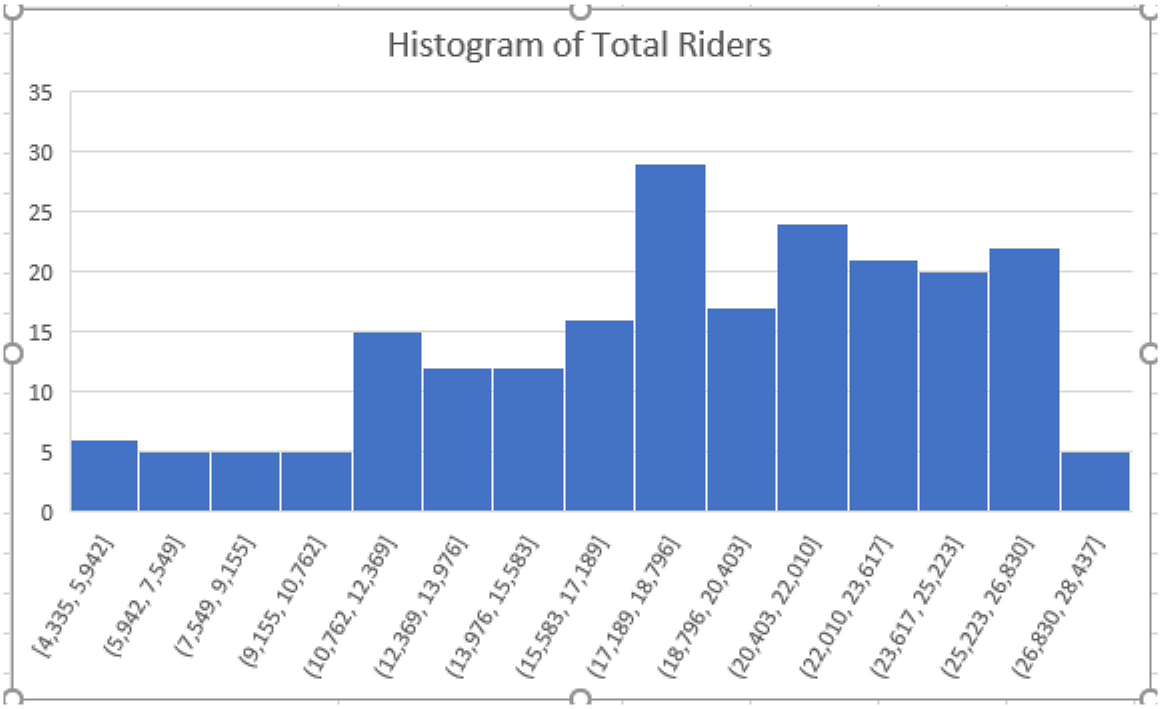
The variables I will be using are

- High Temp
- Low Temp
- Precipitation
- Brooklyn Bridge
- Manhattan Bridge
- Williamsburg Bridge
- Queensboro Bridge
- Total number of Cyclists (pretty self-explanatory)

Temp Data is just the high and low temperatures for the day, bridge data is just the number of riders on a particular bridge that day, and precipitation is the percentage of precipitation received that day. For our analysis, any value of precipitation over 0 will be changed to 1 to indicate raining, 1, or not raining, 0.



	Stdev	Mean	Mode
High Temp	12.51607	74.93364	86
Low Temp	11.64327	61.97243	69.1
Precipitation	0.25935	0.109065	0
Brooklyn Bridge	1131.392	3,031	1916
Manhattan Bridge	1741.402	5,052	3157
Williamsburg Bridge	1906.174	6,161	8231
Queensboro Bridge	1258.036	4,301	4813
Total	5688.746	18,545	18315



This is a histogram of the total amount of riders(cyclists)

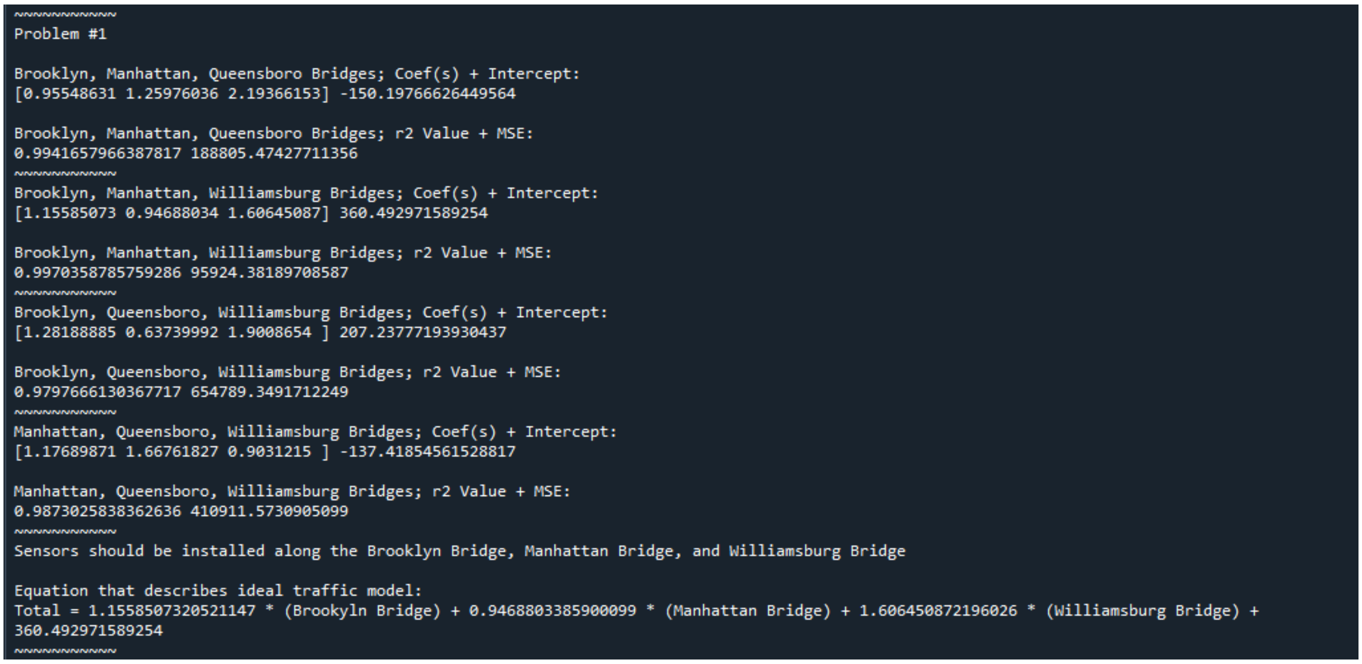
3. Approach :

I decided to do linear regression to determine which bridges the sensors should be installed on. I grouped the data into three groups of bridges, excluding one of the bridges through each group. This left me with four groups of three bridges, with each combination of bridges inside the group of three. Linear regression analysis is used to predict the value of a variable based on another variable, so I thought that would be the best approach for trying to predict the total number of riders based on data from three different bridges. This is also the same reasoning I used for the second part of the project, which is trying to use the temperature data to predict the total number of riders. I used the variables high temp, low temp, and precipitation to try to predict the total amount of riders, through linear regression.

For the last task of determining whether it is raining or not raining. I decided to do a naïve bayes approach. Naïve Bayes is most widely used as a prediction algorithm, so I thought it would be best fit for this analysis to predict whether or not it was raining. I used the variables of the Total amount of riders, and the Precipitation, where I changed all values higher than 0 to 1 for ease of analysis.

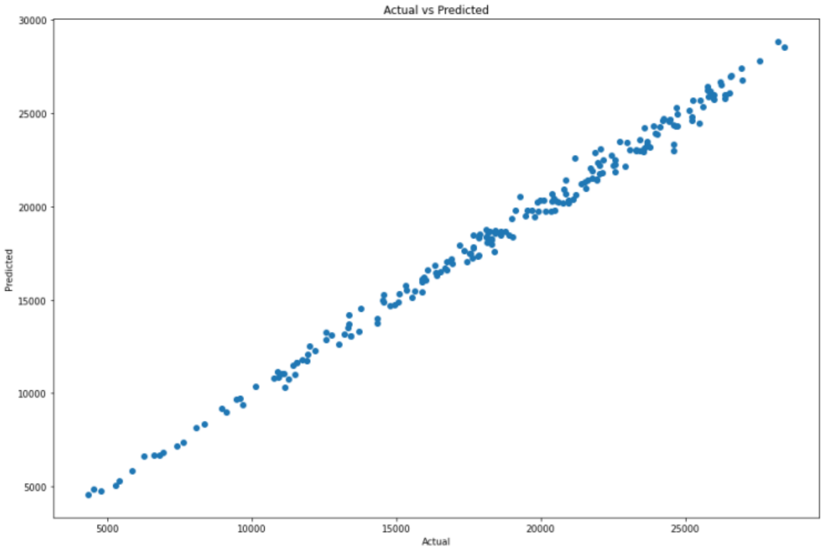
4. Analysis:

Here are the results of the linear regression model for the first problem of my project:

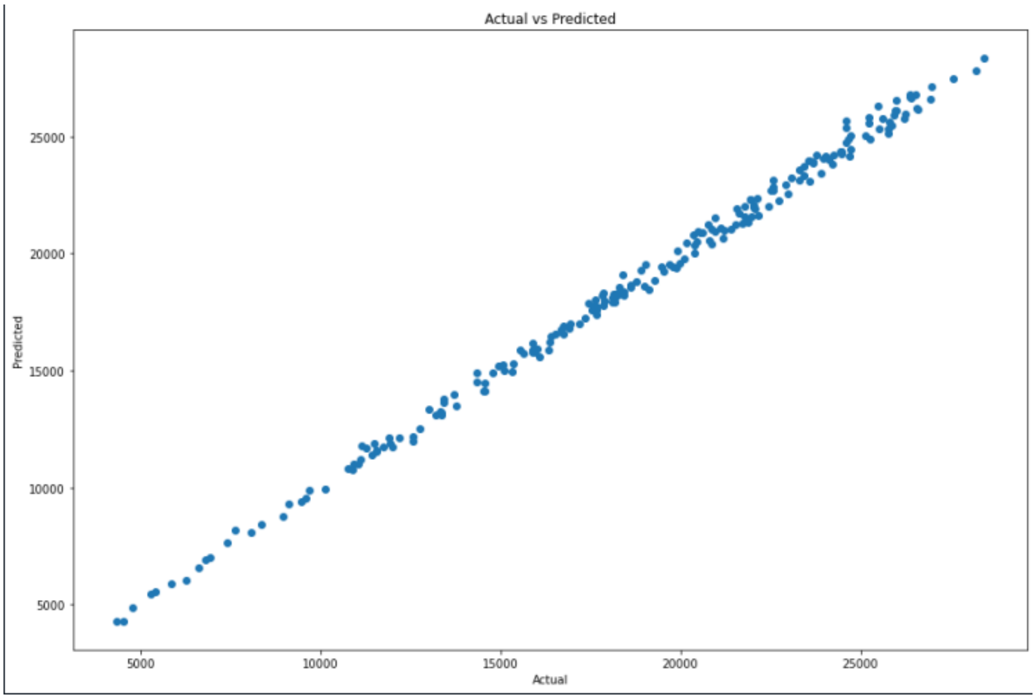


```
~~~~~~~~~~~~
Problem #1

Brooklyn, Manhattan, Queensboro Bridges; Coef(s) + Intercept:
[0.95548631 1.25976036 2.19366153] -150.19766626449564

Brooklyn, Manhattan, Queensboro Bridges; r2 Value + MSE:
0.9941657966387817 188805.47427711356
~~~~~~~~~~~~
Brooklyn, Manhattan, Williamsburg Bridges; Coef(s) + Intercept:
[1.15585073 0.94688034 1.60645087] 360.492971589254

Brooklyn, Manhattan, Williamsburg Bridges; r2 Value + MSE:
0.9970358785759286 95924.38189708587
~~~~~~~~~~~~
Brooklyn, Queensboro, Williamsburg Bridges; Coef(s) + Intercept:
[1.28188885 0.63739992 1.9008654 ] 207.23777193930437

Brooklyn, Queensboro, Williamsburg Bridges; r2 Value + MSE:
0.9797666130367717 654789.3491712249
~~~~~~~~~~~~
Manhattan, Queensboro, Williamsburg Bridges; Coef(s) + Intercept:
[1.17689871 1.66761827 0.9031215 ] -137.41854561528817

Manhattan, Queensboro, Williamsburg Bridges; r2 Value + MSE:
0.9873025838362636 410911.5730905099
~~~~~~~~~~~~
Sensors should be installed along the Brooklyn Bridge, Manhattan Bridge, and Williamsburg Bridge

Equation that describes ideal traffic model:
Total = 1.1558507320521147 * (Brookyln Bridge) + 0.9468803385900099 * (Manhattan Bridge) + 1.606450872196026 * (Williamsburg Bridge) +
360.492971589254
~~~~~~~~~~~~
```

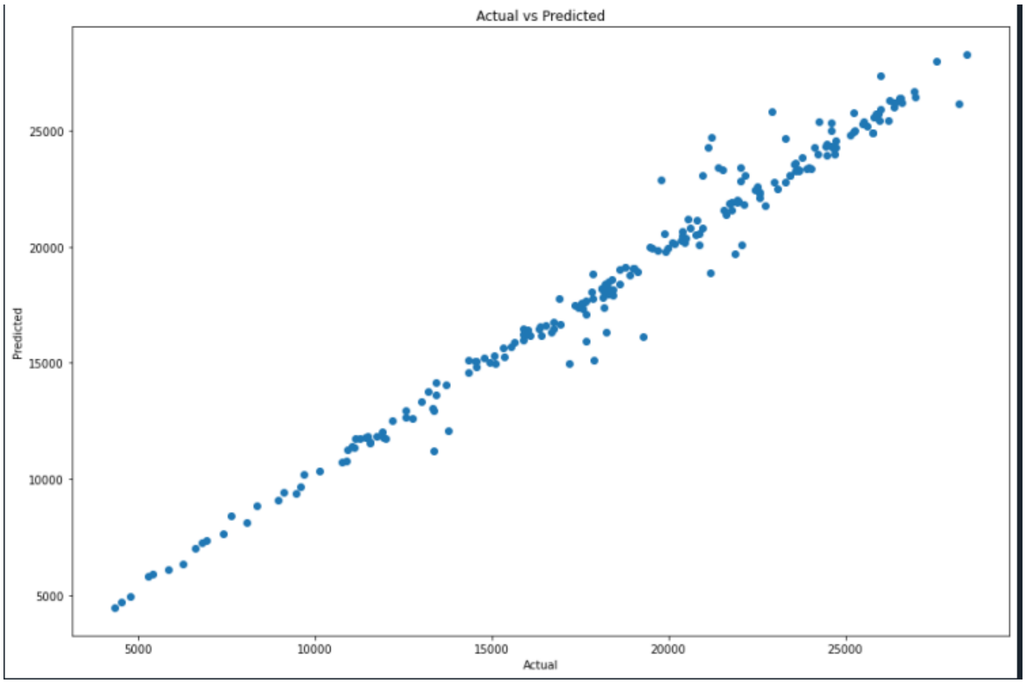
Coefficients for each bridge were found, along with its intercept, r2 value, and most significant error. As we can see from the analysis, the grouped bridges “Brooklyn, Manhattan, and Williamsburg” give us the highest r2 value and the smallest most significant error**. Sensors should then be installed at those bridges to give us the highest prediction of total traffic.** From the regression models shown below, we can also see why this set of bridges are the right ones to install our sensors on, as our test data deviates almost never from our actual data with this set.



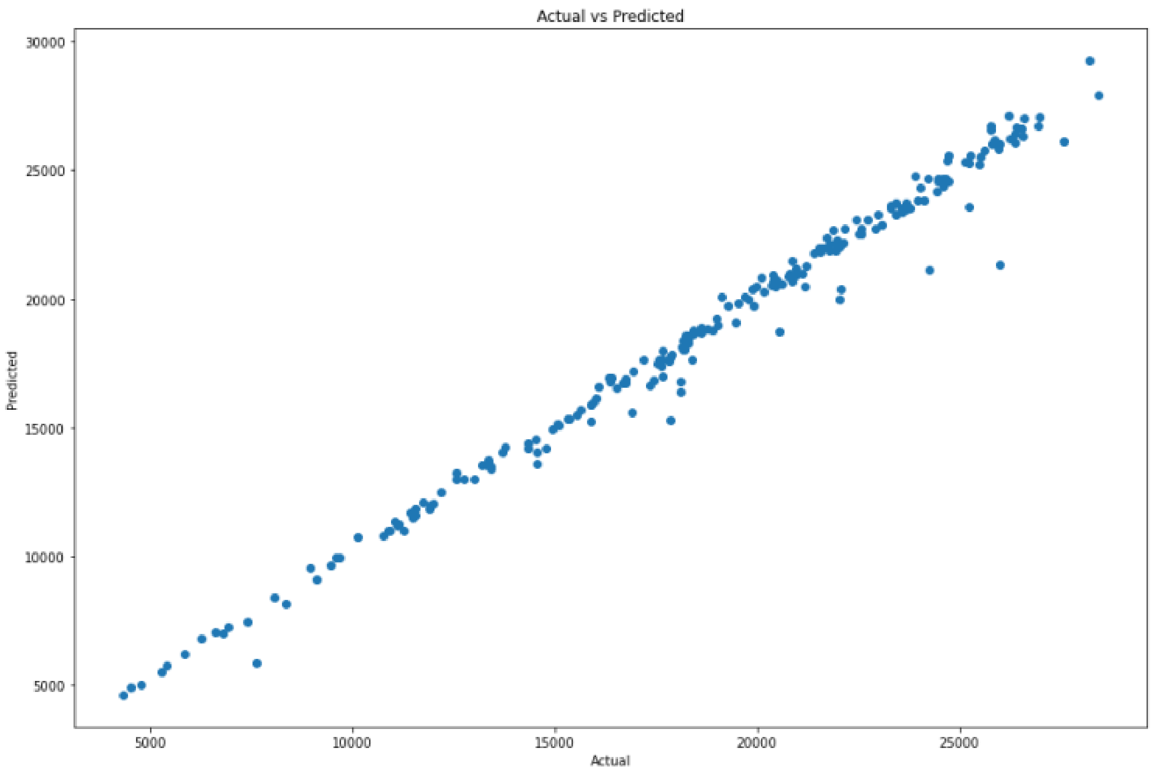
Actual vs Predicted – Brooklyn, Manhattan, Queensboro



Actual vs Predicted – Brooklyn, Manhattan, Williamsburg

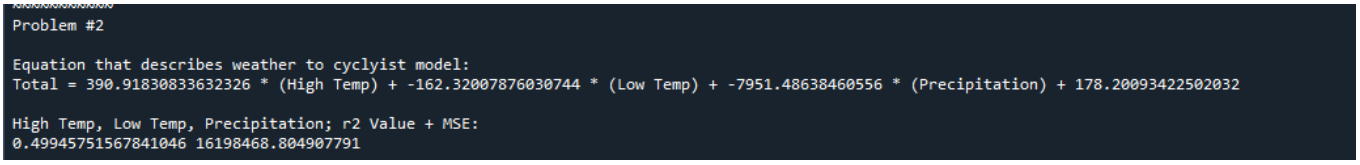


Actual vs Predicted – Brooklyn, Queensboro, Williamsburg



Actual vs Predicted – Manhattan, Queensboro, Williamsburg

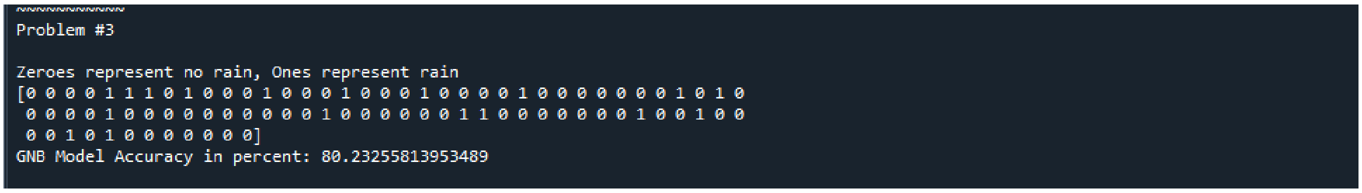
Here are the results from part 2 (Using temp data to predict total amount of riders). The same process was used above, but with only one set of data, the high temp, low temp, and precipitation along with the total number of riders. I did not need to make different groups of data as we are only concerned about a relationship between these input variables and our output variable, total amount of riders.



```
Problem #2

Equation that describes weather to cyclyist model:
Total = 390.91830833632326 * (High Temp) + -162.32007876030744 * (Low Temp) + -7951.48638460556 * (Precipitation) + 178.20093422502032

High Temp, Low Temp, Precipitation; r2 Value + MSE:
0.49945751567841046 16198468.804907791
```

Oh wow! This did not give us the most accurate results. As we can see with a r2 value of 0.499…. and our high MSE of 16198468.80, using the temperature and precipitation does not get us the most accurate prediction of the number of total riders on the bridge. This could be due to any number of things, maybe we need more temperature and precipitation data over more days to find a more accurate equation to describe the total amount of riders, or maybe temperature data would serve better to determine the amount of Joggers on the road as many people do not only recreationally bike, but also use it as a mode of transportation where you may not be able to use worry about how nice its going to be outside that day. Regardless, using the temperature and precipitation data to predict the number of cyclists out that day is not a reliable model.

Here are the results from part 3 (Using total amount of riders and precipitation data to predict if it is raining or not)



```
~~~~~~~~~~~
Problem #3

Zeroes represent no rain, Ones represent rain
[0 0 0 0 1 1 1 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 1 0
 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 1 0 0
 0 0 1 0 1 0 0 0 0 0 0 0]
GNB Model Accuracy in percent: 80.23255813953489
```

Much better! **As we can see from our Naïve Bayes approach, we were able to successfully predict whether or not it was raining! Our model accuracy is an 80 percent, so this is more than enough to use this method to try to predict rainfall.** I would expect accuracy to go up with more data though, easily.