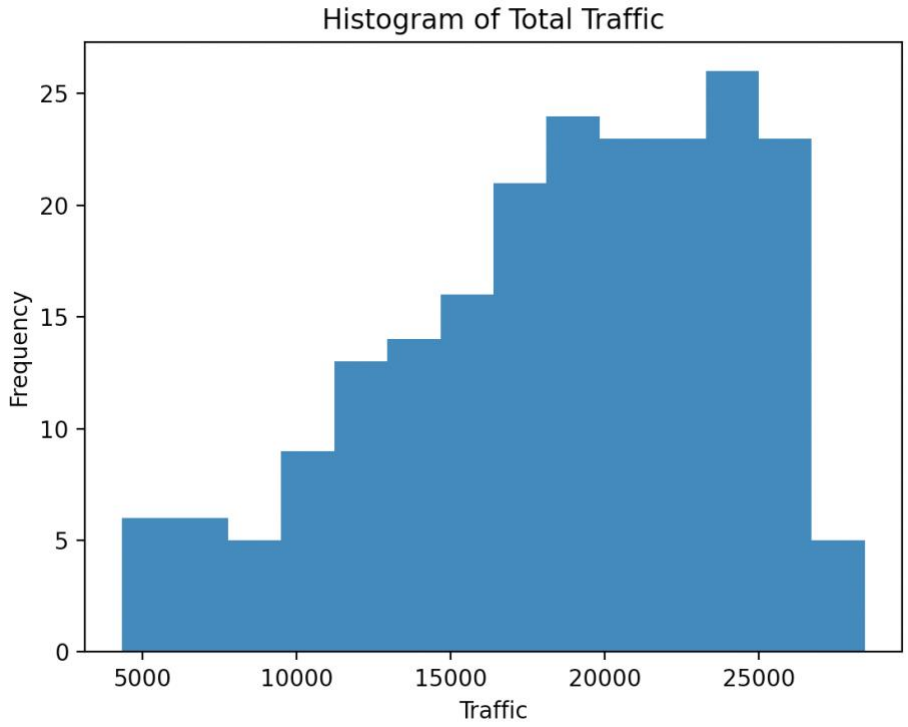
Name: Alex Yan
Purdue Username: ajyan
Github: PebbleBro

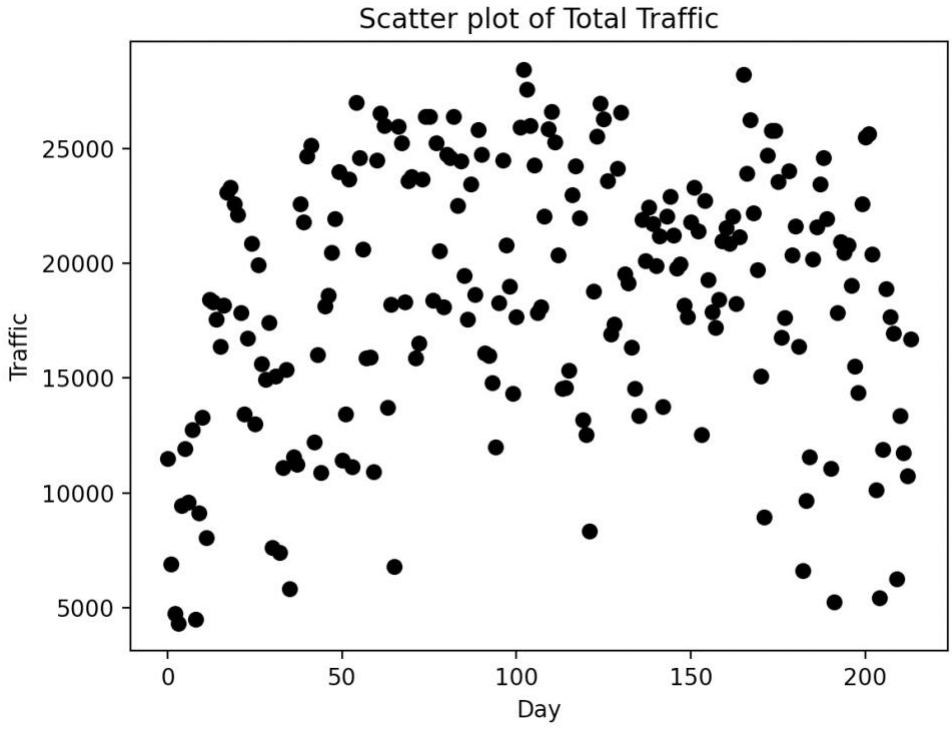
Descriptive Statistics:

The .csv file contains information on bike traffic across 4 bridges in New York City (Manhattan, Williamsburg, Queesboro, Brooklyn), such as high temperature, low temperature, precipitation, and the amount of traffic across each bridge and all bridges.

Table: Summary Statistics

Data	Mean	Standard Deviation
High Temperature (°F)	74.93	12.5
Low Temperature (°F)	61.97	11.6
Precipitation (mm)	0.11	0.26
Brooklyn Traffic	3031	1131
Manhattan Traffic	5052	1741
Williamsburg Traffic	6161	1906
Queensboro Traffic	4301	1258
Total Traffic	18545	5689


Histogram of Total Traffic
Frequency
Traffic
0
5
10
15
20
25
5000
10000
15000
20000
25000


Scatter plot of Total Traffic
Traffic
Day
5000
10000
15000
20000
25000
0
50
100
150
200

This project aims to answer three questions:

1. 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?
2. 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?
3. 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?

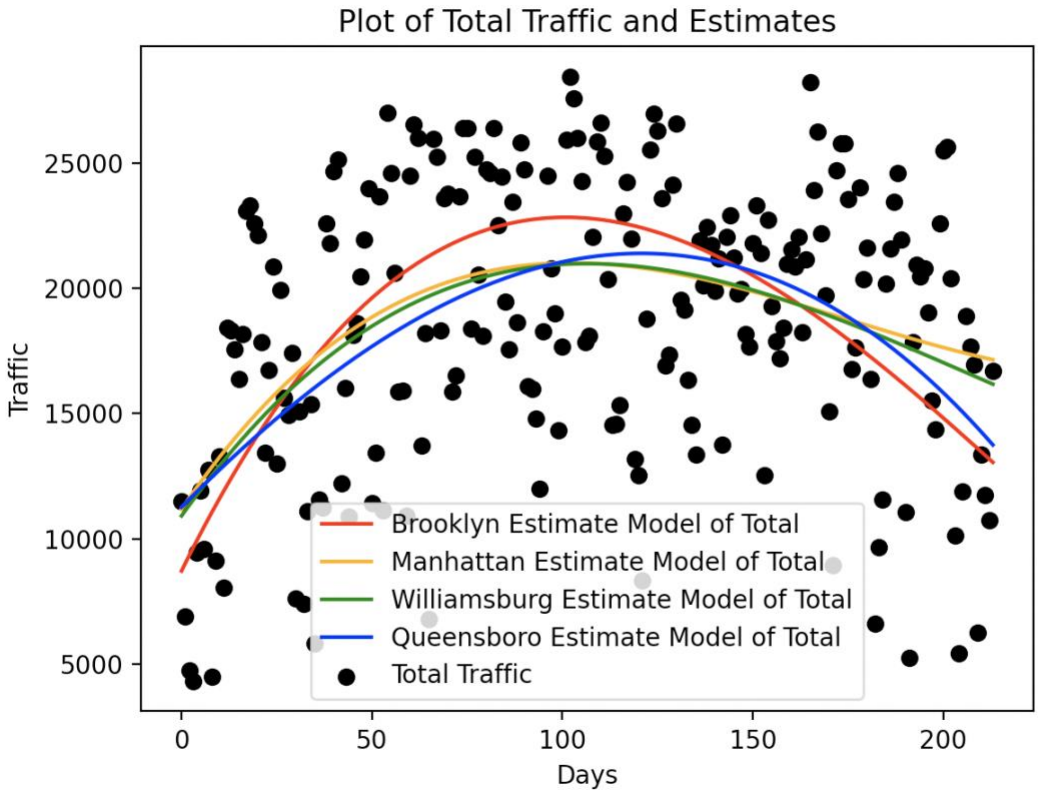
Approach:

For problem one, regression models were created from each bridge and compared to the actual total traffic data. A r-squared value was computed for each model and the lowest r-squared model would correspond to the bridge that provided the least information on the traffic across all bridges and would be chosen as the bridge to not install sensors on. The reason for choosing regression models was based on the scatter plots of the traffic across each bridge and the scatter plot of the total traffic, it could be observed that each scatter plot followed the same approximate shape.

For problem two, regression models were used again for the same reason. The scatter plots of the temperature and precipitation data showed a similar shape and trend as the total traffic. As a result, regression models would be calculated to predict the total traffic using weather.

For problem three, the Gaussian Naive Bayes model is used because the problem is a classification problem.

Problem 1:



Bridge Model	**R-squared Value**
Brooklyn	0.1742
Manhattan	0.2104
Williamsburg	0.2162
Queensboro	0.2081

Judging by the data, it can be observed that Brooklyn Bridge provides the least information about total traffic because it has the lowest r-squared value out of the four models. Even though the r-squared values of every model is very low, this is due to the total traffic being estimated from a single bridge. Once the sensors are installed on three bridges, the model used

to predict the total traffic will be much better because it is based on the traffic across three bridges. The purpose of this problem was to eliminate the least accurate bridge, which was Brooklyn Bridge. For the best prediction of overall traffic, sensors should be installed on the Manhattan, Williamsburg, and Queensboro bridges.

Problem 2:

R-squared value of Regression Model	0.575

Based on the r-squared value of the generated regression model, it can be argued that the city administration can use the next day's weather forecast to predict the total number of bicyclists with limited accuracy. The r-squared value of 0.575 is neither low nor high, so the model is not completely accurate. This may be because the regression model is not the best way to approach this problem, but it may also be because weather is not an accurate predictor of total traffic.

Problem 3:

Table: Actual vs Predicted data on test set

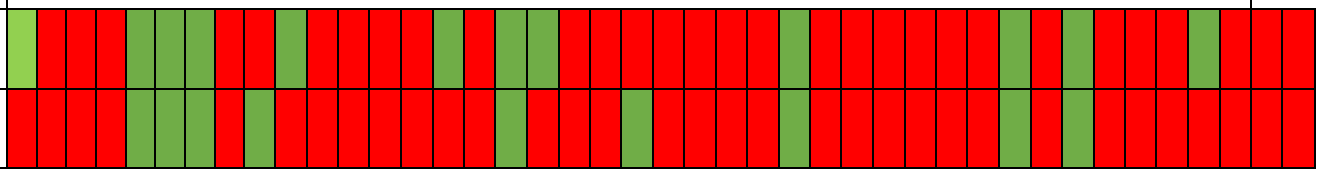
	Raining or not (Green = raining, Red = not raining)
Actual	
Prediction	

Table: Confusion Matrix

	Predicted		
Actual		Raining	Not Raining
	Raining	29	2
	Not Raining	5	7

As seen in the confusion matrix, 36 out of the 43 test data points were predicted correctly.

Based on the raining condition, it can predicted 83.7% of the time whether or not it is raining based on the amount of traffic on the bridges.