Project ECE20875: Python for Data Science Spring 2022

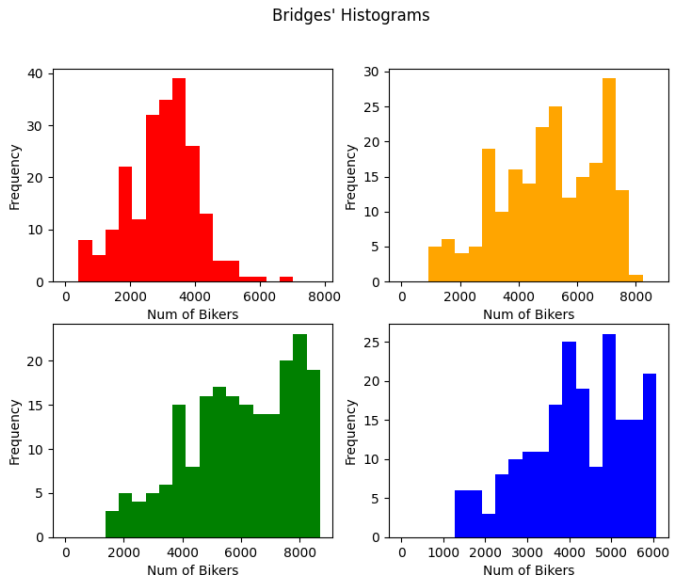
1. Project team information

Mini-Project Spring 2022
ECE20875
Mark Ma – Tequila0322 – ma747@purdue.edu
Tina Xu – Tina925 – xu1493@purdue.edu
Path (data set) chosen: Path 1

2. Descriptive Statistics

For this dataset, the variables as given are number of bicycles, which describes based on various condition the number of changes on bicycles. We also use the variable whether is rainy in order to approach to the right track. We have planned to use the four bridges data by graphing histogram and in order to determine which three of four need to add the sensor. Below is a summary statistics table of the variable we will use. We have listed out the mean, minimum, maximum, difference, and standard deviation for each bridges of the bicycles and the accuracy of the model on determining whether it is rainy based on the existing data.

Variables = Number of Bicycles						
	Mean	Min	Max	Difference	STDEV	Data within 1 STDEV
Brooklyn	3030.701	504	8264	7760	1134.045	73.3645
Manhattan	5052.234	997	9152	8155	1745.4854	58.4112
Williamsburg	6160.874	1440	9148	7700	1910.6431	62.6168
Queensboro	4300.724	1306	6392	5086	1260.9857	32.1495
Variables = Raining or not raining						
Coefficients			[-4.147153e-05, 1.086827]			
Accuracy			0.8037			
r			0.257208			



From the histogram we can see the graph for Manhattan and Williamsburg bridges looks similar and have some same pattern. So we can roughly decide to choose one from those two bridges to put the sensor.

3. Approach

For the first problem we have decided to graph the histogram and the bike pattern based on different days in order to see the pattern of each bridge in order to find which bridge should be put on sensor. According to the graph, if there are similarities between two bridges then we can just choose one out of two to put on sensor because they have similar pattern, and we can only track one of them to get the overall data.

For the second problem we have decided to use cross validation in order to split out the test and train data in order to build a linear regression model to predict. For the forecast the influential category of data would be the high temperature, low temperature, and precipitation. We used those three category and number of bicycles to construct a linear fit model which can use that model to predict the number of bicycles based on forecast.

For the third problem we have decided to create a linear regression model based on the data on number of the bicyclists on the bridges. We have noticed in the descriptive statistics that if there rainy (precipitation exists) we use 1 to represent, and if there's no rain (no precipitation) then we use 0 to represent. We need to check the accuracy of the model in order to decide whether we can use that model or not. If the accuracy is high enough then we can use the model to predict if it is rainy or not based on the number of bicycles.

4. Analysis

For the first problem the instruction is asking us to install sensors on three of the four bridges in order to save budget. We did a couple of analysis using python and graphs in order to determine which of the three bridges should be chosen. We first calculated each bridge's maximum, minimum, mean, and standard deviation. We also calculated the data within one standard deviation. Specifically, for Brooklyn is 73 percent, Manhattan is 58 percent, Williamsburg is 62.6 percent, and for Queensboro is 62.1 percent. This shows that the bridge Brookly, Williamsburg, and Queensboro's data are closer to the mean of data. According to the figure 1.1 which generated by codes it graphs out the four bridges' histograms. From the histogram we can see the graph for Manhattan and Williamsburg bridges looks similar and have some same pattern. This step we simply have our first conclusion on we should put two sensors on Brooklyn and Queensboro, and then choose one from the Manhattan and Williamsburg and we have decided to choose Manhattan which those two are very similar.

Figure 1.1

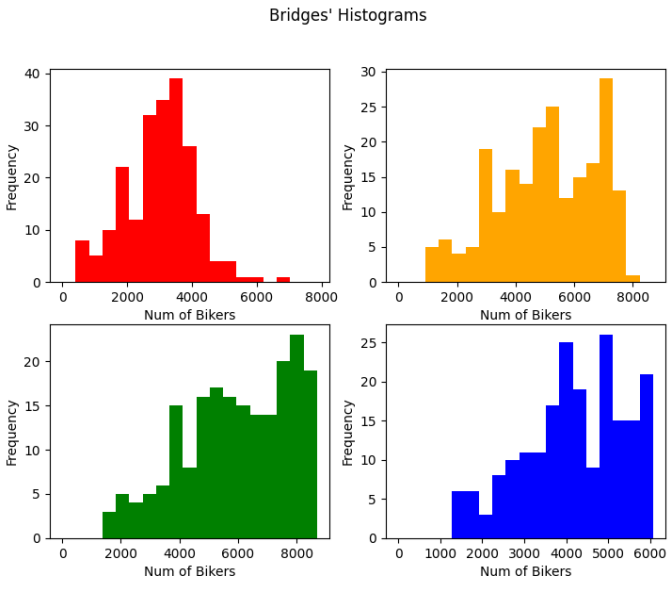
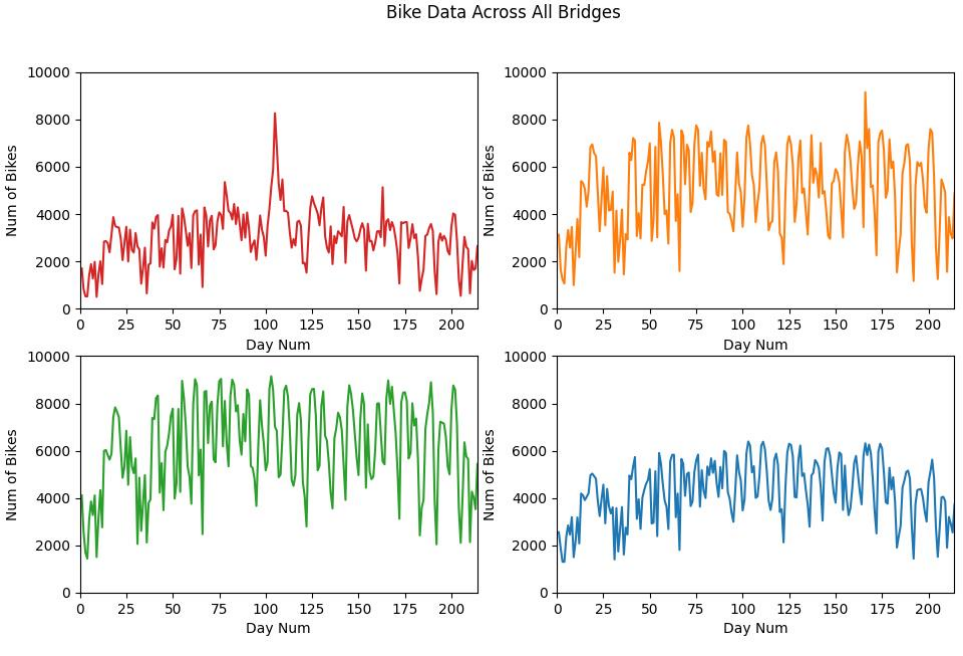


Figure 1.2



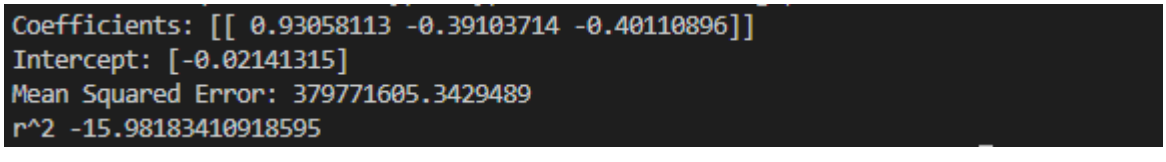
Problem 2:

This problem based on instruction is supposed to be using the next day's weather forecast in order to predict the total number of bicyclists on specific day. We used the data to predict a close estimate of bicycles on the bridges. We used the high temperature, low temperature, and precipitation

as three factor to determine the total number of bicycles. We used the crossed validation and split the data into x train, x test, y train, y test data. Then we normalize the data based on the given sets and then get the normalized data in order to build the linear model.

We used the sklearn python technics to determine the coefficients of the model based on the trained data sets. Which the total number of bikes is 0.93059x1 – 0.391x2 – 0.4011x3, and x1 is high temperature input, x2 is low temperature input, and x3 is precipitation input. However, according to this model the mean square error is huge, and I don't think it can totally represent the prediction for number of bicycles on the bridge. But we tried our best to train a linear model and is the best way we came up to predict the number of bicycles based on the next day's forecast. The outputs according to the code are displayed below in Figure 2.1. Overall, I think the data analysis for this question can only be used as a reference cannot apply as real time use.

Figure 2.1



Problem 3:

For the third problem the instruction is talking about using the number of bicyclists on the bridges in order to predict if it's a rainy day. According to calculation the accuracy of the model is 80 percent and high enough to use it as prediction. We conclude that it is possible to predict if its rainy based on the number of bicycles on the bridges. We first read the data according to the precipitation and total number of bicycles in order to see the connection between both two factors. We then construct the best fit linear model using python techniques.

We have created a linear regression model based on the data on number of the bicyclists on the bridges, and the model is y = -4.147E05 + 1.0868. According to this model when plug in the number of bicycles on the bridge we can get the result on whether it is rainy or not. For the first graph, it represents the data on how many rains fall through the number of changes on bicycles. And for the second graph, for the t value to be 0, it means there's no rain as the prediction, and for the y value be 1, it reveals that there's a rainfall on that specified day based on the number of bicycles.

Figure 3.1

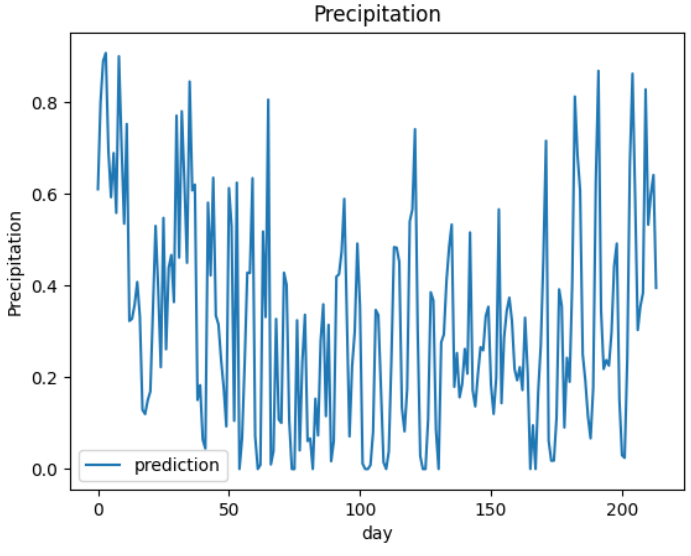


Figure 3.2