Project ECE20875: Python for Data Science Spring 2022

1. Project team information

Mini-Project Spring 2022
ECE20875
Name 1 – alexksiu – aksiu@purdue.edu
Name 2 – shah634 – shah634@purdue.edu
Path (data set) chosen: Path 1: Bike Traffic

2. Descriptive Statistics:

Descriptive Statistic Table:

	High Temp	Low Temp	Precipitation	Brooklyn Bridge	Manhattan Bridge	Williamsburg Bridge	Queensboro Bridge	Total
Mean	**74.94**	**61.92**	**0.12**	**3030.70**	**5052.23**	**6160.87**	**4300.72**	**18544.53**
Standard Deviation	**12.52**	**11.64**	**0.26**	**1131.39**	**1741.40**	**1906.17**	**1258.04**	**5688.75**
Mode	**86.0**	**69.1**	**0.0**	**1916**	**31757**	**8231**	**4813**	**18315**

Figure 1: Descriptive Statistic Table with mean, standard deviation, and mode

The table above indicates the mean, the standard deviation, and the mode for the data set given. This data is valid for April 1st, 2016 to October 31st, 2016. The descriptive statistics table contains the statistics for the high, and low temperatures, the precipitation, the bike usage for the Brooklyn, Manhattan, Williamsburg, and Queensboro bridges, along with statistics for the entire period of April to October.

The Histogram below indicates the frequency of Precipitation in inches for the given dataset:

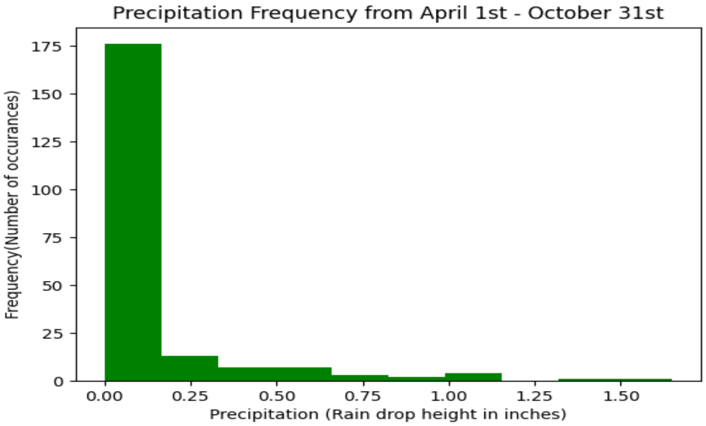


Figure 2: Frequency Histogram of Precipitation in inches between April 1st and October 31st

Problem 1:

- **Approach:**

Our approach assumes that sensors will be utilized the most on bridges with the greatest number of bicyclists. Therefore, we decided to determine on average how many bicyclists would be on each bridge within the given timeframe. The three bridges with the greatest number of cyclists on average in our opinion should be funded to have sensors. A bar chart was used to represent the average number of the bicyclists on each bridge. Our data however only has information regarding the given timeframe, one possible issue with this analysis is that it is a prediction that the average and due to unforeseen factors, it may become invalid.

- **Analysis:**

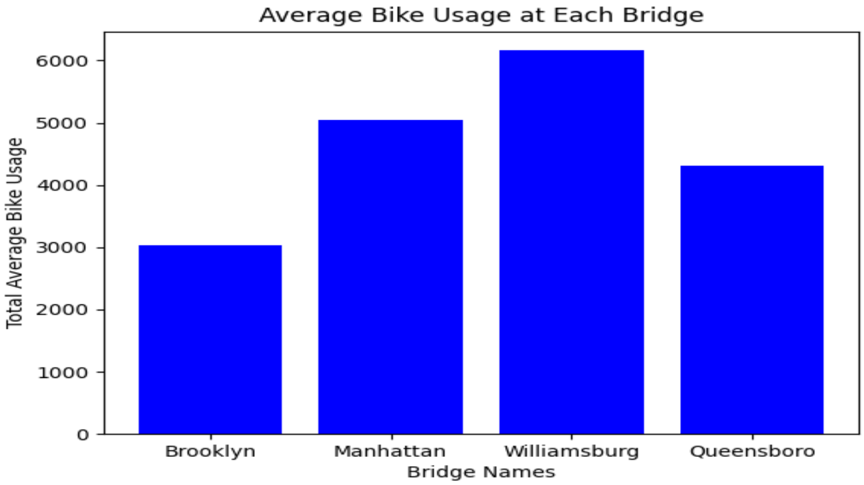


Figure 3: Average Bike Usage at Each Bridge

From Figure 3. above it is obvious the Manhattan, Williamsburg, and Queensboro bridges should be funded for sensors, as they clearly show on average, they have the most bicyclists and hence will ensure the sensors are used to capacity.

<u>Problem 2:</u>

- **Approach:**

For this particular question, we need to determine if there is a correlation between the amount of bike users and the precipitation in inches, the low and high temperatures. As we have more than 30 data points, we can use ridge regression and test-training models. On the X axis we had the day and on the Y axis we had the total amount of bikers/bike usage. We created normalized data and split it into training and testing data our cut off was 20% for testing data and used it to incrementally split and train our model. Once this was complete, we used Ridge regression and a five fold cross validation on the new created test data to better improve our model. This ensures that our model would we as close to the correct data as possible.

- **Analysis:**

Figure 2 below, shows the relation between the actual data and the model we created.

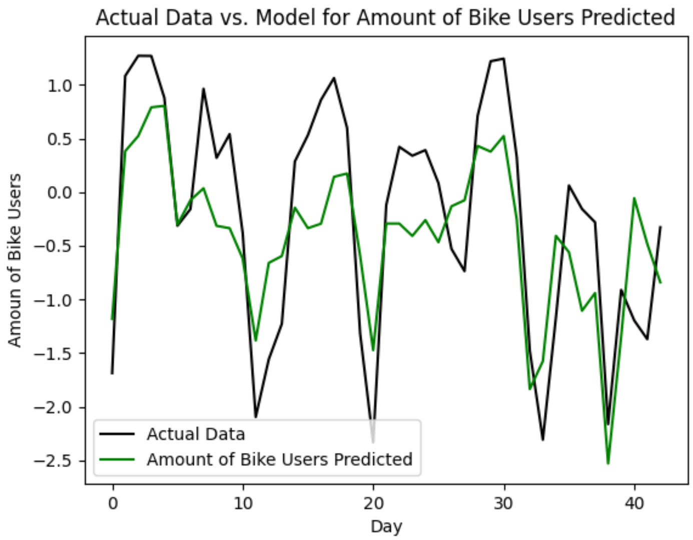


Figure 4: Actual Data vs Model for Amount of Bike Users Predicted

We created a model that would measure how many bikers there are using low and high temperatures and precipitation as correlated factors. We used Ridge regression with k cross validation to accurately create this model.

The model created had the following coefficients:
High Temp : 0.8828700009508329
Low Temp : -0.34909057311731373
Precipitation: -0.3269778410491868
Y intercept: -0.040331726592985596
The coefficients indicate how the variables of high, low temperature and precipitation relate to the number of bikers.

By normalizing the data in the model we created a very low mean square error of 0.456, indicating our model was not skewed. Despite creating a fairly accurate model as shown by figure 2, and our mean square error, our coefficient of determination, 0.595 for the model, only gives us 59.5% confidence that the variables are connected and can only account for this 59.5% of the data.

Hence, as it cannot account for the rest of the data, it is reasonable to assume that the variables of high and low temp along with precipitation, do not correlate and cannot be user to determine the amount of bikers that would be present reliably.

One shortcoming of this approach is that it is a prediction and it may not be valid for future cases due to unforeseen factors. Additionally, if a specific point were to be tested it would have to be within the time frame given in the data set as we can accurately make predictions only for April 1st to October 31st 2016.

<u>Problem 3:</u>

- **Approach:**

To determine whether the data is able to predict if it is raining based on the number of bicycles on the bridge, the approach taken was to create several polyfit models on their correlation. The group decided to use polynomial functions from degrees 1 to 5 using the forward selection procedure of polynomial selection. The polyfit function first creates 5 different polynomial functions of varying degrees and graphs them onto the precipitation and total bikes data. The next step taken was to find the mean squared error value and the coefficient of determination of the model. With these values, the group is able to determine if the precipitation values affect the number of bikes on the bridges.

- **Analysis:**

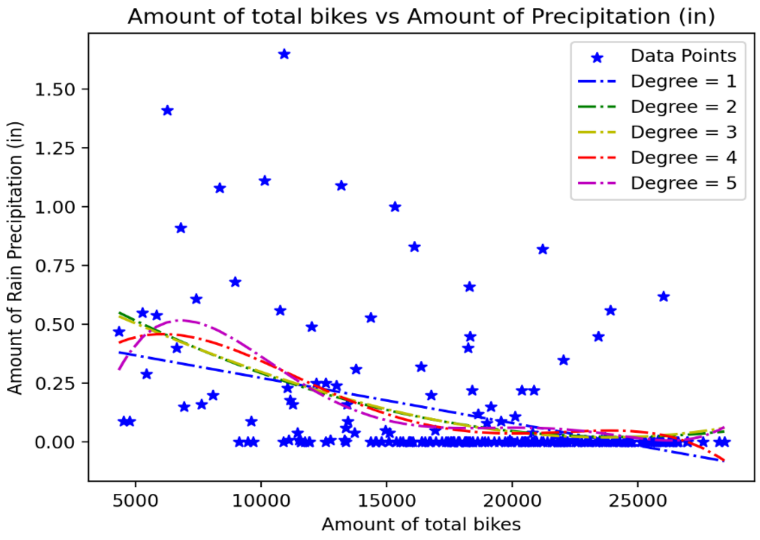
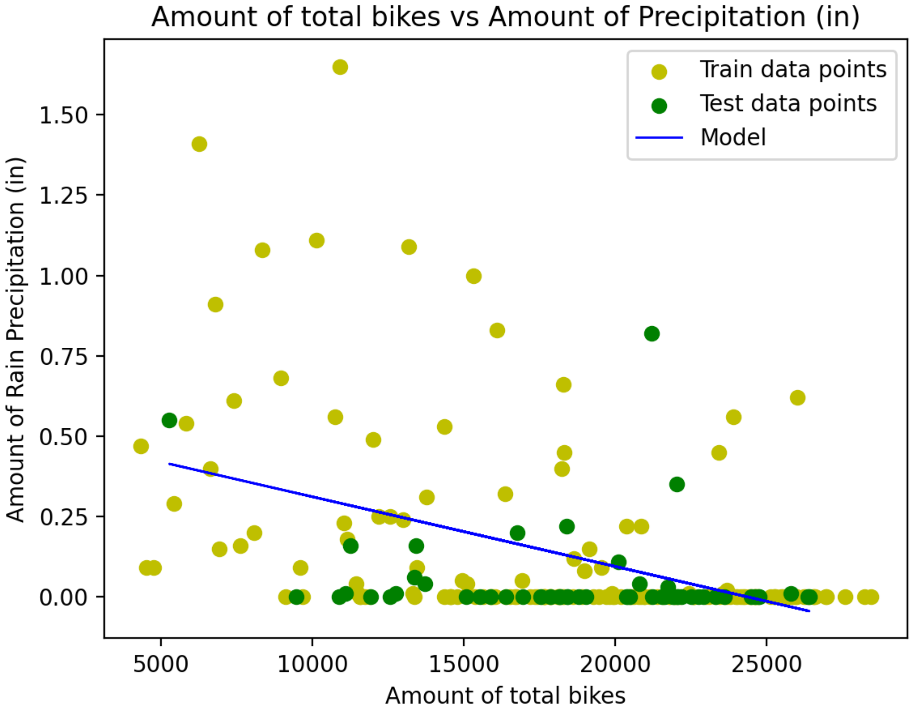
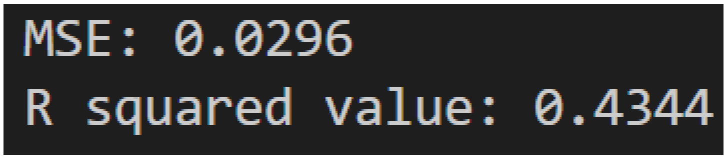


Figure 5: Graph of amount of total bikes vs amount of precipitation (in) with varying polynomial degrees equations

Within the figure shown above, it shows little to zero correlation between the dataset and each of the degree polynomial equations. Using the forward selection procedure, starting from degree 1, the polynomial model gets closer to the model of the dataset, but even at degree 5, it doesn't demonstrate any interaction with the dataset.





```
MSE: 0.0296
R squared value: 0.4344
```

Figure 6: Graph of amount of total bikes vs amount of precipitation (in) with train and test data points and MSE and coefficient of determination

The mean squared error value received from the dataset was 0.0296 and the coefficient of determination was 0.4344. The coefficient of determination shows that there is barely any correlation between the two variables, total bikes and precipitation, further showing the point of no conclusive way of predicting the total amount of bikes based on the amount of precipitation. The mean squared error value shows a low score, but can be explained by the heavy amount of 0 precipitation values across various amounts of total bikes.

To conclude, there is no rational assumption of the correlation between the amount of total bikes on the bridges and of the precipitation occurring on a single day.