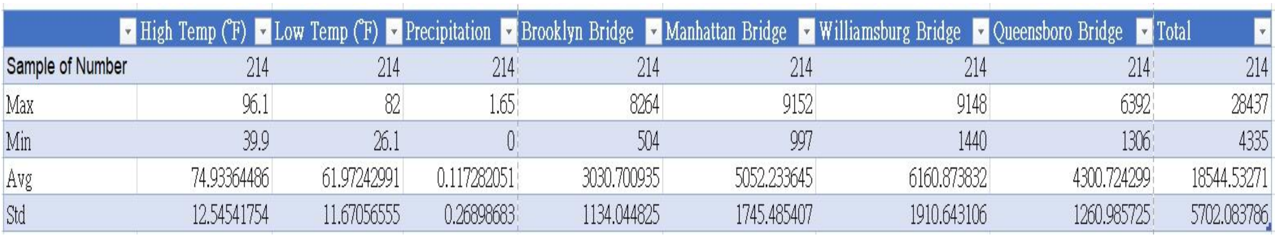
Team Member Name(s):BO-YANG WU   PO-YING HUANG
Purdue Username(s): wu1563 , huan1478
GitHub Username(s): oscarwu100 , Bowen0305
GitHub Team Name: bo-po
Project: Path 1: Bike Traffic

Dataset (15 points): Describe the features, measurement units, number of samples, range of features, and source of the dataset used in the project in one paragraph.

The data we used is original from NYC_Bicycle_Counts_2016_Corrected.csv, which we had also uploaded to GitHub. There are 214 x 10 data in this file. There is Date, Day, High Temp, Low Temp, Precipitation, Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge, Queensboro Bridge, and Total. The units follow is date, day, Fahrenheit, Fahrenheit, mm, bikes, bikes, bikes, bikes and bikes.



	High Temp (°F)	Low Temp (°F)	Precipitation	Brooklyn Bridge	Manhattan Bridge	Williamsburg Bridge	Queensboro Bridge	Total
Sample of Number	214	214	214	214	214	214	214	214
Max	96.1	82	1.65	8264	9152	9148	6392	28437
Min	39.9	26.1	0	504	997	1440	1306	4335
Avg	74.93364486	61.97242991	0.117282051	3030.700935	5052.233645	6160.873832	4300.724299	18544.53271
Std	12.54541754	11.67056555	0.26898683	1134.044825	1745.485407	1910.643106	1260.985725	5702.083786

Methods (30 points): State the three analyses questions from the README. For each question, provide a one paragraph explanation describing the analyses used to answer the question (e.g., why the method was chosen, which features from the data were used to build the model, how the model was verified). Do NOT state the results here.

- Q1:

  We use Pearson’s correlation coefficient, which normally is used to summarize the linear relationship between two data samples, and Spearman’s correlation, which normally is used to summarize the non-linear relationship between two data samples.

  We build the model by using traffic on each bridge and the total traffic. Also, the closer the correlation to 1, the stronger the relationship is.

- Q2:

  We use the linear regression method to build a model for the following relationship pairs:

  1. high temp - total bikes,
  2. low temp-total bikes
  3. avg temp - total bikes
  4. high and low temp - total bikes
  5. precipitation - total bikes
  6. high temp, low temp and precipitation - total bikes.

  We believe that there is some relationship between the values from the weather forecast and the total bikes, the relationship might be causation or confounding. This leads to the next question, how strong is the relationship for each variable. To determine the coefficient and degrees for each variable we think the linear regression method is the best method that fits our needs. By using the linear regression method, we will define 2 / 3 of data from the data set that was given as our training-set and the rest of the data will be our test-set. Then we will build up a linear regression model from degree 1 to 8 for each relationship pair and calculate accuracy and the MSE to find the best model that is going to predict the total bikes based on the weather forecast.

- Q3:

  We use Spearman’s correlation, which normally is used to summarize the non-linear relationship between two data samples. We build the model by using precipitation and the total traffic. Also, the closer the correlation to 1, the stronger the relationship is.

Results (30 points): For each analysis question, describe the outcomes of each analysis and the conclusions you drew from the outcomes. Essentially, you should answer the questions stated in the Methods here. Feel free to include charts, graphs, tables, or any other visual aids to help explain the conclusions drawn from your analyses.

- Q1:

  We should install the sensors on Manhattan Bridge, Williamsburg Bridge, and Queensboro Bridge to get the best prediction of overall traffic.

  Pearson’s correlation coefficient we get in this question is:

Brooklyn Bridge:	0.8744125296971798
Manhattan Bridge:	0.9354741757110536
Williamsburg Bridge:	0.9750891971316352
Queensboro Bridge:	0.963180456707312

  This shows that we should install the sensors on Manhattan Bridge, Williamsburg Bridge and Queensboro Bridge, instead of Brooklyn Bridge. Because the Pearson’s correlation coefficient for Brooklyn Bridge is only around 0.87, which is okay but not as good as the other 3 bridges.

  The Spearman’s correlation coefficient we get in this question is:

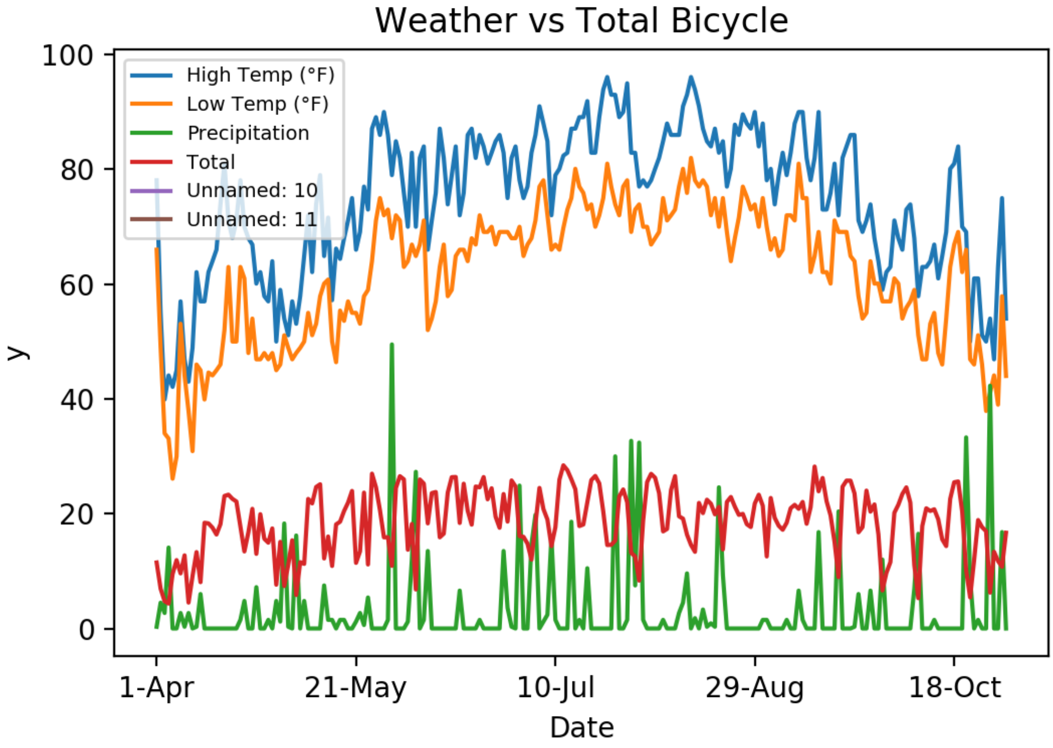
Brooklyn Bridge:	0.9098508922932541
Manhattan Bridge:	0.9204739700934029
Williamsburg Bridge:	0.9750891971316352
Queensboro Bridge:	0.963180456707312

  This also shows that we should install the sensors on Manhattan Bridge, Williamsburg Bridge and Queensboro Bridge, instead of Brooklyn Bridge.

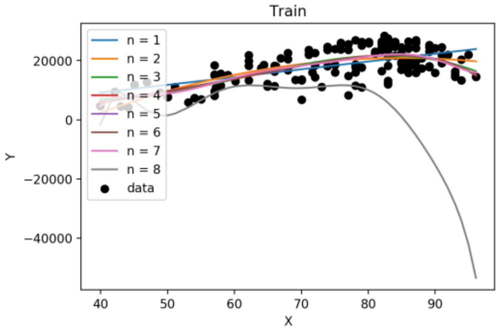
  Comparing the results we got between Pearson’s correlation coefficient and Spearman’s correlation coefficient, we can tell that both results are almost the same. There is a really strong relationship between the traffic on each bridge and the total traffic.

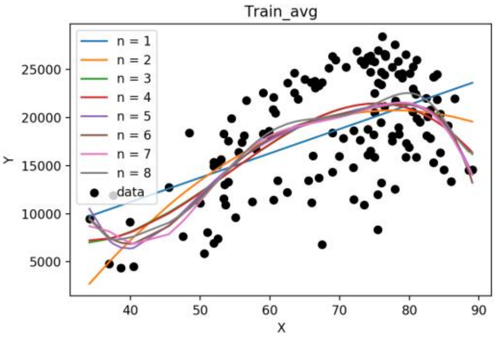
- Q2:

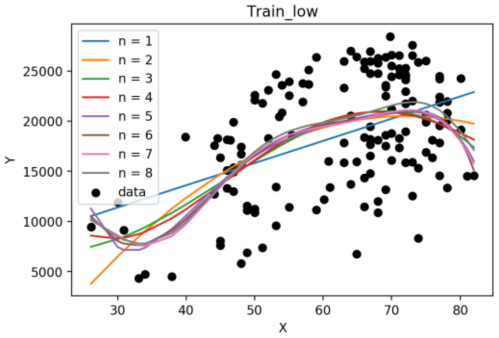
  For this question, we will say YES to this question. We can use the next day’ s weather forecast to predict the number of bicyclists that day.

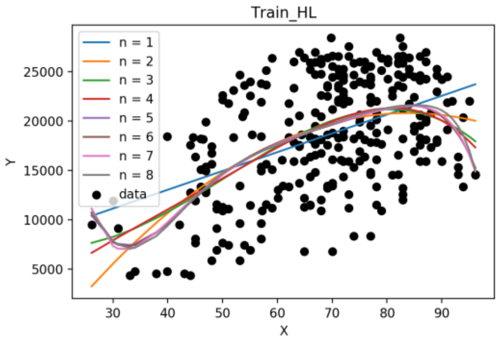


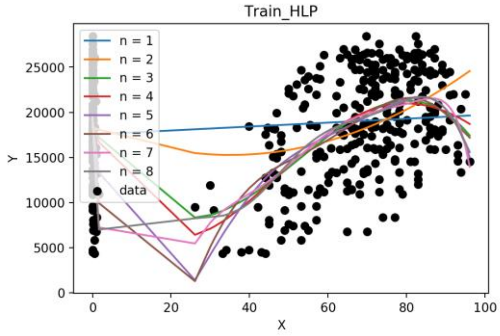
Training

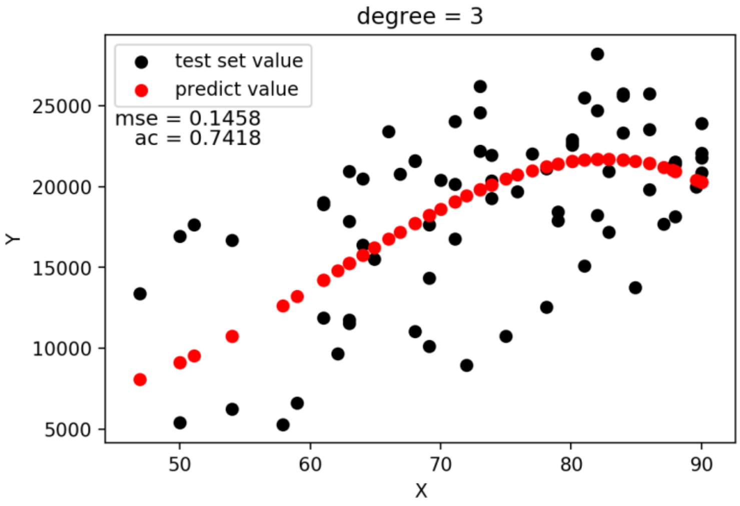










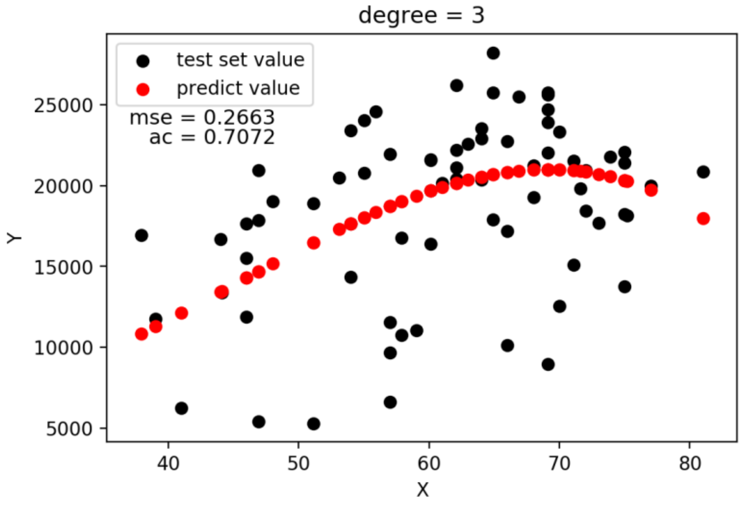


high temp - total bikes:

lowest mse : 0.1492

degree = 3

accuracy = 0.7418

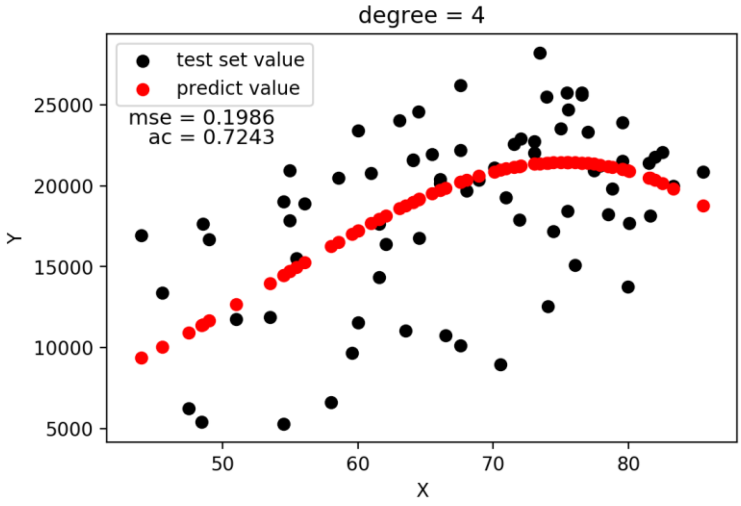


low temp - total bikes:

lowest mse : 0.2663

degree = 3

accuracy = 0.7072

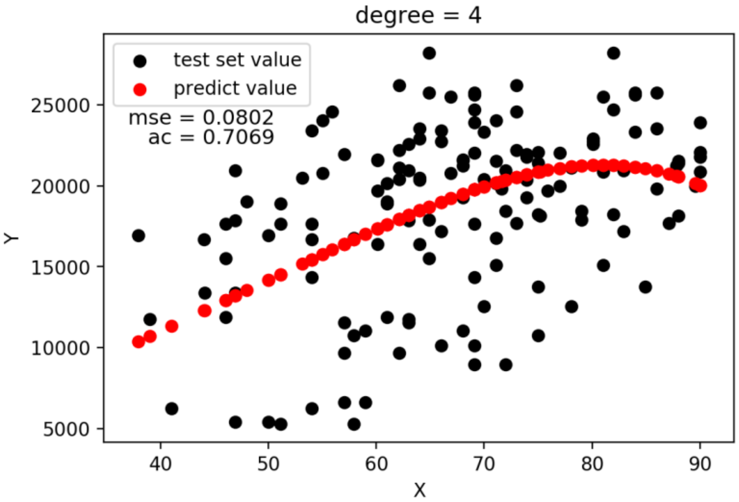


average temp - total bikes:

lowest mse : 0.1986

degree = 4

accuracy = 0.7243

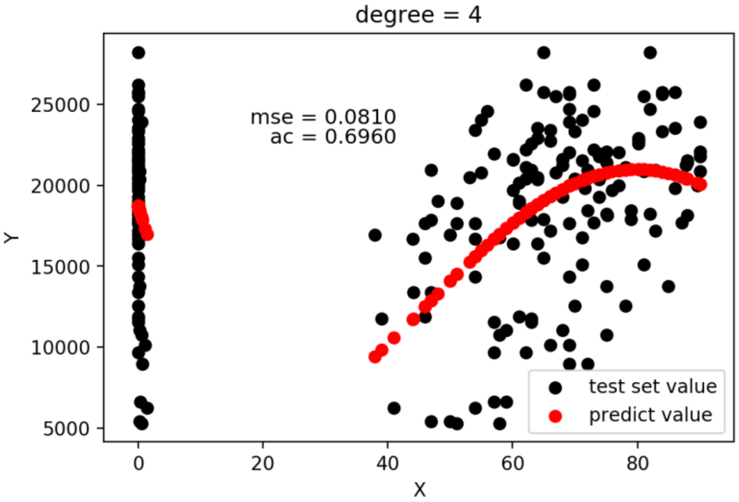


high and low temp - total bikes:

lowest mse: 0.0802

degree = 4

accuracy = 0.7069

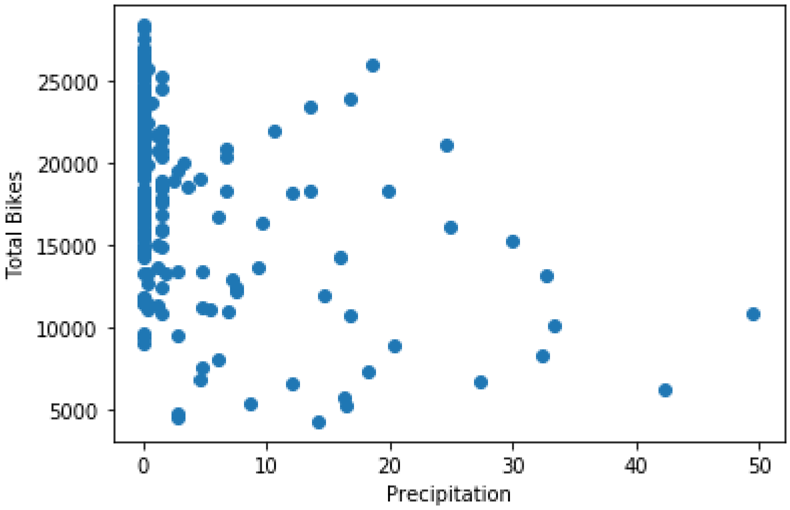


high temp low temp and precipitation

lowest mse: 0.081

degree = 4

accuracy = 0.6960



precipitation versus total bikes

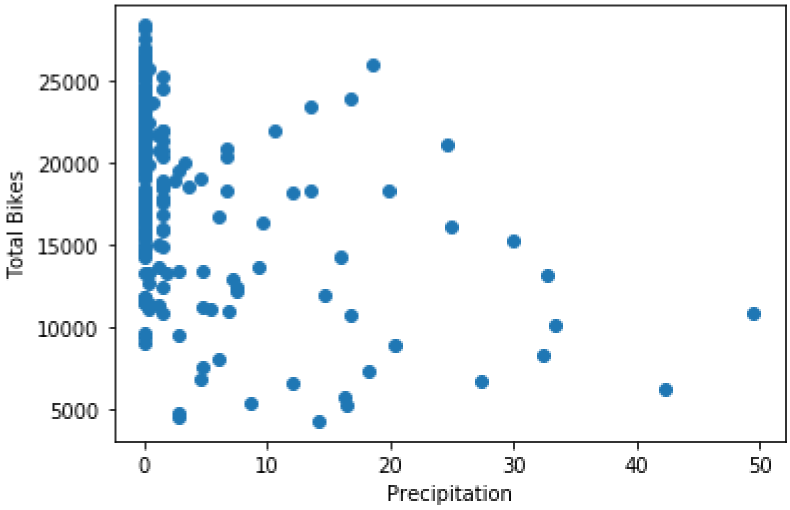
From the graph “Precipitation versus Total Bikes” we didn’t see a clear relationship between them. When going through all the models we find that including precipitation as our variable lowers the accuracy rate, so we decided to not do the model with precipitation only. From all the models we’ve done we find that predicting total bikes based on high temperature only has the best performance with a 74.18% accuracy rate, which is the highest accuracy we had found in this question. We also tried if we can increase the accuracy rate by normalizing the variables, unfortunately the accuracy remains the same.

When we used the model which is trained by high temp, low temp, and precipitation. We found out that the accuracy is 69.60% and the MSE is 0.081. Which is a little bit too low. But we found out that when we only used the model based on high temp and low temp, the accuracy growth a little bit to 70.69 %. But when we used the model only based on high temp,the accuracy increased to 74.18%, which is a good result.

- Q3:

We will say it is a NO for this question. We cannot use this data to predict whether it is raining based on the number of bicyclists on the bridges

We already know that the relationship between Precipitation and the total traffic are low.



According to this graph, we can easily tell that the relationship between Precipitation and the total traffic is low.

The Spearman’s correlation coefficient is only around -0.4, which is only 16% related between overall traffic and precipitation.

But we think maybe it is because there are too many outliers, so we delete some of the data. We keep the total traffic which is under 20000, but the relationship is worse than before. The Spearman’ s correlation coefficient is around -0.12. Which is really bad.

But we think in another way, maybe it is because there is too much precipitation equal to 0 in the dataset, so we delete the data which precipitation is 0. But we found out that it is even worse than before. The Spearman’ s correlation coefficient is only around -0.3, which is not even 10% related between overall traffic and precipitation. But we still did not give up. We keep the data which precipitation is higher than 0.05, we think maybe it is too low. But we found out that it is still worse than before. The Spearman’ s correlation coefficient is only around -0.12, which is not even 1% related between overall traffic and precipitation.