Project ECE20875: Python for Data Science

Group information:

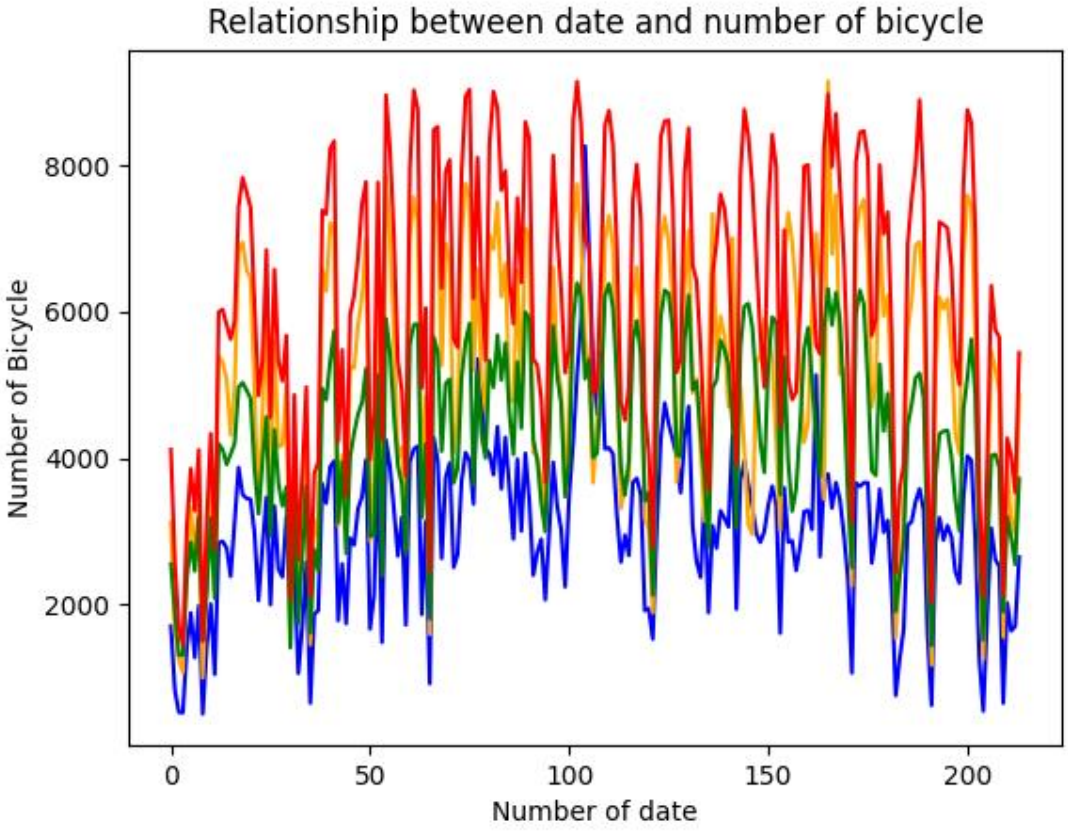
ECE20875
Guanxi Zhou, zhou1139
Yilong Peng, peng280
Path1

Dataset:

The highest temperature is on July 23, the lowest temperature is April 3. Over half of date is 0 precipitation, and the highest precipitation is 1.65 on May 30. The lowest number of the bicycle on Brooklyn bridge is on Apr 9 and the highest is on July 14. The lowest number of the bicycle on Manhattan bridge is on Apr 9 and the highest is on Sep 13. The lowest number of the bicycle on Williamsburg bridge is on Apr 4 and the highest is on July 12. The lowest number of the bicycle on Queensboro bridge is on Apr 3 and the highest is on July 12.

This graph shows the number of bicycles of different bridge on different date.
The red line is Williamsburg bridge.
The orange line is Manhattan Bridge.
The green line is Queensboro Bridge.
The blue line is Brooklyn Bridge.

We can see that the Williamsburg bridge is the highest, and Brooklyn bridge is lowest.



Analysis and Result:

Problem 1 analysis:

For problem 1, we will choose to use linear regression model. We will use three bridges' traffic data to determine which three bridges are the solution. Since we have 4 different bridges, so we will have different combination.

i. Brooklyn, Manhattan, Queensboro.
ii. Brooklyn, Manhattan, Williamsburg.
iii. Brooklyn, Williamsburg, Queensboro.
iv. Manhattan, Williamsburg, Queensboro.

We will build the linear regression model for these four different groups of bridges. We will use the model to find the best fits the total bike traffic data. When we are getting the dataset, we can use it to predict the overall traffic. We will find the r squared for each group, r square is ranges from 0 to 1, 0 is mean that the model does not represent the dataset at all and 1 is mean the model is perfectly representing the dataset, so if the r square is closer to 1, that is mean that group will have a better prediction of overall traffic.

Problem 1 Result:

r^2 Value Table

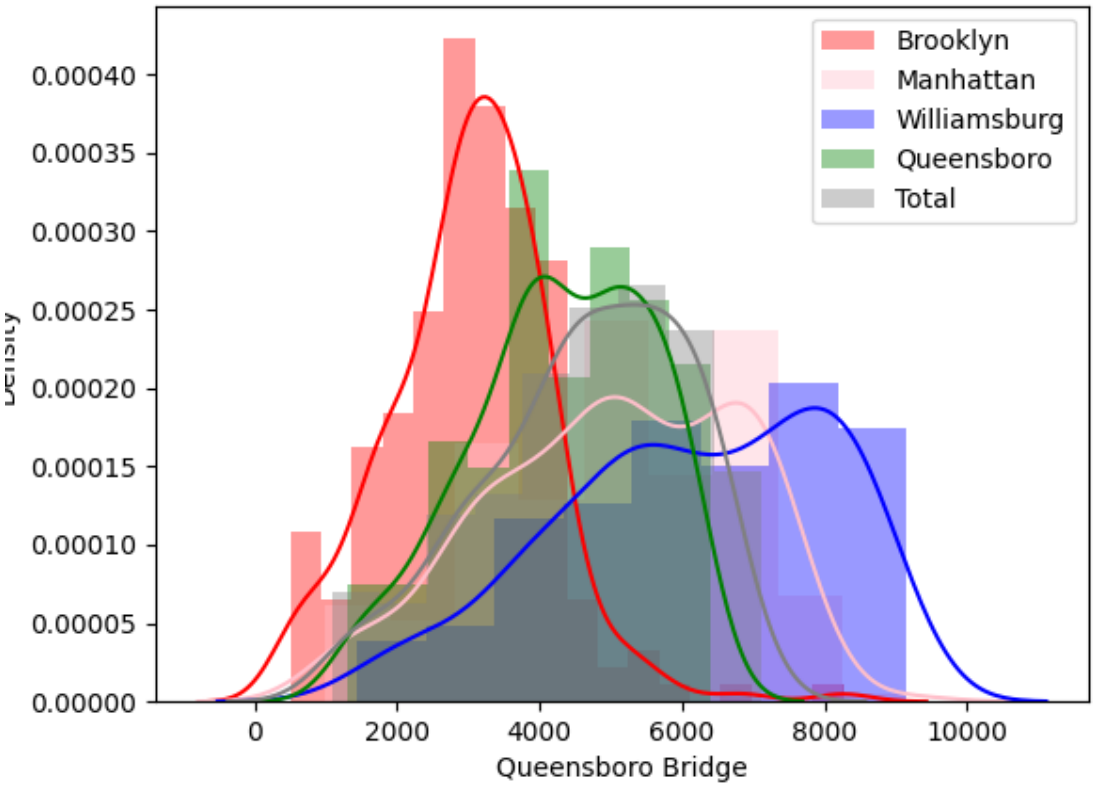
	Brooklyn, Manhattan, Queensboro.	Brooklyn, Manhattan, Williamsburg.	Brooklyn, Williamsburg, Queensboro.	Manhattan, Williamsburg, Queensboro.
r^2	0.988	0.996	0.947	0.982

Model Equation:

Total Traffic = 1.138600078871525*(Brooklyn Traffic) + 0.9471171505682359*(Manhattan Traffic) + 1.6086469611158554*(Williamsburg Traffic) + 382.74566817824234

When we are looking at this data, we can clearly see that the Brooklyn, Manhattan, Williamsburg has the highest r^2 value, which is showing that the prediction score for install the traffic sensors to the bridges was most accurate for these three bridges. We can conclude that if we only have enough budget to install sensors on three of the four bridges, Brooklyn, Manhattan, Williamsburg should we install the sensors on to get the best prediction of overall traffic.

Distribution Plot for Problem 1



Problem 2 analysis:

We can take high temperature, low temperature, and precipitation as input (independent variables). The number of cyclists in four different locations was taken as the output (dependent variable) to set up the model. After setting up the model, we can input different data to predict the number of cyclists that day.

Problem 2 result:

The test size we use in problem2 is 1000, and we got the score is 0.6397986319604876.
The equation we get from the code is:

Total Traffic = 387.36259394223805*(High Temp) + -164.26730430576984*(Low Temp) + -7918.844989943161*(Precipitation) + 575.0528560966159

We can use this equation and the situation of the day we want to test to get total traffic in that day. However, we cannot make a very accurate code to test, because when the test size is 1000, the score we get is 0.64, which means that there still a great possibility of misjudging the number of the total traffic on that day.

Problem 3 analysis:

We will find the average number of bicyclists on each of the Bridges per day in the week. Depending on the number of cyclists on each bridge, the average number of days on different Bridges obtained in step 1 is similar (close to which day in the week) to determine the corresponding days of the week.

Problem 3 result

We integrated total traffic on the different Bridges each day, and finally found the average number of people who crossed all the Bridges each day. These are the average traffic of different day in a week:

Monday	Tuesday	Wednesday	Thurs day	Friday	Saturday	Sunday
19393.70967 7419356	20782.26666 6666666	22422.26666 6666666	2078 1.3	17984.58064 516129	15000.64516 1290322	13716.38709 6774193.

So, we can use these data to compare the total traffic which closest to it, to determine what day it is. But based on the code, it can't all be perfect. Because the total number of people crossing the bridge on Tuesday is like that on Thursday, there may be some error.