Team Members:
Ethan James
Nathan Bi

Usernames (Purdue):
Epjames
bi37

Github Username:
EthanPJames
ChickenMcSpicy

Path Taken: 2

Dataset Description:

By choosing to work with Path 2, we were given a set of data from 2016 that contained various information pertaining to bikers. We were given the date of the data entry, the actual day of the week, the high and low temperatures, the precipitation level, and the number of bikers on each bridge on this particular day. We were also given a column that totaled up the number of bikers on each bridge. For each column, there were 215 data entries which gave us plenty of data to analyze. The data set provided allows us to identify which bridges should have sensors as we will use the number of bikers on each bridge and the total number of bikers over all the bridges. Then using these values along with the precipitation and temperature data, we are able to find out if we can use the temperature and precipitation values to predict the number of bikers on a specific day based on the weather. Finally, we will use these columns along with days of the week to figure out if we can say what day it is based on the number of bicyclists out.

Method:

Part 1:

In part 1, we are asked to install sensors on 3 out of the 4 bridges to get the best prediction of overall traffic. Since we cannot install sensors on all of the bridges, we chose to install sensors on the bridges with the highest number of bikers as compared with the total number of bikers over all four bridges. We then opted to find the correlation between the number of bikers on each bridge as compared with the total number of bikers over all four of the bridges. We first get the total bike count by creating a column 'Total' which is every bridge's bike count added up together. Then, using the correlation function we are able to find the correlation between each bridge's bike count and the total bike count. Then we store each bridge's correlation values in a dictionary in descending order. Since we only want to put sensors on the three bridges, we print out the first three bridges with their correlation values. By using the correlation values, we are able to identify which three bridges best represent the overall traffic across all the bridges since we have essentially compared every bridge which are in a sense normalized that way, they can be compared to each other.

Part 2:

In part 2, we were asked if the city can use the next day's weather forecast(low/high temperature and precipitation) to predict the total number of bicyclists that day. Essentially what we chose to do is use regression to check if we can use the next day's weather to predict the total number of bicyclists. Our intention was to try different degrees of fit and then analyze the r^2 value to determine if there are any models out there that will give us an accurate prediction of the total number of bicyclists that day. We first constructed a linear model to see if that would fit the data by checking its r^2 value. We also then created a while loop that runs through a polynomial fit that alters the degree of the polynomial every time. We then were planning to analyze each r^2 value of every polynomial to see if we get an r^2 above 0.9 as that is a valid threshold to find a model that can predict the number of bicyclists based on weather. We also printed out the coefficients and intercept to the terminal for every model the code checked. Overall, the code is trying to determine which type of regression model (linear or polynomial) fits the data better by comparing their respective R-squared values.

Part 3:

In problem 3, the code is adding a new column "DayOfWeek" which extracts the day of the week from the "Date" column using the pandas to_datetime() function. It then creates a histogram using matplotlib to visualize the total number of bikers. Next, the code groups the dataset by day of the week and calculates the average total bikers for each day. The output of this calculation is then printed. Finally, the code determines the day with the highest average total bikers by finding the max.

Results:

Part 1:

Using the method specified above, we have found that sensors should be installed on the following bridges: 'Williamsburg Bridge', 'Queensboro Bridge', 'Manhattan Bridge'. These bridges had the highest correlation values.

Part 2:

Using the method specified above, we would have concluded that you cannot use the next day's weather to predict the number of bicyclists. There is some correlation, but it is not strong. (More about why in the Analysis section).

Part 3:

No, we cannot use this data to predict what day it is. In our method we found the max average of bikers on each day. If the values were far enough apart to be distinguishable, we could say what day it is based on the number of bikers, but because the values are so close together(the difference between Friday and Saturday is less than 2000) we cannot say that you can predict the day of the week based on the number of bikers. Saturday and Thursday mean are almost indistinguishable. Therefore, we cannot use the number of bikers to conclude what day of the week it is.

Analysis:

Part 1:

Our results indicate that the Williamsburg, Queensboro, and Manhattan bridges all have the highest correlation values with total number of bikers per day. This is why these three bridges should get sensors installed since they will best represent predictions of overall traffic. The correlation values provided:

Correlation of Brooklyn Bridge and Total: 0.8744125296971794

Correlation of Manhattan Bridge and Total: 0.9354741757110537

Correlation of Queensboro Bridge and Total: 0.9631804567073113
Correlation of Williamsburg Bridge and Total: 0.9750891971316346
Indicate that the Brooklyn Bridge had the lowest correlation meaning the other three bridges would better represent a prediction of overall traffic which is why they were assigned the three sensors.

Part 2:

We have concluded that we cannot use the next day's weather to predict the number of bicyclist because we were unable to find a valid model. We have checked many models and found that no model gets over our threshold of an r^2 with 0.9. In general, 0.9 or above for an r^2 value means that you can use the model safely to make predictions and extrapolate. In our case, none of the models did this properly without overfitting which is not a model you would really want to use because then it will be hard to extrapolate from it. The highest r^2 we saw that we could potentially use came from a fourth-degree polynomial with a value of 0.69 which is still relatively low. We also printed the coefficients for each model along with their intercepts and they really did not make much sense which is why we have deduced that you cannot predict the number of bicyclists based on weather. There are definitely other factors that influence the number of bicyclists like what season it is, what month it is, and various other factors that we do not have access to through our data set. Here are some of the results below which back up our decision to say that we cannot predict the number of bicyclist based on weather.

Linear Model

Intercept: 178.20093422504215
Coefficients: [-162.32007876 390.91830834 -7951.48638461]
R-squared: 0.49945751567841035
End of Linear Model

Begin Polynomial Model(Quadratic)

Degree:
0
Intercept(polynomial): 18544.532710280375
Coefficients(polynomial): [0.]
R-squared(polynomia): 0.0
End Polynomial fit

degree
1
Intercept(polynomial): 178.20093422504215
Coefficients(polynomial): [0. -162.32007876 390.91830834 -7951.48638461]

R-squared(polynomia): 0.49945751567841046
End Polynomial fit

degree
2
Intercept(polynomial): -35661.144389328474
Coefficients(polynomial): [0.00000000e+00 -1.02300650e+02 1.44064864e+03 -2.47370586e+04
 -1.72084411e+01 2.78625391e+01 -1.59714331e+02 -1.92375926e+01
  2.85788128e+02 5.62849932e+03]
R-squared(polynomia): 0.5886916351725543
End Polynomial fit

degree
3
Intercept(polynomial): 82557.72456444005
Coefficients(polynomial): [0.00000000e+00 4.71752923e+02 -4.35824010e+03 4.78387883e+03
 -9.69741854e+01 1.39535698e+02 2.01470475e+03 2.13903813e+01
 -3.30723536e+03 8.02669266e+04 1.38342538e+00 -2.24986096e+00
 -1.15858488e+02 1.03557546e+00 1.67004896e+02 6.95460745e+02
 -4.78010361e-01 -4.21740532e+01 -1.27066862e+03 -1.15193405e+04]
R-squared(polynomia): 0.6492642668130242
End Polynomial fit

degree
4
Intercept(polynomial): 73630.40350427867
Coefficients(polynomial): [-4.47465755e-02 4.68393078e+03 -6.12346409e+03 -1.00337964e+06
  3.56883268e+00 -1.92729163e+02 1.21048721e+03 1.68354860e+02
  3.65672409e+04 7.88109060e+05 2.85485573e+00 -1.04329707e+01
  9.61651434e+02 1.35465904e+01 -1.47244994e+03 -3.21083798e+04
 -5.31973358e+00 6.68170772e+01 7.35738517e+03 3.23828548e+04
  4.61364879e-02 -1.81715368e-01 1.53173035e+01 2.99578934e-01
 -3.97910017e+01 -1.52964162e+03 -2.34261333e-01 3.29644756e+01
  2.77910154e+03 1.72568051e+04 6.68863848e-02 -6.57918961e+00
 -1.11130494e+03 -1.62061316e+04 1.60236799e+04]
R-squared(polynomia): 0.6969737731567728
End Polynomial fit

Part 3:

We know based off the data that Wednesday has the greatest number of bikers. However, because each mean number of bikers per day of the week is so close to each other, we cannot accurately predict what day of the week it is. The only day we could possibly predict is Wednesday and we might only be able to predict Wednesday if the number of bikers is over a value of 22,000. Other than that, each day of the week has a number of bikers that we could say is essentially indistinguishable from each other. There is overlap in the data meaning if you picked a random day and found the number of bikers it could be one of many days of the week. Here is the data to prove to you that some values are so close together making those days of the week look indistinguishable from each other.

Day	
Friday	17984.580645
Monday	19393.709677
Saturday	15000.645161
Sunday	13716.387097
Thursday	20781.300000
Tuesday	20782.266667
Wednesday	22422.266667