Project ECE 20875: Python for Data Science Spring 2022

1. **Project Team Information**
   Mini-Project Spring 2022
   ECE 20875 - Section 001
   Name: Husam Chekfa
   Email: hchekfa@purdue.edu
   I worked on this project alone
   Path chosen: Dataset #1 about NYC bikes and bridges
   Program file: nyc.py
   All 3 questions are solved in the same .py file. Each clearly separated, both in the code and in its output.

PATH 1 QUESTIONS:

Question 1: You want to install sensors on the bridges to estimate overall traffic across all the bridges. But you only have enough budget to install sensors on three of the four bridges. Which bridges should you install the sensors on to get the best prediction of overall traffic?

Question 2: The city administration is cracking down on helmet laws, and wants to deploy police officers on days with high traffic to hand out citations. Can they use the next day's weather forecast (low/high temperature and precipitation) to predict the total number of bicyclists that day?

Question 3: Can you use this data to predict whether it is raining based on the number of bicyclists on the bridges (*hint*: The variable `raining` or `not raining` is binary)?

2. **Descriptive Statistics**

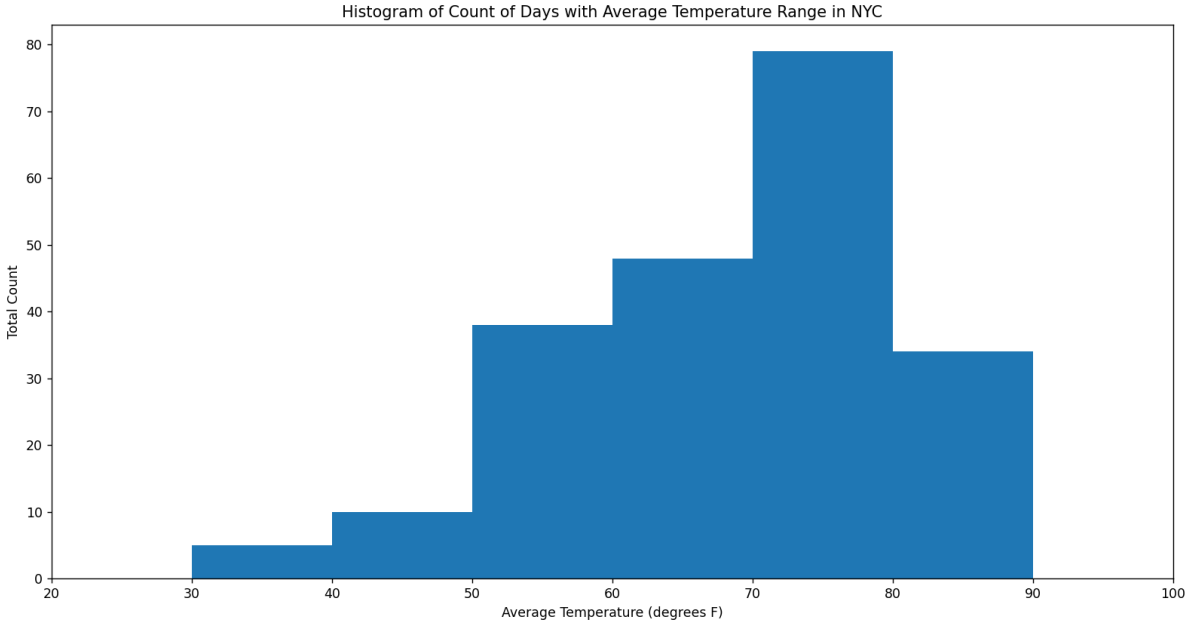
1. I use each column in the given data set in some way, except for the day of the week. The temperatures will be used in question 2 to find a rough estimate for the daily average temperature. The precipitation column will be used in questions 2 and 3 and will be useful in learning the impact on a biker's decision on whether to use their bike or another form of transportation based on the day's weather forecast for rain. The four bridges values are used in question 1 to find which bridge has an average bike count closest to the other three bridges' average. The column for total bike count is used in each question as a control variable and is used for regression and other modeling and helps describe how biker count is affected by NYC's weather. Even the date column is used for random numeric testing as a way to test my models in questions 1 and 2.

2. I will use the given values exactly the way they are given without modifying. The only exception is for the temperature. In question 2, I will find a rough estimate of a day's average temperature by averaging the day's high and low temperatures.
   I will be using averages or means, linear regression, and multivariable linear regression, and the numeric count of histograms to answer these questions.
   In question 1, I will also find the means of each bridge's biker count which are found here and below in my answer for question 1:

Bridge	Average Bike Count of Bridge	Average Bike Count of Other 3 Bridges
Brooklyn	3030	5170
Manhattan	5052	4496
Queensboro	4300	4747
Williamsburg	6160	4127

3. The average temperature in a day has a big effect on the biker count in NYC, as will be seen later in Question 2. I estimate the rough daily average temperature by averaging the day's high and low temperatures.

   Figure 0. A histogram of the given temperature data set. Counts the number of days with an average temperature in a range of 10 degrees F.



   I will be using linear regression and multivariable linear regression in Question 2, where the average temperature is an independent variable. This peak count in the 70-80 degrees F range will be important later in the report.

3. Approach

Question 1

For question 1, I decided to go with a basic approach. I decided to find which bridge had a daily bike average around the value equal to the average of the other three bridges. This way, I could set up the three sensors on the group of three bridges. Then, I can take their average and multiply it by four to find an approximate value of the four bridges' total bike traffic on any given day. The only error here would be on the fourth bridge. I describe why this works with more math below in the Analysis portion.

I took the 214 data points for each bridge and took their average. Then, for each bridge A, I compared its average to the average of the other three bridges, B, C, and D. I found that the single bridge that has an average traffic count approximately equal to the other three bridges' average is the Queensboro bridge. The specific math is found in the Analysis section, but briefly, the Queensboro bridge will not have a sensor and this will result with a daily error of ~2.5%.

Question 2

For question 2, I decided to go with linear modeling and multivariable linear regression. In this problem, I can assume I have access to the exact data of all four bridges, so for counting bikes, I used the "Total" column in the spreadsheet. I will compare the total bike traffic to both temperature and precipitation, since a daily forecast includes both. For temperature, I will use the rough estimated average temperature of the day, found by averaging the high and the low temperatures of the day. This can be further improved by using multivariable linear regression with the feature matrix consisting of the average temperature and precipitation value. The target vector will consist of the known total values of daily bikers found in the spreadsheet.

Question 3

For question 3, I will calculate the percentage chance of a certain range of precipitation for several ranges of bikers. This means that the independent variable will be the biker count and the dependent variable is the precipitation amount. I will choose five ranges of biker counts and another five ranges of precipitation. An important note is there is no rain when precipitation is 0.00. Else, it is sprinkling or raining.

4. Analysis

Question 1

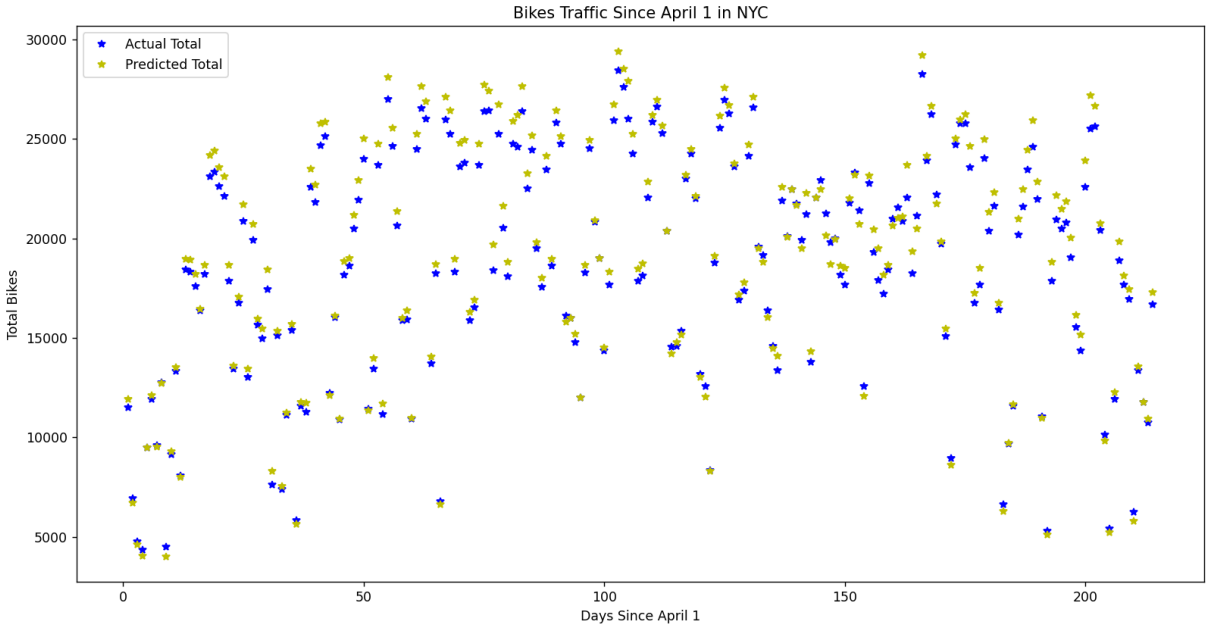
After using the approach from above, I found the combination with the least percentage error is having the Queensboro bridge without a sensor, while the Brooklyn, Manhattan, and Williamsburg bridges would have one each.
Here is the math that brought me to the conclusion:

Table 1. For each bridge, this table compares the bridge's average bike traffic count to the average traffic of the other three bridges. The last column holds the percent error between the two averages.

Bridge with no sensor	Bridges with sensor	No sensor bridge average	Bridges with Sensor Average	Percent Error (%)
Brooklyn	Manhattan, Queensboro, Williamsburg	3030	5170	41.39
Manhattan	Brooklyn, Queensboro, Williamsburg	5052	4496	12.37
Queensboro	Brooklyn, Manhattan, Williamsburg	4300	4747	9.42
Williamsburg	Brooklyn, Manhattan, Queensboro	6160	4127	49.26

The table above lists the four combinations of one bridge having no sensor and the three remaining bridges with a sensor. The combination averages are included, followed by a percent error.
The Queensboro row is highlighted, as it has the lowest percent error of 9.42%.
However, this 9.42% only refers to the error that will be found for the Queensboro bridge traffic value. Since my approach uses the three bridge group's average, their value will be exact. The only error will be found in calculating the Williamsburg daily traffic.
Because of this, the overall error of the four bridges falls from 9.42% to approximately 2.5%.
To ensure this approach worked, I chose several days by random to ensure this method worked, as seen below in Table 2.

Figure 1. A graph comparing the actual daily bikers using the four bridges to the total predicted by the model. Data from the given .csv file. Dates from April 1 to October 31.



In Figure 1, there is a clear graphical comparison between the blue actual total versus the yellow predicted total bikers per day. Graphically, it is clear that most days, the model almost perfectly calculates the true total. But, it is important to confirm the difference numerically, as shown below in Table 2.

Table 2. This table holds the total bike traffic and the model's prediction of four random dates in the dataset. The last column holds the absolute value of the percent error between the actual and model counts.

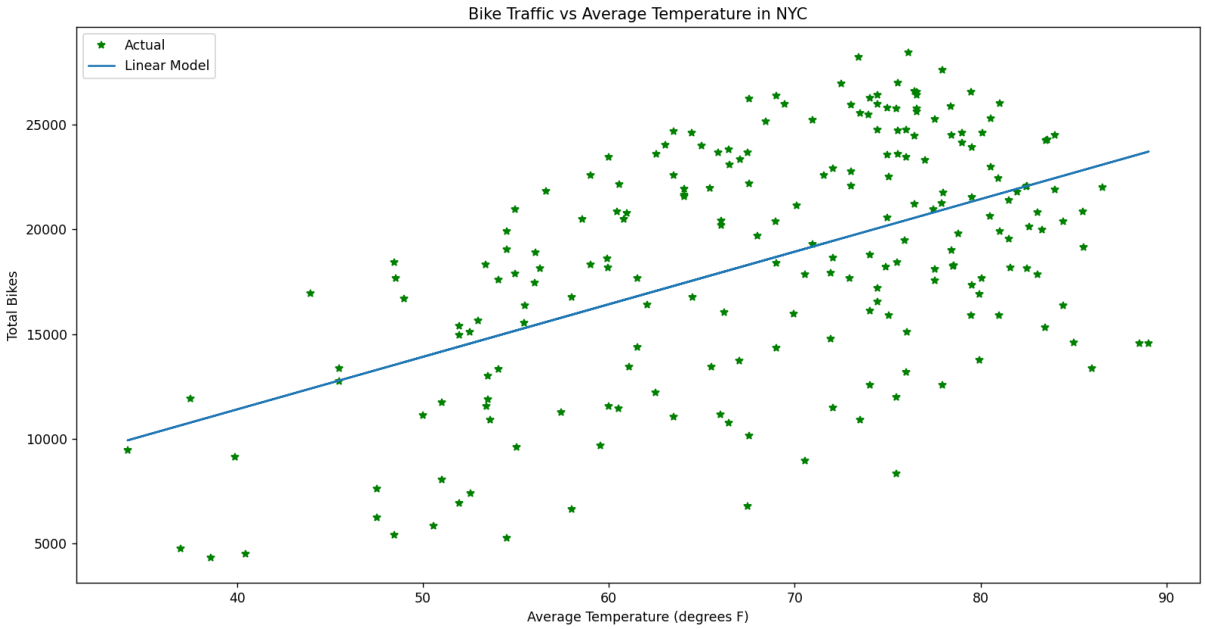
Date (Day-Month)	Actual Bike Count	Model Bike Count	Percent Error (%)
1-Apr	11,497	11,924	3.58
31-May	24,477	25,228	2.98
2-Aug	25,523	26,128	2.32
4-Sep	17,895	19,492	8.19

To ensure the model is reasonable beyond a graphical check, I check four random dates to find the percent error between the actual versus model totals. As seen in Table 2, these predicted values are very close to the actual values. Of course, like any other model, my model can over- or under-estimate the total but the percent error remains low.

Question 2

To begin, I used linear regression to find a generic linear model between total bikes observed versus the average temperature or precipitation of the day.

Figure 2. A graph containing a scatterplot of total bikes observed on the four bridges compared to the day's average temperature. The linear model of the data is superimposed on the graph.

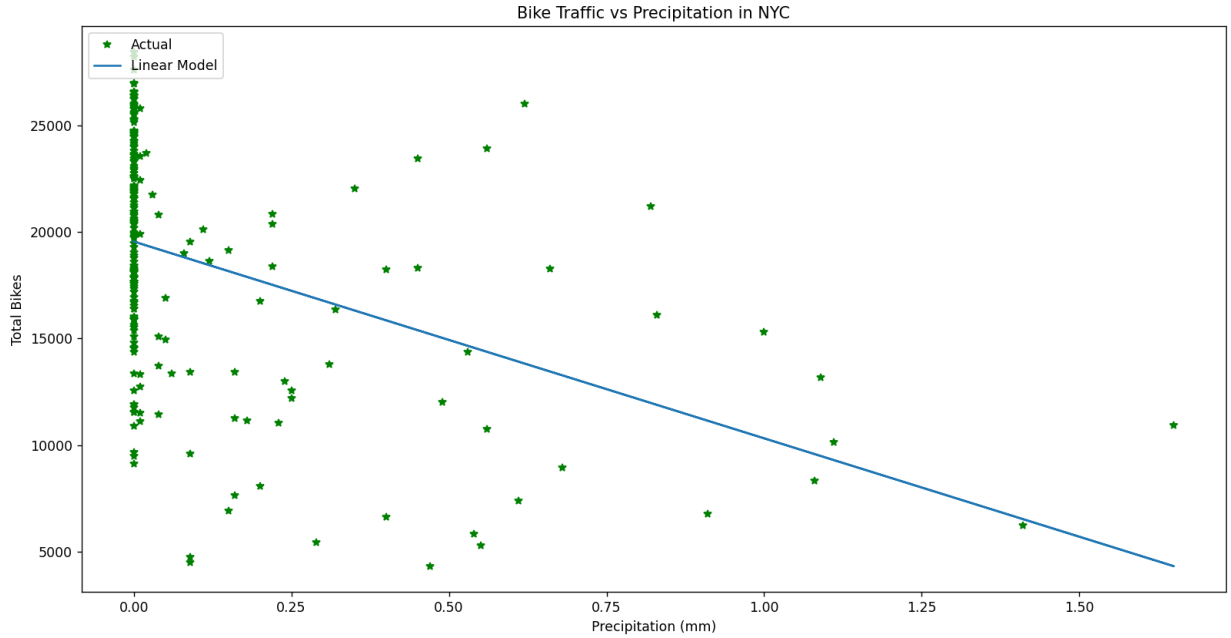


Above, Figure 2 is showing a clear positive relation with temperature and daily bikers. From the graph alone, it is clear there are more bikers in New York City when the average temperature of the day is higher. This observation is furthered when seeing the equation of the linear model. The equation is:

$$y = 250.8 * x + 1378$$

Here, the y value represents the dependent value, total bikes observed. The x value represents the average temperature in degrees F, where the rough average is estimated by averaging the high and low temperatures of the day. This equation makes sense, as there will be few bikers when the temperature approaches freezing. But as temperature approaches room temperature and above, it becomes a better temperature more suitable for biking, thus an increase in bikers is observed with a larger x value.

Figure 3. A graph containing a scatterplot of total bikes observed on the four bridges compared to the day's precipitation. The linear model of the data is superimposed on the graph.



Although the precipitation graph appears to be better suited for an exponential decay model, this is due to the large disparity between points with 0.00 precipitation. The model is accurate for days with more than 0.00 mm of precipitation.
The equation of the linear model is:

$$y = -9228 * x + 19550$$

Again, the y value represents the dependent value, total bikes observed. The x value represents the precipitation in millimeters. This equation makes sense, as there will be few bikers when the rain is coming down heavily. But if there is little to no rain, it becomes a better climate more suitable for biking, thus an increase in bikers is observed with a lower x value.

However, despite these two graphs clearly showing positive and negative relationships between biker count and weather, I cannot only use these. The weather includes both temperature and precipitation, as clearly seen by the massive range in values for the Precipitation value of 0.00 mm in Figure 3.
To reason with this, I decided to use multiple-variable regression. Here, the independent variables are the average temperature and precipitation values of the day. These will be columns one and two of my feature matrix. My target vector holds the dependent variable which is the total bikers of the day.
After using sklearn's multivariable linear regression on the data, I found that it was a fairly accurate model for the given data. Since this process is more complex than that in Question 1, I analyzed the result two ways. First, with four random dates again, then I calculated the average error across all the data.

Before discussing how I found the model to be accurate, I want to display my equation.
The equation is:

$$y = 249.245 * x_1 - 9140.946 * x_2 + 2479.905$$

Here, the x_1 value refers to the average temperature of the day in degrees F and x_2 refers to the precipitation of the day in millimeters. The last term is the y-intercept.
Something important to note is that the values are found with the given data, so a value outside of the temperature and precipitation range found in the spreadsheet will likely be less accurate.

First, I chose four dates by random and calculated the model's bike count.

Table 3. This table holds the total bike traffic and the model's prediction of four random dates in the dataset. The last column holds the absolute value of the percent error between the actual and model counts.

Date (Day-Month)	Actual Bike Count	Model Bike Count	Percent Error (%)
6-Apr	11,919	11,826	0.775
17-May	18,605	17,409	6.425
7-Jun	25,249	21,808	13.625
30-Aug	23,808	21,671	7.020

As seen in Table 3, the model calculated reasonably accurate values for these four dates. April 6th has an especially accurate model calculation, while June 7th has an error of almost 14%.

As mentioned above, selecting four random dates is not enough to observe the true error of this method.
So, I calculated the total error of each date given in the spreadsheet and added them together. Then, I divided the error total by the total days in the spreadsheet. The result was surprisingly accurate: the model's average error is 7.749%.
This shows that my model is a very good way to calculate the expected number of bikes for any given day as long as the average temperature and precipitation values are known beforehand, likely with a weather forecast.

Question 3

For question 3, I need to find whether there is rain on a certain day based on the amount of bikers counted. This means the biker count is the independent variable and the precipitation guess is the dependent variable.
I split the biker counts and precipitation amount into a few ranges. The biker ranges are: 0-9999, 10000-14999, 15000-19999, 20000-24999, 25000 or more bikers. The precipitation ranges are: 0, 0.01-0.24, 0.25-0.49, 0.50-0.74, 0.75 or more mm of rain. An important note that it is not

raining when there is 0 mm of rain and it is raining if there is 0.01 mm or more of rain. This is why the first range of rain only holds the value of 0.

Table 4. This table holds the count of days with the five ranges of bikers and five ranges of precipitation. This contains 214 days of data. The last column contains the percentage of days in the biker count range which had no precipitation.

	Precipitation					
Biker Count	0 mm Count	0.01-0.24 mm Count	0.25-0.49 mm Count	0.50-0.74 mm Count	0.75+ mm Count	0 mm %
0-9,999	3	6	0	7	3	15.8
10,000-14,999	13	14	0	6	3	36.1
15,000-19,999	47	9	0	4	2	75.8
20,000-24,999	57	8	0	3	1	82.6
25,000+	26	1	0	1	0	92.9

The table above shows the amount of days with each amount of bikers and precipitation. The last column clearly shows that the more bikers there are, the less chance of there being rain. I chose to use five ranges, but for a more accurate measure, there would be more ranges, possibly each having a size of 1,000 rather than mine which has a size of 5,000.

An important thing to note is that there is no range with 0% precipitation, even in the 25,000+ biker count range. This could be because of a sudden downpour of rain in the middle of the day after the bridge sensors have already counted a large number of bikers.
However, this table still provides useful values and shows the likeli- or un-likelihood of rain based on the number of bikers.
Based on the values in the table, a generic function could be written which takes an input of the biker count and returns “True” if the count is 15,000 or above. Else, it will return “False”. The bool value will refer to “notRaining”. The boolean holding “True” means it is not likely raining that day, but if it holds “False”, then it is likely raining. Again, a smaller range than 5,000 for biker count would lead to a more accurate daily guess function.