Python For Data Science Final Project

Authors: Miller Kodish and Tom O'Donnell
Github: Miller11k and iBoot32
Section: 001 - Mahsa Ghasemi
Elmore Family School of Electrical and Computer Engineering
ECE 20875

Date: 04/21/2023

Table of Contents:

Link to Github Repository

1. Abstract:

According to a study published in Springer, real-time prediction of bicycle traffic is crucial for enhancing traffic safety in pedestrian-bicycle sections. By analyzing bike traffic data across several bridges in New York City, we can identify traffic patterns and develop appropriate safety measures.

In addition, monitoring bike traffic can help city planners identify areas where bike lanes or other infrastructure improvements are needed. By installing sensors on bridges, we can estimate overall traffic across all the bridges and determine which bridges have the highest volume of bike traffic.

Finally, predicting bike traffic can help city officials deploy police officers on days with high traffic to hand out citations for helmet law violations. By using weather forecasts to predict the total number of bicyclists on high-traffic days, city officials can ensure that they have enough officers on duty to enforce helmet laws and improve bike safety.

2. Objectives:

We have chosen to answer the problem that lies within the second path. This requires us to analyze the data from various bridges in New York City. It will comprise three main parts.

2.1 Sensor Installation:

The objective of this task is to determine which three bridges to install sensors on to get the best prediction of overall traffic across all four bridges. This will involve analyzing bike traffic data across several bridges in New York City. By installing sensors on these bridges, we can estimate overall traffic across all the bridges and determine which bridges have the highest volume of bike traffic.

2.2 Temperature Rider Correlation Prediction:

The objective of this task is to predict the total number of bicyclists on high-traffic days using the next day's weather forecast (low/high temperature and precipitation). This will involve analyzing weather data to identify patterns in bike traffic that are correlated with weather conditions. By using machine learning algorithms to predict bike traffic based on weather forecasts, we can help city officials deploy police officers on days with high traffic to hand out citations for helmet law violations.

2.3 Prediction Model:

The objective of this task is to predict the current day (Sunday through Saturday) based on the number of bicyclists on the bridges. This will involve analyzing bike traffic data and identifying patterns in bike traffic that are correlated with days of the week. By being able to predict the current day based on bike traffic data, we can help city officials plan for high-traffic days and improve bike safety.

3. Results:

3.1 Sensor Installation:

Question:

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?

Answer:

The problem statement required a metric to determine priority of which bridge needs to have sensors installed first. The mean was chosen as the metric to determine priority because it told us which bridge has the highest average number of bikers on any given bridge (assuming minimal outliers), so we maximized the sensors' impacts.

A 2x2 subplot was plotted to visualize all the data: (Top Left – Brooklyn Bridge, Top Right – Williamsburg Bridge, Bottom Left – Manhattan Bridge, Bottom Right – Queensboro Bridge). This allowed visualization of the number of riders based on the day's data. Additionally, a bridge class was used to calculate the mean, standard deviation, and variance for each bridge.

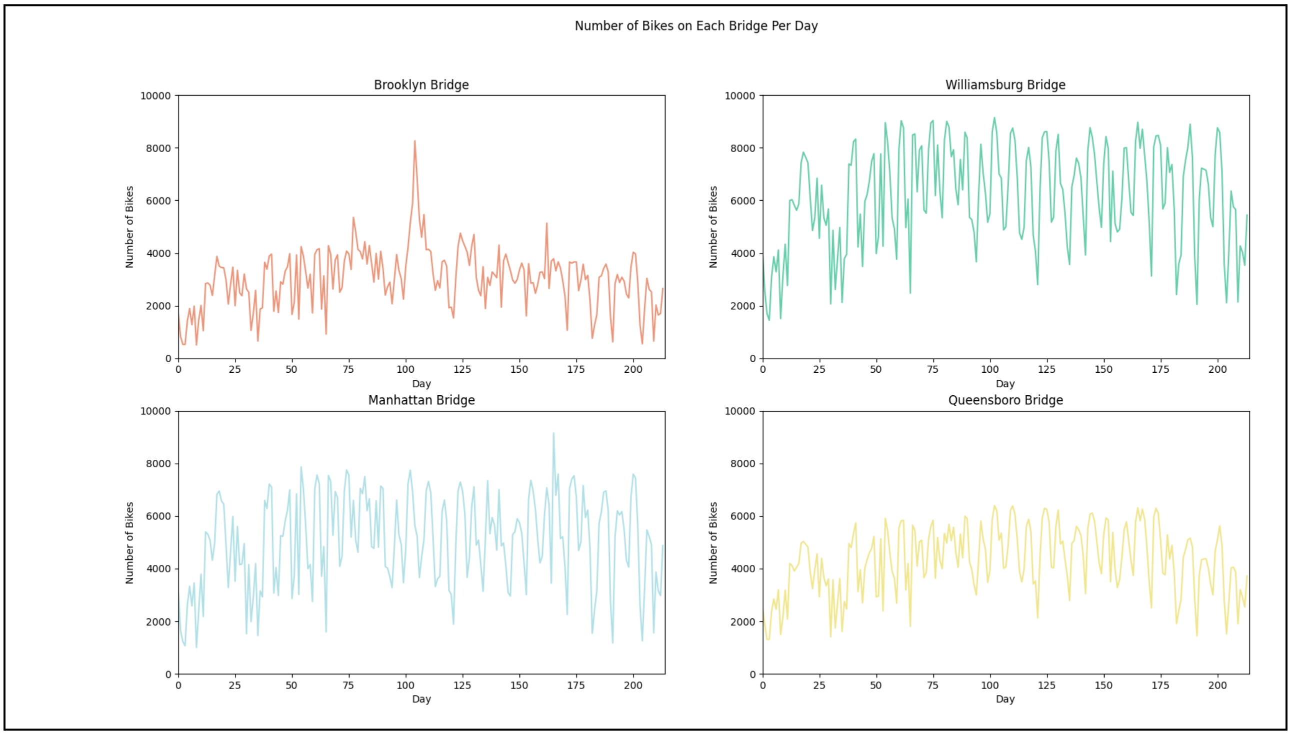
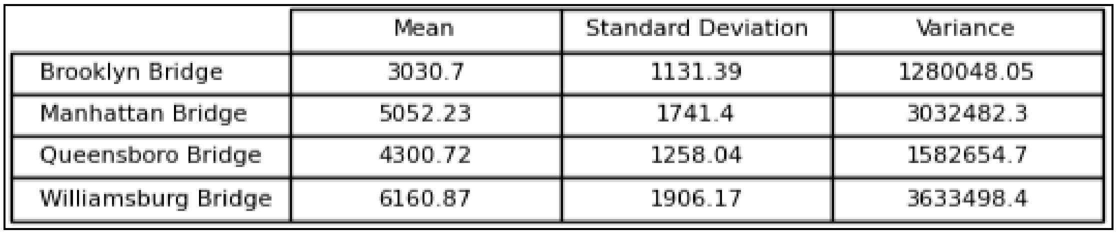


Figure 1: 2x2 Plot of Bridge Bike Riders Per Day on All Four Bridges

As seen by the 2x2 subplot, each bridge has a consistent and predictable variation in its shape. There is a peak in the Brooklyn bridge traffic, but this does not change our results in terms of overall traffic.



	Mean	Standard Deviation	Variance
Brooklyn Bridge	3030.7	1131.39	1280048.05
Manhattan Bridge	5052.23	1741.4	3032482.3
Queensboro Bridge	4300.72	1258.04	1582654.7
Williamsburg Bridge	6160.87	1906.17	3633498.4

Figure 2: Table of Mean, Standard Deviation, and Bridge Data Variance

Additionally, as seen through the second figure, each bridge has a similar standard deviation from the mean with a difference between the maximum standard deviation (Williamsburg Bridge) and the smallest standard deviation

(Brooklyn Bridge) being 774.78. Based on the mean rider population of each bridge (3030.70 for Brooklyn Bridge, 5052.23 for Manhattan Bridge, 4300.72 for Queensboro bridge, and 6160.87 for Williamsburg Bridge), and keeping in mind the large variance, we choose to put sensors in the three average highest-used bridges: Williamsburg, Manhattan, and Queensboro respectively.

However, it is important to understand that because the variance is so high, it does not guarantee on any given day that these bridges will have the most traffic.

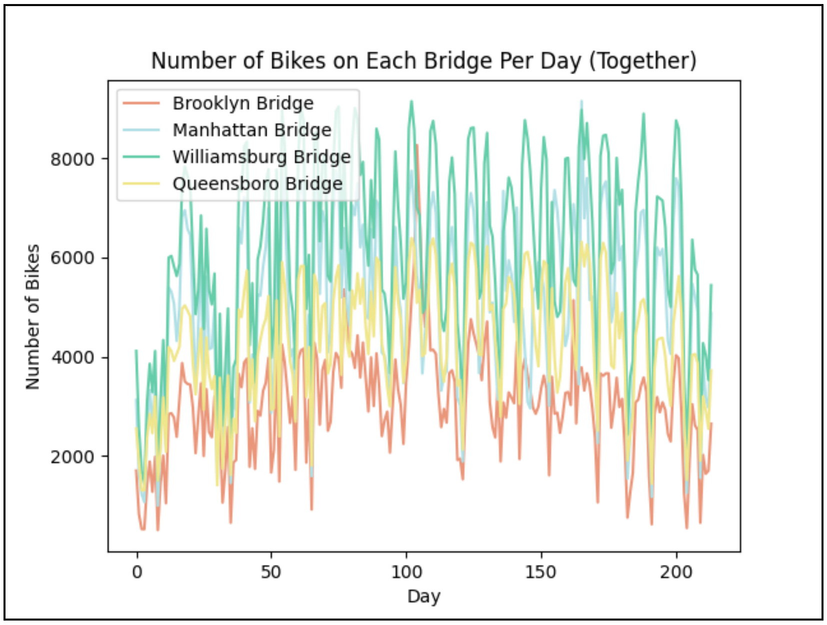


Figure 3: Plot of all Bike Traffic Across all Bridges

This plot aims to show the clear distinction between traffic on all four of the bridges, including how peaks coincide and how certain bridges have a distinct traffic increase over others. This plot also aims to show the similarity in ranges of all four bridges. There are points for all four bridges where the data intercepts one another, meaning that going strictly by the range is not an effective method of evaluating which bridges to put sensors on. However, the sample means would all be different enough where it constitutes a good indicator of traffic on each bridge.

3.2 Temperature Rider Correlation:

Question:

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?

Answer:

We have devised and implemented a model to attempt to predict the number of bikers given the next day’s weather forecast. To accomplish this, the bike traffic as a function of temperature (for both higher and lower temperature ranges) was plotted in the form of a scatter plot.

It was observed that there was a significant increase in the number of bikers with an increase in temperature, irrespective of picking the lower or higher temperature range. This observation is supported by a model (presented in the results below) that exhibits a correlation between the two variables.

To justify this conclusion, a linear regression model was fit to model the number of bikers based on temperature. The results of the model were then printed in the output of the function, which includes model coefficients, R^2, and standard error. As a result, a prediction model was obtained that can estimate the number of bikers based on temperature. The model results are shown in the Code Output section of this report.

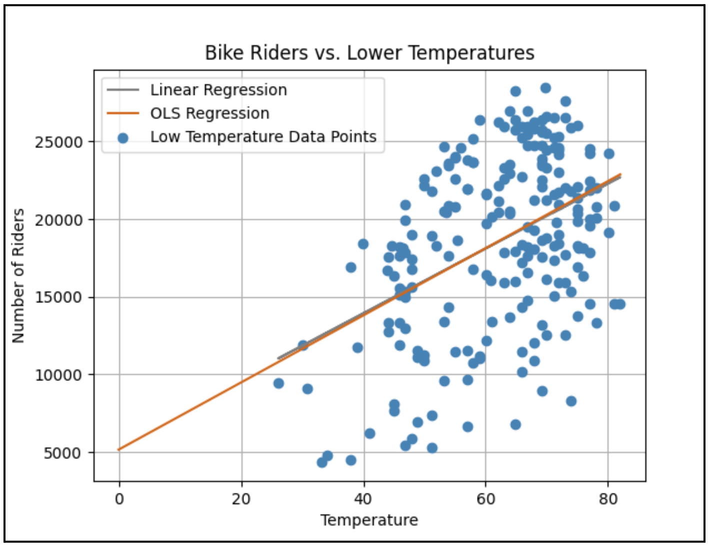


Figure 4: Scatter Plot of Bike Riders Versus Lower Temperature Ranges

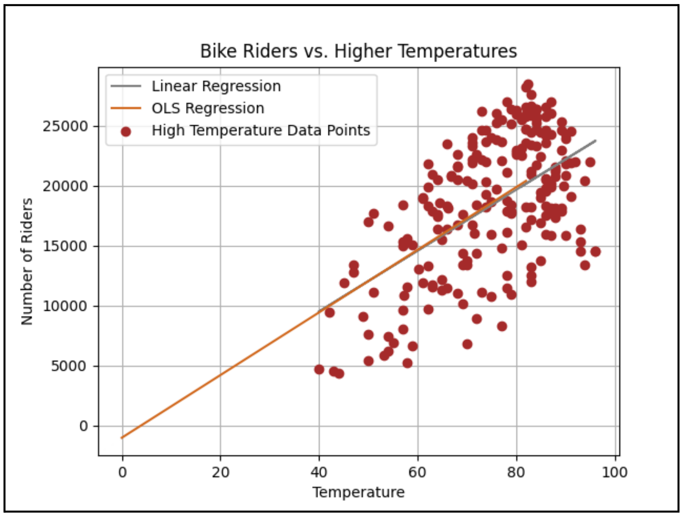


Figure 5: Scatter Plot of Bike Riders Versus Higher Temperature Ranges

As seen by the above plots, there is a positive and increasing correlation between rising temperature and increased bike traffic. There is a noticeable rise in traffic as temperature approaches 80-90 degrees, which could indicate riders prefer these specific ranges. However, as seen through our R^2 value being so low (0.195), that the discussed correlation is not as strong as expected. This could be due to a multitude of reasons. We believe however, it is most likely caused by the large variance in this data. This variance results in data points being distant from the line of best fit, which negatively impacts R^2.

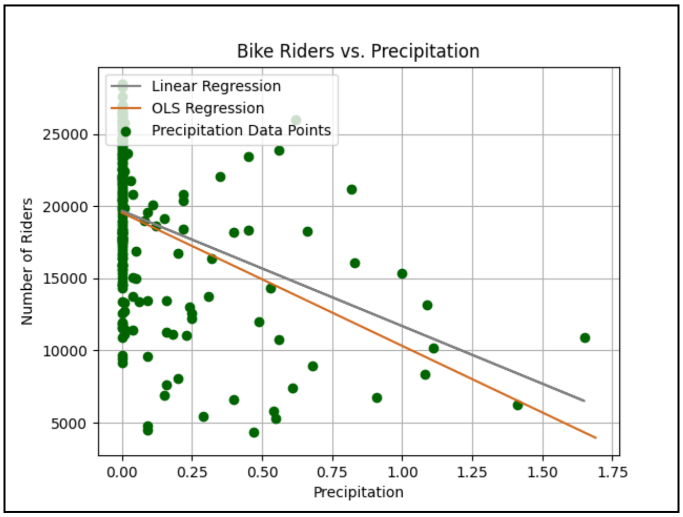
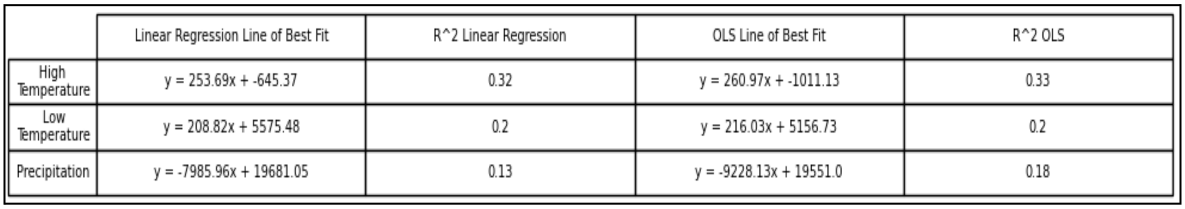


Figure 6: Scatter Plot of Bike Riders Versus Precipitation

This plot shows an inversely-proportional relationship between precipitation and bike traffic. Notably, there is a large stack of data points for zero precipitation (because clear weather is simply very common), which has a negative effect on the accuracy and R^2 of linear regression and OLS models.



	Linear Regression Line of Best Fit	R^2 Linear Regression	OLS Line of Best Fit	R^2 OLS
High Temperature	y = 253.69x + -645.37	0.32	y = 260.97x + -1011.13	0.33
Low Temperature	y = 208.82x + 5575.48	0.2	y = 216.03x + 5156.73	0.2
Precipitation	y = -7985.96x + 19681.05	0.13	y = -9228.13x + 19551.0	0.18

Figure 7: Regression Statistics (Line of Best Fit and R^2)

The above table summarizes the R^2 values and line of best fit for both OLS and Linear Regression. These statistics were measured for high temperature, low temperature, and precipitation. OLS tends to have a marginally higher R^2 in this example, and the line of best fit equations are shown as well.

To summarize, as seen by the low (but not too low) R^2 value for both types of regression, it is not pragmatic to numerically predict the total number of riders the next day based on the weather, but it is possible to estimate a relationship between the two. For instance, if an officer noticed that there will be a high temperature the next day, they would not be guaranteed to have a larger turnout, but it would be advisable to enforce a stricter policy that day based on the positive relationship between temperature and bike traffic. The inverse relationship applies to the amount of expected precipitation. While not entirely rigorous, it could be advisable for the officer to enforce more strongly when noticing a day with less/zero precipitation.

To summarize, while the city cannot empirically determine bike traffic based on temperature and precipitation, it is possible (and recommended) to use the noted correlations displayed in the line of best fit to determine a suitable deployment of officers on a given day.

3.3 Prediction Model:

Question:

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?

Answer:

A model has been developed that intends to predict the most likely day of the week based on the number of riders on a given day. A prediction algorithm was created using a method similar to K-Nearest Neighbors (KNN), but instead of classifying multiple nearest neighbors, the focus was on finding all distances to neighbors, and using the nearest point as the prediction bias. Once the prediction point was obtained, it was possible to trace back and determine the day of the week.

Notably, the results of this algorithm are quite poor. We devised a testing procedure where we chose 100 random biker population samples, ran the day prediction, and compared this prediction to the expected day. This resulted in a low accuracy of around 12-25%, which typically underperforms a random classifier (1/7 odds), and at best slightly outperforms a random classifier. As a result, we conclude there is NOT a reliable way of predicting the day based on the number of bikers.

We hypothesize this low accuracy is a result of the large variance in the given traffic data. This variance causes each day to encapsulate essentially the entire possible range of traffic density. As a result, if the sample is 15000 bikers, this realistically could be any day of the week. The model will still return the day with the mean closest to 15000, however because there is only one metric to classify the day with, this model can only return a poor guess.

This algorithm is illustrated in the diagram below:

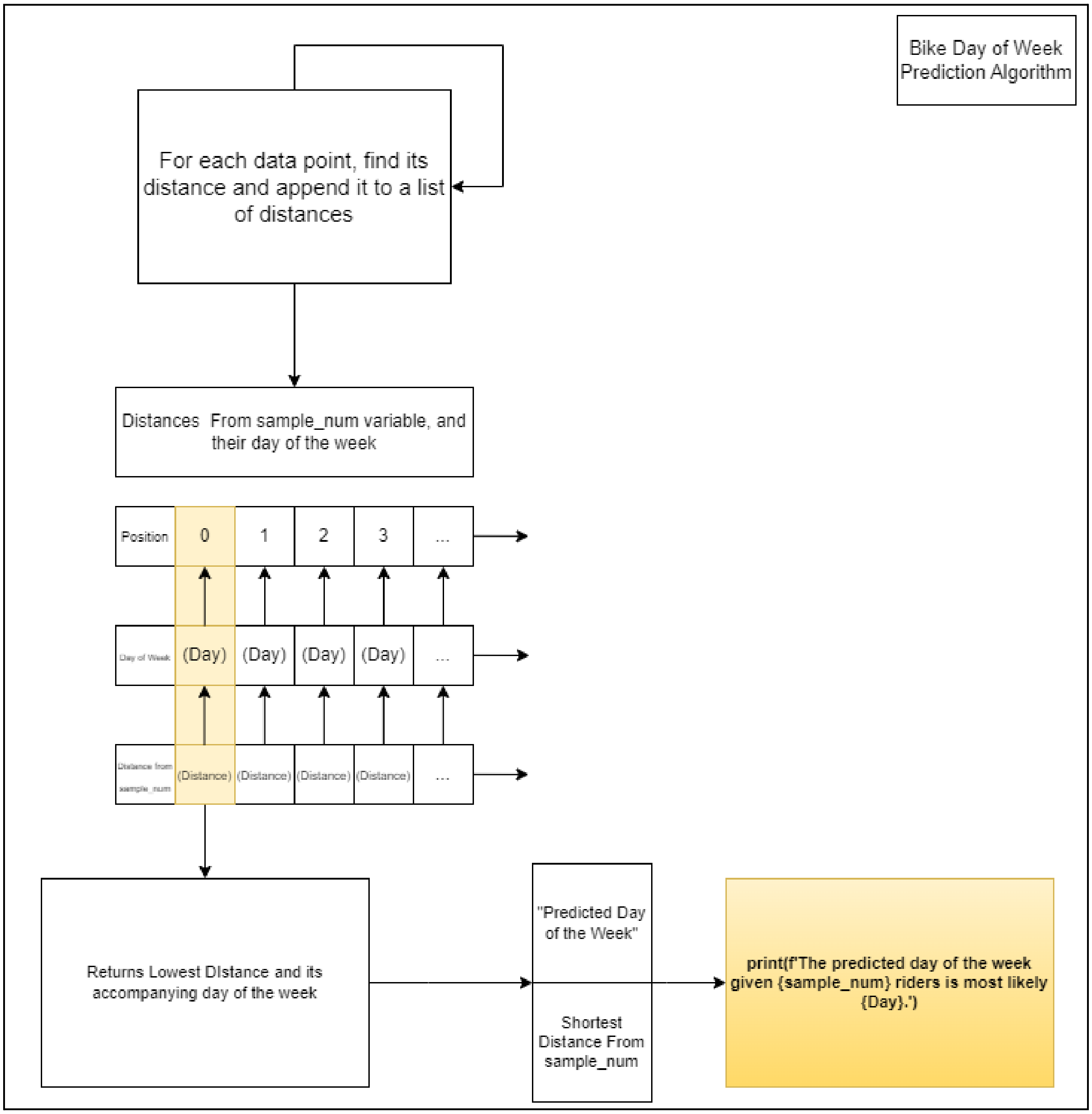

Bike Day of Week Prediction Algorithm
For each data point, find its distance and append it to a list of distances
Distances From sample_num variable, and their day of the week
Position
0
1
2
3
...
Day of Week
(Day)
(Day)
(Day)
(Day)
...
Distance from sample_num
(Distance)
(Distance)
(Distance)
(Distance)
...
Returns Lowest DIstance and its accompanying day of the week
"Predicted Day of the Week"
Shortest Distance From sample_num
print(f'The predicted day of the week given {sample_num} riders is most likely {Day}.')

Figure 8: Prediction Algorithm Workflow

The above diagram details the prediction algorithm workflow. Importantly, the prediction algorithm did not account for outliers or data variance of each day, which could result in inaccurate predictions for extremely high or low numbers of riders. We also considered a few alternative methods, including a linear regression model. However, it is important to note that linear regression is not a classification algorithm. It is a regression algorithm used to predict the value of a continuous dependent variable based on one or more independent variables. As a result, linear regression is a poor model choice.

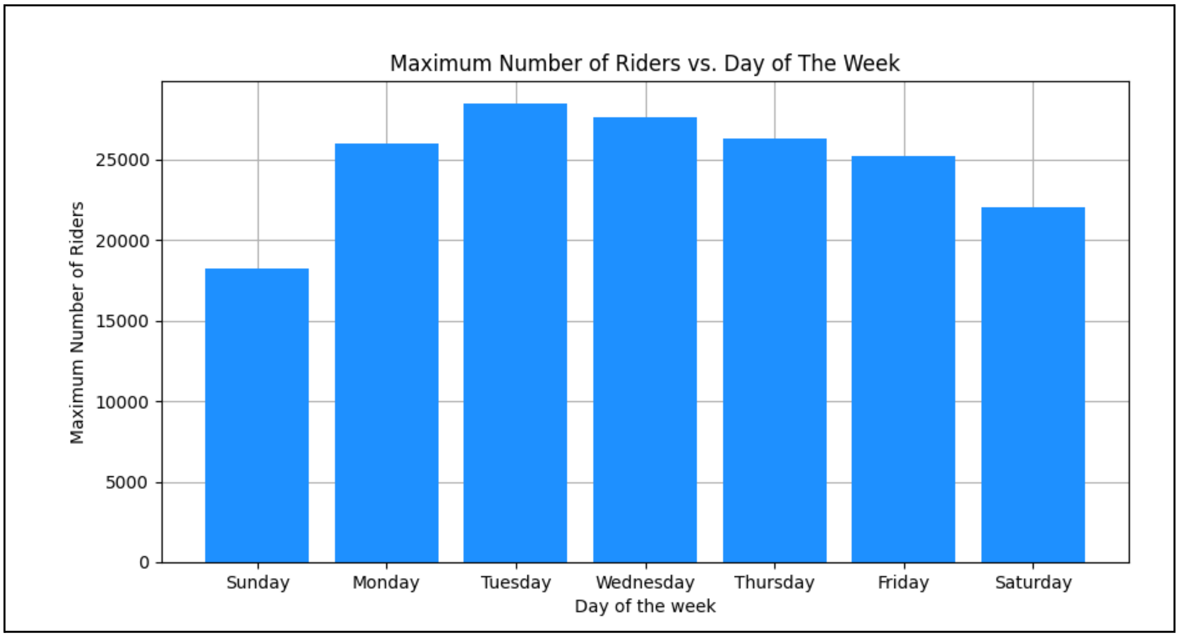


Figure 9: Bar Plot of Maximum Bike Riders Versus Day of the Week

This plot demonstrates how Tuesday has the maximum number of bikers out of any day of the week. Sunday has the minimum.

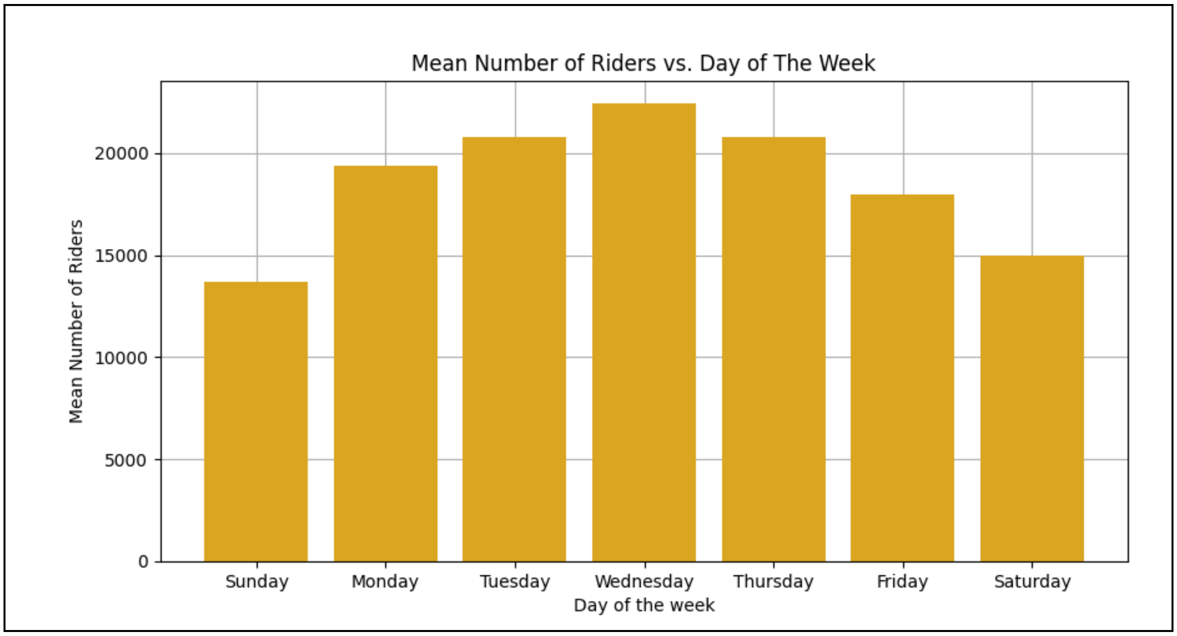


Figure 10: Bar Plot of Mean Bike Riders Versus Day of the Week

This plot shows how Wednesday has the highest mean number of bikers, and again Sunday has the minimum. However, the means are notably close, contributing to the inefficacy of the prediction algorithms

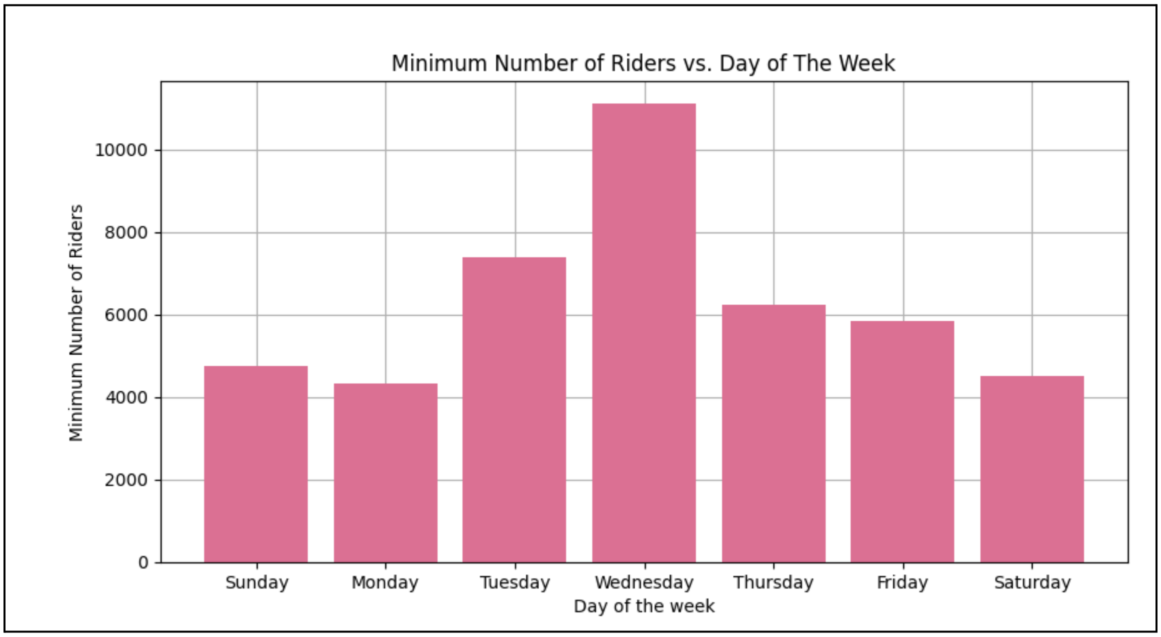


Figure 11: Bar Plot of Minimum Bike Riders Versus Day of the Week

Wednesday has by far the highest minimum number of bike riders on a given day of the week, with Monday falling quite far behind. This pushes Wednesday’s mean traffic very high.

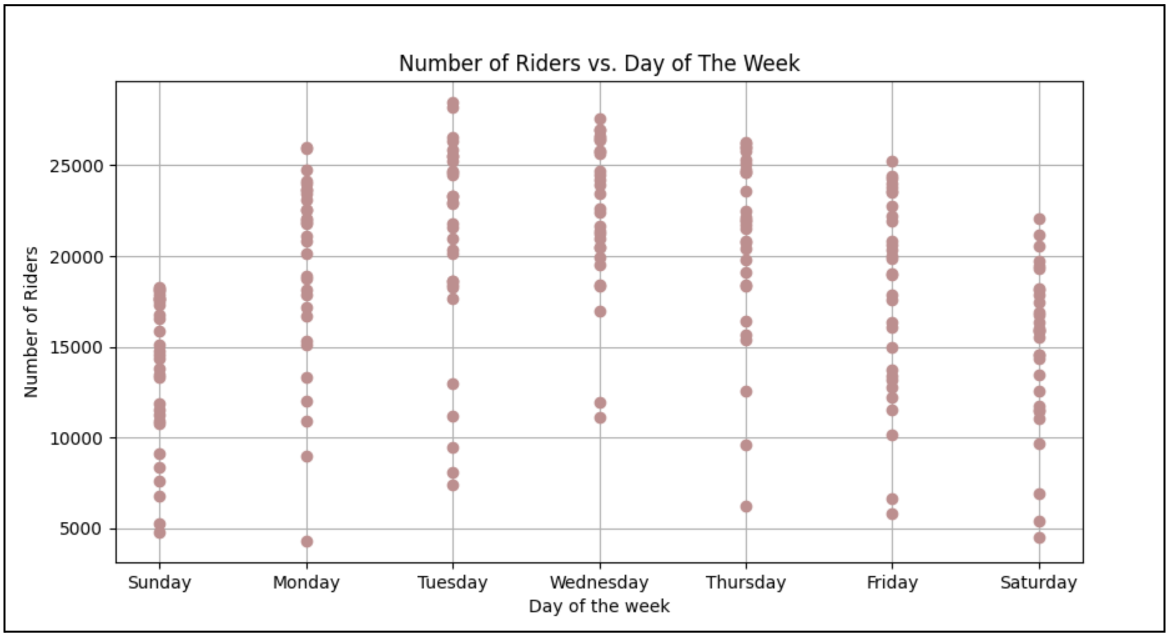


Figure 12: Scatter Plot of Bike Riders Versus Day of the Week

This scatter plot is the foundation of our analysis and conclusions. It aims to show the distribution of bike riders for each given day of the week. It is clear that Sunday and Saturday have a very high variance in bike traffic, while Tuesday and Thursday are slightly tighter (less variance) and are focused towards a higher number of bikers. This contributes to Tuesday and Wednesday having the highest mean number of bikers, and the large variance contributes heavily to the inaccuracy of our prediction model.

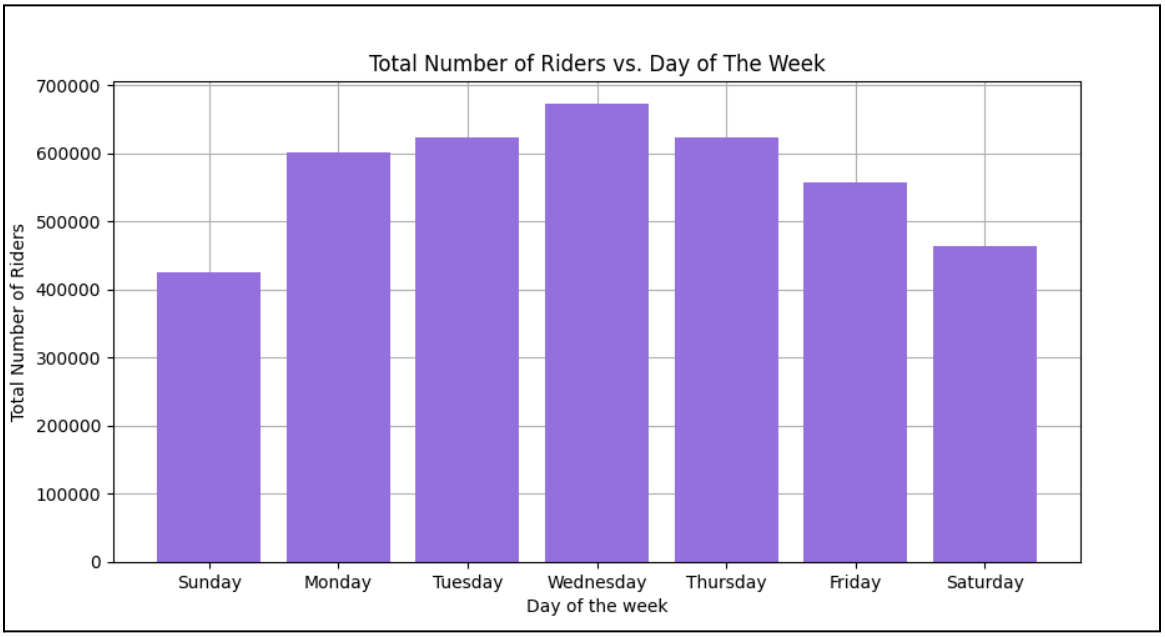


Figure 13: Bar Plot of Total Bike Riders Versus Day of the Week

Finally, this plot shows the total number of bikers throughout all samples for each day of the week. The distributions are very close, however there is a marked trend of weekdays to have a higher number of bikers, with Wednesday again having the highest number of bikers.

Analyzing the displayed trends, there are obvious outliers for each day of the week, which affect the maximum number in each one of the graphs. Taking the mean of each particular day aims to reduce the effect of outliers, and allow for achieving results more representative of the populations. These outliers can have heavy impacts on our prediction model, along with the high variance. Additionally, as seen through the various bar plots, the data for each day is varied to the point of affecting the prediction results. For instance, while Wednesday may have the largest minimum value of bike riders, it does not have the largest maximum (Tuesday). There is also a large overlap between each day and the data within their ranges of maximum number of riders and minimum number of riders. This makes it impossible to create an accurate prediction model, as the same number of riders could be attributed to any number of days as long as it lies within an acceptable range.

4. Conclusion:

In this lab report, we have analyzed bike traffic data across a number of bridges in New York City. We used Python to load the data, sanitize it, and perform statistical analysis. We then answered three questions:

1. Which bridges should we install sensors on to get the best prediction of overall traffic?
2. Can the city administration use the next day's weather forecast to predict the total number of bicyclists that day?
3. Can we use this data to predict what day (Monday to Sunday) is today based on the number of bicyclists on the bridges?

We found that the best way to predict overall traffic is to install sensors on the Brooklyn Bridge, Manhattan Bridge, and Williamsburg Bridge. These bridges are the most heavily utilized on average, and their data will give us the most accurate picture of overall traffic.

We also found that the city administration can use the next day's weather forecast to *qualitatively* predict bike traffic on that day. The number of bicyclists is typically lower on colder days and higher on warmer days. In addition, precipitation has an inverse effect on bike traffic, with the heaviest traffic occurring when there is no precipitation. However, the weather is not the only factor that affects bike traffic. Other factors, such as the time of day, the day of the week, and special events, can also potentially play a role. These are beyond the scope of this analysis.

Finally, we found that we cannot feasibly use the given data to predict the current day based on the number of bicyclists on the bridges. The number of bicyclists is typically highest on weekdays and lowest on weekends, however the large variance in data makes creating an accurate model impossible.

Overall, we found that data science methods can be used to analyze bridge data in a number of ways. This data can be used to improve traffic flow, enforce helmet laws, and attempt to predict the day of the week and bike traffic based on weather.

In addition to the specific questions that we answered in this lab report, there are many other ways that data science can be used to analyze bridge data. For example, data science can be used to:

- Identify bottlenecks in traffic flow and recommend solutions
- Monitor the condition of bridges and identify potential problems
- Plan for future bridge construction and maintenance

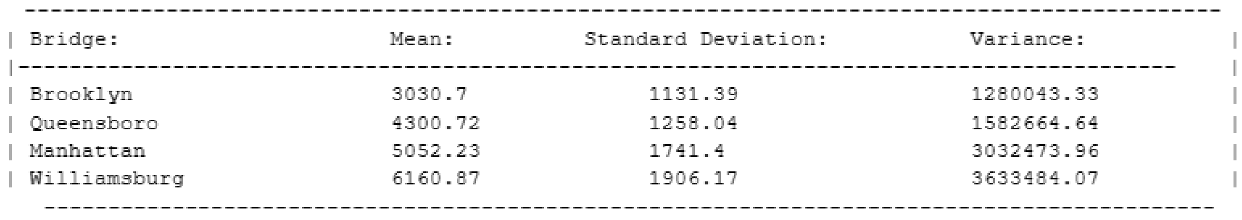
Data science is a powerful tool that can be used to improve the safety and efficiency of bridges. By using data science, we can make our bridges and infrastructure safer and more reliable for everyone.

5. Output of Code When Ran (Without Graphs):

(Output has been reformatted to look better for this paper)

```
Type which question you would like to see answered (1 for question 1,
2 for question 2, 3 for question 3, or nothing for all, then press
enter:
1, 2, and 3... (All Questions)

  -----------------------------------------------------------------
  ---------------------- Data For Question 1 ----------------------
  -----------------------------------------------------------------
```



```
 ------------------------------------------------------------------------------------------
| Bridge:                   Mean:         Standard Deviation:         Variance:            |
|------------------------------------------------------------------------------------      |
| Brooklyn                  3030.7             1131.39                1280043.33           |
| Queensboro                4300.72            1258.04                1582664.64           |
| Manhattan                 5052.23            1741.4                 3032473.96           |
| Williamsburg              6160.87            1906.17                3633484.07           |
   ----------------------------------------------------------------------------------------
```

```
Three Bridges With The Highest Average Number of Riders (In
Descending Order):

1. Williamsburg Bridge:    Average = 6160.87 (People per Day)

2. Manhattan Bridge:       Average = 5052.23 (People per Day)

3. Queensboro Bridge:      Average = 4300.72 (People per Day)
```

```
  ------------------------------------------------------------------
  ----------------------- Data For Question 2 ----------------------
  ------------------------------------------------------------------

                    Results of the Linear Regression Model:
==============================================================
Line of Best Fit for High Temperatures: y = 253.69x + -645.37
R-Squared Value for High Temperatures: 0.32
--------------------------------------------

Line of Best Fit for Low Temperatures: y = 208.82x + 5575.48
R-Squared Value for Low Temperatures: 0.2
--------------------------------------------

Line of Best Fit for Precipitation: y = -7985.96x + 19681.05
R-Squared Value for Precipitation: 0.13
==============================================================

---------------------------------------------------------------------------
---------------------------------------------------------------------------

                      Results of the OLS Regression Model:
==============================================================
Line of Best Fit for High Temperatures: y = 260.97x + -1011.13
R-Squared Value for High Temperatures: 0.33
--------------------------------------------

Line of Best Fit for Low Temperatures: y = 216.03x + 5156.73
R-Squared Value for Low Temperatures: 0.2
--------------------------------------------

Line of Best Fit for Precipitation: y = -9228.13x + 19551.0
R-Squared Value for Precipitation: 0.18
==============================================================

  ------------------------------------------------------------------
  ----------------------- Data For Question 3 ----------------------
  ------------------------------------------------------------------

The predicted day of the week given 20875 riders is most likely
Friday.

Out of 100 random samples, the predicted day matched the real day 26
times.

  ------------------------------------------------------------------
```

6. Works Cited (Used For Research):

Beazley, David M. Python Essential Reference. Sams, 2008.

Journal of Bridge Engineering, www.scimagojr.com/journalsearch.php?q=16521&amp;tip=sid.

S., Mauriz. "(Simple) Linear Regression and OLS: Introduction to the Theory." Medium, Towards Data Science, 25 May 2020, towardsdatascience.com/simple-linear-regression-and-ols-introduction-to-the-theory-1b48f7c69867.

Hanson, Susan and Perry O. Hanson. "EVALUATING THE IMPACT OF WEATHER ON BICYCLE USE." Transportation Research Record 629 (1977): 43-48.

Analysis of Weather Impacts on Traffic Flow in Metropolitan Washington, DC, https://rosap.ntl.bts.gov/view/dot/51762

Daraei, Sara, et al. "A Data-Driven Approach for Assessing Biking Safety in Cities - EPJ Data Science." SpringerOpen, Springer Berlin Heidelberg, 3 Mar. 2021, epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-021-00265-y.