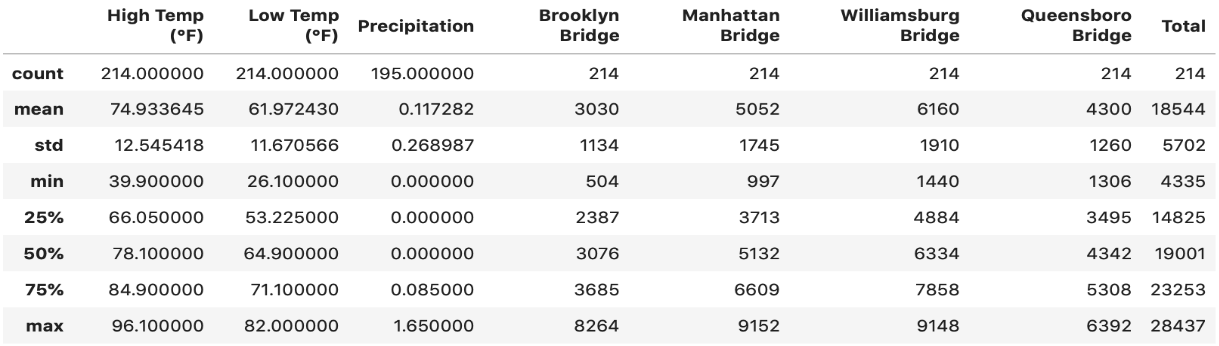
Name: Nomair Bhatti
GitHub Username: bhattin82
Path 1: Bike Traffic

Detailed Data Analysis of Bike Traffic in New York City

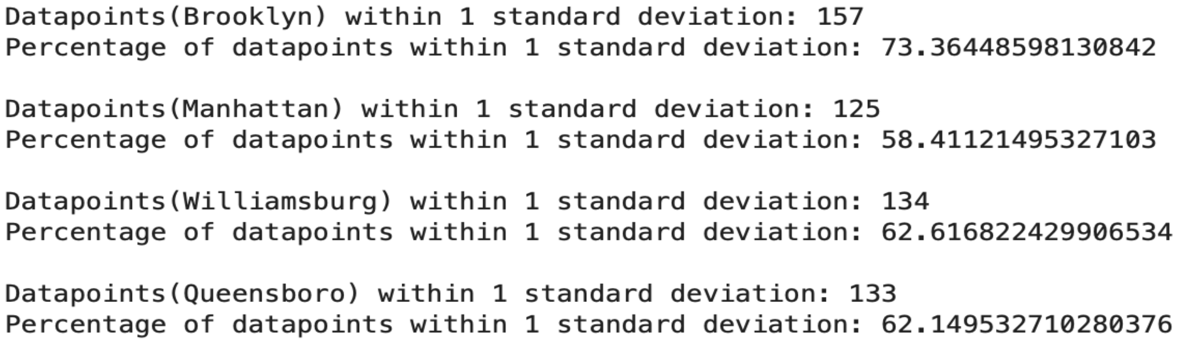
The bicycle dataset consists of 214 rows and 10 columns. The following are the columns: Date, Day, High Temperature (F), Low Temperature (F), Precipitation, Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge, Queensboro Bridge and Total. Each row is a record for that given day. The data collected represents bike usage in New York city only. It is given as a .csv file.

I was presented with the information that there is a need to monitor overall traffic on the bridges of New York City with the help of sensors. However, the budget is limited and only three sensors can be afforded. I had to determine the best three bridges to install those sensors on. I calculated the mean, standard deviation, maximum and minimum values of the cyclists at each bridge and overall (figure 1). I implemented the concept of the normal (gaussian) distribution. For cyclists at each bridge, I defined a range ([mean- standard deviation, mean+ standard deviation]) to determine the number of points that were within one standard deviation (sixty eight percent) of the mean. Later, I took the number of data points and converted them to a percentage for simplicity. Brooklyn, Manhattan, Williamsburg and Queensboro bridge had the following percentage of data points (figure 2) within one standard deviation: 73.36 percent (157 points), 58.41 percent (125 points), 62.62 percent (134 points) and 62.14 percent (133 points). To get a better insight of the data provided, I created bar graphs and histograms. As seen in figure 3, the bar graph shows the average number of cyclists on each bridge. According to it, Williamsburg has the highest number of cyclists on average, followed by Manhattan, Queensboro and then Brooklyn. If I base my decision solely on this bar graph, I will go with Williamsburg, Manhattan and Queensboro. However, I believe that the standard deviation is an imperative measurement, and it cannot be overlooked. I would assign more weight to the data being closer to the mean. This would mean that the data is more consistent. Therefore, I conclude that sensors should be installed on Brooklyn, Williamsburg and Queensboro bridge.



	High Temp (°F)	Low Temp (°F)	Precipitation	Brooklyn Bridge	Manhattan Bridge	Williamsburg Bridge	Queensboro Bridge	Total
count	214.000000	214.000000	195.000000	214	214	214	214	214
mean	74.933645	61.972430	0.117282	3030	5052	6160	4300	18544
std	12.545418	11.670566	0.268987	1134	1745	1910	1260	5702
min	39.900000	26.100000	0.000000	504	997	1440	1306	4335
25%	66.050000	53.225000	0.000000	2387	3713	4884	3495	14825
50%	78.100000	64.900000	0.000000	3076	5132	6334	4342	19001
75%	84.900000	71.100000	0.085000	3685	6609	7858	5308	23253
max	96.100000	82.000000	1.650000	8264	9152	9148	6392	28437

Fig 1. Statistics for the overall data in the file



```
Datapoints(Brooklyn) within 1 standard deviation: 157
Percentage of datapoints within 1 standard deviation: 73.36448598130842

Datapoints(Manhattan) within 1 standard deviation: 125
Percentage of datapoints within 1 standard deviation: 58.41121495327103

Datapoints(Williamsburg) within 1 standard deviation: 134
Percentage of datapoints within 1 standard deviation: 62.61682242990653 4

Datapoints(Queensboro) within 1 standard deviation: 133
Percentage of datapoints within 1 standard deviation: 62.149532710280376
```

Fig 2. Percentage of data points within one standard deviation of the mean

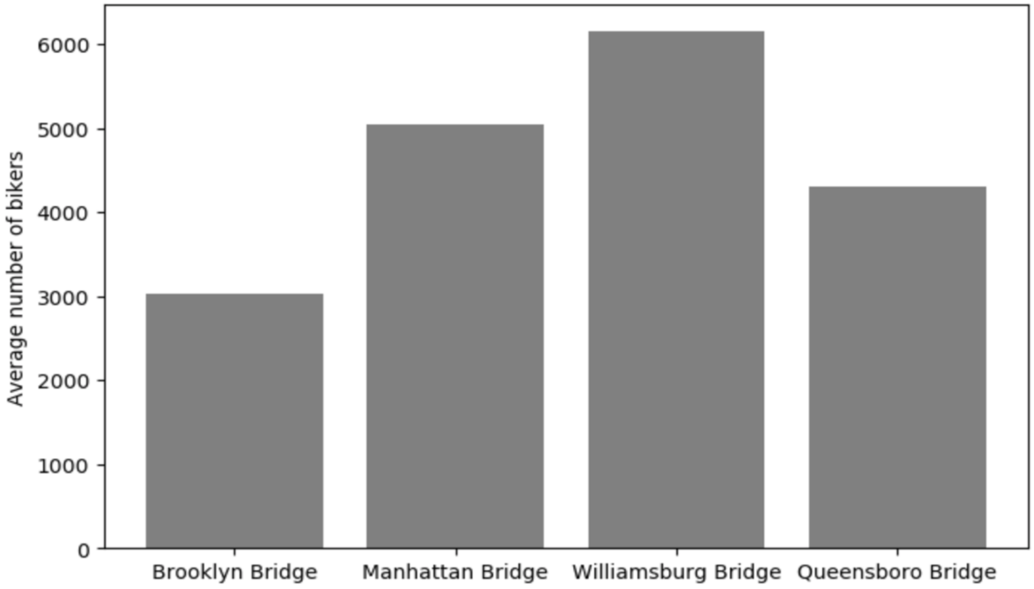


Fig 3. Average number of cyclists on each bridge

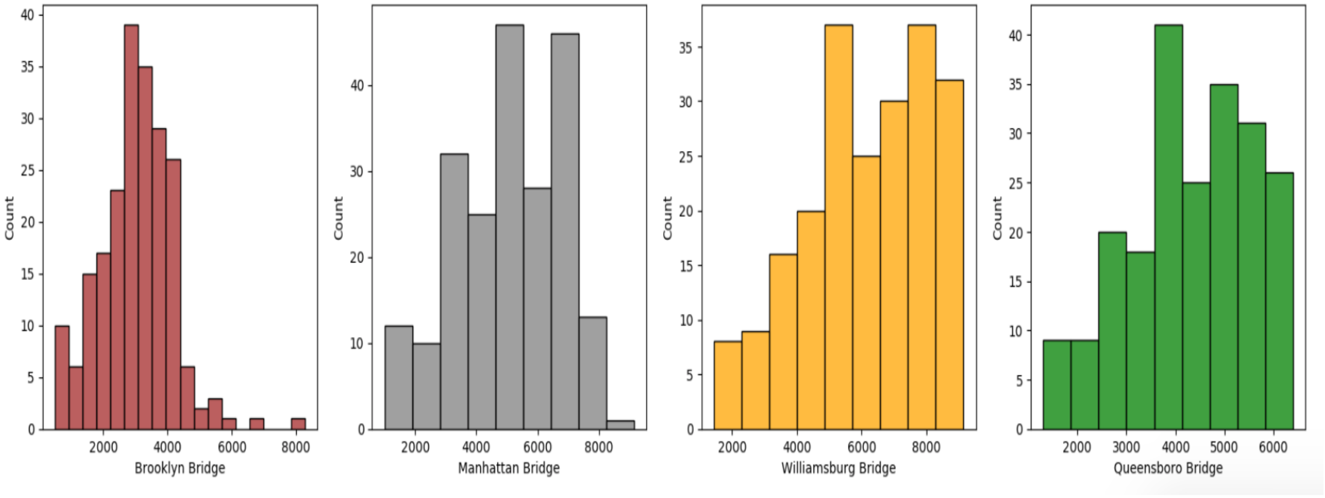
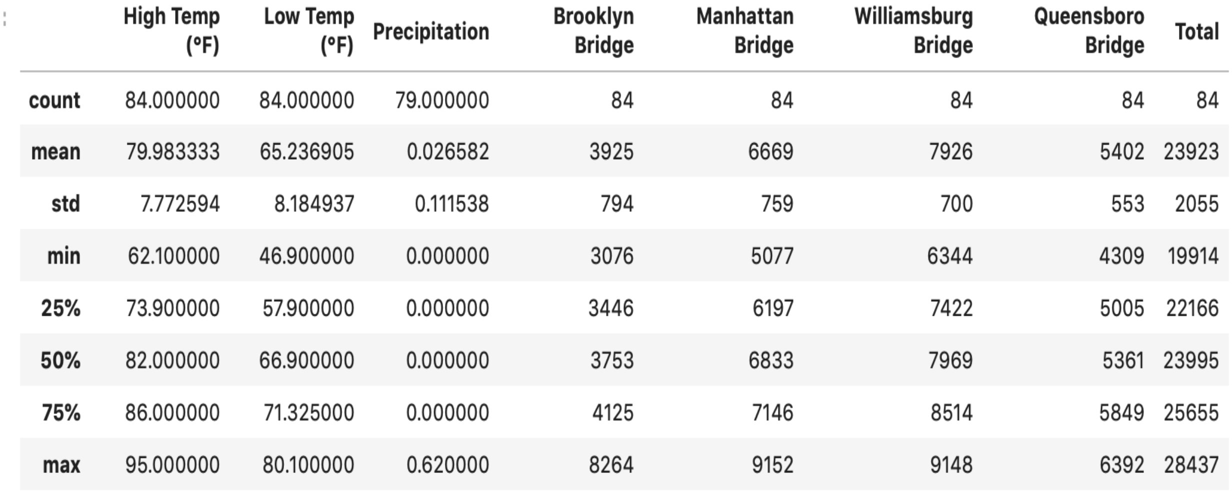


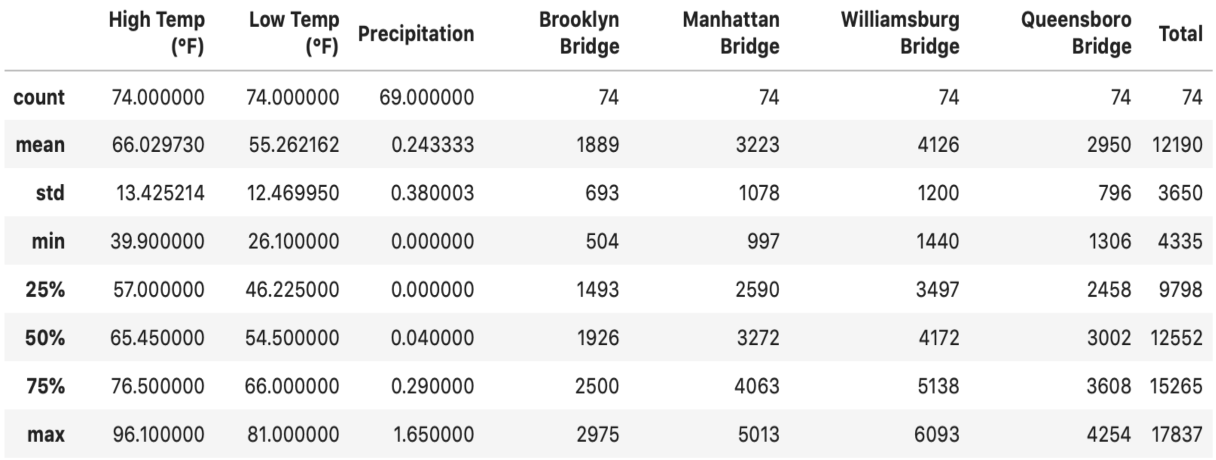
Fig 4. Histograms for cyclists on each bridge

I was also given the information that New York city wants to deploy police officers to hand out citations as they are cracking down on helmet laws. I had to determine if the number of cyclists could be predicted based on the next day's weather forecast. Since I had the mean for the number of cyclists, I applied a condition for each bridge to obtain all the rows where the number of cyclists were greater than or equal to the mean and vice versa. This narrowed down my data, and I was able to see the high and low temperature for those respective days. It was observed that when the number of cyclists were greater than or equal to the mean, the high temperature mean was 79.98 Fahrenheit while the low temperature mean was 65.23 Fahrenheit. This can be seen in figure 5. On the contrary, when the number of cyclists were less than the mean, the high temperature mean was 66.02 Fahrenheit while the low temperature mean was 55.26 Fahrenheit. This can be seen in figure 6. This explains that the number of cyclists increase with an increase in temperature. The data provided was just a sample and to have a representation of the true number of cyclists, I used confidence intervals. To calculate the confidence interval, I used a z-test as the number of data points were greater than thirty. I obtained a z-score of 1.96, and the interval was [23483,24362]. Hence, I am 95 percent confident that the true mean for the total number of cyclists lies between 23483 and 24362 when the temperature is high. Also, I obtained a z-score of 1.96, and the interval was [11358,13021]. Hence, I am 95 percent confident that the true mean for the total number of cyclists lies between 11358 and 13021 when the temperatures are low. Later, the graphs were generated to determine the relationship between the temperature and the number of cyclists. It can be said that the total number of cyclists and the temperature are directly proportional. As the temperature increases, the number of cyclists increase too and vice versa. However, since the correlation values (0.29,0.44) are not so close to 1, the graphs do not account for a lot of the variability in the data. Overall, I would say that the number of cyclists can be predicted based on the next day's weather.



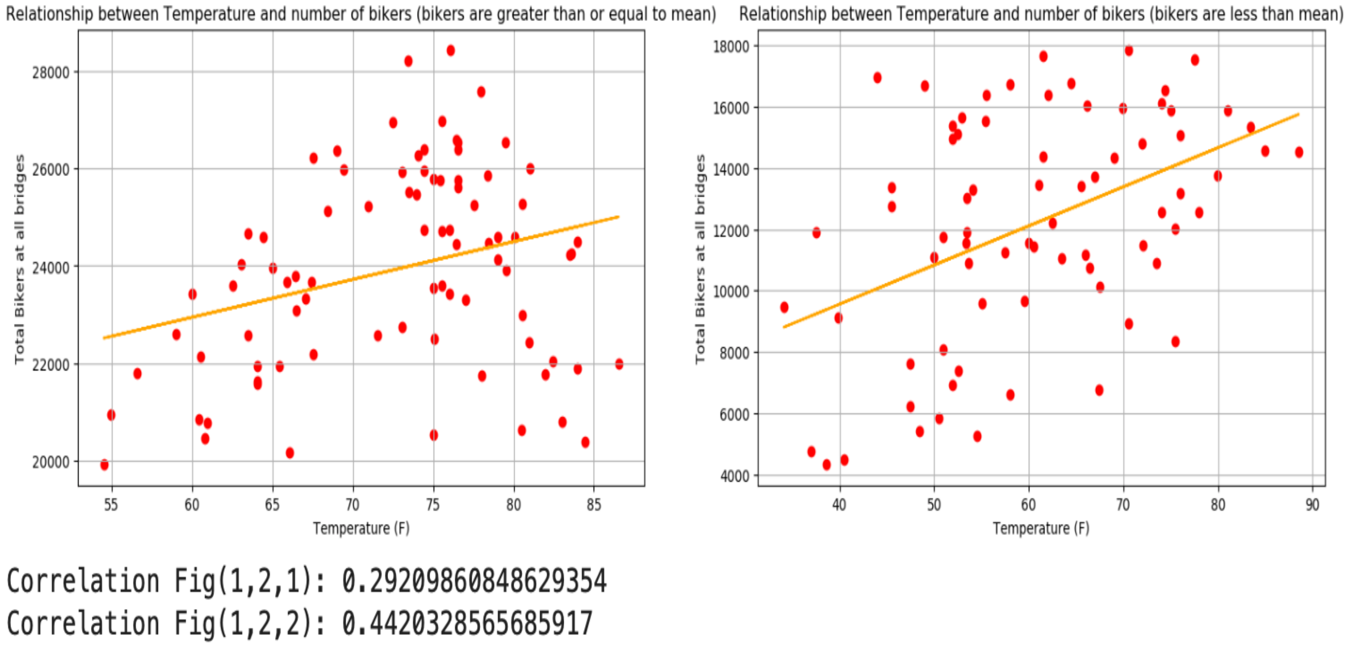
	High Temp (°F)	Low Temp (°F)	Precipitation	Brooklyn Bridge	Manhattan Bridge	Williamsburg Bridge	Queensboro Bridge	Total
count	84.000000	84.000000	79.000000	84	84	84	84	84
mean	79.983333	65.236905	0.026582	3925	6669	7926	5402	23923
std	7.772594	8.184937	0.111538	794	759	700	553	2055
min	62.100000	46.900000	0.000000	3076	5077	6344	4309	19914
25%	73.900000	57.900000	0.000000	3446	6197	7422	5005	22166
50%	82.000000	66.900000	0.000000	3753	6833	7969	5361	23995
75%	86.000000	71.325000	0.000000	4125	7146	8514	5849	25655
max	95.000000	80.100000	0.620000	8264	9152	9148	6392	28437

Fig 5. Statistics when number of cyclists are greater than or equal to the mean (Mean temperatures are higher)



	High Temp (°F)	Low Temp (°F)	Precipitation	Brooklyn Bridge	Manhattan Bridge	Williamsburg Bridge	Queensboro Bridge	Total
count	74.000000	74.000000	69.000000	74	74	74	74	74
mean	66.029730	55.262162	0.243333	1889	3223	4126	2950	12190
std	13.425214	12.469950	0.380003	693	1078	1200	796	3650
min	39.900000	26.100000	0.000000	504	997	1440	1306	4335
25%	57.000000	46.225000	0.000000	1493	2590	3497	2458	9798
50%	65.450000	54.500000	0.040000	1926	3272	4172	3002	12552
75%	76.500000	66.000000	0.290000	2500	4063	5138	3608	15265
max	96.100000	81.000000	1.650000	2975	5013	6093	4254	17837

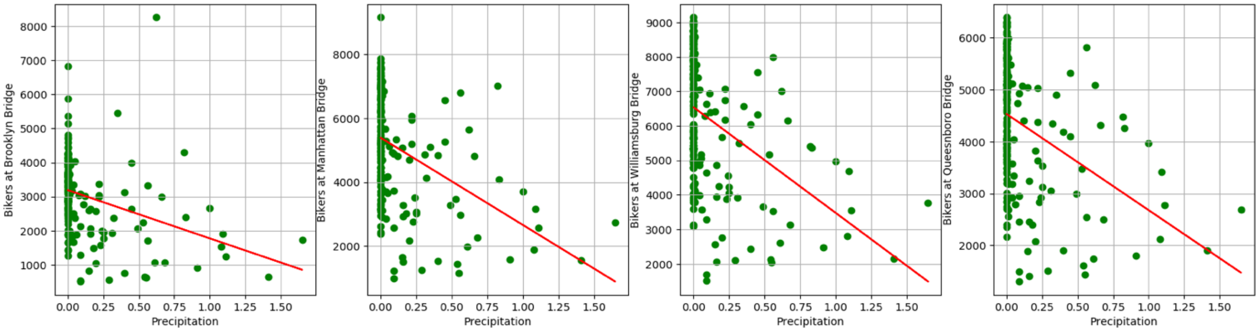
Fig 6. Statistics when number of cyclists are less than the mean (Mean Temperatures are lower)



```
Correlation Fig(1,2,1): 0.29209860848629354
Correlation Fig(1,2,2): 0.4420328565685917
```

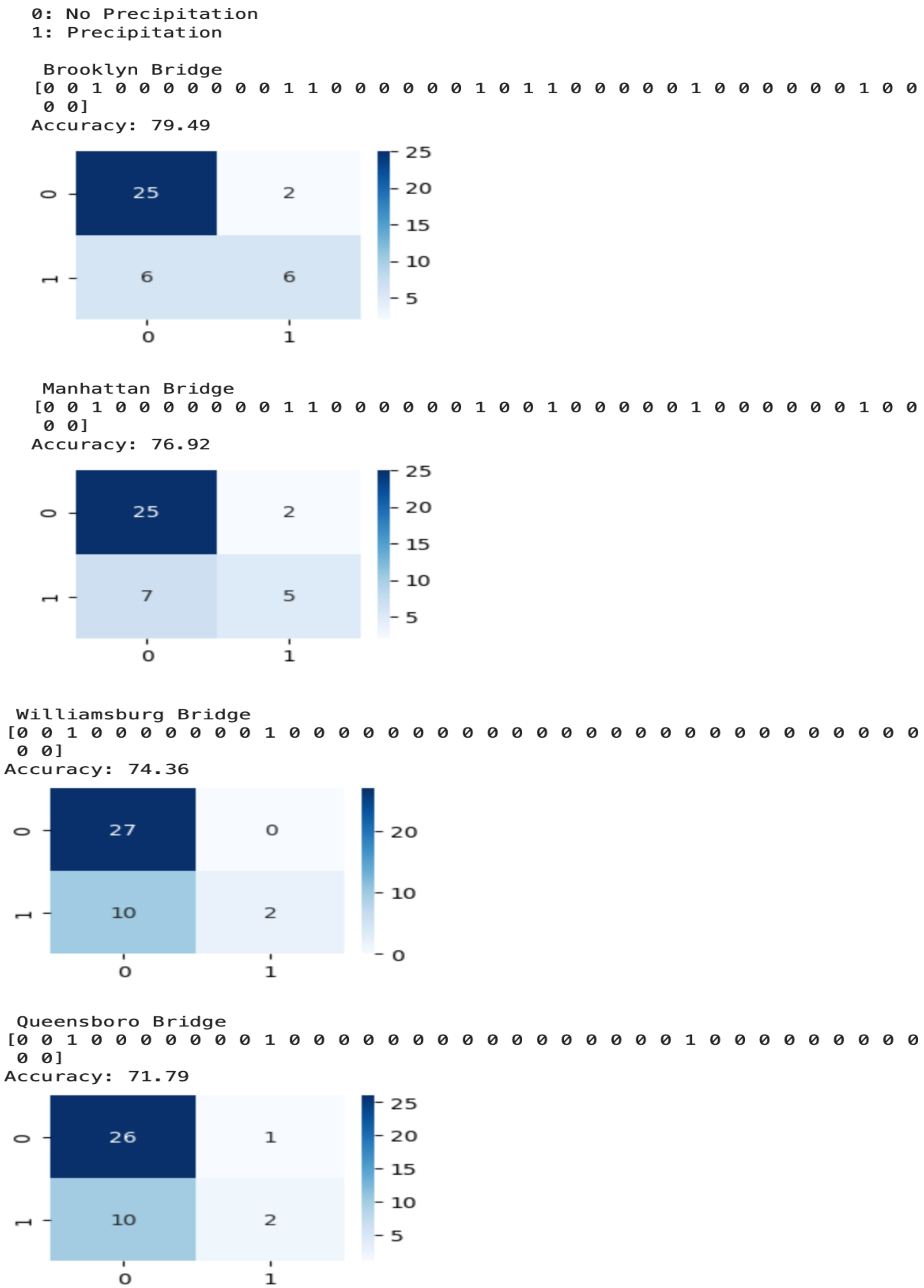
Fig 7. Regression Plots to explain the relationship between temperature and number of cyclists

Lastly, I had to determine if it was raining based on the number of cyclists on the bridges. I used various data science concepts to solve this. I began with linear regression. The plots were generated to determine the relationship between precipitation and the number of cyclists at each bridge. As it can be seen in figure 8, the relationship is inversely proportional. As precipitation increases, the number of cyclists decrease and vice versa. Later, I calculated the linear equation for each plot. The precipitation can be calculated by plugging values in the equation (figure 8). I also used Logistic Regression. This regression technique is an S-shaped curve. Since the precipitation data was in decimals, I had to convert them to integers, so the model was able to understand them (otherwise continuous data error). Hence, I said that all precipitation values that were greater than zero, they were assigned a value of one and vice versa. Then, the model was trained, and the test data assigned to it was twenty percent. Later, it was tested, and the output can be seen in figure 9. For each bridge, a list was generated that had binary values. A value of zero indicates that no precipitation occurred and a value of one indicates that precipitation occurred. The model accuracy for Brooklyn, Manhattan, Williamsburg and Queensboro were as follows: 79.49 percent, 76.92 percent, 74.36 percent and 71.79 percent. Lastly, I used K nearest neighbors. It is another supervised classification algorithm. When a new data point (call it x) is awaiting classification, it looks for the nearest three datapoints (assume neighbors variable is set to 3). For example, let us say that there are two classes: Class A and Class B. The new data point ‘x’ is close to one point from class B and two points from class A. Since class A has more neighbors near to ‘x’, the data point would be classified as class A. Again, the same assumptions were made for this model as for logistic regression. The model accuracy for Brooklyn, Manhattan, Williamsburg and Queensboro were as follows: 69.23 percent, 84.62 percent, 76.92 percent and 71.79 percent. For both logistic regression and K nearest neighbors, the confusion matrices were made. It can be observed that some data was misclassified by both models (False Negative, False Positive). However, they both give us a fair idea of what to expect about precipitation based on the number of cyclists. Overall, it seems that all models did a fair job (linear, logistic, KNN). Therefore, I would say that the data can be used to predict whether it is raining or not based on the number of cyclists.



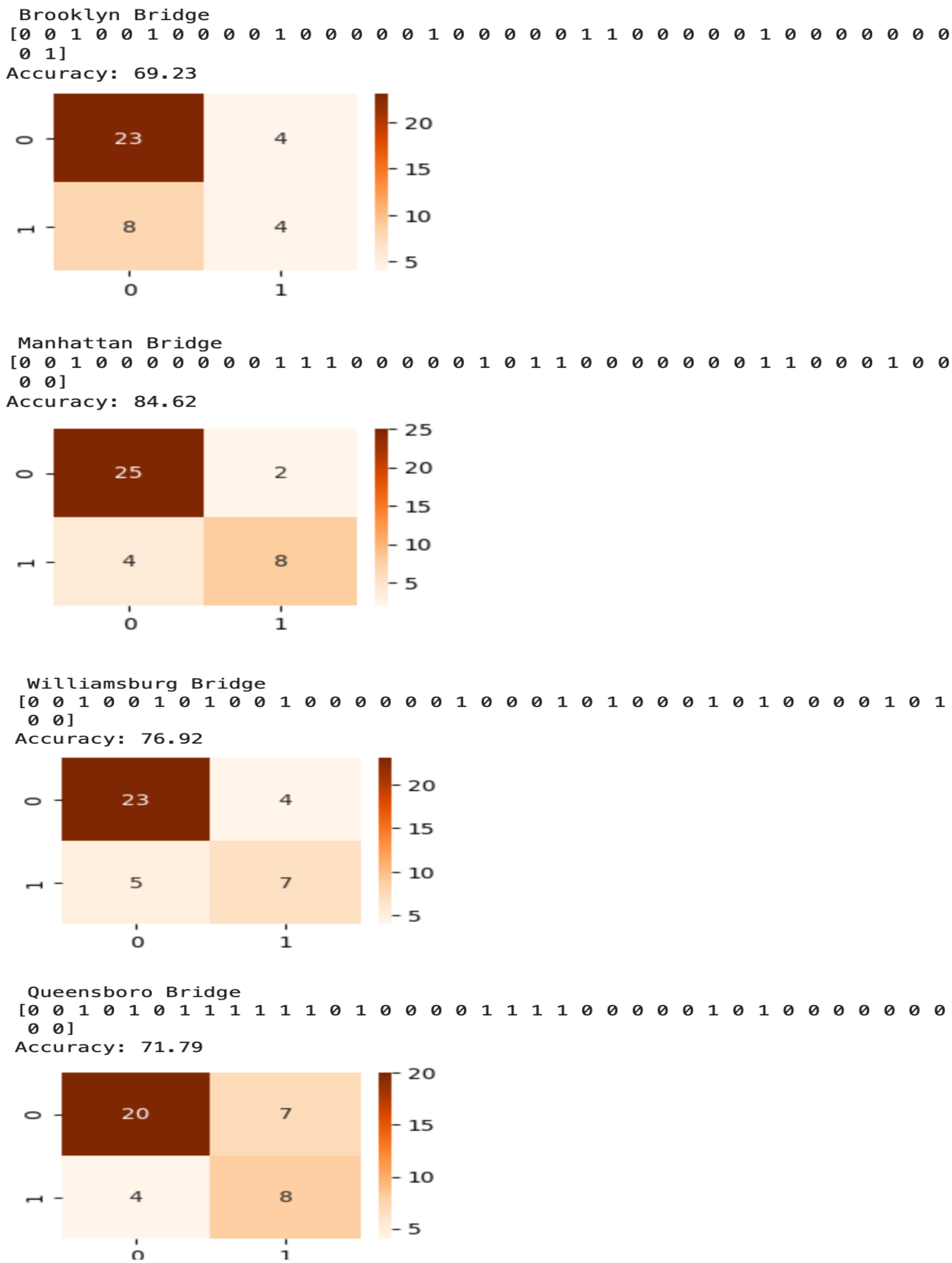
```
Equation 1: brooklyn_bikers= -1418.64*Precipitation + 3188.97

Equation 2: manhattan_bikers= -2722.83*Precipitation + 5389.45

Equation 3: williamsburg_bikers= -3061.78*Precipitation + 6538.28

Equation 4: queensboro_bikers= -1851.49*Precipitation + 4531.12

Correlation Fig(1,4,1): -0.3348207243792219

Correlation Fig(1,4,2): -0.4148010031135251

Correlation Fig(1,4,3): -0.4272161268152351

Correlation Fig(1,4,4): -0.3878380768827879
```

Fig 8. Linear Regression Plots and linear equations to explain the relationship between precipitation and the number of cyclists



```
0: No Precipitation
1: Precipitation

 Brooklyn Bridge
[0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0
 0 0]
Accuracy: 79.49

 Manhattan Bridge
[0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0
 0 0]
Accuracy: 76.92

 Williamsburg Bridge
[0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0]
Accuracy: 74.36

 Queensboro Bridge
[0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
 0 0]
Accuracy: 71.79
```

Fig 9. Logistic Regression outputs(with confusion matrix) that determine precipitation based on number of cyclists on each bridge



```
 Brooklyn Bridge
[0 0 1 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0
 0 1]
Accuracy: 69.23
```

```
 Manhattan Bridge
[0 0 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 1 1 0 0 0 1 0 0
 0 0]
Accuracy: 84.62
```

```
 Williamsburg Bridge
[0 0 1 0 0 1 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0 0 0 1 0 1
 0 0]
Accuracy: 76.92
```

```
 Queensboro Bridge
[0 0 1 0 1 0 1 1 1 1 1 1 0 1 0 0 0 0 1 1 1 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0
 0 0]
Accuracy: 71.79
```

Fig 10. K nearest neighbor outputs (with confusion matrix) that determine precipitation based on number of cyclists on each bridge