Trevor Taka/ttaka@purdue.edu
Lucas Brookes/lbrookes@purdue.edu
Path 1

The data we were given to complete this problem contains 3 feature variables (high temperature, low temperature, and precipitation) and the number of cyclists that travel across 4 major bridges each day over the course of 214 days. The data only contains measurements from April 1st to October 31st.

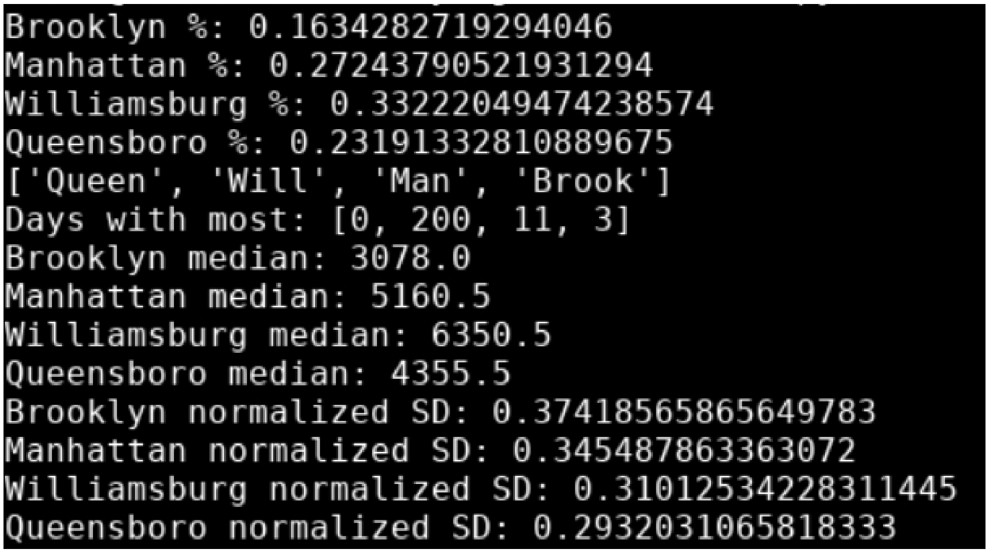
Our first problem asked us to choose one bridge to not put a sensor on to predict overall traffic given that we only have three total sensors for four total bridges. To find the least important bridge, we analyzed different features of each bridge dataset such as the percentage of total cyclists (most used), median, days most used (mode), and normalized standard deviation of the amount of cyclists on each bridge each day. We chose these features of the dataset because we wanted to determine the least used bridge as well as the most varying bridge to remove from our study to get the best picture about overall traffic.

In our second problem we attempted to create a model that could predict the total number of cyclists based on the weather report for the day. To achieve this, We created a linear (first degree) model with three feature variables. The equation uses the high temperature, low temperature, and precipitation for the day to attempt to predict the number of cyclists that will be riding across the bridges on those days.

The last problem required us to create a model to predict whether it is raining or not based on the number of cyclists on the bridges on a given day. Because this model only required one feature variable unlike the previous problem, we decided to test different degree polynomial regression models. We chose this model because we wanted to determine the model that predicts the precipitation most accurately based on the total cyclists. We don't expect any of the polynomial regression models to be way better at predicting precipitation than the others, but we hope at least one of them will be sufficient.

Trevor Taka/ttaka@purdue.edu
Lucas Brookes/lbrookes@purdue.edu
Path 1

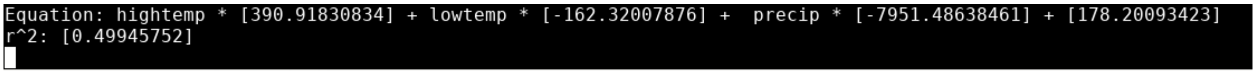
To figure out which three bridges we could monitor to get the best picture of the total population of cyclists for the day, we tried four different methods. Firstly, we summed up the numbers for each day into a total for each bridge and a grand total. From this we found what percentage of the cyclists over the course of the 214 day period were using which bridge. By percentage of total cyclists, the most used bridge was the Williamsburg bridge, and the least used bridge was the Brooklyn bridge. The second method we used was to check each day individually and register what the most used bridge was that day. From this method, we found that the Williamsburg bridge was the busiest on 200 out of the 214 days, followed by 11 days for the Manhattan Bridge, 3 days for the Brooklyn Bridge, and 0 days for the Queensborough bridge. The third method we used was sorting the median, for which we found that the highest was again the Williamsburg bridge, with the lowest being Brooklyn bridge. The last method we used was finding the normalized standard deviation or the percentage that a value deviates away from the mean on average for each bridge. We found that the highest varying bridge was Brooklyn at 37.4% on average while the least varying bridge was Queensboro at 29.3% away from the mean on average. From these methods, we can conclude that we would exclude Brooklyn from the experiment because it is both the least used and least consistent bridge for cyclists given our dataset.



```
Brooklyn %: 0.1634282719294046
Manhattan %: 0.27243790521931294
Williamsburg %: 0.33222049474238574
Queensboro %: 0.23191332810889675
['Queen', 'Will', 'Man', 'Brook']
Days with most: [0, 200, 11, 3]
Brooklyn median: 3078.0
Manhattan median: 5160.5
Williamsburg median: 6350.5
Queensboro median: 4355.5
Brooklyn normalized SD: 0.37418565865649783
Manhattan normalized SD: 0.345487863363072
Williamsburg normalized SD: 0.31012534228311445
Queensboro normalized SD: 0.2932031065818333
```

Figure 1, Percentage of total number of cyclists (first), # of days in which the bridge saw the maximum amount of cyclists for that day (second), median number of cyclists per day on each bridge (third), normalized standard deviation or percentage deviation away from mean of cyclists on each bridge (last)

For problem two, the linear model we generated is shown below in figure 2. The coefficients are 390.9, -162.3, -7951.5, and 178.2 for high temperature, low temperature, precipitation, and y-intercept. When modeled against the actual total number of bikers per day, we can calculate the r-squared of the model to figure out the accuracy of the model. Our model produced an r-squared value of 0.499, which should mean that there is a moderate amount of correlation. Of course, this is to be expected. Though weather can be a predictor of the number of cyclists, there are other factors that were not or could not be accounted for by our model, such as the day of the week, the time of year, holidays, and the status of bicycle routes outside of the bridges we can observe. Considering the factors that are out of our control, the model we have generated is decently good at predicting the total number of cyclists.



```
Equation: hightemp * [390.91830834] + lowtemp * [-162.32007876] +  precip * [-7951.48638461] + [178.20093423]
r^2: [0.49945752]
```

Figure 2, The model and r-squared generated for problem 2

A visual representation of the polynomial fits for first to fifth degree polynomials as well as their corresponding equations (order: 1st to 5th) are shown below. Unfortunately, our model has proven rather inconclusive as to which degree polynomial would best predict precipitation from total cyclists because none of the models have a very high r^2 value, but each of the first to fifth polynomial fits show the general pattern that precipitation decreases as total cyclists increases. The model we created appears to predict precipitation much more accurately when total cyclists is higher and precipitation is lower, showing greater inaccuracy when precipitation is higher and total cyclists are lower. We can conclude that we are generally able to predict whether it is raining based on the total cyclists on the bridges with our model, but it is less accurate as total cyclists approaches 0.

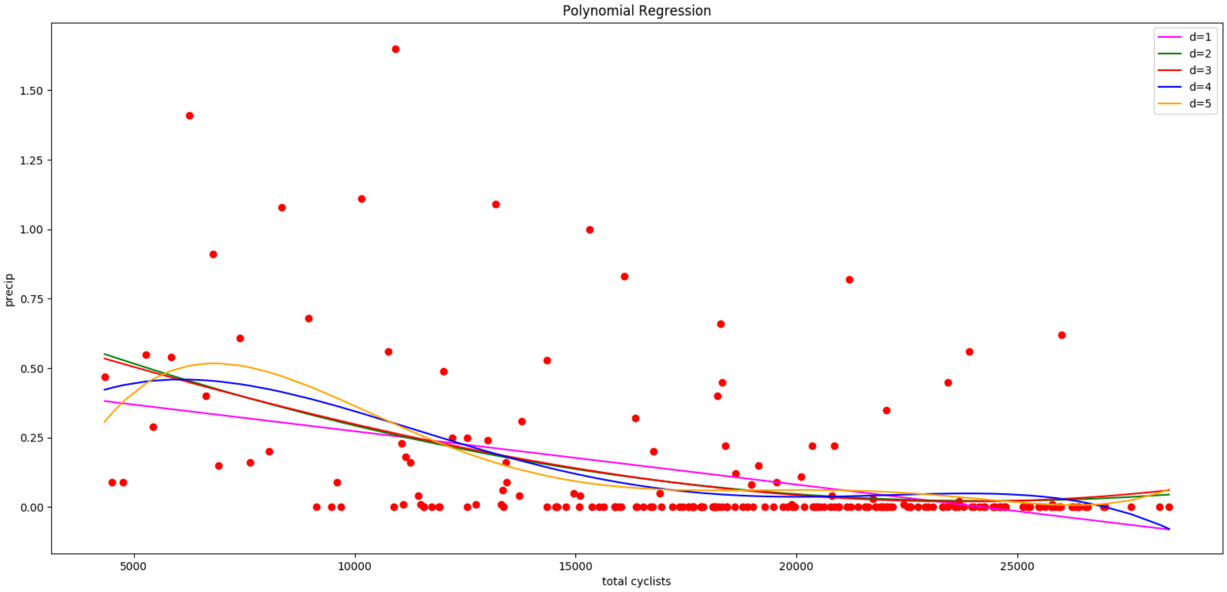


Figure 3, first to fifth order polynomial regression models for problem 3

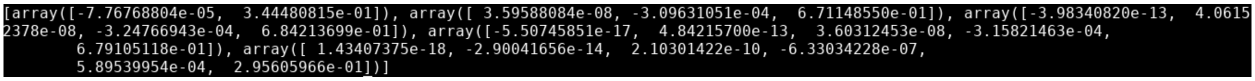

```
[array([-7.76768804e-05,  3.44480815e-01]), array([ 3.59588084e-08, -3.09631051e-04,  6.71148550e-01]), array([-3.98340820e-13,  4.0615
2378e-08, -3.24766943e-04,  6.84213699e-01]), array([-5.50745851e-17,  4.84215700e-13,  3.60312453e-08, -3.15821463e-04,
        6.79105118e-01]), array([ 1.43407375e-18, -2.90041656e-14,  2.10301422e-10, -6.33034228e-07,
        5.89539954e-04,  2.95605966e-01])]
```

Figure 4, polynomial regression equations for first to fifth degree polynomials for problem 3