Team Information

- Luke Canfield
  - Github - LukeCanfield13
  - Email - lcanfiel@purdue.edu
- Andrew Brandon
  - Github - realBrando
  - Email - brandoa@purdue.edu

- Path Taken - Path 1: Bike Traffic

Dataset

The dataset we are working with has nine columns of data about different features of New York City. The first two columns give us the range of all the datasets which is from Friday April 1st to Monday October 31st adding up to 214 days. Two of the columns pertain to the high and low temperatures (fahrenheit) of each day and another column refers to the amount of precipitation received on each given day. The last five columns of data are about the varying number of cyclists on four different bridges within the city with the total number of cyclists being given as the last column of the data. The four bridges are the Brooklyn, Manhattan, Williamsburg, and Queensboro bridge. From these different columns of data we are asked to answer three questions and draw conclusions from all of this information that we have been provided.

Question 1

Analysis: In this question, we were tasked with figuring out which three of the four given bridges we should install sensors on in order to get the best prediction of bike traffic in New York City. We determined that in order to get the best prediction of overall bike traffic we must install sensors on the bridges with the most bike traffic in order to get the most complete representation of overall bike traffic. Our approach for this problem was to analyze the distributions of each bridge based on the data provided to determine the bridges with the highest traffic frequency over the 214 days of data. We did this first by taking each individual day and normalizing the data to get traffic frequencies of each bridge, then averaging out each frequency over the 214 day period to get the average percentage of bicyclists on any given day on each bridge.

We chose this analysis for multiple reasons, but ultimately because we felt this was the best way to determine where to place the three sensors. As was stated before, the sensors can only best do their job when they have the most data to work with, which means they must be placed on the three bridges with the highest traffic frequencies. Building off of this, the best way to find the bridges with the highest traffic frequencies is to calculate the frequency of each bridge for each day, then average out each individual

frequency over the 214 day period. We expect this analysis to tell us four frequency values that should add up to 100%, or 1 if we keep the frequencies as decimals. The bridges with the highest returned frequency values are the ones with the most traffic on them, or the bridges with the highest percentage of the overall bike traffic. The three with the highest values will be the three that should have sensors placed on them.

Results: After performing our analysis, we received solid results that allow us to make a solid recommendation as to which three bridges should have the sensors placed on them. After running our analysis, we received 4 frequency values (that add up to one), and particularly we found one value that was somewhat lower than the rest, giving strong evidence as to which three bridges should have sensors placed on them in order to get the best prediction of overall bike traffic. We found that the three bridges that should have sensors placed on them are the Manhattan, Williamsburg and Queensboro bridges. We found that the average percentage of bicyclists on any given day for the Williamsburg Bridge was 33.3%, Manhattan was 27%, Queensboro was 23.5% and Brooklyn was 16.1%, thus showing that Brooklyn Bridge has the lowest average frequency and thus should not have a sensor placed on it.

Question 2

Analysis: In this question we were asked if we were able to use the next day's weather forecast to predict the number of bicyclists that day. Our approach to answering this question was to use linear regression to (hopefully) find a correlation between weather forecast and the total number of bicyclists. In order to quantify the weather, we squared the high temperature, low temperature and precipitation, then added up the three values and square rooted this sum. This acted as a way to get a somewhat normalized value for weather that could be used. It also works as a value due to the lack of precipitation's effect on bicyclists, as was found in question three. Overall, with this quantity a higher value means better weather, as greater temperature is better, and the value of precipitation was so low in comparison that it did not have an effect on this value. Once we had this quantity for the weather, we were able to run linear regression with the number of bicyclists as the dependent variable and the weather forecast as the independent variable. While the question did say weather forecast, we feel that using the weather data for this test and taking the results from it will also apply to the weather forecast.

With a linear regression, we can use the data given to estimate functions based on the weather forecast to predict the number of bicyclists. We expect to get a linear regression and graph that showcases this. As far as interpreting the data, a line that correlates well with the data and has a high goodness of fit will mean that there is a direct correlation between the weather forecast and the total number of bicyclists.

Logically speaking, we expect there to be at least some sort of correlation as the weather improves, the number of bicyclists will also increase.

Results: After running our analysis, we received somewhat good feedback that our predictions were correct. We found five estimated functions, which were:

$$\hat{y}_1 = 179.88082X + 1039.80897$$

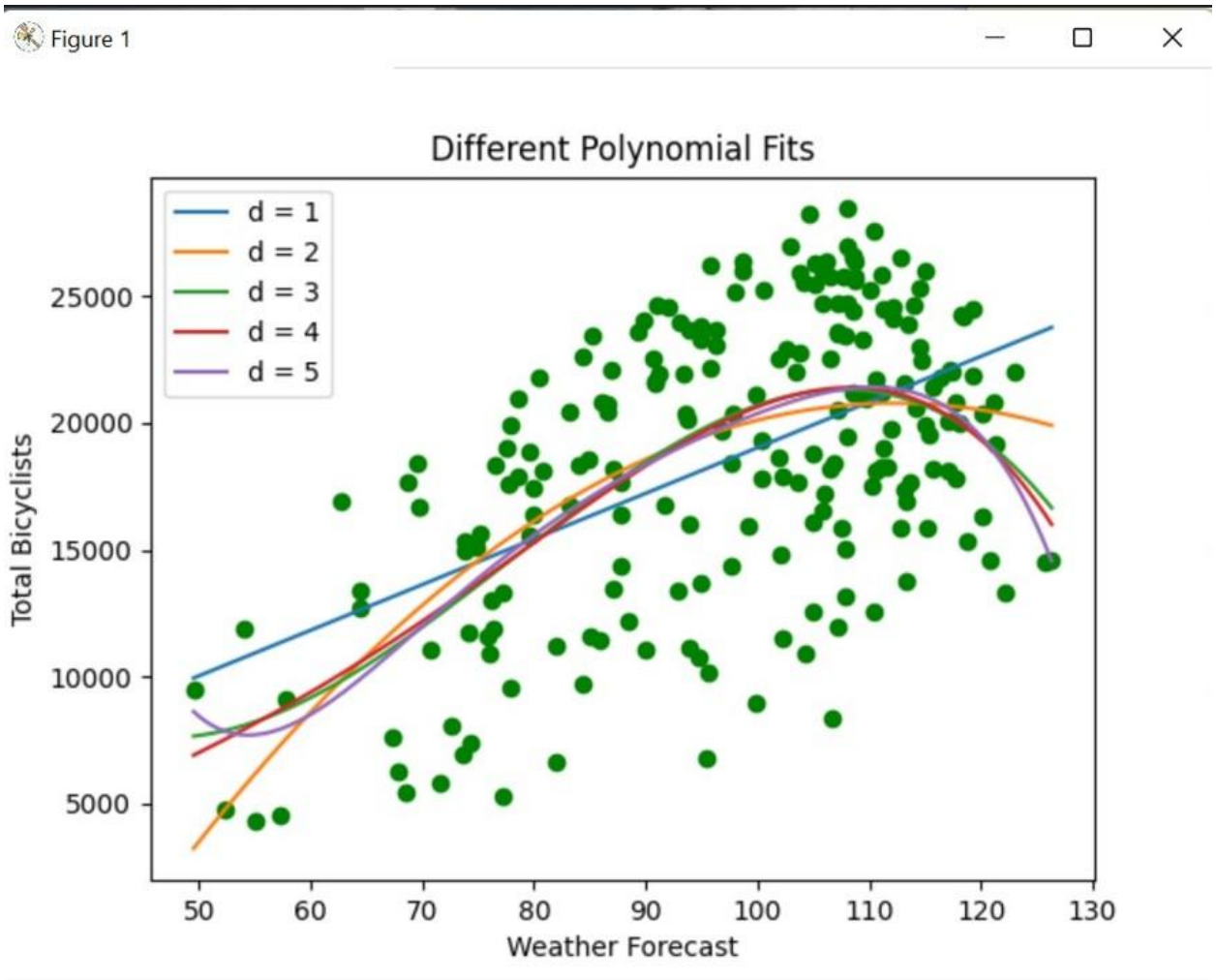
$$\hat{y}_2 = -4.4553X^2 + 1.00052 * 10^3X - 3.53896$$

$$\hat{y}_3 = -1.2369 * 10^{-1}X^3 + 28.76114X^2 - 1.8895 * 10^3X + 4.572 * 10^4$$

$$\hat{y}_4 = -1.08 * 10^{-3}X^4 + 0.26157X^3 - 21.5613X^2 + 9.5388 * 10^2X - 1.2727 * 10^4$$

$$\hat{y}_5 = -1.1477 * 10^{-4}X^5 + 5.01696 * 10^{-2}X^4 - 8.7206X^3 + 7.4949 * 10^2X^2 - 3.1405 * 10^4X + 5.1737 * 10^5$$

Once we had these five estimated functions, we then graphed them against the scatter plot data of the weather forecast and the total number of bicyclists to get this graphical representation:



Looking at the graphical representation of the linear regression, it appears that there is in fact a positive correlation between weather forecast and the total number of bicyclists. As the weather improves (weather forecast number increases), so does the total number of bicyclists. With this positive linear regression, we can say that yes, we can use the weather forecast for the next day to predict the total number of bicyclists.

Question 3

Analysis: In this question we were asked if we were able to predict the weather, specifically if it was raining, based on the number of cyclists on a particular day. Our approach to answer this question was to see if there was in fact a correlation between the two variables, total number of cyclists and precipitation. We decided to test this by running a linear regression on the data to create a linear fit to predict the values of precipitation based on the total number of cyclists. In this scenario, the independent variable is the total number of cyclists and the dependent variable is the level of precipitation. We believe this analysis technique will allow us to answer the question of if we can predict if it is raining based on the number of bicyclists because running a linear regression will allow us to see the correlation between the two variables, if any. By using linear expression we can create a predicted linear function based on the data given, then compare the line we come up with with the actual data and see how good of a fit the predicted line is to the actual data.

We expect the data to tell us if there is in fact a correlation between rain and number of bicycles, specifically if we can predict if it is raining based on the number of bicyclists. This prediction will be based on the goodness of fit of the predicted linear equation to the actual data. If it is a close fit, then the answer would be yes, we can predict if it is raining based on the total number of bicyclists. If not, then we can conclude that there is no correlation between the two variables and we cannot accurately predict if it is raining based on the number of bicyclists.

Results: After performing our analysis, we received an answer to the question that left no doubt in our minds as to what was correct. After running the linear regression, we found five estimated functions, which were:

$$\hat{y}_1 = -1.918027 * 10^{-5}X + 0.464755$$

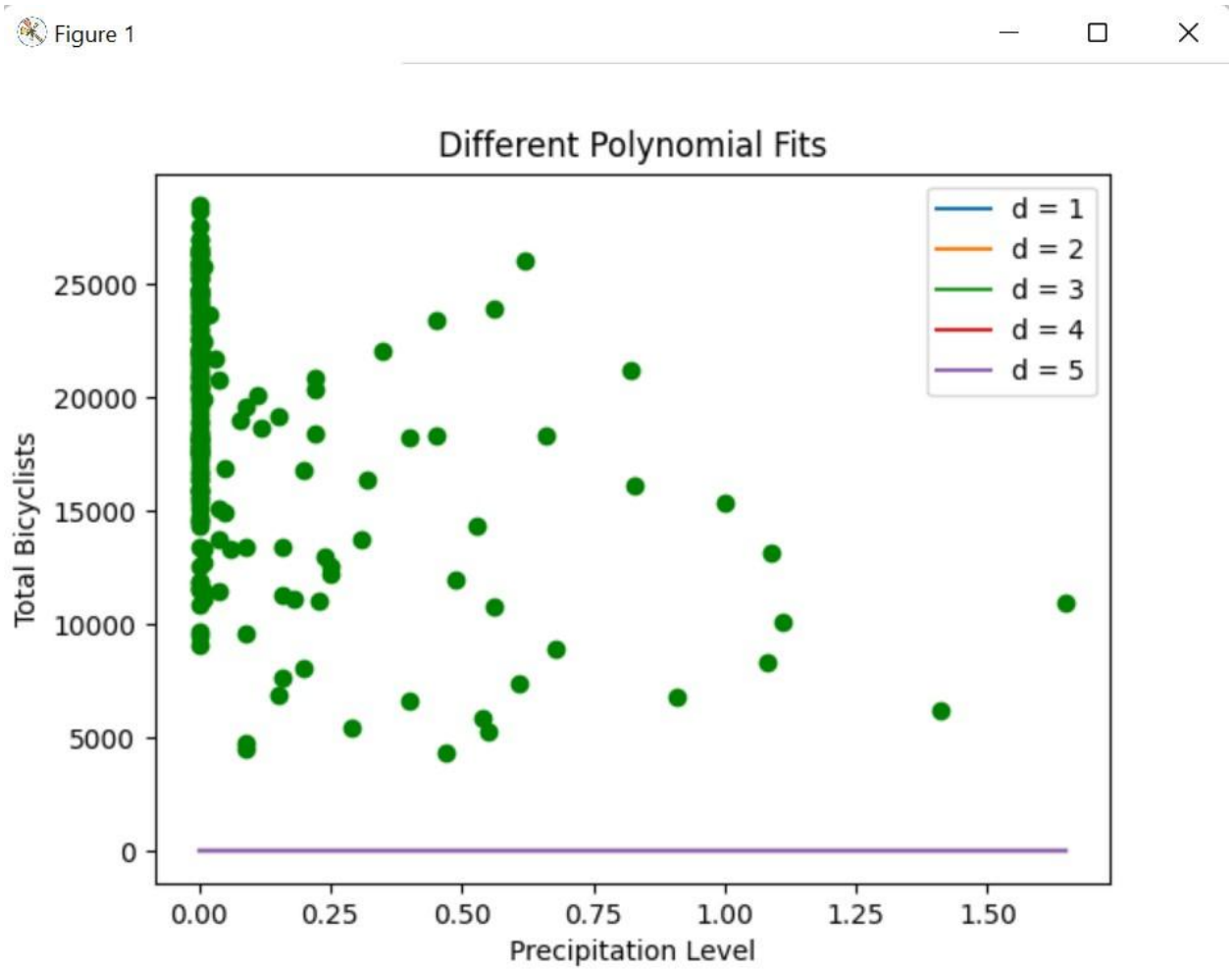
$$\hat{y}_2 = 1.3272 * 10^{-9}X^2 - 6.4483 * 10^{-5}X + 0.805496$$

$$\hat{y}_3 = 1.906 * 10^{-14}X^3 + 3.8024 * 10^{-10}X^2 - 5.02493 * 10^{-5}X + 0.74377$$

$$\hat{y}_4 = -2.1321 * 10^{-17}X^4 + 1.42235 * 10^{-12}X^3 - 3.18315 * 10^{-8}X^2 + 0.000248X - 0.16423$$

$$\hat{y}_5 = 3.305 * 10^{-21}X^5 - 2.95353 * 10^{-16}X^4 + 1.0009 * 10^{-11}X^3 - 1.5728 * 10^{-7}X^2 + 0.00109X - 2.1675$$

Once we had these five equations, we then graphed them against the scatter plot data of precipitation level vs total number of bicyclists in a day, and we received this graphical representation:



Looking at this graphical representation, it is obvious that the estimated functions have very little to no fit to the actual data. There is basically a goodness of fit of zero, as the one line we actually can see is nowhere near the data, which also just looks scattered to begin with and does not appear to follow any particular pattern. With all this in mind, we feel that we can confidently say that we can NOT predict the weather based solely on the total number of bicyclists on a particular day.