ECE 20875 Mini Project Report

Summary:

- Names
  - Wesley Hagemier - whagemei@purdue.edu
  - Ty Brunk - tbrunk@purdue.edu
- Purdue Usernames
  - Wesley Hagemier - whagemei
  - Ty Brunk - tbrunk

- Selected Path
  - Path 1: Bike Traffic

Data Set:

The data set we are dealing with for this project is the bicycle counts on New York City bridges for the months of April thru October in 2016. This data set includes dates, the day of the week for each date, the high and low temperature in Fahrenheit for each day, the precipitation for each day, and the amount of traffic for each of the bridges leading to the city. There are four bridges that have been included in this data set. The four bridges include the Brooklyn Bridge, the Manhattan Bridge, the Williamsburg Bridge, and the Queensboro Bridge. The number of bike passengers passing on each of these four individual bridges has been recorded for each day in the data set. Finally, the data set also includes the combined total number of bike passengers on all four bridges for each given day.

Analysis for Questions:

<u>Question 1</u>

We are to figure out the best prediction for the overall traffic using sensors and placing them on the four bridges. However, we are only given enough sensors to get the traffic on three of the four bridges. We were tasked with figuring out which of the three bridges to add the sensors to predict the overall traffic on all four bridges. We decided that the best way to complete this would be to calculate which bridge has the highest variability in its traffic data. We would then place the sensors on the remaining three bridges. Variability could affect our results when trying to predict the overall traffic on the bridges. Therefore, this path would help us eliminate variability in the data to better predict the overall traffic on all four bridges. It is important that our predictions are

accurate, and removing the variability in the data will make our predictions for overall traffic more accurate.

We can complete this process by first, calculating the mean for the traffic on each bridge. Then, we will calculate the deviation from the mean for each data point. We will take the absolute value of each deviation and sum them all together. Lastly, we would divide by the number of data points for the given bridge to get the average deviation in the data for that bridge. We will complete this process for each of the four bridges. Then, whichever bridge has the largest average deviation from the mean will be the bridge we do not place sensors on.

Question 2

Next, we are tasked with finding out if we can use the next day's forecast to predict the amount of traffic for that day. We decided to do this by constructing a confusion matrix for the data. We will first separate the data into four subgroups: days with a temperature above the mean temperature and low precipitation, days with a temperature above the mean temperature and precipitation, days with a temperature below the mean temperature and low precipitation, and days with a temperature below the mean temperature and precipitation. After taking a look at the data, we saw that when the precipitation for a given day is at or below 0.02, the bike traffic on the bridges is not affected. So we will classify all days at or below 0.02 precipitation levels to be low precipitation. We will calculate the midpoint between the mean traffic for each of these four subgroups. We will use these midpoints as ranges to classify the different subgroups. The highest range for the subgroup with the highest mean amount of traffic, and the lowest range for the subgroup with the lowest mean amount of traffic. Using these ranges we will construct a confusion matrix with our prediction to be the subgroup with which its respective mean is contained in the range of overall traffic. Once the confusion matrix has been calculated, we will then calculate the accuracy of our predictions. Before we conducted any tests, we decided that to accurately predict the next day's traffic based on the weather, we would want the accuracy of classification to be at least 80% or 0.80. This means that we want to be 80% confident in our prediction of the total amount of traffic based on the weather.

Question 3

Lastly, we were asked if it is possible to predict the day of the week based on the number of cyclists on the bridges. To figure this out we decided to construct a Gaussian Mixture Model of the data. With this model, we can train our data to it and then it can predict the day based on the amount of traffic. First, we classified each day as a number. We set Friday=0, Saturday=1, Sunday=2, Monday=3, Tuesday=4, Wednesday=5, and Thursday=6. These will be our y-values corresponding to the x-value which will be the total amount of traffic on that day. We will then take out every

tenth data point to be a part of our testing data, the rest of the data will consist of our training data. We will then use the data to construct the Gaussian Mixture Model with 7 components, one for each day of the week. We will then fit the model to our data. We can then use the model to predict what day it will be. Whichever number 0 through 6 the output is will be the day that the model predicts. We will then compute the accuracy of the predictions. Just like the previous question, we want to be at least 80% accurate in our predictions to say that we can predict the day given the amount of traffic.

Results:

Question 1

Once we carried out our analysis, we decided that we should place the sensors on the Brooklyn, Manhattan, and Queensboro Bridges. We calculate the average deviation from the mean for the amount of traffic for a given bridge on any given day. The mean and average deviations (rounded to the nearest whole number) for the traffic on each of the four bridges are given in Table 1 below.

Table 1: Mean and Average Deviation for the Amount of Traffic on each Bridge

Bridges	Brooklyn	Manhattan	Williamsburg	Queensboro
Mean	3031	5052	6161	4301
Standard Deviation	1131	1741	1906	1258
Average Deviation	854	1445	1612	1044

Based on the data in the table, we chose to not place sensors on the Williamsburg Bridge because it has the highest average deviation. We wanted to remove as much variation in the data as we could, so not placing sensors on the bridge with the largest average deviation would minimize the variety in the data. However, we recognize that this method does not take into account the total amount of traffic for the bridges. The Williamsburg Bridge has the largest amount of traffic compared to the other bridges as seen in the mean traffic for each bridge in Table 1. This means the prediction we make of overall traffic on the four bridges will more likely be smaller than the actual amount of traffic on a given day. By deciding to place the sensors based on the variety in the data, we did not take into account the total amount of traffic on each bridge. This might skew the results of the prediction for the traffic on all four bridges. Despite the lack of a sensor on the Williamsburg Bridge, the bridges share the same trends nonetheless, the others we chose are just more consistent with less volatility in the data.

Question 2

After conducting our analysis, we cannot, with an acceptable confidence rate, accurately predict the total number of bikers that day based on the weather of that day. Once we classified the data into the 4 different subgroups and made a confusion matrix of the confidence at which we can accurately estimate the total number of people crossing the bridges each day, we found the confidence rate to be 45.3% as seen in Table 5 below. This was too low for us to find acceptable. As seen in Table 2, the average number of people crossing the bridges on high precipitation-high temperature days, and low precipitation-low temperature days have similar average values. This caused our ranges to be too close to one another and this affected our results.

Table 2: Mean Amount of Total Traffic Based on Temperature and Precipitation

	High Temp (number of bikers)	Low Temp (number of bikers)
High Precipitation:	17,457	9,866
Low Precipitation:	21,768	18,506

We thought that if we could combine those 2 subgroups of days, it would increase the rate at which we can accurately classify a day in one of our subgroups; thereby allowing us to estimate the number of bikers crossing the bridges for that day based on the weather. As seen in the confusion matrices we've created below, the rate at which our method misclassified each day of the data is too high to deem acceptable. Even after we made this change, the confidence rate was still not close to our expected 80%. Since neither method could accurately classify the amount of traffic for a given day based on the weather, we conclude that we cannot accurately predict the amount of traffic for the following day solely based on the weather forecast.

Table 3: Confusion Matrix With 4 Subgroups (Predicted: Columns) (Actual: Rows)

Predicted:	Low Temp and High Precipitation	High Temp and High Precipitation	Low Temp and Low Precipitation	High Temp and Low Precipitation
Low Temp and High Precipitation	24	7	13	2
High Temp and High Precipitation	4	6	15	18
Low Temp and Low Precipitation	1	8	9	11
High Temp and Low Precipitation	0	9	29	58

Table 4: Confusion Matrix With 3 Subgroups (Predicted: Columns) (Actual: Rows)

Predicted:	Low Temp and High Precipitation	Low Temp and Low Precipitation + High Temp and High Precipitation	High Temp and Low Precipitation
Low Temp and High Precipitation	25	21	2
Low Temp and Low Precipitation + High Temp and High Precipitation	4	36	29
High Temp and Low Precipitation	0	39	58

Table 5: Accuracy (Confidence Rate) of Each of the Two Confusion Matrices

Confusion Matrix:	Four Subgroups	Three Subgroups
Accuracy:	45.3%	55.6%

Question 3

Based on the results of our tests, we could not use the data to predict what day of the week it was. This was because the Gaussian Mixture Model we formed for the data was only able to correctly classify 14% of the days. We wanted to see a much higher number for us to be able to accurately predict the day of the week based on the number of bikers on a given day. We set aside 10% of the data as test data, three points for each day. We used the remaining data as training data to train the model. We used the total amount of traffic as our x-values and the days of the week (0-6) as our y-values. We figured this would be the case from the beginning. When looking at the data, weekdays (Monday through Thursday) had much higher levels of traffic than weekends (Friday, Saturday, and Sunday) with Friday generally having traffic levels in between the numbers closer to those of weekends. However, it was much more difficult to differentiate the number of bikers each weekday because they all had similar numbers. It was the same for the weekends as well. What would have also affected the results are the outliers that are caused by the weather. Potentially, if we set all other variables constant, you could predict if the day is a weekend or a weekday, but not the specific day. The fact is that the Gaussian Mixture Model we created did not predict the

days well enough to be confident in our results. Therefore, we cannot use this data to predict which day of the week it is based on the amount of traffic.