Data set .

The data set analyzed in this project is a collection of the number of cyclists on four different bridges in New York City as well as information about the weather for each date between April 1st, 2019 and October 31st, 2019. The first column of the table contains the date, the second and third contain the high and low temperatures respectively (in degrees fahrenheit). The 4th column contains the amount of precipitation (in inches), as well as extra information in the form of letters. A capital T represents trace amounts of precipitation, and an s in parentheses means the precipitation that day was snow. The next 4 columns provide the number of cyclists on each of the bridges, and the last column contains the total number of cyclists on all of the bridges.

Some preprocessing was necessary to use the data effectively. Since the data is read in as nested list of strings, it needs to be converted into numeric data for use in analysis. The temperature and bridge columns could be converted using the float() operator, but the precipitation column needed more work. To account for trace amounts of precipitation, all T's were replaced with 0.001. Snow only occurred once in the data set, so it would only throw off any regression models. Instead, the (S) was stripped from the string where it occurred. Once these changes were made, all data in the precipitation column was converted to floats.

Chosen Analyses .

The first analysis question posed was which three bridges should be chosen in order to predict the overall number of bikers. I chose the regression model to answer this question, because it would be the best model to find the relationship between a dependent and independent variable(s), which allows us to make the best prediction we desire. To achieve that, I set X feature matrix to be four different combinations of bridges and y to be the total number of bikers. After performing the regression model analysis, the one with the lowest MSE would be the answer to the question. To predict the total number of bikers on a given day, we need to make an equation using coefficients and y-intercept that we determined from the model, and plug the normalized values on the given day into the equation.

The second analysis question asked whether the weather for a day could be used to predict the number of cyclists that day. To answer this question, a linear regression analysis was performed. The goal of this method is to create an equation which will predict the total number of cyclists using the high and low temperatures, as well as the amount of precipitation.

I chose to use this method because regression models have a coefficient of determination, which states how much of the variation in the data can be explained by the model. If a model is based entirely on the weather and is relatively accurate, this measure

can be used to approximate how much of the variation in the number of cyclists can be explained by the weather. Additionally, if the regression model can predict the number of cyclists with enough accuracy, the question has been answered already, as the model created would be able to use the weather to predict the number of cyclists. To determine whether the model accomplishes this goal, (absolute) percent error is used as a measure of accuracy. An average percent error of less than 25% would mean that on average, the model predicts a number of cyclists which is within 25% of the actual value, and this should be sufficient for the scenario posed in the question.

The third analysis question asked whether the number of cyclists on the various bridges could be used to predict whether it is raining or not. I chose to build a classifier to answer this question. Specifically, I used the method of logistic regression, which trains a function to predict which of two categories a data point falls into using a number of features.

I chose this method because the analysis question is asking whether a good classifier can be built. Classifiers offer several metrics of accuracy which can be used to determine how effective they are, so once a classifier has been trained using only the features provided, these metrics can be used to determine whether the information provided is enough to create an effective classifier. The metrics used in this analysis are precision, recall, and accuracy. In the context of this question, a high precision for the rainy state means that if the classifier says it is raining, it probably is, and a high recall for the rainy state means that if it actually is raining, the classifier will probably tell you it's raining. Accuracy measures how often the classifier is able to make correct predictions in general.

I chose to use a logistic regression model as a classifier instead of naïve Bayes because while the distribution of bike traffic for rainy days was roughly normal (see figure 1), the distribution for bike traffic on dry days was much more difficult to model (see figure 2). A logistic classifier should be just as effective, but will be much less complicated to build.

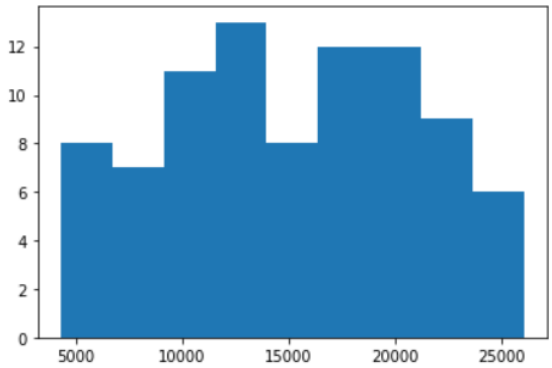

Figure 1: Distribution of Cyclists on Rainy Days

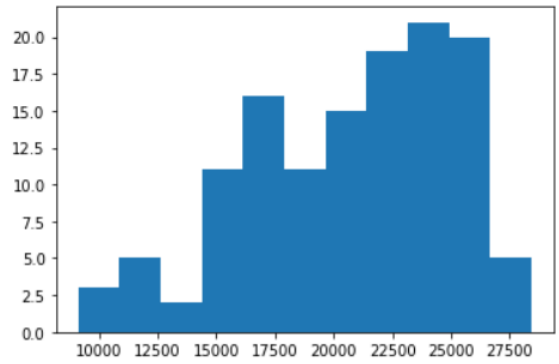

Figure 2: Distribution of Cyclists on Dry Days

Results .

For the first analysis question, it turns out we need to install sensors on Brooklyn, Manhattan, and Williamsburg. After running a linear regression analysis on each combination of three bridges, the results were as follows:

**********Brooklyn Manhattan Williamsburg**********
Best lambda tested is 0.1, which yields an MSE of 95533.60435201984
[[1334.90521028 1586.91254989 3078.61762491]]
[18522.63520063]
**********Brooklyn Manhattan Queensboro**********
Best lambda tested is 5.5, which yields an MSE of 195329.64912304076
[[1121.77349057 2249.8453112 2593.79486426]]
[18566.99163226]
**********Manhattan Williamsburg Queensboro**********
Best lambda tested is 7.5, which yields an MSE of 280055.5442362173
[[2012.29258045 1739.74731361 2029.52630079]]
[18587.28178869]
**********Brooklyn Williamsburg Queensboro**********
Best lambda tested is 4.5, which yields an MSE of 804158.8924318344
[[1443.06775916 3103.21785007 1242.55460626]]
[18491.42495977]

The first combination has the lowest MSE, which is 95533.60 with the lambda value of 0.1. Coefficients are 1334.90, 1586.91, 3078.62, and the y-intercept is 18522.63. Therefore, the equation needed to make the prediction would be 1334.90x+1586.91y+3078.62z+18522.63, where x is equal to cyclists on the brooklyn

bridge minus 3031 (mean of brooklyn sample data) divided by 1131.4 (standard deviation of brooklyn sample data), and so on. MSE is a measure of how far off predictions using a model are on average. Since the combination of Brooklyn, Manhattan, and Williamsburg had the lowest MSE, a model using these bridges predicts total traffic the most effectively, so installing sensors on the Brooklyn, Manhattan, and Williamsburg bridges would be the best way to predict the total traffic.

The second analysis question asked whether the weather forecast could be used to predict cyclist traffic for the purpose of allocating police resources. To answer this question, we ran a linear regression analysis using high temperature, low temperature, and amount of precipitation as features. In order to determine how to implement each feature, scatter plots were created to illustrate the relationship between each feature and the final output. As seen in figures 3 and 4, the relationships between high and low temperatures are roughly linear. Figure 5, however, shows that the relationship between precipitation and total riders is also roughly linear, days with no precipitation and days with only trace precipitation throw off the trend quite a bit. To get around this, we created 3 separate regression models. Additionally, since all features showed roughly linear relationships, no new features needed to be created.

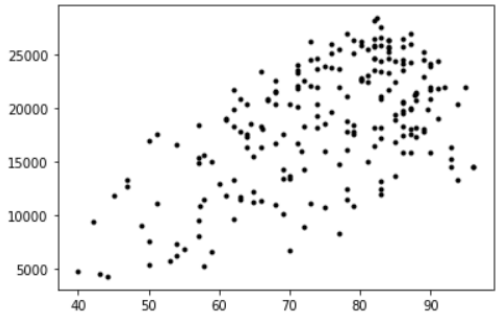

Figure 3: High Temperature vs. Number of Cyclists

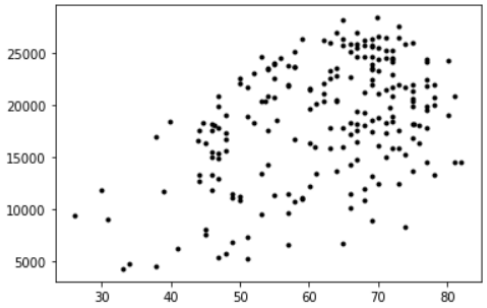

Figure 4: Low Temperature vs. Number of Cyclists

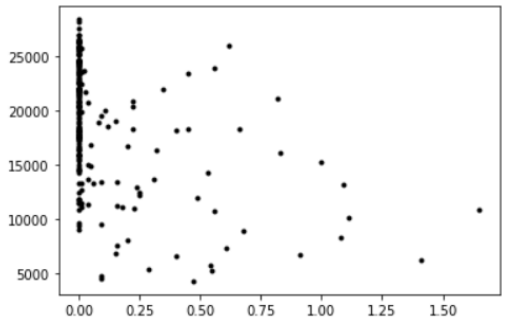

Figure 5: Precipitation (in) vs Number of Cyclists

The three regression models created were as follows:
Dry: 161.93 * (high temp) + 27.28 * (low temp) + 6764.80
Trace: 694.92 * (high temp) - 638.28 * (low temp) + 7018.08
Rainy: 452.99 * (high temp) - 172.47 * (low temp) - 4590.07 (inches of rain) - 6176.24

Next, a function was written which combined all of the models by choosing which to apply on each day. When this model was tested, it had a coefficient of determination of 0.5815, or 58.15%. This means that the model is able to explain 58.15% of the variation in the data. Additionally, the absolute percent error of the model at each point in the provided set was computed, and these values were plotted on a histogram in figure 6 below:

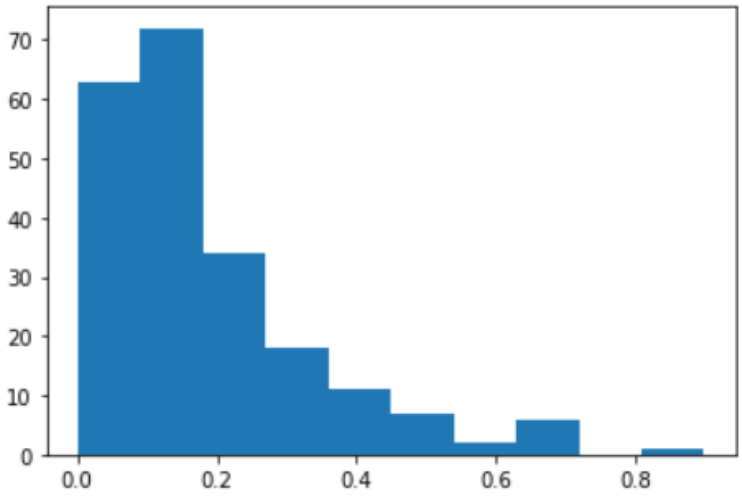

Figure 6: Histogram of Percent Errors for the Combined Regression Model

The average absolute percent error for the combined model was 18.77%. This means that on average, the combined model calculated a value within 18.77% of the true value. As seen in the histogram, roughly 75% of the time, the model was within 25% of the true value.

The original question posed was whether the weather could be used to predict bicycle traffic, and the model created in this analysis suggests that the answer is yes. The linear regression process created a model that could explain about 60% of the variation in

the number of cyclists. This can be taken to mean that, at least within the training data set, the weather explains at least that much of the variation in the number of cyclists, and a better model based on the same data may show that the weather explains even more. Additionally, the model almost always predicted within 25% of the true value, and a better model based on more data could theoretically be slightly better. Being able to reliably guess the number of cyclists within 25% should be sufficient for the police to determine how many officers to deploy on the bridges.

The third analysis question asked whether the number of cyclists on the different bridges could be used to predict whether it was raining. The main method for answering this question is to attempt to build a classifier which takes the number of cyclists on each bridge as inputs. Since the data is difficult to model using a naïve Bayes classifier (see Chosen Analyses), I chose to run a logistic regression analysis to build a classifier using the number of cyclists on each bridge as features. The total number of cyclists could not be used because it is the sum of the other four features, and is therefore not linearly independent of them.

The classifier built using all four bridges as features was the following equation:

$$\frac{1}{1 - e^{-5.03E-6 * a - 3.29E-4*b - 5.17E-4*c + 2.89E-4*d}}$$

Where a is the number of cyclists on the Brooklyn bridge, b is the number of cyclists on the Williamsburg bridge, c is the number of cyclists on the Manhattan bridge, and D is the number of cyclists on the queensborough bridge. When the value of the above function is greater than ½, the model predicts that it is raining.

This model had an overall accuracy of 73%, which is passable. However, the recall is less passable. This is the most important feature of a model designed to predict rain, as false negatives for rain are much worse than false positives. This model had a recall of 0.57, meaning that if it is currently raining, the model as about a 57% chance of telling you it is raining. This isn't great, because if our model has to be wrong, it would be better for it to say it's raining than to say it isn't. To create a more cautious model, one can alter the inputs so that rain is more common than not. Taking only 60 dry days as training data, but still taking all 86 rainy days resulted in an entirely new model, provided below:

$$\frac{1}{1 - e^{7.59E-7 * a - 2.83E-4*b - 8.14E-4*c + 1.5E-3*d}}$$

This new model had a much better recall for rain, at 76%. This would be an acceptable number. Unfortunately, however, the precision dropped down to 50%. This means that while if it is raining, the model will usually correctly tell you it's raining, but

unfortunately this is only because it says it's raining way more often than it should. This was to be expected as the sample data given was biased in favor of rain, however the precision dropped too much for the classifier to be good. Additionally, this new model had an overall accuracy of only 60%.

Taking the accuracy, precision, and recall scores of both models into consideration, I would say that the number of cyclists on bridges can be used to predict whether it is raining more effectively than random guessing, but it probably shouldn't be used for any practical applications. Training the model only on data as provided gave an accuracy of 73%. Since this is a binary classifier, random guessing has an accuracy of 50%, meaning our classifier is definitely better, and if it's raining out, 57% of the time it will correctly determine that. To make it more practical, the model was trained again on biased data. Using this method, the model was able to correctly predict rain 76% of the time, however the accuracy decreased to 60%, which is hardly better than guessing. The tradeoff between recall and precision for rain means that a classifier built on this data has an upper limit on how good it can be. In order to make it correctly tell you it's raining 75% of the time, it must overpredict rain to the point where when the model says it is raining, half of the time it actually isn't.

While the data provided is not enough to build a good classifier, it could still be possible. Since rain and the number of cyclists present depend heavily on the time of year, with more data it may be possible to build individual classifiers for each month or season, which should be more accurate as the overall chance of rain and the expected number of cyclists for rainy and dry days would both become better estimates by removing the influence of the time of year.

Some functions and classes from the sklearn library were used in the analyses presented here.

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011