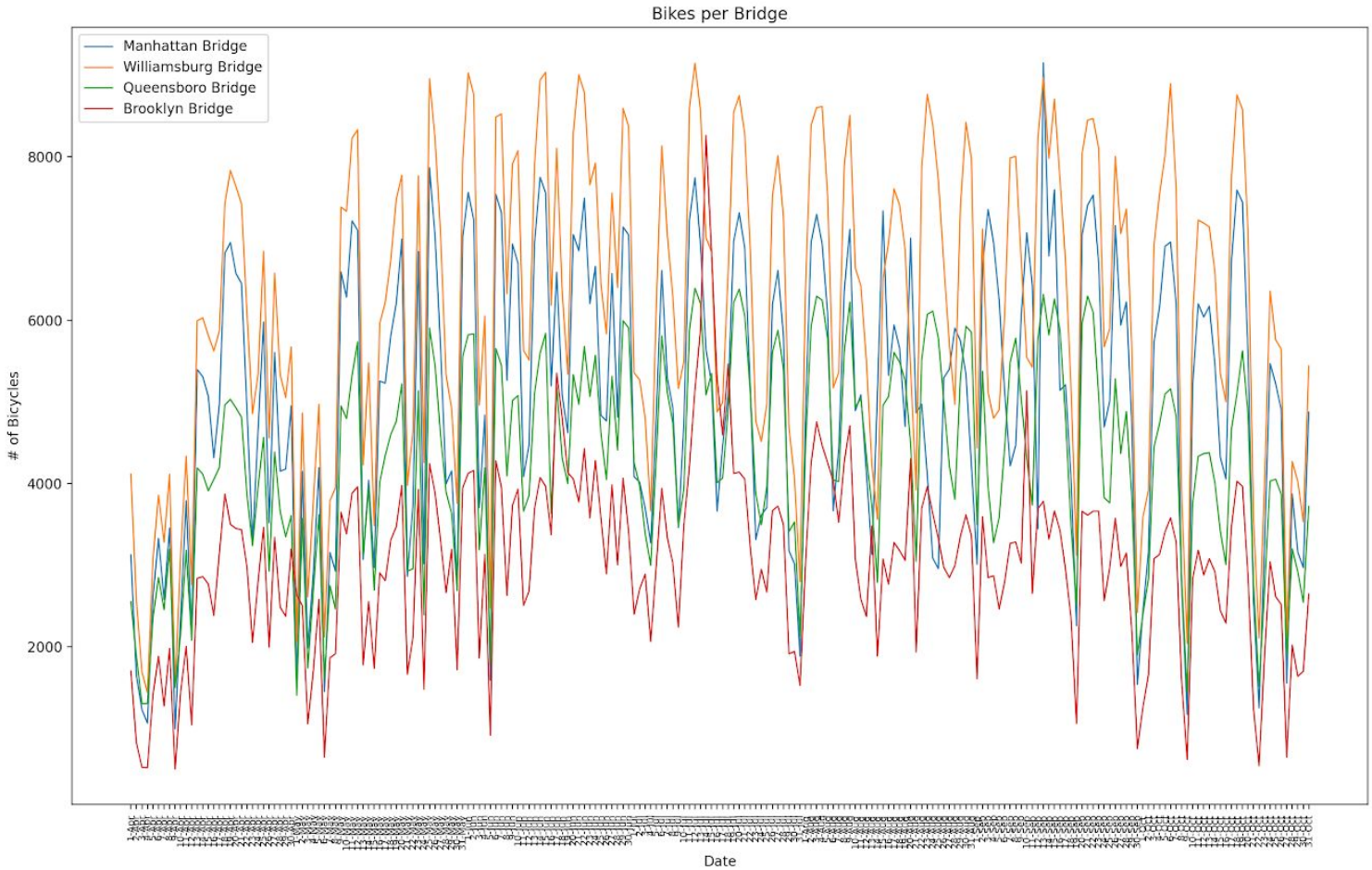
Vivek Khanolkar
vkhanolk
ECE 20875

Path 1: Bike Traffic

In this report I will be analyzing and discussing the various conclusions I have made about the bicycle traffic across a number of bridges in New York City. This data was taken from Kaggle, and the csv file is in the project repository, titled 'NYC_Bicycle_Counts_2016_Corrected_2.csv'. This data takes various logistics such as Temperature, number of Bikers, Precipitation, and date for four bridges.

1. Although we have data from 4 bridges in New York City (Manhattan, Queensboro, Brooklyn, and Williamsburg), we only have a budget that will allow us to put sensors on three of these bridges. To Determine which of the three bridges to put sensors on I first graphed the data with the total number of bicycles per day for each bridge against the date. This was just to get a general sense of what the data looked like, and if I can spot any outliers right off the bat.



It was easy to see that Williamsburg Bridge had much more bikes than the rest, and also seemed the most volatile, however the Brooklyn Bridge also seems to have a decent amount of outliers as well. To confirm this I first calculated the Mean, and Standard deviation for each bridge. At first glance one may say that Williamsburg is the obvious sore thumb, however looking at the bridge with the most outliers may not tell the whole story. Investigating further, I normalized the data around one mean and found how much of each data fell within one standard

```
Brooklyn Bridge Standard Dev:  1134.0448253964707 Mean: 3030.700934579439
Manhattan Bridge Standard Dev:  1745.4854071736424 Mean: 5052.2336448598135
Williamsburg Bridge Standard Dev:  1910.643105842522 Mean: 6160.873831775701
Queensboro Bridge Standard Dev:  1260.9857250379587 Mean: 4300.72429906542
```

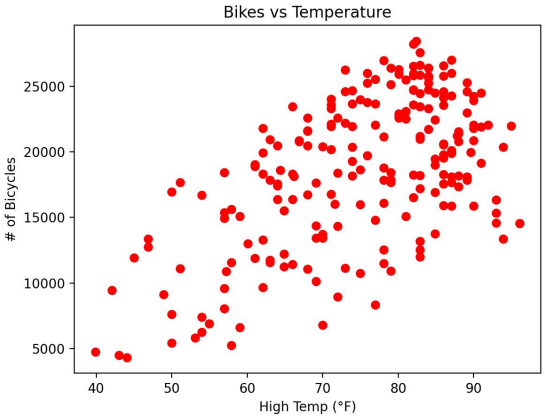
deviation. This made more sense as it was seeing which data were closer in trend i.e which dataset moved like the other. With these results:

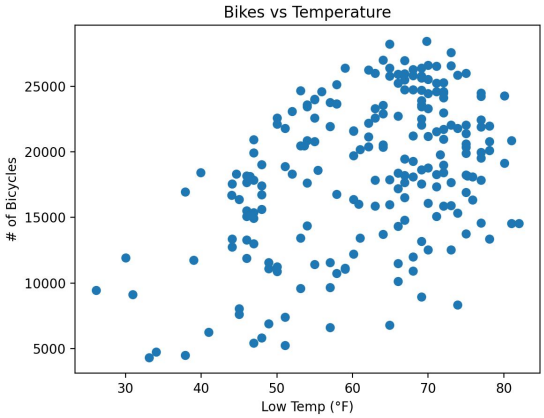
```
Percent within one StDev for Brooklyn: 73.36448598130842
Percent within one StDev for Manhattan: 131.77570093457945
Percent within one StDev for Williamsburg: 194.39252336448598
Percent within one StDev for Queensboro: 256.5420560747663
```

I decided that we should put sensors on Manhattan, Williamsburg and Queensboro as the amount of data in one standard deviation is far below the others.

2. Given the data, the question of can we predict the number of bicycles based on the next day's weather arose. This emerged as a result of the city administration wanting to crack down on helmet law, but only wanting to deploy officers on days of high bicycle traffic.
   In order to find the correlation (if there was any) I first plotted the High and Low Temperature data against the total number of bicycles across all four bridges.

It can be seen that both graphs show a similar distribution of data where the most number of riders are around the 80 degree days. From this it seems feasible that





we CAN make a prediction about bike traffic based on the weather forecast. The next step I took was perform a multivariate analysis, seeing as we have two explanatory variables.

After running a regular linear regression utilizing sklearn tools in Python I was able to get a model of:

$$y = -41176.868 + 1897.149x_1 - 530.779x_2 - 31.325x_3 + 52.250x_4 - 29.770x_5$$

```
R squared value: 0.4585862313204707
Intercept: -41176.86863798903
Coefficients: [1897.14870032 -530.77913883  -31.32537089   52.25041038  -29.77031598]
```

The coefficient of determination (0.458) is not too great, but not bad either, for the fifth degree equation that was calculated. Because of this I ran an ordinary least squares regression using OLS from the statsmodel api. From this I got:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.375
Model:                            OLS   Adj. R-squared:                  0.369
Method:                 Least Squares   F-statistic:                     63.26
Date:                Sun, 02 Aug 2020   Prob (F-statistic):           3.00e-22
Time:                        23:19:56   Log-Likelihood:                -2103.7
No. Observations:                 214   AIC:                             4213.
Df Residuals:                     211   BIC:                             4223.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1526.6832   1884.175     -0.810      0.419   -5240.902    2187.536
x1           483.6102     62.158      7.780      0.000     361.081     606.140
x2          -260.8814     66.817     -3.904      0.000    -392.596    -129.167
==============================================================================
Omnibus:                       22.228   Durbin-Watson:                   1.170
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                8.475
Skew:                          -0.221   Prob(JB):                       0.0144
Kurtosis:                       2.131   Cond. No.                         600.
==============================================================================
```

Unfortunately, the R square value for the OLS regression was even lower. This is likely due to the large variability of the data. This can be seen in the last row of the OLS regression results. The Omnibus, Skew, and JB values all indicate a non-normal distribution and low correlation dataset (values all interpreted with the help of https://www.accelebrate.com/blog/interpreting-results-from-linear-regression-is-the-data-appropriate for definitions and value ranges for interpretation). The OLS regression resulted in an equation of:

$y = -1526.68320352 + 483.61015681 x_1 - 260.88142504 x_2$

Due to these low coefficients of determination I would conclude that it would be somewhat difficult to predict the exact number of bicyclists for the next day. Despite this we are only looking for classifications of 'high' or 'low' traffic. To define high and low I took the top and bottom 6.5% of the total as this would mean it is in the second standard deviation of data.

```
MEAN: 18544.532710280375
Mean of highest 6.5 percent: 26821.714285714286
Mean of lowest 6.5 percent: 6297.571428571428
```

Based on the numbers above it would be reasonable to deploy officers or not depending on if the predicted value from the 5 degree polynomial falls within $\pm 10\%$ of the respective high or low means.

Temperature is not the only attribute of weather, there is also precipitation. So, next, I decided to run a multivariate regression with three explanatory variables: High temp, Low temp, and Precipitation. These were the results:

$y = -35661.144 + 1440.65 x_1 - 102.3006 x_2 - 24737.058 x_3 - 19.237 x_4 + 27.863 x_5 + 285.788 x_6 - 17.208 - 159.714 x_7 + 5628.499 x_8$

```
R squared value three: 0.5886916351725542
Intercept three: -35661.1443893596
Coefficients three: [ 1.44064864e+03 -1.02300650e+02 -2.47370586e+04 -1.92375926e+01
  2.78625391e+01  2.85788128e+02 -1.72084411e+01 -1.59714331e+02
  5.62849932e+03]
```

I was able to get a very high R squared value of 0.588. This indicates that there is a strong correlation between the weather and bike traffic. With this model there is a very good chance of predicting whether bike traffic will be high or low based on the weather forecast, however there is also slight concern about overfitting since it is an 8 degree model. I also ran an OLS regression to see if I could get an even better R squared value, but it was lower, with a value of: 0.499. This value is still good and also came with only a three degree model.

The equation is:

$Y = 178.20093423 + 390.91830834\,x_1 - 162.32007876\,x_2 - 7951.48638461\,x_3$

The summary of the data is below:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.499
Model:	OLS	Adj. R-squared:	0.492
Method:	Least Squares	F-statistic:	69.85
Date:	Mon, 03 Aug 2020	Prob (F-statistic):	2.28e-31
Time:	14:31:23	Log-Likelihood:	-2079.9
No. Observations:	214	AIC:	4168.
Df Residuals:	210	BIC:	4181.
Df Model:	3		
Covariance Type:	nonrobust		

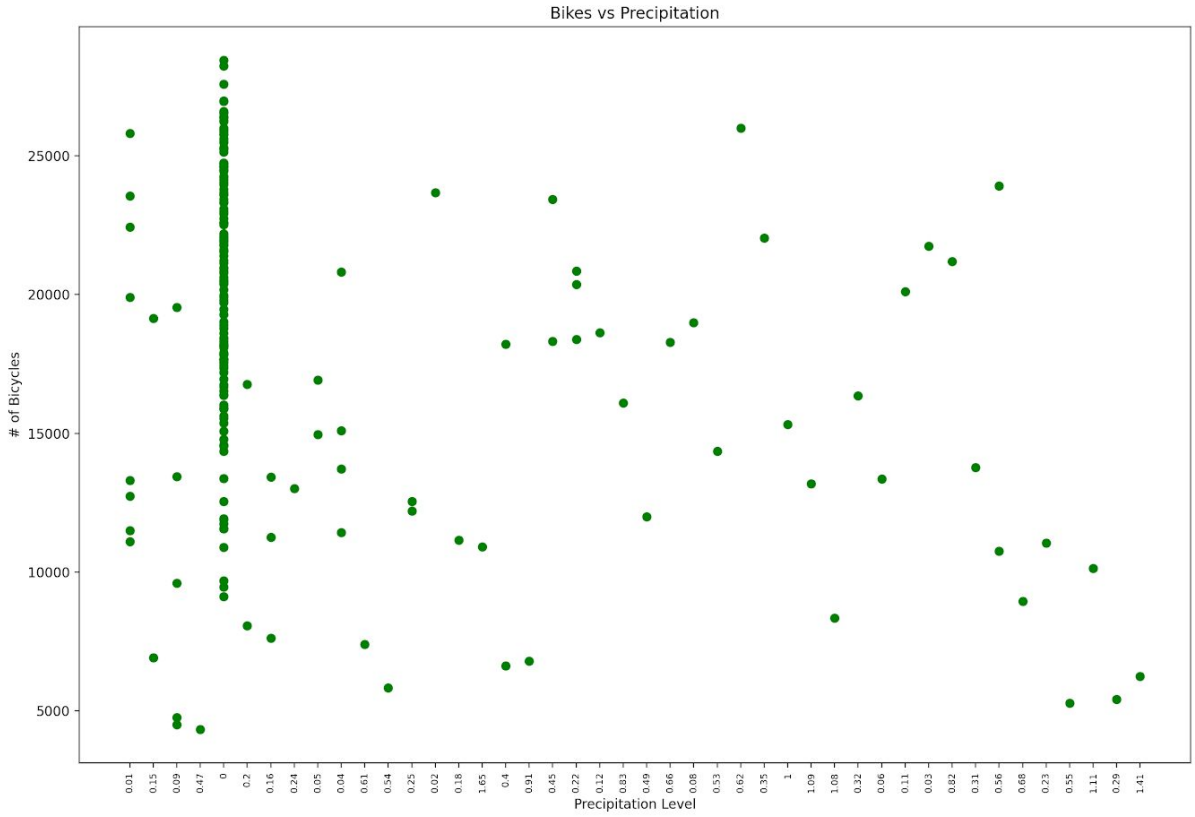
	coef	std err	t	P>|t|	[0.025	0.975]
const	178.2009	1706.345	0.104	0.917	-3185.560	3541.962
x1	390.9183	57.206	6.834	0.000	278.147	503.690
x2	-162.3201	61.461	-2.641	0.009	-283.480	-41.160
x3	-7951.4864	1099.722	-7.230	0.000	-1.01e+04	-5783.576

Omnibus:	10.936	Durbin-Watson:	1.101
Prob(Omnibus):	0.004	Jarque-Bera (JB):	6.251
Skew:	-0.239	Prob(JB):	0.0439
Kurtosis:	2.313	Cond. No.	610.

Finding the correlation between precipitation and bike traffic would also be a useful statistic for this question. This correlation will be discussed more in the next section.

3. Can you use this data to predict whether it is raining based on the number of bicyclists on the bridges?

   To answer this question, as before, I started by plotting the precipitation against the number of bikers, in order to get a general sense of the data.



It can be seen that a large part of the data set contains zero precipitation. When just looking at the zero, it can also be seen that there is still a lot of bike traffic. In order to find a way to predict if it is raining or not depending on the number of bikers I performed a linear regression analysis, with the total number of bikers being the target variable and the precipitation, the target variable. This is because the rain will be depending on the number of bikers for the model. After running the regression I was able to get an equation of:

$$y = 20415.852 - 111356.823 x_1 + 574756.118 x_2 - 1295009.509_3 + 1435353.050 x_4 - 771115.304 x_5 + 159509.850 x_6$$

```
R squared value Precipitation: 0.29188743675499407
Intercept Precipitation: 20415.852674939353
Coefficients Precipitation: [ -111356.82363565   574756.11877149 -1295009.50951372  1435353.05096245
  -771115.30405758   159509.85009463]
```

The reason why I chose a 6 degree model is due to it having the highest r squared value while not having to worry too much about over fitting. To confirm this I computed the r squared value for a degree of 5 (0.28), a degree of 15 (0.30), and a degree of 100 (0.29). It can be seen that a sixth degree model gives us the most bang for our buck. To see if I could attain a better coefficient of determination, I ran an OLS regression which gave an equation of:

$$y = 19551.00239173 - 9228.12818469 x_1$$

The summary of this regression is:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.177
Model:                            OLS   Adj. R-squared:                  0.173
Method:                 Least Squares   F-statistic:                     45.59
Date:                Mon, 03 Aug 2020   Prob (F-statistic):           1.37e-10
Time:                        13:59:10   Log-Likelihood:                -2133.1
No. Observations:                 214   AIC:                             4270.
Df Residuals:                     212   BIC:                             4277.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1.955e+04    384.511     50.846      0.000    1.88e+04    2.03e+04
x1         -9228.1282   1366.665     -6.752      0.000   -1.19e+04   -6534.136
==============================================================================
Omnibus:                        7.670   Durbin-Watson:                   0.850
Prob(Omnibus):                  0.022   Jarque-Bera (JB):                7.100
Skew:                          -0.386   Prob(JB):                       0.0287
Kurtosis:                       2.551   Cond. No.                         3.90
==============================================================================
```

The R-squared value is even lower than the 6 degree model, so it would be better to just stick with the 6 degree model. Now with an R-squared value of 0.29, predicting the number of cyclists based on precipitation may not have accurate

results. There is a definite correlation between the data, however there was a stronger correlation with the Temperature and number of bikers. If we are looking at this from a purely correlative point of view I would say we are there. However, if one is looking for a causal relationship between precipitation and the number of bikers our model might be a stretch.

Overall, I was able to gain a lot of valuable information from these analyses. From part 1 it was shown that Manhattan, Williamsburg and Queensboro are more like each other than Brooklyn is. So it would make sense to put sensors on these three, as the conclusions we arrive at will be more applicable and accurate specifically, for these three. In part 2 we were able to see that there IS a correlation between the number of bikers and the temperature, and an even stronger correlation between the number of bikers and the weather as a whole (temperature and precipitation). Lastly, in part 3 it was deduced that there is a correlation between the precipitation and the bike traffic; however it is not strong enough for a causal relationship.