Ramon Tovar-Yampara/Joseph Wang

ECE 20875 project report

Problem 1:

For the first problem, we want to install sensors across four of the bridges to get the best prediction of overall traffic. However, we can only put sensors on three of the bridges because we have a limited budget. In order to determine which of the three bridges to put sensors on, we did a couple of analyses. The first analysis that we did was to create four different graphs showing the number of bikes used for a specific day at a specific bridge (Brooklyn, Manhattan, Williamsburg, and Queensboro). From figure 1 we can see that the data for the Manhattan, Williamsburg, and the Queensboro are somewhat similar. From this analysis alone we could conclude that these three bridges are the ones we should put sensors on to get the best prediction of overall traffic. However, we did another analysis where we created histograms of all the data showing the frequency for number of bikers on a specific bridge. We calculated the mean, minimum, maximum, difference between the minimum and maximum, standard deviation, data within one standard deviation, and data within two standard deviations for figure 2. The value that we found most important was the data within one standard deviation for each bridge. The data within one standard deviation for Brooklyn was 60%, for Manhattan was 66.666%, for Williamsburg was 66.666%, and for Queensboro was 30%. The higher the percentage for data within one standard deviation meant that the data was very close to the mean of the data. Furthermore, indicating that the data for Brooklyn, Manhattan, and Queensboro is precise. As a result, we suggest putting sensors on Brooklyn, Manhattan, and

Queensboro to get you the best prediction of overall traffic. The reason why we didn't consider important the data within two standard deviations is because all the data (100%) fell within two standard deviations for each of the bridges. Thus, making it meaningless to look at two standard deviations or more.

Figure (1)

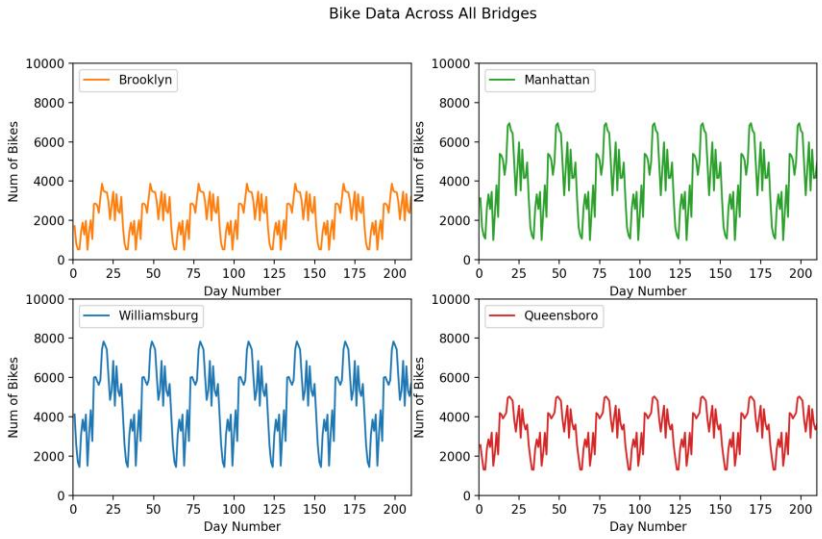
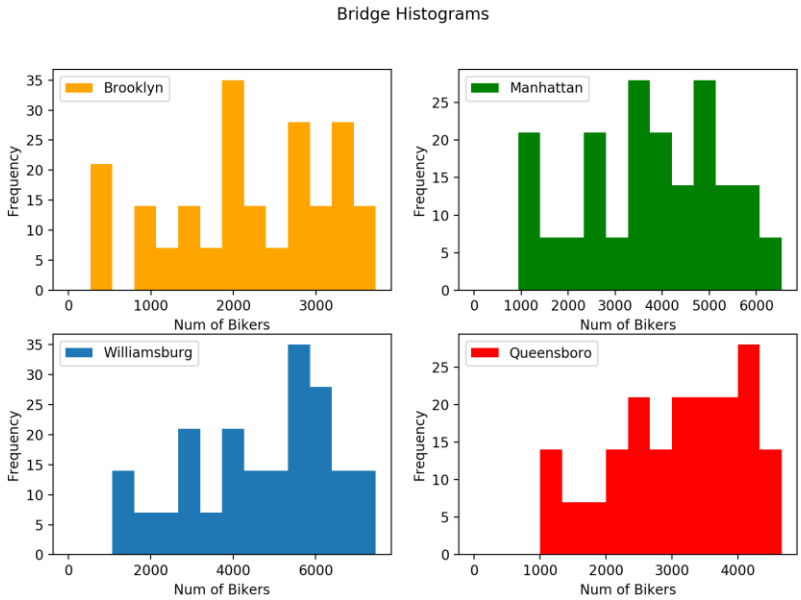


Figure (2)



Problem 2:

The second problem proposed the idea of predicting the number of total riders for a specific day given the high and low temperatures. In short, we found it possible to get a close estimate. The first step in our analysis of this question was first understanding the trends of each feature separately with regards to the total number of bike riders. To do this, we first plotted high and low temperatures to get a visual representation of the data. As seen in figure 3 and figure 4 below, the trend seems to follow a linear, proportional trend. This observation led us to create a 2-feature linear regression model to help predict the data.

We deemed this model to be appropriate for the given data based on a few prerequisites that we checked. The first most obvious prerequisite was that both features shared a linear relationship with the dependent variable. The second prerequisite was that the distribution of the dependent values was roughly normally distributed, meaning they were symmetric, and unimodal. The last prerequisite was that there was no perfect linear relationship for

explanatory values, meaning that the data could not be described by a model with a perfect r^2 value.

In terms of results, our method generated a model with the linear equation y = -7033.9075 + 523.7872*x1 – 218.9591*x2, where x1 is the high temperature input, and x2 is the low temperature input. Our model had an r^2 value of 0.598 and an adjusted r^2 value of 0.594. This means that roughly 60% of the variance of our results can be predicted by the feature values. Although this value is not as close to perfect as we wanted, the model is still great at getting a fair estimate without over modeling the data set provided. A limitation for our model is that it doesn't make fair predictions for extremely low values of Low Temp, and extremely high values for High Temp. As the low and high temp values get closer to positive and negative infinity, you expect the total riders to reach zero. However, since the model is a linear function, this concept would never be realized. For some more in depth results regarding our OLS Regression model, refer to figure 5.

With respect to making actual predictions, we created a graphical user interface that allows the user to enter in a high and low temperature that returns a predicted number of total riders' value. The GUI also features both XY scatter plots that show the linear relationship between both features and the dependent values. Refer to figure 6 for more information.

Figure (3)

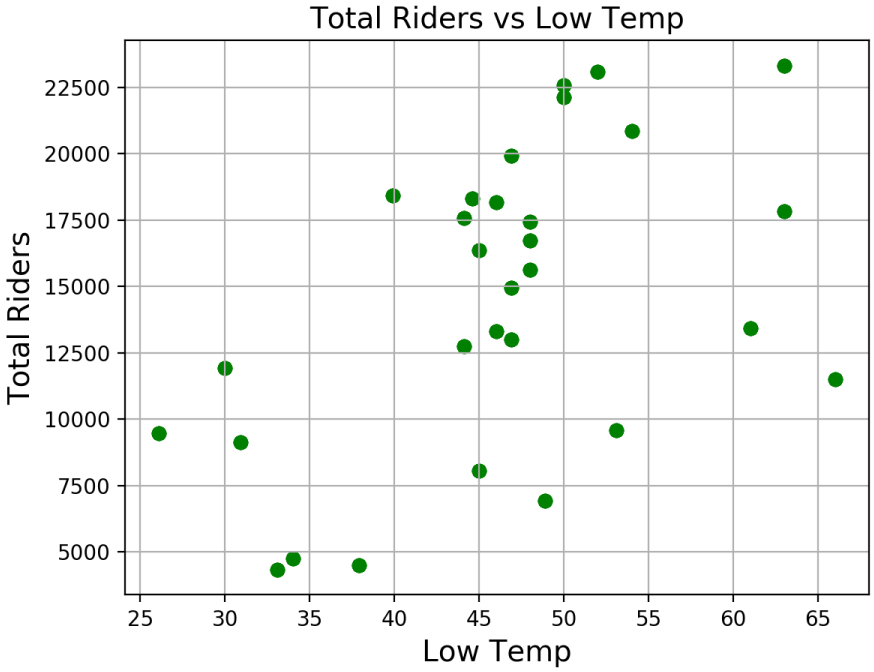


Figure (4)

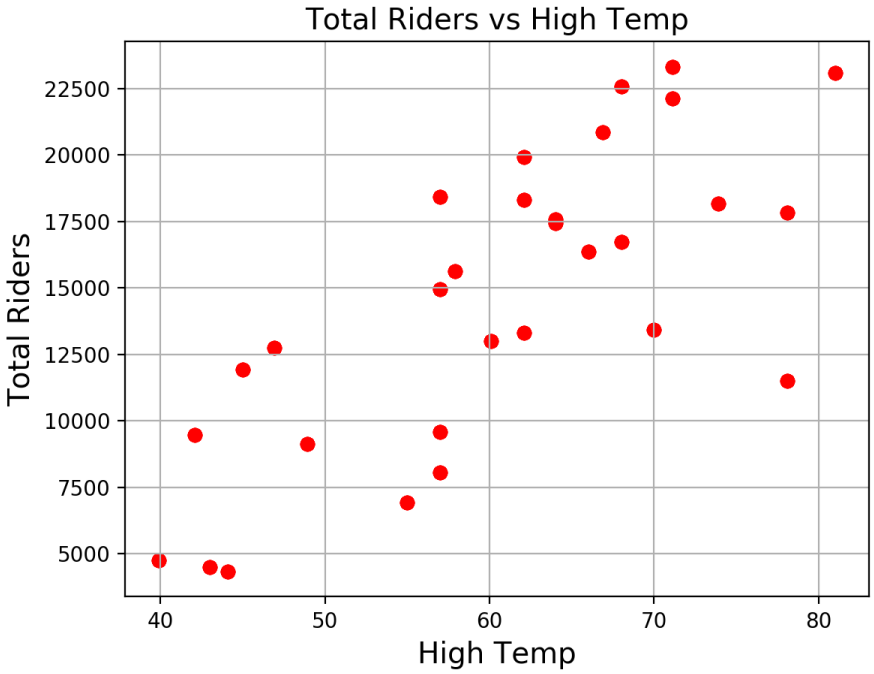


Figure (5)

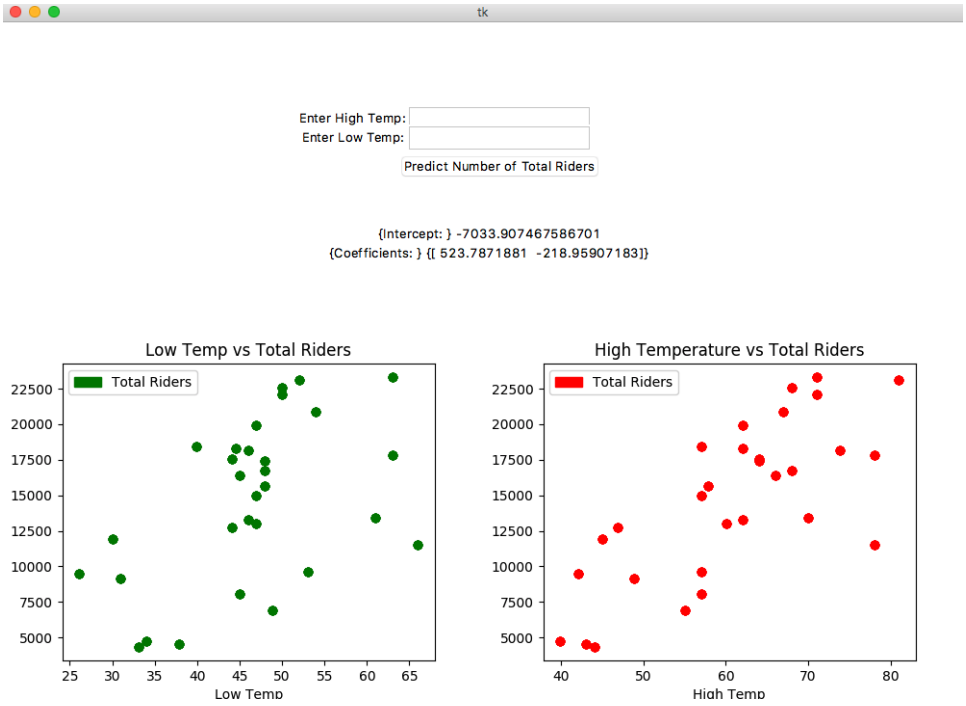
OLS Regression Results

Dep. Variable:	Total	R-squared:	0.598
Model:	OLS	Adj. R-squared:	0.594
Method:	Least Squares	F-statistic:	153.7
Date:	Thu, 05 Dec 2019	Prob (F-statistic):	1.21e-41
Time:	19:14:48	Log-Likelihood:	-2013.1
No. Observations:	210	AIC:	4032.
Df Residuals:	207	BIC:	4042.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7033.9075	1364.191	-5.156	0.000	-9723.397	-4344.418
High Temp	523.7872	38.738	13.521	0.000	447.415	600.159
Low Temp	-218.9591	45.493	-4.813	0.000	-308.647	-129.271

Omnibus:	9.853	Durbin-Watson:	1.508
Prob(Omnibus):	0.007	Jarque-Bera (JB):	4.487
Skew:	0.020	Prob(JB):	0.106
Kurtosis:	2.285	Cond. No.	432.

Figure (6)



Problem 3:

For the third problem, we want to use the data to predict whether it is raining based on the number of bicyclists on the bridges. In order to determine whether it will rain or not, we had to create a regression model using the data from the total number of bikers and the precipitation. One encounter we had with the data is that some data had an S to indicate that is snowing or a T which means that there is no rain. In order to get rid of the S and T, we had to clean the data by erasing the S and changing the T to a 0 in the data. After we had the data, we encountered another problem where in order to plot the data we had to reshape the independent variable (x) because there was only one feature. Then we plotted the total number of bikers' vs precipitation as in figure 7. Finally, we created a regression model and tried various degrees to see which one fit the data the best. In order to determine this, we wanted a regression model that had a r^2 value closest to 1. From our results we found out that the seventh-degree fit best with the data because it had the largest r^2 value of 0.642. The equation of the model was (-9.760*10^5)*x^7+(5.536*10^7)x^6+(-1.203*10^9)x^5+(1.195*10^10)x^4+(-5.871*10^10)x^+(1.372*10^11)x^2+(-1.189*10^11)x^1+(17726.789). We then compared this to the regression model with a degree of one to compare the two. The comparison between the two regression models is shown in figure 8. The equation for this model was (-3.233*10^4)x^1+(1.622*10^4). The seventh-degree regression model is the method you should use in order to predict whether it is raining based on the number of bicyclists on the bridges. You plug in the total number of bikers in the model and it will tell you whether it is raining.

Figure (7)

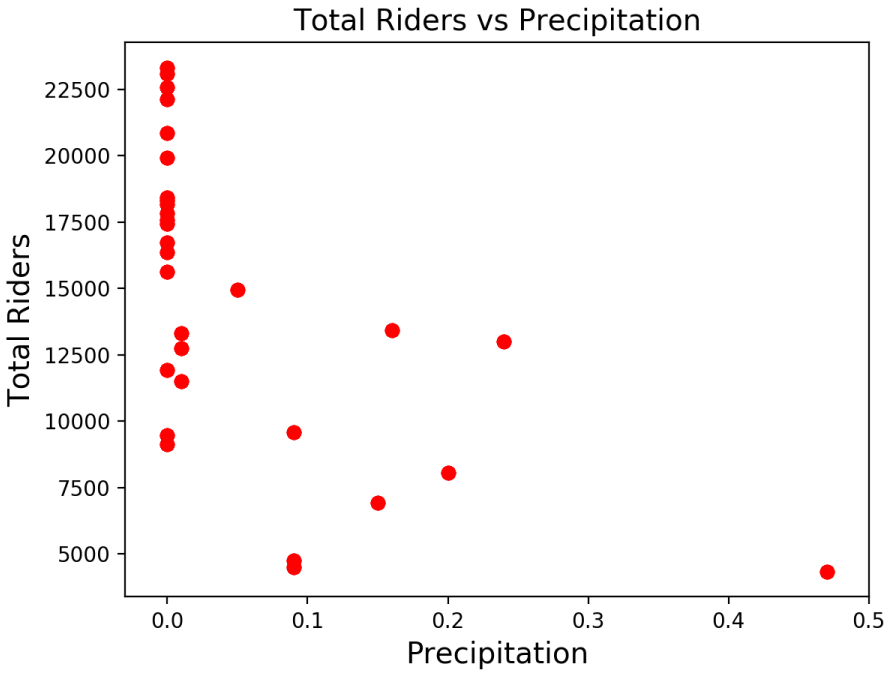


Figure (8)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.349
Model:	OLS	Adj. R-squared:	0.346
Method:	Least Squares	F-statistic:	111.7
Date:	Fri, 06 Dec 2019	Prob (F-statistic):	3.58e-21
Time:	14:40:17	Log-Likelihood:	-2063.6
No. Observations:	210	AIC:	4131.
Df Residuals:	208	BIC:	4138.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.622e+04	349.314	46.437	0.000	1.55e+04	1.69e+04
x1	-3.223e+04	3049.257	-10.568	0.000	-3.82e+04	-2.62e+04

Omnibus:	19.695	Durbin-Watson:	0.750
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9.477
Skew:	-0.320	Prob(JB):	0.00875
Kurtosis:	2.179	Cond. No.	9.84