Project ECE20875: Python for Data Science Spring 2022

1. **Project team information**

Mini-Project Spring 2022
ECE20875
Name 1 – jcpssean - lee3788@purdue.edu
Name 2 – ShaoNingHuang - huan16465@purdue.edu
Path (data set) chosen: 1

2. **Descriptive Statistics**

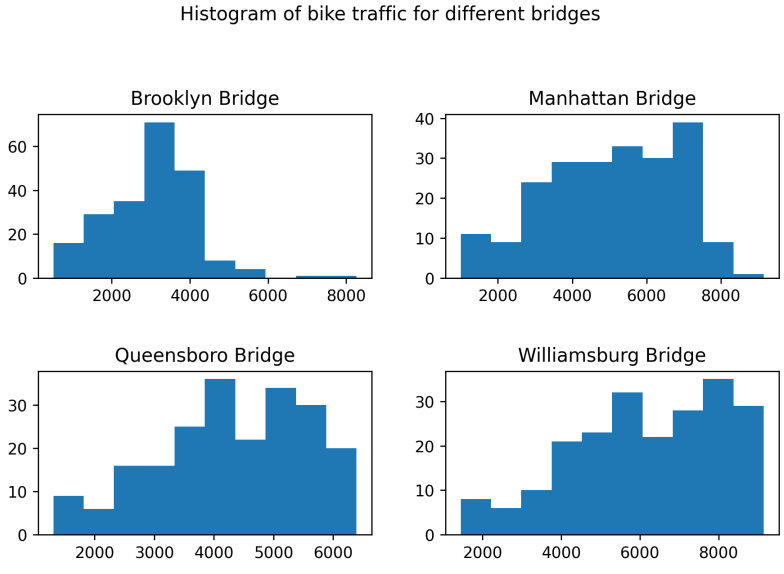
By observing the NYC_Bicycle_Counts_2016_Corrected.csv dataset, there are 8 different variables we can use to tackle the problems. 'High Temp', which is the highest temperature of the day; 'Low Temp', which is the lowest temperature of the day; 'Precipitation', which is the amount of rain or snow of the day; the number of bikes on the 4 individual bridge and the total amount of bikes on the 4 bridges.

Summary Statistics Table:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Date	Day	High Temp	Low Temp	Precipitation	Rain (1 if precipitation > 0)	Brooklyn B	Manhattar	Williamsbu	Queensbo	Total					
2	1-Apr	Friday	78.1	66	0.01	1	1,704	3,126	4,115	2,552	11,497		Brooklyn Bridge	Manhattan Bridge	Williamsburg Bridge	Queensboro Bridge
3	2-Apr	Saturday	55	48.9	0.15	1	827	1,646	2,565	1,884	6,922	Mean	3,031	5,052	6,161	4,301
4	3-Apr	Sunday	39.9	34	0.09	1	526	1,232	1,695	1,306	4,759	Std	1134.044825	1745.485407	1910.643106	1260.985725
5	4-Apr	Monday	44.1	33.1	0.47	1	521	1,067	1,440	1,307	4,335					
6	5-Apr	Tuesday	42.1	26.1	0	0	1,416	2,617	3,081	2,357	9,471					
7	6-Apr	Wednesda	45	30	0	0	1,885	3,329	3,856	2,849	11,919					
8	7-Apr	Thursday	57	53.1	0.09	1	1,276	2,581	3,282	2,457	9,596					
9	8-Apr	Friday	46.9	44.1	0.01	1	1,982	3,455	4,113	3,194	12,744					
10	9-Apr	Saturday	43	37.9	0.09	1	504	997	1,507	1,502	4,510					
11	10-Apr	Sunday	48.9	30.9	0	0	1,447	2,387	3,132	2,160	9,126					
12	11-Apr	Monday	62.1	46	0.01	1	2,005	3,791	4,334	3,182	13,312					
13	12-Apr	Tuesday	57	45	0.2	1	1,045	2,178	2,762	2,082	8,067					
14	13-Apr	Wednesda	57	39.9	0	0	2,840	5,395	5,995	4,192	18,422					
15	14-Apr	Thursday	62.1	44.6	0	0	2,861	5,309	6,030	4,115	18,315					
16	15-Apr	Friday	64	44.1	0	0	2,770	5,072	5,816	3,912	17,570					
17	16-Apr	Saturday	66	45	0	0	2,384	4,316	5,624	4,051	16,375					
18	17-Apr	Sunday	73.9	46	0	0	3,147	4,969	5,867	4,197	18,180					
19	18-Apr	Monday	81	52	0	0	3,871	6,823	7,432	4,964	23,090					
20	19-Apr	Tuesday	71.1	63	0	0	3,501	6,951	7,834	5,032	23,318					
21	20-Apr	Wednesda	68	50	0	0	3,450	6,574	7,639	4,928	22,591					
22	21-Apr	Thursday	71.1	50	0	0	3,436	6,452	7,426	4,813	22,127					
23	22-Apr	Friday	78.1	63	0	0	2,975	4,907	6,093	3,862	17,837					
24	23-Apr	Saturday	70	61	0.16	1	2,055	3,276	4,856	3,239	13,426					
25	24-Apr	Sunday	68	48	0	0	2,798	4,650	5,335	3,957	16,740					
26	25-Apr	Monday	66.9	54	0	0	3,463	5,978	6,845	4,564	20,850					
27	26-Apr	Tuesday	60.1	46.9	0.24	1	1,997	3,520	4,559	2,929	13,005					
28	27-Apr	Wednesda	62.1	46.9	0	0	3,343	5,606	6,577	4,388	19,914					
29	28-Apr	Thursday	57.9	48	0	0	2,486	4,152	5,336	3,657	15,631					
30	29-Apr	Friday	57	46.9	0.05	1	2,375	4,178	5,053	3,348	14,954					
31	30-Apr	Saturday	64	48	0	0	3,199	4,952	5,675	3,606	17,432					
32	1-May	Sunday	50	45	0.16	1	2,634	1,525	2,062	1,408	7,629					

In problem 3, we converted the Precipitation variable into a binary Rain variable which is 1 if the value of precipitation if greater than 0. And we used the mean and standard deviation of the number of bikes for problem 1.

These are the histagram of bike traffic for different bridges which we will be using it for problem 1.



3. Approach

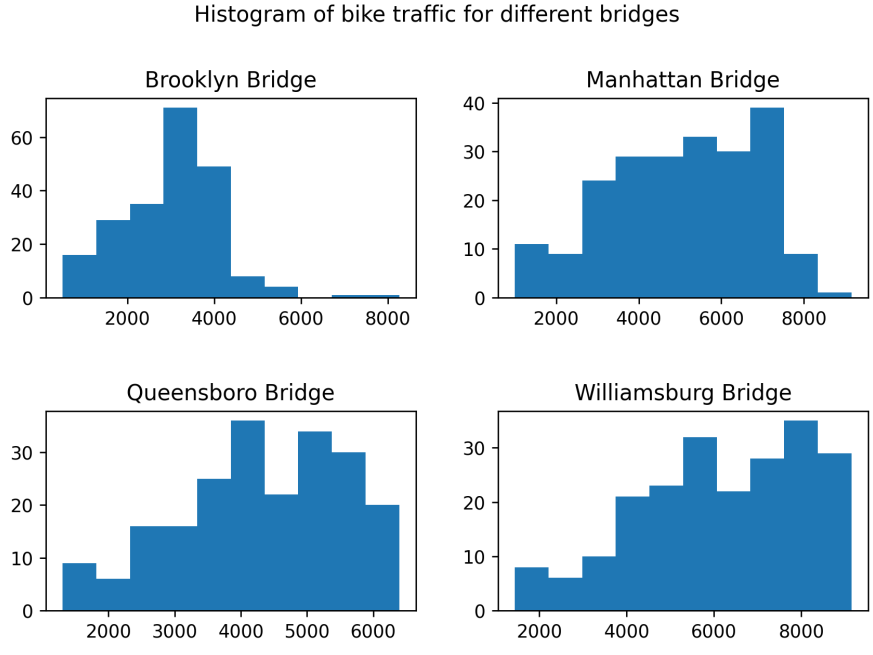
For problem 1, we first subplot the distribution of data, which are the number of bike traffic in each city. We then plot the histogram of the data, number of bike traffic in each city. Finally, we calculated the percentage of data within one standard deviation range of the mean in each city.

For problem 2, we first visualize the relationships between Low temp and Total, High Temp and Total, Precipitation and Total, the temperature difference and Total. We found that except for the temperature difference, the other 3 parameters seem to have significant effect on the number of bikes on that day. So we decided to consider 'High Temp', 'Low Temp', and 'Precipitation' as features and apply linear regression to predict the total number of bikes. We split the data by 80% for training and 20% for testing. Then after the linear regression model is fitted, we apply test it with the testing data and check the error to see if linear regression is a feasible approach to this problem.

For problem 3, we first plot the scatter of total number of bike traffic and the precipitation, and it seems that there was no obvious polynomial model or linear model that cab be used to fit the data. During the process of plotting, we faced a problem that the comma in total number of traffic would cause problem, so we had to get rid of commas in the total number data. We used logistic regression for this problem, we set the data to 1 if the precipitation is greater than 0, else set to 0. Finally, we built our model.

4. Analysis

In problem 1, we assume that we should place our sensors at Manhattan, Queensboro, Williamsburg after we see the plot of the data distribution. Then, the histogram of number of bike traffic in each city solidify our assumption (histograms shown below). However, we still need a quantitative proof to back our assumption. According to the percentage of data within one standard deviation range of each city, we can see that Brooklyn has 72% of the data within one standard deviation, Manhattan has 58% of the data within one standard deviation range, Queensboro has 62% of the data within one standard deviation range, and Williamsburg has 62% of the data within one standard deviation range. Thus, we suggest placing censors at Manhattan, Queensboro, and Williamsburg.



In problem 2, after fitting the linear regression model with the proposed features, we got the model with interception = -404.55833226787945 and the list of coefficients is [401.5002207 -170.78716459 -7171.71486332]. With the values we got, we can predict the total number of bikes with the equation:

$$\widehat{\text{total}} = -404.55833226787945 + 401.5002207 * \text{High} - 170.78716459 * \text{Low} - 7171.71486332 * \text{Precipitation}$$

After obtaining the equation, we yield an accuracy of 0.7496750269174777 within Total_hat and Total (predicted value and ground truth). Therefore, we can conclude that it is feasible to predict the total number of bicyclists that day by observing the weather forecast (low/high temperature and precipitation).

In problem 3, after the result that we use logistic regression to predict the data, we yield an accuracy of 0.7441860465116279 within y_predicted and y_test so we can conclude that we can use the total number of bike traffic to predict whether it is raining or not.