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

1. **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.

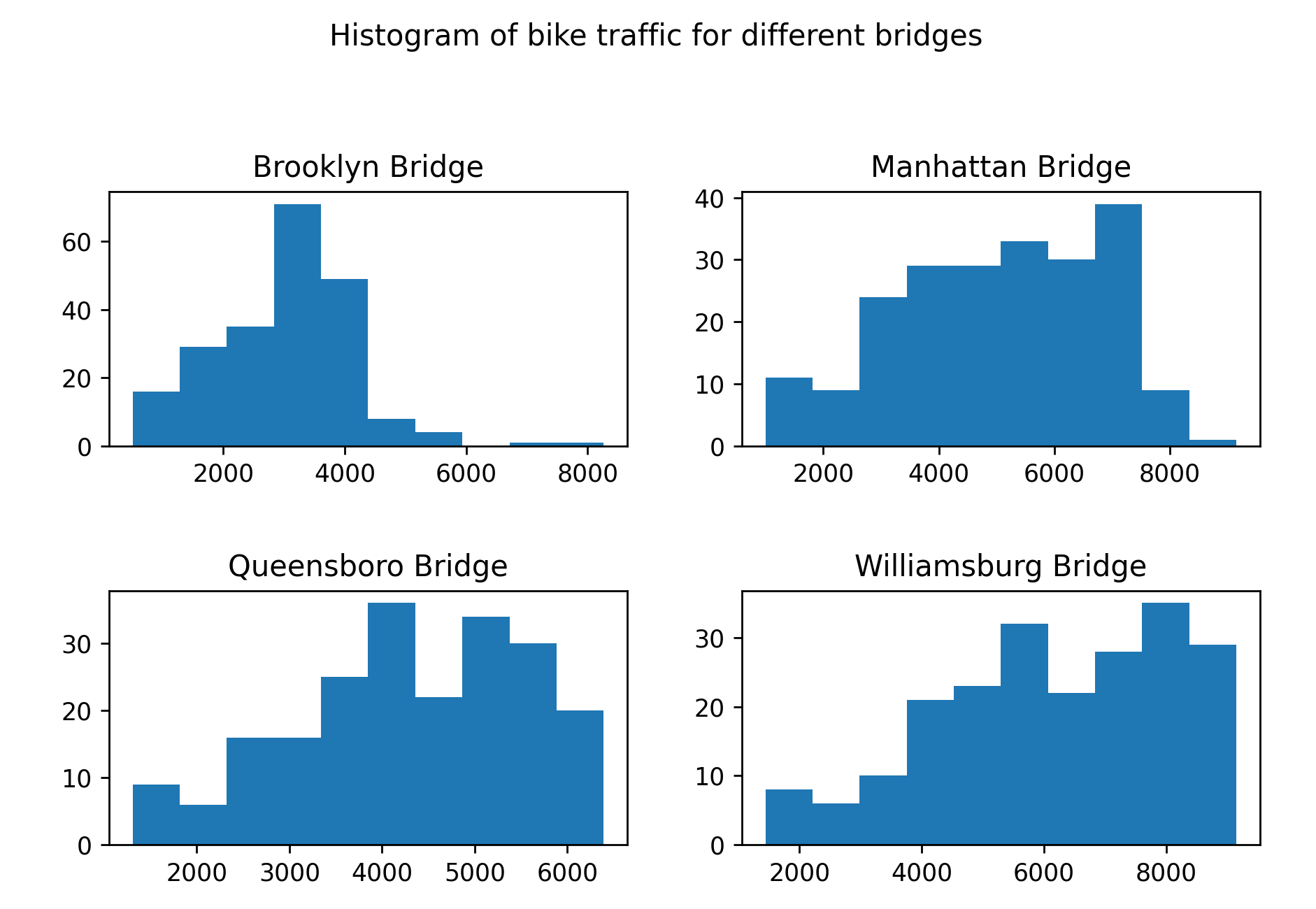
Summary Statistics Table:

一張含有 桌 的圖片

自動產生的描述

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.



1. **Approach**

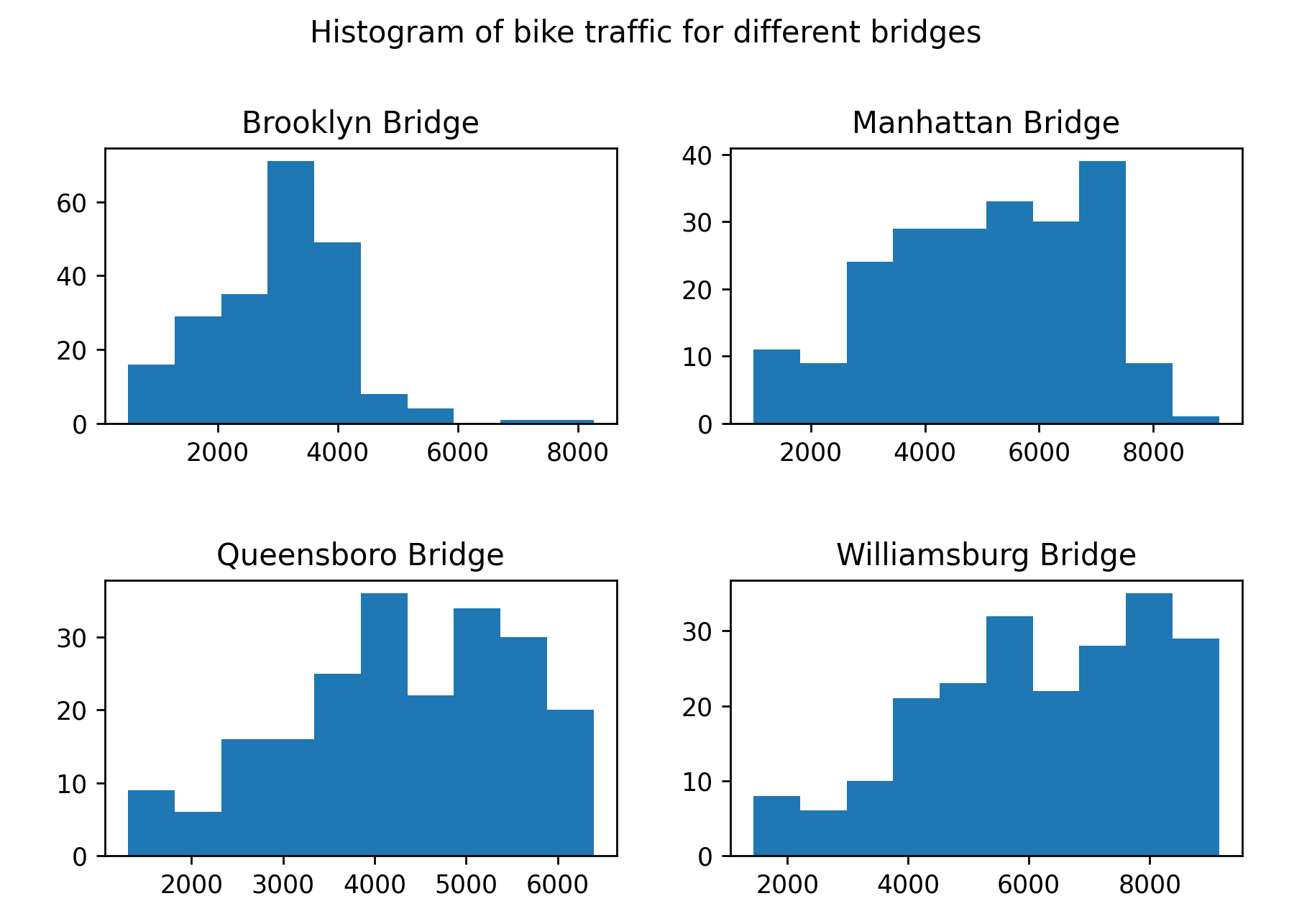
For problem 1, we first subplot the distribution of data, which are the number of bike traffic in each city. We then plot the histogram of the data, number of bike traffic in each city. Finally, we calculated the percentage of data within one standard deviation range of the mean in each city.

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.

1. **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:

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.