Project ECE20875: Python for Data Science Spring 2022

1. Project team information

Mini-Project Spring 2022

ECE20875

Samitha Ranasinghe – Ranasinghe843 – sranasi@purdue.edu

Prajesh Sivakaran – PrajeshS – psivakar@purdue.edu

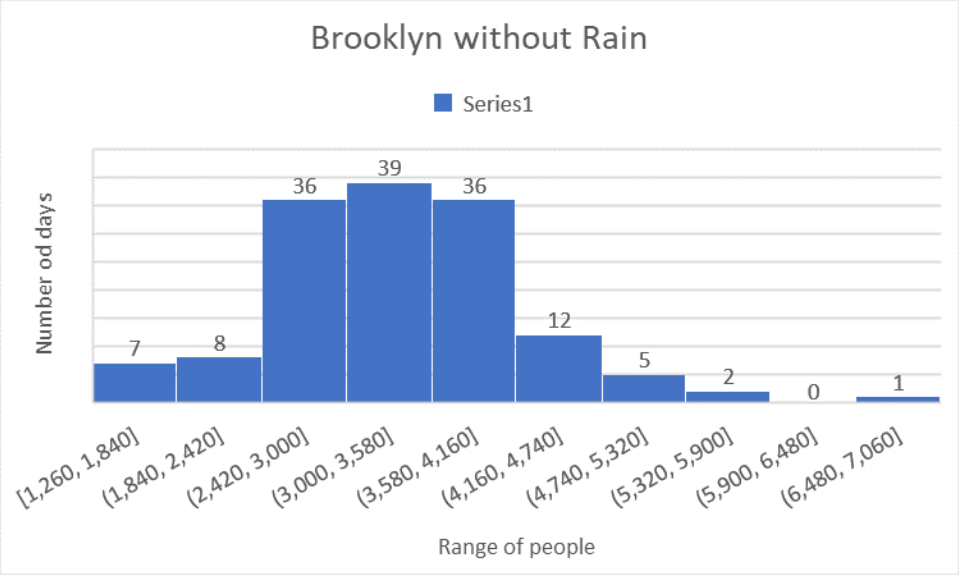
Path: Path 1

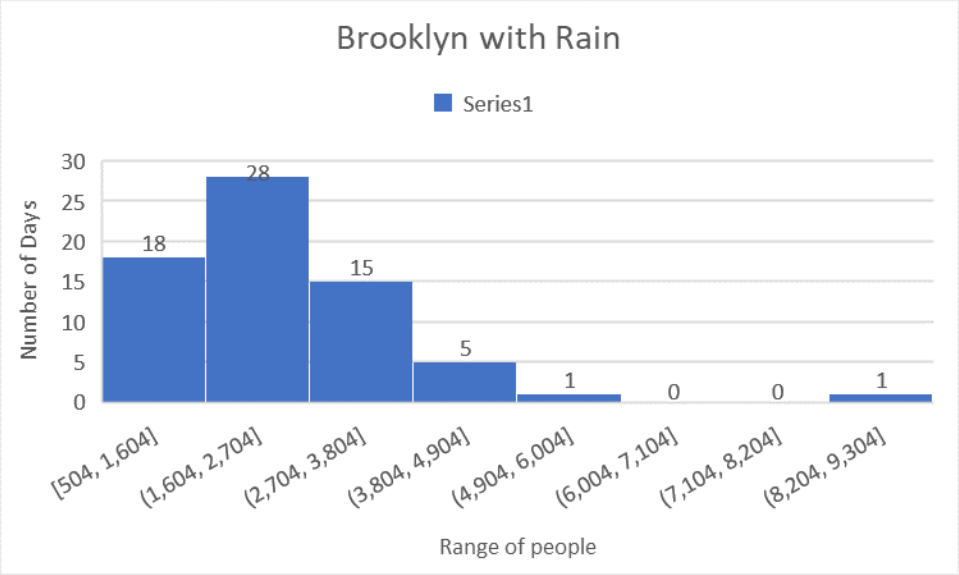
2. Descriptive Statistics

Data Set: NYC_Bicycle_Counts_2016_Corrected

The variables present in the data set are:
Date – The date
Day – The day of the week
High Temp – Highest temperature of the day in degrees Fahrenheit
Low Temp – Lowest temperature of the day in degrees Fahrenheit
Precipitation – Rainfall in inches
Brooklyn Bridge – Bike usage on the Brooklyn Bridge
Manhattan Bridge – Bike usage on the Manhattan Bridge
Williamsburg Bridge – Bike usage on the Williamsburg Bridge
Queensboro Bridge – Bike usage on the Queensboro Bridge
Total – The total for bike usage on four bridges in each day

Variable	Mean without Rain	Mean with Rain
Brooklyn Bridge	3367	2309
Manhattan Bridge	5635	3801
Williamsburg Bridge	6808	4772
Queensboro Bridge	4702	3440
Total	20512	14321





Variable	Mean
High Temp	74.933
Low Temp	61.972
Precipitation	0.109

3. Problem Statements

Problem 1:

You want to install sensors on the bridges to estimate overall traffic across all the bridges. But you only have enough budget to install sensors on three of the four

bridges. Which bridges should you install the sensors on to get the best prediction of overall traffic?

Problem 2:

The city administration is cracking down on helmet laws, and wants to deploy police officers on days with high traffic to hand out citations. Can they use the next day's weather forecast (low/high temperature and precipitation) to predict the total number of bicyclists that day?

Problem 3:

Can you use this data to predict whether it is raining based on the number of bicyclists on the bridges (hint: The variable *raining* or *not raining* is binary)?

4. Approach

Problem 1:

To try to single out the least affecting bridge, we used linear regression to create four models where the features are the traffic of three of the four bridges and the target variable is the total traffic. Each model would contain a unique combination of three bridges so that by choosing the model with the highest accuracy, we can obtain the bridge with traffic that contributes the least to the prediction of the total traffic.

We examined each model's r-squared (r2) value, which shows how the model represents the data we're trying to model, to determine the best fitting response. R-squared is a number that varies from 0 to 1, with 0 indicating that the model does not accurately describe the dataset and 1 indicating that it does.

Problem 2:

In this problem, we're attempting to establish a link between weather and traffic data by modelling the data using linear regression with high temperature, low temperature, and precipitation as features and the total traffic as the target variable. The data is split into training and testing subsets and the analysis done.

Problem 3:

A problem like Problem 2, we are trying to find out a relationship between the total number of cyclists and whether it is raining. Since the variable in question is binary and the data can be classified into two groups based on whether it is raining or not, a Naive Bayes classification method is tested out (a reason is the two graphs obtained for descriptive statistics show two possible gaussian distributions for the classified data). The data is split as usual into training and testing data and the accuracy score taken.

5. Analysis

Problem 1:

The r-squared value that resulted from each model are:

Combination of bridges	Brooklyn, Manhattan, Williamsburg	Brooklyn, Manhattan, Queensboro	Brooklyn, Williamsburg, Queensboro	Manhattan, Williamsburg, Queensboro
r-squared value	0.995728908	0.9881826020	0.9472389433	0.9821507190

As can be seen, the highest r-squared value is given by the model where the Queensboro bridge is missing. Therefore, the sensors should be installed on the Brooklyn, Manhattan, and Williamsburg bridges to get the best prediction of overall traffic.

Best Model Obtained:

Total Traffic = 1.1386000788715267*(Brooklyn)

+ 0.9471171505682368*(Manhattan)

+ 1.6086469611158551*(Williamsburg)

+ 382.7456681782314

Problem 2:

The model obtained in trying to predict the total number of cyclists using next weathers forecast had an r-squared value of 0.41443078726852955 and a mean squared error of 12767729.059751242.

The large mean squared error and the low r-squared value indicate that there is low relationship between the total number of cyclists and that days weather forecast and therefor it Is not recommended for the city administration to use this method to hand out citations.

Model Obtained:

Total = 412.84453725844236*(High Temp)

- 177.35310433687772*(Low Temp)

- 7957.728673516756*(Precipitation)

- 555.2757165284638

Problem 3:

The Gaussian Naïve Bayes model that was used to test the data returned a score of 0.8372093023255814

The model was able to correctly predict 36 out of the 43 test data as seen in the confusion matrix below:

	Predicted Values		
Actual Values		True (1)	False (1)
	True (1)	29	2
	False (0)	5	7

With a model accuracy of 83.72, it is safe to say that this data can be used to predict whether it is raining based on the number of cyclists in the bridges.