Aniketh Bandi, bandi4

Artus Mosquet, amosquet

Section 1

Path 2: Bike Traffic

Repository Link

Dataset

The dataset is structured as a table with about 214 rows and 10 columns. Each column holds a description for a certain type of data. The columns listed are Date, Day, High Temperature, Low Temperature, Precipitation, Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge, Queensboro Bridge, and Total. The Date column is the date the data for that specific row was recorded. The Day column is the day of the week of the corresponding date. High Temperature is the highest temperature recorded on that day. Low Temperature is the lowest temperature recorded on that day. Precipitation is the amount of precipitation in inches recorded for that day. The 4 bridge columns show the count of bikers that pass or use each respective bridge for that day. The data is best viewed using numpy Arrays to hold data of specific columns or recreate the entire dataset, which was done in our code.

Analysis

Part 1 requires the user to figure out out of 4 bridges, which bridge should be excluded from having sensors installed. To find this out, the standard error of each bridge's bike counts had to be found. Once the standard error for each bridge was found, they were all compared to find the bridge with the highest standard error. This bridge would then be the bridge to not have sensors installed. The reason for this is that the sensors are supposed to help estimate overall traffic across the city. When a bridge has a high standard error, it means there's more random probability, and is worse at consistently estimating how traffic will be. So finding the bridge with the highest standard error gives the bridge with the most amount of randomness, making it less reliable for getting a precise measurement of traffic.

Results

Part 1:

Looking at the results for part 1, we see the program returns the string, "Put sensors on every bridge but the Williamsburg Bridge." This means the Williamsburg Bridge was selected as the bridge to not have sensors, and that it had the highest standard error out of all the bridges. It also means the program functions are intended as otherwise it would have returned "Something went wrong." Ultimately this means that the Williamsburg bridge simply had far more random probability in its counts, making any sensor at that bridge far less precise than the 3 other options.

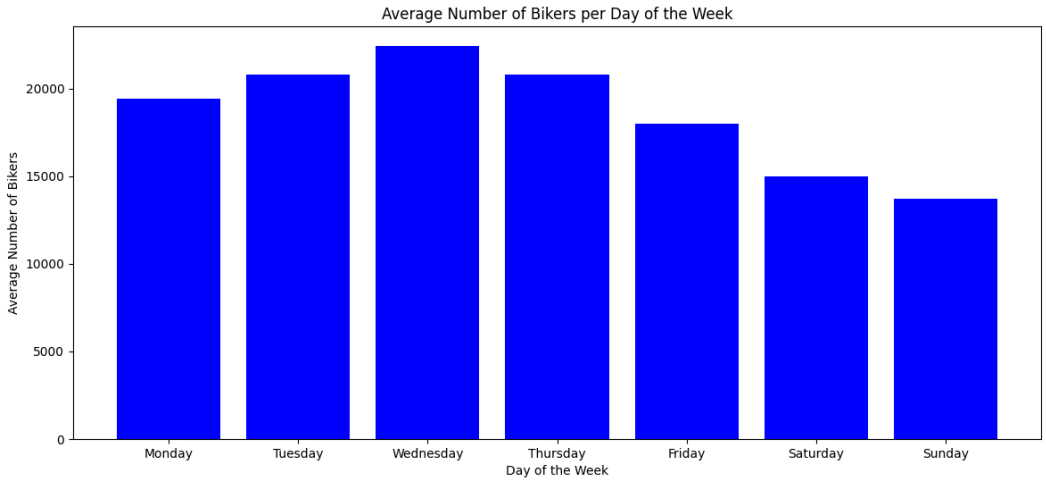
Part 2:

Part 2 asks you to figure out if the next day's weather forecast can be used to determine the total number of bikers on that day. The overall purpose of this is to utilize the weather to decide how many police officers to dispatch to enforce helmet laws for bikers. Given the dataset we have, the idea is to figure out trends in the total number of bikers per day based on the correlating daily data of low temperature, high temperature, and precipitation.

We used a linear regression model to try and determine if this was possible. After running it, we were able to get a correlation coefficient of 0.63, which is good. This means that it would be feasible to predict how much traffic there will be on the bridges to dispatch enough police officers.

Part 3:

For Part 3, the goal is to figure out if there are any trends or patterns related to the total number of bikers due to the day of the week. The end goal of this objective is to see if based on the total number of bikers on a specific day, it is possible to predict the day of the week.



Day of Week	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Averages	19394	20782	22422	20781	17985	15001	13716

The data above does show a trend of weekly bike traffic on the bridges, we start off at ~19,394 people then increase to ~22,422 on Wednesday, then start to go back down with Sunday being the lowest. This trend is to be expected considering people are more likely to be at home over the weekend. The main task here however was to determine if we can use this traffic to predict the current day of the week. This is arguably the most difficult part of this task. We determined that a Logistic Regression would be best for this because we are dealing with categorical values where we need meaningful probabilities for classification. Based on our outputs. Just from a quick glance at the average values per-day of the week, between Friday and Saturday, there's only a ~1,000 person difference. Monday and Tuesday also have only a ~1,400 person difference. These values create a very small margin of error which makes it difficult for us to be able to accurately predict the day of the week. Our model reflects this by only being able

to achieve an accuracy of 25.58%, which is incredibly low. In summary, it would be unwise to use the number of bicyclists to predict the day of the week.