Sebastian Montoya

Professor Brinton

ECE20875

May 1, 2020

Project report – Path 1

**Data:**

Note: I edited the csv such that the (S) and the multiple Ts would not appear. I deleted the "(S)" and converted all the T's into 0.

The data we worked with was well laid out and I would say was self-explanatory. It had the following form:

Date, Day, High Temp (°F), Low Temp (°F), Precipitation, Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge, Queensboro Bridge, Total.

Each of these elements had their own column and it was assigned information for 214 days. To avoid having to do comparison in temperature for High and Low, I created a new column that was calculated by averaging max and min temperature for that day. I then proceeded to add that column into the table given by the CSV.

**Reason for analysis:**

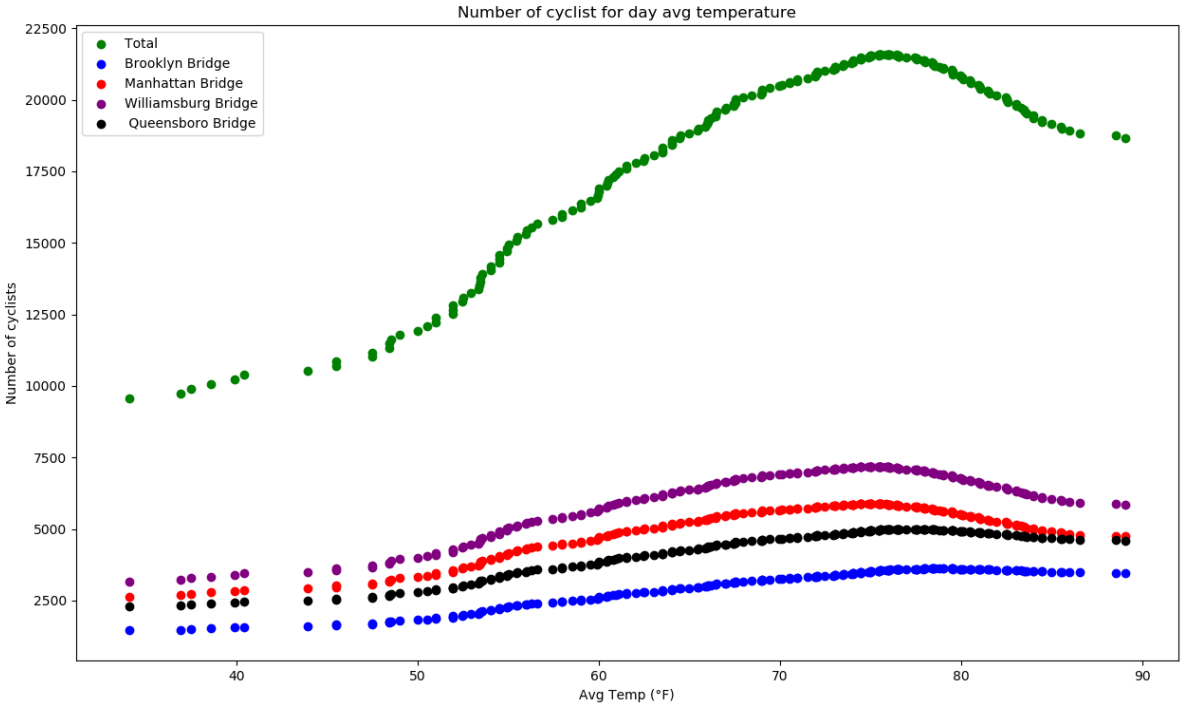
Question 1: I decided to visualize the data because it allows us to observe patterns hidden in the numbers. I also decided to plot number of bikers vs avg temperature because I though it would be the most significant factor influencing traffic and it turned out to be true. Lastly, I decided to make this plot because I wanted to see how different parts of NY would react to temperature changes, which would enable me to draw conclusions about the cites and how their responses related to the overall results.

Question 2: I decided to calculate r score for different factors vs number of bikers because this is an extremely good tool that enables us to see how variables correlate. I also decided to calculate the r score for temperature and precipitation because many times different parameters have different effects on the data. In this case, one parameter could have a stronger relation with the number of bikers than the other. Lastly, I also decided to verify traffic vs day of the week because many times traffic varies depending on day of the week, and it also happened to be the case in this scenario.

Question 3: For this question I also decided to plot the data because, again, it allows us to observe patterns may be hidden in the numbers. (In this case there was none). I also decided to calculate r-score because, again, it is a good tool to analyze how two variables correlate. A good aspect about this question was that it was related to the previous questions, which allowed me to refer to data in the past questions to answer this one.

**Results:**

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?




The plot above is the number of cyclist vs the average temperature for all 4 bridges and their total. For this part, I used the savgol_filter() function in order to smoothen the data given that it had tremendous amounts of noise. As the legend indicates, green is the total number of bikers, blue Brooklyn Bridge, red Manhattan bridge and, purple Williamsburg bridge, and black is Queensboro Bridge.

One thing we can observe on all graphs is that they follow a very similar pattern (all have very similar shape) where the amount of bike traffic increases consistently until around 77 °F, where it starts to decrease and stabilizes at around 85 °F on all 5 lines. If they had different shapes, then it would be harder to argue which places should have cameras because the patterns might change with the temperature and we would be force to make other decisions. Given all Bridges follow a similar pattern, the best 3 places to place the cameras would be the places with the highest traffic: Williamsburg Bridge, Manhattan Bride, and Queensboro Bridge.

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 to predict the number of bicyclists that day?

For this scenario there are number of factors to consider: temperature, precipitation, and day of the week.

Temperature:

| | r | probability |
|---|---|---|
| Total | 0.5214186484476566 | 2.567290020823888e-16 |
| Clean total | 0.5214186484476566 | 2.567290020823888e-16 |
| Brooklyn Bridge | 0.5447224598704488 | 6.206241733154843e-18 |
| Manhattan Bridge | 0.4126040779952521 | 3.328016452654099e-10 |
| Williamsburg Bridge | 0.4754589687627163 | 1.8086501124556565e-13 |
| Queensboro Bridge | 0.576379511426513 | 2.427521371047739e-20 |

The table above is a table that was calculated using the Pearsonr function which returns r, the correlation between the two inputs. The inputs were the avg temperature every day, and the number of bikers for Total, Clean data, and Brooklyn, Manhattan, Williamsburg, and Queensboro Bridges. The r coefficient displayed on the table is the correlations between the labels on the left and the avg temperature. What the coefficients mean is how are the avg temperature and the label on the left are related. The bridge who had the highest correlation with the average temperature was Queensboro, 0.5763. The one with the lowest r value was Manhattan Bridge, 0.4126. What these values mean is that in the case of Queensboro, 0.5763^(2) of the variation is correlated. In the case of Manhattan 0.4126^(2) is correlated. These are decent values, meaning that there is indeed a direct correlation between the weather and the number of bikers outside. In this case, the higher the average temperature, the higher the number of bikers.

Additionally, I believe this number is slightly lower than it would be if temperatures only went up to around 77 °F because after that, the number of bikers decreases according to the plot in the first page. The plot on the first page, Number of Cyclists for day avg temperature, clearly evidences that there is a relationship between temperature and number of bikers as the smooth line created by the scatterplot indicates. What this means is that the relationship between average temperature and number of bikers is not a linear one, but rater a polynomial one. With this in mind, we can say there the weather forecast for temperature could be used to determine number of bikers. In my opinion this is also the most important factor.

Precipitation:

| | r | probability |
|---|---|---|
| Total | -0.4207113087552627 | 1.37339039815829e-10 |
| Clean total | -0.4207113087552627 | 1.37339039815829e-10 |
| Brooklyn Bridge | -0.3388580712586531 | 3.7922272360714444e-07 |
| Manhattan Bridge | -0.4105931111937938 | 4.130033023746216e-10 |
| Williamsburg Bridge | -0.4232468587701603 | 1.0362298020711196e-10 |
| Queensboro Bridge | -0.38802414926865153 | 4.231902725352678e-09 |

The figure above is a table similar to the one in the temperature, but now the difference is that instead of the correlation between temperature and number of bikers, it's between precipitation and number of bikers. In this case, the r values are negative, meaning that as the amount of precipitation increases, the number of bikers decreases. The Bridge that had the greatest correlation was Williamsburg Bridge, -0.4232. The one with the lowest was Brooklyn Bridge, -0.33885. What this means is that if the amount of precipition increased, the fraction of people who would not ride their bikes in Williamsburg would be greater than those in Brooklyn. Again, the range of r between -0.33885 and -0.4232 is not the best, but it does demonstrate that there is a relationship between rain and the amount of people that ride their bikes. As a result, the police could use next days precipitation forecast to have a better estimate of the number of bikers in which the higher amount of precipitation indicates less bikers, but it is the least important factor in determining the number of bikers and should not be considered the sole metric to estimate them.

Day of the Week:

Average number of bikers day of the Week

| Monday | Tuesday | Wednesday | Thursday | Friday | Saturday | Sunday |
|---|---|---|---|---|---|---|
| 19393.71 | 20782.27 | 22422.27 | 20781.3 | 17984.58 | 15000.64 | 13716.38 |

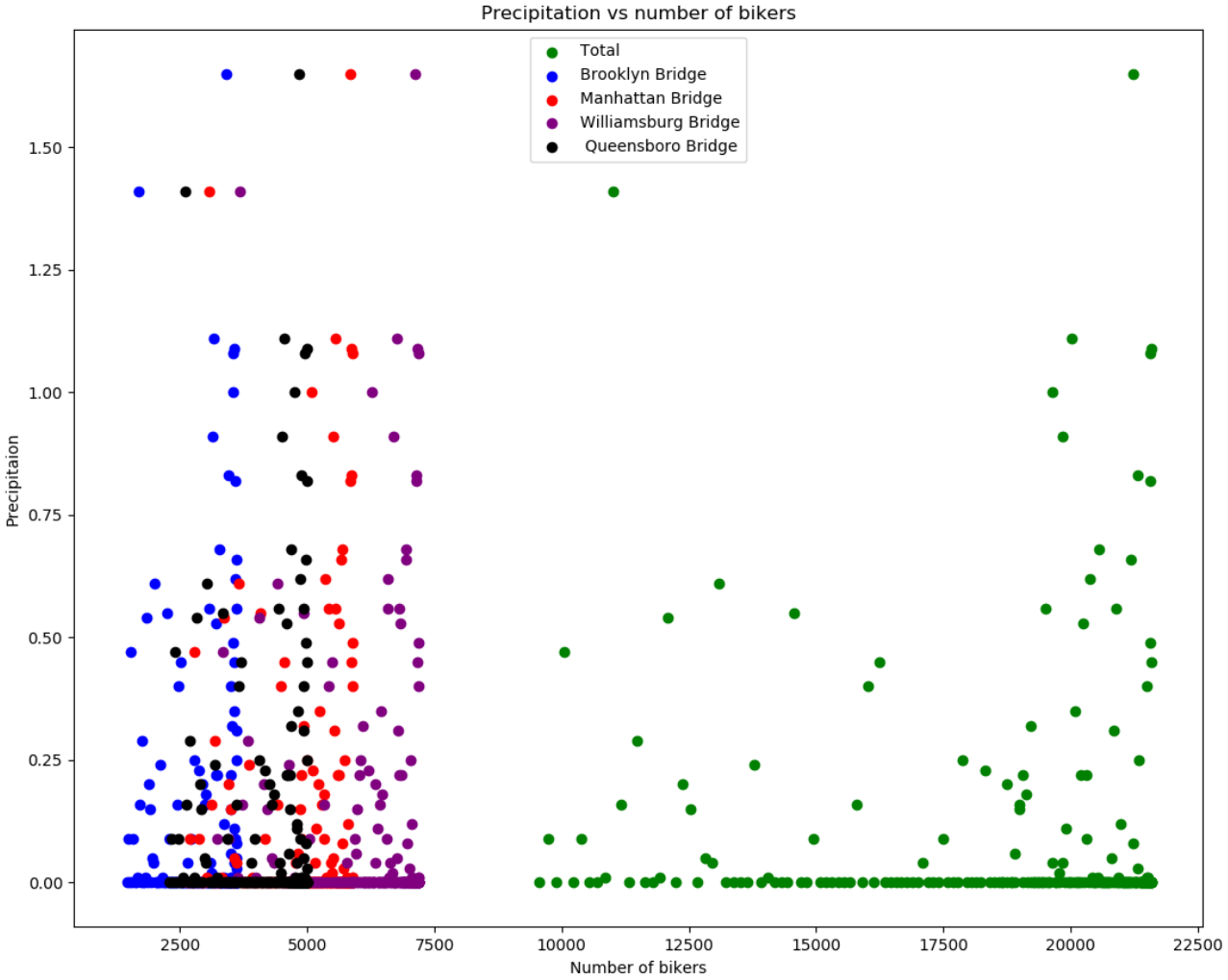
The table above shows the average number of bikers per day of the week. The Data given in the csv has 214 countinous days, If we divide The datapoints (214) by days of the week (7), we get close to 30 days per day of the week. Given that the number of samples per day of the week is at least 30, we can say that they all have a normal distribution by the Central Limit Theorem. Therefore the average number of bikers per day given is acceptable under any circumstance and is significant. What we can observe is thar the number of bikers on Monday through Thursay is relatively the same 21,000 average bikers +- 5%. However, on Friday through Sunday, there is a significant drop in the number of bikers. It drops close to 15% from 21,000(avg number of bikers during weekdays) to around 18,000 on Fridays, Around 25% from 21,000 to 15,000 bikers on Saturdays, and around 33% from 21000 to almost 14000 on Sundays. This clearly demonstrates that the day of the week has a significant effect on the number of bikers. Monday through

Thursay the city administration can expect around 21000 bikers in average, but the number is significantly lower on Friday through Sunday, and thus can plan accordingly.

All around:

All around I would say that the administration can in fact use the weather forecast to predict the number of bikers the next day. Where Friday through Sunday the number is signifincatly lower than that on Moday through Thursday. The number of bikers increases as the temperature goes from lows to around 77 °F, but starts to decrease form there on out, and decreases as the amount of precipitation goes up. I would suggest have a base case scenario where the expected number of bikers depends on day of the week and increases or decreases depending on temperature and decreases as the precipitation forecast increases.

3. Can you use this data to predict whether it is raining based on the number of bicyclists on the bridges?




The plot above corresponds to number precipitation (in units) vs number of bikers. Where the x-axis corresponds to number of bikers and the y axis corresponds to Precipitation. Here, the plot clearly indicates that there is no correlation between Number of bikers and precipitation where, for example, less bikers indicates more rain. This is even after applying a Savgol filter that helps reduce noise significantly and makes some relationships more visible to us. For this question, r metric also applies because r score does not consider which variable is the dependent or independent. So, let us reassess that table.

| | r | probability |
|---|---|---|
| Total | -0.4207113087552627 | 1.37339039815829e-10 |
| Clean total | -0.4207113087552627 | 1.37339039815829e-10 |
| Brooklyn Bridge | -0.3388580712586531 | 3.7922272360714444e-07 |
| Manhattan Bridge | -0.4105931111937938 | 4.130033023746216e-10 |
| Williamsburg Bridge | -0.4232468587701603 | 1.0362298020711196e-10 |
| Queensboro Bridge | -0.38802414926865153 | 4.231902725352678e-09 |

Again, I mentioned that this table indicated that there is a small inverse correlation between precipitation and number of bikers, however, it is small and I mentioned Precipitation is the least significant factor in determining number of bikers. In this case, the converse is also true the amount of bikers is not very significant in determining the amount of precipitation.

One might think that this contradicts the answer to question 2, but I said in that question that Precipitation affected lightly the number of bikers and could be used to fine-tune predictions, but should not be used as the sole metric used to predict bike traffic because r is not large enough to justify doing so. Given that in this case the question asks whether number of bikers could be used to estimate precipitation, I say we cannot because the correlation is not strong enough. Additionally, there were 146 non rainy days, and within those days the amount of traffic varied dramatically. This could help explain why the correlation between precipitation and traffic is much less than that of temperature and traffic. In conclusion, the amount of traffic cannot be used to be used to determine the amount of precipitation.