Section 1 (Group information):

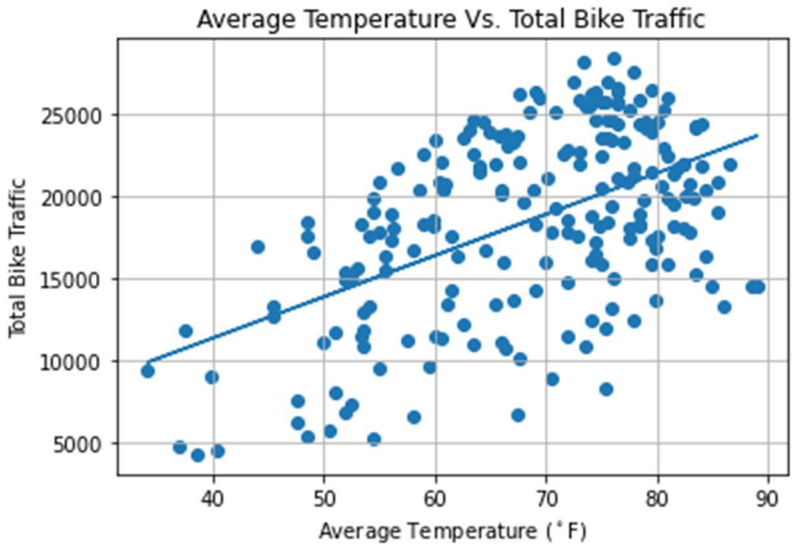
Team Member Names & Usernames:

- Maximilian Drach, mdrach
- Julie Joffe, joffe

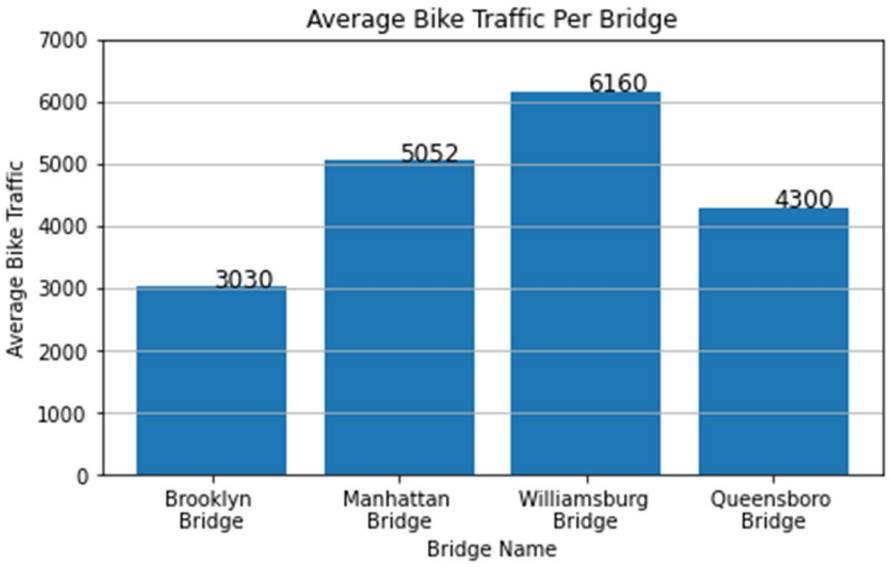
Path chosen: Path 1

Section 2 (Description of the dataset):

The NYC_Bicycle_Counts_2016_Corrected.csv dataset contains the bike traffic data from the four main pedestrain bridges in New York City starting 04/01/2016 and ending 10/31/2016. Included in the dataset, is the date (no year included), the day of the week (Day), the highest temperature recorded on the set day (High Temp), the lowest temperature recorded on the set day (Low Temp), the recorded precipitation on the set day (Precipitation), the recorded amount of cyclist on the Brooklyn Bridge (Brooklyn Bridge), the recorded amount of cyclist on the Manhattan Bridge (Manhattan Bridge), the recorded amount of cyclist on the Williamsburg Bridge (Williamsburg Bridge), the recorded amount of cyclist on the Queensboro Bridge (Queensboro Bridge), and the total recorded cyclist on all the listed bridges on a set day (Total). An additional variable we have added so far is the average temperature, which takes the mean of the lowest and highest recorded temperatures on each day.



Visual 1: scatter plot representation of the correlation between average temperature and total bike traffic on any given day. This graph allows us to see a generally positive correlation between increased temperature and bike traffic.



Visual 2: bar graph representation of average bike traffic, separated by bridge. This graph allows us to see that the Williamsburg bridge has the highest traffic, followed by Manhattan, then Queensboro, and finally Brooklyn.

Section 3 (Methods):

The weather forecast analysis (the combination of the high/low temperatures and precipitation variables) to bike ridership is the basis of how all the other questions will be analyzed. It is pivotal to our analysis because it is primary reason if someone is choosing to ride their bike or not. The first analysis of the weather is correlating the total bike ridership to the average weather temperature and total precipitation. This analysis will be performed using multivariate polynomial regression and supervised learning to properly match up the correlation. We will be using the scikit, NumPy, pandas, and matplotlib libraries to help perform our analysis properly show our results. Properly mapping the correlation of the weather and ridership is important because we can group the days together given certain weather parameters. For example, if it is raining on Friday but sunny 70's on Tuesday, the comparison of the days of the week cannot be fair since the weather conditions are not the same. We will use basic Bayes Theorem (with weather and probability someone will ride on certain bridge given the weather conditions) to get the probability of someone riding given a certain weather condition on all bridges and on given bridges.

1.

After analyzing how weather impacts the cyclist, we can now start to perform analysis on which bridges to set up sensors. Ceteris paribus, the weather might seem to have small impact on bridges market share of the total ridership; different bridges might have different infrastructure built in them to better deal with different weather conditions (shade cover, water fountains, drainage, etc). This means weather will also need to be accounted for when analyzing the individual bridge ridership trends. After controlling for weather, we start analyzing the percentage of total ridership each bridge receives, daily, weekly, monthly, etc. Next, we will come up with a ranking system to rank the most popular, the least ridership affected by weather, and finally find the bridges

that are most representative of the macro ridership trends. The most representative will not be ranked by total ridership but of specially different weighted ranks that take into not just one popular day but consistent popularity of different days. For example, lets say bridge A consistently has 32% of the daily riders, but bridge B has days where it has 90% precent of ridership and day where is it down to 7%. The bridge A would be more useful in the analysis because it is more representative of the total ridership over time. First, we will perform regression analysis of the bridge's individual ridership to weather correlations and see which one is least impacted by the weather. Next, we will find the most representative bridge by multiplying the probability someone will ride on a bridge given the weather conditions times the probability someone will ride that bridge on a given day (aggregate popularity). Then we will add up the new probabilities and then divide it by the total number of days. This rank considers an aggregate popularity but also consistent ridership through different weather environments to create the representative rankings. This ranking is used because it dampens the impact of outliers, it allows smaller bridges that are more consistently used to be monitored, and it more closely achieves the purpose of putting up the sensors: to get a representative view of the bike usage of the bridges. We will be using numpy, scipy, matplotlib, and other python libraries to achieve the ranking outcome.

2.

In order to determine which days police officers should be deployed, we are interested in finding out what the "ideal" temperature range is (what temperature range yields the highest traffic) as well as determining at what point precipitation begins to affect traffic. For the temperature, we will extract and calculate the mean of the recorded high and low temperature values. We will then identify the range(s) with the highest traffic. It is important that we deduce how the weather impacts not only total bike traffic but also bike traffic on individual bridges. We expect this analysis to give us at least 1 distinct range, however we may find that there is no strong correlation for any specific range. For precipitation, we will analyze the daily precipitation and its respective total traffic to try and see if there is a point at which precipitation begins to negatively impact total traffic on all the bridges. Similarly, to the temperature, we will need to consider that precipitation may affect travel on different bridges in very different ways. We could also use the filter() function to see at what precipitation values the total traffic is less than the mean total bike traffic across all the data, as well as bike traffic on each individual bridge. After identifying the "tipping point" we expect to find a single numerical answer of how much precipitation can be present before total bike traffic decreases, but we may not find that our data supports any one answer. By separating out the analysis for overall and individual bridge traffic, we will be able to come to a more in-depth and comprehensive conclusion, and we will be able to avoid extreme generalization of the dataset. Analysis question 2 is looking for a yes or no answer, depending on if at least one of the forecast variable analyses yields a distinct result. This question is entirely dependent on accurate analysis of the weather forecast, as discussed in paragraph 1.

3.

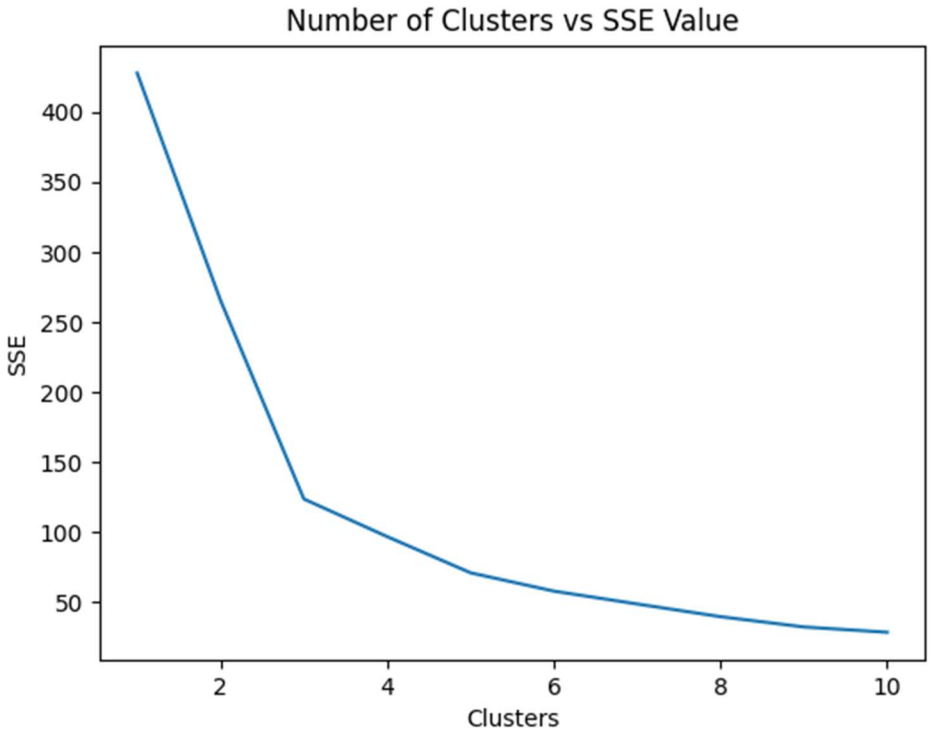
For the third analysis question, the main analysis will be in determining the correlation between day of the week and overall bike traffic on the bridges. We could also see if there is any correlation between days of the week and weather factors, because they have potential (somewhat based on the answer to question 2) to impact daily traffic, possibly in a consistent day-of-the-week

manner. Additionally, we will need to consider how the traffic on each bridge varies on different days of the week, because this could be a very simple and influential indicator as to what day of the week it is. For example, if bridge A consistently has an extreme influx of ridership over the weekends, one of our first steps in narrowing down which day of the week it is (given ridership on any given date) would be by seeing if we can eliminate weekdays based on the number of bikers specifically on bridge A. We can achieve this portion of our analysis by using numpy, matplotlib, and other python libraries to visualize the bike traffic based on these different factors. Analysis question 3 is looking for a yes or no answer.

Section 4 (Results):

1.

Our original method for ranking the bridges, unfortunately was unattainable at our level of expertise due to our lack of knowledge on Conditionality, Bayesian Networks, and advanced regression. Though we did come up with another method for ranking the bridges that didn't involve the standard deviation (prone to outliers) or popularity (not always representative of daily ridership). Our method first involved finding the ideal KMM clustering of both the mean temperature and the precipitation. From our data we found that either 3 or 5 clusters are ideal for analyzing the weather patterns (plot below).



Visual 3: Number of Clusters vs SSE value plot. This plot, generated via MiniProjectPath1.py, shows that either 3 or 5 clusters are ideal for analyzing weather patterns based on our given data.

In our function "rain_temp_cluster" the user has the option to input any number of clusters, but we'd recommend 3 or 5 clusters as seen from the "Number of Clusters vs SSE" plot. After we had our weather clusters, we then found the total ridership for each weather cluster from every respective bridge. This analyzes how the different weather patterns affect the number of riders for each bridge. All these rider cluster amounts were then divided by the total amount of riders for each respective bridge to give us the *cluster rider percentage*. Next, we simply find the share of riders each bridge held for that specific day – the *daily market share of each bridge* or the total amount of riders on specific bridge divided by the total number of riders for that day. To get the daily ranking we multiply the *cluster rider percentage* (given that day's weather cluster) by the respective *bridge's daily market share*. Finally, we sum up all the daily rankings score for the entire data set and the

bridges with the top 3 numbers will be chosen to have the sensors put on them. Total Rankings, Total Yearly Sum, and results are below.

Total Riders by Weather Cluster

0	343165	547342	671057	478689
1	87251	154389	185512	128043
2	166318	299065	360943	239166
3	10648	17237	24386	18679
4	41188	63145	76529	55778

Visual 4: Total riders on each bridge ("BB" represents the Brooklyn Bridge, "MB" represents the Manhattan Bridge, "WB" represents the Williamsburg Bridge, and "QB" for the Queensboro Bridge) by weather cluster.

Cluster Total Rider Percentage by Weather Cluster

Cluster	BB	MB	WB	QB
0	0.529110	0.506246	0.508983	0.520113
1	0.134528	0.142797	0.140707	0.139123
2	0.256438	0.276610	0.273768	0.259863
3	0.016418	0.015943	0.018496	0.020295
4	0.063506	0.058404	0.058046	0.060605

Visual 5: Cluster Total riders on each bridge ("BB" represents the Brooklyn Bridge, "MB" represents the Manhattan Bridge, "WB" represents the Williamsburg Bridge, and "QB" for the Queensboro Bridge) percentage by weather cluster.

Final Rankings

Ranking	Bridge	Total Score for 5 Clusters
1	WB	23.34
2	MB	19.04
3	QB	16.55
4	BB	11.76

Visual 6: Final rankings of each bridge using sum of daily ranking score (as stated above). Our final rankings show that the Williamsburg Bridge was came in first, followed by the Manhattan Bridge, then the Queensboro Bridge, and lastly the Brooklyn Bridge. **Based on this ranking and these results, the sensors should be placed on the Williamsburg, Manhattan, and Queensboro Bridges.**

Growing up in New York, I can say that I would agree with these rankings of which bridges to put the sensors on as they are consistent with my experience of New York traffic. The Williamsburg Bridge has a slightly wider bike lane and generally has the most usage since it connects a big residential area with lower Manhattan's commercial area. The Manhattan Bridge is a very popular bridge for locals because of its exclusive bike lane, nicer views, and drops off straight into Canal Street. It makes sense the Brooklyn Bridge is last, since before 2021 there was no dedicated private bike line, and it was always

packed with tourists on foot, leaving less space for cyclists and making it more difficult for there to be higher amounts of biker traffic.

2.

As stated in part 1, we could not use the conditional probabilities to predict the number of riders so instead we elected to use multivariate polynomial regression. The two independent variables were precipitation and mean temperature, while the total number of riders for each respective bridge was our dependent variable. Unfortunately, our polynomial regression was very inaccurate due to the "line of 0's" for the precipitation. This meant there could be a lot of variation in the number of riders given the same "dry" day. We tried next used Lasso and Ridge linear regression but that also failed due to the "line of 0's". Finally, we noticed that their seemed to be an inverse exponential correlation with precipitation and number of riders: the number of riders goes down as there is more precipitation. We then decided to only apply linear regression on the relationship between the mean temperature and the number of riders. To combine these two predictions, we had to computationally find the weights to apply to each model and found the ideal weights by calculating our lowest MSE and using those weights. Using this new method, **we were able to determine that it is possible to use the next day's weather forecast to predict the total number of bicyclists that day.** The function "rider_prediction" calculates the predicted number of riders given the dictionary input of the coefficient for a specific bridge, the mean temperature, and precipitation of the day (example below). To get the total predicted riders simply add up all the predicted riders for each bridge; you can use the "total_rider_prediction" function and input the bridge dictionary, mean temperature, and precipitation. (Examples below)

Sep-9

```
print(rider_prediction(bridge_dictionary['BB']['coef'], 85.6,.22))
print(total_rider_prediction(bridge_dictionary, 85.5, .22))
>>>2656.6706726662255

>>>16369.601952468372
```

24-May

```
print(rider_prediction(bridge_dictionary['MB']['coef'], 66,.18))
print(total_rider_prediction(bridge_dictionary, 66, .18))
>>> 3381.4459195057484

>>> 12882.515434999956
```

Visuals 6 and 7: inputs and outputs for Sep – 9 and May – 24 showing the predicted number of riders (given the dictionary input of the coefficient for a specific bridge, the mean temperature, and precipitation of the day) and the total rider prediction (given the bridge dictionary, mean temperature, and precipitation inputs)

3.

Question 3 is answered in the "day_of_week" function. From this function, **we determined that it is possible to predict what day of the week it is based on the number of bicyclists**. We first

extracted all of each day's total traffic numbers and created new columns to keep them in. After taking the averages of each of these columns to determine the average ridership on each day of the week, we created a dictionary to store these keys along with a more visually appealing values that could be outputted. We then used the input value and compared it with each of the keys of the dictionary to find which day's average was closest to the given number of riders (Examples below). A limitation of these results is that they use average values, meaning they may not be 100% accurate. If we had additional input of weather description and a much larger dataset, we would be able to include weather analysis. For this hypothetical to work though, we would need a larger dataset to clarify if there is consistent enough weather impacts on the same day every week to affect the average ridership of that day.

14,000 riders:

```
print("Today is: ", day_of_week(df, 14000))
```

```
Today is:  Sunday
```

20,375 riders:

```
print("Today is: ", day_of_week(df, 20375))
```

```
Today is:  Thursday
```

Visuals 8 and 9: Inputs and outputs showing predicted day of the week given the initial data and number of riders as input

Average Riders on Each Day of the Week

Day of the Week	Average Riders
Friday	17984
Saturday	15000
Sunday	13716
Monday	19393
Tuesday	20782
Wednesday	22422
Thursday	20781

Visual 10: Calculated average ridership on each day of the week as calculated by the "day_of_the_week" function. Here we see that the most popular days for cyclists are Wednesdays, averaging 22,422 riders across all four bridges, and the least popular days for cyclists are Sundays, averaging 13,716 riders across all four bridges.