Section 1 (Group information):

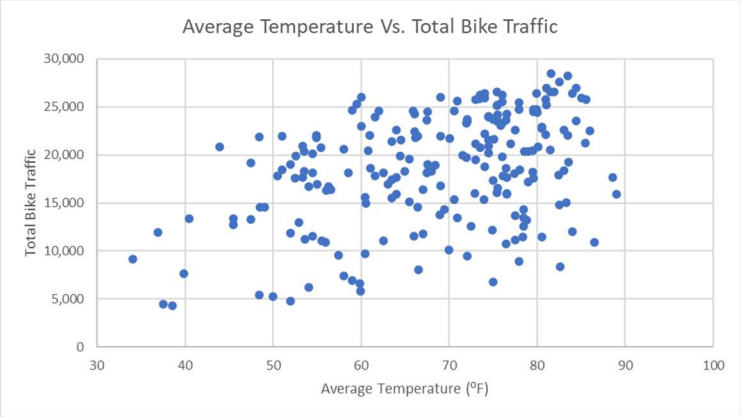
Team Member Names & Usernames:

- Maximilian Drach, mdrach
- Julie Joffe, joffe

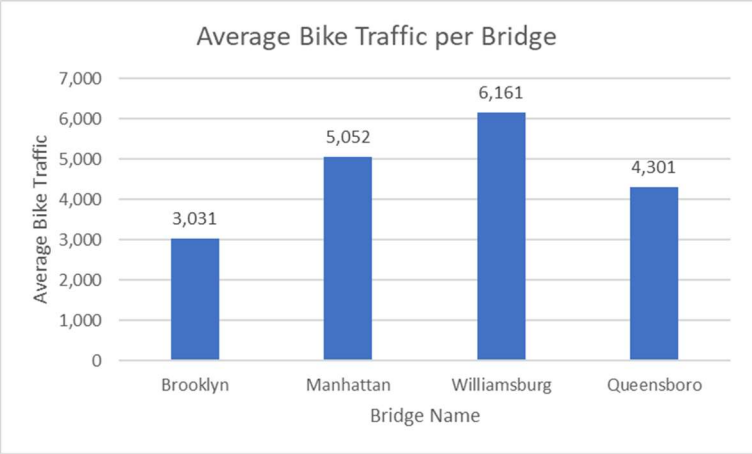
Path chosen: Path 1

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.

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.

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.

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. It is incredibly important that we consider the weather factors for every given day, as mentioned in paragraph 1, because otherwise the comparison between days would not only be unfair but extremely inaccurate. We will also need to 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 libaries to visualize the bike traffic based on these different factors. Analysis question 3 is looking for a yes or no answer.