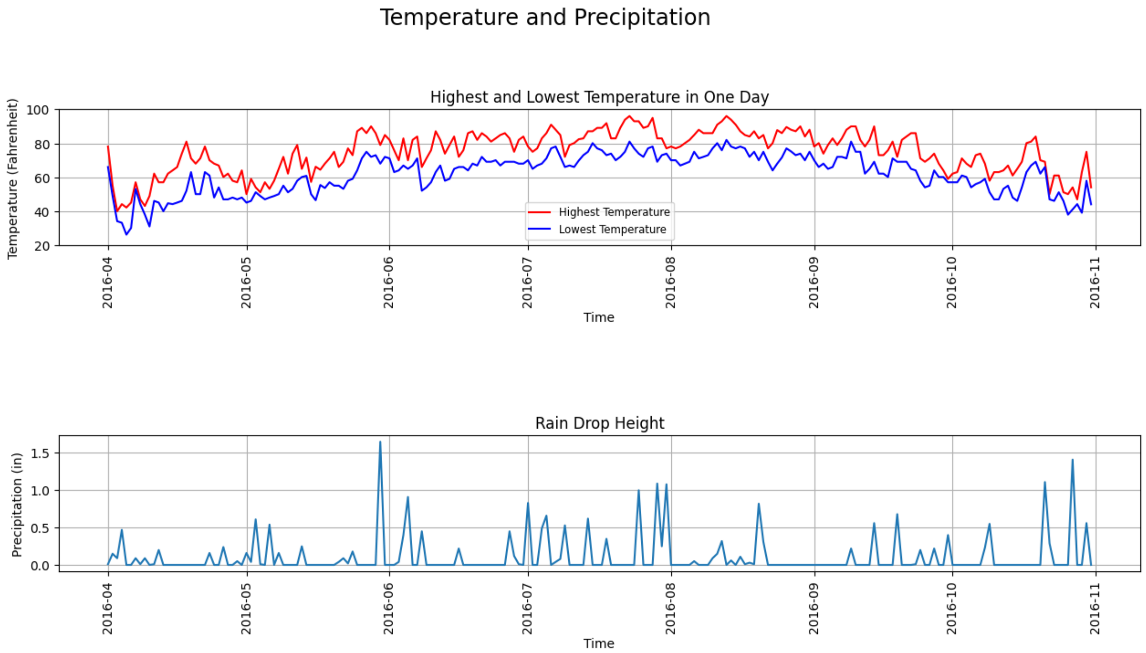
Group Information for Draft 1

Team Members: Eric Zhang (zhan4293) and Andrew Huang (huan1715)
Path: Path 1 Bike Traffic

Description of the Dataset:

The dataset that we will be working with includes various variables that provide specific details on the bike usage across four bridges in New York City. We will be using variables for the date, the day, the highest and lowest temperature in Fahrenheit, the amount of precipitation in inches, and the bike usage at each bridge. The dates and day will help us keep track of a timeline of events. The temperatures can identify the weather and determine if it was a warm, cold, or neutral day. The amount of precipitation signifies whether it was rainy or not, which can help determine the amount of bike usage at the bridges. The usage at each bridge varies from each other, but the Williamsburg Bridge stands out from the others due to its significantly larger usage.

Figure 1



The top graph labeled ‘Highest and Lowest Temperature in One Day’ displays the difference in the temperature variables in the city of New York based on a time scale between April 2016 to November 2016. The lower graph labeled ‘Rain Drop Height’ measures the amount of precipitation in a day and represents the rain variable. The temperature rises during the Summer and settles down in the Fall. Somehow, there’s a significant amount of precipitation during the month of June.

Figure 2

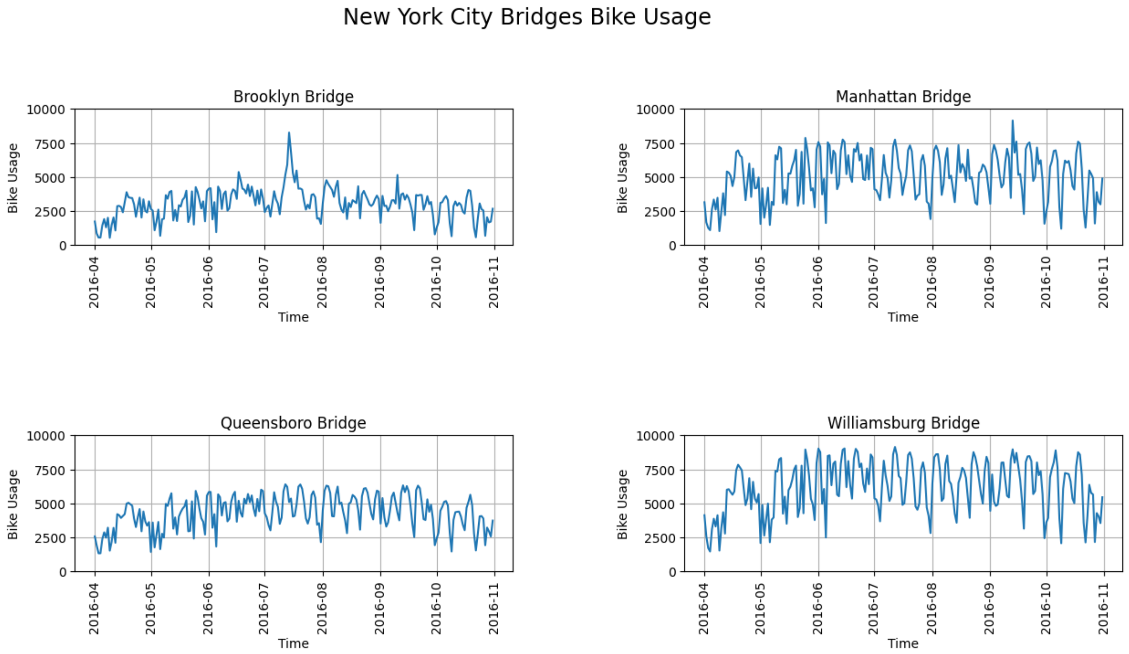


Figure 2 displays the bike usage across four bridges that are located in different areas. Each graph represents a variable according to its name. The Williamsburg Bridge appears to have the largest amount of bike usage with consistency over the 7-month period. The Brooklyn Bridge shows the lowest amount of bike usage amongst the other bridges, but spikes up during the months of July and August, perhaps due to traveling visitors. Manhattan and Queensboro Bridge are both consistent with its bike usage, however, Manhattan does have slightly more usage than Queensboro.

Methods:

1. In order to determine the most accurate prediction of overall traffic, we will calculate the average and standard deviation of bike usage. The three bridges with sensor installations will have the highest average and lowest standard deviation in bike usage. Our decision was to use the method of highest average since there will be a larger amount of bike usage, so the sensors will have a larger dataset to work with. Having a consistent amount of data for the installed sensors to predict the overall traffic at each bridge supports the decision for a lower standard deviation.

2. To help the city administration enforce helmet laws, we will determine whether there is a connection between the following day’s temperature and its total number of bicyclists. If there are correlations between the variables we can possibly generate a model. For the model, we will be using ‘Sklearn’ to create a multiple linear regression model (ridge) because there are multiple independent variables. We will split the data into a train and test set to try and generate a suitable model. We will be using Yellowbrick to visualize our result by generating a qq plot and a residual plot. However, the model may vary, so we would then run several tests to figure out the best parameters (ie. alpha) for the best model. We can then predict the number of cyclists based on our model.
3. Similar to Question 2, we will be using ‘Sklearn’ to determine whether there is a correlation between the day of the week and the number of cyclists on the bridge that day. This one is a little different since we are trying to guess which day of the week it is by the number of bicyclists. Therefore, we would have to first organize the data into different day of the week. We will then use a similar approach as Q2 to predict the day of the week.