For question 1, the dataset we are working with are: Day, High Temp, Low Temp, Precipitation, Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge, Queensboro Bridge.

The question is asking for choosing the three bridges to install sensors on. It means that in order to obtain the overall traffic data, traffic on one bridge will need to be predicted by the data we have. There’s no difference between predicting traffic on one bridge and predicting the overall traffic. We can obtain one another by simply doing addition or subtraction.

We think that multiple factors can affect the traffic. The first factor is the most obvious one, which is traffic on other bridges. If the traffic increase on one bridge, it is likely that traffic on other bridges may decrease. Of course, this is not always true. If there’s a traffic jam on bridge, people may tend to choose its alternative. It doesn’t matter too much for what assumption we’re making right here as the relationship will be revealed by our code. Another factor that may affect the traffic on one bridge is the weather. When it’s raining, it will be less likely for people to ride a bike. Moreover, temperature may also affect people’s behavior. When it’s too hot or too cold outside, people may also not ride a bike. The day in the week may also affect the bike traffic. It may be more likely for people to bike during weekend instead of weekdays. The time in the year may also affect people’s behavior. It is possible that people may ride a bike more often in spring instead of winter. However, the data provided didn’t record the data for one whole year. Thereby, the time in the year will not be in our consideration.

Therefore, the model we build should consider all these factors. In order to build up the model, we will need to preprocess the data. First, there is a comma for numbers having more than three digits, which may cause some problems on python. We choose to eliminate the comma on Excel by choosing another format for it. After that, we still have to transform day in a week to number. We define Monday to be 1, Tuesday to be 2 and so on. We noticed that precipitation is having one data in this form “0.47 (S)”. We choose to ignore the (S), which probably stands for snow. We did this by using regular expression when parsing, so even there are lots of data having the similar form, we are confident that python will help us make the format correct. We also noticed the “T” in precipitation. According to the information online, “T” stands for a very small amount that is very hard to measure with existing unit. Therefore, we replace “T” with zero.

We plan to build up four linear regression models, each treating traffic on one bridge as target variable and all other data as explanatory variables. We will compare the MSE calculated from the four models and choose the model that has the lowest MSE. After finding the model that has lowest MSE, we will choose not to install the sensor on the bridge that is treated as target variable in this case. This is because we are confident that the missing information on one bridge can be accurately predicted by processing the sensors data with the model we have.

We make the assumption that the sensors being installed will collect the same data as shown in the csv file provided, which means data like temperature, precipitation will also be provided. It is true that we can only use the traffic information without considering other factors like weather to build up this model. However, we think that the model will not be as accurate as the one considering multiple factors. Indeed, if the sensors being installed on the bridge will not collect the data like weather information, the less accurate model will be our only option.

For building up the model, I choose to use the similar approach as in hw5 question 2. We will do cross validation by splitting the data we have. After that, we’ll do normalization and test the MSE with different lambda value. After all, we’ll be able to get the MSE value for each model.

After running the code Q1.py, I got plot of MSE versus lambda for each model, which is shown below from figure 1.0 to 1.3.

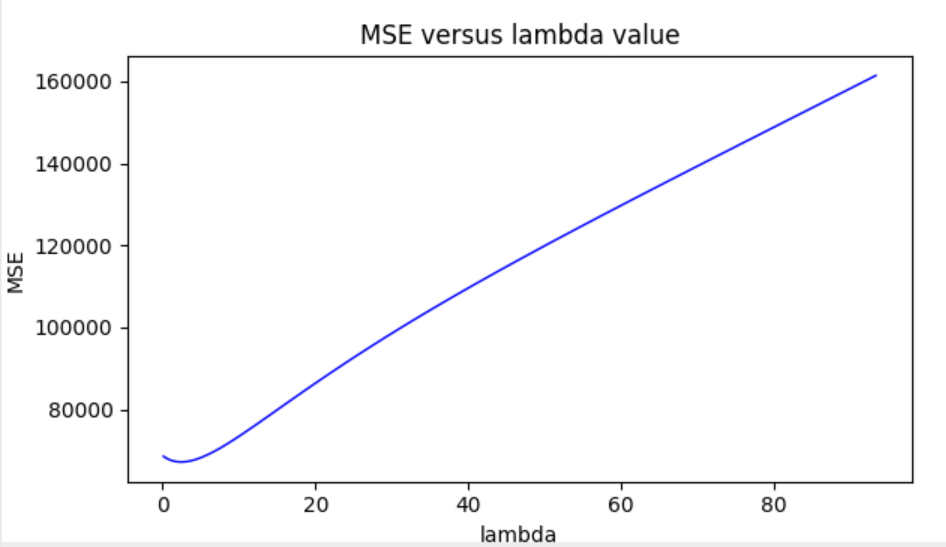


Figure 1.0: MSE versus lambda value plot for treating Queensboro Bridge as target variable

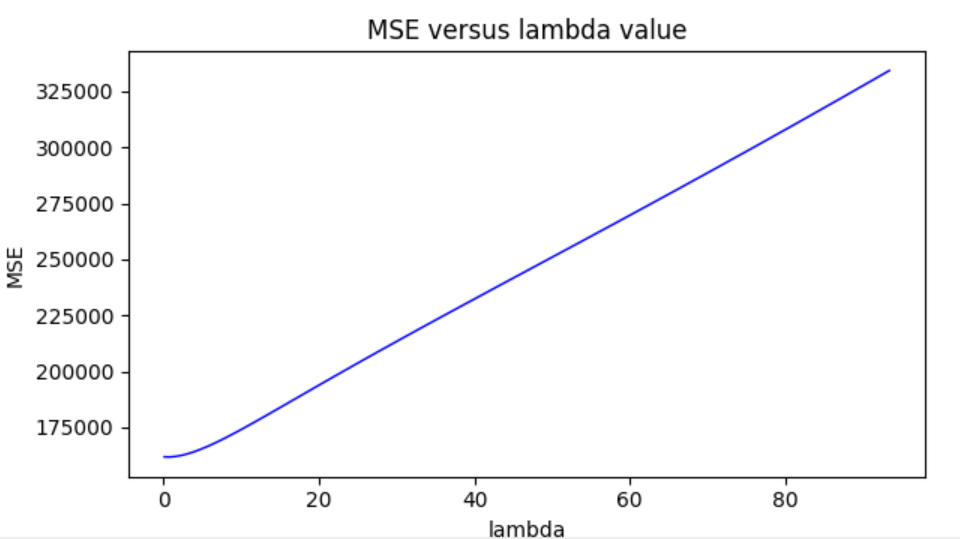


Figure 1.1: MSE versus lambda value plot for treating Williamsburg Bridge as target variable

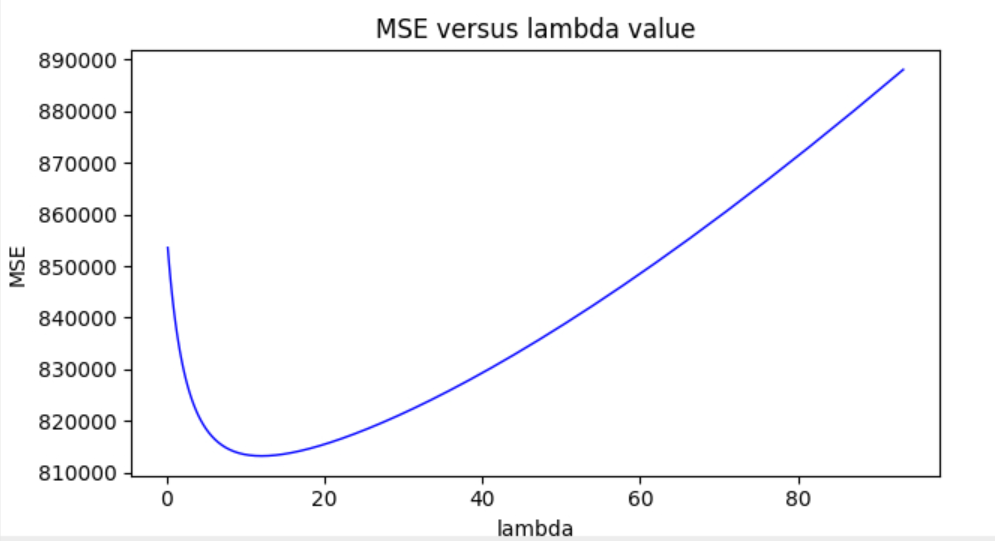


Figure 1.2: MSE versus lambda value plot for treating Manhattan Bridge as target variable

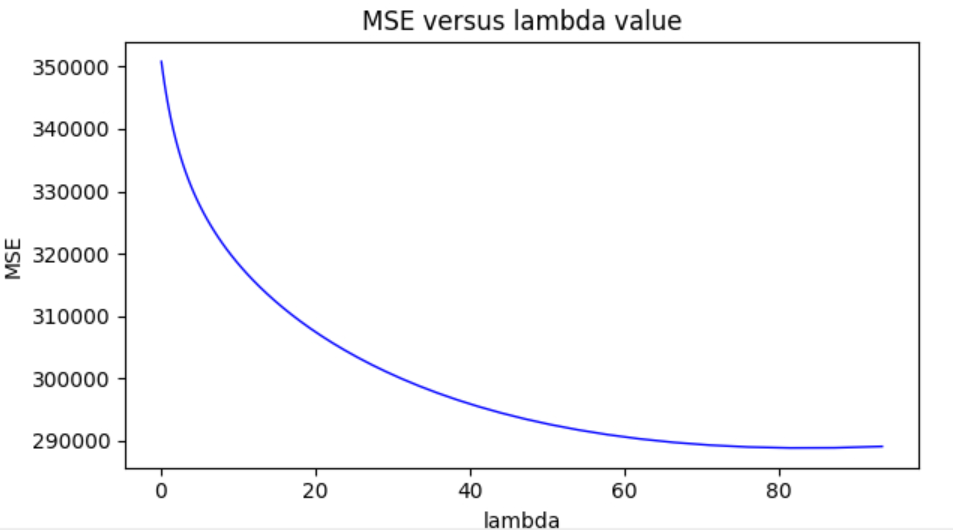


Figure 1.3: MSE versus lambda value plot for treating Brooklyn Bridge as target variable

I got the result as below in figure 1.4.

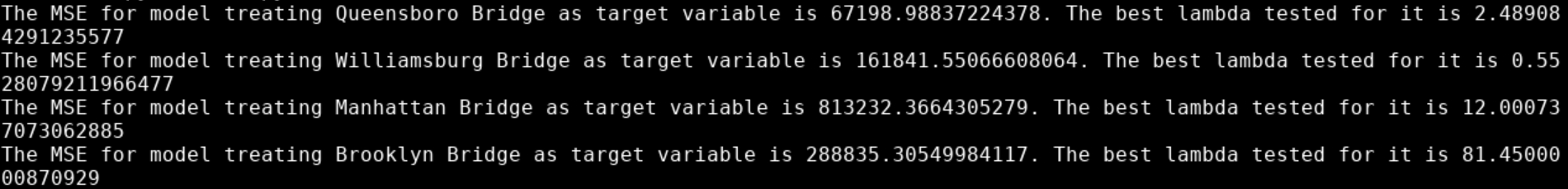


Figure 1.4: Python code result

I can easily tell that the model treating Queensboro Bridge data as target variable will get the lowest MSE among the four models, which is 67198.988. Therefore, I choose not to install sensor on Queensboro Bridge, which means that I’ll install sensors on Brooklyn Bridge, Manhattan Bridge, and Williamsburg Bridge.