Revised Project Proposal:

In our first proposal, we explored the idea of stationing NYPD units based on historical data and intuition. Crime is an extremely complex issue, especially in New York City, a city where crime is of high prevalence. After receiving our first set of feedback we have come to the conclusion that we need to refine our focus even more within the NYPD historical data. Trends have already shown us that high-crime areas remain high-crime areas. We have to ask ourselves, what is a new angle to approach the data? Our stakeholder, the NYPD force, needs a different approach on how to combat crime in New York City.

As a result of this, we have decided to merge our data with a data set that shows legally operating businesses in NYC. Most crime analysis looks at an area's demographic but we want to see if the amount of businesses in an area can help predict and even prevent crime in that area. If we are able to see a trend of crime location and business locations we can not only predict where crime can happen in a new light but, we can try to implement a new plan. That plan could be to open more businesses in areas of higher crime if the data shows us that there is a correlation between crime and there being fewer businesses in an area. To do that we will combine our previous NYPD dataset with our new Legally Operating Business dataset and conduct exploratory data analysis to understand the distribution of crime incidents, business types, and other relevant factors. After exploring our data we can use Random Forest or Gradient boosting to handle the complex relationships between the data.

Part of the feedback given to us raised concerns about biases in reported crime rates. There are many crimes that go unreported ie. gang violence. Obviously, there is no way to track crimes that go unreported but there is a way to try and make up for these unreported crimes. One of these ways is data augmentation. We can over-sample or synthetically generate instances of under-represented crimes to try and balance the dataset. This can help the model learn patterns for under-reported crimes. To do this in Python we can use the machine-learning technique of oversampling and SMOTE to create instances for under-reported crimes.

Combating these biases will be the most challenging obstacle to overcome in our model. Even if we create “fake” examples of data it will not be as accurate as real data and can negatively affect the accuracy of the model. In this case, the pros outweigh the cons to help prevent biases. It is better to over-predict the data than to under-predict when it comes to crime. When it comes to data privacy, since the dataset is public it is not as big of a risk as previously mentioned and will focus more on mitigating the risk of bias.

As the deadline of the project is approaching we decided to make a more detailed project management plan. We have decided that all of us are going to meet one or two times a week once we get back from Thanksgiving break to work on the code of the project together. After each time we meet we will dedicate two people to work on the write-up part of the project. With this in mind, we aim to have fully pre-processed the data, like cleaning the data, dropping columns,

and joining the two datasets together before we come back from back. We also want to begin to feature engineer our columns and explore the data using visualization. When we get back from break we want to within our first week then have trained our data to our model and have evaluated it. Lastly, in the first week of December, we want to hyperparameter-tune the model, address our bias concerns, and prepare for our presentation to the class.