Part 1

So we started off by using data from the NYPD data, we are able to restrict it to a certain time frame and we initially trained and tested our prediction models on the year 2024, since that is the most recent and complete year.

We then adapted the data to our needs. So we converted the dates to a specific format that can be taken by pandas, we dropped the rows with invalid dates. We then converted days of the weeks to number 0-6, 0 being Monday and 6 being Sunday. We did the same for month and also extracted the hour. We also converted if the crime was completed or not completed to 1 and 0. We also converted the boroughs to numbers as well so that it was easier for these models to see patterns. And we also filtered rows that had their crimes serverities labeled.

We then trained different algorithms on this preprocessed data. We chose to train Random Fores, Logistic Regression, SVM, KNN, and Decision Tree models. We chose all of these so we could see how these algorithms would compare and later see if any of these algorithms would be bias leaning. All the models were trained on 80% of the data and tested on 20% of the rest. We predicted the crime occurrence In a location and the crime type in a given area. We evaluated this accuracy using the classification report provided by sci-ki which shows the precision, recall, and f1-score. We also have a table showing the accuracy, macro average, and weighted average.

From these results we see…

We then began back testing…