EDA 1. Looking into the breakdown of violent and non-violent crimes, we began with a basic bar chart to help gain a general overview of the variable. The “arrests by type” chart showed most crimes do not result in arrests, non-violent crimes have the most volume, and violent crimes have the highest arrest rates. We then created a second chart to break down crime frequencies—non-violent crimes in blue and violent crimes in orange. This helped the group gain a deeper understanding of specific crimes that are most common, while reinforcing the previous finding that most arrests for crimes are non-violent.

EDA 2. Continuing with the eda process, we analyzed how crime and arrest counts vary by district. The two graphs displayed work together: the first provides a side-by-side comparison of crime and arrest counts, while the second breaks this down into arrest rates by district-- using red to indicate whether a district falls below the median arrest count. Notably, district 11 has the highest crime and arrest counts but this does not always mean high arrest rates, as shown by district 8 with a similar number of crimes. While these graphs were not directly used in the feature engineering process, they offer valuable context. Should the predictive model underperform, we can revisit district level dynamics to dissect and potentially treat certain districts as outliers, but for now we are ready to move into the modeling phase.

Modeling 1. First, we defined features by dropping columns that lacked predictive value and established arrest as our target variable. Next, we defined the train test split using 25% of the data for training and ensuring our random state was set to 42 for reproducibility. From there we established a preprocessing pipeline to ensure data preservation and proper conversion into a machine-readable format. Similarly, we also established a model pipeline to highlight how the models naturally fit the data before tuning. With this structure, we let the data speak for itself and enable room for modifications should we decide to make changes. Upon examination of the results, we see that logistic regression was the best model overall, however, none of these models were tuned, therefore, we cannot say with certainty this was the best model until tuning efforts are made.

Modeling 2. Within the tuning phase we followed a similar process of defining a pipeline structure, this time incorporating a parameter grid for hyperparameter optimization. Once again, this setup was organized with flexibility to make adjustments easy to implement. Behind the scenes, the code selected the best model within the tuning process, but in this scenario the configurations did not outperform the standard algorithm. At this point we can benchmark the logistic regression model as the most ideal model candidate, however, additional tuning efforts were made, and we cannot make a conclusion until further improvements are officially ruled out or exceed the current model.