The New York Police Department has been recordkeeping and publishing detailed crime data for NYC for many years. We believe a scientific analysis of this data could bring an insight into the type of crime an individual might be a victim of, given the time of the day, location, and the demographics of the individual. It could also help law enforcements predict the likelihood of a crime at a given location and a time of the year, such as the list of likely crimes on December 31st at midnight in Manhattan for a particular demographic.

NYC Open Data is free public data published by New York City Agencies and other partners. This data has been maintained by them and updated annually. The dataset spans the years 2006 through 2020 and contains around 7 million records. For practicality reasons, we limited our time scope to January 2019 through February 2020 which is around 350,000 records. The data itself is well maintained however, it still contains missing values and mixed data types in some columns. It will require minimal cleaning, but some variables will have to be either translated, transformed, grouped, sterilized, or mapped in order to be usable. In particular, many of the features are qualitative (crime type, demographic characteristic) and gaps in the data will need to be accounted for, though the size of the dataset will allow for some instance selection.

After cleaning the data, we will apply more in-depth exploratory data analysis to get a better understanding of the variables along with their correlation. The main goal of the project is to classify crimes based on time, location, victim's sex, and age group in addition to some correlation and patter analysis. Specifically, statistical methods like K-Nearest Neighbor, Random Forest, Support Vector Machine, and Naïve Bayes Classifiers will be implemented. After careful analysis of each model, we will pick the best one that produces the optimal results. The results will be evaluated based on various performance measures like accuracy, F1 score, precision, recall , Cohen's kappa, and ROC curve. All this will be implemented using python as the main programming language, Github as the project environment management, and Streamlit as the GUI design.