# Crime Analysis in the United States

# Proposal

**Team members**

**Overview**

Crime and its impact on public safety and public health are universal concerns. It seems highly aligned with the fluctuation in America's economic status as well. The goal of this project is to identify and analyze the pattern of crime, help law enforcement agencies to establish crime prevention strategies, and make contributions to decrease the adverse impact of crime for American society.

In this project, we will use Communities and Crime Unnormalized Data Set and MSA (the Stanford Mass Shootings of America) dataset to explore if there are any significant relationships between crime and any of different variables. Our crime dataset contains communities in the US. Data combines socio-economic data from the '90 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey, and crime data from the 1995 FBI UCR. Since it does not contain data related to shootings, we use another shooting dataset and combine them together. Our mass shooting dataset is a curated set of spatial and temporal data about mass shootings in America, taken from online media sources. It records shooting incidents with at least 3 or more victims, and the shootings are not identifiably gang, drug, or organized crime-related.

**Methods**

Before processing the data, we are going to start with data cleaning due to the missing values and data format. First, there are many different ways to fix the missing values, such as removing, replacing by the mean value, or replacing by the median value. The decision will depend on the variables. Secondly, there are more than 100 attributes that we can make the dataset tidy by using functions in R packages. Tidy data will make munging and visualizing data easier. Thirdly, data transformation will be applied if the data structure is not ideal. For example, we can use log transformation to see the relationship between two variables more clearly. Finally, the relational data can be applied between two datasets when we want to combine information from both datasets.

After tidying data, we will use different functions, such as conditional statements and control structures to trace the crime variable relationships or crime information we will show, along with necessary data wrangling, and visualize them through R Shiny. R Shiny will help us accomplish the interactive display so that users can choose input parameters with sliders, drop-downs and text fields. Data mining and machine learning techniques will also be used to predict risk scores on specific time, place, or incidents. We will use the behavioral data in our dataset and find the most suitable models for interpretation. Machine Learning may be applied at the final part, using algorithms such as linear regression, logistic regression, additive regression to implement and conduct analysis.