# STA2101 Project Proposal - Willem Atack

## 1 Problem Introduction

In every community, reducing crime would equate to safer communities with increased well-being, health, engagement and economic output. Understanding the root problems that lead to crime allows us to more effectively prevent it. It is well known that factors like education level, income, unemployment and family structure affect crime rates. However, communities have limited resources to allocate toward crime reduction, so understanding what factors are the most significant is vital. In this project, a statistical analysis will attempt to understand what environmental factors or demographics are the most significant predictors of crime.

Of course, not all crime is created equal, and strategies to combat non-violent and violent crime differ. Therefore, it will be tested whether or not different types of crime have different significant predictors. To illustrate why this could be useful, consider if the analysis found that characteristic X tends to be a significant predictor of assault, but not non-violent crime. If community ABC exhibits characteristic X, they could take on a more focused strategy of reducing assault rather than a more general approach.

The demographics and characteristics of regions shift over time, so it would also be useful to build a predictive model which generalizes well to unseen data. Therefore, the project will also focus on building a predictive model which is able to accurately estimate the rates of different crimes on a holdout dataset. This would allow towns to effectively adapt their crime reduction strategies dependent on changes in their demographics.

Finally, though it is not within the course's scope, it would be interesting to utilize unsupervised learning techniques to cluster the regions in the dataset based on their demographics. This may be valuable since if town X and town Y are found to have similar demographics, then they likely have similar underlying pressures contributing to crime. If town X has a higher crime rate than town Y, then it may be valuable for them to consult town Y on how they are able to mitigate these common pressures equating to higher crime.

## 2 Dataset Description

The dataset in question combines socio-economic data from the 1990 U.S. census, law enforcement data from a 1990 Law Enforcement Management survey, and FBI crime data from 1995. It has 2,215 instances of data representing the demographics and crime rates in different communities across the United States.

Each instance has 4 non-predictive features which identify the community. They have 125 predictive features, which include detailed demographic data about the community. This includes, but is not limited to, information on population, population density, age distribution, income distribution, education levels, unemployment rates, family structure data, race, immigration, housing, homelessness, as well as information on the community's police force such as number of police cars, and police department budget. An exhaustive list is not included due to the number of features, though a link to the dataset is included at the end of this section. All of the features are unnormalized, and many instances contain one or more missing values.

It is worth noting that many features are redundant, and can be eliminated before beginning the analysis. In many cases, a feature representing an absolute value, and another feature representing the same quantity as a rate are included. For instance, ”number of people living in areas classified as urban” and ”percentage of people living in areas classified as urban” are separate features. There are many other subsets of features which seem likely to have very strong correlations with each other. For example, it is likely that including both ”percentage of people that have immigrated in the last 3 years” and ”percentage of people that have immigrated in the last 5 years” would add unnecessary model complexity. Of course, this will be formally tested before removing any covariates from the model.

Each instance has 18 possible response variables. These represent the absolute number, and the rate of 9 different types of crime. The types of crime are ”murder”, ”rape”, ””robberies”, ”assaults”, ”burglaries”, ”larcenies”, ”autoTheft”, ”arsons” and ”non-violent”. In the analysis, the rates of each of these different crimes will be considered the response variable in different models.

Link to dataset: https://archive.ics.uci.edu/dataset/211/communities+and+crime+unnormalized

# 3 Analysis Description

As described in the problem description, there are a few questions to be answered in this project. First, is determining what factors are most significant in predicting crime rate, and if these differ based on the type of crime. For this analysis, a variety of methods will be used including hypothesis-testing techniques like t-tests, confidence intervals, and model comparison with F-statistics. An analysis of colinearity is also important and can be used to reduce the number of features. Since the number of covariates in the dataset is very large, testing both criteria-based and variable-selection model selection techniques will be experimented with to reduce the model complexity, while maintaining predictive power.

Once again, due to the large number of covariates, penalized regression techniques will also be explored as a method of feature selection to further reduce model complexity. As the course progresses and more advanced techniques such as Generalized Linear Models are introduced, they will be incorporated into the analysis. Beyond the scope of the course, Bayesian methods can also be used as a method of analyzing covariate significance.

The second part of the project involves optimizing a predictive model for estimating crime rates on unseen data. To evaluate predictive models, k-fold cross validation will be used, and $RMSE$ and $R^2$ values will be compared. To extend the project in this section of the analysis, traditional statistical modelling techniques (such as those taught in class) will be compared with newer machine learning algorithms, such as Random Forest. To reduce dimensionality and overfitting, techniques like PCA will be applied.

Finally, unsupervised learning techniques such as k-means clustering will be used to analyze whether the communities can be naturally grouped.

# 4 Potential Challenges

The main challenge that is expected is that the selected dataset has a very large number of covariates. Therefore, there will need to be a lot of focus during the analysis on determining what features are most significant and hold the most predictive power in order to reduce the model's complexity. Further, among these covariates, there is expected to be a lot of multi-colinearity. Being aware of this and using methods to reduce colinearity will also be important. The large size of the dataset means that there are a lot of research questions that could be asked. Staying focused on the research questions defined above, and potentially altering them through the EDA phase will be necessary.

Other typical problems are expected to be present as well. For one, an initial look at the data shows that there are a reasonable number of missing values, so this will need to be dealt with appropriately. Also, the features all have very different scales (e.g. population vs average years of education). Normalization of the data will be vital. Finally, since the data has a reasonably large number of instances, there are likely to be outliers which could cause overfitting when creating a predictive models, and skew the true significance of the parameters. Formal methods will be used to remove outliers.