Crime and Communities

Group Member 1 Name: _____ **Group Member 1 SID:** _____

Group Member 2 Name: _____ **Group Member 2 SID:** _____

The crime and communities dataset contains crime data from communities in the United States. The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. More details can be found at https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized.

The dataset contains 125 columns total; $p = 124$ predictive and 1 target (ViolentCrimesPerPop). There are $n = 1994$ observations. These can be arranged into an $n \times p = 1994 \times 127$ feature matrix $\mathbf{X}$, and an $n \times 1 = 1994 \times 1$ response vector $\mathbf{y}$ (containing the observations of ViolentCrimesPerPop).

Once downloaded (from bCourses), the data can be loaded as follows.

```
library(readr)
CC <- read_csv("crime_and_communities_data.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
print(dim(CC))
```

```
## [1] 1994  125
```

```
y <- CC$ViolentCrimesPerPop
X <- subset(CC, select = -c(ViolentCrimesPerPop))
```

Dataset exploration

In this section, you should provide a thorough exploration of the features of the dataset. Things to keep in mind in this section include:

- Which variables are categorical versus numerical?
- What are the general summary statistics of the data? How can these be visualized?
- Is the data normalized? Should it be normalized?
- Are there missing values in the data? How should these missing values be handled?
- Can the data be well-represented in fewer dimensions?

YOUR CODE GOES HERE

Regression task

In this section, you should use the techniques learned in class to develop a model to predict ViolentCrimesPerPop using the 124 features (or some subset of them) stored in $\mathbf{X}$. Remember that you should try several different methods, and use model selection methods to determine which model is best. You should also be sure to keep a held-out test set to evaluate the performance of your model.

YOUR CODE GOES HERE