# Stat 4620 Final Project

Group 10

2025-12-04

## Part I: Exploratory Data Analysis

### Introduction to the Dataset

The data come from the Communities and Crime dataset, which combines 1990 U.S. Census socioeconomic variables, law enforcement variables from the LEMAS survey, and crime statistics from the FBI's Uniform Crime Reports. The response variable is ViolentCrimesPerPop, representing the number of violent crimes per 100,000 population in each community. Predictors include demographics, poverty, education, housing, family structure, and policing resource variables.

In the training set, there are 1550 communities and 147 potential predictors plus the response.

### Data Types and Basic Cleaning

```
# Look at the first few variable names
names(train)[1:15]
```

```
##  [1] "X"              "Ęcommunityname" "state"          "countyCode"
##  [5] "communityCode"  "fold"           "population"     "householdsize"
##  [9] "racepctblack"   "racePctWhite"   "racePctAsian"   "racePctHisp"
## [13] "agePct12t21"    "agePct12t29"    "agePct16t24"
```

```
names(train)[(ncol(train)-10):ncol(train)]
```

```
##  [1] "assaultPerPop"       "burglaries"          "burglPerPop"
##  [4] "larcenies"           "larcPerPop"          "autoTheft"
##  [7] "autoTheftPerPop"     "arsons"              "arsonsPerPop"
## [10] "ViolentCrimesPerPop" "nonViolPerPop"
```

```
# Non-predictive ID-like variables (keep for info, drop later for modeling)
id_vars <- c("X", "Unnamed..0", "Unnamed: 0", "Ęcommunityname",
             "state", "countyCode", "communityCode", "fold")
id_vars <- intersect(id_vars, names(train))

# Check classes
sapply(train[1:20], class)
```

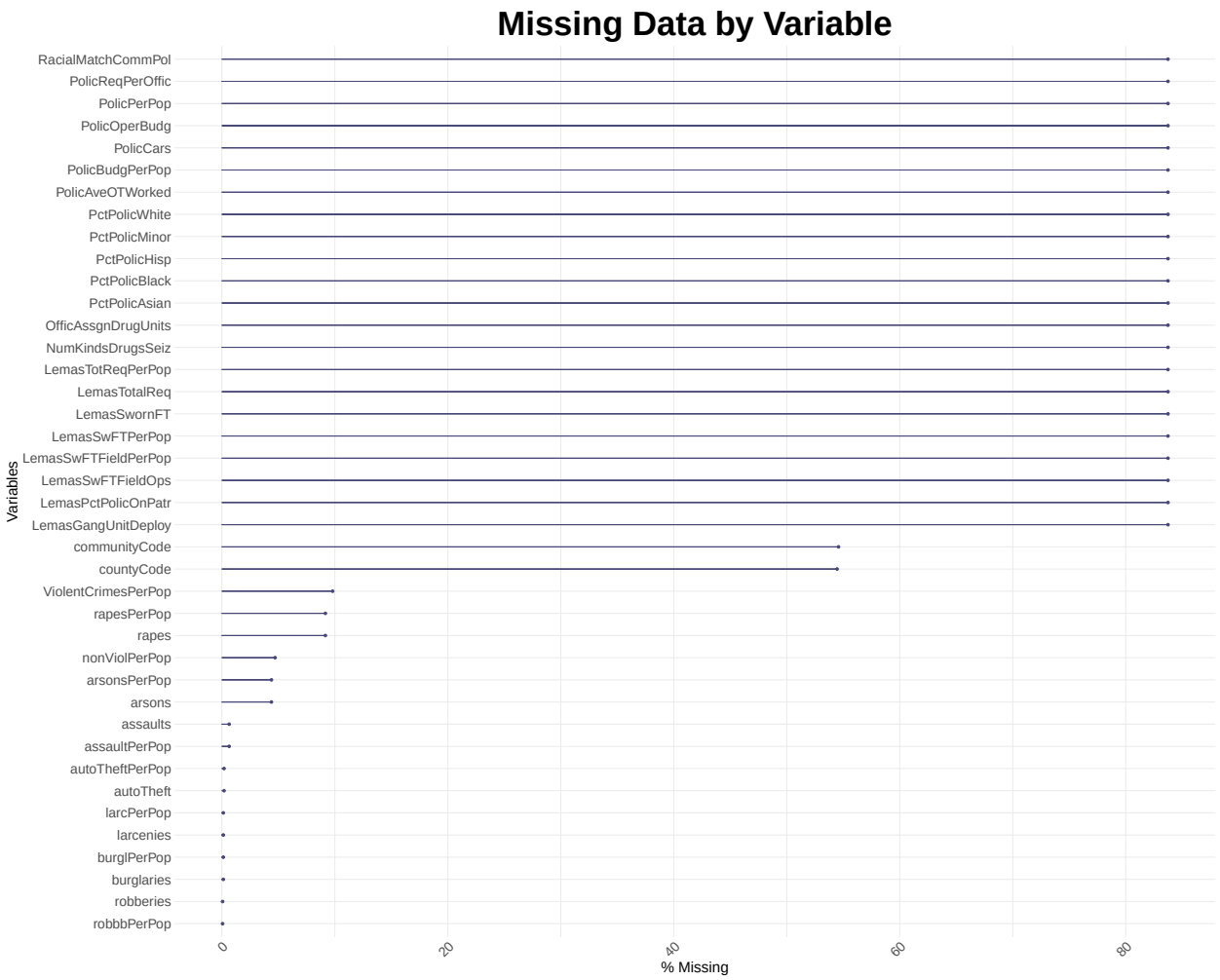
```
##              X Ęcommunityname          state     countyCode  communityCode
##      "integer"    "character"    "character"      "integer"      "integer"
##           fold     population  householdsize   racepctblack   racePctWhite
##      "integer"      "integer"      "numeric"      "numeric"      "numeric"
##   racePctAsian    racePctHisp    agePct12t21    agePct12t29    agePct16t24
##      "numeric"      "numeric"      "numeric"      "numeric"      "numeric"
##     agePct65up      numbUrban       pctUrban      medIncome       pctWWage
##      "numeric"      "integer"      "numeric"      "integer"      "numeric"
```

```r
# Convert character columns (except ID/name fields) to numeric
char_vars <- names(train)[sapply(train, is.character)]
char_to_num <- setdiff(char_vars, c("Ęcommunityname", "state", "countyCode", "communityCode"))

train <- train %>%
  mutate(across(all_of(char_to_num), ~ as.numeric(.)))
```

The dataset contains a mix of identifier variables (state, county code, community code, community name) and a large number of numeric predictors. Some numeric variables were stored as character strings with "?" indicating missing values. These were converted to numeric with "?" treated as NA

## Missing Data Analysis

```r
missing_vars <- names(train)[colMeans(is.na(train)) > 0]

library(naniar)
gg_miss_var(train[missing_vars], show_pct = TRUE) +
  labs(title = "Missing Data by Variable") +
  theme(
    axis.text.x = element_text(size = 20, angle = 45, hjust = 1),
    axis.text.y = element_text(size = 20),
    axis.title = element_text(size = 22),
    plot.title = element_text(size = 50, face = "bold", hjust = 0.5
)
  )
```




To examine missingness in the training dataset, we first identified all variables containing at least one missing value. We then visualized the proportion of missing values for these variables using gg_miss_var() from the naniar package.

This plot clearly shows that most variables in the dataset have no missing values at all, while a small subset contains substantial missingness. In particular, the variables with missingness correspond almost entirely to law enforcement and policing measures.

These variables all exhibit roughly 80–85% missingness, consistent with the original dataset documentation stating that the LEMAS survey only covers police departments with 100+ officers and a random subset of smaller agencies. As a result, many communities simply do not have any LEMAS data available.

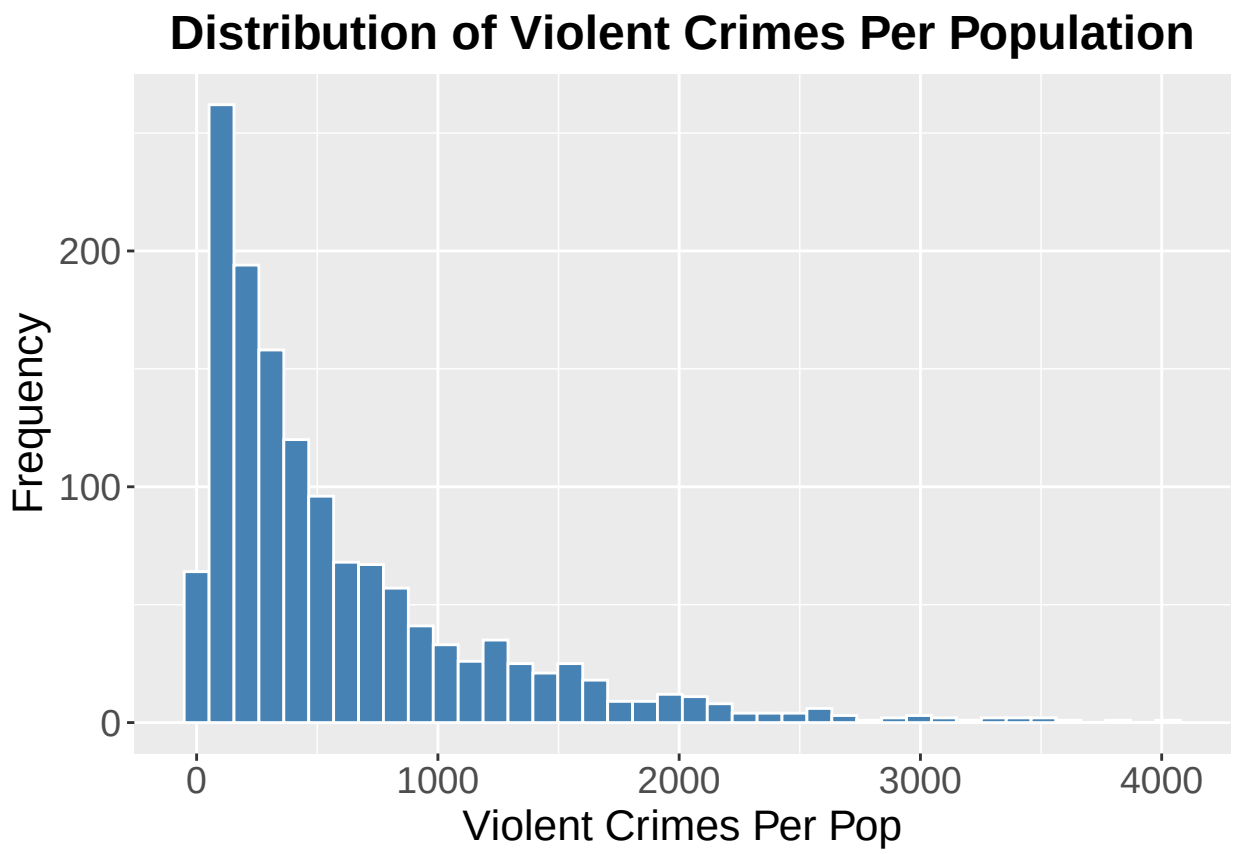
Because such high missingness would either (1) require dropping a large portion of the dataset or (2) require unreliable imputation, these variables will not be included in the modeling stage. All remaining socioeconomic, demographic, and family-structure variables are either complete or nearly complete and therefore suitable for predictive modeling.

## Distribution of the Response Variable

```
ggplot(train, aes(x = ViolentCrimesPerPop)) +
geom_histogram(bins = 40, color = "white", fill = "steelblue") +
labs(
```

```
title = "Distribution of Violent Crimes Per Population",
x = "Violent Crimes Per Pop",
y = "Frequency"
) +
theme(
plot.title = element_text(size = 18, face = "bold", hjust = 0.5),
axis.text = element_text(size = 14),
axis.title = element_text(size = 16)
)
```

```
## Warning: Removed 152 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




The distribution of ViolentCrimesPerPop is heavily right-skewed, with most communities experiencing relatively low violent crime rates while a smaller subset exhibits very high rates. This skewness suggests that linear modeling may produce heteroskedastic residuals and that nonlinear models or transformations may warrant consideration.

### Key Numerical Summaries of Predictors

```
summary(train %>% select(where(is.numeric), -all_of(id_vars)) )
```

As described in the dataset documentation, all numeric variables are normalized to the 0–1 range using equal-interval binning. This means that within-variable comparisons are meaningful (e.g., a value of 0.4 vs 0.8), and between-variable comparisons are not meaningful (e.g., racePctWhite vs medIncome).

Variables related to poverty (PctPopUnderPov), education (PctNotHSGrad), family structure (PctKids2Par), income (medIncome, perCapInc), and unemployment (PctUnemployed) show substantial variation across communities and are likely important for prediction.

## Correlation with the Response

```
numeric_train <- train %>% select(where(is.numeric), -all_of(id_vars))

cor_vec <- sapply(numeric_train, function(x)
cor(x, train$ViolentCrimesPerPop, use = "pairwise.complete.obs")
)

head(sort(cor_vec, decreasing = TRUE), 10)
```

```
## ViolentCrimesPerPop       assaultPerPop         robbbPerPop PctKidsBornNeverMar
##           1.0000000           0.9452283           0.8202987           0.7455162
##         burglPerPop       nonViolPerPop          murdPerPop        racepctblack
##           0.6824353           0.6712011           0.6659152           0.6389150
##     autoTheftPerPop         rapesPerPop
##           0.6132238           0.5822002
```

```
head(sort(cor_vec), 10)
```

```
##        PctKids2Par         PctFam2Par       racePctWhite   PctYoungKids2Par
##          -0.7319027         -0.7025762         -0.6854848         -0.6651604
##        PctTeen2Par         pctWInvInc    PctPersOwnOccup RacialMatchCommPol
##          -0.6632866         -0.5656715         -0.5078119         -0.4726272
##      PctHousOwnOcc          medFamInc
##          -0.4554770         -0.4154574
```

The strongest positive correlations with violent crime include:

- PctIlleg (children born to unmarried parents)
- PctPopUnderPov (poverty)
- PctNotHSGrad (low educational attainment)
- PctUnemployed (unemployment)
- FemalePctDiv / TotalPctDiv (family instability)

The strongest negative correlations include:

- PctKids2Par and PctFam2Par (two-parent households)
- PctHousOwnOcc (homeownership)

- PctSpeakEnglOnly (English proficiency)
- medIncome / perCapInc (income measures)

These relationships align with sociological expectations and provide early insight into promising predictor groups.

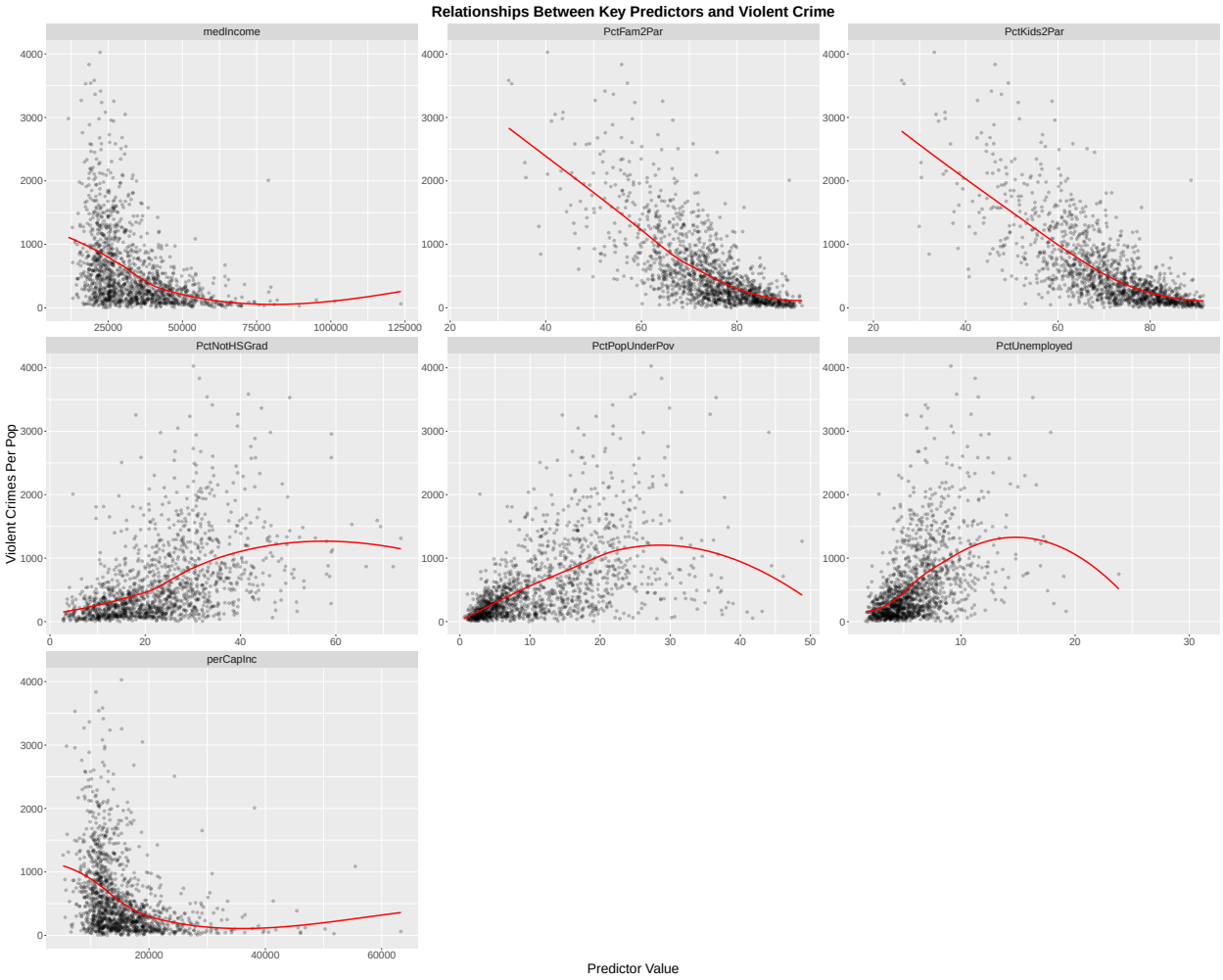
### Visual Exploration of Key Predictors

```
key_vars <- c(
  "PctPopUnderPov",
  "PctNotHSGrad",
  "PctUnemployed",
  "PctKids2Par",
  "PctFam2Par",
  "medIncome",
  "perCapInc"
)

train %>%
select(ViolentCrimesPerPop, all_of(key_vars)) %>%
pivot_longer(cols = -ViolentCrimesPerPop, names_to = "variable", values_to = "value") %>%
ggplot(aes(x = value, y = ViolentCrimesPerPop)) +
geom_point(alpha = 0.25) +
geom_smooth(method = "loess", se = FALSE, color = "red") +
facet_wrap(~ variable, scales = "free") +
labs(
title = "Relationships Between Key Predictors and Violent Crime",
x = "Predictor Value",
y = "Violent Crimes Per Pop"
) +
theme(
plot.title = element_text(size = 22, face = "bold", hjust = 0.5),
strip.text = element_text(size = 16),
axis.text = element_text(size = 15),
axis.title = element_text(size = 20)
)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1064 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1064 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



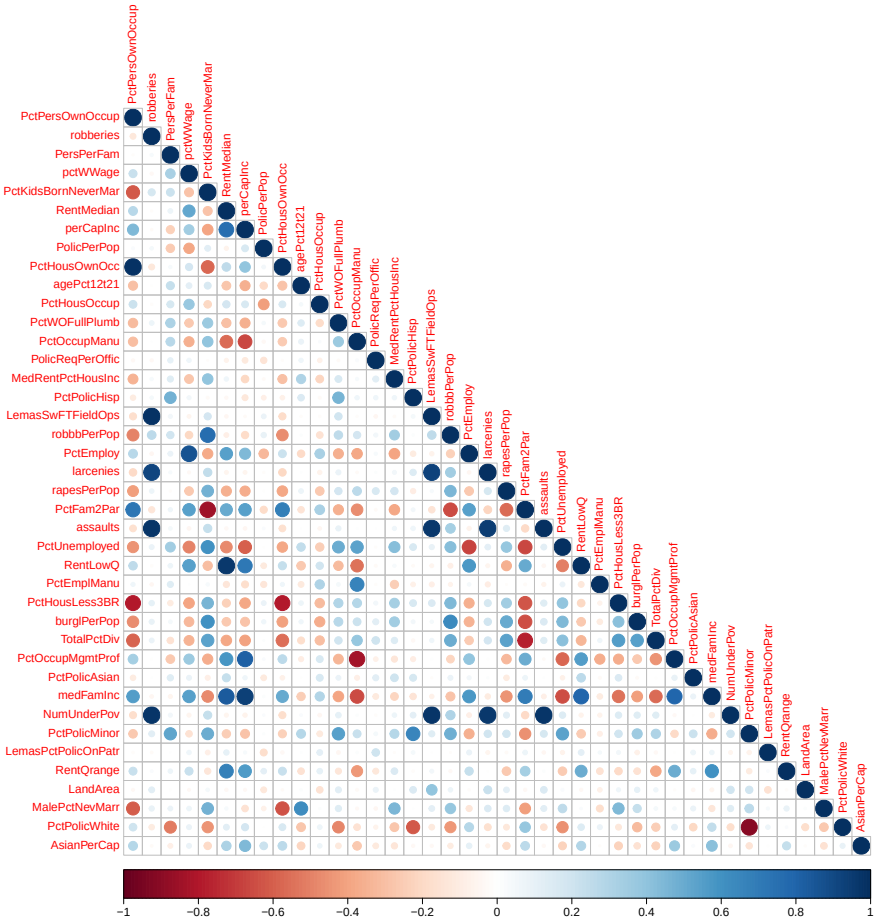

These scatterplots show clear and consistent patterns across key socioeconomic variables:

- Poverty, unemployment, and lower educational attainment (e.g., PctPopUnderPov, PctUnemployed, PctNotHSGrad) are positively associated with violent crime. Communities with higher levels of socioeconomic disadvantage tend to experience higher violent crime rates.
- Family-structure stability and income measures (e.g., PctKids2Par, PctFam2Par, medIncome, perCapInc) are negatively associated with violent crime. Communities with more two-parent households and higher income levels generally have lower violent crime rates.
- Several relationships exhibit nonlinear patterns, indicating that linear effects may not fully capture the underlying trends. This suggests that flexible modeling approaches such as polynomial terms, regression splines, or tree-based models may be beneficial in capturing these dynamics more accurately.

## Collinearity Among Predictors

```
set.seed(1)
subset_vars <- sample(names(numeric_train), 40)

cor_mat <- cor(numeric_train[, subset_vars], use = "pairwise.complete.obs")
```

```r
corrplot(cor_mat, type = "lower", tl.cex = 1.4, cl.cex = 1.4)
```




There is substantial multicollinearity among predictors:

- Poverty, income, education, and employment variables form a tight correlated block.
- Family structure variables also cluster strongly.
- Crime subcategories (e.g., burglary, auto theft, larceny counts) are correlated with one another.

Because of this, ordinary least squares regression will likely suffer from instability and inflated variance. This motivates Ridge regression, LASSO, and Principal Components Regression as more appropriate modeling choices.

## EDA Summary

- Problems with the data: Missingness in policing variables; strong skew in response; heavy collinearity.
- Variable types: Almost all predictors are normalized continuous variables; identifiers excluded from modeling.

- Missing data: Localized ~84% missingness in LEMAS policing variables (excluded).
- Collinearity: Very strong among socioeconomic predictors → requires regularization or dimensionality reduction.
- Important predictors: Poverty, income, education, unemployment, family structure, and housing stability.
- Key figures: Distribution of ViolentCrimesPerPop, missingness map, scatterplots of key predictors, correlation statistics, heatmap.
- Modeling approaches suggested:
  - Ridge regression (handles grouped correlation)
  - LASSO (variable selection)
  - PCR (dimension reduction)
  - Regression splines or polynomial regression (nonlinearity)
  - Tree-based models (nonlinear interactions, no need for collinearity handling)