Unemployment and Crime in the U.S.: Exploring State-Level Predictive Patterns

Feini Pek

Abstract

Understanding the socioeconomic and demographic factors influencing violent crime rates is crucial for developing effective policy interventions. This study evaluates the predictive power of unemployment and other socioeconomic indicators on violent crime rates across U.S. communities using multiple statistical and machine learning models. We compare Ordinary Least Squares (OLS) regression, polynomial regression, LASSO regression, and Random Forest, assessing their predictive accuracy using Mean Squared Error (MSE) and variance explained. Results indicate that polynomial regression and LASSO achieve the lowest test MSE, suggesting that incorporating nonlinear relationships enhances predictive accuracy. While Random Forest provides insights into variable importance, its higher test MSE limits its predictive utility in this context. Our findings highlight the advantages of penalized regression techniques, such as LASSO, in balancing interpretability and predictive performance. These insights contribute to data-driven crime prevention strategies by identifying key predictors of violent crime and refining quantitative approaches for policy analysis.

1. Introduction

Understanding the factors that contribute to violent crime rates is essential for designing effective crime prevention strategies. Socioeconomic and demographic factors, such as unemployment, income, and education levels, have long been associated with crime rates, yet quantifying their predictive power remains a subject of debate. Traditional regression models provide interpretability but may struggle to capture complex, nonlinear relationships among predictors. In contrast, machine learning models, such as Random Forest, can uncover intricate

patterns but often lack transparency, making them less suitable for policy interpretation. This study systematically compares various modeling approaches to determine the most effective method for predicting violent crime rates while balancing accuracy and interpretability.

The central research question guiding this study is: How unemployment level best predicts violent crime rates? Additionally, from an explanatory perspective, how do other relevant predictors—such as employment, income, and education levels—account for variations in violent crime rates across communities? Addressing these questions is crucial for policymakers, law enforcement agencies, and social scientists seeking data-driven strategies to mitigate crime.

Crime prevention has been the focus of extensive academic and policy discussions. Early economic models of crime, such as those proposed by Becker (1968), argue that criminal behavior is influenced by rational cost-benefit calculations, where individuals weigh the risks of punishment against potential rewards. In contrast, sociological theories emphasize the role of structural factors. Sampson and Laub (1993) highlight how family and community cohesion shape crime patterns, while Wilson and Kelling (1982) introduce the "broken windows" theory, which suggests that visible signs of disorder contribute to crime escalation. More recent studies integrate these perspectives, exploring how socioeconomic instability, policing policies, and urbanization interact to influence crime rates (Levitt, 2004; Glaeser & Sacerdote, 1999; Papachristos et al., 2013). Despite these valuable contributions, few studies have systematically compared predictive modeling techniques within the same analytical framework.

This study extends prior research by evaluating a range of statistical and machine learning models to identify the most effective approach for predicting violent crime rates. Specifically, we apply and compare five modeling techniques: base (log) OLS regression, polynomial (log) OLS regression, stepwise log-OLS regression, LASSO regression, and Random Forest**.** Each model is

assessed based on test Mean Squared Error (MSE) and variance explained to determine its predictive effectiveness. By systematically analyzing these methods, we aim to provide empirical insights that can inform crime prevention policies and optimize resource allocation.

By integrating traditional statistical models with machine learning approaches, this study contributes to a growing body of literature that seeks to refine quantitative methodologies in criminology. Our findings will offer a nuanced understanding of how key socioeconomic indicators influence violent crime and highlight the trade-offs between model interpretability and predictive performance. Ultimately, this research aims to support data-driven policy decisions that enhance public safety and reduce crime rates through targeted interventions.

2. Dataset

This study utilizes the Communities and Crime dataset from the UCI Machine Learning Repository, which integrates data from the 1990 U.S. Census, law enforcement agencies, and FBI Uniform Crime Reports. The dataset consists of 1,994 U.S. communities and includes 128 variables capturing demographic, socioeconomic, and law enforcement characteristics. To examine the relationship between unemployment rates and violent crime, we selected ViolentCrimesPerPop (violent crimes per 100,000 residents) as the dependent variable. The key predictors in this study include employment-related factors (PctUnemployed and PctEmploy), education levels (PctLess9thGrade and PctBSorMore), and socioeconomic indicators (perCapInc, PctFam2Par, and PctPopUnderPov). Additionally, the state variable is included to account for regional differences in crime patterns (see Appendix Table1.1).

The dataset contains no missing values in the selected variables, ensuring data completeness for statistical modeling. From exploratory data analysis (EDA), we observe that violent crime rates are highly skewed, with most communities reporting low crime levels while a few exhibit

significantly higher crime rates (Appendix G1.1). Second, regional differences are evident, as some states show notably higher average crime rates, suggesting the influence of local socioeconomic or policy factors (Appendix G1.2). Third, the distribution of unemployment rates indicates that while most communities fall within a moderate range, some experience exceptionally high unemployment, which may contribute to elevated crime levels (Appendix G1.3). Finally, a scatterplot analysis of unemployment rate vs. violent crime rate (Appendix G1.4) suggests a general positive relationship, where higher unemployment is associated with increased violent crime rates. However, substantial variation in the data indicates that unemployment is not the sole determinant of crime, as some communities report high crime rates despite low unemployment.

These initial findings underscore the complexity of crime prediction and highlight the need for further statistical modeling to assess the extent to which employment conditions influence violent crime rates.

3. Principal Component Analysis (PCA)

Before we start our analysis, Principal Component Analysis (PCA) was conducted to examine relationships among predictor variables and assess the degree of multicollinearity. By transforming the original variables into a smaller set of uncorrelated principal components, PCA helps simplify complex datasets and identify redundant predictors. Although we do not directly incorporate principal components into our regression models, this step provides valuable insights into underlying data structures and informs subsequent variable selection.

The scree plot (Appendix G2.1) illustrates the proportion of variance explained by each principal component. The first principal component alone accounts for 66.7% of the total variance, while the second adds 12.2%, with the remaining components contributing

progressively less. The sharp decline in variance beyond the second component suggests that a few key factors drive most of the variability in our predictors, indicating potential redundancy among variables. This finding aligns with expectations, as socioeconomic indicators such as income, poverty levels, and education tend to be highly correlated.

Given the strong correlations detected through PCA, we conducted a Variance Inflation Factor (VIF) analysis before fitting the Ordinary Least Squares (OLS) regression model. Multicollinearity can distort coefficient estimates and reduce interpretability, making VIF analysis essential for selecting independent predictors. A VIF above 5 suggests moderate concern, while values exceeding 10 indicate high redundancy. Our results showed that most variables had acceptable VIF values below 5; however, PctPopUnderPov (10.18) exhibited a strong correlation with other socioeconomic predictors, while perCapInc (7.63) and PctBSorMore (6.02) were also moderately high.

To mitigate multicollinearity, we will implement stepwise variable selection in the next section, ensuring that only the most relevant predictors remain. This process improves model stability, enhances interpretability, and allows for more reliable statistical inference.

4. Methodology

In this section, we implement various prediction models to assess their performance in predicting violent crime rates. To evaluate, we split the dataset into a training set (70%) and a test set (30%), with 1,415 and 579 observations, respectively. This allows the models to learn patterns from most of the data while keeping a separate portion to test their accuracy. The split was done randomly while ensuring that state-level differences were still represented. Comparing model performance on the test data helps us determine which approach best captures the relationship between unemployment, socioeconomic factors, and crime.

4.1 Ordinary Least Squares

We begin with Ordinary Least Squares (OLS) regression as a baseline for predicting violent crime rates. Using stepwise regression with the Bayesian Information Criterion (BIC), we identified six key predictors: state, PctLess9thGrade, PctUnemployed, PctEmploy, perCapInc, and PctFam2Par. The model achieves an adjusted R^2 of 0.5918, explaining 59.2% of the variance in crime rates. Examining the coefficients, PctUnemployed (0.1872) and PctEmploy (0.2168) are highly significant (Appendix M1). While higher unemployment aligns with increased crime, the positive effect of employment is counterintuitive, suggesting additional socioeconomic factors may influence crime rates.

However, diagnostic checks (Appendix G3.1) reveal mild non-linearity and heteroscedasticity, suggesting the model does not fully capture crime rate variation. The QQ plot highlights a heavy right tail, indicating that some communities experience significantly higher crime rates than expected, potentially skewing estimates. To improve normality and variance stability, we apply a Box-Cox transformation, which recommends a log transformation. This adjustment increases adjusted R^2 to 0.6185 (Appendix M2), reduces the magnitude of PctUnemployed (0.1271) and PctEmploy (0.1540), and better captures nonlinear relationships.

The second model shows improved predictive performance, with a Mean Squared Error (MSE) of 0.0104 on the test set. This benchmark allows for comparison with more flexible models, such as polynomial regression, which may better capture complex patterns.

4.2 Ordinary Least Squares with Polynomials

To address the signs of non-linearity we observed above, we applied polynomial regression, which provides greater flexibility in modeling these relationships. To determine which variables would benefit from polynomial terms, we first visualized the relationships between

PctLess9thGrade, PctUnemployed, PctEmploy, perCapInc, and PctFam2Par and the response variable (Appendix G3.2). The scatterplots revealed key nonlinear trends. PctUnemployed showed a sharp initial increase in crime rates that leveled off at higher values. PctEmploy displayed a diminishing effect, where increasing employment reduced crime but at a decreasing rate. perCapInc and PctFam2Par exhibited strong negative relationships with crime, with sharper declines at lower values, reinforcing the need for nonlinear modeling.

Based on these patterns, we introduced quadratic terms for PctLess9thGrade, PctUnemployed, and PctEmploy, and cubic terms for perCapInc and PctFam2Par. This adjustment improved model fit, increasing the adjusted R^2 to 0.6361 (Appendix M3). The polynomial terms for PctUnemployed and PctEmploy were particularly important. The positive first-order coefficients ($PctUnemployed_1$ = 1.3319, $PctEmploy_1$ = 1.3162) indicate that higher unemployment and employment are both associated with increased violent crime rates. However, the negative second-order terms ($PctUnemployed_2$ = -0.2327, $PctEmploy_2$ = -0.7520) suggest a nonlinear effect—unemployment and employment both exhibits diminishing returns in their impact on crime. Specifically, as unemployment rises, crime initially increases but eventually plateaus, indicating that extremely high unemployment may not continue to escalate crime rates at the same rate. Conversely, as employment rises, crime initially declines but then levels off, suggesting that employment alone does not fully mitigate crime risk and that other economic or social factors play a role.

Although diagnostic plots showed improvements, we further assessed whether a response variable transformation was necessary. The Box-Cox procedure recommended a log transformation, which further improved the adjusted R^2 to 0.652 and reduced the Mean Squared Error (MSE) to 0.00999. After transformation, the polynomial term coefficients slightly

decreased ($PctUnemployed_1$ = 1.0343, $PctUnemployed_2$ = -0.1603; $PctEmploy_1$ = 0.9575, $PctEmploy_2$ = -0.5620) (Appendix M4), indicating that the nonlinear relationships remained but were slightly attenuated after adjusting for skewness in the response variable.

These enhancements demonstrate that incorporating polynomial terms and a log transformation better captures the complex relationships between socioeconomic factors and violent crime rates, resulting in more accurate predictions.

4.3 Ordinary Least Squares with Step Function

While polynomial regression provides flexibility in modeling nonlinear relationships, it assumes smooth, continuous changes in crime rates as predictor values increase. However, real-world socioeconomic effects may exhibit abrupt shifts rather than gradual trends. To address this, we apply step functions, which divide continuous predictors into discrete bins, capturing segmented relationships without assuming a specific functional form. This approach helps identify threshold effects while reducing the risk of overfitting from high-degree polynomials.

To determine the optimal binning strategy, we tested 3 to 10 bins and evaluated their Mean Squared Error (MSE) on the test set. The MSE decreased as bin count increased, stabilizing at 9 bins (Appendix G3.3). We applied 9-bin step functions to key predictors (PctLess9thGrade, PctUnemployed, PctEmploy, perCapInc, and PctFam2Par) while ensuring consistent bin edges across the training and test sets. Appendix G3.3.1 provides an example of this binning process for PctUnemployed.

Using these transformed variables, we fitted an OLS model while maintaining the log-transformed response, as previous models demonstrated that this improves predictive performance. This model achieved an adjusted R^2 of 0.6334 (Appendix M5), slightly below the polynomial model but still an improvement over basic OLS. The test set MSE of 0.01036 was

competitive with polynomial regression, confirming that step functions effectively capture complex relationships.

Examining the estimated coefficients provides further insight into how crime rates respond to changes in predictor levels. Crime rates increase more sharply at higher unemployment levels, with bins above 35% showing the strongest effects—coefficients rise from 0.0459 for (0.35, 0.42] to 0.0969 for (0.62, 1]. Similarly, employment rates exhibit a nonlinear pattern, where bins between 37% and 71% are significantly associated with crime, with coefficients ranging from 0.0684 to 0.1067. This suggests that crime rates are lower at moderate employment levels but increase at extreme values, reinforcing the idea that employment alone does not guarantee lower crime rates.

Residual diagnostics indicate some improvements in variance stabilization compared to the polynomial model, but heteroscedasticity remains. The Q-Q plot suggests improved normality, though extreme values persist, likely reflecting outliers or high-crime communities. Additionally, the Residuals vs. Leverage plot identifies a few high-leverage points, suggesting that robust modeling techniques may be worth exploring (Appendix G3.3.2). Moving forward, we compare this approach with LASSO and tree-based models to explore potential further performance gains.

4.4 LASSO Regression

LASSO regression was implemented to enhance feature selection and prevent overfitting by applying L1 regularization, which shrinks some coefficients to zero. We tested three approaches: (1) a base model using original predictors, (2) a log-transformed model incorporating polynomial terms to capture nonlinear relationships, and (3) a log-transformed model with binned step functions. Each model underwent 10-fold cross-validation to determine the optimal λ, balancing model complexity and predictive performance.

4.4.1 LASSO 1: Base Model (see Appendix L1)

The first LASSO model included the original predictors: state, PctLess9thGrade, PctUnemployed, PctEmploy, perCapInc, PctFam2Par, and PctPopUnderPov. Cross-validation selected an optimal λ of 3.11549e-05, yielding an MSE of 0.02135. While all predictors were retained, weaker socioeconomic variables were shrunk toward zero. Notably, PctUnemployed (0.172) and PctEmploy (0.228) remained significant, suggesting that both unemployment and employment rates are positively associated with crime, with employment having a slightly stronger effect. The cross-validation error plot (Appendix G4.4.1) shows a U-shaped curve, where MSE decreases until the optimal λ, then rises as more coefficients shrink.

4.4.2 LASSO 2: Polynomial Features (see Appendix L2)

Building on the stronger predictive performance of polynomial models, we applied LASSO regression using the log-transformed violent crime variable to refine feature selection. This model included quadratic terms for PctLess9thGrade, PctUnemployed, and PctEmploy, and cubic terms for perCapInc and PctFam2Par, capturing potential nonlinear relationships. Cross-validation selected an optimal λ of 4.04813e-05, yielding a lower MSE of 0.00997, confirming that incorporating nonlinear terms improves model fit. The selected predictors emphasized higher-order effects, particularly for PctUnemployed and PctEmploy, reinforcing their complex associations with violent crime rates. Specifically, PctUnemployed's first-order term (1.16) was positive, but its second-order term (-0.15) was negative, suggesting that crime initially rises steeply at lower unemployment levels but plateaus at higher levels. Similarly, PctEmploy's first-order coefficient (1.23) was positive, with a second-order term (-0.67), indicating that while higher employment reduces crime, the effect diminishes at extreme employment levels. As

shown in the cross-validation plot (Appendix G4.4.2), the error curve stabilizes at lower λ values, indicating robust feature selection while maintaining a low MSE.

4.4.3 LASSO 3: Using Binned Data (see Appendix L3)

Given that the OLS model with step functions performed competitively with the polynomial model, we applied LASSO regression using binned predictors and the log-transformed violent crime rate to enhance feature selection and interpretability. Cross-validation identified an optimal λ of 2.50743e-05, yielding an MSE of 0.01035, similar to LASSO 2 (polynomial model). The model retained 48 bins across all variables, but certain bins were excluded, indicating that only specific intervals held predictive value. Notably, the step-function coefficients for PctUnemployed and PctEmploy suggest threshold effects rather than smooth trends. For instance, PctUnemployed between 0.3 and 0.35 had a coefficient of 0.026, while levels above 0.62 had a much stronger effect (0.092), implying that higher unemployment rates have a greater association with crime. Similarly, employment bins exhibited varying effects, with coefficients ranging from 0.02 to 0.10, reinforcing a nonlinear and segmented relationship between employment and crime. As shown in the cross-validation plot (Appendix G4.4.3), the error curve follows a similar pattern to LASSO 2, confirming that binning effectively captures nonlinearities while maintaining a parsimonious feature set. While polynomial modeling provides a smooth approximation, step functions may better highlight abrupt changes in the relationship between socioeconomic factors and crime.

Across all LASSO models, PctUnemployed and PctEmploy remained strong predictors but varied in representation. LASSO 1 retained them as continuous variables, showing a positive association with crime, with employment having a slightly stronger effect. LASSO 2 captured nonlinear trends, where unemployment's effect rose steeply at lower levels but plateaued higher,

while employment showed diminishing returns. LASSO 3 used step functions, revealing threshold effects where specific levels had disproportionate impacts. Both polynomial transformations and binning improved model performance. Choosing between LASSO 2 and LASSO 3 depends on priorities—LASSO 2 captures smooth nonlinear trends, while LASSO 3 highlights clear thresholds, making it more actionable for policy decisions.

4.5 Random Forest

Next, we explore Random Forest due to its ability to model complex relationships between socioeconomic factors and crime rates while handling both categorical and continuous predictors. Unlike traditional regression-based approaches, Random Forest captures nonlinear interactions between variables, making it well-suited for understanding how unemployment and the other key factors may contribute to violent crime rates.

To optimize model performance, we tuned mtry (randomly selected predictors per split) using a grid search from 2 to the total number of predictors (7), evaluated via five-fold cross-validation on RMSE. RMSE decreased up to mtry = 6 (Appendix G4.5a), but further inspection reveals that variance explained peaked at mtry = 4, making it the optimal choice. We also tested ntree (number of trees) and found that OOB (Out-of-Bag) error plateaued around 400 trees (Appendix G4.5b), suggesting diminishing returns. Based on this, we set ntree = 500 to balance accuracy and efficiency. The final model, trained with mtry = 4, ntree = 500, achieved a Mean Squared Residual of 0.0205, explained 61.68% of variance, and had a test MSE of 0.0202, indicating strong generalization.

Performing analysis of feature importance revealed that PctFam2Par (percentage of two-parent households) was the most influential predictor of violent crime rates (Appendix G4.5.1). This is further supported by the decision tree (Appendix G4.5.2), where PctFam2Par was the

primary splitting criterion, suggesting that communities with fewer two-parent households tend to have higher crime rates. This aligns with social theories emphasizing family structure as a protective factor against criminal behavior. Additionally, PctPopUnderPov (percentage of population under the poverty line) and PctUnemployed emerged as highly significant variables, reinforcing the strong link between economic hardship and crime prevalence. The decision tree further illustrates this relationship, with nodes splitting based on PctPopUnderPov and employment levels, indicating their substantial influence on crime rates. State was also a strong predictor, as reflected in both the feature importance plot and the tree structure. This suggests substantial geographic differences in crime prevalence, potentially driven by regional policies, law enforcement effectiveness, and broader economic conditions. The tree structure highlights that states with similar socioeconomic factors can exhibit different crime patterns, emphasizing the role of localized crime determinants. While perCapInc (per capita income) showed moderate importance, its effect was weaker than unemployment and poverty, implying that relative economic deprivation may drive crime more than absolute income. These results highlight the importance of addressing family instability, poverty, and unemployment in crime reduction efforts.

5. Model Comparison and Prediction Result

To assess model performance, we compared test and training Mean Squared Errors (MSE), variance explained (R^2), and adjusted R^2. Each model was selected based on its ability to minimize test MSE while balancing interpretability and generalizability (Appendix T5).

The baseline log-OLS model serves as a simple, interpretable benchmark, achieving a test MSE of 0.0104 and an adjusted R^2 of 0.618. However, its higher training MSE suggests potential

underfitting, limiting predictive power. The stepwise log-OLS model improves slightly (adjusted $R^2 = 0.633$) but is prone to instability, making it less reliable than regularized approaches.

Among OLS models, the polynomial log-OLS model performs best, achieving the lowest test MSE (0.00999) and the highest adjusted R^2 (0.652) by capturing nonlinear relationships. To improve feature selection and prevent overfitting, we applied LASSO regularization to this model. The LASSO polynomial log-OLS model retains a comparable test MSE (0.00997) while removing less relevant predictors, enhancing interpretability without sacrificing accuracy.

The random forest model, though effective at capturing complex interactions, does not outperform the regression-based models, achieving a higher test MSE (0.0202) and explaining 61.68% of variance. While suitable for high-dimensional data, its reduced interpretability and higher test error make it less optimal for this analysis.

Considering trade-offs between accuracy, interpretability, and complexity, the LASSO polynomial log-OLS model emerges as the best choice. It balances predictive accuracy and interpretability, ensuring that only the most relevant predictors are retained. Given the goal of this analysis—to develop a meaningful, interpretable model—the LASSO polynomial log-OLS model provides the most robust and actionable insights.

For our prediction analysis, we selected Observation 239 from the test dataset, representing Westfield City, Massachusetts (State = 25). This community has 27% with less than a 9th-grade education, 30% with a bachelor's degree or higher, 30% unemployment, and 52% employment. Per capita income falls in the 31st percentile, 64% of families are two-parent households, and 20% live below the poverty line. Using the LASSO polynomial model (lasso_model15), we predicted a violent crime rate of 0.312 vs. the observed 0.37, a slight underprediction. High unemployment and low income suggest higher crime risk, while two-parent households and

education levels may act as stabilizers. The underestimation implies unmeasured factors like law enforcement or community engagement may influence crime. Despite this, the model effectively captures key crime trends, making it a reliable tool for understanding unemployment influences on violent crime.

6. Conclusion

This study evaluated models for predicting violent crime rates using socioeconomic factors. OLS regression provided a baseline but struggled with nonlinearities. Polynomial regression improved fit by capturing nonlinear effects, while step functions highlighted threshold effects. LASSO regression enhanced interpretability by selecting key predictors, with the LASSO polynomial model emerging as the best-performing approach, balancing accuracy and complexity. Random Forest captured complex interactions but did not outperform regression models, reinforcing that simpler models can be effective when properly tuned.

Unemployment is a strong predictor of violent crime but follows a nonlinear pattern—crime increases sharply at lower unemployment levels but stabilizes at higher levels. This suggests that while rising joblessness initially fuels crime, extreme unemployment may not further escalate it, possibly due to alternative survival strategies, community adaptation, or saturation effects. Conversely, employment also showed a positive association with crime, implying that job availability alone is insufficient to reduce crime rates. This could reflect unstable or low-wage employment failing to provide economic security, leading to continued social strain.

Despite strong model performance, unmeasured factors like law enforcement and social policies may influence crime. Future research could explore causal relationships, spatial effects, or policy interventions to refine predictions. These findings support data-driven crime prevention and resource allocation strategies.

References

Becker, G. S. (1968). Crime and punishment: An economic approach. Journal of Political Economy, 76(2), 169-217.

Glaeser, E. L., & Sacerdote, B. (1999). Why is there more crime in cities? Journal of Political Economy, 107(S6), S225-S258.

Levitt, S. D. (2004). Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. Journal of Economic Perspectives, 18(1), 163-190.

Papachristos, A. V., Braga, A. A., & Hureau, D. M. (2013). Social networks and the risk of gunshot injury. Journal of Urban Health, 90(6), 1013-1031.

Sampson, R. J., & Laub, J. H. (1993). Crime in the making: Pathways and turning points through life. Harvard University Press.

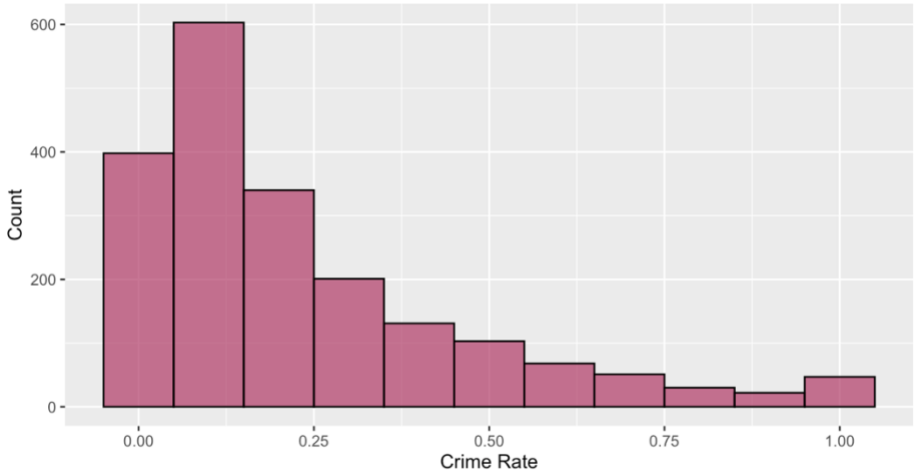
Appendix

Table 1.1 : Community Crime and Demographic Variables

Variable	Description
ViolentCrimesPerPop	Total number of violent crimes per 100K popuation (numeric - decimal)
State*	US state by number (categorical)
PctUnemployed	Percentage of people 16 and over, in the labor force, and unemployed (numeric - decimal)
PctEmploy	Percentage of people 16 and over who are employed (numeric - decimal)
PctLess9thGrade	Percentage of people 25 and over with less than a 9th grade education (numeric - decimal)
PctBSorMore	Percentage of people 25 and over with a bachelors degree or higher education (numeric - decimal)
perCapInc	Per capita income (numeric - decimal)
PctFam2Par	Percentage of families (with kids) that are headed by two parents (numeric - decimal)
PctPopUnderPov	Percentage of people under the poverty level (numeric - decimal)

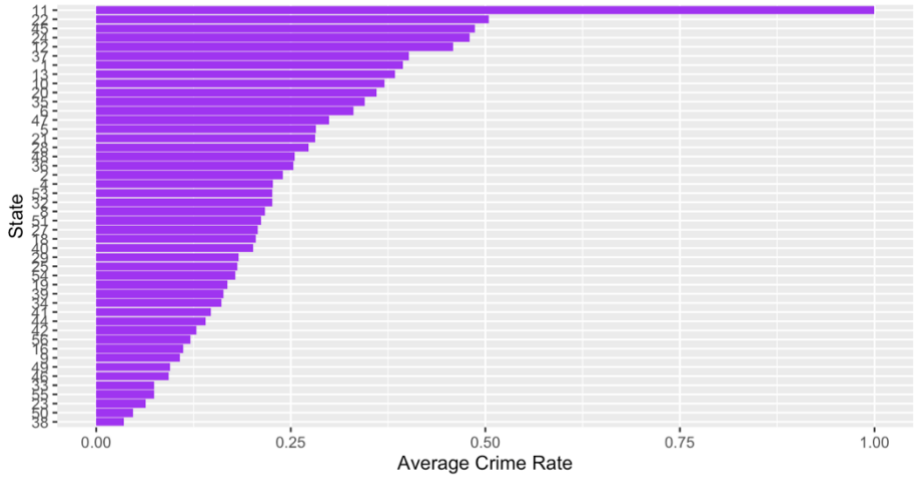
**State follows ANSI numbering system. For more information check: https://en.wikipedia.org/wiki/List_of_U.S._state_and_territory_abbreviations*

G1.1: Distribution of Violent Crimes Per Population



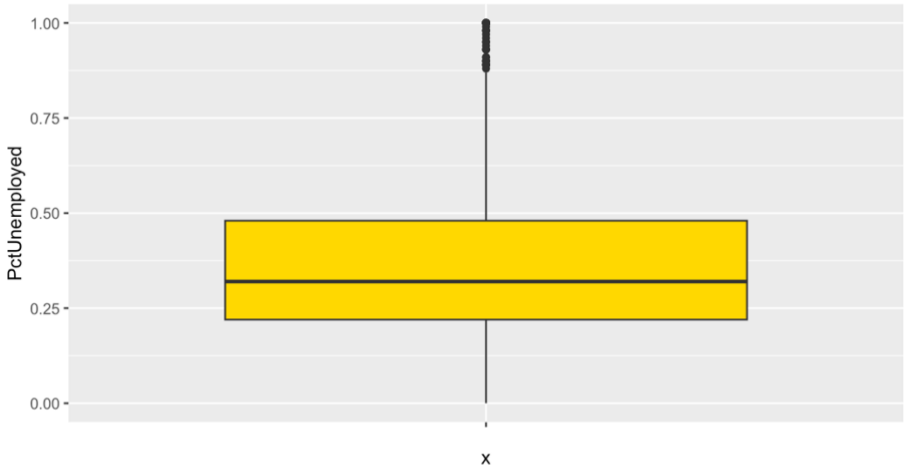
Note: This histogram shows the distribution of violent crimes per population in various communities. The crime rate ranges from 0 to 1, with a higher concentration of communities having lower crime rates. The plot indicates that violent crime rates are relatively skewed, with fewer communities exhibiting higher crime rates.

G1.2: Average Violent Crime Rate by State



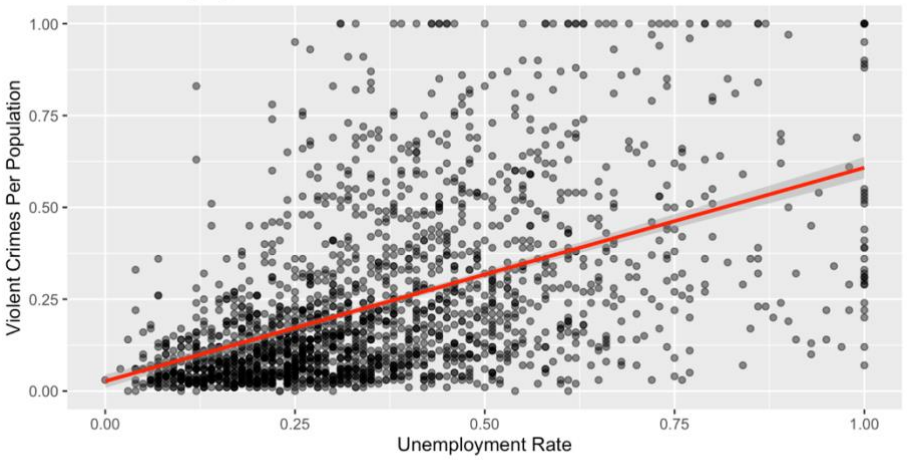
Note: This bar chart displays the average violent crime rate by state (by number). The states are ranked from lowest to highest based on their average violent crime rates. The x-axis represents the crime rate, while the y-axis lists the states. The highest crime rate is concentrated in a few states, while most states have relatively low average violent crime rates.

G1.3: Unemployment Rate Distribution



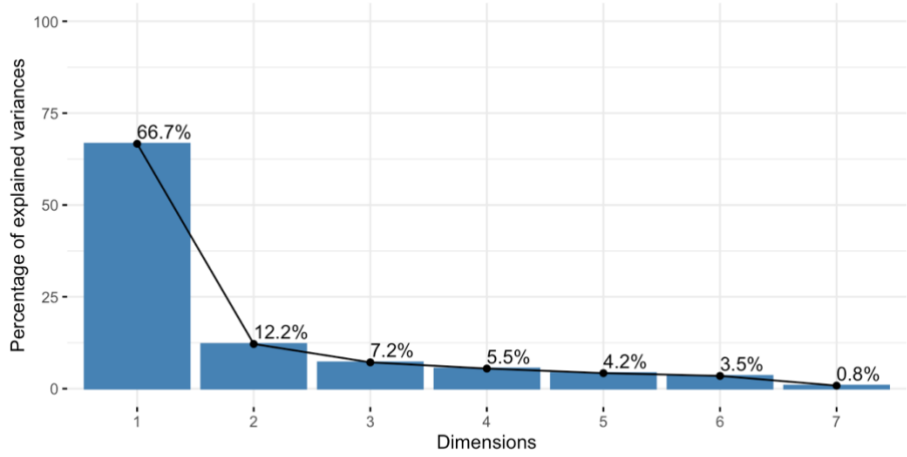
Note: This bar chart displays the average violent crime rate by state. The states are ranked from lowest to highest based on their average violent crime rates. The x-axis represents the crime rate, while the y-axis lists the states. The highest crime rate is concentrated in a few states, while most states have relatively low average violent crime rates.

G1.4: Unemployment Rate vs Crime Rate



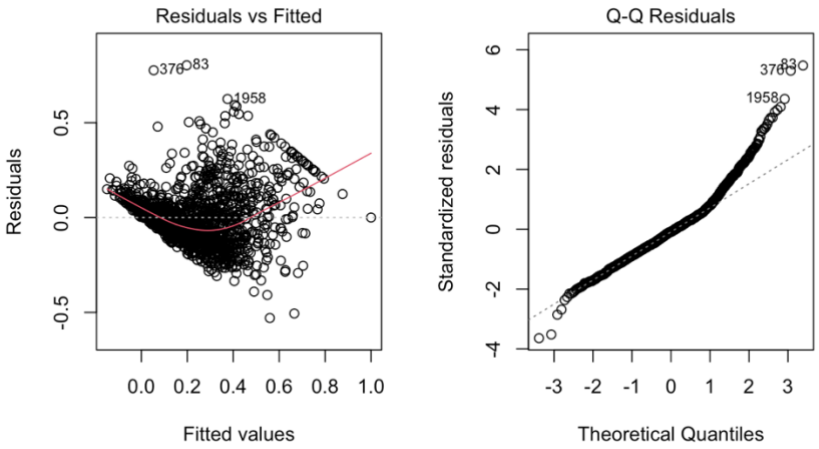
Note: This scatterplot illustrates the relationship between the unemployment rate and the violent crime rate per population across communities. Each point represents a community, with the x-axis showing the unemployment rate and the y-axis showing the violent crime rate. The red regression line suggests a positive correlation, indicating that communities with higher unemployment rates tend to have higher violent crime rates. However, the spread of points suggests variability in this relationship.

G2.1: Scree Plot of Principal Components



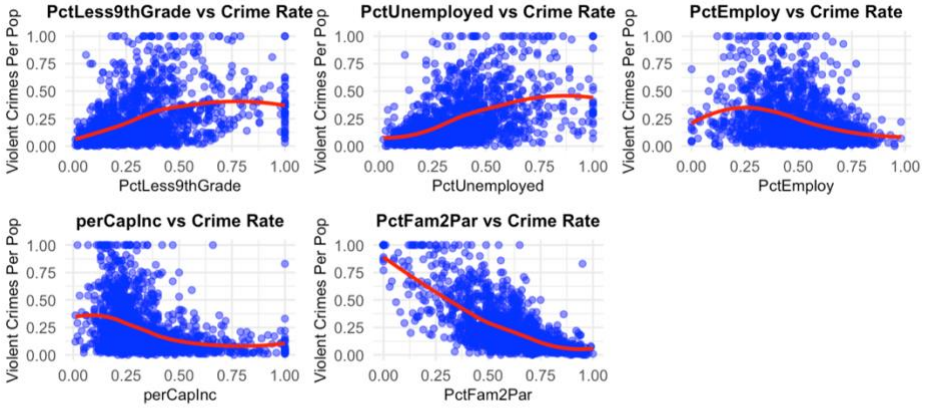
This scree plot displays the percentage of explained variance for each principal component in a principal component analysis (PCA). The first principal component explains the largest proportion of variance (66.7%), with subsequent components contributing progressively less. The sharp decline in variance suggests that only the first few components capture most of the information, indicating an 'elbow' point that can guide dimensionality reduction.

G3.1: OLS Diagnostic Plot



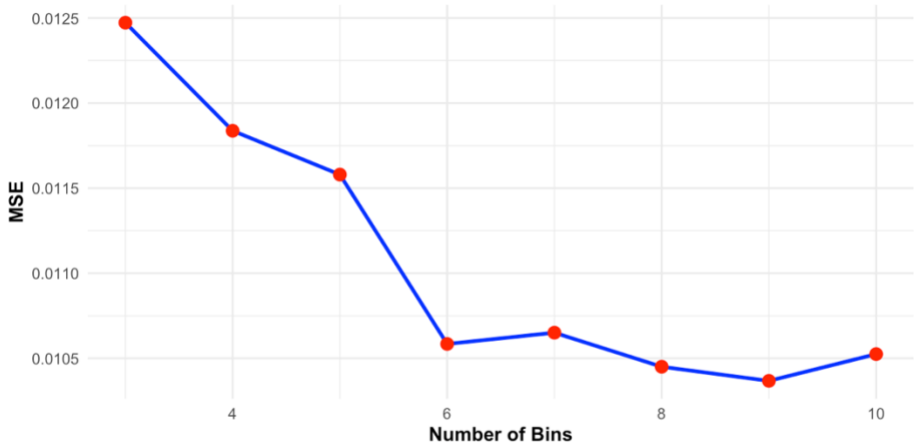
Note: These diagnostic plots assess the assumptions of ordinary least squares (OLS) regression. The 'Residuals vs Fitted' plot checks for non-linearity and heteroscedasticity, where an even spread suggests homoscedasticity. The 'Q-Q Residuals' plot examines normality of residuals; deviations from the diagonal indicate departures from normality.

G3.2: Scatterplot of Variables vs Crime Rate

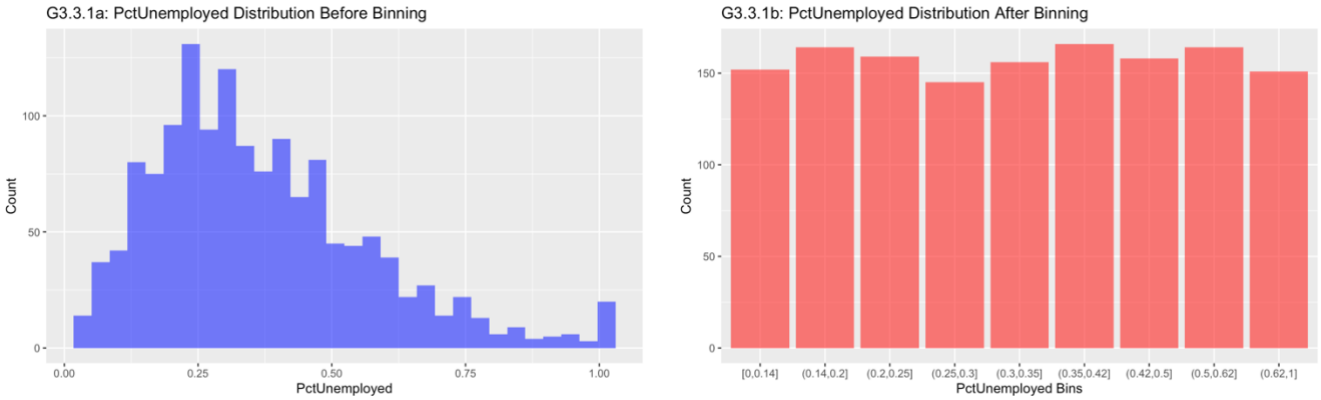


Note: These scatter plots illustrate the relationships between various socioeconomic factors and violent crime rates per
The blue dots represent individual communities, while the red LOWESS smoothing lines reveal trends.
Some variables, such as unemployment and lack of education, show a positive association with crime, whereas incor
exhibit a negative relationship. Notably, several variables, including employment and income, display nonlinear assoc
suggesting that the relationship between socioeconomic status and crime is more complex than a simple linear trend.

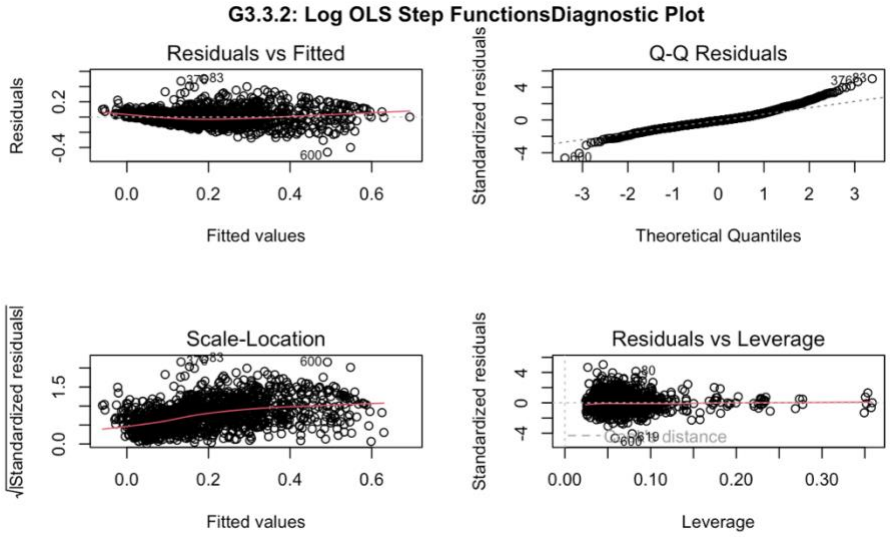
G3.3: Optimal Number of Bins Selection



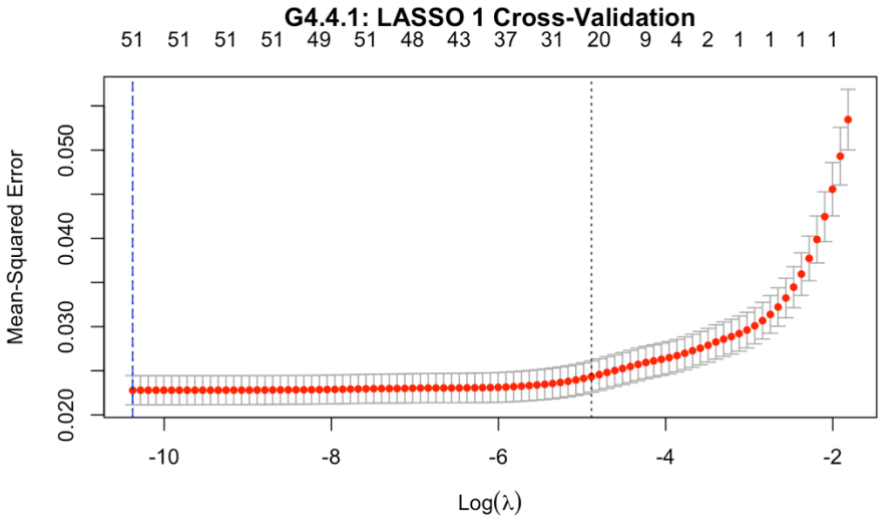
This plot shows the relationship between the number of bins and the Mean Squared Error (MSE).
The MSE decreases as the number of bins increases, reaching a minimum around 8-9 bins,
before slightly increasing again. This suggests that selecting around 9 bins may offer an
optimal balance between accuracy and model complexity.



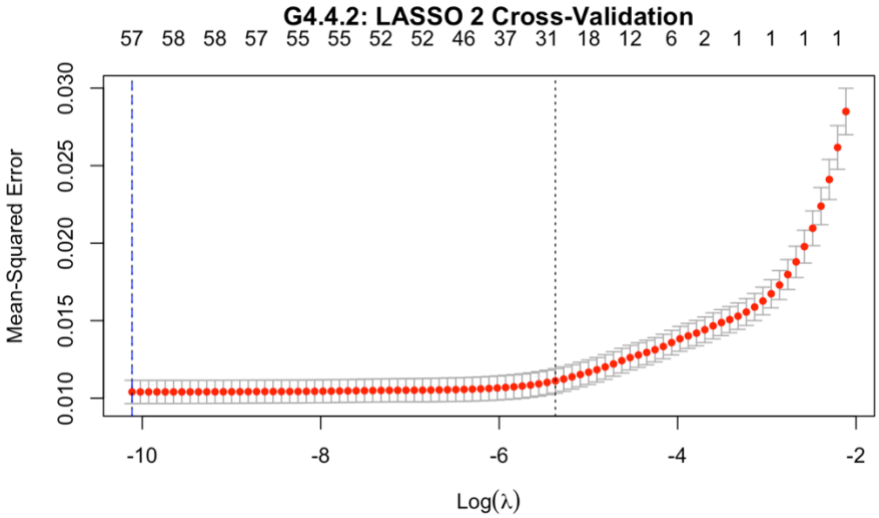
Note: Figure G3.3.1a and G3.3.1b illustrate the distribution of PctUnemployed (percentage of unemployed individuals) before and after binning. The left histogram (G3.3.1a) shows the raw distribution, where unemployment rates vary widely, with a peak around 25% and a long right tail. The right histogram (G3.3.1b) displays the distribution after applying a step function transformation, dividing PctUnemployed into nine bins based on quantiles. This ensures approximately equal-sized groups, facilitating a more interpretable relationship between unemployment and crime rates in the stepwise regression model. The binning process helps capture nonlinear effects and threshold impacts that may not be apparent in a continuous representation.



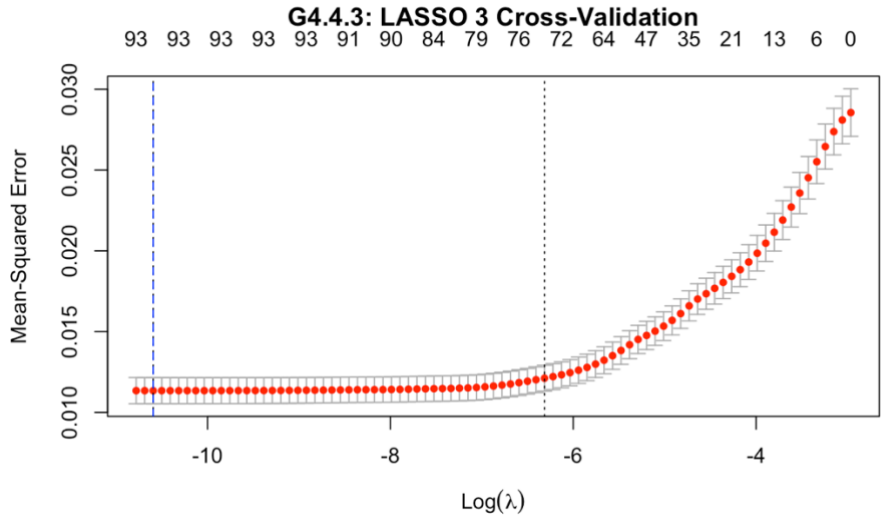
Note: These diagnostic plots evaluate the assumptions of the log-transformed OLS model with step functions. The Residuals vs. Fitted plot indicates mild heteroscedasticity, suggesting variance is not fully stabilized. The Q-Q plot shows that residuals are approximately normal but with some right-tail deviations, indicating potential outliers. The Scale-Location plot suggests relatively stable variance across fitted values, while the Residuals vs. Leverage plot identifies a few high-leverage points, suggesting that certain observations strongly influence the model. While step functions improve variance stabilization, minor assumption violations remain, warranting further robustness checks.



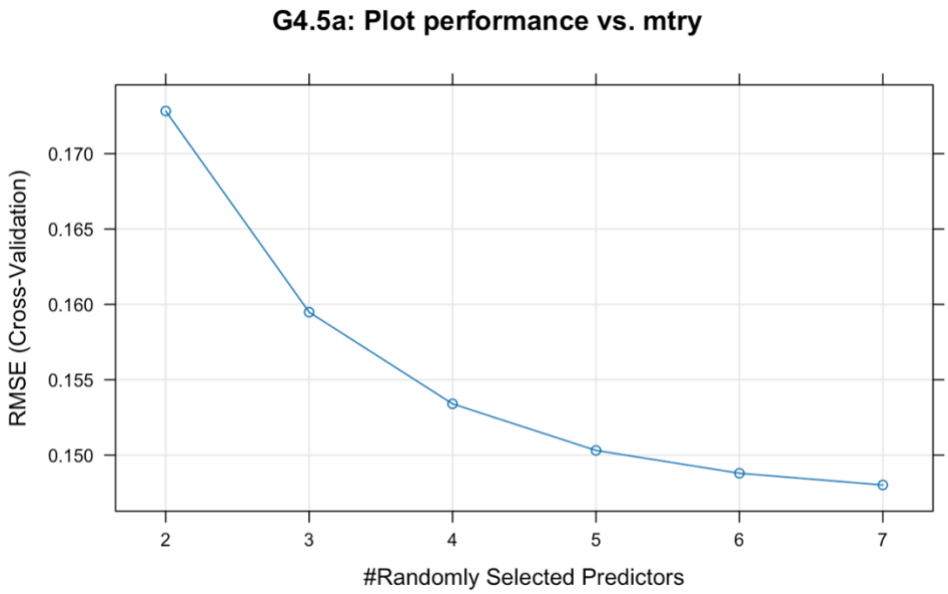
Note: This plot shows the cross-validation errors for different values of λ in the base LASSO model. The x-axis represents the logarithm of λ, while the y-axis shows the Mean Squared Error (MSE). Each red dot represents a tested λ value, with gray error bars indicating variability. The U-shaped curve suggests that very small or large λ values lead to higher errors. The optimal λ (λ.min = 3.11549e-05) is marked by the blue dashed line, representing the point where MSE is minimized before increasing as more coefficients shrink toward zero. This selection balances model complexity and predictive performance



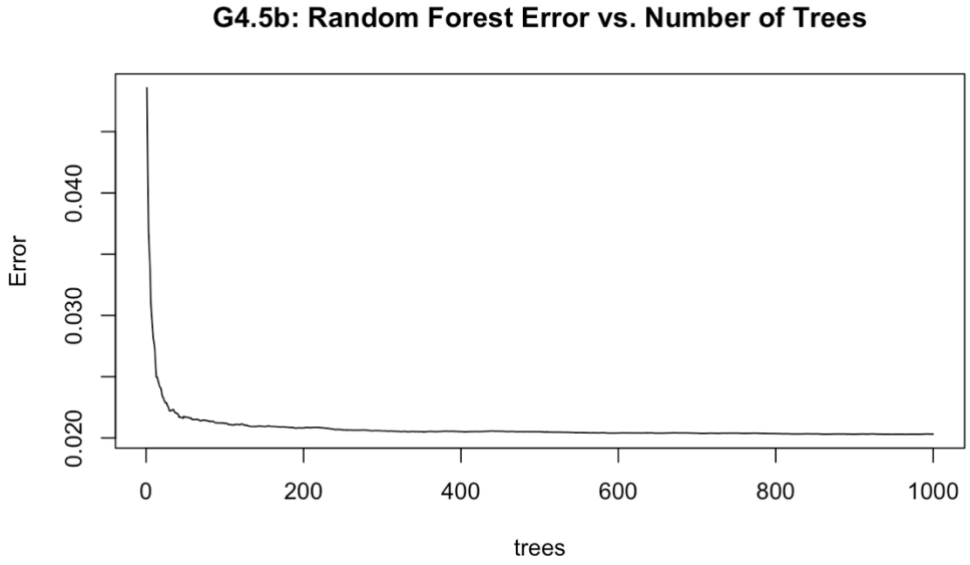
Note: This plot illustrates the cross-validation error for different values of λ in the LASSO regression model with polynomial features. The x-axis represents the log-transformed λ values, while the y-axis shows the Mean Squared Error (MSE). Each red dot corresponds to a tested λ value, with gray error bars indicating variability across cross-validation folds. The blue dashed line marks the optimal λ (λ.min = 4.04813e-05), where MSE is minimized before increasing as more coefficients shrink toward zero. The curve's stability at lower λ values confirms that incorporating polynomial features improves model fit while maintaining robustness.



Note: This plot displays the cross-validation error for different values of λ in the LASSO regression model using step-function predictors. The x-axis represents the log-transformed λ values, while the y-axis shows the Mean Squared Error (MSE). Each red dot corresponds to a tested λ value, with gray error bars indicating variability across cross-validation folds. The blue dashed line marks the optimal λ (λ.min = 2.50743e-05), where MSE is minimized before increasing as more coefficients shrink toward zero. The curve's stability at lower λ values suggests that binning effectively captures nonlinearities while maintaining robust feature selection.

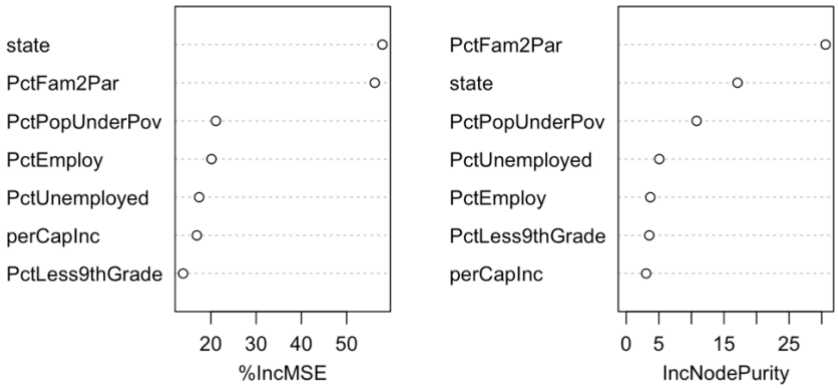


Note: This plot shows the relationship between the number of randomly selected predictors per split (mtry) and the Root Mean Squared Error (RMSE) from cross-validation. As mtry increases, RMSE decreases, indicating improved model performance. However, the rate of improvement diminishes beyond mtry = 6, suggesting diminishing returns. The optimal mtry was chosen as 4, balancing predictive accuracy and model complexity.



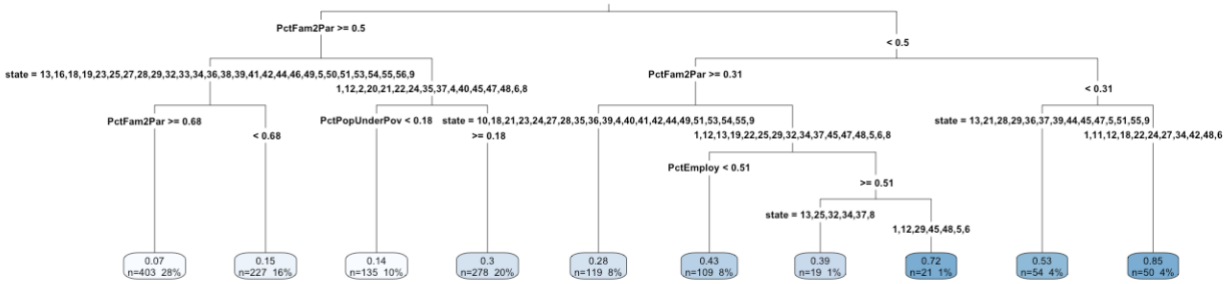
Note: This plot shows the relationship between the number of trees in the Random Forest model and the prediction error. The error decreases sharply as the number of trees increases but stabilizes around 400 trees, indicating diminishing returns. Based on this trend, 500 trees were selected as the optimal value to balance model performance and computational efficiency.

G4.5.1: Feature Importance in Random Forest Model



Note: This figure displays the relative importance of predictors in the Random Forest model for predicting violent crime rates. The left panel shows variable importance measured by the percentage increase in Mean Squared Error (%IncMSE) when a variable is permuted. The right panel presents importance based on IncNodePurity, which quantifies the reduction in impurity at each node where the variable is used. The results indicate that state and PctFam2Par (percentage of two-parent households) are the most influential predictors, followed by PctPopUnderPov (percentage of population under the poverty line), PctUnemployed, and PctEmploy. These findings suggest that both geographical and socioeconomic factors play a significant role in violent crime rates.

G4.5.2: Decision Tree for Predicting Violent Crime Rate



Note: This decision tree provides an interpretable model for predicting violent crime rates based on key socioeconomic and demographic variables. The root node splits on PctFam2Par, suggesting that the proportion of two-parent households is a primary determinant of crime rates. Subsequent splits involve state, PctPopUnderPov, and PctEmploy, highlighting the role of regional differences, poverty levels, and employment status in shaping crime patterns. Terminal nodes display predicted crime rates, with higher values associated with lower family stability, greater poverty, and reduced employment. This tree structure aligns with criminological theories that emphasize the protective effects of stable family environments and economic security against crime.

T5: Model Comparison Table

Model	Test_MSE	Train_MSE	Adj_Rsquared	Rsquared
Base log-OLS	0.0104095	0.0248059	0.618	0.6300
Poly log-OLS	0.0099883	0.0095220	0.652	0.6660
Step log-OLS	0.0103673	0.0248059	0.633	0.6575
LASSO Poly log-OLS	0.0099739	0.0097686	NA	0.6574
Random Forest	0.0202116	0.0205000	NA	0.6168

Note: This table presents a comparison of model performance metrics, including Test MSE, Train MSE, Adjusted R^2, and R^2. Lower Test MSE values indicate better predictive accuracy, while Adjusted R^2 and R^2 measure explanatory power. The Polynomial log-OLS model achieves the lowest Test MSE (0.0099) and highest Adjusted R^2 (0.652), indicating the best balance between fit and generalization.

M1 : OLS

ViolentCrimesPerPop = state + PctLess9thGrade + PctUnemployed + PctEmploy + perCapInc + PctFam2Par

	ViolentCrimesPerPop		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.63	0.54 – 0.72	**<0.001**
state [10]	-0.12	-0.42 – 0.17	0.422
state [11]	0.10	-0.19 – 0.40	0.491
state [12]	-0.01	-0.07 – 0.05	0.764
state [13]	-0.19	-0.27 – -0.11	**<0.001**
state [16]	-0.15	-0.29 – -0.01	**0.036**
state [18]	-0.19	-0.27 – -0.12	**<0.001**
state [19]	-0.12	-0.21 – -0.02	**0.016**
state [2]	-0.19	-0.37 – -0.01	**0.038**
state [20]	-0.06	-0.36 – 0.23	0.668
state [21]	-0.20	-0.28 – -0.11	**<0.001**
state [22]	-0.03	-0.13 – 0.06	0.450
state [23]	-0.33	-0.43 – -0.23	**<0.001**
state [24]	-0.14	-0.25 – -0.02	**0.017**
state [25]	-0.14	-0.21 – -0.08	**<0.001**
state [27]	-0.18	-0.32 – -0.04	**0.014**
state [28]	-0.29	-0.38 – -0.19	**<0.001**
state [29]	-0.18	-0.26 – -0.11	**<0.001**
state [32]	-0.16	-0.32 – -0.01	**0.037**
state [33]	-0.26	-0.35 – -0.17	**<0.001**
state [34]	-0.15	-0.21 – -0.09	**<0.001**
state [35]	-0.10	-0.22 – 0.02	0.107
state [36]	-0.26	-0.33 – -0.18	**<0.001**
state [37]	-0.15	-0.22 – -0.08	**<0.001**
state [38]	-0.27	-0.41 – -0.14	**<0.001**
state [39]	-0.19	-0.26 – -0.13	**<0.001**
state [4]	-0.16	-0.26 – -0.07	**0.001**
state [40]	-0.17	-0.25 – -0.09	**<0.001**
state [41]	-0.24	-0.32 – -0.16	**<0.001**
state [42]	-0.16	-0.23 – -0.10	**<0.001**
state [44]	-0.20	-0.28 – -0.11	**<0.001**
state [45]	-0.00	-0.09 – 0.08	0.918
state [46]	-0.23	-0.35 – -0.11	**<0.001**
state [47]	-0.08	-0.15 – 0.00	0.059
state [48]	-0.11	-0.17 – -0.05	**0.001**
state [49]	-0.12	-0.21 – -0.03	**0.011**
state [5]	-0.15	-0.24 – -0.06	**0.001**
state [50]	-0.32	-0.49 – -0.14	**<0.001**
state [51]	-0.22	-0.30 – -0.14	**<0.001**
state [53]	-0.20	-0.28 – -0.12	**<0.001**
state [54]	-0.22	-0.32 – -0.11	**<0.001**
state [55]	-0.22	-0.29 – -0.15	**<0.001**
state [56]	-0.23	-0.37 – -0.09	**0.001**
state [6]	-0.07	-0.13 – -0.01	**0.014**
state [8]	-0.16	-0.25 – -0.08	**<0.001**
state [9]	-0.20	-0.27 – -0.14	**<0.001**
PctLess9thGrade	0.06	0.00 – 0.12	**0.035**
PctUnemployed	0.19	0.11 – 0.27	**<0.001**
PctEmploy	0.22	0.15 – 0.29	**<0.001**
perCapInc	0.11	0.04 – 0.19	**0.002**
PctFam2Par	-0.80	-0.86 – -0.74	**<0.001**
Observations	1415		
R^2 / R^2 adjusted	0.606 / 0.592		

Table M1: Ordinary Least Squares (OLS) Regression Results for Violent Crime Prediction
This table presents the OLS regression results predicting ViolentCrimesPerPop (violent crimes per 100,000 residents) using socioeconomic factors and state-level fixed effects. The predictors include PctUnemployed (unemployment rate), PctEmploy (employment rate), PctLess9thGrade (percentage with less than a 9th-grade education), perCapInc (per capita income), and PctFam2Par (percentage of two-parent households). The results show that PctUnemployed and PctEmploy are positively associated with violent crime, suggesting that both unemployment and employment correlate with higher crime rates. Significance levels: $p < 0.05$, $p < 0.01$, and $p < 0.001$ are denoted in bold. Adjusted $R^2 = 0.592$, explaining approximately 59.2% of the variance in violent crime rates.

M2: log-OLS

log_ViolentCrimesPerPop = state + PctLess9thGrade + PctUnemployed + PctEmploy + perCapInc + PctFam2Par

	log_crime		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.48	0.42 – 0.55	**<0.001**
state [10]	-0.06	-0.27 – 0.15	0.587
state [11]	0.03	-0.18 – 0.24	0.785
state [12]	0.00	-0.04 – 0.05	0.889
state [13]	-0.12	-0.18 – -0.07	**<0.001**
state [16]	-0.10	-0.20 – -0.01	**0.039**
state [18]	-0.13	-0.18 – -0.08	**<0.001**
state [19]	-0.08	-0.15 – -0.02	**0.015**
state [2]	-0.11	-0.24 – 0.01	0.076
state [20]	-0.02	-0.22 – 0.19	0.885
state [21]	-0.12	-0.18 – -0.07	**<0.001**
state [22]	-0.01	-0.07 – 0.06	0.789
state [23]	-0.24	-0.31 – -0.17	**<0.001**
state [24]	-0.08	-0.16 – -0.00	**0.045**
state [25]	-0.09	-0.14 – -0.05	**<0.001**
state [27]	-0.12	-0.22 – -0.02	**0.019**
state [28]	-0.19	-0.26 – -0.13	**<0.001**
state [29]	-0.12	-0.18 – -0.07	**<0.001**
state [32]	-0.10	-0.21 – 0.01	0.072
state [33]	-0.19	-0.25 – -0.12	**<0.001**
state [34]	-0.10	-0.14 – -0.06	**<0.001**
state [35]	-0.06	-0.14 – 0.03	0.193
state [36]	-0.18	-0.23 – -0.12	**<0.001**
state [37]	-0.09	-0.14 – -0.04	**0.001**
state [38]	-0.21	-0.30 – -0.12	**<0.001**
state [39]	-0.14	-0.18 – -0.09	**<0.001**
state [4]	-0.10	-0.17 – -0.03	**0.003**
state [40]	-0.11	-0.16 – -0.05	**<0.001**
state [41]	-0.17	-0.23 – -0.11	**<0.001**
state [42]	-0.12	-0.16 – -0.07	**<0.001**
state [44]	-0.14	-0.20 – -0.08	**<0.001**
state [45]	0.00	-0.06 – 0.06	0.939
state [46]	-0.17	-0.25 – -0.08	**<0.001**
state [47]	-0.04	-0.09 – 0.02	0.162
state [48]	-0.06	-0.10 – -0.02	**0.005**
state [49]	-0.08	-0.15 – -0.02	**0.011**
state [5]	-0.10	-0.16 – -0.04	**0.001**
state [50]	-0.24	-0.36 – -0.11	**<0.001**
state [51]	-0.15	-0.20 – -0.09	**<0.001**
state [53]	-0.13	-0.18 – -0.07	**<0.001**
state [54]	-0.15	-0.22 – -0.07	**<0.001**
state [55]	-0.16	-0.21 – -0.11	**<0.001**
state [56]	-0.16	-0.26 – -0.06	**0.001**
state [6]	-0.03	-0.07 – 0.01	0.107
state [8]	-0.10	-0.17 – -0.04	**0.001**
state [9]	-0.14	-0.19 – -0.10	**<0.001**
PctLess9thGrade	0.05	0.01 – 0.09	**0.017**
PctUnemployed	0.13	0.07 – 0.18	**<0.001**
PctEmploy	0.15	0.11 – 0.20	**<0.001**
perCapInc	0.07	0.02 – 0.12	**0.009**
PctFam2Par	-0.58	-0.63 – -0.54	**<0.001**
Observations	1415		
R^2 / R^2 adjusted	0.632 / 0.619		

Table M2: Log-OLS Regression Results for Violent Crime Rates
This table presents the results of a log-transformed Ordinary Least Squares (OLS) regression model predicting violent crime rates. The dependent variable is the log of violent crimes per 100,000 residents (log_ViolentCrimesPerPop). Key predictors include unemployment (PctUnemployed), employment (PctEmploy), education levels (PctLess9thGrade), per capita income (perCapInc), and family structure (PctFam2Par). State fixed effects account for regional differences in crime patterns. The coefficients represent the estimated impact of each predictor on the log crime rate, with confidence intervals (CI) and significance levels (p-values). A positive coefficient indicates a direct association with crime, while a negative coefficient suggests an inverse relationship. The adjusted R^2 value (0.619) suggests that the model explains approximately 61.9% of the variance in crime rates. Significant predictors include unemployment, employment, family structure, and per capita income, reinforcing the socioeconomic drivers of crime.

M3: Polynomial OLS

ViolentCrimesPerPop = state + poly(PctLess9thGrade, 2) + poly(PctUnemployed, 2) + poly(PctEmploy, 2) + poly(perCapInc, 3) + poly(PctFam2Par, 3)

	ViolentCrimesPerPop		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.37	0.32 – 0.42	**<0.001**
state [10]	-0.09	-0.37 – 0.19	0.522
state [11]	-0.08	-0.37 – 0.21	0.592
state [12]	0.02	-0.05 – 0.08	0.623
state [13]	-0.23	-0.31 – -0.16	**<0.001**
state [16]	-0.14	-0.27 – -0.00	**0.049**
state [18]	-0.19	-0.26 – -0.12	**<0.001**
state [19]	-0.10	-0.19 – -0.01	**0.034**
state [2]	-0.11	-0.28 – 0.06	0.223
state [20]	-0.03	-0.31 – 0.25	0.839
state [21]	-0.18	-0.26 – -0.10	**<0.001**
state [22]	-0.01	-0.09 – 0.08	0.848
state [23]	-0.32	-0.41 – -0.22	**<0.001**
state [24]	-0.13	-0.24 – -0.03	**0.013**
state [25]	-0.15	-0.21 – -0.09	**<0.001**
state [27]	-0.17	-0.30 – -0.03	**0.015**
state [28]	-0.30	-0.39 – -0.21	**<0.001**
state [29]	-0.16	-0.23 – -0.09	**<0.001**
state [32]	-0.12	-0.27 – 0.03	0.104
state [33]	-0.24	-0.33 – -0.15	**<0.001**
state [34]	-0.17	-0.23 – -0.11	**<0.001**
state [35]	-0.06	-0.17 – 0.06	0.337
state [36]	-0.26	-0.33 – -0.19	**<0.001**
state [37]	-0.16	-0.23 – -0.09	**<0.001**
state [38]	-0.25	-0.38 – -0.13	**<0.001**
state [39]	-0.20	-0.26 – -0.14	**<0.001**
state [4]	-0.11	-0.20 – -0.02	**0.013**
state [40]	-0.14	-0.21 – -0.06	**<0.001**
state [41]	-0.21	-0.29 – -0.13	**<0.001**
state [42]	-0.19	-0.25 – -0.13	**<0.001**
state [44]	-0.24	-0.32 – -0.15	**<0.001**
state [45]	-0.01	-0.09 – 0.07	0.858
state [46]	-0.21	-0.32 – -0.09	**0.001**
state [47]	-0.08	-0.15 – -0.00	**0.037**
state [48]	-0.08	-0.13 – -0.02	**0.010**
state [49]	-0.16	-0.25 – -0.07	**<0.001**
state [5]	-0.14	-0.22 – -0.06	**0.001**
state [50]	-0.27	-0.44 – -0.10	**0.001**
state [51]	-0.21	-0.28 – -0.13	**<0.001**
state [53]	-0.16	-0.24 – -0.09	**<0.001**
state [55]	-0.23	-0.29 – -0.16	**<0.001**
state [56]	-0.21	-0.35 – -0.08	**0.002**
state [6]	-0.04	-0.09 – 0.02	0.195
state [8]	-0.12	-0.20 – -0.04	**0.004**
state [9]	-0.22	-0.29 – -0.15	**<0.001**
PctLess9thGrade [1st degree]	0.83	0.33 – 1.33	**0.001**
PctLess9thGrade [2nd degree]	-0.10	-0.50 – 0.30	0.638
PctUnemployed [1st degree]	1.33	0.69 – 1.98	**<0.001**
PctUnemployed [2nd degree]	-0.23	-0.63 – 0.16	0.250
PctEmploy [1st degree]	1.32	0.85 – 1.78	**<0.001**
PctEmploy [2nd degree]	-0.75	-1.08 – -0.43	**<0.001**
perCapInc [1st degree]	0.28	-0.34 – 0.89	0.379
perCapInc [2nd degree]	-0.35	-0.77 – 0.07	0.099
perCapInc [3rd degree]	0.33	-0.01 – 0.68	0.059
PctFam2Par [1st degree]	-5.34	-5.81 – -4.86	**<0.001**
PctFam2Par [2nd degree]	2.05	1.71 – 2.39	**<0.001**
PctFam2Par [3rd degree]	0.37	0.07 – 0.67	**0.015**
Observations	1415		
R^2 / R^2 adjusted	0.651 / 0.636		

Table M3: Polynomial OLS Regression Results for Violent Crime Rates
This table presents the results of a polynomial OLS regression predicting violent crime rates per 100,000 residents (ViolentCrimesPerPop). State fixed effects control for regional differences, while key socioeconomic predictors are modeled nonlinearly. Quadratic terms (second-degree) are applied to PctUnemployed (unemployment rate), PctEmploy (employment rate), and PctLess9thGrade (percentage of the population with less than a 9th-grade education) to capture diminishing or amplifying effects. Cubic terms (third-degree) are included for perCapInc (per capita income) and PctFam2Par (percentage of two-parent households), allowing for more flexible relationships. The model highlights nonlinear effects, with unemployment's influence steepening at lower levels before plateauing, while family structure and income exhibit more complex trends. The adjusted R^2 of 0.636 indicates an improved fit over linear models, reinforcing the importance of modeling crime as a nonlinear function of socioeconomic conditions.

M4: Log Polynomial OLS

log_ViolentCrimesPerPop ~ state+ poly(PctLess9thGrade, 2) + poly(PctUnemployed, 2) + poly(PctEmploy, 2) + poly(perCapInc, 3) + poly(PctFam2Par, 3)

	log_crime		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.29	0.25 – 0.32	**<0.001**
state [10]	-0.04	-0.24 – 0.16	0.669
state [11]	-0.07	-0.27 – 0.14	0.522
state [12]	0.02	-0.03 – 0.06	0.420
state [13]	-0.15	-0.20 – -0.10	**<0.001**
state [16]	-0.09	-0.19 – 0.00	0.061
state [18]	-0.13	-0.18 – -0.09	**<0.001**
state [19]	-0.07	-0.13 – -0.01	**0.034**
state [2]	-0.06	-0.19 – 0.06	0.295
state [20]	0.00	-0.20 – 0.20	0.992
state [21]	-0.12	-0.17 – -0.06	**<0.001**
state [22]	0.01	-0.05 – 0.07	0.820
state [23]	-0.24	-0.30 – -0.17	**<0.001**
state [24]	-0.08	-0.15 – -0.00	**0.040**
state [25]	-0.10	-0.14 – -0.05	**<0.001**
state [27]	-0.11	-0.21 – -0.02	**0.019**
state [28]	-0.20	-0.26 – -0.14	**<0.001**
state [29]	-0.11	-0.16 – -0.06	**<0.001**
state [32]	-0.08	-0.18 – 0.03	0.150
state [33]	-0.18	-0.24 – -0.11	**<0.001**
state [34]	-0.12	-0.16 – -0.08	**<0.001**
state [35]	-0.03	-0.11 – 0.05	0.467
state [36]	-0.18	-0.23 – -0.13	**<0.001**
state [37]	-0.10	-0.15 – -0.05	**<0.001**
state [38]	-0.19	-0.28 – -0.10	**<0.001**
state [39]	-0.14	-0.18 – -0.10	**<0.001**
state [4]	-0.07	-0.13 – -0.01	**0.031**
state [40]	-0.09	-0.14 – -0.04	**0.001**
state [41]	-0.15	-0.21 – -0.10	**<0.001**
state [42]	-0.13	-0.18 – -0.09	**<0.001**
state [44]	-0.17	-0.23 – -0.11	**<0.001**
state [45]	0.00	-0.06 – 0.06	0.985
state [46]	-0.15	-0.23 – -0.07	**<0.001**
state [47]	-0.04	-0.10 – 0.01	0.108
state [48]	-0.04	-0.08 – -0.00	**0.044**
state [49]	-0.11	-0.17 – -0.05	**0.001**
state [5]	-0.10	-0.15 – -0.04	**0.001**
state [50]	-0.21	-0.33 – -0.09	**0.001**
state [51]	-0.14	-0.19 – -0.08	**<0.001**
state [53]	-0.11	-0.16 – -0.06	**<0.001**
state [55]	-0.16	-0.21 – -0.12	**<0.001**
state [56]	-0.15	-0.25 – -0.06	**0.002**
state [6]	-0.01	-0.05 – 0.03	0.489
state [8]	-0.08	-0.14 – -0.02	**0.008**
state [9]	-0.16	-0.20 – -0.11	**<0.001**
PctLess9thGrade [1st degree]	0.61	0.26 – 0.97	**0.001**
PctLess9thGrade [2nd degree]	-0.07	-0.35 – 0.22	0.640
PctUnemployed [1st degree]	1.03	0.57 – 1.49	**<0.001**
PctUnemployed [2nd degree]	-0.16	-0.44 – 0.12	0.267
PctEmploy [1st degree]	0.96	0.63 – 1.29	**<0.001**
PctEmploy [2nd degree]	-0.56	-0.79 – -0.33	**<0.001**
perCapInc [1st degree]	0.23	-0.21 – 0.67	0.297
perCapInc [2nd degree]	-0.30	-0.59 – 0.00	0.051
perCapInc [3rd degree]	0.27	0.02 – 0.51	**0.032**
PctFam2Par [1st degree]	-3.92	-4.26 – -3.58	**<0.001**
PctFam2Par [2nd degree]	1.21	0.96 – 1.45	**<0.001**
PctFam2Par [3rd degree]	0.32	0.11 – 0.54	**0.003**
Observations	1415		
R^2 / R^2 adjusted	0.666 / 0.652		

Table M4: Polynomial log OLS Regression Results for Violent Crime Rates
This table presents the results of a log-transformed polynomial OLS regression predicting violent crime rates. State fixed effects account for regional differences, while key socioeconomic predictors are modeled nonlinearly. PctUnemployed, PctEmploy, and PctLess9thGrade are modeled with quadratic terms (second-degree), while perCapInc and PctFam2Par are modeled with cubic terms (third-degree). The log transformation stabilizes variance and improves model fit, reflected in an adjusted R^2 of 0.652. The results indicate that unemployment and employment both have nonlinear effects on crime, with unemployment's impact increasing sharply at lower levels before plateauing, and employment showing diminishing returns.

M5: OLS with Step Functions

	log_ViolentCrimesPerPop		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.27	0.19 – 0.35	**<0.001**
state [10]	-0.08	-0.29 – 0.13	0.444
state [11]	0.09	-0.12 – 0.30	0.381
state [12]	0.03	-0.01 – 0.08	0.145
state [13]	-0.13	-0.18 – -0.07	**<0.001**
state [16]	-0.14	-0.24 – -0.04	**0.006**
state [18]	-0.11	-0.16 – -0.06	**<0.001**
state [19]	-0.07	-0.14 – -0.01	**0.035**
state [2]	-0.06	-0.18 – 0.07	0.383
state [20]	0.01	-0.20 – 0.22	0.941
state [21]	-0.10	-0.16 – -0.05	**0.001**
state [22]	0.01	-0.06 – 0.07	0.837
state [23]	-0.23	-0.30 – -0.16	**<0.001**
state [24]	-0.06	-0.14 – 0.02	0.129
state [25]	-0.08	-0.12 – -0.03	**0.001**
state [27]	-0.13	-0.23 – -0.03	**0.013**
state [28]	-0.18	-0.25 – -0.11	**<0.001**
state [29]	-0.10	-0.15 – -0.04	**<0.001**
state [32]	-0.05	-0.16 – 0.06	0.338
state [33]	-0.17	-0.23 – -0.10	**<0.001**
state [34]	-0.09	-0.13 – -0.04	**<0.001**
state [35]	-0.06	-0.15 – 0.02	0.149
state [36]	-0.15	-0.20 – -0.10	**<0.001**
state [37]	-0.09	-0.15 – -0.04	**<0.001**
state [38]	-0.20	-0.29 – -0.10	**<0.001**
state [39]	-0.12	-0.16 – -0.07	**<0.001**
state [4]	-0.08	-0.15 – -0.02	**0.016**
state [40]	-0.10	-0.15 – -0.04	**0.001**
state [41]	-0.14	-0.20 – -0.09	**<0.001**
state [42]	-0.11	-0.15 – -0.06	**<0.001**
state [44]	-0.16	-0.22 – -0.10	**<0.001**
state [45]	-0.01	-0.06 – 0.05	0.855
state [46]	-0.18	-0.27 – -0.09	**<0.001**
state [47]	-0.04	-0.09 – 0.02	0.182
state [48]	-0.05	-0.09 – -0.01	**0.018**
state [49]	-0.11	-0.18 – -0.05	**0.001**
state [5]	-0.10	-0.16 – -0.04	**0.001**
state [50]	-0.18	-0.31 – -0.06	**0.004**
state [51]	-0.14	-0.20 – -0.08	**<0.001**
state [53]	-0.10	-0.16 – -0.05	**<0.001**
state [54]	-0.13	-0.20 – -0.05	**0.001**
state [55]	-0.15	-0.20 – -0.10	**<0.001**
state [56]	-0.15	-0.25 – -0.05	**0.003**
state [6]	-0.00	-0.04 – 0.04	0.880
state [8]	-0.09	-0.15 – -0.03	**0.005**
state [9]	-0.13	-0.18 – -0.08	**<0.001**
PctLess9thGrade bins [>0.1-0.15]	0.01	-0.01 – 0.03	0.405
PctLess9thGrade bins [>0.15-0.19]	0.02	-0.01 – 0.04	0.250
PctLess9thGrade bins [>0.19-0.24]	0.04	0.01 – 0.07	**0.007**
PctLess9thGrade bins [>0.24-0.3]	0.02	-0.01 – 0.05	0.154
PctLess9thGrade bins [>0.3-0.36]	0.02	-0.01 – 0.05	0.238
PctLess9thGrade bins [>0.36-0.44]	0.02	-0.01 – 0.05	0.238
PctLess9thGrade bins [>0.44-0.586]	0.05	0.02 – 0.08	**0.003**
PctLess9thGrade bins [>0.586-1]	0.04	0.01 – 0.07	**0.024**
PctUnemployed bins [>0.14-0.2]	0.00	-0.02 – 0.03	0.730
PctUnemployed bins [>0.2-0.25]	0.01	-0.02 – 0.04	0.493
PctUnemployed bins [>0.25-0.3]	0.02	-0.01 – 0.04	0.287
PctUnemployed bins [>0.3-0.35]	0.03	-0.00 – 0.06	0.069
PctUnemployed bins [>0.35-0.42]	0.05	0.01 – 0.08	**0.006**
PctUnemployed bins [>0.42-0.5]	0.05	0.02 – 0.09	**0.003**
PctUnemployed bins [>0.5-0.62]	0.06	0.03 – 0.10	**0.001**
PctUnemployed bins [>0.62-1]	0.10	0.06 – 0.14	**<0.001**
PctEmploy bins [>0.28-0.37]	0.02	-0.00 – 0.05	0.087
PctEmploy bins [>0.37-0.43]	0.07	0.04 – 0.09	**<0.001**
PctEmploy bins [>0.43-0.49]	0.06	0.03 – 0.09	**<0.001**
PctEmploy bins [>0.49-0.54]	0.11	0.08 – 0.14	**<0.001**
PctEmploy bins [>0.54-0.58]	0.11	0.08 – 0.14	**<0.001**
PctEmploy bins [>0.58-0.64]	0.09	0.06 – 0.12	**<0.001**

PctEmploy bins [>0.64-0.71]	0.10	0.07 – 0.13	**<0.001**
PctEmploy bins [>0.71-1]	0.11	0.07 – 0.14	**<0.001**
perCapInc bins [>0.17-0.21]	-0.01	-0.03 – 0.02	0.694
perCapInc bins [>0.21-0.24]	0.01	-0.01 – 0.04	0.321
perCapInc bins [>0.24-0.28]	0.03	-0.00 – 0.06	0.051
perCapInc bins [>0.28-0.33]	0.06	0.02 – 0.09	**0.001**
perCapInc bins [>0.33-0.38]	0.06	0.02 – 0.10	**0.002**
perCapInc bins [>0.38-0.45]	0.06	0.02 – 0.10	**0.002**
perCapInc bins [>0.45-0.58]	0.05	0.01 – 0.09	**0.026**
perCapInc bins [>0.58-1]	0.05	0.00 – 0.09	**0.038**
PctFam2Par bins [>0.36-0.47]	-0.14	-0.17 – -0.12	**<0.001**
PctFam2Par bins [>0.47-0.54]	-0.21	-0.24 – -0.18	**<0.001**
PctFam2Par bins [>0.54-0.6]	-0.23	-0.26 – -0.20	**<0.001**
PctFam2Par bins [>0.6-0.65]	-0.24	-0.27 – -0.21	**<0.001**
PctFam2Par bins [>0.65-0.71]	-0.26	-0.29 – -0.23	**<0.001**
PctFam2Par bins [>0.71-0.78]	-0.28	-0.31 – -0.25	**<0.001**
PctFam2Par bins [>0.78-0.85]	-0.29	-0.33 – -0.25	**<0.001**
PctFam2Par bins [>0.85-1]	-0.29	-0.33 – -0.25	**<0.001**
PctPopUnderPov bins [>0.06-0.1]	0.01	-0.02 – 0.04	0.546
PctPopUnderPov bins [>0.1-0.15]	-0.00	-0.03 – 0.03	0.994
PctPopUnderPov bins [>0.15-0.21]	0.02	-0.01 – 0.06	0.238
PctPopUnderPov bins [>0.21-0.29]	0.06	0.02 – 0.10	**0.006**
PctPopUnderPov bins [>0.29-0.38]	0.07	0.03 – 0.12	**0.001**
PctPopUnderPov bins [>0.38-0.47]	0.10	0.05 – 0.15	**<0.001**
PctPopUnderPov bins [>0.47-0.6]	0.12	0.07 – 0.17	**<0.001**
PctPopUnderPov bins [>0.6-1]	0.12	0.06 – 0.17	**<0.001**
Observations	1415		
R^2 / R^2 adjusted	0.657 / 0.633		

Table M5: OLS Regression with Step Functions for Predicting Violent Crime Rates
This table presents the results of an OLS regression model using step functions to account for nonlinear relationships between predictors and violent crime rates. The dependent variable is log_ViolentCrimesPerPop, representing the log-transformed violent crime rate per 100,000 residents. Predictors include state-level fixed effects, education levels (PctLess9thGrade), unemployment (PctUnemployed), employment (PctEmploy), per capita income (perCapInc), family structure (PctFam2Par), and poverty levels (PctPopUnderPov), all transformed into discrete bins to capture threshold effects. Estimates represent the change in log crime rates associated with each predictor, with confidence intervals (CI) and p-values provided for significance testing. Statistically significant coefficients ($p < 0.05$) are bolded. The adjusted R^2 value indicates model fit, showing the proportion of variance explained.

L1: Lasso 1

Variable	Coefficient
(Intercept)	0.594
state10	-0.107
state11	0.121
state12	0.007
state13	-0.176
state16	-0.146
state18	-0.177
state19	-0.105
state2	-0.172
state20	-0.050
state21	-0.182
state22	-0.026
state23	-0.310
state24	-0.120
state25	-0.124
state27	-0.161
state28	-0.280
state29	-0.170
state32	-0.147
state33	-0.244
state34	-0.130
state35	-0.092
state36	-0.240
state37	-0.137
state38	-0.264
state39	-0.179
state4	-0.148
state40	-0.156
state41	-0.230
state42	-0.146
state44	-0.180
state45	0.006
state46	-0.226
state47	-0.063
state48	-0.096
state49	-0.107
state5	-0.140
state50	-0.301
state51	-0.208
state53	-0.185
state54	-0.201
state55	-0.207
state56	-0.218
state6	-0.057
state8	-0.149
state9	-0.187
PctLess9thGrade	0.060
PctUnemployed	0.172
PctEmploy	0.228
perCapInc	0.117
PctFam2Par	-0.786
PctPopUnderPov	0.043

Table L1: LASSO Regression Coefficients for Predicting Violent Crime Rates
This table presents the estimated coefficients from the LASSO regression model (L1), which applies L1 regularization to select the most predictive features while reducing multicollinearity. The dependent variable is violent crime rate per 100,000 residents (ViolentCrimesPerPop). The predictors include state-level fixed effects, education levels (PctLess9thGrade), unemployment (PctUnemployed), employment (PctEmploy), per capita income (perCapInc), family structure (PctFam2Par), and poverty levels (PctPopUnderPov). Coefficients represent the direction and magnitude of the relationship between each predictor and crime rates, with positive values indicating an increasing effect on crime and negative values indicating a decreasing effect. Strong predictors include PctUnemployed (0.172) and PctEmploy (0.228), highlighting the complex relationship between employment and crime. PctFam2Par (-0.786) suggests a strong protective effect of two-parent households. LASSO's penalty function shrinks some coefficients toward zero, improving model interpretability and reducing overfitting.

L2: Lasso 2

Variable	Coefficient
(Intercept)	0.254
state10	-0.030
state11	-0.049
state12	0.035
state13	-0.140
state16	-0.085
state18	-0.119
state19	-0.058
state2	-0.049
state20	0.010
state21	-0.104
state22	0.012
state23	-0.221
state24	-0.062
state25	-0.081
state27	-0.098
state28	-0.191
state29	-0.096
state32	-0.060
state33	-0.162
state34	-0.100
state35	-0.023
state36	-0.163
state37	-0.086
state38	-0.181
state39	-0.123
state4	-0.057
state40	-0.079
state41	-0.136
state42	-0.117
state44	-0.150
state45	0.012
state46	-0.142
state47	-0.031
state48	-0.033
state49	-0.092
state5	-0.086
state50	-0.189
state51	-0.125
state53	-0.094
state54	-0.105
state55	-0.150
state56	-0.139
state6	0.000
state8	-0.067
state9	-0.140
poly(PctLess9thGrade, 2)1	0.743
poly(PctLess9thGrade, 2)2	-0.127
poly(PctUnemployed, 2)1	1.160
poly(PctUnemployed, 2)2	-0.151
poly(PctEmploy, 2)1	1.228
poly(PctEmploy, 2)2	-0.667
poly(perCapInc, 3)1	0.466
poly(perCapInc, 3)2	-0.506
poly(perCapInc, 3)3	0.416
poly(PctFam2Par, 3)1	-4.501
poly(PctFam2Par, 3)2	1.452
poly(PctFam2Par, 3)3	0.377
PctPopUnderPov	0.066

Table L2: Lasso 2 model applies LASSO regression with polynomial transformations to predict violent crime rates while controlling for multicollinearity. It incorporates state-level fixed effects alongside key socioeconomic predictors, including education levels (PctLess9thGrade), unemployment (PctUnemployed), employment (PctEmploy), per capita income (perCapInc), family structure (PctFam2Par), and poverty levels (PctPopUnderPov). Quadratic transformations were applied to education, unemployment, and employment to capture diminishing returns or threshold effects, while cubic transformations were used for per capita income and family structure to account for more complex nonlinear relationships. The results show that unemployment and employment have strong nonlinear effects on crime, with significant first-degree terms (PctUnemployed: 1.160, PctEmploy: 1.228) and opposing second-degree effects ($PctUnemployed^2$: -0.151, $PctEmploy^2$: -0.667), indicating that while employment initially reduces crime, its effect diminishes beyond a certain point. The cubic transformation of $PctFam2Par^3$ (-4.501) suggests a complex protective effect of two-parent households. LASSO's regularization ensures model sparsity, prioritizing the most predictive variables while improving interpretability.

L3: Lasso 3

Variable	Coefficient
(Intercept)	0.276
state10	-0.072
state11	0.103
state12	0.042
state13	-0.119
state16	-0.133
state18	-0.104
state19	-0.064
state2	-0.046
state20	0.016
state21	-0.097
state22	0.014
state23	-0.218
state24	-0.052
state25	-0.069
state27	-0.117
state28	-0.173
state29	-0.090
state32	-0.045
state34	-0.079
state35	-0.055
state36	-0.140
state37	-0.085
state38	-0.190
state39	-0.112
state4	-0.075
state40	-0.088
state41	-0.137
state42	-0.100
state44	-0.150
state45	0.002
state46	-0.175
state47	-0.030
state48	-0.043
state49	-0.108
state5	-0.092
state50	-0.176
state51	-0.132
state53	-0.093
state54	-0.118
state55	-0.146
state56	-0.145
state6	0.005
state8	-0.080
state9	-0.118
PctLess9thGrade_bins(0.1,0.15]	0.009
PctLess9thGrade_bins(0.15,0.19]	0.014
PctLess9thGrade_bins(0.19,0.24]	0.037
PctLess9thGrade_bins(0.24,0.3]	0.019
PctLess9thGrade_bins(0.3,0.36]	0.017
PctLess9thGrade_bins(0.36,0.44]	0.018
PctLess9thGrade_bins(0.44,0.586]	0.048
PctLess9thGrade_bins(0.586,1]	0.037
PctUnemployed_bins(0.14,0.2]	0.003
PctUnemployed_bins(0.2,0.25]	0.007
PctUnemployed_bins(0.25,0.3]	0.014
PctUnemployed_bins(0.3,0.35]	0.026
PctUnemployed_bins(0.35,0.42]	0.043
PctUnemployed_bins(0.42,0.5]	0.050
PctUnemployed_bins(0.5,0.62]	0.060
PctUnemployed_bins(0.62,1]	0.093
PctEmploy_bins(0.28,0.37]	0.020
PctEmploy_bins(0.37,0.43]	0.067

PctEmploy_bins(0.43,0.49]	0.060
PctEmploy_bins(0.49,0.54]	0.107
PctEmploy_bins(0.54,0.58]	0.106
PctEmploy_bins(0.58,0.64]	0.087
PctEmploy_bins(0.64,0.71]	0.100
PctEmploy_bins(0.71,1]	0.104
perCapInc_bins(0.17,0.21]	-0.007
perCapInc_bins(0.21,0.24]	0.013
perCapInc_bins(0.24,0.28]	0.029
perCapInc_bins(0.28,0.33]	0.053
perCapInc_bins(0.33,0.38]	0.055
perCapInc_bins(0.38,0.45]	0.057
perCapInc_bins(0.45,0.58]	0.041
perCapInc_bins(0.58,1]	0.041
PctFam2Par_bins(0.36,0.47]	-0.141
PctFam2Par_bins(0.47,0.54]	-0.211
PctFam2Par_bins(0.54,0.6]	-0.234
PctFam2Par_bins(0.6,0.65]	-0.237
PctFam2Par_bins(0.65,0.71]	-0.259
PctFam2Par_bins(0.71,0.78]	-0.282
PctFam2Par_bins(0.78,0.85]	-0.291
PctFam2Par_bins(0.85,1]	-0.295
PctPopUnderPov_bins(0.06,0.1]	0.006
PctPopUnderPov_bins(0.1,0.15]	-0.003
PctPopUnderPov_bins(0.15,0.21]	0.019
PctPopUnderPov_bins(0.21,0.29]	0.054
PctPopUnderPov_bins(0.29,0.38]	0.069
PctPopUnderPov_bins(0.38,0.47]	0.097
PctPopUnderPov_bins(0.47,0.6]	0.111
PctPopUnderPov_bins(0.6,1]	0.112

Table L3: Lasso with step functions
For Lasso 3, the model applied step functions to categorize key predictors into discrete bins. The state-level controls remain included, with coefficients reflecting their respective contributions to violent crime rates. Among socioeconomic factors, PctLess9thGrade (percentage of individuals with less than a ninth-grade education) was divided into bins, with coefficients suggesting a nonlinear relationship—higher education levels appear to have a stabilizing effect on crime, though with diminishing returns. PctUnemployed was also binned, revealing that crime rates increase significantly as unemployment rises, but the marginal effect decreases at higher levels. Similarly, PctEmploy (employment rate) shows a non-monotonic relationship, suggesting that employment alone is not a straightforward deterrent to crime. PerCapInc (per capita income) exhibits mixed effects across bins, with higher income generally associated with lower crime but some inconsistencies. PctFam2Par (percentage of two-parent families) consistently shows a negative association with crime, emphasizing its role as a protective factor. PctPopUnderPov (percentage of population under poverty) demonstrates increasing crime rates in higher bins, reinforcing economic distress as a crime risk factor. This model highlights how different thresholds in socioeconomic variables influence crime rates, making it particularly useful for policy interventions targeting specific population segments.