Big Data Final: Communities and Crime

Anupam Bhattacharjee, Elizabeth Hallenborg, Gabe Lipton Galbraith, Nicolas Vila

May 30, 2023

We pledge our honor that we have not violated the Honor Code during the preparation of this assignment.

Contents

0 Executive Summary

This comprehensive report analyzes the factors contributing to violent crime rates across communities in the United States. The analysis is based on a dataset provided by the UCI Machine Learning Repository called the Communities and Crime Data Set.

The key factors driving violent crime rates were initially explored through data visualization techniques. These revealed two main axes of variables contributing to violent crime: economic conditions (evidenced by median income, level of poverty, and level of unemployment) and family formation (such as whether children are born to married parents and the incidence of divorce).

Various statistical and machine-learning techniques were subsequently employed to deepen the analysis. Linear and Lasso regression techniques identified additional statistically relevant determinants of violent crime rates, including demographic factors (specifically age), housing, and the level of immigration in given communities.

A decision tree model was developed to identify the critical variables predicting violent crime rates, revealing the importance of family structure (PctFam2Par) and the racial composition (racePctWhite) of a community. However, the most robust predictive performance came from a Random Forest model. This ensemble learning technique, which constructs multiple decision trees, effectively addressed overfitting issues and demonstrated strong predictive capability.

Unsupervised learning techniques were also used, with Principal Component Analysis (PCA) and K-Means clustering employed to identify the most influential variables and categorize communities based on these factors. The PCA showed the distinction between structured and unstructured families as highly important to the incidence of violent crime, with the level of immigration also being significant. The K-Means clustering technique effectively grouped communities into clusters based on similarities in their characteristics.

The report concludes that a complex combination of economic conditions, family-level factors, demographics, and immigration levels drive violent crime rates. It challenges simple economic explanations, proposing instead a more nuanced interpretation based on the specific circumstances of individual communities. Furthermore, it emphasizes the predictive power of the Random Forest model over linear regression models, which showed high R-squared values but were susceptible to overfitting considering the high dimensional nature of the dataset.

Future work could include the examination of additional community characteristics, further validation of the models on different datasets, and an exploration of other machine learning models and techniques. The insights from this analysis could inform targeted policy interventions aimed at reducing violent crime rates in specific communities.

1 Introduction to Dataset & Analysis

In this report, we examine US crime data compiled by Michael Redmond, La Salle University, based on data from the US Census (1990), the US FBI Uniform Crime Report (1995), and the US Law Enforcement Management and Administrative Statistics Survey (1990) accessible through the UCI Machine Learning Repository. The dataset includes 147 demographic and crime-related variables for 2215 unique communities across 46 US states and the District of Columbia.

Our project is motivated by a desire to better understand the factors that contribute to violent crime across different communities in the United States. In particular, we lay out three main motivating research questions below.

Research questions:

1. **What factors drive violent crime?** Our goal is to identify, as best as possible, the most important variables for predicting levels of violent crime. This requires significant regularization and out-of-sample testing to achieve dimension reduction and address the issue of significant correlation across our predictor variables.

2. **What models do the best in terms of predicting violent crime?** Using a combination of linear and more complex regression techniques, as well as supervised and unsupervised methods, we'd like to determine what model specifications can best estimate and predict rates of violent crime.

3. **How can we categorize the factors that contribute to violent crime?** Are there specific themes or categories that emerge as we look at the range of predictors that are strongly related to violent crime rates?

1.1 Analysis of Variables & Data Preparation

1.1.1 Data Import and Initial Exploration

The data is imported from a CSV file and an initial exploration is carried out to understand the data types of each variable and to count non-null values for each variable.

1.1.2 Data Cleaning and Pre-processing

Initial Processing

First, in order to create a more robust dataset, we identified and excluded variables where the most frequent value in the variable occurred at least 95% of the time by calculating the percentage of the most frequent variable. Initially, this led to no variables being excluded from the data. We focus on the numeric data for most of our analysis, which corresponds to 145 columns, as the only non-numeric columns are the names of communities (column: communityname) and the states where these communities are located (column: state). The dependent variable for the project is centered around predicting ViolentCrimesPerPop.

Correlation Analysis

- Using the numeric columns in the dataset, we computed the correlation matrices for the data to create a list of strongly correlated variables with the dependent variable under investigation, ViolentCrimesPerPop.

- Identified the variables with strong correlation (both positive and negative) (over 0.5) and moderate correlation (0.3-0.5) with the dependent variable 'ViolentCrimesPerPop'.

- Plotted the scatter plots of these variables versus 'ViolentCrimesPerPop' to better understand the distribution of these variables.

Identification of Highly Correlated Variable Pairs

- Identified pairs of variables that are highly positively correlated with each other (correlation greater than 0.8) and printed.

Missing Value Handling

- Identified and removed variables that have more than 80% missing values - 22 variables were removed because of this.
- For the remaining variables with missing values that are less than 80%, we replaced missing values with the median values of the respective variables.
- Post this transformation, we performed a check to identify if any variables have been significantly biased towards the median due to this imputation.
- Next, we checked for the potential introduction of bias due to the median transformation.
- We removed variables we found to have significantly changed because of bias introduced by the transformation (heuristic: if the standard deviation of the data fell by more than 10%, the variable may be biased and have to be removed). Thankfully, no variables were identified at this threshold.

As the end of this exercise, we are left with 125 variables and 2,215 instances to work with.

2 Exploratory Analysis

We start by exploring the dataset to better understand what variables are highly correlated with elevated rates of violent crime, any possible interactions that might exist between these variables, and how the drivers of violent crime differ across states in the US.

2.1 Variable Correlation with Violent Crime Rates

As a first step, we plot the variables that are highly correlated with high rates of violent crime across the US. In particular, we isolate variables for which the absolute value of the correlation is above 0.5 and generate scatter plots based on the county-level correlation between these variables and violent crime rates.

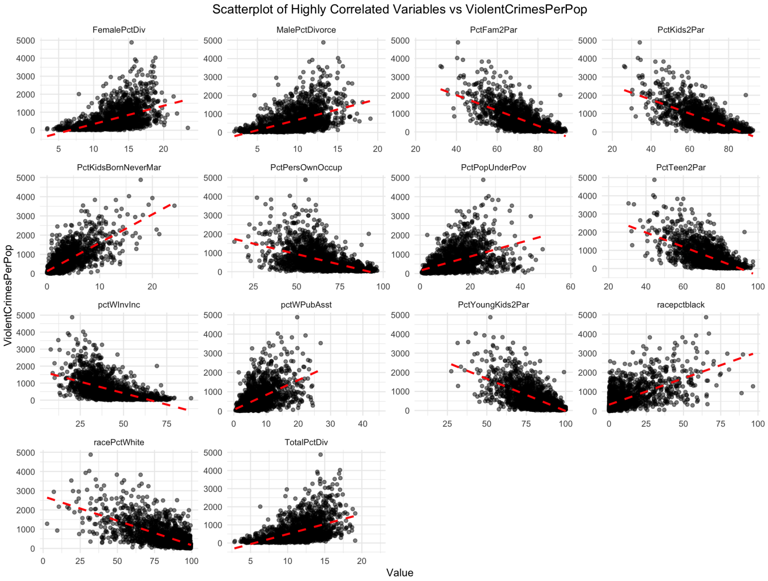


Figure 1: Relationship between highly correlated variables and violent crime rates

From this data, a few important observations can be made. First, it appears that aspects of family formation are strongly related to levels of violent crime. Divorce rates, the prevalence of two-parent households, and marriage rates are all strongly related to levels of violent crime, with divorce rates being positively correlated and two-parent households and marriage rates negatively correlated. Second, measures of income also appear to be important as well, including the percent of the population under the poverty line and the percent of the population on public assistance. Third, whether the racial make-up of a county is Black or White is strongly correlated with violent crime levels. Crucially, given there's likely a high degree of multicolinearity within the dataset, we will need to focus on variable selection via regularization to better tease out which of these might be strong predictors of violent crime rates.

We also plotted those variables that have a somewhat weaker but still meaningful correlation (absolute value between 0.3 and 0.5) with violent crime rates to get a better picture of our high-dimension data set. We find it interesting that some income measures, such as median income and per capita income, have a somewhat weaker correlation with rates of violent crime. This suggests that the drivers of violent crime rates may be more nuanced than simple deterministic economic explanations.

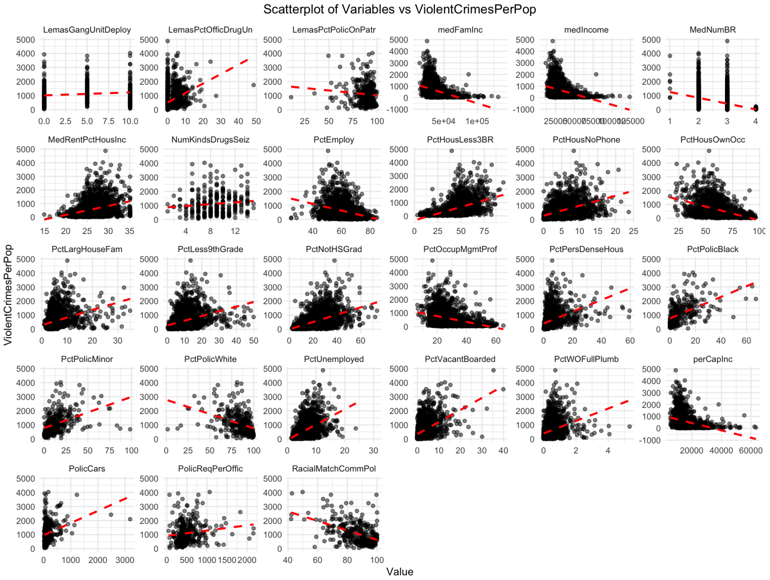


Figure 2: Relationship between moderately correlated variables and violent crime rates

2.2 Outcome Variable Analysis

In addition to examining our independent variables, we also wanted to better understand the statistical behavior and distribution of our dependent variable–violent crime per capita.

2.2.1 Variable Transformation

Based on an initial examination of the distribution of our outcome variable, it is evident that it exhibits a high degree of skewness. In order to ensure the accuracy of our analysis using the selected methods and to enhance the accuracy of our results, we made the decision to perform a log-transformation on the outcome variable. By implementing this transformation, we have observed that the variable distribution becomes a Gaussian, which is preferred for the majority of statistical models utilized in our analysis. This transformation is depicted in Figure 3.

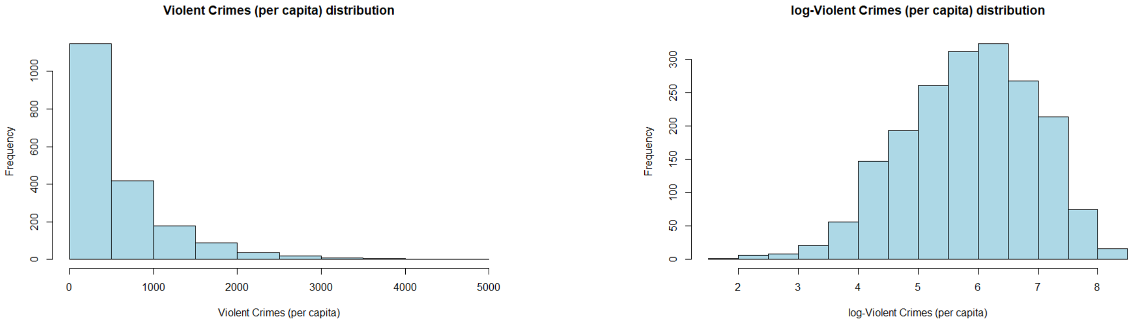


Figure 3: Distribution of the outcome variable before and after performing a log transformation

2.2.2 Regional distribution

To gain further insight into the patterns of violent crime rates in the US, it's informative to visualize the geographical distribution of these rates. This approach allows us to identify areas in which violent crime is concentrated, as well as regions characterized by lower levels of such incidents. By examining the geographic distribution, we can uncover spatial patterns and potentially identify factors contributing to the variations in violent crime rates across different states. This visual exploration will enable us to develop a comprehensive understanding of the spatial dynamics of violent crime, which could aid in the formulation of targeted strategies for prevention and intervention efforts.

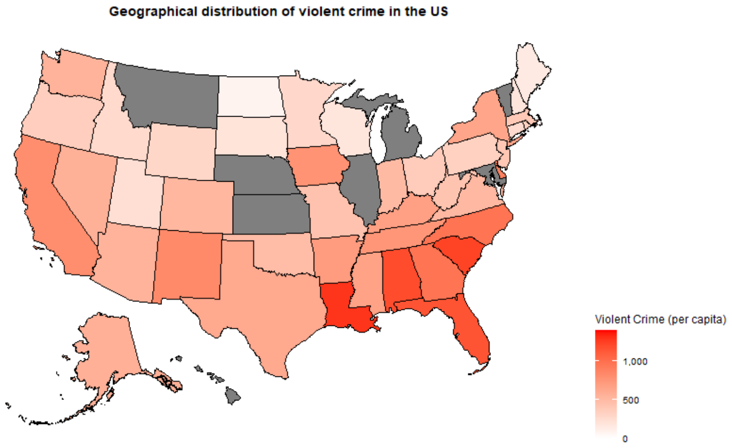

Geographical distribution of violent crime in the US
Violent Crime (per capita)
1,000
500
0

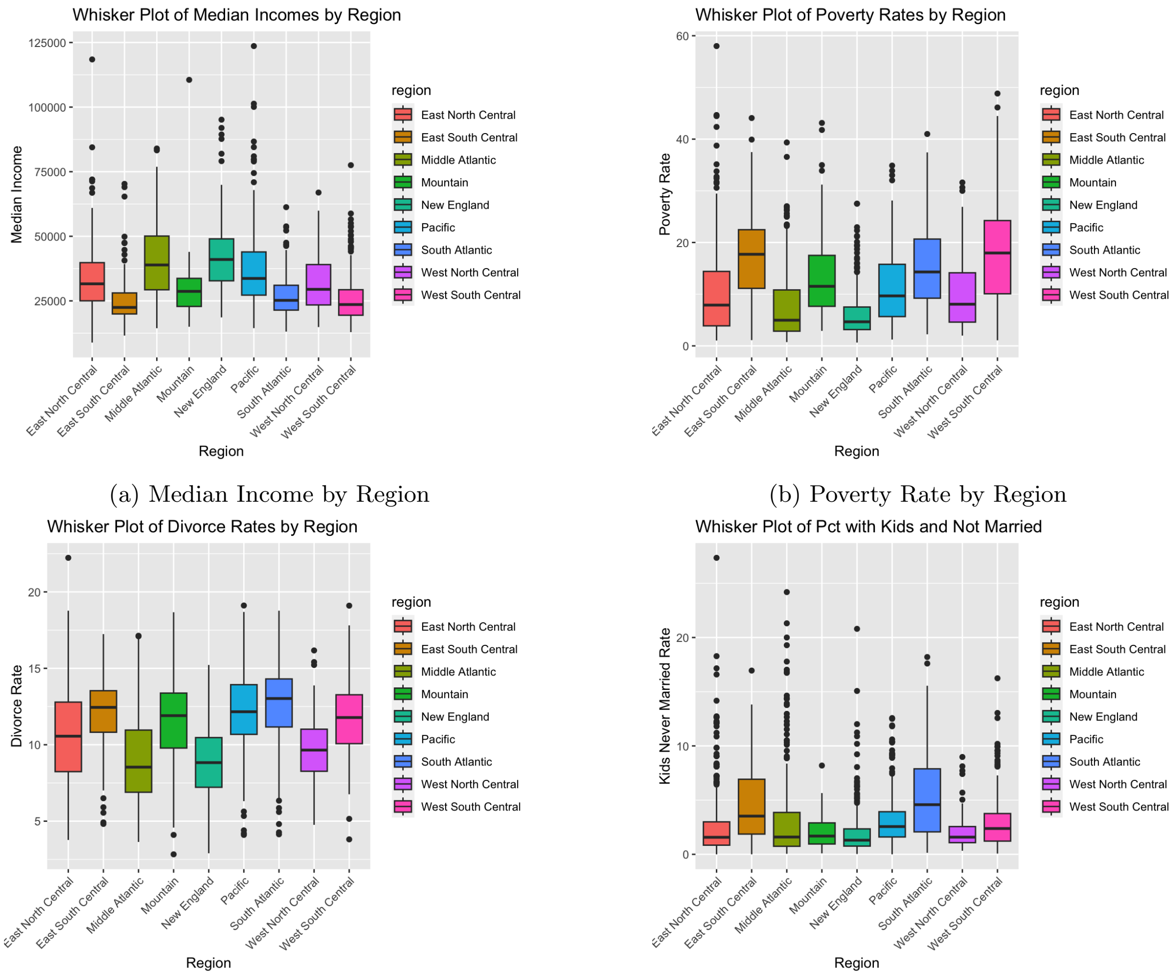
Figure 4: Geographical Distribution of Violent Crimes (per capita)

From the map above, we can tell that the highest violent crime rates on a per capita basis occur in the South of the United States. To better understand the drivers of crime across geographies, we isolate some of the variables that were highly correlated with violent crime in our initial variable analysis, but plot them on a regional basis. In particular, we look at two income variables (median income and the percent of the population below the poverty line) and two family-level variables (divorce rates and percent of people with kids who're not married) to try and further tease out the two main axes that appear to contribute to violent crime rates. To group the data, we classify each community in the data based on their Census region and division classification and then plot the within-region distribution.[1]

Once we group the data by region, a few additional patterns emerge. We can see that the East South Central, South Atlantic, and West South Central regions generally have similar levels of within-region median income (at a level around 25 thousand USD per year), and poverty rates are similarly high for all three regions, though the South Atlantic performs modestly better than the other two regions. Divorce rates appear far more homogeneous across regions, though there are regions such as New England and the Middle Atlantic where divorce rates are noticeably lower. Lastly, the East South Central and South Atlantic regions have higher rates of unmarried parents than the other regions.

Overall, it's difficult to immediately determine whether the income-level or family-level variables will be strong predictors of violent crime rates or the impact of interactions between these two categories of variables. This suggests a rich opportunity for more detailed regression analysis to build on the preliminary spacial analysis and data visualization we've developed so far.

[1] The breakdown of the regions are as follows. Pacific (California, Oregon, Washington), Mountain (Montana, Idaho, Wyoming, Nevada, Utah, Colorado, Arizona, New Mexico), West North Central (North Dakota, South Dakota, Minnesota, Iowa, Nebraska, Kansas, Missouri), West South Central (Texas, Oklahoma, Arkansas, Louisiana), East North Central (Wisconsin, Illinois, Indiana, Ohio, Michigan), East South Central (Kentucky, Tennessee, Mississippi, Alabama), South Atlantic (Florida, Georgia, South Carolina, North Carolina, Virginia, West Virginia, DC, Maryland, Delaware), Middle Atlantic (New York, Pennsylvania, New Jersey), and New England (Connecticut, Rhode Island, Massachusetts, New Hampshire, Vermont, Maine).



(a) Median Income by Region

(b) Poverty Rate by Region

(c) Divorce Rate by Region

(d) Children Not Married by Region

Figure 5: Distribution of Income and Family-Level metrics by region

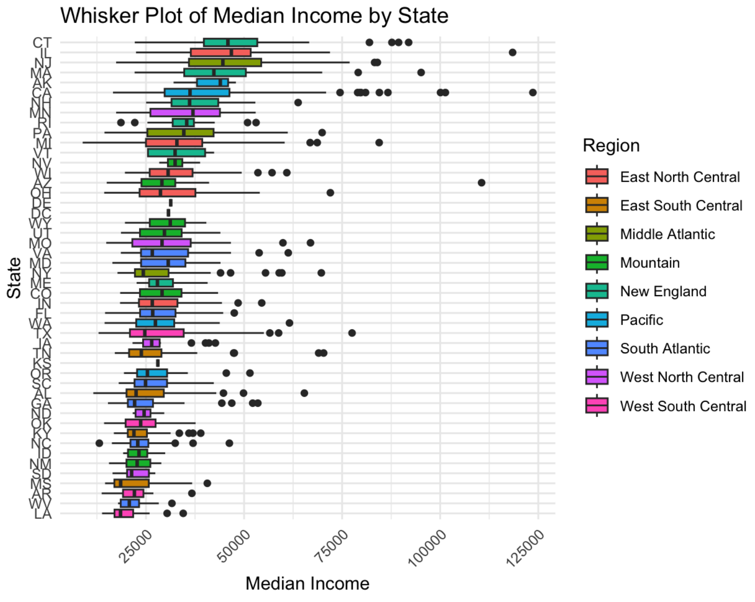


Figure 6: Median Income Distribution by State and Region

3 Regression Analysis

3.1 Linear Regression

Now that we have a cleaned data set as described section 0, we can leverage the data to gain an understanding on how the variables can be used in generalized linear modeling. First, we take every variable in the dataset and run a simple multivariate linear regression to create an initial prediction model for ViolentCrimesPerPop:

R-Squared Value
0.997971

While the model has a high R-Squared value, from the Residuals vs Fitted plot, the horizontal line shows that there is little systematic pattern between the residuals and the fitted points. Moreover, from the concentration of points on the left side of the graph, there is a chance of overestimation where the observed values will be lower than the predicted values. Looking at the Residuals vs Leverage plot, it appears that there are some high-leverage points that seem to be influencing the model significantly.

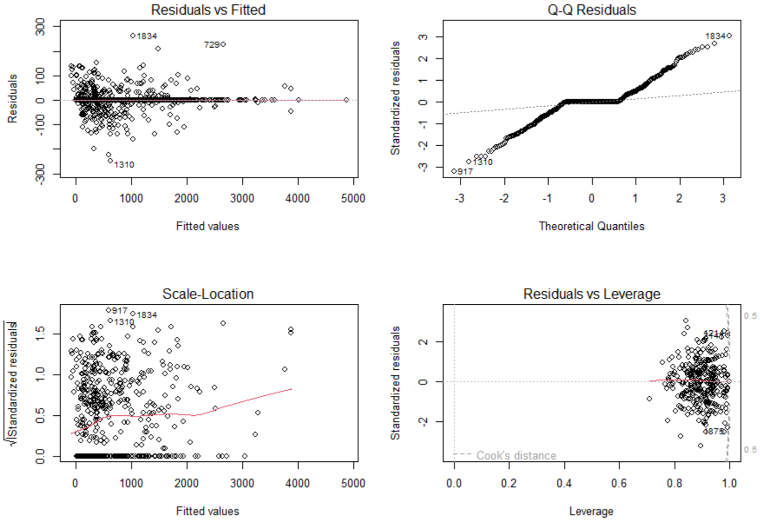


Figure 7: Linear Regression of ViolentCrimesPerPop & all variables in crime data set

As a next step, we're going to remove the constituent violent crime variables that make up ViolentCrimesPerPop and see if it has an impact on the total ViolentCrimesPerPop model. In this case, we are removing 7 columns:

Violent Crime Columns
robbbPerPop
murdPerPop
assaultPerPop
nonViolPerPop
rapesPerPop
autoTheftPerPop
nonViolPerPop

Table 1: Violent Crime Columns in Data Set

Once removed, a generalized linear model is created with a slightly lower but still high R-squared value:

R-Squared Value
0.9956608

We also produce similar residual plots that are slightly less leveraged and with a slightly lower means of overestimation.

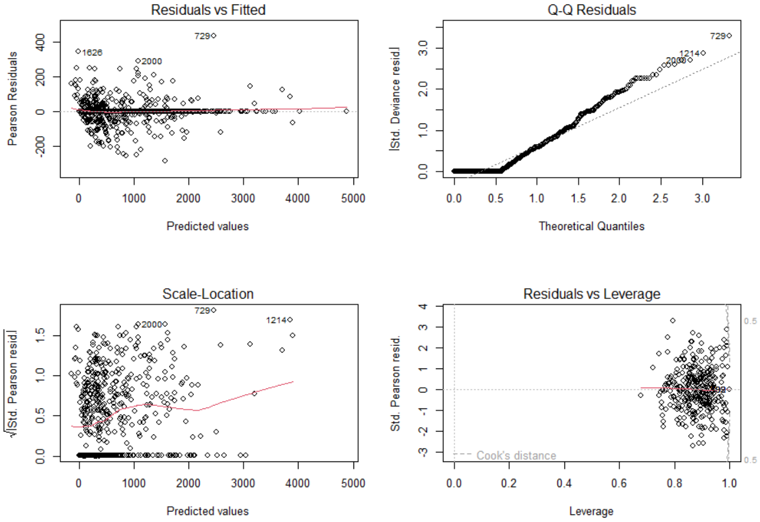
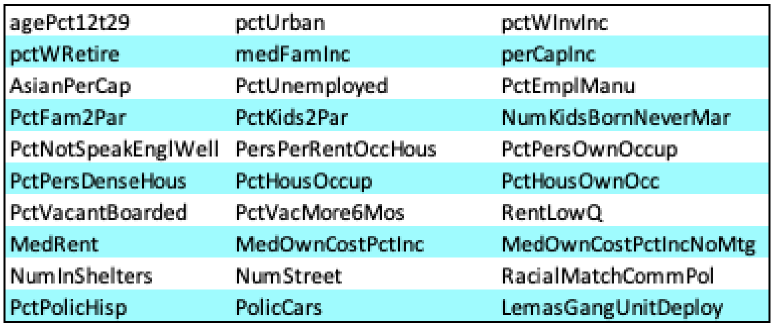


Figure 8: Linear Regression of ViolentCimesPerPop & without other violent crime statistics in crime data set

3.2 FDR Analysis

To achieve further dimension reduction, we use FDR analysis to isolate the p-values from the simple multivariate linear regression of all the community-level variables on ViolentCrimesPerPop to try to get a sense of variables that are highly related and possibly predictive of violent crime rates. Below, we list the thirty variables from the model that are statistically significant at a 95 percent confidence level.



agePct12t29	pctUrban	pctWInvInc
pctWRetire	medFamInc	perCapInc
AsianPerCap	PctUnemployed	PctEmplManu
PctFam2Par	PctKids2Par	NumKidsBornNeverMar
PctNotSpeakEnglWell	PersPerRentOccHous	PctPersOwnOccup
PctPersDenseHous	PctHousOccup	PctHousOwnOcc
PctVacantBoarded	PctVacMore6Mos	RentLowQ
MedRent	MedOwnCostPctInc	MedOwnCostPctIncNoMtg
NumInShelters	NumStreet	RacialMatchCommPol
PctPolicHisp	PolicCars	LemasGangUnitDeploy

Figure 9: Statistically Significant Variables At 95 Percent Confidence

We then use an FDR procedure with a 10 percent cutoff threshold and plot the ordered p-values from the multivariate regression to visualize our cutoff region. Given we have 123 independent variables, our new cutoff is .014. There are now only 18 variables that are statistically significant at this level, which we also list below. It looks like many income variables, such as the unemployment rate and median rent, are statistically significant, but so are things such as the number of streets in a county (a proxy for density) and the percent of the population between 12 and 29 (a proxy for age).

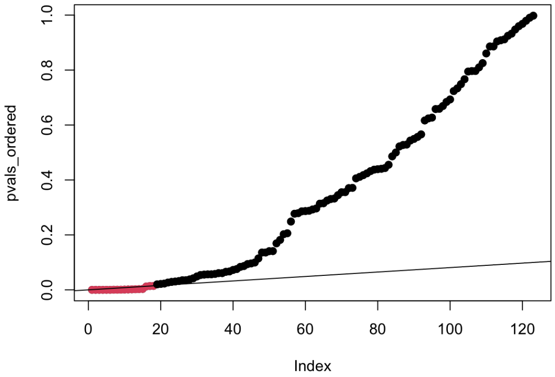
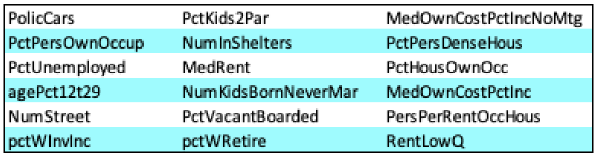


Figure 10: Ordered P-Values and FDR Cutoff Region from Multivariate Regression

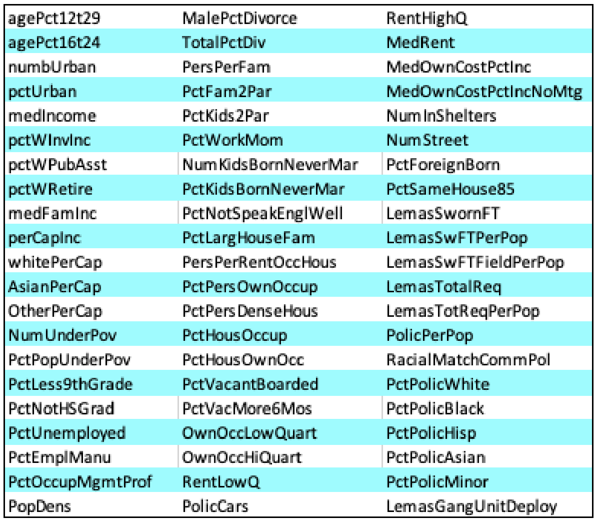


PolicCars	PctKids2Par	MedOwnCostPctIncNoMtg
PctPersOwnOccup	NumInShelters	PctPersDenseHous
PctUnemployed	MedRent	PctHousOwnOcc
agePct12t29	NumKidsBornNeverMar	MedOwnCostPctInc
NumStreet	PctVacantBoarded	PersPerRentOccHous
pctWInvInc	pctWRetire	RentLowQ

Figure 11: Statistically Significant Variables After 10 Percent FDR

3.3 Step-wise Linear Regression

As a next step in variable selection, we run a backward stepwise regression on the full cleaned violent crime dataset. The model selects the 64 variables listed below. The AICc of the backward stepwise regression is 31166. The adjusted R2 is 0.633, compared to around .99 in the simple linear model (see section 3.1). While the backward stepwise procedure likely helped reduce the level of multicollinearity in our model, this probably still isn't the most parsimonious we can achieve using simple linear regression methods.



agePct12t29	MalePctDivorce	RentHighQ
agePct16t24	TotalPctDiv	MedRent
numbUrban	PersPerFam	MedOwnCostPctInc
pctUrban	PctFam2Par	MedOwnCostPctIncNoMtg
medIncome	PctKids2Par	NumInShelters
pctWInvInc	PctWorkMom	NumStreet
pctWPubAsst	NumKidsBornNeverMar	PctForeignBorn
pctWRetire	PctKidsBornNeverMar	PctSameHouse85
medFamInc	PctNotSpeakEnglWell	LemasSwornFT
perCapInc	PctLargHouseFam	LemasSwFTPerPop
whitePerCap	PersPerRentOccHous	LemasSwFTFieldPerPop
AsianPerCap	PctPersOwnOccup	LemasTotalReq
OtherPerCap	PctPersDenseHous	LemasTotReqPerPop
NumUnderPov	PctHousOccup	PolicPerPop
PctPopUnderPov	PctHousOwnOcc	RacialMatchCommPol
PctLess9thGrade	PctVacantBoarded	PctPolicWhite
PctNotHSGrad	PctVacMore6Mos	PctPolicBlack
PctUnemployed	OwnOccLowQuart	PctPolicHisp
PctEmplManu	OwnOccHiQuart	PctPolicAsian
PctOccupMgmtProf	RentLowQ	PctPolicMinor
PopDens	PolicCars	LemasGangUnitDeploy

Figure 12: Statistically Significant Variables in Backward Stepwise Regression

3.4 LASSO

Following our simple linear regression, FDR analysis, and backward stepwise regression, we use a lasso to try and better identify the variables that are closely related violent crime rates. The regularization path of the lasso regression is plotted below. The lasso procedure isolates 69 significant variables, and the minimum AICc is 25313, which compares favorably to the AICc found using the backward stepwise procedure. The deviance of the AICc minimum slice from the lasso regression is .614.

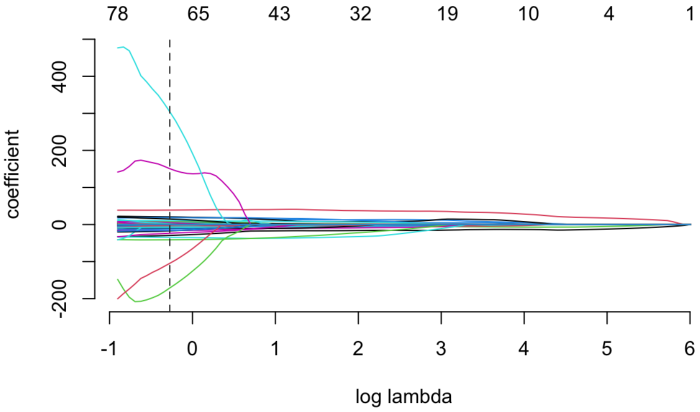
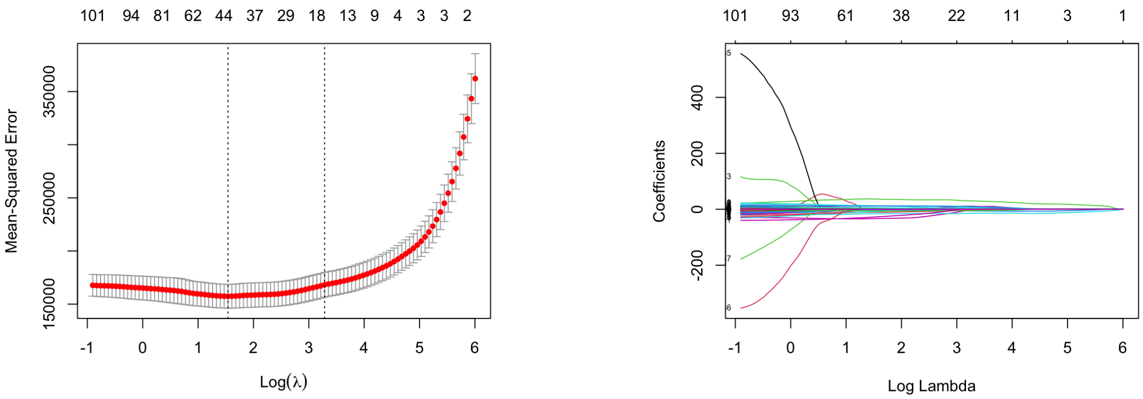


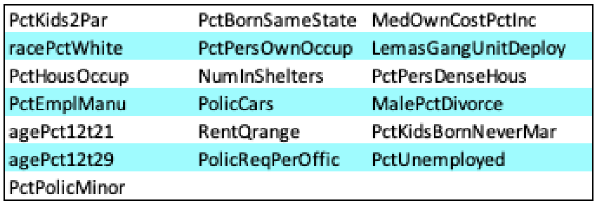
Figure 13: Plot of Lasso Regularization Path for Violent Crime model

We also perform a cross-validated lasso to see if this can possibly help in our effort to isolate a narrower set of prediction variables from our high-dimension dataset. The lasso path of the cross-validated regression is plotted below, as well as the MSE plot. The cross-validated lasso helps considerably in our efforts at dimension reduction, as there are now 19 variables selected compared to 64 in the initial lasso model. The deviance of cross-validated lasso is .3412, which is a considerable improvement from the naive lasso model.



(a) CV Lasso MSE

(b) CV Lasso Regularization Path



PctKids2Par	PctBornSameState	MedOwnCostPctInc
racePctWhite	PctPersOwnOccup	LemasGangUnitDeploy
PctHousOccup	NumInShelters	PctPersDenseHous
PctEmplManu	PolicCars	MalePctDivorce
agePct12t21	RentQrange	PctKidsBornNeverMar
agePct12t29	PolicReqPerOffic	PctUnemployed
PctPolicMinor		

Figure 15: Variables Selected by the CV Lasso

From the list of the selected variables above, we can see that both the income and family-level axes appear (for example, as PctUnemployed and PctKidsBornNeverMarried). But we find it interesting that age (agPct12t29), housing (PctHousOccup), and immigration (PctBornSameState) are also present as well.

3.5 Decision Tree

In this section, we will discuss the application of a decision tree model in our analysis. Decision trees are powerful tools in machine learning and predictive modeling, allowing us to make decisions or predictions based on a sequence of rules or conditions.

To determine the optimal structure of the decision tree, we conducted cross-validation and examined the CV Error(deviance) for different sizes of the tree. The CV Error is a measure of the model's performance on unseen data, and it helps us choose the appropriate complexity level for the tree. The results of the CV Error analysis, illustrated in Figure 16, reveal that the decision tree with six leaf nodes performs the best among the tested tree sizes. The CV Error decreases as the tree grows, indicating improved predictive accuracy, but it eventually plateaus or starts to increase when the tree becomes too complex. The tree with six leaf nodes strikes a balance between model complexity and performance, making it the optimal choice for our analysis.

Figure 17 displays the structure of the selected decision tree. The tree consists of a series of nodes and branches, where each node represents a split based on a specific predictor variable. The tree progressively divides the data into smaller subsets until reaching the terminal nodes, also known as leaf nodes, which provide the final predictions. The structure of the tree reflects the relationships between the predictor variables and the target variable, capturing the key decision-making rules learned from the data. The decision tree model demonstrates out-of-sample (OOS) R^2 value of 51% . These metrics provide an assessment of how well the model captures the variation in the data, with higher values indicating better predictive ability.

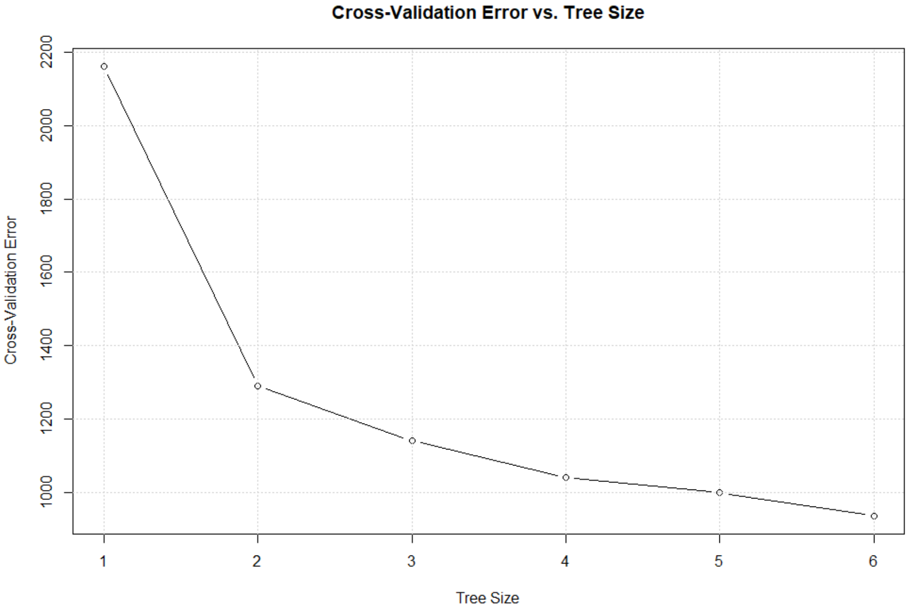


Figure 16: CV-Error for different sizes of the Decision Tree

Based on the decision tree analysis, several conclusions can be drawn. The root node of the tree is 'PctKids2Par', indicating its significance in predicting violent crime rates. It is observed that as the percentage of children with two parents decreases, the likelihood of higher crime rates in a community increases. Additionally, the variable 'racePctWhite' is selected as another important factor by the model. A lower value of 'racePctWhite' suggests a higher predicted rate of violent crime. A similar variable to the first one mentioned, 'PctFam2Par', also demonstrates importance in the analysis. When the percentage of families with two parents is low, there is an associated increase in the predicted rate of violent crime.

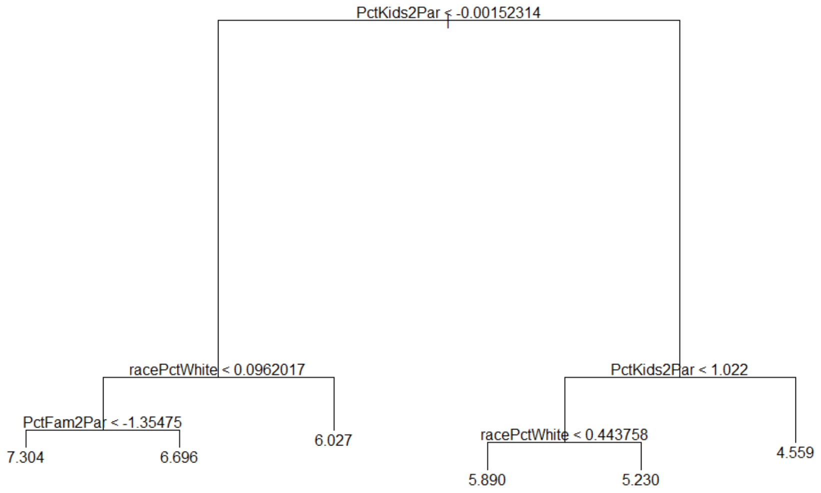


Figure 17: Decision Tree

3.6 Random Forest

While Decision Trees provide insight into the importance of variables, it is crucial to train a Random Forest rather than relying solely on a single Decision Tree, due to the common occurrence of overfitting in the latter. A Random Forest is an ensemble learning technique comprising multiple decision trees, thereby reducing the risk of overfitting through the combined effect of model construction and data selection for each individual tree. As a result, the Random Forest can be regarded as a low variance model that is less susceptible to overfitting. The downside is that as it is a more complex model, we will lose interpretabiliy of the model compared to a regular Decision Tree.

To assess the importance of different variables in our random forest model, we examine Figure 18, which displays variable importance based on node purity. Node purity is an information theory concept that measures the amount of information gained by imposing decision rules on specific variables. A higher value on the x-axis indicates greater importance of the corresponding variable for the final prediction. The variable importance plot provides insights into which predictors contribute most significantly to the model's performance.

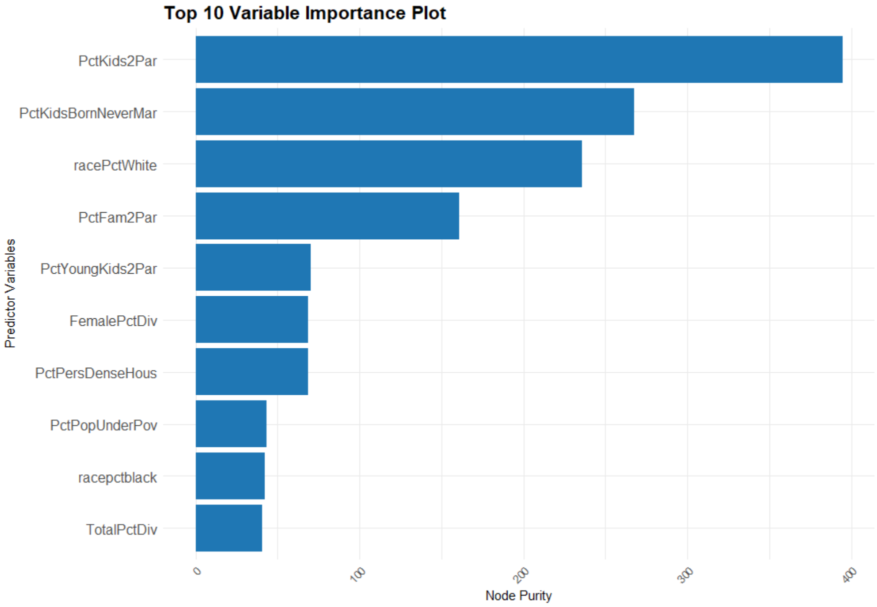


Figure 18: Variable Importance in Random Forest

The random forest model demonstrates favorable predictive performance when applied to predict Violent Crime rates for communities in our test data, which the model has not seen during training. The out-of-sample (OOS) R^2 value is 63%, indicating that the model captures a substantial portion of the variation in the target variable. Additionally, the in-sample (IS) R^2 value is 95%, suggesting that the model fits the training data well. To evaluate the model's predictive accuracy, we consider the mean squared error (MSE) for both the test data and the training data. The MSE for the test data is 0.38, while the MSE for the training data is 0.06. These values provide a measure of the average squared difference between the predicted and actual Violent Crime rates, with lower values indicating better predictive performance. The relatively low MSE values for both the test and training data suggest that the random forest model achieves a satisfactory level of accuracy in its predictions.

Figure 19 illustrates the model's performance by plotting the predicted Violent Crime rates against the actual rates for the test data. The plot demonstrates that the model has learned meaningful patterns from the ensemble of decision trees, which allows it to generalize well to unseen data. The scatter of the points around the diagonal line suggests that the model's predictions align closely with the true values, further indicating its efficacy.

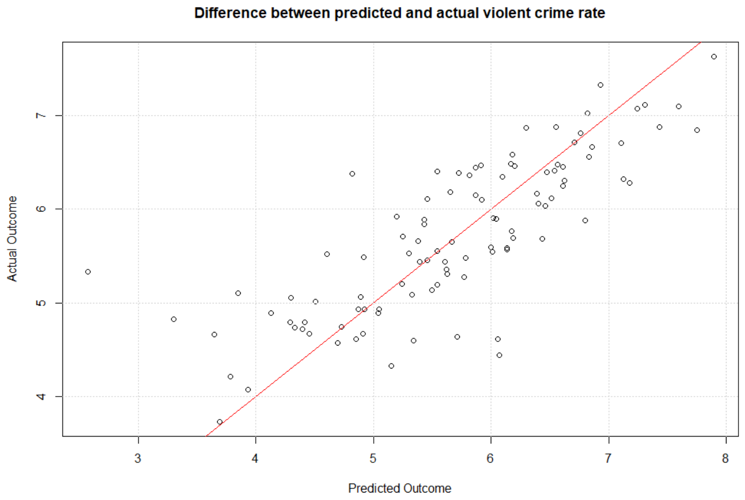


Figure 19: Predicted vs Actual violent crime rates using Random Forest

4 Unsupervised Analysis

In this section, we are trying to identify patterns within the data that will tell us information about the variables, by identifying different groups of communities, without imposing any classification rule or labels. Our goal here will be to perform cluster analysis with K-Means and study the different clusters along with the principal components obtained from PCA. We want to identify what variables are most important, or reproduce most of the data variance, and see if we the model identifies different groups based on the violent crime rates.

4.1 Principal Component Analysis (PCA)

In this section, we will perform Principal Component Analysis (PCA) on the scaled data set. The primary aim is to determine the variables that have the greatest impact on the overall variance by analyzing the loadings of the first two principal components. Following that, in the subsequent section, we will employ the PCA outcomes to visualize clustering, thereby offering valuable insights into the interpretation and significance of the initial principal components within the context of the actual data.

Upon analyzing of Figure 20, we observe that the vast majority of the variance in the data set can be accounted for by utilizing four or five principal components.

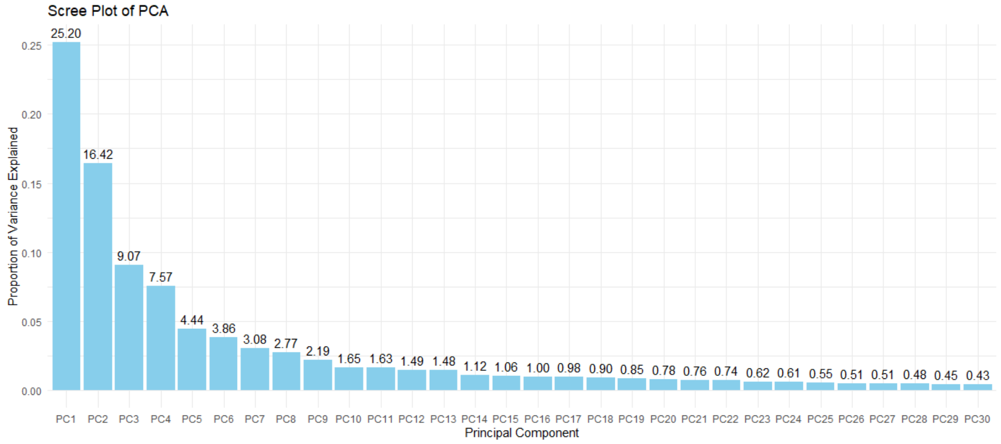


Figure 20: Screenplot of PCA

We can analyze the loadings of the principal components to have a clearer understanding of the components. The charts in Figure 21 provide histograms of the loadings for the first and second principal components, respectively. The loadings represent the meaningful relations between the original variables and the principal components. Peaks on both ends of the histograms indicate variables that contribute significantly to the respective principal component. We see that most of the loadings are around 0 suggesting that they are not very relevant for the principal component.

From these findings, we can draw conclusions about the interpretative meaning of the principal components. PC1 appears to capture the contrast between unstructured families and non-unstructured families, while PC2 reflects the differentiation between individuals born in the US and those who are immigrants. These principal components provide valuable insights into the underlying structure of the data and the factors driving the variation within it.

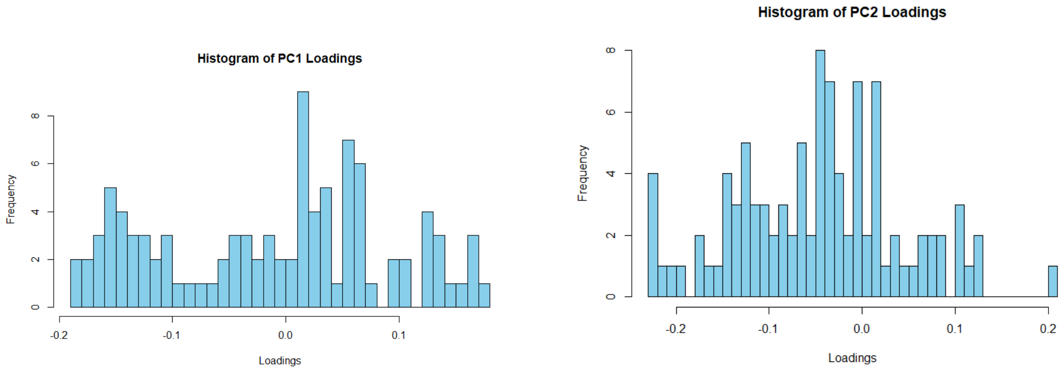


Figure 21: PC1 and PC2 loadings

Highest loadings	Lowest loadings
PctPopUnderPov	medFamInc
PctHousNoPhone	medIncome
PctNotHSGrad	pctWInvInc
pctWPubAsst	PctKids2Par
PctUnemployed	PctFam2Par
PctLess9thGrade	PctYoungKids2Par
PctOccupManu	perCapInc
PctKidsBornNeverMar	MedRent
TotalPctDiv	RentHighQ
PctHousLess3BR	PctOccupMgmtProf

Table 2: Important loadings in PC1

Highest loadings	Lowest loadings
PctSpeakEnglOnly	PctRecImmig10
racePctWhite	PctRecImmig8
PctBornSameState	PctRecImmig5
pctWSocSec	PctForeignBorn
pctWRetire	PctRecentImmig
agePct65up	PctNotSpeakEnglWell
PctWorkMom	PctPersDenseHous
PctVacMore6Mos	racePctHisp
PctPersOwnOccup	PctLargHouseFam
PctWorkMomYoungKids	PctLargHouseOccup

Table 3: Important loadings in PC2

4.2 Cluster Analysis

4.2.1 K-Means

In this section, we present the results of our cluster analysis using the K-Means algorithm on the scaled variables. The goal of this analysis is to identify distinct groups or clusters within our dataset and study different properties about such clusters.

We begin by examining the total within-cluster sum of squares of different K-Means executions to determine the appropriate number of clusters. The total within-cluster sum of squares (WSS) is a measure that quantifies the compactness of the clusters. It is defined as the sum of the squared Euclidean distances between each data point and its cluster center. Mathematically, the WSS can be expressed as follows:

$$WSS = \sum_{i=1}^{K} \sum_{x \in C_i} |x - \mu_i|^2 \tag{1}$$

where K is the number of clusters, C_i represents the ith cluster, x denotes a data point, and μ_i is the centroid of cluster C_i.

To determine the optimal number of clusters, we plot the ”elbow” curve (Figure 22) which shows the relationship between the number of clusters and the WSS. The plot reveals that beyond $k = 4$ or $k = 5$, the improvement in clustering performance becomes less significant. Therefore, we select $k = 4$ as it provides a reasonable balance between capturing meaningful variation in the data and avoiding overfitting.

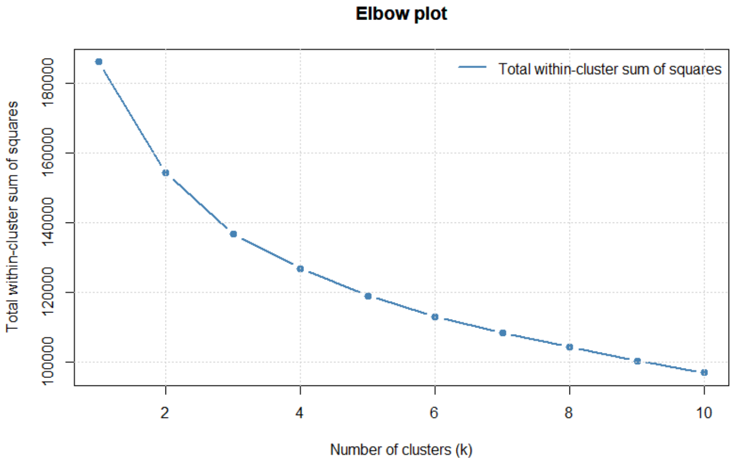


Figure 22: ”Elbow” plot for K-Means clustering

Next, we employ the K-Means algorithm with $k = 4$ to create the clusters. In order to visualize the clusters, we utilize the first principal component (PC1) from the principal component analysis (PCA) performed in the previous section. Figure 23 displays the 2D plot of the clusters using PC1 and PC2. We can observe that the clusters are well-separated and distinct from each other, indicating the effectiveness of the K-Means algorithm in grouping similar data points together projected on the components representing a higher proportion of variance.

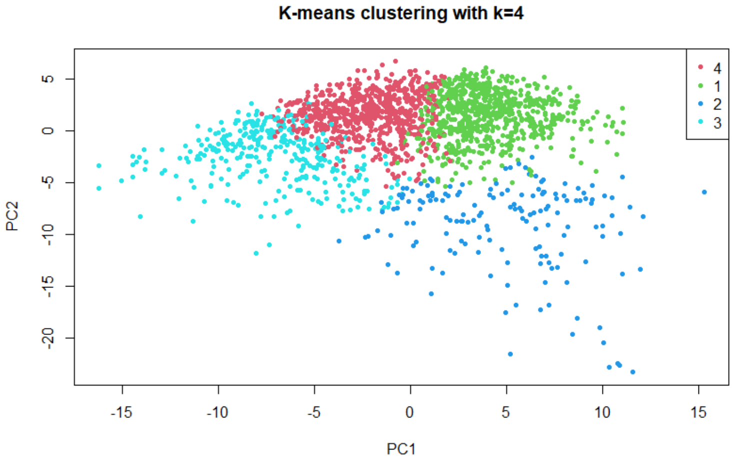


Figure 23: K-Mean clustering visualization on PC1 and PC2

To gain further insights into the characteristics of the clusters, we examine the distribution of the outcome variable, specifically the violent crime statistics, within each cluster. Figure 24 illustrates the variation in violent crime rates across the clusters. We find that clusters 1 and 4 exhibit low violent crime rates, while clusters 2 and 3 have high violent crime rates.

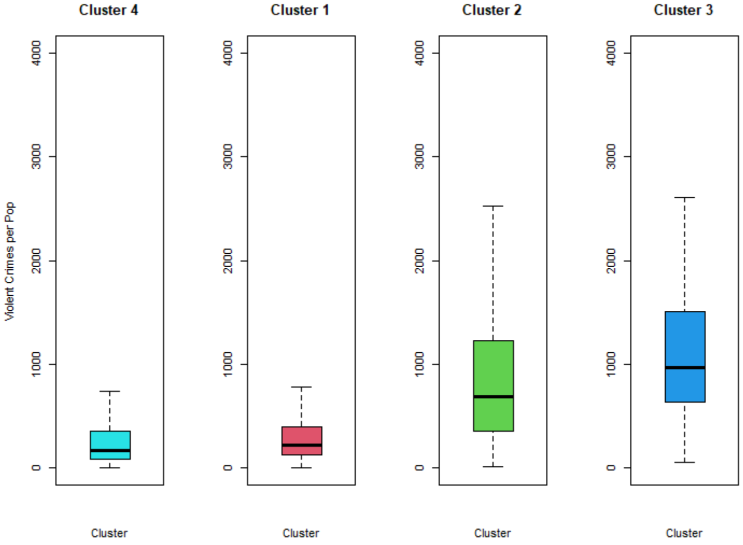


Figure 24: Violent crime statistics within each cluster

The cluster model we have constructed can be employed for classification purposes. Given a new community or set of communities, we can assign them to one of the four clusters to predict whether they are more likely to have a low or high violent crime rate. This clustering approach provides a valuable tool for identifying similar communities based on their characteristics, aiding in crime prevention and policy-making efforts.

5 Conclusion

The prevalence of violent crime is an important social issue whose drivers have been widely studied by social scientists and debated by politicians. Our study aims to utilize a combination of more simple data visualization and advanced data analysis techniques to contribute to a more nuanced understanding of what contributes to higher levels of violent crime within certain communities in the US.

Debates over the determinants of violent crime are often boiled down to overly simplistic, deterministic economic explanations or framed based on politically motivated narratives. After a deep dive into the data, we find that the drivers are far more nuanced and often context specific, based on the specific combination of economic, social, and demographic characteristics of particular communities.

5.1 Results

5.1.1 What factors drive violent crime?

- Our initial data visualization pointed to two main axes of variables that contribute to violent crime–economic and family-based. Specifically, as one might expect, the median income, level of poverty, and level of unemployment in a given community is highly correlated with violent crime. But family formation, including whether children are born to married parents and the incidence of divorce, is also critical too.
- Using a combination of linear regression and lasso regression techniques, we identified other axes of differentiation across high and low violent crime communities. In particular, our cross-validated lasso identified age (demographics), housing, and the level of immigration as other statistically relevant determinants of violent crime rates.
- Using a combination of supervised and unsupervised techniques, we were able to isolate additional variables of importance. Our decision tree analysts identified PctFam2Par and racePctWhite as the two critical nodes for predicting violent crime rates. This reinforces the importance of family-level characteristics and the possible impact of race on community level violence prevalence, though it of course doesn't imply causation between these variables and violence.
- The loadings of our PCA analysis demonstrate that there indeed appears to be a highly important distinction between strcutured and unstrcutured families when it comes to the incidence of violent crime and that the level of immigration to a community is also important.

5.1.2 What models do the best in terms of predicting violent crime?

- In our analysis of various regression models, we found the random forest model to exhibit the most robust and reliable predictive performance when it came to violent crime rates. This model effectively addressed overfitting issues and provided a thorough assessment of variable importance. Additionally, it demonstrated a great generalization capability in predicting violent crime rates for unseen data.
- On the other hand, although the linear regression models displayed high R-squared values, they were prone to potential overfitting and the oversimplification of complex and possibly non-linear data. Despite playing a significant role in enhancing our understanding of the data and the factors driving violent crime rates, these models proved less reliable for accurately predicting violent crimes.

5.1.3 How can we categorize the factors that contribute to violent crime?

- It appears that there are multiple axes of variables that predict violent crime rates, including economic conditions, family-level factors, demographics, and the prevalence of immigration. This belies a simple economic, deterministic explanation of violent crime rates and suggests that there exists more nuanced explanations based on the specific factors of a given community.

- That said, the results of our PCA and K-Means analysis also suggests that some combination of these four factors can go a long way to explaining the majority of violent crime in most communities. This could be helpful in terms of thinking through more targeted policies to reduce the incidence of violent crime in the future.

6 References

1. UCI Communities and Crime Data Set: https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

7 Appendix

7.1 Description of Attributes from Data Set

- **state**:US state (by number) - not counted as predictive above, but if considered, should be consided nominal (nominal)
- **county**:numeric code for county - not predictive, and many missing values (numeric)
- **community**:numeric code for community - not predictive and many missing values (numeric)
- **communityname**:community name - not predictive - for information only (string)
- **fold**:fold number for non-random 10 fold cross validation, potentially useful for debugging, paired tests - not predictive (numeric)
- **population**:population for community
- **householdsize**:mean people per household (numeric - decimal)
- **racepctblack**:percentage of population that is african american (numeric - decimal)
- **racePctWhite**:percentage of population that is caucasian (numeric - decimal)
- **racePctAsian**:percentage of population that is of asian heritage (numeric - decimal)
- **racePctHisp**:percentage of population that is of hispanic heritage (numeric - decimal)
- **agePct12t21**:percentage of population that is 12-21 in age (numeric - decimal)
- **agePct12t29**:percentage of population that is 12-29 in age (numeric - decimal)
- **agePct16t24**:percentage of population that is 16-24 in age (numeric - decimal)
- **agePct65up**:percentage of population that is 65 and over in age (numeric - decimal)
- **numbUrban**:number of people living in areas classified as urban (numeric - decimal)
- **pctUrban**:percentage of people living in areas classified as urban (numeric - decimal)
- **medIncome**:median household income (numeric - decimal)
- **pctWWage**:percentage of households with wage or salary income in 1989 (numeric - decimal)
- **pctWFarmSelf**:percentage of households with farm or self employment income in 1989 (numeric - decimal)
- **pctWInvInc**:percentage of households with investment / rent income in 1989 (numeric - decimal)
- **pctWSocSec**:percentage of households with social security income in 1989 (numeric - decimal)
- **pctWPubAsst**:percentage of households with public assistance income in 1989 (numeric - decimal)
- **pctWRetire**:percentage of households with retirement income in 1989 (numeric - decimal)
- **medFamInc**:median family income (differs from household income for non-family households) (numeric - decimal)
- **perCapInc**:per capita income (numeric - decimal)
- **whitePerCap**:per capita income for caucasians (numeric - decimal)
- **blackPerCap**:per capita income for african americans (numeric - decimal)
- **indianPerCap**:per capita income for native americans (numeric - decimal)

- **AsianPerCap**:per capita income for people with asian heritage (numeric - decimal)
- **OtherPerCap**:per capita income for people with 'other' heritage (numeric - decimal)
- **HispPerCap**:per capita income for people with hispanic heritage (numeric - decimal)
- **NumUnderPov**:number of people under the poverty level (numeric - decimal)
- **PctPopUnderPov**:percentage of people under the poverty level (numeric - decimal)
- **PctLess9thGrade**:percentage of people 25 and over with less than a 9th grade education (numeric - decimal)
- **PctNotHSGrad**:percentage of people 25 and over that are not high school graduates (numeric - decimal)
- **PctBSorMore**:percentage of people 25 and over with a bachelors degree or higher education (numeric - decimal)
- **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)
- **PctEmplManu**:percentage of people 16 and over who are employed in manufacturing (numeric - decimal)
- **PctEmplProfServ**:percentage of people 16 and over who are employed in professional services (numeric - decimal)
- **PctOccupManu**:percentage of people 16 and over who are employed in manufacturing (numeric - decimal)
- **PctOccupMgmtProf**:percentage of people 16 and over who are employed in management or professional occupations (numeric - decimal)
- **MalePctDivorce**:percentage of males who are divorced (numeric - decimal)
- **MalePctNevMarr**:percentage of males who have never married (numeric - decimal)
- **FemalePctDiv**:percentage of females who are divorced (numeric - decimal)
- **TotalPctDiv**:percentage of population who are divorced (numeric - decimal)
- **PersPerFam**:mean number of people per family (numeric - decimal)
- **PctFam2Par**:percentage of families (with kids) that are headed by two parents (numeric - decimal)
- **PctKids2Par**:percentage of kids in family housing with two parents (numeric - decimal)
- **PctYoungKids2Par**:percent of kids 4 and under in two parent households (numeric - decimal)
- **PctTeen2Par**:percent of kids age 12-17 in two parent households (numeric - decimal)
- **PctWorkMomYoungKids**:percentage of moms of kids 6 and under in labor force (numeric - decimal)
- **PctWorkMom**:percentage of moms of kids under 18 in labor force (numeric - decimal)
- **NumIlleg**:number of kids born to never married (numeric - decimal)
- **PctIlleg**:percentage of kids born to never married (numeric - decimal)
- **NumImmig**:total number of people known to be foreign born (numeric - decimal)

- **PctImmigRecent**:percentage of immigrants who immigrated within last 3 years (numeric - decimal)
- **PctImmigRec5**:percentage of immigrants who immigrated within last 5 years (numeric - decimal)
- **PctImmigRec8**:percentage of immigrants who immigrated within last 8 years (numeric - decimal)
- **PctImmigRec10**:percentage of immigrants who immigrated within last 10 years (numeric - decimal)
- **PctRecentImmig**:percent of population who have immigrated within the last 3 years (numeric - decimal)
- **PctRecImmig5**:percent of population who have immigrated within the last 5 years (numeric - decimal)
- **PctRecImmig8**:percent of population who have immigrated within the last 8 years (numeric - decimal)
- **PctRecImmig10**:percent of population who have immigrated within the last 10 years (numeric - decimal)
- **PctSpeakEnglOnly**:percent of people who speak only English (numeric - decimal)
- **PctNotSpeakEnglWell**:percent of people who do not speak English well (numeric - decimal)
- **PctLargHouseFam**:percent of family households that are large (6 or more) (numeric - decimal)
- **PctLargHouseOccup**:percent of all occupied households that are large (6 or more people) (numeric - decimal)
- **PersPerOccupHous**:mean persons per household (numeric - decimal)
- **PersPerOwnOccHous**:mean persons per owner occupied household (numeric - decimal)
- **PersPerRentOccHous**:mean persons per rental household (numeric - decimal)
- **PctPersOwnOccup**:percent of people in owner occupied households (numeric - decimal)
- **PctPersDenseHous**:percent of persons in dense housing (more than 1 person per room) (numeric - decimal)
- **PctHousLess3BR**:percent of housing units with less than 3 bedrooms (numeric - decimal)
- **MedNumBR**:median number of bedrooms (numeric - decimal)
- **HousVacant**:number of vacant households (numeric - decimal)
- **PctHousOccup**:percent of housing occupied (numeric - decimal)
- **PctHousOwnOcc**:percent of households owner occupied (numeric - decimal)
- **PctVacantBoarded**:percent of vacant housing that is boarded up (numeric - decimal)
- **PctVacMore6Mos**:percent of vacant housing that has been vacant more than 6 months (numeric - decimal)
- **MedYrHousBuilt**:median year housing units built (numeric - decimal)
- **PctHousNoPhone**:percent of occupied housing units without phone (in 1990, this was rare!) (numeric - decimal)
- **PctWOFullPlumb**:percent of housing without complete plumbing facilities (numeric - decimal)

- **OwnOccLowQuart**:owner occupied housing - lower quartile value (numeric - decimal)
- **OwnOccMedVal**:owner occupied housing - median value (numeric - decimal)
- **OwnOccHiQuart**:owner occupied housing - upper quartile value (numeric - decimal)
- **RentLowQ**:rental housing - lower quartile rent (numeric - decimal)
- **RentMedian**:rental housing - median rent (Census variable H32B from file STF1A) (numeric - decimal)
- **RentHighQ**:rental housing - upper quartile rent (numeric - decimal)
- **MedRent**:median gross rent (Census variable H43A from file STF3A - includes utilities) (numeric - decimal)
- **MedRentPctHousInc**:median gross rent as a percentage of household income (numeric - decimal)
- **MedOwnCostPctInc**:median owners cost as a percentage of household income - for owners with a mortgage (numeric - decimal)
- **MedOwnCostPctIncNoMtg**:median owners cost as a percentage of household income - for owners without a mortgage (numeric - decimal)
- **NumInShelters**:number of people in homeless shelters (numeric - decimal)
- **NumStreet**:number of homeless people counted in the street (numeric - decimal)
- **PctForeignBorn**:percent of people foreign born (numeric - decimal)
- **PctBornSameState**:percent of people born in the same state as currently living (numeric - decimal)
- **PctSameHouse85**:percent of people living in the same house as in 1985 (5 years before) (numeric - decimal)
- **PctSameCity85**:percent of people living in the same city as in 1985 (5 years before) (numeric - decimal)
- **PctSameState85**:percent of people living in the same state as in 1985 (5 years before) (numeric - decimal)
- **LemasSwornFT**:number of sworn full time police officers (numeric - decimal)
- **LemasSwFTPerPop**:sworn full time police officers per 100K population (numeric - decimal)
- **LemasSwFTFieldOps**:number of sworn full time police officers in field operations (on the street as opposed to administrative etc) (numeric - decimal)
- **LemasSwFTFieldPerPop**:sworn full time police officers in field operations (on the street as opposed to administrative etc) per 100K population (numeric - decimal)
- **LemasTotalReq**:total requests for police (numeric - decimal)
- **LemasTotReqPerPop**:total requests for police per 100K popuation (numeric - decimal)
- **PolicReqPerOffic**:total requests for police per police officer (numeric - decimal)
- **PolicPerPop**:police officers per 100K population (numeric - decimal)
- **RacialMatchCommPol**:a measure of the racial match between the community and the police force. High values indicate proportions in community and police force are similar (numeric - decimal)
- **PctPolicWhite**:percent of police that are caucasian (numeric - decimal)

- **PctPolicBlack**:percent of police that are African American (numeric - decimal)
- **PctPolicHisp**:percent of police that are hispanic (numeric - decimal)
- **PctPolicAsian**:percent of police that are asian (numeric - decimal)
- **PctPolicMinor**:percent of police that are minority of any kind (numeric - decimal)
- **OfficAssgnDrugUnits**:number of officers assigned to special drug units (numeric - decimal)
- **NumKindsDrugsSeiz**:number of different kinds of drugs seized (numeric - decimal)
- **PolicAveOTWorked**:police average overtime worked (numeric - decimal)
- **LandArea**:land area in square miles (numeric - decimal)
- **PopDens**:population density in persons per square mile (numeric - decimal)
- **PctUsePubTrans**:percent of people using public transit for commuting (numeric - decimal)
- **PolicCars**:number of police cars (numeric - decimal)
- **PolicOperBudg**:police operating budget (numeric - decimal)
- **LemasPctPolicOnPatr**:percent of sworn full time police officers on patrol (numeric - decimal)
- **LemasGangUnitDeploy**:gang unit deployed (numeric - decimal - but really ordinal - 0 means NO, 1 means YES, 0.5 means Part Time)
- **LemasPctOfficDrugUn**:percent of officers assigned to drug units (numeric - decimal)
- **PolicBudgPerPop**:police operating budget per population (numeric - decimal)
- **ViolentCrimesPerPop**:total number of violent crimes per 100K population (numeric - decimal) GOAL attribute (to be predicted)

7.2 Variables with Strong Correlations

Variable	Correlation
robbbPerPop	0.9444806
assaultPerPop	0.9157299
PctKidsBornNeverMar	0.8881138
burglPerPop	0.8056104
racepctblack	0.8053283
pctWPubAsst	0.7854172
nonViolPerPop	0.7825537
PctUnemployed	0.7749622
murdPerPop	0.7521416
PctHousNoPhone	0.7437042
PctPopUnderPov	0.7420976
PctLargHouseFam	0.7325530
PctVacantBoarded	0.7319636
PctPolicMinor	0.7254188
PctPolicBlack	0.7125927
autoTheftPerPop	0.7040104
rapesPerPop	0.6913026
FemalePctDiv	0.6794522
TotalPctDiv	0.6762331
MalePctDivorce	0.6526236
PctLargHouseOccup	0.6441014
PctWOFullPlumb	0.6285638
assaults	0.6247085
MalePctNevMarr	0.5920976
PctPersDenseHous	0.5722021
robberies	0.5711050
burglaries	0.5696169
PctNotHSGrad	0.5664521
PersPerFam	0.5600724
arsonsPerPop	0.5177963
NumKidsBornNeverMar	0.5081366
NumInShelters	0.5079753
PctHousOwnOcc	-0.5302890
medIncome	-0.5365731
PctEmploy	-0.5395806
PctHousOccup	-0.5461528
PctPersOwnOccup	-0.5483244
medFamInc	-0.5547477
PctPolicWhite	-0.5741529
RacialMatchCommPol	-0.5819459
pctWInvInc	-0.7503054
PctYoungKids2Par	-0.8153471
racePctWhite	-0.8378503
PctTeen2Par	-0.8415796
PctFam2Par	-0.8464279
PctKids2Par	-0.8549586