

The Effect of Parental Structure on Assaults Across the US

Trinity Turner, Cate Slaven, Nash Balakrishna, Carlo Fairley

University of North Carolina at Charlotte

Dr. Nadia Najjar & Dr. Shannon Reid

Nov 5, 2024

Abstract

Our research explores the relationship between family structure and violent crime rates, with a particular focus on whether communities with higher proportions of two-parent households experience lower rates of violent crime, independent of economic conditions. Using a dataset of demographic and crime statistics from various U.S. cities, our study examines the potential predictive role of family stability on crime rates. By establishing a baseline model and comparing it to advanced regression techniques, the study aims to inform policy implications that support family stability as a preventive measure against crime.

Contents

Introduction	4
Background	5
Data Description	9
Methodology	15
Model Analysis and Results.....	24
Regression Model.....	24
Classification Models.....	25
Conclusion	35
References	37

Introduction

Every 26 seconds, a violent crime takes place, shattering lives in communities across the United States, leaving behind broken families and a lasting ripple of fear (FBI.gov). Murder, rape, robbery, and assault are not just statistics; they are realities that touch every corner of society, draining resources and eroding security. Yet, in the search for solutions, one critical question remains largely unexplored: could something as fundamental as family structure be the key to breaking this cycle of violence? By shifting our focus to the roots of crime, we may uncover a powerful yet underutilized tool for prevention—parental stability.

While poverty, unemployment, and education have been widely studied as drivers of violent crime, family structure has received far less attention. By investigating how two-parent versus single-parent households influence crime rates, this study addresses a critical gap in the literature, offering a fresh perspective on crime prevention strategies. Family stability has traditionally been associated with positive youth outcomes, while disruptions in family structure have been linked to increased risks for delinquency and crime. With the evolution of family configurations in the United States, where single-parent households have risen considerably due to factors like divorce, social norms, and economic pressures, understanding how these changes impact community safety is crucial (Demuth & Brown, 2004; Simons et al., 2004).

Understanding the factors that contribute to violent crime is essential for developing effective prevention strategies and policies. Unlike reactive crime responses, such as: over policing, increasing surveillance and escalating crime punishment, effective deterrents enable proactive resource allocation by identifying communities in need of support before adverse

outcomes arise. The FBI defines violent crime as murder and nonnegligent manslaughter, forcible rape, robbery, and aggravated assault (FBI.gov).

The primary hypothesis of our study is that communities with a higher proportion of two-parent families tend to experience lower rates of violent crime, highlighting the potential influence of family stability on community safety. This hypothesis proposes that the structure and support offered by two-parent households may serve as protective factors, reducing the likelihood of violent behaviors. Additionally, the study hypothesizes that socio-economic conditions may reorient the relationship between family structure and violent crime rates, suggesting that the influence of family stability on crime may vary depending on the economic and sociological environment of the community. By exploring these hypotheses, the study aims to determine whether family structure alone is a significant predictor of violent crime or if economic conditions also play a crucial role in shaping these outcomes. This dual approach will provide insights into the combined impact of family dynamics and economic factors on community safety, potentially informing targeted crime prevention strategies. Using data that combines socio-economic details as well as crime data, we aim to discover if parental structure possesses the capability of predicting violent crime. By employing advanced predictive models, including decision trees and Random Forests, we aim to quantify the relative impact of each predictor on assault rates across different states based on the question of: Can the assault rates be anticipated by parental structure in varying cities across the nation?

Background

To properly give basis to the claims surrounding family structures' effect on assault rates, it is imperative to understand the role of socio-economic characteristics among family structures. Data from the U.S. Census Bureau shows that only 64% of children now live in two-parent

households which is a significant drop from the 90% in 1970. This shift may have profound implications for social cohesion and in turn, crime rates (U.S. Census Bureau, 2020). When referring to social cohesion, it can be characterized as how well integrated communities may be with their neighbors. Communities where families have necessary interaction with each other can create a sense of solidarity constructed in shared norms and beliefs.

Studies show that children from single-parent households are at greater risk for criminal behavior compared to those from two-parent families. This risk is associated with several factors, including: reduced parental supervision, economic hardship, and limited access to social support networks. These findings coincide with the properties of the Social Control Theory, suggesting that these weak parental ties allow children to explore settings that would otherwise be controlled by the network their parents might belong to, often leading them to negative outcomes (Hirschi, 1969).

The lack of dual parental involvement is often linked to inadequate supervision and fewer opportunities for positive role modeling, which social control theory attributes to a decline in community and familial bonds. Social Control Theory provides a theoretical framework for understanding how family structure influences crime at the community level. Originally proposed in the mid 1900's, the theory suggests that communities with lower social cohesion and fewer stable family structures lack the informal social controls necessary to deter crime. Sampson and Groves (1989) expanded this theory by demonstrating that communities with high levels of single-parent households experience weakened community supervision, fewer organized activities for youth, and a diminished capacity for residents to intervene in preventing crime.

With a single parent, there is often less family time due to the financial burden being on one sole provider. In early adolescence and onward young people increasingly spend more time with their peers instead of their parents, as they explore their freedoms and independence (Steinberg and Silk, 2002). This transition to adolescence alone is not a cause for concern, however, newfound freedoms and peer oriented behavior in areas with lower levels of social controls can increase the likelihood of delinquency in adolescence or later on in life (Janessen, Weerman & Eichelsheim, 2016). Empirical studies underscore the association between family structure and crime rates, though the relationship is complex and often influenced by other factors. For example, Mack et al. (2006) found that children in single-parent families are more likely to engage in delinquent behavior due to the increased autonomy and reduced monitoring associated with these family structures. These adolescents in single-parent households are also at greater risk of dropping out of school, which itself is a predictor of criminal activity later in life (Astone and McLanahan, 1991).

Demuth and Brown (2004) further emphasize that the absence of one parent often results in decreased emotional support and stability for children, which are critical for social development and compliance with social norms. Studies on single-parent households reveal a consistent correlation with various forms of youth delinquency, underlining the need to consider parental stability as a factor in crime prevention. The evidence is clear that family attachments are strongly correlated with (non)delinquency. In their famous book 'Unraveling Juvenile Delinquency', Sheldon and Eleanor Glueck(1950) found within their research that affection of the father and the mother for the child were two of the best five predictors of delinquency.

In addition to family dynamics, socioeconomic factors play a critical role in shaping family structure and, subsequently, crime rates. Children raised in single-parent households are

more likely to experience poverty, limited educational resources, and reduced community support—all of which are recognized risk factors for criminal behavior (Sampson, 1987). Economic challenges often place additional strain on single-parent families, making it difficult for parents to provide supervision, emotional support, and engagement with their children's schooling. Simons et al. (2004) observe that single mothers, in particular, may struggle to balance work and family obligations, which can reduce their ability to monitor children effectively, potentially leading to increased risk of delinquency.

In applying Social Control Theory to family structure, recent studies have investigated how shifts in family dynamics affect community safety. For instance, communities with fewer two-parent households may experience reduced collective efficacy—the shared belief in residents' ability to achieve common goals— including when it comes to crime prevention (Sampson, 1997). The loss of family cohesion can contribute to social isolation, where families and individuals have fewer meaningful interactions, leading to lower collective accountability and increased tolerance for deviant behavior (Breivik et al., 2009).

Social Disorganization Theory provides a valuable framework for understanding the influence of family structure on crime at the community level. Originally proposed by Shaw and McKay (1942), this theory posits that communities with low social cohesion and unstable family structures are more prone to crime due to weakened informal socioeconomic controls. Other studies indicate that while family structure is a significant factor, its influence on crime is mediated by socioeconomic status, with poverty and lack of education amplifying the risks associated with single-parent households (Simons et al., 2004).

Long-term monitoring and data-driven policy adjustments would ensure that interventions remain effective and responsive to community needs. Regular data collection and

analysis enable policymakers to identify successful programs, address shortcomings, and allocate resources effectively. Through sustained support for family cohesion, economic stability, and neighborhood engagement, these policy recommendations contribute to building safer, more resilient communities that collectively address the root causes of crime.

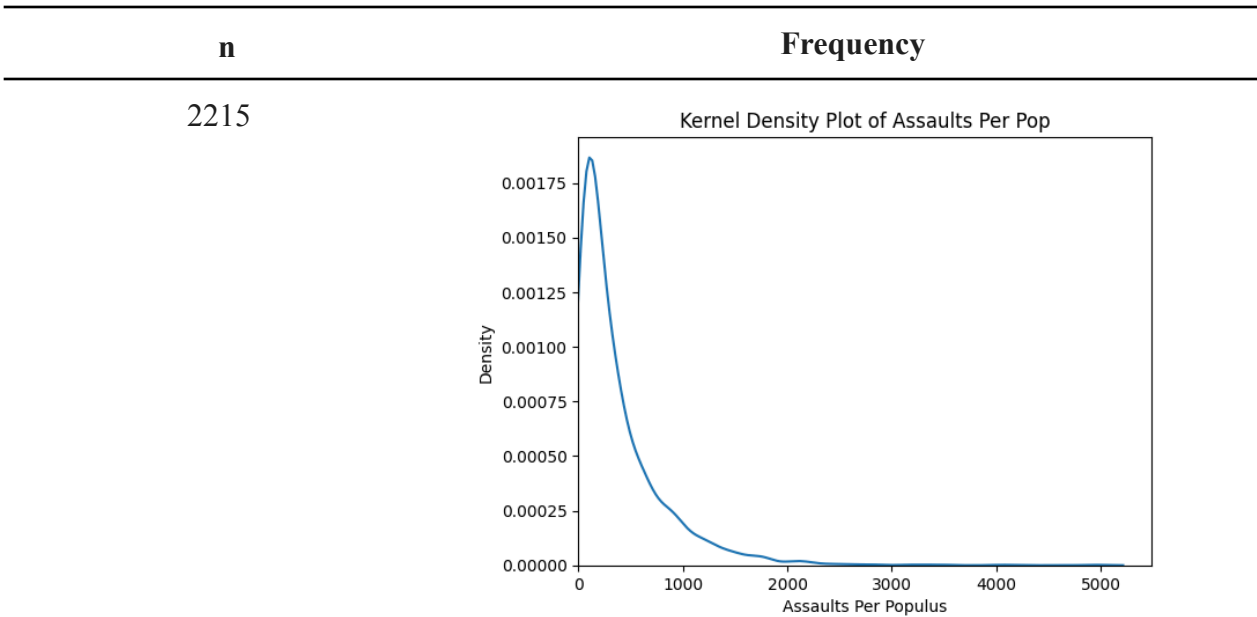
Dataset Description

To effectively investigate the relationship between parental structure and crime we used a dataset that combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. It holds 2216 observations with 147 features detailing socio-economic and personal demographics such as family composition, income, level of education, and instances of different types of crime committed in each city. Most values admitted to the survey are in percentages or measured in general measurements, such as per capita. Violent crime is the addition of all violent crimes listed in the dataset – assaults, murder, rapes and robberies. We acknowledge the official count of rapes from each state was an ongoing dispute, resulting in missing values. The cities with said controversies were omitted from the dataset, which should not affect the rest of our data. Even still, these observations are claimed to be original values. They are incomplete with missing values or non-numeric data entered, we estimated that 13% of the dataset consisted of missing values.

Since we aim to use values that are directly representative of households with two parents to observe their effect on assaults relative to the population of each city as our baseline model, we chose to use the ‘assaultsPerPop’ feature as our dependent variable. The dependent variable, ‘assaultsPerPop,’ is a measurement of persons per 100,000 people and is heavily right-skewed, with most communities reporting fewer than 500 assaults per population. Outliers like Zanesville City, Wyoming, with 4,932 assaults per population, highlight the need for careful handling of

extreme values. Below is a plot of the distribution of values within the assaults per populus in Figure 1. Within our selected features, there were a total of 481 outliers; 'PctFam2Par' with 36, 'PctKids2Par' with 31, 'PctYoungKids2Par' with 36, 'PctTeen2Par' with 58, 'PctKidsBornNeverMar' with 177, and 'assaultPerPop' with 143.

Figure 1
Distribution of Assaults

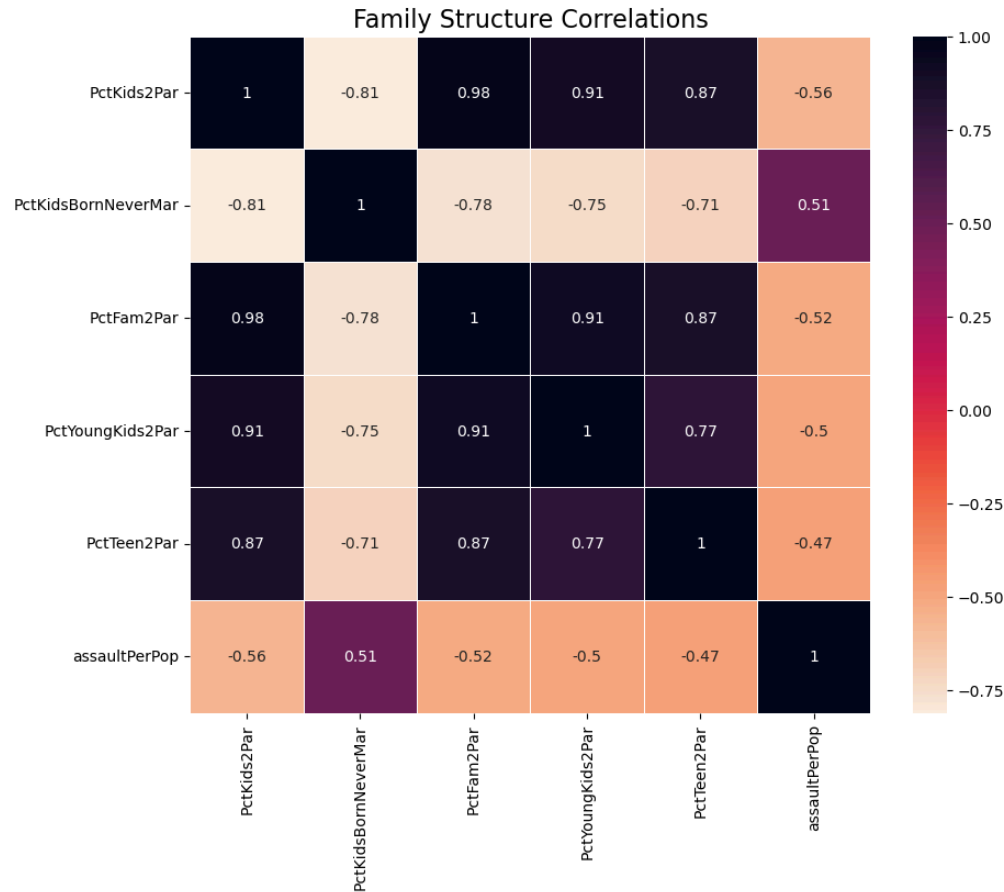


On the x-axis are the varying totals of assaults per populus, with the minimum that can be found in the dataset being 0, and the maximum value being 4932, coming from Zanesville City, Wyoming. As previously discussed, the federal bodies involved in the creation of the dataset had a disagreement that was said to have caused an unidentifiable number of false 0’s. This context is necessary to understand why there is a high frequency of zero values that skew the data. The y-axis shows an estimation of the proportion of observations around a specific amount

of assaults per populus. In order to investigate the influence of parental structure on the rate of assaults, we narrowed our dataset to concentrate on key predictors identified in the existing literature as significant contributors. In order to prepare our dataset for analysis, we implemented extensive data cleaning to ensure the reliability and accuracy of our findings. Our objective is to identify which predictor variables significantly affect the assault rates and how well these chosen variables predict our target variable, with the ultimate aim of informing future crime prevention strategies.

Before pre-processing the dataset there were 147 features, 2215 rows of values, and no discernable missing values. After a closer look, it was discovered that instead of labeling missing data as NaN (Not a Number), the missing values were filled with '?'. In total, there were 44,592 instances in the dataset to imply NaN values. In the case of our dataset, we converted the question marks to NaN values and ensured all the columns were numeric types by coercing non-numeric entries to NaN. The handling of missing and NaN values was a critical step, and we opted for median imputation, a method that fills in gaps while minimizing bias. This approach allowed us to retain as much information in the dataset as possible without losing too much information by filling some of these missing values with the median of that feature. Within our predictors and target, there were only 13 NaN values after preprocessing. The nature of the question “Can the assault rates be anticipated by parental structure in varying cities across the nation?” is indicative of some sort of relationship between whether a household has one or two parents and assault rates in relation to the community population. Below in Figure 2 is the correlation matrix we created to identify the features with the strongest correlation with assaults per populus that represent the parenteral/familial structure.

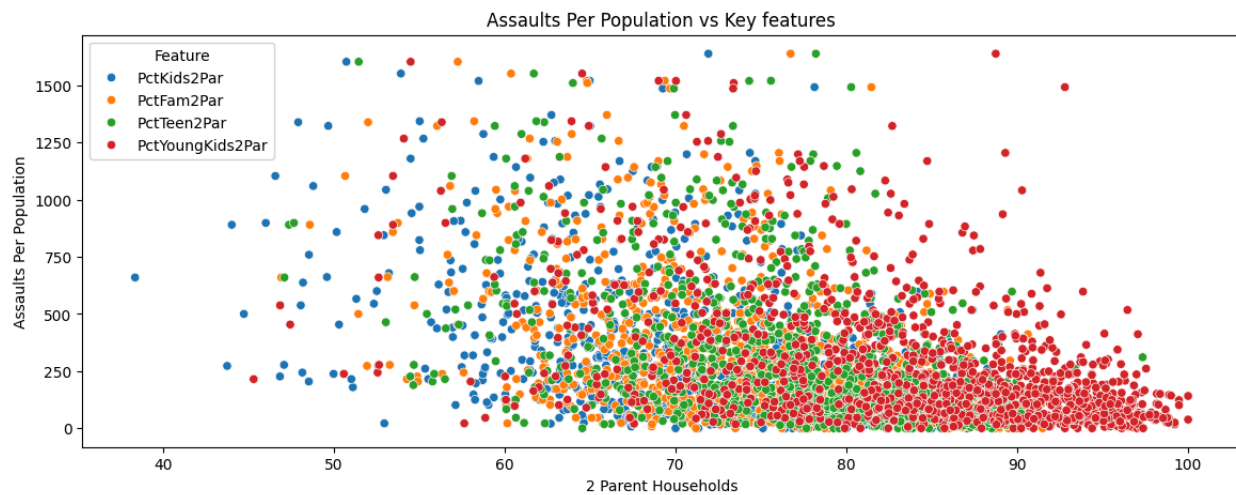
Figure 2
Feature Identification



The variables ‘PctKids2Par’(Percent of kids 6 and under with 2 parents), ‘PctFam2Par’(Percentage of families with 2 Par), ‘PctTeen2Par’(Percentage of Teens with 2 Parents), and ‘PctYoungKids2Par’(Percentage of kids 4 and under with 2 parents) are representative of our question and had a correlation coefficient (r) of -0.47 to -0.51, a moderately strong negative relationship. Out of these variables, only one had a strong positive correlation: ‘PctKidsBornNeverMarr’(representing percent of kids born to parents who never married), which indicates that our hypothesis is acceptable. This means that the higher the assaults per pop, the less likely there are high percentages of 2 parents in a family in a community. To visualize these correlations, we plotted them against the Assaults per Populus response variable, these can be seen below in Figure 3.

Figure 3

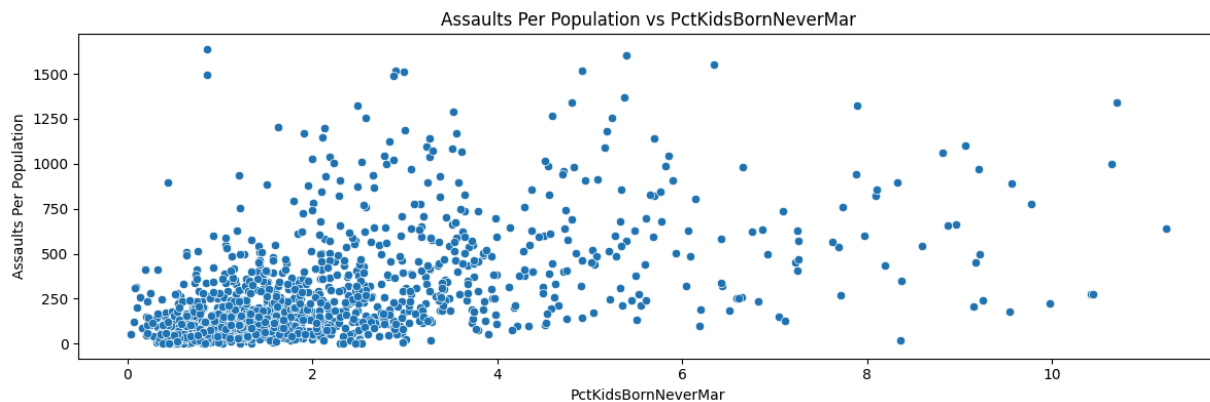
'PctFam2Par', 'PctKids2Par', 'PctYoungKids2Par' and 'PctTeen2Par' vs 'AssaultPerPop'



As assault rates decrease on the y-axis, the amount of families with 2 parents rapidly increases, with there being few instances where there are high numbers of two parent families alongside high assault rates. Since just one of the variables reported was positively correlated with our assault rates, we decided to display its distribution separately in Figure 4.

Figure 4

'PctKidsNeverMarr' vs 'AssaultPerPop'



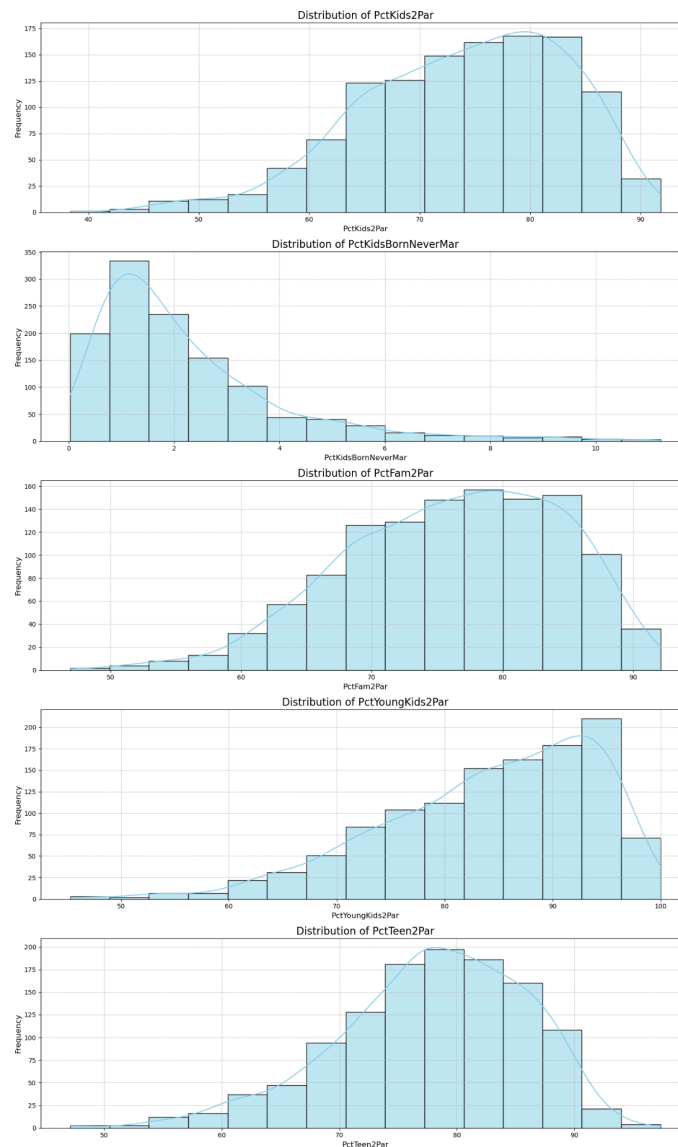
After seeing an indication of a relationship assaults per pop had between the variables: ‘PctFam2Par’, ‘PctKids2Par’, ‘PctYoungKids2Par and PctTeen2Par’, ‘PctKidsBornNeverMar’; we understood that we could move forward with our investigation.

The distribution of these variables individually is important to understanding which models will be the most helpful to use and how exactly to handle outliers, so we decided to take a look.

Below is a visual representation of all predictor variable distributions in Figure 5.

Figure 5

Predictor variable distribution



Seeing the skewed distribution of these variables lets us know to be careful moving forward with our data processing. When we created our new dataset, our sample size decreased to a total of 7182 observations, a relatively large dataset still.

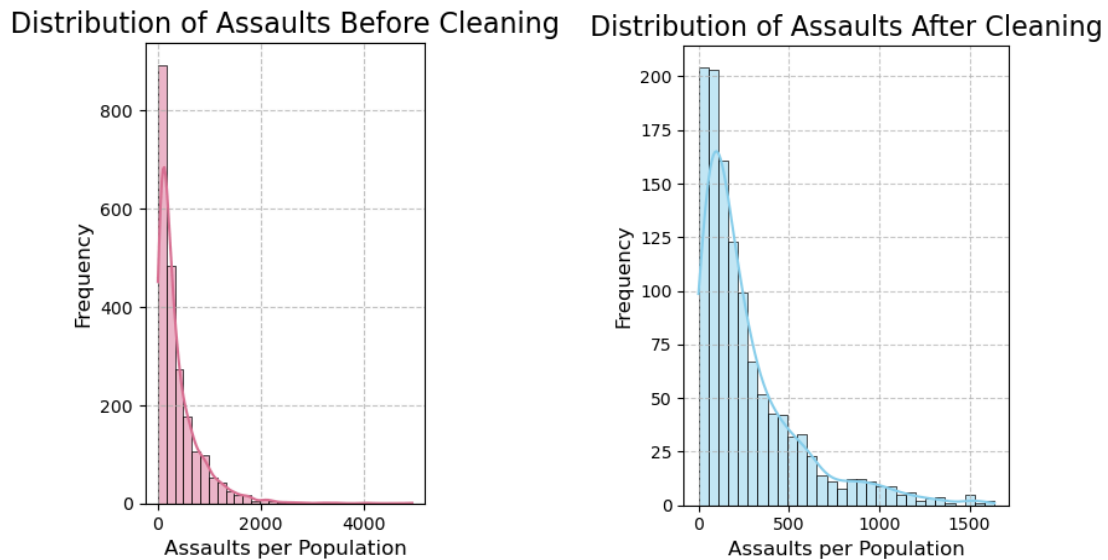
Methodology

Data Processing

Next steps in our processing was identifying and removing outliers of which there were 23,285 total in the entire dataset. Removing or handling outliers is important because outliers can impact the quality, accuracy, and interpretability of our analysis. Outliers can heavily influence statistical measures such as the mean, standard deviation and correlation, even potentially skewing hypothesis testing. To assess whether the data point falls outside of the typical range of values, we employed the Interquartile Range (IQR) method for identifying the outliers in each feature, then calculated and implemented z-scores to remove outliers. A z-score quantifies how far a specific data point is from the mean of the dataset in terms of standardized deviations. For there to be outliers left after this process suggests that the outliers are within 3 standard deviations but still extreme or the data is not normally distributed and is skewed. View Figure 6 below for a comparison of our target feature's distribution before and after this process.

Figure 6

Distribution of Assaults



As can be seen above, the data is still skewed but not as drastically as before. With a skewed distribution, it may impact which model performs best. Linear models such as linear regression and logistic regression tend to perform better with normally distributed features because the models make assumptions about constant variance and linear relationships between features. Tree-based models like decision trees, random forests, or gradient boosting are non-parametric, meaning they do not make assumptions about the population from which it was drawn and therefore do not assume a specific distribution. Unlike regressors, they focus on the ranking or order of the data and are minimally affected by outliers.

After preprocessing the data, a correlation matrix was created to further explore the relationships within our data and to better understand what variables might be most correlated with the target variable, 'assaultPerPop'. This analysis was important for determining which factors exhibited the strongest associations with assault rates and helped reinforce our selection

of predictors for the regression models. With the selection of predictors identified as significant, the dataframe can then be prepared to organize the data to a format suitable for model analysis.

As one of the last steps in preparation for our models and analysis, the dataset is split first into input data (X) to contain all the independent variables used to predict the target variable which are 'PctFam2Par', 'PctKids2Par', 'PctYoungKids2Par', 'PctTeen2Par', 'PctKidsBornNeverMar' and our target variable (y) to contain the outcome we are trying to predict which is 'assaultsPerPop'. The models will help us explore the relationship between the input features and target variable. After establishing the independent and dependent variables the dataset is split 70/30 into training and testing subsets. The training set does as the name implies and is used to train the model which learns from the relationships between X and y using 70% of the data for training. The remaining 30% of data is put aside for the test set to use after the model is trained and to evaluate the model's performance. The test set contains data the model did not use for training which allows assessing how well the model generalizes to unseen data. A smaller test set is typically used on datasets this large, however time limitations do not allow for us to compare multiple times, and using a 70/30 split ensures there is enough data in both sets to properly train and compare.

Then we proceeded to apply linear regression modeling techniques to assess the relationships among the identified predictor variables (X) and our target variable (y). Linear regression is a predictive modeling technique that can provide a trend line that best fits the data based on a few assumptions: there is a linear relationship between the dependent variable and the independent variables, observations are independent of each other, the residuals (errors) have constant variance at all levels of independent variables, the residuals should be normally distributed, and the independent variables should not be highly correlated with each other. Linear

regression is a good tool for predicting continuous outcomes and can represent a measurable quantity.

To further explore the intricate relationships among predictors, we implemented the Random Forest Regressor, known for its ability to capture nonlinear interactions and relationships between variables. This model's ensemble approach combines predictions from multiple decision trees, enhancing accuracy and reducing the risk of overfitting that can arise from using a single decision tree. Like the Linear Regression model, evaluation metrics are used in determining the model's performance.

As one last regression model for comparison, we implemented a K-Nearest Neighbor regression model. The KNN offers a simple but effective method for regression analysis when compared to the more complex models. KNN is a flexible type of model that doesn't require assumptions about data distribution. It's a model that adapts easily and does not have many parameters but it can be prone to overfitting. Just like previously, its performance is evaluated using the evaluation metrics. Each of these regression models has its strengths and limitations, making them suitable for different types of problems and datasets.

In addition to regression models, we decided to implement several machine learning models to enhance our analysis beyond the baseline models and evaluate their performance in predicting assault rates. To adapt to classification models, we manipulated what was being predicted from a continuous value to predict the likelihood of assault rates exceeding a certain threshold, in this case it is the median assault rate. We transformed the target variable into a binary classification problem, where we classified assault rates above the median as one class and those below as another.

Logistic Regression was utilized due to its effectiveness in binary classification tasks. Logistic regression does not predict a continuous outcome, rather, the probability that an input falls into one of the two categories. It estimates the probability that a given input point belongs to a certain class, providing coefficients that indicate the influence of each predictor on the outcome. However, logistic regression assumes a linear relationship between the log-odds of the outcome and the predictors, which may not always hold. This model is intuitive but can be susceptible to overfitting without proper pruning or regularization.

Decision Trees create models based on a series of binary decisions, leading to predictions through feature splits. While intuitive and easy to interpret, Decision Trees can be prone to overfitting without proper pruning or regularization, making them less reliable in complex scenarios. The Random Forest Classifier is an extended version of decision trees in a way, this is a supervised machine learning algorithm that is formed with a group of decision trees. Random Forest utilizes an ensemble method called bagging to create diverse decision trees from random subsets of the training data, combining their predictions to reduce overfitting and to increase accuracy. While random forests do not make strong assumptions on feature distribution, it does assume that observations in the data are independent from each other and that data can be divided into homogeneous subgroups based on the target variable. Random forests do well with large datasets, can handle missing values, and can work with skewed features. It is easy to use and efficient but can be prone to overfitting, though to a lesser extent than single decision trees which makes the random forest classification model seem like a model well suited for our dataset.

Naive Bayes is a classification model that does well with a dataset containing many features. The main assumption of this model is that all features are conditionally independent

from each other which means it assumes none of the chosen features are correlated to each other when predicting the target class. Another assumption is that the data is normally distributed. The algorithm calculates the probability that a given data point belongs to a particular class based on its features. This algorithm is commonly used for text classification where the features are treated as independent even though the features may have some correlation in reality. Our dataset handles mostly numerical data and is very large, so Naive Bayes may not prove to be the most effective model to analyze our dataset.

We chose to measure the models' performance both by the MSE (Mean Squared Error) and the R^2 . Mean Squared Error measures the average difference between the predicted assaults per population and the actual values observed. The R^2 , also referred to as the coefficient of determination, is a numerical representation of the proportion of assaults per population that can be explained by the model given our independent variables. The larger the MSE signifies a larger distance between the model's predictions and true values found within the dataset. The RMSE is these differences before they are squared and summed, making them less resistant to outliers. To further test our models beyond the output of their own scores, we implemented a Cross Validation method for further testing and training with 5 folds. This is a statistical method that splits the data into subsets for continuous training and testing, ensuring the model's ability to predict covers all parts of the data.

Cross-validation, particularly k-fold cross-validation, is a technique used to assess how well a model generalizes to unseen data. In k-fold cross-validation, the dataset is split into k equally sized subsets. The model is trained on k-1 subsets and tested on the remaining subset. This process is repeated k times, with each subset serving as the test set once. Cross-validation provides a more reliable estimate of model performance by reducing the risk of overfitting to a

single training/test split. When we used these evaluation metrics on each model, we used 5 folds in our cross validation which means the dataset is split into 5 parts, and the model is trained and tested 5 times.

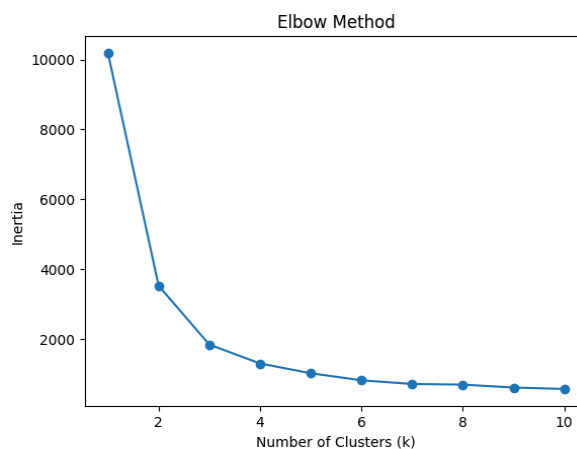
K-Nearest Neighbors Classifier is a supervised learning algorithm and was included as it classifies data points based on the majority class among their nearest neighbors, offering a non-parametric approach. The K-NN method is started with a dataset with known categories and clusters of data points. Then data that has not yet been binned is introduced and classified by how close it is to data clusters or its "nearest neighbors". The initial clustering of the data where we already know the categories is the training data. If the unclassified data point is close to several groups, the point would be categorized with whichever cluster of data has more marks within the K value given. This can also be used for assigning categories to variables used in heat maps. There is no concrete way to choose K so several K values may need to be tested out but it is good to keep in mind that a low K value may be more impacted by outliers while a large K value may be too inclusive.

Lastly, we used the Gradient Boosting Model. Like Random Forests, it is an ensemble technique, however, GBM builds models sequentially. The purpose of this is to correct errors made by previous models by focusing more on misclassified instances. It recursively predicts residuals to ultimately correct the model based on its residual predictor. It is a flexible model that can be used in regression or classification and can produce fairly accurate models that are not as prone to overfitting if the learning rate is not too high. The GBM is capable of capturing non-linear relationships between features which may be useful in analyzing our dataset.

After analyzing the supervised models, we implemented one more method to explore our data. We used K Means Clustering, an unsupervised learning algorithm to group a dataset into K

clusters, where K is a predefined number of clusters. Its goal is to separate and group data by its similarities. For this reason, we used just our predictor variables to identify unknown similarities between them, creating a new feature for analysis that the algorithm might find. The algorithm assigns each data point to the nearest centroid, thus forming the clusters. It then continues to recalculate the centroids by finding the mean of all the data points assigned to each cluster and replotting the centroid. These similarities are predetermined by the algorithm and take careful interpretation to identify their significance in relation to the data. Each cluster has a central point called a centroid that represents the center of the cluster and is the average position of that data assigned to a cluster. This algorithm works well on large data sets, but requires the number of clusters (K) to be specified beforehand. It is also sensitive to non-linear or imbalanced data. For this model it is not necessary to clean outliers, instead we used a robust scaler that scales the data based on the IQR instead of mean and standard deviation. Since the median and IQR are not sensitive to extreme values, this scaling method ensures that the outliers do not dominate the feature scaling process. To specify the number of clusters appropriate for the algorithm, we implemented the Elbow Method, as can be seen below in Figure 7.

Figure 7
The Elbow Method



The point where the inertia slows down dramatically (the elbow point) appears to be at 4 clusters, therefore we determined **K=4** for this data-subset.

Model Analysis and Results

The question “Can assault rates be anticipated by parental structure in varying cities across the nation?” suggests some sort of relationship between assault rates and whether a household has one or two parents.

Regression Models

After deciding to first use a multiple linear regression on our numeric data, we trained them and tested them. Below in table one you can see the results of the metrics previously discussed.

Table 1
Model evaluation

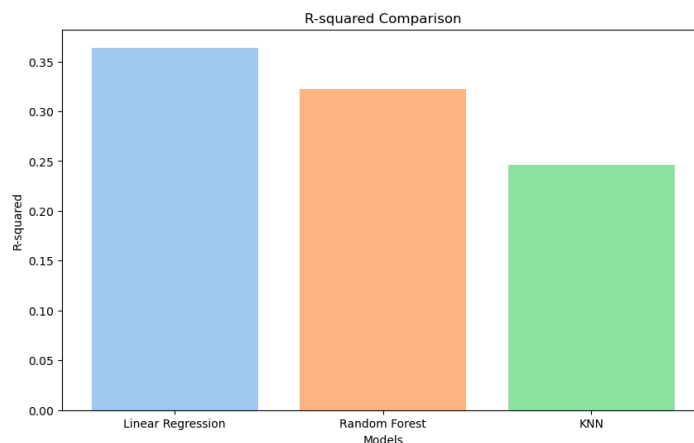
Regression Model	MAE	R ²	MSE:	RMSE:
Linear Regression	159.63	36.39%	55502.65	235.59
Random Forest	166.02	31.76%	59542.30	244.01
KNN	174.65	25%	65793.23	256.50

Performance

The Linear Regression performed the best, achieving an R² value of 0.3639, which means that 36.39% of the variance of assaults per populus is explained by the model. The MSE and

RMSE relative to our dataset are low relative to the range of our values, indicating decent predictive accuracy with room for improvement. The RMSE suggests that our model predicts an average of 235 assaults more or less than our actual assaults based on its given predictors (2 parent households). This is not an unacceptable result considering we have a range of values from 0 – 4,932, indicating a need for more data. In relation to the maximum of our dataset, a typical error of 6.38% could be acceptable. An important note to keep in mind is that the MSE will amplify all errors, including large ones, making it a very sensitive form of measurement. Its cross-validation scores are relatively consistent and suggest the model has covered the data well. Once again taking into consideration that we have many large values in the dataset, the MAE is not entirely unacceptable. Below in Figure 8 is a histogram of a comparison between the R-squared and MAE metrics of all regression models that we employed.

Figure 8
R-Squared and MAE Regression Comparison

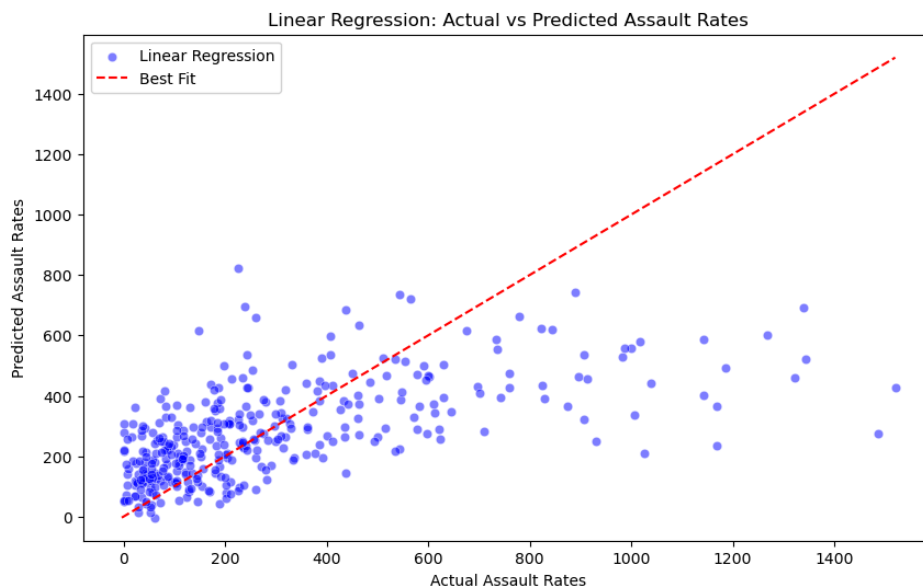


The Random forest did perform second best, likely due to its ability to correct itself when it assumes an error. This would lower its reliability, as this model is prone to heavily reduce overfitting (drawing explanations from the data where there are none), resulting in the lowest R-squared that we see above. This would also explain its slightly higher MSE, suggesting that it

was more deliberate when recursively comparing the predictions of each tree to reduce error. Because this model is aimed toward reducing MSE, it will be less biased by providing less leeway for what is considered acceptable. The KNN performed poorly, doing a poor job of predicting as it does not make use of the same regression equation that Linear and Random Forest regressors do, making it extremely sensitive to large, abnormal data such as this.

Below in Figure 9 is the regression line of best fit, visualizing exactly what the model expected against the actual instances.

Figure 9
Linear Regression- Line of best fit



As mentioned above, the model did a mediocre job of predicting future instances, as most values do not fall around the line of best fit. To test this in the future, we would employ multiple models to be sure that we successfully reduced multicollinearity. Just to emphasize how much better it performed than the KNN regression or the Regression forest models, below in Figure 7 is a histogram of their R-squared side by side.

Hypothesis Testing

We next set up a significance test for each coefficient (predictor variable) of the model to see if each one shows enough significance to support our hypothesis. Below in Table 2 are each of our predictors and their p-values. In the case of our study, the null hypothesis will be that Parental structure does not have a strong enough relationship with Assault rates to influence them in cities across the US, based on this model. The alternative hypothesis will be that it does. For the sake of clarity and reduced multicollinearity, we excluded ‘PctKids2Par’.

Table 2
Model evaluation

Predictor	P-Value	Coefficient
Kids Born to Parents who Never Married	0.000	-39.85
Young kids (6 and under) in 2 Parent Family	0.000	30.50
Teens in 2 Parent Family	0.417	0.35
2 Parent Family	0.001	13.23

The alpha level directly affects the likelihood of a Type I error (false positive) while also influencing the risk of a Type II error (false negative). For larger datasets, it is critical to set a stricter alpha level to minimize errors. As our study is confirmatory, we have chosen an alpha level of 0.10. Since our focus is on identifying a negative correlation, we decided to perform a one-sided z-test and so any p-values exceeding the alpha threshold of 0.10 will be considered statistically insignificant. Predictors like ‘2 Parent Family’ and ‘Young Kids in 2 Parent Family’ have p-values indicating potential relationships with assault rates based on this model. However,

we must also examine the role of each predictor's coefficient, as it helps us understand the direction and strength of the relationship.

For example, the predictor 'Young Kids in 2 Parent Families' has a p-value close to 1, which suggests it is not statistically significant. Although the positive coefficient of 13.23 indicates that increasing the proportion of young children in two-parent families is predicted to lead to higher assault rates, the p-value near 1 suggests that this relationship is likely due to random chance, so we cannot confidently conclude a meaningful connection between this predictor and assault rates in this model. On the contrary, the predictor 'Kids Born to Parents who Never Married' shows a strong negative coefficient of -39.85, but the p-value of 1 indicates that this relationship may not be statistically significant, implying that any association between this variable and assault rates is likely due to random fluctuations. The negative coefficient suggests that an increase in the proportion of children born to unmarried parents correlates with a decrease in assault rates, but the p-value suggests this is not a reliable predictor in this context. The 'Teens in 2 Parent Family' predictor shows a moderate relationship with assault rates, with a p-value below 0.05, suggesting statistical significance. This indicates a meaningful positive relationship between the proportion of teens in two-parent families and assault rates, though the coefficient is relatively small compared to the other predictors.

Based on the results from hypothesis testing, we reject the null hypothesis, which states that parental structure does not influence assault rates, and accept the alternative hypothesis. This suggests that parental structure does indeed have a statistically significant relationship with assault rates, based on the predictors tested. However, these findings should be interpreted with caution. The coefficients and p-values for some predictors, such as 'Kids Born to Parents who

Never Married,' suggest that the relationships are not statistically significant, and their practical significance may be questionable.

These results should be considered alongside other model evaluation metrics and further analysis to ensure robustness and reliability. Any p-values below our threshold of 0.10 are considered statistically significant, but these findings should be viewed with some degree of skepticism, particularly in light of the model's ability to predict. Taken into account with the measurements of our models' ability to predict, these values are to be taken with a grain of salt.

Classification Models

We decided to discretize our data to simplify other complex relationships that might be making our interpretation less clear cut, thus we used models that recognize what could be considered high or low assault rates. We placed all rates into two classes, high and low. High assault rates are those above the median and low assault rates are below the median rate. The accuracy of the model is the proportion of correctly classified instances out of the total instances. An instance where an actual value and a predicted value are aligned is referred to as being true. Accuracy can be misleading on unbalanced datasets. If the majority of the dataset belongs to one class – is above or below the median in this case – then predicting the majority class will give a strong baseline. For this reason we also look to other measurements, such as the models' precision: the proportion of true positives out of the total predictions the model classified as positive. If there are just a few false predictions of assault rates above the median (false positives), then the result will be closer to 100% . How high the precision score of a model is would suggest its predictions of positives are correct, but how can we know the model has correctly identified positives within the data?

The recall or True Positive Rate, is a measurement showing the proportion of true positives out of the total true positives and false negatives (those incorrectly classified as below the median), essentially measuring how many actual positives were correctly identified. The F1-Score is a metric that combines precision and recall into a single value to evaluate the model's performance and is especially useful in the case of unbalanced datasets such as this one. In order to measure how well a model distinguishes between positive and negative classes, we use the Receiver Operating Characteristic curve (ROC) which plots the recall and false positive rate across the different thresholds and the Area Under the Curve (AUC) is the measure of the area under that curve. All measurements can be seen below in Table 3.

Table 3

Model Evaluation

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC	Cross-Validation on Accuracy Mean
Logistic Regression	73.6%	80.7%	67.1%	73.4%	74.2%	74.6%
Decision Tree	66.1%	69.9%	68.7%	68.7%	67%	65.3%
Random Forest	72.2%	77.8%	68.7%	72.7%	72.9%	71.7%
Naive Bayes	72.2%	83.2%	61%	70.4%	73.2%	70.8%
K-Nearest Neighbors	73.6%	77.5%	72.3%	72.3%	73.7%	70.8%

Gradient	73.3%	77.7%	71.3%	74.3%	73.8%	72.3%
Boosting						

Performance

The Decision Tree model achieved an Accuracy of 66.1%, indicating that it correctly classified approximately 66.1% of communities as having high or low assault rates based on family structure predictors such as PctFam2Par. With a Precision of 69.9%, the model was moderately effective in minimizing false positives, while a Recall of 68.7% shows that it identified most high-assault-rate communities. A lower recall than precision suggests the model's inability to identify false positives, therefore it has inefficiently calculated its precision score. The F1-Score of 68.7% reflects this, though its ROC-AUC of 67% shows its limitations in distinguishing between high and low assault rates. This performance suggests that the Decision Tree definitely has room for improvement in generalizing across the dataset. The model's consistent cross-validation accuracy of 65.3% reinforces our assumption that it is improperly identifying data due to its skewed distribution.

The Random Forest model had a stronger performance, with an Accuracy of 72.2%, and a Precision of 77.8%, the highest among all models, demonstrating its ability to minimize false positives. Its Recall of 68.7% goes to show that it successfully identified a large proportion of high-assault-rate communities, however it should be kept into account that a large proportion of the data (around 65%-70%) was below the median. Its F1-Score of 72.7% underscores a strong balance between Precision and Recall, highlighting its false reliability. The ROC-AUC of 72.9%

effectively separates high and low assault rate classifications. By leveraging multiple decision trees, it uses predictors like PctFam2Par and PctKidsBornNeverMar effectively. The cross-validation accuracy being 71.7% demonstrates that this model performs rather well when it comes to identifying the relationship between family structure and community safety.

With an Accuracy of 72.2% and a particularly strong Precision of 83.2%, the Naive Bayes model demonstrated its capacity to successfully reduce false positives. However, when covering all high-assault-rate communities, its limitations were evident by its 61% recall, which I believe feature independence to be the blame. With a strong ROC-AUC of 73.2% and an F1-Score of 70.4%, this trade-off demonstrated how well it could differentiate across classes. Although its assumptions restrict its capacity to simulate more complex patterns in family structure, the cross-validation accuracy of 70.8% validates its dependability as a simple classification tool.

The K-Nearest Neighbors (KNN) model achieved a 73.6% Accuracy and a Precision of 77.5%. Although it is comparable to Naive Bayes, its 72.3% Recall suggests that it is difficult to identify all communities with high assault rates. A fair compromise between reducing false positives and detecting true positives is shown by the 72.3% F1-Score. Although KNN showed dependable discriminatory power with a ROC-AUC of 73.7%, its sensitivity to parameter selection and data scaling indicates that it would need further fine-tuning to match the robustness of ensemble approaches like Random Forest. Its capacity to successfully generalize across the dataset is confirmed by its 70.8% cross-validation accuracy.

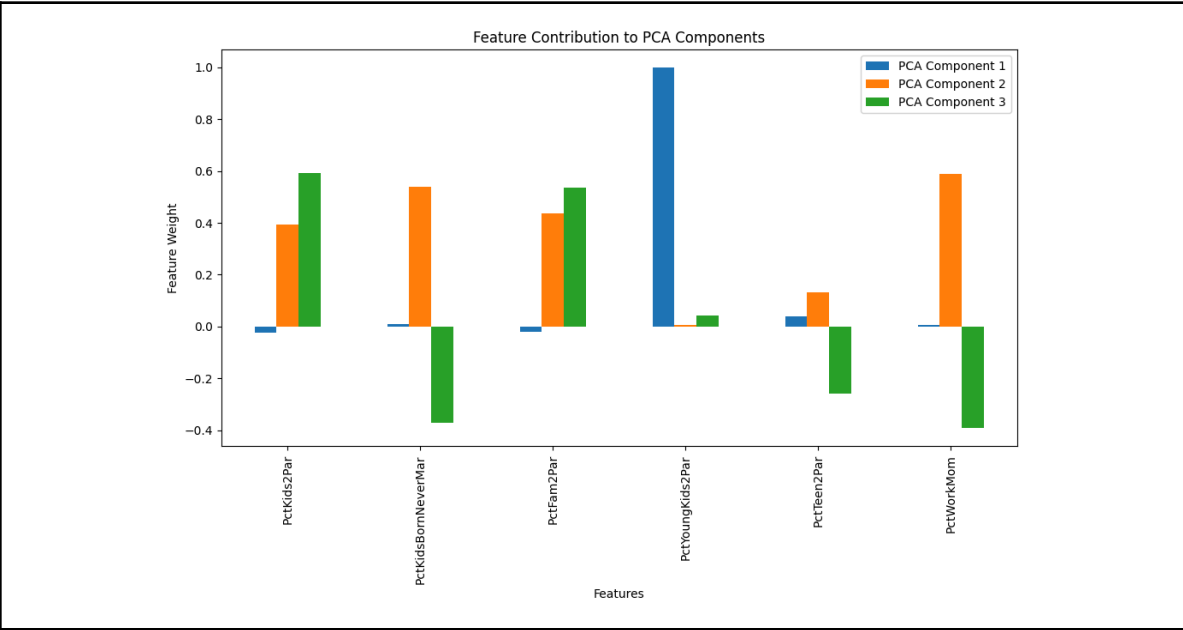
The Gradient Boosting model proved itself to be our strongest performer, achieving an Accuracy of 73.3%, Precision of 77.7%, and Recall of 71.3%. Because of the Gradient

Boosting’s sequential learning process, it was allowed to refine predictions iteratively, making it particularly adept at leveraging predictors like PctFam2Par to classify communities accurately. Its cross-validation accuracy of 72.3% underscores its reliability as the most effective model for understanding the role of family structure in predicting assault rates.

K-Means Clustering-Unsupervised Learning

The goal when it comes to analyzing a K-Means clustering is to understand the groupings made by the algorithm, we must first understand exactly what the PCA components are. PCA components or principal components are new features created as part of the feature reduction. We analyzed the weights that each original feature had on each PCA component to attempt understanding what the algorithm represented them as. Below is a visual representation of the weights each component contributes to the model's PCA components in Figure 10.

Figure 10
Algorithm Analysis



The y-axis represents feature weight in the decided PCA Components and on the x-axis are each variable in the dataset. PCA components are linear combinations of the features, meaning that the dominant feature decides how data is distributed along each PCA component. This graph suggests that the Percent of Young kids with 2 Parents was undeniably the biggest contributor to PCA Component 1, which means that it has the highest weight on that component, the exact weight being 83.79. PCA component 2 was more difficult to gage, as many of the remaining features had similar attributions. The one that appears to be the highest is Percent Working Mom with 0.589 weight. PCA component 3 was most heavily contributed to by Percent Kids with 2 Parents with a weight of 0.592, almost weighing the same as Percent of Families with 2 Parents, at 0.535. For almost all components Percent of Kids Born to Parents that Never Married contributed the least, and even negatively contributed to PCA Component 3.

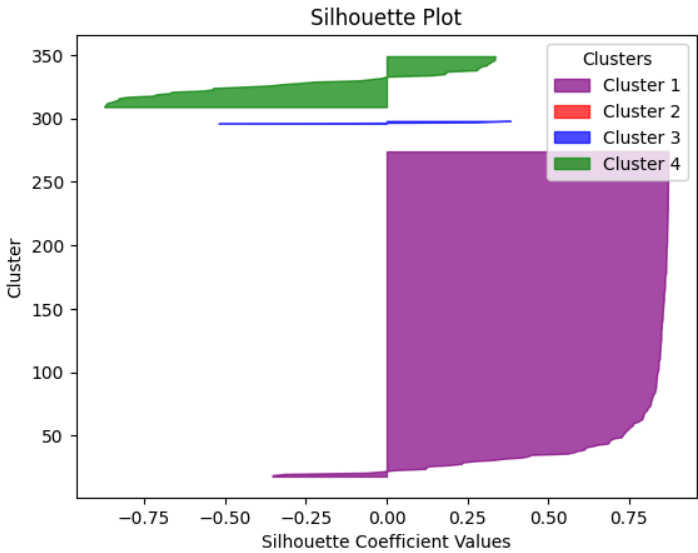
This suggests that there are clear similarities between households with 2 parents, and those without.

Performance

There are two main ways to assess the algorithm's ability to partition the data based on similarities it finds within it, but we chose to focus on two measurements: Silhouette Score and the normalized inertia (as shown in Table 3). The Silhouette score is a metric that measures how well a point is associated with its cluster, providing insight into the quality of the clustering and evaluating how distinct the clusters are. Taking the silhouette score of each point and averaging them together gives us the overall silhouette score. The closer to 1 the score is, the more cohesive the clusters would be. On the other hand, inertia is a measurement of how tightly data points are clustered around their respective centroids, indicating how far the data spans it. This measurement can be sensitive to the scale and number of features within a dataset, so to get a

sense of its magnitude we normalized it by dividing by the number of points in the dataset, which can also be found in Table 4.

Table 4
K-Clustering Evaluation

Algorithm	Silhouette Score	Silhouette Plot	Normalized Inertia
K-Clustering	0.63		3.72

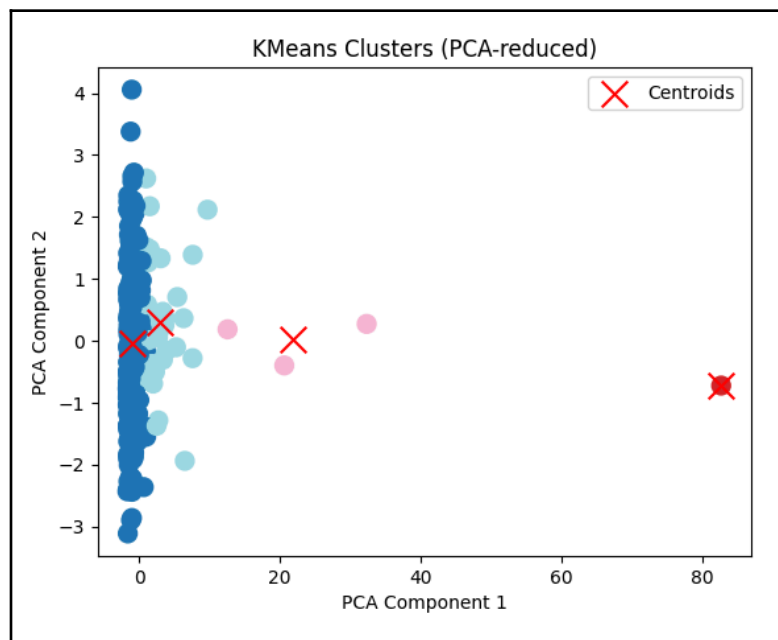
The normalized inertia shown above indicates that the average squared distance of each algorithmically identified point is 3.7 units away. This lets us know that the clustering has semi-successfully partitioned the data, but signifies that the data may be too similar to partition. Our Silhouette Score essentially reassures this indication that there is a need for improvement, as it suggests that the clustering quality meets the minimum requirements. Silhouette score of 0.6 suggests that our clusters are just barely decently formed, hovering just above the halfway

threshold. The plot of silhouette scores indicates that there are too many clusters for the dataset as Cluster 2 is nearly invisible when plotted. While this may seem true, when we experimented with lowering and raising our clusters, it reported unreasonable scores from both inertia and the silhouette. This effect can be observed in Figure 11 below which is a 2D plot of the algorithm clusters.

Figure 11

K-Means Clusters

Visualization 2: 2D Plot



As observed, there are no extraordinary distinctions between clusters which may mean the algorithm found our variables too similar. There is one centroid that does not have a cluster, which may show an outliers' effect on the K-means, despite the robust scaling. It is also important to keep in mind that this dataset is largely left skewed and even a true value can

present as abnormal in its context. Other models previously have indicated this, with many of our classification models showing scores indicative of a skew of positively classified values, such as the Decision Tree and the Naive Bayes. Possible mis-use of a distance metric could have this effect as well, as the K-means clustering uses Euclidean distance that assumes the dataset shape is not spherical. Mistakes such as these would suggest future reevaluation of model choice.

Results

Can the question ‘Can assault rates be anticipated by parental structure in varying cities across the nation?’ be answered effectively by the results mentioned above? It should be stated that generalizing such findings would be inappropriate and inaccurate due to the nature of our studies limitations, however our best performing models, unsupervised learning algorithms, and significance testing suggest a strong relationship between 2 parent households and assault rates. For sociologically centered studies such as these, it is important to understand the effect of outside features that these families may or may not fall into. For example, the k-clusters found separate similarities among our input features like: Income, Gender, etc., that could unknowingly explain further this gap in the relationship. The results of this study led us to fail to reject our null hypothesis, indicating there were strong relationships between many of our predictor variables, specifically being: ‘Teens in 2 Parent Family’ and ‘ were promising and with further research can be used to inform future policies.

Discussion

Our analysis revealed evidence that communities with a higher prevalence of two-parent households exhibited lower assault rates, supporting the hypothesis that family stability serves as a protective factor against violence. This aligns with existing literature that emphasizes the

importance of familial support and supervision in mitigating risk factors associated with delinquency. The policy implications of our study highlights the importance of a multi-level approach to crime prevention that addresses both family stability and socioeconomic challenges within communities.

First, family support programs, such as relationship counseling, parenting resources, and financial assistance like child tax credits, are essential for strengthening family cohesion and alleviating economic stress, particularly in single-parent households where financial strain often hinders effective supervision. By enabling parents to invest more in their children's education and development, such programs help create a more stable and supportive environment, reducing the risk of delinquency.

Community-based family services can further enhance social cohesion by providing critical resources directly within neighborhoods. Family-oriented social work and community centers offering after-school care, mental health services, and educational workshops reduce social isolation and connect families to vital support networks, fostering resilience in areas with high single-parent prevalence. Educational initiatives, especially within schools located in high-crime areas, can build supportive environments for children from diverse family backgrounds. School-based programs that emphasize parent engagement, youth mentorship, and extracurricular opportunities give at-risk youth structured activities that reduce their likelihood of engaging in criminal behavior.

Addressing economic inequalities is another vital aspect of crime prevention, as stable employment and job training programs directly impact family stability and resilience. For single parents and low-income families, subsidized child care can significantly reduce financial stress and improve work-life balance, enabling parents to focus on fostering positive outcomes for their

children. At the community level, promoting neighborhood safety through initiatives like Crime Prevention Through Environmental Design (CPTED) and neighborhood watch programs enhances collective efficacy, or the shared belief in a community's ability to maintain social order. Well-designed public spaces, improved lighting, and resident-led safety initiatives create an environment where neighbors work together to prevent crime.

These insights are particularly relevant for stakeholders such as law enforcement agencies, state legislators, and community organizations. Law enforcement can utilize this information to develop community outreach programs aimed at strengthening family units, while state legislators can advocate for policies that support family stability through funding for educational and community-based programs. Families, especially those with children, stand to benefit from initiatives that promote parental involvement and provide resources to support dual-parent structures.

The findings of this study not only emphasize the critical role of parental structure in influencing violent crime rates but also highlight the potential of leveraging data-driven approaches for targeted interventions. Communities with stronger parental stability, indicated by a higher prevalence of two-parent households, demonstrated lower assault rates. These insights must be contextualized within broader societal dynamics, such as socioeconomic disparities and community-level variables.

The integration of evidence-based programs like Functional Family Therapy (FFT), LifeSet, and Communities That Care (CTC) underscores the tangible benefits of addressing family instability proactively. These programs offer scalable solutions tailored to the unique challenges of different communities. FFT's focus on improving communication and reducing negativity within families has shown significant reductions in youth recidivism. Similarly,

LifeSet addresses vulnerabilities faced by youth such as housing insecurity and mental health challenges, leading to improved social and economic outcomes. CTC takes a broader community approach, emphasizing collective action and the use of localized data to tackle risk factors for youth violence and delinquency.

Scaling these interventions requires overcoming challenges, including funding constraints, workforce training, and cultural adaptation to diverse communities. Cross-sector collaboration between criminal justice, education, and social services can maximize resource efficiency, while technology-driven monitoring and evaluation tools enhance program effectiveness. Nonetheless, ethical considerations must remain central. Programs should be framed as supportive rather than punitive to prevent stigmatization of vulnerable groups, such as single-parent families. This approach ensures that interventions are empowering, equitable, and culturally sensitive.

The clustering and supervised learning models applied further demonstrate the utility of familial and socioeconomic data as indicators for community-level assault risks. Caution is warranted to avoid oversimplifying these relationships. Family structure is significant but is only one of many interrelated factors contributing to crime rates. Socioeconomic conditions, education access, and social cohesion must be integrated into a holistic framework for understanding and preventing violence. Regional and cultural variations demand that policies and programs be adaptable to the specific needs of each community.

The temporal scope of the dataset highlights the need for updated and comprehensive data to validate findings in contemporary contexts. Longitudinal studies could track how changes in family dynamics, economic conditions, and policy interventions impact crime trends over time, offering a clearer roadmap for future policymaking.

Potential improvements in this data exploration could include a more thorough pre-processing to normalise the distribution of the data's features through the use of log transformation in addition to the robust scaler. The logistic and linear regression models appeared to perform the best among the machine learning models used but this is most likely due to how skewed the distribution of the data. Such models function under the assumption that the data is normally distributed and can be prone to overfitting, this could result in poorer generalization to unseen data and is a potential explanation of what happened within our models. Given the opportunity to continue this data exploration, improvements to the study could be done through additional data collection and the use of more recent data. This dataset is over 2 decades old and can not effectively be applied to many crime statistics that we see today. It is understood that this amalgamation of data from 1990-1995 could pose a hindrance to obtaining the full story on crime in these cities in states across the US.

Conclusion

This study investigated the relationship between family structure and violent crime rates, combining supervised and unsupervised learning models to uncover meaningful patterns in the data. The findings confirmed a negative correlation between two-parent households and assault rates, supporting the hypothesis that family stability serves as a protective factor against violent crime. However, the relatively low R^2 scores from regression models indicate that family structure alone does not fully explain crime rate variability. Socioeconomic conditions, community cohesion, and other unmeasured variables contribute significantly to these outcomes.

K-Means clustering added depth to the analysis by identifying four distinct community types based on family stability, economic factors, and assault rates. These clusters largely aligned with supervised model predictions but also revealed intermediate and anomalous patterns, highlighting complexities not captured by regression models. For example, some clusters indicated high assault rates despite moderate family stability, suggesting that other contextual factors may influence crime.

These insights reinforce the importance of comprehensive, evidence-based interventions targeting both family stability and socioeconomic challenges. Programs such as Functional Family Therapy, LifeSet, and Communities That Care have demonstrated success in strengthening family cohesion and reducing youth violence. Coupled with investments in education, after-school programs, and community mentorship, such initiatives can address the systemic roots of violent crime while fostering community resilience. Supporting low-income families through financial assistance, affordable childcare, and counseling services further strengthens these efforts.

This dual modeling approach demonstrates the value of integrating supervised and unsupervised techniques to guide policy and community interventions. By systematically leveraging these insights, policymakers can proactively address the root causes of violence and foster safer, more supportive environments. Importantly, care must be taken to avoid stigmatizing single-parent households, recognizing instead that family structure is just one of many interconnected factors influencing crime rates. Future research should explore additional predictors, such as mental health resources, cultural dynamics, and neighborhood cohesion, to refine these models further and support more equitable policy solutions.

References

- Farrington, D. P., & Loeber, R. (2000). Epidemiology of juvenile violence. *Child and Adolescent Psychiatric Clinics of North America*, 9(4), 733-748.
- Farrington, David. "Chapter 5: Families and Crime." *Crime and Public Policy*, Oxford University Press, 2011.
- Laub, J. H., & Sampson, R. J. (1993). Turning Points in the Life Course: Why Change Matters to the Study of Crime. *Criminology*, 31(3), 301-325.
<https://doi.org/10.1111/j.1745-9125.1993.tb01132.x>
- Leeman, Robert F., et al. "Perceived parental permissiveness toward gambling and risky behaviors in adolescents." *Journal of Behavioral Addictions*, 3(2), 2014, pp. 115-123.
- McCord, Joan, and Geoffrey Sayre-McCord. *Crime and Family: Selected Essays of Joan McCord*. Temple University Press, 2007.
- Sanvictores, Terrence, and Magda D. Mendez. "Types of Parenting Styles and Effects On Children." StatPearls, StatPearls Publishing, 18 September 2022.
- Kroese, J., Bernasco, W., Liefbroer, A. C., & Rouwendal, J. (2020). Growing up in single-parent families and the criminal involvement of adolescents: a systematic review. *Psychology, Crime & Law*, 27(1), 61–75. <https://doi.org/10.1080/1068316X.2020.1774589>
- Janssen, H. J., Weerman, F. M., & Eichelsheim, V. I. (2017). Parenting as a Protective Factor against Criminogenic Settings? Interaction Effects between Three Aspects of Parenting and Unstructured Socializing in Disordered Areas. *Journal of Research in Crime and Delinquency*, 54(2), 181-207. <https://doi.org/10.1177/0022427816664561>

Steinberg , L. S., & Silk, J. S. (2002). *Handbook of Parenting Volume 1 Children and Parenting*. LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS.

United States Census Bureau. (2020). 2020 Census Results. Census.gov.

<https://www.census.gov/programs-surveys/decennial-census/decade/2020/2020-census-results.html>

Hirschi, T. (1969). *Causes of Delinquency*. Routledge.

