Visualizing and Modeling United States Community Data against Crime Statistics

Annie Edmonds Harry Shield

1. Data Description

Annie

1.1. Overall Objective

Our project utilizes a dataset that focuses on quantifying community diagnostics and crime statistics in communities in the United States. The primary objective is to use machine learning models to predict violent crime rates using the demographic and socioeconomic features provided in the dataset. By predicting our target variable, ViolentCrimesPerPop, we want to identify which features most strongly influence or predict violent crime rates and how these factors contribute to variation across communities.

1.2. Dataset Overview and Attributions

The dataset integrates information from three primary sources: socioeconomic data from the 1990 U.S. Census, law enforcement data from the 1990 Law Enforcement Management and Administrative Statistics (LEMAS) survey, and crime data from the 1995 FBI Uniform Crime Reporting (UCR) Program. The creators of this dataset included the most relevant attributes from the three sources, excluding any unrelated variables that had no connection to crime in U.S. communities. Overall, the combination of these three sources produces a dataset that is a length of 1994 instances, with a comparison of 127 features. Of the 127, five are non-predictive categorical values used for identifying U.S. states, counties, communities, and cross-validation numbers. The other 122 features are predictive measures that contain numerical data for making predictions.

The five non-predictive categorical values dataset provides geographic identifiers that allow for both broad and localized analysis of crime statistics. The 122 predictive measures include demographic variables such as age and race, as well as socioeconomic indicators, including unemployment rates and education levels. Additional personal data points include marital status, number of children, immigration status, English proficiency, and family dynamics. Law enforcement-specific variables include officers per capita, total LEMAS per population, and department demographics such as officer race. Additional data points cover law enforcement budgets, drugs seized, and overall crime rates per population.

In total, the dataset contains 127 variables, making data cleaning a challenging task. To manage this, selecting a subset of 25 features that include a mix of predictive and non-predictive variables will enable us to adequately clean and create models, as well as visualizations, to represent the communities in the dataset. These chosen features are: Harry

- state, county, community, communityname, population, householdsize, racepctblack, racepctWhite, racepctAsian, racepctHisp, agePct12t21, agePct12t29, agePct16t24, agePct65up, PctPersDenseHous, medIncome, PctPopUnderPov, PctForeignBorn, PolicOperBudg, ViolentCrimesPerPop, LandArea, PopDens, PolicPerPop, LemasTotalReq, NumInShelters

Each variable will have its own cleaning and processing pipeline, as outlined below. However, as most variables are continuous, they all may require transformation of different types. Widely, the dataset includes missing values that must be addressed during our data preparation stage to better equip the later stages of model production and feature visualization. The missing values are identified by a question mark within the dataset, which we will coerce to NaN values during the cleaning stage.

Most of the data are derived directly from the three mentioned sources; however, the creators of this dataset computed additional variables for better data understanding. For instance, the per capita violent crime variable is calculated using population data. There is also a sum of crime variable that aggregates major violent offenses in the U.S., including murder, rape, robbery, and assault.

The predictive values are represented on a normalized scale Annie
between 0 and 1, therefore will not need as much cleaning, other than to handle missing values. In order to normalize the values between 0 and 1 for consistency, the authors utilized an unsupervised, equal-binning method. The unsupervised aspect is defined as the data was only adjusted in accordance with that feature's values, without input from other columns or target values, such as violent crimes. The equal-binning aspect of the method sets the largest value at 1.00 and the smallest value at 0.00, then uses equal sized proportions to set the different values that fall between the minimum and the maximum. This alters how we can visual-

ize and model our data, because we do not know what the initial numbers are from this data set, only the normalized ones. Therefore we can only make assumptions in comparison to the values found within each feature. It is also worth noting that this normalization does not help maintain relationships between values of attributes, meaning that comparing values for related variables, such as whitePerCap and blackPerCap for a community, would not yield noteworthy insights. This adds additional challenges when we are in the process of cleaning and preparing the dataset to make meaningful comparisons between related variables and produce visualizations to demonstrate their relationships.

Harry

The non-predictive variables, state, county, community, and communityname use numeric codes and strings to represent their values. The state feature uses readily available Federal Information Processing Standards (FIPS) numeric codes. These have values ranging from 1 to 56 (skipping some numbers) to represent the 50 states. In the cleaning process, it may be helpful to create a dictionary to map the state names, as they are more useful in data visualization. This is similar to the county feature, which has a code specific to each state, using the same FIPS numeric codes. Although changing this to county names is less useful, because they are less recognizable from a visualization standpoint, this could be helpful, as they are more recognizable than numbers. Lastly, the community codes are the least beneficial, other than helping to provide a possible basis for missing data. If there is a communityname present, then a community code may be produced. As well as if the state FIPS code or county FIPS code is missing, the communityname might provide enough information to fix the missing values in these features.

2. Methods

Annie

The goal of this project is to analyze the relationship between socioeconomic and demographic factors and the rate of violent crimes across different US communities. We aim to develop models that estimate the ViolentCrimesPerPop variable (the rate of violent crimes per population) based on specific community-level aspects. We strive to identify which features, socioeconomic and demographic factors contribute the most to violent crimes within different communities. We will use four different models for our project: linear regression, K-Nearest Neighbor regression, K-Means clustering, and a decision tree model.

For each of our models, we will need to prepare the data by cleaning data, handling missing values, transforming data types, and encoding categorical values. We will then complete an exploratory data analysis to understand the scale of our data, patterns, relationships, and any potential significant outliers. In this step, we will create visualizations and plots to demonstrate any major features and patterns easily. For the linear regression, K-Nearest Neighbor, and the decision tree models, the data will be split into training and testing sets (80% training and 20% testing). Since K-Means clustering does not predict a target variable, splitting the data is unnecessary. The next step is model training, which involves fitting the model to the training data. For linear regression, this entails minimizing the sum of squared errors between predicted and actual values. For K-Means, the algorithm iteratively assigns points to the nearest centroid and updates the centroids until convergence. For decision trees, the algorithm will recursively split features to reduce variance, producing terminal nodes that provide predictions for the future.

To produce the exploratory analysis and model each algorithm, various libraries must be utilized within Python to ensure the reliability and validity of data modeling. Firstly, matplotlib is used to visualize histograms, line plots, and heat maps. It also ensures that data visualizations can be correctly labelled, titled, and customized to effectively encapsulate the correlation between different data variables. An extension for exploratory analysis is the usage of the seaborn package, which allows for further visualization. To make the algorithmic models, the scikit library is imperative. From the library sklearn.linear_model, LinearRegression is imported to create the linear regression model. From the library sklearn.neighbors, KNeighborsClassifier, and KNeighborsRegressor are imported for KNN regression. From the library sklearn.cluster, KMeans is imported to help with the kMC algorithm. Lastly, from the library sklearn.tree, DecisionTreeClassifier, and DecisionTreeRegressor are imported for decision tree modeling. Harry

2.1. Linear Regression:

The linear regression model helps quantify how an independent variable (X) or variables affect a dependent variable (Y). It is utilized by assuming the interaction between variables is a linear relationship. By calculating a slope and intercept coefficient, as well as an error residual, once trained, the model can provide a predicted Y value from a new given X value. In our project, we use this algorithm to estimate the ViolentCrimesPerPop variable, which in this case is our dependent (Y) variable. By applying the linear regression model to specific demographic variables (X) as mentioned above, we can predict the quantity of violent crimes per population and determine how these variables impact the probability of violent crime in specific populations compared to others.

Analyzing a linear regression model, the R^2 value, which ranges from 0.00 to 1.00, provides insight into how well the regression line fits the data. It is calculated by dividing the variance explained by the model by the total variance within the data. A value closer to 1 indicates a higher likelihood of correlation between variables X and Y. Another critical metric is the root mean squared error (RMSE), which quantifies how far the model's predicted values fall from the actual observed values. Overall, it helps to create a more applicable analysis to real-world data by quantifying the accuracy of the model. Since the goal of linear regression is to minimize the SSE, using first-order conditions for optimization can also be a way to validate the model. Other options to expand the range of variables that the model can use to explain the variation in the data are maxmin normalization, z-score normalization, logarithmic transformation, or inverse hyperbolic sine transformation.

Annie

2.2. K-Nearest Neighbor:

The KNN regression algorithm predicts continuous numerical values by using the target values of similar data points. This algorithm computes the distance from the new point to each observation in the training dataset. It then identifies the k closest data points (k nearest neighbors) based on those distances and associates each of them with a known outcome value. Next, the algorithm computes the average of these k nearest outcome values, and this average becomes the predicted value for the new data point. This means that KNN regression predicts outcomes based on similarities among feature values, making it a beneficial tool for predicting quantities based on similar observations. This algorithm is valuable to our project as it enables us to predict the rate of violent crimes in a community by analyzing rates in comparable communities. These similarities are determined by factors such as population size, socioeconomic status, and demographic characteristics.

Harry

One important part of model validation for KNN regression is feature normalization and scaling. Since the distances in KNN depend on the size and scale of the variables, scaling one variable while leaving others unchanged often changes predictions. One solution is to use MaxMin normalization to reduce the effect of a change in values on the model's performance. The KNN regression is validated using a training and test set, where some data is used for model training and the remaining data serves as a test to assess the model's closeness to the actual data points. These individually compute residuals, which come together as a sum of squared residuals (SSE), showing how well the model fits the dataset. These also help quantify the discrepancy between the i-th prediction and the actual value.

Annie

2.3. K-Means Clustering:

Given the 1,994 instances in our dataset, the K Means clustering algorithm is particularly useful as it can handle very large datasets. The K-Means clustering algorithm starts with initialization and randomly selects k points as the centroids. Then, it calculates the distance of each observation to each centroid, assigns each data point to the closest centroid, and updates the centroid's value to the average of all assigned observations. These steps are repeated until convergence is reached. Although this algorithm does not target the ViolentCrimesPerPop variable, it can help to identify patterns and groupings in the data. By grouping data points with similar characteristics or values, we can determine the most similar communities. This will help us better understand crime patterns and how the demographic and socioeconomic profiles of communities affect crime rates. By clustering similar communities, we can broaden our insights to compare the average violent crime rate by cluster, instead of by each community individually.

Harry

Similar to KNN model validation, K-Means also benefits from feature normalization to ensure that all dimensions are similar in magnitude. MaxMin normalization can also be applied to enhance the model's validity. The K-Means clustering algorithm can be validated by computing the sum of squared error (SSE), which measures how well the model fits the data. kMC tries to minimize the SSE because this means that the clusters are tighter. A scree plot is an important tool for determining the number of clusters (k) that are needed for the most efficient kMC fit. To form a scree plot, the SSE is plotted against the number of clusters (k), allowing us to visualize how adding or removing a cluster within the dataset affects the overall fit measure. To determine the optimal k value, the elbow of the scree plot offers valuable insight. When visualizing the SSE at k-1, k, and then k+1, there should be a significant drop. Once the drop is not as substantial, this signals there is no marginal benefit of adding an additional cluster. Therefore, increasing k further past this point would not improve the fit. In the event that there is no elbow on the scree plot, there are no discrete clusters present within the dataset. Another validation method involves examining the group statistics of the clusters to describe their condition based on their assignment. This can be done by using the .groupby('g_hat').describe() method, which organizes summary statistics for each feature of the cluster. This would provide a better understanding of how the clusters differ from each other.

Annie

2.4. Decision Tree:

A decision tree model splits the data to minimize discrepancies in outcomes within groups. This model consists of a set of decision nodes that represent decision points, a set of edges that represent data-driven choices, and a set of terminal nodes or outcomes at the bottom of the tree representing predictions. This method aims to build a predictive tree for data and future outcomes. This model predicts ViolentCrimesPerPop by analyzing community-level socioeconomic and demographic factors through the analysis of similar dataset groups. At each decision node, the algorithm chooses the feature and split that most reduces variance. This process continues until terminal nodes are reached, simultaneously creating branches that represent decisions that group homogeneous communities. For categorical variables, Gini impurity can be used to evaluate the homogeneity of the split node and the effectiveness of the split. However, for numeric variables, the aim of minimizing SSE can be an effective way to evaluate the splits.

Harry

Decision tree models can be defined by their tendency to be classified as overfitting or underfitting as an overall model. If a tree has too many branches, it ends up providing specific data to which few of the data sets refer. This indicates that the model has overfit the dataset, making it less applicable to new data points due to its specificity. On the other hand, if a model has insufficient branches and the data is too broad, it can be said to have underfit the dataset. As a result, a new data point is unlikely to be correctly fitted because of the model's broad nature. It is crucial to strike a balance between overfitting and underfitting to ensure that models accurately fit the data and, consequently, can effectively predict future data. There are multiple methods that can be used for decision trees to avoid overfitting and underfitting. Truncating the tree to limit its depth, setting a lower bound on the impurity for a terminal node, and limiting the number of cases at a terminal node can help prevent overfitting. Whereas, programming the tree to avoid splits that make the outcome populations too pure would help prevent underfitting.

2.5. Model Validation

To evaluate model performance, we implemented model validation procedures across all the predictive approaches described above. For the supervised learning models, including linear regression, K-nearest neighbors (KNN), and decision trees, the dataset was split into training and testing subsets, using a common 80:20 method. This approach allowed assessment of whether patterns learned during training could accurately predict unseen data. For each model, error metrics such as root mean squared error (RMSE) and an R^2 value were computed to quantify the accuracy of the predictive model. In addition, multiple values of k were tested in the KNN model to examine change in neighborhood size, while the decision tree was evaluated for depth and complexity. Because K-means clustering is an unsupervised method, model validation was performed on a cluster basis. Collectively, these procedures provided a consistent framework for assessing performance and determining the reliability of each model.

Annie

2.6. Model Comparison

Each model contributes to a different objective within our project. The linear regression model predicts violent crime rates by assuming a linear relationship between different features and the target variable. This model shows how demographic and socioeconomic features influence crime rates. The K-Nearest Neighbor model also predicts crime rates, but does so by averaging the violent crime rates of similar communities. This captures the prediction of violent crime rates through the similarity of communities, rather than through linear relationships. In contrast, the K-Means Clustering model groups communities into clusters based on similar features. This model does not directly predict crime rates, but it does reveal hidden feature patterns and compares the average crime rate by community clusters, rather than by individual communities. Finally, the decision tree regression model predicts violent crime rates by splitting the data at certain thresholds that best reduce variance. This produces insights into which demographic or socioeconomic factors most strongly differentiate violent crime rates.

3. Results

Harry

Before modeling, distributions of race, age, and other numerical variables were plotted to guide necessary transformations. Race features are skewed (racepctblack, racePctAsian, racePctHisp low; racePctWhite high, Figure 1), while age features are more centrally clustered (Figure 2).

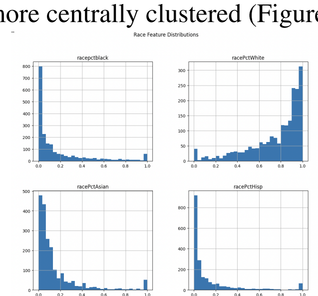

Figure 1. Race Feature Distributions

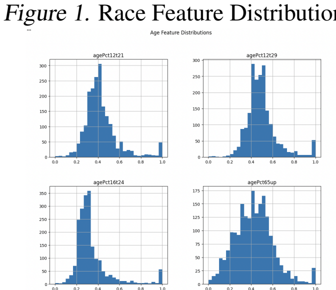

Figure 2. Age Feature Distributions

We then created simple visualizations to explore potential patterns and relationships between demographic features and our target variable, ViolentCrimesPerPop. Figure 3 displays scatterplots of ViolentCrimesPerPop against race-related features, while Figure 4 shows the same for age-related features.

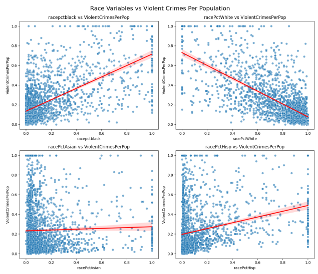

Figure 3. Race vs. ViolentCrimesPerPop

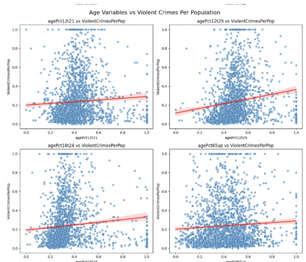

Figure 4. Age vs. ViolentCrimesPerPop

Annie

Appendix A shows scatterplots of ViolentCrimesPerPop against socioeconomic, law enforcement, and geographic variables, which display high variability. This suggests that multiple features, rather than any single variable, contribute to predicting crime rates. This is where the strengths and limitations of different models are important to understand and study when analyzing how a model predicts the crime rate and which variables it uses.

Correlations of numeric features with ViolentCrimesPerPop were calculated and ranked to identify the most influential predictors and establish a baseline for model comparison. The correlations are presented in Table 1.

Table 1. Feature Correlations with ViolentCrimesPerPop

Feature	Correlation
ViolentCrimesPerPop	1.000
racepctblack	0.631
PctPopUnderPov	0.522
PctPersDenseHous	0.456
NumInShelters	0.370
population	0.367
racePctHisp	0.203
PopDens	0.201
LandArea	0.197
PctForeignBorn	0.104
agePct12t29	0.059
agePct16t24	0.029
medIncome	-0.209
agePct65up	-0.067
agePct12t21	-0.080
racePctAsian	-0.213
householsize	-0.235
state	-0.311
racePctWhite	-0.685

3.1. Linear Regression Model

The analysis began with a linear regression to model the relationship between socioeconomic and demographic factors and ViolentCrimesPerPop. This model provides coefficients for each feature, indicating their contribution to the prediction. The first regression used a least absolute shrinkage and selection operator (LASSO) hedonic model, including one-hot-encoded state indicators. Table 2 presents the resulting coefficients, showing the strength and direction of the association of each feature with ViolentCrimesPerPop. In particular, Mississippi appears as the state with the strongest negative correlation with the crime rate.

Table 2. Feature Correlations with Lasso Coefficient

	Feature	LASSO Coefficient
0	MISSISSIPPI	-0.246149
1	FLORIDA	0.161913
2	SOUTH CAROLINA	0.118021
3	MASSACHUSETTS	0.116139
4	MARYLAND	0.106675
5	CALIFORNIA	0.100682
6	racepctblack	0.092495
7	VIRGINIA	-0.092302
8	NEW MEXICO	0.082628
9	GEORGIA	-0.080927

Other states included in the analysis generally show a positive correlation with the crime rate. Among the top ten features, racepctblack stands out, showing a potential positive linear relationship, though the correlation is not strong. This LASSO model helps interpret linear relationships and identify factors that increase or decrease crime rates, with corresponding training and testing root mean squared error (RMSE) and R values. An additional model incorporating principal component analysis (PCA), with the same one-hot encoding for states, was also fitted, producing updated training and testing RMSE and R^2 values. Table 3 summarizes the performance of the two linear modeling approaches in predicting ViolentCrimesPerPop.

Harry

Table 3. Training and Testing RMSE and R^2 for Each Model

Model	Train RMSE	Test RMSE	Train R^2	Test R^2
LASSO	0.1350	0.1277	0.6724	0.6595
PCA	0.1492	0.1359	0.6002	0.6145

In Table 3, it is evident that the LASSO hedonic model has a lower RMSE and a higher R^2 on both the training and the testing sets compared to the PCA model. This suggests that the LASSO model allows for more accurate predictions of ViolentCrimesPerPop.

While linear regression is helpful in identifying linear relationships, it automatically assumes that all features are linearly related to ViolentCrimesPerPop. The model is also sensitive to outliers, which can disproportionately influence the line of best fit. As a result, the linear regression results from this dataset may not fully capture the complexity of all features and their effects on the predicted crime rate.

Annie

3.2. K-Nearest Neighbor Model

The K-Nearest Neighbors (KNN) model predicts ViolentCrimesPerPop by averaging the values of the most similar communities based on socioeconomic and demographic factors. Unlike linear regression, KNN does not assume a linear relationship between features and the target. Figure 5 and Figure 6 show scatterplots for k = 3 and k = 300, illustrating that as k increases the predictions do not drastically change. Scatterplots for the intermediate k values are provided in Appendix B.

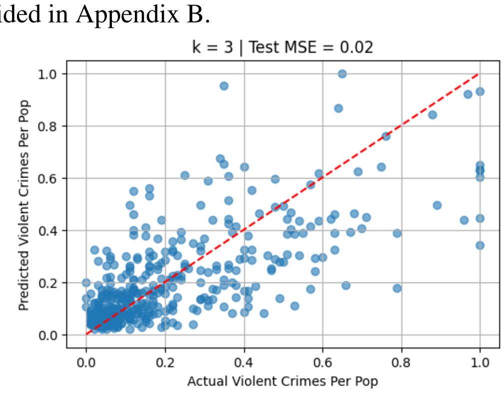

Figure 5. k = 3

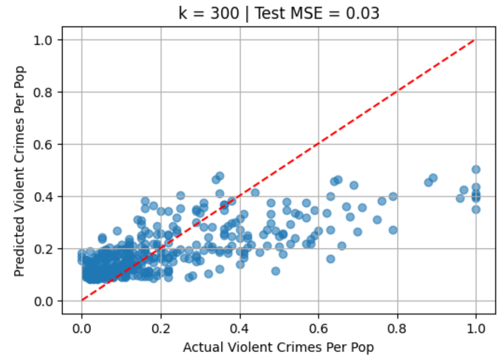

Figure 6. k = 300

Across all tested values of k, the KNN regression model produces nearly identical test MSE values. This is reflected in the scatterplots, which show very similar prediction patterns for each k-value. The target variable, ViolentCrimesPerPop, has low variance, with most data points clustered in a narrow range. As a result, KNN predictions are largely similar to the overall mean, making the model relatively insensitive to the choice of k. These results suggest that KNN may not be an effective approach for this dataset, as using nearest neighbors does not substantially improve predictions.

3.3. K-Means Clustering Model

Harry

Another model that uses a k value is K-means clustering, which helps identify patterns and group communities. This model clusters communities based on similarities in socioeconomic and demographic factors. The number of clusters was selected using an elbow scree plot, shown in Figure 7.

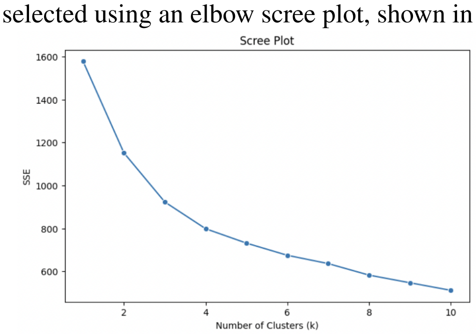

Figure 7. Scree Plot to Determine Optimal k Value

Based on the elbow of the scree plot, k = 3 was chosen as the optimal number of clusters for the K-means model.

Scatterplots in Figure 8 visually illustrate how the chosen k value (k = 3) captures meaningful patterns among the communities compared to other k values.

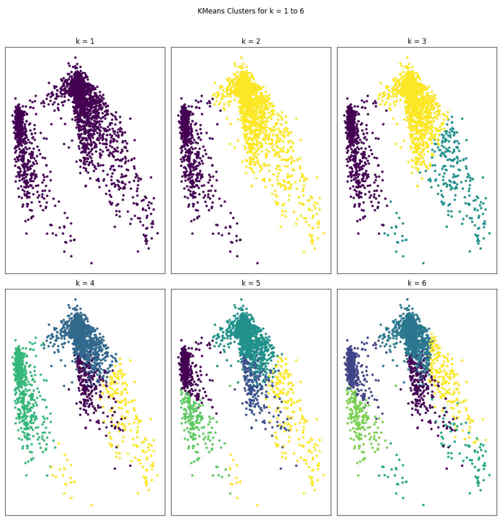

Figure 8. Clusters of K-Means from k = 1 to k = 6

Figure 8 shows k = 3 balances underfitting and overfitting, capturing meaningful community patterns without excessive segmentation.

Table 4. Mean Socioeconomic and Demographic Features by K-Means Cluster (k = 3)

Cluster	population	householdsize	racePctBlack	racePctWhite	racePctAsian	racePctHisp
0	0.01	0.45	0.19	0.78	0.07	0.11
1	0.15	0.62	0.23	0.82	0.05	0.48
2	0.86	0.44	0.16	0.81	0.14	0.39
	agePct12t21	**agePct12t29**	**agePct16t24**	**agePct65up**	**PctPersDenseHous**	**medIncome**
0	0.11	0.51	0.37	0.46	0.16	0.91
1	0.48	0.57	0.30	0.42	0.21	0.97
2	0.39	0.31	0.33	0.43	0.61	0.92
	PctPopUnderPov	**PctForeignBorn**	**PolicPerBudg**	**LandArea**	**PopDens**	**PolicePerPop**
0	0.22	0.29	0.08	0.05	0.13	0.22
1	0.30	0.69	0.15	0.07	0.07	0.22
2	0.33	0.18	0.07	0.07	0.22	0.22
	LemasTotalReq	**NumInShelters**				
0	0.12	0.10				
1	0.21	0.08				
2	0.09	0.03				

Table 4 summarizes the mean socioeconomic and demographic features for each cluster at k = 3, highlighting factors associated with ViolentCrimesPerPop.

We then calculated the mean ViolentCrimesPerPop for each cluster to explore which community characteristics are associated with higher crime rates. Cluster 1 had the highest mean at 0.468, followed by Cluster 0 at 0.214, and Cluster 2 at 0.179.

A key limitation of the K-means model is its unsupervised nature, as it does not directly predict ViolentCrimesPerPop. Additionally, the choice of k clusters can be subjective, and different values of k imply different assumptions about the dataset and the community groupings. Finally, the resulting clusters may not perfectly correspond to crime outcomes, limiting their usefulness for predicting ViolentCrimesPerPop.

3.4. Decision Tree Model

Annie

The final model applied to this dataset was a decision tree, which can capture non-linear relationships and feature interactions while identifying important factors for predicting violent crime rates. The decision tree model is shown in Appendix C.

The decision tree model achieved a test R^2 of 0.566 and an RMSE of 0.144, indicating that this model has moderate predictive performance. Residual analysis in Figure 10 shows that the residuals are mostly centered around zero, which implies that the model can capture most of this.

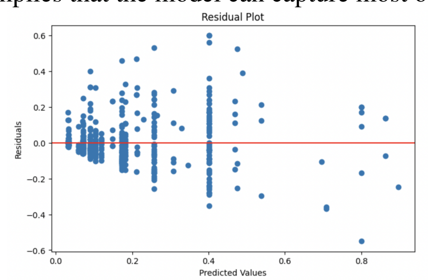

Figure 9. Residual Plot

Feature importance analysis in Table 5 shows the most influential predictors on the ViolentCrimesPerPop.

Table 5. Feature Importance Analysis

Feature	Importance
racePctWhite	0.617153
PctPopUnderPov	0.143874
PolicPerPop	0.052035
state_12	0.030132
racepctblack	0.028939
householdsize	0.018898
community	0.013642
state_28	0.010015
state_39	0.009711

Variables, such as racePctWhite, PctPopUnderPov, PolicePerPop, and racePctBlack were the most predictive demographic and socioeconomic features of the target variable. Therefore, the decision tree model helps to provide insight into which socioeconomic and demographic factors are most associated with violent crime rates in communities.

4. Conclusion

Using our dataset, we visualized and modeled United States community data alongside crime statistics to examine which factors contributed most to our target variable, violent crimes per population (ViolentCrimesPerPop). Our goal was to analyze the relationship between socioeconomic and demographic factors and violent crime rates in different

United States communities.

During the exploratory data analysis, we discovered key patterns and relationships in our dataset. Race features demonstrated skewed distributions (racePctBlack, racePctAsian, racePctHisp low; racePctWhite high), and age features were more centrally clustered. Visualizations created to examine potential patterns and relationships between demographic features, and ViolentCrimesPerPop showed high variability, suggesting that there were multiple contributing factors to high crime rates. Finally, correlation analysis of numeric features and ViolentCrimesPerPop showed strong positive correlations with features, such as percentage of the population who is Black (racePctBlack), percentage of population under the poverty line (PctPopUnderPov), and percentage of the population in density house (PctPersDenseHous). Whereas, it showed a strong negative correlation between features, such as percentage of the population who is white (racePctWhite), state, size of the household (householdsize), and percentage of the population who is Asian (racePctAsian).

Our four models aimed to further analyze the relationship between these socioeconomic and demographic factors and our target variable, ViolentCrimesPerPop. The linear regression model used two different regression mechanisms: LASSO and PCA. The LASSO model had lower training and testing RMSE and higher R^2, indicating it made more accurate predictions than the PCA model. The LASSO regression with one-hot encoded states identified feature contributions, such as states including Mississippi, Virginia, and Georgia, having negative relationships with ViolentCrimesPerPop. Whereas, states such as Florida, South Carolina, Massachusetts, Maryland, California, and New Mexico all had positive relationships with ViolentCrimesPerPop. Notably, racePctBlack stood out as the only racial feature in the feature correlation analysis, showing a positive linear relationship with the target variable. We identified several limitations in our linear regression model: it assumes linear relationships and is sensitive to outliers.

Our next model, K-Nearest Neighbors (KNN), predicted ViolentCrimesPerPop by averaging values of similar communities. We tested five k values (3, 10, 25, 50, 100, and 300) and found that the predictions were largely similar across k values, due to low variance in the target variable. Therefore, the KNN was insensitive to the choice of k, showing minimal improvement over mean-based prediction and was therefore not a very effective model for our dataset.

The third model, K-Means Clustering, clustered communities based on socioeconomic and demographic similarities. We determined our optimal k-value (k=3) through the elbow scree plot. Cluster 1 had the highest average ViolentCrimesPerPop (0.468), indicating the highest violent crime rate among the three clusters. This cluster contained communities with higher poverty rates (PctPopUnderPov) and higher proportions of foreign-born residents (PctForeignBorn). Cluster 0 had an average ViolentCrimesPerPop of 0.214. This cluster consisted of communities with a lower population density (population/PopDens). Finally, cluster 2 had the lowest average ViolentCrimesPerPop at 0.179. Cluster 2 had a high PctPersDenseHous and population, but otherwise had fairly moderate demographics, with feature means generally between those of clusters 0 and 1.

Finally, the decision tree model captured non-linear relationships and feature interactions. The model had a moderate predictive performance (Test R^2 = 0.566, RMSE = 0.144) and residuals mostly centered around zero, indicating a reasonable fit. The top predictors for this model were racePctWhite, PctPopUnderPocv, PolicPerPop, and racePctBlack.

Although our models did not pinpoint a single definitive set Harry
of features contributing to high crime rates in United States communities, the dataset includes a range of demographic and socioeconomic features. This provides a holistic view of the factors that contribute to violent crime rates and shows that it is not just one feature that affects our target variable. Furthermore, using four different models helped us address gaps and limitations across them, allowing us to capture both linear and nonlinear relationships. The LASSO regression identified linear relationships and provided correlation coefficients for features, enabling us to interpret which factors have positive or negative relationships with our target variable early on. Decision tree models then identified interactions between variables that linear models might have missed, while K-means clustering revealed community groupings and patterns. Together, exploratory data analysis and these model approaches offered complimentary perspectives, strengthening our understanding of how community characteristics relate to higher crime rates.

Outside of this project, it could be important to expand analysis by incorporating geographical modeling to demonstrate how spatial patterns in community features affect overall ViolentCrimesPerPop. Other models could work to integrate data from beyond the 1990s. Finding census data from more recent years could help show how possible early-warning indicators for communities might have changed over time, as well as how the ViolentCrimesPerPop changed based on these indicators over time. Lastly, it could be interesting to investigate policy and how it may be affecting the statistics in these datasets. Furthermore it would be imperative to investigate if these indicators are being properly, or poorly used in resource allocation, policing strategies, or community development.

Both

5. References

Redmond, M. (2002). Communities and Crime [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C53W3X.

6. Appendix

6.1. Appendix A: Scatterplots of ViolentCrimesPerPop Across Socioeconomic, Crime, and Geographic Features

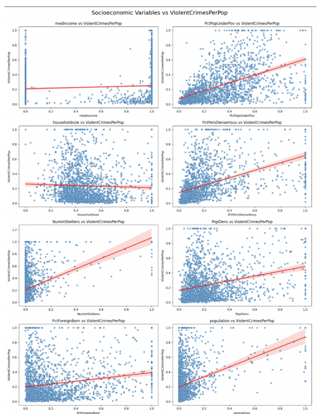

Figure 10. Socioeconomic Features vs. ViolentCrimesPerPop

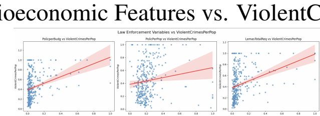

Figure 11. Crime Features vs. ViolentCrimesPerPop

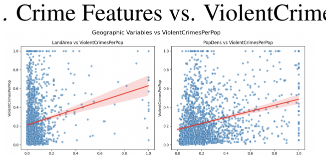

Figure 12. Geographic Features vs. ViolentCrimesPerPop

6.2. Appendix B: Scatterplots of KNN Predictions for Intermediate k Values

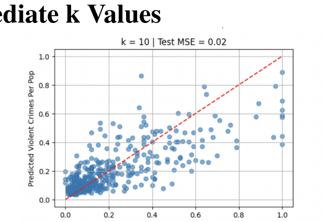

Figure 13. k=10

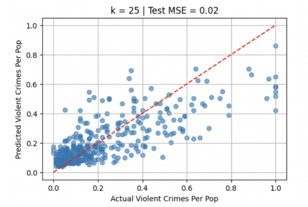

Figure 14. k=25

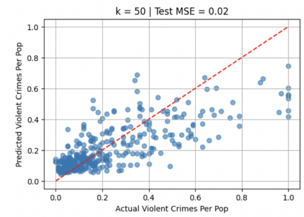

Figure 15. k=50

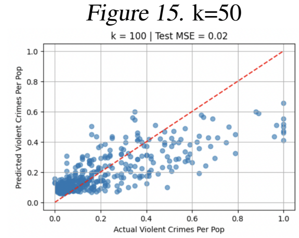

Figure 16. k=100

6.3. Appendix C: Decision Tree

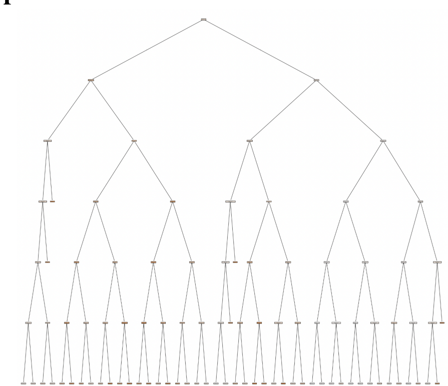

Figure 17. Decision Tree