Visualizing and Modeling United States Community Data against Crime Statistics

Annie Edmonds Harry Shield

1. Data Description

Annie

Our project utilizes a dataset that focuses on quantifying community diagnostics and crime statistics across communities in the United States. This 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.

Harry

In total, the dataset contains 127 variables, making data cleaning a challenging task. To manage this, selecting a subset of 25 features that includes 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:

- 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.

Annie

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.

Harry

The predictive values are represented on a normalized scale 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 visualize 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. Annie 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.