Crime in US Communities

Team 51 Progress Report

JENNY CHEN

JONATHAN CASCIOLI

MAKRAND KALYANKAR

MEIR WEINER

XIAOLU SU

Overview

The reduction of crime in US communities, especially violent crime, is something that every local government should prioritize. Our project explores the factors that most influence the rate of violent crime in order to help local governments allocate funding and enact policies to create the safer communities.

Progress

Data Selection

As part of the project, we started out with crime data as our base dataset. We had intended to enrich the data with features that communities deploy that may have effect on number of crime incidents occurring in the community. This required us to collect features available through various agencies and publicly available datasets.

Planned Parenthood:

We were unable to find planned parenthood data from the specific year of our main crime dataset (1995). We gathered the data by scraping the planned parenthood website. This marks an area where our project, as mentioned in the proposal, must dive into the scope of being theoretical. More on this later in the challenges section. A simple binary variable (0,1) made up our planned parenthood data.

Education Data:

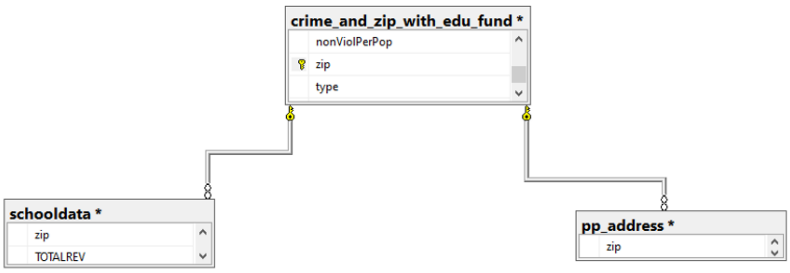
Our Educational data source was gathered from the National Center for Education Statistics (NCES) from 1995-1996 and contained information on:

School Information	*Finance*	*State Level Sum:*	*School Counts*	*Grade & Grad counts*
School District School Codes Student Counts etc	Financial Aid Revenue Taxes Employee Salaries Employee Benefits etc	Financial State Summary	Dropouts Ethnicity Counts School Staffing	Grade (1-12) Counts Teacher Counts Librarians School Admins Etc

This marks an area where due to time constraints we were unable to incorporate all of the data we thought would be valuable. We joined financial ratios of federal, state, and local aid contributions into our main dataset as we thought these held the most promise and hypothesized would have the greatest explanatory strength.

Data Cleaning:

Our additional data sets were reported on the zip code level or some sort of government statistical code (NCES). The base crime data was reported at community level with non-standardize community name. A significant amount of time was spent just enabling the data to be joined to our main dataset. The team enriched this main dataset with Zip code information based on the existing State-Community combination. Once Zip codes were associated with communities, we could experiment with relevant feature data.



We followed industry best practices in cleaning the data:

- We leveraged RDBMS inherent ability to enforce domain constraints i.e. column data types. This guarantees that the data for the field is of data types we define it to be. Any data not meeting the criteria was type casted.
- Removed any duplicate rows originating from join condition
- Normalized missing data to NULL instead of mixture of NA, “empty string”, NULL so there is a consistent way to deal with them

Further Data Preparation and Exploratory Data Analysis:

Once the team joined the core raw dataset with 2 additional sources, we started an investigation in python as part of our exploratory data analysis. This led to further data preparation. We discovered there were almost 200 features, and 11,516 rows with lots of missing values.

Filtering Features:

First, we went though and manually chose all the values that were relevant to our business problem and ethical to use. For example, we removed all columns related to race. Our remaining explanatory variables made up several categories:

1. Age Demographics
2. Social program / assistance
3. Poverty indicators
4. Education History
5. Education Spending
6. Employment
7. Family Demographics
8. Housing Occupancy
9. Policing

With Violent Crime Rates Per population unit as our response variable, as we had originally posed in or proposal.

We did some additional manipulations and aggregations to get the data to be unique at the community level. Discovering that when we incorporated our zip code data to our original dataset that it created duplication of rows. After the aggregation and cleaning to this point, we held 1886 unique communities / observations.

Feature Standardization:

From here we started a further null value analysis, removing additional features that contained greater than 30% null values.

After the final cleaning and standardization process, we landed on 1,414 observations with 38 explanatory variables.

Statistical Exploration:

With our final features, we performed some light exploratory data analysis. First, we noticed that the distribution of the response variable was right-skewed(fig 1), implying that there are only few communities with huge crime rates.

We then looked at some of the explanatory variables that we hypothesized to be most important, and compared them to our response variable (Violent Crimes per Population Unit). These consisted of:

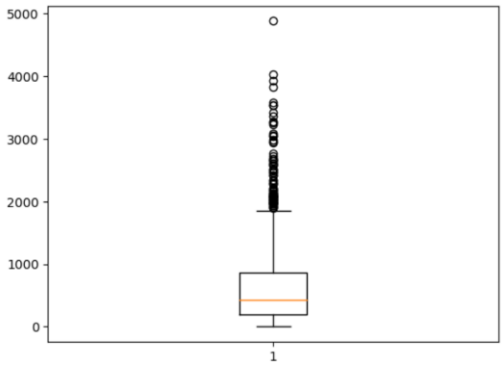


Fig 0: Box and Whisker Plot for Violent Crimes Per Population Unit

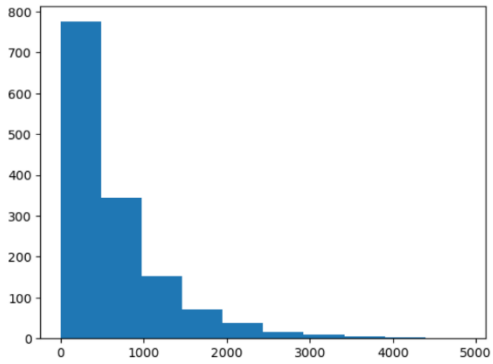


Fig 1: Violent Crimes Per Population Unit Histogram

- % with less than 9th Grade Education (fig 2, fig8)
- Access to Planned Parenthood (fig 3)
- % of population using public assistance (fig 4,fig 9)
- % of homes that are non-vacant (fig 5, fig 10)
- % of population that is unemployed (fig 6, fig 11)
- % of education funding by local government(as opposed to federal or state) (fig 7,fig 12)

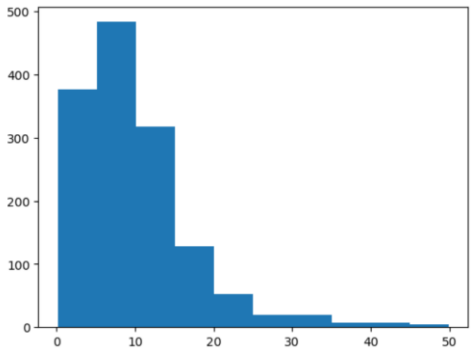


Fig 2: Less than 9th Grade Education Histogram

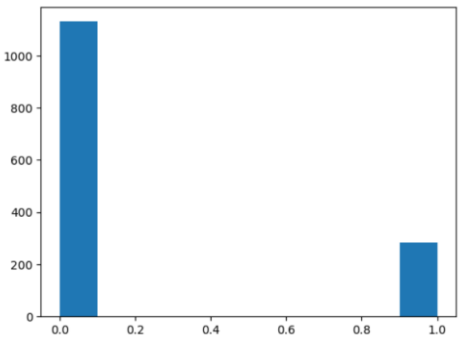


Fig 3: Access to Planned Parenthood Histogram

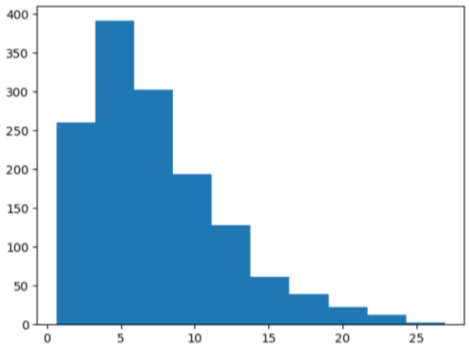


Fig 4: % of population using public assistance Hist.

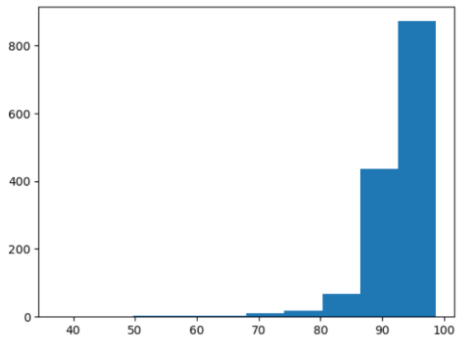


Fig 5: % of homes that are non-vacant Histogram

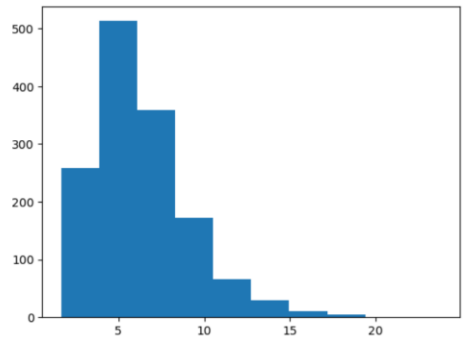


Fig 6: % of population unemployed Histogram

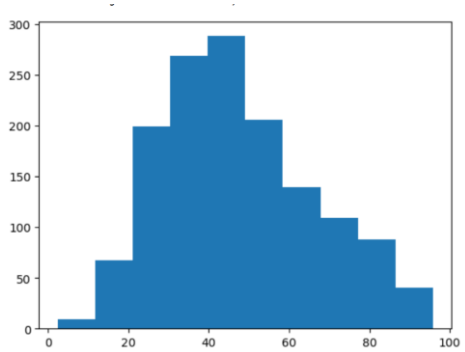


Fig 7: % of Education by Local Government Hist.

**Note: for each of the following figures the format is a scatter plot where:*

- *explanatory variable – x axis vs*
- *response variable Violent Crimes per Population Unit – Y axis*

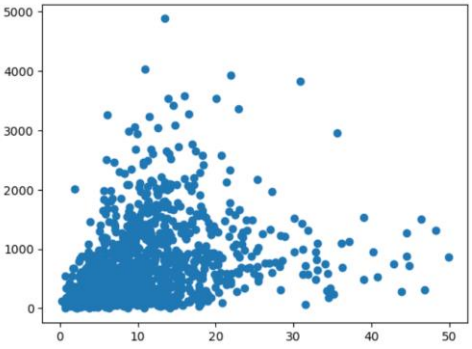


Fig 8: Less than 9th Grade Education

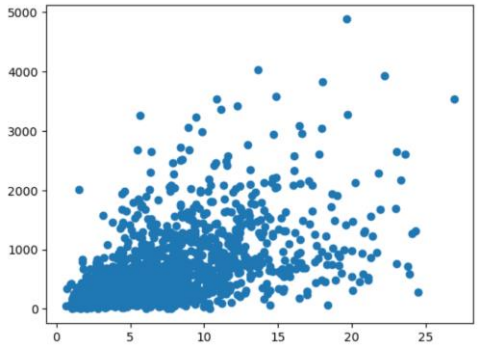


Fig 9: % of population using public assistance

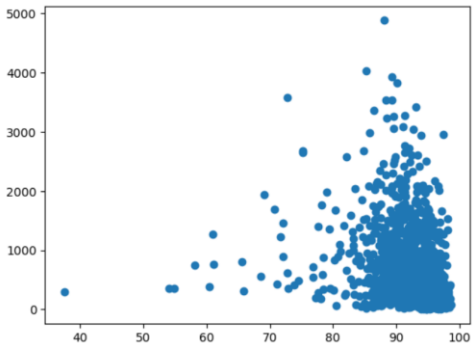


Fig 10: % of homes that are non-vacant

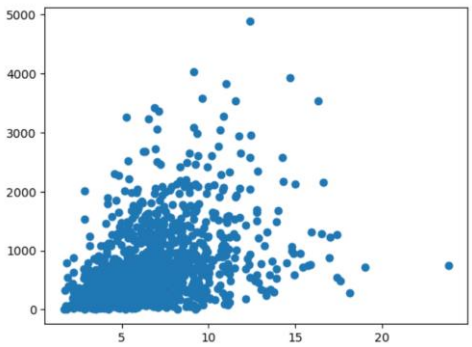


Fig 11: % of population un-employed

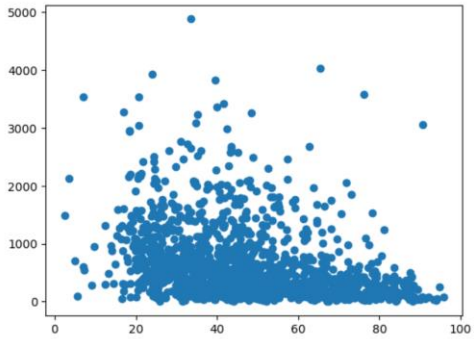


Fig 12: % of Education by Local Government.

Here, our 2 key findings were that (1) many of the features do not have linear or linear- transformable relationships with the response, and (2) there is a decent amount of correlation amongst the explanatory variables themselves. Both findings suggest that a multivariate linear regression may not be the best model for our data.

Summary of Challenges we have faced

Preliminary Data Processing:

Though much of the data that we were interested was publicly available, our team was perhaps overly ambitious in what we wanted to accomplish and include. Though we accomplished a significant amount of quality work, learning about the problem and gathering additional data sources; the data was quite messy. This led to us spending substantial time cleaning and processing the data.

Specifically joining the data and normalization.

Theoretical Limitations:

As mentioned above, we were unable to gather overlapping data from the 1995 timeframe. We cannot quantify the significance of this on our output. However, it must be acknowledged. Specifically, this affected us in terms of zip-codes and planned parenthood data. Our zip-code dataset used to clean and join our data together was modern. We discovered in our research, as one would logically conclude, that overtime zip-codes have been added or decommissioned. Our planned parenthood data was modern as well. As we revealed in our research, in order to understand the effect of social programs on a specific

period of time, you must understand the state of social programs at a times significantly before (such as 18 years) the period in which you are analyzing (source 1). This is due to the impact that such social programs might have had on individuals when they were children, or an earlier point in their life when they were exposed to the program.

Time limitations:

All of these different challenges that we encountered highlighted for the team the real level of detail and effort that must be put into such a complex problem of this nature. Time that based on the course constraints, simply wouldn't allow us to do true justice to the posed problem. Therefore, while we can generate a quality analysis, and practice the appropriate methodology in the confines of our project, our project can only serve as a template for another to dedicate time and gather "hard" real world conclusions.

Remaining Work

We still plan to test a linear regression as our first model, although we do not expect it to perform well in terms of prediction accuracy. As the final dataset still has almost 40 features, we will include variable selection in our analysis. Additionally, we will run a regression tree on the data, with hyperparameter optimization through manual search or a more sophisticated method based on research.

Research (Work Cited)

Our group attempted to conduct background research on our hypothesized variables to better understand the intricacies of the problem.

1. Effect of Legalized Abortion on Crime:
    - https://pricetheory.uchicago.edu/levitt/Papers/DonohueLevittTheImpactOfLegalized2001.pdf
2. Effect of Homeless Rates on Crime
    - https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-9125.2006.00042.x
3. School Finance Reforms on Juvenile Crime Rates
    - https://academic.oup.com/aler/article-abstract/24/1/1/6591238
4. Crime Reducing Effects of Education:
    - https://academic.oup.com/ej/article-abstract/121/552/463/5079723