Caleb Kartchner
Robiul Islam

CS5830-Project2

Introduction

In this project, we explore crime and housing data for Austin, Texas, collected in 2015. The purpose of the analysis is to investigate whether there is a relationship between the prevalence of large households (5+ members) and crime rates across different zip codes and finding correlation between poverty levels and crime rates. Understanding these relationships can help city planners and law enforcement agencies allocate resources more effectively to areas with higher crime potential. The dataset contains 43 columns, including attributes related to housing, reported crimes, economic statistics as many others. The analysis employs statistical methods like Pearson correlations, scatterplots, and t-tests to uncover significant patterns.

Link to presentation slides:
https://docs.google.com/presentation/d/1owF93-qNdV0ssP5Tzu4rOFMP0McrCLlSdEvlGnaz6k4/edit?usp=sharing
Link to project folder: https://github.com/A02276327/CS5830_project2

Dataset

The primary dataset contains crime and housing statistics for Austin, Texas, in 2015. It includes important columns such as zip codes, highest offense descriptions, household sizes, and clearance status of crimes. We focused on analyzing the zip codes, crime types, poverty and household sizes to investigate correlations. Additional data from population densities was also available to compute crime per capita, adding depth to the analysis. This dataset is ideal for exploring community and urban crime patterns in relation to socioeconomic variables.

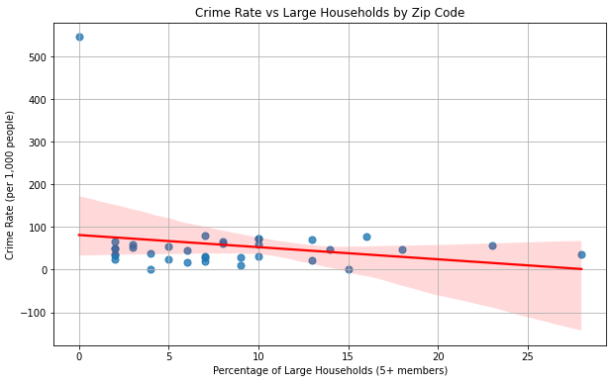
Analysis technique

For the analysis of crime statistics by zip code, we used several key techniques to ensure the results were accurate and meaningful. Population data was incorporated to normalize crime rates, providing a more accurate reflection of crime relative to the number of residents in each area. Data cleansing was crucial, involving the conversion of string values to integers and the removal of symbols such as '%' and '$' to ensure consistency. Scatterplots were employed to visually detect trends in the data, and statistical tests were used to verify the significance of any observed relationships. Lastly, when analyzing the total number of crimes, we considered both the specific time frames and the dataset as a whole to maximize the depth and accuracy of the analysis.

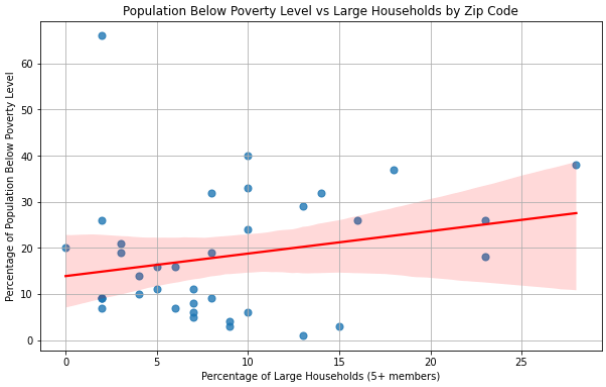
Results

Analysis 1

We tried to find if there was correlation between the percentage of large households (5+ members) with poverty and crime rates. Initially it made sense that larger families would be poorer so areas with higher percentage of large households would be more susceptible to poverty. With increased poverty and the hardships of raising many children we made the assumption that areas with larger households would correlate with higher crime. Looking into this data we got the following scatter plots:



Trend Line: y = -2.8462x + 81.3323
R-squared: 0.0414
Pearson Correlation= -0.20
P-value = 0.24 -> NOT SIGNIFICANT



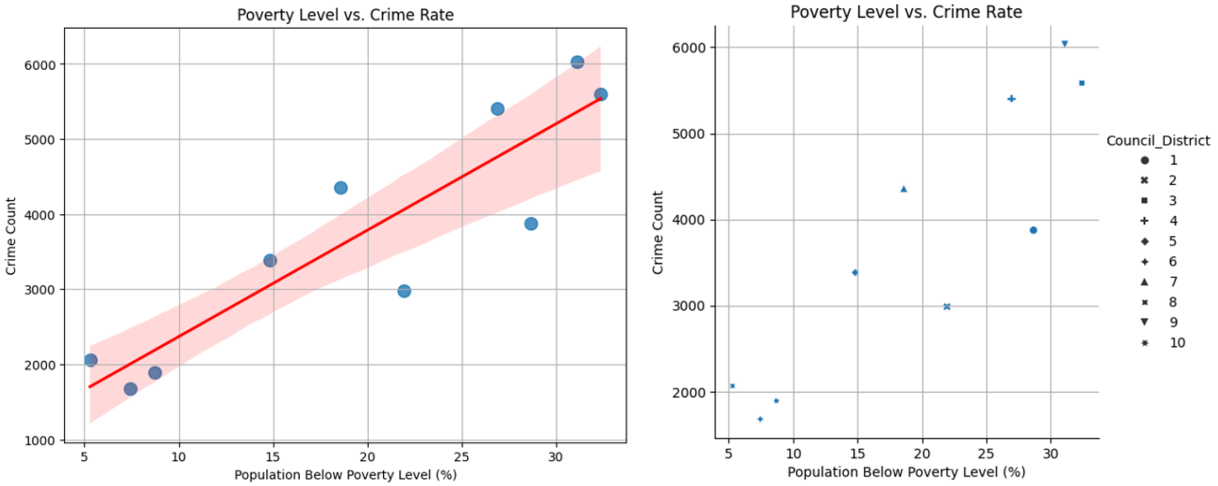
Trend Line: y = 0.4878x + 13.8758
R-squared: 0.0542
Pearson Correlation: 0.2328
P-value: 0.1656 -> NOT SIGNIFICANT
T-test Statistic: -3.7751

The results were surprising in a number of ways. Firstly, there was no statistically significant result meaning that overall, having a large household does not in fact play a large role in crime rates or poverty. In retrospect, this makes a lot of sense as many other factors such as socioeconomics likely play a larger role in crime statistics. The results however still showed at least a partial correlation in larger households contributing to increased poverty levels. Government assistance for larger households could help prevent further poverty increases which as we will prove later in this report, plays a significant role in crime increase. There is also a small correlation with crime and larger households but with the p-value being so high it is highly unlikely that this is a worthy enough problem to entail government intervention. There are more impactful measures that can be introduced.

Analysis 2

In this analysis, it was tried to determine if there is a relationship between the percentage of people living below the poverty level and the crime rate, alongside the influence of council districts. It was considered the average crime rate per council district. As expected, it was found that areas with higher poverty levels tend to have higher crime rates, as individuals often commit crimes to meet their needs. A Pearson correlation test was conducted to examine the linear

relationship between poverty levels and crime rates. The resulting coefficient value of 0.9004431591463176 with a p-value of 8.38e-07 indicates a significant positive correlation. Therefore, it can be concluded that as poverty levels rise in an area, crime rates increase, with more individuals living below the poverty line committing crimes.



Additionally, the scatter plot shows that in some geographical areas, such as districts 6, 8, and 10, where the average poverty level is lower, the crime rate is also lower compared to areas with higher poverty levels. This data could help the city of Austin allocate resources more effectively by identifying areas with high poverty levels that are prone to increased crime. Targeted interventions, such as poverty reduction programs and enhanced policing, can be prioritized in these districts to reduce crime rates and improve community well-being.

Technical

Data Preparation

The dataset required extensive cleaning and formatting to ensure consistency and accuracy in the analysis. This involved converting string values to integers, removing symbols such as '%' and '$', and normalizing crime rates using population data to reflect crime per capita. Additionally, null values were removed to ensure that the regression analysis would not be skewed by incomplete data. Including null data often meant receiving error during statistical significance calculations.

Analysis

The analysis techniques employed were chosen to uncover meaningful patterns and relationships within the dataset. Pearson correlations were used to measure the strength and direction of linear relationships between variables, such as household size, poverty levels, and crime rates. Scatterplots provided a visual representation of these relationships, making it easier to detect trends and outliers. T-tests were conducted to determine the statistical significance of

the observed correlations. These methods are suitable for the dataset as they allow for a comprehensive examination of the relationships between socioeconomic factors and crime rates, which are inherently quantitative.

Analysis process

The analysis process involved several steps to ensure robust and reliable results. Initially, exploratory data analysis was performed to understand the distribution and characteristics of the data. This included generating summary statistics and visualizations. During the analysis, multiple iterations were conducted to refine the models and techniques used. For instance, initial scatterplots and correlation tests revealed no significant relationship between large households and crime rates, prompting a deeper investigation into other factors such as poverty levels. Adjustments were made to account for potential confounding variables, and alternative approaches, such as considering different time frames and normalizing crime rates by population, were explored to enhance the accuracy of the findings. Despite some failed attempts, these iterative adjustments ultimately led to more meaningful insights and conclusions.