Project 2: Crime and Housing

Shylee Oler and Ezri Brimhall

CS5830: Data Science in Practice

Dr. John Edwards

February 2nd, 2023

Introduction

In this project, we focused on the intersection of socioeconomic factors and crime rates in Austin, Texas. Crime statistics and socioeconomic indicators are crucial to understanding community well-being, public safety, and resource allocation. Our dataset comprises crime data for the year 2015 in Austin, Texas, including various criminal offenses, clearance statuses, demographic data, and socioeconomic data for each zip code. Each analysis employs Pearson correlation coefficients or t-tests to uncover potential relationships and disparities. Understanding these relationships is vital for policymakers, law enforcement agencies, and community members to devise targeted interventions, allocate resources efficiently, and foster safer communities.

Our findings suggest intriguing associations between certain socioeconomic factors and crime rates, shedding light on potential areas for intervention and policy focus. For instance, we observe correlations between median household income and crime rates, as well as between unemployment rates and crime. By discovering these patterns, our analysis aims to contribute to informed decision-making for enhancing Austin's public safety and community well-being.

Project slides may be found at the following link:
https://docs.google.com/presentation/d/1XCqX-cUUOByTQ6Tb5bVfVuBbTwpZxaymkZKw_jUqwgc/edit?usp=sharing

A project folder (with a Jupyter notebook containing analyses and charts) may be found here:
https://github.com/ezri-brimhall/cs5830-project2

Dataset

The primary dataset we used for this project was a table of crime statistics for the area of Austin, Texas during the year 2015. This data includes basic information about the crimes such as the type of crime, location information such as zip code, and demographic information for the entire zip code. We also used a supplemental dataset that contained population statistics for each zip code.

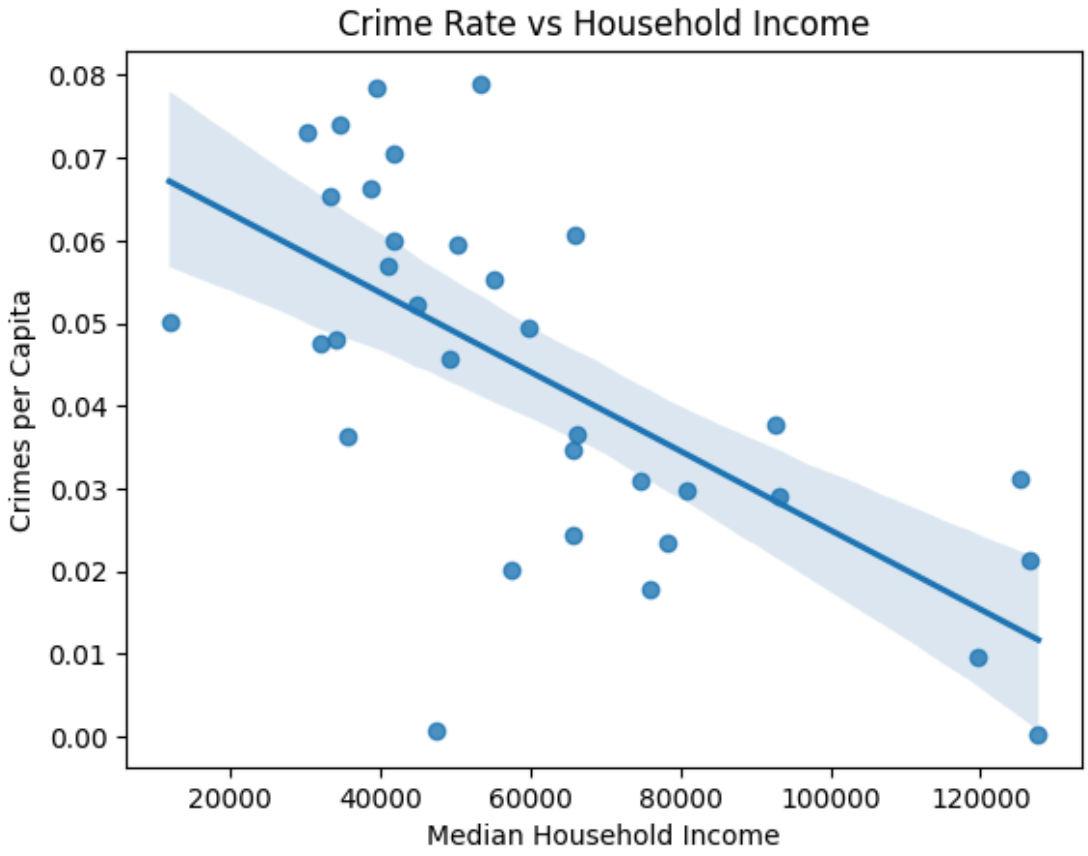
Analysis Technique

Each analysis involved plotting the variables we were interested in on a scatter plot, along with a regression line, to visually inspect the data. Most of the variable combinations we were interested in showed a strong correlation on this scatter plot, which we then confirmed using Pearson correlations. Importantly, there was one zip code which stood out as a major outlier, in that it had an enormous number of reported crimes relative to its population; more than one crime for every two people living there. This zip code was removed from the dataset for all of our analyses as it was the sole outlier, with all other zip codes showing a per-capita crime rate below 1 in 10.

Results

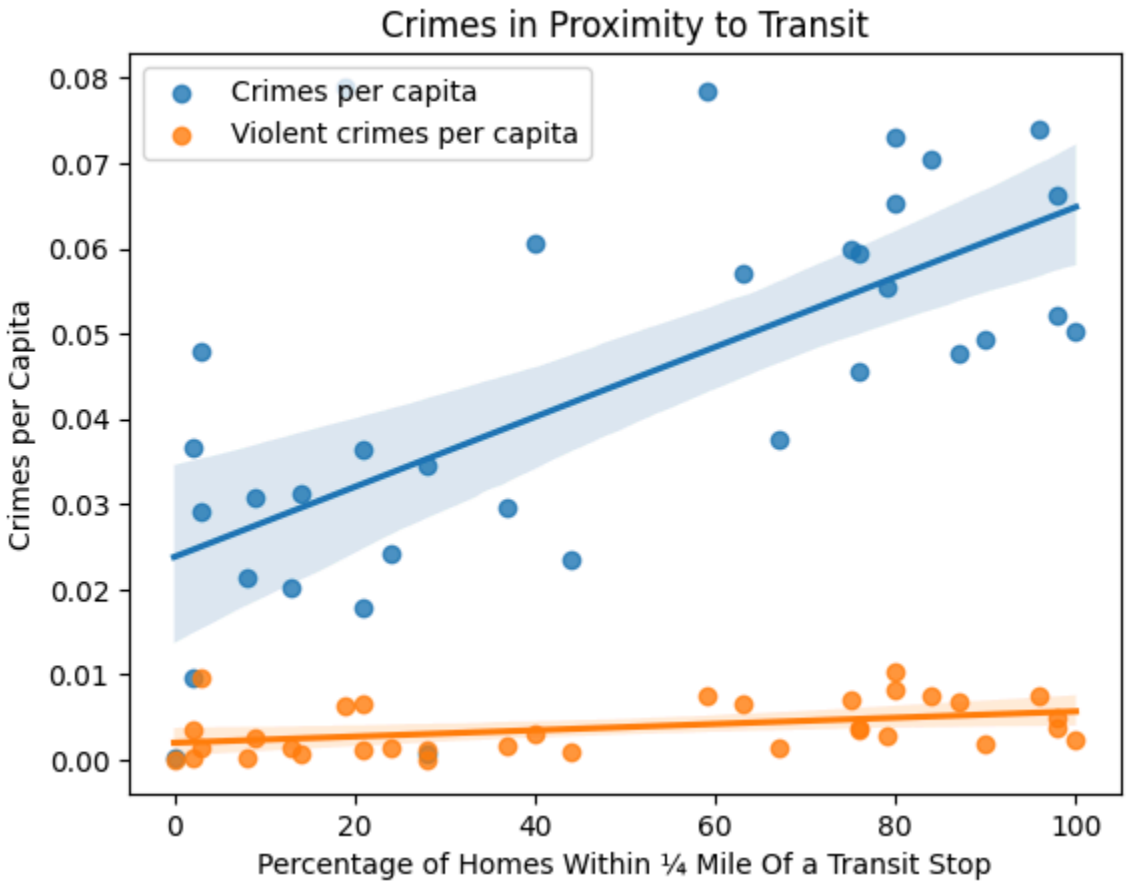
Household Income to Per Capita Crime Rate

This first analysis confirmed that a fairly well-known trend is present in Austin, Texas, and represented in the crime dataset. These variables are very likely correlated, with a Pearson correlation statistic of -0.6584 and a p-value of 2.313×10^{-5}, indicating a very low chance of a coincidence. This is the result we were expecting, as households with less money may need to turn to crime even just to make ends meet. There are likely other factors involved as well, but conclusions involving those factors cannot be drawn from this analysis alone.



Proximity to Transit

This analysis revealed a surprising insight into the relationship between proximity to public transit and crime rates. It shows a strong positive correlation between the percentage of homes within a quarter-mile of a transit stop, and the per-capita crime rates. This correlation was found both for all crimes (a Pearson correlation value of 0.6628 and a p-value of 1.945×10^{-5}) and when restricting to just violent crimes (with a Pearson correlation value of 0.4267 and a p-value of 0.0119). These results could potentially be explained by a number of factors, such as a higher population density near transit, or transit being more common in lower-income neighborhoods. However, we don't believe that it stems from non-violent crimes committed on public transit, as we initially considered, as we were unable to find a correlation between proximity to transit and the percentage of crimes which were violent (p-value of 0.2867).

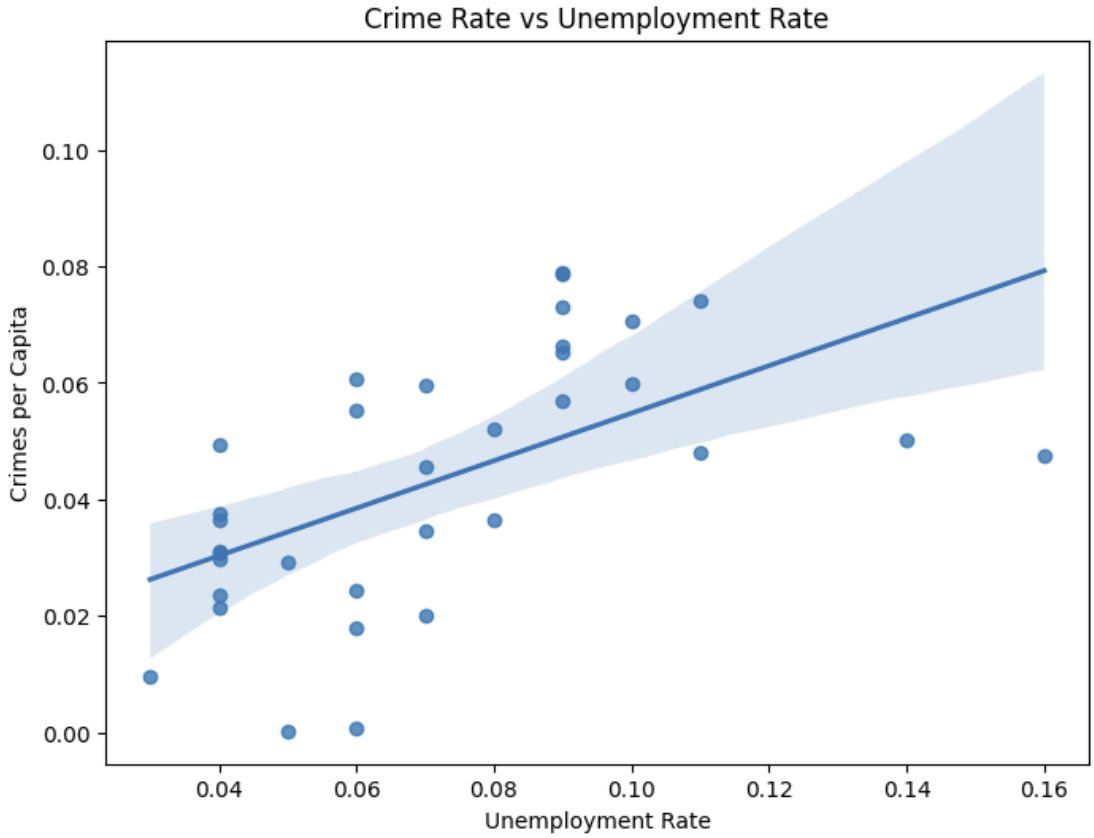


Unemployment Rate to Per Capita Crime Rate

This analysis revealed insights into the relationship between unemployment rates and per capita crime rates in Austin, Texas. The statistical analysis below indicates a significant negative correlation between crime rate and unemployment rate. The mean crime rate is 0.0434 with a standard deviation of 0.0217, while the mean unemployment rate is higher at 0.0721 with a standard deviation of 0.0304. The calculated t-statistic of -4.469 suggests that the observed difference in means between crime and unemployment rates is substantial and unlikely to have occurred by random chance alone, supported by a very low p-value of 3.16e-05. Thus, it can be inferred that as unemployment rates increase, crime rates tend to decrease, suggesting a potential inverse relationship between these two variables. These findings emphasize the need for nuanced and context-specific approaches in tackling crime and fostering safer and more resilient communities.

```
Crime Rate Mean: 0.043403398060387496
Unemployment Rate Mean: 0.07205882352941177
Crime Rate Standard Deviation: 0.02172624097342629
Unemployment Rate Standard Deviation: 0.030429193694164906

T-statistic: -4.468874022899825
P-value: 3.1608465384074884e-05
```

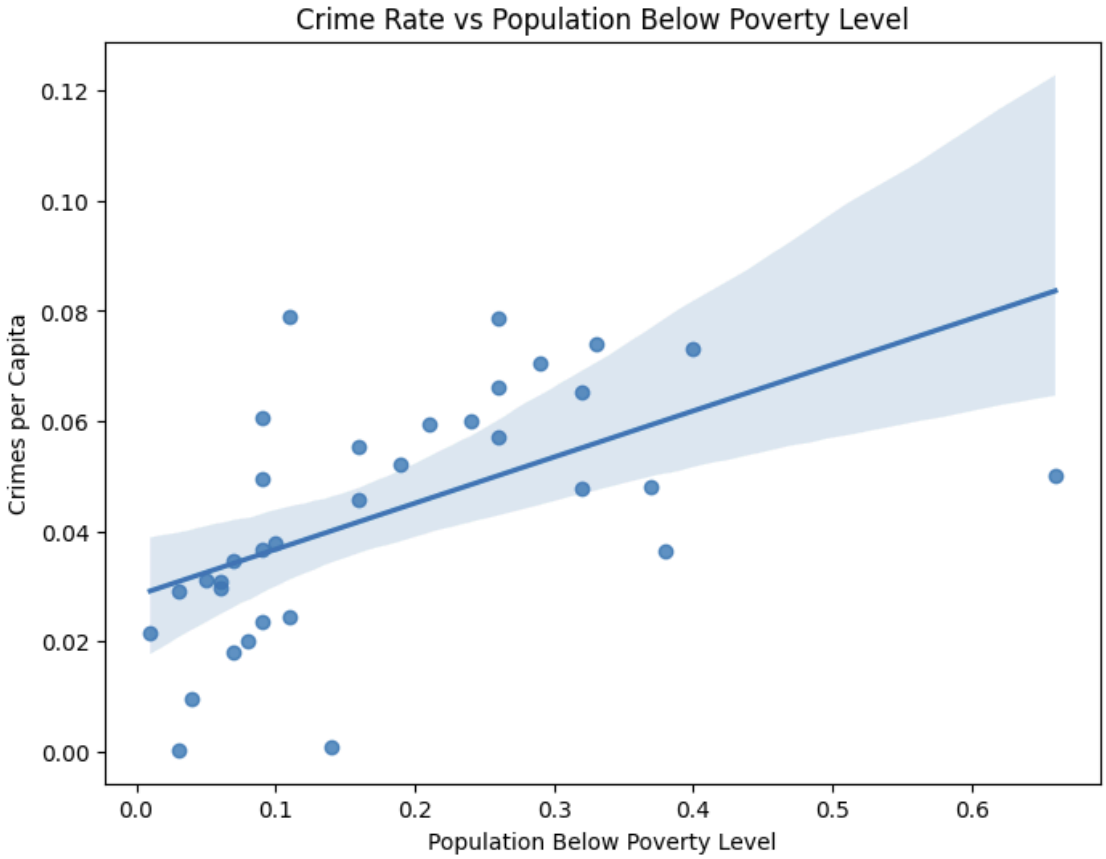


Population Below Poverty Level to Per Capita Crime Rate

The analysis uncovered significant insights regarding the relationship between the population below the poverty level and per capita crime rates in Austin, Texas. The results reveal a significant relationship between crime rate and the proportion of the population living below the poverty level. With a mean crime rate of 0.0434 and a mean poverty level of 0.1803, it indicates that a notable portion of the population is affected by crime, and approximately 18% of the population is living below the poverty line. Crime rates exhibit moderate variability around their mean, while the poverty level shows greater variability across different areas. The t-statistic of -5.51, coupled with an extremely small p-value of 6.37e-07, strongly rejects the null hypothesis and supports the presence of a statistically significant association between higher poverty levels and higher crime rates. Thus, these findings highlight the socioeconomic factors at play in influencing crime rates.

```
Crime Rate Mean: 0.043403398060387496
Population Below Poverty Level Mean: 0.18029411764705885
Crime Rate Standard Deviation: 0.02172624097342629
Population Below Poverty Level Standard Deviation: 0.14322022233456244

T-statistic: -5.510216823244732
P-value: 6.373093062820474e-07
```



Technical

Before conducting the analysis, the dataset underwent several steps to ensure its suitability for analysis. This involved merging two datasets: crime data for Austin in 2015 and demographic information by zip code. Data cleaning included converting population and numeric columns from string to integer or float format, handling missing values, and filtering out outliers. It was also vital to calculate "crimes per capita" to use for all analyses. Additionally, categorical variables were created, and scaling was performed where necessary to ensure the compatibility of variables.

For this analysis, we employed statistical techniques such as Pearson correlation coefficients and t-tests to explore the relationships between socioeconomic variables and crime rates. Pearson correlation is suitable for examining linear relationships between continuous variables, making it ideal for assessing the associations between median household income and crime per capita. Also, associations between homes within a quarter mile of a transit stop and crime per capita T-tests were used to compare means between groups, such as crime per capita and unemployment rates and crime per capita and population below the poverty level, providing insights into potential disparities and their statistical significance. Where T-tests were implemented, we chose to include averages and standard deviations between the data sets to better understand the results. The regression plots generated using the seaborn library visually depict these relationships.

The project began with exploratory data analysis to understand the distribution and relationships between variables. Several iterations of data preprocessing were performed to handle missing values, and outliers, and ensure data integrity. Failed attempts primarily revolved around fine-tuning visualization techniques and ensuring the validity of each analysis. Overall, the iterative nature of the analysis process facilitated a deep understanding of the dataset and yielded valuable insights into the dynamics between socioeconomic factors and crime rates in Austin.