Project 2 Report - Crime & Housing

Nicolas Holden & Caitlin Thaxton

Introduction

Our analysis highlights the relationship between indicators of financial health and overall crime rates for ZIP Codes in Austin, TX using data on reported crimes from 2015. This information can provide valuable information to the city leaders of Austin as they seek solutions to reduce crime across the city. By using scatterplots, correlations, and t-tests, we show that ZIP Codes with poor financial health (i.e. those with lower household incomes, lower rent, and lower home values) tend to have higher overall rates of crime. However, caution should be exercised when interpreting these results or implementing any new policies. Many other factors and reasons could be contributing to the relationships described in this analysis. It is clear, however, that wealth and overall crime rates are connected.

Presentation Slides

https://docs.google.com/presentation/d/1SewMIVFVgho9Xl4Y4XQFXZIjfAmOoUpvi2arZdg3pys/edit?usp=sharing

Project Folder

https://github.com/Needle13/cs5830_project2

Dataset

The dataset used in our analysis is a collection of reported crimes and housing data from Austin, TX in 2015. The data set contains 38,414 reported crimes for the city of Austin, TX with various details for each crime, such as the type of crime, where it occurred, and when it happened. Additionally, the data set contains demographic information about each ZIP Code in Austin such as the median household income, median rent, and the median home value. We also used a dataset with population and population density information for each ZIP Code and used this to calculate the crime per 100,000 people. The combination of these data sets allowed us to compare financial health indicators, such as median household income, to crime rates for each ZIP Code in Austin.

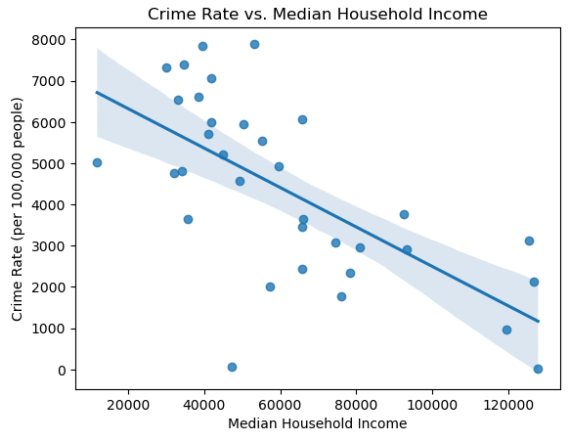
Analysis Technique

To analyze our data, we used scatterplots, Pearson correlations, summary statistics, and a t-test. Our first goal was to analyze whether there is a correlation between crime rate and financial well-being. We plotted the crime rate against financial well-being metrics in scatterplots with linear regression lines to visualize these relationships. We also calculated the Pearson correlations to determine if the correlations were significant. This analysis was appropriate because we were measuring the correlations between two variables in one population.

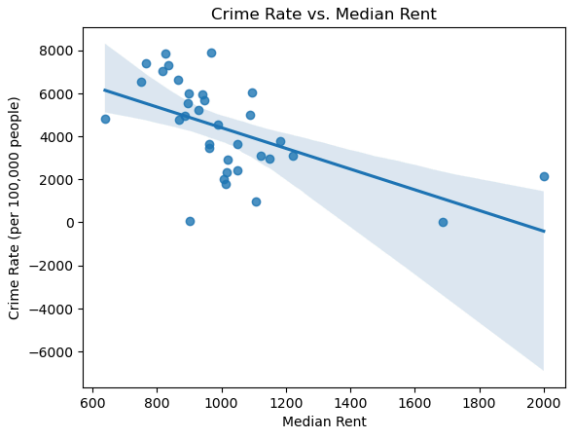
Our second goal was to analyze whether there is a significant difference in crime rates between lower income and higher income ZIP Codes. For this analysis, we sorted the data into two categories based on income. We plotted this distribution and calculated summary statistics (minimum, maximum, mean, median, and standard deviation) to visualize whether there appeared to be a difference in the distributions. We then performed a t-test to determine whether the difference was statistically significant. This analysis was appropriate for this data because we were comparing the distributions of two populations over one variable.

Results

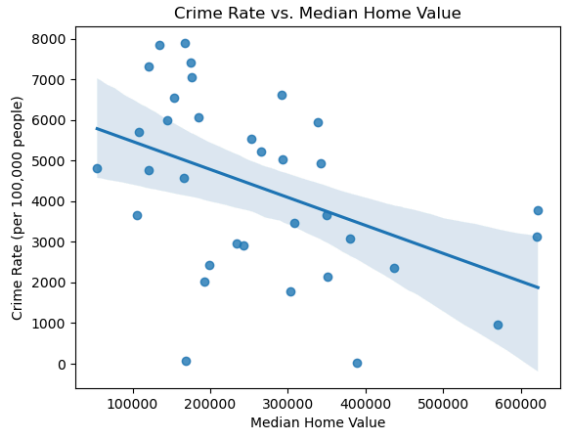
Our first few analyses examined the correlation between crime rate and financial well-being. We selected median household income, median rent, and median home value for our financial well-being metrics. We then plotted these metrics against the crime rate (normalized to be per 100,000 people). We found that all three of these metrics had a statistically significant negative correlation with the crime rate, although some were stronger than others. This means the crime rate decreases as financial well-being increases. Median household income had the strongest correlation, followed by median rent, then by median home value (see Figures 1, 2, and 3).



PearsonRResult(statistic=-0.6583575925756773, pvalue=2.3126276203441627e-05)



PearsonRResult(statistic=-0.5526589116740486, pvalue=0.0007006143274077552)



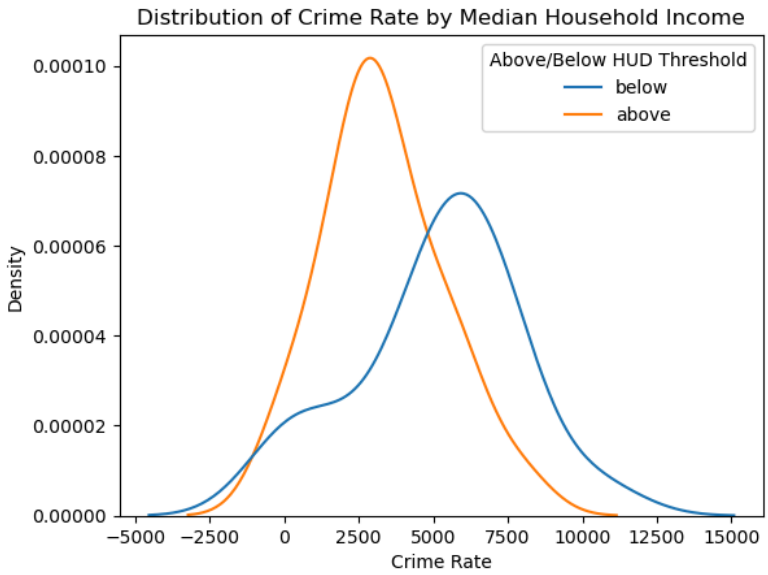
PearsonRResult(statistic=-0.45440739160137295, pvalue=0.0069393713621015585)

Figure 1 (top left). Crime Rate vs. Median Household Income scatterplot with linear regression and Pearson r correlation test result

Figure 2 (top right). Crime Rate vs. Median Rent scatterplot with linear regression and Pearson r correlation test result

Figure 3 (bottom left). Crime Rate vs. Median Home Value scatterplot with linear regression and Pearson r correlation test result

We wanted to further quantify the difference in crime rates between high income and low income areas. By using the Department of Housing and Development's (HUD) threshold for low income (80% of the area median income) [1], we classified each ZIP Code as either lower or higher income. The lower income ZIP Codes were primarily found in the south and eastern sections of the city. The distributions of crime rates for ZIP Codes above and below the threshold are compared in Figure 4. Based on the distributions, it appears that lower income areas have slightly higher crime rates and more variance than those of higher income.



T-Test Results: statistic=-2.41, p-value=0.02, df=38

Figure 4. Distributions of Crime Rates by Above or Below the HUD 80% Low Income Threshold with T-Test results

Using a t-test, we found some evidence that this difference is statistically significant (p-value=0.02). In 2015, the lower income ZIP Codes had an average of 1809 more crimes per 100,000 people than the higher income ZIP Codes. The summary statistics for lower and higher income ZIP Codes are shown in Table 1.

	Mean	Standard deviation	Min	Median	Max
Above HUD Threshold	3,317	1,991	28	3,081	7,894
Below HUD Threshold	5,126	2,728	11	5,414	10,538

Table 1. Summary Statistics for ZIP Codes Above and Below the HUD 80% Low Income Threshold

Of extra note is ZIP Code 78701, which is the very center of the city. This ZIP Code had a crime rate of 54,552 per 100,000 people, very far above any other observation, and was removed during our analysis. Several reasons could have contributed to this ZIP Code's very high crime rate, including: a relatively low population in the city center, a higher percentage of businesses and shops as locations for theft and other crime, a higher presence of tourists as easy targets for theft and other crimes, and the possible presence of gangs or other organized crime.

Our analysis results indicate that ZIP Codes with poor financial health, such as those with lower household incomes, lower rent, and lower home values, tend to have a higher overall rate of crime. These results may have safety, social, and political implications. Crime rates could possibly be reduced by initiatives to raise the median household income in lower income areas, or by increasing crime-prevention policies in those same areas. However, caution should be exercised, since actual crime rates may be substantially different from reported crime rates. Many other factors could also be contributing to the effect we have shown. It may be that lower income areas are already policed more heavily, thus creating a bias in the data. Additionally, this analysis indicates only a correlation between financial health and crime, which is not necessarily a causal relationship. However, it is clear that crime rates and wealth are connected.

Technical

These data sets required a significant amount of cleaning. We first cleaned our columns of interest by removing any commas, percentage signs, or dollar signs and converting them to float data types. We also removed any rows where the ZIP Code of the crime was not reported. We then counted the total number of crimes reported for each ZIP Code while keeping financial demographic data in the process. We then merged the data with the population information of each ZIP Code to be able to use that later for normalization. We also had to remove 6 ZIP Codes because they did not have any data in the financial demographic columns of interest (median household income, median rent, and median home value). We did not have a chance to take a closer look at those ZIP Codes to see what caused the gap in the data.

When calculating the crime rates, we decided to normalize by population by dividing by the ZIP Code population then multiplying by 100,000. This made the crime rates more comparable but also introduced the ZIP Code 78701 as a large outlier. We decided to address that ZIP Code separately and then remove it from the rest of our analysis to remove its impact from subsequent statistical tests. To divide ZIP Codes into lower or higher income areas, we used HUD's definition of low income, which is below 80% of the median income for a metropolitan area. The median household income for Austin, TX in 2015 was $57,689, which set our threshold at $46,151.20. This produced a roughly equal split of the city with 21 ZIP Codes above the threshold and 19 below.

The analysis techniques we used were appropriate for our dataset and goals. The scatterplots and Pearson correlations allowed us to compare two variables in the population. The t-test and summary statistics allowed us to split the population into two groups and compare the distributions.

Our analysis process was fairly straightforward. We started by doing some exploratory data analysis to see what we could potentially analyze in the data. Then we pinpointed a focus for the analysis on financial health. This led to us cleaning the data as described above. The biggest adjustments we needed to make throughout the analysis process were to normalize the crime rates by population and to drop the 78701 ZIP code.

Sources

[1] https://www.hudexchange.info/faqs/crosscutting-requirements/section-3/general/how-are-low-income-and-very-low-income-determined/