CS 5830/6830
Sajan Neupane
Preston Hall

Project 2: Crime

INTRODUCTION:

This analysis aims to uncover how socioeconomic factors influence crime rates within different neighborhoods and assess the impact of the time it takes to close cases on overall crime rates. We defined the time it takes to close a case to be the difference in days between the report date and the clearance date. By understanding these relationships, we seek to provide valuable insights for policymakers and law enforcement agencies to enhance public safety and allocate resources more effectively. Our motivation is to identify key factors that contribute to crime and improve strategies for crime prevention and management in various districts.

For a detailed overview of our findings, please refer to our presentation slides here and access our project files here.

DATASET:

We used the "crime and housing data" dataset for 2015 in Austin, Texas, complemented by population data, for our analysis. This dataset includes detailed records of individual crime incidents, such as crime keys, report and clearance dates, and the zip codes where crimes occurred. It also provides socio-economic attributes like median household income, poverty levels, unemployment percentages, and median home values, linked to the corresponding zip codes. The population data allows for the calculation of crime rates by normalizing crime counts against the population of each zip code. This comprehensive dataset enables us to explore the relationship between crime dynamics, socio-economic conditions, and demographic factors across various zip code or districts.

ANALYSIS TECHNIQUE:

In our analysis, we employed regression analysis and the Mann-Whitney U test to explore crime-related data. Regression analysis, utilizing Pearson correlation coefficients (with a p-value threshold of 0.05), was used to quantify the relationships between crime rates and socio-economic factors such as poverty levels, median household income, unemployment, and home values. This technique effectively highlighted significant predictors of crime rates through scatter plots with regression lines. For comparing the time it takes to close cases between districts, we applied the Mann-Whitney U test, a non-parametric test suitable for data that may not follow a normal distribution. This test helped us determine if there were significant differences in the time it takes the close cases between districts, providing a robust approach to understanding socio-economic impacts on crime and operational differences in crime management.

RESULTS:

Linear Regression Analysis:

Our analysis of the relationships between crime rates and various socio-economic factors revealed the following (Figure 1):

1. Crime Rate vs. Population Below Poverty Level: The correlation coefficient is $r = 0.55$ with a p-value of 0.0007. This p-value is less than 0.05, indicating that the correlation is statistically significant. Thus, we reject the null hypothesis of no relationship between crime rates and the percentage of the population below the poverty level and accept the alternate hypothesis that there is a significant relationship between these variables.
2. Crime Rate vs. Median Household Income: The correlation coefficient is $r = -0.66$ with a p-value of 0.00002. The p-value is less than 0.05, confirming that this correlation is statistically significant.

3. Crime Rate vs. Unemployment: The correlation coefficient is r = 0.57 with a p-value of 0.0004. The p-value is less than 0.05, indicating that this correlation is statistically significant.
4. Crime Rate vs. Median Home Value: The correlation coefficient is r = −0.45 with a p-value of 0.007. The p-value is less than 0.05, showing that this correlation is statistically significant.

The analysis indicates that crime rates tend to be higher in neighborhoods with lower median household incomes, higher poverty levels, and higher unemployment rates. Conversely, neighborhoods with higher median home values and higher median household incomes generally experience lower crime rates. This suggests that addressing socio-economic challenges, such as poverty and unemployment, and improving economic conditions in lower-income areas could be crucial for reducing crime. These findings provide actionable insights for policymakers and community planners to develop targeted interventions aimed at enhancing neighborhood safety and economic well-being.

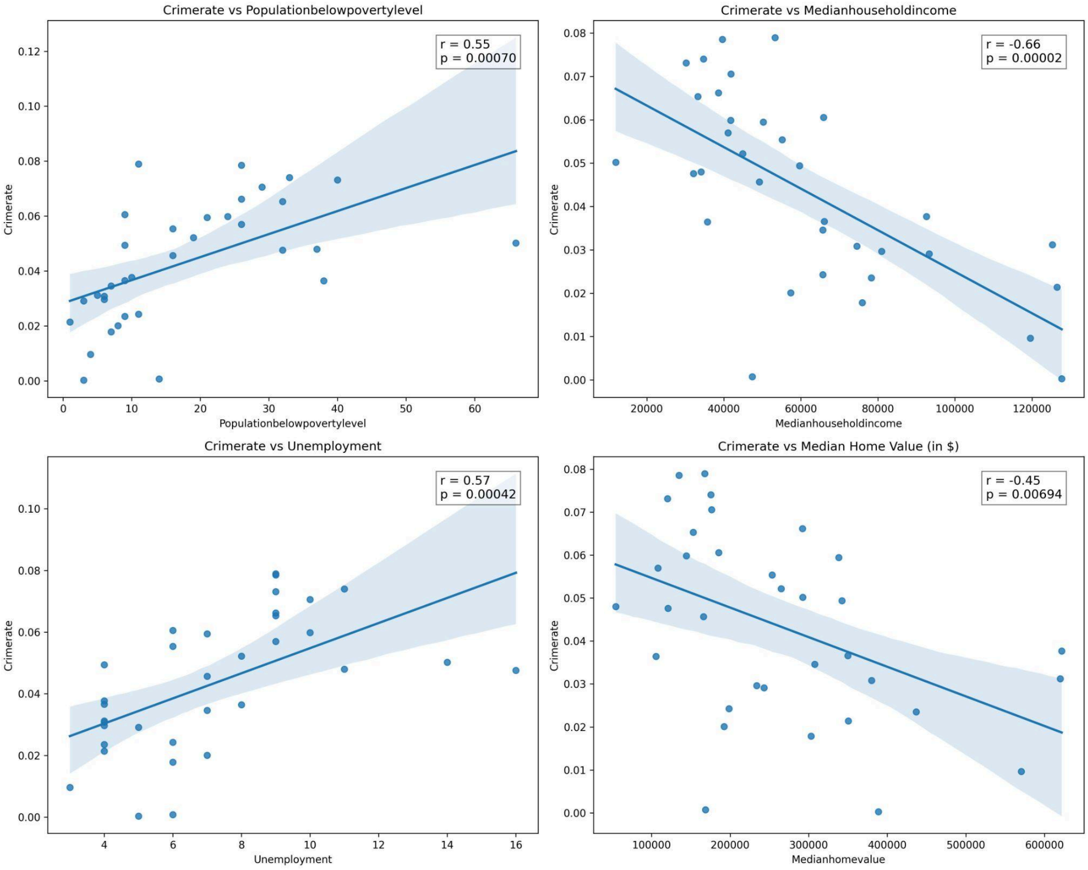


Figure 1: *Scatter plot showing crime per capita correlations with median household income, poverty levels, unemployment, and home value in Austin, TX neighborhoods (2015).*

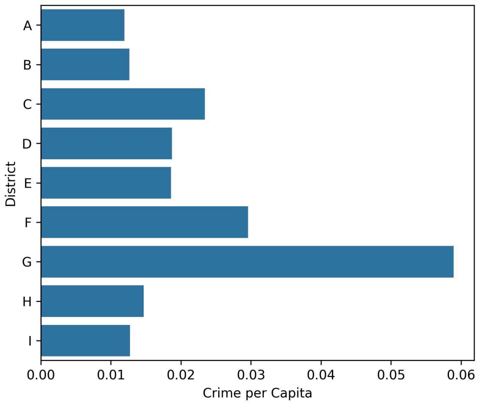


Figure 2 (left): *Crime rates per capita across different police administration districts in Austin, TX (2015).*

Crime by District:

Crime rates are highest in District George and lowest in District Adam. **Figure 2** shows the distribution of crime rates per capita across different districts. **Figure 3** illustrates that, despite having the lowest total crime count, District George has the highest crime rate per capita due to its smaller population and area.

For District Adam, the mean time to clear a crime is approximately 17.94 days, with a standard deviation of 35.07 days, indicating significant variability in clearance times. In contrast, for District G, the mean time to clear a crime is about 14.99 days, with a standard deviation of 32.42 days.

Figure 4 shows the distribution of the number of days required to close cases in these districts. The high standard deviations in both districts suggest considerable variability, with cases taking a wide range of times to resolve. This variability could be due to differences in resources, case complexity, or procedural efficiency.

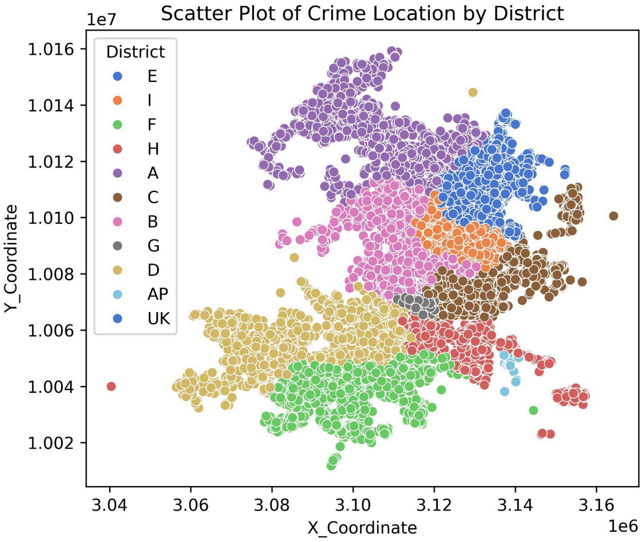


Figure 3: *Geographic distribution of crime locations, categorized by police districts.*

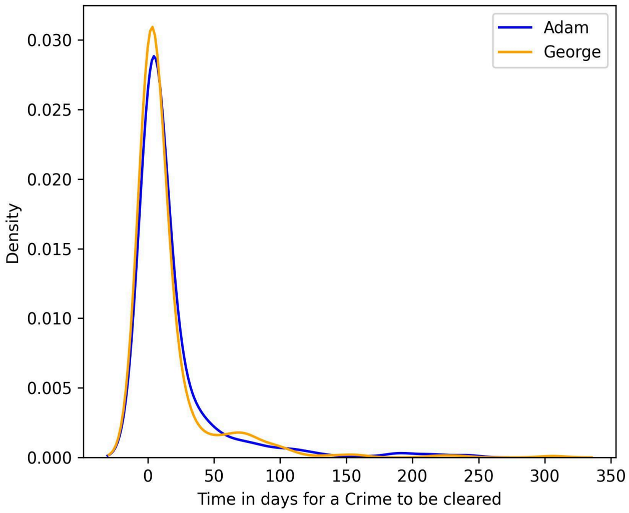


Figure 4: *Distribution of elapsed time in days for a case to be closed in District Adam and George*

To determine if the observed differences in report/clearance dates between District Adam and District George are statistically significant, we conducted a Mann-Whitney U test. This non-parametric test is suitable for comparing distributions when the data does not necessarily follow a normal distribution. The test assesses whether there is a significant difference in the distributions of the time it takes to close a case between the two districts. The results of this test will help us understand whether the differences observed are likely due to random variation or if there is a meaningful distinction in the time taken to close cases in these districts.

The one-sided Mann-Whitney U test was conducted to compare the time required to close cases between District Adam and District George. The test yielded a U statistic of 117,850.0 and a p-value of 2.83 × 10^-6. Since the p-value is significantly less than the significance level of 0.05, we reject the null hypothesis that there is no difference in the time required to close cases between the two districts, and accept the alternative hypothesis that the time it takes to close cases in District Adam is significantly longer than in District George.

Even though the crime rate is highest in District George, the time required to close cases is shorter compared to District Adam. This could be attributed to George's central location in the heart of the city, where a larger number of officers are likely deployed, enhancing the efficiency of crime resolution. Additionally, the nature of crimes in more crowded areas like George may be more visible and detectable, leading to quicker clearance.

TECHNICAL:

The dataset was cleaned and formatted to ensure accuracy in the analysis. This involved removing non-numeric characters from monetary values, converting percentages to numeric values, and handling missing data by dropping rows where essential information was absent. Zip code data was standardized for merging with population information, and crime rates were calculated by dividing the total number of crimes by the population for each zip code. Anomalous data points, particularly those with unusually high crime rates, were identified and removed to prevent bias in regression analysis

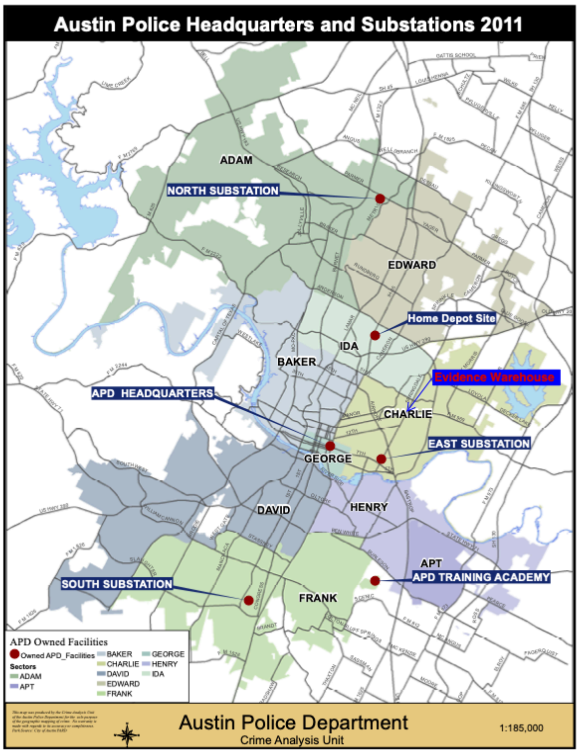
.

Two key analysis techniques were used: Pearson correlation and the Mann-Whitney U test. Pearson correlation was employed to explore the relationship between crime rates and socio-economic factors, such as median household income and poverty levels. This method is appropriate because it assesses the linear relationship between continuous variables, providing insights into how socio-economic conditions are associated with crime rates. Given that the data did not meet the normality assumption required for parametric tests, the Mann-Whitney U test was used to compare the time it takes to close cases in District George (G) and District Adam (A). This non-parametric test is suitable for comparing two independent samples with non-normal distributions, as it assesses whether one sample tends to have larger values than the other. The time it takes to close a case was measured as the number of days between the crime report date and the clearance date, and only crimes marked as "C" in the clearance status field were included in this analysis.

The analysis began with visual exploration of crime locations and clearance times through scatter plots and density plots, with a map of the districts used as an aid (See appendix A). Pearson correlation analysis was conducted to examine the strength and direction of the relationships between crime rates and socio-economic variables. During regression analysis, an outlier with an abnormally high crime rate was detected and removed to avoid distorting the results. The Mann-Whitney U test was then used to compare the time taken to close cases between District G, which had the highest crime rate per capita, and District A, which had the lowest. This comparison was crucial for understanding differences in clearance times across districts with varying crime rates. Initial attempts highlighted the impact of the outlier on regression results, emphasizing the importance of its exclusion for accurate analysis.

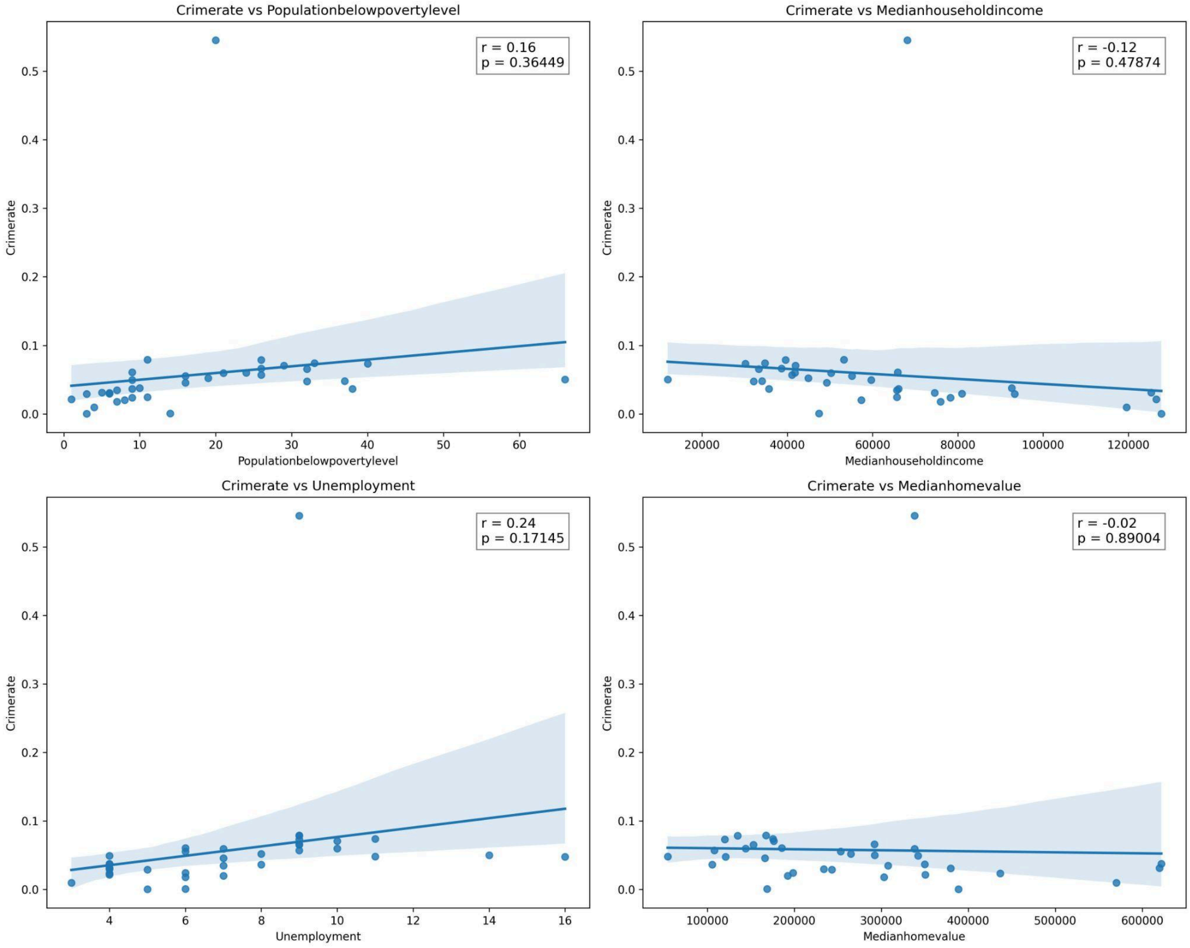
APPENDIX A:

While working on our analysis through the police districts, we found the following map of the districts especially helpful in getting an idea of their size, proximity, and geographic relationship to one another. This was especially helpful when considering why the districts might have had differences in the crime rates and the time it takes to close cases.



APPENDIX B:

When plotting the linear regressions between the four socioeconomic factors we analyzed, we found a statistical outlier for each of them. To prevent biasing our analysis, we removed this from the plots in the main report. Below are the same linear regression plots as we used in the main report, but with this outlier still present, as well as their correlation coefficients and p-values. It is clear after comparing these statistics with the linear regression plots in our main report, that removing the outlier was necessary for the analysis to yield accurate results.


Crimerate vs Populationbelowpovertylevel
r = 0.16
p = 0.36449
Crimerate
Populationbelowpovertylevel
Crimerate vs Medianhouseholdincome
r = -0.12
p = 0.47874
Crimerate
Medianhouseholdincome
Crimerate vs Unemployment
r = 0.24
p = 0.17145
Crimerate
Unemployment
Crimerate vs Medianhomevalue
r = -0.02
p = 0.89004
Crimerate
Medianhomevalue