Austin Crime Data 2015 Analysis Report

Satyanarayana Vinay Achanta
Brad Leavitt

Introduction:

Data analysis on the Austin housing and crime dataset provides much information about crime in the Austin, Texas, area. Our analysis of this dataset can help predict important analytics like a crime that happens the most, in which zip code or district has more crimes and has to be taken more care of, predict patterns of crimes happening, so it helps the police to analyze and catch the criminals faster, reasons making people commit a crime, heighten awareness of what conditions may be dealt with to lessen the influence of crimes hopefully, crimes taking a long time to get cleared and can analyze ways to clear it faster etc. Austin dataset is a significant sample and may provide insight into the nature of crimes in general. Through our analysis, we have looked at conditions that may affect the crime rates and types of crimes committed.

GitHub Link
Presentation Slides Link

Dataset:

Our dataset only contained one large C.S.V. file and a smaller supplementary one having zip code and population density.

Crime-housing-austin-2015.csv: This file contains every reported crime in Austin, Texas. This included 43 columns of data for each report. The sample size was also large, with over 38,500 incidents tracked in 2015. That's, on average, over 100 crimes a day!

- **AustinZipCodes.csv**: this file gave the report of population density per square mile.

Analysis Technique:

We started with data exploration, went through all the tables we needed, and looked at the columns required for each analysis. Additionally, we looked for NaN values and missing values. We grouped the rows based on the essential columns, such as Zip Code Crime and Highest NIBRS UCR Offense Description, using the data grouping technique after combining the two tables. Then, to count the number of rows, calculate the maximum count from the rows, and calculate the sum, we used the necessary aggregation functions count, max, and sum. Seaborn is used to plot the graph and create charts once the necessary data is available. We calculated correlation and significance using pearsonr and used t-test, average, standard deviation. To explore for trends and patterns, we employed a variety of charting approaches, including bar, displot, regplot and scatter plots. The analyses we made were these:

1. **Analysis-1:** Crime rates compared to Unemployment. poverty and income, population per zip code.

- All contributing factors were aggregated based on the zip code they were in and grouped into averages. The crime count itself was simply the total count per zip code.
- All variables were compared and given correlation values to each other so we could not only compare whether they possibly influenced crime but influenced each other. For the measures made comparing against crime, we used a Pearson correlation function to get the p-values confirming the accuracy of each measure.

2 **Analysis-2:** Count and types of crimes

- For the crimes that did happen, we were hoping to observe the nature of the crimes. What type were they? We used averages and visuals to find helpful information about them.

3 **Analysis-3**: Time from Crime report to clearance

- We wanted to see whether different areas of Austin were similar in how often crime shows up along with how well it's dealt with. We ran t-tests between a few areas to see if the distribution of handling crime is consistent across those areas in Austin, Texas

4 **Analysis-4:** Top 5 zip codes with the highest number of Crimes of each type.

- More area-type visualizations to show the count of crime types per zip code. A multi-bar chart is helpful for comparing similar plots of data across different categories. Zip codes, in our case

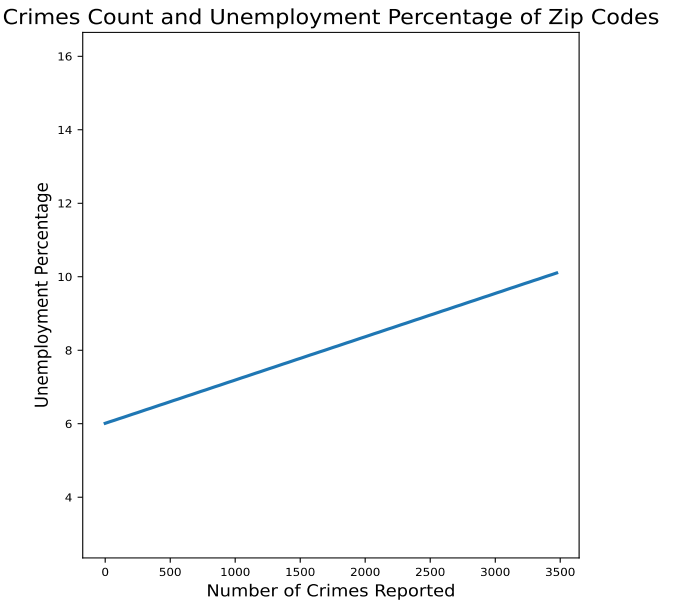
5 **Analysis-5:** Crime count in each district

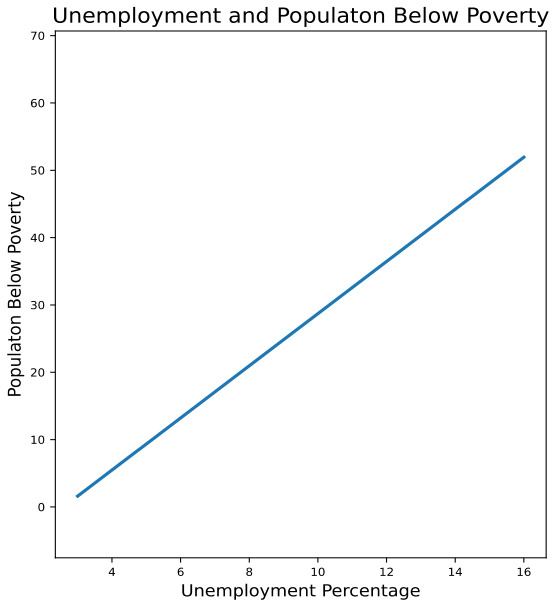
- The occurrence of crimes in different areas should also be measured to provide evidence if crimes are similar in different districts.
- A T-test and distribution plot was made to show numerically and visually how similar they are.

Results:

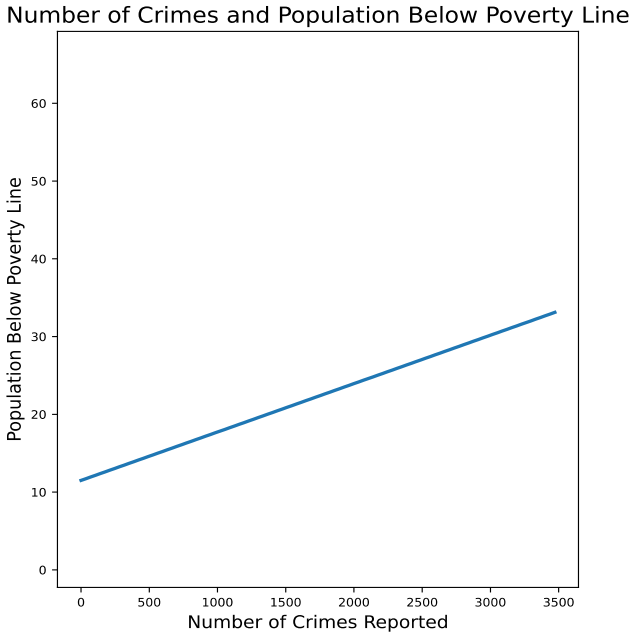
Analysis-1: Crime rates compared to Unemployment. Poverty and income, population per zip code.

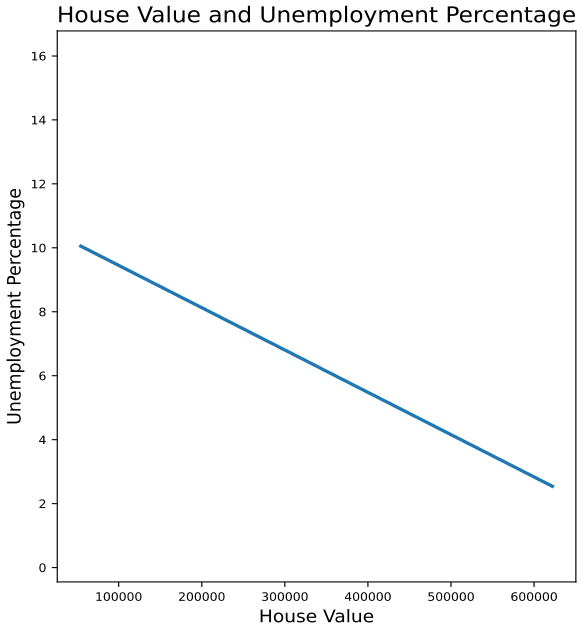
We found through Pearson correlation tests that Crime rates positively correlate with the percentage of unemployed people in an area. (See figure below to the left) Test results give us an r-value of .368 which is decently correlated. A p-value of .029 shows the accuracy of their correlation is significant. Speculation can be made that the unemployed aren't making money through jobs, so they are more likely to get by on illegal means. We measured another correlation between unemployment and how many are in poverty. (Below to the right).





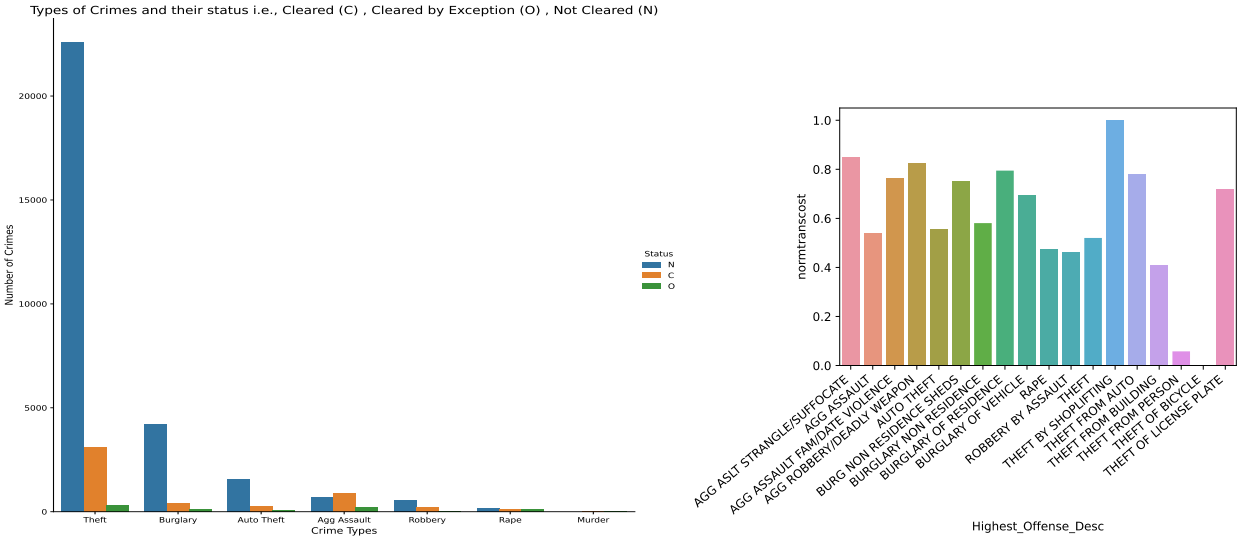
The t-test resulted in an r-value of .827, very positively correlated, supported by a p-value of 9.48×10^{-10}. Very small, meaning the two are strongly correlated. Unemployed people are in fact living in poverty. Those desperate enough may turn to crime.





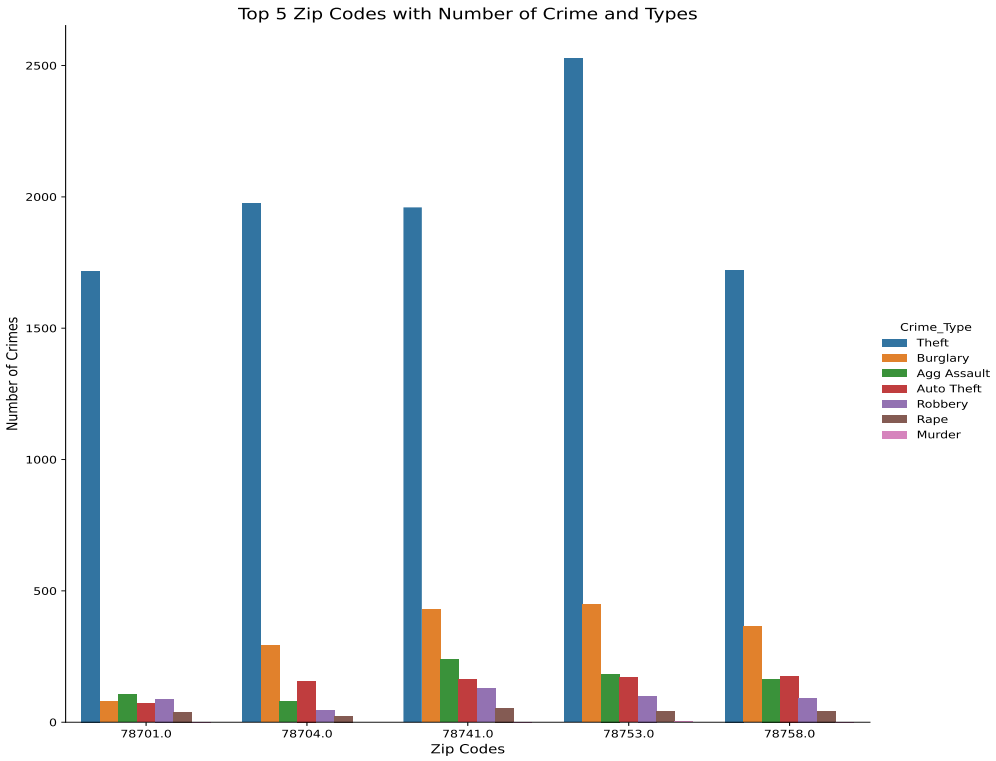
As with the other comparisons in this analysis, correlations were made to show the effects of poverty and unemployment. More crimes are reported as the number of impoverished increases. (top left) r-value = .415, p-value = .013. Housing values increase as unemployment goes down. (top right) r-value = -.622, p-value = .0000657. Our strongest correlation so far! This analysis is used to get the reasons triggering the people to make crimes and the government can investigate these and provide help to people to reduce crimes by providing basic necessities.

Analysis-2: Types of Crimes and their Frequency with Crime Status.



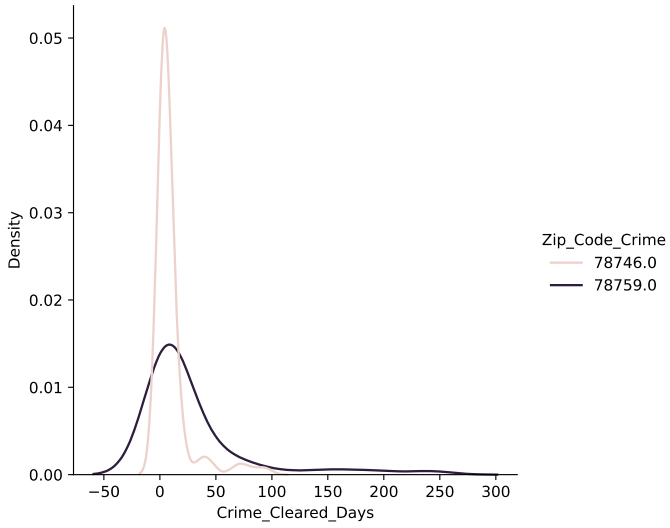
In analyzing the status of each crime, It was found that more people are caught and arrested while committing theft. This makes sense because theft is the most significant problem by a longshot just by the volume of theft. Local police are probably trained better to handle it better than others because of this as well. We also wanted to see how the cost of transportation affected how crime is distributed. Normalizing the data, we see, shoplifting is most likely to occur when transportation cost is high. Other forms of theft and burglary are high as well. This analysis can help police get insights about crime cases are not cleared more and pending, so they can concentrate on these for public safety.

Analysis-3: Top 5 Zip Codes with the highest number of Crimes of each type



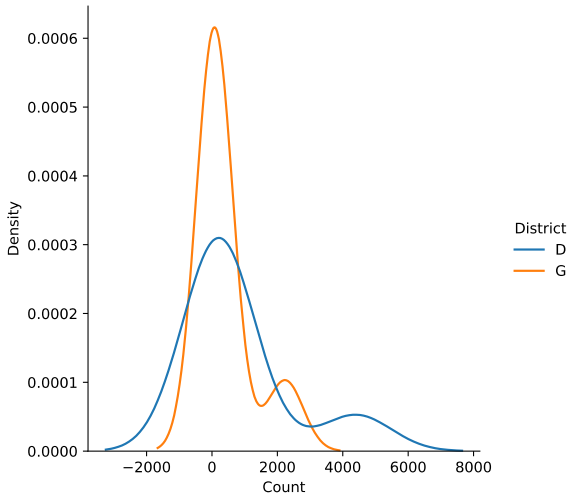
The top five crime-filled zip codes have very similar crime distributions other than the left two. They still have a ton of theft, yet the more personal violent crime is less compared to other zip code areas. Aggravated assault, rape, etc. This analysis will help security departments on which zip codes the police are working great and which zip codes more concentration is needed. In this case top 5 zip with more crime shows that a lot of investigation and work is needed in these zip codes and solve lot of pending crime cases.

Analysis-4: Number of Days taken for Crimes to get cleared based on Zip Code



Getting the averages to clear a crime, it took the longest to catch and arrest those who commit rape (70+ days), murder (36+ days), and robbery/burglary (32+ days) A t-test was performed to show how close the average time to catch and arrest a perpetrator between a couple of zip codes (see above figure). The results: statistic = 4.36, p < .001. Since the p-value is much less than .05 we have reasoned that there is a difference in that area of clearing crimes. This analysis will help get information about average clear time of a crime case and cases that are taking more time to solve based on zip code.

Analysis-5: Crime Count in Each District



We also wanted to check if there was a difference in the types of crime committed between districts. Choosing a couple of districts, we aggregated the count of each crime type in each of them. We then ran a t-test to determine a difference in distribution, with a statistic of .626 and a value of .543. We find little evidence of a difference between crime type distributions per district.

Technical:

We did data exploration and preprocessing, checked for features we would like to compare and analyze, and removed missing values. We had to remove '$', '%', ',' to use the numbers and did data preprocessing as well. We used the pandas' package in Python to load all the necessary files. Afterward, we used the group by function to group the data for each analysis and various aggregation functions (sum, mean, max, count) as needed. We did end up merging the two tables provided for population density and zip codes.

Producing numerical data about the crimes themselves was a little difficult. We needed to use aggregation to get the data we wanted. Normalization was an interesting part of the analysis that shed some light on how seemingly similar valued features can be scaled to become usable. The t-test is effective when finding if distributions are the same or similar. Finding proper distributions to test is harder, especially with less numerical data right off the bat.

We tried getting days from date to calculate on which days the crime happened the most. Tried extracting the days using different functions but did not work. Tried getting the regression line for crime rate and disable people percentage but the r-value is low, so did not include It in the analysis.