# CS5830

## Project 1 Report

Joshua Hatch, Nate Taylor

### Introduction:

This analysis was conducted over the “Austin Crime Report 2015” dataset and used a supplemental zip-code dataset for population information. Our analysis explored the well known correlation between poverty and crime as well as looking at how an increase in the poverty rate correlates with crime. In this case we found that both correlate with crime, though unintuitively the change in poverty rate correlates negatively. For reasons explained in our results section we would not recommend using the second value as a predictor for crime.

Additionally we focused on the different categories of crime, and how the population density of where the crimes took place differs across categories. Our findings indicate that the differences between average population densities of different crimes, while not large, is statistically significant. Our analysis would be primarily useful to inform residents of Austin living in the represented zip codes on types of crimes to expect in their area based on the population density and poverty rate. It would also inform decisions of the Austin area law enforcement in deciding which zip-codes to allocate finite police resources to and help potential home buyers and investors in the Austin region understand potential crime climates of prospective neighborhoods.

### Important Links:

Google Slides
Github Repo

### Dataset:

The “Austin Crime Report 2015” dataset comprises a collection of crime reports in the Austin area from 2015. While each report contains many different values, we only used a subset of these. In addition we the Zip-Code dataset provided in class to access the population totals and population density of the Austin zip-codes. It is important to note that the year that this data was collected was not provided, and because of this most of the results using this data should be received cautiously.

### Analysis Technique:

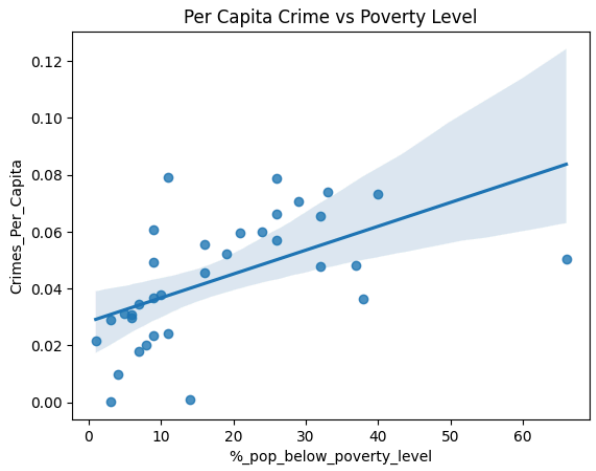
When evaluating the correlations between the percentage of population below poverty, the change in that percentage and the crime per capita we aggregated the crimes per capita at the zip-code level. We used a regression plot to show the linear regression of these two values when plotted with per capita crime, and we used Pearson correlation tests to find the r and p values. It should be noted that for this portion of the analysis, we dropped the zip code 78701 because it was an extreme outlier. When this value was included neither of our poverty tests showed statistically significant correlation.

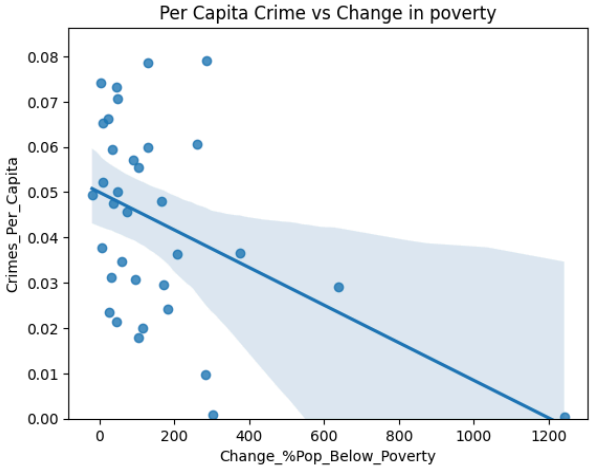
We also grouped the data based on crime classification, performing several aggregations on the different data coinciding with each crime. As a surface-level contextual

analysis, a total count of each crime was taken, as well as the percentage for each clearance status of the totals for each crime. These were both represented as simple bar plots. The first dive into crime classification versus average population density of zip codes in which they were committed yielded a bar plot, which we sorted in order for easy visual parsing. Next, the distributions of population densities of the zip codes in which each crime was committed were analyzed. A normalized distribution plot showed a general trend in the distributions of all crimes, while individual distribution plots showed the mean population density and standard deviation of individual crimes. Finally, a series of t-tests were performed on the data for each pair of crimes in order to determine which apparent differences in mean population densities were statistically significant. The three pairs with the smallest p-values (and most significant differences) were represented on normalized distribution plots, along with indicators for their different mean values.

## Results:

We predicted that the percentage of people under the poverty level would correlate positively with per capita crime, and we found a r-value of .55, with a p-value of .0007. This is a significant result and didn't surprise us. We hypothesized that a community that experiences economic decline would also see increased levels of per capita crime. Our correlation test of percent change in population below poverty level resulted in an r-value of -.44, and p value of .008, which is also significant. However we found this to be an unintuitive result. Why would communities that are getting poorer experience less crime per capita? If we examine our data the answer becomes clear.



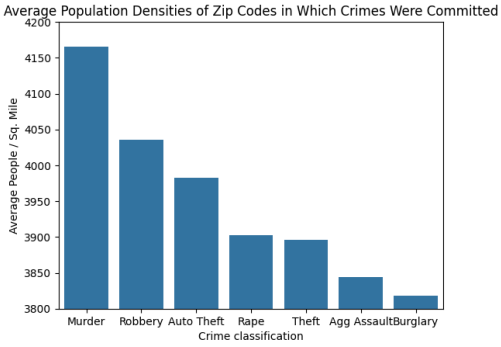



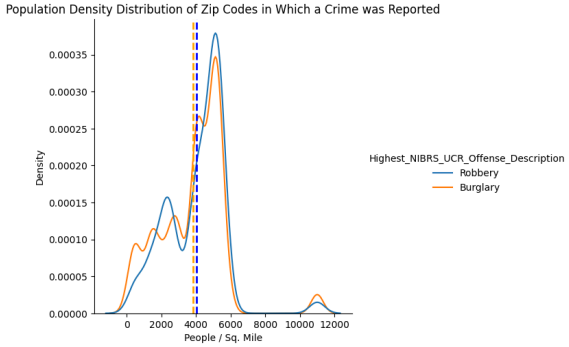

.

The data points that drive most of this correlation all exist at the upper bound of change in percentage. If we examine these zip codes we see that they all have low poverty levels. If we consider that an increase from 1% to 2% poverty is a 100% change in the original percentage, while an increase from 75% to 100% is only a 33% change in percentage. This results in the wealthiest zip codes dominating the analysis. In our case all of the wealthiest zip-codes became a little poorer so the correlation was negative, but if they had instead gotten only a little wealthier we would likely have a strongly positive correlation. Ultimately this is a very bad choice for predicting crime rates, and should not be extrapolated upon.

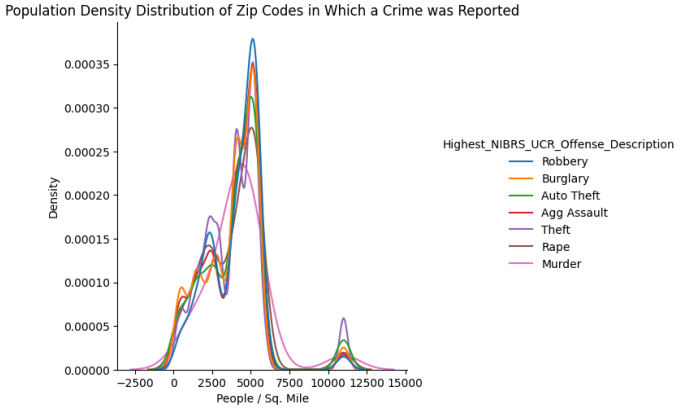
We extracted a few shallow results from our contextual analyses of crime totals, finding that thefts made up the vast majority of crimes reported at 28,274. While there was still plenty of data for most of the rest of the classifications, rape and murder reports lacked sample size (murders totalling only 23). This resulted in a lack of confidence in result findings related to these two crimes in particular. A look at the percentages of clearance statuses for each crime showed a few trends. Violent crimes like murder and aggravated assault showed much higher clearance by arrest percentages compared to the totals for these crimes, suggesting that perhaps these cases were higher profile/priority. Burglary, theft, and auto theft had very high uncleared percentages, and rape was the only crime with a higher number of cleared by exception cases than cleared by arrest. These statistics may have been due to a multitude of factors, so these results are generally inconclusive.

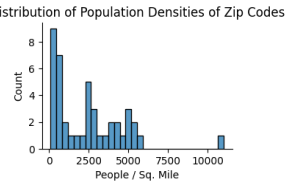
Taking the average of the population densities of zip codes in which crimes were committed showed that interestingly, violent crimes didn't seem to follow a trend, with murder having the highest mean population density of about 4166 people / sq. mile and aggravated assault the second lowest at about 3845. Burglary proved to have the lowest mean of about 3818 people / sq. mile – a result that seems to make sense considering it is safer for a criminal to break and enter a building undetected, and more homes and buildings may be remote/isolated from observation at lower population densities. Through t-testing, all differences between crimes excepting murder and rape proved to be generally statistically significant, with the greatest significant differences being between burglary and other crimes and robbery and other crimes. The p-value for the average population density of robberies versus burglaries was .0014, making this difference of 217 people / sq. mile the difference with the most confidence.



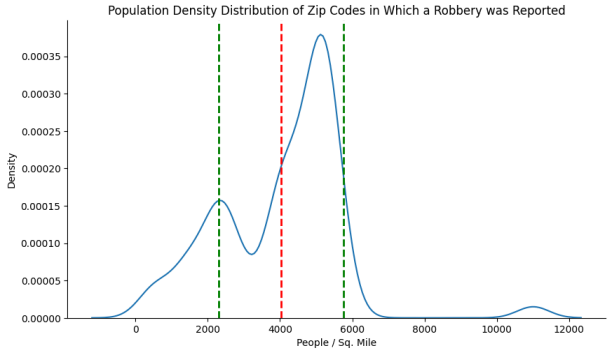



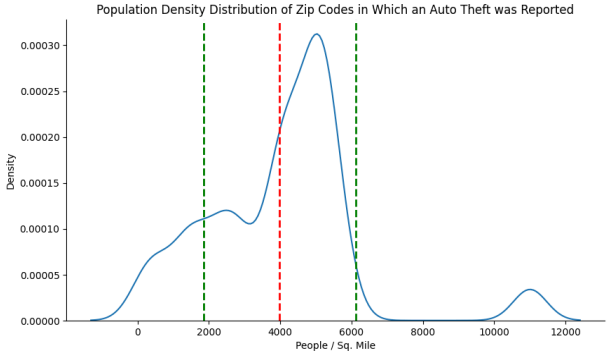

The normalized distribution of population densities for each crime showed a general pattern, with the mean distribution generally skewed toward higher population densities. Distribution plots showed four distinct peaks in the distributions at population densities of around 0, 2,500, 5,000, and 11,000 people / sq. mile. Further investigation showed that these peaks were at least in part due to the distribution of population densities of the zip codes included in the dataset. With a higher number of zip codes at or around each of these peak population densities, more data for crimes at or around these population densities in particular was



represented than other ranges of population density. Analyzing the data for each crime classification showed that the distribution of each crime's population densities had standard deviations ranging from about 1726 to about 2131. Robberies were the most tightly concentrated in population density, while auto thefts were the most spread out across population densities.




## Technical:

From the "Austin Crime Report 2015" dataset we used the zip-codes where a crime took place, the clearance status of the crime, the description of the type of crime committed, the percentage of the population below poverty, and the change in that percentage from the years 2000-2012. Prior to using our dataset for the poverty analysis we dropped zip codes that contained null values for either poverty metric. We also dropped one outlier zip-code 78701 from this analysis.

As discussed in our results we found the percent change in poverty to be a bad metric for our correlation analysis. A better metric would likely be the actual poverty level in 2000 minus the poverty level in 2012, however we didn't have access to this data, and because the real poverty level data we had in the dataset presumably came from 2015 we felt it would be irresponsible to calculate it based off of the 2015 data.

Because population density was the focus of many of our analyses, it was worth checking the distribution of population densities of the zip code data provided, which did prove to have an effect on results. There was one outlier zip code with an abnormally high population density, but since the zip code also had one of the higher populations and provided a large portion of crime reports, we felt that excluding it from the analyses would change the data too drastically for it to be representative of the unaltered dataset. After merging population density information with the crimes dataset based on zip code, data entries with no information about population densities were dropped, since they had nothing to contribute to population density aggregations.

For separate crime classifications, different data were aggregated through counts, means, and standard deviations. Crime clearance statuses were summed by crime and then divided by totals for each crime to find the crime-specific percentages of clearance statuses. The mean and standard deviation of population densities of zip codes where crimes were committed were taken as the data was grouped by crime. Lastly, t-tests were performed on each pair of crimes in order to confirm statistical significance of differences in mean population densities.