Population and Crime Analysis by Michael Childs & Ben Shaw

Introduction: For this project, we analyzed a dataset containing information about crime in the city of Austin. We looked for correlations between crime rates & the population as well as the population density. We also compared the crime rates for zip codes originating in two different areas of the city of Austin. Finally, we also looked for correlations between unemployment rates & poverty rates on the total number of crimes committed in various Austin zip codes.

Dataset: The main dataset is titled "Austin Crime Report 2015," & is available here as a .csv file, with each row corresponding to a crime committed in the year 2015. Each row has a description of the crime, the location of the crime (zip code & usually an address), & information about the area in which the crime took place, such as the population. Also available to us is a dataset with additional information about a given zip code such as the population density: the average number of people per square mile.

Analysis Technique: We wish to analyze the correlation between the Austin population and number of crimes, as well as the correlation between Population density and the number of crimes. In order to do this, we make use of the Pearson correlation coefficient and associated p-value. We also create scatterplots to visualize any correlation. In comparing the crime/population of zip codes from two different regions, we will make use of the t-test. We will use the t-statistic and associated p-value to analyze the statistical difference between crime/population values of zip codes from different regions. We will also compare the mean values of each distribution. We also wish to analyze the correlation between unemployment rates & poverty rates (aka the % of the population under the poverty line) on the total number of crimes committed across Austin. By using the Pearson correlation on unemployment vs poverty, unemployment vs total crimes, and poverty vs total crimes, we can likewise identify if there are statistically significant results for these elements.

Results: Our first result is that there appears to be a correlation between the population and the number of crimes committed. We obtained a correlation coefficient of approximately 0.818 and associated p-value of approximately $6.0 \cdot 10^{-11}$.

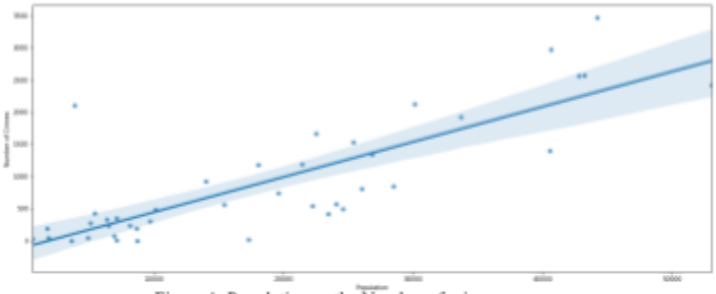

Figure 1: Population vs the Number of crimes.

Next we examined the correlation between the population density and the number of crimes. We obtained a correlation coefficient of approximately 0.598 with p-value of approximately $3.0 \cdot 10^{-5}$.

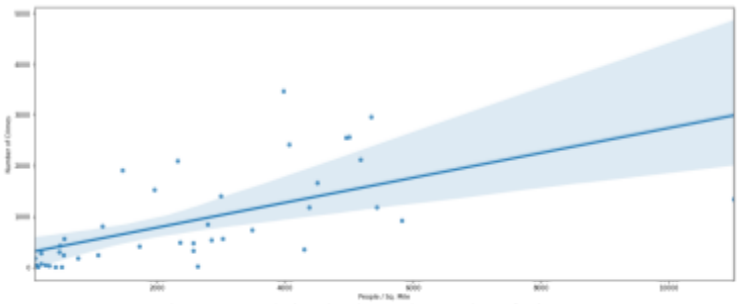

Figure 2: Population density vs the Number of crimes.

Next, we grouped the zip codes by geographical location. A map is given in Figure 3. The zip codes were grouped based on whether they were below/right of the diagonal line drawn from the top right corner to the bottom left or whether they were above/left of the line.

Figure 3: A map of the zip Codes of Austin City, downloaded from:
https://www.maxlcaman.com/mortgage-resources/texas-zip-code-maps/city-of-austin-zip-code-map/.

A Histogram of each distribution is shown in Figure 4 below.

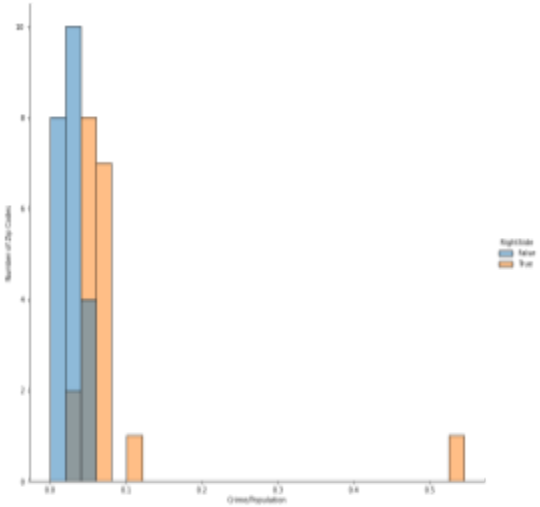

Figure 4: A Histogram of the right and left Zip codes.

A t-test was performed which yielded a t value of approximately 2.477, along with a p-value of 0.018. This informs us that there may be a statistically significant difference in the crime/population values of the different groups of zip codes. We find that the mean crime/population value of the left group is 0.026, while for the right group it is 0.086.

Our second group of results was given by the analysis of poverty & unemployment on total crime rates. The Pearson Correlation coefficient for these is 0.7577 (very high; the closer to 1, the more likely for correlation) with a p-value (statistical significance of the correlation) value of $8.753 \cdot 10^{-08}$ (extremely significant). We created a scatterplot with a regression line to visually identify this correlation (the closer to the line, the better the correlation):

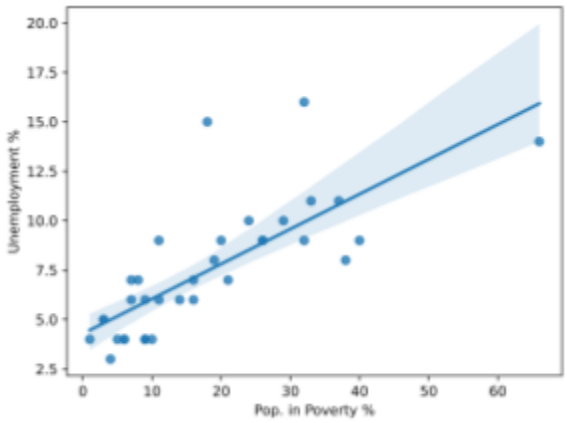

Figure 5: Scatter Plot with Regression Line for Unemployment vs Poverty Percentages across each zip code.

From this figure, we see that there is indeed a very strong correlation between Unemployment and Poverty (it is proposed that being unemployed means no income which contributes to poverty). Next, we analyzed poverty vs the total amount of crimes committed in each zip code. Our Pearson Correlation Coefficient was 0.4114 with a p-value of 0.0127. This is interesting, as the likelihood for correlation is a great deal lower than the previous of 0.7577, and the p-value gives a measurably lower probability of statistical significance than $8.753 \cdot 10^{-08}$ and thus provides a weaker argument that poverty and crime occurrence is directly correlated than that of unemployment and poverty.

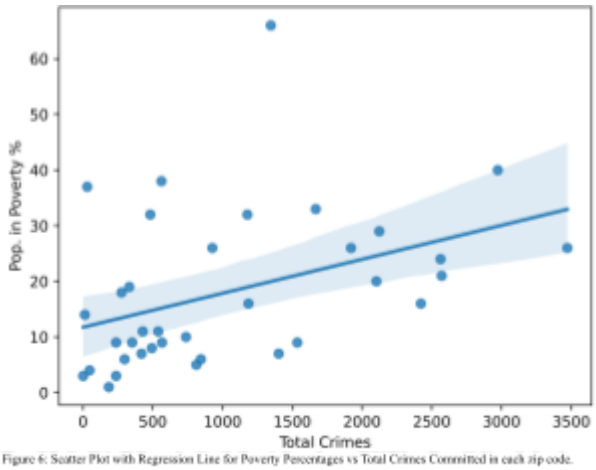

Figure 6: Scatter Plot with Regression Line for Poverty Percentages vs Total Crimes Committed in each zip code.

This visualized plot validates the numerical result. Finally, we performed a similar analysis of the correlation of unemployment and total crimes which resulted in a Pearson Correlation Coefficient of 0.279 and a p-value of 0.0992. These provided the lowest confidence of correlation of the 3 analyses, and thus there is a lower chance of a correlation between crimes and unemployment than that of crimes and poverty.

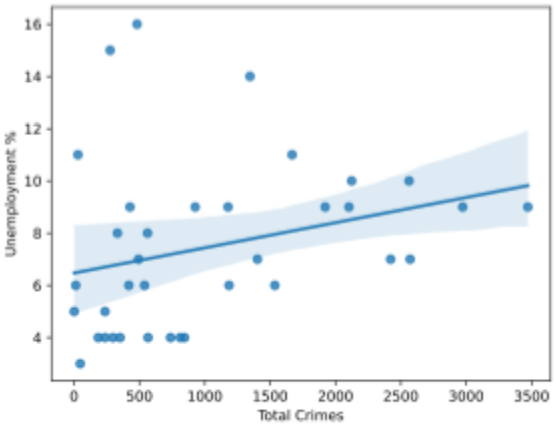

Figure 6: Scatter Plot with Regression Line for Poverty Percentages vs Total Crimes Committed in each zip code.

This is fascinating, and it is proposed that although poverty may be a year-to-year state of being and thus individuals may enter a life of crime because of this, unemployment is likely a temporary status for individuals and is less likely to induce crime. So, although poverty may lead to unemployment, unemployment likely doesn't lead to crime (on a general level)..

Technical: The data itself did not present any major obstacles in the analysis. However, the values for the population and population density were given as strings and needed parsing so that the data in those columns were numeric. At one point, dates were being examined, requiring reformatting of the dates–however, this project was cut from the final report and presentation.
As can be seen in figure 4, there is an outlier in the right group. Taking this value out, the mean for the right group is 0.06. The new t-value becomes 5.98, with associated p-value of approximately $6.0 \cdot 10^{-7}$. Thus, the t-test still demonstrates the potential difference between the crimes/population values for the two different groups. However, the sample size for each group was approximately 20, so we believe these findings are rather limited.

The previously alluded to (yet abandoned) analysis was that of comparing the seasonality of crimes from one zip code to another. It was thought that a histogram could be created for a given zip code detailing the number of crimes for a time of year: perhaps crimes would be more likely to be committed in January for a particular zip code, while a different zip code could have a peak in crimes in the summer. However, a t-test was not attempted for comparing the distributions as the number of crimes committed per zip code appeared to be relatively constant, not having the appearance of a normal distribution.

Finally, for the analysis of unemployment, poverty, and total crimes committed, the data was similarly accessed as previously mentioned, but uniquely required extra string processing due to percentage signs present in the data. The statistic of "total crimes" was not a given part of the dataset & was manually computed with a count of all crimes in each zip code & then compared with the respective statistics. This comparison was a challenge as several zip codes were missing values for unemployment and/or poverty, & thus needed to be excluded from the dataset. However, these zip codes were still present in the sum of total crimes committed, & thus figuring out the best way to join these respective dataframes was a challenge but accomplished. Lastly, an attempt was made to analyze each of these statistics in regards to an ANOVA test & their crime-clearance statuses. However, this attempted analysis proved to be outside the scope of the project or the pursued results thereof, & was thus omitted.

Links:https://github.com/kaisermikael/cs6830_project2 - GitHub repository.
https://docs.google.com/presentation/d/1b1ep8oF7_CqQJvUSjyKJw-4mxy4zOoYY2BFWdUUWh-A/edit?usp=sharing - Presentation slides.