Michael Childs
Ben Shaw

Population and Crime

Introduction

For this project, we analyzed a dataset containing information about crime in the city of Austin. We found correlations between crime rates and the population, as well as the population density using the Pearson correlation coefficient: these correlations are supported by small p-values. We also compared the crime rates for zip codes originating in two different areas of the city of Austin. Using a t-test, we found that the crime rates for the areas were different, and that the difference was statistically significant.

Dataset

The main dataset is entitled "Austin Crime Report 2015," and is publically available at https://data.world/dash/austin-crime-report-2015. The data comes as a .csv file, with each row corresponding to a crime committed in the year 2015. Each row contains information about the nature of the crime (a description), the location of the crime (zip code and, most often, an address), and information about the area in which the crime took place, such as the population of the corresponding zip code.

Also available to us is a dataset which contains 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 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.

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 $6.0 \cdot 10^{-11}$. A scatterplot is given in Figure 1.

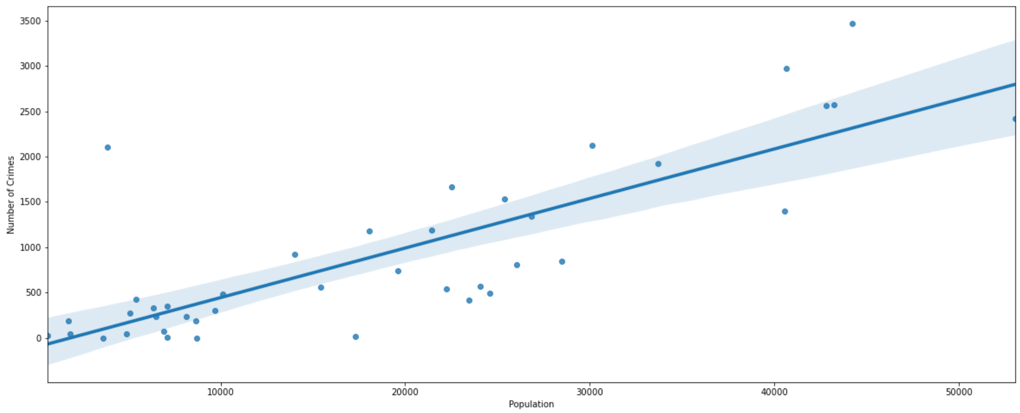

Figure 1: Population vs the Number of crimes.

Next we examined the correlation between the population density and the number of crimes. We obtained a correlation coefficient of approximately 0.598 with an associated p-value of approximately $3.0 \cdot 10^{-5}$. A scatterplot is given in Figure 2.

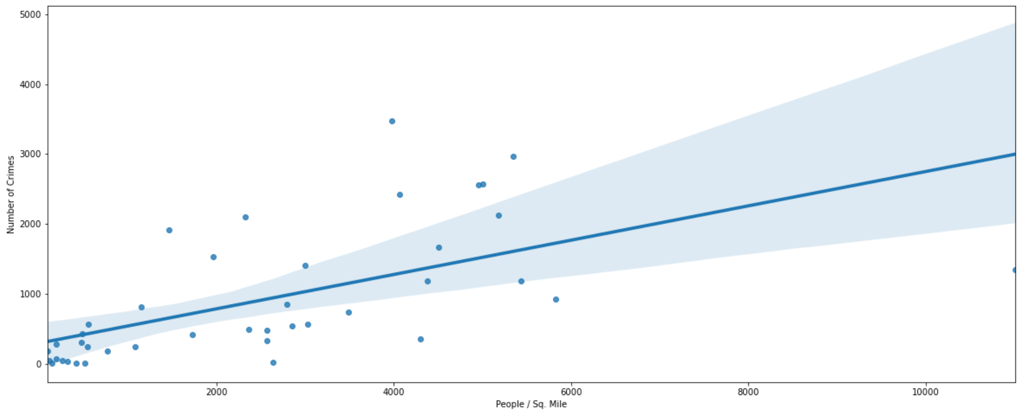

Figure 2: Population density vs the Number of crimes.

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.

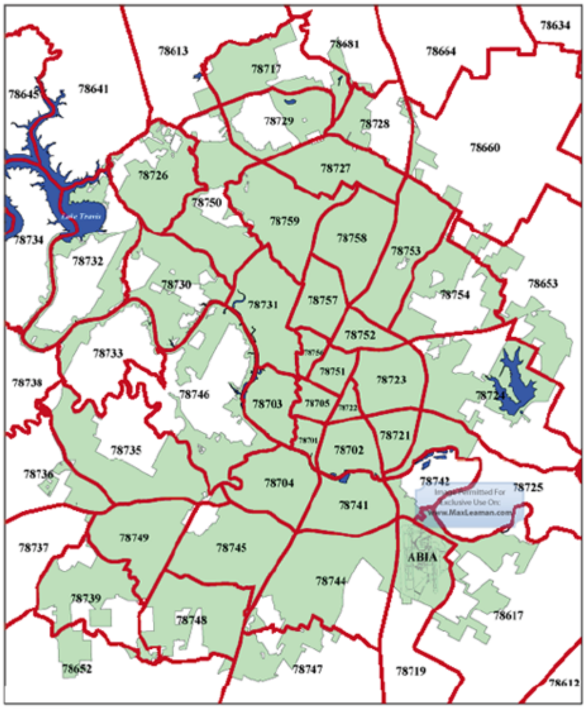

Figure 3: A map of the zip Codes of Austin City, downloaded from https://www.maxleaman.com/mortgage-resources/texas-zip-code-maps/city-of-austin-zip-code-map/.

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

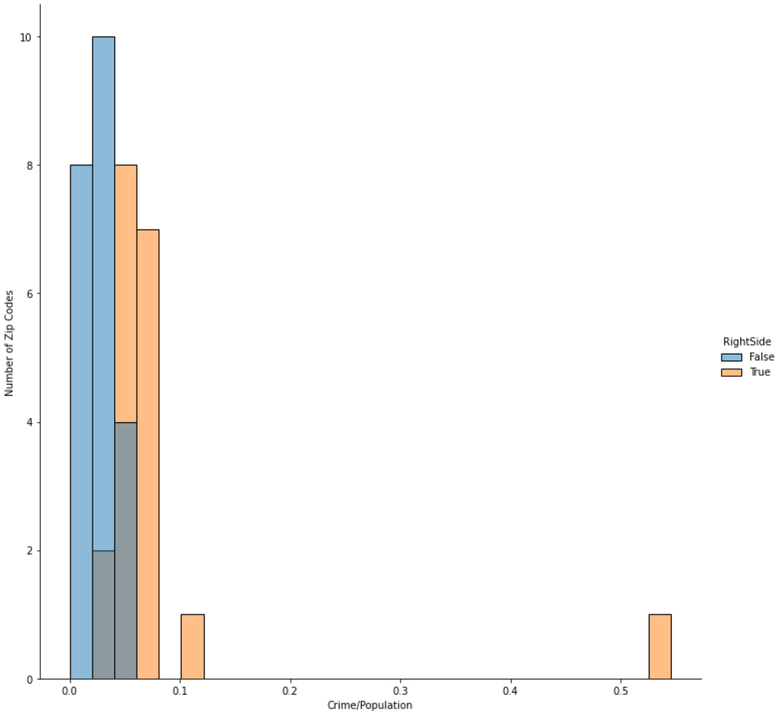

Figure 4: A Histogram of the right and left Zip codes.

A t-test was performed, which test yielded a t value of approximately 2.477, along with a p-value of 0.018. Strictly speaking, the t-test 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.

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, begging the need to parse/convert 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.

One analysis that was abandoned–alluded to previously–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 (comparing zip codes), as the number of crimes committed per zip code appeared to be relatively constant, not having the appearance of a normal distribution.

Links

https://github.com/kaisermikael/cs6830_project2 - GitHub repository.

https://docs.google.com/presentation/d/1b1ep8oF7_CqQJvUSjyKJw-4mxy4zOoYY2BFWdUUWh-A/edit?usp=sharing - Presentation slides.