Introduction

In this report, we aim to provide a comprehensive analysis of Crime and Housing in Austin, Texas in 2015. To accomplish this, we utilize a dataset composed of specific crimes, dates, zip codes, and various housing information for both renters and owners. We did not have a specific approach to analyze the data but we started with asking questions and figuring out which of those questions would be most useful with the given dataset. This way of analyzing gave us the ability to view the dataset in a way which was both interesting and useful to help understand and hopefully decrease crime. The insights obtained from this study can be used to inform future decisions and drive positive change.

GitHub repo: https://github.com/ablibranix/project2

PowerPoint: Presentation

Dataset

This dataset contains information about the date and type of specific crimes, housing rates for various occupations, zip codes for the respective crimes and housing areas. From the dataset, specific key attributes were selected which were vital to our analysis. These attributes were Zip_Code_Housing, Zip_Code_Crime, Report_Date, Medianhouseholdincome, Highest_NIBRS_UCR_Offense_Description, Zip Code and Population. Without these attributes, we would not be able to complete our analysis nor would the dataset be complete.

Analysis Technique

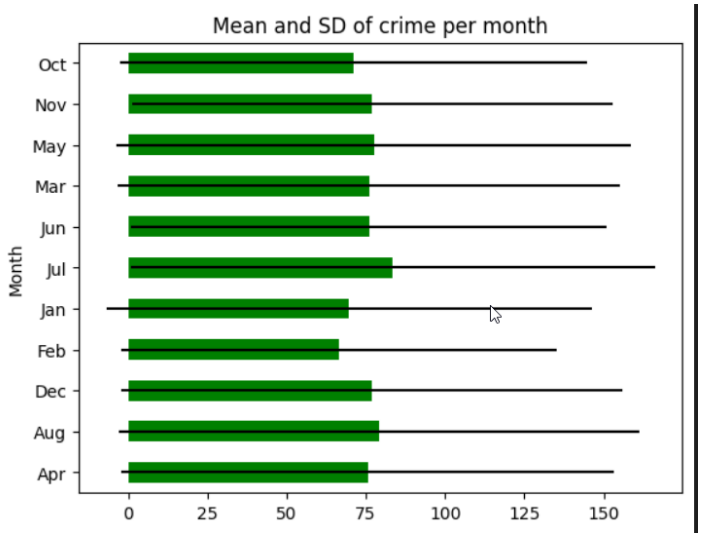
The analysis technique used was the process of question and answer. We analyzed the data, came up with some questions which interested us and we wanted to learn more about. We then performed the specific statistical methods required and answered some of our questions. This type of analysis was great for our dataset since there was a lot of information presented and being able to select the most useful columns and format it to answer our question was very useful.

Results

Crime per Month

As we looked through the dataset, we became interested in the potential variations in crimes committed per month, and whether there seemed to be any visual difference in how much crime

occurred in different months. For each month, we found the total number of crime reports made in each zip code, then calculated the mean and standard deviation. We found the following:

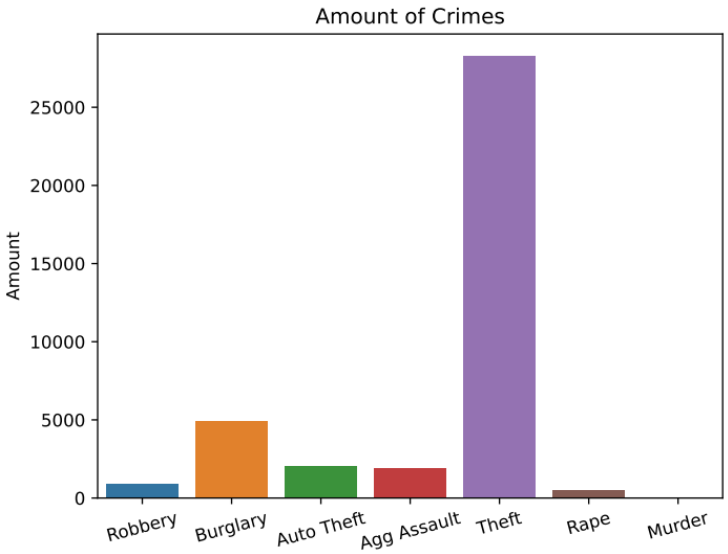


	Count	
	mean	std
Month		
Apr	75.714286	77.695241
Aug	79.309524	82.075066
Dec	76.860465	79.168771
Feb	66.558140	68.811787
Jan	69.681818	76.547834
Jul	83.500000	82.648315
Jun	76.000000	75.057121
Mar	76.071429	79.269036
May	77.488372	81.203851
Nov	77.073171	75.907967
Oct	71.139535	73.574576

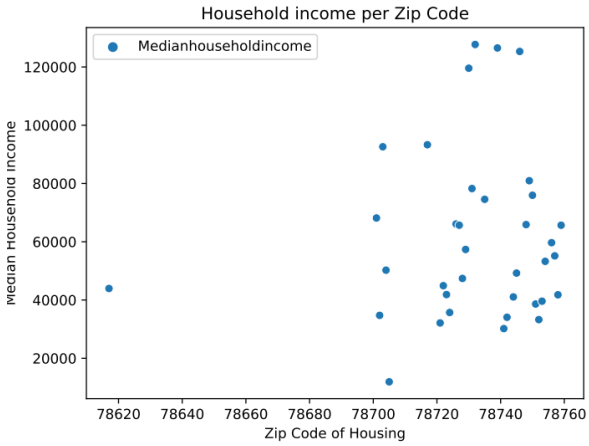
Interestingly, July has a much higher mean number of reports than does a month like February. The huge standard deviation is likely a result of differing populations of zip codes as well as certain zip codes having higher crime rates. February could have the lowest crime rate perhaps because it is the shortest month, but it could also be because February is too cold for crime, because people are too head over heels, or because of Lent.

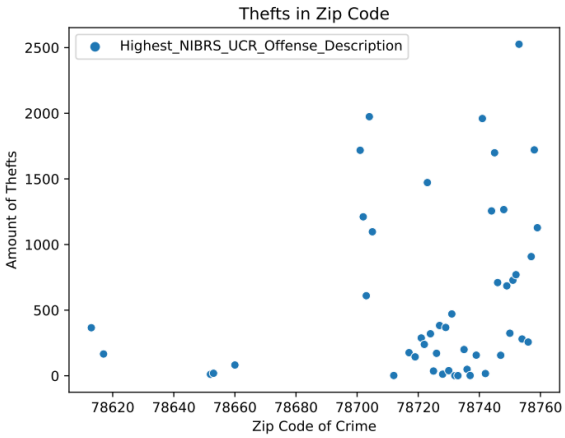
Housing vs Crime

One of the many questions which were derived from the dataset was is there a correlation between housing income and crime statistics. From this question, the analysis of Highest NIBRS UCR Offense, Median household income, and their respective zip codes. The first analysis is to see what the most common crime is and after formatting the data, Theft, over all the zip codes, is the most common.



With this information, the dataset was changed to include just theft crimes since they were the majority of crimes. To compare crimes to income, the income for each zip code was required. The information for median income and theft locations was filtered to only include zip codes with theft as the highest crime description and then plotted the points.

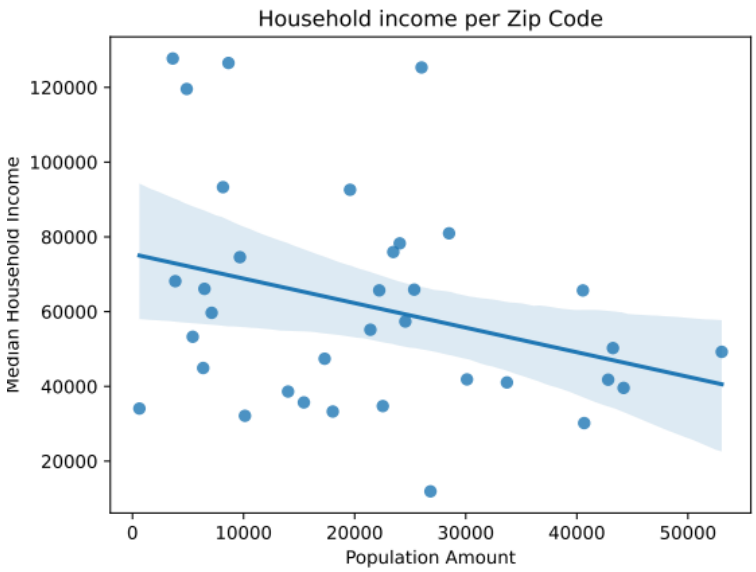




As you can see from the scatter plots above, the zip codes with higher income rates also had a higher amount of crime. This can be seen specifically in the zip codes between 78740 and 78760. A Pearson correlation test was done to see the correlation between median household income and the housing or crime zip code. There was a positive correlation of 0.0388 with a p-value of 1.40835649699256e-10. With a low p-value, we can reject the null hypothesis and conclude the relationship between zip code and household income exists and is statistically significant.

Population vs Household Income

While looking through the given dataset, the household income and population information was questioned if they had a strong relationship. The information was formatted to exclude null values and then graphed to show the regression of household income as the population expanded through Austin, Texas. From the graph, as the population grew, the household income decreased.



A t-test was performed to see the difference between the two pieces of data. From the test, we found a t value of -7.405 and a p-value of 2.631e-10. Since our p-value is significantly smaller than 0.05, we can reject the null hypothesis of no difference and say with confidence that the population of an area has an effect on the median household income for that area.

Technical

Since much of this data was not just numerical information, some adjusting was required in order for some columns to be in format which could be compared to other columns.The analysis technique we chose was to look at the dataset and, based on the column names and the small portion of the data viewed, ask questions which we were interested in and then format the data in a way to answer these questions. This analysis technique worked well for us and the dataset since there was so much information. It would have been difficult to analyze the whole dataset or just jump in without knowing some of the information we would be utilizing.

The most difficult part of analyzing this dataset was formatting the data types in a way which could be compared to other data. Some additional adjustments were removing null values, adjusting data types, and shortening the dataset to only the needed columns or rows. With this adjustment, the analyses were easier to complete. I think a different approach we would have taken was better understanding t and Pearson tests so this portion of the analysis would not have been as difficult or possibly incorrect.