Colby Wight A01632572; Tanner Wheeler A01770306
Data Science Incubator CS 5830/6830
Slides:
https://docs.google.com/presentation/d/1Bu8hzsaJZ3C4mP44wsamaCRdSueU0P-zaIdXGjxAygs/edit#slide=id.g7ce7e97532_0_25

Project 2

Introduction

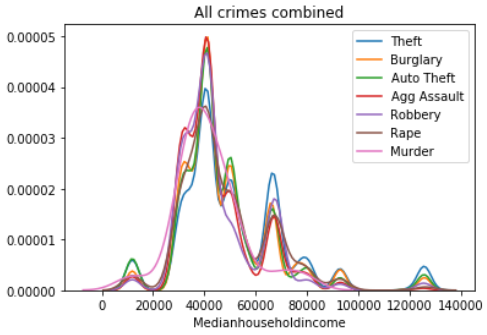
This report uses statistical methods to analyze the crime gathered in Austin, Texas. Our analysis will inform local authority of similar patterns in crimes reported using household income and home values in Austin. The techniques used in this analysis include Pearson R values, t-tests, descriptive statistics (mean, standard deviation, median), and graphs.

Dataset

The dataset provided for this analysis is from https://data.world/dash/austin-crime-report-2015. This dataset contains all of the crimes reported from January 1, 2015 to December 31, 2015. Each report contains the type of crime, the filing and closing dates, the location, the result of the report, and 37 other attributes describing the area of the crime. This analysis focuses on the column Highest_NIBRS_UCR_Offense_Description in the dataset. This column details the highest crime of the report. Analysis is done grouping the data into reports of each crime using the highest crime of the report. Dollar signs and percentages are used in the dataset. These symbols needed to be removed from the data. The columns Medianhomevalue and Medianhouseholdincome were also used throughout the analysis. Reports that did not contain values for these columns were also removed.

Analysis technique

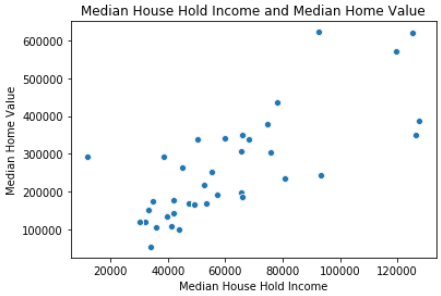
The crimes in the dataset are compared and analyzed for each median home value area and median household income area. Scatter plots are used to visualize the crime's correlation with respect to these two areas. An R-value is computed on each of these scatterplots. Histograms are another way the crimes and areas are visualized. The histograms are visualized both with bins and kernel density curves. Kernel density curves are used to compare thefts in different median household income areas and the other crimes. We tried to visualize the differences between all of the crimes in one kernel density curve, but the data was difficult to distinguish between the crimes.

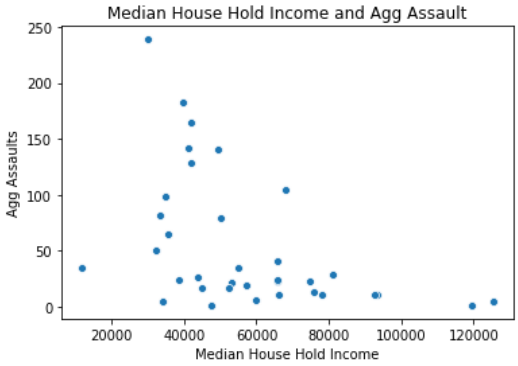


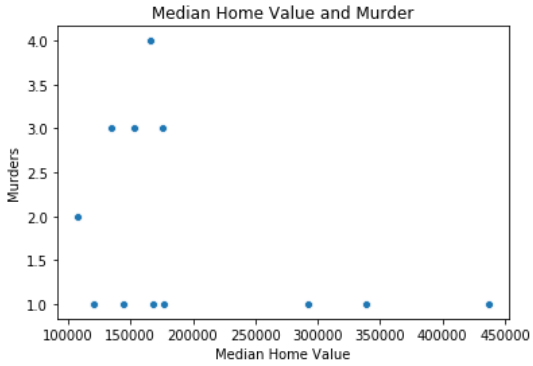
Instead thefts were compared with each crime separately. A t-test was computed for pair of crimes. We also used standardization to make sure the scatter plots we were making were being shown properly.

Results

We learned a few things about where crimes take place with regard to the area's median home value. In the scatterplots we made we learned that the two variables (income and home value) are positively linearly correlated as one might be able to guess, but we didn't learn a whole lot more about the different crimes using this. We found a very strong correlation of .98 with a very small p-value to reinforce this.

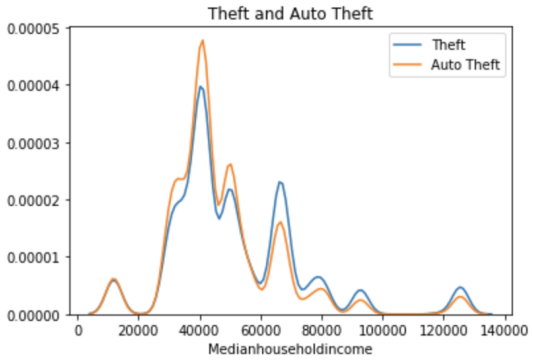






In the other scatter plot examples above, we see a negative correlation between both home value and income when compared to each crime. As with the murder scatterplot we found that some correlations are not statistically significant. If we were to receive more data we might be able to confirm these correlations. Creating a histogram for each crime we saw that they all followed a similar pattern too. Most of the crimes happened in areas where the median income was lower.

With the kde of theft plotted against other crimes we saw some interesting things. While the theft vs auto theft plot look similar we noticed that auto theft was always higher in poorer neighborhoods while theft was higher in richer neighborhoods. Why could this be? Maybe cars are easier to steal in poorer neighborhoods while, thieves go to richer neighborhoods to steal diamonds or other smaller expensive things. This could be used to help cops to be aware that cars may be more likely to be stolen in poor neighborhoods to keep a lookout or to give rich people more peace of mind about their nice cars. See the plot below:



The evidence of the difference between theft and auto theft can be further seen by looking at a t-test. We tested whether there is a difference in the mean of the median household income between the two crimes. With a p-value of 5.31e-15 we have strong evidence we can reject the hypothesis that the means are the same and conclude that there is indeed a difference.

Finally, in our analysis of the crime clearance rates we got the following table showing the clearance rates for all the crimes:

	Clearance_Status			
	size	sum	mean	std
Highest_NIBRS_UCR_Offense_Description				
Agg Assault	1839	1140	0.619902	0.485543
Auto Theft	1988	362	0.182093	0.386018
Burglary	4852	543	0.111913	0.315291
Murder	18	17	0.944444	0.235702
Rape	450	238	0.528889	0.499720
Robbery	829	276	0.332931	0.471547
Theft	26673	3611	0.135380	0.342136

In this table, size is the total number of crimes committed, sum is the total number of those crimes that were cleared, and the mean is the proportion of the total crimes that were cleared. We conducted multiple t tests to see if the means were the same, but even using the more strict bonferroni significance threshold we found that the means were all very different. We see that different crimes not only happen in different areas, but they are handled differently, and some may be easier to solve than others.

In the future, one thing to think about is that maybe a larger number of crimes happened in neighborhoods with lower income because there were just more areas with lower incomes here. It would be beneficial to do a study to see how the amount of people living in the different income areas is distributed as well. So we can see if the crimes are actually happening just as much in less populated richer neighborhoods.