Project 2 - CS5830
Caden Maxwell, Jack Agren, Peter Kurtz

Jupyter Notebook: GitHub
Presentation: Google Slides

Introduction

In this project, we investigated four questions with relation to crime. Our goal was to determine how certain variables impacted crime in order to understand when and where they occur. Analysis like this is important for agencies that work to stop crime, rehabilitate those who commit crime, and help potential victims.

The questions are the following:

(1) Is there a difference in the number of crimes between seasons?
(2) Does median household income of a certain area related the amount of crime committed?
(3) Is the percentage of affordable housing (or rental housing) related to the amount of crime committed?
(4) Does the proportion of the severity of crimes committed change by median income in an area?

Results and Methods for Question 1: We found that there is a significant difference between winter and fall. We suspected that the variation in temperature affects the crime rate. In order to find this difference, we conducted an ANOVA test.

Result and Methods for Question 2: We found that there is a correlation between median household income and amount of crimes committed. We suspected that there would be a negative correlation between income and crimes and this was correct. We used a Pearson correlation test to find the correlation.

Result and Methods for Question 3: Using Pearson's correlation test, we found that there is a moderate positive correlation between the percentage of affordable housing and the total crime rate, and a moderate positive correlation between the percentage of affordable housing and burglary rate . When looking at rental housing, we found a moderate positive correlation to the total crime rate, and a high positive correlation to burglary rate.

Result and Methods for Question 4: Using a correlation test, we found that minor crimes become a larger proportion of the crimes committed in a given area as the income for that area increases, while the proportion of more severe crimes decreases.

Dataset

We used two datasets in our analysis. The first dataset is a combination of demographic data in the Austin, TX area and crimes committed in those same areas in the year 2015. There were 42 different attributes for each crime, including prosecution status, category, highest charge, and date reported. The data for each zip code primarily deals with income and affordability of housing and transportation. The second dataset has population densities and population for each zip code. This was used to find the crime-per-capita for a given zip code.

Analysis

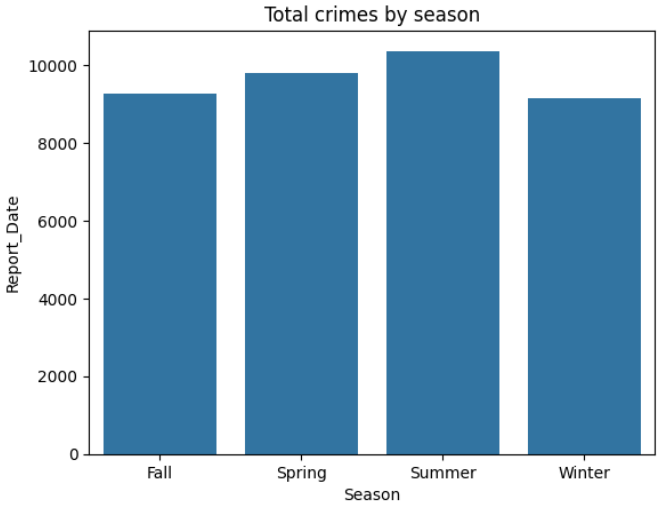
All of these analyses used were relatively simple statistical tests to find evidence for each of our hypotheses. If the tests succeeded, then we could conclude that our hypotheses were likely valid; otherwise, we failed to get any evidence for that hypothesis. Statistical tests can tell us if 2 variables are related or if 2 groups are different, which is useful for determining how and where crime happens, or, just as important, what things do not impact crime rates.

Results

Question 1:

We found that there is a significant difference between the number of crimes in the summer vs winter, but not between other combinations of seasons. Fig 1.1 shows the total number of crimes in each season, and it can be seen that winter and summer do have a reasonably large distance between them. The average crimes per day for winter was 101.68, and for summer was 110.2. Reasons for this could include perpetrators not wanting to be out in the cold, heat making people more irritable, or an increase in tourists during the summer providing easier victims. Further analysis would be needed to determine if these variables do impact crime rate.

Fig 1.1



Question 2:

We found that there is a correlation between median household income and amount of crimes committed. Fig 2.1 is a scatter plot of the median household income by the crime per population. Each point in the plot represents an area for a zip code and crimes per population commited. Note that there is an extreme outlier in the data. This outlier had a zip code of 78701 which is near downtown Austin. We hypothesize that since it is near downtown, not many people live there and that a city's downtown area has a lot of crime that occurs or that it is heavily enforced. We then decided to take out the outlier in this analysis. Fig 2.2 is the scatter plot without the outlier.

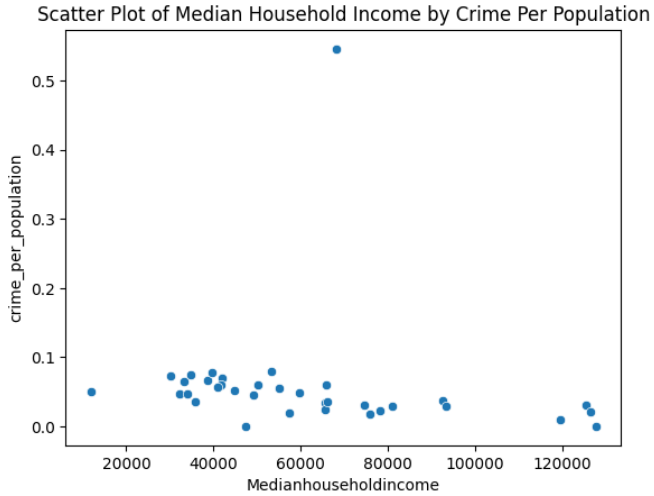


Figure 2.1

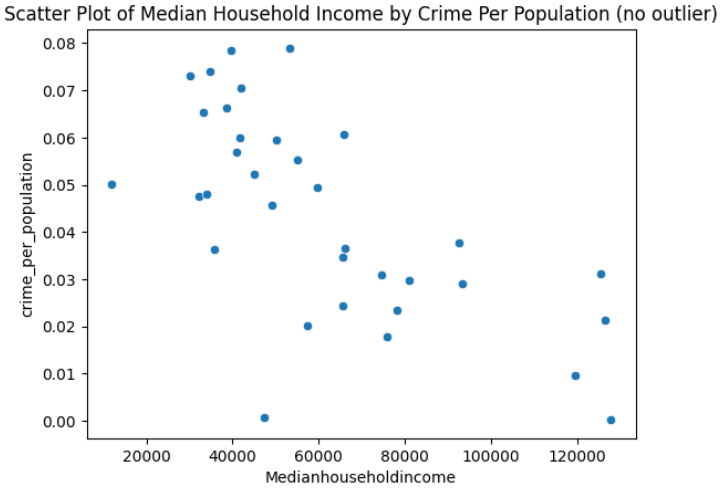


Figure 2.2

We then conducted a Pearson's correlation test of the dataset without the outlier. This gave us a Pearson's correlation statistic of -0.6583 and a p-value of 0.00002. Because of this p-value, we concluded that there is a correlation between crime per population and median

household income. We further concluded that the correlation is negative which means as the household median income goes up, crime goes down. This gives us reason to believe that low income areas tend to have more crime. This fact can help us know where to allocate resources for low income areas so that people won't need to turn to crime.

Question 3:

To investigate the relationship between housing affordability and crime rate, we used data from both the population density dataset and the crime dataset. Namely, we used the population of each zip code to get the crime rate per 1000 people, and compared that against home and rental affordability in different ways. For each potential relationship, we used a Pearson's correlation test to get a Pearson correlation coefficient and a p-value.

First, we looked at the housing affordability for those with a salary of less than $50,000. Against the total crime rate for each zip code, we saw a statistically insignificant Pearson correlation coefficient of -0.031 (p-value of 0.86). However, when looking at the scatterplot, we noticed the same obvious outlier zip code 78701, downtown Austin. We reasoned that there is likely a high rate of theft there due to tourism, which may contribute to a large percentage of the total crime for that zip code.

Upon removal of this outlier, we saw a statistically significant Pearson correlation coefficient of 0.46 (p-value: 0.0061), which means that there is a moderate positive correlation between housing affordability and total crime rate.

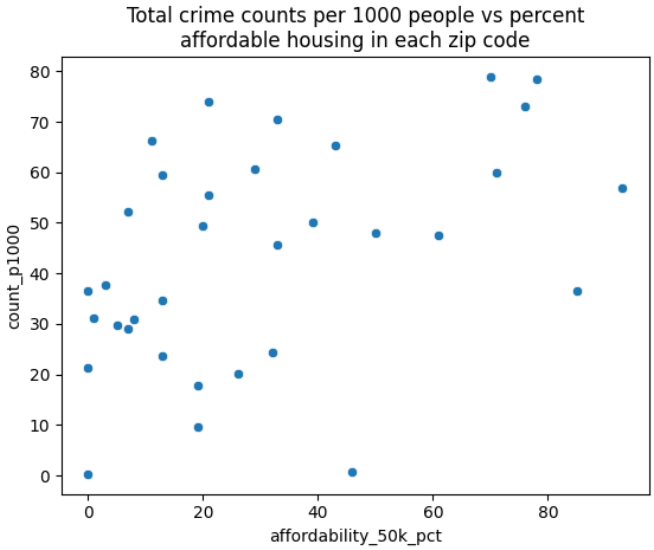


Figure 3.1

Similarly, we found a moderate positive correlation between rental housing affordability for those making under $25,000, with a Pearson correlation coefficient of 0.557 and a p-value of 0.00062.

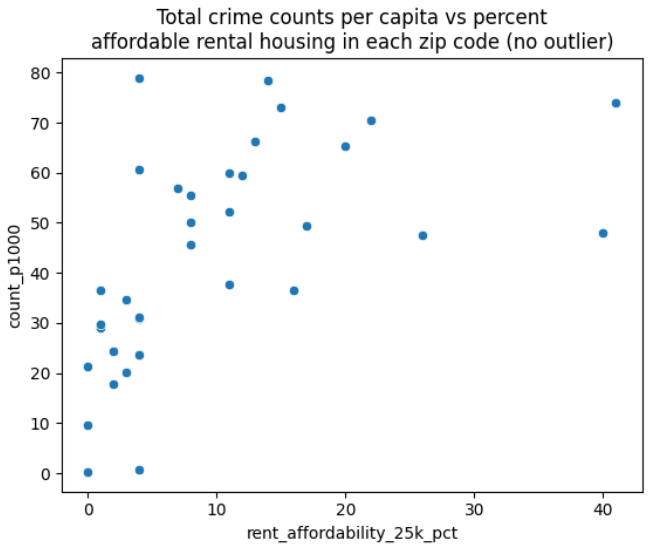


Figure 3.2

While looking at both housing and rental housing affordability, we thought it might be interesting to view specific crimes. Most notably, burglary rate per 1000 people was highly positively correlated with rental housing affordability, with a Pearson correlation coefficient of 0.69 and a tiny p-value of 0.000017.

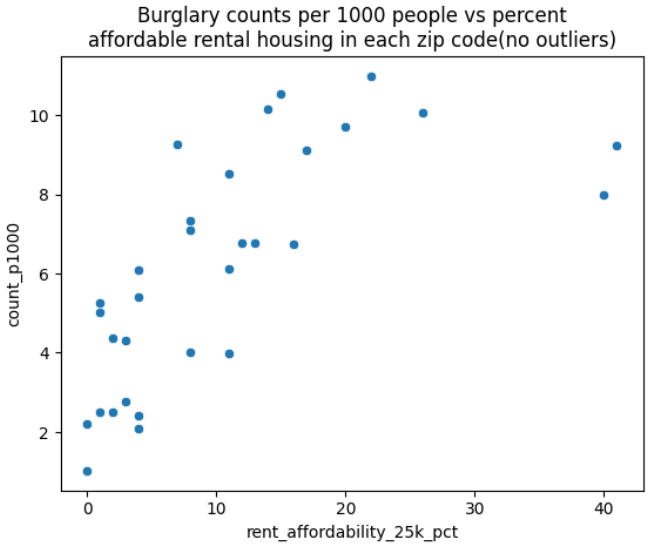


Figure 3.3

These results are all very interesting, as we initially hypothesized that a higher housing and/or rental housing affordability would ensure a lower crime rate in the area. However, as seen in these tests, it is clearly the opposite. It's hard to say *why* this is happening, but one reason may be that a higher percentage of housing affordability may signify a population with a low average income and likely a high proportion of individuals in poverty or economic stress. It can be reasonably presumed that a population with these characteristics may be more inclined to commit crime out of desperation. In any case, these results are notable and may be cause for further investigation into the subject by state officials.

Question 4:

We used tests for correlation to find if the proportion of each type of crime changed with increasing income. The charts below show the proportion of each severity level vs. the median income. Looking at the plots, it seems like severity level 1 becomes a higher proportion of crimes in richer areas, while the more severe crimes become less of the proportion in nicer areas.

This initial guess is backed up by the Spearman correlation test we used (statistic=.66, p=.00001). The others looked generally negatively correlated, which was backed up by tests for severity levels 2, 3, and 5. These results line up with our expectations, as richer areas tend to be safer and have nicer stuff to steal. One interesting follow up to this would be to investigate the value of the things that are stolen. Perhaps there are more thefts in richer areas, but they are small shoplifting kind of offenses. Another would be to see if those committing the crimes are from outside the richer zip code; people who are more desperate might go for the bigger score.

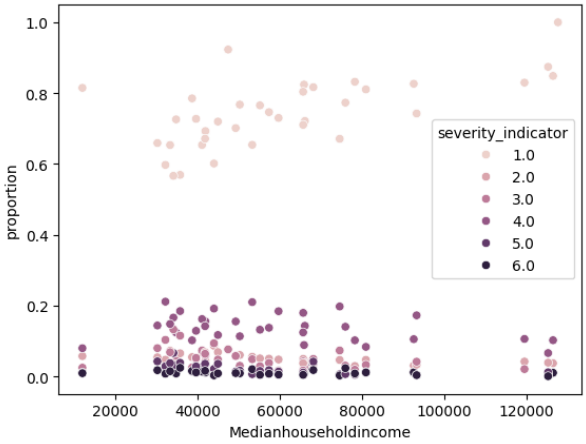


Figure 4.1

The plot below has severity level 1 omitted in order to see the other levels more clearly.

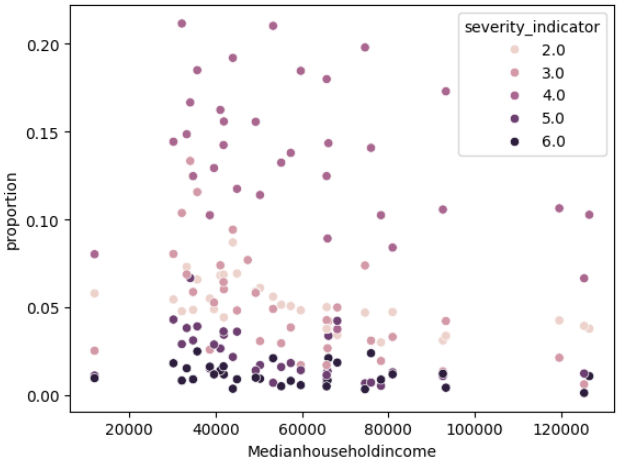


Figure 4.2

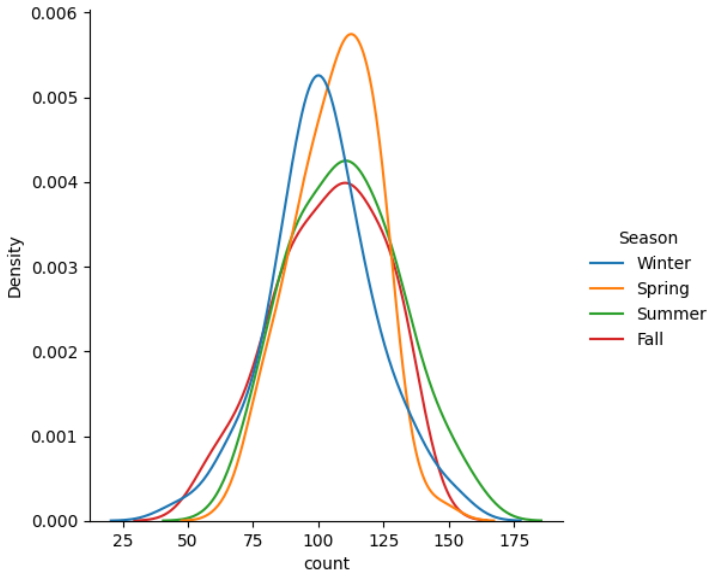
Technical

In order to conduct the analysis that we did, we had to implement the 'groupby()' function for all three analyses. This allowed us to view the amount of crimes committed in different groups such as seasons and zip codes. For two of the analyses, we merged the crime data set and the population density dataset in order to get crimes per capita instead of simply total crimes which would not have been very informative, as there can be large differences in population from zip code to zip code.

Question 1: Crime by season

To separate the crimes by season, we added a 'Season' column to the data, then created a new table with total crimes on each day and the season that day is in. A quick bar plot (fig. 1.1) showed mild differences between the seasons, but it wasn't clear if it was significant or not. To test this, we did an ANOVA with the season as the category and found that at least 1 season is likely different from the others ($p=0.023$). Then, we used Tukey's HSD to find that winter and summer were significantly different ($p=0.017$). It's worth noting that the ANOVA is valid here since each category is the same size, each category's data are about normally distributed, and the variances are about the same (see plot 1.1).

Fig 1.1



Question 2: Crime per population by minimum household income

We had to format the data so that each area with zip code would be associated with the number of crimes committed in said area. To do this we used a group by statement. Then we used a merge in order to pair the population densities with the correct zip code from the other dataset. We also had to change the column population to a numeric value. We also removed an outlier that is described above in the Results section.

We chose a Pearson correlation Statistic since the scatterplot showing crime per capita by median household income was linear. The main complication with this analysis was making sure the data was correctly formatted and that we used the correct variable to find crime per capita. No alternative approaches were obvious to us.

Question 3: Affordable housing vs. crime committed.

In order to get the data into a form that was easily usable in our correlation tests, we had to do a lot of wrangling and some cleaning. First, we dropped all columns aside from the zip code column and the housing and rental affordability columns. We got rid of duplicates, got the crime counts by type for each zip code by performing a 'groupby()' on the original dataset, and merged that into our affordability dataframe by zip code. Finally, we merged the population column from the population dataset with our affordability dataframe and created a new column with crime counts by crime type for each zip code.

For the two of our analyses regarding affordability and total crime rate, we had to perform another 'groupby()' on the dataset by zip code, rental affordability, and housing affordability to get a sum of crime counts for each zip code, then performed a Pearson correlation test, which showed that both housing and rental housing affordability were positively correlated with crime rate per 1000 people. We deemed the Pearson correlation test as a suitable analysis technique since it gives us both a p-value and a correlation coefficient, which are very useful in deciding *if* and *by how much* the data was related, respectively. Scatterplots were suitable for data visualization in this analysis since we were looking for a relationship between two groups.

We did make a mistake in the wrangling process for this analysis. When comparing housing affordability to crime rate, we forgot to perform 'groupby()' on the crime types to sum their crime counts, so we were actually comparing all different types of crime and their counts together on the same scatterplot, which gave us a very uncorrelated and uninteresting results, since each type of crime happens at very different rates (theft vs murder, for example).

With this analysis in particular, we think we could have dug a little deeper into why the correlations were the way they were, instead of leaving it largely up to interpretation. We should have looked at other socioeconomic and demographic factors to uncover why those zip codes with lower incomes had a larger crime rate than others.

Question 4: Median income vs. severity of crimes committed

This analysis required us to decide which crimes were more severe than others. We decided to use the maximum sentence possible for each crime, and give them each an ordinal rank. We had several options for what exactly to test here, but in the end we decided to simply check the correlation for the proportion of each severity level with respect to the median income of the area. Since some areas have the same median income, we took those areas and averaged their proportions for each severity level.

Since one of our variables was ordinal, the Pearson correlation test wouldn't work well here, so we used a test called Spearman's rank correlation (read more here), which is

non-parametric and can account for the ordinal data. In addition, since we were doing several tests with the same variable, we needed to account for potential false positives, so we took p>.0083 as significant. Below are the results for all tests:

Severity: 1 statistic=0.6604890604890606,pvalue=1.1673972681492266e-05
Severity: 2 statistic=-0.7197860962566844, pvalue=1.6032240034257506e-06
Severity: 3 statistic=-0.615478647987722, pvalue=0.00010698657347025012
Severity: 4 statistic=-0.33017570664629486, pvalue=0.0565079955493883
Severity: 5 statistic=-0.6337976539589442, pvalue=9.845207963691642e-05
Severity: 6 statistic=-0.24303519061583576, pvalue=0.18014315430197464