CS 6830- Data Science Incubator

Maanav Choubey, Lexy Simmons, Jasper Swensen

Project 2 Report

Introduction

Our focus for this project is on housing and crime statistics for Austin, Texas, in 2015. In our analysis of these data, we look at a number of different things, such as the top 20 crimes and how often they happen, the zip codes with the highest crime rates, how crimes are spread out across Austin on a map, and the relationship between arrests (via clearance_status) and median household income. In doing so, we hope to gain a deeper understanding of crime in Austin and its trends. We mainly wanted to focus on the audience of anti-crime organizations. This way we can help them find out what crimes are most common in which areas. Github Presentation

Dataset

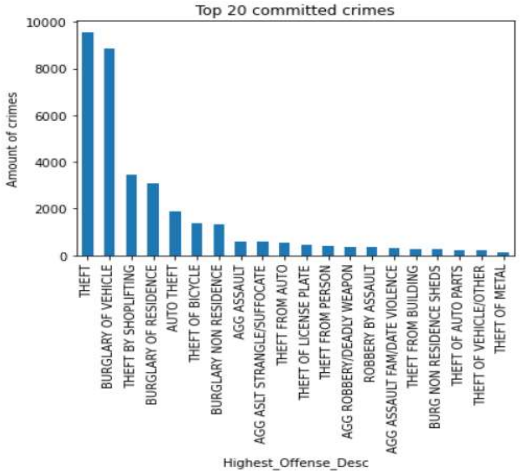
Our data set has a lot of information about crime and housing in the city of Austin, Texas, in 2015. We have access to data on the types of crimes committed, the locations of the crimes, and the median income levels of the perpetrator's zip code. By analyzing this data, we aimed to gain insight into the geographical distribution of crime as well as its distribution across different income levels.

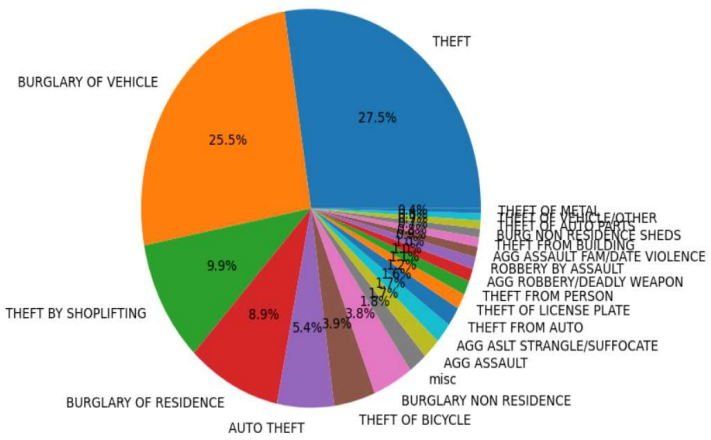
Analysis technique

One of the things we decided to investigate was the frequency of the top 20 crimes so that we could plot them on a map. We decided to look at the zip codes with the highest crime rates so we could learn more. This helped us gain a better understanding of the neighborhoods where people resided and the crimes that were committed there. Finally, to get a complete picture, we mapped the distribution of crimes throughout Austin, allowing us to see where the highest crime rates were. Our crime markers now include information about the specific offenses that have been committed, and we've clustered them together on the big picture to show us something about the area; these clusters shift as you zoom in and out to examine different neighborhoods. We also looked at how often people were arrested (as indicated by clearance_status) in relation to other variables, such as average income. Here, we used the dataset's "Medianhouseholdincome" and "Clearance_Status" attributes to determine how many people in each income bracket had each clearance status type, and then we counted them all up in a dataframe. Once the data was prepared, using pandas' in-built functions made it easy to generate a scatter plot showing the correlation between different types of clearance status and income, as well as to calculate the Pearson coefficients for these correlations.

Results:

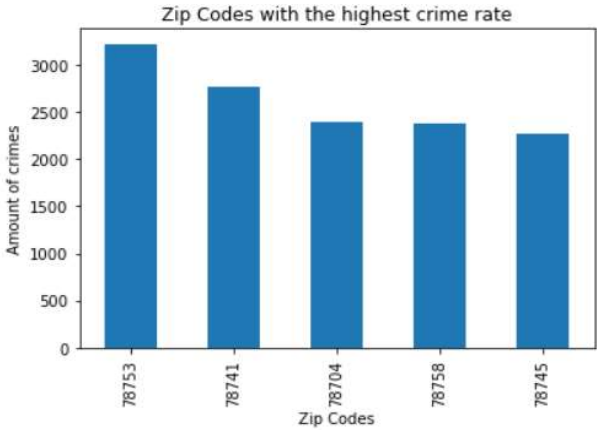
The top 20 crimes committed and their frequency

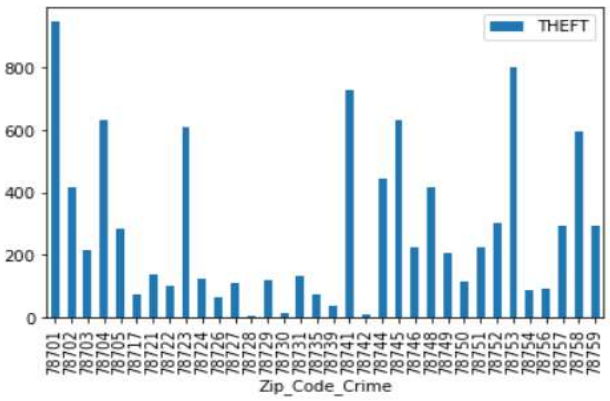




When we plot the data on a map, these visualizations will help us make sense of where crimes are happening.

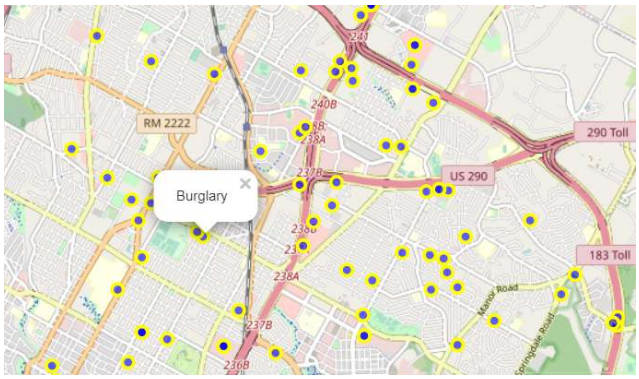
Zip codes with the highest crime rates

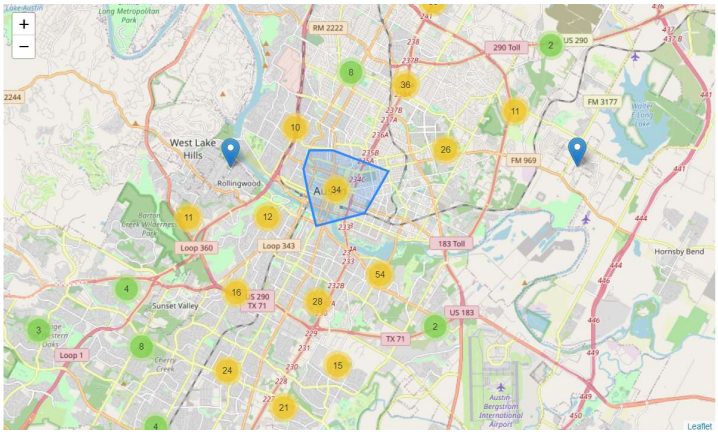




First, we look at the five zip codes where the most crimes are happening. We also look at the zip codes where the most common crime, theft, happens and how often it happens. Visualizing it on a map helps us get a clearer picture of the whole situation.

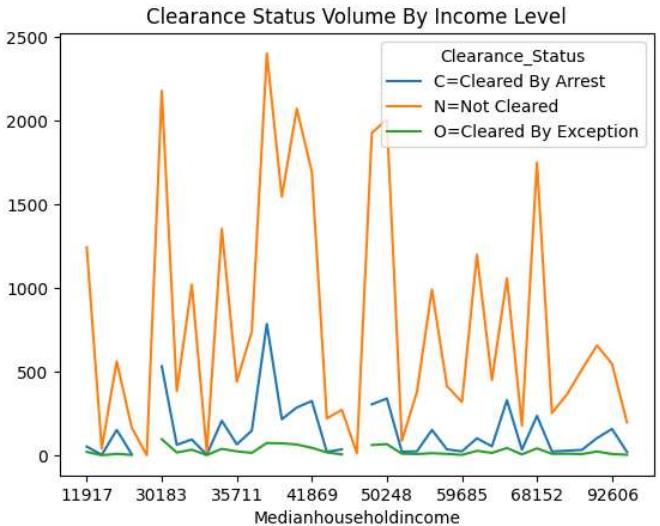
The mapped distribution of crimes across Austin

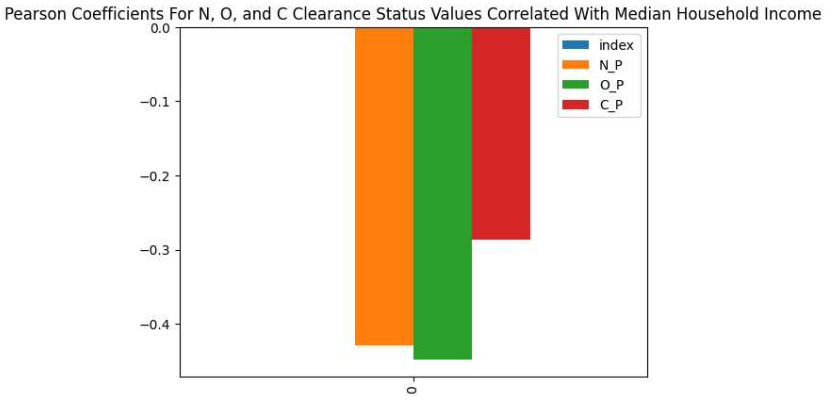




When we finally plot our data on a map of Austin, each crime will be represented by a marker that, when hovered over, displays the type of offense. We also grouped these crimes together so that the number of crimes in particular blocks is highlighted in blue and changes as the map is zoomed in or out.

The relationship between arrests (as measured by clearance_status) and the median household income.





As median household income rises, clearance statuses decrease arrests. Crimes cleared "by exception" and uncleared follow this tendency. Our dataset did not include any information concerning offenses cleared by exception or not cleared at all, nor did the website from which it was sourced. The trends are there regardless of clearance status, but more information would have helped us understand them. Arrests drop when median household income rises because the Pearson coefficients for all three clearance statuses are negative.This trend could be due to a number of factors, including:

- Bias from police officers against poorer individuals
- A general decrease in crimes committed as income level increases
- Bias from police officers in favor of wealthier individuals

The difference between crime type and the median household incomes and the crime

Statistic: Ttest_indResult(statistic=-7.357881278681037, p-value=1.9224080438891666e-13)

The mean for the median household income in areas where robberies occurred: $45724.16

The mean for the median household income in areas where robberies occurred: $51343.32

We tested to see the difference between the mean of household incomes for all of the documented “robbery” crimes and all of the documented “theft” crimes. Our results indicated that there was a massive difference in the average median household income, and there was a very low, nearly impossible, chance that this statistic was inaccurate.

Technical

To plot our map we had to figure out how to convert the coordinates from NAD83 to UTM. However once we got things set up it was relatively smooth sailing and we were able to complete every analysis we had originally planned to. We believe the analysis techniques chosen were appropriate for this dataset. In general, our data was complete with few missing entries or row values. We used industry standard statistical tools like pandas to set up each analysis, and used techniques like calculating the Pearson coefficient to determine the significance of our analyses.

For each of our analyses, the process was largely the same, but varied in some aspects. Of course for each analysis it was necessary to import the dataset from the csv, create a pandas dataframe, and aggregating the various attributes we wanted to examine for that given analysis. However the process for mapping the distribution of crime was slightly different and required some additional steps including setting up folium, plotting the coordinates on the Austin map, initializing the markers and their size on the map with added labels**.** Some analyses like the relationship between instances of arrest and income level required more data wrangling than we originally anticipated to get things into a format that worked well and which pandas was happy with.

For the analyses that included income rates, we needed to process the data from a string of dollar amounts into a numeric value so that we could actually analyze that information. Additionally, some results kept putting out a “nan” result until we realized that we needed to drop “na” rows.

Presentation:
https://docs.google.com/presentation/d/1CbdqmRbrqTn-oo2hVERvN0bOnyL_532mccHlScGcXTE/edit?usp=sharing
Github:
https://github.com/lexykj/cs5830_project2