CS5830/6830 Project 2 Report

Megh KC and Hari Chandana Kotnani

Introduction

For this project, we analyzed an Austin housing and crime dataset that focuses on the individual crimes reported in Austin, Texas in the year 2015. We used python in conjunction with pandas to explore and analyze the dataset. The data can be analysed to identify areas with high crime rates and analyze the factors contributing to these patterns. In our analysis, we have examined the various factors that may influence the incidence of crime and the nature of crimes committed in a particular area. Our analysis will help local government to target their resources more effectively and improve public safety. Additionally, it can inform policy decisions aimed at reducing poverty and addressing the root causes of crime. Presentation link, Github link.

Dataset

The crime-housing-austin-2015 dataset and the Austinzipcodes dataset are two datasets that contain information related to crime and housing in Austin. The crime-housing-austin-2015 dataset has 43 columns and includes details about crimes, crime descriptions, zip codes, and housing. The Austinzipcodes dataset provides information about the population density per square mile and is in CSV format, which makes it simple to use with the pandas' library in Python for data analysis and preparation. We specifically used the zip code crime, highest UCR offence description, population below the poverty level, unemployment report date, and clearance date to obtain our analysis.

Analysis Techniques

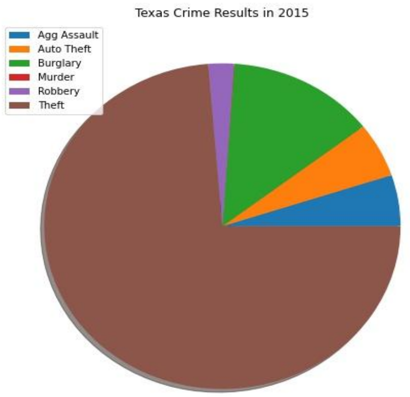
Once we selected the relevant data, we used various functions from the Pandas library, such as group by, count, mean, statistical, and aggregate functions, to prepare the data for analysis. We then applied various statistical techniques, including calculating Pearson correlations (p-value), averages, standard deviations, and t-tests, to gain insights into the data. To visualize the results, we used various plotting techniques, including regression, bar, pie, and scatterplots. These simple visualizations helped to gain valuable insights into the dataset. The various analyses performed aimed to uncover patterns and relationships in the data.

1. Which crime had the highest occurrence among all crimes reported in different zip codes of Austin?
    a. For the first analysis, we calculated the total number of each type of crime by grouping the crime data. We then created a pie chart to visualize the count of each crime type and to compare the occurrences of different crimes. This analysis aimed to determine the distribution of crime types in the dataset.

2. What was the rate of crimes committed per person in the population, also known as crimes per capita, in different zip codes of Austin?
    a. For the second analysis, we determined the crime rate per capita by dividing the number of crimes by the population in each zip code. We then plotted this crime per capita rate against the different zip codes in a scatterplot. The scatterplot showed that there was one zip code with an outlier, which was removed to obtain a more optimized plot.
3. What type of crime had the highest rate of occurrence per person in the population, also known as crime type per capita, in certain zip codes of Austin?
    a. For the third analysis, we separated the crime-type data into individual crime reports. For each crime, we grouped the data by zip code to determine the number of times the crime was reported in each zip code. We then calculated the per capita rate of each crime in each zip code. To visualize the results, we created two bar plots.
4. Is there a relationship between the population living below the poverty line and the number of crimes committed per person? Is there a relationship between the number of crimes committed and the number of unemployed individuals?
    a. For the fourth analysis, we took the poverty level data in the dataset and aggregated it by calculating the average poverty level. We then grouped this average poverty level with the mean of crimes per capita and plotted it against the poverty level. Finally, we calculated the Pearson correlation to determine the relationship between the poverty level and crime rate in the areas under consideration.
    b. Here, we grouped the crimes by zip code and calculated the average unemployment percentage for each zip code. We then plotted the number of crimes against the average unemployment percentage and calculated the Pearson correlation.
5. What is the difference in the time it takes to obtain the clearance status between two types of crime? And, what is the comparison of the number of crimes reported in two different zip codes?
    a. As a fifth analysis, the goal was to determine the difference between the means of two groups using a t-test. The two groups were based on two crime types, theft, and theft from the person, and their clearance times were calculated by subtracting the cleared date from the reported date.
    b. For the other analysis, two random zip codes were selected and the number of crimes reported in each zip code was plotted.

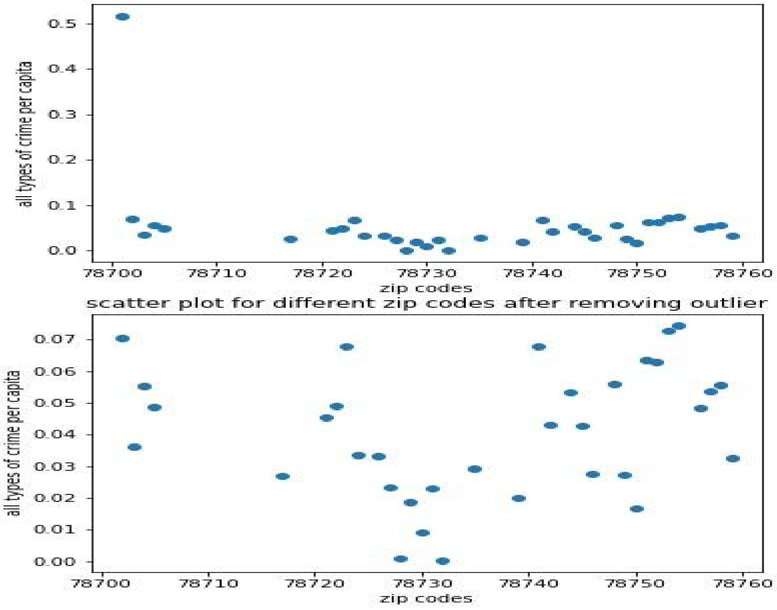
Results

Analysis 1:



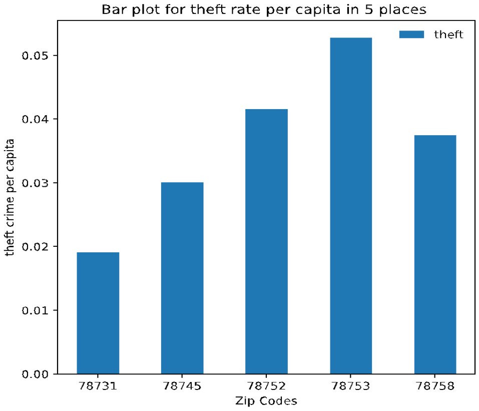
According to our analysis, theft is the most frequently committed crime in Austin, while murder is the least committed.

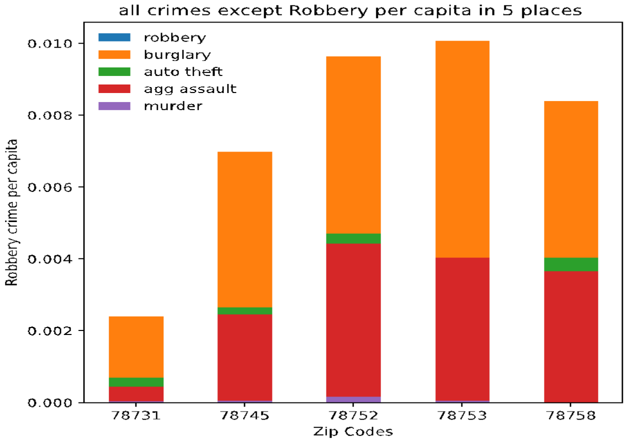
Analysis 2:



Here, the scatterplot shows a relationship between zip code and crime rate per capita. The plot showed that zip code 78701 had a higher crime rate per capita compared to other zip codes, with a rate of 0.516, indicating that one crime is committed for every two people in that area. Interestingly zip code 78701 has a smaller population in the dataset. After removing the high crime rate value from zip code 78701 as an outlier, a new chart was plotted. The resulting chart showed that the remaining values were more similar in range. The scatterplot is carefully interpreted as zip codes are considered categorical variables. A comparison of the two plots is shown above.

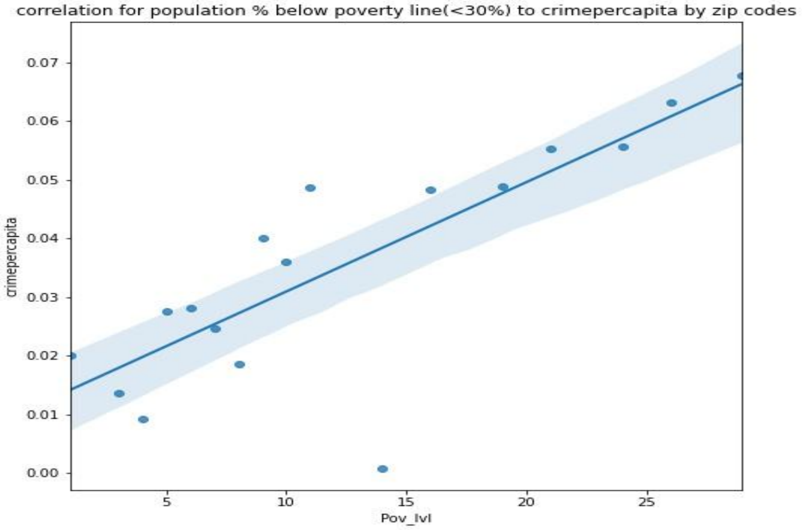
Analysis 3:





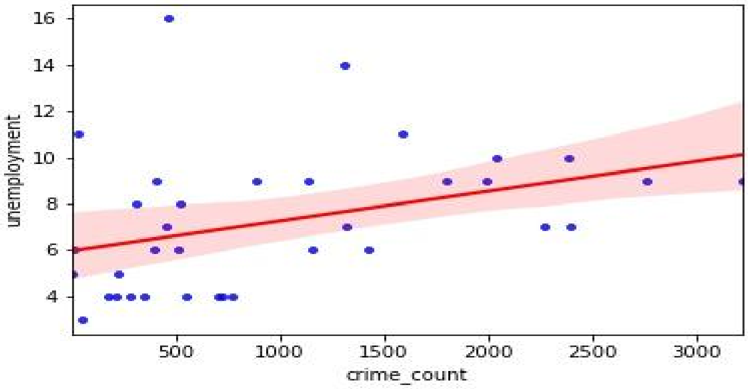
The ‘Highest_NIBRS_UCR_Offense_Description’ column which has types of crimes was divided into separate categories for each type of crime, and the crime rates per capita were calculated for all six crimes. The results showed that theft was almost ten times more prevalent than the other crimes, with murder being the least common. The analysis was conducted for Travis County, Austin, Texas, in 2015 and focused on five different zip codes: 78731, 78745, 78751, 78753, and 78758. we created bar graphs to compare the crime rates in these areas, and it was observed that zip code 78753 had the highest rate of crime for all types.

Analysis 4(a):



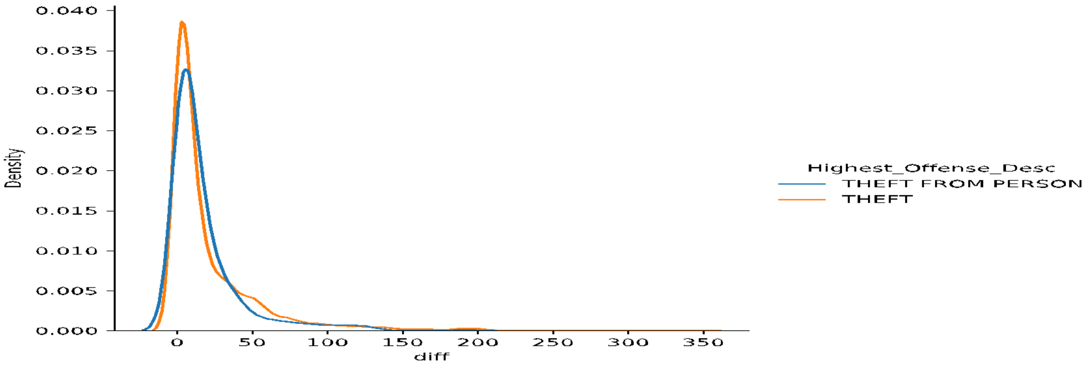
We conducted a Pearson correlation analysis between the poverty level and crime per capita and found a strong correlation (0.80) between the two. This indicates that as the poverty level increases, crime per capita also increases. The p-value (8.73e-05) supports the statistical significance of this relationship. It can be inferred that a rise in the number of individuals living below the poverty line leads to an increase in the occurrence of criminal activities.

Analysis 4(b):



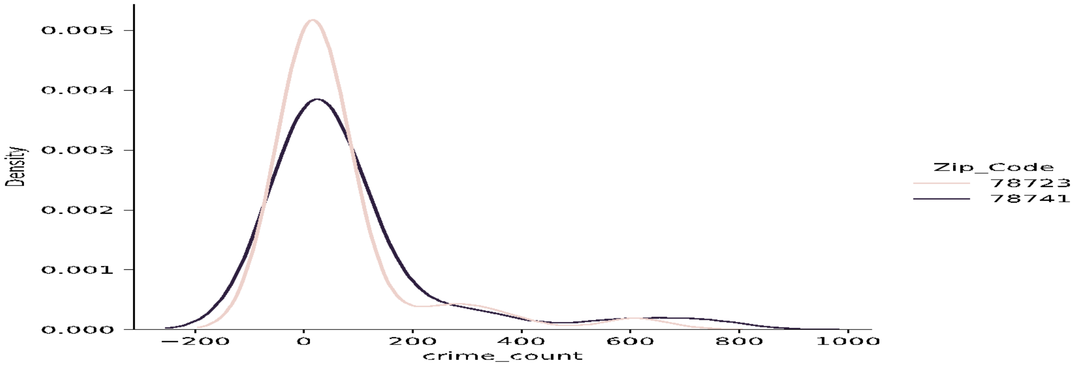
We conducted a Pearson correlation analysis between the number of crimes and unemployment means. The results of the Pearson correlation analysis showed that there is a moderate relationship (0.40) between the number of crimes and the mean unemployment rate, which means that as unemployment goes up, so does the number of crimes. The significance of this relationship is supported by a low p-value (0.02). This suggests that as unemployment increases, people may turn to criminal activities due to financial difficulties

Analysis 5:



A.

We performed a t-test to compare the clearance time for two types of crimes: theft and theft from a person. We calculated the mean and standard deviation for each crime type and found that the mean clearance time for theft was 21.3 days with a standard deviation of 33.0, while for theft from a person, it was 17.4 days with a standard deviation of 24.8. Our t-test results showed that there was a significant difference between the two crime types, with a p-value of 0.01 and a statistic value of 2.33. The low p-value indicates that there is a low probability of observing such a difference in the clearance times due to chance. Therefore, we can conclude that the clearance times for these two types of crimes are significantly different.



B.

We conducted a t-test to compare the number of crimes reported in two zip codes, 78723 and 78747. We calculated the mean and standard deviation for each zip code and found that the mean number of crimes in 78723 was 61.8 with a standard deviation of 130.0 and in 78747, the mean number of crimes was 86.25 with a standard deviation of 168.1. Our t-test showed a p-value of 0.76 and a statistic value of -0.29. In conclusion, based on the t-test results, we do not have sufficient evidence to conclude that there is a significant difference in the number of crimes reported between zip codes 78723 and 78747.

Technical

The Pandas libraries were utilized to load the dataset, remove rows with missing values, and replace '%' and '$' from selected data-frames. Some of the core tools we used to analyze the data were data grouping, summation, averaging, and other forms of aggregation. To obtain meaningful results, the crime dataset and population datasets were merged to gather information on the types of crimes committed, details related to the crimes, and population statistics in different zip codes of Austin. Descriptive statistics of the merged dataset were found by grouping the six types of crimes according to their zip codes, as the zip code provides population information that can be used to calculate the crime rate (crime per capita). The grouped data was then analyzed for crime per capita by adding a new column to the dataset. The statistics of the dataset (mean, median, mode, and standard deviation) are calculated using a built-in function in Python. In one attempt, which got failed, we tried to understand the relationship between poverty levels and non-native individuals, we combined non-white Latinos and Hispanico-Latinos and plotted a graph against poverty levels. The results of the graph seem to indicate that the correlation between poverty levels and this group is not significant and weakly correlated. This suggests that other factors may play a larger role in poverty levels for this population.