CS 5830 - Project 2

Anny Chen, Zion Steiner

INTRODUCTION

The purpose of this report is to communicate general statistical measures of interest found in the *Austin Crime Report 2015* dataset. In addition to the type and location of criminal reports, the dataset includes housing statistics associated with where these crimes took place. The aim of this analysis is to determine how crime is geographically distributed within Austin, TX, in 2015. Descriptive statistics such as correlations, means, standard deviations are used to supplement the findings of a hypothesis test. As a result, we found evidence that suggests crime is more common in the eastern half of Austin than the western half.

Presentation link:
https://docs.google.com/presentation/d/199_QlGP2VxgMai--4H5K80zQmTuex1Srj5goa65AS9o/edit?usp=sharing

DATASET

The *Austin Crime Report 2015* dataset is mainly about crime and housing statistics for Austin, Texas in 2015. This dataset informs us of the crime position, crime data, clearance status and the wealthy status of the crime position. The dataset domain of crime and housing dataset is criminal statistics. Crime and Housing dataset is a specific part of criminal statistics, which includes specific time and position crime information. The format of the dataset is a comma-separated values file. Before using the dataset, we transform percentage or dollar amounts to the float type for further analysis, and change the string type of crime date to date type.

ANALYSIS TECHNIQUE

The proportion of total reported crimes occurring in each zip code was calculated to give a measure of region-specific criminal activity. Calculate the Pearson's correlation coefficient between crime proportion and other columns in table 2, and find crime proportion and median household income is loose negative correlation. Given the coordinates for where each crime occurred, we were able to visually analyze which areas of the city experienced more reported crime. Plotting standard crime proportion side-by-side with standard median household income shows that crime is more common among lower-income areas(scatterplots of standardized data). Fig. 2 shows that the split occurs roughly at x coordinate = 3100000, with higher income to the west and higher crime levels to the east. A two-sided hypothesis test was conducted to compare the mean crime proportion between zip codes on either side of the city. With a t-statistic of -62.0 and a p-value of ~0.000, we have evidence suggesting that less crime occurs in the western half of Austin than the eastern half.

Statistics detailing the distribution of crime proportion by zip code are shown in Table 1. These show there is a spread of ~9% between zip codes with the lowest crime occurrences and those with the highest. The minimum proportion is 0.0026%, so at least one zip code experiences almost no reported crime and another experiences 9% of all crime in the city.

RESULTS

First, the crime proportion of each zip was calculated to represent regional crime severity, which is shown in Fig 1, and described in Table 1.

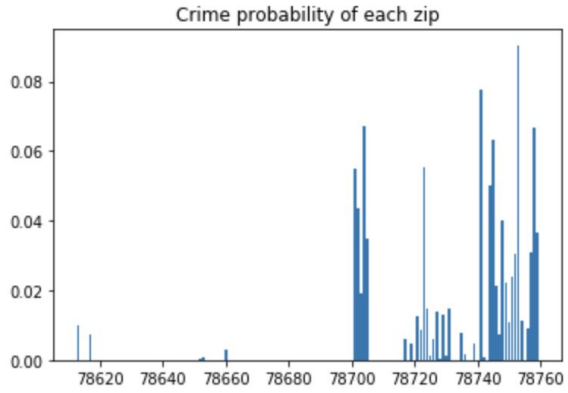


Figure 1. Crime probability varies from zip to zip

count	38573
mean	0.046828
std	0.025235
min	0.000026
25%	0.024032
50%	0.049802
75%	0.066445
max	0.090011
Name: crime_prop, dtype: float64	

Table 1. The spread between the zip codes with the highest and lowest occurrences of crime is 9%

From Fig 1 and Table 1, we observed that crime probability varias seriously from zip to zip. To figure out the related factors to crime probability, Pearson correlations between crime proportion and all other factors are calculated, in Table 2(only best six and worst two items are shown).

Pearson correlations with crime_prop	Pearson's correlation coefficient	p value
crime_prop	1	0
Ownerunitsaffordabletoaverageteacher	0.540611	0
Homesaffordabletopeopleearninglessthan$50000	0.520392	0
HispanicorLatinoofanyrace	0.518378	0
Ownerunitsaffordabletoaverageartist	0.500552	0
……	……	……
Medianhomevalue	-0.387767	0
Medianhouseholdincome	-0.401332	0
Name: crime_prop, dtype: float64		

Table 2. Pearson correlations coefficient and p value between each column(housing related) and crime proportion

From Table 2, we figured out owner units affordable to average teacher has most high Pearson's correlation coefficient with crime proportion(0.5406) and median house holding income has most low Pearson's correlation coefficient(-0.4013). In other words, owner units affordable to average teacher may have loose positive correlation with crime proportion, and median house holding income may have loose negative correlation. It is quilt understandable median house holding income has loose negative correlation with crime proportion, since poverty leads to high crime rates. However, we cannot image to relate teacher affordable units with crime proportion.

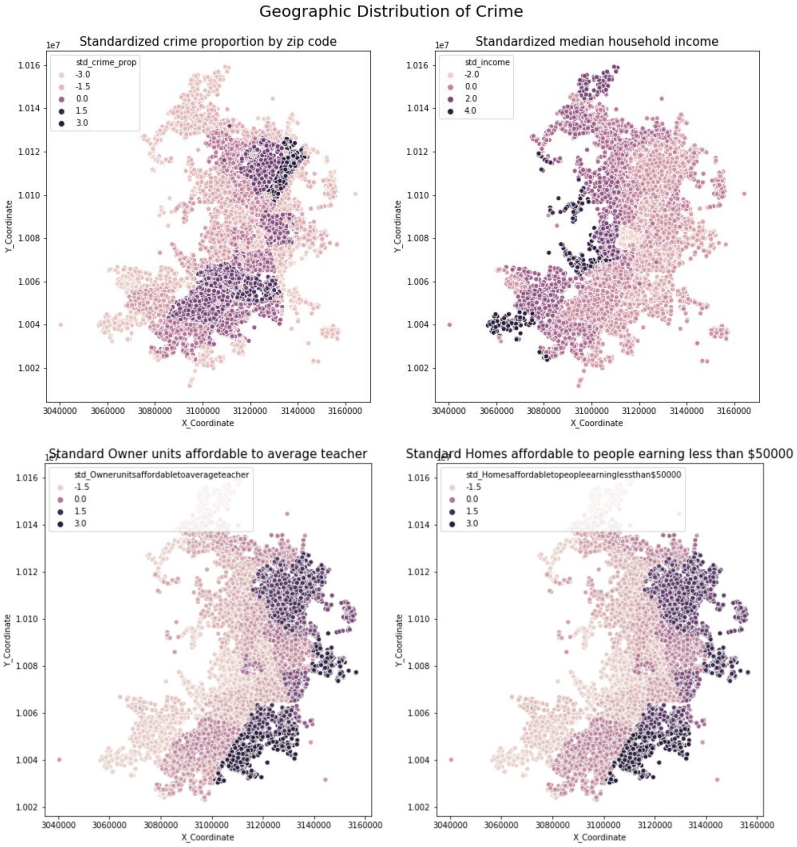
Therefore, we decide to calculate the Pearson's correlation coefficient with owner units affordable to average teacher, which is shown in table 3.

Pearson correlations with Owner units affordable to average teacher	Pearson's correlation coefficient	p value
Ownerunitsaffordabletoaverageteacher	1	0
Homesaffordabletopeopleearninglessthan$50000	0.9967	0
Ownerunitsaffordabletoaverageartist	0.929891	0
HispanicorLatinoofanyrace	0.82655	0
……	……	……
Medianhouseholdincome	-0.593866	0
Medianhomevalue	-0.724997	0
Name: Ownerunitsaffordabletoaverageteacher, dtype: float64		

Table 3. Pearson correlations coefficient and p value between each column(housing related) and owner units affordable to average teacher

From table 3, we find that homes affordable to people earning less than $50000 may have positive correlation with owner units affordable to average teacher(0.9967), and median home value may have negative correlation(-0.7250). In other words, teacher average salary is quilt low, which may explain why owner units affordable to average teacher has most high Pearson's correlation coefficient with crime proportion.

To prove our conjecture, we will plot scatterplots of standardized data, crime proportion, median household income, owner units affordable to average teacher and homes affordable to people earning less than $50000 in xy coordination(Fig 2).



From Fig 2, we can find another interesting point, which is criminal activity may be concentrated higher in the eastern half of Austin. To prove this assumption, we use t test technique. With a t-statistic of -62.0 and a p-value of ~0.000, we have evidence suggesting that less crime occurs in the western half of Austin than the eastern half in 2015.

Finally, according to the one year crime and housing dataset, we suggest the officers pay more attention to the eastern half of Austin and may pay

attention to the problem of low teachers' salaries.

Fig 2. Criminal activity is concentrated higher in the eastern half of Austin