CS5830_Project2 Report
By Logan Liddiard and Jensen Judkins

Links:

Github - https://github.com/Logsterx7/cs5830_project2
Presentation - CS5860 - Project 2

Introduction:

Through this project we as a team decided that burglary was going to be our main focus. We would like to unravel some of the hints that the data may have left us in order to decipher factors of homes and areas more susceptible to it. Within this project we decided to cover several factors that we thought may have had an effect on the likelihood of an area being burglarized. These areas include median household income, median home value, month of the year, median rent in relation to crime, and coordinate relevance to burglary.

Dataset:

This dataset was provided to us from data.world (we can assume the person who uploaded it found the dataset from public information given to the public by the city of Austin, TX). Using this dataset we are able to extract all of the necessary key data points needed for our research. This includes Median Household Income, Highest Offense Description, Median Home Value, Report Date, Zip Code Housing, Population Below Poverty Level, Location, and X and Y coordinates of the crime. This dataset includes many more points that are not used or referenced in our study that may however have unseen correlations to our research.

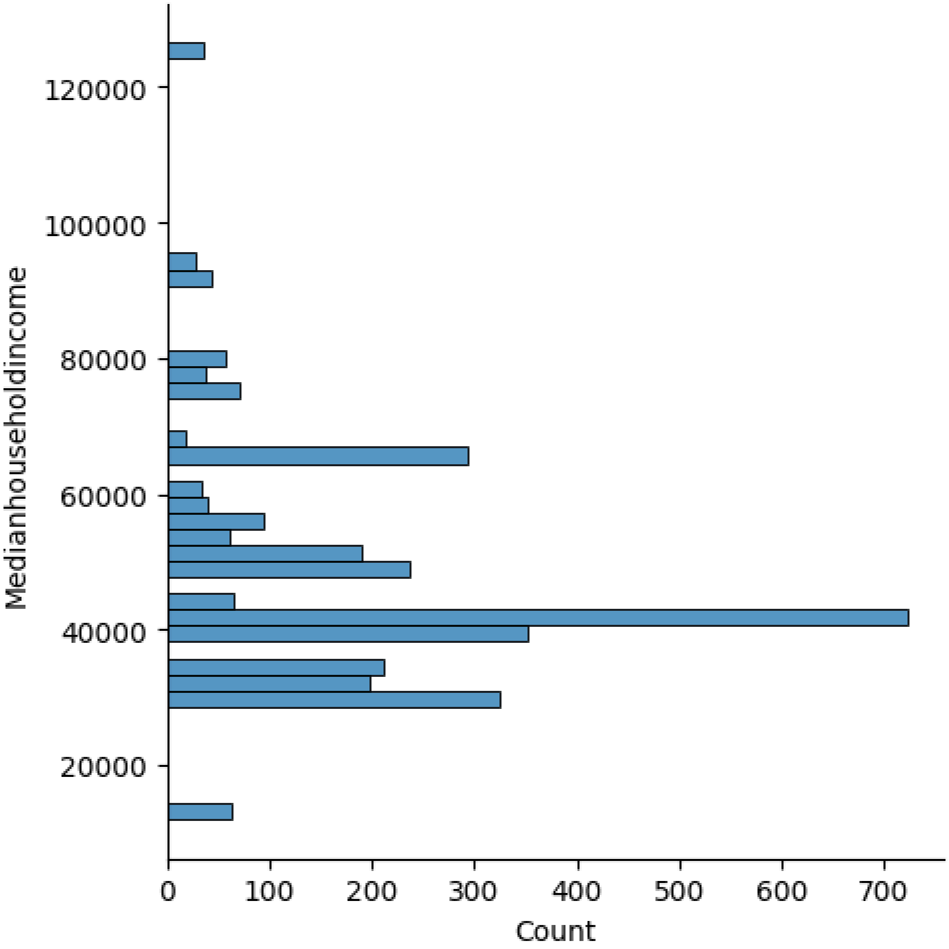
Analysis:

For this dataset there were a couple of techniques used. The first being we wanted to only see burglaries so we focused the scope of our project mostly on that. There was a lot of trial and error for getting the proper data but as time went on I think we managed to pull all of the right stuff. At one point there were a few too many outliers so we had to clean up the data by getting rid of NaN values and then finding the sweet spot to analyze deeper. Once we found the majority of burglaries were at locations between two points we mostly focused on there. Then we could broaden our horizons to something similar like theft and compare our findings. It was all very interesting to discover because we could use a T test to compare the two and find out despite being two separate crimes they correlate in location value fairly well. Then we looked at locations of crime and compared that as well by overlaying and taking advantage of scatter plots.

Results:

First Analysis: An analysis on median household income and number of burglary crimes reported.

Our first analysis is whether or not there is a correlation between the median household income of a burglarized home, and the count of those burglaries pertaining to the household income. Intuition made us believe that the lower the household income, the higher the chance of a burglarized home. We see this correlation when looking at our chart below.

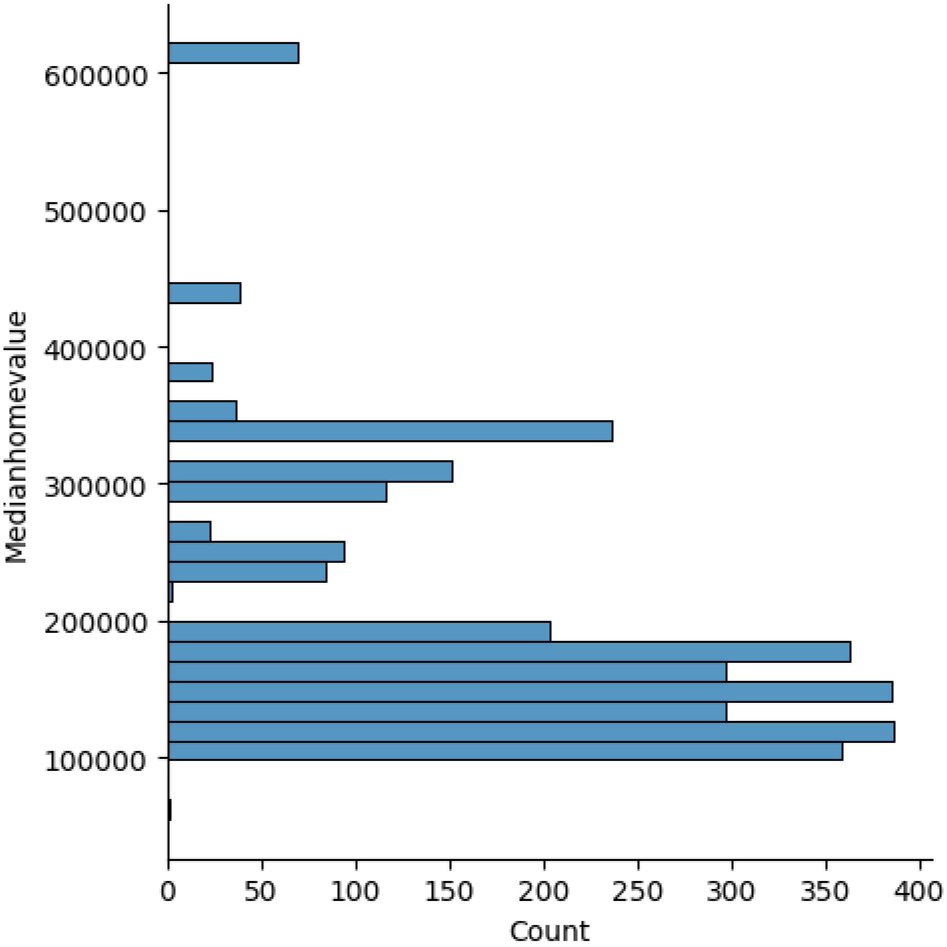


When doing further research we learned that the median household income for all of Austin Texas is $67,195 [1]. So we see that there is a large difference in the count of burglary crimes found below that $67,195 income, especially when the MHI is about $40,000. Something to note is that it is said the “magic number to live comfortably in Austin is $53,225” [2]. So this rough $13,000 missing from living comfortably may be turning some of the population that are less fortunate than others to crime, burglarizing their neighbors to make up that difference.

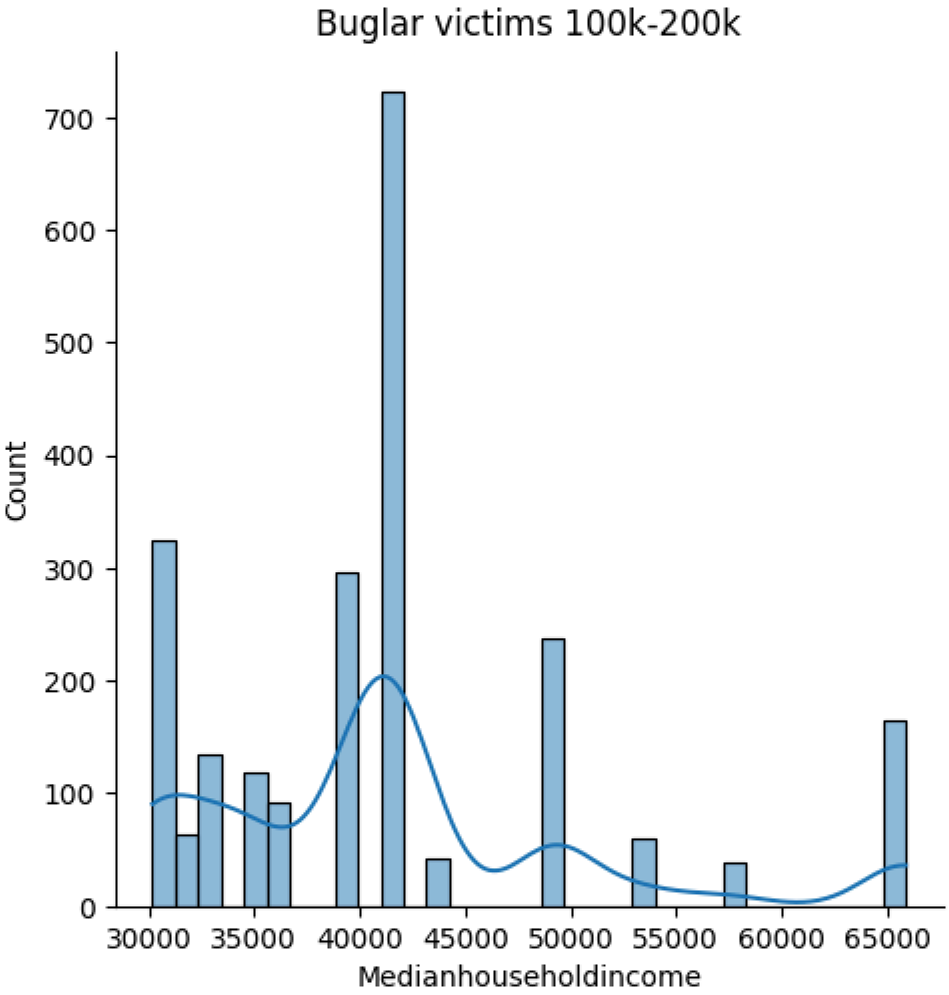
Thus we can take away that if a person's household makes less than the median household average, their home is more likely to be burglarized.

Second Analysis: An analysis on median home value and count of burglarized homes.

Our second analysis dove deeper into burglary crime by seeing if the house (rough) value played a factor in whether or not it would be burglarized. Our assumption was that the squabble would be stealing from the squabble or that houses with lower value would be stolen from more than houses of more value. We, the researchers, figured that houses that are worth more would contain security features that other houses of lower value may not have, such as neighborhood watch programs, fences, home security measures (Ring doorbells, security alarms, door alarms, better/more expensive locks). When looking at the chart below we see that there is a correlation in which houses worth between $200,000-$100,000 reported a much higher amount of burglary cases than other valued homes.



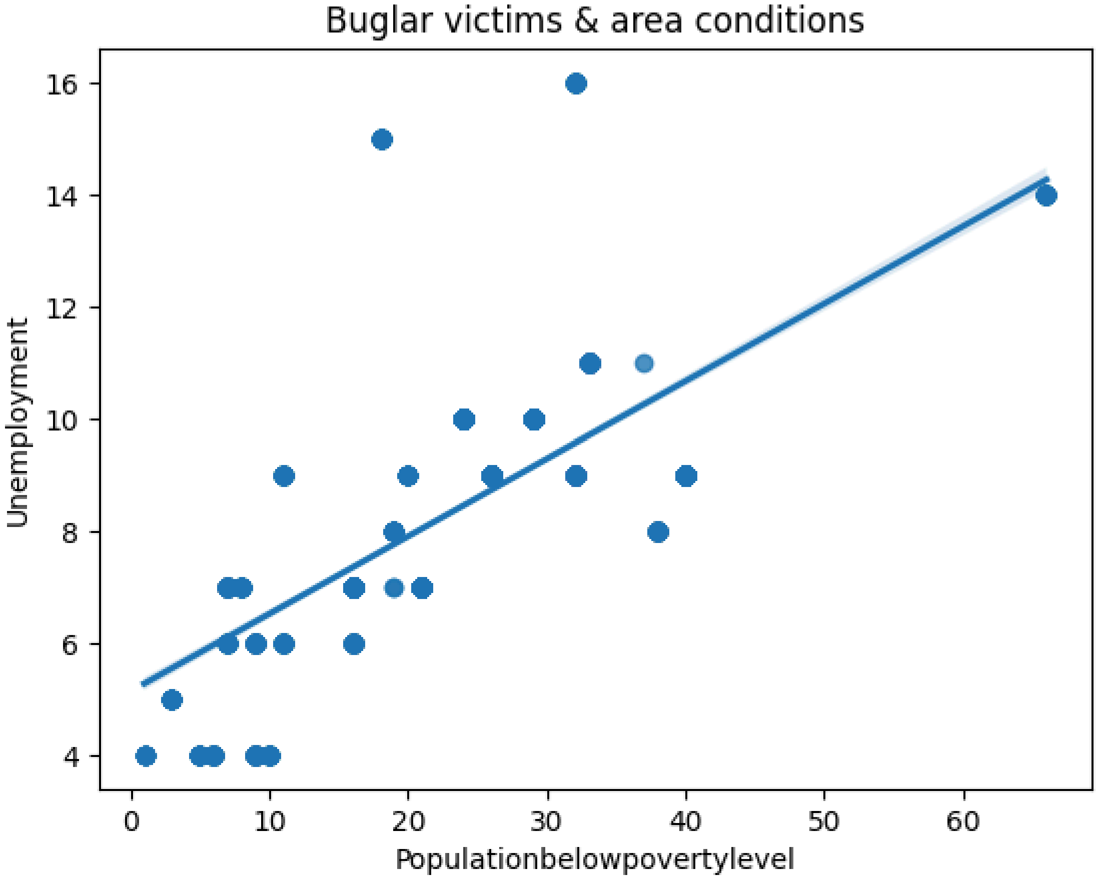
So we decided to dive in further to this since this is where the majority of these crimes are being committed. Upon closer look we decided it would be a good idea to see what the average household income of these. We figured if so many crimes are taking place in this range, then what are the odds that there is something worth stealing here? The truth is that the chart below shows that most people aren't making much. When we pulled the mean of this data we were left with a value of $41,524. Even accounting for inflation, this isn't a lot and is still under the poverty line. The data was pulled from the chart below after all the outliers were taken out so we could get a better scope on this data. Our standard deviation before pooling this data closer together was a larger value but once we honed in on this specific set it narrowed it down by almost half showing that our data set is a lot closer together when taking out more of our outliers.



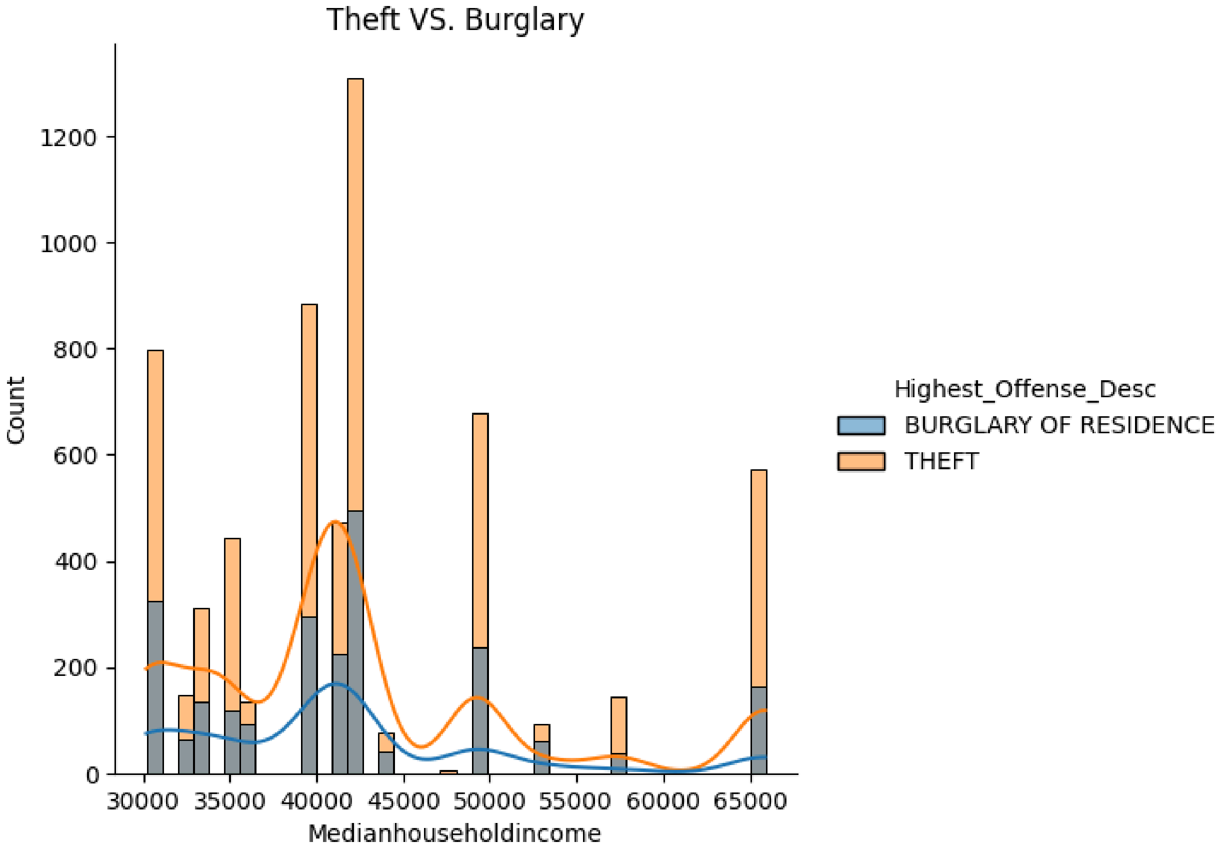
We do however still see the spike and overall increase in volume of burglarized homes being that of homes below the poverty level with a very slight uptick near $65,000 which could be accounted for as "Big ticket homes". Overall our hypothesis was confirmed by the data showing that low income homes are being exploited the most in the city of Austin and that burglars in particular may be targeting them for the reasons we explained above such as gang presence, fewer security measures, less pedestrians to report crimes, and more just due to the neighborhood the houses find themselves in.

Third Analysis: An analysis on the living conditions of burglary.

Two big questions came to mind when thinking of this analysis, and it was what is community like and how does this compare with crimes on a similar level. So by taking out our dirty data, or things that had a NaN value for our data set, we could now look at a more dialed version. Using a scatterplot as seen below we took our dataset and decided to see how its correlation between unemployment and population below poverty level looked. As you can see there is a direct correlation between the two that will have a near absolute 0 p-value. Representing a positive correlation of the two. We can see that as unemployment levels go up for that area so does the population below poverty level. Now that may seem obvious but we now have a set of data that reinforces this idea.

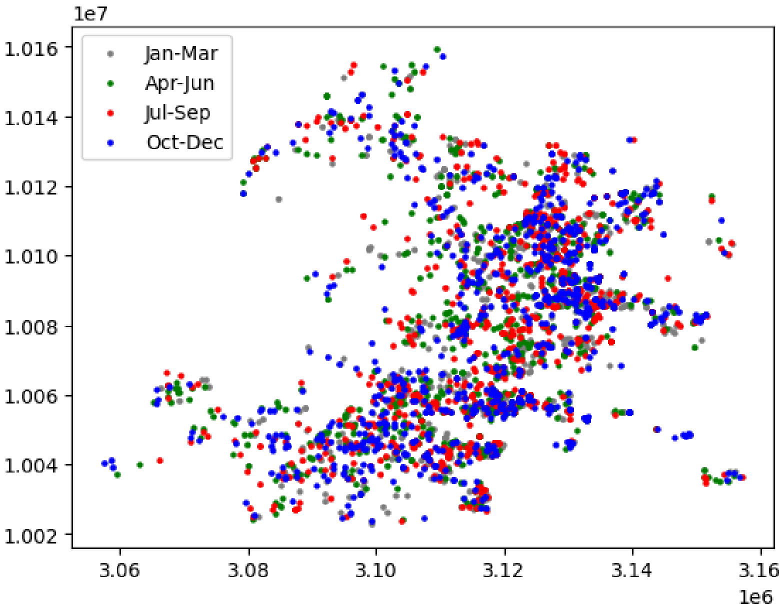


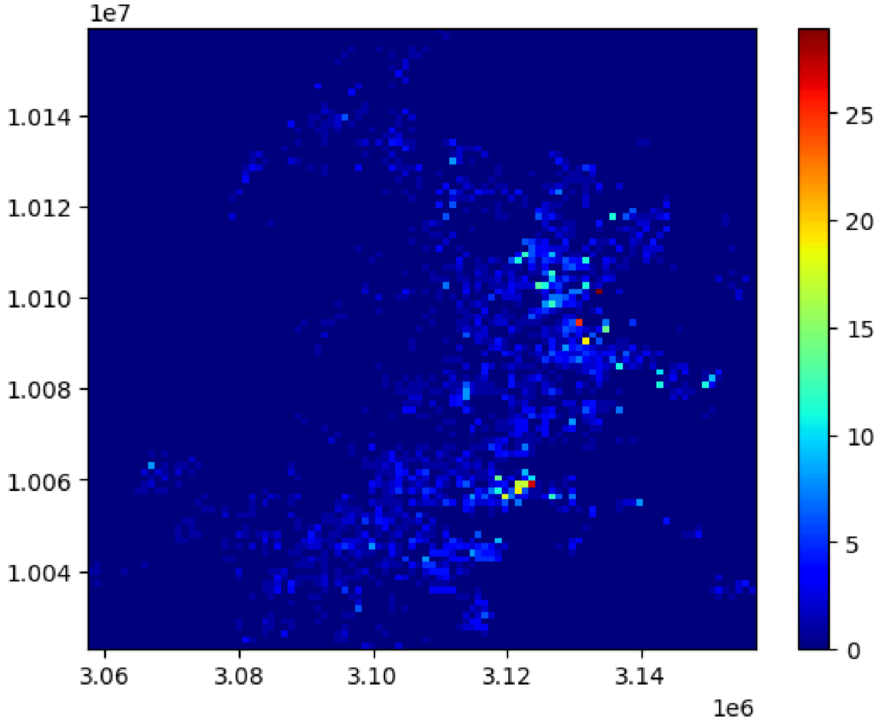
The next part which I personally found really interesting was we took a look at another crime to see if our population below poverty level and wages of these people had any correlation with another similar crime. For this the closest thing we can find in our data set to Burglary of residence is theft. Which is perfect since there was almost double the data for this crime than there was bruglary's. This is most likely because it is much easier to steal a bike sitting on the street than it is to barge into someone's home. Nonetheless, we inspected the data of thefts and found similar trends in this data set so to make things even we ran it through the same filter as before, there were a lot of thefts reported in areas with median home values of 100k to 200k just as before. The average for median home values in this dataset was $149,153 which was actually close to the same as our previous value for burglaries $145,697. When we compiled median household incomes for both of these data sets it ended up being super similar with a lot of the same trends with only minor differences. Below is an overlapping graph of the two. When a T test, which allows us to see the correlation between the two, was run on these a very strong correlation was found. Showing us that despite there being two different crimes these types of crimes are indeed being committed in areas that are below the poverty level.



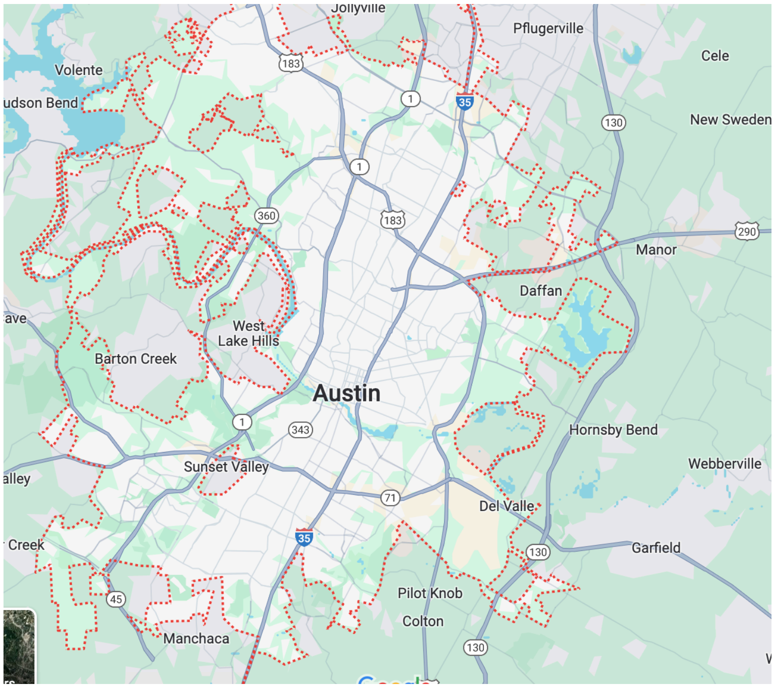
Fourth Analysis: An analysis on mapping X Y coordinates to reported burglaries.

The fourth analysis was based on the idea that certain areas, particularly those areas below poverty level, could be seen and highlighted on a graph. Thus giving Austin police officers areas to concentrate their efforts. We used burglarized homes and their X and Y coordinates to do such and put it both in a scatter plot followed by a heat map to show the concentration areas. In addition to that we broke down one of the graphs to be color coded to the time of year that those houses were 'hit', to see if certain areas became more popular to burglarize during certain points in the year.



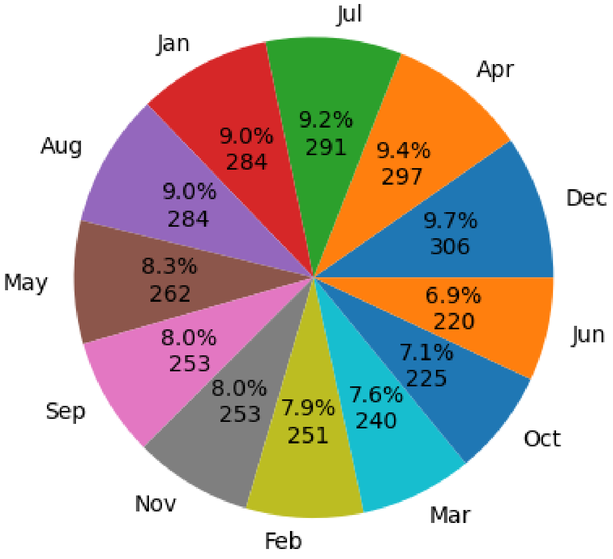


What we end up seeing is that there is little to no correlation between the month of the incident and the area the incident was located in. But what we do find in the second chart is areas more likely to be burglarized, and that correlates to suburban areas. Lastly something we noticed was the big blank spot of little to no crimes to the west of the city. It turns out this is actually a mountain range with little to no housing and that is why there are so few data points found there. Just a little interesting tid bit that we otherwise wouldn't have ever seen.
(Below is a map of Austin for reference confirming this)



Fifth Analysis: An analysis on whether or not time of year correlates to chances of a house being burglarized.

Our last thought on trying to find patterns that would indicate the likelihood of a home being burglarized was the time of year. Both researchers having watched Home Alone, we figured that the holiday season would see a large increase in the number of homes burglarized due to the fact that many people travel during the holidays, thus leaving their homes unattended.



However as we can see with the chart there is nor marginal difference between each month. All staying between 200 and 300 burglaries per month with the exception of December which thus confirmed our hypothesis, but this very well could have been an exemption due to the year the data was taken. If we wanted to confirm this further we would need multiple years of data to compare but as for now there seems to be very little correlation to month and odds of being burglarized.

Technical:

For data preparation we had spitballed a few ideas back and forth. Once it was settled that we wanted to mostly work with burglary data we decided it was time to get our hands dirty. It was easy to filter them out at first but without the crutch of the first assignment, it was a little harder to navigate. I ended up opening the cvs file manually with vim and browsed the raw data that way to explore columns, what values looked like, and what to do with that data. Once I felt comfortable enough I could start to actually look at it. There was certainly a lot of trial and error. We should have read the assignment requirements more thoroughly first because we ended up having to redo a good portion of the project. Once we found the actual data we needed it was easy to display averages and means and such and from there we thought we needed to focus on house values, income, burglaries, location, poverty level percentages and such. So we cleaned up NaN values, then after displaying on our code that there was too wide of a distribution we focused on the 100k-200k mark.

As far as analysis's a lot of it was to make graphs and analyze it. We would often look at the graph we would make and ask ourselves. Well, does this tell us anything? And if so, what? Other than requirements that needed to be met, it was totally necessary to use scatter plots and look at means. It helped us see not only where things were happening, but we were able to find solid correlations with data that way. A lot of it felt that it came up naturally after playing around with it for a while. The process was to play around, find something interesting, latch on to it, and then run tests. All while asking ourselves if it made sense and was interesting as we went along. We found it overall super interesting to find this data and to see the correlations using real techniques such as looking at p values and t tests, checking averages and such. Our result was reached by finding a thread of information and then following it to the end. Trimming off excess elements that didn't help us find what we were looking for and eventually finding evidence for everything.

Citations:

[1]: https://www.deptofnumbers.com/income/texas/austin/
[2]:
https://austin.culturemap.com/news/city-life/04-22-16-income-salary-needed-to-live-in-austin/