Crime in Los Angeles

2020-Present

Ledy De La Rosa
Computer Science
University of Colorado
Boulder
Boulder,Co,USA
wide7181@colorado.edu

Ledy De La Rosa
Computer Science
University of Colorado
Boulder
Boulder,Co,USA
wide7181@colorado.edu

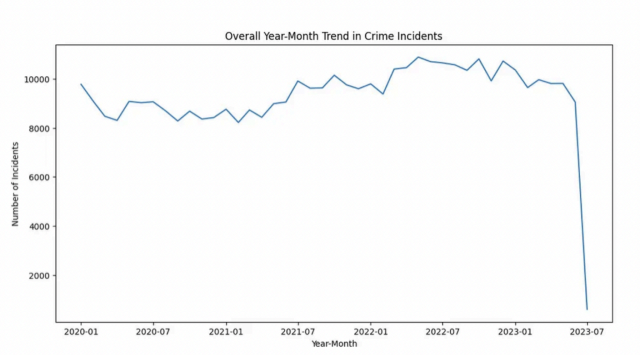
Ledy De La Rosa
Computer Science
University of Colorado
Boulder
Boulder,Co,USA
wide7181@colorado.edu

PROBLEM STATEMENT

The data that I chose to work with is about crime in Los Angeles, California from 2020 to 2023. The data has many different attributes including, crime committed, victims age, victims sex, and victims race. The primary reason why I chose this dataset is because I wanted to do research on a topic that genuinely interested me. I have always been fascinated in the criminal justice system, specifically how crimes are investigated and how we can gather data to come up with better solutions that could keep our communities safe. My data has great attributes with useful information regarding the victims of these cases. I think we can answer many questions such as which race may experience a certain crime more, does age or sex affect the likelihood of being a victim of a crime. If we are able to link a certain crime or a higher predisposition of being a victim of a crime, we can potentially use this information to come up with strategies to protect these individuals. For example, I know throughout history we have seen warnings on the news for dangerous killers or serial rapists on the loose who target a specific population such as females with long hair or young boys who are white. Even recently in New York City we had a case of a young male going around the city punching the elderly and the community was on high alert caring for their older relatives.
If we can find a relation between any specific attribute and the type of crime committed, I believe we can have a great impact on keeping our streets safer. Overall, the main goal of my project is to identify which crime affects who the most specifically, which sex, which race, and what age *to help come up with strategies to ensure the safety of our communities.*

LITERATURE SURVEY

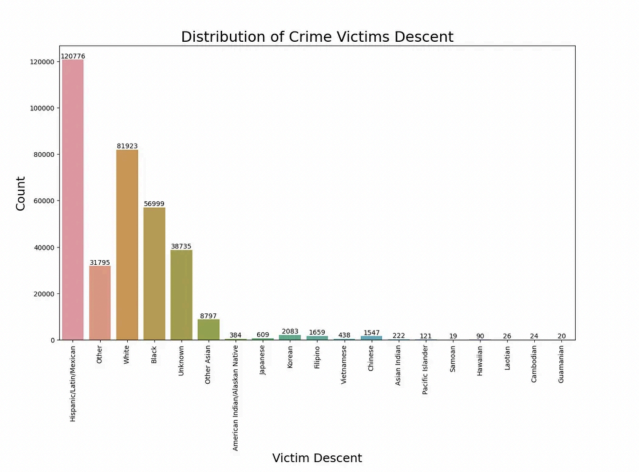
Since crime is a topic of great interest by any city or country, it is expected to have multiple articles and analysis of the data I am using. For my specific project, I will be focusing on two of the multiple currently available which are: Case Study: LAPD Crime Data from 2020 to Oct 2023 Analysis using Python by Ronaldo Pangarego and Exploratory Data Analysis: Los Angeles Crime (2020–2023) by Nguyenphantuan. I chose these two articles because they utilize a very similar dataset as I am using but each with a different focus. One focuses on criminal behavior patterns while the other closely mirrors my project's goal of identifying vulnerable groups while also focusing on questions such as why is crime occuring and when.
The case study performed by Ronaldo Pangarego has valuable information I can use to further answer my questions such as trends in crime. He provided the following chart which shows the crime trends from 2020 to 2023:



The case study concluded that they " confirmed this analysis with distribution of crime incidents each year below. Specifically, there is a 5% increase in incidents from 2020 to 2021, and 11% increase from 2021 to 2022." This is important information because we can assume that the 5% from 2020 - 2021 doubled in 2021 to 2022 due to covid restrictions going down. We were living through a pandemic which caused most businesses to shut down and people to stay home which would have resulted in lowering of crimes compared to the following years. This is crucial if we want to understand our crime trends and provide an explanation to inconsistencies in our findings.

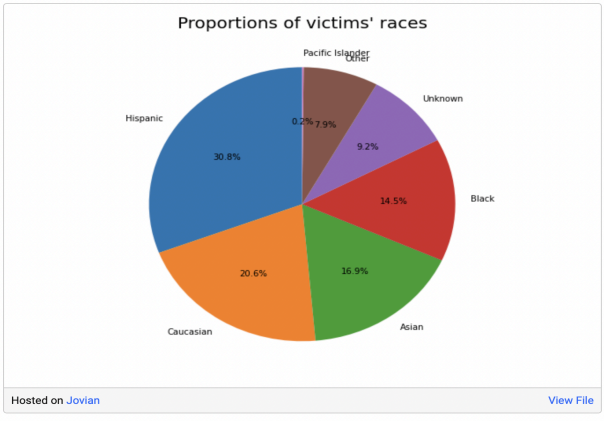
The article provides crucial information that I can use to broaden my search and get a better picture of criminal patterns in Los Angeles. One of the topics he talked about were the days of the week and how on certain days, Friday and Saturday specifically, tend to be the day most crimes are occuring while Sunday and Tuesday reported less crimes. This can be beneficial for my project since my dataset has an attribute that shows the timestamps the crimes occurred which can help us solidify our findings. He also reported that the most common crimes in Los Angeles were found to be vehicle theft, battery assault, and identity theft which can be used to look for any correlation between these popular offenses and their victims.

The most important aspect of this case study are the findings on how crime affects different ethnicities and sex which are illustrated below:



The study concluded that hispanics have the highest number of victims with white as its second. It is important to note that 120,776 crimes involved a hispanic victim while 81,923 were white. The study also concluded that 48.6% of the victims of these crimes were male, 41.3% were female and 10.1% were unknown while the average age was 27.

The study conducted by Nguyenphantuan gives us a different perspective on our topic. The data used for this analysis consisted of 28 different attributes including demographic information about the victim and crime incidence with more than 650,000 entries. This analysis provides great information which will be of great addition to my project. The most important information I gathered from his report is the correlation between crime and race. The following chart was provided to illustrate his findings:



Here we can see that his results match appropriately with our previous report since once again we see a larger risk for hispanics and whites compared to the rest of the races. He also mentions that male and females tend to be at the same risk for crimes which will be great information to integrate when answering questions about sex and crime. I plan on utilizing these findings while also trying to further answer questions such as whether or not females are at higher risk for sexual assault than males. The data is only showing the number of crimes reported per sex group without indicating which crimes most affect each sex, leaving room for further research.

PROPOSED WORK

I plan on integrating additional data to achieve more concrete findings. I am also working on using the same data I reviewed as prior work since they focused on the same timeframe and location as I do. I also plan to do dimensionality reduction to remove any attributes I do not seem to find useful for our project such as "division of records number" and the district the crime was reported to. I also plan on replicating many of the methods used in prior analysis of my data such as python Pandas. This is a great tool for dissecting the data and easily extracting useful information such as the average age of victims and their sex. Lastly, I plan on visualizing data using Tableau. I found great tutorials on youtube that go into great details about the many different tools Tableau has available for visualizing data which I think is one of the greatest ways to convey a message clearly and efficiently.

DATASET

The dataset I'll be working with is Crime_Data_from_2020_to_Present.csv which can be found at https://www.kaggle.com/datasets/middlehigh/los-angeles-crime-data-from-2000/data?select=Crime_Data_from_2020_to_Present.csv. The data consists of 28 unique attributes and 944,000 rows. Some of the attributes are victims age, victims sex, race, timestands, location of crime, crime committed, etc. which is ideal for my project since I'll have enough information to develop strong theories and more definitive results.The data has been downloaded and uploaded to jupyter notebook where I have been working on applying and practicing multiple techniques I've found useful to use.

EVALUATING METHODS

I plan on using cross-validation to evaluate my methods. This involves using previous work and comparing their findings to my results. Like mentioned before, there are multiple data analysis reports available online that closely mirrors my topic of interest. Therefore, I believe this approach would be ideal for me. I also plan on cleaning the data and making sure any noise or outliers are removed to increase the validity of my results.

TOOLS

For my tools, I plan on using Pandas and Tableau. Pandas will be great when working with statistics and manipulating my data. This will facilitate my data mining since I can use this tool to efficiently have access to more interesting findings which is what data mining is all about! I also plan on using Tableau for data visualization. As stated above, data visualization allows the reader/audience to quickly get a sense of how the data is looking without having to go into great detail. Sometimes it is more effective to show a well structured and designed chart than to have bullet points.

MILESTONES

I plan on having the data entirely cleaned and new data integrated by the deadline of project 3. I am also working on having some type of interesting knowledge that I can report about my data by this time as well.