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

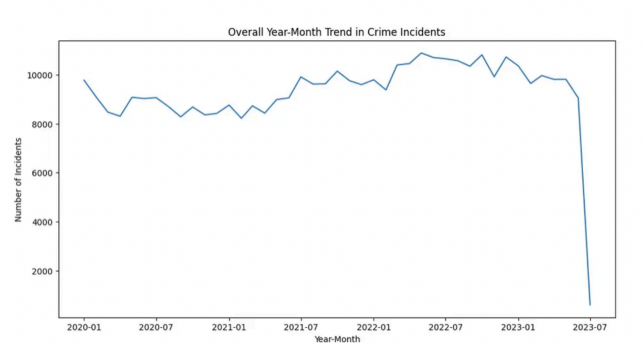
The data that I chose to work with is about crime in Los Angeles, California from 2020 to 2023. The data has many different attributes including, crime committed, victims age, victims sex, and victims race. The primary reason why I chose this dataset is because I wanted to do research on a topic that 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.

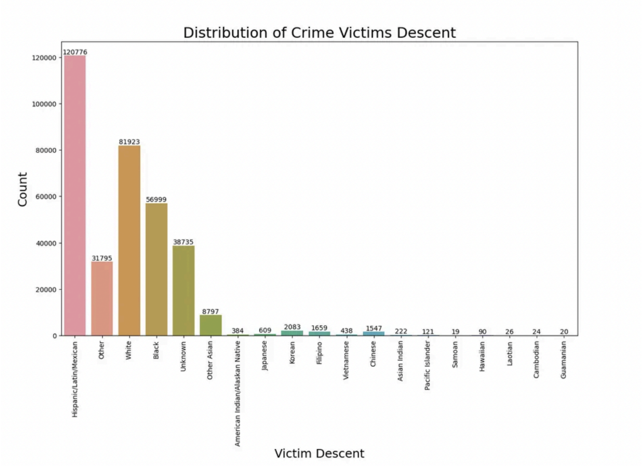
The case study performed by Ronaldo Pangarego has valuable information I can use to further answer my questions such as trends in crime. He provided the following chart which shows the crime trends from 2020 to 2023:



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.

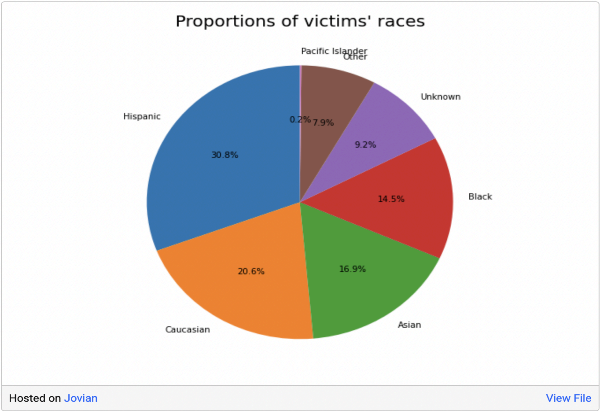
The article provides crucial information that I can use to broaden my search and get a better picture of criminal patterns in Los Angeles. One of the topics he talked about were the days of the week and how on certain days, Friday and Saturday specifically, tend to be the day most crimes are occuring 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 Microsoft Power BI. 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 Power BI 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 COMPLETED

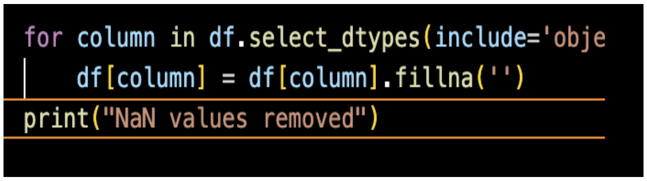
I had planned on having the data entirely cleaned and new data integrated by the deadline of project 3. I had also planned on working on having some type of interesting knowledge that I can report about my data by this time as well.

I am currently working on the data using visual code and Python. I ran into some technical difficulties since I had to redownload visual code to a new Laptop I purchased. I was not familiar with pandas before working on this project. I had worked with it in the past but only on Jupyter notebook or simple programming assignments so this took me some time to learn and get adjusted to.

The first thing I started working with was cleaning my data. At this moment, I removed any attribute that will not serve any value to my final report or will not answer any of my questions. The attributes I removed consisted are the following: Area, District number, Part 1-2, Crime CD, Mocodes, Status Desc, Crime cd 1, Crime cd 2, Crime cd 3, Crime cd 4, Status, Status Desc, Latitude, Longitude, Premis, Cross Street, Location. After removing these attributes, I am left with only 11 that I believe will provide the best insights to what I am searching for. The rest of the attributes focused more on identifying the district and criminal codes. In addition, I removed any duplicates from the data to increase accuracy of my findings.

One of the challenges I am currently facing are NaN values. One of the attributes in my data is weapons used. I believe we can use this information to better understand how violent or personal the crimes are. For example, a murder case that involves a knife instead of a gun is more personal. However, the data is very inconsistent and has many values with no entries or NaN values. I already removed them by using the following Python code:

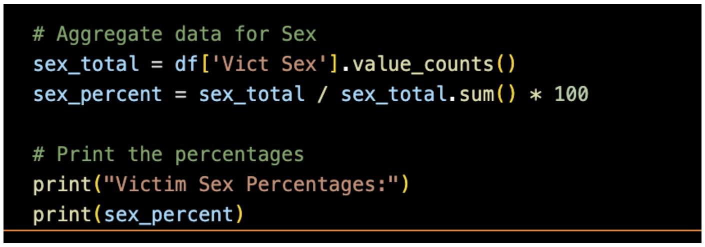
for column in df.select_dtypes(include='object').columns:



```
for column in df.select_dtypes(include='obje
    df[column] = df[column].fillna('')
print("NaN values removed")
```

This worked as intended but I am currently deciding whether I should continue with the data provided or integrate new data with clearer information on weapons use.

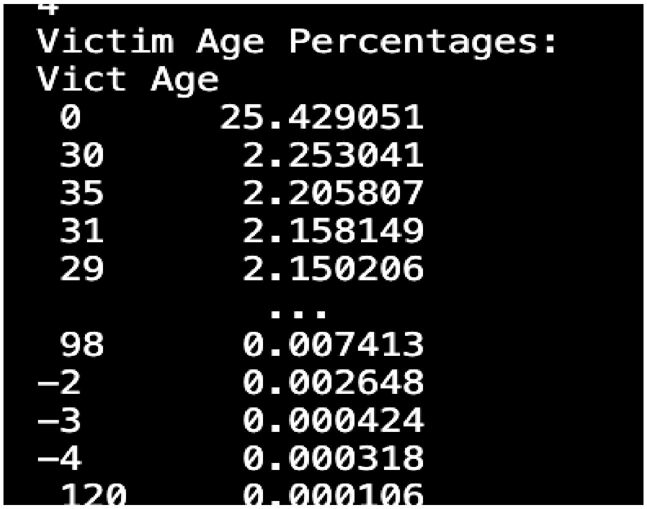
I am also working with data aggregation. I started with the three most viable attributes in my data which are Victim's age,sex, and descent. I approached this step using the following code:



```
# Aggregate data for Sex
sex_total = df['Vict Sex'].value_counts()
sex_percent = sex_total / sex_total.sum() * 100

# Print the percentages
print("Victim Sex Percentages:")
print(sex_percent)
```

This works as intended but I do not get my desired outcome at the moment. For example, when I run this code, my results are the following:

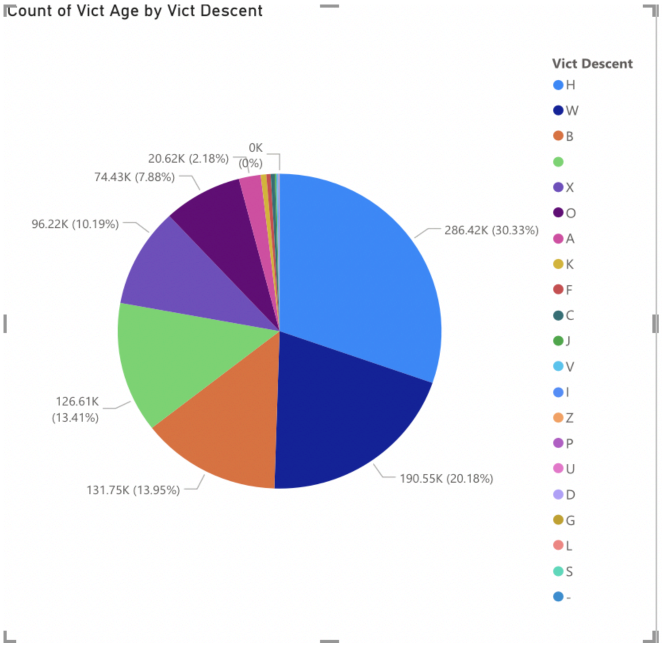


```
4
Victim Age Percentages:
Vict Age
 0        25.429051
 30        2.253041
 35        2.205807
 31        2.158149
 29        2.150206
             ...
 98        0.007413
-2         0.002648
-3         0.000424
-4         0.000318
 120       0.000106
```

Here we can clearly see that I have negative values for age. I have to further clean the data and make sure the values are appropriate for what I am trying to measure. This has been one of my challenges since the data is so large I can not see all the values myself. However, I have found amazing resources on youtube where data analyst/data scientists cover these topics in detail which I plan on learning more about to solve this problem.

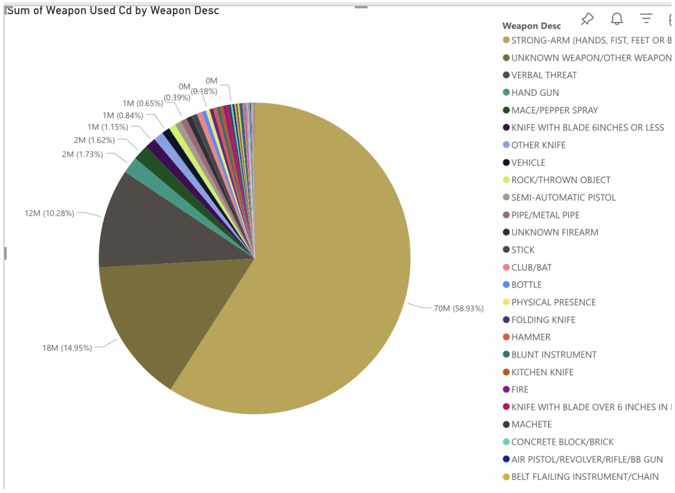
For data Visualization, I decided to work with Microsoft Power BI instead of Tableau since it was available for free as a student. I started exploring the many features and tools available. So far I see why visualizing data is so important. Even though my data is not entirely clean, I uploaded it and created a few charts to practice and just from doing this I found many weak points in my data. One of these points was found under the 'Date occurred' and 'Date reported' columns. I selected a clustered bar chart to show the differences between the two using 'date occurred' on the y axis and 'date reported' on the x axis. Instead of a chart, all I could see were unique symbols/characters on the screen. After doing some research, I learned this is because the values in my data may not all be under the same format and cause data corruption. I am currently working on fixing these issues.

Below is a chart that illustrates the total number of victims for the different ethnicities:



Even if this data is not entirely clean, we can still see that it is closely mirroring the data I reported under the Literature work section of this paper. This leads me to believe that I will have similar results to some of the research that has been conducted before on similar data.

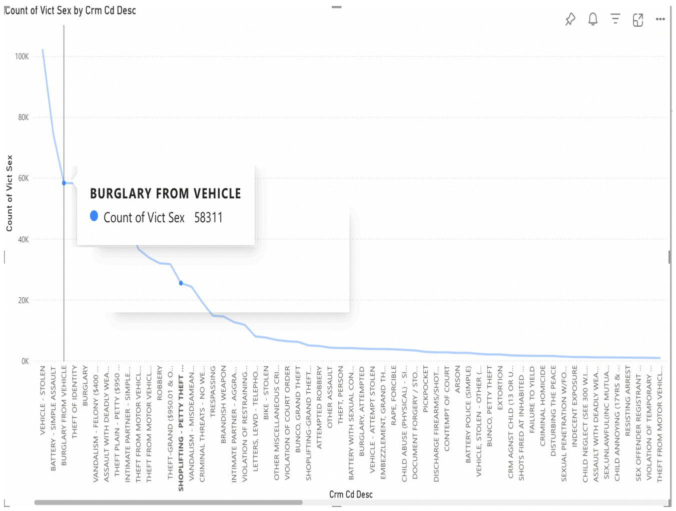
I also tried creating a pie chart with the 'Weapons Used' attribute since I already know I had some issues previously with NaN values. The results surprised me since I got a pretty clear picture of the sum of weapons used with strong arms (Fist, hands, feet, etc.) being at the top with 58.93%. Here is the chart:



MILESTONES TO DO

There are still a few things that need to get done to finish my project. One of these issues is properly cleaning the data. Like stated above, one the challenges I am currently facing is making sure the data is clean. I understand that I should start getting comfortable with working with large datasets and not expect to see every single entry for the data that I work with. However, because this is my first time working with data, I am stuck on finding the values I need to remove. For example, I know I have NaN values under my “Weapons” column because I printed out the first few values of my dataset and I saw it but for other attributes such as ‘date occurred’, I only noticed the entries had different formats when I uploaded the data to a chart. Because of this, I plan on using Power BI to help me find the outliers since that is what has worked best for me so far.

Another part that I am working on is getting comfortable with Power BI. I have certain ideas on what I want the charts to look like but I am not getting these results. For example, I have been trying to create a bar chart showing the correlation between specific crimes and the number of females versus males who fell victim to these crimes. However, Power BI shows the total number of victims of the crime regardless of sex. Here is the chart for reference:



While I was expecting to see something like “Trespassing, F: 23456 M: 27839”, I am getting “Trespassing Count of Vict Sex 68876”. After conducting some research, this could be due to having the column as a measure not a dimension but the version I am using of Power BI does not allow me to change these features. There could be a way but I am working on figuring that out at the moment or using a different interface.

Lastly, after all the other steps are complete, I still need to do the interpretations of my results and find answers to my questions. I also need to dive deeper into the data to find interesting patterns like figuring out whether or not males are more at risk of being victims of murder than females.

RESULTS SO FAR

I currently do not have any concrete results. I have certain expectations based on what I’ve seen so far such as hispanics being the number one target of most crimes and male and females having percentages close in value when evaluating how vulnerable each sex might be to crime in general. I will continue work on the data with the goal of finding interesting discoveries.