# Crime in Los Angeles

2020-2023

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

**ABSTRACT**

This report will cover the foundational knowledge I gathered while analyzing crime data on Los Angeles from 2020-2023. The main reason why I chose to work with this specific dataset is because of my great interest in helping our communities using data analysis. I firmly believe we can find the answers to most of our questions regarding the safety of our neighborhoods and citizens if we take a closer look at what is right in front of us, the immense amount of data available today.

The main goal of my analysis is to find patterns and trends within Los Angeles that could be useful in many different areas such as crime prevention, law creation, police force placements within neighborhoods, and to caution individuals. The way my report is set out, I have four different clusters or categories specifically with questions regarding all four. The categories are crime,victim, weapon and rape. I decided to focus on rape specifically out of all the other crimes because I would like to help the women in our communities by finding concrete statistics that we could potentially use in the future to decrease the chances of another female being raped. I hope to contribute to this goal by evaluating my results and coming up with reliable solutions.

My data consists of 10 attributes which are: date reported, date occurred, time occurred, area name, crime committed, victim's age, victim's sex, victim's descent, weapons used and location where crime occurred such as store, bus or school. All these attributes are key sources to the overall success of this report. I plan on utilizing and analyzing the data to generate a better picture of what truly went down in Los Angeles during these years. The questions I will be focusing on are the following:

Crime:

- Which crime has the highest average?
- During which month do we see the highest crime rates?
- Which area experiences the most crime?
- Which hour is most prone to crime?

Victim:

- Which age group experiences more crime?
- How does the victim's descent affect their experiences with crime?
- Which area do females experience the most crimes?
- Which sex has a higher predisposition to crime?

Weapons:

- Which area has the highest number of crimes involving a weapon?
- Average number of weapons used
- Which age group is most likely to be a victim of a crime involving a weapon?

Rape:

- Which hour rape occurs most often?
- Which area we see rape happening the most?

- Which month/area has higher averages of rape?

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.

**INTRODUCTION**

I carefully selected all my questions to reflect my real intentions with this project. I have always been interested in criminology in relation to crime prevention. One of the questions I wish to answer is which crime has the highest average? I believe it is important to know which crime specifically is disturbing the community. This is truly important because if we don't understand what our neighborhood is experiencing, we can not provide the proper tools and resources to fix the problem. For example, if you have a community who is the main target of aggravated assault in 2024 while identity theft was the #1 crime back in 2020. If there is no recent data available, we might be still operating on a mission to prevent identity theft which in itself requires different approaches than aggravated assault, diminishing the effectiveness of our criminal justice system as a whole.

My next question regarding crime is during which month do we see the highest crime rates and which hour? This is critical if we want to refine our results. What I mean by this is that if we really want to make a difference when combating crime, we need to be as specific as possible. If we manage to find the month and hour in respect to a specific crime we are interested in, we have a strong foundation to work with. Even if the time and month do not correlate as the year goes by, we still have enough data to come up with solutions. My last question regarding crime is which area experiences the most crime? Even though any crime and any area is important, I think knowing which area is contributing to crime the most can be greatly beneficial. This is because we can gather a stronger police force in these areas to maintain order and to reduce crime.

The next category I will be looking into is victims. I have done previous work in the past while completing my Forensic Psychology degree in regards to victims of different crimes. This is an area I am passionate about and wanted to include in this project. The first question I am looking to answer is how does the victim's descent affect their experiences with crime? As stated previously, any crime is important to look into and the victim deserves a fair process but I do want to take a closer look at how ethnicity/race might have an effect on the likelihood an individual will be a victim of a crime. As we have seen in the past, there are a vast amount of stories where a person is falsely accused of committing a crime without proper investigations due to implicit biases.

Next question is which age group/sex experiences more crime? Similar to my explanation above, the most we narrow our findings, the more we can help. If we can find trends between age or sex and crime, we create a strong foundation to future investigations to prevent these crimes. Lastly, I want to know which area do females experience the most crimes? I am particularly interested in knowing how females are exposed to crime. Knowing which areas tend to affect females the most could help us not only increase our protection in these places and set new regulations, but also encourage precaution for women in these regions.

Next, I had questions regarding weapons. The questions are pretty similar to what I explained above involving questions such as Which area has the highest number of crimes involving a weapon?, Average number of weapons used, Which age group is most likely to be a victim of a

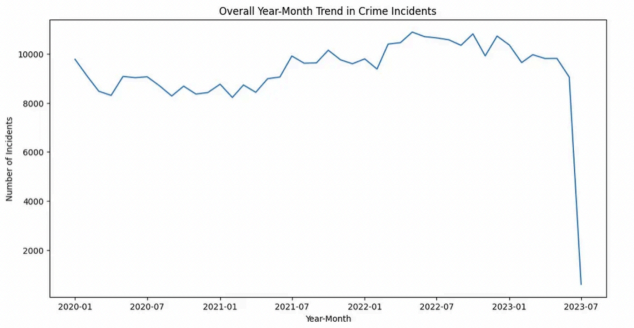
crime involving a weapon? This is an important and controversial category in my report because the findings could either support the right to have a gun or not. When it comes to weapons, specifically guns, the incidents could go right or left. Let's say there was a specific case where a woman without a weapon was struck in the head with a gun and became unconscious. Some might say that if she had a weapon, she would have defended herself while others might think the fact that the offender had a weapon shows evidence as to why we should make guns illegal. That is why I think it is important to look further and see which age group is most affected by weapons? Is it children, teenagers, adults or the elderly? The answer to this question could represent different reactions from the general public.

Lastly, my last category is rape with questions including Which hour rape occurs most often? Which area we see rape happening the most? Which month/area has higher averages of rape? I specifically chose these questions because they had the most relevance in regards to my dataset. I initially was not going to focus directly on rape. However, after conducting more research and analysing my data I noticed I was intrigued by the results of rape in relation to any attribute more than the rest of the crimes. This might be because as females, this is the one crime we all, whether consciously or unconsciously, feel predisposed to. Gathering data that reflect which hour, which area and which month we see rape happening the most can be highly valuable when implementing cautionary measures for the safety of our communities.

## RELATED WORK

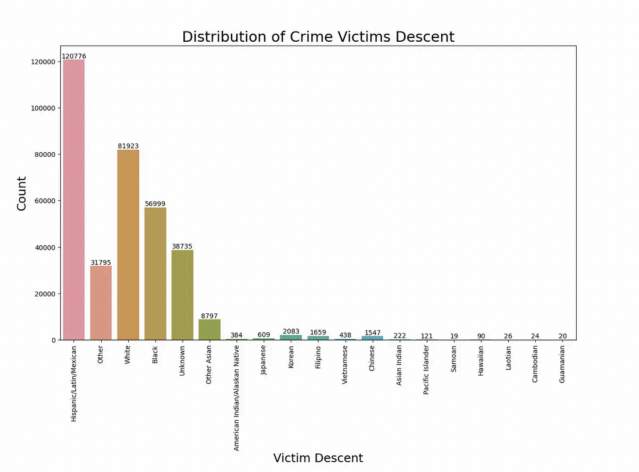
Since crime is a topic of great interest by any city or country, it is expected to have multiple articles and analysis of the data I am using. For my specific project, I will be focusing on two of the multiple currently available which are: Case Study: LAPD Crime Data from 2020 to Oct 2023 Analysis using 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:*



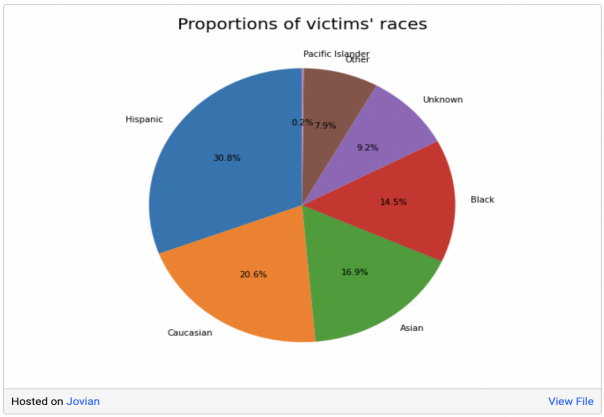

The case study concluded that they " confirmed this analysis with distribution of crime incidents each year below. Specifically, there is a 5% increase in incidents from 2020 to 2021, and 11% increase from 2021 to 2022." This is important information because we can assume that the 5% from 2020 - 2021 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.

**DATASET MAIN TECHNIQUES**

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. However, out of all these attributes, I only worked with 10 which are: date reported, date occurred, time occurred, area name, crime committed, victim's age, victim's sex, victim's descent, weapons used and location where crime occurred such as store, bus or school. I decided to only focus on these because they were more relevant in comparison to the rest in regards to helping me find valuable insights from the data to answer my questions.

**MAIN TECHNIQUES**

The main techniques I used to complete my project were data cleaning, data aggregation, data transformation and data visualization.
For data cleaning, I used Python, more specifically the Pandas library. This is the first time I ever got to analyze data other than reading research papers or case studies and drawing conclusions from there so I went with tools that I felt were at a good level of understanding for my knowledge of data thus far. There were other tools I tried but took me a very long time to understand or I did not understand at all. I chose to work with Python because I already had general knowledge and experience with this language. I had never worked with Pandas but since I was familiar with the syntax and overall structure of Python, picking up Pandas was fairly easy. I did not have to do too much coding for this project because the attributes I worked with were considerably cleaned.

The first thing I started working with was cleaning my data. I removed any attribute that would not serve any value to my final report or would not answer any of my questions. The attributes removed were 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 was left with only 10 that I believed would provide the best insights to what I was searching for. The rest of the attributes focused more on identifying the districts and criminal codes. In addition, I removed any duplicates from the data to increase accuracy of my findings.
Another part of the data cleaning process consisted of filling in missing values. This is important because a lot of attributes in my data were missing entries which prevented me from properly displaying data when working with visuals. For example, one of the attributes in my data is weapons used. I believe we could 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 was very inconsistent and had many values with no entries or NaN values which first needed to be addressed before drawing any conclusions using this data.
I also worked with data aggregation with the help of Python. There were certain questions I needed answers for but my data was not enough to provide me the details necessary. For example, I wanted to know what age group is most likely to experience crime. My dataset has an attribute for the victim's age but it was not categorized into groups. This is when I decided to create my own groups using bins and labels in Python. I created a total of 11 different age groups ranging from '0-10', '11-20', '21-30',

'31-40', '41-50', '51-60', '61-70', '71-80', '81-90', '91-100', '100+'. This helped me visualize my data more effectively as well as discovering more reliable information. This turned out to be very useful when deciding whether or not my data had been contaminated. For example, in my final results, the average age group for victims of a stolen vehicle were '0-10'. This is not accurate since a child does not own a vehicle or drive a car. The discrepancy occurred because part of cleaning my data was substituting negative values with zero. When I printed out the average age of the victims, I ended up with multiple negative numbers which is clearly an issue since no one is negative years old. So I decided to fill in these values with zero which ended up causing a massive average of entries to be reported as the number one target for many crimes.

Another technique I utilized was data transformation. This was a quick and easy process where I dissected the 'TIME OCC' attribute in my data which illustrated at what time these crimes were happening. I created a new subcolumn under 'TIME OCC' named 'Hour OCC' where I could directly access the exact hour these crimes were taking place. I also had to make sure they were under the HHMM format to make sure the hour was accurately reported in military time or 24-hour clock time as that is how the time was entered in the dataset.

Lastly, I utilized Power BI to help me with data visualization. This was definitely the most helpful tool during my entire project. This was by far the best way for me to see for myself what I was creating and what knowledge could be mined from this dataset.

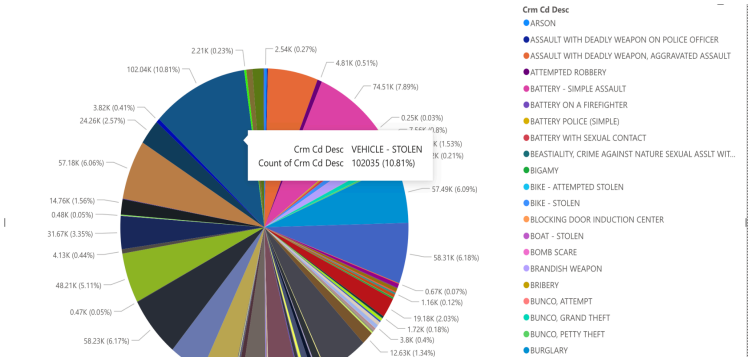
Initially, I thought this was going to be the last step in my project where I would bring my nicely clean data and create some charts but I was incorrect.

The first time I uploaded my data after cleaning it, I realized there were so many other factors that needed to be addressed that I was not able to see given how large my dataset is. I worked with Power BI as a guide to what needed my attention. This is where I found out I had all the missing values and where I also realized my 'TIME OCC' attribute was not entirely converted to the HHMM format causing me false results. This is also relatively easy to learn, especially with the endless resources available on youtube and Google where you can find tutorials to help you create your first project.

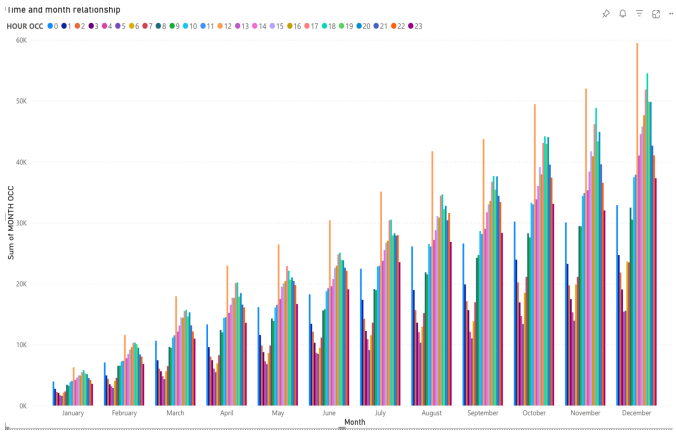
## KEY RESULTS

Using these different strategies helped me gain valuable insights that accurately answered all my questions. Here is a detailed explanation of the different answers to my previous questions under the four categories I came up with using my data.

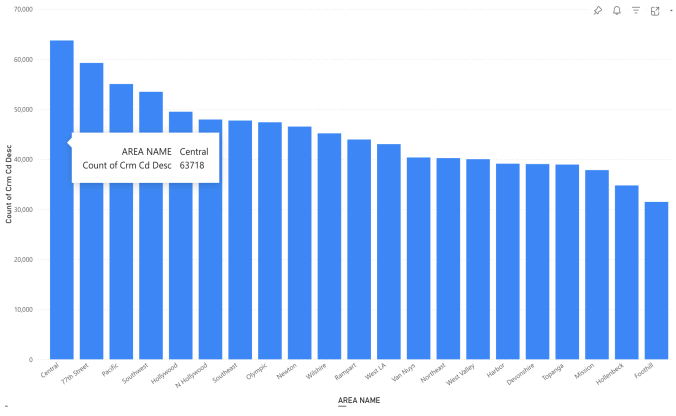
*Crime*




When analyzing the average of crime and trying to figure out which crime had the highest number of reports from 2020 to 2023, based on the pie chart illustrated above, we can see the crime with the highest average is vehicle theft with an average of 10.81% and battery coming in as second with a combined average of 8.2%




In relation to which month do we see the highest crime rates, based on the graph above, we can see that the month of December tends to be the most dangerous and heavier on crime. We can only observe how as the year goes by, crime increases almost double. For this particular question, I don't think my data was accurate and could have been contaminated by different factors when conducting the reports or entering the data into the system. I would have to further look into this question and potentially find additional data collected around the same time to compare it to for accuracy.

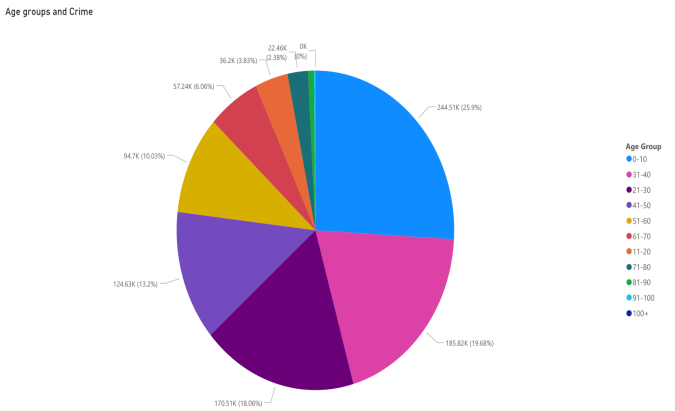



When looking into which area experiences the most crime, Central LA was the number one contender while Foothill was the area with the lowest crime average. In regards to the hour where most crimes were reported, it was concluded that 2200 or 10:00 pm were the hour of each month where crime in general happened the most.

So far in our analysis, just by diving deeper into our crime data we have created a great foundation for further investigation. We now know vehicle theft is prevalent in Los Angeles as well as the area most affected and time of day. Next, I moved on to answer my questions regarding victims.

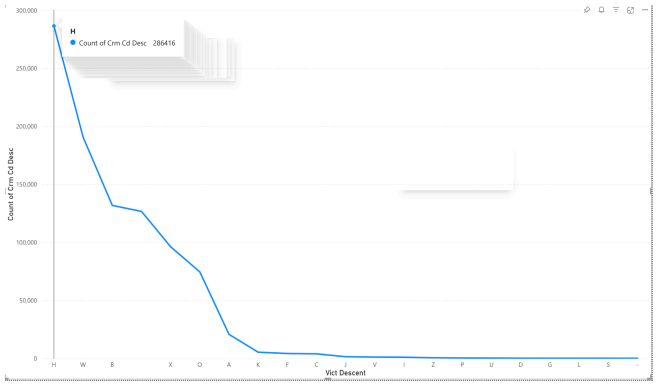
*Victims*

This was the part of my project that interested me the most and I was able to find significant discoveries that could help us better understand crime in Los Angeles. First, I started with my initial question of which age group experiences more crime. The results are illustrated below:



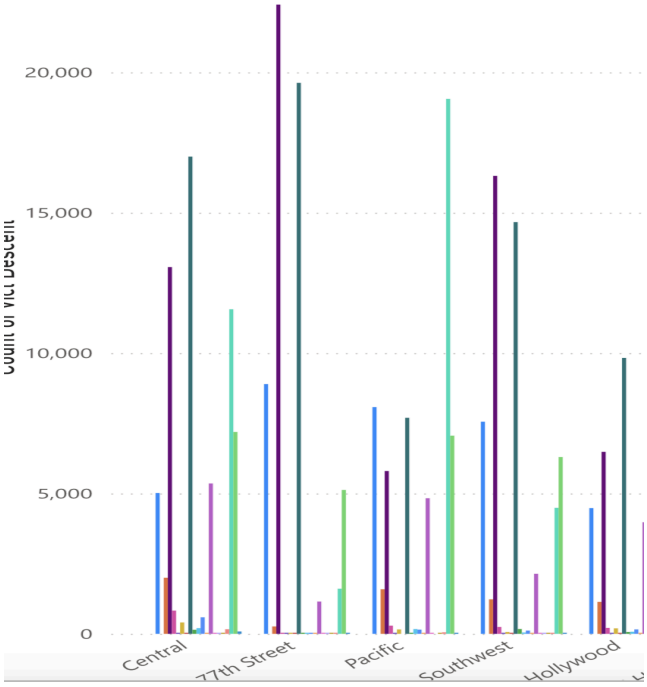

We can see that the age group most affected by crime is [31-40] with [21-30] as second. In addition to this, it was concluded that age 30 overall had the most encounters with crime. Another interesting finding is that in Central LA, which previously we had described as the area

with most crime happening, most individuals who live here are in the age group of [31-40] which we could assume is the reason why both crime is happening so often here and why such a big part of the population around this age is involved in crime. Another aspect I wanted to look into was how does the victim's descent affect their experiences with crime or is there a prevalence for crime given a certain ethnicity. The data I used to investigate this question is below:

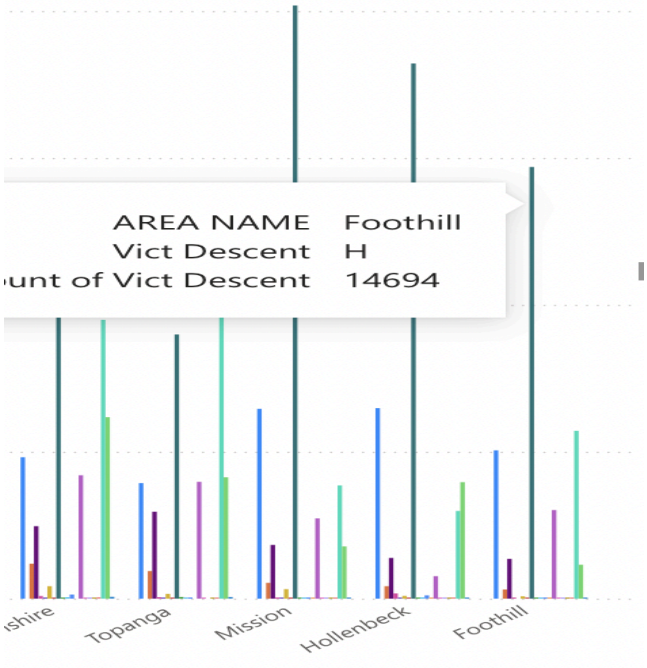



Here we can see a drastic difference between the main two races up on the board. Hispanics are the top race involved in multiple different crimes while White is the runner up. It is important to note that my data also shows that Los Angeles has a large population of Hispanics compared to other races which might be the reason why we see such a big difference as well as why we might see hispanic as the number one ethnicity involved in multiple crimes.
If you have a large number of individuals from a certain race in comparison to the rest, then the statistics are going to show a large involvement from this group in separate occasions causing disproportion in our data. So we would have to investigate further whether or not hispanics are more violent or engaged more in criminal activity or if the reports are simply due to larger size than true propensity for crime.

While conducting this part of my research, I decided to also take a look at the victim's descent in areas of both high and low crime rates to further investigate this question. The results for highest crime rates are listed below:

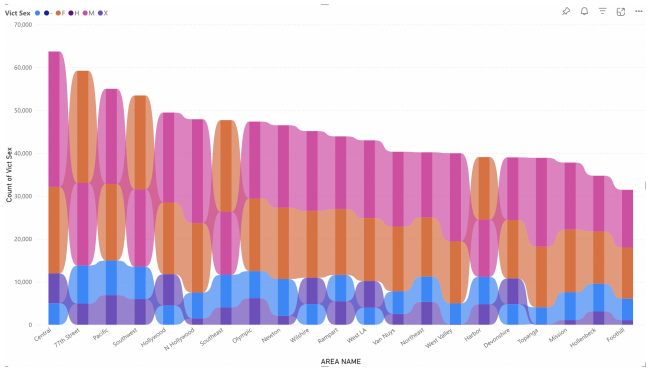



The results for lowest crime rates are listed below:



Here we can see how in both areas, Hispanic is the predominant race regardless of how much crime takes place in the area. We can also see that the averages for most areas are relatively similar, meaning we can safely assume Hispanics take up a large part of the population of LA, helping us understand why there might be a larger number of Hispanics involved in crime than the other races.

In regards to which area we see most females being victims of crimes I put together the following table:

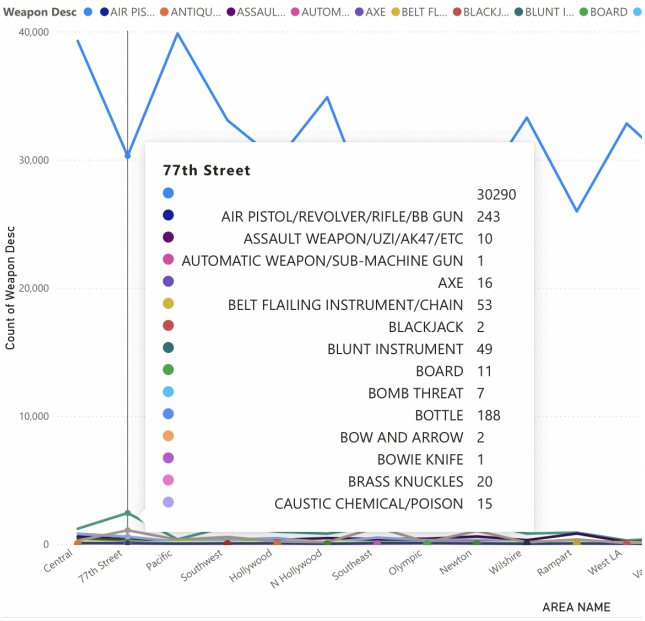



The color orange represents females while pink represents males. We can see that most cases involving women take place in 77th street while for men in Central LA, which we have established is the place with the highest crime. Just by looking at this graph we can also assume that men have a higher disposition to crime than women in most areas. This is a different conclusion than the prior work that had been done with this data since previous reports showed a small difference between the two sexes.

*Weapons*

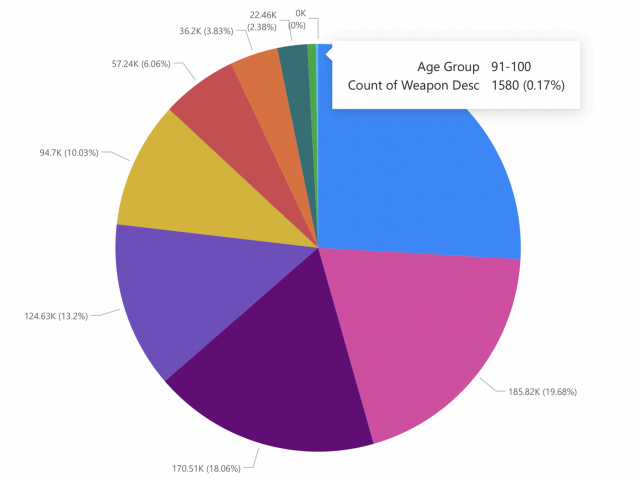
The next area of interest for this project was weapons. The whole purpose of this category was to try to find any correlation between weapons and crime and get a better look at under what circumstances are they mostly present. As stated before, this part of the project was mainly included for clarity and to find future areas within crime where we could potentially look into. There are so many different theories and my findings can be interpreted as either weak or strong evidence depending on the perspective applied.

The first question I looked into was the area with the highest number of crimes involving a weapon. Below are the results:




To my surprise, the area reporting most crime involving weapons is 77th street with Central LA second. Even though this is a good start, this chart and dataset alone are not enough to provide concrete evidence supporting this claim. As seen above, the blue line which takes up for most of the values gathered from this were NaN values. As I started previously, I had to remove NaN values from the Weapons attribute in order to process the data in Power BI. As a result, all these invalid entries compromised the data.

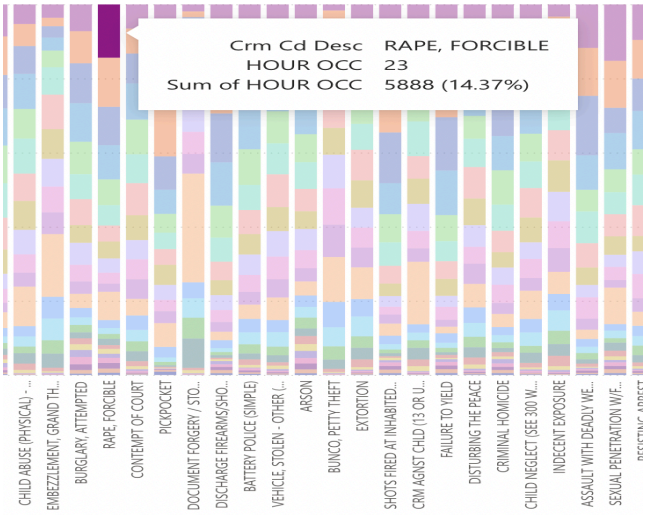
My next and final question regarding weapons consisted of finding which age group was most likely to be a victim of a crime involving a weapon. The results are as follows:



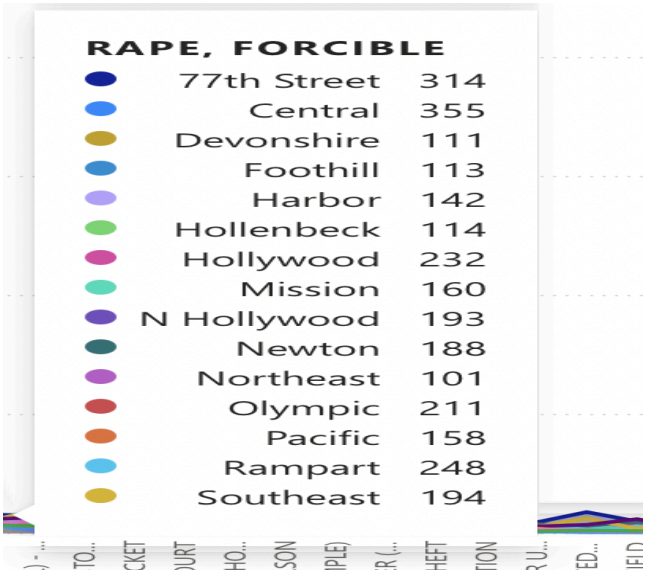

Here we can see the age group with the lowest encounters with weapons were between 91-100 while the group with the highest encounters with weapons were between 31-40. Once again, this data has to be approached with caution due to the missing values but the results still align with our previous findings.

*Rape*

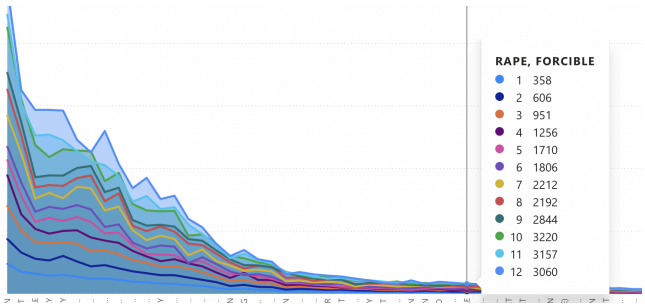
The last few question on my research focused on rape specifically with one of my three questions being which hour rape occurs the most. Here are the results:




The chart illustrates the hour of 2300 or 11:00pm to be the most common time rape crimes occur with 14.37%, while 2200 (10:00 pm) second with 13.26%. Given the proximity in values, It is reasonable to assume that the hours between 10:00 PM and 11:00 PM (2200-2300) are the most dangerous, with a higher likelihood of rape crimes occurring during this time. With these findings in mind, lets take a look at the results for our second question: Which area we see rape happening the most?




We see here that 77th street reported the most rape crimes of the group which is interesting considering this was also reported as the area with most weapons used during time of crime. Another surprise from this chart is Foothill being on the higher spectrum of the list considering this is where the least amount of crimes are taking place and the area with the highest percentage of elderly individuals. Lastly, I will show data providing insights to which month has higher averages of rape.



Here we can see the month with the most rape reporting is December. However, due to the increasing flow of the data, I suspect my dataset does not accurately measure this attribute. I tried fixing this by using the binning method, which worked great for my other attributes, but the data still showed the same findings. I suggest further investigation of old and recent data to back this up. Lastly, while playing around with the different tools and values on Power BI, I was able to figure out which sex reported rape crimes more often and the results were as expected with females reporting 3,672 case and males only 31. Even though this outcome was expected, there are other factors to consider. One of which is the stigma associated with men reporting rape or any type of sex associated crimes where they are the victim. We should also consider that this may not fully represent the scope of crimes I reviewed since many cases go unreported.

**APPLICATIONS**

From completing this project, I gained valuable knowledge that I believe are a great contribution when coming up with strategies to keep our communities in Los Angeles safe. First, There is a great correlation between all the data analyzed. For example, we now know that the areas where most crime is being reported is in Central LA and 77th Street. In addition to this, we know both areas proved to have viable access to weapons potentially causing the large numbers in crimes involving weapons in the area. We also know that the average age group here is [31-40] which we also see is the number one group reported as victims of crimes with its most dangerous hour being 2200 which also turned out to be one of the common hours where rape crimes are being committed. With this information we can start a new initiative such as growing our police presence in these communities but most importantly, I think we should develop educational programs that provide extracurricular activities to the new generations coming up to potentially prevent and decrease future crimes.

These findings could also be used as precaution measures for potential rape victims. We now know the most dangerous hours regarding rape are between 2200 and 2300. We also know 77th street and Central is where rape mostly occurs which can be great information for those who live in these neighborhoods or have to pass by as part of their daily commute. We should also pay closer attention to Foothill and why one of the areas with lesser crime still hits the top of the list concerning rape.

In conclusion, many precautionary measures can be drawn from my conclusions to determine which crimes impact specific groups the most, particularly, which sex, which race, and what age to help come up with strategies to ensure the safety of our communities.