Scala/Spark Project on LA County Crime Dataset

Authors: Frank Tucci, Luke Franks, Winton Gee, Victor Phan

California Polytechnic State University, San Luis Obispo

CSC 369

Stanchev

March 14, 2024

Introduction

The purpose of this report is to use a scala/spark distributed computing pipeline for discovering interesting and important information regarding crime in LA county. The dataset was obtained from a data.gov dataset that is publicly accessible and regularly updated by LAPD. There are many records in this dataset, so our group uses a roughly two year slice from the beginning of 2020 to the beginning of 2022 for analysis. For the common areas of crime, rdd's were used to find the most dangerous and safe areas in LA County. Methods used were similar to what was done in the labs and assignments like: Split(), ReduceByKey(), SortBy(), Explode(), and Take(). Similarly, the weapon trends were analyzed the same way. ScalaFX was then used to visualize the data with bar graphs.

Data Features

The dataset contains features such as date and time occurred, date and time updated, approximate location, the type and weapon used (if at all) of the crime, and demographic information, such as the age, sex, and descent.

Preprocessing

There are some occasional missing values that we had to fill in the gender field, if unknown, and many crimes do not involve a weapon at all, which we created a "None" value for. Not all fields were useful, such as the codes used internally for things like crimes and weapons, that we discarded, using descriptions instead. It also appears that LAPD uses 12:00pm as a default value for dates that do not have times for them, and the 1st of the month for crimes with no date or time, which dirties the data that does occur during these times.

Demographic Information

The victim demographics include information about crimes which involve victims. Not all crimes have victims so those would be filtered out. The information contains the gender and average age. To differentiate between which crimes include a victim, we can simply check whether there is data for the gender because those without information on gender would not include that data, we can apply a "F" filter to check if the victim was a female. To determine the average age, we would map the age and gender of the victim, and we can figure out average age through, sum of female ages / number of females.

Total Crimes	410567
Crimes with Victims	78%
Average Age of Male Victims	37

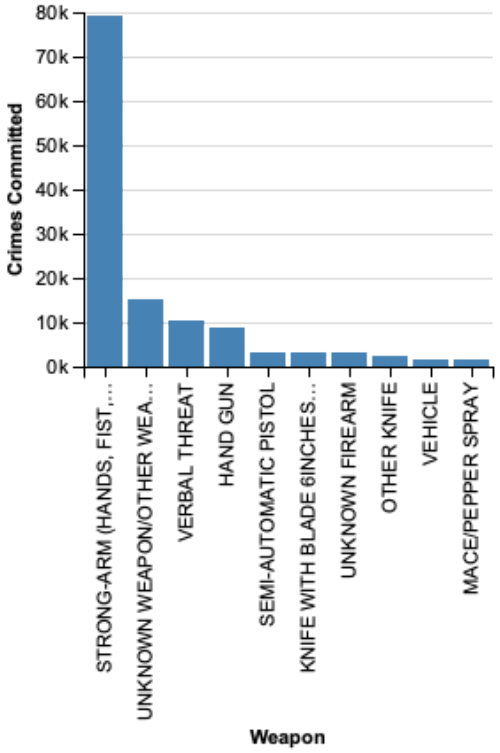
Average Age of Female Victims	38

Most Common Weapons

Results for the top weapons per crime:

1. STRONG-ARM (HANDS) - ~80,000crimes
2. Unknown WEAPON/ OTHER WEAPON: ~15,000 crimes
3. VERBAL THREAT ~10,000 crimes
4. HANDGUN ~ 10,000 crimes
5. KNIFE ~ 5,000 crimes

The results for this were a little tricky to read as some fields within the data set were empty and not filled out. So, instead of the weapons type category being parsed, the sections next to it were scanned instead like the status description or premise description. In the results above, I ignored those fields if they popped up in the results.

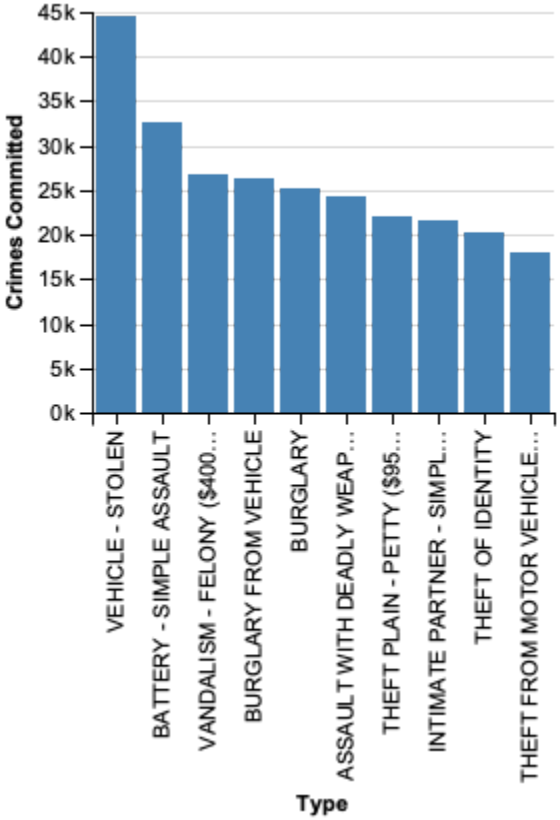


Most Common Crimes

The most common crimes committed were:

1. VEHICLE - STOLEN - 44601 crimes
2. BATTERY - SIMPLE ASSAULT - 32638 crimes
3. VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS) - 26679 crimes
4. BURGLARY FROM VEHICLE - 26269 crimes
5. BURGLARY - 25131 crimes

Each crime description was included in the report. It appears that the most common crimes are vehicle and burglar



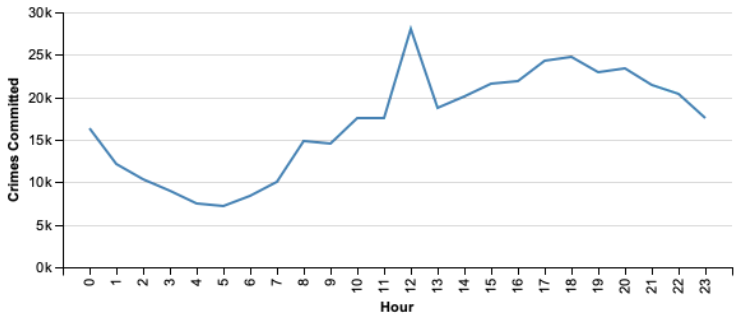
Most Dangerous Days in LA County

To assess a crime's danger level, we focused on incidents involving weapons. We organized the data by weapon involvement and noted the dates for future reference. Grouping by date allowed us to consolidate rows with matching dates, and assigning an integer value facilitated comparisons of dangerous crime occurrences per day. Sortby with the count of the list, is for sorting by the number of crimes that involve weapons, multiply by -1 to make the data in descending order. Using 'take 10' narrowed our focus to the top 10 most perilous dates, streamlining the analysis by omitting less significant dates.

1. 11/01/2021 - 752 crimes
2. 10/12/2021 - 742 crimes
3. 11/29/2021 - 727 crimes
4. 11/08/2021 - 723 crimes
5. 09/27/2021 - 721 crimes
6. 08/30/2021 - 720 crimes
7. 10/18/2021 - 718 crimes
8. 07/05/2021 - 716 crimes
9. 11/15/2021 - 712 crimes
10. 12/07/2021 - 708 crimes

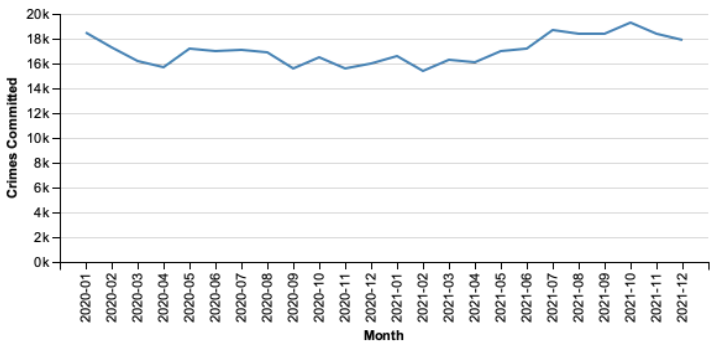
Hourly and Monthly Patterns in Crime

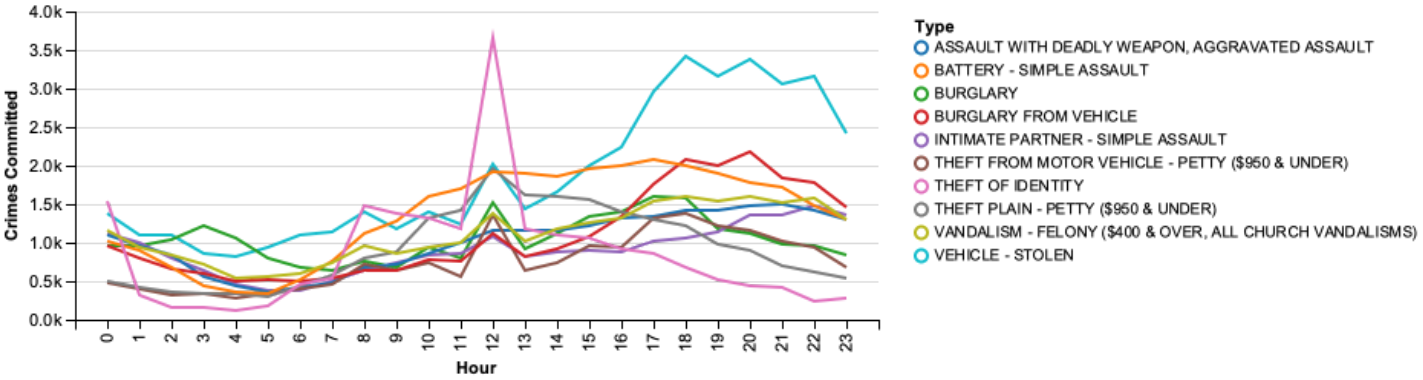
Next, we looked at if there were crime patterns that existed throughout the day, or by month. We used the scala keyBy function to group dates and times together and reduceByKey to get the number of crimes in the dataset.



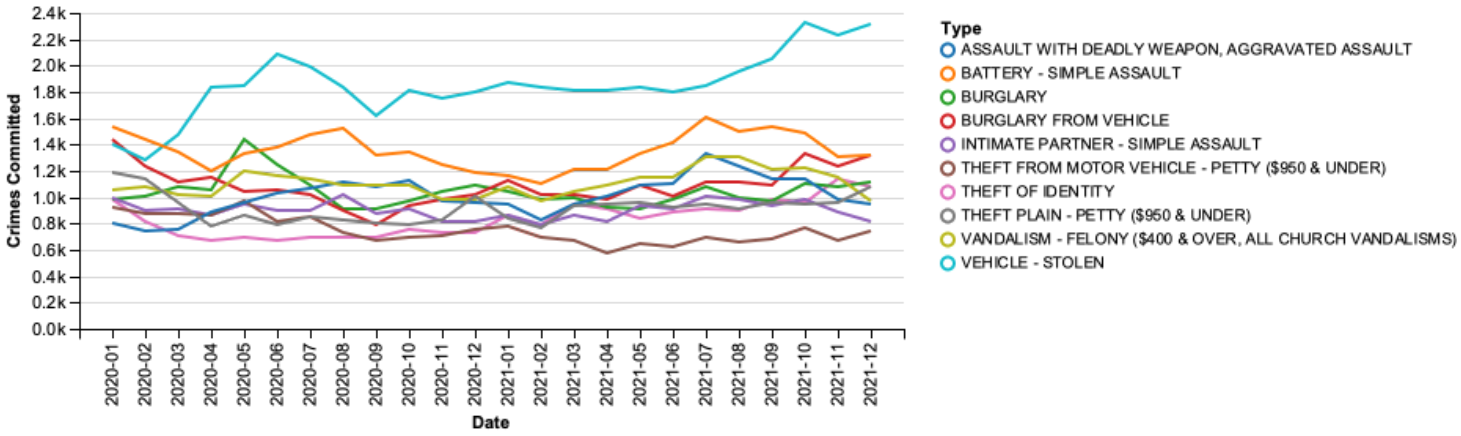
It looks like the number of crimes decreases during the early morning hours, and spikes in the afternoon. The spike at 12:00 is the default value and should not be taken literally.

There is little sway in month-to-month numbers. A slight decrease and increase may be due to lower levels of crime reporting during the 2020 COVID shutdown.





This graph, showing the types of crimes that occur over a 24 hour day, show that most categories of crimes that involve a confrontation, such as battery, drop in the early morning. Crimes that involve theft, such as vehicle theft and burglary, occur much more often late at night. Crimes that do not happen physically, such as identity theft, do not seem to change much throughout the day.

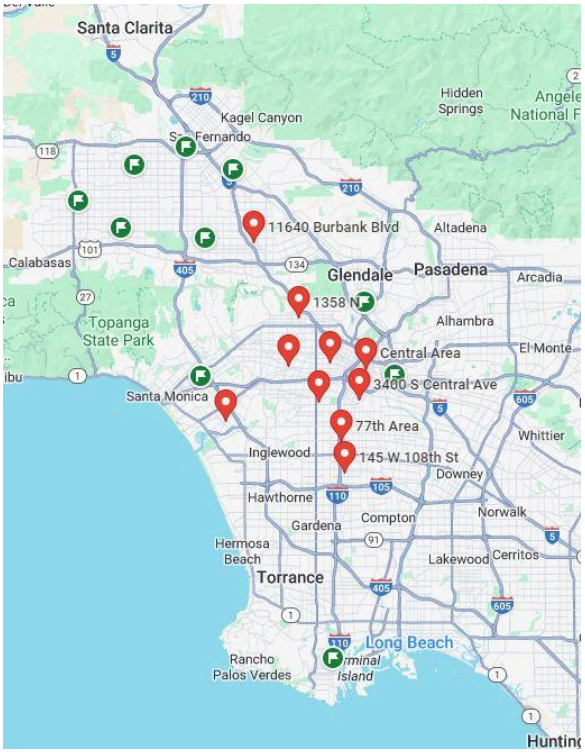


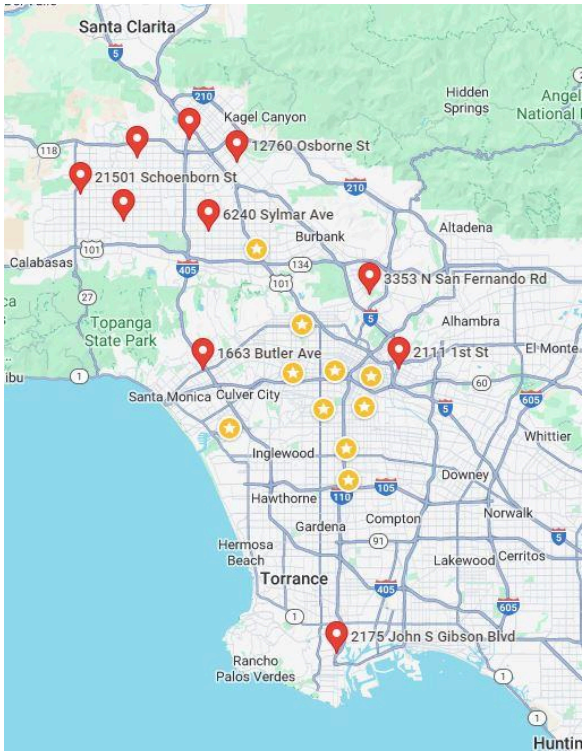
It looks as if vehicle theft, which is already the most common crime in LA county, is increasing as the graph continues, suggesting that this is a worsening problem. Generally, crimes stay pretty consistent.

Dangerous and Safe Areas in LA County

LA county has a total of 21 police stations and they are all coded 1 through 21 in the data set.

Top 10 Most Dangerous Areas	Top 10 Safest Areas
Central Area	Foothill Area
South Bureau	Hollenbeck Area
Pacific Area	Mission Area
Southwest Area	Devonshire Area
Hollywood Area	Topanga Area
Olympic Area	Harbor Area
Southeast Area	West Valley Area
N. Hollywood Area	Van Nuys Area
Newton Area	Northeast Area
Wilshire Area	West LA Area





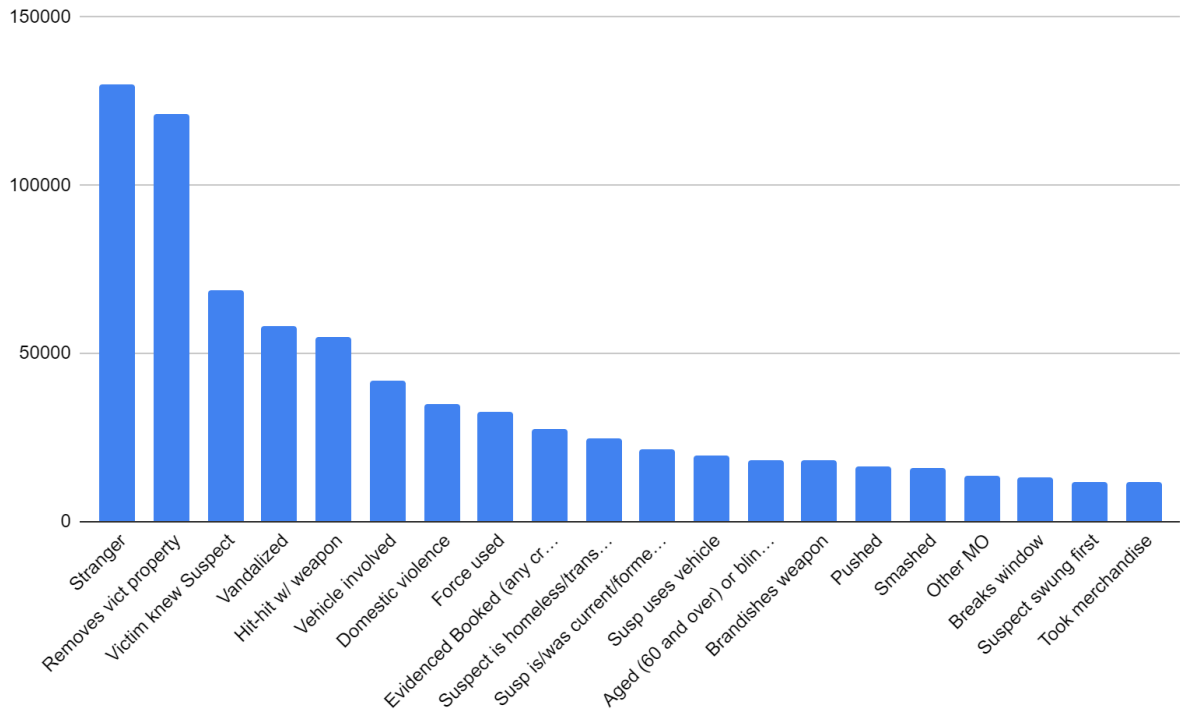
Most Commonly Reported MOs

The amount of times MOs are reported and the way they are reported differ significantly over time and is highly dependent on the case. The program takes in and organizes the Modus-Operandi codes and iterates through the occurrences in the dataset. It references the most recently published LAPD MO codes.

The bar chart reflects the translation of the MO codes into useful reports that are assessed through crime reports. The significance behind the dataset is how often some of these codes are reported. For example, ‘Stranger’ and ‘Removes vict property’ hold large differences among the rest of the MO codes, with over 120,000 counts each. ‘Victim knew suspect’ is the only one closest to 70,000 counts. Meanwhile, the rest of the codes do not come close to these top datasets. Another essential component of the codes is the varying description

Modus_Operandi	Occurrences
1822	129908
0344	121038
0913	68830
0329	57891
0416	54933
1300	41779
2000	34717
0400	32384
1402	27683
2004	24631
1814	21460
1309	19811
1202	18150
0334	18093
0444	16581
1609	15902
1501	13412
1307	13287
0446	11732
0325	11723

only showing top 20 rows



Problems

- Problems with missing or misleading default data.
- Not as much documentation available for scala when compared to Java, Python.
- We used scala/spark to look at trends in crime location, demographics, and methods.
- The most common crimes are vehicle and burglary related.
- Police stations that handle more crimes are in more populous areas.
- Theft is much more common during the night and there is little variation in monthly patterns.
- Crime habits commonly involve anonymity and erratic behaviors
- Anything else