Crime in Los Angeles

An Analysis from 2010 to 2023

Matthew Biller
University of Colorado

Problem Statement

With a population of 3.849 million people, Los Angeles, California is a vibrant and unique city. Home to numerous communities and neighborhoods, LA offers a home for anyone, bringing communities and nationalities together from across the world.

However, the City of Angels also has a dark side. With a crime rate of 36 per one thousand residents, Los Angeles has the highest crime rate in the United States compared to communities of any size [1].

This research project, covering the timeframe of 2010 to 2023, will be an analysis conducted to analyze all documented crimes that occurred. The intent is to uncover knowledge on the types, frequency, and location of crimes occurring within this 13-year timeframe and determine if any interesting information can be concluded from the research.

What knowledge is being looked for?

To generate any interesting conclusions from the data mined on historic crime in LA, interesting questions must be asked first:

- What crimes occurred the most between 2010 and 2023? The least?
- Was there a change in the type of crime that occurred the most/least frequently over a period of time? Every 5 years?
- What areas in Los Angeles had the most criminal activity?
    - What were those crimes?
    - What was the demographic of these areas?
- What factors may have contributed to certain crimes in certain areas?

Answering these questions will require determining if any interesting patterns can be found in the crime data, which may lead to a greater understanding of crime patterns as a whole.

Another interesting topic that will be explored is COVID-19 and the impact of a global pandemic on the crime in Los Angeles:

- Did COVID-19 change the frequency of any crimes?
- Did the crime rate in Los Angeles become higher? Lower?
- Did certain neighborhoods see a change in criminal activity with COVID-19?

While this data will be analyzed as a whole, there will also be a focus on the timeframe of 2020–2022 to determine if COVID-19 had any significant impacts on crime in Los Angeles.

Key Takeaways

The most important takeaways for this project will be:

- To determine if any interesting information, patterns, statistical evaluations, or connections can be concluded from the Los Angeles crime data.
- Evaluate crime in neighborhoods
- Understand if COVID-19 impacted crime and created any unique findings
- To develop a greater understanding of criminal patterns in major cities.

Literature Survey

DataLA:

An interesting study from the data team of Karen Bass, Mayor of Los Angeles, was found using the same dataset to explore crime trends in Los Angeles from 2010 to 2019. The study takes a dive into the crime rates per year for the specified timeframe, focusing on crimes per 10,000 residents.

An interesting thing to note is that this study has a primary focus on the demographics of victims, focusing on the change in crime rates with respect to the ethnicity of victims, and the patterns that emerged over the nine-year period.

This study also mentions COVID-19 and a general acknowledgment that COVID-19 may have impacted the crime rates of 2020 and 2021, however COVID-19 was only briefly mentioned in this regard. [2]

CAP Index:

CRIMECAST, a company that has developed a crime index scoring system called CAP, published their findings for the city of Los Angeles on the changes of crime rate at the beginning of COVID-19.

While the scope of their study is limited to only retail theft and to March of 2020, detailed percentages of various crime types in relation to retail theft are noted, with a week-by-week breakdown of percentage change. A general summary is also included as to why COVID-19 may have caused these changes in crime rates related to retail theft. [3]

Proposed Work

Data Cleaning:

The first step of the data cleaning process will be the merging of two datasets into one. The datasets for this project are currently divided into 2010-2019, and 2020-2023. The two sets will be merged together and presented as one set in assorted order from 2010 to 2023.

It has been noted in the dataset that null locational attributes are labeled as "0°, 0°". These values will be replaced with locations from the same neighborhood, block, street, cross street, etc. if possible, otherwise all null locations will be evaluated separately from the crime entries with locations, but will still be considered in the overall analysis.

Data Preprocessing:

Part of the data preprocessing that will occur is determining which attributes of the dataset are necessary. Several "code" attributes (city codes or codes from the Los Angeles Police Department) will not be relevant to this study and will be ignored.

Another import preprocessing step will be to ensure that the provided Lat/Lon are able to be used in reference for mapping and visual representation. The locations provided are only to the block-level to maintain privacy.

Data Grouping:

Determining groupings for crime entries will be a very important part of the data preprocessing and overall processing. Data entries may be grouped by year, type of crime, location, severity, etc., most likely several groupings will emerge to allow for a deeper evaluation of the data for finding patterns and interesting information.

Differences in Research:

While the DataLA study does use the same dataset, key differences include only using the 2010–2019 dataset with no use of the newer 2020–2023 dataset. Some comparisons to the changes in crime rates will be made between this research project and my own, mainly to cross-reference percent changes year-over-year.

This research also includes demographics of victims, which will not be a factor in my research project. I will only be focusing on the "neighborhood-level" or higher.

While the CRIMECAST analysis does look into COVID-19 and its' impact on crime rates in Los Angeles, they focus their findings to only retail crime and within the first four weeks of COIVD-19, March

2020. My research will involve various crime types and span a thirteen-year period.

Data Set:

Title: "LA Crime Data"

Type: CSV

Provided By: Kaggle (link)

With ~2.7 million entries, this dataset provides incidents of crime between 2010 to 2023 in Los Angeles, CA. All data entries were transcribed from original crime reports.

The dataset provides 28 different attributes that include items such as date, location, type of crime, type of weapon, severity, victim information, etc.

Evaluation Methods:

To fully evaluate the data for interesting patterns and relationships between attributes, several evaluation methods will be used:

Metrics:

Percentage change will be one of the most important metrics to calculate. Crime types will be viewed at various temporal levels (month-to-month, year-to-year, and year binning) to gain insight into the occurrences of crimes, their changes over time, and patterns that may emerge from that.

Clustering:

Clustering will be used to determine if similar crimes occurred in certain neighborhoods or areas around Los Angeles. Identifying hotspots of criminal activity could return interesting results, those hotspots will then be analyzed over different periods of time to look for patterns or changes in behavior of the types of crimes being committed throughout the city.

Data Visualization:

An overlay of world imagery will be used to demonstrate visually where crimes, or groupings of crimes, are occurring throughout Los Angeles. Imagery will be provided by Maxar/Google Earth and will have points, as well as hotspot areas overlaid as a layer to show where crimes occurred.

Tools:

Python will be used as the codebase for the data preprocessing and data analysis, packages Numpy and Panads will be used for statistical evaluations.

Mapbox will be used to visually represent the crime data, overlaid with satellite imagery. Data points and heat maps will be built with Mapbox's API and included in the final report to visualize emphasize the occurrence of crimes.

Milestones:

In three weeks: I would like to have the majority of my data analysis complete. This will involve organizing crime types, performing statistical analyses, and determining any clustering in the data.

In four weeks: I intend to have the majority of my analysis complete, I will spend this week looking for any/all interesting informational and connections that are a result from my research.

In five weeks: I would like to have the data visually represented to be included in my final report.

In six weeks: I intend to have my final report through the rough draft stage and nearing completion. I will continue refining my report, building my final presentation, and prepare my final submission for the project.

Milestones Completed:

The biggest milestone so far has been the completion of preprocessing the data. With ~2.8 million entries divided between two CSV sheets, preprocessing took a considerable amount of time.

Initially, the two CSVs were merged together to create one master sheet. This created a CSV that was unable to be opened, due to size, in anything but a generic text editor such as Notepad. While this made viewing the data more difficult, it was still possible to visually see which attributes needed to be kept and which needed to be eliminated within the master sheet. Of the 28 original attributes, 15 of them were ignored in the master sheet, with most of those attributes being coded letters specific for information pertaining to law enforcement, or empty areas for extended crime descriptions that weren't included in police reports. These attributes weren't necessary for this project and generated more data than necessary, allowing for them to be removed from the master sheet without diluting any necessary information for analysis or final reporting.

Another aspect to preprocessing the data were including appropriate latitude and longitude values for data entries that didn't have an included location. To properly do this, a sample mean of latitude and longitude were taken from the 'Area' that the cell belonged too.

For example, an attribute within the mater CSV was labeled 'Area Name', with names such as "Newton", "Pacific", and "Hollywood" (all pertaining to police station names). Data entries were divided based on their area name, and an average Lat/Lon was calculated for each 'Area Name'. If a cell didn't contain its' own locational data, the mean location for that specific area was applied to that cell.

Another completed milestone is having the data organized into manageable groups to conduct analyses. The data is binned based on years, with this research project spanning a thirteen-year time period, the data were binned into two-year groupings: 2010/11, 2012/13, 2014/15, 2016/17, 2018/19, 2020/21, 2022/23. Having two year bins allowed for the data to be broken down enough to see if interesting findings can be concluded in smaller timeframes, as well as creating an easier analysis for processing and analyzing the data for the entire timeframe. Additionally, the data were also binned into five and four year groupings: 2010/14, 2015/19, 2020/23, allowing for a higher level of analysis to be conducted.

The organization and binning of this data was done by creating lists in Python, setting conditions within a for-loop to check a specific column of the master sheet, and store data entries into their appropriate lists based on the year that the crime occurred. Those lists were then converted into DataFrames using the Pandas package, DataFrames are two-dimensional data structures that contain row and column information, with the column information included from the master data sheet. Finally, each DataFrame was iterated through to create dictionaries, these dictionaries stored the type of crimes and the number of times each crime occurred for every year bin.

The most recent milestone has been the beginning of a more in-depth data analysis to determine if interesting findings can be concluded from the data. With entries being divided into years and totals being returned for each type of crime committed, a percent change analysis has been done for each type of crime committed. While simple, conducting a percent change analysis across the year bins, focusing on the change in types of crime, has already yielded interesting results.

Milestones To-do:

With the data preprocessing complete and having the ability to conduct analyses on the dataset, one of the most import milestones to focus on moving forward will be the analysis of 2020 and the impact of COVID-19 compared to the rest of the dataset. Given the restrictions that were introducing during the COVId-19 pandemic, it's possible that changes in criminal activity and the types of crimes that occurred may have changed for 2020 and the following years.

To properly conduct this analysis, 2020 will be separated into its' own year group (outside of the two and five-year bins), lists and dictionaries will be made to store data for the crimes that occurred during 2020, and in which areas throughout Los Angeles. Independent measurements will be taken to analyze the percent change in types of crime, comparing 2020 to 2019 and 2021, 2022, and 2023.

Finally, data visualization will be included, displaying types of crimes that occurred and how frequently, compared to surrounding years. Data visualization will be a useful addition to emphasize any changes in criminal activity that occurred during the pandemic-era. Any interesting findings will be included in the final report under its' own subsection, where 2020 and the impacts of COVID-19 will be discussed and compared to the surrounding years.

Another milestone to complete will be the continuation of analyzing the data on a deeper level. While the percent change calculations have been useful and interesting, it would be beneficial to divide the data even further into subsets beyond what's already been done. For example, having a 2010/11 bin that holds criminal reports subcategorized by their Area Name ("Newton", "Pacific", and "Hollywood", etc). While this may seem simple, there are 23 unique area names within the master data file, dividing each year list into 23 separate subgroups has been tricky and taxing on the machine.

To remedy this, data will be subcategorized on an as-needed basis. For example, if there are no changes in which 'Area' had the most criminal activity for an extended period of time, then it wouldn't be useful to impose extra work on the machine to break the data down even further. However, if certain year-groupings return a different result that show interesting changes in which 'Area' had more or less criminal activity, then a more in-depth analysis will be performed. This process will save time and resources while still ensuring that interesting results aren't overlooked during the data analysis process.

With the data being divided into very usable groupings, another milestone to focus on will be pattern evaluation. Pattern evaluation will be done by looking at the year-over-year outputs for each type of crime, and determine if any interesting results can be seen from the findings.

- Did any particular type of criminal activity have a dramatic increase or decrease from surrounding years?
- How did one particular type of crime change over time?
- Can any interesting outliers be determined over periods of time?
    - Did a certain type of criminal activity become an outlier?
    - Did that same criminal activity become more "normal"?

Once the majority of the data analysis is complete, data visualization will be one of the last milestones that needs to be completed, aside from writing the final report.

The Python package Datashader is incredibly useful for processing and plotting millions of points of data and returning well-made and easy to understand visualizations. Ideally, several different plots will be made: one for showcasing the changes of criminal activity over the thirteen-year period, one showcasing criminal changes by 'Area' over different time periods, and a final visualization for any uniqueness for 2020 and COVID-19.

After the data visualization, the final milestone will be writing the final report and creating a presentation video to showcase the project and any interesting findings. This is intended to be completed during the final week for the project and will be fully submitted on-time.

Results So-Far:

With a percent change analysis conducted on the types of crime and changes within 'Areas' throughout Los Angeles, interesting results have already started to come to the surface. The first interesting finding comes from the percent change between types of crimes for all of Los Angeles:

From 2010 to 2021, the five highest crimes that occurred for all of LA were: Battery, Burglary from Vehicle, Burglary (non-vehicle), Vehicle Theft, and Petty Theft. In total there are ~135 unique types of criminal activity reported for each two-year bin. While not all types of criminal activity will be included in the final report, any interesting findings based on percent change or area will be included and expanded on.

A breakdown of the changes in each crime from the five-year bins of 2010/14 to 2015/19:

- Battery **decreased** by 4.765% from 97,610 counts to 92,959
- Burglary from Vehicle **increased** 7.65% from 78,103 counts to 84,081.
- Burglary (non-vehicle) **decreased** 4.5% from 75,564 counts to 72,167.
- Vehicle Theft **increased** 15.13% from 74,328 counts to 85,575.
- Petty Theft **increased** 3.27% from 73,748 counts to 76,162.

Another interesting finding was a crime that increased dramatically starting in 2021. Using the two-year bins, from the years of 2020/21 to 2022/23, Identity Theft **increased** 35.92% from 19,524 counts to 26,537. For the years of 2022/23, Identity Theft become the second highest crime for all of Los Angeles with 26,537 counts. Vehicle Theft was the highest with 31,637 counts. This dramatic increase will be interesting to look into further and see if any determinations can be made as to why Identity Theft increased so dramatically for this two-year period.

An interesting finding that isn't related to the type of crime, but rather the location of crimes has also been uncovered:

From 2010 to 2022, the area of 77th Street, which is indicative of the 77th Street Police Station, located slightly south of downtown Los Angeles, had the highest number of crimes per year, averaging ~26,000 crimes for every two-year bin. The highest crime for 2020/21 was Vehicle Theft with 3,362 counts.

However, for 2022/23, the area of Central, which is indicative of the Central Police Station and located directly downtown in Los Angeles, had the most criminal activity with 22,737 counts while 77th Street only had 18,546 counts. The highest crime reported for Central was Identify Theft with 2,491 counts. This number may be higher now as the criminal reports for this project were only retrieved until March 2023.

More interesting results will continue to be uncovered as further data analyses are performed. Using a percent change formula to return interesting findings has been the most useful data mining tool so-far.

REFERENCES

[1] Neighborhood Scout, Los Angeles, CA Crime Rate (link)

[2] DataLA, 2022 A Data-Driven Exploration of Crime Trends in Los Angeles. Medium.com. Link

[3] CAP Index, 2020. COVID-19 & Crime - CAP's Perspective on crime & Loss in the Age of COVID-19. CAPINDEX. Link

Link to GitHub