National College of Ireland

National College of Ireland

Project Submission Sheet

Student Name:	Zuu Zuu Kyaw Shwe
Student ID:	x24106585
Programme:	MSCDAD_C **Year:** 2024 - 2025
Module:	Data Intensive Scalable Systems
Lecturer:	Jaswinder Singh
Submission Due Date:	9th August 2025
Project Title:	Scalable Analysis of LAPD Crime Data
Word Count:	3192

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Zuu
Date:	9th August 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

Data Intensive Scalable Systems

Project

Your Name/Student Number	Course	Date
Zuu Zuu Kyaw Shwe / 24106585	Data Intensive Scalable Systems	9th August 2025

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
Chatgpt	Clarify concepts and provide additional insights on complex topics	https://chatgpt.com/

Description of AI Usage

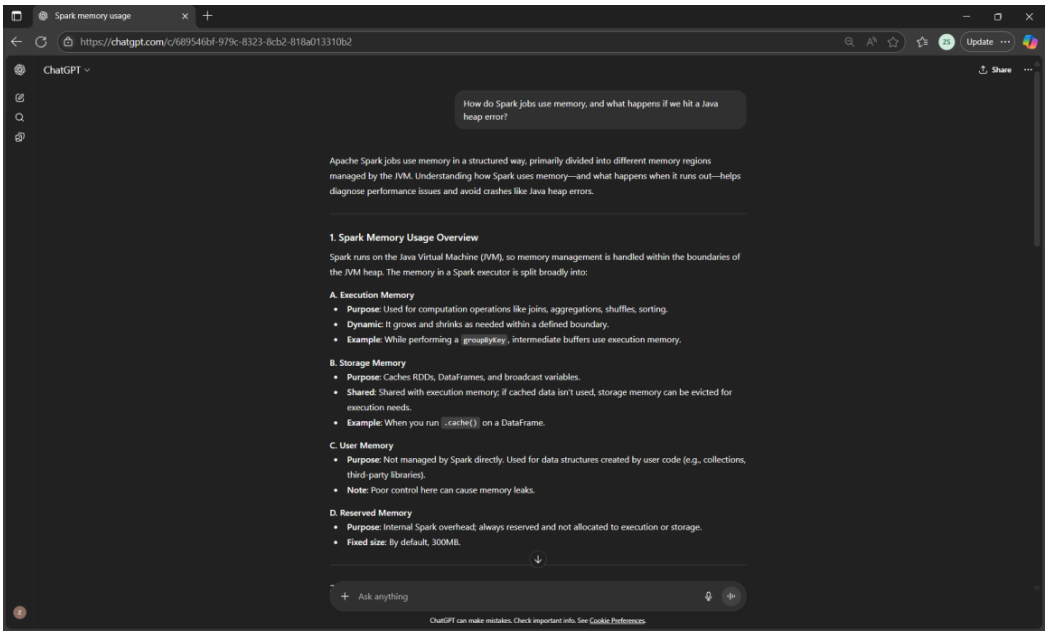
This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

ChatGPT	
For some of the questions and concepts explored, I sought help from OpenAI's ChatGPT, which helped provide additional understanding.	
How do Spark jobs use memory, and what happens if we hit a Java heap error?	Apache Spark jobs use memory in a structured way, primarily divided into different memory regions managed by the JVM. Understanding how Spark uses memory—and what happens when it runs out—helps diagnose performance issues and avoid crashes like Java heap errors.

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence: ChatGPT



Scalable Analysis of LAPD Crime Data

Zuu Zuu Kyaw Shwe
School of Computing (MSCDAD-C)
National College of Ireland
Dublin, Ireland
x24106585@student.ncirl.ie

***Abstract*—This project is the scalable analysis of Los Angeles Police Department (LAPD) crime data between 2020 and 2024 that is delivered with the help of the end-to-end data pipeline which includes Apache Spark, PostgreSQL, and Tableau. The objective was to identify crime trends across temporal, location, type, demographic, and weapon-related dimensions, thereby supporting data-informed urban safety strategies. More than one million crime records were processed, cleaned, and aggregated to extract meaningful patterns at scale.**

The analysis revealed stable crime levels over the period of 2020 to 2023, followed by a sharp decline in 2024, likely attributable to changes in LAPD reporting procedures. Simple assault and vehicle theft became the most common reported crimes, and the spatial analysis described some consistent hotspots in the Central and 77th Street divisions.Demographic analysis revealed that female victims were more common among teens, while male victims were more common among middle-aged individuals. Most crimes involved no weapon, underscoring the potential value of non-lethal prevention and intervention measures.

The pipeline structure was modular, which allowed it to easily transform, integrate, and visualize data. Regardless of the constraints based on data quality and missing data in 2024, the project still shows the potential of the publicly available data to be utilized to create scalable, actionable evidence-based information required to support the planning of public safety.

***Index Terms*—Keywords: Crime Analytics, Apache Spark, HDFS, Big Data, Urban Crime**

I. Introduction

The fast expansion of open government data offers researchers and policymakers with previously unexplored possibilities in studying urban phenomenon utilizing data-driven methods. The project is devoted to crime analysis in Los Angeles based on the data published by the Los Angeles Police Department (LAPD) [1]. Having millions of recorded crime cases and metadata, including geolocation, timestamp, demographics of victims, and the type of weapons, this data set offers a valuable foundation for understanding the distribution and nature of crime in a major metropolitan context.

The goal of this project is to perform a scalable end-to-end analysis on LAPD crime data (2020-2024) leveraging the power of Apache spark, and find relevant actionable insights regarding the crime trends and victimisation patterns. This analysis supports the development of informed strategies for urban safety and resource allocation, particularly in the context of changing socio-political conditions, pandemic-related disruptions, and shifts in law enforcement operations.

The analysis aims to:

- Examine temporal crime trends from 2020 to 2024, identifying significant fluctuations and anomalies.
- Identify the most prevalent categories of crime based on frequency and type.
- Analyse the geographic distribution of incidents to locate high-crime areas or hotspots.
- Explore victim demographics particularly age and sex to understand differential vulnerability.
- Investigate the usage of weapons in criminal incidents and assess the prevalence of armed versus unarmed crimes.

These goals are unified under the central research question: *How have crime patterns in Los Angeles evolved over time, by location, offence type, and victim demographics between 2020 and 2024, and what insights can be extracted using scalable data processing techniques to inform future urban safety strategies?*

It is also necessary to mention that despite the fact that the dataset goes all the way up to 2025, the year was omitted in analysis because of failings to report and inconsistency linked to LAPD switching to an internal system. This data inclusion could lead to distorting of the longitudinal trends and the temporal comparison.

Overall, this project demonstrates the value of applying big data technologies to real-world datasets for civic insight generation. In such a way, it also overcomes the technical, interpretive, and ethical concerns of work with large-scale data on the public safety.

II. Related Work

Scalable data processing frameworks, that allow for the process of crime data to be analyzed, have accumulated a lot of interest, and this has been experienced in both academic and applied studies in the recent years. A number of works have also shown that the use of big data technologies including Apache Spark and Hadoop can be used to analyze scaled-up datasets in an efficient manner to derive valuable patterns that can be used in the areas of public safety and law enforcement.

Kumar *et al.* [2] examined how to integrate big data analytics and visualisation tools such as Tableau to identify spatial and temporal abnormalities related to crime. They emphasised the collaboration of geospatial data together with visual dashboards, which impacted our choice of Tableau to aid in the exploratory analysis and visual storytelling opportunities. Nonetheless, they primarily emphasized in their work on exploratory analysis but failed to show an end-to-end

automated pipeline which is one of the key contributions of our project.

Ahmed *et al.* [3] came up with a crime analysis application based on Hadoop and Spark, which operates with the help of Zeppelin notebooks, which allow to query and visualize the data interactively. Although their architecture succeeded in handling large amount of records of crimes, there was no particular focus on data cleaning and quality assurance. Our pipeline addressed this by incorporating a dedicated Spark-based transformation stage. In addition, our analysis extended beyond interactivity to include structured exporting to PostgreSQL for persistent storage and Tableau integration.

Stec and Klabjan [4] explored the use of deep learning models for forecasting crime events, focusing on the prediction of specific crime categories over time. While their approach was advanced in terms of modeling, it relied on clean and well-structured input data. Our project, although not centered on prediction, supports such future modeling efforts by ensuring a high-quality and scalable data preparation pipeline, making it complementary to rather than in competition with prior predictive approaches.

Xiong [5] used Support Vector Machines, Random Forest, XGBoost, and Gradient Boosted Decision trees to forecast future occurrence of four major types of crimes in the City of Chicago, and it showed very good predictive performance on actual data. This paper demonstrates the operational capability of the machine learning in preventing crime but also shows that it is dependent on curated inputs. With our pipeline being rigorously preprocessed, integrated and aggregated, it could be used as a baseline to be used in other similar predictive deployments.

Mandalapu *et al.* [6] conducted a systematic review of more than 150 works on the prediction of crime with the help of machine learning and deep learning. They emphasized the repetitive issues like integration of heterogeneous data, variable formats of reporting, and insufficient preprocessing that may negatively impact model performance. These are our project pain points directly as automated ingestion, normalization and storage are implemented, thus giving reproducible and high-quality datasets to be used both in analysis and prediction.

Overall, the existing literature has much to offer in the domain of visualization, mass computation, or predictive models, but most are limited by partial pipelines, manual preprocessing, or assumptions about data quality. Our work can be characterized by its use of a reproducible, fully automated, end-to-end workflow, that spans ingestion, transformation, storage, and visualization, and closes the gap between raw, messy public crime datasets and actionable, scalable analytical outputs.

III. Methodology

This section outlines the methodology adopted for the implementation of our end-to-end LAPD crime data analytics pipeline. It details the dataset used, the sequence of data processing steps, the technology stack chosen, and the design patterns applied. This was done to build a scalable, automated workflow for ingesting, cleaning, transforming, analyzing, and visualizing large-scale urban crime data.

A. *Dataset Description*

The dataset used in this study was obtained from the Los Angeles Open Data portal, which provides public access to records maintained by the Los Angeles Police Department (LAPD) [1]. The dataset captures more than a million crime records from January 2020 to date, alongside details like crime classification, timestamp, victim's age, area of the crime, weapon employed, and the current status of the crime.

The complexity of the dataset is due to the size of the dataset as well as the structure of the dataset. It covers a period of 5 years and contains various characteristics, such as timestamps, geographic divisions, crime types, weapons, and victims that makes it highly dimensional. Also, the dataset has quality issues including missing values, irregular formatting, and unclear category names. The cleaning and transformation process based on Spark addressed most of these problems, standardizing textual fields, substituting nulls with placeholders, and combining date-time data in order to analyse temporal trend. The combination of scale, variety, and remaining quality considerations necessitated the use of a distributed processing framework like Apache Spark to enable efficient ingestion, transformation, and analysis.

The selection of this dataset was motivated by the dataset's low accessibility barriers and LAPD's crimes data's rich granularity alongside its temporal and spatial relevance. Furthermore, LAPD data provides a valuable opportunity to analyze urban crime trends over time and across various demographic and geographic dimensions, making it ideal for a data-intensive project that emphasizes real-world impact.

B. *Technologies and Design Justification*

Apache Spark was adopted to run the project because it could execute distributed, in-memory processing on large data due to the required need in the processing of millions of records on crime. **Python** was used to script because it is flexible and has broad library support and ease of integration across stages in the pipeline. **Hadoop Distributed File System (HDFS)** was used as it offers scalable and fault-tolerant storage that persistently provides raw and processed data.

To store aggregated results in an organized manner, a proven reliable database with the capabilities of SQL had to be chosen and **PostgreSQL** was selected due to these features, compatibility with data visualization tools. **Tableau** was used for creating data visualizations and performing chart-based analysis to extract insights.

All scripts were executed within a virtualized **Ubuntu** environment, with data stored in the Hadoop Distributed File System (HDFS) to enable horizontal scaling and persistence. The workflow followed an Extract–Transform–Load (ETL) design pattern to promote modularity, enabling each stage—data ingestion, transformation, aggregation, and visualization—to

be developed, tested, and executed independently. This modular design facilitates maintenance, makes debugging easy and make it easy to extend the pipeline in future.

C. *Data Processing Pipeline*

The data pipeline consists of the following sequential components:

- **Data Ingestion:** The data were retrieved through HTTP GET endpoint in an API by using a Python script. The information of 2020-2024 was queried and stored in JSON format. Data of 2025 was not used in analysis as no more than 100 reports were available, which makes the year not statistically representative.
- **Data Storage:** The HDFS was used to load the raw JSON file and this allowed reliable distributed access for Spark processing.
- **Data Cleaning and Transformation:** PySpark script was applied to enable selection of appropriate fields beside the merging of date with the time field into a single timestamp field, a new field of year and month was also extracted. The standardization of textual characteristics involved the trimming and the uppercase conversion, whereas missing values in categorical fields were fixed onto values of the same placeholders (e.g. UNKNOWN, NO WEAPON).
- **Data Analysis:** Different aggregations were conducted via five PySpark scripts with each script being assigned to one dimension of analytical work. The advantage of this modular approach is that analysis components can be executed separately, that maintainability is easier, and that debugging or later extensions are made simpler.
  - **Trend Analysis:** Crimes grouped by year and month.
  - **Crime Type:** Frequency of crime descriptions.
  - **Crime by Area:** Spatial distribution based on LAPD divisions.
  - **Victim Profile:** Breakdown by gender and age group.
  - **Weapon Usage:** Analysis of weapon descriptions.
- **PostgreSQL Export:** All analysis outputs were written to PostgreSQL tables using the JDBC connector.
- **Visualization:** Tableau connected to PostgreSQL for data visualization.
- **Automation:** The pipeline was automated via a shell script, which allows sequential execution of tasks from ingestion to export.

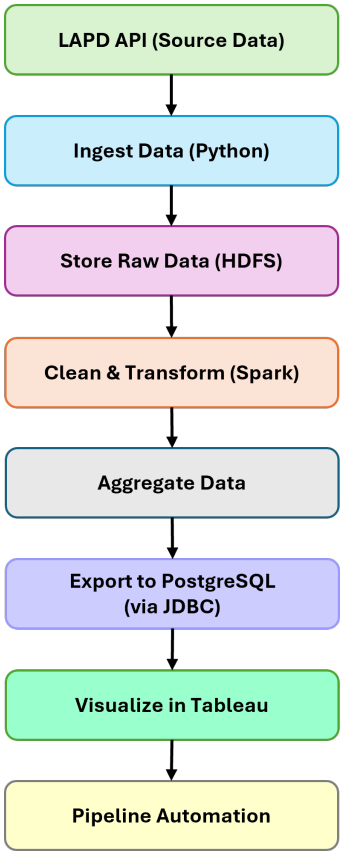


Fig. 1: Data Pipeline Overview

Figure 1 presents a high-level overview of the complete data processing pipeline. It demonstrates how the project integrates diverse technologies into a cohesive and automated architecture for scalable urban crime analytics.

D. *Implementation Challenges and Solutions*

There were a number of implementation issues. Disparities in data was the main issue especially in areas like vict_sex and weapon_desc where there would be incidences where data would be missing or being incorrectly formatted. This was countered by ensuring that it was filtered and standardized earlier during transformation.

The use of the larger dataset size also brought about performance problems during process within Spark particularly aggregation. Early dropping of unnecessary columns was implemented in the pipeline to enhance performance before grouping operations. Memory settings of Spark executor were optimized in order to prevent bottlenecks.

Finally, repeatability and modularity of the pipeline was taken care of via shell based automation. Although designed with potential for scheduling, the pipeline was executed manually for demonstration purposes. All these measures helped to achieve a strong and scalable analysis framework.

IV. Results

This section presents the outcomes of the analytical pipeline applied to Los Angeles Police Department (LAPD) crime data

from 2020 to 2024. The results are organized according to the core research question: how have crime patterns evolved across time, location, type, and victim demographics, and what insights can be derived to support urban safety strategies?

A. *Temporal Crime Trends*

To understand fluctuations in crime over time, the month wise crime data were merged into the total crime over the five years period. Overall crime volume, as seen in Figure 2, was rather constant over the period between 2020 and 2023 with small seasonal declines. The decline in reported incidents is noticeable beginning with the end of 2023. Nevertheless, the significant adjustment in 2024 can be viewed with caution because LAPD shifted to an improved form of reporting early in 2024, which resulted in insufficient release of data. The consistency of the crime rates until 2023 implies that it has systematical patterns, which may prove beneficial to predictive modeling. Incompleteness of its data in 2024 elucidates the need to ensure quality in external data prior to extrapolating conclusions.

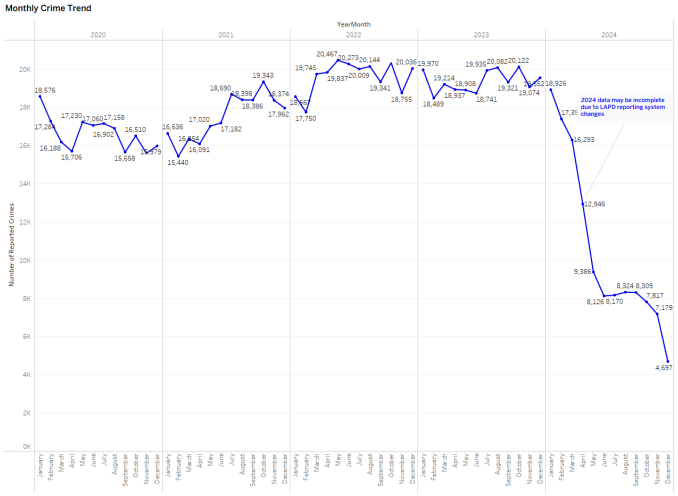


Fig. 2: Monthly Crime Trend

B. *Prevalence of Crime Types*

The most frequent 20 categories of crimes are ranked in figure 3. As revealed in the analysis, the occurrence of crime that was mostly reported in the period includes the crime Vehicle-Stolen with over 115,000 recorded cases. This is followed by interpersonal offenses, including Battery-Simple Assault, as well as property crimes such as Burglary from Vehicle and Theft of Identity. These crimes are mostly non-lethal and are of huge concern to the community and economically. Diversity of the leading 20 crimes also indicates the necessity to keep in mind both property protection and interpersonal violence prevention being a crucial area of focus in city policing. The vehicle crimes and assaults prevalence call not only on technological solutions (e.g., vehicle tracking systems) but also community intervention activities associated with de-escalation and violence prevention efforts.

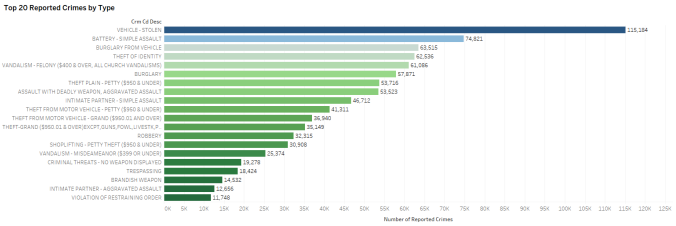


Fig. 3: Top 20 Reported Crime Types

C. *Geographic Distribution of Crime*

In order to check the spatial variation, the crime data was categorized according to LAPD division (area name). Figure 4 shows that the more incidents were committed in the areas of Central, 77th Street, and Pacific divisions on a regular basis. Such divisions are also related to highly populated and business-oriented areas. In urban areas, spatial clustering of crime identifies the urban hot spots and provides the possibility to make data-driven policy choices when it comes to taking area-specific measures and planning the infrastructure.

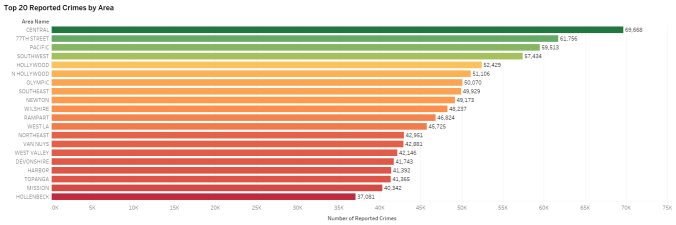


Fig. 4: Crime Count by LAPD Area Name

D. *Victim Demographic Patterns*

Figure 5 presents victim sex distribution across six age groups. In the Child (0–12) category, over half (52.42%) were recorded as ”Unknown,” and 31.67% as Non-Binary, indicating inconsistencies or ambiguity in gender reporting for young victims. From Teen (13–17) through Senior (65+) categories, the distribution stabilizes, with Female victims more prevalent among Teens (59.26%), and a relatively balanced sex ratio observed in the remaining age groups. The Middle-Aged group showed the largest deviation, with Males comprising 56.45%. The given findings allow considering that some age-specific and gender-sensitive interventions can be required to prevent crimes and support their victims efficiently.

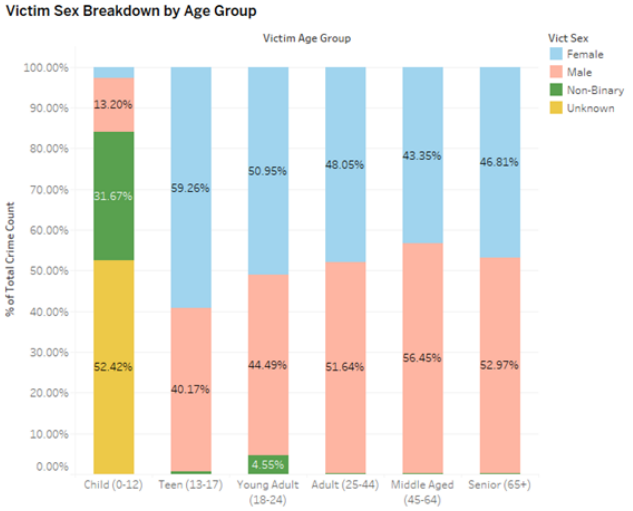


Fig. 5: Victim Sex Breakdown by Age Group

E. Weapon Usage Analysis

Figure 6 displays the 20 most frequently recorded weapon categories in reported crimes. The largest single category, "No Weapon," accounts for over 677,000 incidents. While this includes many cases involving physical force or intimidation without weapons, it may also capture offences where weapons were irrelevant to the crime type (e.g., fraud, vandalism). The second most frequent category, "Strong-Arm," represents bodily force using hands, fists, or similar means. Firearms, such as handguns and semi-automatic pistols, appear in the dataset but represent a relatively small share of total weapon-related records. Overall, these results suggest that most reported incidents during the study period were unarmed or involved non-lethal means. This suggests that community interventions and conflict resolution programs may be more effective than strategies solely focused on weapon control.

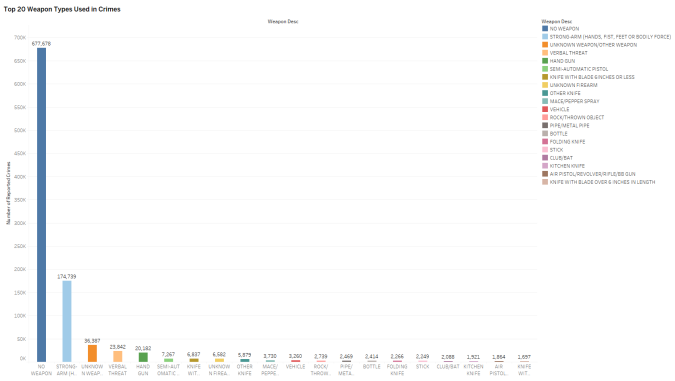


Fig. 6: Top 20 Weapon Types Used in Reported Crimes

V. Conclusions & Future Work

This project provided the scalable end-to-end analysis of the LAPD crime data (2020-2024) in Apache Spark, PostgreSQL, and Tableau and demonstrated patterns over time, geography, crime type, weapon type, and victim demographics. The most important findings included a stable crime rate until late 2023, with a sharp drop in 2024 likely due to reporting changes. The most common reported crimes were vehicle theft and simple assault, and the spatial analysis indicated that the hotspots are consistent throughout the more dense divisions such as Central division and 77 th Street division. Gender analysis carried out through demography revealed some gender imbalances within particular age ranges. The majority of crimes did not involve a weapon, and this fact means that the community-based prevention measures may be discussed along with weapon-oriented enforcement.

As much as these insights help in a better comprehension of the dynamics in urban crimes, the analysis had various limitations. The extent and coherence of LAPD records was not consistent over the demographic levels and particularly in the reporting of victim sex and weapon usage. Moreover, more remarkable consequences affected the 2024 data due to a change in in LAPD's internal reporting systems, which restricts the ability to draw firm conclusions for the final year in scope.

This study could be a stepping stone into which future work could be done in a number of ways. To begin with, integrating socioeconomic, environmental, or mobility datasets (e.g., census or transport data) could help uncover underlying factors behind spatial and temporal crime patterns. Second, the machine learning models (e.g. clustering, time-series forecasting etc.) could be integrated to allow predictive policing insights, which is especially valuable regarding variable resources allocation. Third, it would be possible to implement continuous monitoring and alerting structures by improving real-time or near real-time ingestion pipelines. Finally, collaborating with local stakeholders to improve the granularity and standardization of data—especially around victim demographics—could strengthen both analytical accuracy and policy relevance.

In conclusion, this project illustrates the value of combining scalable technologies and open data to illuminate crime dynamics in complex urban environments. Not in predictive nature, the outputs of analytics are the basis of more responsive and data informed approaches to public safety in Los Angeles and the possibility applies to other metropolitan regions.

VI. Link

- Project Presentation Video Link: https://youtu.be/PzcFx FtFLfQ

References

[1] City of Los Angeles, "Crime Data from 2020 to Present," *Los Angeles Open Data Portal*, [Online]. Available: https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8. [Accessed: Aug. 3, 2025].

[2] A. V. Kumar, P. D. Jeyakumar, and V. D. Prasad, "Crime Data Analysis Using Big Data Analytics and Visualization Using Tableau," in *Proc. 6th Int. Conf. Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2022, pp. 1689–1693, doi: 10.1109/ICECA55472.2022.10012241.

[3] S. Ahmed, N. Aamer, and H. Patil, "Crime Data Analytics Using Hadoop, Spark and Zeppelin," *Int. J. Adv. Innov. Res.*, vol. 7, no. 1(III), pp. 34–40, Jan.–Mar. 2020.

[4] A. Stec and D. Klabjan, “Forecasting Crime with Deep Learning,” *arXiv preprint*, arXiv:1806.01486, Jun. 2018. [Online]. Available: https://arxiv.org/abs/1806.01486. [Accessed: Aug. 3, 2025].

[5] Y. Xiong, “Research on Crime Occurrence Prediction Using Machine Learning Methods—Considering Four Types of Crime in Chicago,” *Int. J. New Developments in Engineering and Society*, vol. 9, no. 1, pp. 7–14, 2025, doi: 10.25236/IJNDES.2025.090102.

[6] V. Mandalapu, L. Elluri, P. Vyas, and N. Roy, “Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions,” *arXiv preprint*, arXiv:2303.16310, Mar. 28, 2023. [Online]. Available: https://arxiv.org/abs/2303.16310. [Accessed: Aug. 3, 2025].