Beyond the Badge: Leveraging Big Data Technologies for Crime Analysis in Los Angeles County (2020-2025)

Amogu J. Uduka
Electrical Engineering
Team 5, GMU ECE552 Big Data Technologies, Spring 2025

***Abstract*—This research project uses big data technologies to analyze crime patterns and trends in Los Angeles County. A comprehensive crime dataset from 2020-2025, including more than 1 million incident reports, geographic coordinates, demographic information, and temporal data was examined by integrating MongoDB and Apache Spark for distributed data processing. This methodology focuses on large-scale data aggregation and statistical analysis to identify spatiotemporal crime hotspots and evaluate the effectiveness of various policing strategies. The project creates interactive visualization dashboards for law enforcement agencies, revealing previously undetected correlations between environmental factors and criminal activity. Key findings demonstrate significant variations in crime patterns across Los Angeles neighborhoods, providing insight into seasonal trends and geographic distributions of different types of offense. This research provides actionable information for resource allocation, community policing initiatives, and policy development, ultimately contributing to more effective crime prevention strategies throughout Los Angeles County.**

***Index Terms*—MongoDB, Apache Spark, Los Angeles County, Crime Patterns**

I. Introduction

Crime in the United States encompasses a wide range of behaviors that violate state or federal statutes. Due to this extensive scope, a meaningful discussion requires breaking the subject into more specific categories [1]. According to recent poll data, public concern about crime has increased significantly in the United States, with approximately 58% of American adults now considering crime reduction a top priority for the President and Congress. This represents a substantial increase from 2021, when the Biden administration began and only 47% of US adults viewed crime reduction as a top priority [2].

The costs of crime are challenging to quantify, with impacts that extend to victims and their families, communities, and the broader society. Researchers are yet to establish a consensus methodology to measure crime-related costs. Generally, these costs fall into two categories: (1) direct costs stemming from criminal incidents and the public expenditures required to maintain the criminal justice system, and (2) indirect costs encompassing non-tangible damages and lost opportunities affecting both individuals involved in the criminal justice system and society as a whole [3].

Los Angeles (often abbreviated as L.A.) is the most populous city in California and the second-largest in the United States, after New York City. As of 2023, the city proper is home to approximately 3.82 million residents, though more recent estimates indicate a population exceeding 4 million (specifically 4,015,546). The broader Los Angeles metropolitan area boasts a population of around 12.9 million as of 2024. In addition to its demographic prominence, L.A. is the central hub of Southern California's economy, finance, and cultural life, known for its rich ethnic and cultural diversity [4].

When it comes to crime, Los Angeles reports roughly 29,400 violent crimes annually, equating to a violent crime rate of 732 per 100,000 people. This includes approximately 258 homicides, 2,274 sexual assaults, 9,652 robberies, and 17,216 aggravated assaults each year [5]. Given its large and dynamic population, the city presents both challenges and opportunities for public safety. Big Data can play a transformative role in helping law enforcement agencies develop more efficient and proactive strategies. By analyzing crime patterns, identifying high-risk zones, predicting potential incidents, and optimizing resource deployment, data-driven insights can lead to smarter policing, improved response times, and ultimately, safer communities. This report seeks to address the following questions;

- *What are the peak times and days for criminal activity?*
- *Are there seasonal patterns in specific types of crimes?*
- *How have crime rates evolved over time for different crime types?*
- *Which areas have the highest concentration of crime?*
- *What victim demography is associated with what crime?*

II. Literature Review

In [6], their analysis revealed notable patterns in criminal activity in the Chicago area, including peak occurrences during summer and weekends. The researchers identified significant disparities in crime rates in various communities and districts. In addition, they developed a predictive classification model using date, time, and location data to forecast crime types, with findings included in their presentation. The results demonstrate the value of crime analysis and modeling approaches. Their investigation shows that criminal activity in Chicago has been on an upward trajectory since 2020, with vehicle theft emerging as the primary driver of this increase. The researchers identified critical areas requiring intervention through detailed mapping of crime patterns in community areas and police districts, highlighting zones where concentrated enforcement efforts could produce significant reductions in criminal incidents. Although acknowledging certain limitations, they successfully developed a predictive model with considerable potential to anticipate criminal activity patterns, including identifying likely locations and timeframes for more serious offenses.

According to [7], the Los Angeles Mayor's Office reported significant improvements in public safety during 2023, with homicides dropping 17% in the city and all geographic offices of the LAPD experiencing reductions in violent crime. Mayor Bass implemented a dual approach to community safety that addressed immediate concerns while developing long-term preventative measures, including strengthening the police department's recruitment efforts and deploying specialized response teams like Crisis and Incident Response through Community Lead Engagement (CIRCLE), which handled nearly 10,000 nonviolent calls. The comprehensive strategy yielded notable results in multiple categories, including a 10% decrease in crimes involving unhoused individuals, a 26% reduction in gang-related homicides and a 10% decrease in shooting victims. These improvements reflect the administration's commitment to both traditional policing methods and innovative community-based interventions designed to create sustainable safety improvements across Los Angeles neighborhoods.

The evolution of modern police services has seen widespread adoption of data analytics technologies. [8] describes this transformation as encompassing both targeted monitoring of specific individuals and broad-spectrum data collection across populations. Law enforcement agencies now increasingly rely on algorithmic tools for predictive policing–both location–based and person–centric—-representing a fundamental shift toward data-driven operations presented as more objective and efficient than traditional methods. This technological advancement has prompted significant scholarly concern regarding possible algorithmic biases that may reinforce historical patterns of inequality, alongside broader questions about privacy rights and civil liberties. Current legal structures have struggled to address these rapid technological developments, raising important questions about police authority, data management practices, and oversight mechanisms. Scholars emphasize the need for a transparent evaluation of these predictive systems and advocate for inclusive development processes that incorporate diverse community voices and interdisciplinary expertise.

III. Data Review

The data analyzed in this report was pulled from the Los Angeles Police Department Data Portal: Crime Data from 2020 to Present. The crime data dataset is a rich source of information for understanding crime patterns over time in Los Angeles. Includes detailed data on crime incidents and locations, which can be instrumental in analyzing crime patterns over time. Such data are crucial for urban planners, law enforcement authorities, and researchers interested in studying urban mobility, public safety, and the effectiveness of police strategies. The data set consists of 1,005,199 rows, 27 columns which includes information about the date and time stamps, crime types, detailed descriptions, and geographical regions. This can help identify peak crime hours, the areas most affected, and the effectiveness of various control measures. It also includes data on the location, dates, and consequences of each reported crime. This information is vital for public safety analysis, helping to identify crime hotspots, common types of offenses, and the impact of neighborhood conditions on safety. The data set can be used by city authorities to improve public safety measures, by researchers to study factors that contribute to local crimes, and by policymakers to develop more effective public safety laws and regulations. Table I provides a complete description of all columns available in the data set, while Figure 4 shows the data schema. This data set is valuable for data analysis, as it provides empirical evidence that can be used to improve public safety management in Los Angeles.

IV. Technical Setup

This document explains a data processing pipeline for analyzing Los Angeles crime data, as illustrated in Figure 1. The pipeline encompasses data acquisition, storage, transformation, and analysis.

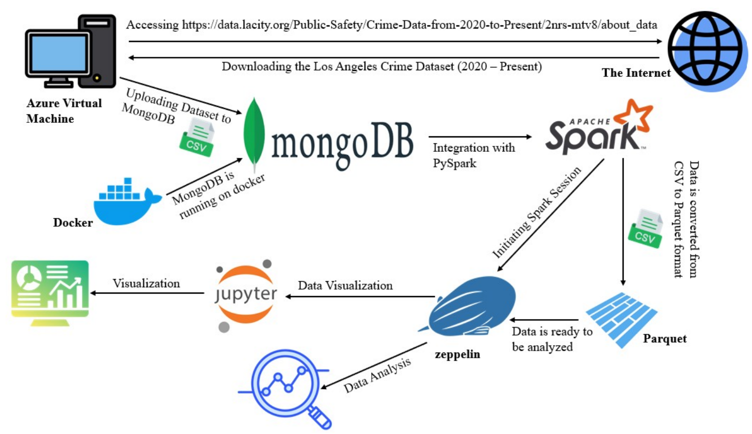

Fig. 1. System Architecture

```
// Rename columns to replace special characters and fix column names
val cleanDF = crimeDF.columns.foldLeft(crimeDF) { (df, colName) =>
  // Replace spaces and other invalid characters with underscores
  val newColName = colName.replaceAll("[.\\s{}()\\[\\]\\\\,-]", "_")
  df.withColumnRenamed(colName, newColName)
}
```

Fig. 2. Regular Expression for Cleaning

The data processing pipeline is implemented on an Azure Virtual Machine with the following specifications: Processor: Intel(R) Xeon(R) Platinum 8370C CPU @ 2.80GHz 2.79GHz Installed RAM: 16 GB, System type: 64-bit operating system, x64-based processor, Windows 10 Pro. The system begins by accessing crime data from the Los Angeles city data portal (data.lacity.org), specifically downloading the "Crime Data from 2020 to Present" dataset. After acquisition, the data is

TABLE I
TABLE 1: COMPLETE STRUCTURE OF LAPD CRIME DATASET (2020-PRESENT)

Column Name	Description
DR_NO	Division of Records Number: Official file number made up of a 2 digit year, area ID, and 5 digits
Date Rptd	MM/DD/YYYY
DATE OCC	MM/DD/YYYY
TIME OCC	In 24 hour military time.
AREA	The LAPD has 21 Community Police Stations referred to as Geographic Areas within the department. These Geographic Areas are sequentially numbered from 1-21.
AREA NAME	The 21 Geographic Areas or Patrol Divisions are also given a name designation that references a landmark or the surrounding community that it is responsible for. For example 77th Street Division is located at the intersection of South Broadway and 77th Street, serving neighborhoods in South Los Angeles.
Rpt Dist No	A four-digit code that represents a sub-area within a Geographic Area. All crime records reference the 'RD' that it occurred in for statistical comparisons.
Crm Cd	Indicates the crime committed. (Same as Crime Code 1)
Crm Cd Desc	Defines the Crime Code provided.
Mocodes	Modus Operandi: Activities associated with the suspect in commission of the crime.
Vict Age	Two character numeric
Vict Sex	F - Female M - Male X - Unknown
Vict Descent	Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian
Premis Cd	The type of structure, vehicle, or location where the crime took place.
Premis Desc	Defines the Premise Code provided.
Weapon Used Cd	The type of weapon used in the crime.
Weapon Desc	Defines the Weapon Used Code provided.
Status	Status of the case. (IC is the default)
Status Desc	Defines the Status Code provided.
Crm Cd 1	Indicates the crime committed. Crime Code 1 is the primary and most serious one. Crime Code 2, 3, and 4 are respectively less serious offenses. Lower crime class numbers are more serious.
Crm Cd 2	May contain a code for an additional crime, less serious than Crime Code 1.
Crm Cd 3	May contain a code for an additional crime, less serious than Crime Code 1.
Crm Cd 4	May contain a code for an additional crime, less serious than Crime Code 1.
LOCATION	Street address of crime incident rounded to the nearest hundred block to maintain anonymity.
Cross Street	Cross Street of rounded Address
LAT	Latitude
LON	Longitude

uploaded as Comma Separated Values (CSV) to a MongoDB database, which is deployed within a Docker container to provide isolation and portability.

The transformation process involves integration with Apache Spark, which converts the data from CSV format to Parquet format. Parquet is a columnar storage format that improves query performance. The analysis environment consists of Zeppelin as the central analysis platform, which connects to the processed data. From Zeppelin, two main activities occur: data visualization, which was used alongside a Jupyter Notebook and data analysis (performing statistical analysis and data exploration).

This architecture follows a modern data engineering approach with several key benefits: separation of storage (MongoDB) and compute (Apache Spark), format optimization (CSV to Parquet conversion), containerization (Docker) for consistent deployment, and interactive analysis capabilities (Zeppelin and Jupyter Notebook). The workflow represents a complete Extract, Transform, Load (ETL) pipeline specifically designed for analyzing Los Angeles crime data.

V. DATA PREPARATION

Crime data is first downloaded from [9], and then extracted from MongoDB using Spark connectors. The workflow begins by importing necessary libraries from the MongoDB Spark connector and establishing a connection to a MongoDB instance located at "mongodb://127.0.0.1:27018/sampleDataset.crimeData". The data is then loaded into a Spark DataFrame structure, which enables efficient distributed processing of the crime records. Data cleaning is a crucial step in this process, with the code specifically addressing problematic column names. The implementation uses a folding operation to systematically rename columns, replacing spaces and special characters with underscores through a regular expression pattern as seen in Figure 2. This standardization ensures compatibility with downstream analysis tools that may have strict naming requirements. Once cleaned, the data undergoes transformation to create a more analysis-ready structure while preserving the original information integrity. The final steps involve persisting the prepared data in Parquet format, an efficient columnar storage format ideal for analytical workloads. The code writes the cleaned DataFrame to a specified path and then verifies the output by reading it back and displaying sample rows. The schema output confirms the data structure includes geographic information (latitude/longitude), timestamps, crime classifications, and location descriptors - all properly typed as integers, strings, or doubles with appropriate nullability settings. This comprehensive preparation pipeline creates a foundation for subsequent crime data analysis in Spark. The Data flow diagram is presented in Figure 6.

VI. DATA ANALYSIS AND RESULTS

The analysis of crime data across Los Angeles provides valuable insights into temporal and spatial patterns of criminal activity. The findings are organized into the following sections:

- Daily Distribution of Criminal Activity
- Seasonal Variations in Crime Rates
- Geographic Concentration of Criminal Activity
- Frequency Distribution of Crime Types
- Crime Victimization by Demographic Groups
- Intersection of Crime Types and Demographics
- Intersection of District and Victim Demographics
- Machine Learning Analysis

daily distribution, seasonal variations, and geographic concentration. [10] presents the github repository for this project.

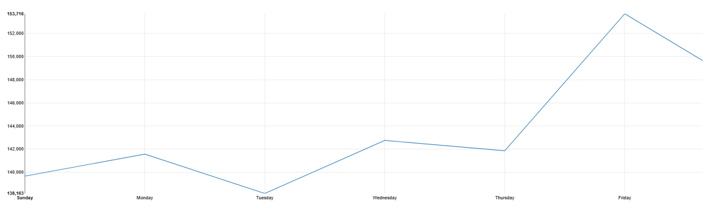

Fig. 3. Daily Crime Count

A. Daily Distribution of Criminal Activity

The analysis of crime distribution by day of week (Figure 3) reveals a distinct weekly pattern with **Friday** showing the highest crime rates (153,716 incidents), representing a peak in criminal activity. **Saturday** follows with approximately 147,000 incidents, while **Tuesday** recorded the lowest crime frequency (138,163 incidents).

This weekly pattern demonstrates a 10.9% difference between the highest and lowest days, suggesting a gradual buildup of criminal activity during the workweek that culminates on Friday before declining over the weekend. **Wednesday** and **Thursday** show moderate levels (approximately 142,000 incidents), creating a characteristic midweek plateau.

The data indicates that law enforcement resource allocation should be adjusted to accommodate the higher demand for services toward the end of the work week, particularly on Fridays when criminal activity reaches its peak.

B. Seasonal Variations in Crime Rates

The monthly distribution analysis Figure 5 demonstrates significant seasonal variations in crime rates throughout the year. **January** recorded the highest crime count (92,776 incidents), followed by a gradual decline through spring and early summer months.

The seasonal pattern reveals an inverse relationship between temperature and overall crime rates in Los Angeles, with cooler months (January–March) experiencing higher crime rates than warmer months. The lowest crime rates occur in **December** (78,224 incidents), potentially influenced by holiday periods. This represents a 15.7% difference between the highest and lowest months.

The data also shows a secondary peak in **March** (approximately 88,000 incidents) and minor fluctuations in the latter half of the year, with small increases in **July–August** and **October**. These patterns suggest the need for seasonal-specific

```
df.printSchema()

root
 |-- AREA: integer (nullable = true)
 |-- AREA_NAME: string (nullable = true)
 |-- Crm_Cd: integer (nullable = true)
 |-- Crm_Cd_1: integer (nullable = true)
 |-- Crm_Cd_2: integer (nullable = true)
 |-- Crm_Cd_3: integer (nullable = true)
 |-- Crm_Cd_Desc: string (nullable = true)
 |-- Cross_Street: string (nullable = true)
 |-- DATE_OCC: string (nullable = true)
 |-- DR_NO: integer (nullable = true)
 |-- Date_Rptd: string (nullable = true)
 |-- LAT: double (nullable = true)
 |-- LOCATION: string (nullable = true)
 |-- LON: double (nullable = true)
 |-- Mocodes: string (nullable = true)
 |-- Part_1-2: integer (nullable = true)
 |-- Premis_Cd: integer (nullable = true)
 |-- Premis_Desc: string (nullable = true)
 |-- Rpt_Dist_No: integer (nullable = true)
 |-- Status: string (nullable = true)
 |-- Status_Desc: string (nullable = true)
 |-- TIME_OCC: integer (nullable = true)
 |-- Vict_Age: integer (nullable = true)
 |-- Vict_Descent: string (nullable = true)
 |-- Vict_Sex: string (nullable = true)
 |-- Weapon_Desc: string (nullable = true)
 |-- Weapon_Used_Cd: integer (nullable = true)
 |-- _id: struct (nullable = true)
 |    |-- oid: string (nullable = true)
```

Fig. 4. DataFrame Schema

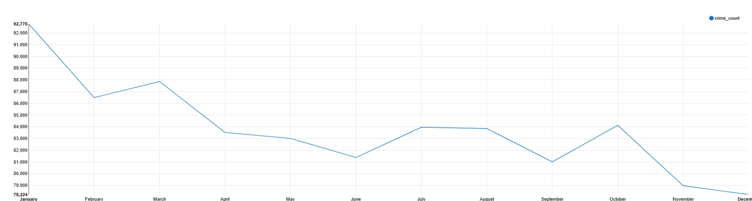

Fig. 5. Monthly Crime Count

crime prevention strategies, with heightened vigilance during the early months of the year.

C. Geographic Concentration of Criminal Activity

The spatial analysis shown in Figure 7 reveals substantial variation in crime distribution across Los Angeles districts.

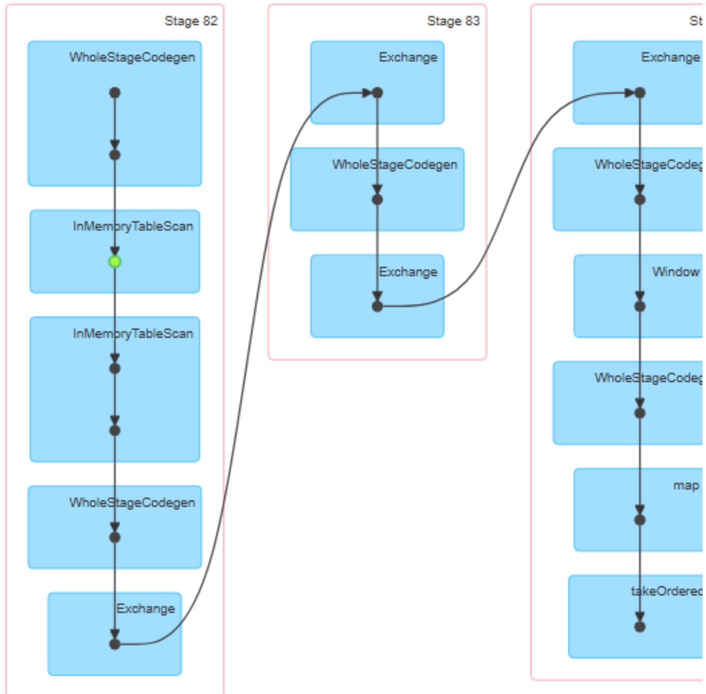

Fig. 6. Data Flow

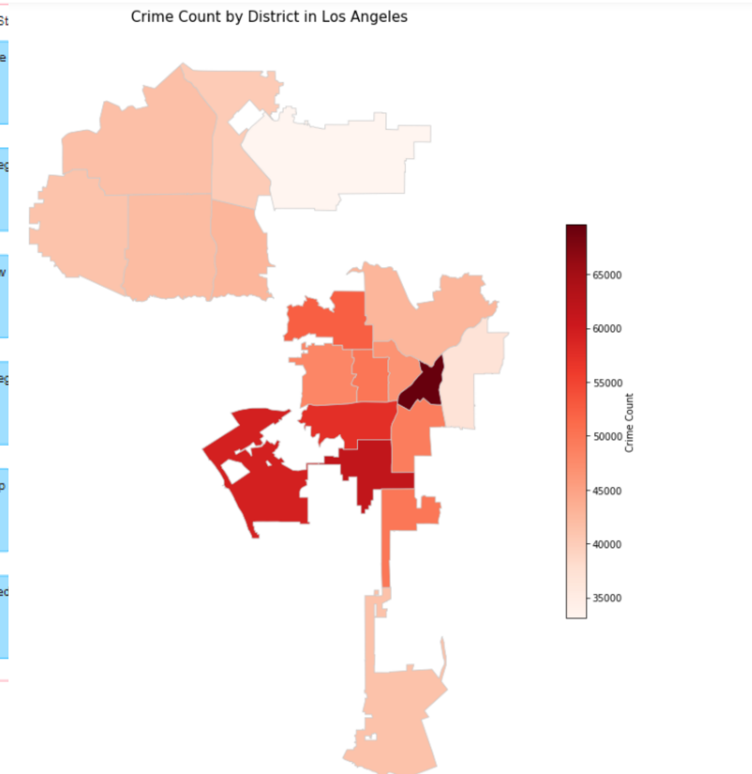

Fig. 7. District Mappings by Crime Count

The choropleth map demonstrates that central and southern districts experience significantly higher crime rates (55,000–65,000 incidents), compared to northern and eastern districts (35,000–40,000 incidents).

One district in particular recorded over 65,000 criminal incidents, making it the most crime-intensive area in the city. The western coastal regions and several northeastern districts show the lowest crime rates (under 35,000 incidents). This geographic distribution highlights pronounced disparities, with some districts experiencing nearly twice the criminal activity as others.

The spatial pattern reveals a concentration of crime in the central and southern regions, with a clear north-south divide in criminal activity levels. This finding suggests the need for targeted interventions in high-crime districts.

D. Frequency Distribution of Crime Types

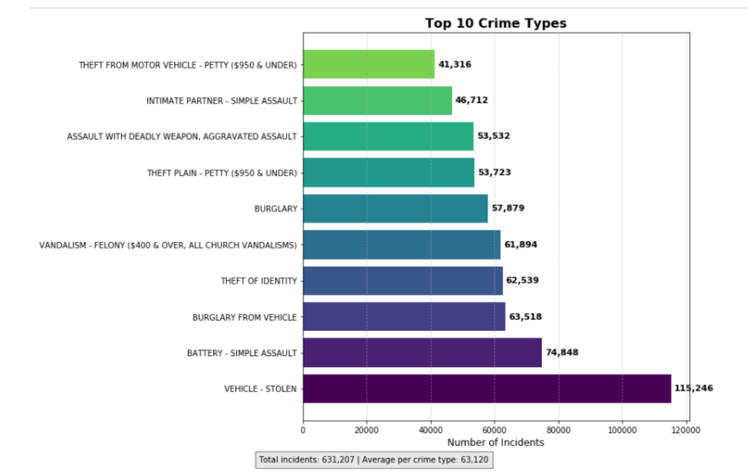

Fig. 8. Top 10 Crime Type

As shown in Figure 8, vehicle theft represents the most prevalent crime type with 115,246 incidents, significantly exceeding other categories. This figure is approximately 54% higher than the second most common crime, battery-simple assault, which recorded 74,848 incidents. The substantial gap between vehicle theft and other crime categories suggests that vehicular crimes constitute a disproportionate share of the total criminal activity in the region.

The data reveals three distinct tiers of crime frequency:

- High-frequency crimes (> 70,000 incidents): Vehicle theft and battery-simple assault
- Medium-frequency crimes (50,000-70,000 incidents): Burglary from vehicle, theft of identity, vandalism-felony, burglary, theft plain-petty, and assault with deadly weapon
- Lower-frequency crimes (¡50,000 incidents): Intimate partner-simple assault and theft from motor vehicle-petty

In particular, the total recorded incidents across all the top 10 crime types amounts to 631,207, with an average of 63,120 incidents per crime category. This high volume underscores the significant challenge faced by law enforcement agencies in addressing these prevalent offenses.

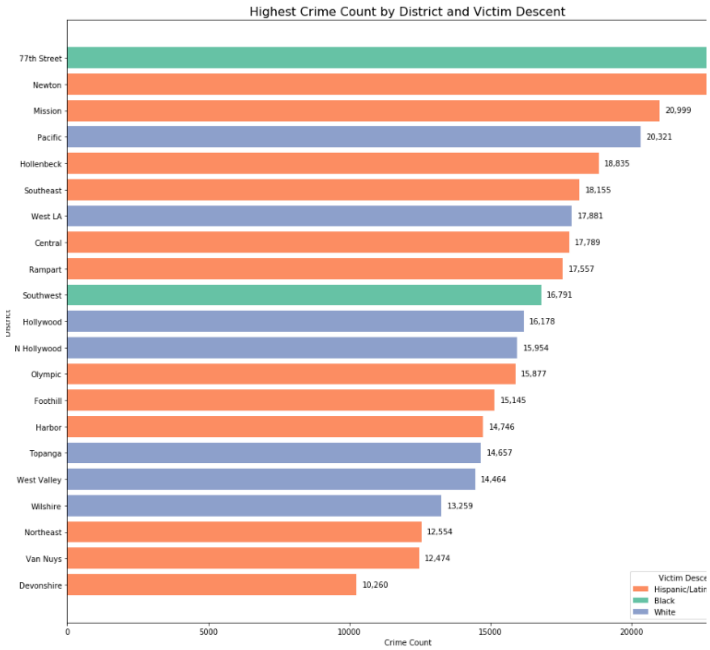

Fig. 9. Victim Demographics by District

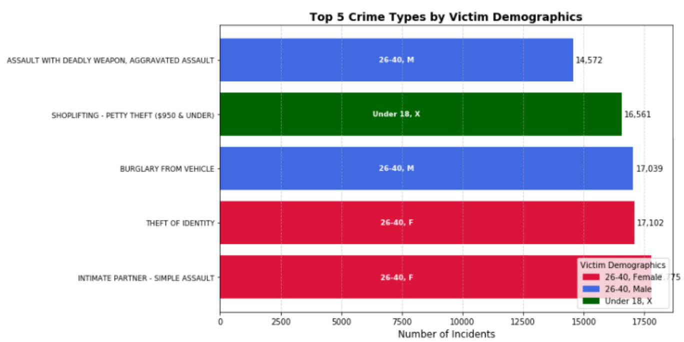

Fig. 10. Top 5 Crime Type by Victim Demographics

E. Crime Victimization by Demographic Groups

Figure 10 offers valuable insight into the demographic patterns of victimization in five major types of crime. Several significant observations emerge:

- Gender-specific patterns: Females in the 26-40 age group are disproportionately affected by both intimate partner-simple assault (17,175 incidents) and theft of identity (17,102 incidents). These crimes show a marked gender disparity in victimization.
- Age-specific vulnerability: Individuals under 18 years of age show particular vulnerability to shoplifting-petty theft, with 16,561 recorded incidents. This suggests that juvenile victims represent a significant demographic in certain property crimes.
- Male victimization patterns: Males in the 26-40 age bracket are primarily affected by vehicle burglary (17,039 incidents) and assault with deadly weapon/aggravated assault (14,572 incidents). This indicates that violent and property crimes targeting males follow different patterns than those primarily affecting females.

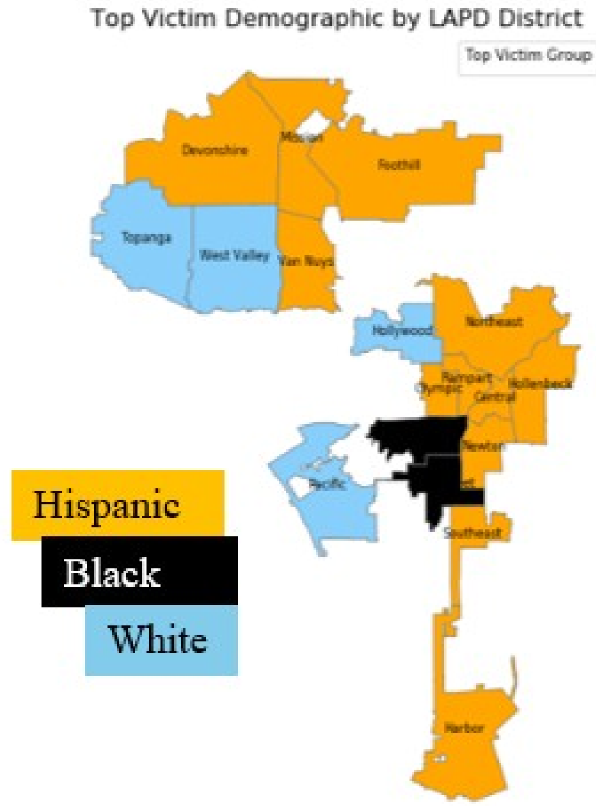

Fig. 11. Map Showing Victim Descent

The demographic analysis reveals that adults aged 26-40 constitute the most frequently victimized age group in four of the five major crime categories. This finding has significant implications for targeted crime prevention strategies and resource allocation.

F. Intersection of Crime Types and Demographics

When comparing the two figures (8 & 10), several notable correlations emerge. Although vehicle theft appears to be the most common crime in general, it is not among the top five crimes when categorized by victim demographics. This might suggest that vehicle theft affects a broader demographic spectrum rather than concentrating within specific age or gender groups.

Similarly, intimate partner-simple assault ranks relatively low in the overall frequency (8th position), but appears as one of the top five crimes when analyzed by demographic distribution, with a clear concentration among women aged 26-40 years. This highlights how certain crimes may not appear predominant in aggregate statistics but reveal significant patterns when examined through demographic lenses.

G. Intersection of District and Victim Demographics

As depicted in both Figure 9 and Figure 11, the crime count analysis reveals significant patterns between districts and victim demographics, with the 77th Street district recording the highest crime count (22,995) primarily affecting Black victims, followed closely by Newton (22,763) and Mission (20,999) districts where Hispanic/Latin/Mexican victims predominate. Pacific district ranks fourth with 20,321 incidents, mostly affecting White victims. The data illustrate a clear demographic distribution of victimization: Hispanic/Latin/Mexican victims represent the majority in ten districts, white victims in six districts, and black victims in just two districts (77th Street and Southwest). The districts with the lowest crime counts (under 13,000) include Wilshire, Northeast, Van Nuys, and Devonshire, with Devonshire showing the lowest overall count of approximately 10,260 incidents, primarily affecting Hispanic/Latin/Mexican victims.

H. Machine Learning Analysis

1) Sex Prediction Model Performance: The sex prediction model achieved an accuracy of 60.4% across the test dataset, significantly outperforming the baseline random classifier. Feature importance analysis revealed that spatiotemporal factors exerted the strongest influence on victim sex prediction. Weekend status (0.2257) emerged as the single most predictive feature, followed by geographic coordinates (longitude: 0.1135, latitude: 0.0430). Demographic features such as age group (0.0069) contributed modestly to predictive power, while time-of-day features (hour: 0.0000, night status: 0.0000) demonstrated negligible impact.

The confusion matrix (Table II) illustrates classification patterns across the 146,661 victim records in the test dataset. The model correctly identified 39,046 female victims (55% of actual females) and 16,882 male victims (23% of actual males). A systematic bias toward predicting female victims was observed, as evidenced by the misclassification of 56,715 male victims as female.

TABLE II
SEX PREDICTION CONFUSION MATRIX

Sex	Pred. Female	Pred. Male	Pred. Other
F	39,046	31,860	0
M	56,715	16,882	0
X	1,347	796	3
H	7	5	0

Crime-specific analysis revealed context-dependent prediction patterns. Homicide cases predominantly yielded male victim predictions regardless of actual sex, while sexual assault cases consistently yielded accurate female victim predictions (Table III). This suggests crime-type-specific sex targeting patterns effectively captured by the model.

2) Demographic-Crime Type Correlations: Analysis of crime patterns across demographic groups revealed substantial variation in victimization profiles. Table IV presents the three most frequent crime types by victim sex. Female victims

TABLE III
SAMPLE SEX PREDICTIONS BY CRIME TYPE

Crime Type	Area	Hour	Actual Sex	Predicted Sex
Criminal Homicide	Central	12	F	M
Criminal Homicide	Central	20	M	M
Criminal Homicide	Central	9	M	M
Criminal Homicide	Central	14	F	M
Rape, Forcible	Central	3	F	F
Rape, Forcible	Central	19	F	F
Rape, Forcible	Central	0	F	F
Rape, Forcible	Central	23	F	F

disproportionately experienced identity theft (35,715 cases) and intimate partner violence (35,204 cases), while male victims were predominantly affected by battery (38,825 cases) and aggravated assault (36,961 cases). It is worthy to note that the Central area was the area used for this analysis due different areas had different relationships

TABLE IV
MOST COMMON CRIMES BY VICTIM SEX

Sex	Primary Crime Type	Count
F	Theft of Identity	35,715
F	Intimate Partner - Simple Assault	35,204
F	Battery - Simple Assault	34,801
M	Battery - Simple Assault	38,825
M	Assault with Deadly Weapon, Aggravated Assault	36,961
M	Burglary from Vehicle	34,823
X	Shoplifting - Petty Theft	1,923
X	Vandalism - Felony	1,015
X	Burglary	966

When stratified by age group (Table V), victims under 18 predominantly experienced battery (4,089 cases) and child abuse (3,403 cases), while the 26-40 age cohort was most frequently targeted for vehicle burglary (30,301 cases) and identity theft (27,358 cases).

TABLE V
MOST COMMON CRIMES BY AGE GROUP

Age Group	Primary Crime Type	Count
Under 18	Battery - Simple Assault	4,089
Under 18	Child Abuse (Physical) - Simple Assault	3,403
Under 18	Assault with Deadly Weapon, Aggravated Assault	2,365
18-25	Battery - Simple Assault	10,958
18-25	Intimate Partner - Simple Assault	9,879
18-25	Burglary from Vehicle	9,435
26-40	Burglary from Vehicle	30,301
26-40	Theft of Identity	27,358
26-40	Battery - Simple Assault	25,043
41-60	Battery - Simple Assault	23,964
41-60	Theft of Identity	19,268
41-60	Burglary from Vehicle	17,121
Over 60	Battery - Simple Assault	9,848
Over 60	Burglary	8,744
Over 60	Theft of Identity	8,015

Combined age-sex analysis revealed particularly pronounced victim profiles (Table VI). Young women (18-25) were disproportionately affected by intimate partner violence (8,104 cases), while middle-aged men (26-40) experienced the highest incidence of vehicle burglary (17,039 cases). Young men (18-25) faced elevated rates of aggravated assault (6,176 cases), and for victims of unknown sex, shoplifting was the predominant crime type across all age groups.

TABLE VI
SELECTED CRIME PATTERNS BY AGE-SEX COMBINATION

Age-Sex	Primary Crime Type	Count
18-25_F	Intimate Partner - Simple Assault	8,104
18-25_M	Assault with Deadly Weapon, Aggravated Assault	6,176
26-40_F	Intimate Partner - Simple Assault	17,779
26-40_M	Burglary from Vehicle	17,039
41-60_F	Battery - Simple Assault	11,015
41-60_M	Battery - Simple Assault	12,872
Over 60_F	Battery - Simple Assault	4,400
Over 60_M	Battery - Simple Assault	5,430
Under 18_F	Battery - Simple Assault	2,193
Under 18_M	Battery - Simple Assault	1,863

These findings demonstrate significant demographic stratification in crime victimization patterns across Los Angeles, with implications for targeted prevention strategies and resource allocation.

I. Implications for Crime Prevention

These findings point to several focus areas for crime prevention efforts:

1) Vehicle-related security measures should be prioritized given the prominence of vehicle theft and burglary from vehicles.
2) Targeted interventions that address intimate partner violence focusing on women aged 26-40 could address significant vulnerability.
3) Identity theft protection initiatives would benefit from targeting women in the 26-40 age bracket.
4) Youth-oriented programs that address shoplifting among people under 18 years of age could serve as effective preventive measures.

The observed patterns suggest that crime prevention strategies would benefit from demographic targeting rather than focusing solely on crime types in isolation. The substantial variation in victimization patterns between demographic groups indicates that one-size-fits-all approaches may have limited effectiveness in reducing crime rates.

VII. CONCLUSIONS

The temporal and spatial patterns identified in this analysis offer valuable insights for law enforcement and public policy. The combined findings suggest that:

- Resource allocation should be adjusted to account for higher Friday crime rates (153,716 incidents) and early-year seasonal peaks, with January showing the highest crime count (92,776 incidents).
- Targeted interventions should be developed for high-crime districts in central and southern Los Angeles, where some areas experience nearly twice the criminal activity compared to northern and eastern districts.
- Crime prevention strategies should be seasonally adjusted to address the 15.7% higher rates observed during winter months compared to December.
- Demographic-specific approaches should target the distinct victimization patterns identified: women aged 26-40 experiencing higher rates of identity theft (35,715 cases) and intimate partner violence (35,204 cases), while men in the same age group face elevated risks of vehicle burglary (17,039 cases) and aggravated assault (14,572 cases).
- Vehicle-related security measures should be prioritized given the prominence of vehicle theft (115,246 incidents) and burglary from vehicles, which significantly outpaced other crime categories.
- Youth-oriented programs should address the vulnerability of individuals under 18 years to shoplifting-petty theft (16,561 incidents).

The machine learning model for demographic prediction demonstrated promising results with 60.4% accuracy, revealing that spatiotemporal factors—particularly weekend status (0.2257) and geographic coordinates—exerted the strongest influence on victim demographic prediction. This suggests potential for predictive policing applications that could anticipate criminal activity patterns.

These distinct temporal, spatial, and demographic patterns demonstrate the importance of data-driven approaches to understanding and addressing crime in urban environments. The significant variations across time, space, and victim demographics highlight the need for tailored, context-specific interventions rather than one-size-fits-all approaches to crime prevention and public safety.

By leveraging big data technologies like MongoDB and Apache Spark for distributed data processing, law enforcement agencies can transform raw crime statistics into actionable intelligence. This technological approach enables more efficient resource allocation, targeted community policing initiatives, and ultimately more effective crime prevention strategies throughout Los Angeles County.

REFERENCES

[1] S. R. Department. (2024) Crime in the u.s. statista. [Accessed: Apr. 15, 2025]. [Online]. Available: https://www.statista.com/topics/2153/crime-in-the-united-states/#topicOverview

[2] J. Gramlich. (2024) What the data says about crime in the u.s. Pew Research Center. [Accessed: Apr. 15, 2025]. [Online]. Available: https://www.pewresearch.org/short-reads/2024/04/24/what-the-data-says-about-crime-in-the-us/

[3] BJS. (2022) Cost of crime. Bureau of Justice Statistics. [Accessed: Apr. 15, 2025]. [Online]. Available: https://bjs.ojp.gov/topics/costs

[4] Wikipedia. (2023) Los angeles. Wikipedia. [Accessed: Apr. 15, 2025]. [Online]. Available: https://en.wikipedia.org/wiki/Los_Angeles

[5] S. . D. P.C. (2024) Most dangerous cities in california based on fbi violent crime data. Spolin Dukes P.C. [Accessed: Apr. 15, 2025]. [Online]. Available: https://www.spolinlaw.com/criminal-defense/most-dangerous-cities-in-california/

[6] H. Frields and M. Risaldar, "Ece 552 big data: Analysis of chicago crime," 2025, unpublished manuscript, Department of Electrical and Computer Engineering.

[7] Office of Mayor Karen Bass. (2023) Lapd releases end of year crime statistics for the city of los angeles 2023. City of Los Angeles. [Accessed: Apr. 16, 2025]. [Online]. Available: https://mayor.lacity.gov/news/lapd-releases-end-year-crime-statistics-city-los-angeles-2023

[8] S. Brayne, "Big data surveillance: The case of policing," *American Sociological Review*, vol. 82, no. 5, pp. 977–1008, 2017.

[9] U.S. Department of Justice, “Crime data from 2020 to present,” https://catalog.data.gov/dataset/crime-data-from-2020-to-present, 2025, accessed: April 29, 2025.

[10] A. J. Uduka, “GitHub - AmoguJUduka: Crime Data Repository,” https://github.com/AmoguJUduka, 2025, accessed: April 29, 2025.