Crime Detection Analysis using PySpark

A report on
Big Data Analytics Lab Project
[CSE-3263]

Submitted By
Jayasuryan Mutyala - 210962009
Prahlad Menon - 210962057

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MANIPAL INSTITUTE OF TECHNOLOGY,
MANIPAL ACADEMY OF HIGHER EDUCATION

April 2024

Crime Detection Analysis using PySpark

Jayasuryan Mutyala[1], Prahlad Menon[2]

[1]Computer Science and Engineering (Artificial Intelligence and Machine Learning) & MIT Manipal, India

[2] Computer Science and Engineering (Artificial Intelligence and Machine Learning) & MIT Manipal, India

[1]jsmu.dev@gmail.com; [2]prahladmenon02@gmail.com;

Abstract*— This project leverages big data analytics to revolutionize crime detection and prevention, enhancing public safety and security by analysing diverse data sources like crime reports and demographic information. Employing data processing, machine learning algorithms, and geospatial analysis, including K-means and K-modes clustering, it identifies crime hotspots in urban and disadvantaged neighbourhoods. The system offers real-time monitoring and predictive modelling, enabling law enforcement agencies to make evidence-based decisions. Innovative visualizations, such as a safety index map and police station cluster centres, provide actionable insights for targeted interventions. This approach not only elucidates complex crime dynamics but also promotes collaboration among stakeholders, enhancing urban resilience and quality of life in Los Angeles and beyond.

Keywords: KMeans clustering, KModes Clustering, Geospatial Analysis, Predictive Modelling.

I. INTRODUCTION

In an era where data-driven decision-making is critical, by meticulously curating and analysing a comprehensive dataset encompassing crime reports, demographic information, and temporal-spatial details, we aim to construct a robust system that not only identifies but also predicts criminal hotspots within urban environments. Many police departments employed artificial intelligence algorithms and big data tools to assist them in forecasting crime scenes. For instance, police departments in Seattle, Los Angeles, and Atlanta have experimented with predictive police programs, which aim to pinpoint the future location of crime hotspots. In parallel, the Chicago Police Station employed a heat list-based algorithm to classify individuals as offenders based on violent offenses or patterns of abuse. To prevent crime, modern technologies can handle massive amounts of data that old analytic approaches cannot handle. As a result, big data and artificial techniques are used to analyse crime data. The proposed methodology emphasises greatly on data cleaning techniques, ensuring the integrity and accuracy of our analysis. This involves refining unique identifiers for crime reports, standardizing time, and date formats, and implementing strategies to deal with missing or inconsistent data. Through the deployment of advanced machine learning algorithms - K-means and K-modes clustering, the model learns the diverse patterns of criminal activity. This can be further integrated with law enforcement agencies and policymakers providing actionable insights, facilitating targeted interventions and strategic resource allocation. This project prioritizes scalability, performance, and ethical considerations, is a critical step in using big data to protect our communities and support security and justice values.

II. LITERATURE REVIEW

This paper[1] discusses the challenges of crime prediction and the potential of big data analysis to address them. The authors propose a model that uses recurrent neural networks (RNNs) to predict the type, time, and place of crimes. The model was tested on a dataset of crimes in Chicago and achieved promising results. The authors conclude that their model can be used to help law enforcement agencies predict and prevent crime. Crime prediction is a challenging task, as it is often difficult to identify patterns in crime data. However, big data analysis has the potential to overcome these challenges by providing access to large amounts of data that can be used to identify patterns and trends. The authors of this paper propose a model that uses RNNs to predict crime. RNNs are a type of artificial neural network that are well-suited for sequential data, such as crime data. The model was tested on a dataset of crimes in Chicago and achieved promising results.

[1] H. Hashim and A. T. Abdulameer, “Crime Prediction Using Big Data Analysis,” Sep. 2021

The authors conclude that their model can be used to help law enforcement agencies predict and prevent crime. However, it is important to note that the model is still in its early stages of development, and more research is needed to validate its effectiveness. In addition, the use of big data for crime prediction raises a number of ethical concerns. For example, it is important to ensure that the data is used fairly and does not discriminate against certain groups of people. Overall, this article is a promising step forward in the field of crime prediction. However, more research is needed to validate the effectiveness of the model and to address the ethical concerns that it raises.

The authors have highlighted the diverse sources and voluminous nature of crime data that necessitate advanced processing techniques. Big data in crime analysis encompasses various dimensions, including spatial, temporal, and categorical features, which pose challenges in traditional data management and analysis systems (Liu et al., 2018). Recent studies emphasise the role of advanced technologies, such as machine learning algorithms and data mining techniques, in enhancing crime detection and prevention efforts. PySpark, with its distributed computing capabilities, offers an efficient platform for implementing these algorithms on large-scale crime datasets (Haque et al., 2020). Crime data classification involves identifying patterns and trends within vast datasets to facilitate proactive law enforcement strategies. PySpark facilitates the implementation of sophisticated pattern recognition algorithms, such as clustering and association rule mining, to uncover hidden insights from crime data (Li et al., 2019) Traditional crime scene analysis methods have limitations in handling the increasing volume and complexity of crime data. PySpark-based frameworks enable real-time data processing and visualisation, empowering investigators with actionable insights for swift and effective crime resolution(Chen et al.,2021).

This paper[2] on Crime prediction has seen significant advancements, particularly with the integration of machine learning (ML) and artificial intelligence (AI) techniques. With crime analysis being a critical aspect of criminology, researchers have sought to leverage ML algorithms to extract crime patterns from large datasets efficiently. Such efforts are crucial due to the complexities inherent in crime prevention, including diverse crime types, motives, and handling methods. Machine learning models have emerged as powerful tools for predicting future crimes, allowing police departments to optimise their resources by identifying hotspots based on various factors such as time, type, and location. Traditional methods like hotspot analysis, while reactive, are being supplemented or replaced by AI-driven approaches that analyse past crime data to forecast future events. These advancements enable proactive measures in crime prevention and resource allocation, contributing to enhanced community safety and economic growth.

The author showcases the diversity of approaches in crime prediction research, ranging from generic crime prediction to specific crime types and categories. Various ML algorithms and techniques have been employed, with studies often comparing different models to identify the most effective ones for a given dataset. Moreover, recent surveys have provided comprehensive overviews of crime prediction methods, exploring factors such as socioeconomic, spatial-temporal, demographic, and geographic attributes. The research also highlights the importance of selecting appropriate models based on dataset compatibility and the need for field testing to assess the usability of proposed approaches. Additionally, systematic reviews have focused on AI strategies for crime prediction, analysing model performance metrics, dataset utilisation, and validation approaches. By covering research from 2008 to 2021, this review underscores the evolving landscape of crime prediction and the increasing reliance on ML and AI technologies to address complex societal challenges related to crime prevention and public safety.

The integration of big data analytics into crime detection and prevention strategies has transformed law enforcement practices in recent years. This literature review synthesises insights from studies spanning from 2010 to 2020, exploring the utilisation of big data analytics in crime detection and prevention. With the exponential growth of digital data sources, including social media, sensor networks, and surveillance cameras, law enforcement agencies now have unprecedented opportunities to analyse vast amounts of data to anticipate and combat criminal activities. By leveraging big data analytics techniques, such as machine learning algorithms, data mining, and geographic information systems (GIS), actionable intelligence can be extracted from diverse data streams to identify crime patterns and allocate resources effectively.

[2] F. Dakalbab, M. Abu Talib, O. Abu Waraga, A. Bou Nassif, S. Abbas, and Q. Nasir, "Artificial intelligence & crime prediction: A systematic literature review," Social Sciences & Humanities Open, vol. 6, no. 1, p. 100342, 2022.

Machine learning algorithms serve as the backbone of crime pattern analysis, enabling the development of predictive models that forecast crime hotspots, identify trends, and optimise resource allocation. From decision trees to neural networks, a variety of machine learning techniques have been employed to analyze crime data and extract valuable insights. Complementing machine learning, data mining techniques uncover significant correlations and clusters within large datasets, offering insights into the spatial and temporal dynamics of criminal activities. Geographic information systems (GIS) provide a spatial perspective to crime pattern analysis, facilitating visualisation and analysis of crime data in relation to geographical features.

Despite the promise of big data analytics in crime detection and prevention, several challenges persist. Data privacy concerns, data quality issues, algorithmic bias, and ethical considerations surrounding predictive analytics in law enforcement necessitate careful attention and mitigation strategies. Addressing these challenges requires multidisciplinary collaboration between researchers, law enforcement agencies, policymakers, and other stakeholders to ensure the responsible and effective use of data-driven approaches. This literature review aims to provide a comprehensive overview of existing research, identify gaps in the literature, and offer insights into future directions for leveraging big data analytics in enhancing public safety and security. Through synthesising insights from diverse studies, this review contributes to the ongoing discourse on the role of advanced analytics in crime detection and prevention.

In conclusion, the integration of big data analytics and machine learning techniques into crime detection and prevention strategies represents a significant advancement in modern law enforcement practices. Through the synthesis of insights from various studies spanning from 2010 to 2020, this literature review underscores the transformative potential of leveraging diverse data sources, such as social media, sensor networks, and surveillance cameras, to anticipate and combat criminal activities. Machine learning algorithms, including decision trees, neural networks, and recurrent neural networks (RNNs), have emerged as powerful tools for analysing crime data and extracting actionable intelligence. Complemented by data mining techniques and geographic information systems (GIS), these approaches enable the development of predictive models that identify crime hotspots, uncover hidden patterns, and optimise resource allocation.

However, despite the promise of big data analytics in crime detection and prevention, several challenges persist. Ethical considerations surrounding data privacy, algorithmic bias, and the fair use of predictive analytics in law enforcement necessitate careful attention and mitigation strategies. Moreover, ensuring data quality and addressing issues related to interdisciplinary collaboration remain critical for the responsible and effective implementation of data-driven approaches. Moving forward, addressing these challenges requires continued multidisciplinary collaboration between researchers, law enforcement agencies, policymakers, and other stakeholders.

By providing a comprehensive overview of existing research, identifying gaps in the literature, and offering insights into future directions, this review contributes to the ongoing discourse on the role of advanced analytics in enhancing public safety and security. Ultimately, the responsible and ethical use of big data analytics holds immense potential to empower law enforcement agencies in their efforts to combat crime, safeguard communities, and uphold the principles of justice and security in the digital age.

III. METHODOLOGY

A. Summary

Our methodology for crime detection across Los Angeles follows the given steps below:

- Data cleaning
- Exploratory Data Analysis
- Data Visualization
- Kmeans clustering model
- Analysis of Kmeans silhouette score
- KModes clustering model
- Insights drawn from KModes and KMeans clustering.

The data preprocessing and cleaning phase was done to improve the quality of the Los Angeles (2010–2019) crime data dataset obtained from Kaggle for our study, we carried out a thorough Extract, Transform, Load (ETL) procedure. We verified the uniqueness of crime identifiers, standardizing date and time formats, and fixing inconsistencies or missing values in multiple critical attributes, including {Mocodes}, {Vict Age}, {Vict Sex}, and `Vict Descent}, were all part of the process. We also separated geographic coordinates, removed unnecessary columns, and built dimensional data frames to optimize data administration. The result was cleaned data and dimensional tables exported as CSV files. Robust data-driven insights are supported by this methodological approach, which greatly enhances the dataset's integrity and makes it easier for it to be prepared for further analytical study and database integration.

In the data visualization phase the Critical findings include the percentage of force-related crime categories (coded 0400), the identification of 137,845 force-related occurrences, and the examination of crimes perpetrated by strangers (coded 1822). The survey also reveals differences in the length of time it takes to report different kinds of crimes, including higher waits for sexual offenses. An essential component of the visualization is the mapping of all crimes, which provides a clear and thorough geographical overview of crime patterns by color-coding serious Part 1 crimes (like murder and robbery) and less serious Part 2 crimes (like simple assaults and drug laws).

In the modelling building phase where KMeans and KModes clustering is used to analyse crime data from Los Angeles. Using KModes, incidents are ideally clustered into five groups according to their geographic closeness. The results are visually displayed on a map, which may help in identifying high-crime regions and potential police station locations. On the other hand, KModes clustering, which makes use of 10 clusters, groups data points with comparable characteristics to address the categorical factors in the dataset, such as the type of crime and the weapon used. By separating spatial patterns with KMeans and category similarities with KModes, this split method enables a thorough examination of the data and provides a full picture of crime trends and occurrences in the city.

B. Dataset

The dataset consists of historic crime data in Los Angeles provided in Kaggle as a part of the Police Violence and Racial Equity dataset. This dataset includes comprehensive records of crimes that were recorded in the Central Los Angeles area between January 1st, 2010, to 2019. Each entry includes detailed information on a particular criminal incident, such as the Division of Records number (DR_NO), the time of the crime (TIME OCC), the date and time the report was filed (Date Rptd), and the day the crime was committed (DATE OCC). Together with the reporting district number (Rpt Dist No), the dataset additionally identifies the specific region by number (region) and name (AREA NAME).

The FBI's Uniform Crime Reporting (UCR) hierarchy divides crimes into two categories: Part 1 (more serious offenses, such as violent crimes) and Part 2 (other crimes, such as vandalism or drug offenses) (Part 1-2). A crime code (Crm Cd), a descriptive title (Crm Cd Desc), and, if relevant, modus operandi codes (Mocodes) that offer more information on the crime's commission are used to further categorize each crime.

Victim data is broken down by age (Vict Age), sex (Vict Sex), and descent (Vict Descent), offering demographic details on the people who are impacted. The site type (Premis Cd and Premis Desc) of the crime, including whether it happened at a single-family home, on the street, or somewhere else, is also described in the dataset. The weapon used (Weapon Used Cd and Weapon Desc), the case status (Status and Status Desc), and the precise location of the crime (LOCATION, Cross Street, LAT, LON) are supplied when applicable.

The dataset shows trends in criminal activity in the Central region, including victim demographics, incident locations, and common offenses (such as theft, burglary, and assault). For law enforcement to allocate resources and devise strategies, as well as for policymakers and researchers to comprehend the dynamics of urban crime, this knowledge is essential.

# DR_NO	Date Rptd	DATE OCC	# TIME OCC	# AREA	AREA NAME	# Rpt Dist No
001307355	02/20/2010 12:00:00 AM	02/20/2010 12:00:00 AM	1350	13	Newton	1385
011401303	09/13/2010 12:00:00 AM	09/12/2010 12:00:00 AM	0045	14	Pacific	1485
070309629	08/09/2010 12:00:00 AM	08/09/2010 12:00:00 AM	1515	13	Newton	1324
090631215	01/05/2010 12:00:00 AM	01/05/2010 12:00:00 AM	0150	06	Hollywood	0646
100100501	01/03/2010 12:00:00 AM	01/02/2010 12:00:00 AM	2100	01	Central	0176
100100506	01/05/2010 12:00:00 AM	01/04/2010 12:00:00 AM	1650	01	Central	0162
100100508	01/08/2010 12:00:00 AM	01/07/2010 12:00:00 AM	2005	01	Central	0182
100100509	01/09/2010 12:00:00 AM	01/08/2010 12:00:00 AM	2100	01	Central	0157
100100510	01/09/2010 12:00:00 AM	01/09/2010 12:00:00 AM	0230	01	Central	0171

Fig. 1 Dataset preview from Kaggle.

```
root
 |-- DR_NO: integer (nullable = true)
 |-- Date_Rptd: string (nullable = true)
 |-- DATE_OCC: string (nullable = true)
 |-- TIME_OCC: integer (nullable = true)
 |-- AREA: integer (nullable = true)
 |-- AREA NAME: string (nullable = true)
 |-- Rpt_Dist_No: integer (nullable = true)
 |-- Part_1-2: integer (nullable = true)
 |-- Crm_Cd: integer (nullable = true)
 |-- Crm_Cd_Desc: string (nullable = true)
 |-- Mocodes: string (nullable = true)
 |-- Vict_Age: integer (nullable = true)
 |-- Vict_Sex: string (nullable = true)
 |-- Vict_Descent: string (nullable = true)
 |-- Premis_Cd: integer (nullable = true)
 |-- Premis_Desc: string (nullable = true)
 |-- Weapon_Used_Cd: integer (nullable = true)
 |-- Weapon_Desc: string (nullable = true)
 |-- Status: string (nullable = true)
 |-- Status_Desc: string (nullable = true)
 |-- Crm_Cd_1: integer (nullable = true)
 |-- Crm_Cd_2: integer (nullable = true)
 |-- Crm_Cd_3: integer (nullable = true)
 |-- Crm_Cd_4: integer (nullable = true)
 |-- LOCATION: string (nullable = true)
 |-- Cross_Street: string (nullable = true)
 |-- LAT: double (nullable = true)
 |-- LON: double (nullable = true)
```

Fig. 2 Schema of PySpark Data Frame.

C. Data cleaning

The cleaning phase begins with handling the DR_NO column, which includes distinct IDs for every crime report, was the first step. This was a critical step in guaranteeing data integrity, enabling the elimination of possible duplicates, and laying the groundwork for precise analysis. The date and time columns (Date Rptd, DATE OCC, and TIME OCC) in the dataset were then examined. The choice to convert the TIME OCC column to a 24-hour format and standardize these columns to a more database-friendly date format (YYYY-MM-DD) further emphasizes on data consistency and analytical readiness.

The issue of missing or erroneous values in a number of columns, including Mocodes, Vict Age, Vict Sex, and Vict Descent, was also resolved throughout the cleaning process. Through the process of adding placeholders for missing data and fixing incorrect values (such negative ages), the dataset was made more reliable for further studies in addition to being more complete. This meticulous attention to detail made sure that there were no distortions from missing or erroneous entries, and that every variable in the dataset appropriately mirrored the underlying facts.

A major structural optimization was achieved by reorganizing the dataset into dimensional data frames, especially for geographic locations. This approach improved the dataset's navigability and analytical usefulness by facilitating the separation and deduplication of spatial data. The cleaning procedure set the stage for more complex geographical analysis and visualizations by generating unique dimensional tables for attributes like location, allowing a closer look at Los Angeles crime trends.

The cleaned dataset and the newly created dimensional tables were saved as CSV files. This marked the end of the cleaning process and guaranteed that the dataset was ready for additional analysis, database integration. This thorough and meticulous cleaning procedure turned the dataset from a raw collection of crime reports into a polished, accessible resource.

D. Data Visualization

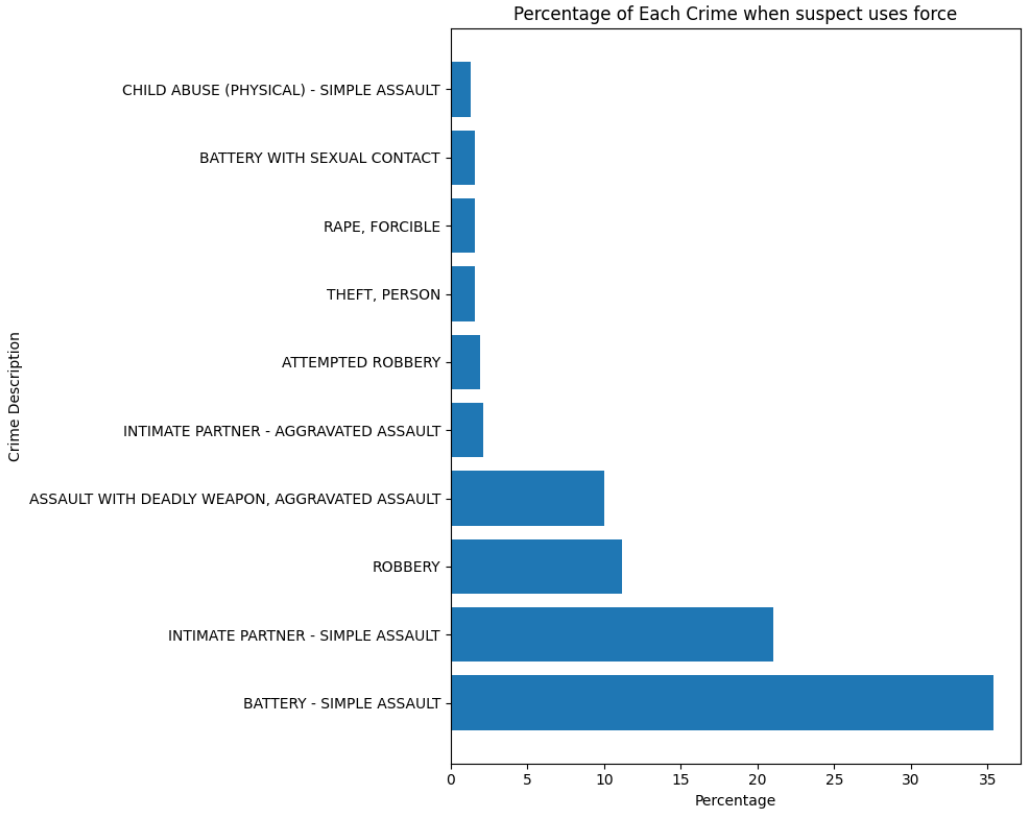

Fig. 3 Visualization of results where the percentage of each crime type where suspect uses force

The horizontal bar chart displays the percentage of various crimes committed when the suspect uses force.

Key insights from fig 3:

Battery - Simple Assault:

- This crime has the highest percentage, indicating that it is the most common when force is used by the suspect, with a percentage around 35%. This could suggest that simple assault is a prevalent issue and often involves physical force.

Intimate Partner - Simple Assault:

- The second most frequent crime involving force concerns incidents between intimate partners, which are slightly less than the percentage of simple assaults. This underscores the serious issue of violence in intimate relationships.

Robbery:

- Represented with a significant percentage, robbery is the third most common crime involving the use of force. This aligns with the nature of robbery, which typically involves the threat or use of force.

Assault with Deadly Weapon, Aggravated Assault:

- With a percentage near 15%, this severe crime involving force is also alarmingly common. The use of a deadly weapon indicates the potential for serious injury or fatality.

Intimate Partner - Aggravated Assault:

- This category has a slightly lower percentage than the previous, suggesting that aggravated assaults in intimate relationships are also a significant concern.

Attempted Robbery:

- Attempted robbery has a smaller percentage compared to completed robberies, which might indicate that a proportion of attempted robberies are thwarted or do not succeed.

Theft, Person:

- The use of force in thefts directly from a person is relatively less common but still significant, which may include snatching or forceful taking of property.

Rape, Forcible:

- Forcible rape has a relatively lower percentage compared to other violent crimes involving force. However, the impact of such crimes is profound and traumatic.

Battery with Sexual Contact:

- Similar to forcible rape, this crime involves a sexual component and the use of force, although it represents a smaller percentage of the total.

Child Abuse (Physical) - Simple Assault:

- This crime has the lowest percentage on the chart but is particularly concerning because it involves minors. The fact that it is represented at all indicates that child abuse involving force is an issue that needs attention.

These insights demonstrate that simple assaults, including those against intimate partners, are the most common types of crimes where suspects use force. Violent crimes, particularly those with aggravated elements or involving deadly weapons, make up a significant portion as well. The data can be instrumental for law enforcement and social services in prioritizing resource allocation, prevention programs, and victim support services. It highlights the need for continued attention to interpersonal violence and robbery due to their prevalence and the harm they cause to individuals and communities.

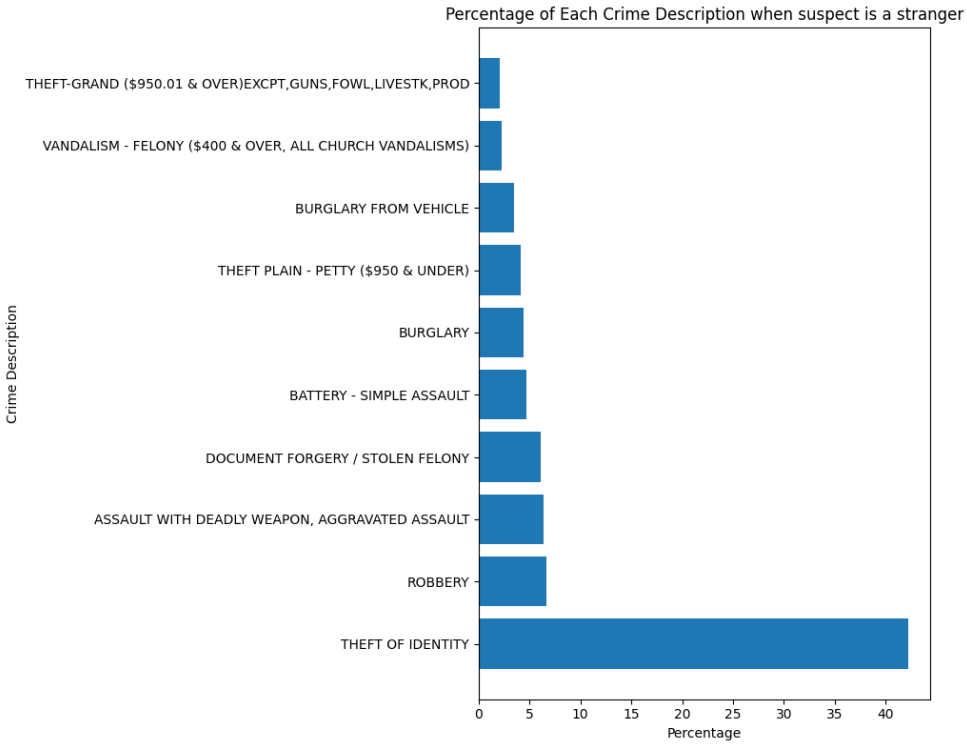

Fig. 4 Visualization of results where the percentage of each crime type where suspect is a stranger.

The horizontal bar plot visualises the percentage of each crime description when the suspect is a stranger.

Key Insights for fig 4:
Theft of Identity

- This category significantly outnumbers other crimes, which could indicate that it's easier for strangers to commit this crime without personal interaction or that there is a rising trend in identity-related crimes due to increased online activity.
- The most common crime in this dataset is when the suspect is a stranger, with the percentage approaching 40%.
- The high percentage might also reflect the difficulty in securing personal information in the digital age, necessitating stronger measures for identity protection.

Robbery

- As the second-most frequent crime committed by strangers, it suggests a high rate of confrontational crimes where the perpetrator doesn't know the victim, potentially indicating areas with higher rates of street crime.
- The fact that robbery accounts for about 15% could imply that it is a significant concern for law enforcement and might require targeted interventions such as increased patrols in hot spots.

Assault with Deadly Weapon, Aggravated Assault

- This type of violent crime being the third most common suggests that when strangers commit violent acts, they can often be severe, potentially reflecting societal or community stressors that escalate conflicts.
- The rate of approximately 10% also underscores the need for community education on conflict resolution and the importance of accessible emergency services.

Document Forgery/Stolen Felony and Battery - Simple Assault

- These two crimes having similar percentages indicate that non-violent financial crimes and personal physical assault are both prevalent issues when dealing with crimes by strangers, possibly requiring diverse approaches in prevention.
- Both account for approximately 7.5% to 8% each.
- The relatively high percentage of these crimes may reflect both the opportunistic nature of document-related crimes and the prevalence of personal disputes escalating to physical violence.

Burglary, Theft Plain - Petty ($950 & Under), and Burglary from Vehicle

- The presence of these three types of theft in the middle range of the chart suggests that theft of property is a common intention among crimes by strangers, with various methods and targets.
- Their percentages being close to each other might indicate that law enforcement could focus on property crime prevention strategies, like encouraging better home and vehicle security measures.

Vandalism - Felony ($400 & Over, All Church Vandalisms)

- Being less common could suggest that while still present, crimes of this nature are not as prevalent as theft or assault, which might be due to the specific intent required or the opportunities to commit such acts.
- The rate of around 4% may reflect a threshold of tolerance within communities for reporting such crimes or could be indicative of the effectiveness of existing deterrents against property damage.
- Theft-Grand ($950.01 & Over) Except Guns, Fowl, Livestock, Prod

- The fact that this crime is the least frequent could suggest that larger-scale thefts are more challenging to commit or that they are more effectively policed and prevented.
- The close to 3% prevalence indicates that while significant, the opportunities to commit grand theft without being detected might be less frequent or that such crimes are more likely to be reported and investigated with rigor due to the higher value involved.

Based on this data, it can be concluded that identity theft is more frequent than other crimes when the victim is unaware of the suspect. Most offenses in this category are crimes related to theft, such as burglaries, robberies, and various types of theft. This could indicate that strangers are more prone to commit property crimes than personal crimes. Though they are less frequent, violent crimes like assault are nonetheless serious. Law enforcement and community awareness campaigns that concentrate on preventative measures for the crimes most likely to be perpetrated by strangers may find this information to be helpful.

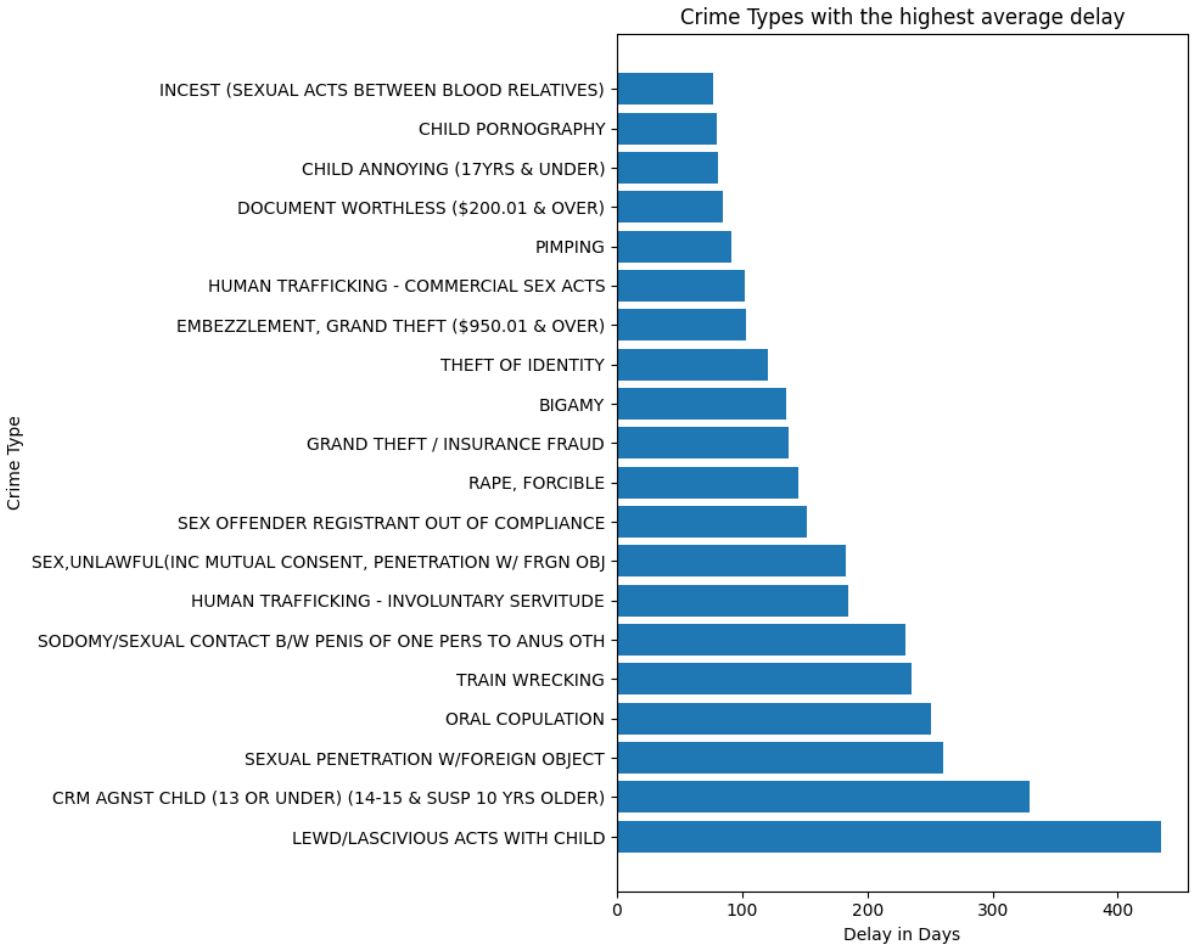

Fig. 5 Visualization of each crime type with the highest average delay.

The horizontal bar chart provides various crime types with their associated average delay in days. This delay refers to the time between the crime being committed and it being reported, or between the report and the subsequent action taken by authorities.

Key insights from fig 5:

Lewd/Lascivious Acts with Child:

- This crime has the highest average delay, significantly longer than the others, with the delay approaching 400 days. This could indicate a reluctance to report these crimes promptly due to their sensitive nature, or it could suggest challenges in discovering or proving these offenses.

CRM Against Child (13 or under) (14-15 & Susp 10 yrs Older):

- Crimes against children with suspects significantly older also show a very high average delay, over 300 days. This might suggest similar reporting challenges as the previous point or reflect the time needed to investigate these allegations thoroughly.

Sexual Penetration w/Foreign Object:

- Another serious sexual crime with a high average delay, also over 300 days. The delay could be due to the complexity of such cases, the availability of evidence, or the time victims take to come forward.

Oral Copulation:

- This crime's average delay is over 250 days, which could indicate the time needed for the victims to report such crimes or for authorities to compile a viable case.

Train Wrecking:

- Although a less common crime, it has a significant average delay, which could be due to the rarity and complexity of these cases, making investigation and attribution time-consuming.

Sodomy:

- The average delay here is just over 200 days. As with other sexual crimes, the delay may be influenced by the factors mentioned earlier, like the time it takes for victims to report and for authorities to build cases.

Human Trafficking - Involuntary Servitude:

- This crime shows an average delay over 200 days. The complexities involved in uncovering and investigating human trafficking cases, especially those involving involuntary servitude, could contribute to this delay.

Sex Offender Registrant Out of Compliance:

- This type of offense has an average delay around the 200-day mark. It suggests that tracking non-compliant sex offenders is a challenging and time-consuming process.

Rape, Forcible:

- This serious crime shows an average delay nearing 200 days. The delay could result from the time it takes for victims to report the crime, the collection of evidence, and the investigation process.

Grand Theft / Insurance Fraud:

- This crime type has an average delay approaching 150 days. The complexity in uncovering fraudulent activities might contribute to such delays.

Bigamy:

- Surprisingly, this crime has a similar delay to theft and fraud-related crimes, which might reflect the time taken to discover and provide evidence for such cases.

Theft of Identity:

- Interestingly, this crime has a lower average delay than many of the others, though still significant. It might indicate that identity theft can be quicker to report and prove than other types of crimes.

Embezzlement, Grand Theft ($950.01 & Over):

- This financial crime shows an average delay of about 100 days, potentially due to the time required to notice and verify the embezzlement or grand theft.

Human Trafficking - Commercial Sex Acts:

- The average delay is slightly less than that for embezzlement, which might be related to the difficulties in detecting and proving such crimes.

Pimping:

- Shows an average delay close to those related to human trafficking. The reporting and investigation processes for such crimes can be complex, contributing to the delay.

Document Worthless ($200.01 & Over):

- A crime involving fraudulent or worthless documents shows a moderate average delay, which could be due to the time taken to detect such activities.

Child Annoying (17yrs & Under):

- The average delay is lower compared to more serious offenses, yet still notable. The term 'annoying' could cover a range of behaviours, affecting the delay in reporting.

Child Pornography:

- The delay for child pornography is also on the lower side compared to other sexual crimes, possibly due to the digital nature of the evidence that might be quicker to report.

Incest (Sexual Acts Between Blood Relatives):

- This crime has the lowest average delay on the chart, which is still substantial. This might indicate that such cases, while difficult, may come to light quicker than other types of sexual crimes due to family dynamics.

The overall insight is that sexual crimes, particularly those against children, tend to have the longest delays in this dataset, potentially due to the sensitive nature of these crimes and the challenges involved in reporting and investigation. Financial crimes and crimes against property have shorter delays but are still significant, likely due to the complexities in detecting and prosecuting these offenses. This information is crucial for law enforcement and support services in understanding the dynamics of crime reporting and the potential need for improved processes to encourage timely reporting and efficient investigations.

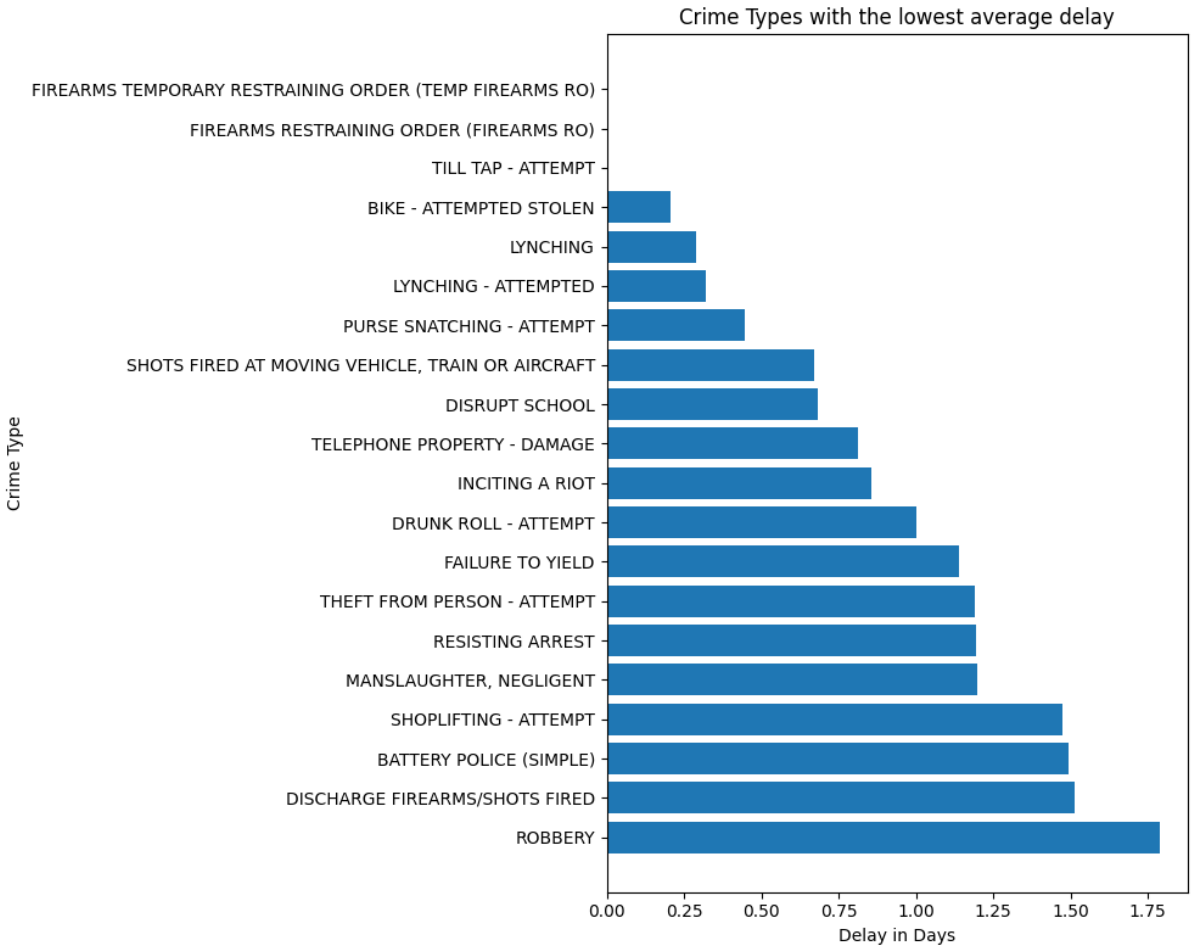

Fig. 6 Visualization of each crime type with the lowest average delay.

This horizontal bar chart depicts various crime types with the lowest average delay in days, likely indicating the time between the occurrence of the crime and its reporting or response by law enforcement.

Key insights from fig 6

Robbery:

- This crime type has the shortest average delay, which is close to zero. This suggests that robberies are reported almost immediately, likely due to their violent and urgent nature which demands a quick response.

Discharge Firearms/Shots Fired:

- Similarly, incidents involving firearms are also reported very quickly, which makes sense given the immediate danger they present and the high priority law enforcement places on such reports.

Battery Police (Simple):

- Offenses against police officers, such as simple battery, have a very low average delay. These are typically reported instantly because they occur in the presence of law enforcement.

Shoplifting - Attempt:

- Attempts at shoplifting are quickly noticed and reported, reflected in the low average delay. Retail establishments often have measures in place to detect and report these incidents promptly.

Manslaughter, Negligent:

- Negligent manslaughter cases have a short delay, likely because these incidents are usually accidental and immediately evident to those involved or to witnesses, leading to rapid reporting.

Resisting Arrest:

- Incidents of resisting arrest are reported with little to no delay because they happen in the presence of law enforcement during an attempted arrest.

Theft from Person - Attempt:

- Attempted theft from a person, such as pickpocketing, is often quickly detected and reported by the victim or onlookers.

Failure to Yield:

- This is a traffic-related offense that is noticed immediately during traffic stops or accidents, leading to a low average delay in reporting.

Drunk Roll – Attempt:

- An attempt to rob an intoxicated person is quickly reported, either by the victim or bystanders, hence the low average delay.

Inciting a Riot:

- The act of inciting a riot is likely to be reported rapidly due to its public and potentially volatile nature.

Telephone Property - Damage:

- Damage to telephone property is often detected quickly by the service providers or users, leading to swift reporting.

Disrupt School:

- Incidents that disrupt school activities are usually reported with minimal delay because they occur in a structured environment with established protocols for incident reporting.

Shots Fired at Moving Vehicle, Train or Aircraft:

- Like other firearms-related incidents, these are serious and reported almost immediately due to the potential for mass harm.

Purse Snatching - Attempt:

- Attempts at purse snatching are quickly noticed and thus reported rapidly, often while the crime is in progress or immediately after.

Lynching - Attempted and Lynching:

- These are grave offenses that, when identified, are reported with little delay, although they are relatively rare in occurrence.

Bike - Attempted Stolen:

- Attempted bike thefts are usually spotted and reported quickly, especially in public or monitored areas.

Till Tap - Attempt:

- An attempt to steal from a cash register is typically noticed immediately by employees or security systems.

Firearms Restraining Order (FIREARMS RO) and Firearms Temporary Restraining Order (TEMP FIREARMS RO):

- Incidents involving breaches of firearms restraining orders are likely reported quickly due to the clear and present danger they represent.

The common thread among these crimes is the immediacy of their impact and the presence of witnesses or victims who can report them quickly. They tend to be either public, immediately dangerous, involve law enforcement directly, or occur in monitored environments. This information is valuable for understanding the responsiveness of reporting systems and can be used to assess the efficiency of law enforcement response and the public's awareness of different types of crimes.

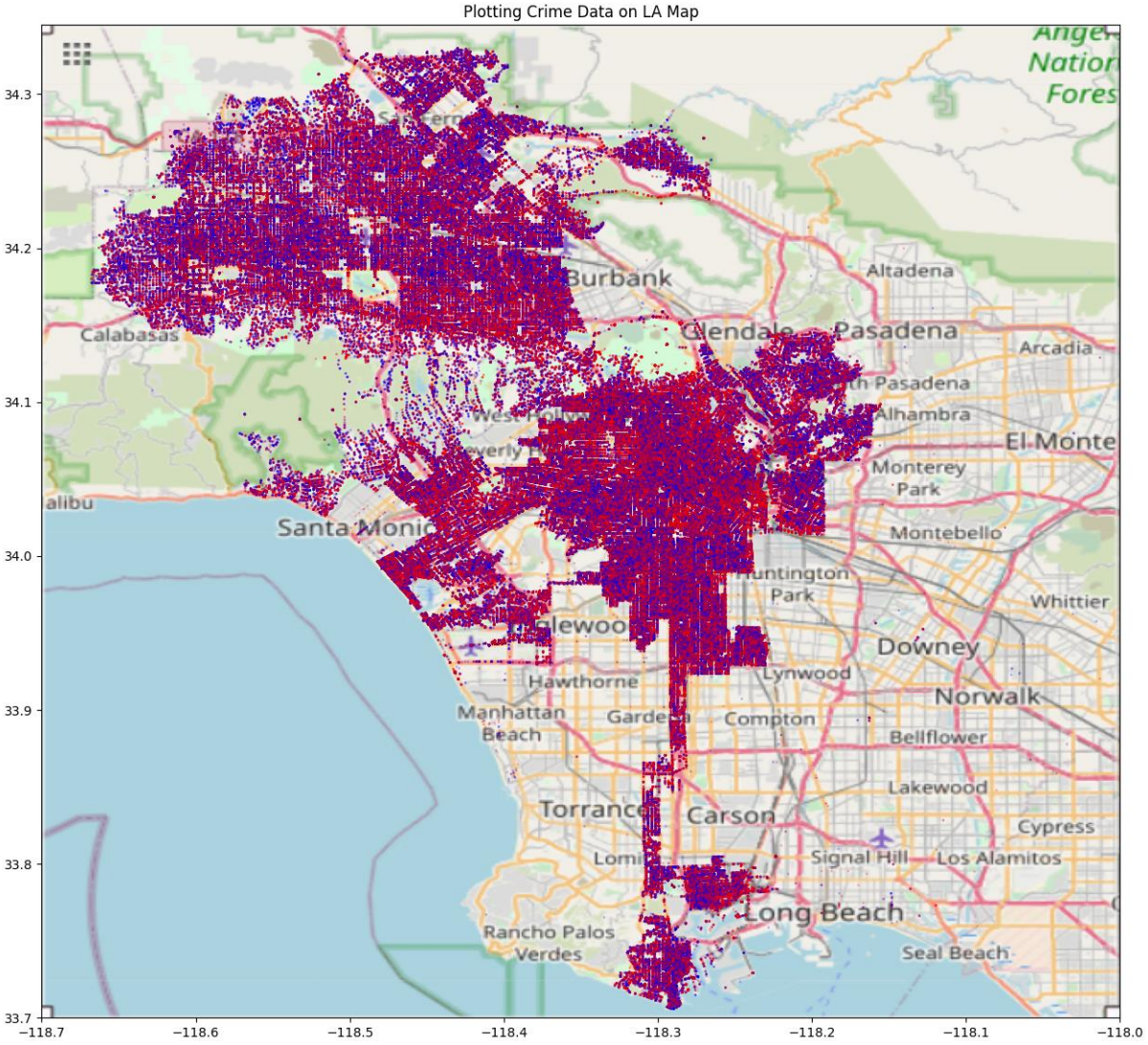

Fig. 7 Plotting all crimes on the LA map. Red dots are part-1 crimes (murder, manslaughter, sex offenses, robbery, aggravated assault, burglary, motor vehicle theft, and arson) and blue dots are part-2 crimes (Simple Assaults, Forgery/Counterfeiting, Embezzlement/Fraud, Receiving Stolen Property, Weapon Violations, Prostitution, Sex Crimes, Crimes Against Family/Child, Narcotic Drug Laws, Liquor Laws, Drunkenness, Disturbing the Peace etc.).

E. KMeans and KModes clustering

Kmeans clustering

KMeans clustering algorithm is applied to geographic data represented by latitude and longitude coordinates. The VectorAssembler joins the coordinates into a single vector per data point, and the StandardScaler normalizes these features. This is the first step in the process of preparing the data. Then, using this pre-processed data, KMeans is used to explore different cluster sizes (k) and choose the best configuration based on the silhouette score, a measure that compares an object's similarity to its own cluster against other clusters to assess cluster cohesion and separation.

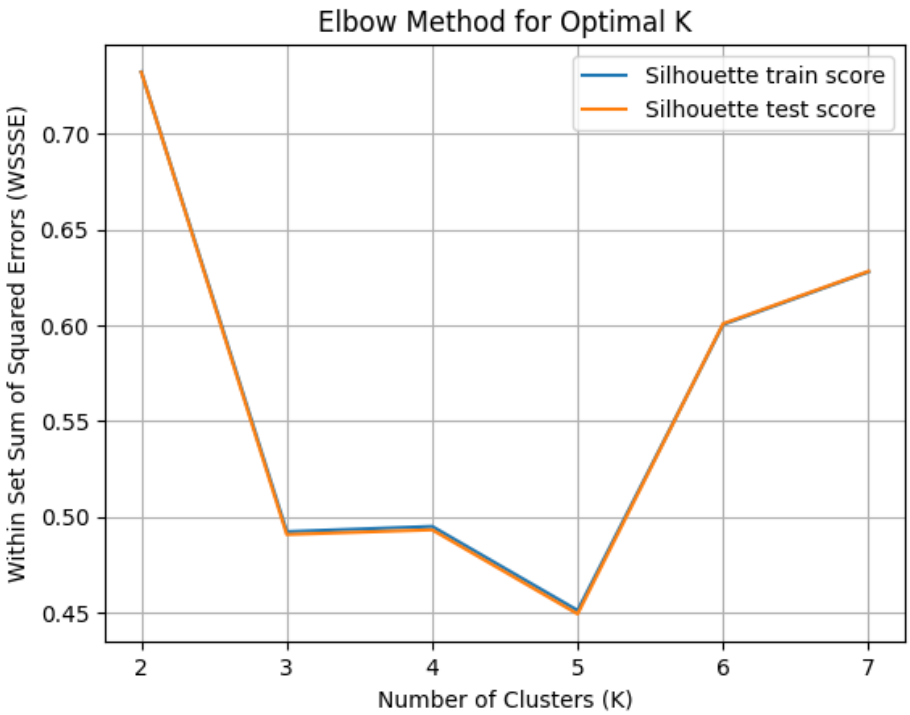

Fig. 8 Using the elbow method the determine the optimal k value.

```
Silhouette Score for k=2 | Train score: 0.7322942428377445 Test score: 0.7321367584818453
Silhouette Score for k=3 | Train score: 0.49225226047092585 Test score: 0.4907743987839934
Silhouette Score for k=4 | Train score: 0.4949913578659054 Test score: 0.4932126213544947
Silhouette Score for k=5 | Train score: 0.451123712957093 Test score: 0.4492901444280098
Silhouette Score for k=6 | Train score: 0.6003993627878642 Test score: 0.6008392474224924
Silhouette Score for k=7 | Train score: 0.6279229300047391 Test score: 0.6281913898139837
```

Fig. 9 Comparing the silhouette scores for k values ranging from k = 2 to 7.

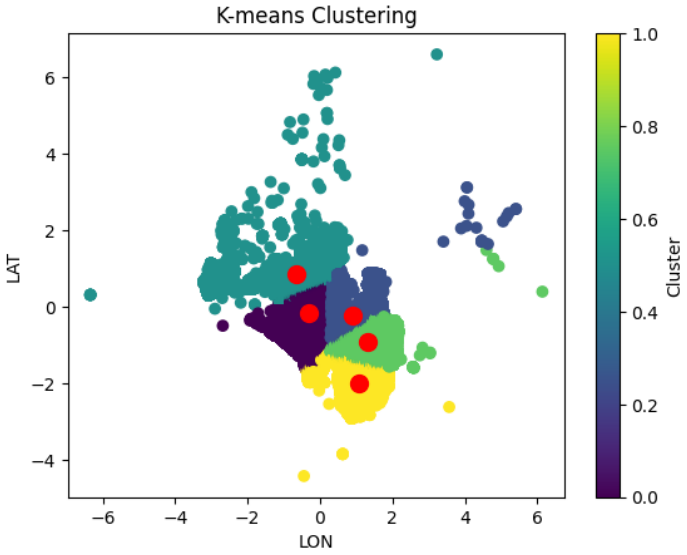

Fig. 10 Visualizing Kmeans clustering (k = 5) with the Cluster Centres

KModes clustering

KModes clustering, is a technique used to group similar data points based on categorical variables, for example crime types, weapon types, and premises descriptions. Unlike other clustering methods like KMeans that work well with numerical data, KModes is designed specifically for categorical data. After preparing and filtering your crime dataset to focus on the most significant categories, the KModes clustering algorithm is used to find clusters among crimes. This is accomplished by choosing 10 clusters to categorize the data into, based on similarities in the descriptions of crimes, weapons used, and premises involved. The algorithm iteratively assigns each crime record to the cluster whose characteristics (centroids) are most similar to that record, with the goal of minimizing differences within clusters. The result is a model that groups the crime records into 10 distinct clusters, providing insights into common patterns within the crime data based on the selected features.

IV. RESULTS AND DISCUSSION

A. *KMeans clustering*

Geographical Clustering

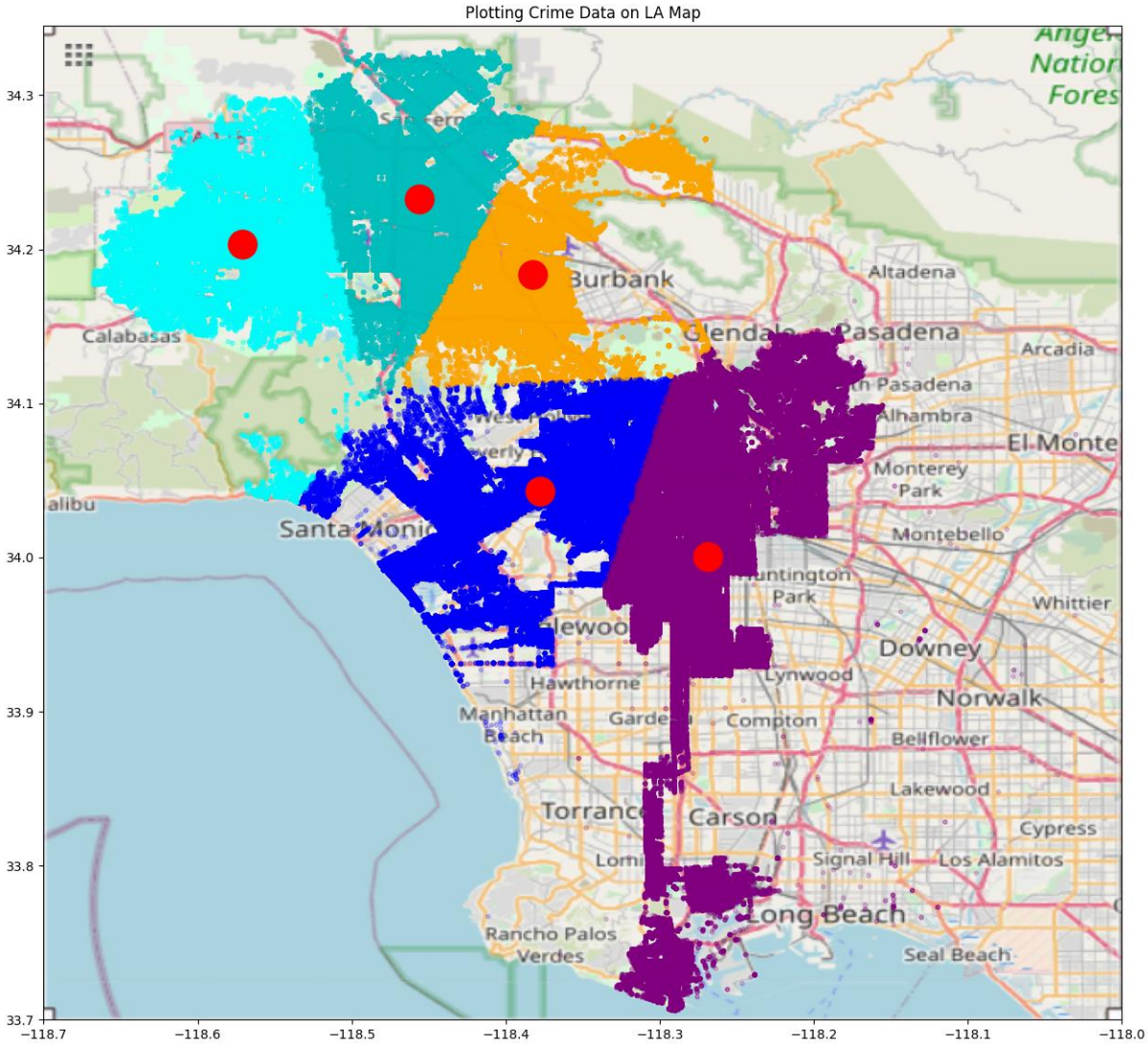

Fig. 11 Visualizing Geographical clustering (k = 5) using Kmeans with the Cluster Centres (Potential locations for Police stations)

Geographical clustering visualizes crime data in Los Angeles using k-means clustering, depicting each cluster with a unique colour and its centroid with a red dot. These centroids suggest optimal locations for police stations, aiming to minimize the Euclidean distances between the incidents and the stations, which could potentially enhance law enforcement efficiency. This clustering not only aids in resource allocation but also

reflects historical crime patterns, indicating that areas with high past crime rates might continue to be hotspots, thereby guiding preventive measures and strategic planning.

Clustering various districts criminality

To classify districts in Los Angeles into zones of varying criminality—from safe to moderately criminalized areas to places with significant criminal activity—a methodical approach was taken in our investigation. First, a data frame was created to include the most common categories of crimes. A pivot table was then created, combining the incidences of different crimes in every district to provide a thorough picture of the criminal environment at the district level. The generated data set was carefully subjected to the k-means clustering technique after the aggregation. In order to divide the districts into clusters, each cluster reflecting a different type of criminal activity, this step was essential. Districts with comparable crime profiles were grouped together into a single cluster, although the goal was to keep the types and rates of crimes within each cluster as different from one another as feasible. This grouping made it easier to spot trends in criminal activity around Los Angeles, which helped to define zones according to how safe and how likely they were to commit crimes.

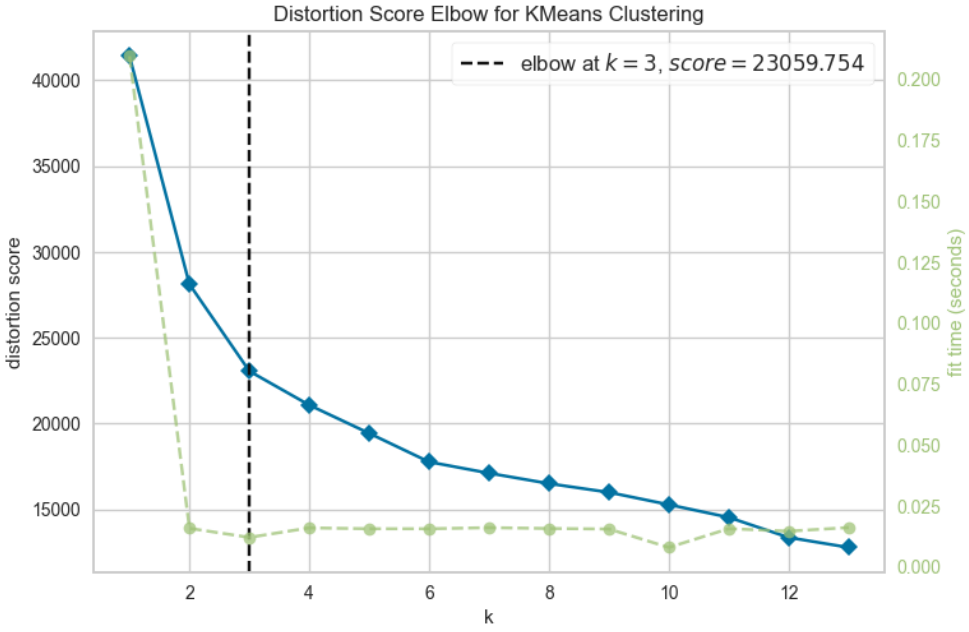

Fig. 12 Computing the optimal k value elbow score

```
1  kmeans = KMeans(n_clusters=3, random_state=0, n_init=100).fit(scaled_distr_crimes)
2  print(kmeans.cluster_centers_)
```

✓ 0.6s

```
[[ 0.09226446  0.07092313  0.14133334  0.18051322  0.06674877  0.16329954
  0.3858777   0.27604864  0.34980315  0.30135842  0.2995502   0.23835234
  0.23834879  0.26074983  0.32792119  0.01591345  0.10205252  0.14950052
  0.05170763  0.02166876  0.3849472   0.4243269   0.44695913  0.10579279
 -0.01788885  0.08642549  0.18236621  0.34246983  0.27095056  0.33237724
  0.25826602  0.33983463]
[-0.61794762 -0.58869423 -0.67693695 -0.61881498 -0.24423276 -0.68368179
 -0.68007925 -0.57927699 -0.72530387 -0.65971637 -0.61999668 -0.63062375
 -0.76178242 -0.62703054 -0.66013093 -0.04134805 -0.50886047 -0.63320274
 -0.57934101 -0.19243036 -0.6591668  -0.85300899 -0.77096181 -0.50552551
 -0.47327094 -0.37506782 -0.49956674 -0.84065776 -0.78465741 -0.75524232
 -0.6017231  -0.73183728]
[ 2.13086897  2.10907594  2.14670081  1.72330425  0.7010972   2.07230848
  1.02036789  1.114353    1.3758506   1.32976837  1.17359703  1.50276139
  2.04629154  1.3835194   1.20772789  0.09720157  1.63321497  1.92741723
  2.1598435   0.69646303  0.93804331  1.55783997  1.11239582  1.60196887
  2.04455187  1.15159406  1.22051354  1.88803675  1.98918945  1.58110265
  1.29021917  1.4493689 ]]
```

Fig. 13 Cluster centers

Key insights drawn from the cluster centers:

First Centroid (Safer Districts):

- The values for all columns are below 0.
- This indicates that the first centroid corresponds to districts with fewer reported crime types.
- Districts associated with this centroid are considered to be the safest, characterized by the lowest frequencies of criminal activity.

Second Centroid (More Dangerous Districts)

- Most column values are above 1.
- This suggests that the second centroid represents districts with a higher incidence of various crime types.
- Districts falling into this category are deemed more dangerous, signifying areas with the highest crime reports and diverse criminal activity.

3. **Third Centroid (Moderately Safe Districts):**

- The values for all columns hover slightly above or sometimes below 0.
- This centroid describes districts that exhibit an average level of crime types reported.
- Such districts are considered moderately safe, indicating a median level of criminal occurrences compared to the extremes identified by the first and second centroids.

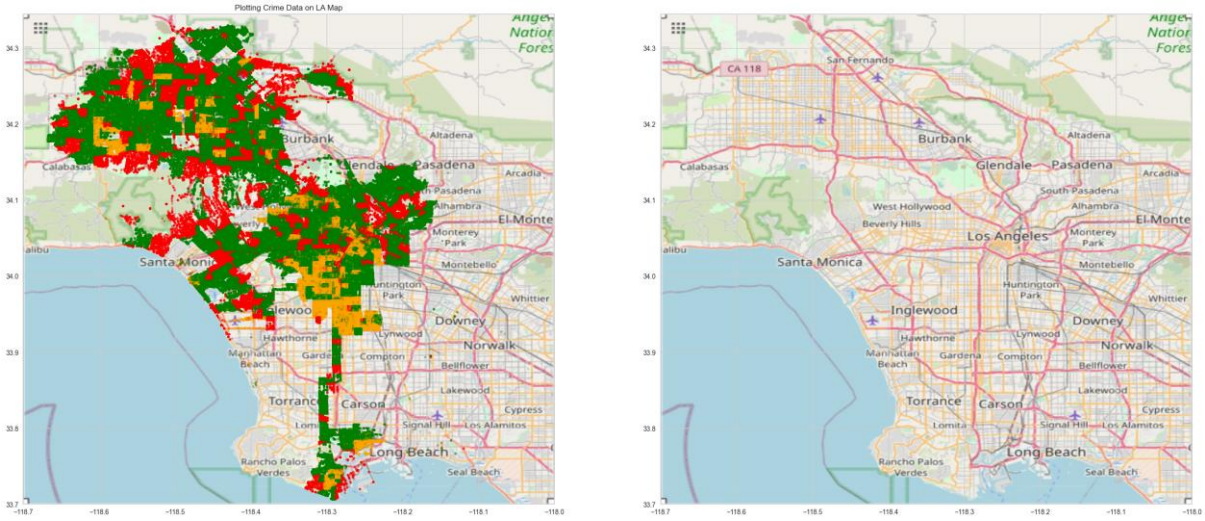

Fig. 14 Plotting Safe, Moderately Safe and Dangerous districts on the LA map

Key insights from the map:

Three Levels of Crime Incidents:

- The map visualizes crime data across different districts in Los Angeles from 2010 to 2019. The districts have been clustered into three categories based on the sum of crime incidents, which are represented by three different colors - green, orange, and red.

Safety Classification:

- **Green Points (Safer Zones)**: The green points indicate districts with a lower than average number of crimes. This conclusion is drawn from the fact that these districts are associated with the cluster whose centroid has negative values after normalization. This suggests these areas are relatively safe. Areas such as Bel Air, Beverly Hills, West Hollywood, Brentwood, Mulholland Dr/Sepulveda Blvd, and Palisades Dr/Ave De Santa Ynez fall into this category.

- **Orange Points (Moderate Safety Zones)**: The orange points denote districts with an average number of crime incidents. These areas are neither particularly safe nor inherently dangerous. This interpretation comes from the clustering that places these districts in the centroid range around zero. Harbor Blvd, Forest Lawn Dr, Dolanco Junction, and Tampa Ave are examples of such districts.

- **Red Points (Less Safe Zones)**: The red points are found in districts with a higher-than-average number of crimes, indicating these areas are less safe. These districts correlate with the cluster that has high positive centroid values. Downtown, South Park, Central City, and Fashion District are examples of areas that fall under this category.

Unmarked Regions:

- Areas without any colour coding could indicate either very low crime rates (to the point of being non-existent) or a lack of reported data. For instance, the Marvin Braude Mulholland Gateway Park, which is unmarked and known to be less residential and more recreational, might naturally have lower crime rates. In contrast, the absence of data in a residential area like Torrance could imply that crime data is missing rather than an actual absence of crime.

B. *KModes clustering*

```
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 1, iteration: 1/100, moves: 72160, cost: 569777.0
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 2, iteration: 1/100, moves: 78078, cost: 552661.0
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 3, iteration: 1/100, moves: 169467, cost: 549651.0
Run 3, iteration: 2/100, moves: 121650, cost: 535611.0
Run 3, iteration: 3/100, moves: 165, cost: 535611.0
Best run was number 3
[['BATTERY - SIMPLE ASSAULT'
  'STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)'
  'SINGLE FAMILY DWELLING']
 ['INTIMATE PARTNER - SIMPLE ASSAULT'
  'STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)'
  'MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)']
 ['INTIMATE PARTNER - SIMPLE ASSAULT'
  'STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)'
  'SINGLE FAMILY DWELLING']
 ['ROBBERY' 'STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)' 'SIDEWALK']
 ['CRIMINAL THREATS - NO WEAPON DISPLAYED' 'VERBAL THREAT'
  'SINGLE FAMILY DWELLING']
 ['ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT'
  'STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)' 'STREET']
 ['BATTERY - SIMPLE ASSAULT'
  'STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)' 'SIDEWALK']
 ['BATTERY - SIMPLE ASSAULT' 'UNKNOWN WEAPON/OTHER WEAPON'
  'MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)']
 ['VANDALISM - MISDEAMEANOR ($399 OR UNDER)'
  'STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)' 'OTHER BUSINESS']
 ['ROBBERY' 'HAND GUN' 'SIDEWALK']]
```

Fig. 15 Results of KModes clustering

Key Insights from the output of KModes clustering:

- **Most Common Crime Types:** The top crime types based on the clustering centroids include "Battery - Simple Assault", "Intimate Partner - Simple Assault", "Robbery", "Criminal Threats - No Weapon Displayed", "Assault with Deadly Weapon, Aggravated Assault", and "Vandalism - Misdemeanor".
- **Prevalent Weapon Types:** "Strong-Arm (hands, fist, feet or bodily force)" is the most common weapon description across several crime types, indicating that physical force without the use of weapons is frequently reported. "Hand Gun" also appears specifically in robbery incidents occurring on the sidewalk, highlighting a specific concern for armed robbery in public pedestrian areas.

- **Common Premises for Crimes:** Crimes frequently occur in "Single Family Dwelling" and "Multi-Unit Dwelling (Apartment, Duplex, etc)", as well as on the "Sidewalk" and "Street". This indicates that both residential and public spaces are primary locations for these types of reported crimes.
- **Clustering Iterations and Cost:** The k-modes algorithm went through several iterations with different initializations, settling on the third run as the best one due to the lowest cost, which is a measure of dissimilarity within the clusters. The cost decreased significantly from the first to the third run, suggesting a better clustering was achieved with each initialization.
- **Significance of Cluster Centroids:** Each centroid represents the mode of the attributes within a cluster. For instance, the most common profile in one cluster is a "Battery - Simple Assault" with "Strong-Arm" as the weapon description occurring at a "Single Family Dwelling". From this clustering, law enforcement could infer that most common assaults do not involve a weapon and occur in residential settings.

The presence of a "Hand Gun" in robberies specifically on sidewalks suggests a focus area for patrols. The data-driven approach in deploying resources and creating prevention strategies can be based on these insights, targeting the most common types of crime in the most frequent locations with the appropriate response plans.

V. CONCLUSIONS

In conclusion, this project effectively used k-means clustering to assess and classify Los Angeles districts according to reported crime types, resulting in a detailed knowledge of the city's fluctuating safety and criminal activity levels. The Machine Learning model was able to divide districts into three categories: safer districts, somewhat safe districts, and more risky districts. This was accomplished by combining crime statistics into a comprehensive dataset and then using k-means clustering. The features of these groups were clearly identified through the interpretation of centroid centres, which offered valuable information about the relative safety of various locations.

The key component of our analysis was the deliberate use of the K-Means clustering algorithm, which divided LA districts into three safety tiers based on the frequency of crimes. This was illustrated using a geographical map with colour codes: green represented low-crime districts, orange represented moderate-crime zones, and red represented high-crime zones. Uncoloured areas indicated either little crime or no data. The study also highlights how machine learning methods, such k-means clustering, may be used to turn unprocessed data into insightful knowledge. Policymakers, law enforcement organizations, and community leaders can improve urban safety and citizens' quality of life by using such sophisticated analytical techniques to help them make well-informed decisions.

In addition, it is worth noting that additional machine learning algorithms such as decision trees could have been implemented for categorising the different crime types per region, alternatively predicting the further probability future crime occurrences. Integration of multiple data sources – social media applications like Instagram, Twitter, Reddit and news websites could have helped in obtaining a larger dataset.

This project demonstrates the revolutionary potential of big data and machine learning in addressing complex societal issues, while also adding to our understanding of the distribution of crime and safety levels throughout Los Angeles. It presents a picture of a safer, more secure urban future where decision-making is deliberate, strategic, and proactive by opening the door for creative solutions that make use of technology and data analytics.

VI. FUTUREWORK

As mentioned previously we can integrate real-time and streaming data sources, such as social media and environmental elements, future advancements could greatly increase the effect of this initiative, building on the foundation it created. Investigating more complex machine learning models than k-means clustering, like decision trees and neural networks, may offer more insightful predictions and reveal the root reasons of criminal activity. Enhancing the dataset and analysis will involve including comparative urban studies, involving communities through crowdsourcing data, and addressing ethical and privacy concerns. In addition to deepening our knowledge of the dynamics of urban safety, these developments hold the possibility of providing practical solutions for resource allocation, proactive crime prevention, and the development of safer communities—all while negotiating the complicated issues of privacy and moral data usage.

REFERENCES

[1] H. Hashim and A. T. Abdulameer, “Crime Prediction Using Big Data Analysis,” Sep. 2021.

[2] F. Dakalbab, M. Abu Talib, O. Abu Waraga, A. Bou Nassif, S. Abbas, and Q. Nasir, "Artificial intelligence & crime prediction: A systematic literature review," Social Sciences & Humanities Open, vol. 6, no. 1, p. 100342, 2022.

[3] R. Kumar and B. Nagpal, "Analysis and prediction of crime patterns using big data," International Journal of Information Technology, vol. 11, no. 4, Dec. 2018. DOI: 10.1007/s41870-018-0260-7.

[4] A. Ravi and M. Tech, “Issue 7 www.jetir.org (ISSN-2349-5162),” *JETIR2107201 Journal of Emerging Technologies and Innovative Research*, vol. 8, 2021, Available: https://www.jetir.org/papers/JETIR2107201.pdf

[5] G. Journals, “Big Data Prediction on Crime Detection,” *www.academia.edu*, Accessed: Mar. 03, 2024. [Online]. Available: https://www.academia.edu/32899528/Big_Data_Prediction_on_Crime_Detection

[6] G. A. AL-Rummana, Abdulrazzaq Alahdal, and G. N. Shinde, “The Role of Big Data Analysis in Increasing the Crime Prediction and Prevention Rates,” Jan. 2021, doi: https://doi.org/10.1002/9781119711629.ch10.

[7] M. Menkudle and R. Potpelwar, “Big Data Analytics and Crime Patterns Detection and Prevention,” vol. 8, no. 9, pp. 2320–2882, 2020, Available: https://ijcrt.org/papers/IJCRT2009316.pdf