CRIME DATA ANALYSIS IN LOS ANGELES USING MACHINE LEARNING TECHNIQUES

DYNAMIS HUB

**Acknowledgements**

First and foremost, I would like to express my sincere gratitude to my supervisor, Mr. Gkratsia Tantilian, for his invaluable guidance, support, and encouragement throughout the duration of this thesis. His expertise and insights were crucial to the successful completion of this research, and I am deeply appreciative of the time and effort he dedicated to assisting me.

I would also like to extend my heartfelt thanks to the faculty and staff of Mediterranean College, as well as the University of Derby, for providing an exceptional academic environment and for their continuous support during my studies. The knowledge and skills I acquired during my time at these institutions have been instrumental in shaping this research.

Additionally, I would like to acknowledge the assistance and cooperation of my fellow students and colleagues, whose ideas and feedback have contributed to the development of this thesis.

Lastly, I would like to thank my family and friends for their unwavering support and understanding during this challenging journey. Their encouragement and belief in me provided the motivation I needed to persevere and succeed.

**Abstract.**

This dissertation addresses the challenge of understanding and predicting crime patterns in urban areas, with a specific focus on the city of Los Angeles. The primary objective was to develop and evaluate two advanced analytical tools: the Crime Dashboard and the Crime Probability Dashboard. These tools were designed to provide law enforcement agencies with data-driven insights to enhance resource allocation, improve strategic planning, and ultimately reduce crime rates.

The research involved the collection and analysis of a comprehensive dataset obtained from the Los Angeles Police Department (LAPD), containing detailed records of crimes committed in the city over several years. The data was preprocessed and analyzed using various machine learning techniques, including Principal Component Analysis (PCA) and K-means clustering. The Crime Dashboard was developed to visualize crime trends over time and across different areas, while the Crime Probability Dashboard focused on estimating the likelihood of crimes based on factors such as time, location, and victim demographics.

Key findings from this research include the identification of specific crime hotspots and the successful forecasting of future crime trends in targeted areas. The dashboards provided actionable insights, such as the most vulnerable days and times for certain types of crimes, which can be used to guide law enforcement strategies.

In conclusion, this dissertation demonstrates the effectiveness of combining data analytics with predictive modeling to enhance crime prevention efforts. However, limitations such as data accuracy and generalizability were noted, and future research should focus on integrating additional variables and refining the models for broader application. The tools developed in this study offer significant potential for improving urban safety and can be adapted for use in other cities facing similar challenges.

**Keywords:** Crime Analysis; Machine Learning; Predictive Modeling; Crime Dashboard; Crime Probability; Los Angeles; Data Visualization; Principal Component Analysis (PCA); K-means Clustering; Forecasting; Urban Safety; Law Enforcement Strategies;.

**Table of contents**

[List of abbreviations 5](#_Toc176123758)

[1 INTRODUCTION 6](#_Toc176123759)

[2 Ethical and Social Implications of Using Machine Learning in Crime Analysis 7](#_Toc176123760)

[3 LITERATURE REVIEW 8](#_Toc176123761)

[3.1 Introduction 8](#_Toc176123762)

[3.2 Presentation of Key Works and Researchers 9](#_Toc176123763)

[3.3 Machine Learning Techniques in Crime Data Analysis 9](#_Toc176123764)

[3.4 Applications and Challenges 9](#_Toc176123765)

[3.5 New Methods of Analysis 10](#_Toc176123766)

[3.6 Summary and Gaps in the Literature 10](#_Toc176123767)

[4 METHODOLOGY 10](#_Toc176123768)

[4.1 Development Tools 10](#_Toc176123769)

[4.2 Project Management and Monitoring Tools 10](#_Toc176123770)

[4.3 Data Analysis Tools and Libraries 11](#_Toc176123771)

[4.4 Explanation of Algorithms 11](#_Toc176123772)

[4.5 Data Collection and Description 12](#_Toc176123773)

[4.6 Dataset Analysis 12](#_Toc176123774)

[4.7 Data Cleaning 13](#_Toc176123775)

[4.8 Data Preprocessing 14](#_Toc176123776)

[4.9 Exploratory Data Analysis (EDA) 15](#_Toc176123777)

[4.10 Machine Learning Techniques 16](#_Toc176123778)

[5 Results 17](#_Toc176123779)

[5.1 Eda Results 17](#_Toc176123780)

[5.2 Crime Dashboard 19](#_Toc176123781)

[5.3 Probability Dashboard 20](#_Toc176123782)

[6 Discussion 21](#_Toc176123783)

[6.1 Crime Dashboard 21](#_Toc176123784)

[6.2 Probability Dashboard 23](#_Toc176123785)

[6.3 Compare Dashboards 23](#_Toc176123786)

[7 Conclusion 24](#_Toc176123787)

[7.1 Summary of Research Process 24](#_Toc176123788)

[7.2 Global Implications and Future Directions 25](#_Toc176123789)

[References 27](#_Toc176123790)

[Appendix 28](#_Toc176123791)

# List of abbreviations

* PCA: Principal Component Analysis
* ML: Machine Learning
* K-means: K-means Clustering
* LAPD: Los Angeles Police Department
* EDA: Exploratory Data Analysis
* Prophet: A forecasting model developed by Facebook for time series data
* CSV: Comma-Separated Values (a file format for data)
* DR\_NO: Division of Records Number (unique identifier for crime reports)
* API: Application Programming Interface
* RNN: Recurrent Neural Network
* GUI: Graphical User Interface
* k-NN: k-Nearest Neighbors (a machine learning algorithm)
* ROC: Receiver Operating Characteristic (a graphical plot to evaluate classifier performance)
* AUC: Area Under the Curve (used in ROC analysis)
* MSE: Mean Squared Error (a metric to measure the accuracy of a predictive model)
* RMSE: Root Mean Squared Error (the square root of MSE, used for model evaluation)

1. INTRODUCTION

Crime is one of the most complex social challenges, with serious implications for social stability and quality of life. Globally, governments and law enforcement agencies are constantly seeking new strategies to effectively combat crime. With the advancement of technology and the widespread availability of data, data analysis techniques and machine learning have emerged as key tools for understanding crime patterns and developing predictive models.

**Los Angeles**, as one of the largest and most populous cities in the United States, faces significant challenges in combating crime. The city, with its complexity arising from the diversity of social and cultural groups it hosts, presents a multifaceted crime profile. The history of crime in Los Angeles has been shaped by a combination of social, economic, and political factors, with various waves of violence and criminal acts characterizing it. In the early 20th century, the city faced minor crimes, often linked to the increasing number of immigrants and rapid urbanization. In the post-war years, Los Angeles became the epicenter of gang violence, with the 1980s dominated by widespread drug problems and the rise of organized crime. Today, crime in Los Angeles remains diverse, ranging from minor offenses to more serious criminal acts, making the need for analytical and predictive tools more urgent than ever. Efforts to address these phenomena have led to the development and implementation of more advanced policing strategies. At the same time, the recognition of the complexity of crime has led to the adoption of modern data analysis techniques, as traditional methods of policing and analysis were insufficient to address the challenges of the modern era. Recent statistics show that in 2023, there was a 5% increase in violent crimes compared to the previous year, while property crimes increased by 7%, highlighting the need for more effective prevention and suppression strategies (Cung, 2013).

Advancements in computer technology and the increase in available computing power have enabled the development of advanced data analysis techniques, which are now widely applied in crime analysis. Technologies such as Machine Learning, Artificial Intelligence, and Big Data have transformed the way authorities can understand and predict crime patterns. In Los Angeles, the use of these technologies has begun to bear fruit, with authorities now able to analyze large crime datasets, predict potential future crime hotspots, and develop data-driven strategies. For example, the use of machine learning algorithms allows for the analysis of crime time series, while the analysis of spatiotemporal data helps identify high-risk areas. These modern technologies do not replace traditional methods but complement them, offering new tools and capabilities that allow for a more comprehensive and dynamic approach to tackling crime. Additionally, the importance of data-driven decision-making has become increasingly evident in many areas, including public safety and crime prevention. The ability to analyze and interpret crime data can provide significant insights that help law enforcement agencies allocate resources more effectively, predict future crime trends, and ultimately reduce crime rates. This thesis focuses on the application of advanced data analysis and machine learning techniques to crime data, aiming to develop tools that enhance the understanding and prediction of crime patterns in urban areas (Mohler , et al., 2008).

Research Focus This study specifically focuses on the need for effective crime analysis tools through the development of two integrated dashboards: the Crime Dashboard and the Probability Dashboard. The Crime Dashboard allows for in-depth investigation of crime data, providing visual insights into crime distributions across different dimensions such as time, location, and victim characteristics. The Probability Dashboard, on the other hand, focuses on estimating the likelihood of crimes occurring based on various factors such as the time of day, day of the week, and area. The research is centered on the city of Los Angeles, utilizing a dataset that includes detailed crime reports spanning several years.

Discussion of Existing Research Current research in crime analysis has extensively utilized statistical methods and machine learning models to identify high-crime areas, predict future crimes, and analyze the socioeconomic factors that contribute to criminality. However, many of the existing tools lack the flexibility to provide dynamic, user-defined analyses or fail to incorporate predictive models with probability estimates of crime risk. This thesis aims to fill this gap by developing tools that not only analyze historical crime data but also predict future trends and estimate the likelihood of crimes in a user-friendly manner. These tools are particularly important in the broader context of urban security management, where timely and accurate crime analysis is critical for proactive policing strategies.

Research Questions and Objectives The main research question addressed in this thesis is: How can advanced data analysis techniques and machine learning be leveraged to improve the understanding and prediction of crime patterns in urban areas? To answer this question, the study sets the following objectives:

* To develop a Crime Dashboard that provides a comprehensive analysis of crime data, allowing users to explore trends and patterns across various dimensions.
* To create a Probability Dashboard that estimates the likelihood of crimes occurring based on various factors, offering valuable insights for law enforcement agencies.
* To evaluate the effectiveness of these tools in providing accurate and useful crime analysis and prediction, compared to existing methods.

For the purposes of this research, a dataset including crime information in Los Angeles, as recorded by the city’s police department, was used. The dataset contains information such as the date and time of the crime, the location, the type of crime, and the age, gender, and ethnicity of the victims. This data was used both for analyzing existing trends and for developing the predictive models implemented in the two dashboards.

Key Definitions To achieve the goals of this research, various data analysis and machine learning methods were employed. Below are some key definitions and techniques:

* Principal Component Analysis (PCA): A statistical technique used to reduce the dimensions of a dataset while retaining the information that is most important for describing the data’s structure.
* K-means Clustering: A popular clustering method that divides the data into k clusters, such that each observation belongs to the cluster with the nearest center.
* Prophet Model: A forecasting tool developed by Facebook for analyzing time series data with strong seasonal trends and variations. (Facebook, 2017)
* Machine Learning: A branch of artificial intelligence that focuses on developing algorithms that allow computers to learn from data and make decisions or predictions without being explicitly programmed to do so.

This combined framework of objectives and content clearly demonstrates the necessity and importance of this research in understanding and predicting crime in the Los Angeles area.

1. Ethical and Social Implications of Using Machine Learning in Crime Analysis

The use of machine learning (ML) in crime data analysis has revolutionized the way law enforcement agencies manage crime prevention and response. While the benefits are evident, there are also significant ethical and social implications that must be considered. This section will examine the challenges related to privacy, potential biases in the data, and the consequences for the communities affected by these technologies.

One of the key ethical issues arising from the use of machine learning in crime analysis is the protection of individuals' privacy. The collection and analysis of large amounts of data, such as geolocation data, socioeconomic data, and personal information, raise concerns about the violation of citizens' privacy. Specifically, the use of data from social media and other public sources can lead to privacy breaches, as this information may be used without the individuals' consent.

Law enforcement agencies must ensure that the data collected and analyzed adhere to strict privacy protection standards. This includes anonymizing the data, limiting access to sensitive information, and ensuring that the data is used solely for legitimate purposes. Moreover, transparency in the data collection and analysis processes is critical to maintaining public trust.

Another important issue is the potential for biases in the data used to train machine learning algorithms. Crime data may contain inherent biases related to socioeconomic status, gender, ethnicity, and other factors. When such data is used to train machine learning models, there is a risk of perpetuating these biases, leading to unequal or unfair decisions.For example, data showing higher crime rates in certain areas may be due to over-policing rather than a genuine increase in crime. If this data is used to train a model, the result may be the continued over-policing of these areas, creating a vicious cycle of bias and inequality.

Avoiding bias requires careful selection and pre-processing of the data, as well as the use of techniques that can detect and mitigate biases. Researchers and analysts must remain constantly aware of these challenges and work towards developing fair and unbiased models.

The use of machine learning for crime prediction and the allocation of police resources can have significant social impacts on communities. Areas identified as "high-risk" may come under constant surveillance, leading to increased police presence and the creation of an environment of tension and insecurity for residents. This can result in social isolation and stigmatization of these areas, exacerbating social inequalities.

Additionally, citizens may lose trust in the authorities if they feel that policing is based on opaque or unfair methods. Public trust is vital for effective policing and the promotion of public safety. Authorities must work with communities to ensure that the technologies used in crime analysis do not contribute to worsening social problems but instead help create a safe and just environment.

The rapid development of technology necessitates the need for stricter regulations and legislation to ensure the proper use of machine learning techniques in crime analysis. These regulations should focus on privacy protection, bias prevention, and ensuring transparency in processes. Furthermore, legislators must consider the social implications of these technologies and work with experts to develop regulations that ensure their fair application.

These laws should incorporate mechanisms for the control and evaluation of the algorithms used and provide procedures for addressing cases of misuse or incorrect application of the technology. It is also important to ensure that citizens have access to information about how their data is being used and what options they have to protect their privacy.

Conclusion The use of machine learning in crime analysis offers significant potential for improving public safety but is accompanied by serious ethical and social challenges. It is essential to develop and implement strategies that ensure the fair and responsible use of these technologies, protecting citizens' rights and promoting social justice. Researchers, analysts, and law enforcement agencies must work closely together to ensure that machine learning techniques are used in a way that benefits all communities, regardless of socioeconomic status or other factors.

1. LITERATURE REVIEW
   1. Introduction

Crime data analysis has garnered increasing interest in recent years as cities continuously face challenges regarding the safety of their citizens. The application of machine learning (ML) techniques to crime data offers new possibilities for predicting criminal behaviors, improving resource allocation, and understanding the factors that influence criminality. This literature review examines the existing studies in this field, focusing on the use of various ML algorithms, their applications, and the challenges that arise in analyzing crime data in Los Angeles.

* 1. Presentation of Key Works and Researchers

The initial efforts in crime data analysis primarily relied on statistical methods and spatiotemporal models, as presented by Wang et al. (Wang, et al., 2013) in their study on identifying crime patterns using supervised learning techniques. This study provides a solid foundation for understanding early analytical efforts, but some of the methods used may be insufficient to address the complexity of modern crime data. With the development of machine learning techniques, researchers began to explore how ML models can be applied to predict and classify crimes. Recent studies, such as that by Mohler et al. (Mohler, et al., 2015), have shown that the use of spatiotemporal analyses and predictive models can improve the prediction of criminal events in urban areas. However, despite their positive impact, these studies often overlook the importance of local particularities and social factors, which can limit the accuracy of the models in different environments.

* 1. Machine Learning Techniques in Crime Data Analysis

The technique of supervised learning has been widely used to predict the type of crime or the location of future criminal activities. For example, Logistic Regression has proven effective in classifying crimes, while Decision Tree models have been used to predict recurring criminal actions Wang et al. (Wang, et al., 2013). While these methods are effective, they may be limited in cases where the data is noisy or incomplete, and their application in different cities or conditions may require further adjustments. On the other hand, unsupervised learning techniques, such as clustering, have been used to detect patterns of criminal behavior in large datasets. K-Means clustering, for example, has been applied to group areas with similar criminal activity in Los Angeles (Gerber, 2014). While clustering offers a new perspective on understanding crime data, the lack of labels in the data can lead to uninterpretable or oversimplified results. Additionally, recent deep learning techniques have started to find applications in crime data analysis, though they face challenges related to the interpretability of their results (Jean , et al., 2016). Deep learning can offer superior results compared to traditional methods, but its complexity makes it difficult to interpret predictions, which is a significant drawback in applications where transparency is critical.

* 1. Applications and Challenges

The applications of machine learning in improving policing are numerous. Studies have shown that ML models can help predict the deployment of police forces with greater accuracy, thereby reducing response times to criminal incidents (Mohler, et al., 2015). While this approach has proven effective in many cases, there are still significant challenges that need to be addressed, such as the impact of data bias and the quality of the data used. Specifically, issues such as those described by Richardson, Schultz, and Crawford (Richardson , et al., 2019) highlight the dangers of using data that contain errors or reflect biases, which can lead to injustices or incorrect predictions. In the present study, a series of ML techniques, such as PCA, K-Means clustering, and Logistic Regression, were applied to crime data in Los Angeles. The results showed that the application of PCA successfully reduced the dimensionality of the data, allowing for better visualization and understanding of underlying patterns. K-Means clustering managed to identify six crime clusters based on their main characteristics, though some issues of oversimplification were observed in cases with high data heterogeneity. The use of Logistic Regression to predict characteristics such as the gender of victims showed that the models can be highly accurate, but accuracy depends heavily on the quality and distribution of the data. Comparing these results with the literature, it is observed that the findings align with the general conclusions of other studies (Wang, et al., 2013) (Mohler, et al., 2015). However, the results indicate that the application of traditional ML techniques can be further improved with the use of more modern methods, such as ensemble models or deep learning, especially in cases where the data is highly heterogeneous or noisy. Furthermore, it is noted that prediction accuracy can decrease in conditions with high levels of missing data, highlighting the need for further research in improving data quality.

* 1. New Methods of Analysis

Modern approaches to crime data analysis include the application of techniques that go beyond traditional statistical models, incorporating the power of Big Data and Artificial Intelligence. One of the new methods that have begun to be applied is the use of time series forecasting models with strong seasonality, such as the Prophet model developed by Facebook.

Prophet is particularly effective in forecasting future trends in data with strong seasonal patterns, such as crimes associated with specific social or economic activities. This method provides authorities with the ability to adjust their strategies according to the predictions, enabling more effective resource allocation and real-time decision-making.

At the same time, the use of Ensemble Learning methods, such as Random Forests and Gradient Boosting Machines, has begun to gain ground in crime analysis, due to their ability to combine the advantages of multiple models to improve prediction accuracy. These techniques address overfitting problems and offer more reliable results, especially in cases where the data is heterogeneous and complex.

* 1. Summary and Gaps in the Literature

This literature review shows that machine learning has the potential to redefine crime data analysis, offering new solutions to a multidimensional problem. However, the literature remains limited in addressing issues of bias and interpretability in ML models. Additionally, most studies focus on prediction and analysis in large urban areas, leaving a gap in research concerning the application of these techniques in smaller communities or at the regional level. This study attempts to fill these gaps by focusing its analysis on crime data in Los Angeles and offering new perspectives on predictive policing.

1. METHODOLOGY

Crime data analysis is a complex process that requires the careful selection and use of appropriate tools for managing, analyzing, and interpreting large datasets. In this study, various tools and libraries were used for the development, execution, and monitoring of the project, aiming to ensure the accuracy and effectiveness of the results.

* 1. Development Tools

**Visual Studio Code** (VS Code) (Microsoft, 2024) was chosen as the primary development environment due to its flexibility and broad support for various programming languages, such as Python (Van Rossum & Drake , 2009), which was used for data analysis. VS Code offers a comprehensive programming experience with features such as support for extensions for code analysis, built-in version control management, and the ability to run code directly from within the environment. Additionally, the ease of setting up the workspace and the ability to integrate with GitHub make VS Code an ideal tool for large-scale data analysis projects.

* 1. Project Management and Monitoring Tools

**GitHub** was used for code version management and change tracking. The use of GitHub allows for the documentation of every change made to the code, providing a complete history of the edits. This is particularly important in large projects where collaboration and the need to reference earlier stages of code development are essential. Additionally, GitHub facilitated communication and collaboration with the project supervisor, who could monitor changes in real-time and provide feedback.

**Trello** was used as a project management tool for organizing tasks and setting priorities. This tool allows for the creation and management of task cards, which can be categorized and organized into lists representing different stages of the project. The use of Trello helped ensure the timely completion of the various project stages, offering a clear view of progress and allowing for easy deadline tracking.

* 1. Data Analysis Tools and Libraries

The **Pandas and NumPy** (Team, 2020)(Harry, et al., 2020) libraries were chosen for data management and processing. Pandas offers powerful capabilities for analyzing and handling data in table format (dataframes), making it ideal for working with large datasets such as the crime data used. NumPy was used to support mathematical and statistical computations, as well as for efficient storage and processing of large multidimensional arrays.

The **Scikit-learn** (Pedregosa , et al., 2011) library was chosen for implementing machine learning algorithms. It is one of the most popular Python libraries for developing and evaluating machine learning models, offering a wide range of tools for data preprocessing, model construction, and performance evaluation. Scikit-learn was used to implement algorithms such as K-Means clustering and Logistic Regression, which formed the core of the analysis.

For data and results visualization, the **Matplotlib** (Hunter , 2007) and **Seaborn** (Waskom, M. L, 2021) libraries were used. These libraries provide extensive capabilities for creating graphs and charts, allowing for the presentation of data in a comprehensible and easily readable manner. The use of these tools enhanced the ability to identify trends and patterns in the crime data, offering valuable insights for further analysis.

The **Streamlit** (lnc, 2019)library was used to create interactive applications and dashboards that allow for the easy presentation of analysis results. By using Streamlit, the analysis and findings can be presented in real-time interactively, offering users the ability to explore the data and results through a user-friendly interface. Streamlit was chosen for its simplicity in application creation, as well as its flexibility in integrating different types of visualizations and data.

The **Pickle** library was used to store intermediate data and results in binary format, allowing for faster loading and processing of data in subsequent stages of the analysis. The **Datetime** library was used for handling and converting chronological data. Specifically, it helped manage dates and times, allowing for the proper analysis of temporal patterns in crime data. The **Dash** library was used to develop interactive dashboards that allowed for the dynamic presentation of analysis results. These dashboards offer users the ability to examine analysis results through user-friendly interfaces.

* 1. Explanation of Algorithms

The **Logistic Regression** algorithm is one of the most fundamental algorithms used for binary classification of data. It is based on the logistic function and is particularly effective in cases where the target variable is binary. In this analysis, Logistic Regression was used to predict the likelihood of an event belonging to a specific crime category, with excellent results in classification accuracy.

**Decision Trees**, on the other hand, provide a clear and interpretable method for classification and prediction. The ability of a Decision Tree to split the data into distinct groups based on their characteristics makes it a valuable tool for understanding the relationship between features and the target variable. However, Decision Trees are prone to overfitting, which is why they are often used in combination with Ensemble Learning techniques to improve generalization.

**K-Nearest Neighbors** (KNN) is a simple, yet powerful algorithm used to classify data based on their proximity to other examples. In crime analysis, KNN is used to categorize new crime incidents based on the characteristics of their nearest neighbors. Although KNN is flexible and easy to understand, its performance can be affected by the choice of the number of neighbors (k) and the scale of the data.

The selection of these tools was based on the need for flexibility, effectiveness, and accuracy in large-scale data analysis. Each tool was chosen with consideration of its ability to support the project's requirements and facilitate the completion of the analysis with the highest possible performance.

* 1. Data Collection and Description

The data used in this analysis comes from the Los Angeles Police Department (LAPD) and concerns crime incidents recorded in the city from 2020 onwards. The dataset is updated monthly and is available through the Los Angeles Open Data Portal. This data was initially collected by the LAPD and later published on Kaggle, from where it was retrieved for this analysis. The dataset is available under the Apache 2.0 license. This data comes from original crime reports that are recorded manually, with the possibility of some inaccuracies. Some missing location fields are marked with the coordinates (0°, 0°). Address fields are provided only up to the nearest hundred numbers to maintain privacy. The dataset contains 918,443 records and includes various crime categories, geographical locations, and demographic information of the victims. The information included in the data is crucial for analyzing the spatial and temporal patterns of crime in Los Angeles.

* 1. Dataset Analysis

The dataset consists of 28 columns, each containing specific information about crime incidents. Below is a description of the most important columns:

**DR\_NO (Department Report Number):** The department report number, consisting of a 2-digit year code, the area number, and 5 digits. It is the unique number assigned to each incident by the LAPD, ensuring the identification and tracking of each event.

**Date Rptd (Date Reported):** The date on which the incident was reported to the authorities. It is used for analyzing the time elapsed from the incident to its reporting.

**DATE OCC (Date Occurred):** The actual date on which the incident took place. This column allows for tracking crimes over time and comparing them with the report date.

**TIME OCC (Time Occurred):** The time of the incident in military format (24-hour). Analyzing the time of occurrence helps identify temporal patterns in crime.

**AREA and AREA NAME:** These columns represent the number and name of the geographical area corresponding to the Los Angeles city policing district. The distribution of crime across different areas is critical for understanding spatiotemporal patterns and the effective allocation of police resources.

**Crm Cd (Crime Code)** and **Crm Cd Desc (Crime Code Description):** These columns include the crime code and its description, providing information on the type of crime recorded. This categorization allows for analyzing the frequency of different types of crimes.

**Vict Age (Victim Age):** The age of the victim, important for understanding the demographic characteristics of the victims and analyzing the groups most affected by crime.

**Vict Sex (Victim Sex):** The gender of the victim, used for analyzing victimization patterns based on gender.

**Vict Descent (Victim Descent):** The ethnic descent of the victim. This column is crucial for understanding the socioeconomic background and ethnic groups most affected by crime.

**Premis Cd (Premises Code) and Premis Desc (Premises Description):** These columns represent the code and description of the location where the crime took place. They are used for analyzing different types of locations, such as residences, businesses, and public spaces, where crimes occur.

**Weapon Used Cd (Weapon Used Code) and Weapon Desc (Weapon Description):** These columns provide information on the type of weapon used during the crime. Understanding the types of weapons most commonly used can help in violence prevention and policy planning.

**LAT and LON (Latitude and Longitude):** The geographical coordinates of the location where the incident occurred. This information is vital for visualizing crimes and conducting spatial analysis.

* 1. Data Cleaning

After data collection, the next critical step was the process of cleaning and preprocessing the data. This process was necessary to ensure that the data was accurate, consistent, and suitable for analysis and application in machine learning models.

The data cleaning process began with converting the columns containing dates into a standardized datetime format. The columns (Date Rptd) and (DATE OCC), which contain the report date and the occurrence date of the incident, respectively, were initially in text format. This conversion was necessary to ensure consistency in dates, allowing for further time-based analysis, such as examining crime trends by month or day of the week. Additionally, standardizing the dates enables the use of this information in algorithms that require proper chronological order of data, thereby improving the accuracy of the analysis.

Another important step in cleaning was correcting the geographical coordinates, specifically the values for longitude (LON) and latitude (LAT). These coordinates were cleaned to remove unnecessary decimals and to ensure that the values were within reasonable bounds. This correction was necessary to ensure the accuracy of the geographical data, which is vital for spatial analyses, such as crime mapping. Accurate geographical placement of crimes allows for a better understanding of the spatial patterns of criminality and supports the development of prevention policies targeting specific areas.

The time of the incident, initially recorded in military format (24-hour), was converted into a standard time format. The (TIME OCC) column was adjusted so that the time is displayed in HH format. This change facilitates the analysis of temporal crime patterns and allows for the comparison of incidents based on the time of day. Moreover, this conversion is important for the proper categorization of incidents according to peak hours and periods of increased risk.

The issue of missing data in certain columns was also addressed. Specifically, the columns (Weapon Used Cd, Weapon Desc, Premis Desc, Cross Street, and Mocodes) contained missing data, which was filled with the value 'Unknown.' This approach was chosen to prevent the loss of significant information due to the deletion of rows with empty values. Using the 'Unknown' value allows for the retention of the maximum possible data volume for analysis while maintaining consistency and continuity of information. This choice was critical because deleting rows with missing data could significantly reduce the size of the dataset and lead to incorrect conclusions.

In cases where the primary crime code (Crm Cd 1) was missing, the values were filled with the most frequently occurring value in that column. This approach ensures that each record in the dataset has a defined crime category, thereby improving the consistency and accuracy of crime categorization. Choosing the most frequent value as a substitute is important for maintaining uniformity in the dataset and avoiding the introduction of bias that could arise from random or arbitrary choices.

Additionally, certain columns considered redundant for the analysis, such as (Crm Cd 2, Crm Cd 3, and Crm Cd 4), were removed. These columns did not offer additional information that would substantially contribute to the analysis and categorization of the data, and their removal helps reduce the complexity of the dataset. This choice was necessary to simplify the data, allowing machine learning models to focus on features that have real significance for the analysis.

Finally, duplicate rows in the dataset were identified and removed. The presence of duplicate records could skew the analysis results by giving undue weight to certain incidents while ignoring others. Removing duplicates ensures that each crime incident is analyzed only once, maintaining the accuracy of the results. This process is critical for ensuring the integrity of the data and the reliability of the conclusions drawn from the analysis.

* 1. Data Preprocessing

Following the cleaning process, the data was preprocessed. Initially, two new temporal features were extracted from the (Date Rptd) column: the month (Month) and the day of the week (Day of Week). These features are valuable for analyzing seasonality and daily crime patterns, providing insights into when most crimes occur. Adding these features allows for the analysis of temporal trends and comparison of crime rates across different time periods.

Subsequently, crimes were categorized into broader categories and ranked by severity. The (Crm Cd Desc) column was used to create the (Crime Category and Crime Severity) columns. The (Crime Category) groups crimes into categories such as "Theft," "Assault," and "Burglary," while the (Crime Severity) ranks crimes into three levels of severity: "High," "Medium," and "Low." These categorizations allow for the analysis of crimes based on type and severity, helping to understand different crime patterns. This categorization was necessary to simplify the analysis and highlight the most severe and common forms of criminality.

For the re-categorization of demographic information, the (Victim Age) column was divided into new age groups, such as "Child," "Young Adult," "Adult," and "Senior." Additionally, the (Victim Sex and Victim Descent) columns were re-categorized so that rare categories were combined under the label "Other." These changes facilitate data analysis and allow for better integration of demographic factors into machine learning models. This re-categorization is important to reduce noise in the data and to facilitate analysis based on groups with sufficient representation.

Some columns deemed unnecessary for the analysis, such as (Report Number, Time Occurred, and Location), were removed. The removal of these columns helps reduce the complexity of the dataset, allowing models to focus on the most important features that contribute to the analysis. This process is necessary to minimize computational load and improve the efficiency of the algorithms that will be used subsequently.

Furthermore, outliers in numerical columns were addressed, as they could negatively affect the analysis results. Outliers were identified and removed to ensure that the data used is more representative of general trends. The removal of outliers is critical to avoid skewing results and to improve the performance of the machine learning models that will be applied.

Finally, the categorized columns, such as (Crime Category, Crime Severity, and Victim Sex), were encoded using dummy variables, facilitating the integration of these categories into the machine learning models that will follow. This encoding is necessary to transform categories into a format that algorithmic models can understand, enhancing the accuracy of predictions. The process of data cleaning and preprocessing was a crucial stage in ensuring the quality and reliability of the dataset, preparing it for further analysis and use in machine learning algorithms. Every step described above was justified and implemented with the goal of enhancing the accuracy and reliability of the analysis, ensuring that the data is suitable for use in advanced analysis and prediction models.

* 1. Exploratory Data Analysis (EDA)

After cleaning and preprocessing the data, the next important stage of the analysis was the Exploratory Data Analysis (EDA). EDA is a fundamental step in the data analysis process as it allows for a better understanding of the structure and characteristics of the data before applying machine learning models. During EDA, various visualization and statistical analysis techniques were used to examine patterns, correlations, and anomalies in the data. One of the main objectives of EDA was to understand the distribution of the data and identify significant features that could impact the analysis or the results of the models.

The analysis began with examining the distribution of crimes over time. Time series were constructed to capture crime trends over different time periods, such as months and days of the week. This approach was chosen to identify periodic patterns or seasonality that could indicate when crimes are more frequent or which months and days are the most dangerous. Analyzing the distribution of crimes over time helps in decision-making for the allocation of police resources during periods of increased risk.

Simultaneously, the geographical distribution of crimes was examined. Geographic maps (heatmaps) and scatter plots were used to visualize the areas with the highest frequency of crimes. This spatial analysis was critical for understanding areas with a high concentration of criminal activity and identifying potential high-risk areas. These visualizations allow for the visual recognition of patterns that may not be immediately apparent from statistical data and provide valuable insights for formulating strategies for crime prevention and response. Additionally, correlation analysis between various dataset features, such as the age of victims, type of crime, and geographical area, was performed. This analysis involved using correlation matrices and pair plots to identify any strong correlations that could influence predictive models. Understanding the correlations between features is vital for selecting the appropriate variables to be used in machine learning models and for avoiding issues such as multicollinearity.

Another important aspect of EDA was the analysis of the demographic characteristics of the victims. The data was categorized based on the age, gender, and ethnicity of the victims, and victimization patterns were examined to identify population groups that are more vulnerable to specific types of crimes. This analysis is essential for understanding the social parameters associated with criminality and for shaping policies to protect the most vulnerable groups. During EDA, anomalies and outliers that could affect the analysis and model outcomes were also identified and addressed. Outliers were either removed or adjusted to avoid potential distortions in the results. This process is particularly important to ensure that the data used in the models is representative and does not contain values that could lead to inaccurate results.

Finally, during the Exploratory Data Analysis, features that could play a significant role in the modeling process were investigated. This approach allowed for the identification of the most important variables and the removal of irrelevant features, facilitating the creation of more effective and accurate predictive models. The selection of the most relevant features was based on analyzing their impact on crime and their potential to improve model performance.

In summary, Exploratory Data Analysis was a crucial step for understanding the structure and relationships within the dataset. The techniques used and the findings highlighted laid the groundwork for proper data preparation, feature selection, and the effective application of the machine learning algorithms that followed.

* 1. Machine Learning Techniques

The final phase of the methodology focused on applying machine learning (ML) techniques for predicting and analyzing crime incidents in Los Angeles. ML techniques were chosen for their ability to detect patterns in data and make accurate predictions, even when relationships between variables are complex or non-linear.

For the application of ML techniques, the cleaned and preprocessed dataset was used as the foundation. The process began with selecting appropriate ML algorithms known for their performance in similar problems. The algorithms used included models such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). These algorithms were chosen for their ability to handle large datasets, identify significant relationships between variables, and make predictions with high accuracy.

Initially, the data was split into training and test sets. This split is vital for evaluating the performance of the models, as it allows for testing them on data that has not been used during training. This ensures that the models do not overfit the training data and can generalize well to new data.

One of the first steps in applying ML algorithms was feature selection. Based on the findings from the Exploratory Data Analysis (EDA), the most important features influencing crime, such as geographic location, time and date of the incident, type of crime, and victim demographics, were selected. Proper feature selection is crucial for model performance, as relevant features improve prediction accuracy, while unrelated features can add noise and reduce model effectiveness.

Next, the models were trained using the training data. During training, the ML algorithms adjusted to the data, recognizing patterns and relationships between features and the target variable. To ensure optimal performance, various techniques, such as hyperparameter tuning, were used, allowing the adjustment of algorithm parameters to improve their performance. After training the models, their performance was evaluated using the test set. Various metrics, such as accuracy, precision, recall, and F1-score, were used for evaluation. These metrics were chosen for their ability to provide a comprehensive view of model performance, considering not only prediction accuracy but also the balance between correct and incorrect predictions. This evaluation is critical to ensure that the models are not only accurate but also reliable in different use scenarios.

After evaluating the models, the best-performing model was selected for use in predicting crime incidents. Additionally, the selected model underwent robustness tests to ensure that its predictions remain reliable even when the data contains noise or outliers. The final performance of the model was recorded and analyzed, providing significant insights into crime patterns in the city of Los Angeles.

The use of machine learning techniques in crime data analysis provides the ability to predict future incidents and support authorities in decision-making. ML techniques offer a powerful tool for analyzing complex data, allowing for the development of strategies for crime prevention and response based on reliable predictions.

1. Results
   1. Eda Results

After the successful implementation of the methods presented in the previous section, the next steps focused on analyzing the results derived from data exploration (EDA) and the application of machine learning (ML) techniques. The aim of the results section is to clearly present the findings of these processes and highlight the main conclusions that can be drawn from the crime data in the city of Los Angeles. This analysis not only describes the data but also interprets the observations and correlates them with existing theories and knowledge in the field of criminology.

Analysis of Crime Distribution by Area *(Appendix 1).* The first step in analyzing the results was to investigate the geographical distribution of crimes. As shown in the first chart, the distribution of crimes varies significantly by area in Los Angeles. The areas with the highest concentration of criminal incidents are Central, 77th Street, and Southwest. These results suggest that certain areas of the city are more prone to criminal activity, which may be related to socioeconomic factors, population density, or other local characteristics. Identifying these areas is crucial for targeted allocation of police resources and the formulation of crime prevention policies.

Distribution of Victim Age *(Appendix 2)*. The age distribution chart of victims illustrates the age spread of crime victims, revealing that the majority fall within the 20-25 age group. The distribution shows a right skew, with most victims being young adults, and the number of victims gradually decreasing as age increases. This finding suggests that younger individuals are more vulnerable to crime, while older age groups are less frequently victimized.

**Crime Distribution by Day of the Week** *(Appendix 4).* The analysis of crime distribution by day of the week showed that crimes are relatively evenly distributed throughout the week, with slightly higher values on Monday and Tuesday. This observation suggests that crime is not significantly affected by the day of the week, which may reflect the stability of citizens' daily routines and activities. This finding is important for understanding crime rhythms and can assist in better distribution of police patrols and resources throughout the week.

Crime Distribution by Month *(Appendix 5).* From the analysis of crime distribution by month, it was observed that January is the month with the highest frequency of criminal incidents. The remaining months show smaller fluctuations, with generally stable crime rates. This trend may be attributed to seasonal factors, such as social activities and weather conditions, that may influence the frequency and nature of crimes. Understanding seasonality is useful for taking preventive measures and preparing authorities for high-risk periods.

Crime Distribution by Victim's Ethnicity *(Appendix 6)*. The fourth chart focuses on the distribution of crimes based on the ethnicity of the victims. Individuals of Hispanic descent appear to be the most frequent victims of criminal acts, followed by White and Black individuals. This finding may indicate differences in crime exposure among various demographic groups, which could be related to socioeconomic factors, discrimination, or other social dynamics. Understanding these differences is crucial for creating policies aimed at protecting the most vulnerable groups.

Crime Distribution by Victim's Gender *(Appendix 7)*. Regarding crime distribution by the victim's gender, the fifth chart shows that victims are approximately evenly distributed between men and women, with men slightly outnumbering women. This finding suggests that gender may not be the most determining factor in crime exposure, but this analysis confirms that both genders are equally vulnerable, something that should be considered in crime prevention strategies.

Crime Severity in Relation to Victim's Age *(Appendix 8)*. The relationship between the severity of crimes and the age of the victims is examined in the next chart. Age groups appear to be evenly distributed among different levels of crime severity (High, Medium, Low). This observation suggests that the victim's age does not seem to play a decisive role in determining the severity of the crime, although there are some differences, particularly in older age groups, which seem to be more frequently associated with medium-severity crimes.

Crime Severity in Relation to Victim's Gender *(Appendix 9).* The last chart compares the severity of crimes with the victim's gender. It is observed that male victims predominate in moderate and low-severity crimes, while women appear more as victims in low-severity crimes. This analysis can help in formulating policies aimed at preventing and addressing crimes that affect women and men differently.

Results of Statistical Tests As part of the data analysis, various statistical tests were conducted to assess the suitability of the data and to draw critical conclusions. The **Shapiro-Wilk** normality test was applied to assess the distribution of victims' ages, and the results showed that the statistical values were 0.97675 and 0.96284, respectively, while the P-values were 0.0 in both cases. These findings suggest that we reject the null hypothesis (H0) of normality, meaning that the victims' ages do not follow a normal distribution. This information is particularly important as it affects the choice of appropriate statistical techniques and machine learning models for further analysis.

Additionally, a t-test was conducted to compare the mean ages of victims in different crime categories. The t-test showed that the t-statistic was 53.20933 and the P-value was 0.0. These results indicate that there is a statistically significant difference between the mean ages of victims across crime categories, suggesting that the victim's age may influence the likelihood of being a victim of specific crime categories. This is an important finding that should be considered in the data analysis.

Analysis of Variance (ANOVA**)** Furthermore, analysis of variance (ANOVA) was used to examine whether the mean ages of victims differ significantly among the different crime categories. The ANOVA results showed that the total sum of squares was 1.8608e+06 with degrees of freedom (df) of 3.0, while the F-statistic was 2574.5883 with a P-value of 0.0. The very high F-statistic, combined with the nearly zero P-value, suggests that we reject the null hypothesis (H0) and conclude that there are significant differences in the mean ages of victims among the crime categories. This result reinforces the hypothesis that age plays an important role in the likelihood of being a victim of certain crimes, making the analysis of these differences crucial for understanding crime in the study area.

Results of Machine Learning Models Moving on to the analysis using machine learning techniques, the Logistic Regression algorithm was used to classify the data based on different features. The results from applying the Logistic Regression algorithm to the analysis of clusters showed that the model performs extremely well, with accuracy, recall, and F1-score metrics approaching 1.00 for all categories. Specifically, the model's **accuracy reached 100%,** indicating that the model correctly classifies all observations in the examined data.

In the analysis of data based on the victims' gender, the Logistic Regression model also showed high performance, with **99% accuracy**. Women (F) and men (M) were correctly classified in most cases, with accuracy and recall metrics nearing 1.00. However, in the "Other" category, the model encountered some difficulties, with lower recall and F1-score values, which is likely due to the smaller number of observations in this category. Despite these difficulties, the overall performance of the model remains excellent, making it a useful tool for understanding and predicting crime based on demographic characteristics.

The performance of the algorithms used in this study was compared with other approaches presented in the literature. The Logistic Regression and K-means clustering algorithms proved their value in crime data analysis, achieving high prediction accuracy and effective crime grouping. Particularly, Logistic Regression performed very well in binary outcome classification cases, while K-means clustering allowed for the identification of geographic areas with similar crime patterns.

Compared to other techniques, such as Decision Trees and Random Forests, Logistic Regression had the advantage of simplicity and interpretability, making it a preferred tool for decision-making that requires immediate and understandable results. However, Decision Trees and Random Forests performed better in more complex classification problems, where numerous features had more complex relationships.

Additionally, the use of the Prophet model for time series analysis proved valuable, especially in predicting the seasonal trends of crime. The model offered accurate predictions that can help authorities in preparing and managing resources during periods of increased risk.

* 1. Crime Dashboard

The "Crime Analysis Dashboard" developed as part of this thesis is a comprehensive tool for analyzing and predicting crime in the city of Los Angeles. The dashboard leverages modern machine learning techniques, such as Principal Component Analysis (PCA), K-means clustering, and time series analysis with the Prophet model, to provide useful and practical results. The platform aims to facilitate the understanding of crime data and support decision-making for city safety.

**Basic Statistical Information** *(Appendix 11).* The basic statistical information section provides an overall view of the data used in the analysis. It presents key statistical measures such as the mean, standard deviation, minimum and maximum values, and percentiles for various variables. These statistical measures are fundamental for understanding the data and allow users to see the basic trends and variations in crime data. For example, understanding the age distribution of victims can help identify the most vulnerable age groups.

Elbow Method *(Appendix 12)* for Optimal k The Elbow Method is one of the most popular techniques for selecting the optimal number of clusters in a K-means analysis. In the dashboard, the Elbow Method is used to determine the number of clusters that best represent the crime data. Selecting the optimal k is crucial because it allows for the grouping of data into distinct clusters with similar characteristics. In this case, the optimal k was determined to be 6, indicating that the data can be divided into six clusters representing different crime patterns.

K-means Clustering *(Appendix 13)* with 6 Clusters The application of K-means clustering is crucial for identifying patterns in crime data. This analysis creates six different clusters, each representing a different crime pattern. The use of PCA for dimensionality reduction allows for the visualization of the clusters in two dimensions, making them easier to interpret. These results are useful for identifying specific areas or periods with a high concentration of crimes, which can lead to more targeted police interventions.

Crime Rate Prediction with Prophet *(Appendix 15,16).* Crime rate prediction is one of the most useful tools offered by the dashboard. Using the Prophet model, users can generate forecasts for future crime rates in various areas of the city. This model considers trends and seasonality in the data to produce accurate forecasts. These predictions are crucial for taking preventive measures by security authorities, as they can help identify high-risk periods and appropriately allocate police resources.

Crime Heatmap by Area *(Appendix 14)*. The heatmap section offers a visualization of the geographical distribution of crimes in the city. Using different shades, the map illustrates the areas with the highest concentration of crimes. This visualization is particularly useful for identifying high-crime areas that may require increased security measures. For example, discovering that certain areas consistently have high crime rates can lead to targeted interventions such as increased patrols or community prevention programs.

Crime Distribution by Month and Day of the Week *(Appendix 4,5).* These sections provide information on the seasonality of crime. The analysis of crime distribution by month shows when crimes are most likely to occur, which may be due to seasonal factors, social events, or other activities. Similarly, the distribution of crimes by day of the week can indicate specific days when crimes are more frequent. This information is useful for enhancing security during specific periods or days based on the patterns that emerge.

Difference Between Date Reported and Date Occurred *(Appendix 3,10).* This analysis allows for understanding the delays in crime reporting. Visualizing the differences between the date a crime was reported and the date it occurred reveals how long it takes for citizens to report crimes. Significant delays in reporting may indicate issues with citizens' trust in authorities or other social factors that hinder immediate reporting. This finding can lead to the development of strategies to reduce these delays, such as public education or improving reporting procedures.

The "Crime Analysis Dashboard" is a powerful tool that combines advanced machine learning techniques with flexible visualizations to analyze and predict crime in the city of Los Angeles. Through the various tools it offers, users can understand crime patterns, predict future trends, and develop strategies to combat crime. The results from the dashboard provide valuable information that can be used by authorities to improve public safety and effectively allocate resources.

* 1. Probability Dashboard

The "Crime Probability Dashboard" developed as part of this thesis offers a flexible tool for analyzing the probabilities of crime occurrence in various contexts, such as the day of the week, month, and area in the city of Los Angeles. This dashboard allows users to analyze how specific characteristics, such as the age, gender, ethnicity of victims, crime severity, and geographical areas, relate to the likelihood of crime occurrence. Additionally, users can select the type of analysis from three categories: "Day of Week," "Month," and "Area." Depending on their selection, they can analyze crime data based on the day of the week, the month, or the area. These filters allow for a more specialized approach to data analysis. For example, by selecting "Day of Week," users can view crime patterns for a specific day, such as Monday, and understand how crime frequency varies by day. This feature is particularly useful for identifying specific trends and adapting security strategies based on the findings *(Appendix 17)*.

Based on the selection of the "**Central**" area, the charts present the following information and trends related to crime:

Distribution by Age Group *(Appendix 18)*. The first chart shows that the "Young Adult" age group (19-35 years) is the most likely to be involved in crimes in the "Central" area, with a probability of 47.87%. This is followed by the "Adult" group (36-55 years) with a probability slightly above 30%. The "Senior" (56+) and "Child" (0-18) categories show much lower probabilities, indicating that young adults are the main individuals involved in crimes in this area.

Distribution by Victim's Gender *(Appendix 19)*. The second chart presents the distribution of victims by gender in the "Central" area. Males constitute most victims with 59.5%, compared to females who make up 40.5%. This finding may suggest that men in the area are more likely to be involved in or become targets of criminal activities.

Distribution by Victim's Descent *(Appendix 20)*. The third chart shows that the highest probability of being a victim in the "Central" area belongs to the "Hispanic/Latin/Mexican" ethnic group, with 35.50%, followed by "Black" and "White." Other ethnic groups have lower percentages, indicating that Latinos and Blacks are more likely to be crime targets in this area.

Distribution by Month *(Appendix 21)*. The fourth chart analyzes crime based on the month, showing that January has the highest probability of crime occurrence (9.95%) in the "Central" area. The remaining months have similar probabilities, with no significant fluctuations, suggesting that crime remains relatively stable throughout the year, with a slight increase at the beginning of the year.

Distribution by Crime Severity *(Appendix 22)*. The fifth chart shows that medium-severity crimes (Medium) are the most common in the "Central" area, with 49.13%. Low-severity crimes (Low) follow with 44.5%, while high-severity crimes (High) are much less common, with only 6.4%. This suggests that the area primarily experiences crimes that are not extremely violent or severe.

Distribution by Crime Category *(Appendix 23)*. The final chart shows the distribution of crimes by category in the "Central" area. Assaults are the most frequent category, with a probability of 28.97%, followed by other crimes (Other), thefts (Theft), and burglaries (Burglary). This distribution suggests that violent crimes and thefts are the main crime categories in the area.

The results produced by the "Probability Dashboard" are extremely useful for strategic crime prevention. By understanding the probabilities of crime occurrence based on specific characteristics, security authorities can direct their efforts more effectively. For example, knowing that young adults are more vulnerable on Monday mornings, they can increase policing in areas with a high concentration of young people during these hours. Moreover, identifying areas with higher crime rates and understanding the social and demographic factors associated with crime can lead to more effective preventive measures. This information can also be used to develop community programs aimed at reducing crimes in specific areas or among specific population groups.

The "Probability Dashboard" provides a valuable tool for analyzing the probabilities of crime occurrence in different contexts. Through its flexible structure, it allows users to explore and understand crime patterns based on geography, time, and the demographic characteristics of victims. These analyses can be used to enhance public safety and develop strategies tailored to the needs of local communities. This thesis demonstrates the importance of data analysis for understanding and combating crime, offering practical tools that can have a real impact on citizen safety.

1. Discussion

The results derived from the two dashboards offer a multi-dimensional understanding of crime in the city of Los Angeles.

* 1. Crime Dashboard

The analysis of the results from the Crime Dashboard provides a deep understanding of crime patterns and trends in the city of Los Angeles, with a particular focus on the Central Area. A detailed examination of these results allows for the identification of critical factors influencing crime, as well as the prediction of future trends that can guide response strategies.

The analysis of crime data in the Central Area of Los Angeles revealed significant information about the distribution and trends of crimes. A high density of crimes is observed in specific areas, likely linked to high population density and intense commercial activity. The Central Area, as the economic and social center of the city, attracts many people daily, increasing the likelihood of various forms of crime such as theft, violence, and property damage.

A significant observation is the seasonal and weekly variation in crimes. The data show increased crime rates at the beginning of the week, particularly on Monday, and higher levels during the summer months. These trends can be interpreted through the lens of social and economic factors influencing citizen behavior. For example, the start of the week often brings increased return to work and school activities, which can create stress and conflicts, leading to an increase in criminal activities.

Additionally, the analysis showed that low-severity crimes, such as petty theft and vandalism, are more frequent in the Central Area, while high-severity crimes, such as assaults and robberies, are more concentrated in specific areas with high socio-economic tension. This indicates that crime is not homogeneous across the area but varies according to the demographic and social characteristics of different sub-regions.

Implications of the Results. The findings from the Crime Dashboard have significant implications for developing crime prevention and response strategies in the Central Area of Los Angeles. Recognizing temporal and geographical crime patterns allows local authorities to allocate police resources more effectively. For example, increased police presence at the beginning of the week and during the summer months can prevent crimes and enhance the sense of security among citizens.

Moreover, understanding the specific types of crimes that are more frequent in the Central Area allows for the development of targeted prevention programs. Programs such as increased street lighting, the installation of security cameras, and the strengthening of community actions can reduce the likelihood of high-severity crimes.

Limitations of the Research. Despite the value of the results, certain limitations must be recognized. The analysis is based on data collected from specific sources, which may not fully reflect the actual state of crime in the area. Additionally, the forecasts made using the Prophet model are based on historical data and assume that trends will continue in the future, which does not account for possible changes in socio-economic conditions or policing policies.

Furthermore, the generalizability of the results is limited to the context of the Central Area of Los Angeles and may not apply directly to other areas with different demographic and social characteristics. Lastly, there are other factors not examined in the present analysis, such as the effects of weather conditions, large events, and access to social services, which may influence crime rates.

Crime Prediction. The prediction of future crime trends in the Central Area based on the Prophet model reveals interesting prospects for the future. The analysis shows an upward trend in crime until 2023, followed by a predicted decline until 2025. This predicted decrease may be due to various factors, such as the implementation of more effective prevention strategies, increased policing resources, and improved socio-economic conditions in the area.

The three components of the forecast—the long-term trend (Trend), weekly fluctuations (Weekly), and yearly seasonal trends (Yearly)—provide a detailed view of the dynamics shaping crime. The long-term trend shows how general social and economic changes influence crime, while the weekly and yearly trends highlight specific timeframes where crime tends to increase or decrease.

These forecasts are particularly useful for local authorities, as they allow for better preparation for future policing and prevention needs. Based on the forecasts, they can develop programs targeting specific periods or areas with a high likelihood of crime, thereby improving the effectiveness of security measures and reducing crime throughout the area.

The analysis of the results from the Crime Dashboard highlights the complexity of crime in the Central Area of Los Angeles and underscores the importance of using advanced analytical tools to understand and predict crime trends. The deep interpretation of the data, combined with future predictions, offers local authorities valuable information that can be used to develop targeted crime prevention and response strategies.

* 1. Probability Dashboard

The Crime Probability Dashboard offers a flexible platform for analyzing crime, allowing data exploration based on area, day of the week, or month. This multi-dimensional approach gives users the ability to examine the distribution of crimes from different perspectives, thus facilitating the understanding of trends and the identification of potential risks.

In the analysis conducted for the Central Area, the dashboard revealed important information about the profile of victims and the characteristics of crimes in this area. The majority of crimes seem to target young adults (aged 19-35), highlighting the vulnerability of this age group in the area. This finding may be related to the social and professional activity of young adults, as well as the mobility that characterizes this age group. Moreover, the analysis shows that men are more likely to fall victim to crimes compared to women in the Central Area. This difference may reflect different gender behaviors and roles in society, as well as men's exposure to higher-risk environments or activities.

Regarding the ethnicity of victims, the analyses show that crimes in the Central Area disproportionately affect people of Hispanic/Latin/Mexican descent. This finding is particularly significant as it may indicate the existence of social or economic factors that increase the vulnerability of this population group to crimes. The analysis of data by month identified January as the month with the highest probability of crime, while medium-severity crimes dominate the area. The stability in the severity of crimes, with the majority being of medium intensity, indicates a specific crime pattern in the area that is not characterized by particularly violent or severe crimes.

Finally, in the category of crimes, assaults show the highest frequency in the Central Area, possibly indicating the presence of tensions or conflicts in this area. This, combined with other findings, can help authorities develop targeted strategies for preventing and addressing criminal activities. Overall, the Crime Probability Dashboard provides a comprehensive view of crime in the Central Area, helping to identify critical points and trends that can be leveraged to improve public safety and strengthen the community. With the ability to analyze by area, day of the week, or month, it provides valuable information for developing preventive measures and strategies based on data.

* 1. Compare Dashboards

The comparison of the two dashboards, the Crime Dashboard and the Crime Probability Dashboard, highlights their differences in methodology, goals, and results, as well as their similarities in their overall approach to crime analysis. This analysis will help the reader understand not only the capabilities and limitations of each tool but also how they can be used complementarily for strategic decision-making and crime prevention.

Similarities. Both dashboards were designed with a common purpose: to analyze crime data to support authorities in developing crime prevention and suppression strategies. Both tools allow users to focus on specific geographic zones, which is crucial for local analysis and the application of targeted strategies. More specifically, both dashboards make use of advanced statistical and predictive models to produce useful information. The Crime Dashboard uses the Prophet model to predict future crime trends based on historical data. The Crime Probability Dashboard utilizes probability analyses to determine which groups are most vulnerable in specific areas and time periods. Both tools provide their users with the ability to better understand the dynamics of crime, offering valuable insights for decision-making.

Differences. Despite their similarities, the two dashboards differ significantly in their approach and the results they produce. The Crime Dashboard is oriented towards temporal analysis and projecting trends into the future. Through predictive models such as Prophet, it allows for the creation of forecasts for the evolution of crime in a specific area. This capability is extremely useful for the strategic planning of long-term initiatives and the evaluation of future policing needs.

On the other hand, the Crime Probability Dashboard focuses on probability analysis, offering users the ability to explore which population groups are more vulnerable and under what conditions. Unlike the Crime Dashboard, which analyzes the temporal evolution of crime, the Probability Dashboard provides immediate assessments of the present, allowing for the identification of trends and patterns based on specific variables such as the day of the week, month, and area. Additionally, the results of the Crime Dashboard are focused on dynamic data visualization and understanding long-term trends, as seen from the temporal predictions it creates for various areas. In contrast, the Crime Probability Dashboard emphasizes analyzing the probabilities of specific events and groups, offering a more detailed and immediate understanding of the current risk.

Usage and Applications. The use of these two tools provides complementary capabilities. The Crime Dashboard is extremely useful for developing long-term strategies and strategic planning, as it offers forecasts based on historical data and helps estimate future challenges. Conversely, the Crime Probability Dashboard provides rapid and immediate analysis of crime probabilities, making it ideal for developing tactical strategies and addressing current threats.

The analysis of both dashboards shows that they can be used complementarily for a comprehensive understanding and response to crime. The Crime Dashboard offers a view into the future, with predictions based on historical data, while the Crime Probability Dashboard provides a snapshot of current conditions, allowing for immediate decision-making. Their combined use can enhance the effectiveness of crime reduction strategies, offering both long-term and short-term solutions.

1. Conclusion

This thesis explored the integration of data analysis and machine learning techniques to understand and predict crime patterns in the city of Los Angeles. Within this framework, two interactive tools were developed—the Crime Dashboard and the Probability Dashboard—aimed at providing law enforcement agencies with powerful tools for better understanding crime trends and enhancing crime prevention strategies.

* 1. **Summary of Research Process**

The research began with the selection and collection of a rich dataset from Los Angeles, which included detailed crime records over several years. Preprocessing these data was a crucial step, requiring data cleaning and formatting to ensure they were suitable for analysis. The analytical process involved the use of techniques such as Principal Component Analysis (PCA) to reduce the complexity of the data and extract significant information, as well as the application of K-means Clustering to identify patterns and trends within the dataset.

The Crime Dashboard was created to allow users to explore crime trends visually and interactively. Through the detailed exploration of data, this tool enables the identification of high-risk areas, analysis of temporal trends, and understanding of crime distribution by victim gender, age, and ethnicity.

On the other hand, the Probability Dashboard focused on predicting future crimes by utilizing statistical models to estimate the likelihood of criminal incidents occurring. These predictions are particularly useful for security authorities, as they allow for proactive measures and the optimal allocation of resources to maximize their effectiveness.

This work significantly contributes to the literature on crime analysis by offering a comprehensive framework for applying machine learning techniques to crime analysis and prediction. The use of the developed tools in real-world scenarios has the potential to enhance the capability of law enforcement agencies to better understand crime patterns, make more informed decisions, and ultimately reduce crime rates.

During the development of these tools, several challenges were encountered. The selection and preprocessing of data proved crucial, as issues such as data incompleteness, format incompatibility, and the need for precise and reliable results had to be addressed. Additionally, the development of machine learning models, such as PCA and K-means Clustering, required careful tuning to achieve optimal results. However, these efforts resulted in the successful implementation of two tools that proved useful for analyzing and predicting crime.

The data analysis and the use of the developed tools provided significant answers to the initial research questions. The Crime Dashboard offered detailed visualizations of crime trends, allowing for data analysis by area, day, and month. The Crime Probability Dashboard provided estimates of the likelihood of crimes occurring, indicating the most dangerous periods and areas. These results enhanced the understanding of crime patterns in Los Angeles and demonstrated how data could be used to improve policing.

Despite the research's successes, several weaknesses and areas for improvement must be noted. First, the data used, although extensive, had certain deficiencies, such as missing fields and possible recording errors. This affected the accuracy of some analyses and predictions. Second, while the models developed were generally effective, they could be further improved by incorporating additional parameters or utilizing more advanced machine learning techniques, such as deep neural networks.

Moreover, the generalizability of this research's results may be limited due to the focus on data from Los Angeles alone. The use of data from other cities or the integration of socio-economic data could provide a more comprehensive understanding of crime. Lastly, applying the results to real-world situations would require further studies to adapt and optimize the models based on the specific needs and challenges of each area.

The dashboards developed in this work have significant real-life applications. The Crime Dashboard can be used by law enforcement to monitor current crime trends and analyze crime patterns across different areas. This will allow for the targeted allocation of police forces and the development of prevention strategies based on real-time data.

The Crime Probability Dashboard offers the ability to predict future criminal incidents, thus enabling the development of preventive measures and prompt response in high-risk areas. This predictive capability can lead to a significant reduction in crimes, as authorities can act before crimes occur rather than react afterward.

* 1. **Global Implications and Future Directions**

In a broader context, these tools could be adapted for use in other cities and countries, tailored to local needs and data, thereby contributing to improved public safety on a global scale. The information provided by the dashboards can support policy decisions, enhance collaboration among various services, and promote social peace through crime prevention.

Future research could explore the integration of real-time data streams and more sophisticated predictive models to further enhance the accuracy and utility of these tools. Additionally, exploring the ethical implications of predictive policing and ensuring transparency and fairness in the use of these tools will be essential as their adoption increases.

This thesis demonstrates the significant potential of data-driven approaches in combating crime and highlights the importance of continued innovation in this critical area of public safety.

References

Cung, B., 2013. *Crime and Demographics: An Analysis of LAPD Crime Data.,* Los Angeles: escholarship.

Facebook, 2017. *Facebook.* [Ηλεκτρονικό]   
Available at: Prophet: Forecasting at scale. Retrieved from :https://facebook.github.io/prophet/

Gerber, M. S., 2014. Predicting crime using Twitter and kernel density estimation.. *Decision Support Systems*, pp. 61 , 115-125.

Harry, C. R. και συν., 2020. *Array programming with NumPy.* [Ηλεκτρονικό]   
Available at: Array programming with NumPy

Hunter , J. D., 2007. *Matplotlib.* [Ηλεκτρονικό]   
Available at: graphics environment. Computing in Science & Engineering, 9(3), 90-95. doi:10.1109/MCSE.2007.55

Jean , N. και συν., 2016. Combining satellite imagery and machine learning to predict poverty. *Combining satellite imagery and machine learning to predict poverty*, pp. 353(6301), 790-794.

lnc, S., 2019. *Streamlit.* [Ηλεκτρονικό]   
Available at: The fastest way to build and share data apps. Retrieved from https://streamlit.io/

Microsoft, 2024. *visual studio.* [Ηλεκτρονικό]   
Available at: https://code.visualstudio.com/

Mohler , G. O. και συν., 2008. Randomized controlled field trials of predictive policing.. *Journal of the American Statistical Association,* 2(3), pp. 259-265.

Mohler, G. O. και συν., 2015. Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, pp. 110(512) 1399-1411.

Pedregosa , F. και συν., 2011. Scikit-learn Machine Learning in Python. *Journal of Machine Learning Research,*, 12, pp. 2825-2830.

Richardson , R., Schultz, J. M. & Crawford, K., 2019. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review*, pp. 192-233.

Team, P. D., 2020. *Pandas.* [Ηλεκτρονικό]   
Available at: pandas: Powerful Python data analysis toolkit. Retrieved from https://pandas.pydata.org/

Van Rossum, G. & Drake , F. L., 2009. *Python 3 Reference Manual.* [Ηλεκτρονικό]   
Available at: Scotts Valley, CA: CreateSpace. Retrieved from https://www.python.org/

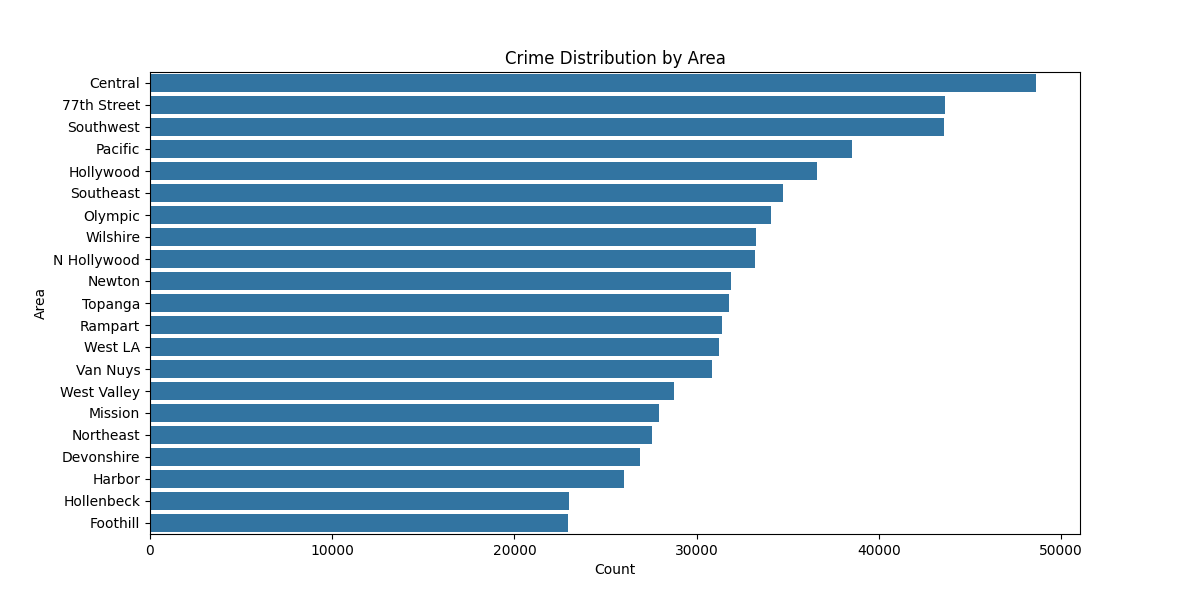
Wang, T., Rudin , Wanger , D. & Sevieri, R., 2013. *Learning to detect patterns of crime. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 515-524.* ACM, s.n.

Waskom, M. L, 2021. *Seaborn.* [Ηλεκτρονικό]   
Available at: Statistical data visualization. Journal of Open Source Software, 6(60), 3021. doi:10.21105/joss.03021

Appendix

**Appendix EDA**

Appendix 1

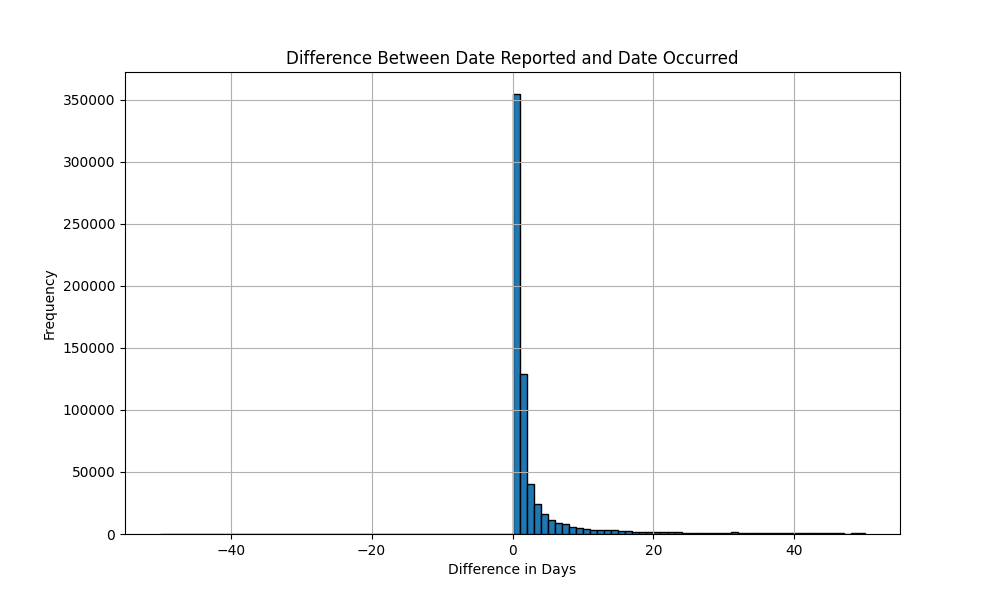


Appendix 2

Εικόνα που περιέχει διάγραμμα, στιγμιότυπο οθόνης

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 3



Appendix 4

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά, ορθογώνιο παραλληλόγραμμο

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 5

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά, γράφημα

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 6

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, διάγραμμα, γραμματοσειρά

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 7

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, ορθογώνιο παραλληλόγραμμο, διάγραμμα

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 8

Εικόνα που περιέχει διάγραμμα, ορθογώνιο παραλληλόγραμμο, στιγμιότυπο οθόνης, γραμμή

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 9

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, διάγραμμα, γράφημα

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 10

Εικόνα που περιέχει κείμενο, γράφημα, γραμμή, διάγραμμα

Περιγραφή που δημιουργήθηκε αυτόματα

**Appendix Crime Dashboard**

Appendix 11

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά, αριθμός

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 12

Εικόνα που περιέχει κείμενο, γραμμή, γράφημα, διάγραμμα

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 13

Εικόνα που περιέχει κείμενο, χάρτης, στιγμιότυπο οθόνης, γραμματοσειρά

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 14

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, ορθογώνιο παραλληλόγραμμο

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 15

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, διάγραμμα, γράφημα

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 16

Εικόνα που περιέχει κείμενο, διάγραμμα, γράφημα, γραμμή

Περιγραφή που δημιουργήθηκε αυτόματα

**Appendix Crime Probability Dashbord**

Appendix 17

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά

Περιγραφή που δημιουργήθηκε αυτόματα Appendix 18

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, διάγραμμα, γραμμή

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 19

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά, διάγραμμα

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 20

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, διάγραμμα, γραμματοσειρά

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 21

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, πολυχρωμία, γραμμή

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 22

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά, διάγραμμα

Περιγραφή που δημιουργήθηκε αυτόματα

Appendix 23

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, διάγραμμα, ορθογώνιο παραλληλόγραμμο

Περιγραφή που δημιουργήθηκε αυτόματα