Spatio Temporal Crime Hotspot Identification and Prediction

Project report submitted in partial fulfillment of the requirements for the award of the degree of

MASTER OF SCIENCE

IN

STATISTICS WITH DATA SCIENCE

Submitted by

VIDYA S
(220011024018)

Under the supervision of

Dr.STEPHY THOMAS

Department of Statistics
Bishop Chulaparambil Memorial College,Kottayam
Mahatma Gandhi University,Kottayam

JULY 2024

Declaration

I, Vidya S, student of M.Sc Statistics with Data Science hereby declare that the project dissertation titled "Spatio Temporal Crime Hotspot Identification and Prediction" which is submitted by me to the Department of Statistics, B.C.M College, Kottayam, in fulfillment of the requirement for awarding of the Master of Science degree, is not copied from any source without proper citation. This work has not previously formed the basis for the award of any degree, diploma, fellowship or other similar title or recognition.

Date: 15-07-2024

Place: Kottayam **Vidya S**

CERTIFICATE

This is to certify that the project titled "SPATIO TEMPORAL CRIME HOTSPOT IDENTIFICATION AND PREDICTION," submitted by Vidya S in partial fulfillment of the requirements for the degree of Master of Science (M.Sc.), is a record of the original research carried out by the student under my guidance and supervision. To the best of my knowledge, this work has not been submitted, either in whole or in part, for any other degree or diploma at this university or any other institution.

Date: 15-07-2024

Place: Kottayam

Dr. STEPHY THOMAS
(HEAD OF THE DEPARTMENT)
Department of Statistics
B.C.M. College, Kottayam

Acknowledgement

This project is the result of many months of strenuous effort, during which I have received advice and support from numerous individuals.

First and foremost, I am deeply grateful to **Dr. Stephy Thomas**, my guide and Head of the Department of Statistics at B.C.M College, Kottayam. His constant guidance, support, and insightful feedback were crucial to the successful completion of this project.

I also wish to express my sincere thanks to all the faculty members of the Department of Statistics. Their support, motivation, and expert advice were essential in overcoming the challenges faced during this endeavor.

Additionally, I extend my gratitude to everyone who contributed to this project, both directly and indirectly. Your help has been greatly appreciated.

Finally, I would like to thank my parents and well-wishers for their unwavering support and encouragement. Their constant presence and belief in me have been a source of strength throughout this journey.

VIDYA S

Contents

List of Figures

List of Tables

CHAPTER 1
Introduction

Crime remains a pervasive and complex challenge faced by communities and law enforcement agencies worldwide. Understanding and predicting crime patterns is crucial for effectively deploying resources, developing crime prevention strategies, and enhancing public safety. Traditional crime analysis methods often fall short in capturing the dynamic and multifaceted nature of criminal activity, which exhibits both spatial and temporal variability. This project aims to address these challenges through advanced spatio- temporal analysis techniques, specifically focusing on the identification and prediction of crime hotspots.

Spatio-temporal crime hotspot identification involves analyzing crime data to detect areas with unusually high concentrations of criminal activity. This requires sophisticated statistical methods to account for the geographic distribution and temporal patterns of crimes. Getis-Ord Gi* statistics, a widely used spatial analysis tool, will be employed to identify clusters of high crime intensity. By evaluating the spatial arrangement of crime incidents, this method helps in pinpointing areas that require targeted interventions.

Predicting future crime hotspots involves forecasting how crime patterns will evolve over time. For this purpose, the Prophet forecasting model—a robust tool for time series analysis—will be utilized. PROPHET, developed by Facebook, excels at handling time series data with strong seasonal effects and missing values, making it suitable for predicting crime trends based on historical data.

The integration of this methods—spatial hotspot identification and temporal forecasting—provides a comprehensive framework for understanding and anticipating crime patterns. By combining spatial and temporal dimensions, this project seeks to enhance the accuracy of crime predictions and support proactive law enforcement strategies.

The significance of this project lies in its potential to improve crime prevention and resource allocation. Through detailed analysis and forecasting, it aims to provide actionable insights for law enforcement agencies, urban planners, and policymakers. This approach not only helps in identifying high-risk areas but also in anticipating future crime trends, thereby contributing to a more strategic and informed approach to crime management.

Overall, this project represents a critical step towards leveraging advanced statistical techniques for the betterment of public safety. By exploring the intersections of space and time in crime data, it promises to offer valuable contributions to the field of criminology and statistics.

This project investigates the use of advanced spatio-temporal analysis techniques for identifying and predicting crime hotspots, focusing on the dynamic and multifaceted nature of criminal activity. Employing Getis-Ord Gi* statistics for spatial analysis to detect high-crime areas and the Prophet model for temporal forecasting to predict future crime trends, this project aims to enhance the accuracy of crime predictions. By integrating these methods, the study provides a comprehensive framework for understanding and anticipating crime patterns, offering actionable insights for law enforcement, urban planners, and policymakers to improve crime prevention and resource allocation strategies

CHAPTER 2
Literature Review

This chapter provides a concise overview of key advancements in spatio-temporal crime prediction and detection, focusing on recent research contributions and identified gaps in the field.

The paper by Butt et al. (2020) provides a comprehensive systematic literature review (SLR) on spatio-temporal crime hotspot detection and prediction. This review critically examines the methodologies and advancements in the field, emphasizing the integration of data mining and machine learning techniques for crime analysis. It highlights the evolution of spatio-temporal crime prediction models, which are crucial for effective law enforcement and market research applications.

The authors categorize existing approaches into several domains, including time series analysis, spatial statistics, and machine learning models. They discuss various methods such as spatial point pattern analysis, time-series forecasting, and hybrid models that combine spatial and temporal data to predict crime hotspots. The review also addresses the challenges associated with data quality, model accuracy, and the practical implementation of these techniques in real-world scenarios.

Overall, the paper provides valuable insights into the current state of spatio-temporal crime prediction and identifies key areas for future research, including the need for more robust and scalable models that can handle diverse and complex crime data

The paper by Zhang et al. (2022) presents a study on interpretable machine learning models for crime prediction, published in Computers, Environment and Urban Systems. The authors address the increasing need for transparent and understandable machine learning models in the context of crime forecasting. As predictive models become more complex, there is a growing emphasis on making these models interpretable to enhance trust and facilitate practical use by law enforcement agencies.

The study evaluates several machine learning algorithms that offer varying levels of interpretability, such as decision trees, linear models, and rule-based methods. The authors explore how these models can be utilized to predict crime hotspots and analyze their performance in terms of accuracy and comprehensibility. They also discuss the trade-offs between model complexity and interpretability, offering insights into how simpler models can sometimes provide more actionable information compared to more complex, black-box models.

Zhang et al. also emphasize the importance of integrating domain knowledge into model design to improve interpretability and ensure that the predictions are meaningful and actionable for stakeholders. The paper contributes to the field by providing a framework for selecting appropriate models based on the specific needs of crime prediction tasks, ultimately aiming to bridge the gap between advanced predictive analytics and practical application in crime prevention.

The paper by Yu et al. (2020) explores an advanced approach for crime prediction by integrating historical crime data with movement data of potential offenders through a spatio-temporal cokriging method. Published in ISPRS International Journal of Geo-Information, this study addresses the complexity of predicting crime hotspots by leveraging both spatial and temporal data to enhance predictive accuracy.

The authors propose a novel spatio-temporal cokriging model that combines historical crime records with the movement patterns of potential offenders. This approach allows for a more nuanced understanding of crime dynamics by incorporating multiple data sources, improving the model's ability to forecast crime hotspots. The use of cokriging, an extension of kriging methods that accounts for multiple variables, provides a robust framework for integrating different types of data and capturing the spatio-temporal dependencies in crime patterns.

Yu et al. highlight the effectiveness of their model through a series of experiments, demonstrating its superior performance compared to traditional methods. They discuss the practical implications of their findings for law enforcement and urban planning, emphasizing the model's potential to aid in more effective crime prevention strategies. The study contributes significantly to the field by providing a method that not only improves predictive accuracy but also offers valuable insights into the spatial and temporal aspects of crime.

Research Gap

The paper by Butt et al. (2020) identifies a critical research gap in the need for more robust and scalable models that can handle the complexity and diversity of crime data. While they provide a comprehensive review of existing methodologies, they highlight significant challenges related to data quality and model accuracy. This indicates a pressing need for the development of advanced techniques that can integrate various data types and sources more effectively, thereby enhancing the reliability and applicability of crime prediction models.

Zhang et al. (2022) focus on the importance of model interpretability in the context of crime prediction, addressing the growing need for transparent and understandable machine learning models. Their study points out the trade-offs between model complexity and inter-

pretability, emphasizing that simpler models can sometimes provide more actionable information than complex, black-box models. This gap suggests that future research should aim to develop machine learning models that balance high accuracy with interpretability, ensuring that these models can be practically used by law enforcement agencies. Incorporating domain knowledge into the design of these models is also crucial to improve their practical utility.

Yu et al. (2020) present an advanced approach to crime prediction by integrating historical crime data with the movement patterns of potential offenders using a spatio-temporal cokriging method. While their model demonstrates superior performance compared to traditional methods, there is still a need for further refinement of these advanced data integration techniques to enhance predictive accuracy and practical application. This study highlights the potential of combining multiple data sources to capture the spatio-temporal dependencies in crime patterns more effectively, suggesting that future research should continue to explore and develop such integrated approaches.

In summary, the primary research gaps identified across these studies include the need for more robust and scalable models, the development of interpretable and practically useful machine learning models, and the refinement of advanced data integration techniques. Addressing these gaps will be crucial for improving the accuracy and applicability of crime prediction models, ultimately contributing to more effective crime prevention strategies and resource allocation.

CHAPTER 3
Basic Concepts

This chapter explores key techniques in data analysis, focusing on crime hotspots and time series analysis. We start by identifying areas with disproportionately high crime rates through spatial analysis. Next, we delve into time series analysis to understand and forecast data trends over time, examining components such as trends, seasonality, and randomness. The chapter also covers forecasting methods and introduces Plotly for creating interactive visualizations, including choropleth maps. Finally, we discuss the K-nearest neighbors (KNN) algorithm, a fundamental machine learning technique for classification.

3.1 Crime Hotspot

A specific location or area where crime rates are significantly higher than those in surrounding areas.These areas often experience a disproportionate amount of criminal incidents relative to their size or population.Hotspots are typically identified through spatial analysis of crime data.

3.2 Time Series Analysis

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. Typically to understand underlying patterns and forecast future values.

3.2.1 Components of a Time Series

- Secular Trend or Trend: The long-term movement or direction in the data over a period. It shows the general direction in which the data is moving, either upward, downward, or stable.
- Seasonality variation: Regular, predictable variations that recur at specific intervals such as daily, weekly, monthly, or yearly.
- Cyclic variation: Long-term oscillations that are not of fixed period. These cycles are often influenced by economic or business cycles, and unlike seasonality, the duration of cycles is not fixed.
- Random or irregular movement: The random variation in the data that cannot be explained by trend, seasonality, or cyclic patterns. It represents the noise or irregular fluctuations in the data.

3.2.2 Objective of Time Series

Time series analysis is a statistical technique that deals with time-ordered data. Its main objectives include:

- Description:

  Summarize Data: Provide a comprehensive summary of the dataset's key characteristics, such as trends, seasonal patterns, and cyclic behavior.

Visualization: Use plots and charts to visualize the data over time, making it easier to identify patterns, anomalies, and changes.

- Modelling:

  Modeling Relationships: Develop models to understand the underlying relationships within the data. This can involve identifying the impact of various factors on the observed series.

  Statistical Tests: Perform tests to check for stationarity, autocorrelation, and other statistical properties that help explain the data's behavior.

- Forecasting:

  Predict Future Values: Use historical data to predict future values. This is crucial for planning, decision-making, and strategy development in various fields such as finance, economics, and business operations.

  Estimate Confidence Intervals: Provide confidence intervals for predictions to quantify the uncertainty associated with forecasts.

- Control:

  Monitor and Control Processes: In industrial and economic contexts, time series analysis helps in monitoring and controlling processes to maintain quality and efficiency.

  Anomaly Detection: Identify unusual patterns or anomalies that may indicate potential problems or opportunities.

3.3 Forecasting

Time series forecasting predicts future values based on historical data collected at regular intervals. This involves using statistical models and methods—such as ARIMA, exponential smoothing, and

machine learning techniques—to analyze past patterns, trends, and seasonality. By examining these historical patterns, forecasting aims to estimate future values and provide insights for decision-making in various fields like finance, economics, and inventory management. Accurate forecasting aids in anticipating trends, planning resources, and making informed decisions based on expected future conditions.

Time series analysis involves developing models to understand the data and its underlying causes. Although forecasts are not always precise due to the fluctuating nature of time series data and external factors, specialized time-series algorithms are crucial. Unlike general machine learning techniques—such as random forests and gradient boosting—which can extrapolate data but may not handle patterns outside the training domain well, time-series algorithms are designed to extend patterns beyond the data they were trained on, making them essential for effective forecasting.

3.4 Visualization using Plotly

Plotly is a powerful Python library for creating interactive and high-quality visualizations. It is particularly useful for exploratory data analysis and presenting results in a visually appealing manner.
Basic Concepts of Plotly Visualization

- Interactive Visualizations: Plotly allows for interactive charts and graphs where users can zoom, pan, and hover over data points to get more information. This interactivity enhances the user experience and helps in better understanding the data.
- Plotly Express: A high-level interface in Plotly that simplifies the creation of common types of visualizations with less code. It is useful for quickly generating plots like scatter plots, line charts, bar charts, and more.

- Plotly Graph Objects: A lower-level interface that provides more control and customization options for creating complex visualizations. It allows you to build plots from scratch and customize every aspect of the visual.
- Dash by Plotly: An open-source framework for building interactive web applications with Python. It integrates with Plotly to create interactive dashboards and data applications.

3.4.1 Choropleth Maps

- A choropleth map is a type of thematic map where areas are shaded or patterned in proportion to the value of a variable. It helps in visualizing how a variable changes across a geographic area, such as crime rates across different neighborhoods or regions.
- Plotly Implementation:
  - Data: Choropleth maps require data that includes geographic boundaries (such as regions or countries) and corresponding values.
  - Visualization: Plotly's plotly.express.choropleth function makes it straightforward to create choropleth maps. You provide it with data, location identifiers (e.g., country codes or state names), and values to display.
- . Key Parameters for Choropleth Maps
  - Locations: The column in your data that contains geographic identifiers (e.g., country names, state abbreviations).
  - Color: The column in your data that contains the values to be visualized.
  - Location Mode: Specifies the type of geographic identifier used (e.g., 'country names', 'ISO-3').

- Color Scale: Determines how values are mapped to colors. Plotly offers several color scales like 'Viridis', 'Cividis', and 'Blues'.
- Title and Labels: Customizable options to add titles and labels to enhance the readability of the map.

Using Plotly for visualization involves leveraging its capabilities to create interactive and visually appealing charts and maps. Choropleth maps are a powerful way to display spatial data and highlight regional differences. By understanding these basic concepts, we can effectively utilize Plotly to present your analysis results in a clear and engaging manner

3.5 K-nearest neighbors (KNN)

KNN is one of the most basic yet essential classification algorithms in machine learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning it does not make any underlying assumptions about the distribution of data . We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

The K-NN algorithm works by finding the K nearest neighbors to a given data point based on a distance metric, such as Euclidean distance. The class or value of the data point is then determined by the majority vote or average of the K neighbors.

Choosing k

The parameter k, representing the number of neighbors to consider, is a crucial factor in the performance of k-NN. A small k may lead to a noisy decision boundary, while a large k may oversmooth the boundary. The choice of k should be determined through experimentation or validation

Euclidean Distance

The Euclidean distance between two points $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ is given by the formula:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

Alternatively, using summation notation, it can be written as:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

CHAPTER 4

Methodology

The methodology of this study focuses on analyzing and predicting spatio-temporal crime hotspots using a combination of spatial statistics and time series forecasting techniques. The approach integrates data collection, data integration, preprocessing, exploratory data analysis (EDA), spatial analysis, and temporal forecasting 3

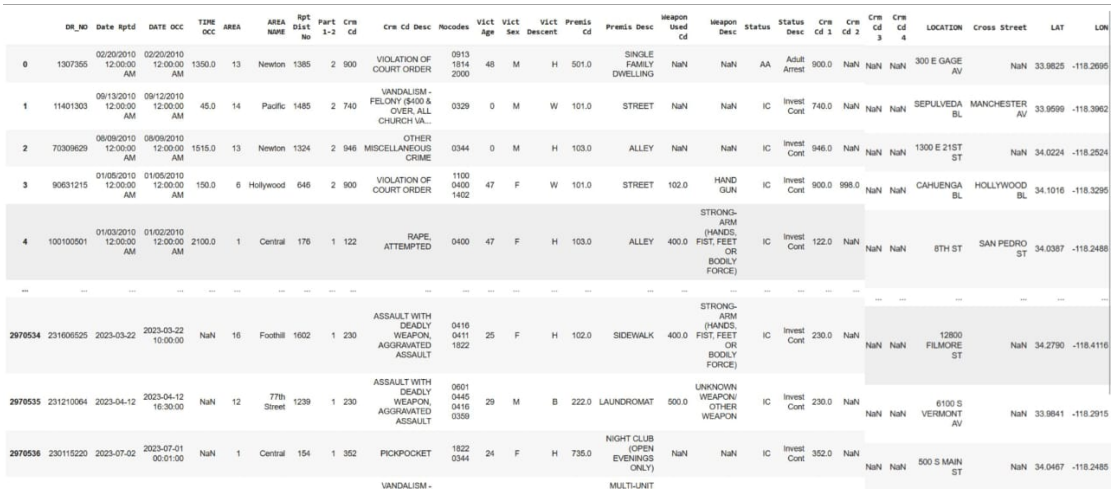
4.1 Data Collection

Data Sources: The data for this research project is sourced from Kaggle and consists of two primary datasets of crimes in Los Angeles over the years January 2010 to December 2023:

- Dataset 1: `Crime_Data_from_2010_to_2019.csv.zip`
  Contains crime records from January 2010 to December 2019 consists of (2117589, 28)
- Dataset 2: `Crime_Data_from_2020_to_Present.csv.zip` consists of (852950, 27)
  Includes crime records from January 2020 to December 2023.

4.2 Data Integration

- Unzip and load the crime datasets from CSV files.
- Standardize column names across datasets to ensure consistency. .



	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	Mocodes	Vict Age	Vict Sex	Vict Descent	Premis Cd	Premis Desc	Weapon Used Cd	Weapon Desc	Status	Status Desc	Crm Cd 1	Crm Cd 2	Crm Cd 3	Crm Cd 4	LOCATION	Cross Street	LAT	LON
0	1307355	02/20/2010 12:00:00 AM	02/20/2010 12:00:00 AM	1350.0	13	Newton	1385	2	900	VIOLATION OF COURT ORDER	0913 1814 2000	48	M	H	501.0	SINGLE FAMILY DWELLING	NaN	NaN	AA	Adult Arrest	900.0	NaN	NaN	NaN	300 E GAGE AV	NaN	33.9825	-118.2695
1	11401303	09/13/2010 12:00:00 AM	09/12/2010 12:00:00 AM	45.0	14	Pacific	1485	2	740	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VA...	0329	0	M	W	101.0	STREET	NaN	NaN	IC	Invest Cont	740.0	NaN	NaN	NaN	SEPULVEDA BL	MANCHESTER AV	33.9599	-118.3962
2	70309629	08/09/2010 12:00:00 AM	08/09/2010 12:00:00 AM	1515.0	13	Newton	1324	2	946	OTHER MISCELLANEOUS CRIME	0344	0	M	H	103.0	ALLEY	NaN	NaN	IC	Invest Cont	946.0	NaN	NaN	NaN	1300 E 21ST ST	NaN	34.0224	-118.2524
3	90631215	01/05/2010 12:00:00 AM	01/05/2010 12:00:00 AM	150.0	6	Hollywood	646	2	900	VIOLATION OF COURT ORDER	1100 0400 1402	47	F	W	101.0	STREET	102.0	HAND GUN	IC	Invest Cont	900.0	998.0	NaN	NaN	CAHUENGA BL	HOLLYWOOD BL	34.1016	-118.3295
4	100100501	01/03/2010 12:00:00 AM	01/02/2010 12:00:00 AM	2100.0	1	Central	176	1	122	RAPE, ATTEMPTED	0400	47	F	H	103.0	ALLEY	400.0	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	IC	Invest Cont	122.0	NaN	NaN	NaN	8TH ST	SAN PEDRO ST	34.0387	-118.2488
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2970534	231606525	2023-03-22	2023-03-22 10:00:00	NaN	16	Foothill	1602	1	230	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	0416 0411 1822	25	F	H	102.0	SIDEWALK	400.0	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	IC	Invest Cont	230.0	NaN	NaN	NaN	12800 FILMORE ST	NaN	34.2790	-118.4116
2970535	231210064	2023-04-12	2023-04-12 16:30:00	NaN	12	77th Street	1239	1	230	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	0601 0445 0416 0359	29	M	B	222.0	LAUNDROMAT	500.0	UNKNOWN WEAPON/ OTHER WEAPON	IC	Invest Cont	230.0	NaN	NaN	NaN	6100 S VERMONT AV	NaN	33.9841	-118.2915
2970536	230115220	2023-07-02	2023-07-01 00:01:00	NaN	1	Central	154	1	352	PICKPOCKET	1822 0344	24	F	H	735.0	NIGHT CLUB (OPEN EVENINGS ONLY)	NaN	NaN	IC	Invest Cont	352.0	NaN	NaN	NaN	500 S MAIN ST	NaN	34.0467	-118.2485
										VANDALISM -						MULTI-UNIT											

Figure 4.1: Dataset

- Data Merging: The two datasets are combined into a single DataFrame to ensure a comprehensive dataset that spans from January 2010 to December 2023. This is achieved using the pandas library's concat function and named it as "data".
- Column Renaming: Map and rename columns in the newer dataset to match those in the older dataset, ensuring uniformity in data structure

Now the dataset contains (2,970,539 , 28) The columns of the dataset include various attributes such as:

`DR_NO` Division of Records Number: Official file number made up of a 2 digit year, area ID, and 5 digits.. (type: int)
`Date Rptd` The date when the crime was reported. (type: datetime64[ns])
`DATE OCC` The date when the crime occurred. (type: datetime64[ns])
`TIME OCC` The time when the crime occurred in 24 hour military time. (type: datetime64[ns])
`AREA` The LAPD has 21 Community Police Stations referred to as Geographic Areas within the department. These Geographic Areas are sequentially numbered from 1-21. (type: int)
`AREA NAME` The 21 Geographic Areas or Patrol Divisions are also given a name designation that references a landmark or the sur-

rounding community that it is responsible for. For example 77th Street Division is located at the intersection of South Broadway and 77th Street, serving neighborhoods in South Los Angeles (type: object)

`Rpt Dist No` A four-digit code that represents a sub-area within a Geographic Area. All crime records reference the "RD" that it occurred in for statistical comparisons. (type: int)

`Part 1-2` The part number of the crime incident. (type: int)

2 types: Part 1 (Serious Crime) and Part 2 (Less Serious Crime)

`Crm Cd` The code corresponding to the crime committed. (type: int)

`Crm Cd Desc` The description of the crime code. (type: object)

`Mocodes` Modus Operandi (method of operation):Activities associated with the suspect in commission of the crime. (type: object)

`Vict Age` The age of the victim. (type: int)

`Vict Sex` The gender of the victim. (type: object)

`Vict Descent` The ethnic descent of the victim. (type: object) Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian

`Premis Cd` The type of structure, vehicle, or location where the crime took place. (type: object)

`Premis Desc` The description of the premises where the crime occurred. (type: object)

`Weapon Used Cd` The code representing the weapon used in the crime (if applicable). (type: object)

`Weapon Desc` The description of the weapon used in the crime (if applicable). (type: object)

`Status` The status of the case. (type: object)

`Status Desc` The description of the status of the case. (type: object)

`Crm Cd 1` Indicates the crime committed. Crime Code 1 is the primary and most serious one. Crime Code 2, 3, and 4 are respectively less serious offenses. Lower crime class numbers are more serious.(if applicable). (type: object)
`Crm Cd 2` May contain a code for an additional crime, less serious than Crime Code 1. (type: object)
`Crm Cd 3` May contain a code for an additional crime, less serious than Crime Code 1. (type: object)
`Crm Cd 4` May contain a code for an additional crime, less serious than Crime Code 1. (type: object)
`LOCATION` Street address of crime incident rounded to the nearest hundred block to maintain anonymity.(type: object)
`Cross Street` The cross street of rounded address (if applicable). (type: object)
`LAT` The latitude coordinate of the crime incident location. (type: float64)
`LON` The longitude coordinate of the crime incident location. (type: float64)
The merged dataset has a shape of (2,970,539, 28), indicating that it contains 2,970,539 records and 28 columns

4.3 Data Preprocessing

- Date Conversion: Convert the date columns (Date Rptd and DATE OCC) to datetime objects for accurate temporal analysis.
- Handle missing values appropriately by either dropping or imputing them
- Feature Engineering: Extract year, month, and hour from the DATE OCC column to facilitate time series analysis.
- Handling Missing Values: Remove rows with missing latitude (LAT) or longitude (LON) values to ensure data quality.Replacing nan values with 0

- Aggregation: Aggregate data by AREA to compute the mean latitude and longitude, and count the number of occurrences (Counts) for better plotting of hotspot.
- Spatial Data Preparation: Convert the aggregated DataFrame into a GeoDataFrame to enable spatial analysis.

4.4 EDA

EDA stands for Exploratory Data Analysis. It is an approach to analyzing and summarizing the main characteristics of a dataset in order to gain insights and better understand the data. EDA involves various techniques and visualizations to explore the data, identify patterns, relationships, and outliers, and generate hypotheses for further analysis.

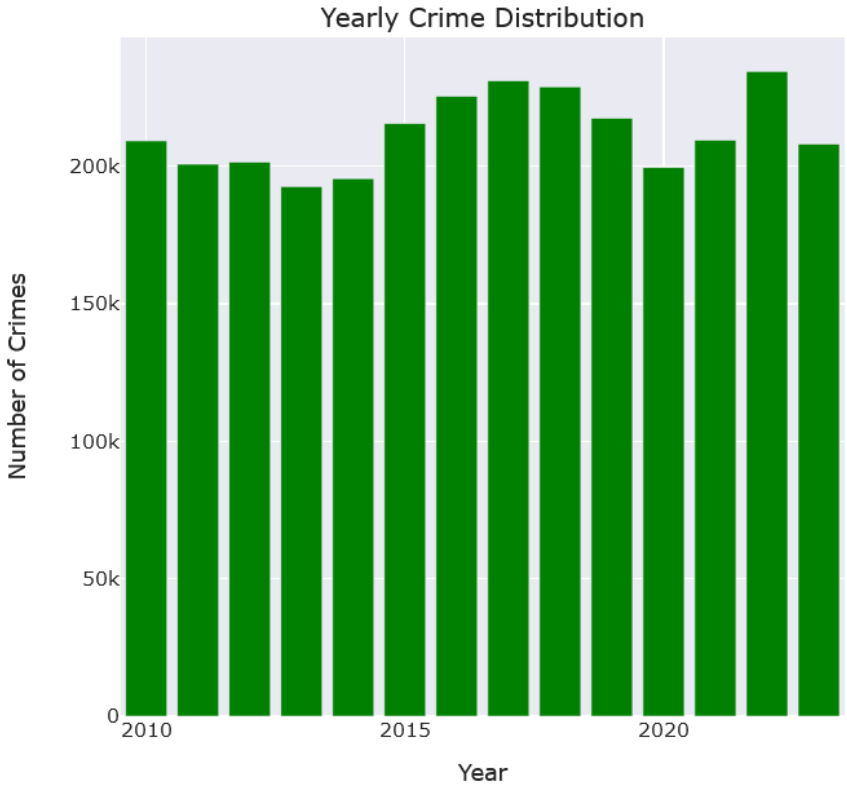


Figure 4.2: Crime Count : Year

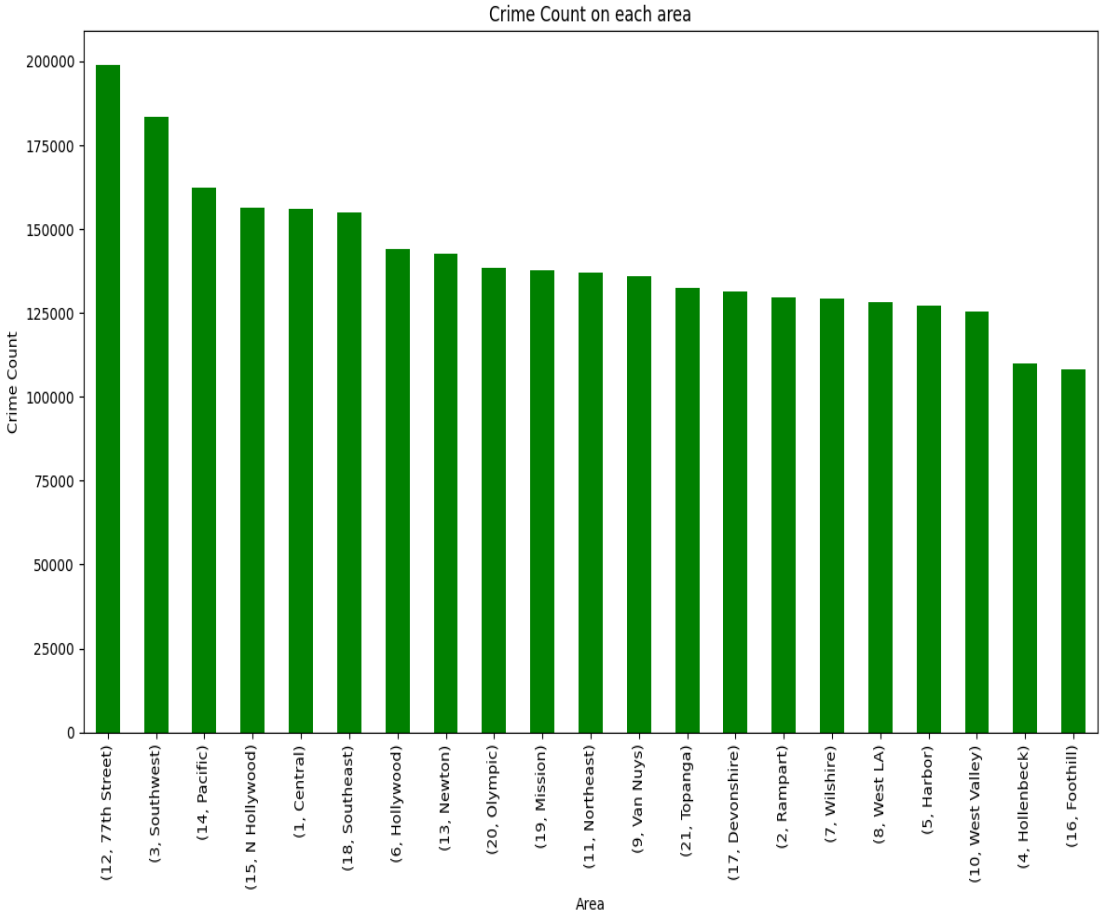


Figure 4.3: Crime Count : Area

4.5 Spatial Analysis

Spatial analysis in this context refers to the process of examining the geographical patterns of crime incidents to identify areas with high concentrations of criminal activities. This type of analysis leverages spatial statistics and geographic information systems (GIS) to detect and understand the spatial distribution and clustering of crimes.

- Create Spatial Weights Matrix: Use the K-nearest neighbors (KNN) method to define spatial relationships between geographic points.
- Calculate Getis-Ord Gi* Statistic: Compute the Getis-Ord Gi* statistic to identify spatial clusters of high or low crime rates.

Getis-Ord Gi* Statistics

The Hot Spot Analysis tool calculates the Getis-Ord Gi* statistic for each feature in a dataset. The resultant z-scores and p-values tell you where features with either high or low values cluster spatially. This tool works by looking at each feature within the context of neighboring features. A feature with a high value is interesting but may not be a statistically significant hot spot. To be a statistically significant hot spot, a feature will have a high value and be surrounded by other features with high values as well. The local sum for a feature and its neighbors is compared proportionally to the sum of all features; when the local sum is very different from the expected local sum, and when that difference is too large to be the result of random chance, a statistically significant z-score results. When the FDR correction is applied, statistical significance is adjusted to account for multiple testing and spatial dependency.

The Getis-Ord Gi* statistic is used to identify clusters of high or low values in spatial data. The statistic for a point i is given by:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij}x_j - \bar{X}\sum_{j=1}^{n} w_{ij}}{S\sqrt{\frac{[n\sum_{j=1}^{n} w_{ij}^2 - (\sum_{j=1}^{n} w_{ij})^2]}{n-1}}}$$

where:

- x_j is the attribute value at location j,
- w_{ij} is the spatial weight between locations i and j,
- $\bar{X}$ is the mean of the attribute values,
- S is the standard deviation of the attribute values,
- n is the number of locations.

The result of the Getis-Ord Gi* statistic for each point indicates

the intensity of clustering for high or low values. Positive values of G_i^* indicate clustering of high values, while negative values indicate clustering of low values.

- Visualization:The spatial distribution of crime hotspots is visualized using Plotly's scatter plot with mapbox, where each point represents an area, and the color of the points reflects the Gi* statistic values. A continuous color scale (Jet) is used to indicate the intensity of the hotspots and cold spots.The hover_name attribute is set to display the area name, and the size of the points is determined by the crime counts.

AREA	**AREA NAME**
1.0	Central
2.0	Rampart
3.0	Southwest
4.0	Hollenbeck
5.0	Harbor
6.0	Hollywood
7.0	Wilshire
8.0	West LA
9.0	Van Nuys
10.0	West Valley
11.0	Northeast
12.0	77th Street
13.0	Newton
14.0	Pacific
15.0	N Hollywood
16.0	Foothill
17.0	Devonshire
18.0	Southeast
19.0	Mission
20.0	Olympic
21.0	Topanga

Table 4.1: List of Areas and their Names

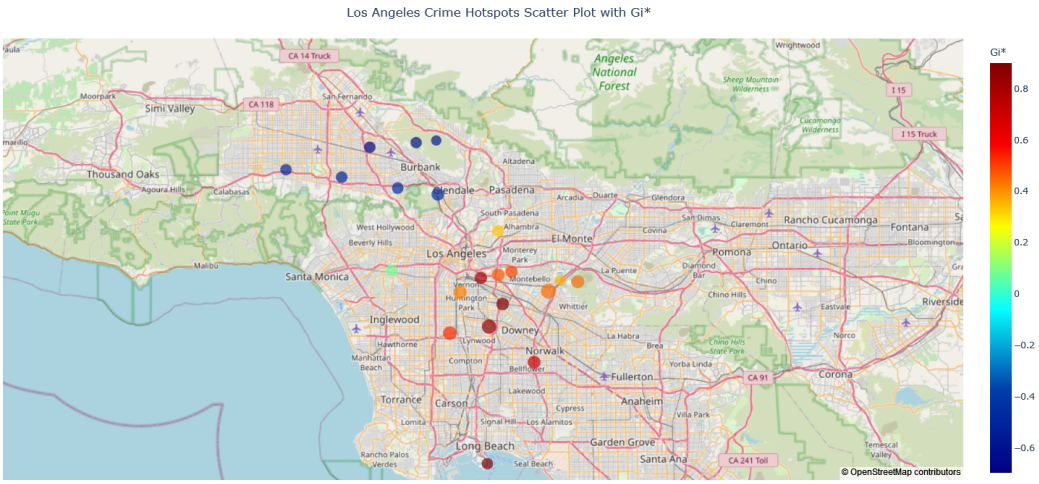


Figure 4.4: Spatial Analysis using Getis-Ord Gi* statistics

(Note:The above plot is an interactive plot which can be viewed correctly in soft copy.)

From the plot, it is evident that areas 5, 12, 13, and 7 are prone to crimes, while area 8 shows a moderate crime count. Specifically, area 5, with a crime count of 36,450, is identified as a hotspot, whereas area 15, with a crime count of 43,908, is classified as a cold spot. This suggests that the statistical method used does not rely solely on crime counts to identify hotspots. Instead, it considers multiple variables, assigning appropriate spatial weights to classify areas into hotspots and cold spots.

4.6 Temporal Analysis and Forecasting

- Data Aggregation by Month: The dataset is aggregated by year and month to count the number of crimes for each month. This aggregation transforms the data into a time series format, suitable for time series analysis and forecasting.

- **Prophet Model**

  Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly,

weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.Prophet is open source software released by Facebook's Core Data Science team.Prophet also imposes the strict condition that the input columns must be named as ds (the time column) and y (the metric column).

In this project, the monthly data is reformatted to suit the Prophet model requirements. The date column is renamed to 'ds' and the counts column to 'y'.

A Prophet model is fitted to the monthly crime counts. Prophet is a forecasting tool specifically designed to handle time series data with strong seasonal effects and missing data.

The Prophet model is used to forecast crime counts for the next 12 months. The forecasted values ($\hat{y}$), including the lower($\hat{y}$ lower) and upper ($\hat{y}$ upper)confidence intervals, are printed and plotted to visualize the predicted crime counts.

- Visualization:The forecasts from Prophet models are visualized, showing the observed crime counts, forecasted values, and confidence intervals. Prophet plots the observed values of our time series (the black dots), the forecasted values (blue line) and the uncertainty intervals of our forecasts (the blue shaded regions).This visualization helps in understanding the trends and patterns in crime data and the accuracy of the forecasts.

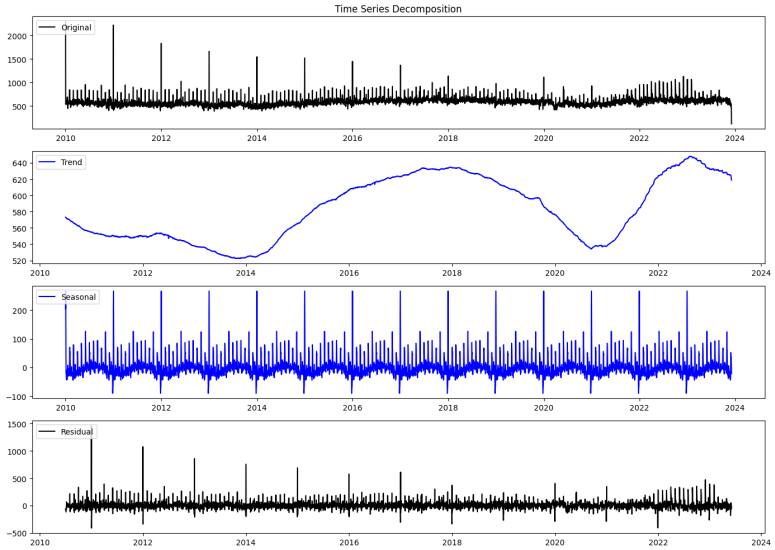


Figure 4.5: Components of Time Series

Table 4.2: Prophet Forecast Values

Index	Date	$\hat{y}$	$\hat{y}$ lower	$\hat{y}$ upper
71	2023-11-30	13850.96	10649.03	17251.14
72	2023-12-31	36214.15	32885.15	39424.34
73	2024-01-31	33444.29	30066.32	36672.29
74	2024-02-29	21163.39	17889.30	24616.83
75	2024-03-31	22643.39	19245.56	26142.99
76	2024-04-30	23518.91	20322.32	26709.54
77	2024-05-31	24680.69	21356.10	28011.47
78	2024-06-30	25720.14	22415.36	29067.79
79	2024-07-31	25470.37	22089.16	28727.83
80	2024-08-31	24678.28	21267.71	28045.21
81	2024-09-30	23318.36	19848.32	26490.16
82	2024-10-31	20647.33	17418.90	23980.28

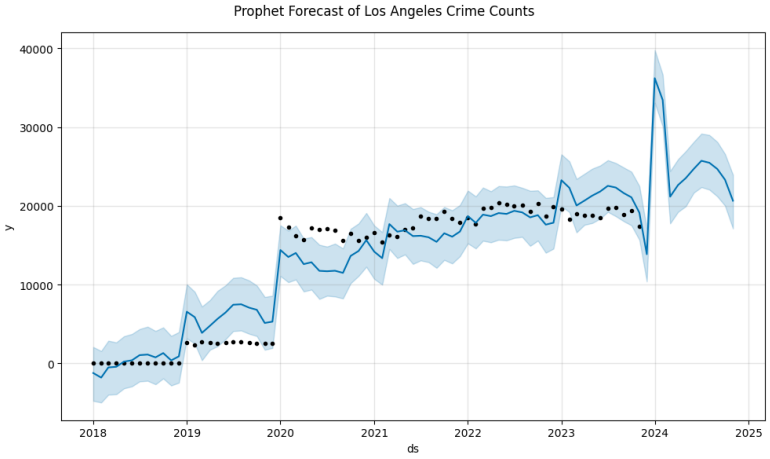


Figure 4.6: Prophet forecasting

In Prophet forecasting, the model sometimes limits the historical data shown in the forecast plot (like in Figure 4.5) to maintain clarity and focus on the forecast period. Here are a few reasons why this might happen:

- Default Plot Settings: Prophet's default plotting settings may truncate historical data to make the forecast period more prominent and to avoid cluttering the plot with excessive historical data.
- Forecast Horizon: The plot is designed to highlight the forecasted future period, which can lead to a reduction in the amount of historical data displayed, especially if the forecast horizon is long relative to the historical data.
- Data Density: If you have a large amount of historical data, displaying all of it on the plot might make it less readable. The plot might then automatically adjust the visible historical range to keep the forecast results clear and interpretable.
- Plot Size: The size of the plot can also impact how much historical data is shown. If the plot area is constrained, it might limit the visible historical data to fit the plot space.

  Prophet RMSE: 2610.048

  The RMSE for the Prophet model is 2610.048. This value indicates the average deviation between the predicted and actual crime counts. A lower RMSE value would signify a better fit, meaning the predictions made by the Prophet model are closer to the actual observations.

  The Prophet model, with an RMSE of 2610.048, provides a quantitative measure of its forecasting accuracy for crime trends in Los Angeles. This metric helps in understanding the reliability of the model's predictions, aiding in better decision-making and strategic planning for crime prevention and resource management.

CHAPTER 5

Conclusion and Future Work

In this project, a comprehensive analysis of crime data from Los Angeles, spanning January 2010 to November 2023, was conducted to understand crime patterns and forecast future trends.

The hotspot analysis (Figure 4.4) revealed that areas 5, 12, 13, and 7 are particularly prone to crimes, while area 8 has a moderate crime count. Notably, area 5, with a crime count of 36,450, is identified as a hotspot, whereas area 15, with a crime count of 43,908, is classified as a cold spot. This indicates that the statistical method used for hotspot detection does not rely solely on crime counts but considers multiple variables and assigns appropriate spatial weights to classify areas into hotspots and cold spots.

The Prophet model (Figure 4.5, Table 4.2, Figure 4.6) provided robust forecasts with uncertainty intervals, effectively handling seasonality and trend changes. It offered a clear view of expected future crime trends and demonstrated flexibility in modeling seasonal effects and holiday impacts. This comprehensive approach to analyzing and forecasting crime data can significantly aid in strategic planning and resource allocation for law enforcement agencies.

Future Work

Based on the conclusions, several areas for future work can be identified to enhance the analysis and its applications:

- Incorporate Additional Variables:

  Socioeconomic Factors: Integrate socioeconomic variables (e.g., unemployment rates, education levels) to better understand the factors influencing crime rates.

  Weather Data: Include weather conditions to analyze their impact on crime patterns.

- Refine Spatial Analysis:

  Dynamic Spatial Models: Explore advanced spatial models that account for changes over time and better capture dynamic patterns of crime.

  Higher Resolution Data: Use higher resolution spatial data to refine hotspot detection and improve the accuracy of crime mapping.

- Enhance Forecasting Models:

  Hybrid Models: Combine ARIMA, Prophet, and machine learning models (e.g., LSTM, XGBoost) to improve forecast accuracy and capture complex patterns.

  Real-time Data Integration: Implement real-time data streaming for dynamic forecasting and immediate response to emerging crime trends.

- Policy and Intervention Analysis:

  Impact Evaluation: Assess the effectiveness of interventions and policy changes based on the forecasted crime trends and observed outcomes.

Resource Allocation: Develop optimized strategies for resource allocation and crime prevention based on the insights from spatial and temporal analyses. Public Engagement:

- Visualization Tools: Create interactive dashboards and tools for public and policy makers to visualize crime trends and hotspots effectively.

  Community Outreach: Engage with local communities to understand their perceptions of crime trends and incorporate their feedback into the analysis.

- Longitudinal Studies:

  Extended Time Frames: Extend the analysis to cover more recent data beyond June 2021 to capture ongoing trends and refine forecasting models.

  Comparative Studies: Conduct comparative studies with other cities or regions to identify unique patterns and factors influencing crime.

By addressing these areas, the analysis can be further improved, leading to more accurate predictions, better resource allocation, and more effective crime prevention strategies.

Bibliography

- https://facebook.github.io/prophet/
- https://www.kaggle.com/code/prashant111/tutorial-time-series-forecasting-with-prophet#3.-Installation-of-Prophet-
- *Introduction to Time Series and Forecasting*, Second Edition,Peter J. Brockwell & Richard A. Davis

  http://home.iitj.ac.in/~parmod/document/introduction%20time%20series.pdf
- Yu, H.; Liu, L.; Yang, B.; Lan, M. *Crime Prediction with Historical Crime and Movement Data of Potential Offenders Using a Spatio-Temporal Cokriging Method.* ISPRS Int. J. Geo-Inf. 2020, 9, 732.

  https://doi.org/10.3390/ijgi9120732
- U. M. Butt, S. Letchmunan, F. H. Hassan, M. Ali, A. Baqir and H. H. R. Sherazi, *"Spatio-Temporal Crime HotSpot Detection and Prediction: A Systematic Literature Review,"* in IEEE Access, vol. 8, pp. 166553-166574, 2020, doi: 10.1109/ACCESS.2020.3022808.

  https://ieeexplore.ieee.org/stamp/stamp.jsptp=&arnumber=9187772&isnumber=8948470
- Xu Zhang, Lin Liu, Minxuan Lan, Guangwen Song, Luzi Xiao, Jianguo Chen,*Interpretable machine learning models for crime prediction,* Computers, Environment and Urban Systems, Volume 94,2022,101789,ISSN 0198-9715.

  https://www.sciencedirect.com/science/article/pii/S0198971522000333

APPENDIX A

Appendix

- **DATASET**

  https://www.kaggle.com/datasets/sumaiaparveenshupti/los-angeles-crime-data-20102020/data

  https://www.kaggle.com/datasets/sahirmaharajj/crime-data-from-2020-to-present-updated-monthly/data

- **PYTHON CODE**
  - PREPROCESSING

    https://colab.research.google.com/drive/1iP3lPORsHUN5FVnaxUFTdMUy7B7jF-4x?usp=sharing

  - ANALYSIS

    https://colab.research.google.com/drive/1tQjfaTP_aX53l2R-S_6SFechWEB1wLLL?usp=sharing