Predictive Policing: Utilizing Machine Learning Algorithms To Enhance Crime Prevention Strategies

Hanna Bodnar
hannabodnar@gatech.edu

April 28, 2024

1 Problem Statement

Crime prediction and prevention have long been focal points in enhancing public safety and law enforcement strategies. In this project, I delve into the realm of criminology by leveraging machine learning algorithms to analyze and predict crime patterns. By harnessing the power of data-driven insights, I aim to contribute to the development of proactive measures that aid law enforcement agencies in allocating resources efficiently and mitigating criminal activities. Through the utilization of advanced techniques, including feature engineering, model selection and evaluation, I explore the potential of predictive analytics in guiding strategic decision-making for crime prevention efforts. This report presents my methodology, findings and implications, showcasing the transformative potential of machine learning in addressing contemporary challenges in criminology and public safety.

2 Data Collection and Preprocessing

2.1 Data Source

The dataset used in this analysis was obtained from the Los Angeles Police Department's (LAPD) Crime Data from 2020 to Present repository on the Los Angeles Open Data Portal[1]. This dataset provides detailed information on reported crimes in the city of Los Angeles, including type of crime, location, date, and time of occurrence.

2.2 Data Description

The dataset consists of 932,000 observations and 28 variables. Key variables include "Date Occurred", "Time Occurred", "Type of Crime", "Area", "Location", "Weapon Used", and "Victim Age". The data contains both categorical (e.g., crime type, location) and numeric (e.g., report district) variables.

2.3 Data Cleansing

Prior to analysis, the data was thoroughly cleansed to address missing values, and outliers. Out of the 28 variables, 8 were removed due to the vast amount of missing information that could not be imputed. For the remaining data, missing values were either declared as "Unknown" for categorical variables, or "0" for numeric variables. To streamline the dataset and reduce complexity, premise description and crime descriptions were aggregated under broader categories. This aggregation process involved grouping similar descriptions into more generalized types, facilitating a more concise representation of the data for analysis.

2.4 Feature Engineering

Feature engineering was performed to extract additional information from the raw data and create new variables. For example, the "Date Occurred" variable was dissected to create individual variables for year, month and day of the crime. Additionally, categorical variables such as "Victim Sex", and "Victim Descent" were encoded using label-encoding to prepare them for modeling.

3 Methodology

My approach comprises two key components: exploratory data analysis (EDA) and model development. Through EDA, I comprehensively examined the crime dataset to gain insights into the spatial and temporal distribution of crime incidents, identify potential patterns or trends, and inform subsequent modeling efforts. Subsequently, I employed advanced modeling techniques to develop predictive models capable of forecasting future crime occurrences and identifying factors contributing to crime risk. By integrating EDA and model development, I aimed to uncover underlying patterns and relationships in the data, improve my understanding of crime dynamics, and generate actionable insights to support evidence-based decision-making in crime prevention and law enforcement efforts.

3.1 Exploratory Data Analysis

Before delving into the analysis and visualizations, I conducted Exploratory Data Analysis (EDA) to gain a preliminary understanding of the dataset. EDA involves examining the structure, characteristics, and relationships within the data to identify patterns, anomalies, and potential insights. Through EDA, I explored various summary statistics, distributions, and correlations, laying the groundwork for our subsequent analyses and informing our approach to data exploration.

Table 1: Most Crimes Committed

	Crime Desc	Count
0	VEHICLE - STOLEN	94544
1	BATTERY - SIMPLE ASSAULT	70096
2	BURGLARY FROM VEHICLE	54108
3	BURGLARY	53824
4	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)	52894
5	THEFT OF IDENTITY	50435
6	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	49516
7	THEFT PLAIN - PETTY ($950 & UNDER)	45340
8	INTIMATE PARTNER - SIMPLE ASSAULT	42899
9	THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)	34144
10	THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER)	31479
11	ROBBERY	29958
12	THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD	29333
13	SHOPLIFTING - PETTY THEFT ($950 & UNDER)	23423
14	VANDALISM - MISDEAMEANOR ($399 OR UNDER)	22700

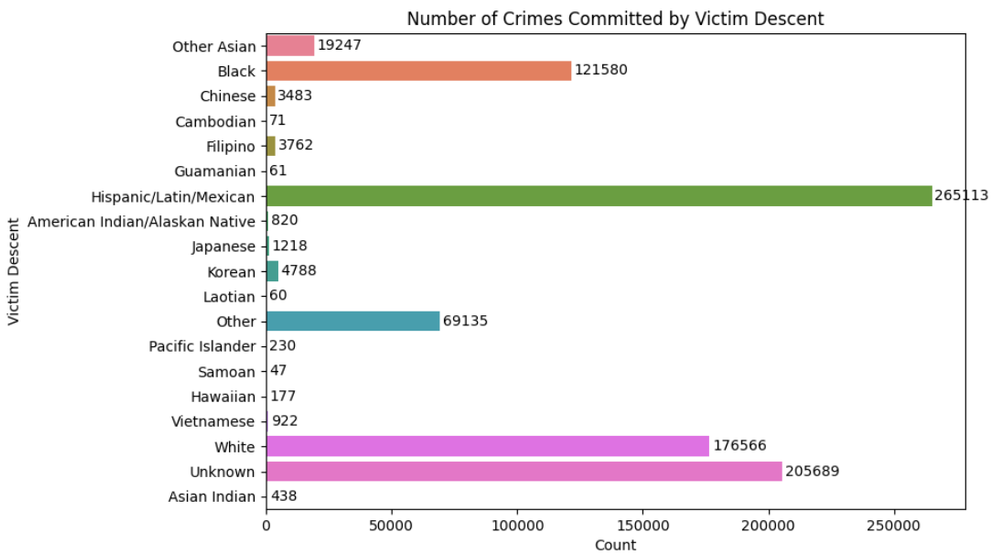


Figure 1: Victim Descent

Visualizing the distribution of descent among victims can illuminate disparities in crime victimization across different demographic groups. Based on Figure 1, individuals of Hispanic, Latin, or

Mexican descent have the most risk of being a victim of crime. Aside from "Unknown" individuals who are more at risk are of white and black descent, respectively. According to the US Census Bureau, the Hispanic population makes 48.4%, followed by White at 28.1% and Black at 8.6%.

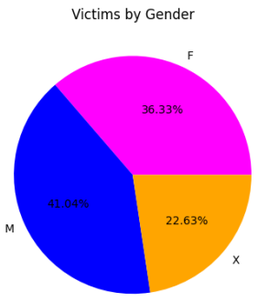


Figure 2: Sex Distribution

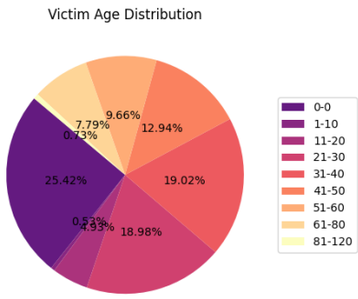


Figure 3: Age Distribution

In the analysis of victim demographics, the data reveals a predominance of male victims across various crime types, surpassing female victims and those of unknown sex. This observation underscores potential gender-specific patterns in victimization, suggesting areas where targeted interventions or preventative measures may be warranted. Furthermore, the victim age distribution exhibits the age range of 31-40 as having the highest percentage of victims after those categorized under "unknown" age. This may indicate a demographic segment that is disproportionately affected by crime incidents. Understanding these demographic nuances is pivotal, as it informs the development of tailored strategies to address the specific vulnerabilities and risk factors associated with different demographic groups.

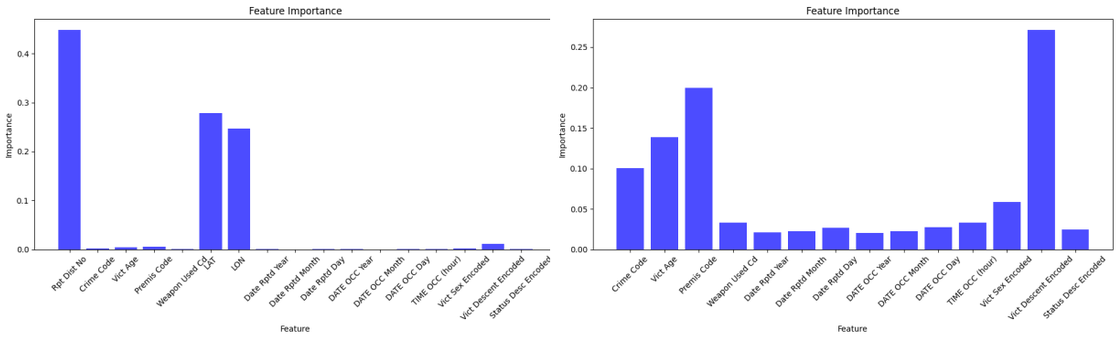
3.2 Model Development and Evaluation

The preceding subsections delve into the development and evaluation of various predictive models. Each subsection is dedicated to a different modeling approach, including Random Forest, spatial analysis, and time-series modeling. By exploring multiple models, I aim to gain a comprehensive understanding of the spatial, temporal, and predictive dynamics underlying criminal activity. This understanding will enable us to derive actionable insights for enhancing public safety and addressing crime-related challenges.

3.2.1 Random Forest

In the context of this analysis, the Random Forest model serves as a robust tool for predicting "Area Code" based on a set of input features derived from reported crime data. By leveraging a diverse set of predictors such as "Date Rptd", "TIME OCC", "DATE OCC", "Rpt Dist No", "Crime Code", "Vict Age", "Vict Sex", "Vict Descent", "Premise Code", "Weapon Used Cd", "Status Desc", "LAT", and "LON", the model aims to accurately classify the geographical area associated with each reported crime incident. The Random Forest algorithm offers several advantages, including its ability to handle large and high-dimensional datasets, handle both categorical and numerical variables, and provide insights into feature importance for interpretability. As such, it represents a valuable tool for exploring the underlying patterns and relationships within the crime data.

The model was trained and evaluated systematically to ensure robust performance and reliable predictions. First, the data was split into training and testing sets, with 70% of the data allocated for training and 30% reserved for evaluation. To optimize performance, hyper-parameters such as number of trees, maximum depth of trees, and minimum number of samples per leaf were tuned using a random grid search approach. This involved systematically searching through a predefined hyper-parameter space and selecting the combination that maximized model performance, as assessed by cross-validation. Regarding cross-validation, 5-fold cross-validation was applied to evaluate the model's generalization performance. The data was partitioned into five subsets; each was used as a validation set, and the model was trained on the remaining data. This process was repeated 5-times with performance metrics averaged across all folds to provide a robust estimate of the model's performance.



(a) All Geographic Features

(b) No Geographic Features

Figure 4: Model Feature Importance Based on Inclusion of Geographic Features

Model	Accuracy
All Features	99.77%
W/out District Code	97.65%
W/out District Code, Longitude and Latitude Coordinates	18.24%

Table 2: Model Accuracy Based on Geographic Features

The results of the Random Forest models for predicting "Area Code" reveal insights into the importance of different variables and their impact on model performance. The model trained on all variables achieved the highest accuracy, suggesting that the combination of predictors provides a robust representation of the data and effectively captures the underlying patterns related to "Area Code" prediction. The model trained on all the variables except "Reported District Code" achieved a slightly lower accuracy compared to the model with all variables. This suggests that "Reported District Code" may not be as influential in predicting "Area Code" as other variables. However, the high accuracy indicates that the remaining predictors still provide valuable information to predict "Area Code" accurately. In contrast, the model trained without any geographic variables exhibited a significant drop in accuracy to 18%. The dramatic decrease in accuracy suggests that these variables play a crucial role in predicting "Area Code". The "Latitude" and "Longitude" coordinates are fundamental in determining the location and, consequently, the "Area Code" associated with a reported crime.

3.2.2 Spatial Autocorrelation Analysis

Spatial Autocorrelation refers to the degree of similarity between neighboring spatial units (e.g., geographic regions, points) concerning a particular variable of interest. It is a fundamental concept in spatial analysis that helps assess the spatial dependence and clustering patterns in geographical data. In my analysis, spatial autocorrelation plays a crucial role in understanding the spatial distribution of crime incidents and identifying clusters or hotspots of criminal activity within the study area. By examining the spatial relationship between crime occurrences across different geographical locations, we can uncover spatial patterns that may need to be evident through traditional statistical methods. Spatial Analysis provides insights into the underlying spatial processes driving crime dynamics, such as the presence of spatial clusters or spatial outliers. It helps us identify areas with familiar crime rates or characteristics, enabling law enforcement agencies and policymakers to prioritize resources and interventions more effectively.

In preparation for spatial analysis, a series of preprocessing steps were taken to manipulate geographic data and establish spatial relationships between crime incidents and LAPD police stations. Geopandas, a Python library for geographic manipulation, was initially used to handle spatial data and perform spatial operations. Within this framework, latitude and longitude coordinates from crime incidents were stored as geometry points, representing spatial locations within a geographic coordinate system. Simultaneously, geographic data for LAPD police stations was obtained from GeoHub L.A. City, a comprehensive platform offering access to various spatial datasets for Los Angeles. The dataset was then imported into Geopandas for further processing, facilitating the extraction of coordinates for each police station. Subsequently, the Euclidean distance between each crime incident and the nearby police station was calculated using Geopanda's spatial operations. Doing so allowed for the aggregation of crime incidents by the closest police station, offering insights into the spatial distribution of law enforcement coverage. To establish spatial relationships

between police stations, K-Nearest Neighbors (KNN) spatial weights were computed based on the proximity of police stations to one another. By assigning weights to neighboring police stations according to their distance, this approach facilitated the computation of Moran's statistic and the identification of spatial clustering patterns.

Moran's scatterplot visualizes the relationship between the values of the variable of interest for each spatial unit and the corresponding spatial lag, which represents the average value of the variable among neighboring spatial units. The scatterplot is divided into four quadrants based on the positive or negative values of the variable and its spatial lag. In the upper-right and lower-left quadrants, where high and low values of the variable tend to be surrounded by similar values, positive spatial autocorrelation (clustering) is indicated. Conversely, in the upper-left and lower-right quadrants, where high values are surrounded by low values and vice versa, negative spatial autocorrelation (dispersion) is indicated. Moran's I, a statistic quantifying the strength and direction of spatial autocorrelation, complements the scatterplot interpretation. Positive Moran's I values accompany clustering, while negative values indicate dispersion. A value close to zero suggests spatial randomness.

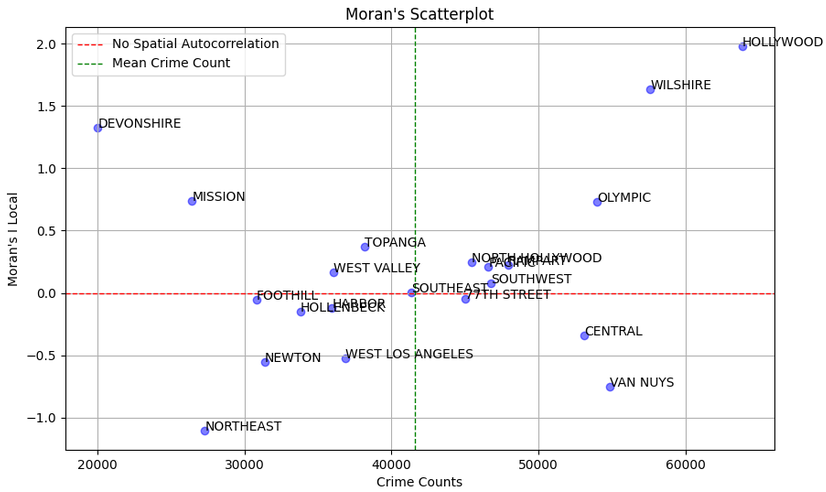


Figure 5: Moran's Scatterplot for Spatial Autocorrelation

Based on the scatterplot in Figure 5, LAPD divisions in the upper-right quadrant, including Hollywood, Wilshire, Olympic, North Hollywood, Pacific, and Rampart, exhibit negative spatial autocorrelation, suggesting high crime rates within these divisions but with neighboring divisions displaying relatively lower crime rates. Conversely, divisions such as Central and Van Nuys, located in the lower-right quadrant, demonstrate positive spatial autocorrelation, indicating lower crime rates surrounding similarly low-crime neighboring divisions. This clustering pattern implies potential areas with effective crime prevention measures or community interventions. Divisions in the lower-left quadrant, including Northeast, Newton, and West LA, also exhibit negative spatial autocorrelation, signifying high crime rates within the divisions amidst lower-crime neighboring areas. Meanwhile, divisions in the upper-right quadrant, such as Devonshire, Mission, Topanga, and West Valley, display positive spatial autocorrelation, indicating lower crime rates surrounded by neighboring divisions with similarly low crime rates, suggesting areas of effective law enforcement strategies. Notably, divisions positioned near the y=0 axis, such as Foothill, Harbor, Hollenbeer, Southeast, and 77th Street, reflect a near absence of spatial autocorrelation, suggesting spatial randomness or heterogeneity in crime patterns within these divisions.

To assess the statistical significance of the observed spatial autocorrelation patterns identified in the Moran scatterplot, significance testing was conducted using a significance threshold of 0.05. This threshold indicates that any Moran's I values exceeding the critical threshold at the 0.05 significance level are considered statistically significant, suggesting non-random spatial patterns in the distribution of crime incidents. The results of the significance testing reveal that several LAPD districts exhibit statistically significant spatial autocorrelation in their crime rates. Specifically, districts including Wilshire, Hollywood, Topanga, and West LA surpass the significance threshold of 0.05, indicating robust spatial clustering or dispersion patterns in crime incidents within these areas.

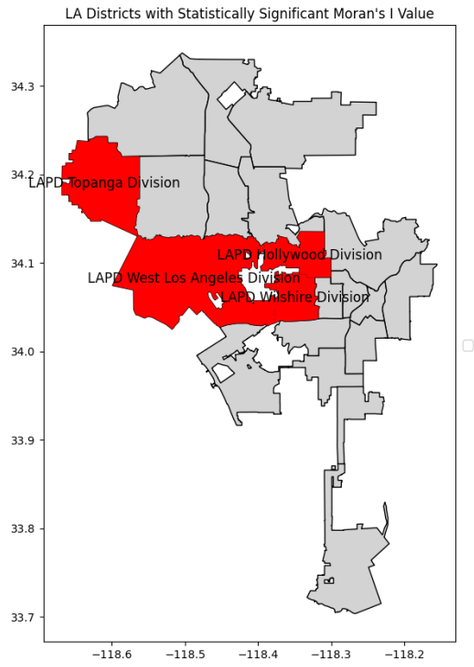


Figure 6: LA Districts with Statistical Significance (P-Value < 0.05)

Identifying Wilshire, Hollywood, Topanga, and West LA as districts with statistically significant spatial autocorrelation in crime rates holds important implications for crime prevention and law enforcement strategies. These districts may be focal points for targeted interventions addressing localized crime hotspots or vulnerabilities. For instance, in districts exhibiting positive spatial autocorrelation (clustering), such as Wilshire and Hollywood, concentrated efforts could be directed towards implementing community policing initiatives, enhancing neighborhood watch programs, and improving environmental design to deter criminal activities and promote community safety. Conversely, in districts demonstrating negative spatial autocorrelation (dispersion) like Topanga and West LA, strategies may focus on identifying and addressing underlying socio-economic factors contributing to crime dispersal and ensuring equitable access to law enforcement resources across diverse neighborhoods.

3.2.3 Time-Series Analysis

Time series analysis plays a pivotal role in understanding the temporal dynamics of crime data and uncovering patterns, trends, and seasonality in criminal activities. In this project, time series analysis serves multiple purposes, each contributing to a comprehensive understanding of crime patterns and informing evidence-based decision-making in crime prevention and law enforcement efforts.

The transformation of raw crime data into a structured weekly time series involved several sequential steps to facilitate comprehensive analysis and modeling. Beginning with data preparation, where raw crime data was refined to ensure consistency and completeness, I extracted date information from each incident. Subsequently, the entire period covered by the dataset was partitioned into discrete weekly intervals, enabling me to capture temporal trends with granularity. Within these intervals, the counts of crime incidents were aggregated, consolidating individual occurrences into weekly sums. This aggregation process provided a concise representation of crime activity over time. Finally, I organized these aggregated counts into a structured time series format and established a valuable analytical framework for exploring temporal patterns, detecting trends, and forecasting future occurrences.

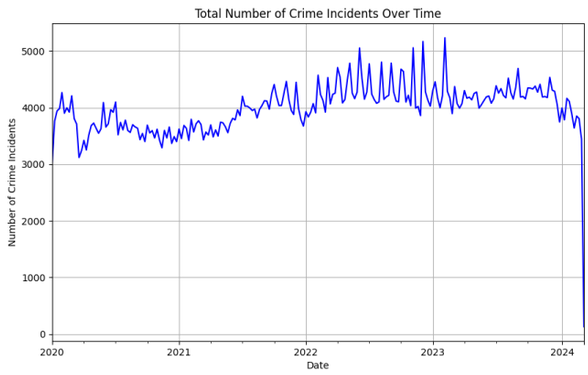


Figure 7: Number of Crimes Over Time

The time series plot in Figure 7 reveals a consistent number of crimes ranging between 3000 and 5000 over time, underscoring the stability of criminal activity within the observed period. This temporal consistency suggests a resilience in crime patterns, with no discernible upward or downward trend evident. The stable baseline of crime incidents serves as a reference point for understanding typical levels of criminal activity, informing resource allocation strategies, and setting realistic targets for crime reduction initiatives. While reflecting a stable crime environment, this insight also highlights the need for ongoing efforts to address underlying socio-economic factors and strengthen community partnerships to promote public safety and enhance community well-being.

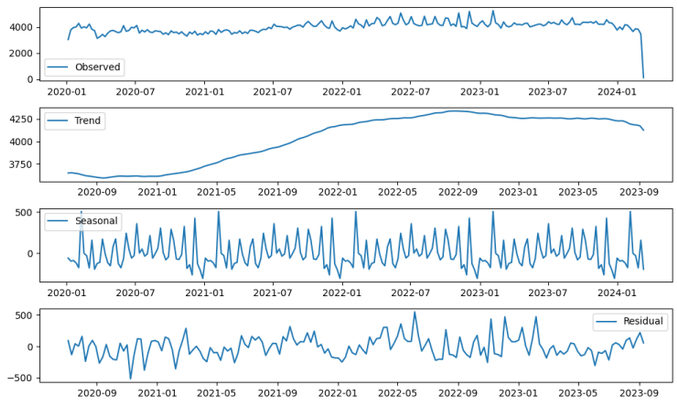


Figure 8: Seasonal Decomposition of Time-Series

The seasonal decomposition analysis, with the y-value representing the number of reported crimes, shows compelling insights into the temporal dynamics of criminal activity. Firstly, the consistent increase in the magnitude of the observed values between 2022 and 2023 signifies a sustained upward trend in crime rates over this period. This trend suggests a concerning escalation in criminal activity, indicative of potential societal challenges or shifts in underlying socio-economic factors. Additionally, the identification of a yearly seasonal pattern underscores the presence of recurring fluctuations in crime rates tied to specific times of the year, such as seasonal changes, holidays, or other temporal factors influencing criminal behavior. Moreover, an increasing trend component highlights a systematic rise in crime rates over time, indicative of underlying societal or structural factors contributing to long-term growth in criminal activity.

The time series in Figure 9 regarding victim descent reveals compelling insights into the dynamics of crime victimization across different demographic groups. Notably, between 2022 and 2023, significant increases in the magnitude of reported crimes are observed, particularly among individuals of Hispanic, White, and Black descent. This escalation in reported crimes underscores the disproportionate impact of criminal activity on these demographic groups, highlighting potential disparities in victimization rates and vulnerabilities within the community. Of particular concern is the Hispanic victim descent category, which not only exhibits the most substantial increase in magnitude but also starts above all others, suggesting a higher incidence of crimes targeting individuals of Hispanic descent compared to other demographic groups. This trend emphasizes the urgent need to address specific challenges and vulnerabilities faced by Hispanic communities, such as language barriers, immigration status, or socioeconomic factors contributing to increased crime victimization. Additionally, the observed slight increase in trend for the Unknown victim descent

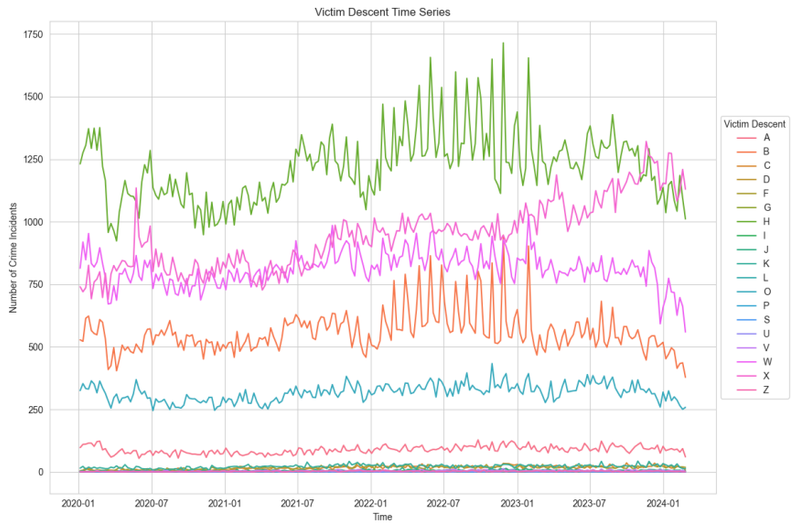


Figure 9: Victim Descent Time-Series

category starting from July 2022 raises questions about the accuracy and completeness of victim information, underscoring the importance of robust data collection and reporting practices to ensure comprehensive and equitable crime reporting.

I employed a rigorous approach to model selection, utilizing grid search and 5-fold cross-validation to identify the optimal parameters for the non-seasonal order and seasonal order parameters in the SARIMA model. Despite these efforts, my attempts with SARIMA modeling yielded very large Mean Squared Error (MSE) values, indicative of poor model performance. Additionally, I explored alternative modeling techniques, including Seasonal-Trend decomposition using LOESS (STL) and exponential smoothing. However, these models also produced exceedingly large MSE values, with exponential smoothing demonstrating the lowest MSE at approximately 51000. It's important to note that the dataset required aggregation to create a time series, which may have excluded other relevant factors and resulted in a smaller dataset. Consequently, the reduced number of observations limited the accuracy of my forecasts and contributed to the large MSE values observed across all modeling approaches.

In light of the challenges encountered with traditional modeling techniques, I adopted an exploratory approach to gain insights into independent factors' temporal patterns and trends. Through time series visualization, I plotted the trends of these factors over time, allowing me to identify any discernible patterns or anomalies. This qualitative analysis provided valuable insights into the dynamics of these independent factors and their potential influence on the observed outcomes. While this approach does not replace quantitative modeling, it complements our analysis by offering a qualitative understanding of the temporal dynamics within the dataset.

4 Discussion

The comprehensive analysis of crime data has yielded valuable insights into the spatial and temporal dynamics of criminal activity within our study area. I have identified significant patterns and trends in crime occurrences through a combination of advanced modeling techniques, including Random Forest, SARIMA, and spatial autocorrelation analysis. Random Forest modeling provided a robust framework for predicting crime occurrences based on diverse predictor variables, enabling me to identify critical factors influencing crime hotspots and spatial distributions. However, despite its predictive capabilities, Random Forest modeling encountered challenges in accurately capturing complex temporal patterns and trends, particularly compared to traditional time series models such as SARIMA. My SARIMA modeling efforts revealed notable increases in the magnitude of reported crimes among Hispanic, White, and Black demographic groups, highlighting potential disparities in victimization rates within the community. Additionally, spatial autocorrelation analysis unveiled spatial clustering and patterns of crime occurrences, indicating the presence of localized hotspots and areas of heightened criminal activity. Despite these significant findings, my analysis has limitations. Data aggregation for time series modeling may have led to the omission of critical variables and reduced the accuracy of our forecasts. Furthermore, spatial autocorrelation analysis is sensitive to the modifiable areal unit problem (MAUP) and the edge effect, which may have influ-

enced our spatial clustering results. Additionally, the reliance on historical crime data introduces inherent biases and limitations, and future research endeavors should explore integrating real-time data and novel data sources to enhance predictive accuracy and inform proactive crime prevention strategies. By addressing these limitations and leveraging interdisciplinary insights, we can advance our understanding of crime dynamics and develop more effective interventions to promote public safety and well-being in our communities.

5 Conclusion

In conclusion, my analysis of crime data has provided valuable insights into the spatial and temporal dynamics of criminal activity in our study area. Through exploratory data analysis, spatial autocorrelation analysis, and time series modeling, I have uncovered patterns, trends, and factors influencing crime occurrences, laying the groundwork for evidence-based decision-making in crime prevention and law enforcement efforts.

My findings underscore the importance of a multifaceted approach to addressing crime, encompassing targeted interventions, community engagement, and resource allocation strategies. Building upon the insights gained from our analysis, I offer the following recommendations for enhancing public safety and reducing crime rates:

1. **Community-Centered Initiatives**: Foster community partnerships and grassroots initiatives aimed at addressing underlying socio-economic factors contributing to crime, promoting neighborhood cohesion, and empowering residents to actively participate in crime prevention efforts.

2. **Targeted Enforcement**: Utilize predictive modeling and spatial analysis techniques to identify high-crime areas and allocate law enforcement resources strategically. Implement targeted patrols, hotspot policing, and proactive interventions to deter criminal activity and enhance public safety in vulnerable communities.

3. **Early Intervention Programs**: Develop and implement early intervention programs targeting at-risk individuals, particularly youth, to steer them away from criminal behavior and provide access to education, employment opportunities, and social support services.

By implementing these recommendations in a coordinated and collaborative manner, we can work towards creating safer and more resilient communities, where all residents can thrive free from the fear of crime. Our commitment to evidence-based strategies, community engagement, and proactive interventions will be instrumental in achieving our shared goal of building a safer and more secure future for all.

References

[1] *Crime Data from 2020 to Present*. en. Apr. 2024. URL: https://catalog.data.gov/dataset/crime-data-from-2020-to-present (visited on 04/11/2024).

[2] Suchismita Sahu. *DECISION BOUNDARY FOR CLASSIFIERS: AN INTRODUCTION*. en. Feb. 2024. URL: https://medium.com/analytics-vidhya/decision-boundary-for-classifiers-an-introduction-cc67c6d3da0e (visited on 03/11/2024).

[3] *6.6 STL decomposition — Forecasting: Principles and Practice (2nd ed)*. URL: https://otexts.com/fpp2/stl.html (visited on 04/27/2024).

[4] *City of Los Angeles Hub*. en-us. URL: https://geohub.lacity.org/ (visited on 04/26/2024).

[5] *Crime Data from 2020 to Present — Los Angeles - Open Data Portal*. URL: https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data (visited on 04/22/2024).

[6] *Los Angeles, CA — Data USA*. en. URL: https://datausa.io/profile/geo/los-angeles-ca (visited on 04/26/2024).

[7] *Spatial Autocorrelation (Global Moran's I) (Spatial Statistics)—ArcGIS Pro — Documentation*. URL: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/spatial-autocorrelation.htm (visited on 04/26/2024).

[8] *Statistical Data*. en-us. URL: https://www.lapdonline.org/statistical-data/ (visited on 04/25/2024).

[9] *U.S. Census Bureau QuickFacts: Los Angeles city, California*. en. URL: https://www.census.gov/quickfacts/fact/table/losangelescitycalifornia/PST045223 (visited on 04/26/2024).

[7] [3] [6] [4] [1] [5] [9] [8] [2]