**Exploring Crime Analysis with LAPD Leveraging Machine Learning for Public Safety**

PROJECT REPORT

Submitted in the partial fulfilment of the requirements for award of the

6 Months Online Certification Course

In

Machine Learning with Python Programming

By

Naveen rampa [21R11A0589]

J Sinduja [21R11A0574]

E Saikiran [22R15A0514]

Lakshanya S [21R11A05H3]

A Bhavya Sri [21R11A05A5]



Geethanjali College of Engineering and Technology

Cheeryal Village, Keesara Mandal, Hyderabad,Telangana 501301

March 2024

**ABSTRACT**

Law and security officials face significant challenges in managing and reducing crime rates. To address this issue, we propose a Crime Analysis and Prediction System that leverages data analytics and machine learning techniques to analyze historical crime data, identify trends, and predict future crime occurrences. The system integrates various data sources, including crime reports, demographic information, and spatial data, to provide insights into crime patterns and hotspots.

The project focuses on three main components: exploratory data analysis (EDA), predictive modeling, and visualization. In the EDA phase, we conduct comprehensive analyses of crime data to understand temporal and spatial patterns, identify high-crime areas, and explore correlations with socio-economic factors. Using machine learning algorithms, we build predictive models to forecast crime occurrences based on historical data and contextual variables. The system also incorporates interactive visualizations, such as maps and charts, to facilitate intuitive exploration of crime trends and predictions.**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapter No** | **Chapter Name** | **Page No** |
|  | **ABSTRACT** |  |
| **1** | **INTRODUCTION** | 1 |
| **2** | **PROBLEM STATEMENT** | 2 |
| **3** | **OBJECTIVES** | 2 |
| **4** | **METHODOLOGY** | 3 |
|  | **4.1 Data Source** | 3 |
|  | **4.2 Data Visualization** | 4 |
|  | 4.2.1 General Visualization | 4 |
|  | 4.2.2 Analysis Based on Gender | 9 |
|  | **4.3 Data Pre-Processing** | 18 |
|  | 4.3.1 Replacing Null values | 18 |
|  | 4.3.2 Handling data format inconsistencies | 22 |
|  | 4.3.3 Handling unexpected values | 22 |
|  | 4.3.4 Enhanced data featuring for further analysis | 24 |
|  | 4.3.5 Data Preparation for Predictive Model Building | 25 |
| **5** | **Research Problems** | 26 |
| **6** | **Final Implementation** | 43 |
|  | **Conclusion** | 47 |

1. **INTRODUCTION**

Crime is a pervasive and multifaceted phenomenon that has been a longstanding concern for societies around the world. It encompasses a wide range of behaviors that violate societal norms and laws, including theft, assault, burglary, homicide, and various forms of white-collar crime. The impact of crime extends beyond individual victims to affect communities, economies, and the overall quality of life.



Understanding the dynamics of crime is essential for developing effective strategies for prevention, intervention, and law enforcement. Crime patterns can vary significantly based on factors such as geographical location, socio-economic conditions, cultural norms, and demographic characteristics. By analyzing these patterns and identifying underlying trends, researchers and policymakers can gain valuable insights into the root causes of crime and tailor interventions accordingly.

**2. PROBLEM STATEMENT**

The city of Los Angeles faces persistent challenges related to crime, with law enforcement agencies striving to enhance public safety and reduce criminal activity. To address these challenges, it is crucial to leverage the wealth of information available in the dataset comprising detailed records of reported crimes in Los Angeles from 2020 to 2024.

The key variables in the dataset provide valuable insights into the nature and characteristics of crimes occurring in the city, including the type of crime, date and time of occurrence, location details, victim demographics, and more. By analyzing this data comprehensively, we aim to develop a predictive model to identify high-risk areas and forecast crime trends in Los Angeles based on historical crime data

**3. OBJECTIVES**

1. Uncover Crime Patterns: Analyze the dataset to uncover patterns in crime occurrence, including temporal trends, spatial distribution, and correlations between different types of crimes.
2. Understand Victim Demographics: Explore the demographics of crime victims, including age, gender, and ethnicity, to identify vulnerable populations and assess the impact of crime on different communities.
3. Identify Spatial-Temporal Trends: Conduct spatial and temporal analysis of crime data to identify patterns and trends over time and space, including seasonal variations, weekly/daily patterns, and hotspots of criminal activity.
4. Identify Factors Influencing Crime Severity: Investigate potential factors influencing crime severity, such as the type of crime, location, victim demographics, and socio-economic indicators, to understand the drivers of criminal behavior and prioritize prevention efforts.
5. Provide Actionable Insights: Generate actionable insights and recommendations for law enforcement agencies, policymakers, and community stakeholders to develop targeted crime prevention strategies, allocate resources effectively, and improve public safety in Los Angeles

**4. METHODOLOGY**

**4.1 Data Source**

The dataset contains comprehensive records of reported crimes in Los Angeles, covering the period from 2020 to 2024. Each entry in the dataset provides detailed information essential for understanding and analyzing crime trends and patterns.

The data set provided to us comprised of 28 features and 925720 observations.

|  |  |  |
| --- | --- | --- |
| **S.NO** | **Feature Name** | **Description** |
| 1 | **DR\_NO** | **The** Department Report Number is represented as DR\_NO which is generated for every crime that is reported to the Los Angeles department. |
| 2 | **Date Rptd:** | The date when the crime was reported to the police department is recorded as Date Rptd and the date lies between 2020 to 2024. |
| 3 | **DATE OCC:** | The actual date when the crime occurred is recorded under this column. |
| 4 | **TIME OCC** | The time the crime occurred is represented in a 24-hour format. |
| 5 | **AREA:** | The Los Angeles Police Department has 21 community police stations within the geographical location sequencing from 1-21. |
| 6 | **AREA NAME** | The 21 divisions also have an area name in Los Angeles |
| 7 | **Rprt Dist. No** | Code that represents a sub-area within a Geographic Area this is usually prefixed by area code. |
| 8 | **Part code** | Indicate whether the crime is serious or less offensive. |
| 9 | **Crm Cd** | Indicates the crime committed |
| 10 | **Crm Cd Desc** | Defines the Crime Code provided |
| 11 | **Mocodes** | Modus Operandi code provides additional details about the crime. |
| 12 | **Vict Age:** | Indicates the age of the victim. |
| 13 | **Vict Sex** | F: Female M: Male X: Unknown |
| 14 | **Vict Descent** | The Descent Code in the dataset provides information about the ethnicity or descent of individuals involved in reported crimes. |
| 15 | **Premise Cd** | Type of structure the crime took place in (vehicle, building, parking lot, etc.) |
| 16 | **Premise Desc** | Describes premise Cd |
| 17 | **Weapon Used Cd** | Code for Type of weapon used in crime |
| 18 | **Weapon Desc** | Description of the weapon. |
| 19 | **Status** | Status code of the case |
| 20 | **Status Desc** | Status of the case description |
| 21 | **Crm Cd 1, Crm Cd 2, Crm Cd 3, Crm Cd 4** | Additional status codes associated with the case. |
| 22 | **LOCATION** | Crime location |
| 23 | **Cross Street** | Cross street from the crime location |
| 24 | **LAT:** | The latitude of the location is recorded under this column |
| 25 | **LON:** | The longitude of the location is recorded under this column |

**4.2 Data Visualization**

**4.2.1 General Visualizations**

**4.2.1.1 Analysis of Top Crimes**

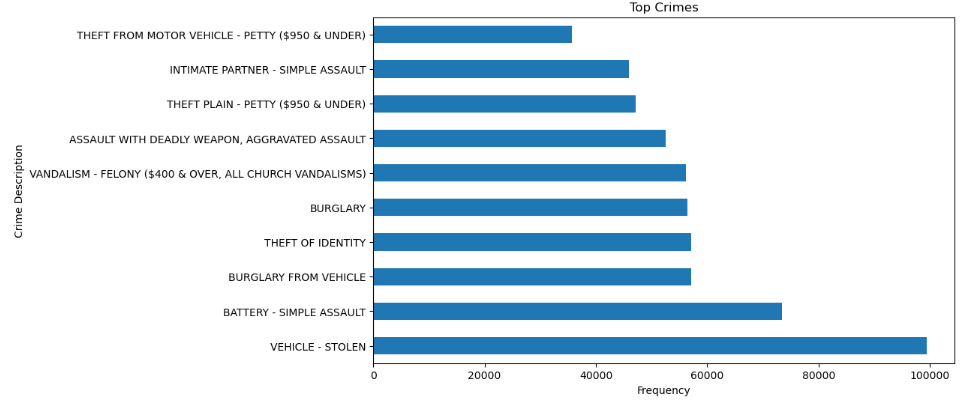


Fig. 4.2.1.1 Top 10 crimes

**Observation :** The analysis reveals that crimes related to "Vehicle Stolen," "Battery - Simple," and "Burglary from Vehicle" exhibit the highest frequency of occurrence compared to offenses such as "Theft from Motor Vehicle - Petty ($959 and under)" and "Intimate Partner - Simple Assault."

**4.2.1.2 Analysis of Top weapons used**

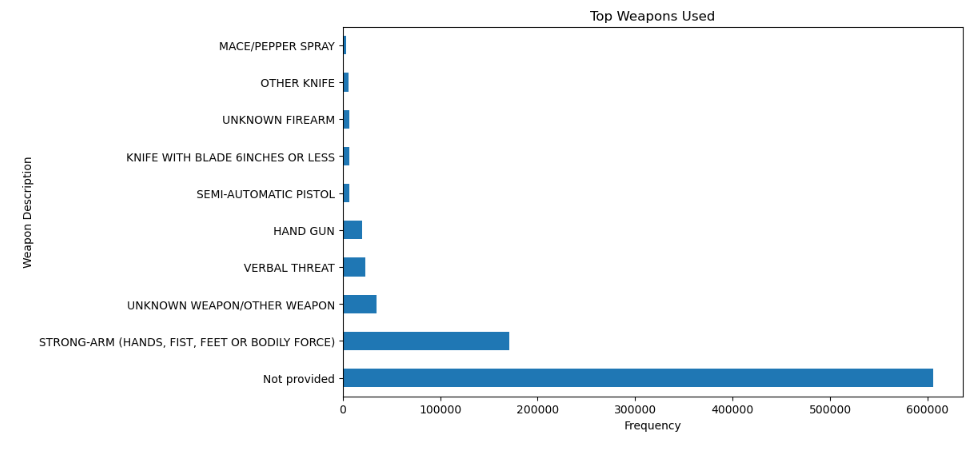


Fig. 4.2.1.2 Top 10 weapons used

**Observation :** The analysis of top weapons used in crimes highlights a significant number of cases where the weapon used was not provided, comprising approximately 600,000 instances. Following closely behind are occurrences involving physical force, categorized as "Strong Arms," with around 180,000 cases. These observations underscore the need for further investigation into the prevalence of unreported weapons and the implications for law enforcement strategies.

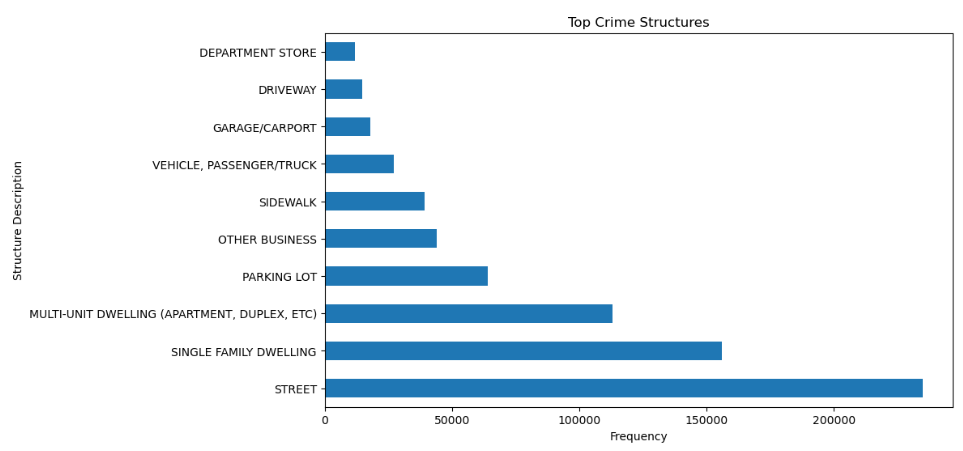
**4.2.1.3 Analysis of Top Crime Structures**  


Fig. 4.2.1.3 Top 10 crime structures

**Observation:** The analysis of top crime structures reveals that street crime structures have the highest frequency of occurrence, followed by single-family dwellings and multi-unit dwellings. Conversely, driveway-related crimes are among the least reported, followed by incidents within department stores. This highlights the varying vulnerabilities of different types of structures to criminal activity, emphasizing the importance of targeted prevention measures for specific locations.

**4.2.1.4 Analysis of Status of cases**

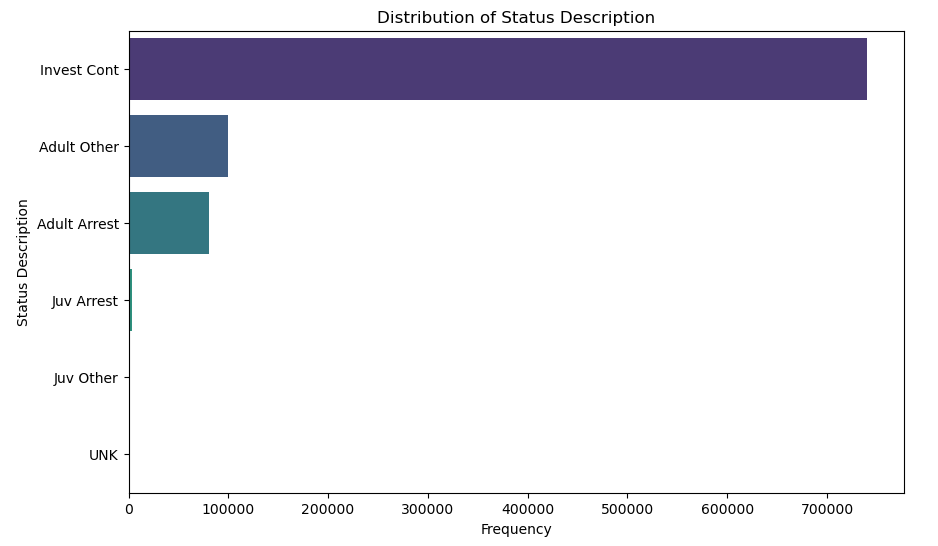


Fig. 4.2.1.4 Distribution of status description

**Observation:** In the analysis of the top status descriptions, it's evident that "Invest Cont" cases dominate with a count exceeding 700,000, indicating that a significant portion of reported cases are actively under investigation. This suggests a considerable workload for law enforcement agencies in Los Angeles in processing and resolving ongoing investigations. Following "Invest Cont," "Adult Other" cases are notable, with approximately 120,000 occurrences, indicating various outcomes beyond arrest for cases involving adults. "Adult Arrest" cases, where arrests have been made, follow with around 80,000 instances. Interestingly, the count for "Juv Arrest" cases is minimal compared to the other categories, suggesting relatively fewer arrests involving juveniles.

**4.2.1.5 Analysis of Victim Descent**

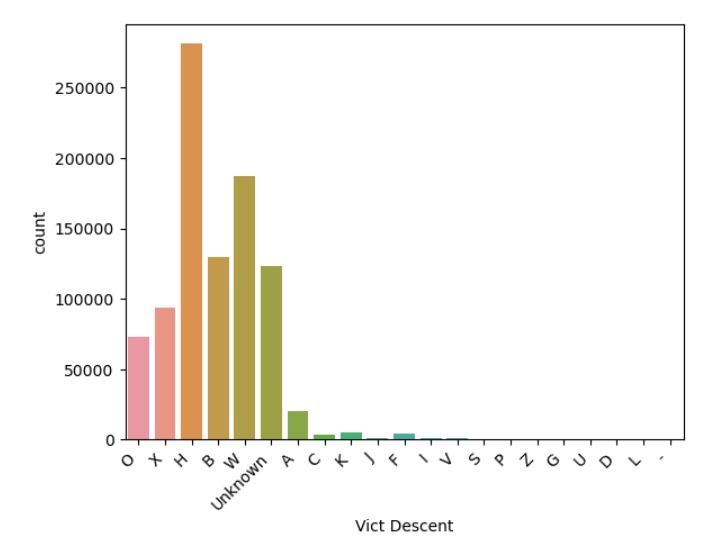
****

Fig. 4.2.1.5 Analysis of Victim Descent

**Observation:** The analysis of top crimes by victim descent reveals a notable hierarchy, with crimes primarily affecting individuals of Hispanic/Latin/Mexican descent (H) ranking the highest. Following are crimes impacting victims of White (W) and Black (B) descent. Instances where the victim's descent is unknown and those categorized as "X" also feature prominently among the top crimes. Crimes affecting victims of "Other" (O) descent are observed to a lesser extent.

**4.2.1.6 Analysis of top crimes location**

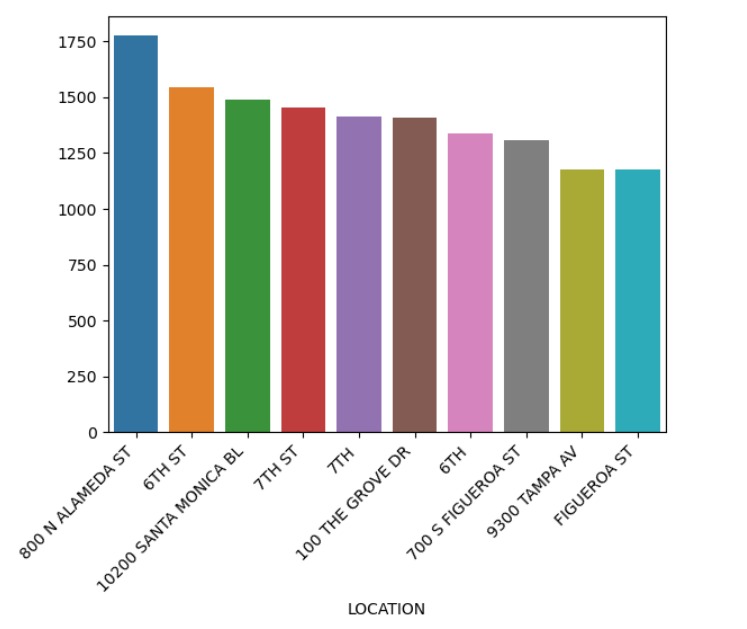
****

Fig. 4.2.1.6 Top 10 crimes locations

**Observation:** The analysis of top crime locations reveals several key areas with consistently high crime frequencies. Notably, locations such as 800 N Alameda St, 6th St, and 10200 Santa Monica Blvd emerge as hotspots for criminal activity, suggesting concentrated areas of concern for law enforcement and community safety efforts. The presence of recurring locations like 7th St and 100 The Grove Dr underscores persistent challenges in maintaining security and addressing crime in specific neighborhoods or commercial districts. Additionally, the inclusion of major thoroughfares like Figueroa St and Tampa Ave among the top crime locations highlights the significance of transit hubs and urban corridors as potential focal points for crime prevention initiatives.

**4.2.2 Analysis based on Gender**

**4.2.2.1 Analysis of Structure of crime**

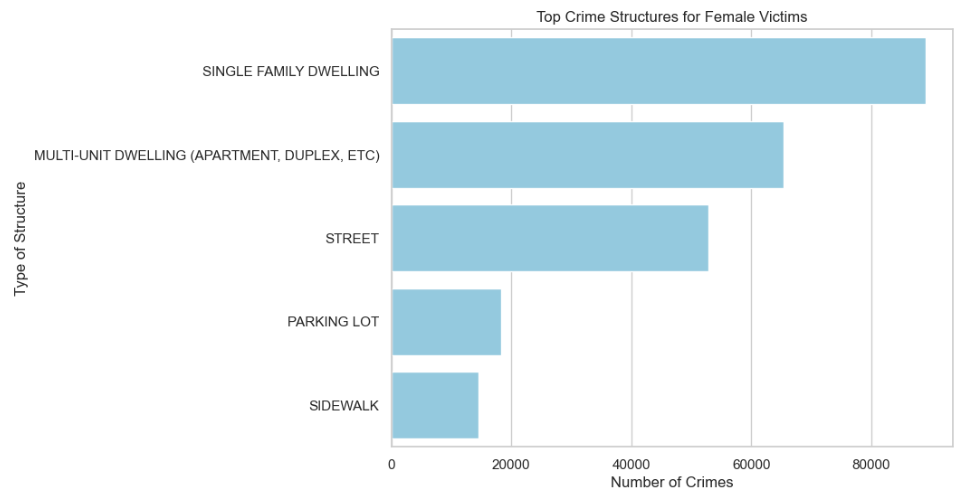


Fig. 4.2.2.1.1 Top crime structures for Female victims

**Observation:** Analysis of crime structures for female victims reveals distinct patterns. Incidents in single-family dwellings suggest domestic violence, while those in multi-unit dwellings indicate potential harassment. Street-related crimes highlight risks in public spaces, parking lot incidents signal threats in vehicle contexts, and sidewalk-related offenses underscore vulnerabilities outdoors. Crimes in other businesses expose women to risks in commercial settings. These patterns emphasize the diverse challenges faced by female victims in various environments.

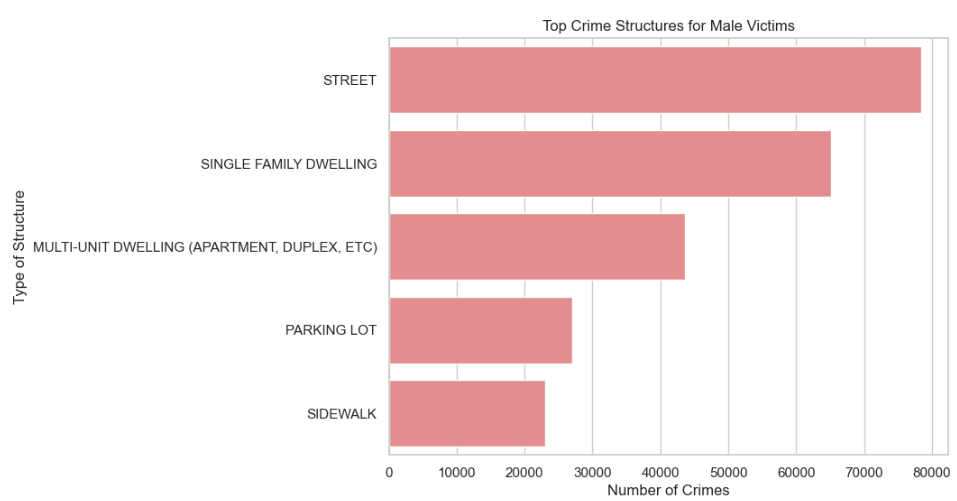


Fig. 4.2.2.1.2 Top crime structures for Male victims

**Observation:** Analysis of crime structures for male victims reveals distinctive trends. Street-related incidents suggest risks of confrontations or assaults in public areas, while crimes in single-family dwellings may reflect domestic disturbances or property offenses. Incidents in multi-unit dwellings could signify tensions in shared housing environments, and parking lot crimes highlight risks in vehicle-related contexts. Sidewalk-related offenses expose vulnerabilities to opportunistic crimes outdoors, and incidents in other businesses underscore risks in commercial settings. These patterns illustrate the varied challenges faced by male victims across different environments.

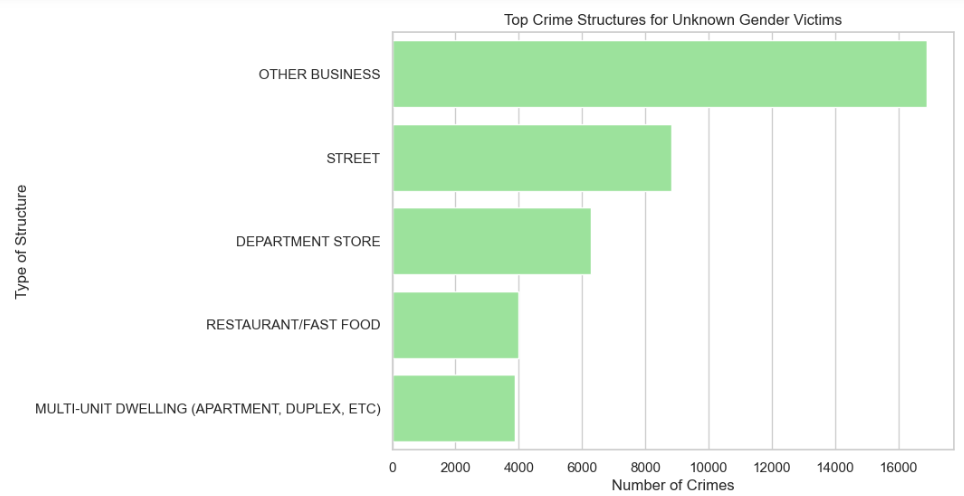


Fig. 4.2.2.1.3 Top crime structures for Unknown victims

**Observations:** For unknown victims, the prevalence of street-related crimes suggests a heightened risk of confrontations, muggings, or assaults in public spaces. Incidents in other business establishments indicate exposure to various offenses in retail or commercial settings, while occurrences in department stores signify vulnerabilities to criminal activity in large retail establishments. Crimes in restaurant or fast-food venues reflect risks in dining environments where victim identity is unknown, and incidents in multi-dwelling units may signify tensions or criminal activity in shared housing environments.

**4.2.2.2 Analysis of serious and unserious crimes**

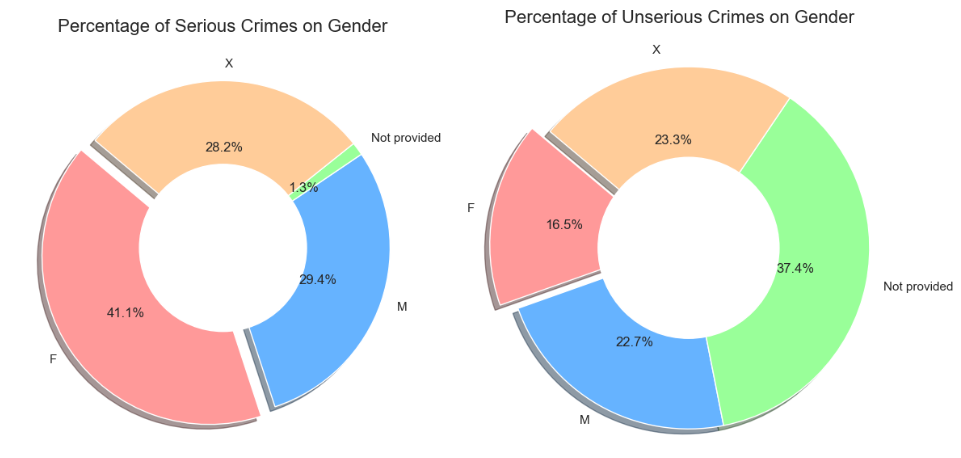


Fig. 4.2.2.2 Percentage of distribution of serious and unserious crimes

**Observation:** The analysis reveals a notable gender disparity in the distribution of serious and unserious crimes. Among females, a higher proportion (41.1%) of reported crimes are classified as serious, indicating a greater prevalence of severe offenses compared to males (29.4%) and cases with unspecified gender (28.2%). Conversely, in unserious crimes, males and cases with unspecified gender have a higher representation (22.7% and 23.3% respectively) compared to females (16.5%). The significant percentage (37.4%) of unspecified gender in the category of unserious crimes suggests a lack of data or reporting specificity, potentially obscuring the true gender distribution in less severe offenses.

**4.2.2.3 Age distribution among genders**

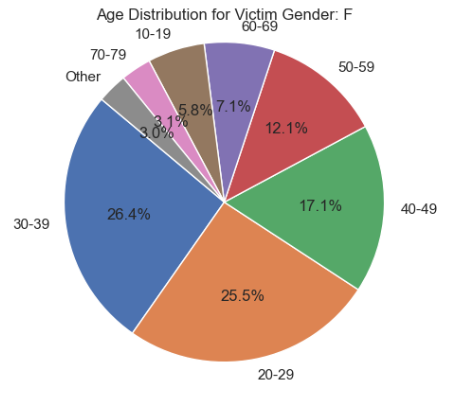


Fig. 4.2.2.3.1 Age distribution for females

**Observation:** The age distribution for female victims indicates a concentration of incidents among individuals aged 20-39, comprising approximately 52% of reported cases. This age group appears to be the most vulnerable, suggesting potential risks associated with young adulthood and early middle age. Notably, individuals aged 40-59 also represent a significant proportion (32%) of female victims, indicating continued susceptibility to crime across multiple life stages.

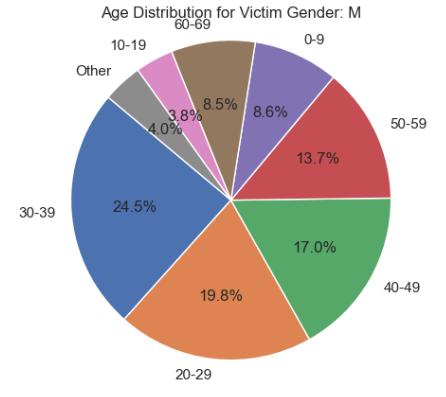


Fig. 4.2.2.3.2 Age distribution for Males

**Observation:** Similar to females, males aged 20-39 constitute a substantial portion (44.3%) of reported victims, highlighting a vulnerability among young to middle-aged men. Additionally, individuals aged 40-59 account for 30.7% of male victims, indicating sustained risks into middle adulthood. Notably, children under 10 years old represent 8.6% of male victims, underscoring vulnerabilities among young boys to criminal victimization

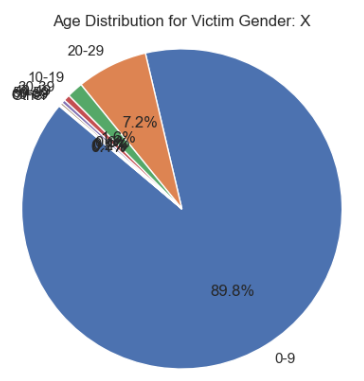


Fig. 4.2.2.3.3 Age distribution for Unknown

**Observation:** The age distribution for unspecified gender (X) reveals a stark contrast, with the vast majority (89.8%) of reported victims fall into the 0-9 age category. This suggests a disproportionate vulnerability among very young individuals of unspecified gender, potentially indicating incidents involving children or infants.

**4.2.2.4 Analysis of top crimes based on gender**

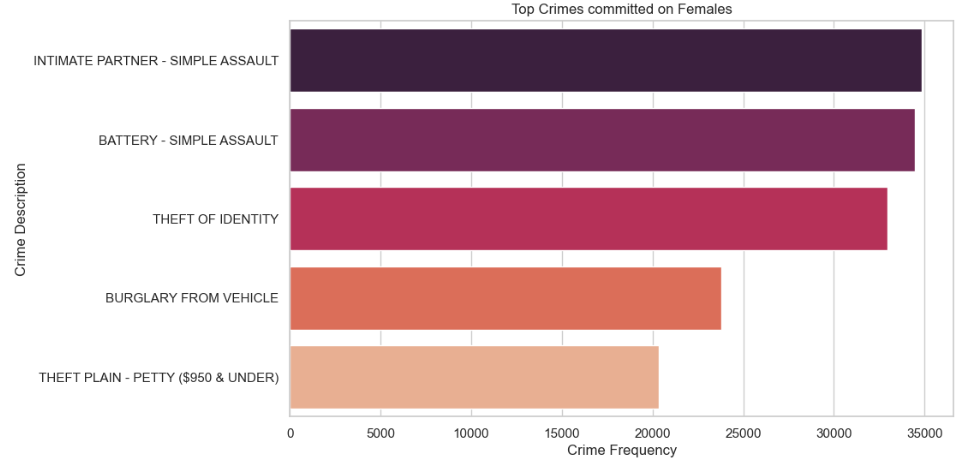


Fig. 4.2.2.4.1 Top crimes committed on females

**Observation:** The top crimes reported among females predominantly include offenses related to interpersonal conflicts or domestic disputes, such as intimate partner simple assault and battery simple assault. The prominence of identity theft and burglary from vehicles also suggests vulnerabilities to property-related offenses among female victims. These patterns may reflect the intersection of gender-based violence and property crime, highlighting the need for targeted interventions to address both personal safety and property security for women.

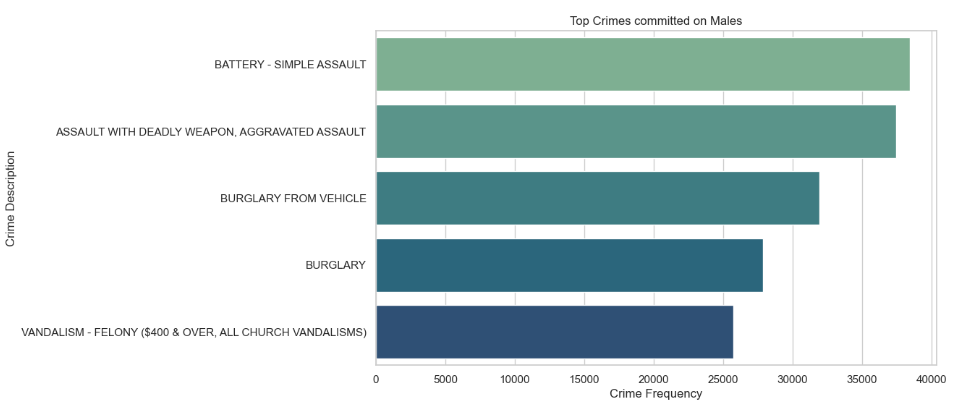


Fig. 4.2.2.4.2 Top crimes committed on males

**Observation:** Among males, battery simple assault emerges as the most prevalent offense, indicating a higher likelihood of physical altercations or assaults involving men. Additionally, crimes involving deadly weapons or aggravated assaults rank prominently, underscoring risks associated with violent confrontations or criminal acts among male victims. The occurrence of burglary-related offenses further suggests vulnerabilities to property crimes among men, emphasizing the diverse nature of victimization experienced by males.

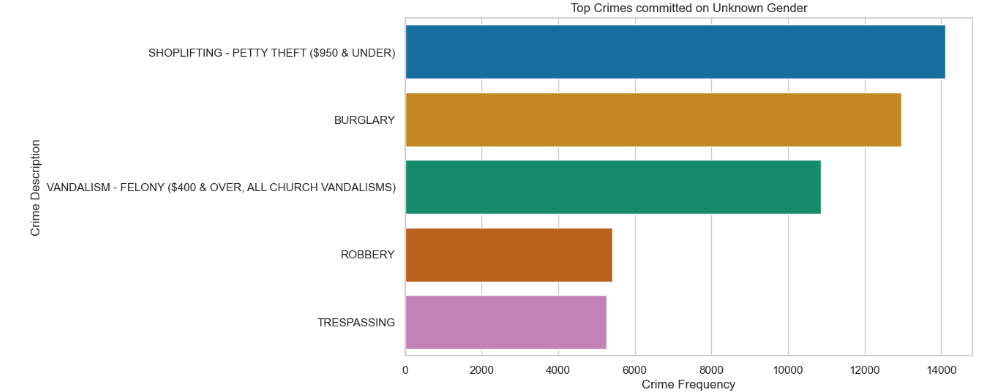


Fig. 4.2.2.4.3 Top crimes committed on unknown gender

**Observation:** Crimes reported for individuals of unknown gender primarily include offenses such as shoplifting, burglary, and vandalism, suggesting a mix of property-related offenses and crimes against persons. The prevalence of shoplifting and vandalism may indicate opportunistic crimes or minor offenses occurring in public spaces. The presence of burglary and robbery among top offenses also suggests potential risks of more serious criminal acts within this category, underscoring the need for further investigation into victim demographics and crime contexts.

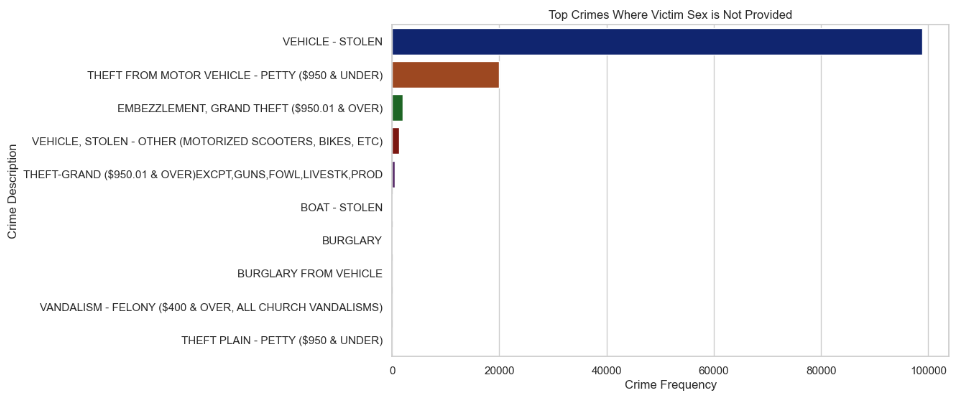


Fig. 4.2.2.4.4 Top crimes committed on Not provided

**Observation:** Victims with unspecified gender (not provided) predominantly experience crimes related to theft and property offenses, including vehicle theft, theft from motor vehicles, embezzlement, and grand theft. The prevalence of these offenses highlights vulnerabilities to property crimes among individuals for whom gender information is not available. The diversity of offenses suggests a range of criminal activities targeting property or financial assets, indicating the need for enhanced security measures and crime prevention strategies.

**4.2.2.5 Analysis of top weapons based on gender**

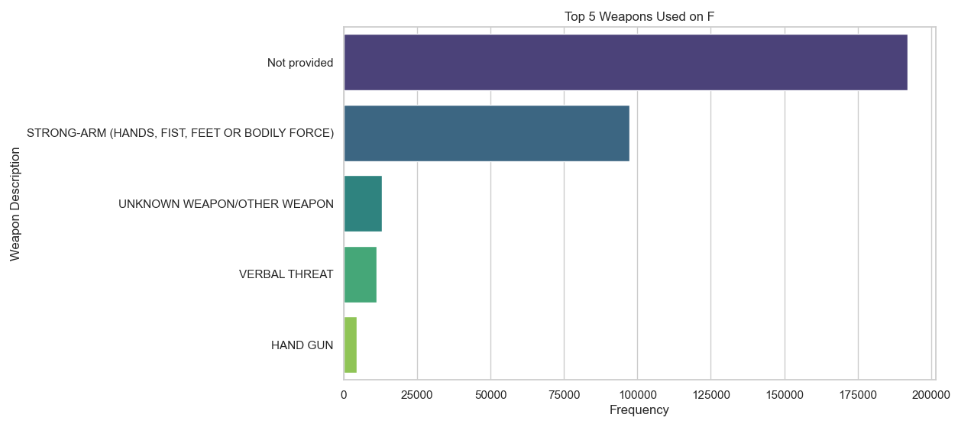


Fig. 4.2.2.5.1 Top weapons used on females

**Observation:** The prevalence of "not provided" as the top weapon used among female victims suggests a lack of detailed information or disclosure regarding the specific weapons involved in crimes against women. This opacity hinders efforts to understand the nature of victimization and the mechanisms of violence experienced by female individuals. Additionally, the prominence of strong-arm tactics and bodily force underscores the prevalence of physical assaults or confrontations involving female victims.

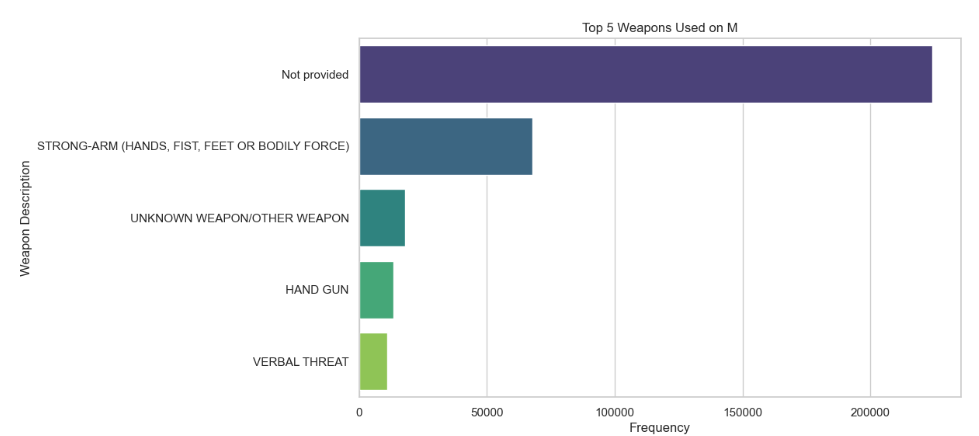


Fig. 4.2.2.5.2 Top weapons used on males

**Observation:** Similar to females, the predominance of "not provided" as the top weapon used among male victims reflects a lack of clarity or disclosure regarding weapon types in reported crimes involving men. The prominence of strong-arm tactics and bodily force suggests a propensity for physical altercations or assaults among male individuals.

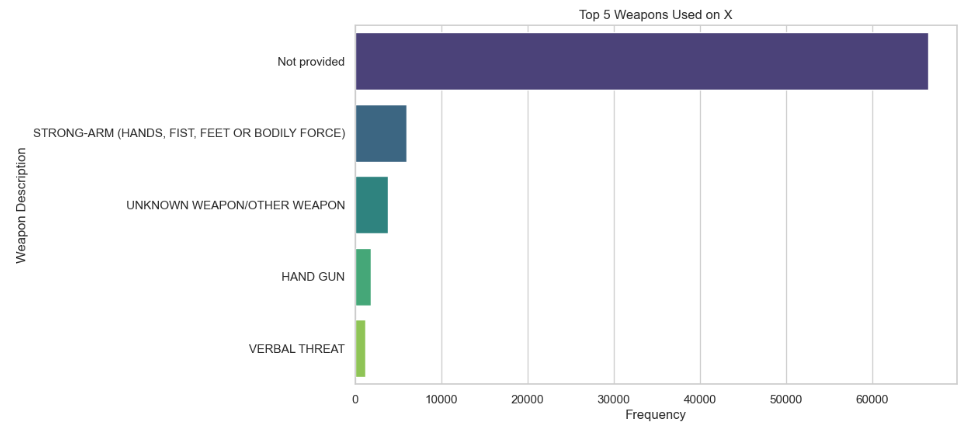


Fig. 4.2.2.5.3 Top weapons used on unknown gender

**Observation:** Among victims with unknown gender, the prevalence of "not provided" as the top weapon used indicates a lack of detailed information or disclosure regarding weapon types in reported crimes. This ambiguity complicates efforts to understand the dynamics of victimization and the mechanisms of violence experienced by individuals with unspecified gender. The prominence of strong-arm tactics, unknown weapons, and verbal threats suggests a range of potential risks and vulnerabilities for victims of unknown gender, emphasizing the need for further investigation into the nature and context of these incidents.

**4.2.2.6 Analysis of Victim Descent based on gender**

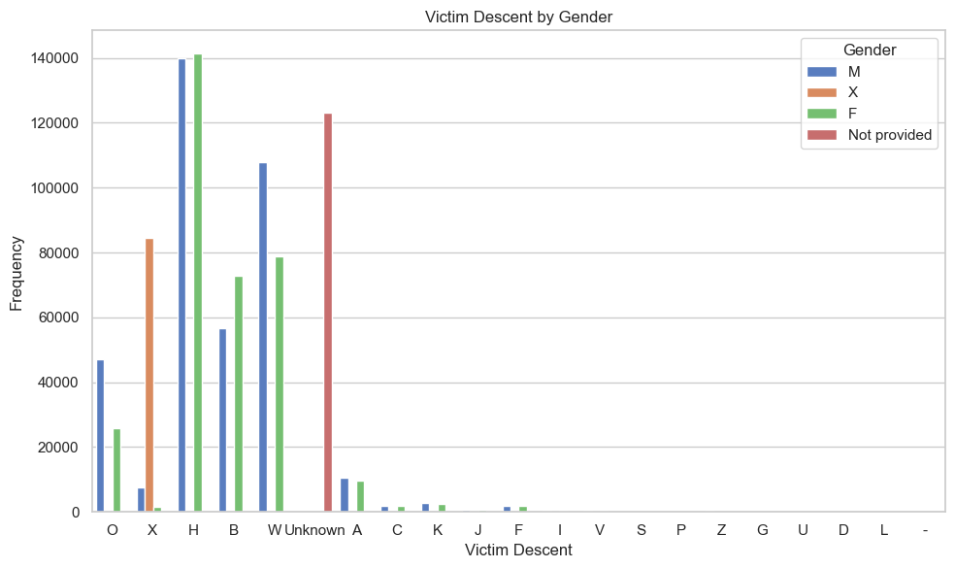


Fig. 4.2.2.6 Analysis of victim descent by gender

**Observations:** For Hispanic/Latin/Mexican descent (H), male victims slightly outnumber females, with approximately 140,000 instances each. Among Black (B) victims, there are fewer male instances (below 60,000) compared to females (below 80,000). White (W) victims show a higher prevalence among males (over 100,000) than females (below 80,000). For victims identified as "Other" (O), there's a notable prevalence among males, with just over 40,000 instances, compared to females, with just over 20,000 instances.

**4.3 Data Pre-Processing**

**4.3.1 Replacing Null values -**

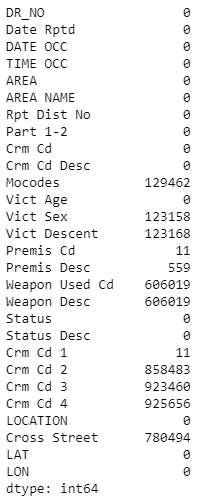


Fig. 4.3.1.1 Sum of all null values

Mocodes: This column has 129,462 missing values, indicating that data on the Modus Operandi (MO) codes for crimes is not available for a significant portion of the dataset. To handle this we have decided to replace the null values with a dummy values “unknown”.

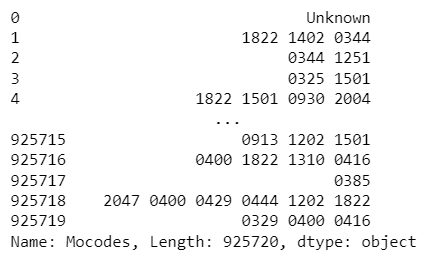


Fig. 4.3.1.2 Dataset after handling Mocodes null values

Cross Street: There are 780,494 entries missing data on the cross street from the crime location, indicating that specific intersection details are absent. Given the inability to assume or replace these missing values accurately, we have opted to fill them with a dummy value, "Unknown." This approach ensures consistency in the dataset while acknowledging the absence of precise cross street information for these entries.

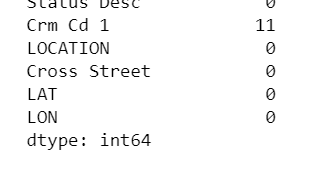


Fig. 4.3.1.2 Dataset after handling Cross street null values

Victim Sex: With 123,158 entries lacking information on the sex of the victim, it's evident that details regarding the gender of the affected individuals are missing. Since it's inappropriate to make assumptions or impute values for such sensitive data, we've chosen to fill these missing values with a placeholder, "Not provided."



Fig. 4.3.1.3 Dataset after handling victim null values

Victim Descent: The dataset contains 123,168 entries without information on the descent or ethnicity of the victim. Given the sensitivity and potential implications of ethnicity data, making assumptions or imputing values would not be appropriate. Therefore, to maintain the integrity of the dataset while acknowledging the absence of specific victim descent information for these entries, we have filled these missing values with a placeholder, "Unknown."

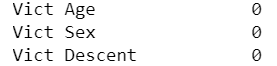


Fig. 4.3.1.4 Dataset after handling Victim Descent null values

Premis Cd: The dataset exhibits 11 missing values in the Premis Cd column, suggesting that certain entries lack information regarding the premise code, which denotes the type of structure where the crime occurred. Given the relatively small number of missing values in this column, we have decided to handle them by dropping the corresponding rows from the dataset.



Fig. 4.3.1.5 Dataset after handling Premise code null values

Premis Desc: Among the dataset entries, 548 instances lack a description of the premise code. This absence indicates that crucial details regarding the location or setting of the crime are missing for these cases. Initially, an attempt was made to map the Premis Cd values to their corresponding Premis Desc descriptions. However, this endeavor was unsuccessful due to a discrepancy where certain Premis Cd values did not have corresponding descriptions.

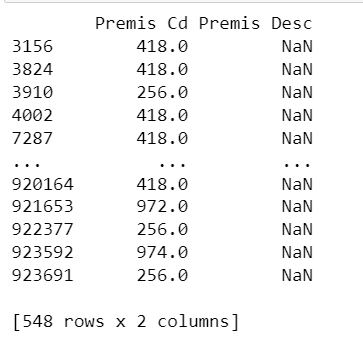


Fig. 4.3.1.5 Mapping of premise code and premise description

Subsequently, as a resolution, it was decided to drop the rows with missing Premis Desc values. This action ensures data integrity and accuracy by eliminating instances where crucial information about the crime location is unavailable.

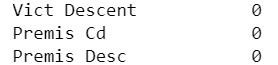


Fig. 4.3.1.6 Dataset after handling Premise description null values

Weapon Used Cd and Weapon Desc: These columns have a substantial number of missing values, with 606,019 entries lacking information on the type of weapon used in the commission of the crime. These columns were populated with placeholder values 'Not provided' to address the large number of missing values in these fields. Dropping or replacing these entries was not feasible due to the significant number of NaN values present. Therefore, utilizing a dummy node helped maintain the completeness of the dataset while acknowledging the absence of weapon-related information in these instances.

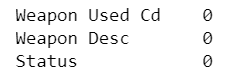


Fig. 4.3.1.7 Dataset after handling Weapons used code and weapon description null values

Crm Cd 1, Crm Cd 2, Crm Cd 3, Crm Cd 4: The columns Crm Cd 1, Crm Cd 2, Crm Cd 3, and Crm Cd 4 contained missing values, suggesting that additional crime codes associated with certain cases were unavailable. Since Crm Cd 2, Crm Cd 3, and Crm Cd 4 had relatively few values, it was decided to drop these columns as they did not provide significant data. Additionally, Crm Cd 1 had only 11 NaN values, which were dropped to maintain data integrity.

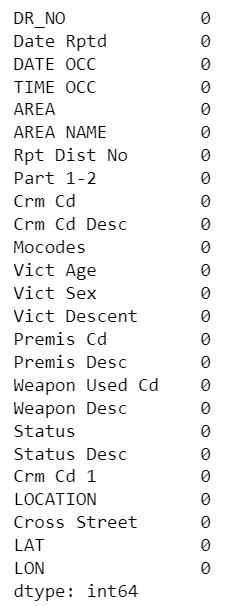


Fig. 4.3.1.8 Dataset after handling crime codes 1,2,3,4

**4.3.2 Handling Data Format Inconsistencies**

During exploratory data analysis (EDA), we identified columns with data format inconsistency.

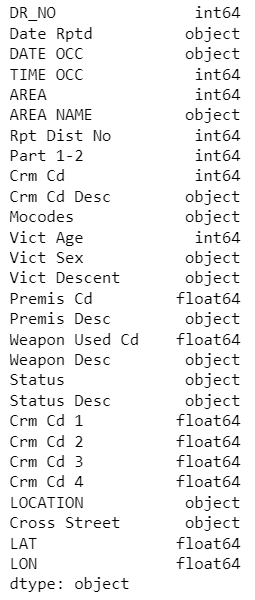


Fig. 4.3.2.1 Datatypes of dataset

Date Rptd (Date Reported): This column is stored as an object (string) data type instead of a datetime format. Storing dates as strings can limit the ability to perform date-based operations and analysis.

DATE OCC (Date Occurred): Similar to the Date Rptd column, the DATE OCC column is also stored as an object (string) data type instead of datetime format, which may hinder date-related analysis and visualization.

TIME OCC (Time Occurred): Although represented as an integer data type, time values are typically stored in datetime format to facilitate time-based analysis and interpretation. Storing time as integers without context may lead to ambiguity and difficulty in analyzing temporal patterns.

We have addressed the inconsistencies in data formats for the date and time columns by employing conversion methods. As a result, the "Date Rptd," "DATE OCC," and "TIME OCC" columns have been successfully converted to the datetime64 data type.

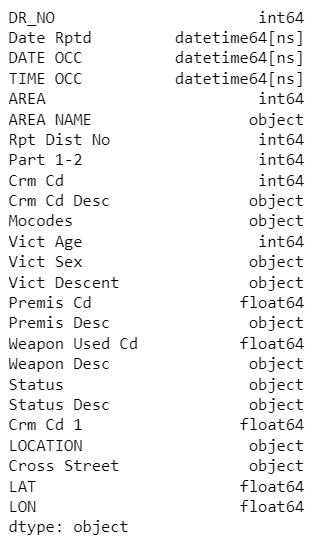


Fig. 4.3.2.1 Dataset after handling inconsistent datatypes

**4.3.3 Handling Unexpected values**

Age: Unexpected negative values and values exceeding realistic age ranges were observed in the victim age data.

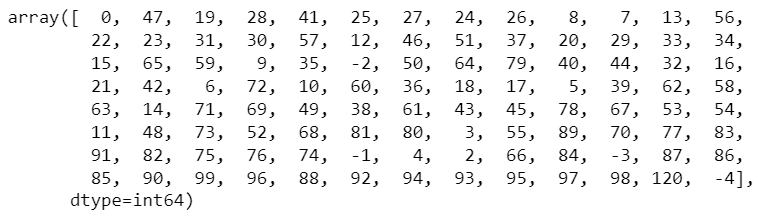


Fig. 4.3.3.1 Unexpected values in “Age” column

Since these values are erroneous and cannot be valid ages,therefore we removed all rows where the victim age was recorded as a negative value. Negative ages are logically impossible and likely indicate data entry errors or missing values.

Vict Sex: The "Vict Sex" column in the dataset contained unique values such as M (Male), X (Unknown), F (Female), Not provided, H, and -. However, there was no information provided regarding the meanings of "H" and "-", which appeared in a total of 107 rows.



Fig. 4.3.3.2 Unique values in “Vict Sex” column



Fig. 4.3.3.3 Number of rows with unexpected values of “Vict Sex”

Due to the lack of clarity on these categories and their potential impact on the analysis, a decision was made to drop the rows associated with these values from the dataset.Also since “X” and “Not provided” signify the same meaning of containing unknown values we decided to combine them as on category.

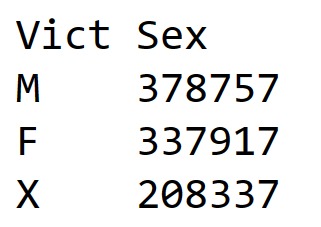


Fig. 4.3.3.1 Dataset after handling unexpected rows

**4.3.4 Enhanced Data Features for further analysis**

To facilitate more efficient exploratory data analysis (EDA) and streamline the model-building process the following changes were made:

* Numeric Encoded Columns Created for Categorical Variables: 'Location\_Num', 'Vict\_Sex\_Cat', 'Cross\_Street\_Cat', 'Vict\_Desc\_Cat', Derived from Columns 'LOCATION', 'Vict Sex', 'Cross Street', and 'Vict Descent'.

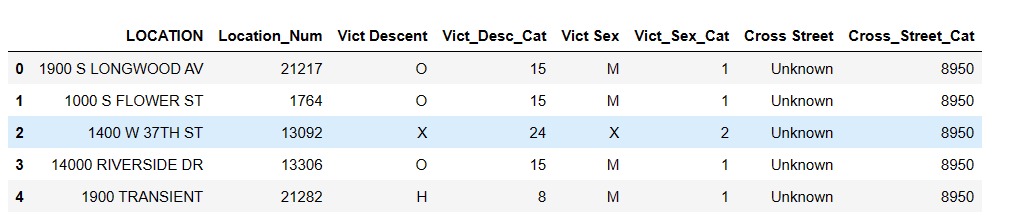


Fig. 4.3.4.1 Dataset after generating new feature columns

* Three new columns,'Day of the Week', 'Month' and 'Year'—have been incorporated into the dataset. The addition of these temporal features allows for deeper insights into the temporal distribution of crimes, enabling analysts to identify patterns, trends, and seasonality more effectively.

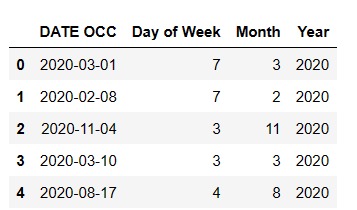


Fig. 4.3.4.2 Dataset after generating 'Day of the Week', 'Month' and 'Year'

Day of Week : {1: "Monday",2: "Tuesday",3: "Wednesday",4: "Thursday",5: "Friday",6: "Saturday",7: "Sunday"}

Month : {1: "January",2: "February",3: "March",4: "April",5: "May",6: "June",7: "July",8: "August",9: "September",10: "October",11: "November",12: "December"}

* In order to augment the temporal granularity of the dataset, two new columns—hours and minutes—have been derived from the existing 'TIME OCC' column.

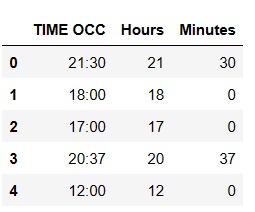


Fig. 4.3.4.3 Dataset after generating “Hours” and “Minutes”

**4.3.5 Data Preparation for Predictive Model Building**

Due to the dataset's size, we opted to utilize the first 5000 rows for modeling purposes, and then split the data into an 80:20 ratio.

**5.Research Problems**

**5.1 Predicting Crime Type**

**Problem statement:**Predicting the type of crime based on factors such as location and day of the week is crucial for law enforcement agencies to allocate resources effectively and implement targeted crime prevention strategies.

**Feature and Target variables :** According to the given dataset we have selected 'Hours','Month','Day of Week','AREA','Premis Cd','Part 1-2','Vict\_Sex\_Cat' columns as feature variables and ‘Crm Cd’ as target variable.

**Models :**

1. Random Forest Classifier: The Random Forest Classifier achieved an accuracy of 41%, indicating its capability to predict the type of crime with moderate success. The model's performance was assessed through macro and weighted average metrics, with macro average F1-score of 0.13 and weighted average F1-score of 0.36, respectively.

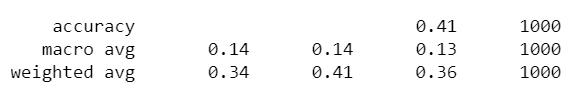


Fig. 5.1.1 Result Accuracies of Random Forest classifier

2.Gradient Boosting Classifier :Similarly, the Gradient Boosting Classifier also attained an accuracy of 41%, showcasing comparable predictive performance to the Random Forest model.

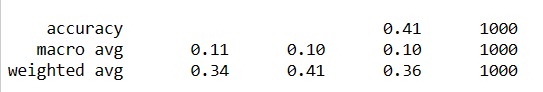


Fig. 5.1.2 Result Accuracies of Gradient Boosting classifier

3.SVM Model :On the other hand, the SVM model exhibited lower accuracy at 33%, suggesting limited effectiveness in classifying crime types based on the given features. The macro average F1-score for the SVM model was notably low at 0.01, indicating poor performance across the different classes.

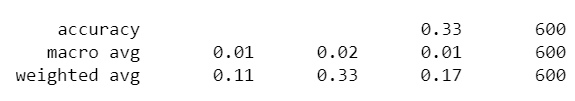


Fig. 5.1.3 Result Accuracies of SVM model

**Conclusion :** We opted for the Random Forest Classifier due to its ability to handle high-dimensional data and capture complex relationships between features and target variables effectively.

**5.2 Crime Severity Prediction**

**Problem Statement:**The problem statement is to investigate whether factors such as the type of weapon used, victim demographics, and other relevant features can predict the severity of crimes, particularly in terms of injuries sustained.

**EDA :**

* **Correlation Heatmap:**

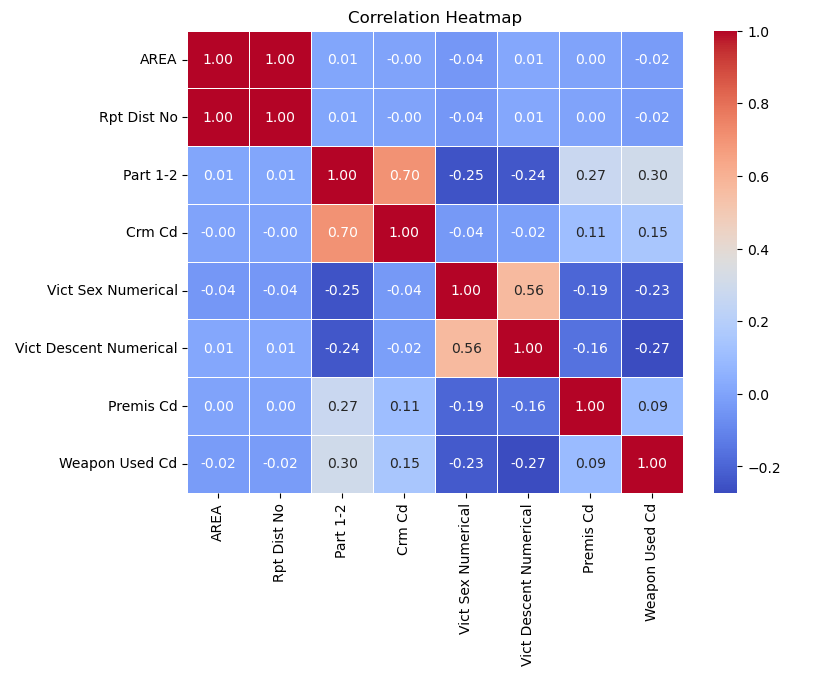


Fig. 5.2.1 Correlation heat map

**Observation :** From the above heatmap we can see that we have some correlation between “Part 1-2” and “Crm cd” of 0.70 which is followed by a correlation values of 0.56 between “Vict Descent Numerical” and “Vict Sex Numerical” suggesting us on which variables to choose for the model.

* **Distribution of Crime Severity by Area**

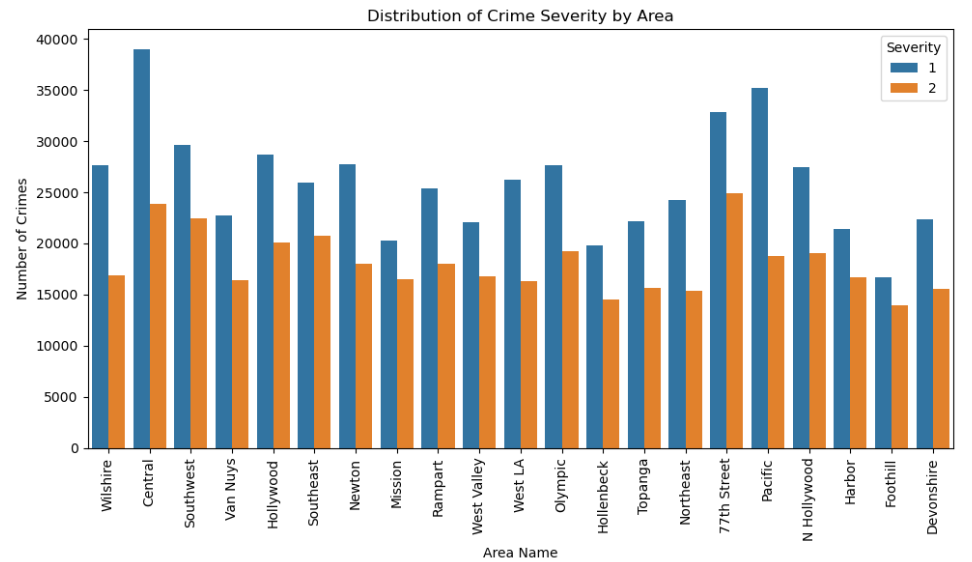


Fig. 5.2.2 Distribution of crime severity by area

**Observation :** “ Central” and “Pacific” are the two top most areas with high crime severity of type “1” while “77th Street” and “Central” are the top two most areas where crime type “2” has occurred. It can also be noted that crime “2” occurs across most of the areas consistently.

**Feature and Target Variables :** The following are the feature variables, ‘Weapon Used Cd', 'Crm Cd', 'Vict Age', 'Vict Sex Numerical', 'Vict Descent Numerical', 'AREA' and “Part 1-2” was chosen as our target variable.

**Models:**

1. Random Forest Classifier :The Random Forest Classifier achieved near-perfect precision, recall, and F1-score for both classes, demonstrating exceptional performance with an accuracy of 99.99%.

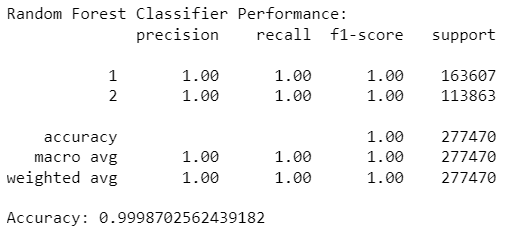


Fig. 5.2.3 Result Accuracies of Random Forest classifier

1. SVM Model :Similarly, the Support Vector Machine (SVM) Classifier displayed good precision and recall, yielding an accuracy of 76%.

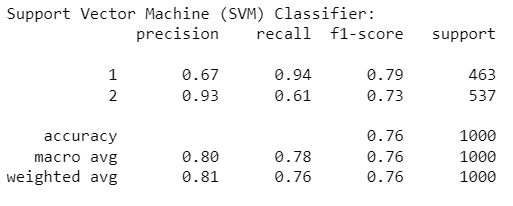


Fig. 5.2.3 Result Accuracies of SVM model

1. Gradient Boosting Classifier: However, the Gradient Boosting Classifier outperformed both models, achieving perfect precision, recall, and F1-score for both classes, resulting in an accuracy of 100%.

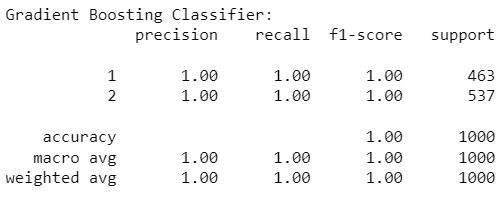


Fig. 5.2.3 Result Accuracies of Gradient Boosting classifier

**Conclusion:** We chose the Gradient Boosting Classifier due to its superior performance across all metrics, indicating its robustness in accurately classifying crime types.

**5.3 High-Crime Area Identification**

**Problem Statement:**Developing a model utilizing geo-spatial data to predict areas with a higher propensity for crime.

**Feature and Target variables :** “LAT” and “LON” were the variables used as features variables and “AREA” is the variable used for target variable.

**Models:**

1.Random Forest Classifier : Demonstrated high accuracy similar to the Gradient Boosting Classifier with an accuracy with an accuracy of 0.98.

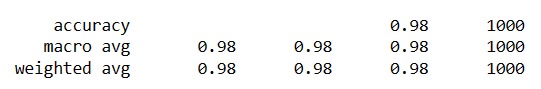


Fig. 5.3.1 Result Accuracies of Random Forest classifier

2.SVM Model :The Support Vector Machine (SVM) classifier achieved an accuracy of 0.08, with very low precision and recall scores, indicating poor performance in classifying the data.

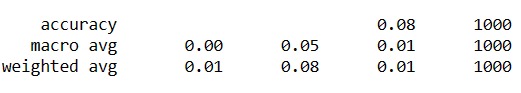


Fig. 5.3.2 Result Accuracies of SVM model

3.Gradient Boosting Classifier: Achieved high accuracy of 0.98 with precision, recall, and F1-scores close to 1.00, indicating excellent performance in classifying the data.

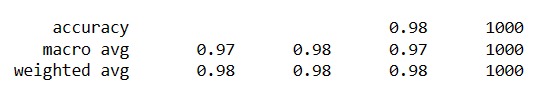


Fig. 5.3.3 Result Accuracies of Gradient Boosting classifier

**Conclusion :** We chose the Random Forest classifier due to its robustness against overfitting, ability to handle large datasets with high dimensionality, and capability to capture complex relationships within the data while maintaining computational efficiency.

**5.4 Crime Trend Analysis**

**Problem Statement:**Using historical crime data, this project aims to identify seasonal and temporal trends in crime patterns, providing insights into variations in criminal activity over time.

**EDA:**

* **Analysis of Monthly distribution of crime**

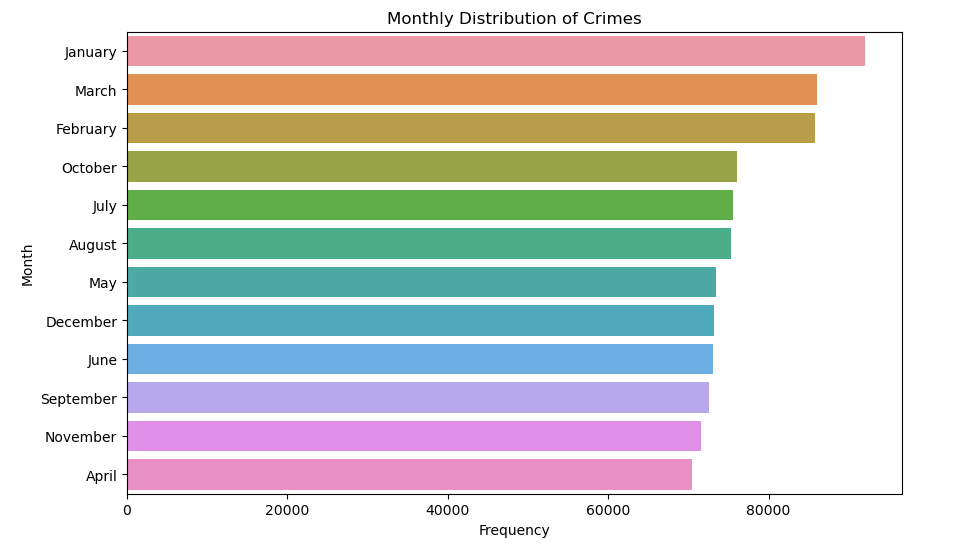


Fig. 5.4.1 Monthly distribution of crimes

**Observation :** The analysis of monthly distribution of crimes reveals that January experiences the highest frequency of reported crimes, followed closely by March and February. Interestingly, there is minimal discrepancy between the frequencies of occurrence in February and March, indicating a consistent level of criminal activity during these months. In contrast, November and April exhibit the lowest frequencies of reported crimes. While there exists a discernible hierarchy in terms of the most occurring months to the least, the differences in their frequencies are relatively marginal. This suggests a relatively stable pattern of criminal activity throughout the year in the dataset, with slight fluctuations between certain months.

* **Analysis of Weekly Distribution of crimes**

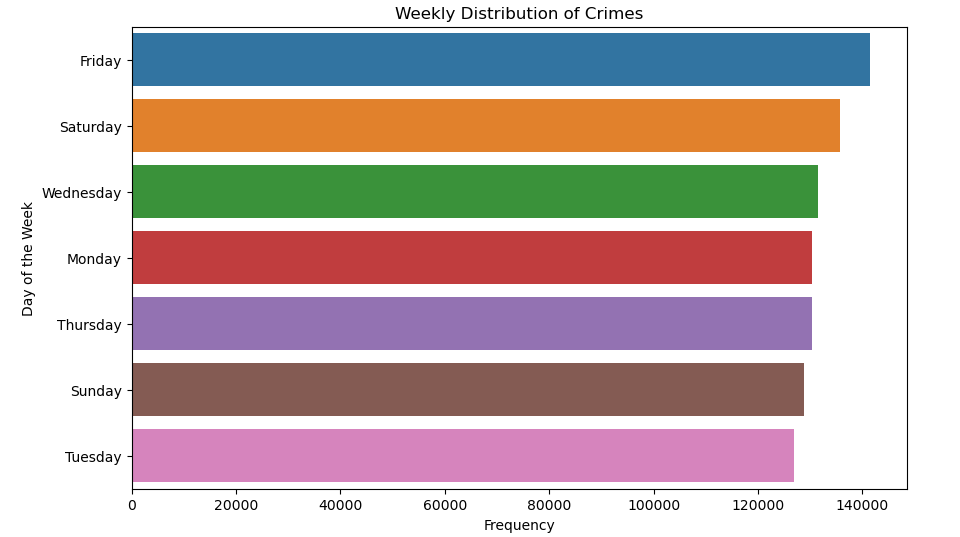


Fig. 5.4.2 Weekly distribution of crimes

**Observation :** The analysis of the weekly distribution of crimes reveals that Friday experiences the highest frequency of reported crimes, followed closely by Saturday. Interestingly, Monday and Thursday exhibit very similar frequencies of occurrence, suggesting comparable levels of criminal activity on these days. In contrast, Tuesday and Thursday have the lowest frequencies of reported crimes. Notably, the differences in frequencies between these weekdays are relatively small, indicating a consistent pattern of criminal activity throughout the week. This suggests that criminal activity in the dataset is distributed relatively evenly across different days of the week, with Friday emerging as the peak day for reported crimes.

* **Temporal Analysis by Gender**

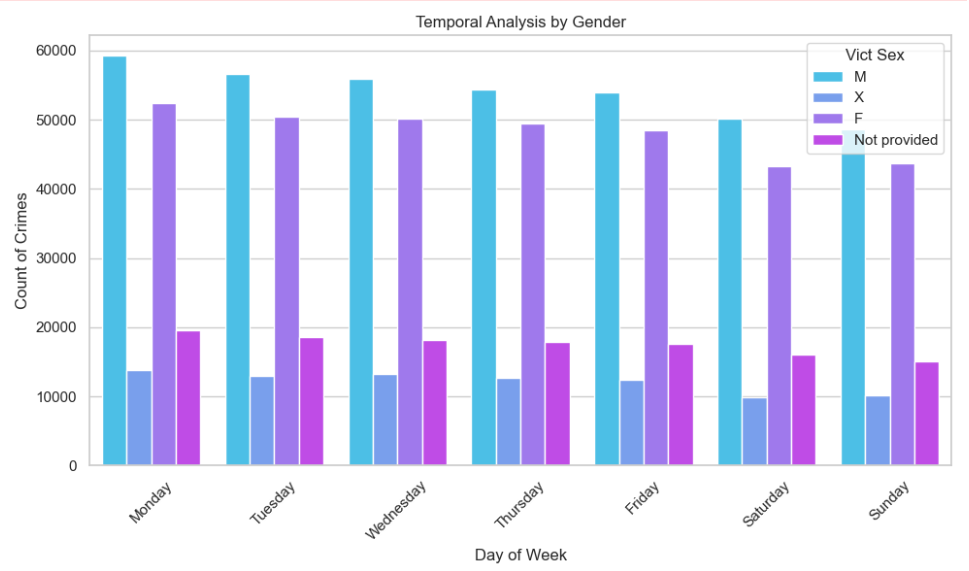


Fig. 5.4.3 Temporal analysis by gender

**Observation :** The temporal analysis of gender by week highlights a consistent trend where male occurrences are consistently higher than females, "Not provided," and "X" throughout each day of the week. Specifically, on Monday, male occurrences are nearly 60,000, slightly higher than females at approximately 50,000. Following Monday, occurrences decrease minimally but do not increase again. This suggests a persistent disparity in crime occurrence between genders, indicating potential gender-specific vulnerabilities or differential involvement in criminal activities throughout the week.

* **Temporal Distribution of crime severity over years**

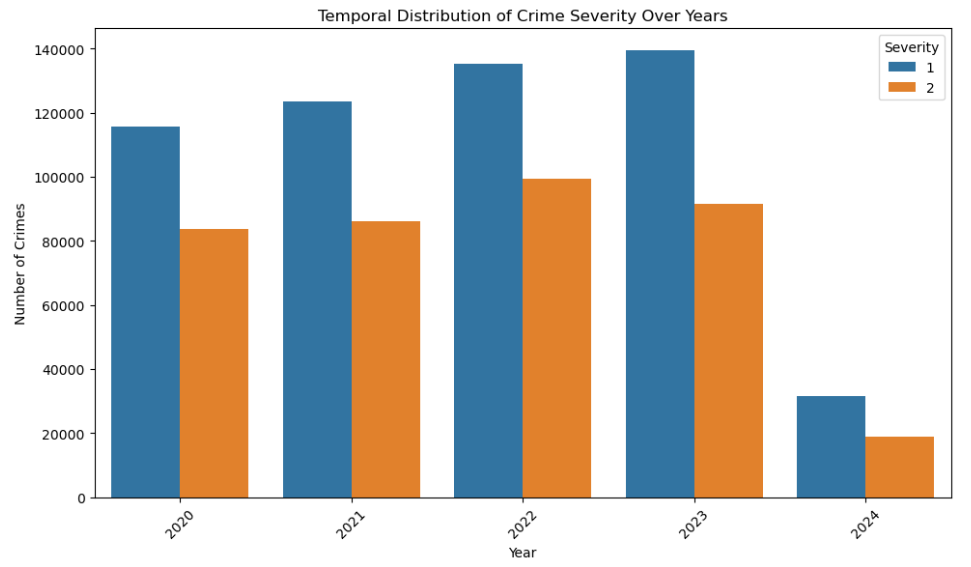


Fig. 5.4.4 Temporal distribution of crime severity over years

**Observation:** Examining the temporal distribution of crime severity across multiple years reveals fluctuating patterns. In 2023, the prevalence of unserious crimes, categorized as type "1," appears to be notably high, indicating a potential increase in less severe criminal activities during that period. Conversely, in 2022, serious crimes classified as type "2" seem to dominate, suggesting a shift towards more severe criminal incidents during that year.

* **Analysis of Seasonal variation in crime severity**

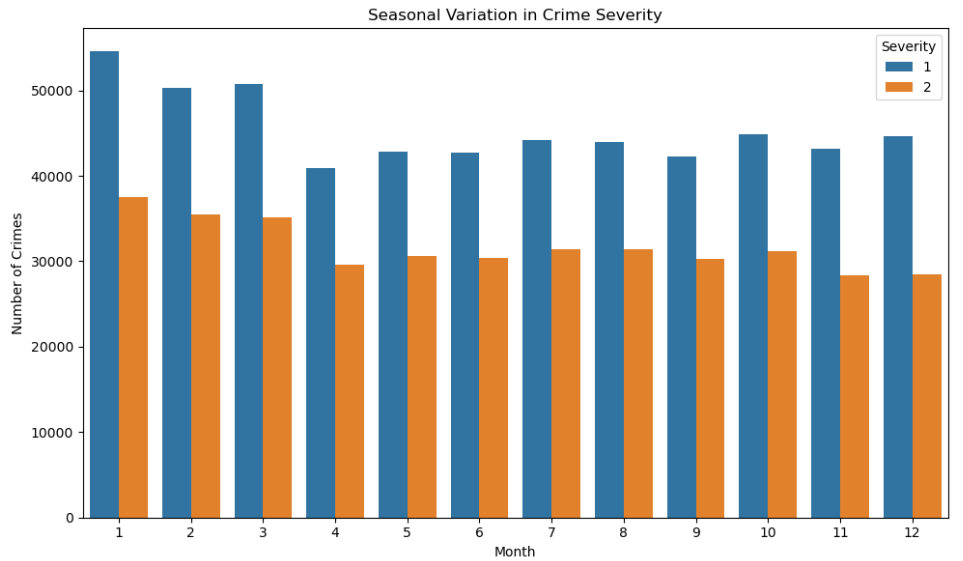


Fig. 5.4.5 Analysis of seasonal variation in crime severity

**Observation :** The analysis of seasonal variation in crime severity reveals distinct patterns for different crime types. For crime severity type "1," the highest occurrences are observed in January, with a notable decrease in April. Conversely, crime severity type "2" peaks in January, closely followed by February, indicating a consistent prevalence during the early months of the year. These findings suggest potential seasonal trends in criminal activities, with variations in occurrence across different months.

* **Crime Analysis by Time of Day**

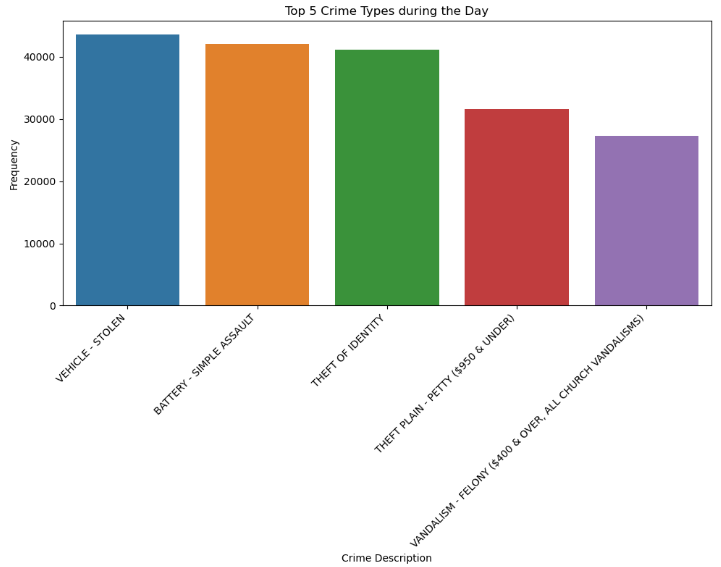


Fig. 5.4.6 Top 5 crimes during the day

**Observation :** The prevalence of crimes such as vehicle theft, battery (simple assault), identity theft, petty theft, and felony vandalism during daytime hours suggests distinct patterns of criminal activity. These offenses are often opportunistic and may occur when individuals are more likely to be out and about, engaging in daily activities. Vehicle theft, for instance, might occur in parking lots or residential areas where cars are left unattended during daytime hours. Simple assaults and theft-related crimes could occur in public spaces, retail establishments, or residential areas where people are more active. Identity theft may involve fraudulent activities during business hours when financial institutions and businesses are open, while felony vandalism might occur in both public and private properties during daylight hours when visibility is high.

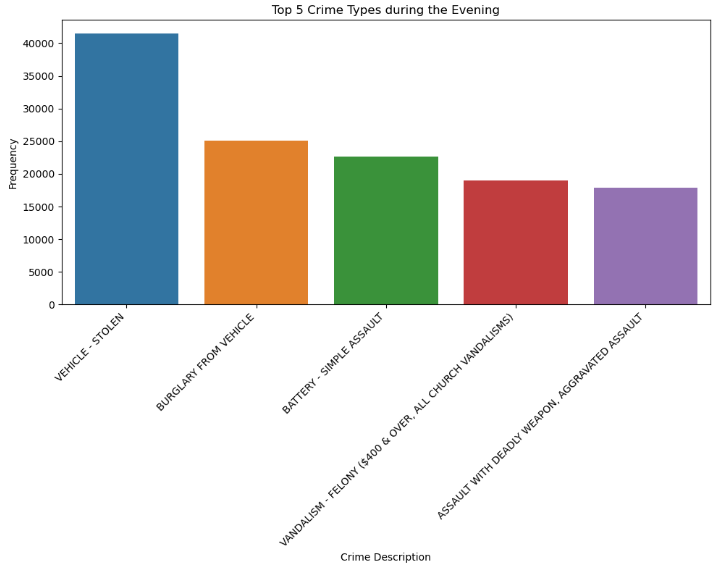


Fig. 5.4.7 Top 5 crimes during evening

**Observation** :The occurrence of crimes such as vehicle theft, burglary from vehicles, battery (simple assault), felony vandalism, and aggravated assault during evening hours suggests distinct patterns of criminal activity. These offenses often occur when individuals are returning home from work or engaging in leisure activities, presenting opportunities for criminal behavior

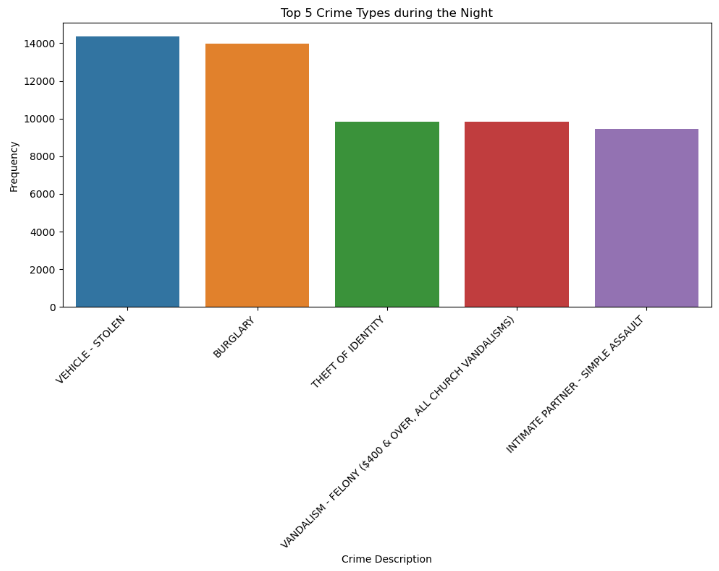


Fig. 5.4.8 Top 5 crimes during night

**Observation :** Nighttime serves as an opportune moment for criminals to engage in illegal activities under the cover of darkness. Vehicle theft and burglary often target unattended vehicles or unoccupied residences, taking advantage of reduced visibility and fewer witnesses. Identity theft may occur during the night through various means, including online fraud or theft of personal information from unsecured locations. Felony vandalism incidents, particularly those targeting churches, may occur during the night when there is minimal surveillance or security presence. Additionally, intimate partner simple assault incidents during nighttime hours might occur within private residences or domestic settings, where conflicts escalate under the cover of darkness

**Conclusion :** In conclusion, our analysis of temporal trends in crime occurrence reveals distinct patterns across different time dimensions. Monthly distribution indicates that crimes peak in January, with consistent frequencies in March and February, while November and April witness lower activity. Similarly, the weekly distribution shows Friday as the peak day for crimes, with minimal discrepancies between other weekdays. Gender-wise analysis highlights consistent male predominance in crime occurrences throughout the week, particularly on Mondays. Temporal distribution of crime severity fluctuates over years, with 2023 showing a higher prevalence of unserious crimes and 2022 dominated by serious crimes. Seasonal variation indicates peaks in January for both crime severity types, with a slight decrease in April for less severe crimes.n addition to our analysis of temporal trends in crime occurrence, the examination of crime patterns by time of day sheds light on distinct criminal behaviors observed during different hours. Daytime sees a prevalence of opportunistic crimes, while evenings witness offenses linked to leisure activities and nighttime harbors clandestine activities under reduced visibility. Understanding these temporal nuances can aid law enforcement agencies in deploying resources effectively to combat crime and ensure public safety.

**5.5 Resource Allocation Optimization**

**Problem Statement:**By leveraging historical crime data and geographical information, the objective is to predict areas with a higher likelihood of crime occurrence, thus allowing law enforcement to proactively deploy resources to prevent and address criminal activities in those locations.

**EDA:**

* Unique Areas in the dataset



Fig. 5.5.1 Unique area values in the dataset

**Observation :**The above list contains total of 21 areas falling under Los Angeles.

* **Crime Distribution by Area**

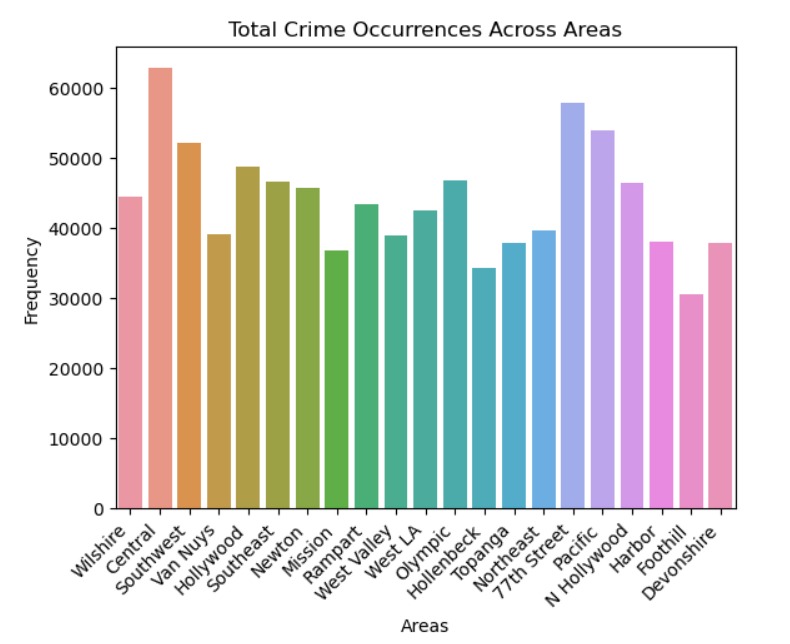


Fig. 5.5.2 Crime distribution by area

**Observation :**The data provides insight into the distribution of crimes across various areas within the jurisdiction. Central stands out with the highest number of reported crimes, totaling 62,822 incidents. Following closely behind is Pacific with 53,936 reported crimes, and 77th Street with 57,809 incidents. Other notable areas with relatively high crime rates include Wilshire, Southwest, and Hollywood. Conversely, Foothill and Mission areas have comparatively lower crime rates, with 30,602 and 36,792 reported incidents, respectively. This analysis highlights the spatial variation in crime occurrence, which can inform law enforcement agencies' resource allocation and strategic planning efforts.

* **Unique Crime Incidents by Area**

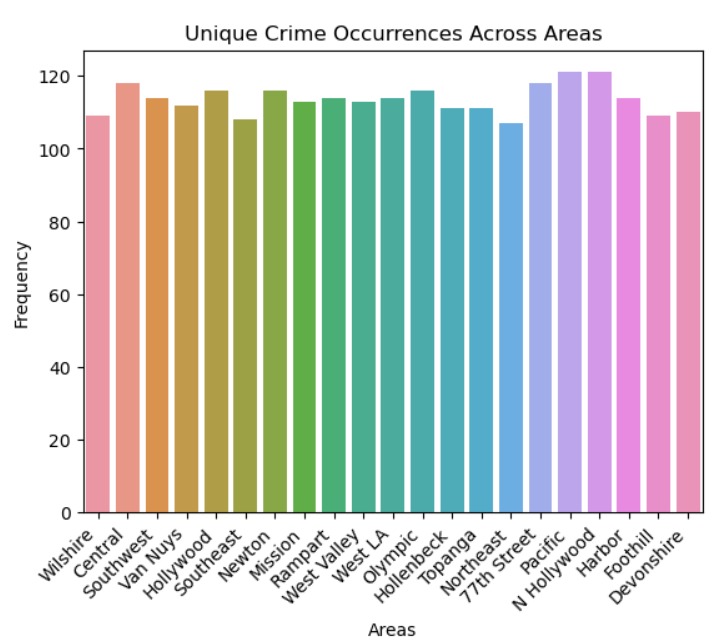


Fig. 5.5.3 Unique crime occurences acroos area

**Observation :**The data reveals the distribution of unique crime incidents across different areas within the jurisdiction. Pacific and N Hollywood areas have the highest number of unique crime incidents, each with 121 reported incidents. They are closely followed by Central and 77th Street, both with 118 unique incidents. Hollywood, Newton, Rampart, West LA, and Olympic areas also show a relatively high number of unique incidents, each with 116 reported cases. Conversely, Northeast and Foothill areas have the lowest number of unique incidents, with 107 and 109 reported cases, respectively.

**Conclusion :** Resource Allocation Optimization based on Crime Distribution

**Areas with High Occurrence of Unique Crimes:**

1. Pacific: 121 incidents
2. N Hollywood: 121 incidents
3. Central: 118 incidents
4. 77th Street: 118 incidents
5. Hollywood: 116 incidents
6. Olympic: 116 incidents
7. Rampart: 114 incidents
8. West LA: 114 incidents
9. Newton: 116 incidents

**Most Common Crimes:**

Vehicle theft, battery (simple assault), and identity theft are the most common crimes across these areas.

**Unique Crime Patterns:**

* Pacific and N Hollywood show a high incidence of vehicle theft, battery, and identity theft, indicating a recurring pattern of property and personal crimes.
* Central and 77th Street exhibit a similar trend, with a notable presence of violent crimes like battery (simple assault) and aggravated assaults.
* Hollywood and Newton have a significant occurrence of property crimes such as theft and burglary, suggesting a need for targeted interventions to address property-related offenses.
* Optimize resource allocation in Hollenbeck and Foothill with lower frequency of crime occurrence by reallocating personnel and resources to high-incidence areas where they are most needed.

**Recommendation:**

Allocate additional police and security resources to areas with high occurrence of common crimes, focusing on proactive patrols and targeted enforcement efforts to deter criminal activities.Implement community engagement programs and neighborhood watch initiatives to enhance collaboration between law enforcement agencies and local residents in crime prevention efforts.

**5.6 Predicting Victim Gender with Machine Learning**

**Problem Statement :**The aim of this study is to develop a predictive model using machine learning techniques to determine the gender of crime victims based on contextual variables extracted from historical crime data.

**Features and Target variables:** The features variables selected for the model are 'Crm Cd','AREA','Time Period','Vict\_Desc\_Cat','Vict Age' and ‘Vict\_Sex\_Cat’ is was taken as target variable.

**Models:**

1. Logistic Regression :The logistic regression model achieved an accuracy of 63% on the dataset. It exhibited moderate precision, recall, and f1-score for each class, with class 2 (unknown) having the highest scores. While the model performed reasonably well, it showed some limitations in accurately predicting classes 0 (male) and 1 (female).

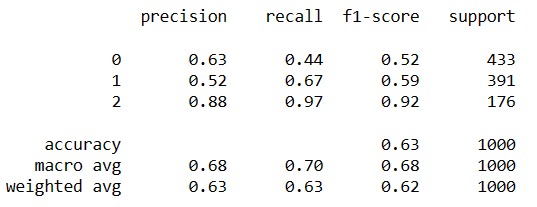


Fig. 5.6.1 Result accuracies of Logistic Regression

1. Random Forest Classifier:With an accuracy of 67%, the random forest classifier outperformed logistic regression. It demonstrated balanced precision, recall, and f1-scores across all classes, with class 2 achieving near-perfect scores. The model's ability to handle complex relationships and capture non-linear patterns in the data contributed to its superior performance.

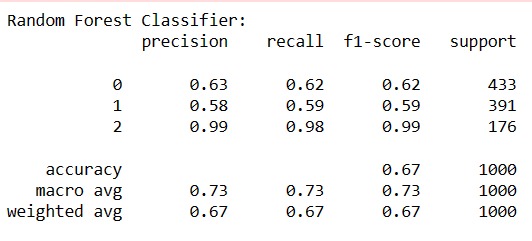


Fig. 5.6.2 Result accuracies of Random Forest Classifier

1. SVM model: The SVM classifier achieved an accuracy of 64% on the dataset. While it demonstrated competitive precision, recall, and f1-scores, particularly for class 2 (unknown), it showed some imbalance in predicting classes 0 and 1. SVMs are known for their ability to handle complex data and high-dimensional spaces, but in this case, they may have struggled to capture the nuances of the victim gender prediction task.

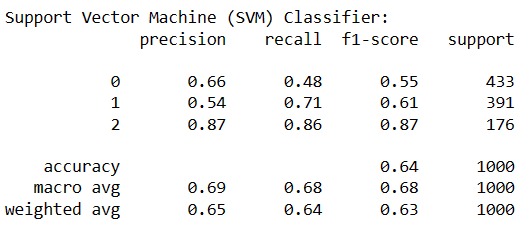


Fig. 5.6.3 Result accuracies of SVM model

1. Gradient boosting :The gradient boosting classifier yielded the highest accuracy of 68% among the models evaluated. It exhibited robust precision, recall, and f1-scores across all classes, with balanced performance for both male and female victim predictions.

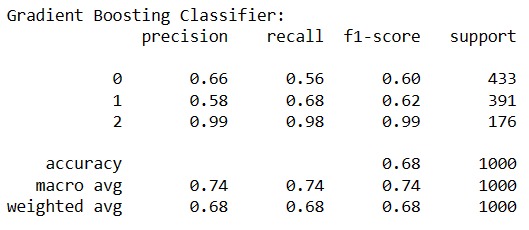


Fig. 5.6.4 Result accuracies of Gradient boosting classifier

**Conclusion :**Gradient boosting was selected as the preferred model due to its superior overall performance compared to the other classifiers. It consistently achieved the highest accuracy and exhibited balanced precision, recall, and f1-scores across all classes.

**6.Final Implementation**

For the final implementation, a web application is developed using Flask. The models, which were built using Jupyter, are transformed and saved into pickle files. These pickle files are then loaded onto the API. To use the web application and obtain results, the following steps need to be followed.

1. Launch the web app by executing the Python code, which will deploy the application on localhost.

WhatsApp Image 2024-04-30 at 7.26.01 AM

1. Open the deployed web app in a browser. If the deployment is successful the following webpage is rendered.Here we have four prediction options and on clicking any one of them you will be directed to that particular form.

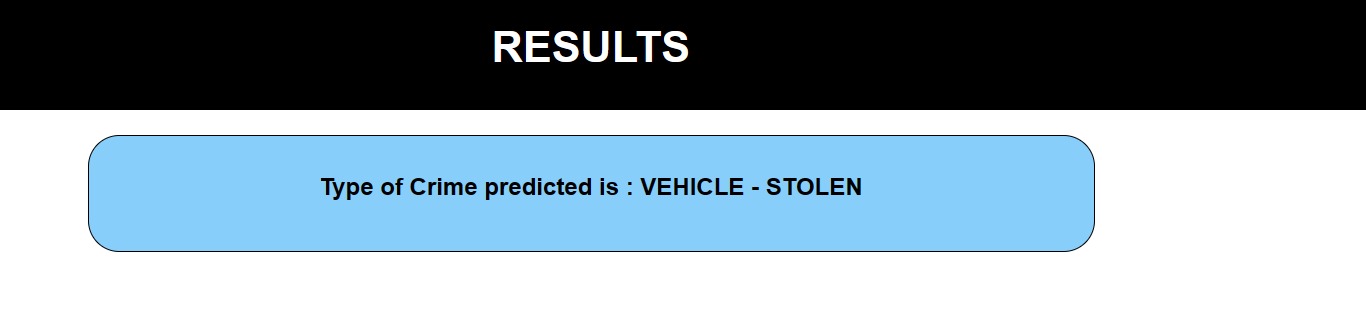


**2.1 Predict Crime Type**

* First step is to fill out all the details required in the form.For example, the below image contains the all the required field data.

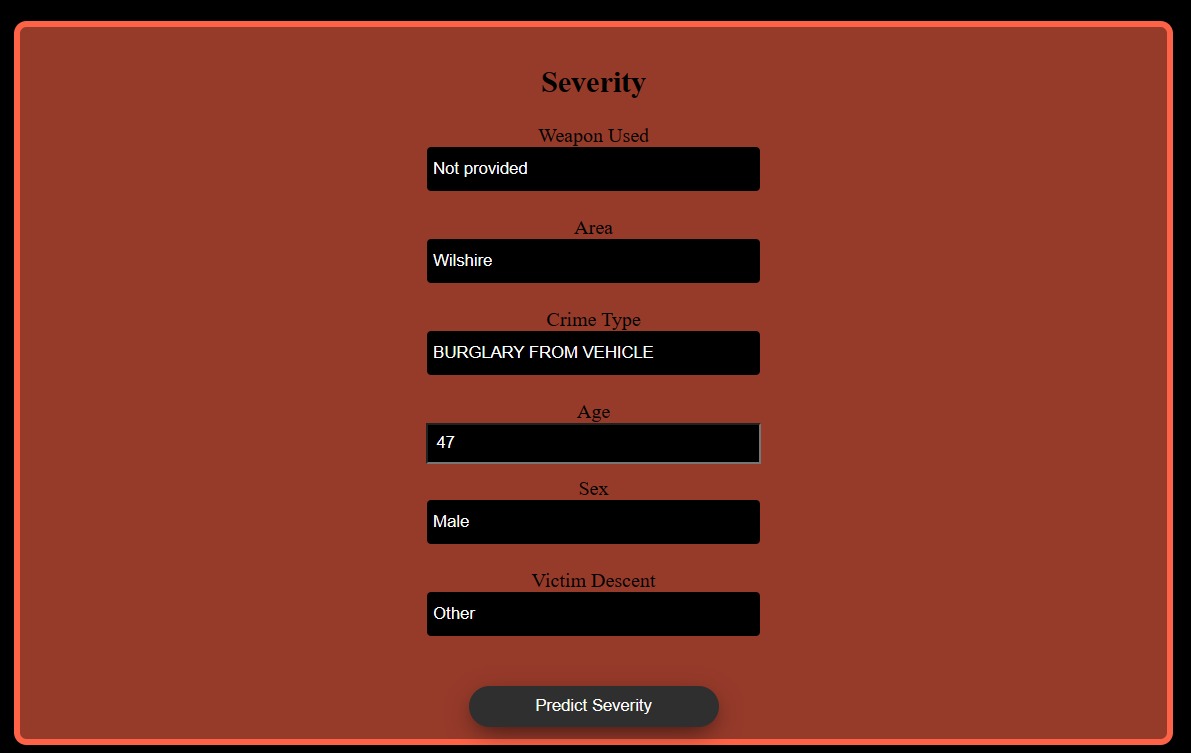


* Results : The below image is the result generated

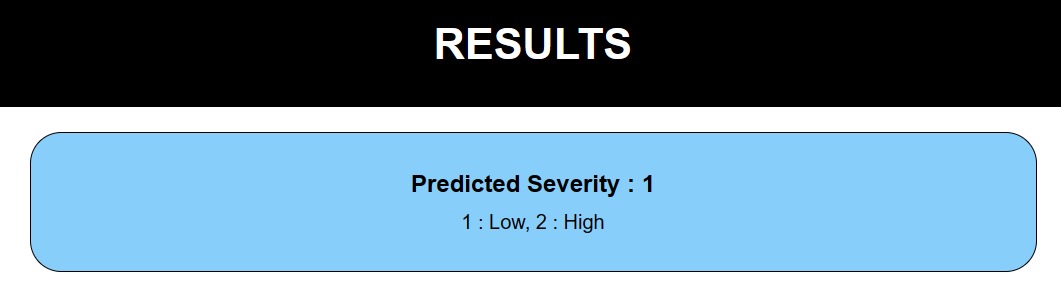


**2.2 Predict Severity**

* First step is to fill out all the details required in the form.For example, the below image contains the all the required field data



* Results : The below image displays the result generated.



**2.2 Predict Location**

* First step is to fill out all the details required in the form.For example, the below image contains the all the required field data.

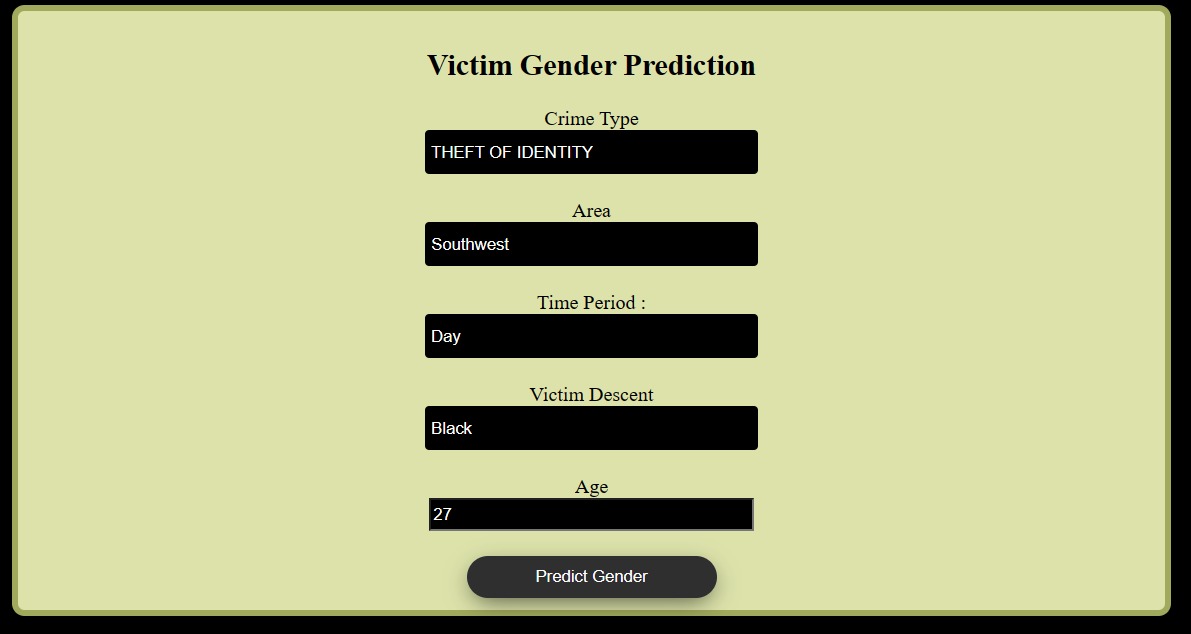


* Results : The below image displays the result generated.

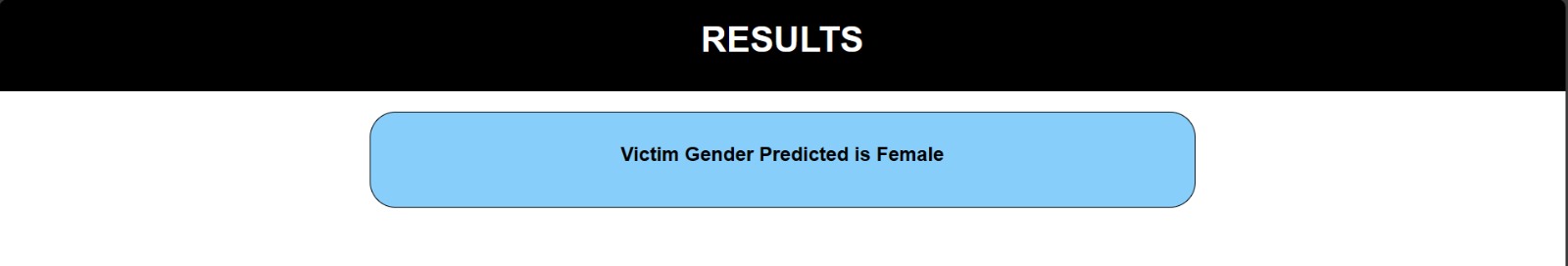


**2.4 Predict Gender**

* First step is to fill out all the details required in the form.For example, the below image contains the all the required field data.



* Results : The below image displays the result generated.



**Conclusion**

Through our comprehensive analysis of crime data, we have uncovered valuable insights into the spatial, temporal, and contextual dynamics of criminal activities. Our findings underscore the presence of distinct patterns and trends that can significantly inform law enforcement strategies and resource allocation efforts aimed at enhancing public safety and preventing crime. In our spatial analysis, certain areas such as Pacific, N Hollywood, and Central consistently exhibit higher incidences of common crimes like vehicle theft, battery, and identity theft. These identified hotspots necessitate targeted policing efforts and heightened surveillance to effectively mitigate criminal activities and safeguard community well-being. Additionally, our temporal analysis reveals notable peaks in criminal activity during specific periods, notably January and Friday, highlighting the critical importance of strategically deploying resources during these high-risk periods to deter criminal behavior and bolster response capabilities.