## Capstone-Project-4: Exploring Crime Analysis with LAPD Leveraging Machine Learning for Public Safety

PROJECT REPORT

**Data Science with Python Programming**

**By**

**Team -4**

**INDUSTRIAL PROJECT BASED LEARNING**



**Department of Computer Science and Engineering**

**Accredited by NBA**

**Geethanjali College of Engineering and Technology**

**(UGC Autonomous)**

(Affiliated to J.N.T.U.H, Approved by AICTE, New Delhi)

Cheeryal (V), Keesara (M), Medchal.Dist.-501 301.

**Team Details :**

**Team Leader:**

Shivanoor Vignesh 21R11A0597

**Team Members:**

G.Sandhya Rani 21R11A0569

R.Dhathri 21R11A0545

Hasini 21R11A0518

A.Pooja 21R11A0501

PSBN Sriya 21R11A0595

N Sai Charan 21R11A0588

MD.Akbar Ali 21R11A0586

# ABSTRACT

Crime is a complex phenomenon that affects communities worldwide, posing significant challenges for law enforcement agencies and policymakers. The Los Angeles Police Department (LAPD) often faces the task of predicting crime before it happens to better safeguard the community. The LAPD has a historical record of over a million incidences. Statistical analysis of some of these incidences can aid the LAPD in crime prevention by identifying patterns and trends. This capstone project delves into an in-depth analysis of LAPD crime data from 2020 to 2024, utilizing machine learning techniques to extract actionable insights for public safety enhancement in Los Angeles. The study comprehensively investigates temporal patterns, spatial distributions, crime type relationships, victim demographics, severity analysis, case statuses, and location details. Geospatial analysis is employed to visualize crime hotspots, enabling precise interventions and resource allocation. Ultimately, this endeavor contributes to bolstering public safety efforts and cultivating a safer environment across Los Angeles.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S No** | **Title** | **Page No** |
|  | **ABSTRACT** | 2 |
| **1** | **INTRODUCTION** | 4 |
| **2** | **PROBLEM STATEMENT** | 6 |
| **3** | **OBJECTIVES** | 6 |
| **4** | **Literature Survey** | 7 |
| **5** | **METHODOLOGY** | 8 |
|  | 5.1 Data Source | 8 |
|  | 5.2 Data Pre processing | 10 |
|  | 5.2.1 missing entries | 10 |
|  | 5.2.2 Removal of redundant colomuns | 10 |
|  | 5.2.3 Replacing the null values | 10 |
|  | 5.2.4 date and Time data processing | 11 |
|  | 5.3 Exploratory Data Analysis | 12 |
| **6** | Algorithms | 32 |
|  | 6.1 Predicting Crime Type | 32 |
|  | 6.2 Predicting Crime Severity | 32 |
|  | 6.3 Predicting Crime Area | 32 |
| **7** | **Implementation** | 32 |
|  | 7.1 Predicting Crime Type | 32 |
|  | 7.1.1 RandomForest | 32 |
|  | 7.1.2 Support vector Machine | 32 |
|  | 7.1.3 Gradient Boosting: | 33 |
|  | 7.2 Crime Severity Prediction | 34 |
|  | 7.2.1 Random Forest | 34 |
|  | 7.2.2 Support Vector Machine | 34 |
|  | 7.2.3 Gradient Boosting | 34 |
|  | 7.3 High-Crime Area Identification | 35 |
|  | 7.3.1 Random forest | 35 |
|  | 7.3.2 Gradient Boosting | 35 |
|  | 7.3.3 Support Vector Machine | 35 |
| **8** | **Final Implementation** | 37 |
| **9** | **Conclusion** | 44 |
| **10** | **Future Scope** | 44 |

# INTRODUCTION

Crime has always been an existing factor in every city but is an even bigger issue in larger cities such as Los Angeles. Crimes are the most serious threat to the human society and is increasing now-a-days in most of the cities in various parts of the world. Data mining algorithms plays an important role in predicting the number and type of crime events likely to take place in the future. Also, various other technologies and hi-tech methods may help Law Enforcement agencies to track the pattern of crimes and predict the probability of crime incidents. Though predictions cannot be 100% accurate, yet the probability for its occurrence can be detected.

The primary goal of the Los Angeles Police Department (LAPD) is to safeguard the lives and properties of the Los Angeles community, with a key objective of reducing crime rates each year. However, understanding the limits of crime reduction and identifying the factors that influence crime occurrence remains a significant challenge. This capstone project aims to analyze LAPD crime data from 2020 to 2024 to uncover actionable insights that can support law enforcement agencies and policymakers in their efforts to combat crime and improve community well-being.

The study will focus on various key variables, including the Department Report Number (DR\_NO), the date and time of crime occurrence and reporting, geographical location, crime type, modus operandi, victim demographics, and weapon usage. By examining these variables, we aim to identify trends, correlations, and factors influencing crime occurrence in Los Angeles. This dataset serves as a valuable resource for analyzing crime patterns, victim demographics, and spatial-temporal trends, thereby informing targeted interventions and resource allocation strategies to enhance public safety.



Specifically, the analysis will focus on temporal patterns, spatial distributions, and crime type relationships. Temporal analysis will uncover seasonal variations and daily/weekly patterns in crime occurrence, while spatial analysis will identify high-crime areas and potential contributing factors. Crime type relationships will help understand the interconnectedness of different types of criminal activity.

Through this comprehensive analysis, this project seeks to contribute to evidence-based decision-making processes aimed at enhancing crime prevention strategies and fostering a safer and more secure environment for all residents of Los Angeles. By identifying crime patterns, predictors, and relationships, we hope to provide actionable insights that can help the LAPD better prepare for and respond to criminal activity, ultimately improving public safety and community well-being.

# PROBLEM STATEMENT

Crime in Los Angeles poses a significant threat to community safety and well-being, impacting the quality of life for residents. Despite efforts by law enforcement agencies, understanding the intricate dynamics of crime remains a challenge. Crime is a multifaceted challenge that presents significant obstacles for law enforcement agencies and policymakers worldwide. In Los Angeles, the Los Angeles Police Department (LAPD) grapples with the daunting task of predicting and preventing crime to safeguard the community effectively. Los Angeles, like many major cities, faces a complex crime landscape. To improve public safety and resource allocation, the Los Angeles Police Department (LAPD) requires deeper insights into crime patterns, victim demographics, and the factors influencing crime.This project aims to leverage a rich crime dataset provided by the LAPD, spanning from 2020 to 2024. By performing a comprehensive data analysis, we will uncover valuable insights to support the LAPD in achieving the following objectives:

* **Identify patterns and trends:** Uncover trends in crime occurrence over time (seasonal variations, daily/weekly patterns) and spatial patterns (high-crime areas).
* **Understand victim demographics:** Analyze victim demographics like age, sex, and ethnicity to identify potential risk factors.
* **Explore crime type relationships:** Investigate how different crime types co-occur or are linked.
* **Predict crime and victim characteristics:** Develop machine learning models to predict crime types, severity, high-crime areas, and potentially even victim gender.

# 3.OBJECTIVES

* To uncover insights into crime patterns, victim demographics, spatial-temporal trends, and factors influencing crime severity.
* To provide actionable insights for law enforcement agencies and policymakers to enhance crime prevention strategies and improve public safety in Los Angeles.
* Analyze crime trends over time: Identify seasonal variations (e.g., property crimes might increase during summer) or daily/weekly patterns.
* Explore spatial patterns: Investigate high-crime areas and identify potential factors (e.g., demographics, socio-economic indicators).
* Understand crime type relationships: Analyze how different crime types might be linked or occur together.

**4. Literature Survey**

Temporal Analysis of Crime Patterns: A Review of Literature - This survey delves into existing studies that analyze how crime rates fluctuate over time, focusing on seasonal and daily/weekly trends. By understanding these patterns, researchers aim to uncover insights into the temporal dynamics of criminal behavior.

Spatial Analysis of Crime: A Comprehensive Review - This literature review explores research on the spatial distribution of crime, including the identification of crime hotspots and the examination of geographical factors influencing crime occurrence. Through this analysis, researchers seek to gain a deeper understanding of how location plays a role in criminal activity.

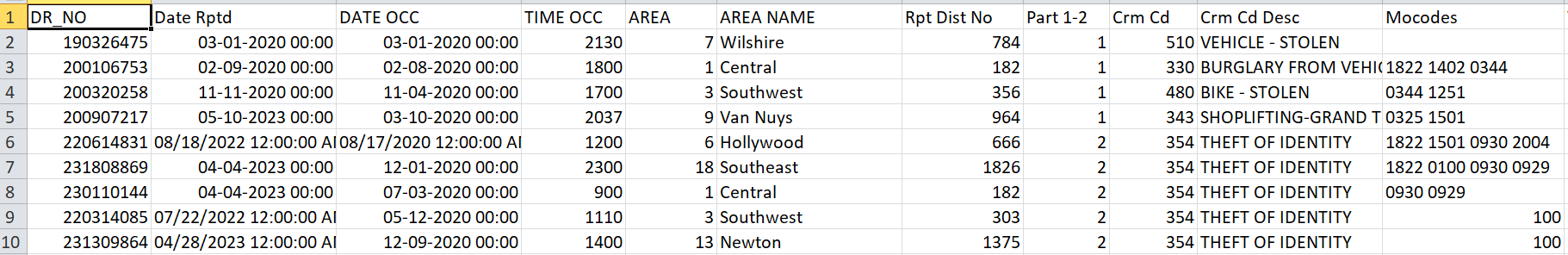
Understanding Crime Type Relationships: A Survey of Literature - This survey synthesizes existing literature on the relationships between different types of crimes, including their modus operandi. By exploring patterns in criminal behavior, researchers aim to elucidate the interconnectedness of various crime categories and their implications for law enforcement strategies.

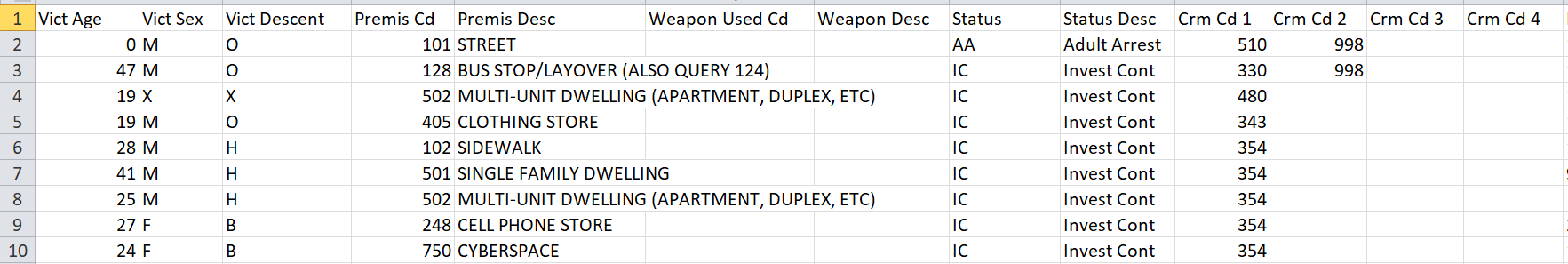
Demographic Profiling of Crime Victims: A Review of Research - This literature review examines studies that profile crime victims based on demographics such as age, gender, and ethnicity. By identifying trends and disparities in victimization rates, researchers seek to inform more equitable and targeted crime prevention approaches.

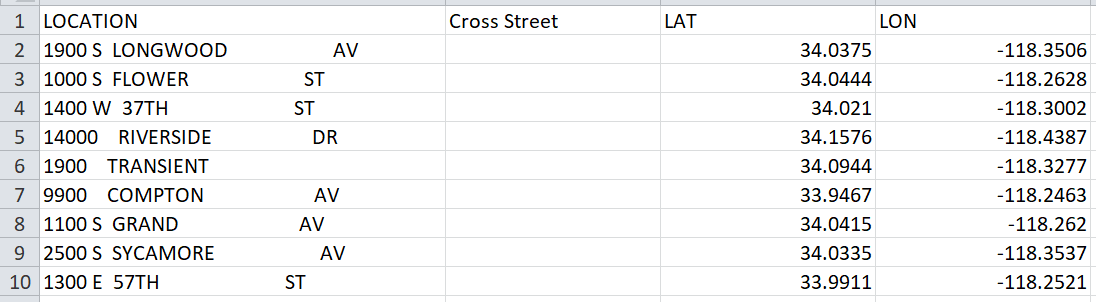
Geospatial Visualization Techniques in Crime Analysis: A Comprehensive Review - This survey investigates the use of geospatial visualization techniques in crime analysis, including mapping crime hotspots and clusters. By utilizing these techniques, law enforcement agencies and policymakers can better allocate resources and implement interventions to address crime effectively.

**5.METHODOLOGY**

**5.1 Data Source**







This data is extracted from a US Government website Data.gov.

https://catalog.data.gov/dataset/crime-data-from-2020-to-present

The daily crime occurrence in Los Angeles state is recorded and updated on the website starting from 2020 till the latest.

Columns/ Variables - 28

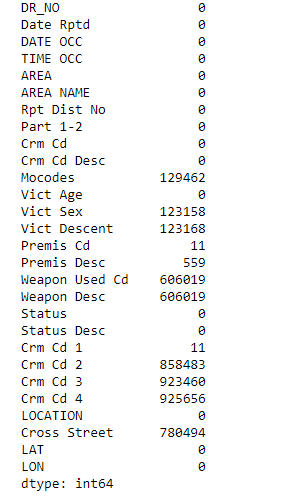
Rows/ Observations – 925721

* **Brief description of the data source**

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: Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian
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

**5.2 Pre processing the data**

**An examination of the dataset was conducted to identify missing entries**



All the null values has been replaced in each respective columns and some columns has been not replaced

**5. 2.2 Removal of Redundant Columns :**

* Due to a large number of missing entries, columns 'Crm Cd 2', 'Crm Cd 3', and 'Crm Cd 4' were removed from the dataset. Specifically, 'Crm Cd 2' had 858,483 missing values, 'Crm Cd 3' had 923,460 missing values, and 'Crm Cd 4' had 925,656 missing values out of a total of 925,721 observations. These columns were considered non-essential for further analysis and were dropped to simplify the dataset and improve its usability.

**5.2.3 Replacing the Null values**

* Replacement in 'Weapon Used Cd' and 'Weapon Desc' Columns:
* Action: Identified and replaced null values in 'Weapon Used Cd' and 'Weapon Desc' columns with 'Unknown'.
* Reasoning: By using 'Unknown', we maintain data completeness without making assumptions about the weapon used, ensuring accurate representation in analyses involving weapon details.
* Replacement in 'Vict Descent' Column:
* Action: Replaced null values in 'Vict Descent' with 'X'.
* Reasoning: Using 'X' denotes unknown or unspecified descent, preventing bias in demographic analyses towards non-disclosed data.
* Replacement in 'Vict Sex' Column:
* Action: Replaced "unknown" with 'P' in the 'Vict Sex' column.
* Reasoning: This replacement standardizes the treatment of unknown gender data, using 'P' to avoid confusion with typical gender indicators ('M' for Male, 'F' for Female, 'X' for unspecified), ensuring clear data interpretation.
* Replacement in 'Cross Street' Column:
* Action: Replaced null values in 'Cross Street' with 'Unknown'.
* Reasoning: 'Unknown' maintains dataset integrity while indicating the absence of specific location data, important for geographical analysis of incidents.

**5.2.4 Date and Time Data Processing**

* Conversion to DateTime Format: The 'Date Rptd' and 'DATE OCC' columns were converted to DateTime format using pd.to\_datetime(), ensuring standardized handling of date information.
* Extraction of Date Components: New columns ('day\_rptd', 'month\_rptd', 'year\_rptd', 'day\_occ', 'month\_occ', 'year\_occ') were created to store the day, month, and year components of both reporting and occurrence dates, allowing for detailed temporal analysis.
* Splitting Time Components: The 'TIME OCC' column was split into 'hour' and 'minute' components using str.split(). These components were then converted to integers ('occ\_hour', 'occ\_minute') to facilitate numerical analysis of time-related trends.
* These transformations were implemented using Pandas DataFrame methods and string manipulation techniques to extract and format temporal information, followed by dropping the original columns.

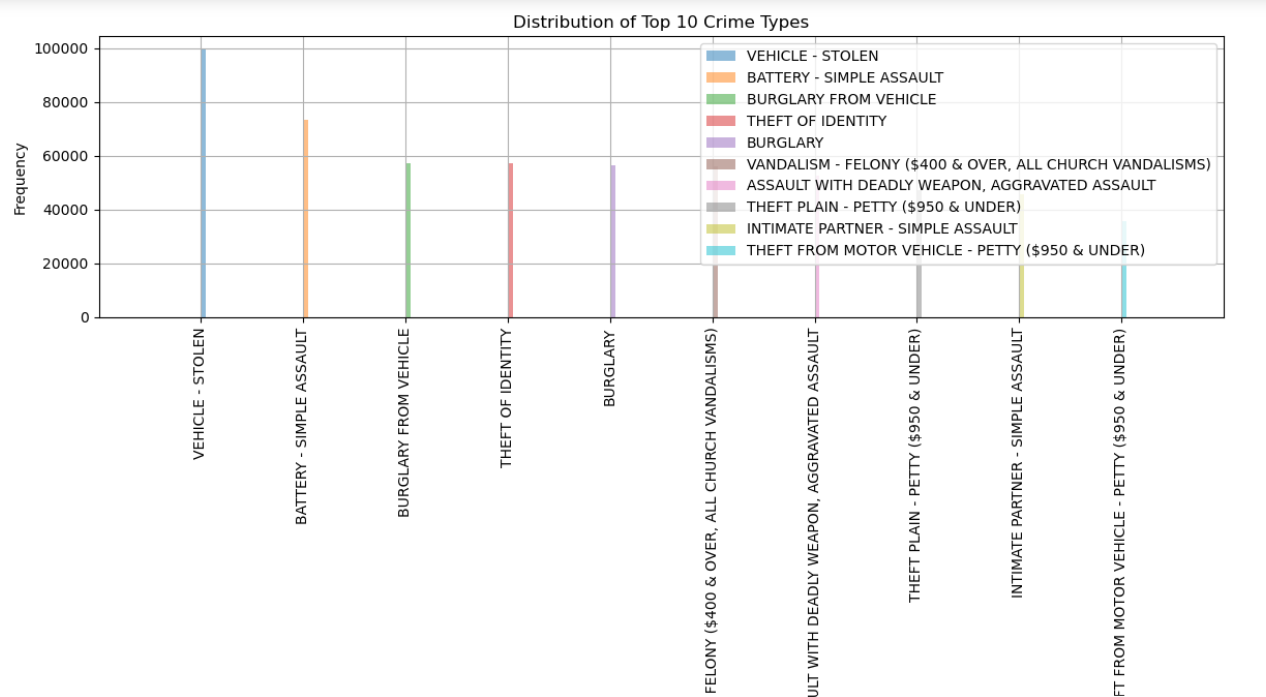
**5.3 Exploratory Data Analysis**

Exploratory Data Analysis (EDA) serves as a crucial initial phase in the machine learning workflow, aiming to comprehensively examine and comprehend the data's properties before engaging in modeling endeavors. Through EDA, analysts seek to elucidate fundamental data characteristics, including distribution, inter-variable correlations, and discernible patterns or irregularities. This preliminary exploration is pivotal as it furnishes a profound insight into the dataset, facilitating informed decisions regarding feature manipulation, data preprocessing, and model selection.

By undertaking EDA, researchers can pinpoint and rectify any missing or erroneous data, outliers, or inconsistencies, thus ensuring a more robust foundation for subsequent machine learning tasks.

**Exploratory Data Analysis:**

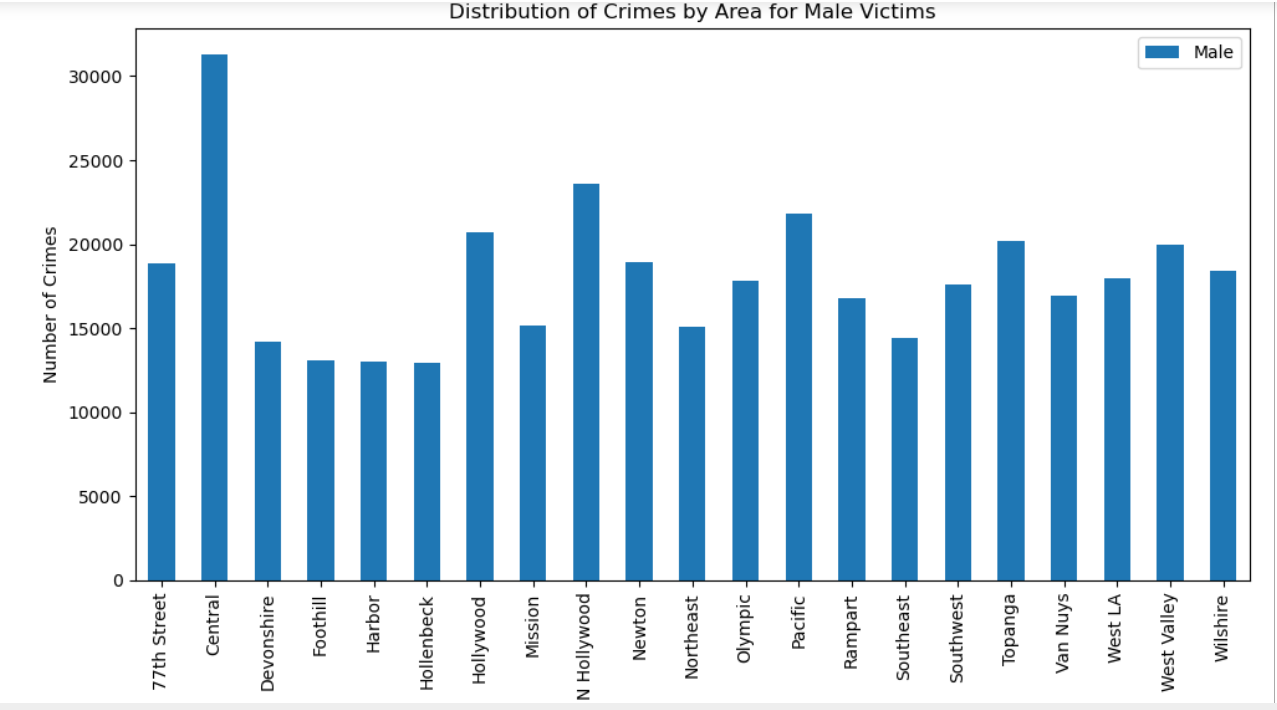
**1.Distribution of Top 10 Crime Types:**



**Observation:**

The line graph shows the distribution of the top 10 crime types. The crime type with the highest number of deaths is "VEHICLE-STOLEN" at around 100,000. The number of deaths then decreases steadily for the remaining crime types.

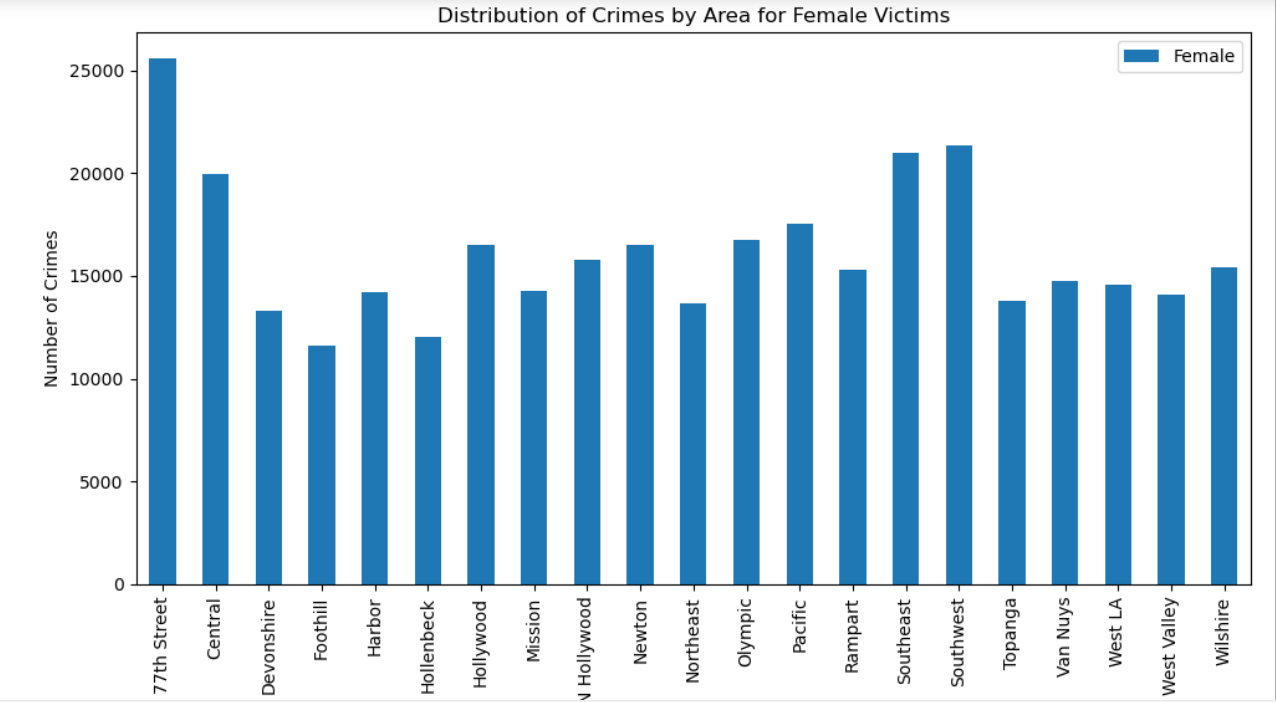
**2. Distribution of Crimes by Area for Male Victims:**



**Observation:**

The above graph represents the distribution of crimes by area for males victims. As we can observe that ‘Central’ area has many of the male victims.

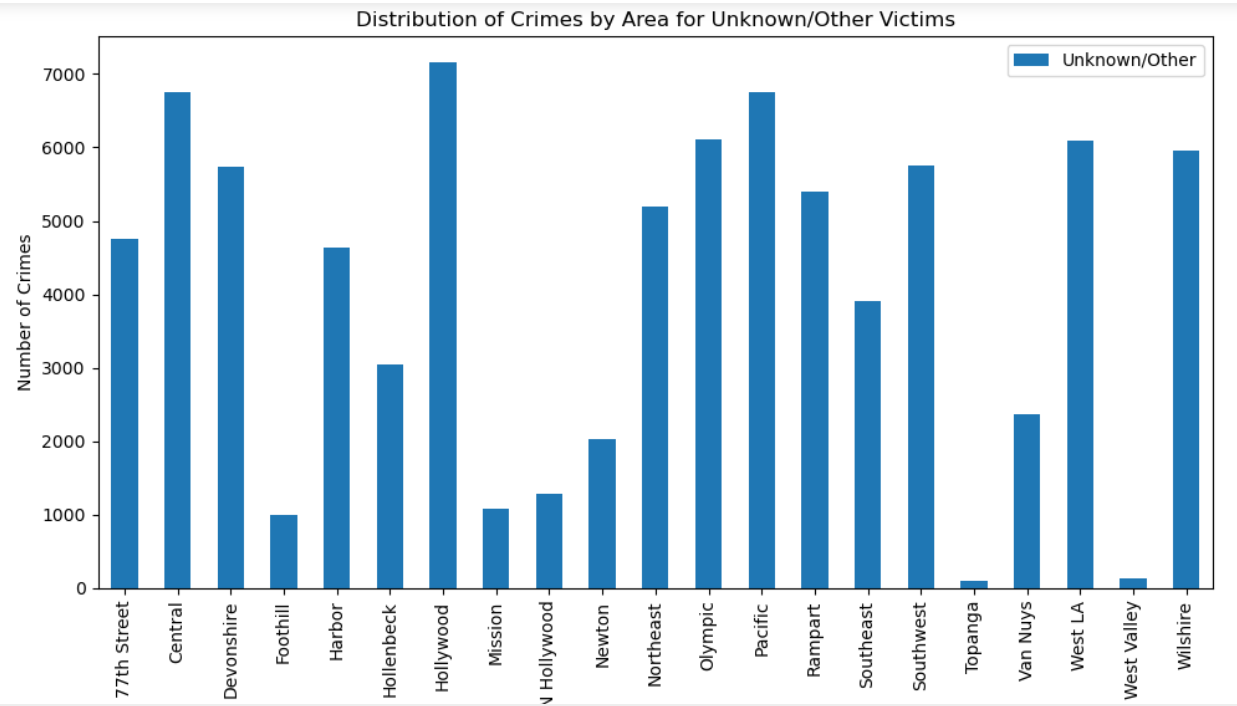
**3. Distribution of Crimes by Area for Female Victims:**



**Observation:**

The above graph represents the distribution of crimes by area for females victims. As we can observe that ‘77th Street’ area has many of the female victims.

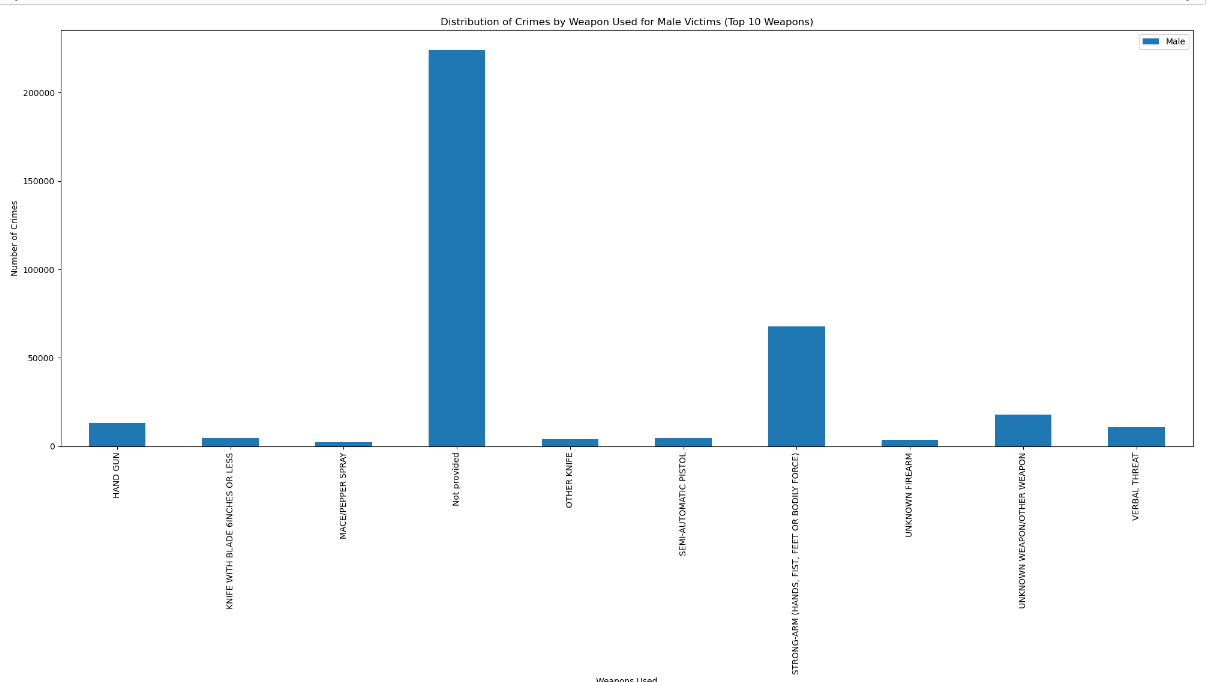
**4. Distribution of Crimes by Area for Unknown/Other Victims:**



**Observation:**

The above graph represents the distribution of crimes by area for Unknown/Other victims.As we can observe that ‘Hollywood’ area has many of the Unknown/other victims.

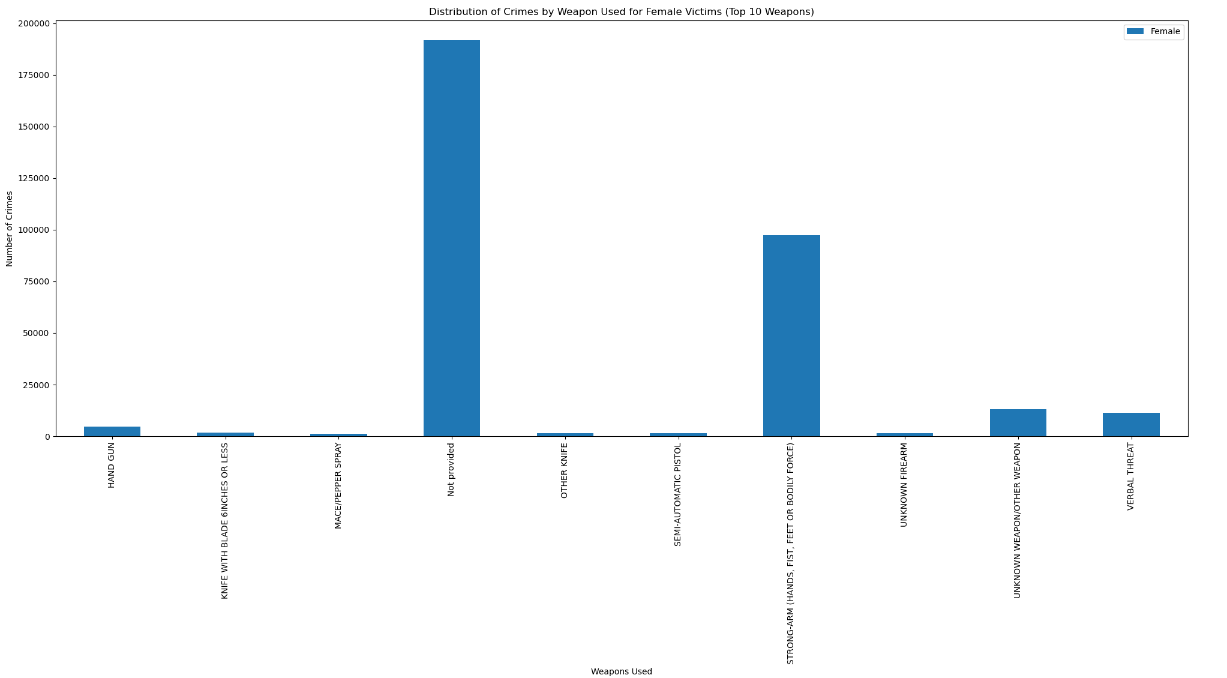
**5. Distribution of crimes by Weapons Used for Male Victims:**



**Observation:**

The graph above illustrates the distribution of crimes by the ten most frequently used weapons for male victims. From the graph, it's evident that the highest number of crimes have occurred with the weapon labeled "Strong-Arm". Therefore, based on the data provided, "Strong-Arm" has caused the most crimes for male victims.

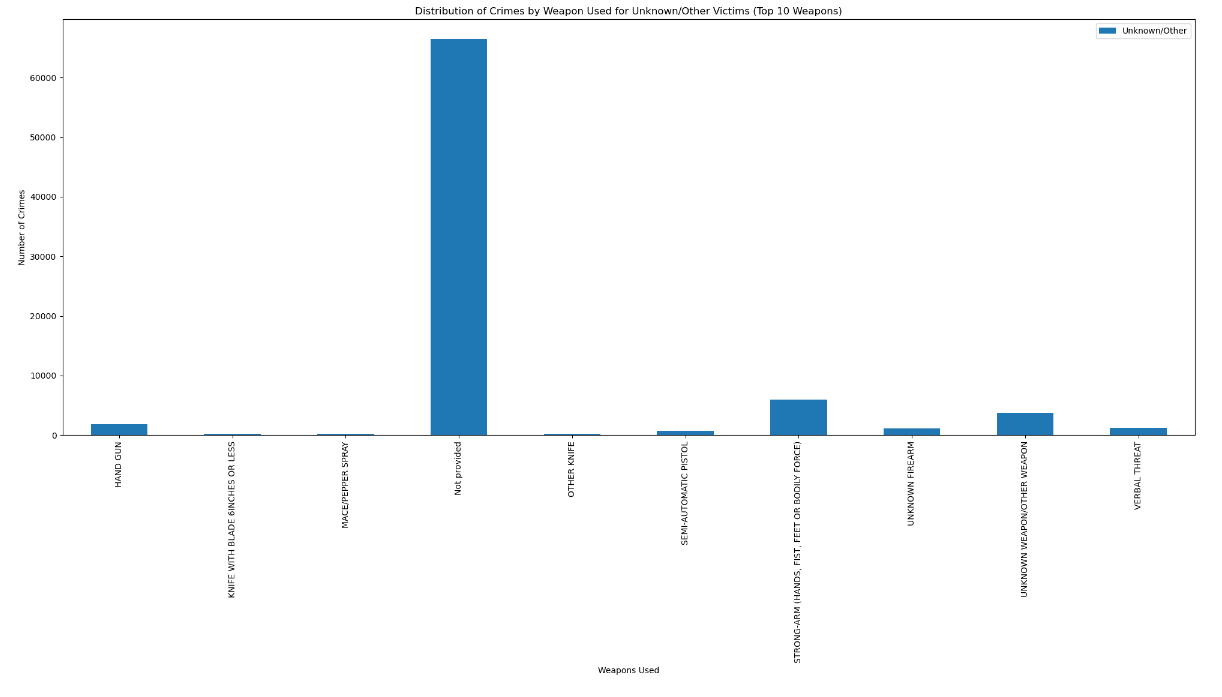
**6.Distribution of Crimes by Weapons Used for Female Victims:**



**Observation:**

The graph represents the distribution of crimes for weapons used in incidents involving female victims. Notably, the graph closely resembles the distribution observed for male victims. Consequently, we can conclude that the weapon labeled "Strong-Arm" has caused the most crimes for female victims, mirroring the pattern observed for male victims.

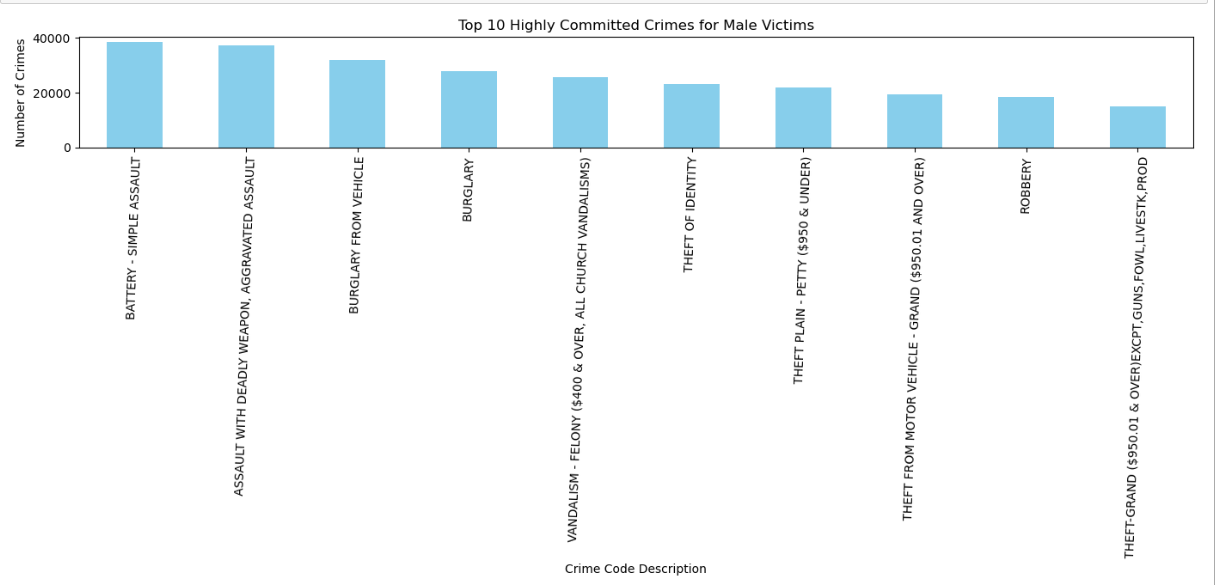
**7.Distribution of Crimes by Weapons Used for Unknown/Other Victims:**



**Observation:**

From the graph above, it's evident that the weapon labeled "Strong-Arm" has been associated with a higher number of crimes.

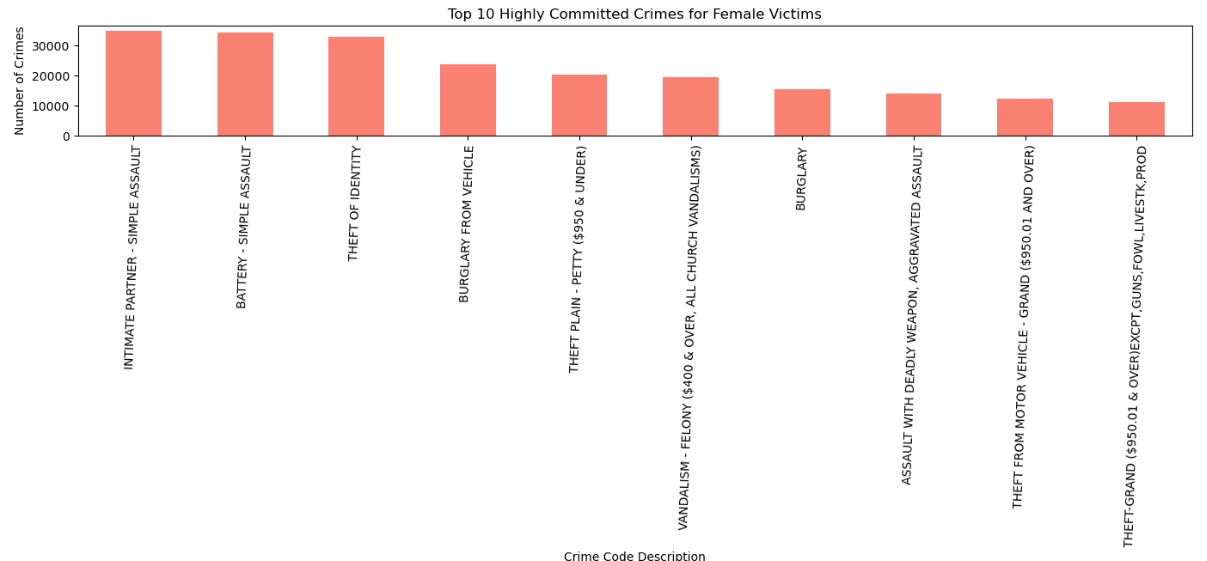
**8. Distribution of Highly committed crimes for male victims:**



**Observation:**

The presented graphs illustrate the distribution of highly committed crimes for male victims. It's apparent from the graphs that the crime code description labeled "Battery - Simple Assault" has accounted for the highest number of crimes involving male victims.

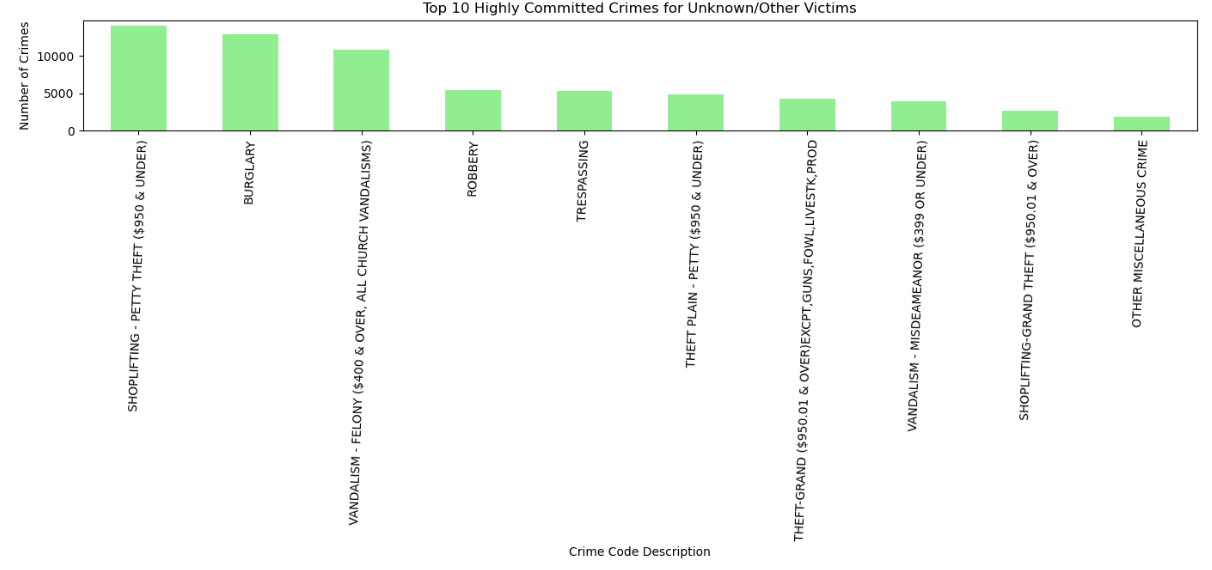
**9. Distribution of Highly committed crimes for female victims:**



**Observation:**

The presented graphs illustrate the distribution of highly committed crimes for female victims. It's apparent from the graphs that the crime code description labeled "Intimate Partner-Simple Assault" has accounted for the highest number of crimes involving female victims followed by “Battery -Simple Assault” next.

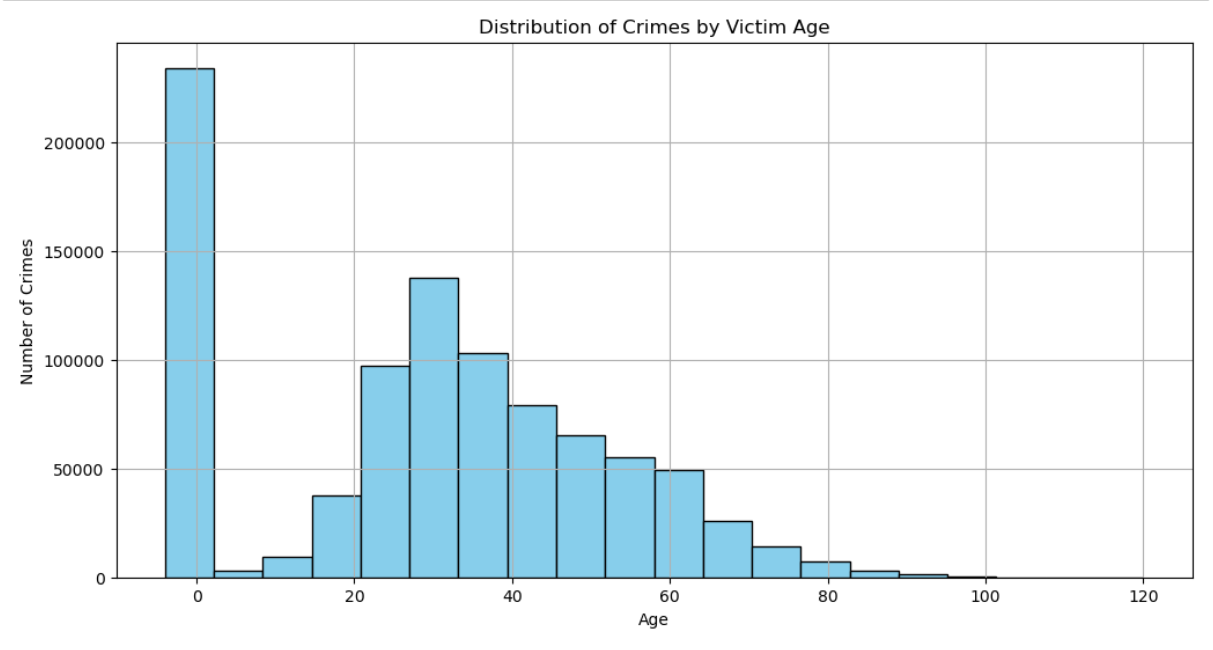
**10.Distribution of Highly committed crimes for Unknown/Other victims:**



**Observation:**

The presented graphs illustrate the distribution of highly committed crimes for unknown/other victims. It's apparent from the graphs that the crime code description labeled "SHOPLIFTING-PETTY THEFT" has accounted for the highest number of crimes involving Unknown/Other victims followed by “BURGLARY” next.

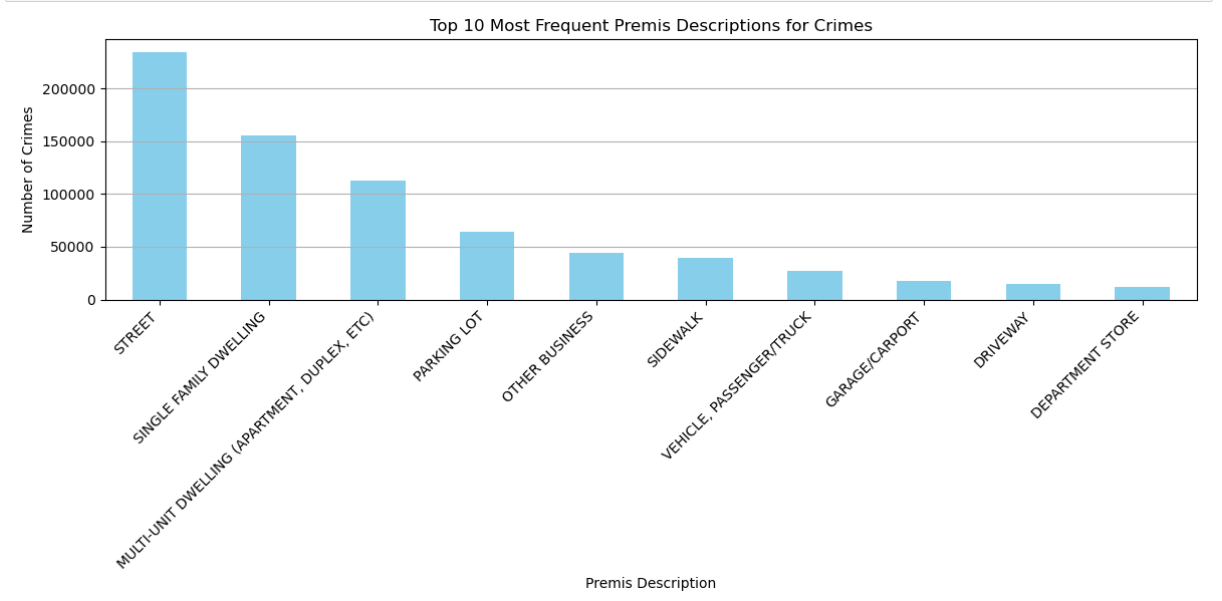
**11.Distribution of Crimes by Victim Age:**



**Observation:**

The above graph depicts the Distribution of crimes by victims age group.If we consider age groups of male,femal and other victims as whole,we can observe that the around 30 age grouped people are victims for most of the crimes.

**12. Distribution of Premis Description for crimes:**



**Observation:**

The above is a line graph titled "Top 10 Most Frequent Premis Descriptions for Crimes." The x-axis shows the premise description, and the y-axis shows the number of crimes.

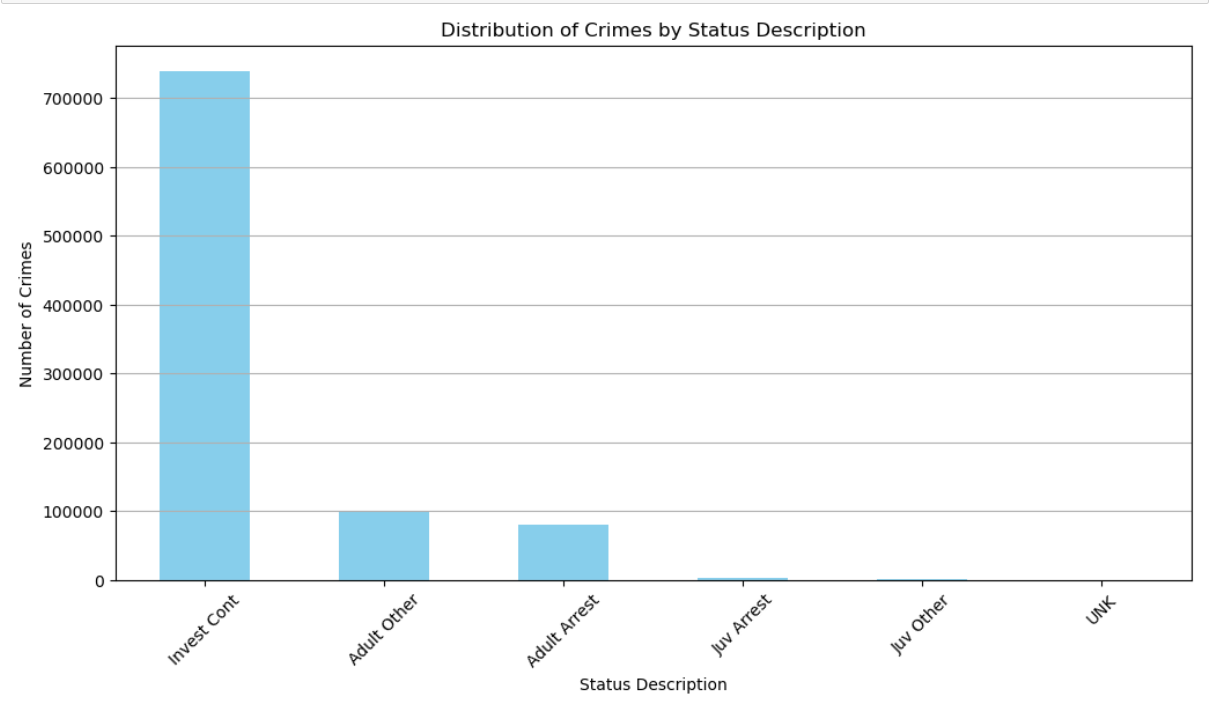
· The most frequent crime premise is on the street, with over 200,000 crimes.

· This is followed by single-family dwellings and multi-unit dwellings (apartment, duplex, etc.), both with around 150,000 crimes.

· Parking lots and other businesses are the next most frequent locations for crimes, each with around 100,000 crimes.

·The least frequent crimes on the list are those that occur in garages/carports, driveways, and department stores, each with less than 50,000 crimes.

**13. Distribution of Crimes by Status Description:**



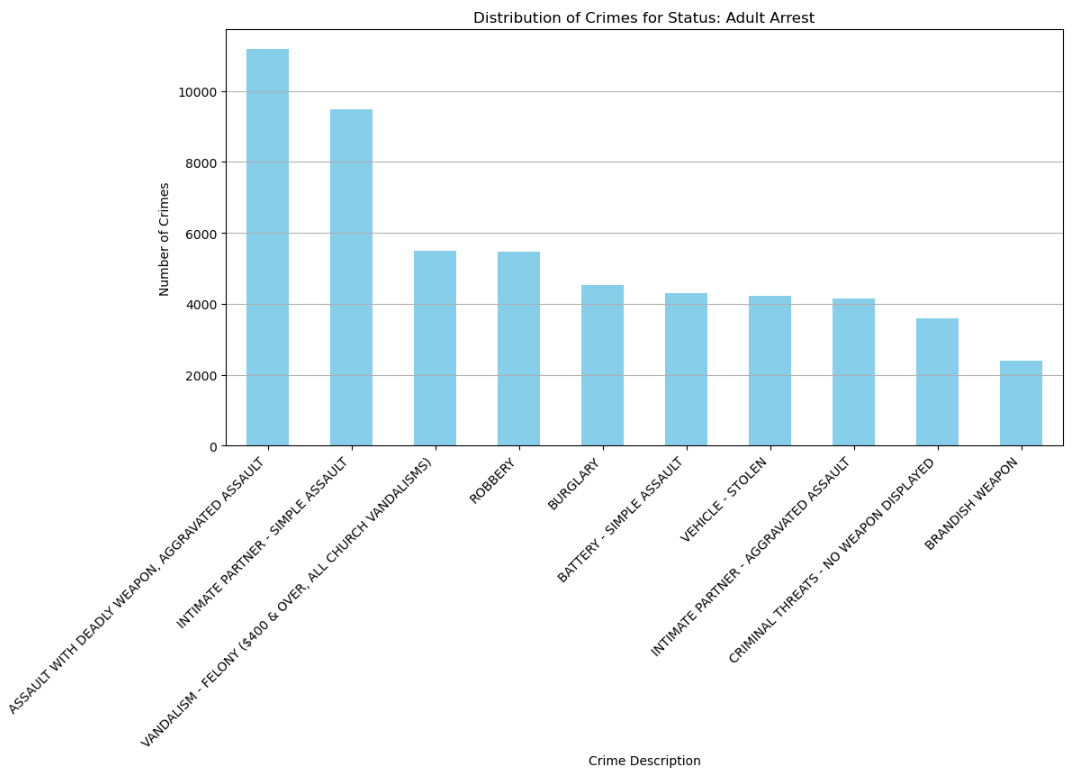
**Observation:**

The graph shows the distribution of crimes by status description.AS we can observe that,the most common status description is "Investigation Continues" with over 700,000 crimes.

There are significantly fewer "Adult Other" and "Adult Arrest" crimes compared to "Juv" categories.

" juvenile Arrest" and " juvenile Other" are the least frequent categories.

**14. Distribution of Crimes for Status: Adult Arrest-**



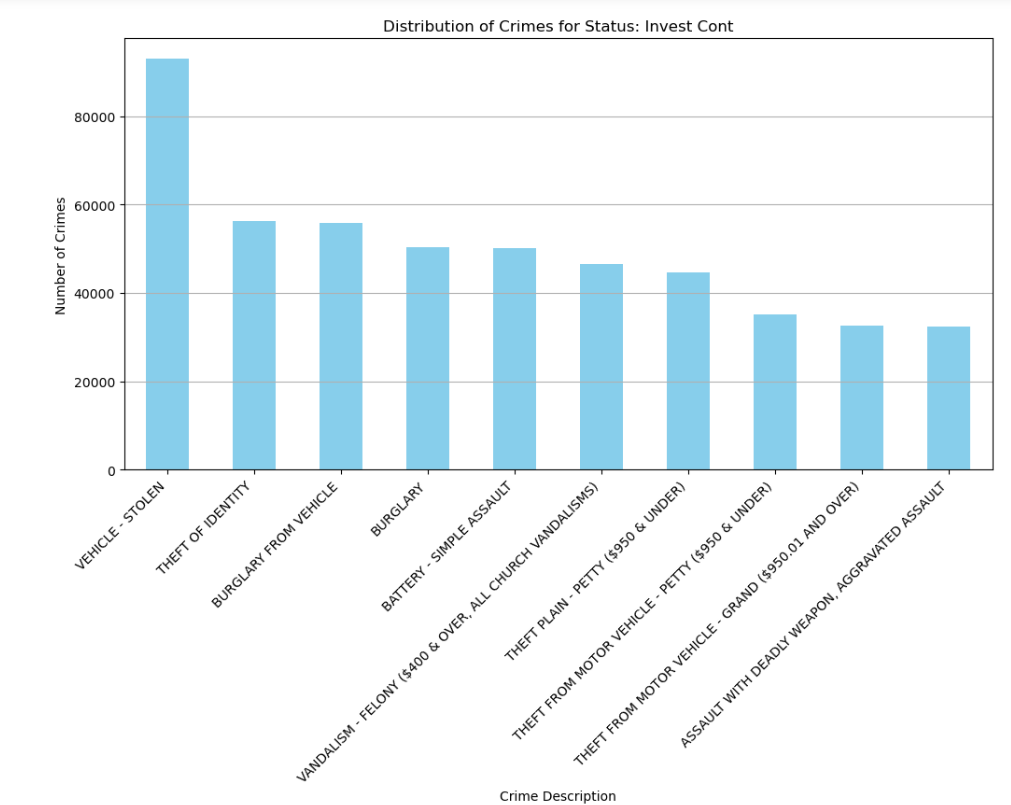
**Observation:**

· The graph titled "Distribution of Crimes for Status: Adult Arrest" shows the number of crimes arrested for by adults in different crime categories.

· Assault with a deadly weapon is the most common crime, followed by Intimate Partner-Simple Assault ,vandalism felony and robbery.

· There are fewer arrests for intimate partner aggravated assault , criminal threats with no weapon displayed and Brandish Weapon compared to other crimes.

**15.Distribution of Crimes for status:Investigation Continues**



**Observation:**

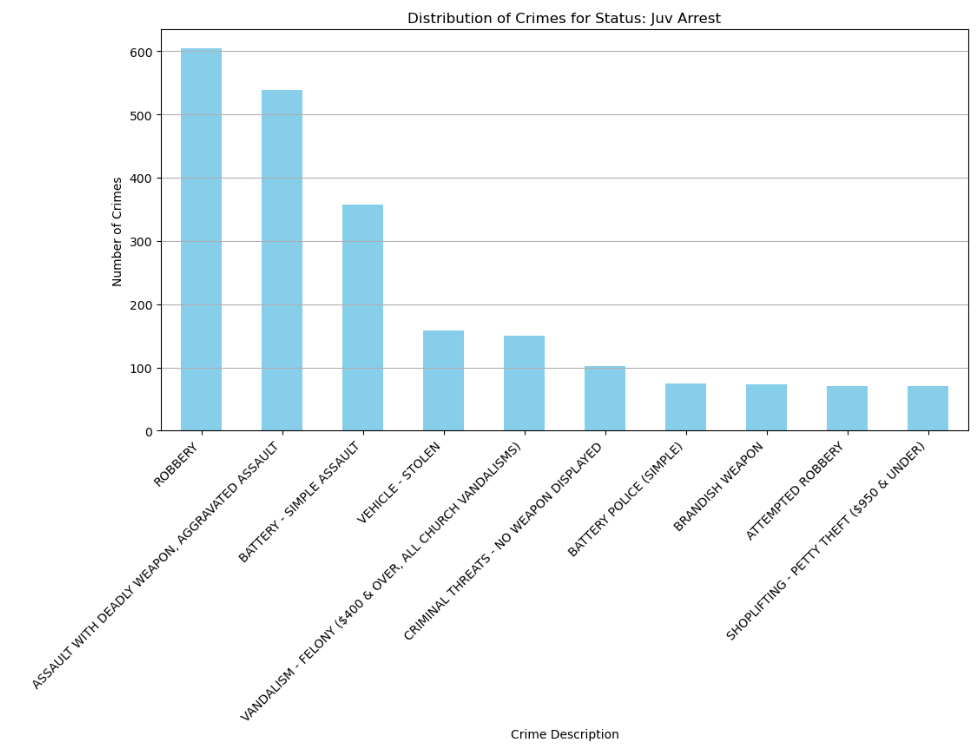
· The graph titled "Distribution of Crimes for Status: Invest Continues" shows the number of crimes committed where the investigation status is 'Invest Continues'. 'Invest Continues' likely refers to "investigation continues".

· The most common crime type is "VEHICLE STOLEN" at around 80,000.

· There are a significant number of thefts, including thefts of identity, thefts from motor vehicles, and plain thefts.

· There are also a substantial number of assaults and burglaries.

**16. Distribution of Crimes for Status: juvenile arrests**



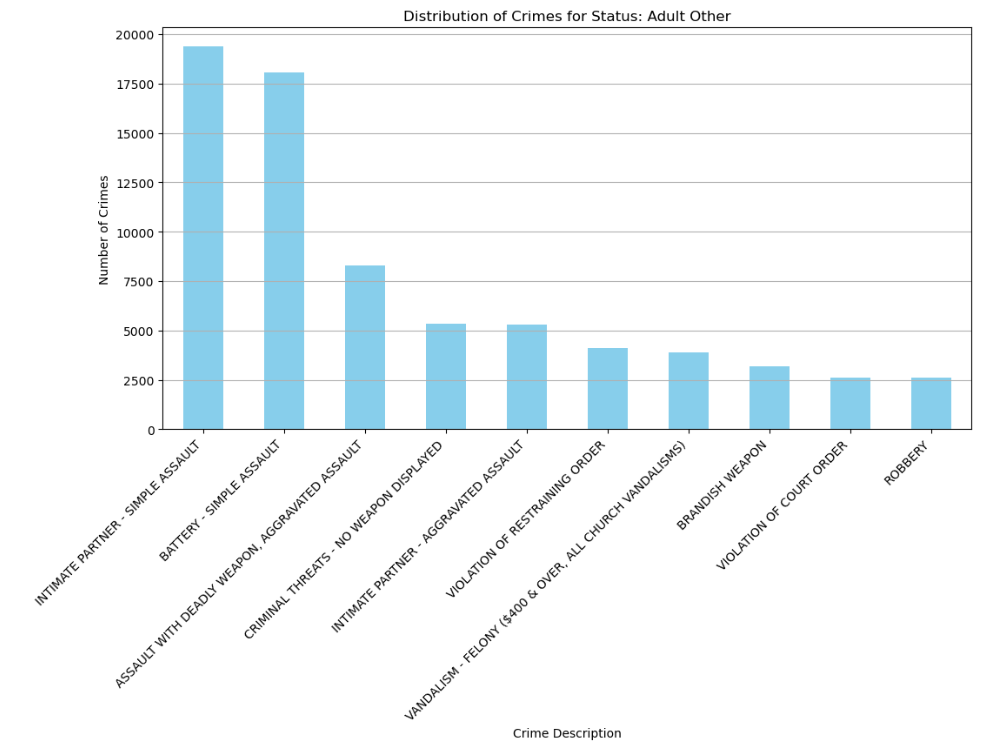
**Observation:**

· The graph titled "Distribution of Crimes for Status: Juv Arrest" shows the number of juvenile arrests for various crime types.

· Simple assault is the most common crime for juvenile arrests, followed by vandalism felony and shoplifting.

·There are fewer juvenile arrests for robbery, arson, and criminal damage compared to other crimes.

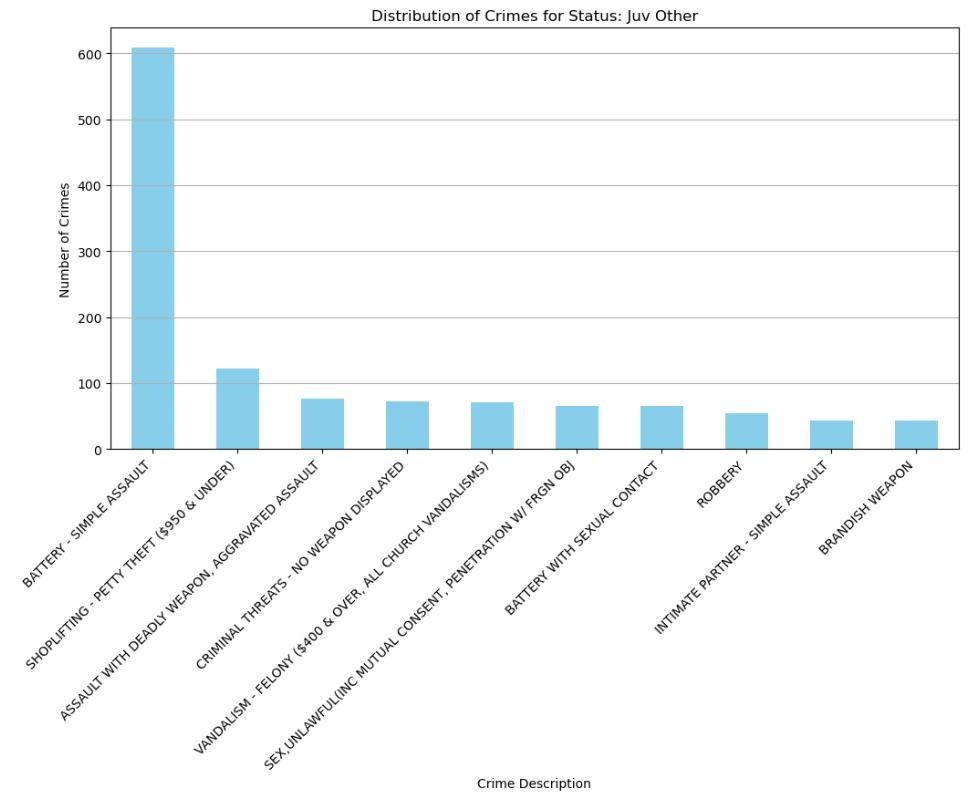
**17. Distribution of Crimes for Status :Adult Other-**



**Observation:**

The above graph says that Distribution of Crimes for Status :AdultOther where Intimate Partner -simple Assuault has highest number of crimes more than 17500 nearly 20000 , and least criimes has the robbery which has count of 2500

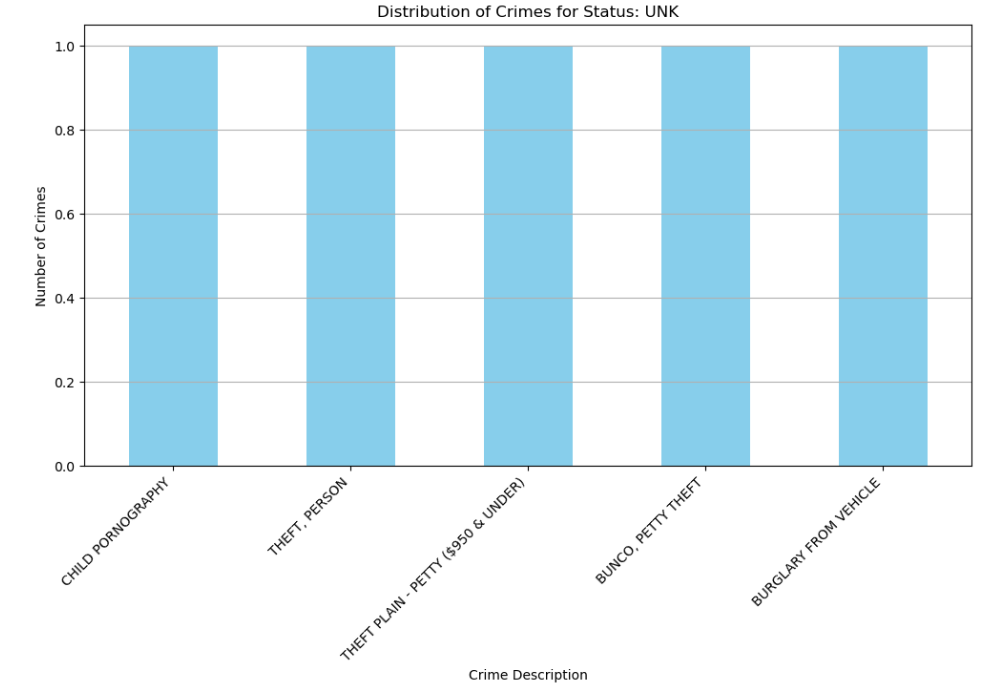
**18.Distribution of Crimes for Status: juvenile Other**



**Observation:**

* The above graph says that Distribution of Crimes for Status: juvenile Other
* The y-axis represents the Number of crimes ranging from 0 to 600 and x-axis lists various crime Description The most frequent crime listed is **“BATTERY-SIMPLE ASSAULT,”** with over 500 incidents.
* All other crimes listed have significantly fewer incidents, all under 100.

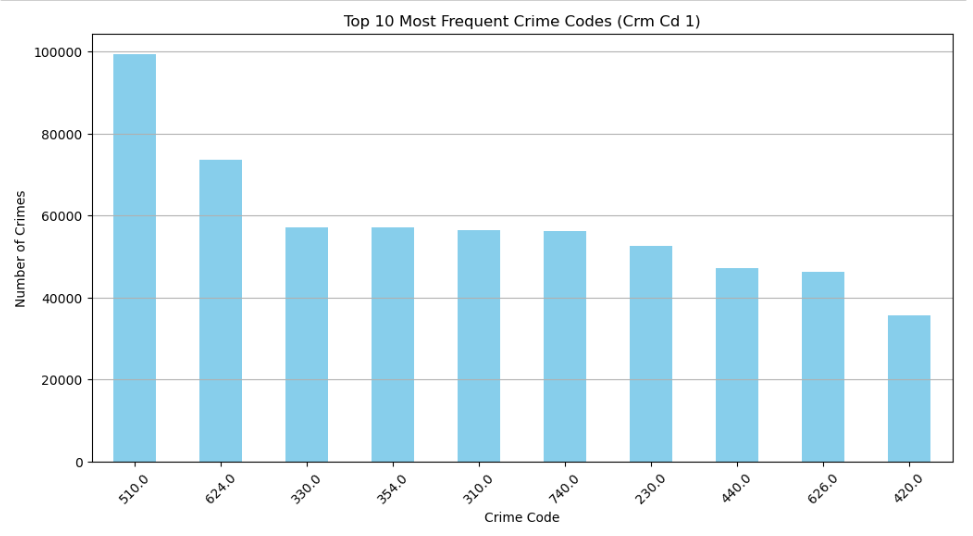
**19.Distribution of Crimes for Status:Unknown**



**Observation:**

* The above graph says about Distribution of Crimes for Status: Unknown
* The chart shows an equal distribution of crimes across six categories: CHILD PORNOGRAPHY, THEFT (PERSON), THEFT FROM AUTO ($200 AND UNDER), BURGLARY - POTTY THEFT, and BURGLARY FROM VEHICLE.
* Each category has the same height in bars, indicating that the number of crimes reported for each type is similar.

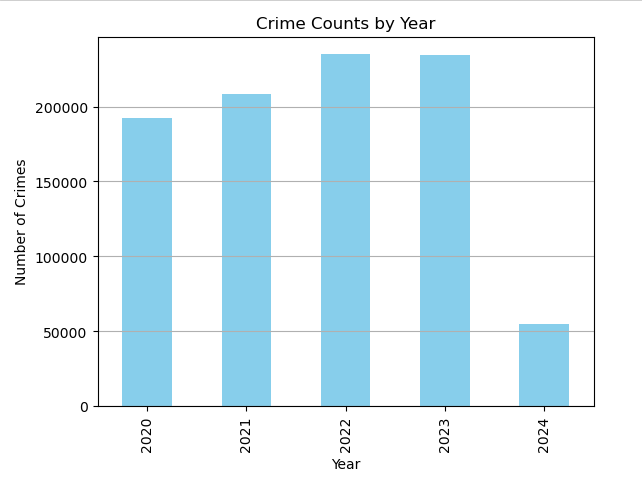
**20.Distribution of most frequent crime codes:**



**Observation:**

The graph above illustrates the distribution of the top 10 most frequent crime codes. It's noticeable that crime code 510.0 has notably higher occurrences compared to others.

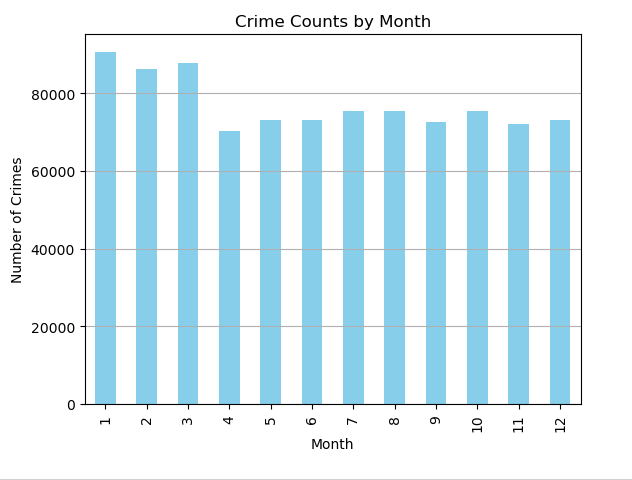
**21.Distribution of Crime Counts by Year:**



**Observation:**

The graph above displays the number of crimes over the years. It is evident that the years 2022 and 2023 exhibit nearly identical occurrences of crimes, with both years displaying the highest crime rates compared to other years.

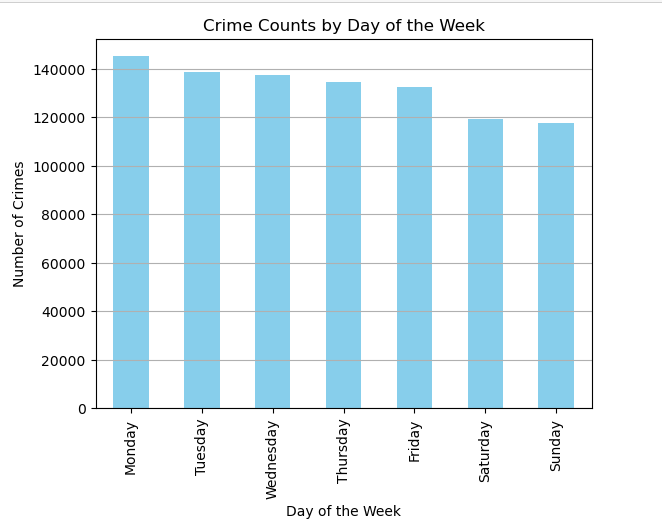
**22. Distribution of Crime Counts by Month:**



**Observation:**

The graph above illustrates the number of crimes over the months. It's notable that the month of January stands out with significantly high crime counts.

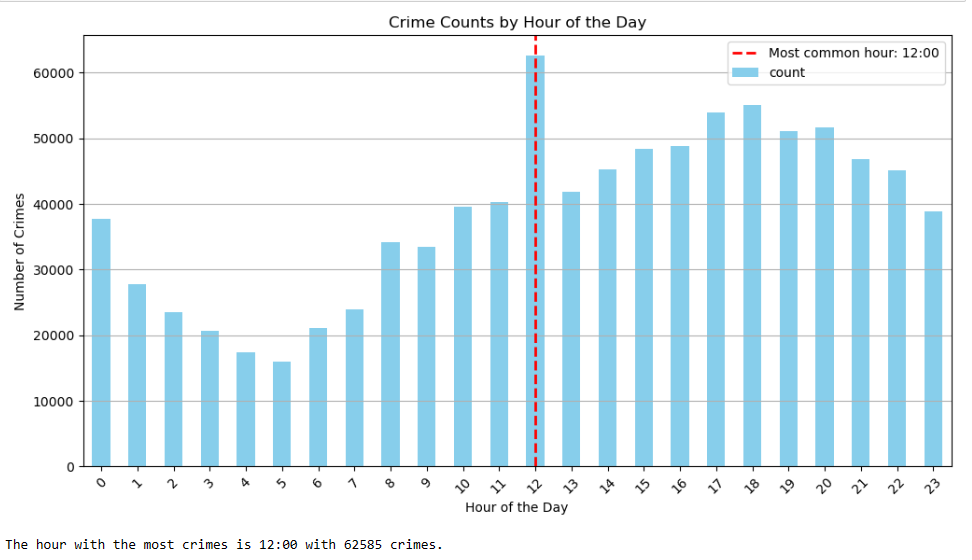
**23. Distribution of Crime Counts by Day of the week:**



**Observation:**

The graph above shows the distribution of crime counts by day of the week. It's notable that Monday stands out with significantly higher crime counts compared to other days.

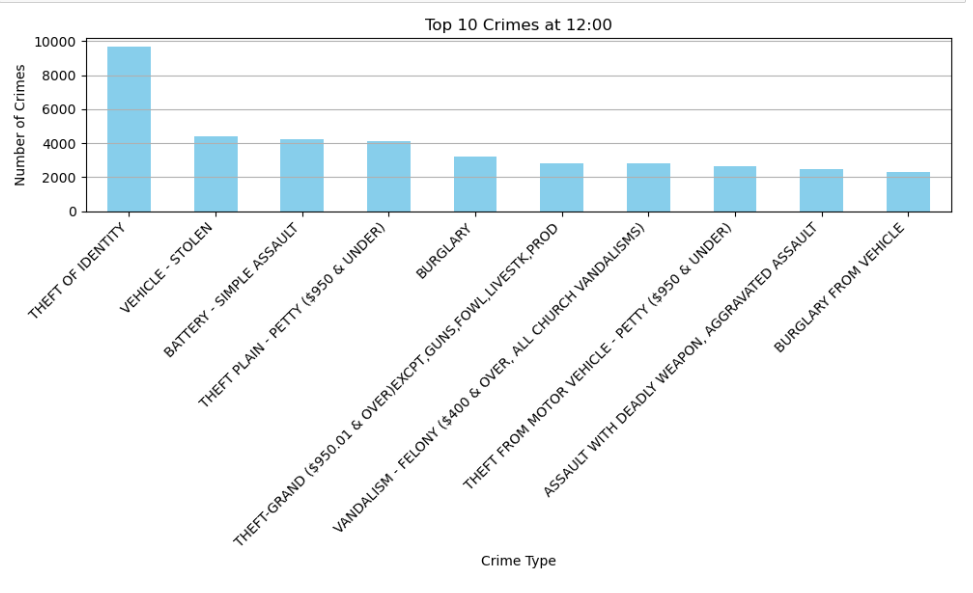
**24 . Distribution of Crime Counts by Hour of the Day:**



**Observation:**

The graph above illustrates the crime counts by hour of the day. It's evident that the most common hour for the highest crime occurrence is at 12 o'clock with 62585 crimes.

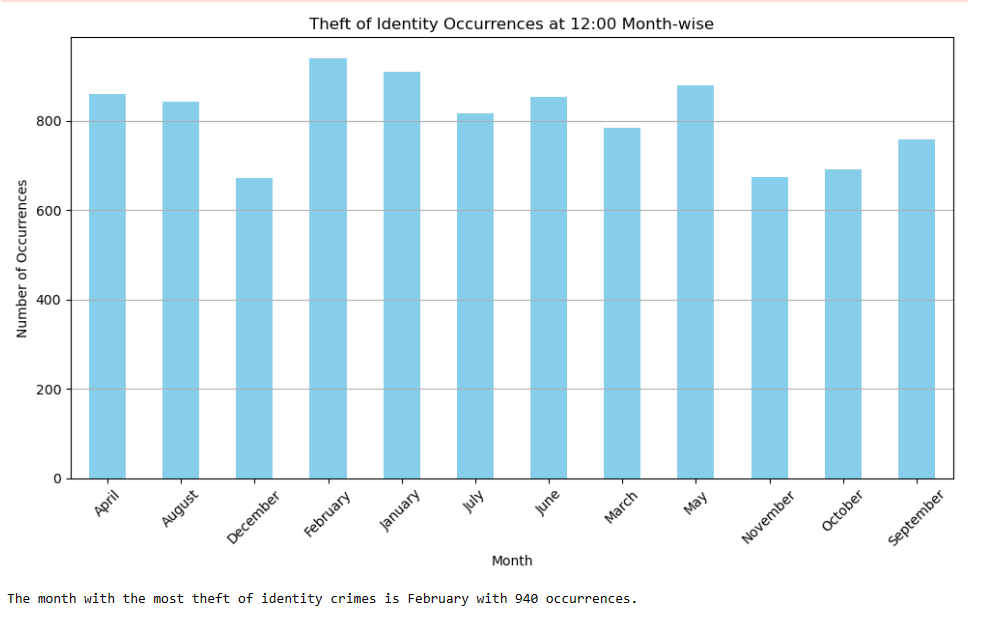
**25. Distribution of Top 10 Crimes at 12:00 :**



**Observation:**

The graph above shows the distribution of top crimes at 12:00 o'clock. It's apparent that the crime type labeled "THEFT OF IDENTITY" has the highest number of occurrences.

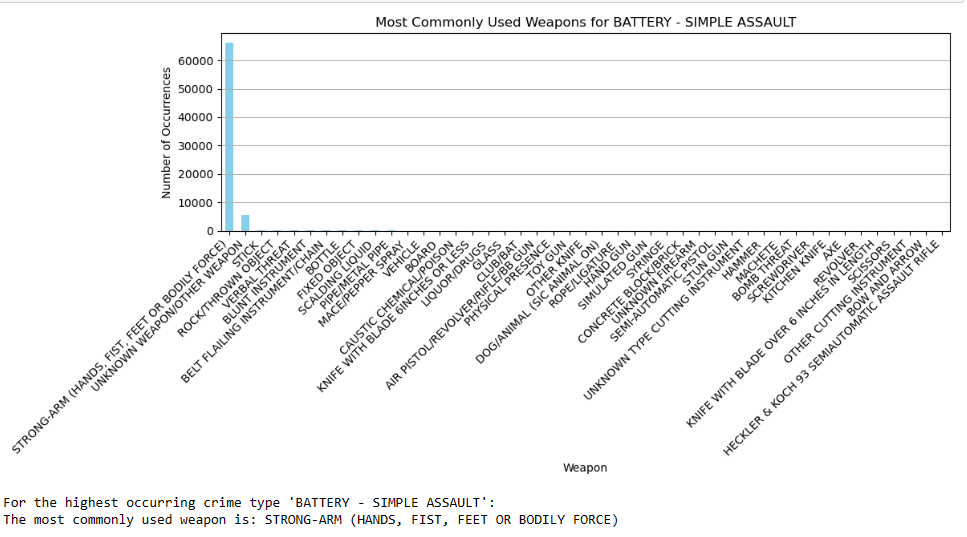
**26. Distribution of The Theft of Identity Occurrences at 12:00 Month-wise:**



**Observation:**

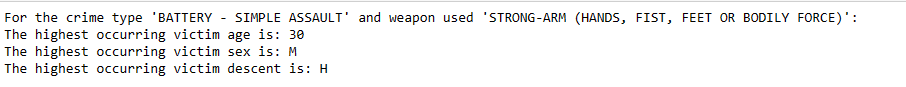
The graph above illustrates the occurrences of the crime type "Theft of Identity" at 12:00, broken down by month. It's evident that the month with the highest number of theft of identity crimes is February, with 940 occurrences.

**27.Distribution of Most Commonly Used Weapons for Battery-Simple Assault:**



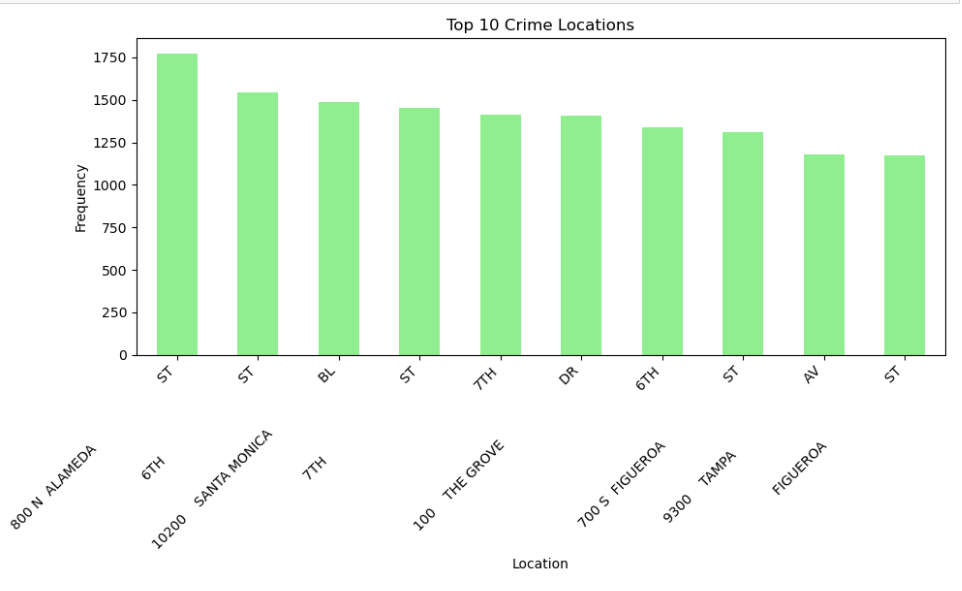
**Observation:**

The above graph depicts the distribution of most commonly used weapon for Battery-Simple Assault.As we can see that weapon named “STRONG-ARM(HANDS,FIST,FEET OR BODILY FORCE” has significantly more occurrences.



And for highest occurrences of the crime type “BATTERY - SIMPLE ASSAULT” and weapon used “STRONG-ARM(HANDS,FIST,FEET OR BODILY FORCE” ,the highest occurring victim age is 30,victim sex is male and descent is H.

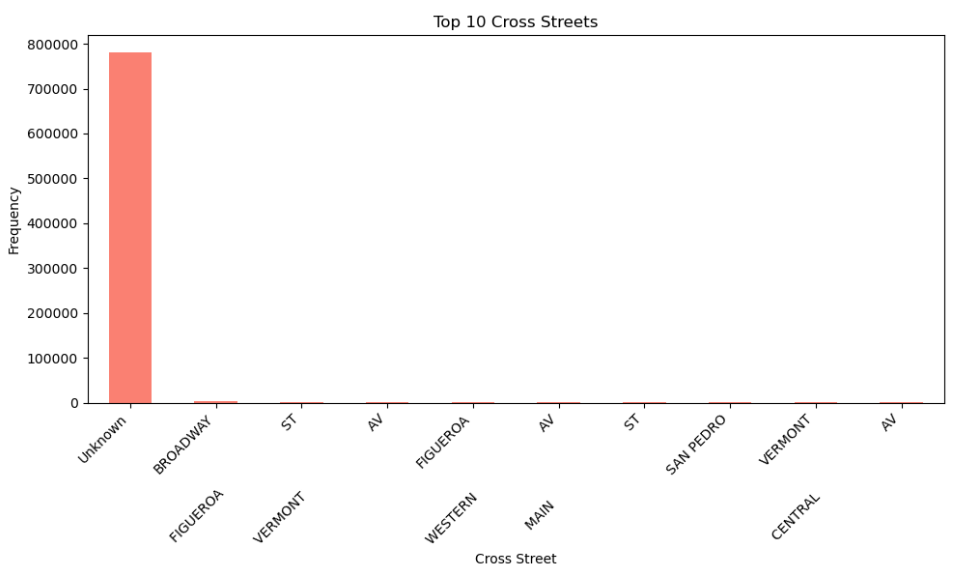
**28. Distribution of Top 10 Crime Locations:**



**Observation:**

The graph above illustrates the locations where most crimes have taken place. It's apparent from the graph that the location labeled "800 N ALAMEDA ST" has significantly more crimes, followed by "6TH ST".

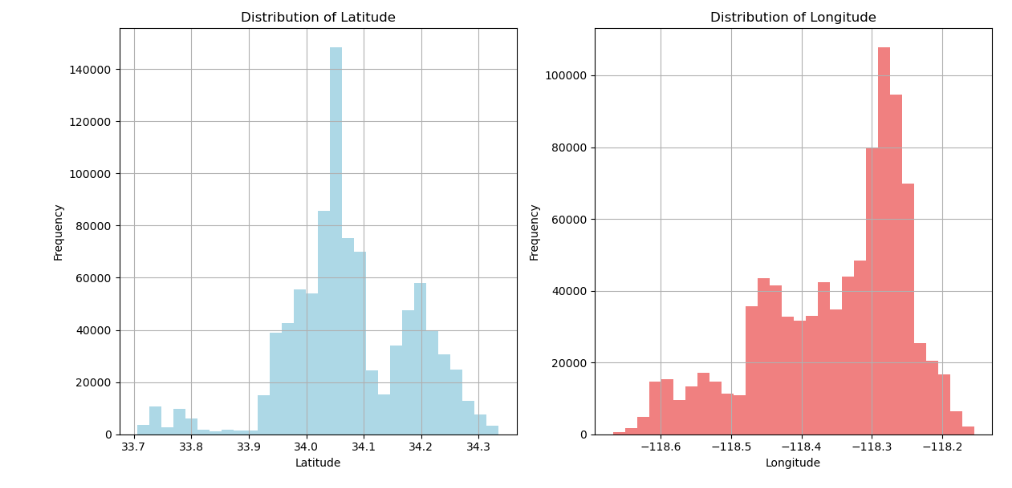
**29. Distribution of Top 10 Cross Streets:**



**Observation:**

The above graph represents the highest crimes in Cross Streets.It is evident that from “BROADWAY” have comparetively more number of crimes than “FIGUEROA ST” and many more.

**30.Distribution of Latitude and Longitude:**



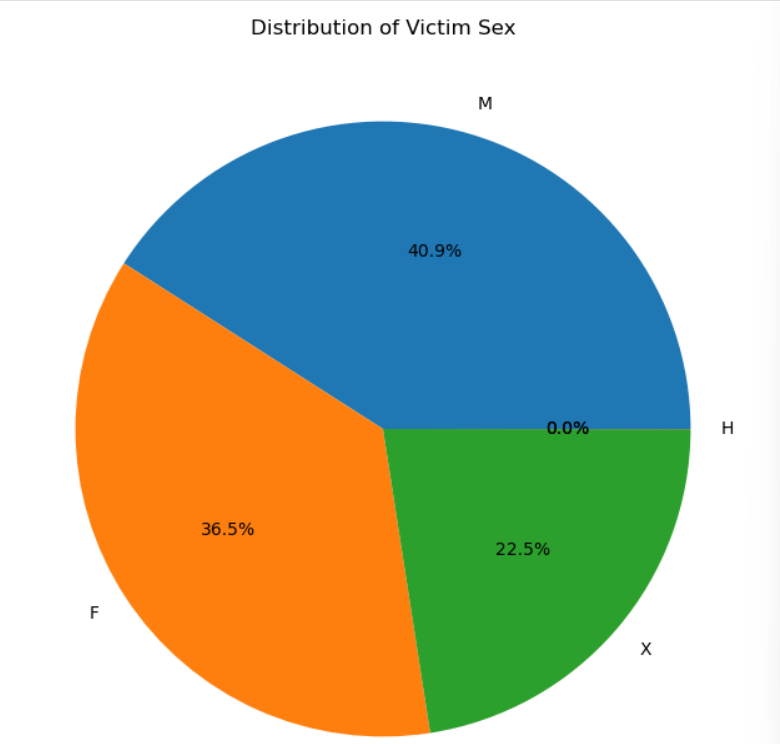
**Observation:**

The above graph says that :

Latitude Distribution: The frequency of latitudes peaks between approximately 33.8 and 34.2. This indicates that a significant number of incidents occurred around these latitudes. The distribution appears to be skewed to the right, with fewer incidents occurring at latitudes higher than 34.2.

Longitude Distribution: The frequency of longitudes peaks between approximately -118.6 and -118.2. This suggests that a large number of incidents occurred around these longitudes. Similar to the latitude distribution, the distribution of longitudes appears to be skewed, with fewer incidents occurring at longitudes lower than -118.6

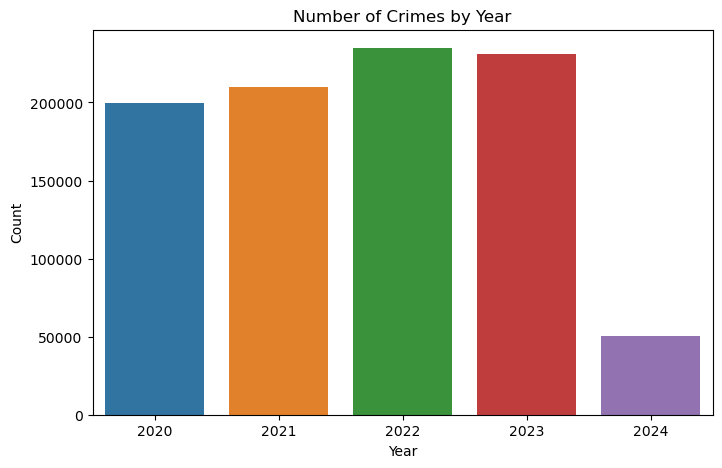
**31. Distribution of Victim Sex**



**Observation:**

The above pie chart says that **Women are more likely to be victims of sexual assault than men.** The pie chart shows that 40.9% of victims are female (F), 36.5% are male (M), and 22.5% are unspecified (X).

**32. Number of Crimes By Year**



**Observation:**

The above bar graph shows that from 2020 to 2023, the number of reported crimes remained relatively stable, with each year averaging around 200,000 incidents, but in 2024, there was a drastic reduction in crime rates, plummeting to below 50,000 incidents.

### 6.ALGORITHMS

Predicting Crime Type

* Random Forest
* Support Vector Machine (SVM)
* Gradient Boosting

Crime Severity Prediction

* Random Forest
* Support Vector Machine (SVM)
* Gradient Boosting

High-Crime Area Identification

* Geospatial Random Forest

### 7.Implementation

# 7.1 Predicting Crime Type

# 7.1.1 Random Forest:

# 

# The dataset is split into features (X) and target (y). Only the first 5000 rows of data are considered (X[:5000] and y[:5000]).

# The testing set size is set to 20% of the total data, and a random seed of 42 is specified for reproducibilit . The dataset was divided into training (80%) and testing (20%) sets.

# A Random Forest classifier is initialized with 100 trees (n\_estimators=100) and a random seed of 42. Model Training: The classifier is trained on the training data

# The accuracy of the model is evaluated using the score method, which calculates the accuracy on the test data (X\_test, y\_test).

# The computed accuracy is approximately 27.4%.

# 7.1.2 Support Vector Machine (SVM):

# 

# The dataset is split into features (X) and target (y). Only the first 5000 rows of data are considered (X[:5000] and y[:5000]).

# The testing set size is set to 20% of the total data, and a random seed of 42 is specified for reproducibility. The dataset was divided into training (80%) and testing (20%) sets.

# The code initializes a Support Vector Machine (SVM) classifier using the SVC class from sklearn.svm. The classifier is configured to use the radial basis function kernel with regularization parameter C=1.0 .A random seed of 42 is specified for reproducibility .

# The SVM classifier is trained on the training data.

# The computed accuracy is approximately 20.7%.

# 7.1.3 Gradient Boosting:

# 

# The dataset is split into features (X) and target (y). Only the first 5000 rows of data are considered (X[:5000] and y[:5000]).

# The testing set size is set to 20% of the total data, and a random seed of 42 is specified for reproducibility. The dataset was divided into training (80%) and testing (20%) sets.

# The code initializes a Gradient Boosting Classifier using the GradientBoostingClassifier class. The classifier is configured with

# Number of estimators= 100

# A learning rate=0.1

# Maximum tree depth=3

# A random seed = 42.

# The Gradient Boosting Classifier is trained on the training data.

* The computed accuracy is approximately 24%.

**Overall Observations:**

Based on the observed accuracies, the Random Forest classifier demonstrated the highest accuracy among the three methods evaluated. The Random Forest Classifier effectively classified instances in the synthetic dataset, indicating its suitability for classification tasks.

So the below model is built based on this classifier.

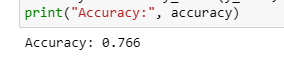
**7.2 Crime Severity Prediction**

**7.2.1 Random Forest:**



* + The code begins by preprocessing the dataset by replacing values in the "Vict Sex" column with numeric representations (2 for Female, 1 for Male, 0 for Unknown, and 3 for other cases).
  + Similarly, the "Vict Descent" column values are replaced with numeric representations based on a predefined mapping dictionary.After preprocessing, the dataset is filtered to include only selected columns .
  + The dataset is split into training and testing sets. The testing set size is set to 20% of the total data. The dataset was divided into training (80%) and testing (20%) sets.
  + A Random Forest classifier is initialized with 100 estimators. The classifier is trained on the training data.
  + The Random Forest classifier achieves a high accuracy of approximately 99% on the test data.

**7.2.2 Support Vector Machine (SVM):**



* + The testing set size is set to 20% of the total data, and a random seed of 42 is specified for reproducibility. The dataset was divided into training (80%) and testing (20%) sets.
  + A Support Vector Machine (SVM) classifier is initialized with a radial basis function (RBF) kernel using the SVC class.
  + The classifier is configured with a regularization parameter C=1.0, auto-scale gamma parameter, and a random seed of 42.
  + The SVM classifier is trained on the training data.
  + The SVM classifier achieves an accuracy of approximately 76.6% on the test data.

**7.2.3 Gradient Boosting:**

* + The testing set size is set to 20% of the total data, and a random seed of 42 is specified for reproducibility. The dataset was divided into training (80%) and testing (20%) sets.
  + A Gradient Boosting Classifier is initialized using the GradientBoostingClassifier class.
  + The classifier is configured with number of estimators=100, a learning rate= 0.1, a maximum tree depth =3.
  + The Gradient Boosting Classifier achieves an exceptionally high accuracy of approximately 99.9% on the test data.

**Overall Observations:**

The Random Forest classifier achieves an accuracy of 99%, the SVM classifier achieves an accuracy of 76.6%, and the Gradient Boosting classifier achieves an accuracy of 99.9%.

The best method among the three classifiers appears to be the Gradient Boosting Classifier, which achieved the highest accuracy of 99.9%.

So the below model is built based on this classifier.

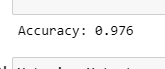
**7.3 High-Crime Area Identification**

**7.3.1 Random Forest:**



* + Only the first 5000 rows of data are considered for analysis.
  + The target variable (y) is defined as the 'AREA' column from data, limited to the first 5000 rows.
  + The dataset is split into training and testing sets, 80% of the data is allocated for training and 20% for testing. The random state is set to 42 for reproducibility.
  + A Random Forest classifier is initialized with 100 estimators and a random state of 42.
  + The accuracy of the Random Forest classifier on the test data is computed and printed, yielding an accuracy of approximately 97.9%.

**7.3.2 Gradient Boosting:**



* + Only the first 5000 rows of data are considered for analysis.
  + The dataset is split into training and testing sets with 80% of the data allocated for training and 20% for testing.
  + A Gradient Boosting Classifier is initialized with the following hyperparameters:
    - Number of boosting stages to be run. Number of estimators=100.
    - The shrinkage parameter to control the contribution of each tree. Learning rate is 0.1.
    - Maximum depth of the individual trees is 3.
    - A random seed for reproducibility, set to 42.
  + The calculated accuracy is printed, which is approximately 97.6%.

**7.3.3 Support Vector Machine (SVM):**



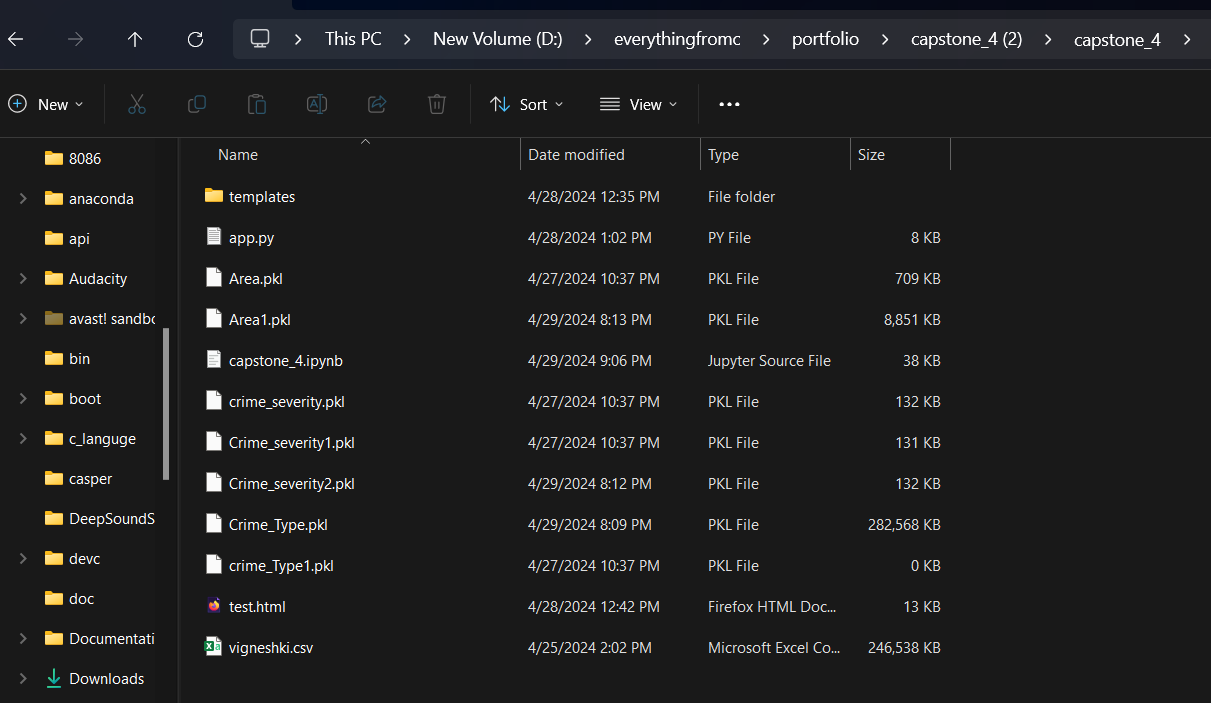
* + Only the first 5000 rows of data are considered for analysis.
  + The dataset is split into training and testing sets with 80% of the data allocated for training and 20% for testing.
  + An SVM classifier with a radial basis function (RBF) kernel is initialized with the following hyperparameters:
    - kernel: The type of kernel used for the SVM. Here, it's set to 'rbf'.
    - C: The regularization parameter, set to 1.0.
    - gamma: The kernel coefficient for 'rbf', set to 'scale'.
    - random\_state: A random seed for reproducibility, set to 42.
  + The calculated accuracy is printed, which is quite low at approximately 8.2%.

**Overall observations:**

Both Random Forest and Gradient Boosting classifiers perform well on the dataset, achieving accuracies of around 97.9% and 97.6%, respectively. However, the SVM classifier performs poorly, with an accuracy of only 8.2%. Therefore, Random Forest and Gradient Boosting classifiers seem to be more suitable for this particular task. The Random Forest Classifier effectively classifies instances in the dataset, demonstrating its utility in classification tasks. So the below model is built based on this classifier.

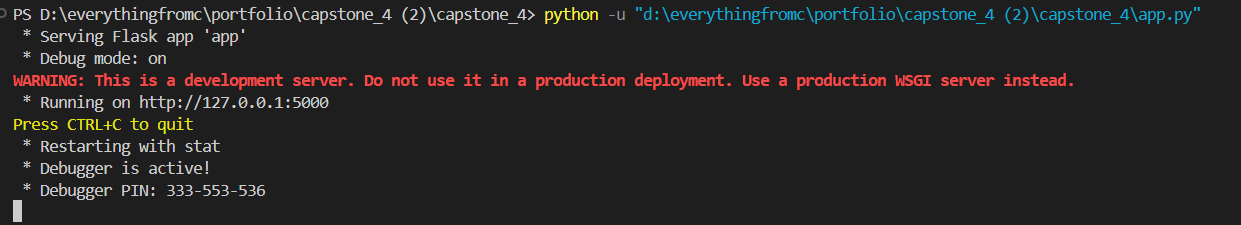
**8.IMPLEMENTATION**

**For predicting the crime area, crime type, crime severity**

****

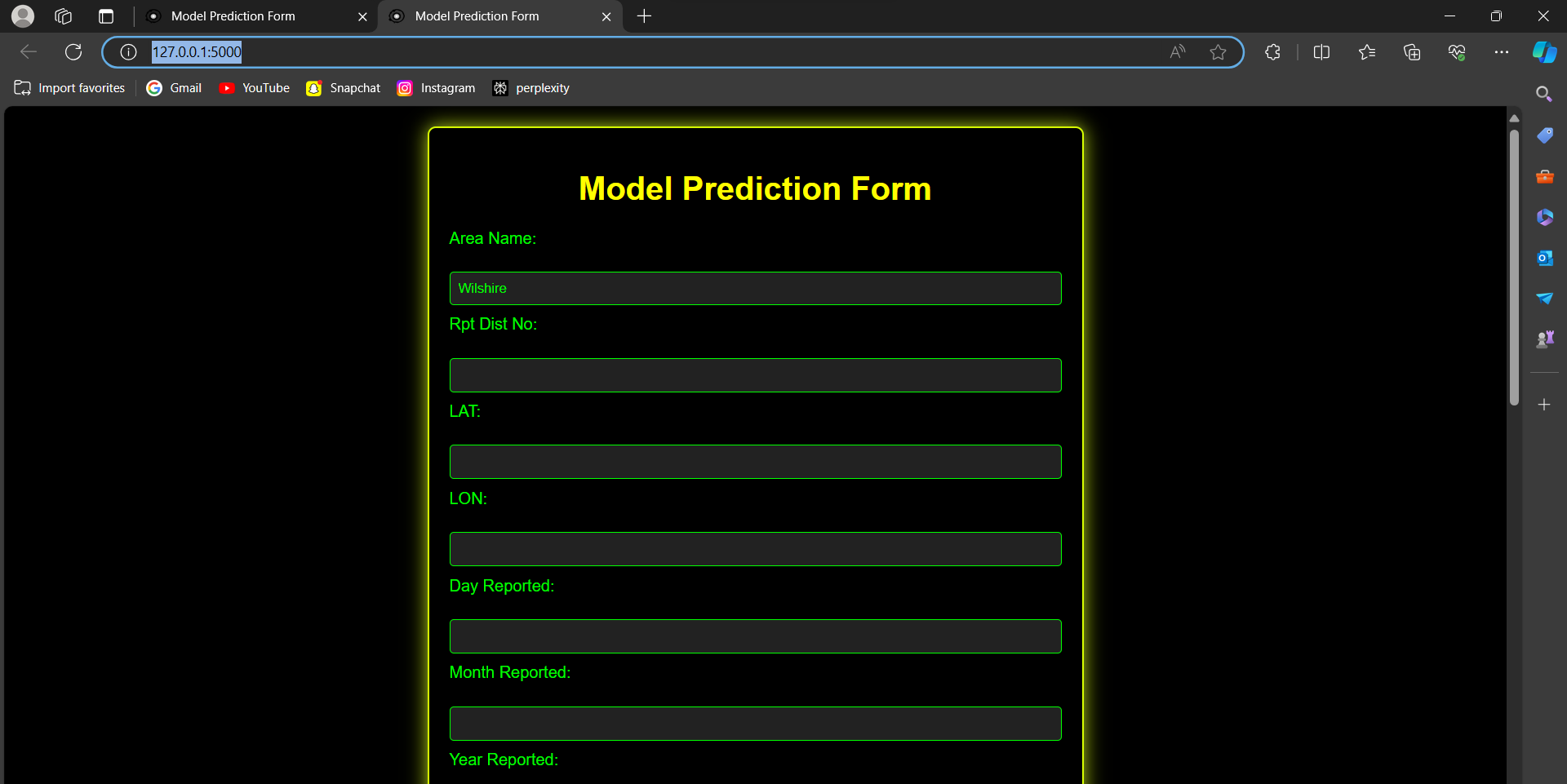
Open with vs code or spyder

Run in the vs code terminal

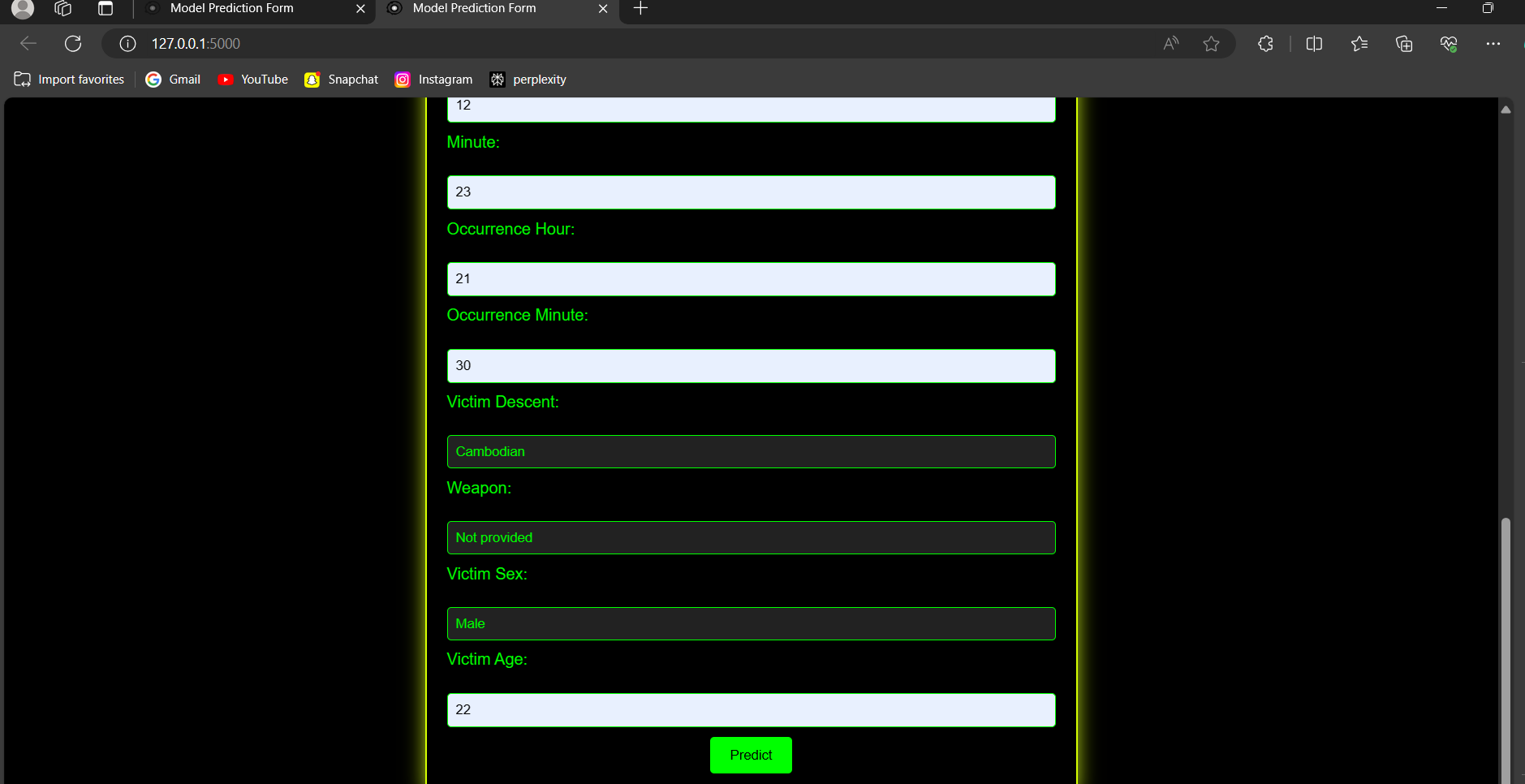


Click on the link <http://127.0.0.1:5000>

Open in the browser

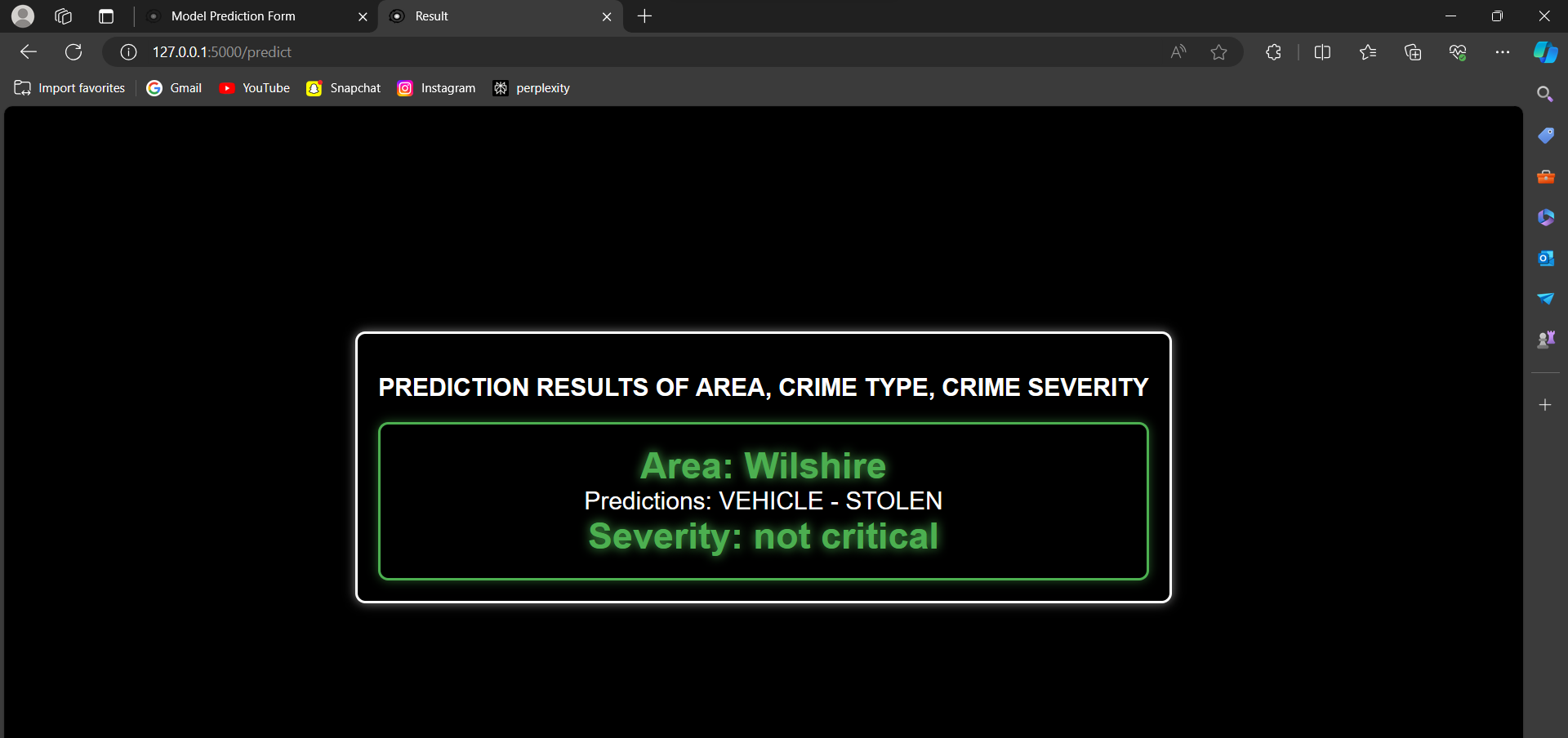


Fill the form to predict the area with high crime and , crime severity , crime type .



After filling the all the labels and select and button predict to get the values

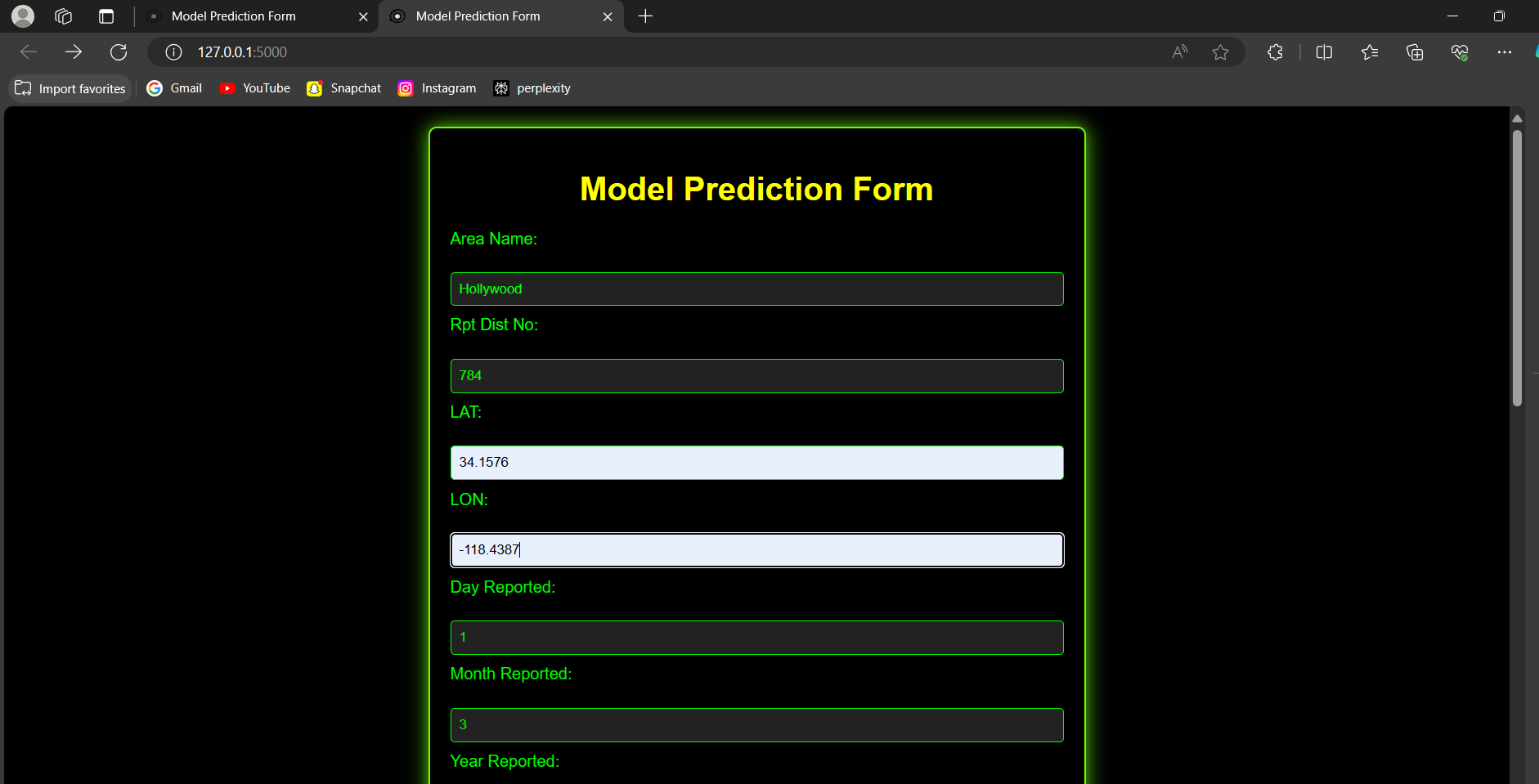
Click on the predict button

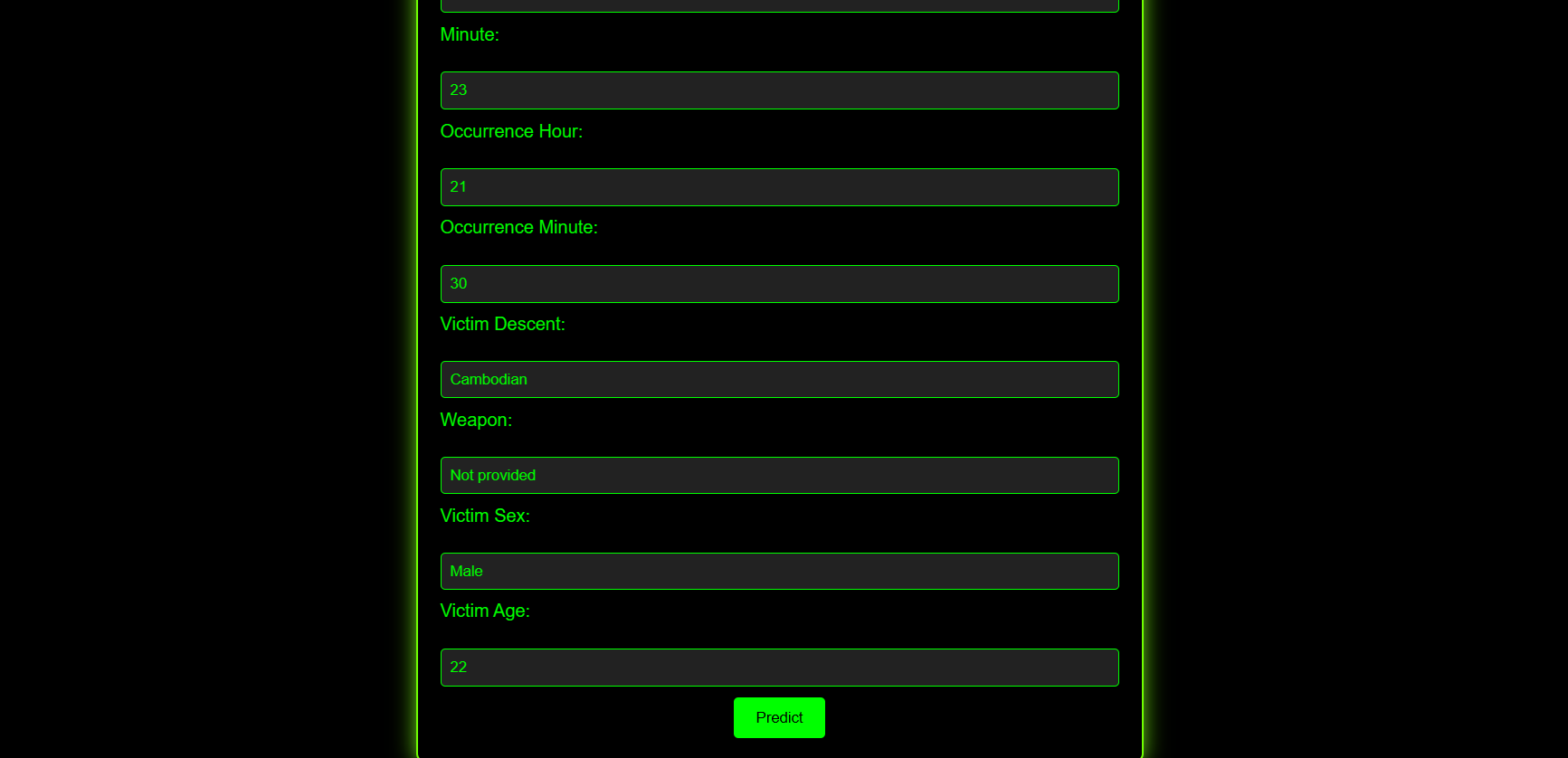


After clicking on the predict button then Area: Wilshire , Predictions :VEHICLE-STOLEN , Severity: Not-CriticalThese are the predicted values for the area of whilshire and location number of 784 and occurred on 1/3/2020 to 4/3/2020 .

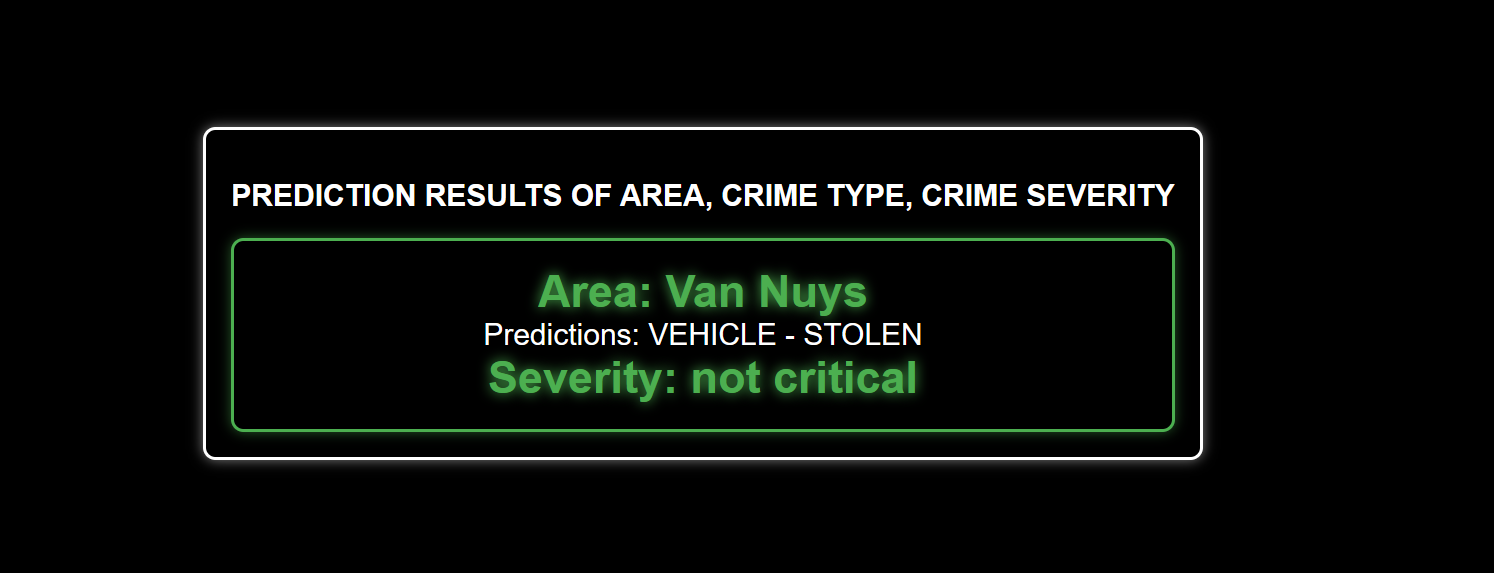
The above flask implementation gives us research problem 1,2,3

Predicting another crime area, crime type



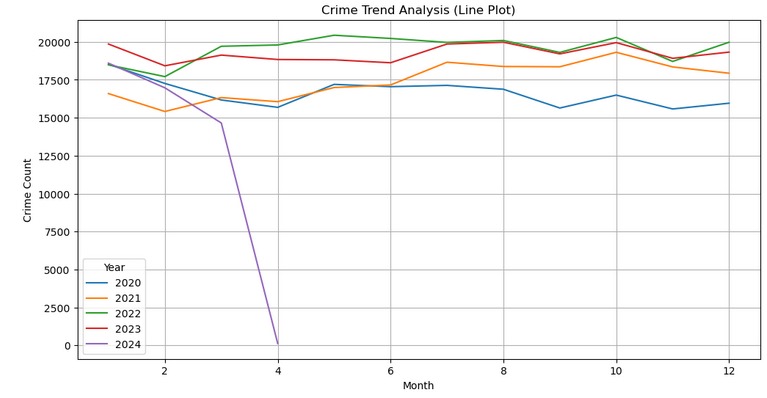
****

Click on predict button

**s**

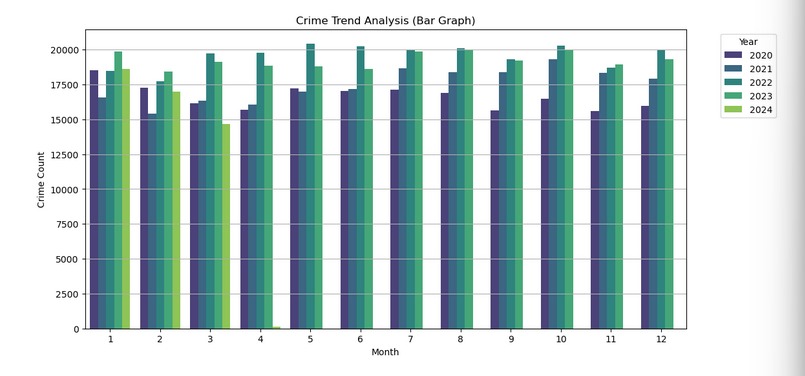
**Research Question 4: Crime Trend Analysis**

* Time Series Analysis: Analyze crime data over time to identify seasonal trends (e.g.,
* property crimes might increase during summer) or weekly/daily patterns.
* Visualization: Create charts and graphs to visualize changes in crime rates over time
* or across different locations.
* Benefits: This analysis can help predict future crime trends and inform resource
* allocation strategies.



**Observation:**

* The graph illustrates the crime counts over a 12-month period, with each line representing a different year (from 2020 to 2024).
* In 2020, there was a notable and abrupt decline in crime counts after the second month. However, from 2021 to 2024, the crime counts fluctuated but remained relatively stable compared to the drastic decrease in 2020.
* While there are peaks and troughs in the years following 2020, there is no consistent upward or downward trend.
* Year 2020:The most significant change occurred in 2020, characterized by a sharp decrease in crime counts after the second month. This decline could be attributed to various factors such as changes in law enforcement practices, social circumstances, or external influences.
* Such as in year 2020 the covid-19 attack across the world so there was decline in crime rate in 2020



* The bar graph represents crime trends over a period of five years, from 2020 to 2024.
* The crime count fluctuates monthly but does not show a consistent trend of increase or decrease over the years.

**Research Question 5: Resource Allocation Optimization**

Approach: This builds upon the previous research problems.

● Use crime prediction models (crime type and high-crime areas) to identify areas

with a higher likelihood of specific crimes.

● Consider crime severity predictions to prioritize resource allocation for potentially

more serious crimes.

Benefits: By allocating resources (e.g., patrols) based on predicted crime hotspots and

severity, law enforcement can potentially deter crime and improve public safety.

resource\_allocation\_score = crime\_types\_pred \* high\_crime\_areas\_pred \* crime\_severity\_pred

**output:**  
array([ 9204, 13146, 8736, 4956, 1416, 7788, 4248, 8112, 1872,

7788, 13320, 6018, 2832, 3540, 6120, 5610, 3248, 6510,

6372, 4248, 12520, 4590, 5664, 7788, 4620, 2832, 14280,

708, 4140, 4602, 9893, 1452, 21420, 6260, 19728, 4620,

4340, 745, 7650, 2040, 2220, 1248, 5664, 3894, 9912,

11268, 6630, 3310, 18704, 2478, 12520, 9690, 21216, 4650,

2832, 1020, 36120, 7080, 12744, 1248, 20400, 1260, 24960,

5008, 26480, 11232, 10560, 3744, 5616, 2124, 9282, 708,])

**9. Conclusion**

This capstone project has conducted a comprehensive analysis of LAPD crime data from 2020 to 2024, leveraging machine learning techniques to extract valuable insights for enhancing public safety in Los Angeles. Through exploratory data analysis, the study has uncovered temporal patterns, spatial distributions, crime type relationships, victim demographics, and severity trends. The application of machine learning models, including Random Forest, Support Vector Machines, and Gradient Boosting Classifiers, has enabled accurate predictions of crime types, severity levels, and high-crime areas. These predictions, combined with geospatial visualization techniques, can inform targeted interventions and resource allocation strategies for law enforcement agencies.

this project has demonstrated the potential of data-driven approaches in crime analysis, providing the LAPD and policymakers with actionable insights to enhance public safety efforts. By identifying crime hotspots, predicting crime types and severity, and optimizing resource allocation, the findings of this study can support evidence-based decision-making processes aimed at fostering a safer and more secure environment for the residents of Los Angeles.

**10. Future Scope**

The purpose of this study is to examine crime analysis through the applicability of data mining methods in the process of crime prediction and prevention. The future scope of the project involves further exploration of advanced predictive modeling techniques, integration of socio-economic data, expansion of community engagement initiatives, development of real-time data analysis systems, evaluation of intervention strategies, longitudinal studies, technological advancements, and integration with enterprise software solutions. These efforts aim to improve public safety, reduce crime rates, and enhance the effectiveness of law enforcement strategies in Los Angeles and beyond.