ISDS 577 - Seminar in Information Systems Implementation, Spring 2024

DATA ANALYTICS ON CRIME IDENTIFICATION AND PREVENTION WITH MACHINE LEARNING IN LOS ANGELES

Analytics Project – Final Report

ISDS 577 – Group 5

Jeevan Gowda Hemanth Kumar

Sai Krishna Mallineni

Anwesh Reddy Malgireddy

Bala Avinash Allam

Table of Contents

Executive Summary

One of the most challenging tasks in the criminal justice system is identifying crimes since precise identification necessitates more investigation and analysis of the incident. The current approach uses a manual survey and examination of papers that cannot be identified because of the variety of offenses that make identification more difficult. There is also very little precision in the findings regarding detecting crimes. The suggested approach, which uses data analytics to analyze criminal activity, will aid in both identification and future criminal prevention. The various factors obtained from the crime are used as input, and data analytics combined with metrics helps to precisely determine the location of the crime and its place in a detailed framework. The data will be analyzed with the aid of data cubes, which facilitate the analysis of criminal data from multiple perspectives related to victim-perpetrated crimes. Applying statistics and analytical techniques to identify the areas with the greatest crime rates and then allocating increased policing resources to those areas is known as predictive policing. Using data mining techniques, one can find patterns related to theft, homicide, and domestic abuse. Police investigators can now quickly sort through vast amounts of data, including text, photos, and video, by employing powerful data analytics.

Big data analytics can be used to predict financial crimes, including insider trading, insurance fraud, money laundering, and criminal care fraud. Both organized and unstructured data will be analyzed using the suggested methodology in an effort to find any possible criminal evidence. Furthermore, charges against fraudsters might be brought using the insights gleaned from his data.

Data Collection & Preparation

Data Collection

The crime dataset of Los Angeles was collected from Kaggle. Attached is the link to our dataset:

https://www.kaggle.com/datasets/shayalvaghasiya/los-angeles-crimes

Data Legal Privacy

No legal or privacy concerns are associated with our data because it is public data collected through the Kaggle dataset. It is also under the Community Data License Agreement, which allows the collaborative sharing of data that we have seen proven to work in open-source software communities.

Data Format

The data will be provided in common-separated value (CSV) format.

Integrated Datasets

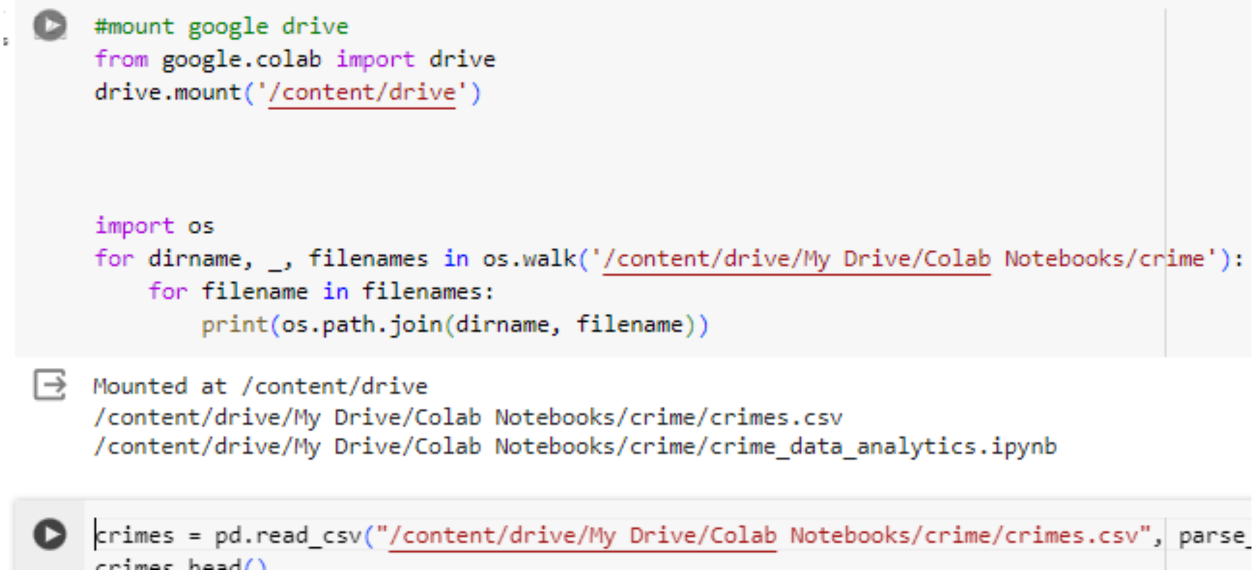
Based on our research questions, there will be one dataset for the project. It will be the crime dataset from Kaggle; however, we will split it into training and validation sets when used for prediction. No additional datasets will be integrated with our baseline one for analysis.

Data Cleaning

The crime dataset has many missing values and variables that are irrelevant to our investigation; therefore, cleaning it up will be important. In addition, we decreased the amount of our dataset, using only the information from 2022 to 2023 for this research. Nonetheless, since 21 indicator variables were evaluated in the study, we shall investigate the necessity of each variable in our model. We might need to decide which rows to exclude since they are duplicates. For this analysis and classification, we might have to create only two outcome variables to categorize our anticipated variables.

Additionally, because the dataset has a significant imbalance in the outcome variable, we might need to investigate ways to alleviate any classification overfitting. To clean the data, we'll utilize Python, Google Colab, and a variety of data analytics codes.

Import Dataset:



```
#mount google drive
from google.colab import drive
drive.mount('/content/drive')



import os
for dirname, _, filenames in os.walk('/content/drive/My Drive/Colab Notebooks/crime'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

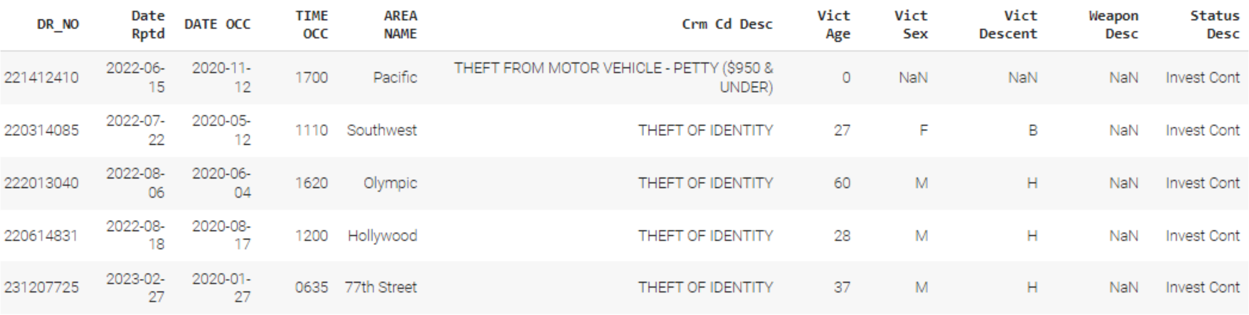
```
Mounted at /content/drive
/content/drive/My Drive/Colab Notebooks/crime/crimes.csv
/content/drive/My Drive/Colab Notebooks/crime/crime_data_analytics.ipynb
```

```
crimes = pd.read_csv("/content/drive/My Drive/Colab Notebooks/crime/crimes.csv", parse_
crimes.head()
```

Figure 0.1: Python code for importing the dataset

The project code and the dataset are loaded into the Google Drive folder located within the Google Colab folder. 247,989 crime records from 2020 to 2023 are included in the collection. There are 12 feature variables in this dataset. A typical crime in Los Angeles that affects millions of people annually and puts a heavy financial strain on the economy is crime data analytics.

Results of loading dataset:

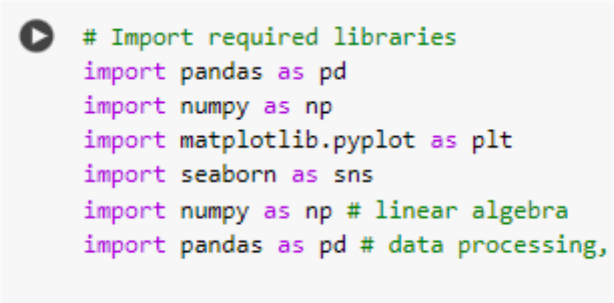


DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA NAME	Crm Cd Desc	Vict Age	Vict Sex	Vict Descent	Weapon Desc	Status Desc
221412410	2022-06-15	2020-11-12	1700	Pacific	THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)	0	NaN	NaN	NaN	Invest Cont
220314085	2022-07-22	2020-05-12	1110	Southwest	THEFT OF IDENTITY	27	F	B	NaN	Invest Cont
222013040	2022-08-06	2020-06-04	1620	Olympic	THEFT OF IDENTITY	60	M	H	NaN	Invest Cont
220614831	2022-08-18	2020-08-17	1200	Hollywood	THEFT OF IDENTITY	28	M	H	NaN	Invest Cont
231207725	2023-02-27	2020-01-27	0635	77th Street	THEFT OF IDENTITY	37	M	H	NaN	Invest Cont

Figure 0.2: Results of importing the dataset.

The city of Los Angeles, California, is well-known for its breathtaking coastline, palm palms, and status as the center of the entertainment sector. Some of history's most recognizable songs and movies have come from Hollywood. But crime can be an unwelcome reality in any busy metropolis. Since 2, 47,989, crime data has been gathered. This research used a 2020 Kaggle-provided CSV dataset that included responses from 247,989 people and 12 features. These qualities are either variables calculated based on individual responses or questions posed to participants.

Initially, the library files are loaded into the application.



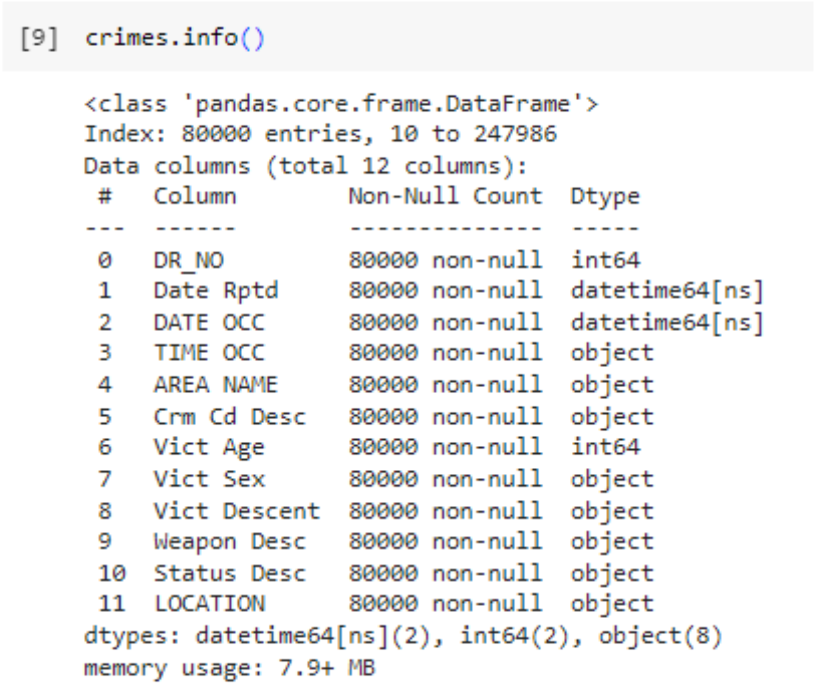
```
# Import required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np # linear algebra
import pandas as pd # data processing,
```

Figure 0.3: Import the library files

After loading the library files, the dataset is mapped with the Google Colab notebook folder, and the dataset file is passed into the application.

Dataset Attributes/ Variables

There are many variables in our dataset. We will explore each of the crime indicators to assess the predictability of crime. The variables include DR_NO, Date Rptd, DATE OCC, TIME OCC, AREA NAME, Crm Cd Desc, Vict Age, Vict Sex, Vict Descent, Weapon Desc, Status Description, LOCATION



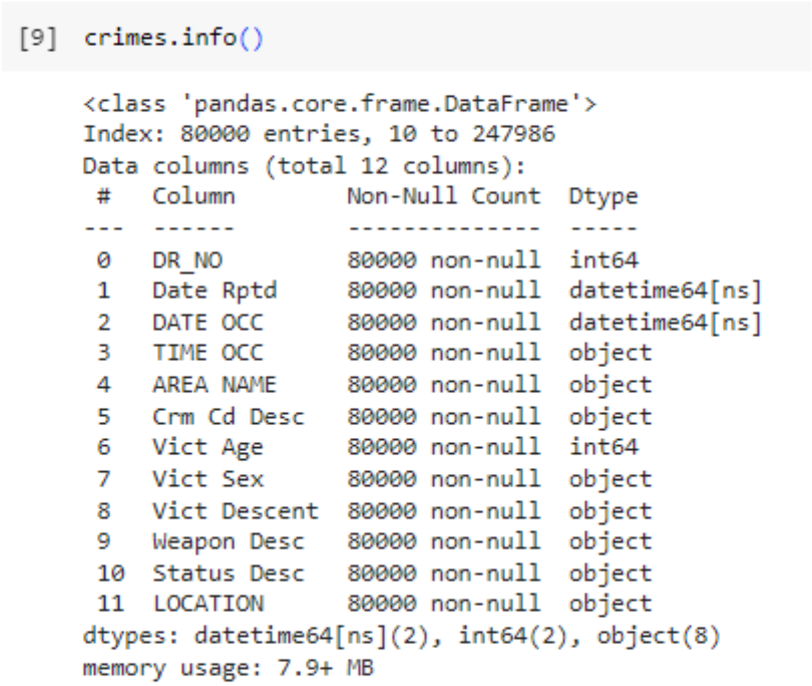
```
[9]  crimes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 80000 entries, 10 to 247986
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   DR_NO         80000 non-null  int64
 1   Date Rptd     80000 non-null  datetime64[ns]
 2   DATE OCC      80000 non-null  datetime64[ns]
 3   TIME OCC      80000 non-null  object
 4   AREA NAME     80000 non-null  object
 5   Crm Cd Desc   80000 non-null  object
 6   Vict Age      80000 non-null  int64
 7   Vict Sex      80000 non-null  object
 8   Vict Descent  80000 non-null  object
 9   Weapon Desc   80000 non-null  object
 10  Status Desc   80000 non-null  object
 11  LOCATION      80000 non-null  object
dtypes: datetime64[ns](2), int64(2), object(8)
memory usage: 7.9+ MB
```

Figure 0.4: Dataset Attribute properties

A total of 12 column attributes are present in the dataset, and each column contains a specific unique property.

The metric values of each property can also be identified with the Python Describe function.



```
[9] crimes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 80000 entries, 10 to 247986
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   DR_NO         80000 non-null  int64
 1   Date Rptd     80000 non-null  datetime64[ns]
 2   DATE OCC      80000 non-null  datetime64[ns]
 3   TIME OCC      80000 non-null  object
 4   AREA NAME     80000 non-null  object
 5   Crm Cd Desc   80000 non-null  object
 6   Vict Age      80000 non-null  int64
 7   Vict Sex      80000 non-null  object
 8   Vict Descent  80000 non-null  object
 9   Weapon Desc   80000 non-null  object
 10  Status Desc   80000 non-null  object
 11  LOCATION      80000 non-null  object
dtypes: datetime64[ns](2), int64(2), object(8)
memory usage: 7.9+ MB
```

Figure 0.5: Results of Describe Function

Final Dataset Size (Number of variables, etc.)

The baseline Crime dataset has 2,47,989 observations, each representing a different individual Los Angeles crime data. It has 12 attribute indicators used for prediction.

Data Analysis

Research Question #1: Are there any crime indicators that belong proportionately to those with/at risk for crime? Contrastingly, are there any crime indicators that belong proportionately to those without crime?

Variables/ Attributes

Area name, Crime code, Vict Age, Vict Sex, Weapon Description, Location.

Managerial Decision-making

Answering the first part of the research question about crime indicators that belong proportionately to those risks for crime and identifying the underlying reasons will enable criminology field operations of police to optimize patient crime records, manage the crime conditions, and find the crime indicators that help to find the crime early stage. Additionally, understanding the specific causes of crime will allow for targeted crime attributes and early identification.

Lastly, analyzing how the crime indicator identifies crime with various attributes across different locations.

ANALYSIS:

To gain a comprehensive understanding, the analysis will proceed as follows:

1. Health Indicators of Crime Risk analysis:

    a) Area name

    b) Crime code

c) Vict Age

d) Vict Sex

2. Crime Analysis:

a) Crime hourly analysis

b) Crime Description

3. Graph char visualization

4. Result Analysis

Area name Analysis:

Throughout the process of this project, we will categorize Area Name Check into primary classifications. More than 20% of records belong to the no risk of crime category, and 80 % of records are at risk of crime.

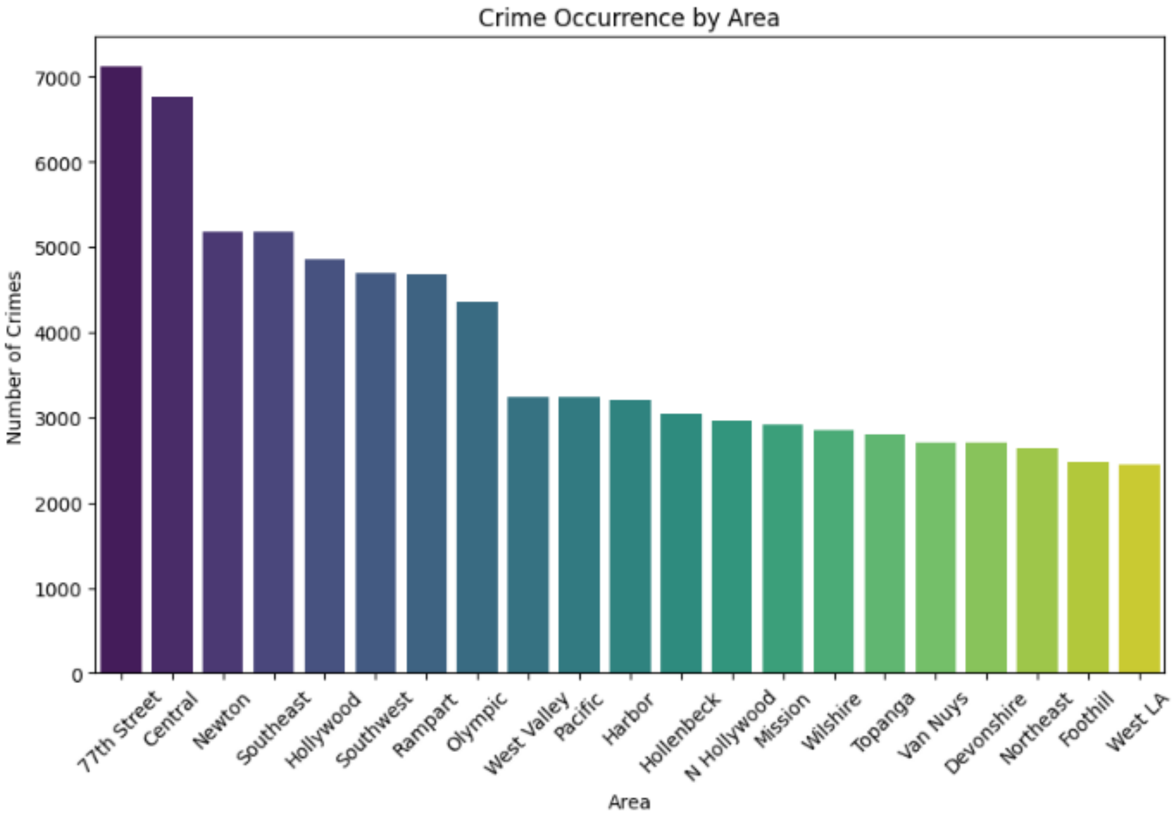
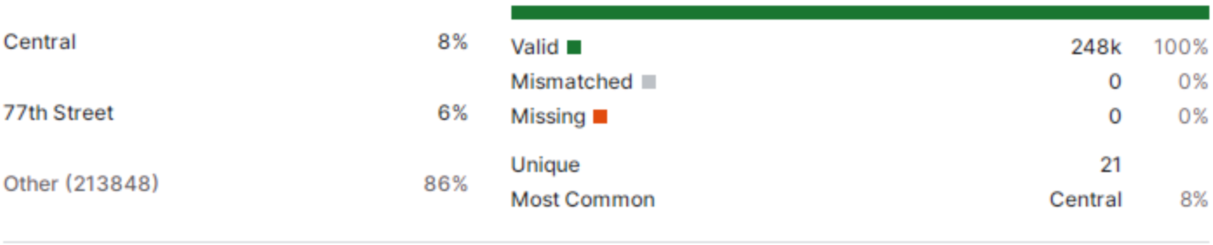


Figure 0.6: Area name with number of Crimes

The area's name attributes cause a huge increase in the risk factor of crime. The 77th street has the highest crime records of 7000 crimes. The 21 Geographic Areas or Patrol Divisions are also given a name designation that references a landmark or the surrounding community that it is responsible for. For example, the 77th Street Division is located at the intersection of South Broadway and 77th Street, serving neighborhoods in South Los Angeles.



Central 8%

77th Street 6%

Other (213848) 86%

Valid 🟩 248k 100%

Mismatched ⬜ 0 0%

Missing 🟥 0 0%

Unique 21

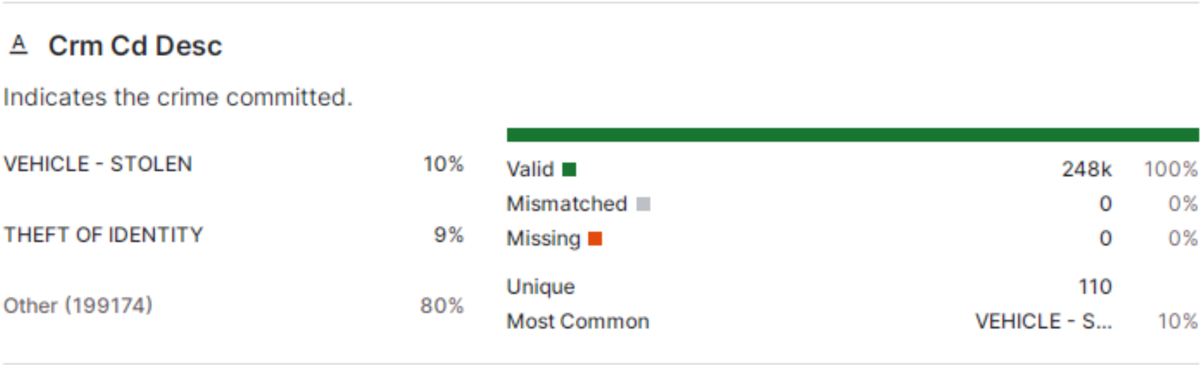
Most Common Central 8%

0.7: 77th Street High Risk to Crime

The 77th Street paves the path for the higher crime risk in Los Angeles records.

Crime Code Description Data Analysis:

The Vehicle Stolen is the highest crime record in the state of Los Angeles.



A Crm Cd Desc

Indicates the crime committed.

VEHICLE - STOLEN 10%

THEFT OF IDENTITY 9%

Other (199174) 80%

Valid 🟩 248k 100%

Mismatched ⬜ 0 0%

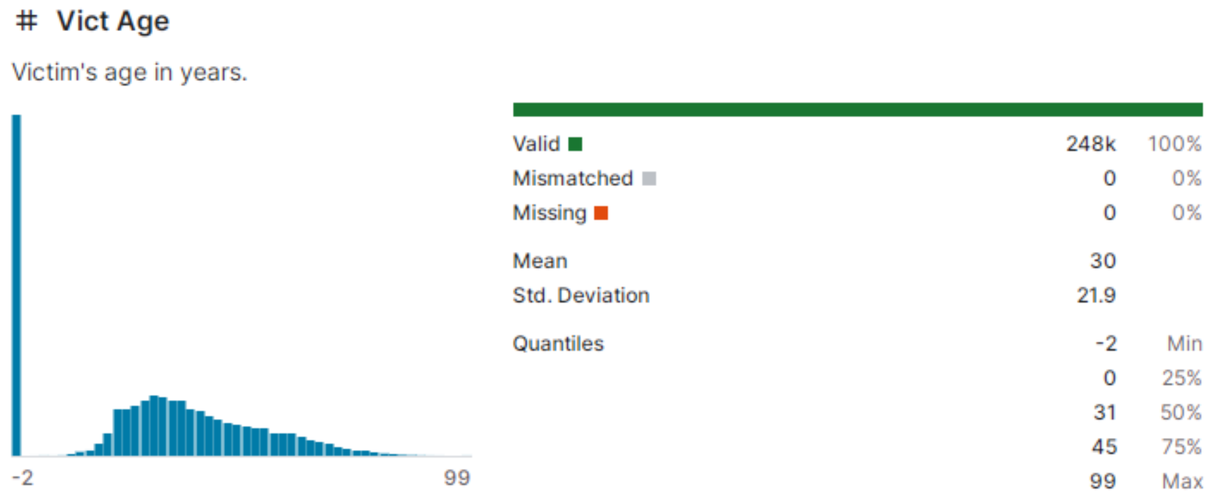
Missing 🟥 0 0%

Unique 110

Most Common VEHICLE - S... 10%

0.8: Crime Code Description to Crime

Victim Age Description Data Analysis:



# Vict Age

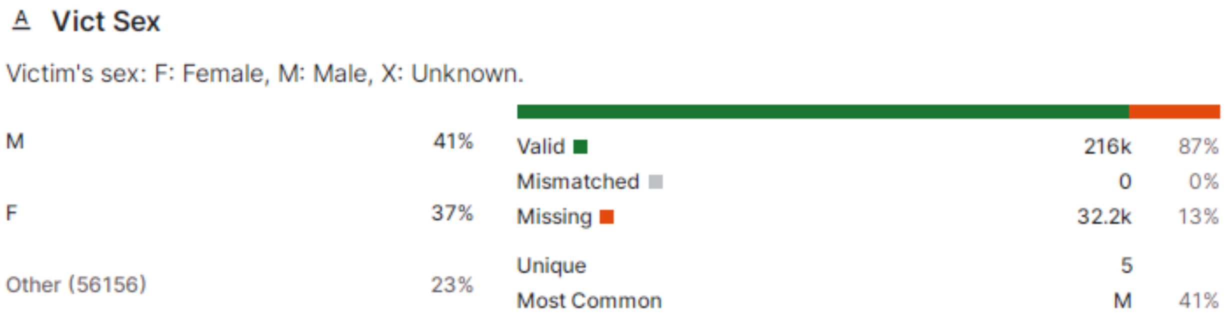
Victim's age in years.

Valid 🟩	248k	100%
Mismatched ■	0	0%
Missing 🟥	0	0%
Mean	30	
Std. Deviation	21.9	
Quantiles	-2	Min
	0	25%
	31	50%
	45	75%
	99	Max

0.9: Victim Age Risk to Crime

The adult age group has the highest records of crime. The graph shows the age-wise data with an increase in crime.

Victim Age Description Data Analysis:



A Vict Sex

Victim's sex: F: Female, M: Male, X: Unknown.

M	41%
F	37%
Other (56156)	23%

Valid 🟩	216k	87%
Mismatched ■	0	0%
Missing 🟥	32.2k	13%
Unique	5	
Most Common	M	41%

1.0: Victim Sex Factor Risk to Crime

The gender-wise data analysis on crime shows that male has 41% of records of crime, whereas Female has a 37% crime ratio.

Below are the indicators of victim sex:

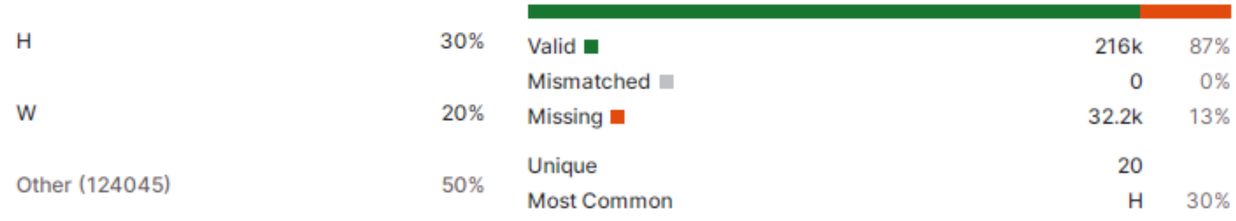
F: Female,

M: Male,

X: Unknown.

Victim Descent in Crime Data:

The Victim Descent in accordance with the crime records as given below:



H	30%
W	20%
Other (124045)	50%

Valid 🟩	216k	87%
Mismatched ▨	0	0%
Missing 🟥	32.2k	13%
Unique	20	
Most Common	H	30%

1.1: Victim Descent Attribute Factor Analysis

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

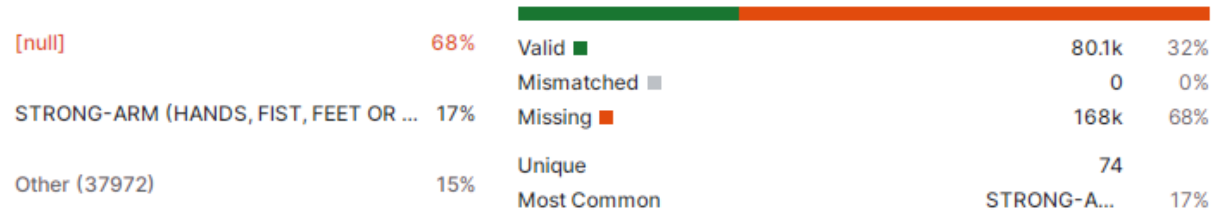
V – Vietnamese

W – White

X – Unknown

Z - Asian Indian

Weapon Description Factor with Crime:

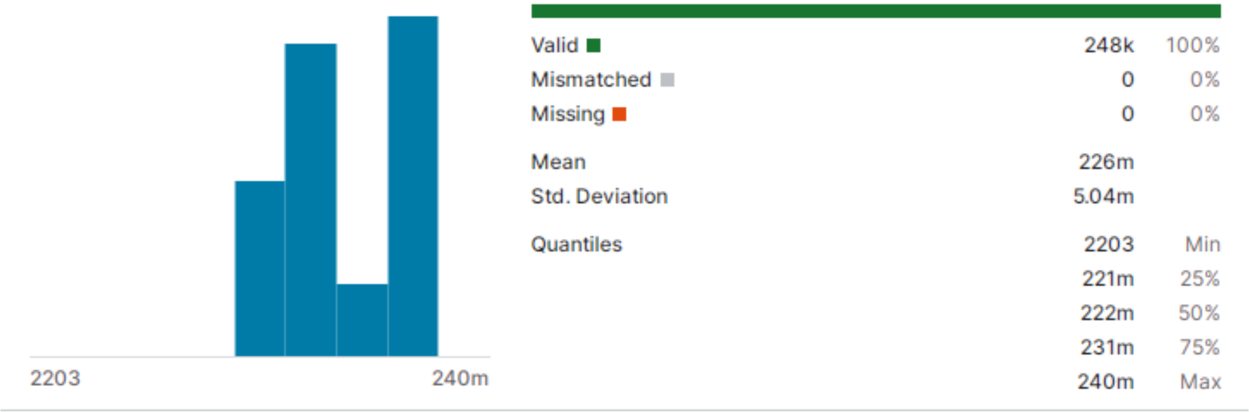


[null]	68%
STRONG-ARM (HANDS, FIST, FEET OR ...	17%
Other (37972)	15%

Valid 🟩	80.1k	32%
Mismatched ⬜	0	0%
Missing 🟥	168k	68%
Unique	74	
Most Common	STRONG-A...	17%

Strong arms (hands, fists, feet) are 17% of crime records, and other types of weapons used for crime are 15%.

Division of Record Number:

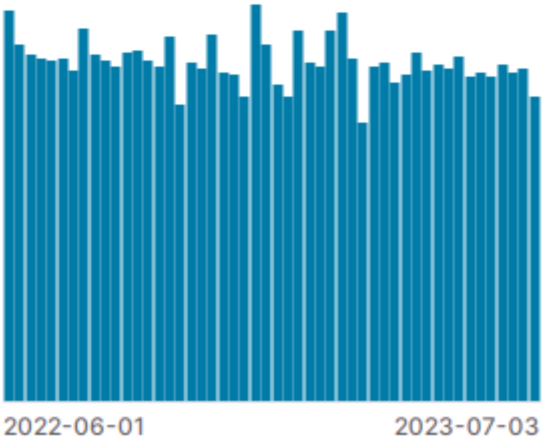


Valid 🟩	248k	100%
Mismatched ▨	0	0%
Missing 🟥	0	0%
Mean	226m	
Std. Deviation	5.04m	
Quantiles	2203	Min
	221m	25%
	222m	50%
	231m	75%
	240m	Max

1.3: Division of Records Attribute Factor Analysis

Division of Records Number: Official file number comprises a 2-digit year, area ID, and 5 digits.

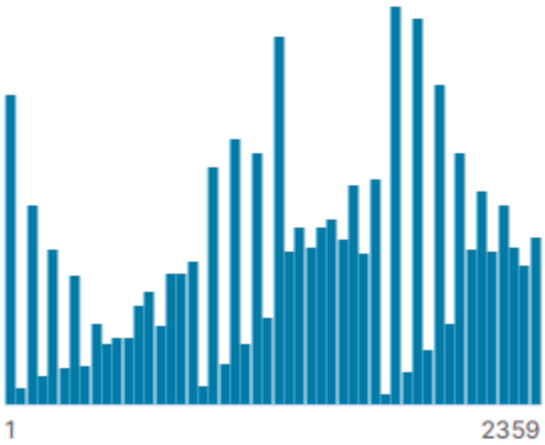
Date of Crime Reported:



1.4: Date of Occurrence Attribute Factor Analysis

The year 2022 and 2023 has the highest records of crime registered in Los Angeles.

Time of occurrence of crime:



1.5: Time of Occurrence Attribute Factor Analysis

In 24-hour military time is considered for the graphical analysis.

Conclusion and Recommendations:

The research clearly indicates that the crime indicators affect the crime identification. The Victim Descent, DR No, and Victim Descent attribute factors make less chance of crime risk, which has been analyzed with the different types of data analytics with Python programming.

Research Question #2: Is there a correlation between the Victim Age and Victim Sex? Is there a significant association between crime reasons and crime factors such as crime description, crime location, and crime purpose related to age and sex?

Variables/ Attributes

Area name, Crm Cd Desc , Vict Age, Vict Sex, Weapon Desc, Status Desc, Location

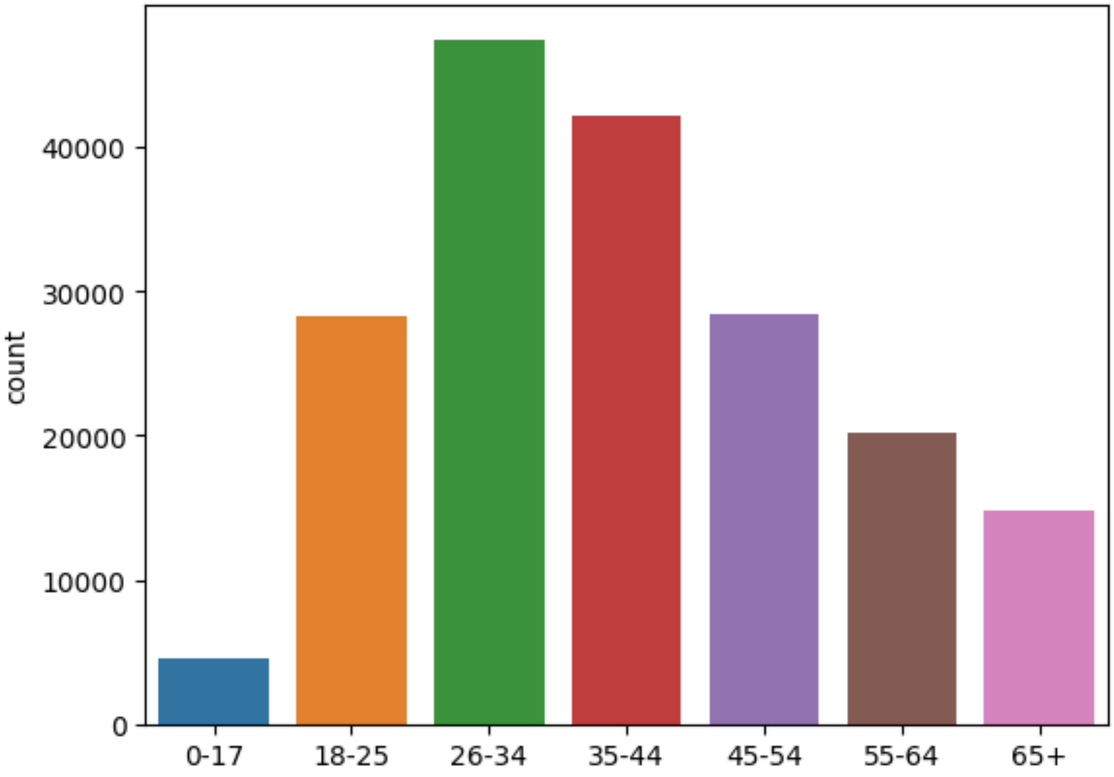
Managerial Decision-making

By addressing the Research question, it will be possible to identify the underlying causes of crime decrease, as well as how physical and mental elements related to criminal presence connect to it. These insights can then be used to inform strategic decision-making across a range of crime prediction-related domains. This data-driven strategy will improve the accuracy of victim age predictions, optimize crime records, and focus on the specific crimes committed there. Furthermore, the results will be used to develop various data analytics algorithms for detecting criminal activity.

Victim Age Analysis with Crime:

Throughout the process of this project, we will set the relationship of age relates to the presence of crime.

More than 45000 records belong to the no-risk of crime group, with ages 26 to 34, and records of 10000 have a low risk of crime for old people of age more than 65.



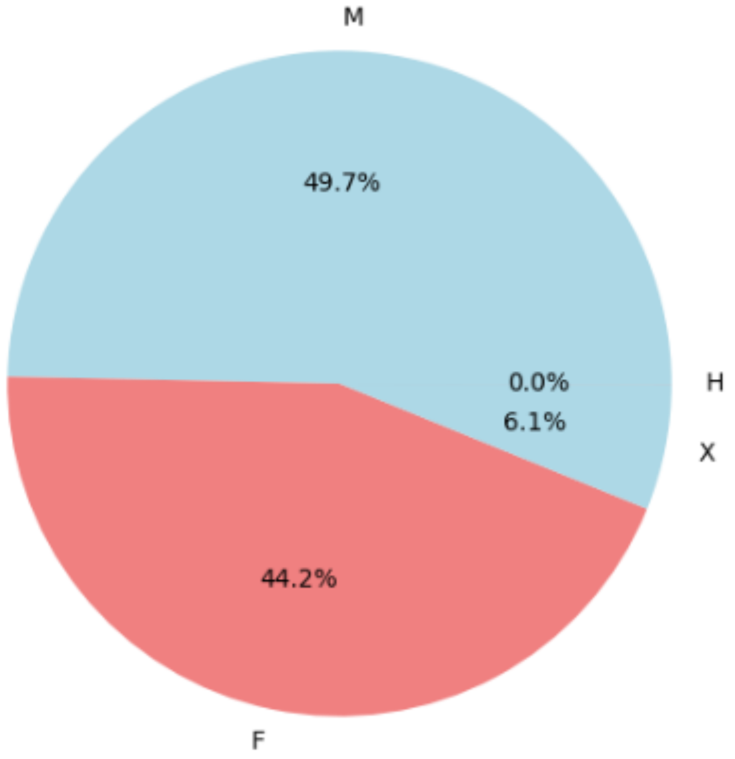
1.6: Age Relate to Crime

The graph denotes the Age, with the range of 26-34 and 35-44 ages having the highest crime records.

Victim Sex factor Analysis with crime:

Throughout the process of this project, we will set the relationship of Victim sex with related to the crime records.

More than 49% records of Males have higher crime records, and males have 44% of crime records in the city.

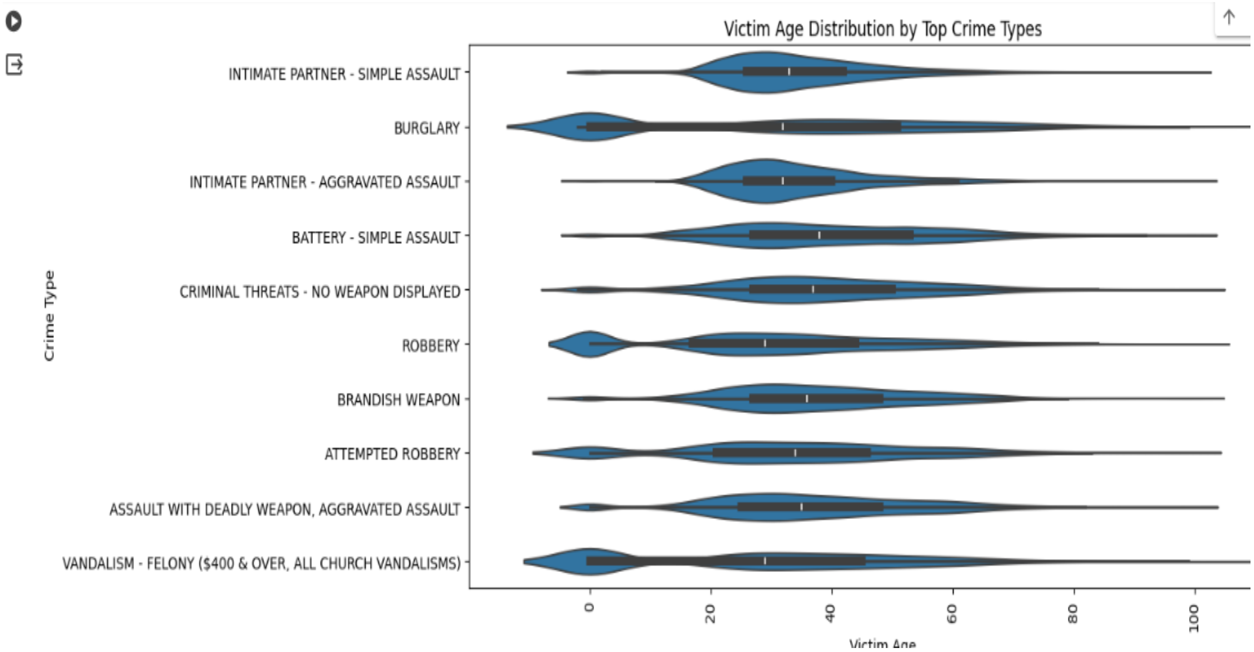


1.7: Victim Sex Relate to Crime

Significant Association between Crime Status and Victim Sex:

There is a significant association between crime status and factors such as victim age and victim sex group.

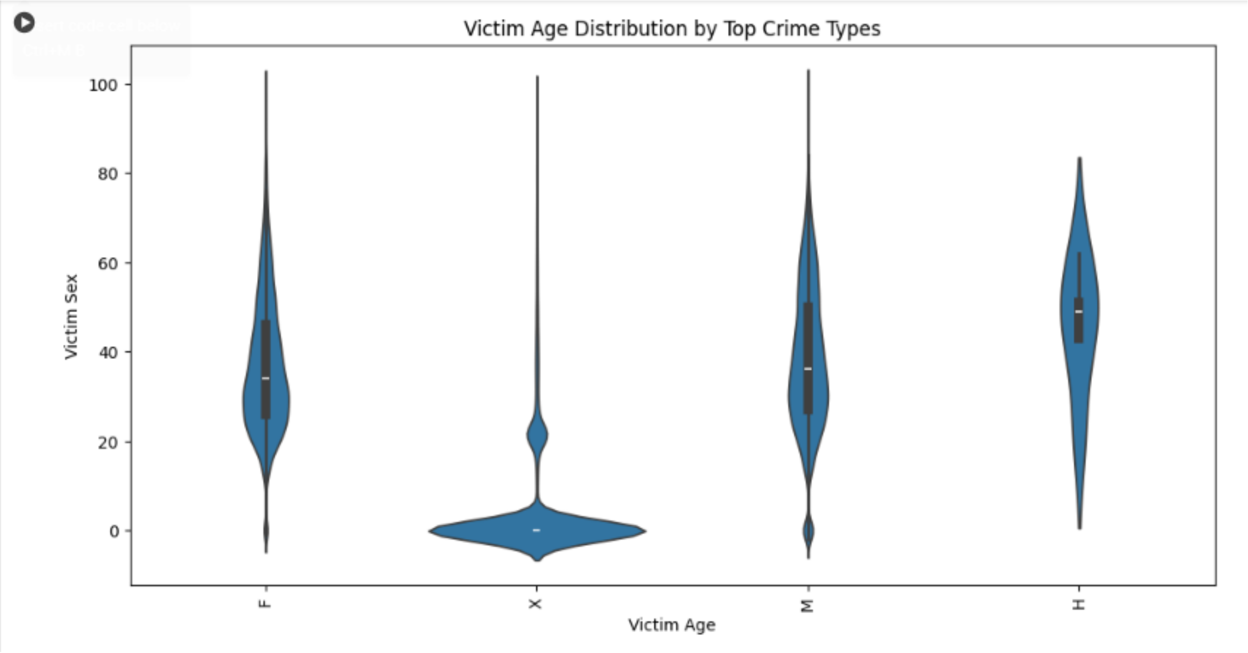
Association Victim Sex with Crime:



1.8: Victim Sex Relate to Crime

In the analysis of victim age group 325-35 has many types of crime types with respect to the data analytics.

Association Victim Sex with Victim Age to Crime:



1.9: Association of Victim Age with Victim Sex to Crime

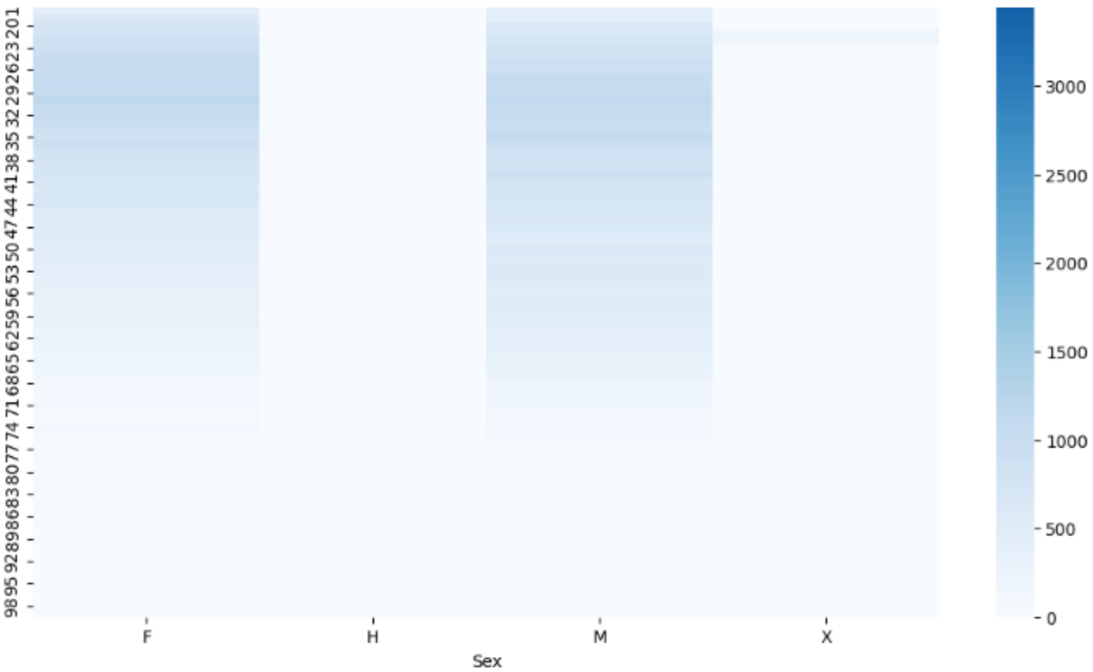
F - Female

M- Male

X- Not known

The analysis of victim age and victim sex related to the crime shows that females and males in the age group of 25-40 have a higher risk of crime.

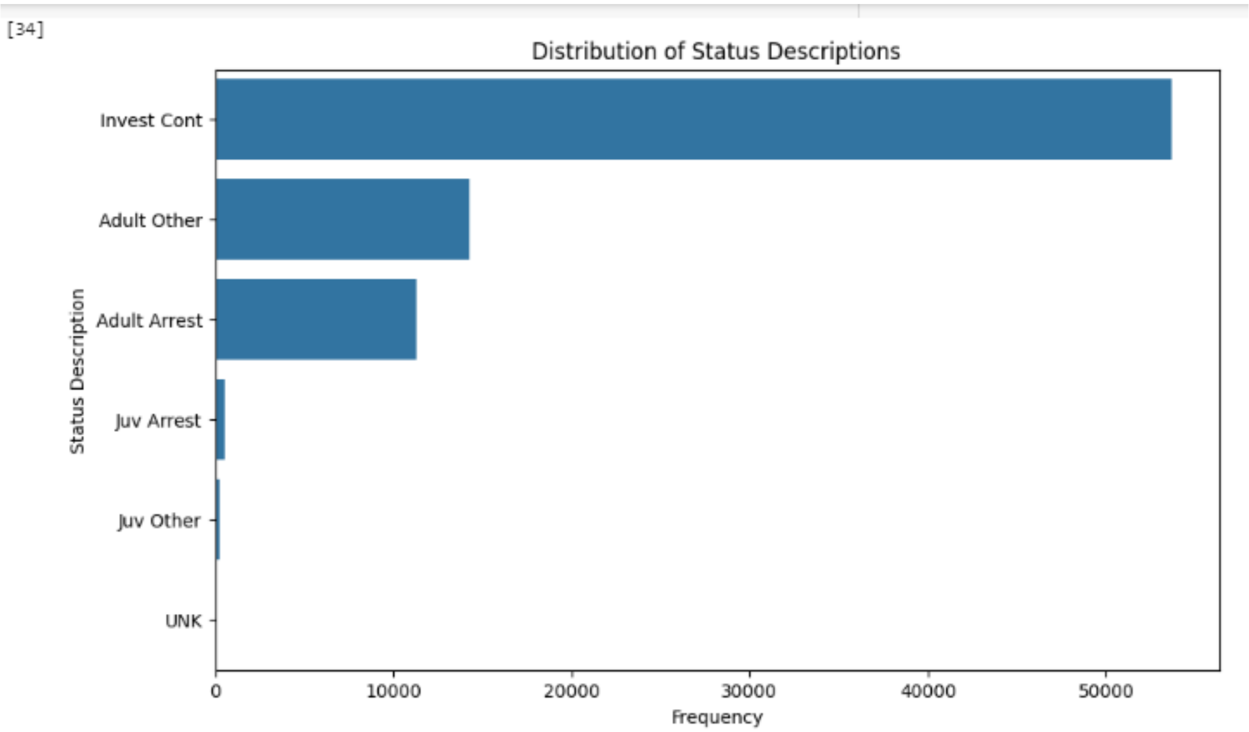
Association between Ages related to sex with crime:



2.0: Association of Age Relate to Crime

The graph denotes the age related to the sex

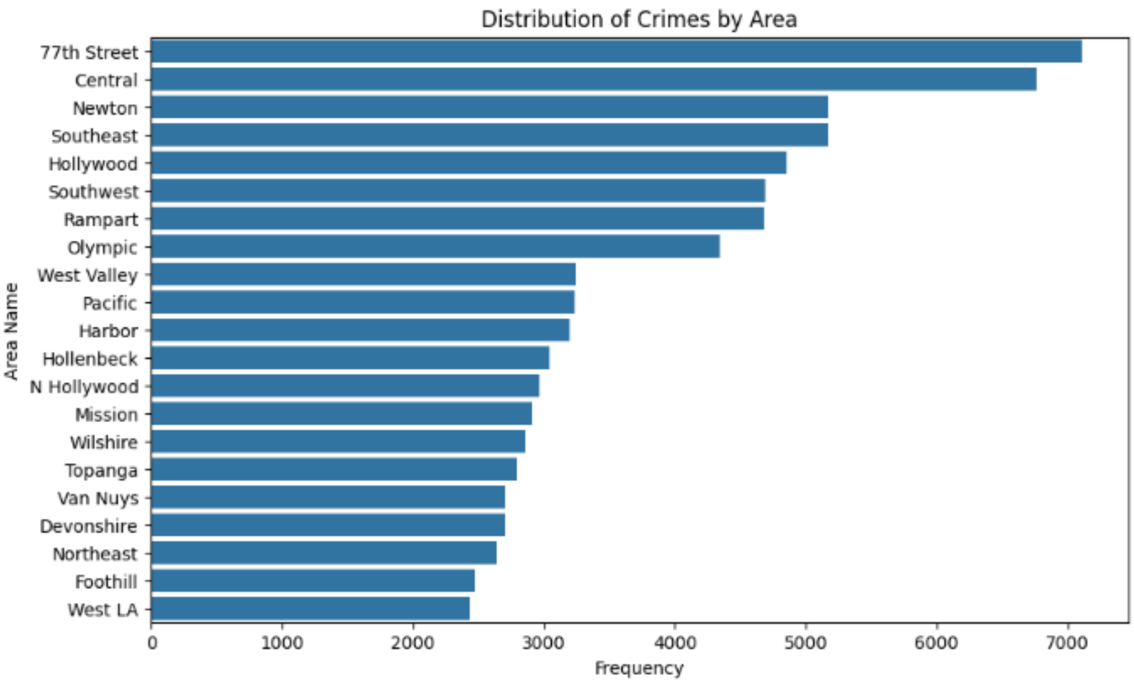
Association between crime status descriptions:



2.1: Association of status description to Frequency of crime

In the analysis, the Invest cont. has the highest crime records of more than 50000, and Adult others have records of nearly 20000.

Association between Crime Locations with Crime:

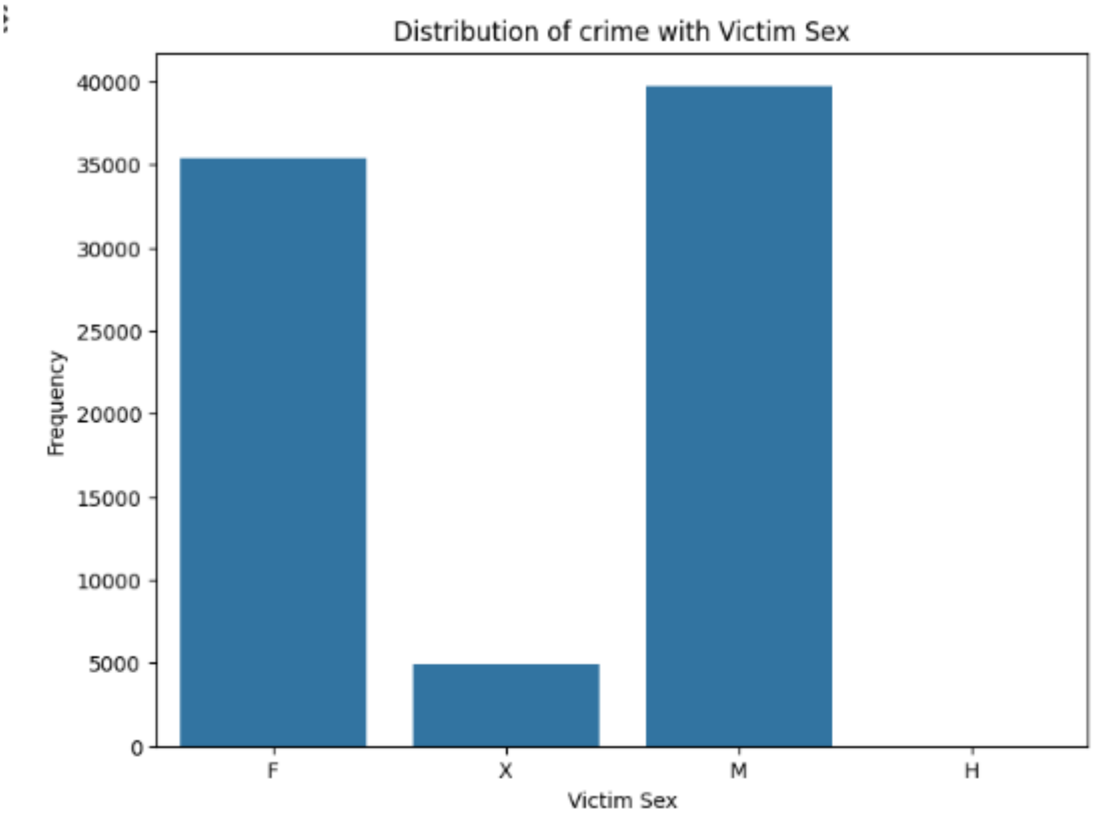


2.2: Association of Crime Location Relate to Crime

77th Street- higher crime records

West LA – lower crime records

Association of Victim Sex Relate to Crime



2.3: Victim Sex Relate to Crime

M-Male

F-Female

X-unknown

The male gender has higher crime records as compared with the female records.

Conclusion and Recommendations:

Extensive Research has shown a strong relationship between victim age and sex when it comes to crime identification. By addressing the Research question, it will be possible to identify the underlying reasons for crime prediction and how elements such as age, sex, and crime rate connect to crime's influence. These insights can then be used to inform strategic decision-making regarding the many aspects of crime prediction. This data-driven strategy will improve predictive analytics and optimize criminal records, and Physical crime records provide further detail on the specific offense that was committed there. Furthermore, the results will be used to develop various data analytics algorithms for detecting criminal activity. The probability of crime can be considerably decreased by considering variables like victim age, victim sex, area location, victim type, and victim sex, according to thorough data analytics performed with Python programming.

Research Question #3: What is the role of crime attributes description such as crime description, crime location, and date of occurrence in the prevalence of the crime?

Variables/ Attributes

Date Rptd, Date occ, Time occurred, Crm Cd Desc, Vict Age, Vict Sex, Weapon Desc, Status Desc, Location

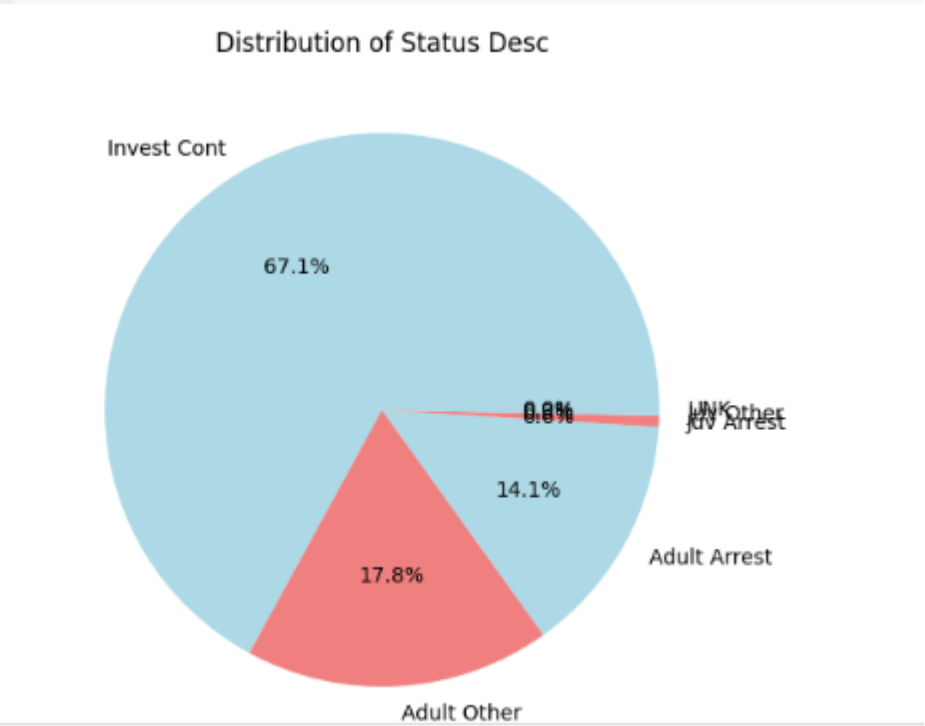
Managerial Decision-making

Answering this Research question will yield important information for criminology managers making decisions. Police can more effectively optimize their crime identification and prevention efforts by knowing the relationship between the crime description and the likelihood of crime. The field of criminology can use this information to recognize crimes and take proactive steps to improve crime identification accuracy, which could shorten the time needed to detect crimes and increase accuracy overall. Additionally, proactive decision-making can benefit from this study in order to reduce delays in the identification process, control crime in its early stages, and improve the effectiveness of crime prevention. In the end, data-driven decisions can be made to improve the accuracy levels of crime identification by finding patterns and factors contributing to crime identification.

Significant a significant role of crime description with Crime:

The role of crime attributes description such as crime description, crime location, and date of occurrence in the prevalence of the crime.

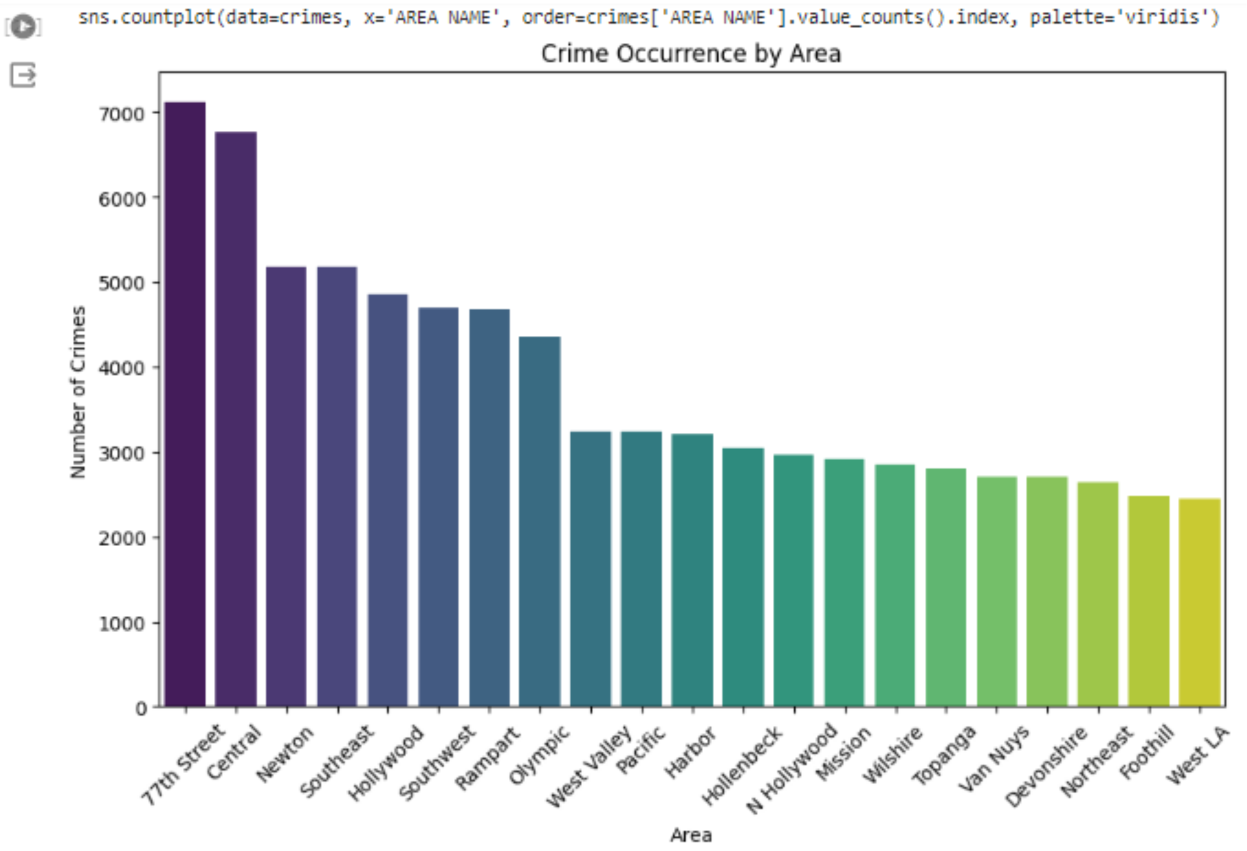
Role of Status Description in the prevalence of crime:



2.4: Role of Status Description with Relevance to Crime

The graphical chart shows that the Adult Arrest has the higher Crime description status whereas the Adult other crimes are 14%.

Role of Crime Location in the prevalence of crime:



```
sns.countplot(data=crimes, x='AREA NAME', order=crimes['AREA NAME'].value_counts().index, palette='viridis')
```

2.4: Role of Crime Location with Relevance to Crime

The dataset shows that 77th Street has the highest crime records, and the second highest is the Central Location of Los Angeles.

Below are the most common areas of Los Angeles where the crime rate is always higher in count:

I. Newton

II. South East

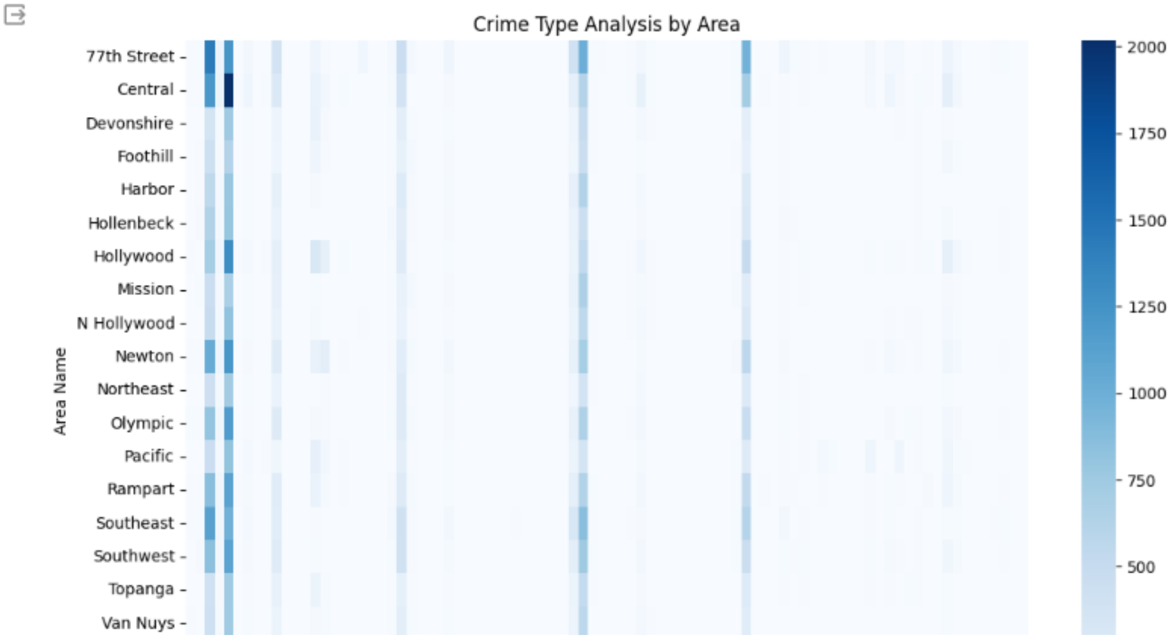
III. Hollywood

IV. Southwest

V. Rampart

VI. Olympic

Nearly 80% of the crime data are from the 77th Street, and the lowest crime rate is in the area of Van Nuys of Los Angeles.

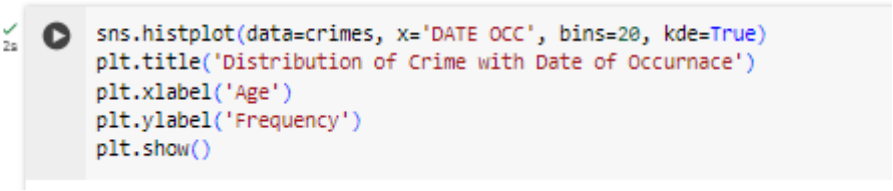


2.5 Crime analysis by area

Role of Date of Occurrence in the prevalence of crime:

The Date of Occurrence of the crime will be checked to make the identification of crime

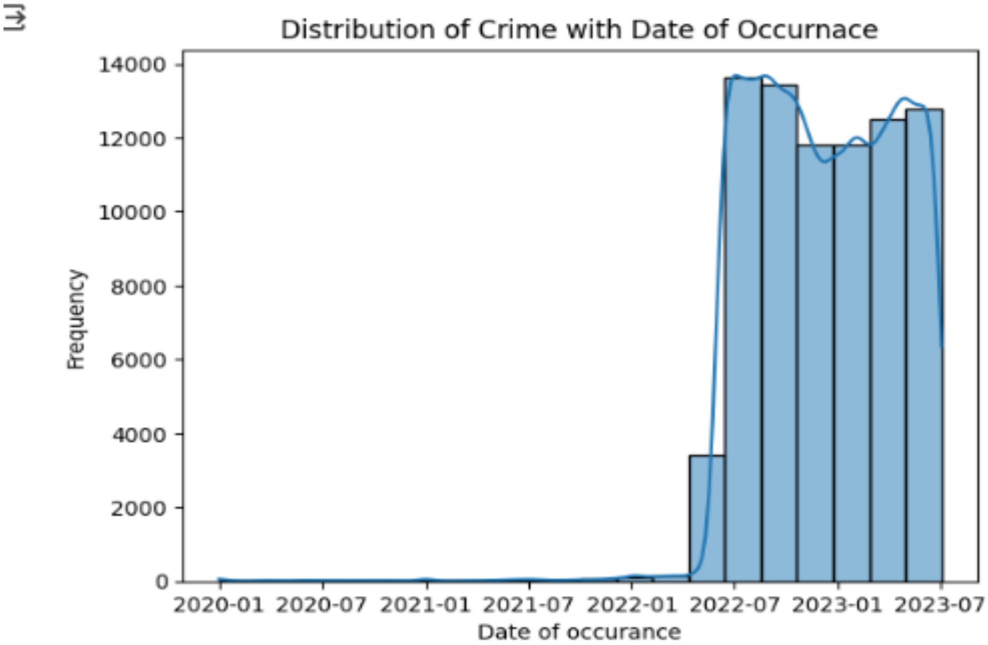
The below Python code is used to analyze visual graphs:



```
sns.histplot(data=crimes, x='DATE OCC', bins=20, kde=True)
plt.title('Distribution of Crime with Date of Occurnace')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```

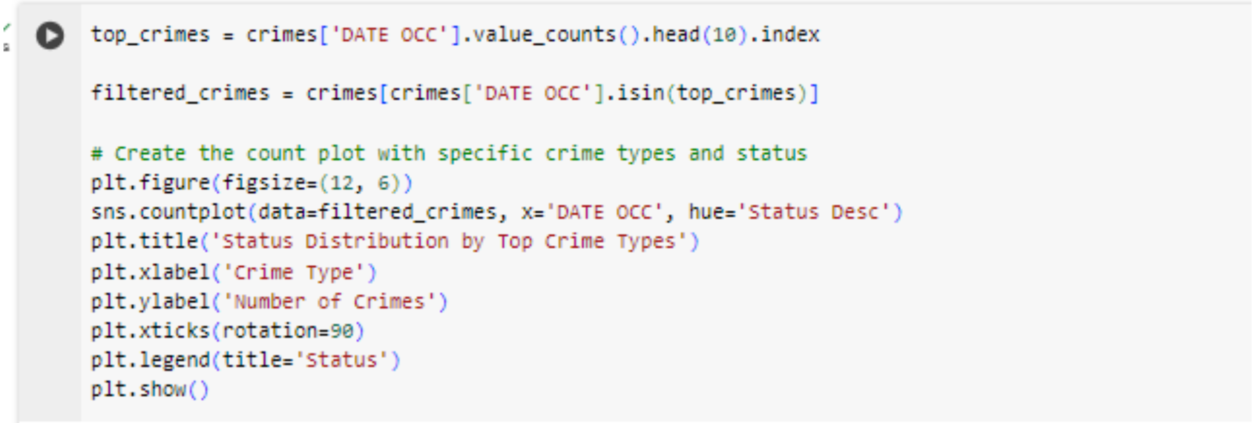
2.6 Code to plot Date of Occurrence with relevance to Crime

The histogram plot is used to denote the date of occurrence of the crime. The mean values are calculated with the offset points ploted with the graph. The results of the code execute the below graph visual:



2.7 Role of Date of Occurrence with Relevance to Crime

The code below shows the date of occurrence of the crime with the crime type, number of crimes, and status.



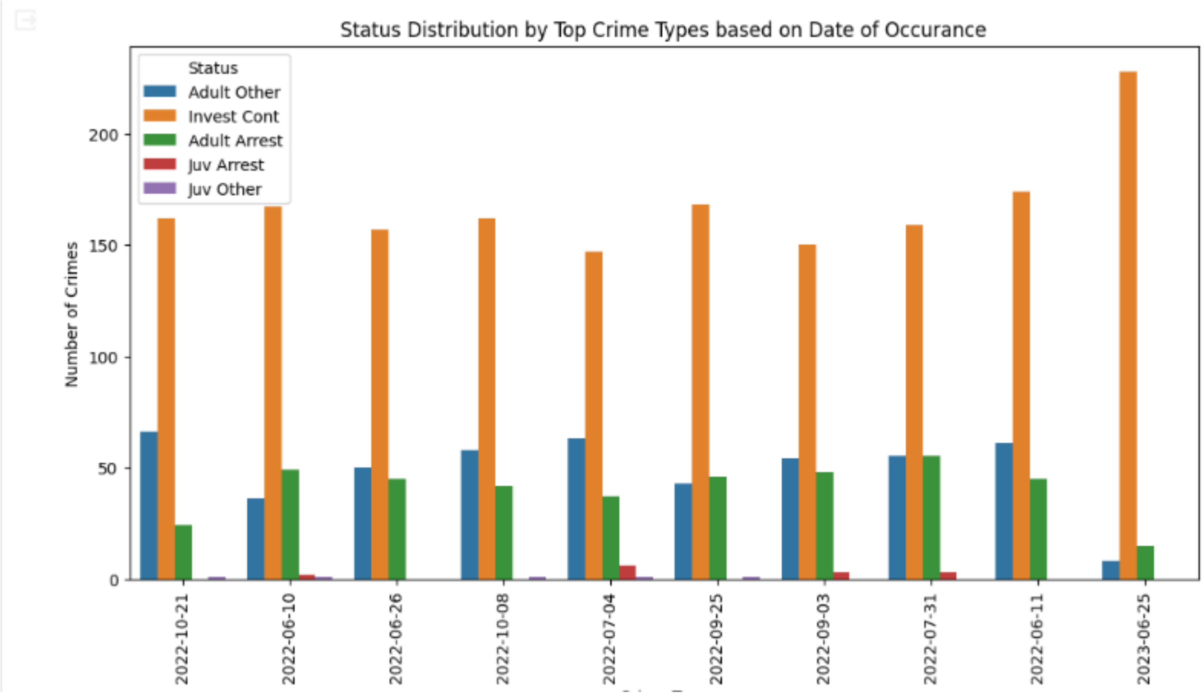
```
top_crimes = crimes['DATE OCC'].value_counts().head(10).index

filtered_crimes = crimes[crimes['DATE OCC'].isin(top_crimes)]

# Create the count plot with specific crime types and status
plt.figure(figsize=(12, 6))
sns.countplot(data=filtered_crimes, x='DATE OCC', hue='Status Desc')
plt.title('Status Distribution by Top Crime Types')
plt.xlabel('Crime Type')
plt.ylabel('Number of Crimes')
plt.xticks(rotation=90)
plt.legend(title='Status')
plt.show()
```

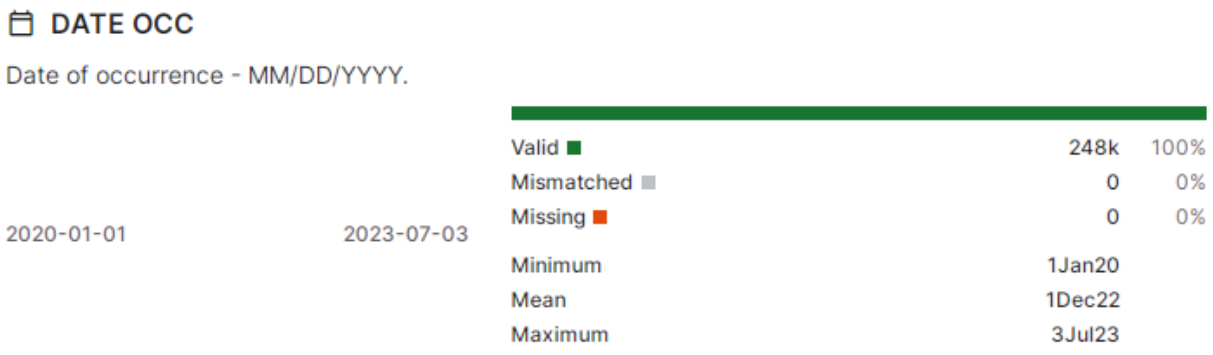
2.8 Code of Date of occurrence to Crime types

The application of technology and mathematics to solve the mysterious riddles hidden in data makes the topic of data science crime so fascinating! When it comes to anticipating illegal activity, data scientists are similar to detectives in that they carefully collect all relevant information and apply it to solve the case. They gather crime data from various sources, then clean it up and organize it so they can start analyzing it to forecast criminal activity.



2.9 Date-wise occurrence of Crime Analysis

In total, there are 247989 records present in the dataset. The higher crime records are mostly registered on 25/06/2023 and 11/06/2022. The Invest cont. is the higher crime data on 25/06/2023. Adult Arrest is a common type of crime occurrence on all types of days.



DATE OCC

Date of occurrence - MM/DD/YYYY.

2020-01-01 2023-07-03

Valid 🟩	248k	100%
Mismatched ▨	0	0%
Missing 🟥	0	0%
Minimum	1Jan20	
Mean	1Dec22	
Maximum	3Jul23	

3.0 Date occurrence analysis on crime

The date of occurrence based on the analysis shows the max value as 3/07/2023, the min value as 1/01/2020, and the mean value as 01/12/2023.

Conclusion and Recommendations:

The application of technology and mathematics to solve the mysterious riddles hidden in data makes the topic of data science crime so fascinating! When it comes to anticipating illegal activity, data scientists are similar to detectives in that they carefully collect all relevant information and apply it to solve the case. They gather crime data from various sources, then clean it up and organize it so they can start analyzing it to forecast criminal activity.

Research Question #4: Can machine learning models accurately predict crime based on the survey data from this dataset? Based on previous findings, are there any factor groupings that should not be included in the prediction/classification model?

Variables/ Attributes

Dr_no Area name, Crm Cd Desc, Vict Age, Vict Sex, Weapon Desc, Status Desc, Location, Date Rptd, Date occ, Time occ.

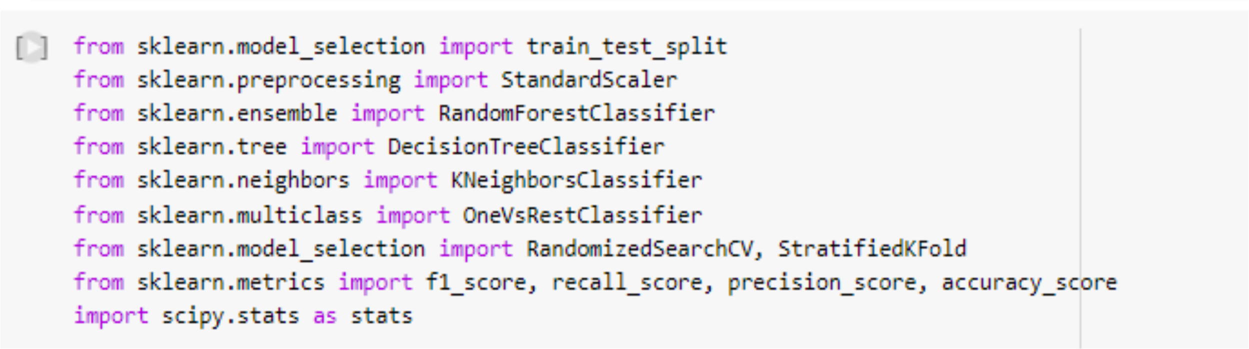
Managerial Decision-making

Police can make more informed decisions by understanding which elements influence crime identification. For example, they can provide precise information about the intent of the crime, provide accurate explanations for its occurrence, manage the crime, and produce results with higher overall accuracy. In addition to the aforementioned, the police will be able to implement more competitive preventative efforts and potentially lower crime rates in the future with the help of victim and crime reason discoveries. Machine learning will be useful in identifying crimes in their early stages and in pinpointing the precise causes of their occurrence.

Machine Learning Models in Crime Analysis:

The machine learning algorithms help in the detection of crime based on the dataset available. The existing dataset will be trained and given as input for crime identification and detection in the future.

Importing the Machine Learning Library Files:

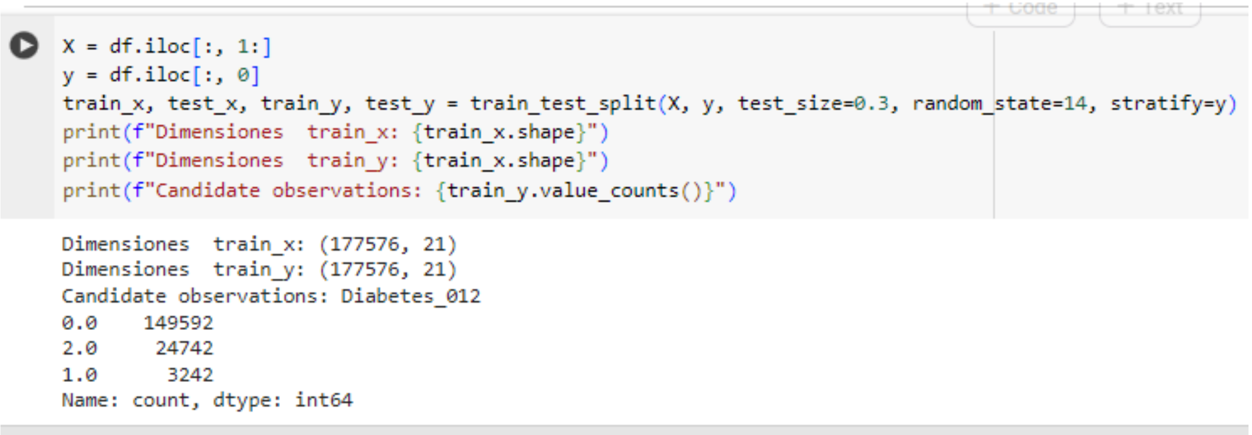


```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.multiclass import OneVsRestClassifier
from sklearn.model_selection import RandomizedSearchCV, StratifiedKFold
from sklearn.metrics import f1_score, recall_score, precision_score, accuracy_score
import scipy.stats as stats
```

3.1 Import the machine learning library files

Sklearn is a highly practical and resilient Python machine-learning library that offers a broad range of effective tools for statistical modeling and machine learning, such as classification, regression, clustering, and dimensionality reduction through a unified Python interface. This library, which has been predominantly written in Python, is constructed on the foundation of NumPy, SciPy, and Matplotlib.

Split the Train and Test data:

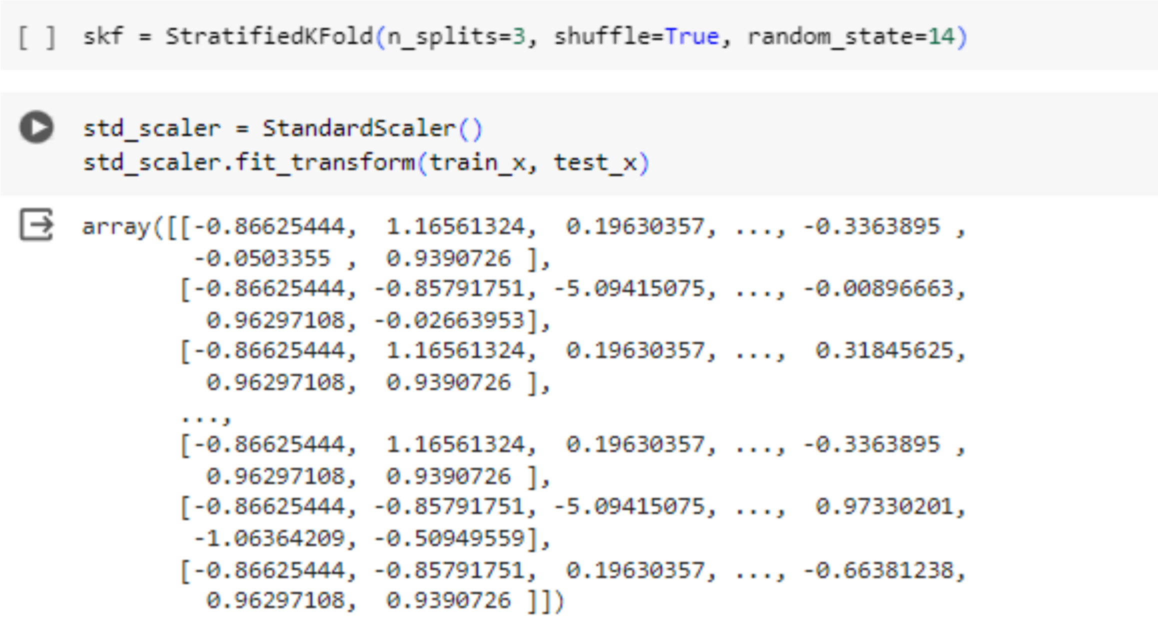


```
X = df.iloc[:, 1:]
y = df.iloc[:, 0]
train_x, test_x, train_y, test_y = train_test_split(X, y, test_size=0.3, random_state=14, stratify=y)
print(f"Dimensiones  train_x: {train_x.shape}")
print(f"Dimensiones  train_y: {train_x.shape}")
print(f"Candidate observations: {train_y.value_counts()}")
```

```
Dimensiones  train_x: (177576, 21)
Dimensiones  train_y: (177576, 21)
Candidate observations: Diabetes_012
0.0    149592
2.0     24742
1.0      3242
Name: count, dtype: int64
```

3.2 Splitting the training and test data

The total records are taken and split into training and testing. The dimensions of train_x contain 30% of records with 12 columns, and the dimensions of train_y contain 70% of records with 12 columns.



```
[ ] skf = StratifiedKFold(n_splits=3, shuffle=True, random_state=14)
```

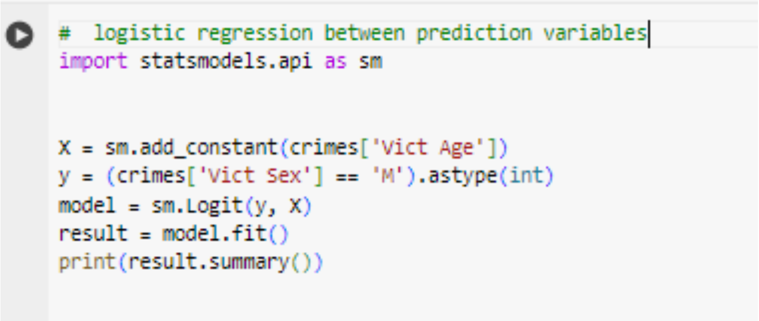
```
std_scaler = StandardScaler()
std_scaler.fit_transform(train_x, test_x)
```

```
array([[-0.86625444,  1.16561324,  0.19630357, ..., -0.3363895 ,
        -0.0503355 ,  0.9390726 ],
       [-0.86625444, -0.85791751, -5.09415075, ..., -0.00896663,
         0.96297108, -0.02663953],
       [-0.86625444,  1.16561324,  0.19630357, ...,  0.31845625,
         0.96297108,  0.9390726 ],
       ...,
       [-0.86625444,  1.16561324,  0.19630357, ..., -0.3363895 ,
         0.96297108,  0.9390726 ],
       [-0.86625444, -0.85791751, -5.09415075, ...,  0.97330201,
        -1.06364209, -0.50949559],
       [-0.86625444, -0.85791751,  0.19630357, ..., -0.66381238,
         0.96297108,  0.9390726 ]])
```

3.3 Standard Scalar Data Analysis

The results of the standard scalar show the array data binding with the values.

Implementing Logistic Regression Classifier:

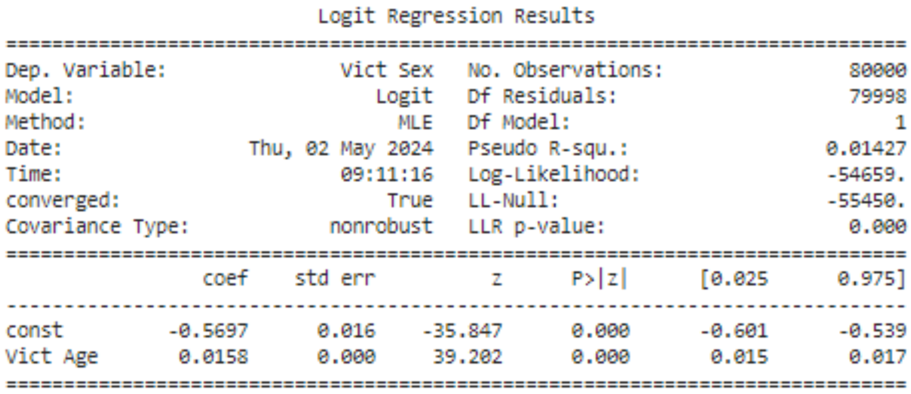


```
#  logistic regression between prediction variables
import statsmodels.api as sm


X = sm.add_constant(crimes['Vict Age'])
y = (crimes['Vict Sex'] == 'M').astype(int)
model = sm.Logit(y, X)
result = model.fit()
print(result.summary())
```

3.4 Logistic Regression Classification

The logistic regression model is applied in the machine learning algorihtm model.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:               Vict Sex   No. Observations:                80000
Model:                          Logit   Df Residuals:                    79998
Method:                           MLE   Df Model:                            1
Date:                Thu, 02 May 2024   Pseudo R-squ.:                 0.01427
Time:                        09:11:16   Log-Likelihood:                -54659.
converged:                       True   LL-Null:                       -55450.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.5697      0.016    -35.847      0.000      -0.601      -0.539
Vict Age       0.0158      0.000     39.202      0.000       0.015       0.017
==============================================================================
```

3.9 Results of Logistic Regression

Conclusion and Recommendations:

Extensive Research has shown a strong relationship between the characteristics of crime and its identification. By concentrating on these criminal variables, people can proactively protect their lives and reduce their chance of being victims of crime by utilizing in-depth data analytics through Python programming. The machine learning model is used to detect crimes from the provided dataset by training and testing the data. The high accuracy scores show the prediction is clear and accurate. According to the earlier study portion conclusions, attributes like DR_NO Date Reported and location are not required for the suggested approach techniques. It is advised that the accuracy of the dataset be further enhanced.

Research Question #5: How does the data analytics help identify and control the crime with respect to the Area Name, Victim Age, Victim Sex, weapons, and how crime can be controlled?

Variables/ Attributes

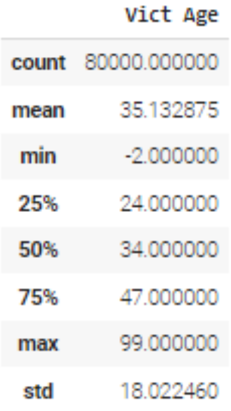
Location, Dr_no Area name, Crm Cd Desc, Vict Age, Vict Sex, Weapon Desc, Status Desc, Date Rptd, Date occ, Time occ.

Managerial Decision-making

Because it defines data analytics as identifying crime and facilitates quick access to the crime field, this question can aid in making better managerial judgments. This can make it easier for police to apply specific characteristics like age, sex, and crime type identification and optimize criminal identification for higher accuracy. With this knowledge, one can take a proactive approach that can facilitate the detection of crimes more smoothly and reliably and enable the implementation of the necessary preventive steps to control future crimes.

Managing Crime With Respect To the Age Group and Gender:

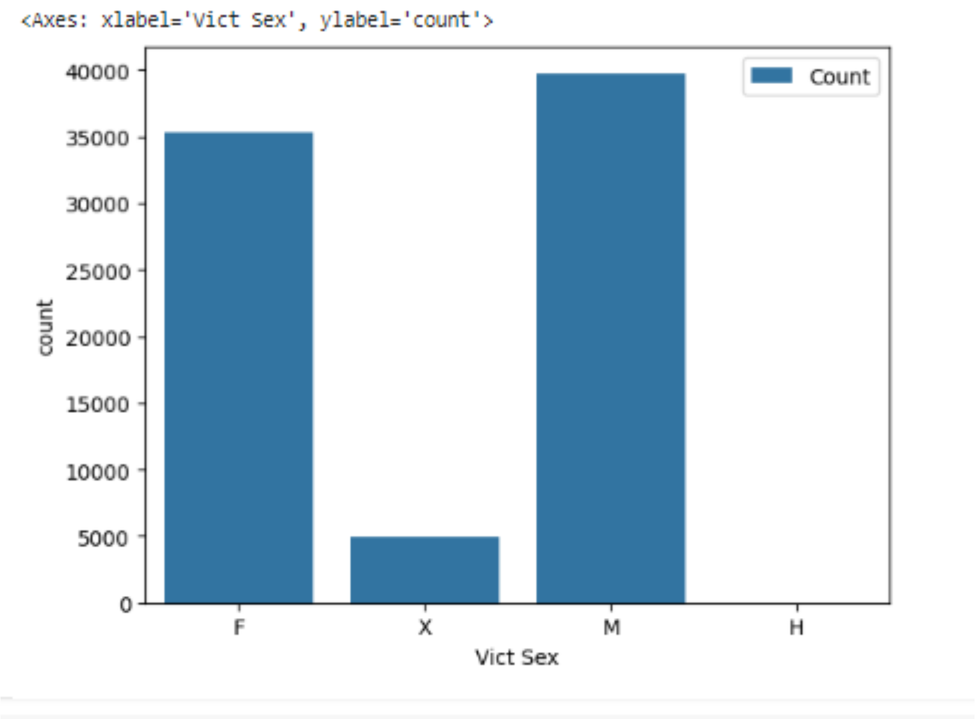
The crime is identified mostly based on the age group only. The data analytics helps to find accurate analysis with respect to the age groups.



	Vict Age
count	80000.000000
mean	35.132875
min	-2.000000
25%	24.000000
50%	34.000000
75%	47.000000
max	99.000000
std	18.022460

4.2 Metrics of Victim Age

The metrics show that the mean value of Age is 35.132875, the min is -2.0, and the max is 99.0.



```
<Axes: xlabel='Vict Sex', ylabel='count'>
```

4.3 Data Visualization of Victim Gender

The indicators of the data are:

F ➔ Female

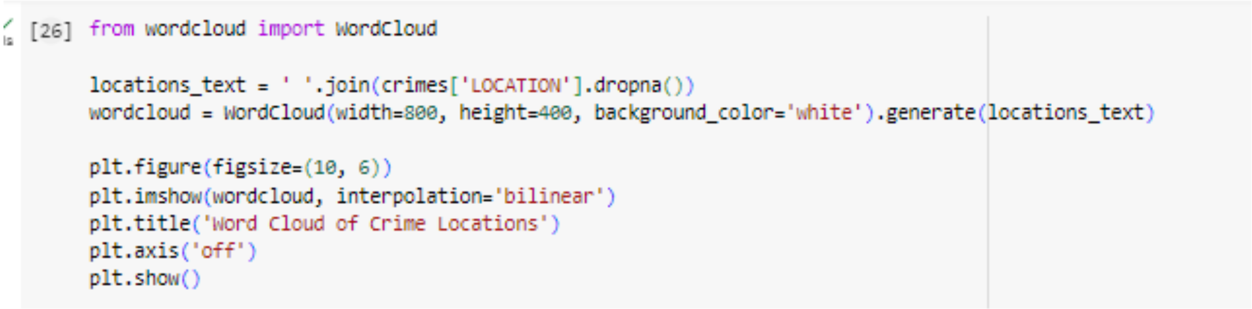
M ➔ Male

X →not known

The data visualization of the gender shows that Females are 35000 and the male category are around 40000. The age data has been mapped with the gender to analyze the crime presence.

Managing Crime With Respect To the Location:

The below code is used to map the location of the crime with the large dataset using a word cloud.

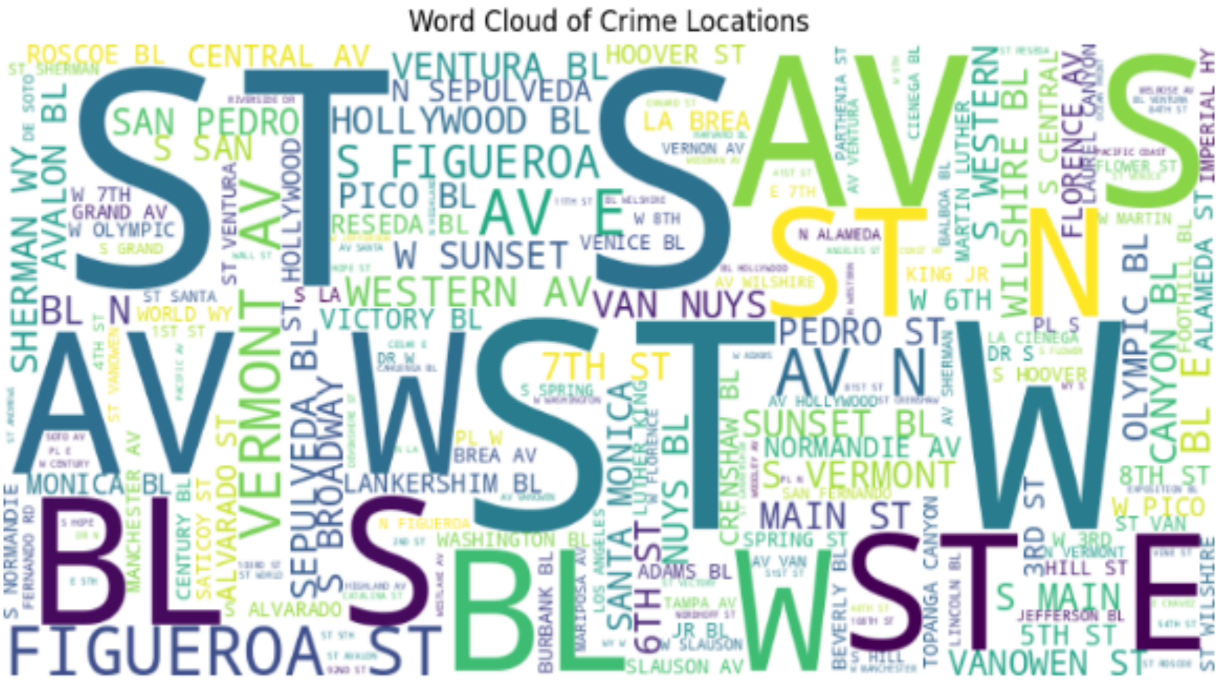


```
[26] from wordcloud import WordCloud

     locations_text = ' '.join(crimes['LOCATION'].dropna())
     wordcloud = WordCloud(width=800, height=400, background_color='white').generate(locations_text)

     plt.figure(figsize=(10, 6))
     plt.imshow(wordcloud, interpolation='bilinear')
     plt.title('Word Cloud of Crime Locations')
     plt.axis('off')
     plt.show()
```

4.4 Code with data analysis of Location with word cloud

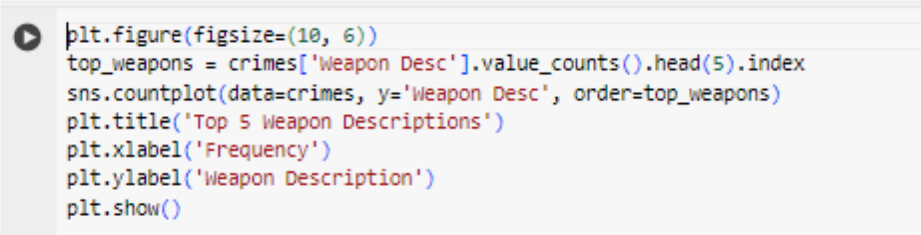


4.4 WordCloud with Locations

The graph visualizes the presence of crime with respect to the locations using word cloud with metrics.

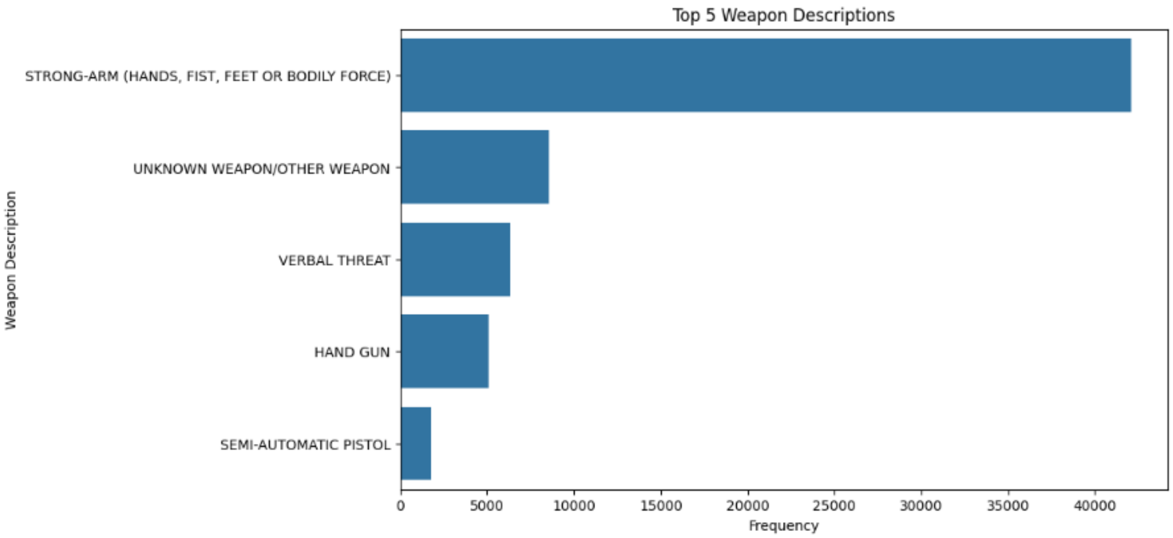
Managing Crime With Respect To the Weapons:

Analyzing weapons used in crime plays an important factor in the data analytics to identify the crime factors.



```
plt.figure(figsize=(10, 6))
top_weapons = crimes['Weapon Desc'].value_counts().head(5).index
sns.countplot(data=crimes, y='Weapon Desc', order=top_weapons)
plt.title('Top 5 Weapon Descriptions')
plt.xlabel('Frequency')
plt.ylabel('Weapon Description')
plt.show()
```

4.5 Code of Weapons in Crime



The top 5 weapons were identified with the help of graphical visualization. The strong arms used are Hands, Fist, Feet, or Badly Force. The string arms are more than 40000, and hand guns used in crime are nearly 7500.

Recommendation:

Based on our precise machine learning system, the following are our suggestions for those making decisions on crime predictions:

1. Integration with Operations: Integrate the prediction model into routine operational operations for crime prediction in order to provide accuracy in crime detection.

2. Resource Allocation: Utilize the staff member to distribute resources as effectively as possible in order to get precise data that aids in the prevention of crime.

3. Strategic Planning: Incorporate the forecasts into the processes for strategic planning so that the people in charge may anticipate any disruptions and prepare the required modifications in advance. Improving overall operational efficiency in crime prediction may be necessary to achieve this.

Appendices Appendix: The Accuracy machine learning algorithm models

KNN Random Classifier:

Accuracy: 80%

REFERENCES

Data

- Crime Health Indicator data:

  https://www.kaggle.com/datasets/shayalvaghasiya/los-angeles-crimes

Software and Packages

- Tool: Python Colab
- Software: Python
- Libraries: ggplot, pandas, machine learning libraries

Contextual Resources

1) KS: Author", Definition and Types of Crime Analysis" International Association of Criminalists. (2014). Definition and types of crime analysis [White Paper 2014-02]. https://www.iaca.net/Publications/Whitepapers/iacawp_2014_02_definition_types_crime_analysis.pdf

2). Ubon Thongsatapornwatana, "A survey of data mining techniques for analyzing crime patterns", Defence Technology (ACDT) 2016 Second Asian Conference on, pp. 123-128, 2016.

3).S. Sathyadevan, M. Devan, and S. Surya Gangadharan, "Crime analysis and prediction using data mining," in Networks Soft Computing (ICNSC), 2014 First International Conference on, Aug 2014, pp. 406– 412.

4)..Isuru Jayaweera, Chamath Sajeewa, Sampath Liyanage, Tharindu Wijewardane, Indika Perera “, Crime Analytics: Analysis of Crimes Through Newspaper Articles,” Moratuwa Engineering Research Conference (MERCon), 2015

5).D.Usha, Dr.K.Rameshkumar,” A Complete Survey on Application of Frequent Pattern Mining and Association Rule Mining on Crime Pattern Mining,” Volume 3, No.4, April 2014 International Journal of Advances in Computer Science and Technology

6)International Association of Crime Analysts - Wikipedia

https://en.wikipedia.org/wiki/International_Association_of_Crime_Analysts.

7).P. Thongthai and S. Srisuk,” An Analysis of Data Mining Applications in Crime Domain,” Computer and Information Technology Workshops, 2008. CIT Workshops 2008. IEEE 8th International Conference on Computer and Information Technology Workshops

8). Chung-Hsien Yu 1, Max W. Ward1, Melissa Morabito 2, and Wei Ding1,” Crime Forecasting Using Data Mining Techniques .”2011 11th IEEE International Conference on Data Mining Workshops

9). Manish Gupta, B. Chandra, and M. P. Gupta,” Crime Data Mining for Indian Police Information System.”Indian Institute of Technology Delhi, Hauz Khas, New Delhi. India 110 016.

10). Revathy Krishnamurthy, J. Satheesh Kumar”, Survey of data mining techniques on crime data analysis,” Vol 01, Issue 02, December 2012 International Journal of Data Mining Techniques and Applications ISSN: 2278-2419