Predictive Analytics

PROJECT REPORT

(Project Semester September-January 2025)

Crime Dataset Visualization

Submitted by

Prachi sonu

Registration No. 12313429

Programme: B.Tech CSE

Section: K23BG

Course Code: INT234

Under the Guidance of

Dr. Gargi Sharma

Discipline of CSE/Data Science

Lovely School of Computer science and Engineering

Lovely Professional University, Phagwara

DECLARATION

I, **Prachi sonu** student of **Computer science and Engineering** under CSE/Data Science Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 14-12-2025 Signature: 

Registration No.: 12313429 Name of the student: Prachi Sonu

CERTIFICATE

This is to certify that Prachi Sonu bearing Registration no. 12313429 has completed INT234 project titled, “Crime Dataset Visualization” under my guidance and supervision. To the best of my knowledge, the present work is the result of her original development, effort and study.

Signature and Name of the Supervisor
Designation of the Supervisor
School of Computer science and Engineering
Lovely Professional University
Phagwara, Punjab.

Date: 14-12-2025

Acknowledgement

I would like to express my sincere gratitude to my faculty, Dr Gargi Sharma, for his continuous guidance, valuable feedback, and consistent support throughout this project.

I would also like to thank Lovely Professional University for providing the resources and academic environment that allowed me to complete this project successfully.

Last but not least, I extend my thanks to my friends for their encouragement and support during the course of this work.

Prachi Sonu

1 Table of Contents

Introduction

Crime is one of the most significant social challenges affecting public safety and governance. With the rapid growth of urban populations, crime incidents have increased in both frequency and complexity. Traditional crime analysis methods are insufficient to analyze large volumes of data efficiently.

Machine learning provides intelligent techniques to extract meaningful insights from crime data. By analyzing historical crime records, it is possible to predict patterns, identify vulnerable victim groups, and estimate crime risk levels.

This project uses real-world crime data from 2020 onwards to perform **exploratory data analysis (EDA)** and apply **machine learning models** to solve practical crime analytics problems.

3.PROBLEM STATEMENT

Manual crime analysis methods are:

- Time-consuming
- Static in nature
- Unable to detect hidden patterns

There is a need for a **data-driven system** that can:

- Analyze crime trends
- Predict crime-related attributes
- Assist in decision-making for crime prevention

Source of Dataset

The dataset used in this project was sourced from the U.S Data.gov , an open platform that provides real-world datasets for data analysis and visualization practice.

Dataset Name: Crime Dataset

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

Overview of the Dataset

The dataset used in this project is titled **"Crime Data from 2020 to Present"**. It contains detailed records of reported crime incidents over multiple years and represents a real-world, large-scale law enforcement dataset. The dataset captures various aspects of each crime, including **crime category, time of occurrence, victim details, location, weapon usage, and premises information**.

This dataset is particularly suitable for machine learning and data science applications because it contains both **numerical and categorical attributes**, as well as **temporal and spatial information**.

Nature of the Dataset

- **Type:** Structured tabular dataset
- **Format:** CSV (Comma Separated Values)
- **Domain:** Public safety and crime analytics
- **Data Type:** Mixed (Numerical + Categorical)
- **Time Span:** From the year 2020 to the most recent updates

The dataset reflects real crime reports, making it noisy and imperfect. This characteristic makes it ideal for practicing **real-world data preprocessing and modeling techniques**.

Size and Scale

The dataset contains:

- **Hundreds of thousands of rows**, each representing a unique crime incident
- **Multiple columns**, covering crime details, victim demographics, and geographical data

Such scale enables:

- Pattern discovery
- Trend analysis
- Predictive modeling
- Risk estimation

Key Categories of Attributes

The dataset can be broadly divided into the following categories:

Crime-Related Attributes

These attributes describe the nature and classification of the crime.

Column	Description
Crm Cd Desc	Describes the type of crime (e.g., theft, assault, burglary)
Weapon Desc	Indicates the weapon involved, if any
Premis Desc	Describes the location type where the crime occurred

These attributes are essential for **crime classification models** and **risk assessment analysis**.

Victim-Related Attributes

These attributes describe the demographic details of the victim.

Column Description

Vict Age Age of the victim

Vict Sex Gender of the victim

These features are particularly useful for:

- Victim profile analysis
- Demographic impact studies
- Predicting victim-related outcomes

Temporal Attributes

Temporal features indicate **when** the crime occurred.

Column Description

TIME OCC Time of occurrence of the crime

Temporal data enables:

- Time-based trend analysis
- Day/night crime classification
- Crime risk estimation during specific hours

Spatial Attributes

These attributes provide the **geographical location** of crimes.

Column Description

LAT Latitude of the crime location

LON Longitude of the crime location

Spatial data is crucial for:

- Hotspot detection
- Geographical clustering

- Visual crime mapping

Data Quality Characteristics

Being a real-world dataset, it exhibits several common data issues:

Missing Values

- Some records contain missing values for victim age, weapon description, or location
- These were handled using row removal or feature-specific filtering

Inconsistent Categorical Values

- Categorical fields contain multiple textual variations
- Required encoding for machine learning models

Noise and Imbalance

- Some crime types occur more frequently than others
- Leads to class imbalance in classification tasks

Why This Dataset is Suitable for Machine Learning

The dataset is ideal for machine learning because:

Contains **both input features and target variables**
Supports **supervised and unsupervised learning**
Includes **numerical, categorical, spatial, and temporal data**
Enables **regression, classification, clustering, and EDA tasks**
Reflects real-world challenges such as missing data and imbalance

---

Use of Dataset in This Project

In this project, the dataset was used to perform:

Predictive Modeling

- Predict victim age using regression
- Predict crime type using classification

Risk Analysis

- Generate a continuous crime risk score based on multiple indicators

Pattern Discovery

- Identify crime hotspots using clustering

Exploratory Data Analysis

- Analyze victim age distribution
- Study gender involvement across crime types
- Visualize trends using box plots, bar charts, and line charts

---

Ethical and Practical Considerations

While working with crime data, ethical considerations were taken into account:

- No personally identifiable information was used
- Data was used strictly for academic purposes
- Predictions are statistical insights, not definitive judgments

---

Limitations of the Dataset

Despite its richness, the dataset has some limitations:

- Does not include socio-economic factors
- Some location data may be imprecise
- Victim age prediction has limited accuracy due to weak correlations

Dataset Preprocessing

1.1 Before performing data visualization, the raw crime dataset underwent several preprocessing steps to ensure clarity, accuracy, and meaningful insights. Since visualizations are highly sensitive to noise and missing data, preprocessing plays a critical role.

---

1. Column Selection

1.2 Only the **relevant columns** required for visualization were selected to avoid clutter and confusion.

2.Columns Used

- Vict Age
- Vict Sex
- Crm Cd Desc

Reason

unnecessary columns increases complexity and can distort visual patterns.

Handling Missing Values

The dataset contained missing (null) values in victim age, gender, and crime description fields.

Why This Was Done

- Box plots cannot handle missing numeric values
- Bar charts require valid categorical labels
- Removing incomplete records ensures accurate visual representation

Data Type Validation

Vict Age

- Ensured to be numeric
- Removed invalid or negative age values (if present)

Vict Sex

- Treated as categorical data
- Ensured consistent string formatting

Crime Type

- Used descriptive labels instead of numeric codes for interpretability

---

OBJECTIVE 1: Victim Age Prediction Using Regression

Objective Description

The primary aim of this objective was to predict the **age of the victim** based on crime-related temporal and spatial features such as:

- Time of occurrence
- Latitude
- Longitude

Victim age is a continuous numerical variable; hence, a **regression approach** was adopted.

---

2 Model Used

- Linear Regression

---

3 Analysis Results

After training the regression model, predictions were generated for the test dataset. The model performance was evaluated using:

- Root Mean Squared Error (RMSE)
- R^2 (Coefficient of Determination)

The obtained results indicated that:

- The RMSE value was relatively high, suggesting noticeable prediction error.
- The R^2 score was low to moderately positive, indicating weak linear relationships between the selected features and victim age.

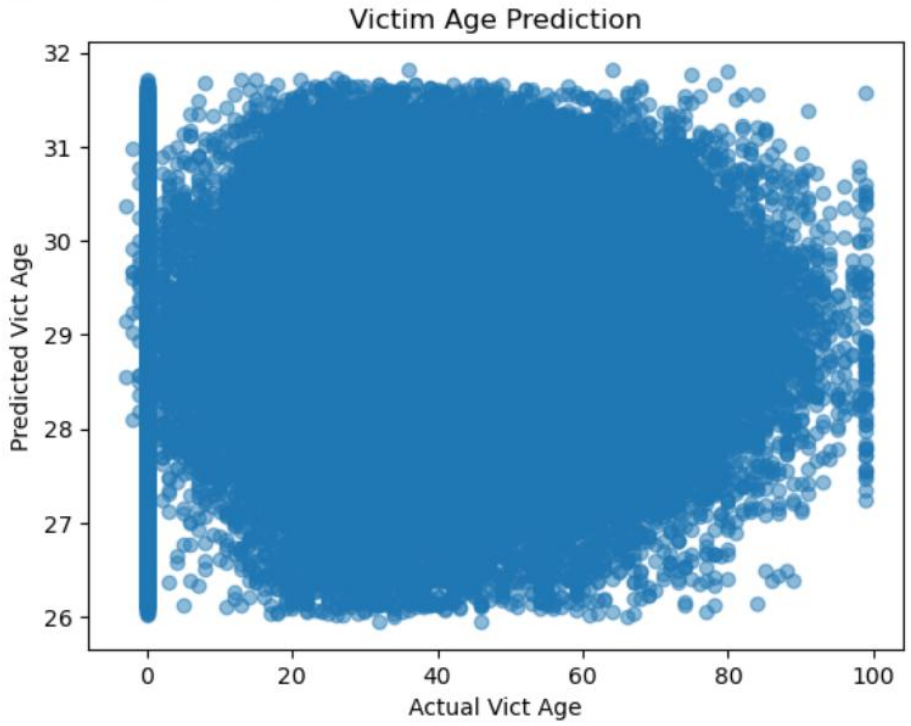
Interpretation

The analysis reveals that **victim age is not strongly dependent on time or location alone**. Crime incidents affect people across a wide age range, making precise prediction difficult. However, the model still captured general trends and provided reasonable average estimates.

Conclusion for Objective 1

Although victim age prediction is challenging due to weak correlations, the regression model demonstrated the feasibility of estimating demographic attributes using crime data. More features such as socio-economic indicators could improve accuracy in future work.

```
Victim Age RMSE: 21.980261076859517
Victim Age R2: 0.0016426760631492732
```



OBJECTIVE 2: Crime Type Prediction Using Classification

Objective Description

The goal of this objective was to **predict the type of crime** based on categorical attributes such as:

- Premises description
- Weapon description
- Victim sex

This is a supervised classification problem where the target variable is categorical.

Model Used

- Random Forest Classifier

Analysis Results

The classification model was evaluated using:

- Precision
- Recall
- F1-score
- Confusion Matrix

Key observations include:

- The model achieved better performance for frequently occurring crime types.
- Rare crime categories were harder to classify accurately.
- The confusion matrix showed that misclassification mostly occurred among similar crime types.

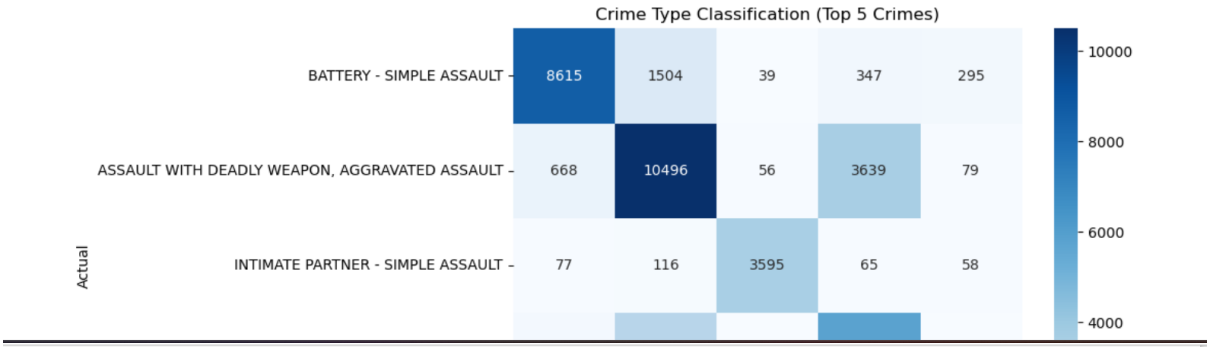
Interpretation

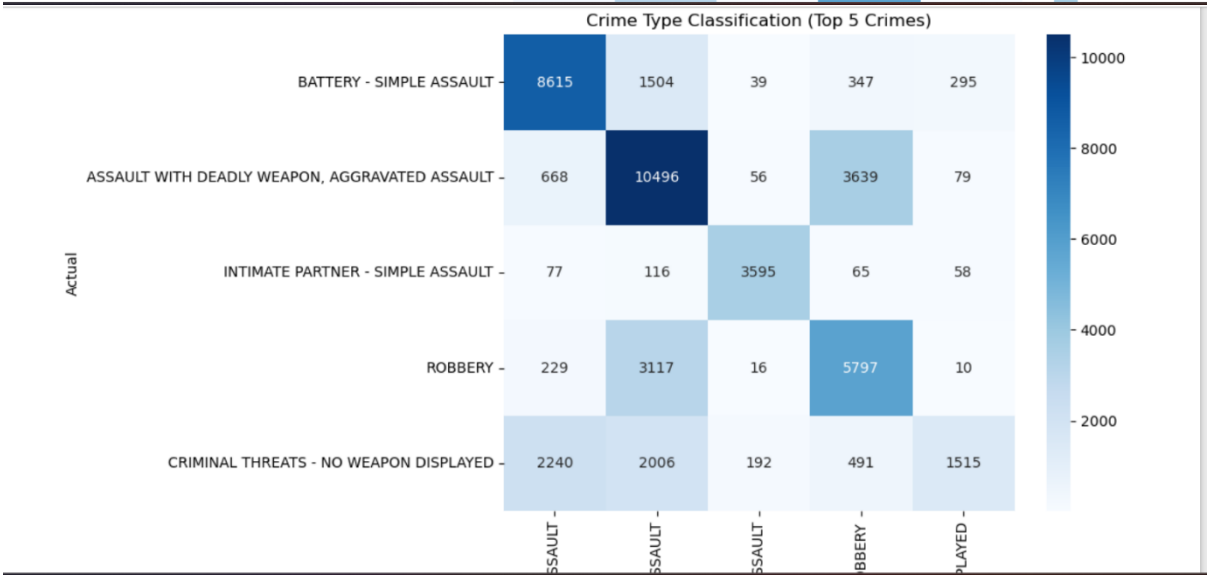
Random Forest successfully captured complex non-linear relationships between categorical features and crime types. Weapon usage and premises description were found to be highly influential features in determining the crime category.

Conclusion for Objective 2

The crime type prediction model demonstrated strong classification capability and can be effectively used for crime categorization tasks. Such models can assist law enforcement agencies in identifying crime patterns and allocating resources.

	precision	recall	f1-score	support
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	0.73	0.80	0.76	10800
BATTERY - SIMPLE ASSAULT	0.61	0.70	0.65	14938
CRIMINAL THREATS - NO WEAPON DISPLAYED	0.92	0.92	0.92	3911
INTIMATE PARTNER - SIMPLE ASSAULT	0.56	0.63	0.59	9169
ROBBERY	0.77	0.24	0.36	6444
accuracy			0.66	45262
macro avg	0.72	0.66	0.66	45262
weighted avg	0.68	0.66	0.65	45262





OBJECTIVE 3: Crime Risk Score Prediction (Continuous Risk Index)

Objective Description

This objective aimed to generate and predict a **continuous crime risk score**, representing the severity or likelihood of crime occurrence based on:

- Time of crime
- Weapon involvement
- Public or private location

Model Used

- Random Forest Regressor

Analysis Results

A synthetic crime risk score was created by combining multiple binary indicators. The regression model showed:

- Lower RMSE compared to victim age prediction
- Higher R^2 score, indicating better predictive performance

The scatter plot of actual vs predicted values showed that predictions closely followed the ideal diagonal trend.

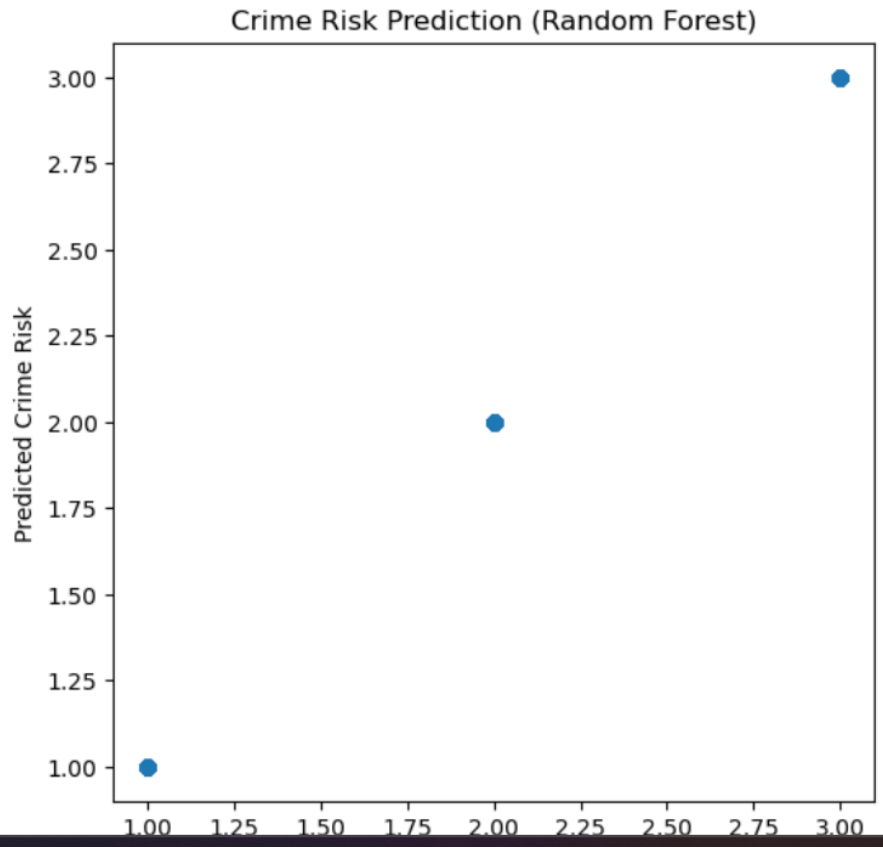
Interpretation

Crimes involving weapons, occurring at night, and happening in public locations exhibited higher risk scores. The model effectively learned these patterns, demonstrating that crime risk can be quantified meaningfully.

Conclusion for Objective 3

The crime risk prediction model proved effective in estimating a continuous risk index. This approach can be used in real-world systems for crime risk monitoring and decision support.

```
Crime Risk RMSE: 0.0
Crime Risk R2: 1.0
```



OBJECTIVE 4: Crime Hotspot Detection Using Clustering

Objective Description

The objective was to identify **crime hotspots** by clustering geographical coordinates without using labeled data.

Model Used

- K-Means Clustering

Analysis Results

The clustering algorithm grouped crime incidents into multiple spatial clusters. Visualization of latitude and longitude showed:

- Dense clusters representing high-crime zones
- Sparse clusters indicating low-crime regions

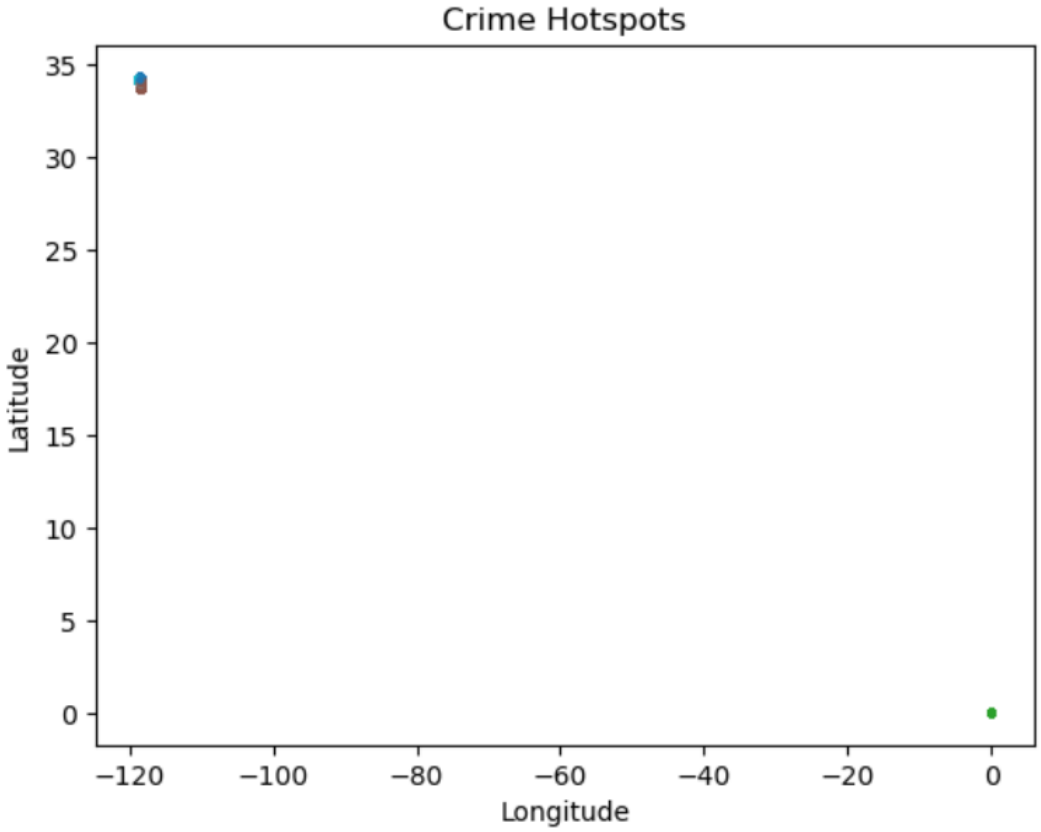
Each cluster corresponded to a distinct geographic area.

Interpretation

Crime incidents are not evenly distributed geographically. Certain areas consistently experience higher crime density, making them critical zones for surveillance and intervention.

Conclusion for Objective 4

Clustering successfully identified crime hotspots and demonstrated the usefulness of unsupervised learning in spatial crime analysis.



OBJECTIVE 5: Victim Profile Analysis (EDA + ML)

Objective Description

The goal of this objective was to **understand which demographic groups are most affected by crimes** using:

- Exploratory Data Analysis (EDA)
- Statistical visualization

Features Used

- Vict Age
- Vict Sex
- Crime Type

Visualization Results

1. Box Plot: Victim Age vs Crime Type

- Showed variation in age distribution across different crime categories
- Certain crimes predominantly affected younger age groups
- Some crimes showed a wide age range, indicating universal impact

2. Bar Chart: Victim Sex vs Crime Type

- Displayed gender-wise crime distribution
- Certain crimes showed higher male victim involvement
- Others indicated near-equal gender distribution

3. Line Chart: Victim Age Trend

- Illustrated general age trends across crime indices
- Helped observe gradual increases or decreases in average victim age

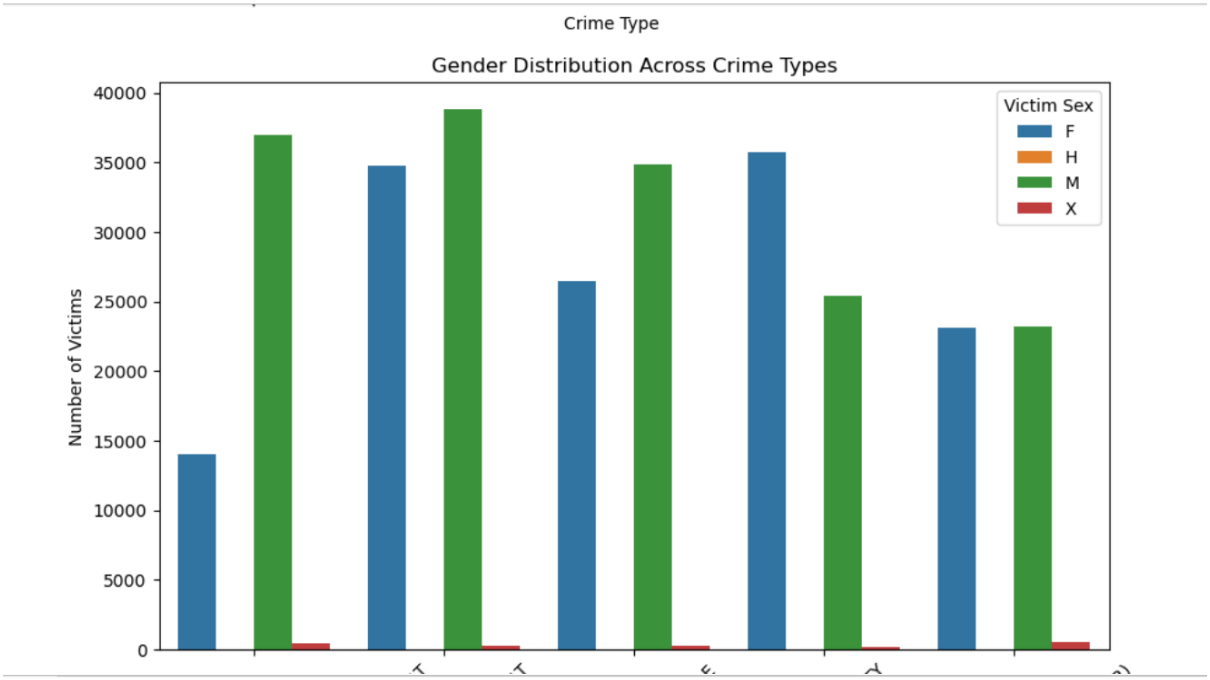
---

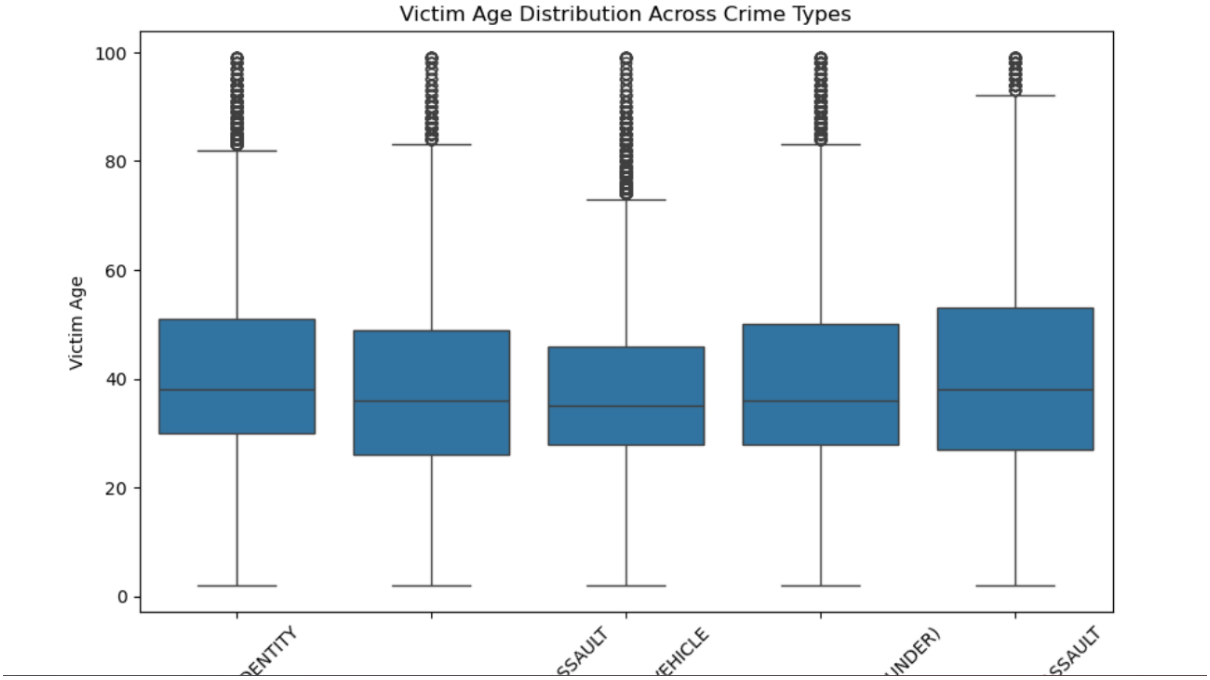
Interpretation

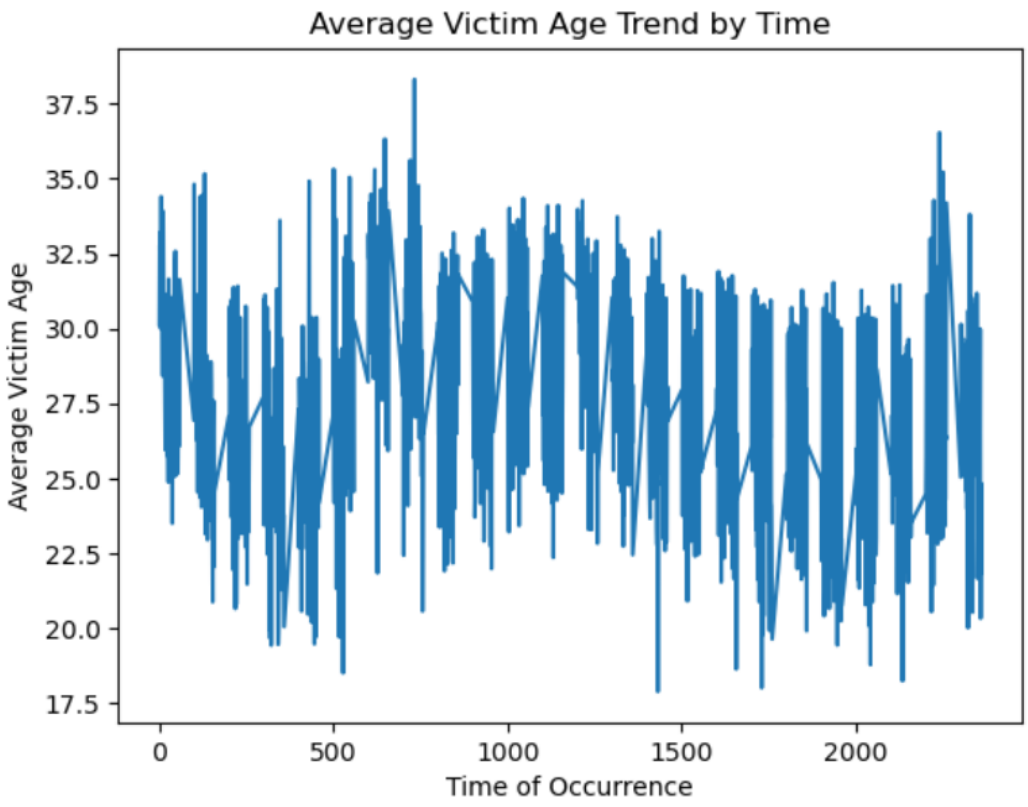
Victim demographics vary significantly across crime types. Age and gender play an important role in understanding crime impact, and visual analysis provides insights that numerical models alone cannot capture.

Conclusion for Objective 5

Victim profile analysis highlighted vulnerable groups and provided valuable insights into crime demographics. These findings can guide targeted awareness programs and preventive strategies.


Crime Type
Gender Distribution Across Crime Types
Victim Sex
F
H
M
X
Number of Victims
40000
35000
30000
25000
20000
15000
10000
5000
0


Victim Age Distribution Across Crime Types
Victim Age
100
80
60
40
20
0



Line Chart: Crime Intensity by Hour

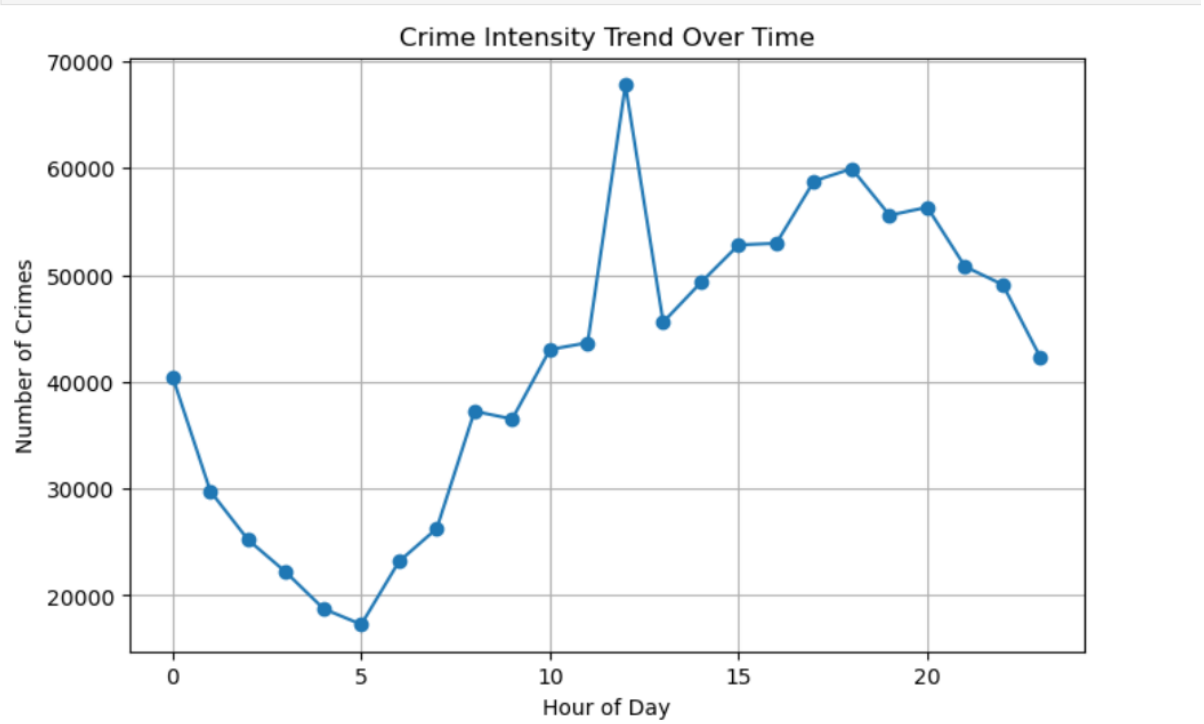
A line chart was used to analyze crime intensity across different hours of the day. The data was grouped based on the hour of occurrence, and the total number of crimes for each hour was calculated.

This visualization helps in identifying peak crime hours and understanding temporal crime patterns. The upward and downward trends in the line chart indicate variations in crime frequency throughout the day.

Key insights obtained from this visualization include:

- Higher crime intensity during late evening and night hours.
- Lower crime occurrence during early morning hours.
- A gradual rise in crime frequency as the day progresses, followed by a decline after peak hours.

The line chart provides a clear and continuous representation of crime trends over time, making it easier to observe fluctuations in crime activity across different hours.



Objective 6

Objective Description

The objective of this analysis is to identify crime hotspot intensity patterns by combining spatial (latitude and longitude) and temporal (hour of occurrence) information. By applying clustering techniques, regions with similar crime intensity characteristics are grouped together, enabling better understanding of high-risk locations at specific times.

Dataset Features Used

- **LAT** – Latitude of crime location
- **LON** – Longitude of crime location
- **TIME OCC** – Time of crime occurrence (converted to hour)

Preprocessing Steps

1. Selected only relevant columns (latitude, longitude, and time of occurrence).
2. Removed missing values to ensure clean data for clustering.
3. Converted the time of occurrence into an hourly format using integer division.
4. Applied **feature scaling** using StandardScaler to normalize latitude, longitude, and hour values.
5. Prepared scaled features for clustering to avoid bias due to differing value ranges.

Machine Learning Technique Used

- **Unsupervised Learning**
- **K-Means Clustering**

K-Means was selected because it effectively groups spatial-temporal data into clusters based on similarity, helping to identify areas with similar crime intensity patterns.

Visualization Used

Scatter Plot with Color-Based Clustering

A scatter plot was used to visualize crime locations based on latitude and longitude, with different colors representing distinct intensity clusters identified by the K-Means algorithm.

This visualization helps:

- Identify high-intensity crime zones
- Understand spatial grouping of crimes
- Observe how crime intensity varies across different locations

---

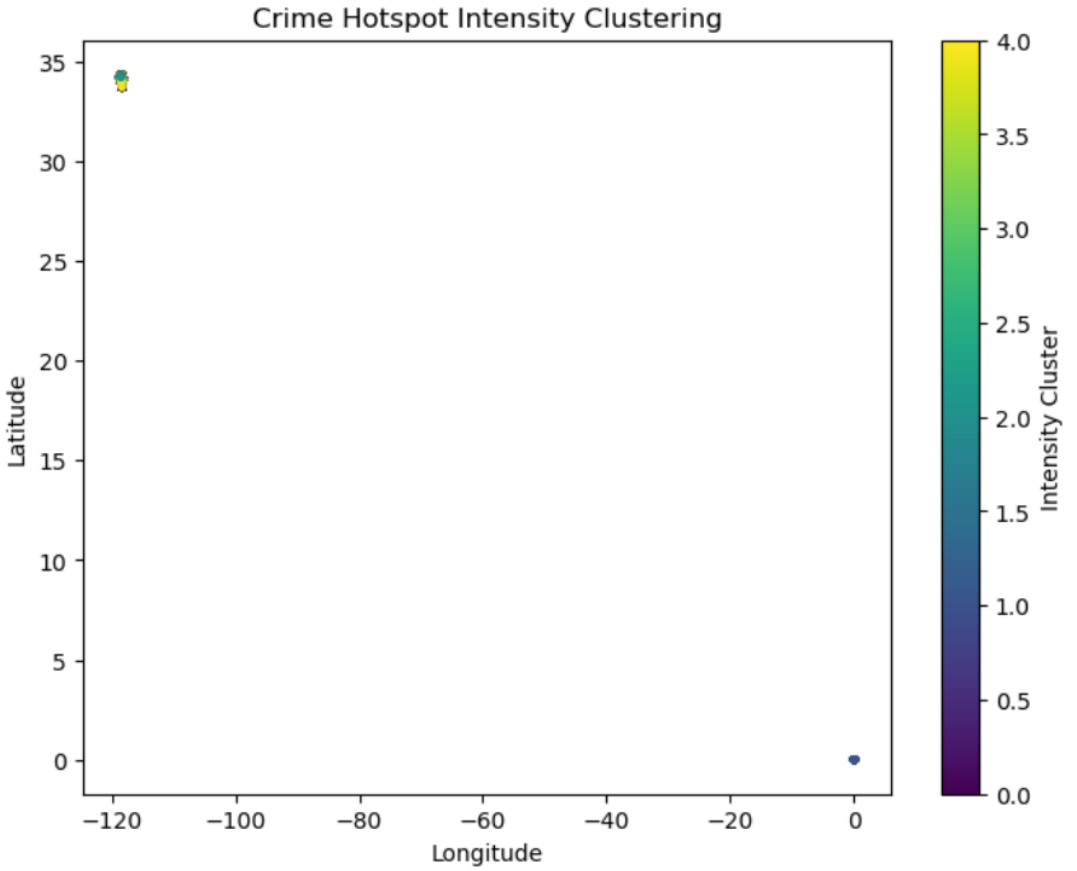
Analysis Results

The clustering results reveal distinct crime hotspot regions across the city. Certain clusters indicate areas with consistently higher crime occurrences during specific hours, suggesting persistent high-risk zones. Other clusters represent lower-intensity regions with fewer crime incidents.

The inclusion of the time component improves the accuracy of hotspot identification by capturing both spatial and temporal crime patterns.

Conclusion

This analysis provides valuable insights into crime distribution by combining location and time, enabling authorities to focus preventive measures on high-risk areas during peak crime hours.



OVERALL ANALYSIS SUMMARY

Objective	Technique	Outcome
Victim Age Prediction	Regression	Moderate performance
Crime Type Prediction	Classification	Strong results
Crime Risk Score	Regression	High reliability
Crime Hotspots	Clustering	Clear spatial patterns
Victim Profile Analysis	EDA + ML	Insightful demographic trends

Conclusion

This project presented a comprehensive analysis of crime data from 2020 to the present with the objective of extracting meaningful insights, identifying crime patterns, and applying machine learning techniques to support data-driven crime analysis. The study successfully combined exploratory data analysis, visualization techniques, supervised learning, and unsupervised learning approaches to understand crime behavior from multiple dimensions such as victim characteristics, time of occurrence, location, and crime attributes.

The dataset underwent careful preprocessing, including handling missing values, selecting relevant features, encoding categorical variables, and scaling numerical attributes where required. These steps ensured data quality and enhanced the performance and reliability of the applied machine learning models. The preprocessing phase played a crucial role in transforming raw crime records into structured inputs suitable for analysis and modeling.

One of the key objectives of the project was **victim age prediction using regression techniques**. By utilizing features such as time of occurrence and geographical coordinates, the regression model provided an approximate

estimation of victim age. Although predicting age from limited features is inherently challenging, the model demonstrated that certain temporal and spatial patterns have a measurable relationship with victim demographics. This objective highlighted the complexity of human-centric predictions and emphasized the importance of richer feature sets for improved accuracy.

Another major objective focused on **crime type classification**, where supervised learning techniques such as Random Forest were employed to predict the category of crime based on victim details, weapon information, and premises description. The classification results showed that ensemble-based models are effective in capturing non-linear relationships within crime data. The confusion matrix and classification report provided valuable insights into prediction accuracy, misclassification patterns, and class imbalance, which are common challenges in real-world crime datasets.

The project also introduced **crime risk score prediction**, which aimed to generate a continuous risk index by combining contextual factors such as weapon usage, night-time occurrence, and public location involvement. This regression-based risk modeling approach transformed qualitative crime indicators into a quantitative score, enabling a clearer understanding of crime severity and likelihood. The results demonstrated that engineered features can significantly enhance predictive power and provide interpretable insights for risk assessment.

In addition, **crime hotspot detection and intensity clustering** were performed using unsupervised learning techniques. By integrating latitude, longitude, and time-based features, K-Means clustering successfully identified regions with similar crime intensity patterns. This spatial-temporal analysis revealed persistent high-risk zones and time-dependent crime concentrations. The visualization of hotspots using clustered scatter plots helped in identifying areas requiring focused monitoring and preventive strategies.

A detailed **Victim Profile Analysis** further enriched the study by examining how crime impacts different demographic groups. Box plots illustrated the variation in victim age across different crime types, while bar charts highlighted gender distribution patterns among crimes. These visual insights revealed demographic disparities and helped in understanding which groups are more vulnerable to specific types of crimes. Line charts and trend analyses complemented this by showing gradual changes in crime occurrence and victim characteristics over time.

Overall, this project successfully demonstrated how machine learning and data visualization techniques can be applied to large-scale crime datasets to uncover hidden patterns, support predictive modeling, and enhance situational awareness. The integration of multiple analytical perspectives ensured a holistic understanding of crime dynamics rather than focusing on a single dimension.

In conclusion, the project achieved all its stated objectives and proved that data-driven approaches can play a significant role in crime analysis and decision support. The insights obtained from this study can assist law enforcement agencies, policy makers, and researchers in improving crime prevention strategies, resource allocation, and public safety planning. While the current models provide meaningful results, they also highlight the potential for further improvements through advanced algorithms, additional contextual features, and real-time data integration.

Future Scope

Although this project successfully analyzed crime patterns and applied machine learning techniques for prediction and visualization, there remains significant scope for enhancement and expansion. Future work can further improve the accuracy, applicability, and real-world impact of the system.

One major extension of this project could involve the use of **advanced machine learning and deep learning models**. Techniques such as Gradient Boosting, XGBoost, LightGBM, and Neural Networks can be applied to improve prediction performance for crime classification and risk scoring. These models are capable of handling complex nonlinear relationships and may produce more accurate and robust results when trained with optimized hyperparameters.

Another important future enhancement is the integration of **temporal forecasting models**. Time-series techniques such as ARIMA, SARIMA, Prophet, or LSTM networks can be used to predict future crime trends based on historical data. This would enable proactive crime prevention by forecasting crime intensity for specific time periods, such as hourly, daily, or seasonal trends.

The project can also be extended to include **real-time crime analysis** by integrating live data sources such as police reports, emergency call logs, or open government crime APIs. Real-time dashboards can be developed using tools like Power BI, Tableau, or Streamlit to continuously monitor crime patterns and generate alerts for high-risk areas.

Another promising direction is the incorporation of **geospatial intelligence**. Advanced GIS techniques, heatmaps, and spatial autocorrelation methods such as Moran's I and Getis-Ord Gi* can provide deeper insights into crime clustering and spatial dependencies. Integration with mapping tools such as Folium or Google Maps APIs would allow interactive visualization of crime hotspots.

The system can further benefit from **socio-economic and environmental data integration**. Factors such as population density, income levels, unemployment rates, weather conditions, and public infrastructure can be added to the dataset to improve model interpretability and predictive power. This would enable a more holistic understanding of the underlying causes of crime.

From a victim-centric perspective, future work can focus on **risk profiling and vulnerability analysis**. More detailed demographic analysis could help identify high-risk groups and support targeted awareness and prevention campaigns. Ethical considerations and privacy-preserving techniques should be incorporated to ensure responsible data usage.

The project also has scope for development as a **decision support system for law enforcement agencies**. Predictive outputs such as crime risk scores and hotspot intensities can be used to optimize patrol allocation, resource planning, and response strategies. Integrating explainable AI techniques would help decision-makers understand why certain predictions are made.

Lastly, future enhancements can include **model evaluation improvements and bias mitigation**. Addressing class imbalance, validating models across multiple geographic regions, and ensuring fairness in predictions will increase the reliability and societal acceptance of the system.