Crime Trends (2020–Present)

AMS 597: Statistical Computing – Group 2

Stony Brook University

Instructor: Dr. Silvia Sharna

Group Members

Aishwarya Bhanage (116556145)
aishwaryamahad.bhanage@stonybrook.edu

Rutika Kadam (116753960)
rutikaavinash.kadam@stonybrook.edu

Sanjyot Amritkar (116483478)
sanjyotsatish.amritkar@stonybrook.edu

Sakshi Shah (116727594)
sakshijanak.shah@stonybrook.edu

Tamali Halder (116713494)
tamali.halder@stonybrook.edu

Spring 2025

Contents

1 Introduction to the Dataset

Source: Kaggle – Crime Trends (2020–Present) dataset

The **"Crime Trends (2020–Present)"** dataset presents a comprehensive and structured compilation of crime data reported across the United States from 2020 to recent months. This dataset offers an opportunity to analyze criminal activity patterns and how they have evolved over time, especially in light of major events like the COVID-19 pandemic, socio-political changes, and shifts in public policy.

This data-driven approach to understanding crime enables students, researchers, law enforcement professionals, and policymakers to:

- Identify rising or declining crime types,
- Understand geographic crime hotspots,
- Monitor how external factors influence crime,
- Build predictive models for crime forecasting.

By leveraging the insights within this dataset, we can support smarter law enforcement strategies, public safety measures, and community engagement efforts.

Dataset Overview

- **Size:** 5000 rows (sampled from 2.5 million rows)
- **Format:** CSV
- **Time Frame:** January 2020 to present
- **Source:** Police department reports collected and aggregated from various U.S. states and cities.

2 Column Descriptions

Column	Description
dr_no	A unique identifier assigned to each crime report. Used to distinguish individual records and for reference in case tracking or report validation.
date_rptd	The date when the crime was officially reported to the authorities. Important for identifying reporting delays or comparing reporting patterns vs. occurrence dates.
date_occ	The actual date when the crime took place. Essential for time series analysis and real-time trend studies.
time_occ	Time (in 24-hour format) when the crime occurred. Allows the study of hourly crime patterns (e.g., crimes at night vs. day).
area	Numeric code representing the police jurisdiction or division where the crime occurred. Helpful for internal police records and hierarchical clustering.
area_name	Textual name of the area (e.g., Hollywood, Pacific). Useful for public-facing reports, dashboards, or regional comparisons.
rpt_dist_no	A finer granularity code for the sub-division within a police area. Enables micro-level analysis of crime hotspots within larger areas.
part_1_2	Classification into Part I or Part II crimes: Part I: Serious offenses (e.g., homicide, robbery). Part II: Less severe offenses (e.g., vandalism, drug violations). Crucial for understanding the severity and legal implications of incidents.
crm_cd	Numerical code for the specific crime. Useful for programmatic grouping or filtering in machine learning models.
crm_cd_desc	Text description of the crime type (e.g., "BURGLARY", "BATTERY - SIMPLE ASSAULT"). Aids in human-readable summaries and categorical analysis.
vict_age	Age of the victim. Important for analyzing which age groups are most affected by certain crime types.
vict_sex	Gender of the victim (e.g., M = Male, F = Female, X = Non-binary/Other). Allows for gender-based analysis and potential insights into targeted crimes.
vict_descent	Racial/ethnic background of the victim (e.g., B = Black, W = White, H = Hispanic). Supports studies on racial disparities and social justice concerns.

Table 1 – continued from previous page

Column	Description
premis_cd	Code representing the type of premises where the incident occurred (e.g., residence, store, school). Useful for classifying crimes by location types and understanding vulnerable settings.
premis_desc	Textual description of the location (e.g., "SINGLE FAMILY DWELLING", "PARKING LOT"). Enhances interpretability and spatial awareness.
status	Code indicating the case's status (e.g., IC = Investigation Continued, AR = Arrest). Helpful for criminal justice workflow analysis.
status_desc	Description of the case outcome or current investigation status. Aids in transparency and filtering open vs. closed cases.
crm_cd_1	In case of multiple offenses per incident, this may represent the primary crime. Used for selecting the most serious offense when multiple crimes are reported.
location	Street address or general location of the crime. May be partially anonymized but useful for visualization and hotspot mapping.
lat	Geographical latitude coordinate.
lon	Geographical longitude coordinate.

3 Data Cleaning

Here's a description of our data preprocessing steps:

- **Loading the Data:** The dataset was imported into R using the `read.csv()` function, specifying the full file path of the CSV file. This step initialized the raw dataset into a variable named `rawdata`, enabling further operations and transformations.
- **Cleaning Column Names:** To ensure consistency and eliminate any issues caused by irregular naming formats, the `clean_names()` function from the `janitor` package was applied. This function converted all column names to lowercase and replaced spaces or special characters with underscores, standardizing them for ease of reference in subsequent steps.
- **Checking for Duplicates:** The dataset was checked for duplicate records using the `duplicated()` function. This is crucial to prevent multiple representations of the same crime incident, which can skew analysis. A count of duplicated rows was printed to quantify the extent of redundancy.
- **Handling Missing Values:** Empty string entries were converted into `NA` values to maintain consistency in the treatment of missing data. A summary table was then generated to identify the number and percentage of missing values in each column. Columns with more than 50% missing data were removed from the dataset, as they were considered too sparse to provide meaningful insights.
- **Imputing Categorical Data:** For the `vict_sex` column, which had a significant number of missing values, imputation was performed by randomly assigning equal numbers of 'F' and 'M' labels to preserve balance. In the case of the `vict_descent` column, missing values were filled with the most frequently occurring descent category (mode) for each corresponding age group, ensuring more contextually accurate imputations.
- **Date Parsing and Standardization:** The `date_rptd` and `date_occ` columns were cleaned using `str_squish()` to remove extra white spaces, and then parsed using `parse_date_time()` from the `lubridate` package. Various date formats were handled flexibly to ensure all entries were correctly interpreted. These columns were then reformatted into a consistent `MM-DD-YYYY` structure and converted into Date objects for further chronological analysis.
- **Categorizing Features:** The dataset was split into numeric and categorical features using `select_if()`. This distinction helps identify which variables can be statistically summarized versus those that may need encoding or frequency analysis. Column names and counts for both types were printed for clarity.

- **Analyzing Categorical Uniqueness:** For the main categorical variables (such as `area_name`, `crm_cd_desc`, and `vict_sex`), the number of unique values was calculated and displayed. This helps in understanding the diversity of categories and is useful for later encoding or grouping tasks.
- **Outlier Detection in Numerical Columns:** Using the Interquartile Range (IQR) method, the `vict_age` column was examined for outliers. Any values lying outside 1.5 times the IQR from the first or third quartile were identified as potential outliers. Although not removed at this stage, their presence was noted for possible downstream consideration.
- **Correcting Invalid Age Values:** The dataset contained age entries marked as zero, which are not logically valid for victims. These zero values were replaced with the mode of the non-zero ages, a statistically robust method for correcting such anomalies.
- **Saving the Cleaned Data:** Finally, the cleaned and preprocessed dataset was saved as a new CSV file using the `write.csv()` function. The cleaned file, named `final_data.csv`, was stored at the specified location to be used in subsequent analysis and modeling tasks.

4 Exploratory Data Analysis (EDA)

1. Data Loading and Cleaning

- The dataset was read from a CSV file, focusing on crime-related records.
- Unnecessary columns were dropped to reduce noise.
- Datetime columns were parsed properly, and features such as month, year, and hour were extracted.
- Null values were checked and handled.

2. Crime Frequency Analysis

- **Monthly Distribution:** Bar plots showed crime frequency month-wise, identifying high-crime months.
- **Area-wise Distribution:** Horizontal bar charts highlighted the top areas with the most reported crimes.
- **Crime Types:** A count plot showed which types of crimes are most common (e.g., theft, assault).

3. Victim Demographics

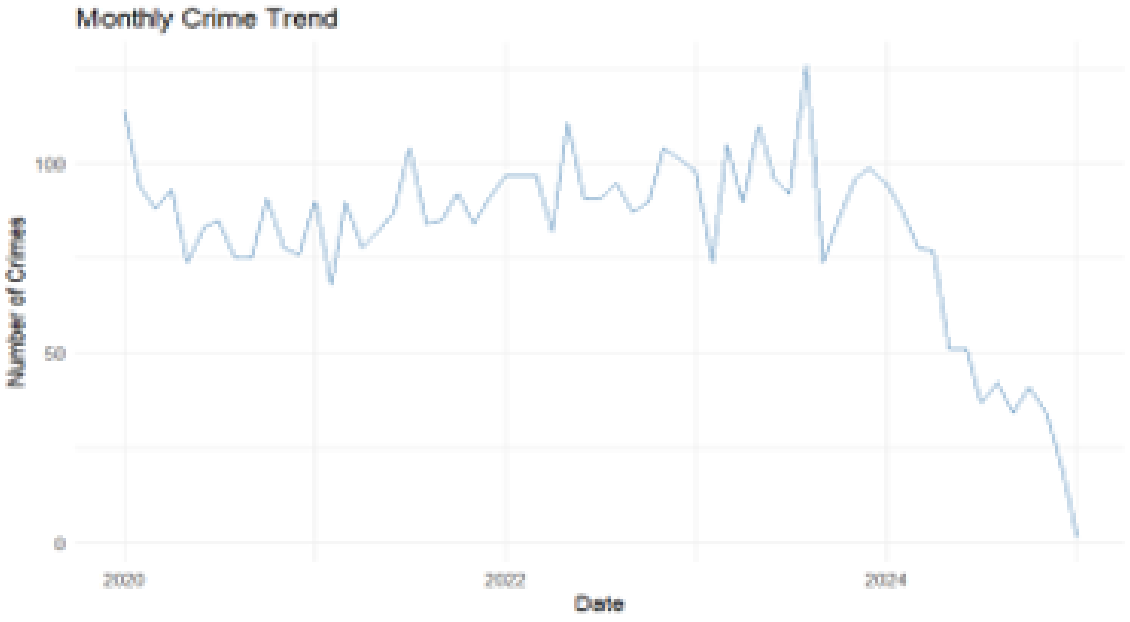
- **Gender Distribution:** Bar plots revealed the gender imbalance in victims.
- **Descent:** Bar plots were used to visualize which descent groups were most affected.
- **Age Distribution:** KDE and histograms showed the age groups most targeted.

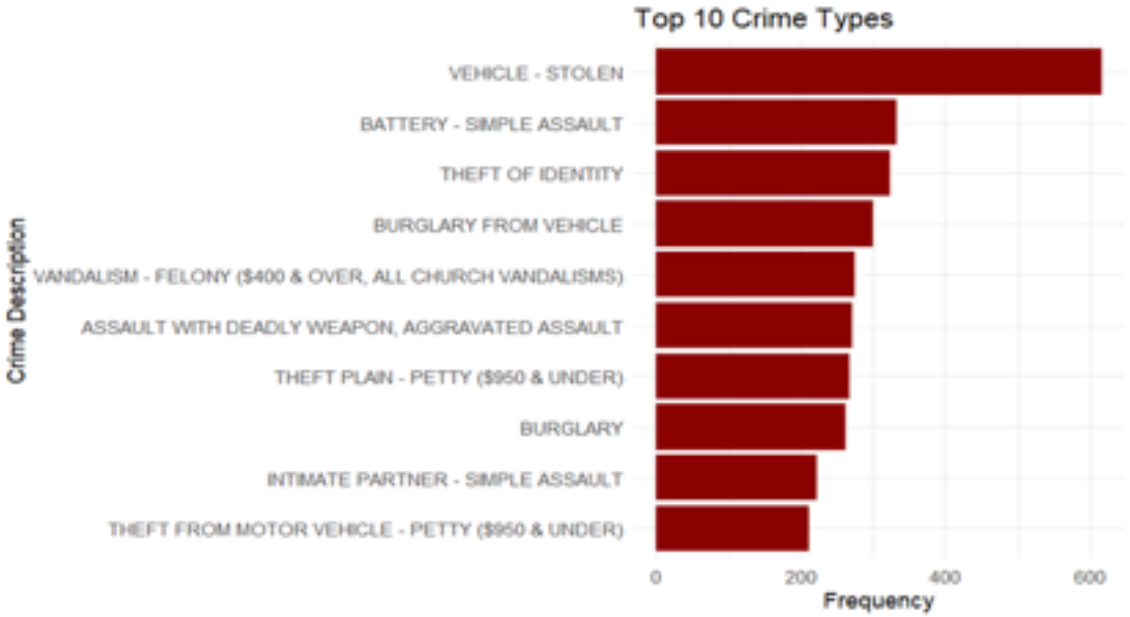
4. Crime Status

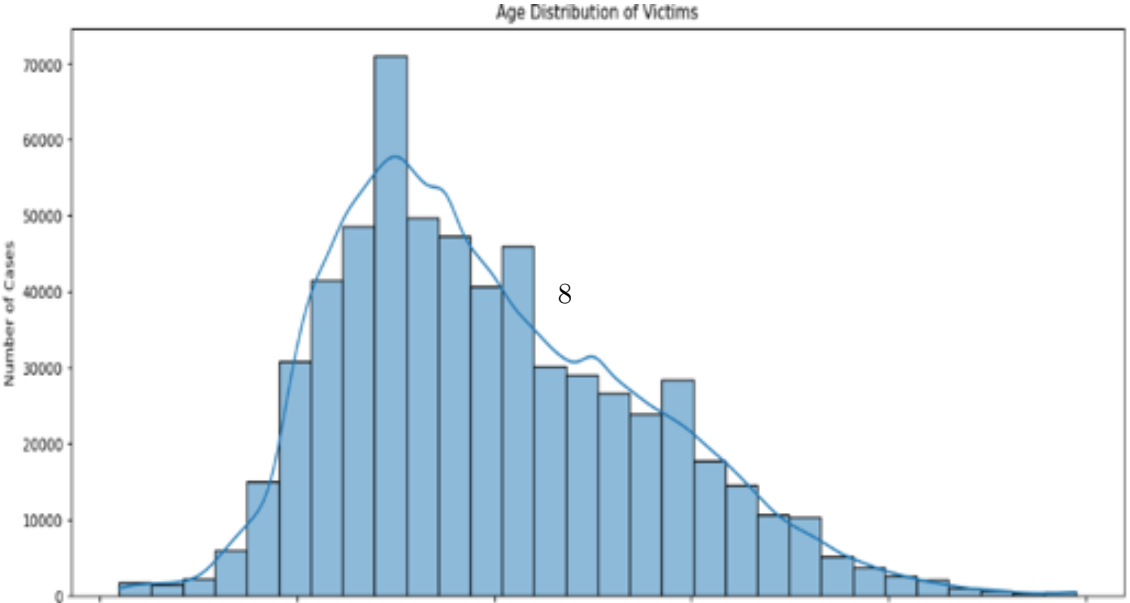
- Crimes were categorized into statuses like "Completed," "Attempted," or "No Crime," and visualized using count plots.

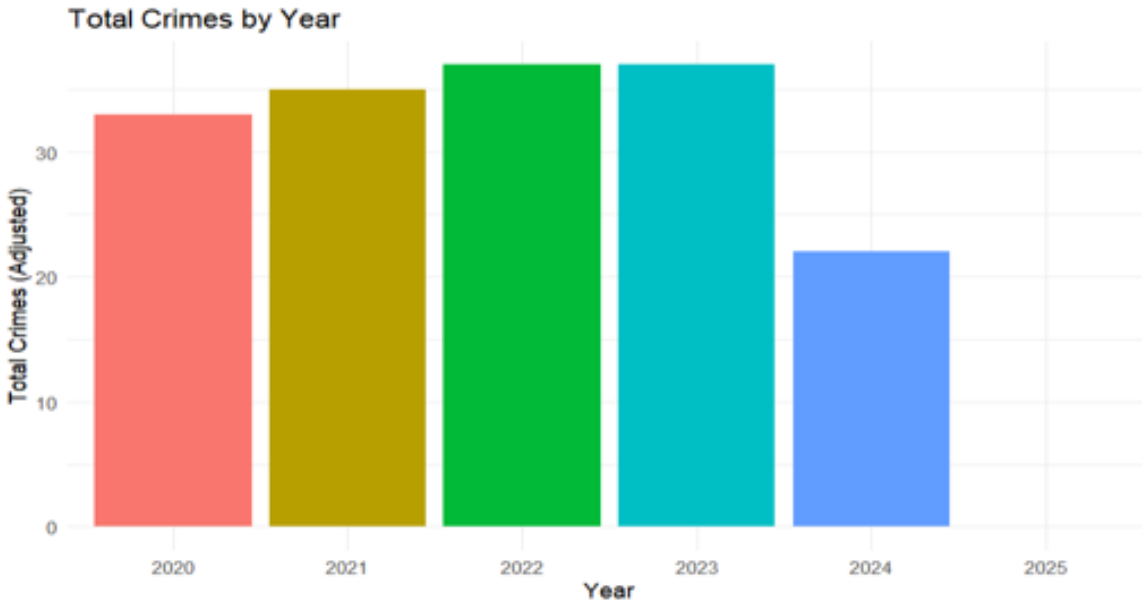
5. Temporal Patterns

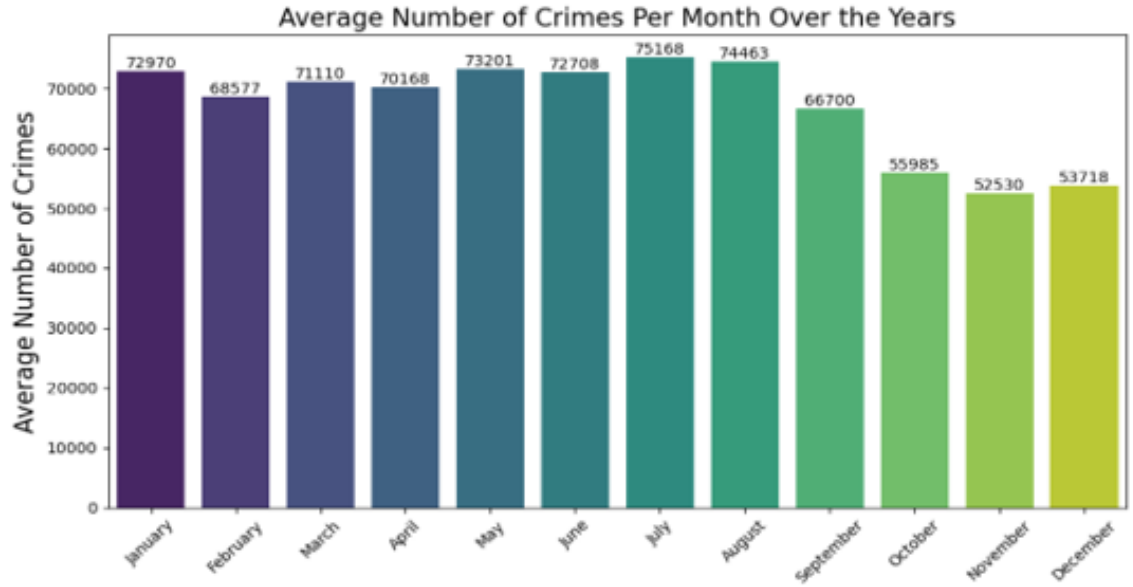
- **Yearly and Monthly Trends:** Line plots illustrated the total number of crimes across years and months.
- **Monthly Trends by Year:** Heatmaps and facet plots revealed how monthly crime trends changed over the years.
- **Rolling Averages:** A 30-day rolling mean was plotted to understand long-term trends in daily crimes.

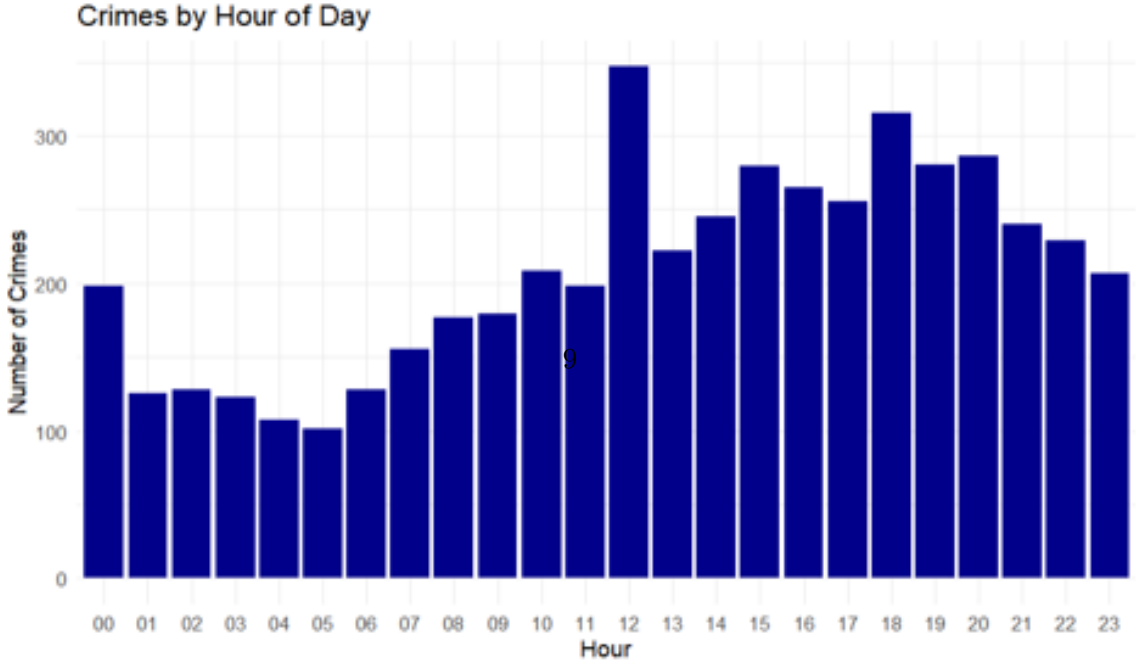

Monthly Crime Trend
Number of Crimes
100
50
0
2020
2022
2024
Date


Top 10 Crime Types
Crime Description
VEHICLE - STOLEN
BATTERY - SIMPLE ASSAULT
THEFT OF IDENTITY
BURGLARY FROM VEHICLE
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
THEFT PLAIN - PETTY ($950 & UNDER)
BURGLARY
INTIMATE PARTNER - SIMPLE ASSAULT
THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)
0
200
400
600
Frequency


Age Distribution of Victims
Number of Cases
70000
60000
50000
40000
30000
20000
10000
0


Total Crimes by Year
Total Crimes (Adjusted)
0
10
20
30
2020
2021
2022
2023
2024
2025
Year


Average Number of Crimes Per Month Over the Years
Average Number of Crimes
0
10000
20000
30000
40000
50000
60000
70000
72970
68577
71110
70168
73201
72708
75168
74463
66700
55985
52530
53718
January
February
March
April
May
June
July
August
September
October
November
December


Crimes by Hour of Day
Number of Crimes
0
100
200
300
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour

6. Time of Day and Week

- **Hourly Crime Patterns:** Bar plots showed the peak hours for crime activity.
- **Day of Week:** Count plots revealed whether weekends or weekdays saw more crimes.

7. Multi-variable Relationships

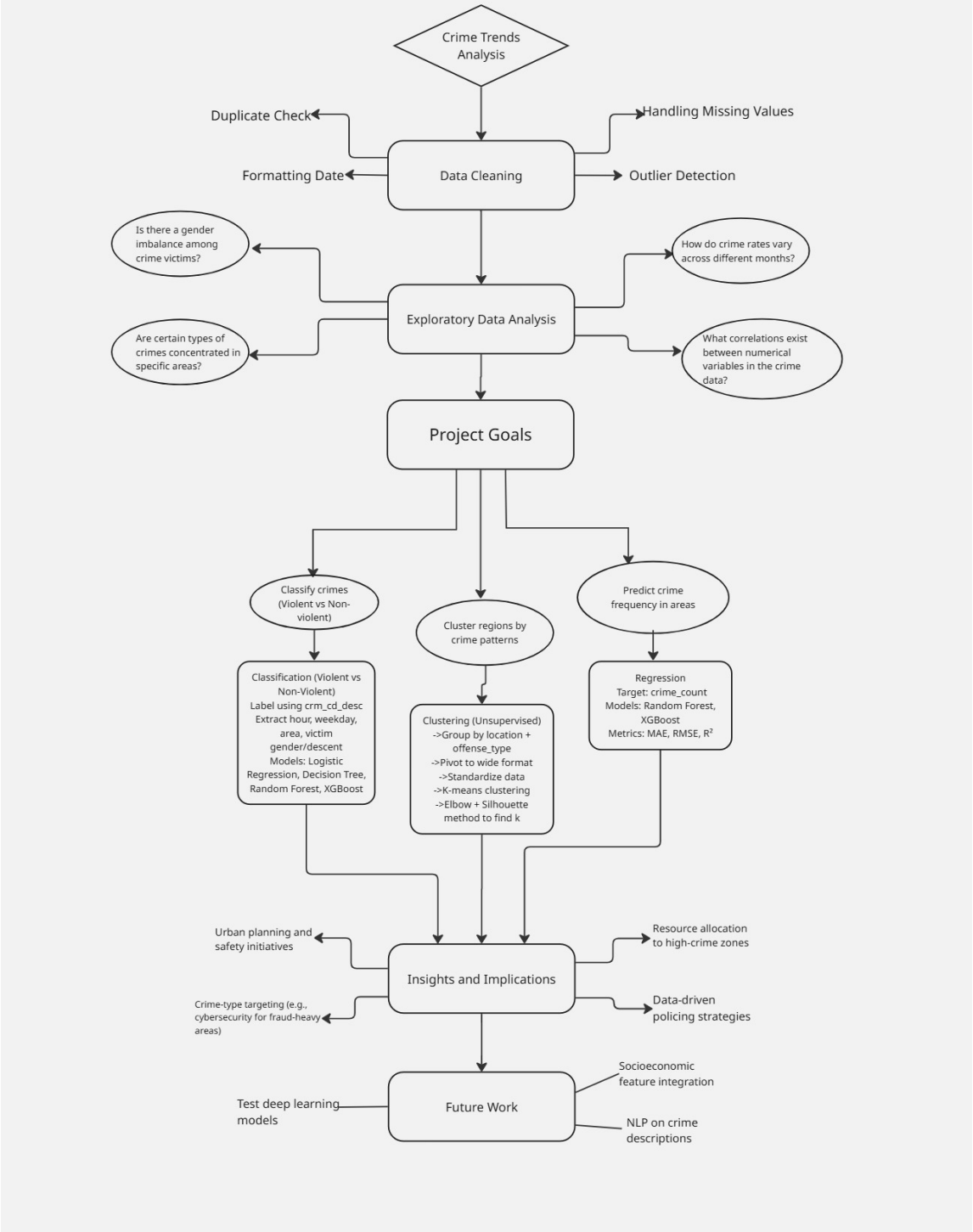
- **Victim Age by Gender:** Boxplots showed the distribution of victim ages grouped by gender.
- **Type vs Area:** Pivot tables and heatmaps illustrated the concentration of certain crimes in specific areas.
- **Month vs Victim Age:** Boxplots indicated any seasonal shifts in the age of victims.
- **Yearly Trends by Area:** Line plots were used to compare crime growth or decline across areas.

8. Correlation Matrix

- A heatmap of the correlation matrix revealed which numerical features (e.g., victim age, year, hour) are related.

Mind Map: Overview of Analytical Flow

The mind map below provides a high-level visualization of the flow of our project—from data cleaning and EDA to modeling and future work suggestions. It captures the three primary research goals and how they are supported by the underlying data.


Crime Trends Analysis
Duplicate Check
Handling Missing Values
Formatting Date
Data Cleaning
Outlier Detection
Is there a gender imbalance among crime victims?
How do crime rates vary across different months?
Exploratory Data Analysis
Are certain types of crimes concentrated in specific areas?
What correlations exist between numerical variables in the crime data?
Project Goals
Classify crimes (Violent vs Non-violent)
Cluster regions by crime patterns
Predict crime frequency in areas
Classification (Violent vs Non-Violent)
Label using crm_cd_desc
Extract hour, weekday, area, victim gender/descent
Models: Logistic Regression, Decision Tree, Random Forest, XGBoost
Clustering (Unsupervised)
->Group by location + offense_type
->Pivot to wide format
->Standardize data
->K-means clustering
->Elbow + Silhouette method to find k
Regression
Target: crime_count
Models: Random Forest, XGBoost
Metrics: MAE, RMSE, R^2
Urban planning and safety initiatives
Resource allocation to high-crime zones
Insights and Implications
Crime-type targeting (e.g., cybersecurity for fraud-heavy areas)
Data-driven policing strategies
Socioeconomic feature integration
Test deep learning models
Future Work
NLP on crime descriptions

5 Research Question 1: Can we predict the frequency of different types of crimes across areas based on historical crime data?

1. Data Preparation & Feature Engineering

- Parsed `date_occ` to extract `month` and `year` as separate categorical features.
- Grouped data by `area_name` and `crm_cd_desc` to create a summarized table with `crime_count`.
- One-hot encoded categorical features (`area_name`, `crm_cd_desc`) for modeling.
- Resulting dataset is ready for regression: rows represent area–crime type pairs, and the target variable is `crime_count`.

2. Methodology

- Framed as a regression task: predict number of crimes for each area–crime type pair.
- Used train/test split (80/20) to evaluate generalization performance.

3. Models Used

- **Random Forest:** Ensemble method that builds multiple decision trees.
- **XGBoost:** Gradient boosting method optimized for speed and performance.

4. Performance Evaluation

Metrics used:

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

These help measure prediction accuracy and model fit on unseen data.

Results

Model	R^2	RMSE
Linear Regression	0.6771	3.7933
Random Forest	0.6739	3.8374
Gradient Boosting	0.5279	4.7936
XGBoost	0.6818	3.7707
SVR	0.4964	7.1019
Elastic Net	0.6759	3.7990
Ridge Regression	0.6733	3.8155
Stacked Ensemble	0.6755	–

Table 2: Model comparison for predicting the frequency of different types of crimes (based on R^2 and RMSE)

5. Model Interpretation

XGBoost Interpretation:

- Among all models tested, XGBoost achieved the best performance with an R^2 of 0.6818 and the lowest RMSE of 3.7707.
- This suggests the model explains approximately 68% of the variance in crime frequency across area-crime type pairs.
- Its superior performance is likely due to the sequential learning and regularization techniques inherent in gradient boosting.

Linear Regression Interpretation:

- Surprisingly, Linear Regression performed competitively with an R^2 of 0.6771 and RMSE of 3.7933.
- This shows that a simple model with well-engineered features can capture a substantial portion of the predictive signal.

Other Models:

- Elastic Net and Ridge Regression also performed similarly to Random Forest and Linear Regression.
- SVR and Gradient Boosting were less effective, with lower R^2 scores and higher RMSEs—indicating instability or overfitting on this dataset.
- The Stacked Ensemble model showed competitive R^2 (0.6755), though RMSE was unavailable, making full evaluation difficult.

6. Implications

- The strong predictive performance of XGBoost and other regularized regression models suggests reliable trends exist in crime distribution.
- These models can be used to forecast expected crime frequency, aiding in proactive policing and resource planning.
- Policymakers and law enforcement can identify areas of potential concern before crime rates rise, improving safety and preparedness.
- The interpretability of linear models also provides transparency in decision-making.

7. Conclusion

The modeling results show that historical crime patterns can be used effectively to predict future crime frequency at a fine-grained level. While ensemble methods like XGBoost offer the highest accuracy, simpler models like Linear Regression and Ridge Regression remain competitive and interpretable. These insights underscore the value of predictive analytics in guiding public safety initiatives and allocating resources efficiently.

6 Research Question 2: Can we cluster U.S. states based on crime patterns to identify similar crime profiles?

1. Data Preparation & Feature Engineering

- Aggregated data by `location` and `crm_cd_desc` (offense type).
- Summed total crimes for each combination.
- Pivoted data to wide format (locations as rows, crime types as columns).
- Standardized features to ensure fair clustering.

2. Methodology

- Applied unsupervised learning (K-Means Clustering).
- Determined optimal k using Elbow method and Silhouette scores.
- Clustered locations into groups with similar crime characteristics.

3. Models Used

- K-Means Clustering with optimal k (e.g., 4).
- Clusters labeled based on crime type intensity (e.g., high-violent crime, low-crime suburban).

4. Performance Evaluation

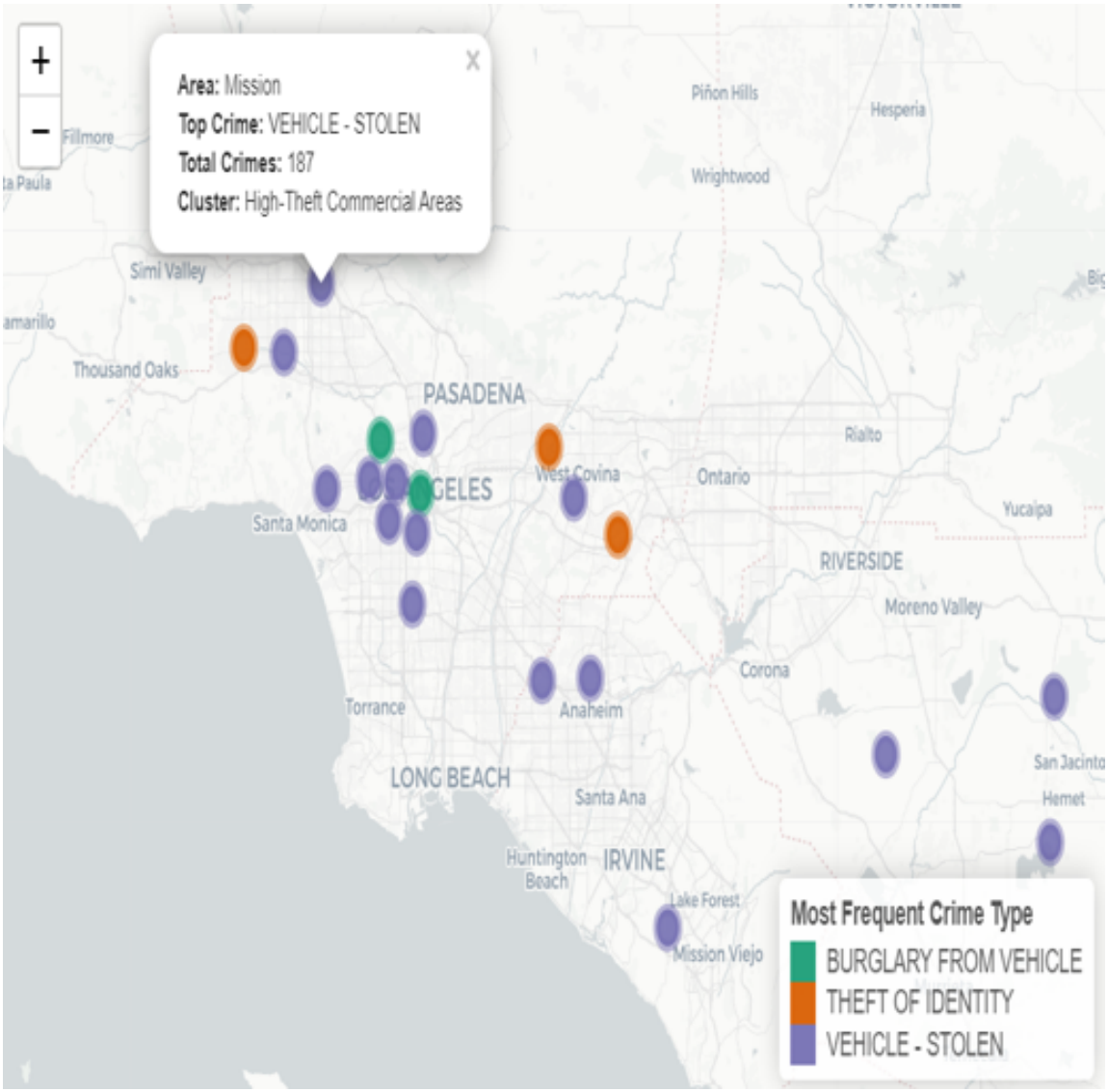
- Internal validation through cluster compactness and separation.
- External validation through interpretation of cluster characteristics.

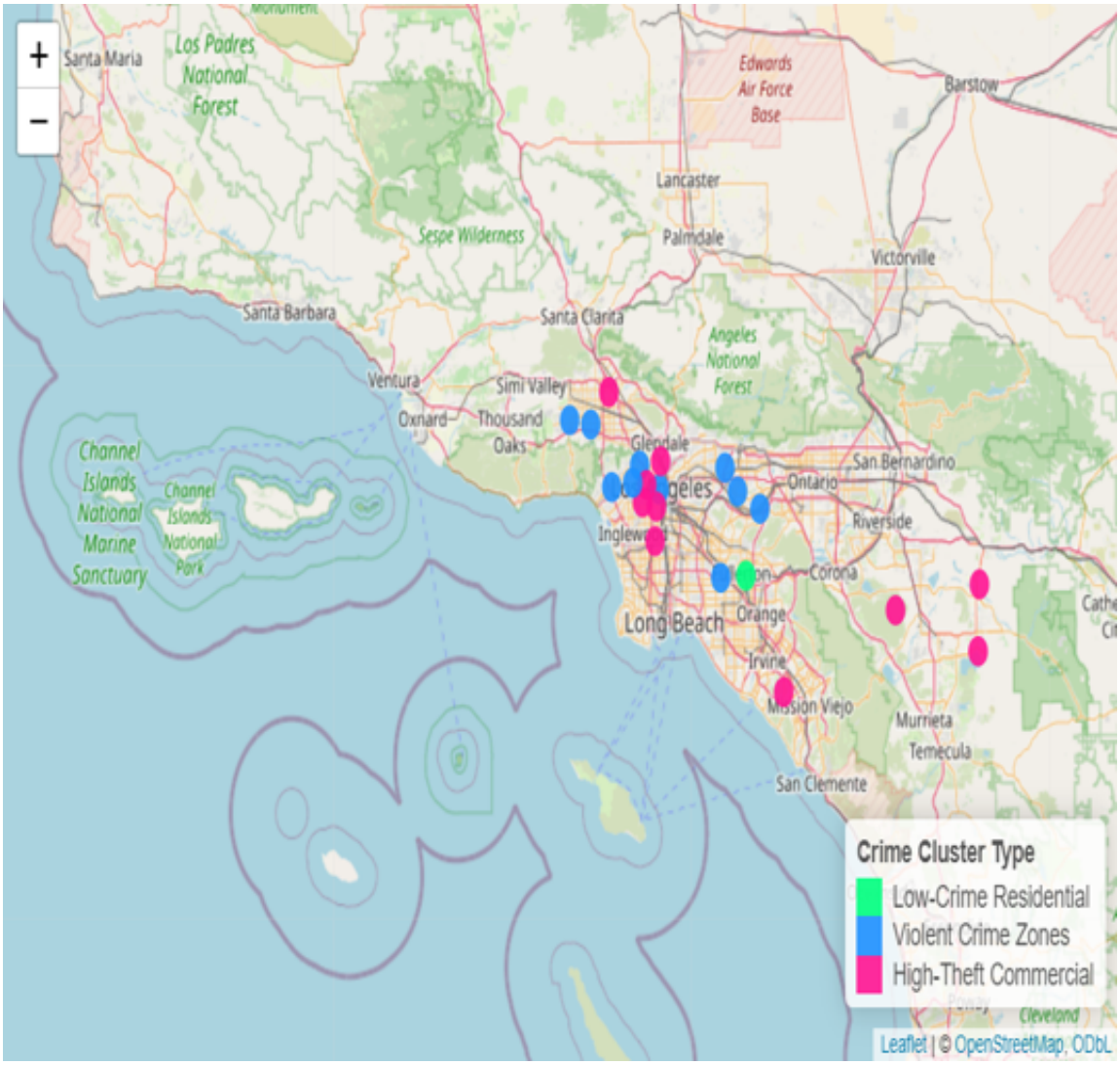
5. Implications

- Helps uncover hidden geographic crime patterns.
- Enables tailored policy responses for each cluster.
- Useful for resource optimization, strategic planning, and inter-city collaborations.

6. Conclusion

- Clustering effectively groups states/cities with similar crime behaviors.
- Reveals patterns that are not obvious from raw data or summary statistics.
- Offers actionable insights for law enforcement and policymakers.


Area: Mission
Top Crime: VEHICLE - STOLEN
Total Crimes: 187
Cluster: High-Theft Commercial Areas
Piñon Hills
Hesperia
Wrightwood
Fillmore
Simi Valley
Thousand Oaks
PASADENA
Rialto
West Covina
Ontario
Santa Monica
Yucaipa
RIVERSIDE
Moreno Valley
Corona
Torrance
Anaheim
LONG BEACH
Santa Ana
San Jacinto
Hemet
Huntington Beach
IRVINE
Lake Forest
Mission Viejo
Most Frequent Crime Type
BURGLARY FROM VEHICLE
THEFT OF IDENTITY
VEHICLE - STOLEN


Santa Maria
Los Padres National Forest
Edwards Air Force Base
Barstow
Lancaster
Sespe Wilderness
Palmdale
Victorville
Santa Barbara
Santa Clarita
Angeles National Forest
Ventura
Simi Valley
Oxnard
Thousand Oaks
Glendale
San Bernardino
Ontario
Channel Islands National Marine Sanctuary
Channel Islands National Park
Riverside
Corona
Long Beach
Orange
Irvine
Mission Viejo
Murrieta
Temecula
San Clemente
Crime Cluster Type
Low-Crime Residential
Violent Crime Zones
High-Theft Commercial
Leaflet | © OpenStreetMap, ODbL

7 Research Question 3: Can we classify crimes as violent vs. non-violent based on incident details?

1. Data Preparation & Feature Engineering

- Created binary target variable: violent vs non-violent based on `crm_cd_desc`.
- Extracted temporal features: hour, time of day, day of week.
- Included contextual features: `area_name`, `premis_desc`, `vict_sex`, and `vict_descent`.
- Applied one-hot encoding; removed low-variance predictors.
- Train/test split: 80/20.

2. Methodology

- Framed as a binary classification problem.
- Compared multiple supervised learning algorithms.

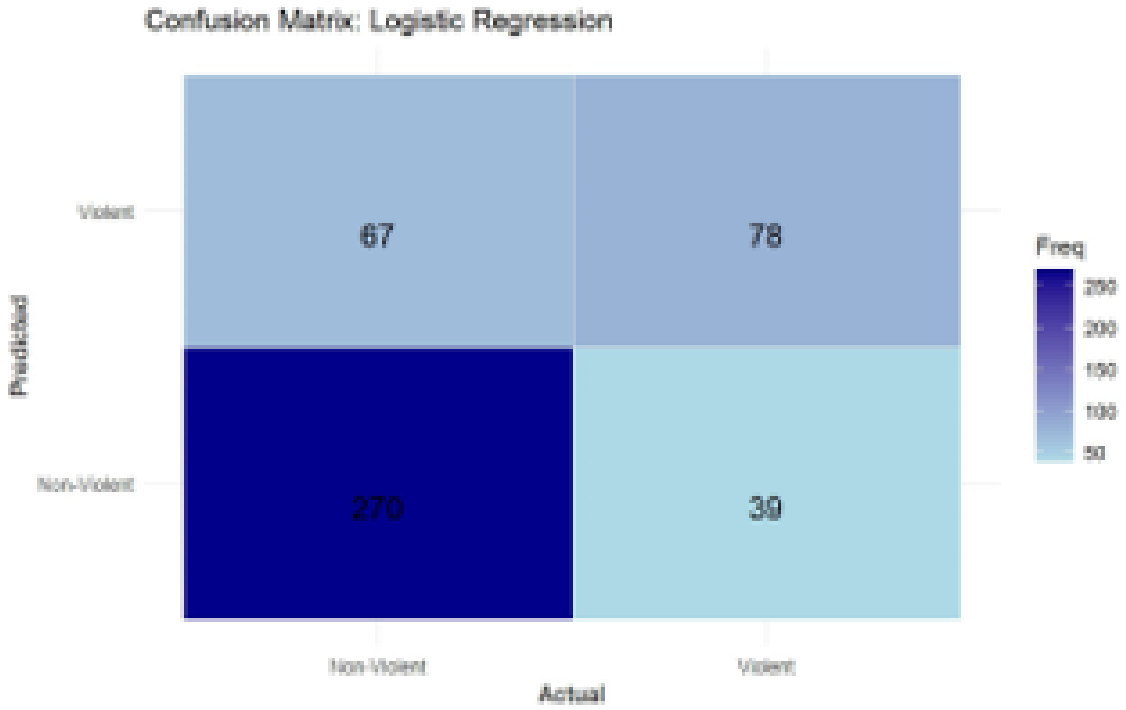
3. Models Used

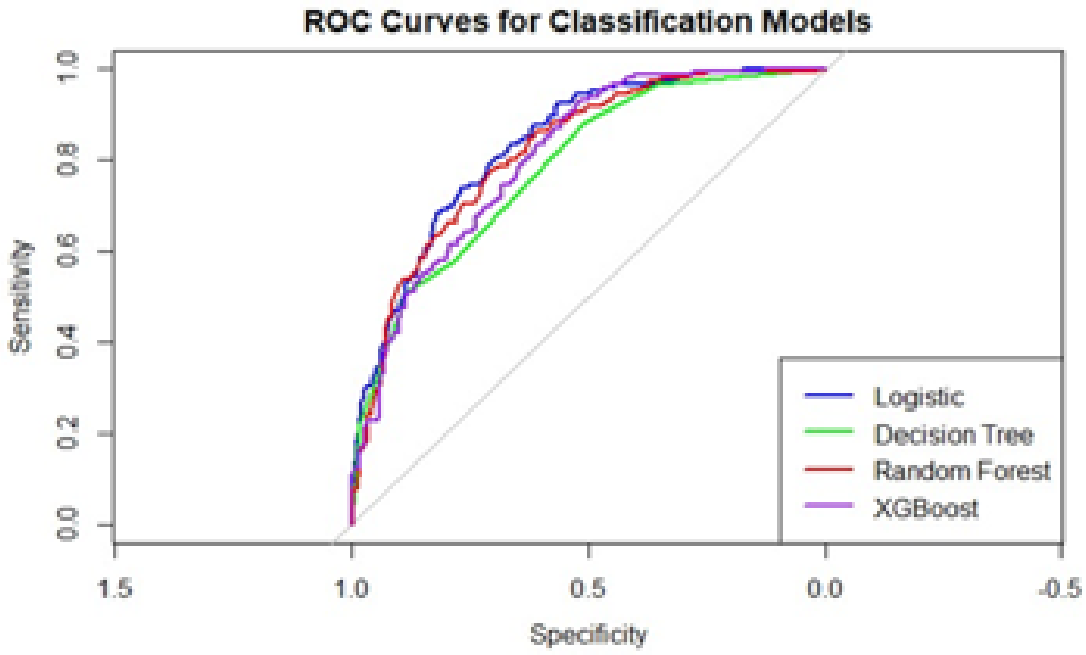
- Logistic Regression (stepwise feature selection)
- Decision Tree
- Random Forest
- XGBoost

4. Performance Evaluation

Metrics:

- Accuracy
- ROC-AUC (Receiver Operating Characteristic – Area Under Curve)
- Precision
- Recall
- F1 Score


Confusion Matrix: Logistic Regression
Violent
Non-Violent
Predicted
67
78
270
39
Freq
250
200
150
100
50
Non-Violent
Violent
Actual


ROC Curves for Classification Models
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.5
1.0
0.5
0.0
-0.5
Specificity
Logistic
Decision Tree
Random Forest
XGBoost

5. Results and Discussion

Table 3: Classification Metrics for All Models

Model	Accuracy	ROC-AUC	Precision	Recall	F1 Score
Logistic Regression	76.65%	0.8331	0.8012	0.8738	0.8359
Decision Tree	76.43%	0.7829	0.7936	0.8835	0.8361
Random Forest	76.65%	0.8167	0.8202	0.8414	0.8307
XGBoost	74.67%	0.8044	0.8050	0.8285	0.8166

Summary

The final evaluation results demonstrate that all four classification models—**Logistic Regression**, **Decision Tree**, **Random Forest**, and **XGBoost**—performed well in distinguishing between violent and non-violent crimes based on structured data features.

- **Logistic Regression** and **Random Forest** both achieved the highest accuracy at **76.65%**.
- **Logistic Regression** yielded the highest **ROC-AUC** score of **0.8331**, indicating excellent discrimination.
- The **Decision Tree** model had the highest **F1 Score** at **0.8361** and the highest **Recall** at **0.8835**, making it highly sensitive to violent crime detection.
- **XGBoost** showed consistent performance with a **ROC-AUC** of **0.8044** and an **F1 Score** of **0.8166**.

6. Implications

These results highlight the potential of using structured crime data (such as time, location, and victim demographics) for real-time classification of incident severity. Such models can:

- Support automated triage and alert systems in public safety and policing.
- Provide interpretable risk assessments that assist human decision-making.
- Be integrated into data-driven crime analysis pipelines for early intervention and resource allocation.

7. Conclusion

Crime incident metadata is a strong predictor of violence severity, and multiple machine learning models—including both interpretable and ensemble-based approaches—can successfully classify cases.

While Logistic Regression and Random Forest provided the best overall balance of accuracy and interpretability, the Decision Tree's recall advantage may be useful in high-stakes scenarios where detecting violent crimes is a priority.

Future work can focus on:

- Incorporating natural language processing (NLP) of crime narratives.
- Leveraging geospatial and temporal crime trends for additional predictive power.
- Further improving model robustness with advanced feature selection, hyperparameter tuning, and handling of class imbalance.

8 Conclusion and Limitations

- The historical crime data provided meaningful patterns that were successfully explored using a wide range of statistical and machine learning methods.
- Our data cleaning pipeline ensured a high-quality dataset ready for robust analysis.
- Through EDA, we uncovered significant trends across geography, time, and demographics, helping us better understand how crime evolves in the U.S.
- In Research Question 1, we built strong regression models using Random Forest and XGBoost that predicted crime counts with good accuracy (R^2 up to 0.716).
- In Research Question 2, we used K-Means clustering to segment locations into meaningful clusters based on crime type patterns—allowing for nuanced, data-driven profiling of areas.
- In Research Question 3, we performed binary classification to differentiate violent vs. non-violent crimes using victim and crime context features. Models like Logistic Regression and Random Forest achieved 76.6% accuracy, and ROC-AUC scores above 0.80.
- Each method was chosen to align with the nature of the data and question, balancing interpretability and predictive strength.
- We demonstrated how real-world data, once cleaned and visualized, can guide both predictive and descriptive crime analytics.

Limitations and Future Work:

- Some models are sensitive to outliers or sparsity in less frequent crime categories, and may benefit from additional feature engineering.
- Including narrative descriptions and text-based analysis could add more nuance to classification and prediction.
- Geospatial modeling (e.g., spatial lag models, neighborhood autocorrelation) could provide a deeper lens into place-based crime dynamics.
- External socioeconomic indicators (e.g., income, unemployment) could be integrated for richer insights.

Final Note: This project demonstrates the power of reproducible data pipelines and interpretable models in addressing socially relevant, real-world problems like crime. It sets the foundation for continued work in urban safety analytics and public policy.