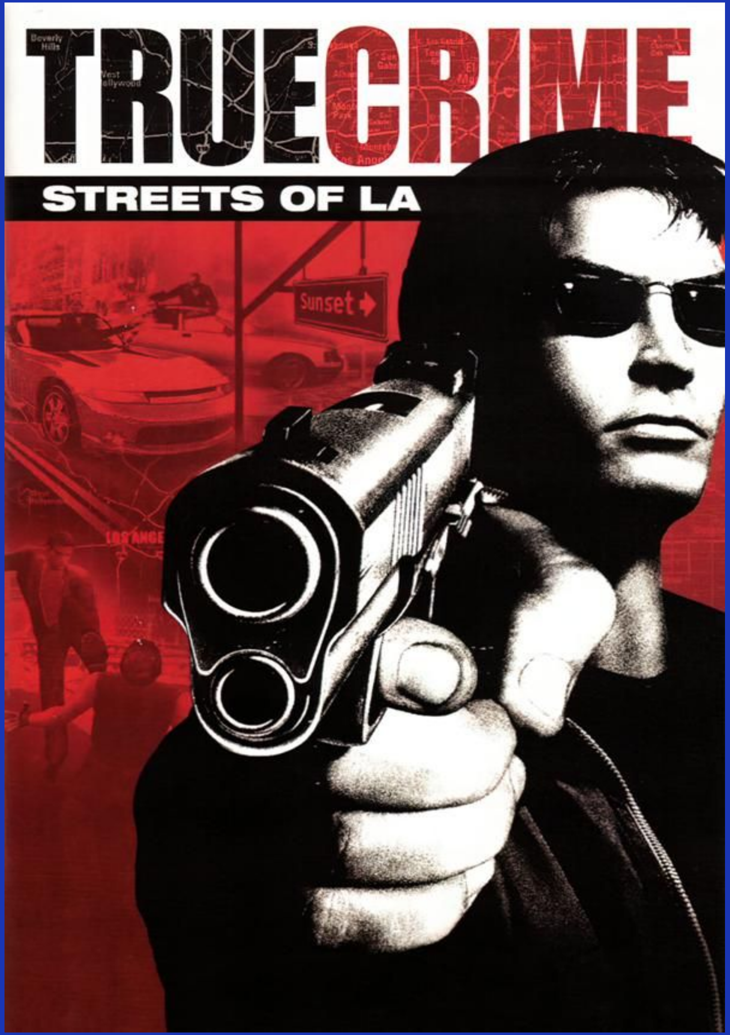
Data Driven Detectives

Los Angeles Crime Investigation

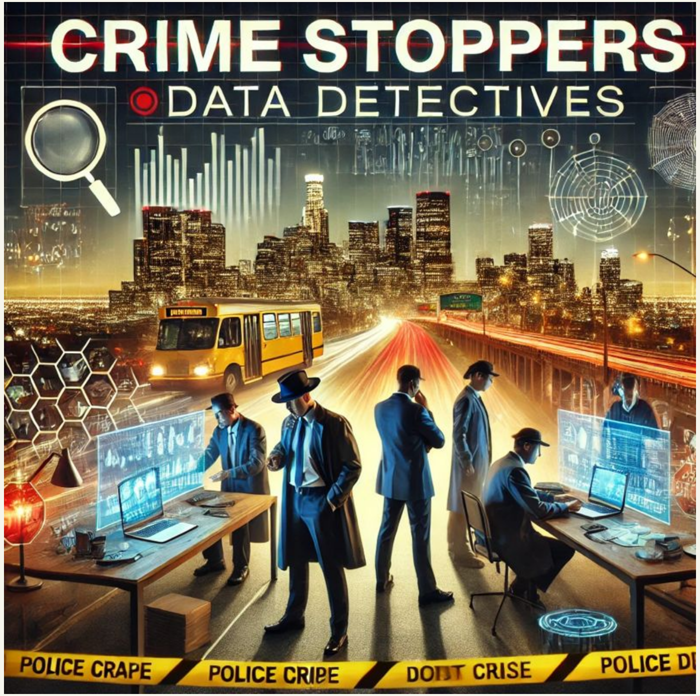
Team Members:

Abhinaysai Kamineni
Lasya Raghavendra
Neeraj Magadum
Aakash Hariharan
Amogh Ramagiri



Synopsis of Data

- Data talks about close to 1 million data records & 28 columns which include various levels of key features.
- Data Cleaning and Pre-processing consists of excluding NOISE and redundant columns from the dataset.
- Timeline of the Datasets talks about various levels of factors and exceptional circumstances for variety of crimes.
- Additional Steps also included in considering coordinate validation and temporal verification steps.



SMART QUESTION 1

How accurately can we predict the time taken to report different types of crimes in Los Angeles, considering factors such as crime type, location, and victim demographics, to identify potential reporting delays that might affect crime investigation efficiency

What?

Question Objective

To predict the time difference between when a crime occurs (Date_Occc) and when it's reported (Date_Rptd) by analyzing:

- Crime type patterns
- Location-based reporting behaviors(LAT and LON)
- Demographic influences on reporting speed(eg Sex, Age, Victim Descent)

Why?

Why This Matters

1. Investigation Efficiency
    - Quick reporting
    - Helps identify areas with systematic reporting delays
    - Allows better resource allocation for investigations

2. Pattern Analysis
    - Different crimes may have different reporting patterns
    - Certain locations might show consistent delays
    - Demographic factors could influence reporting behavior

HOW ?

Implementation Approach

1. Feature Engineering
   - Calculate reporting delay: Date_Rptd - Date_Occ
   - Create delay categories:
     - Same Day (≤1 day)
     - Within Week (≤7 days)
     - Within Month (≤30 days)
     - Over Month (>30 days)
2. Key Variables
   - Geographic: AREA.NAME, LAT, LON
   - Crime-specific: Crm.Cd
   - Demographic: Vict.Age, Vict.Sex, Vict.Descent

XGBoost Model Parameters

Core Parameters

- objective: "multi:softmax" (4-class classification)
- num_class: 4 (Same Day, Within Week, Within Month, Over Month)
- eta: 0.3

Complexity Control

- max_depth: 6
- min_child_weight: 1

Overfitting Prevention

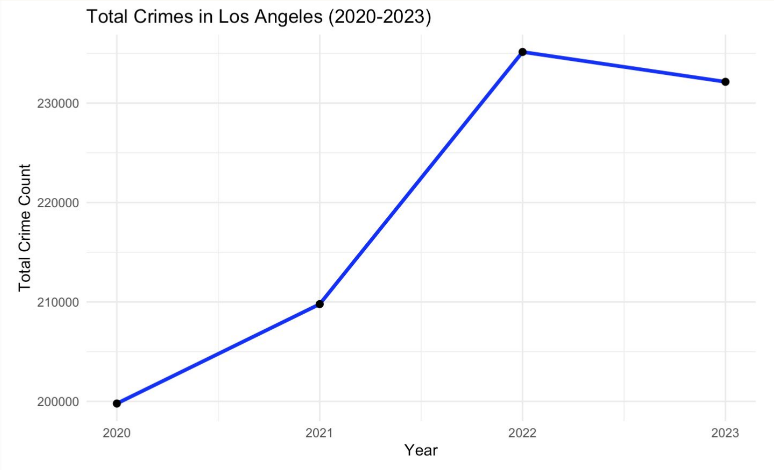
- subsample: 0.8
- colsample_bytree: 0.8

Model achieved 71.03% accuracy

SMART QUESTION 2

How can we analyze the increase in crime rates across Los Angeles neighborhoods from 2020 to 2023 to identify victim demographic patterns, across high-crime areas on a periodic basis?

- **Crime Increase:** Total crimes in Los Angeles rose from **2020 to 2022**, peaking at over **230,000** incidents.
- **2023 Decline:** A slight drop in crime was observed in **2023**.



- **Monthly Trends in 77th Street**: Crime incidents dropped steadily from **January 2020** (1305 crimes) to **March 2020** (1012 crimes).
- **Fluctuations & Stabilization**: After a low point in **April 2020** (1064 crimes), crime counts stabilized between **1076–1098** in the following months. Monitoring these trends helps reveal neighborhood-specific patterns.

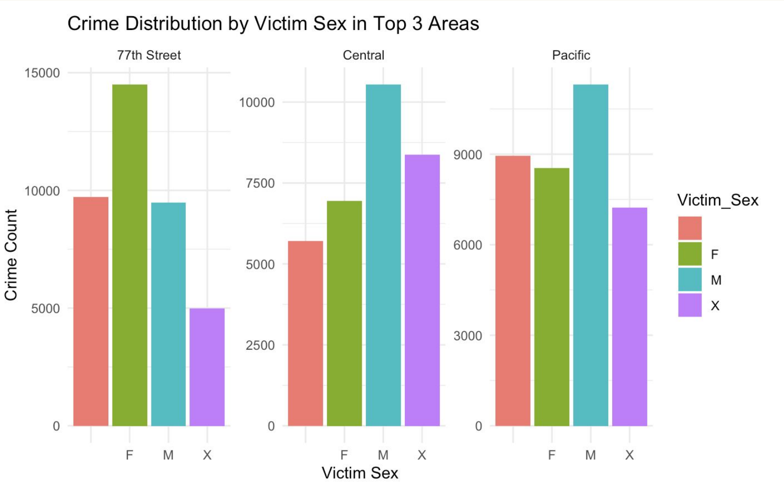
A tibble: 6 × 3

Area_Name <chr>	Month <date>	Crime_Count <int>
77th Street	2020-01-01	1305
77th Street	2020-02-01	1104
77th Street	2020-03-01	1012
77th Street	2020-04-01	1064
77th Street	2020-05-01	1076
77th Street	2020-06-01	1098

6 rows

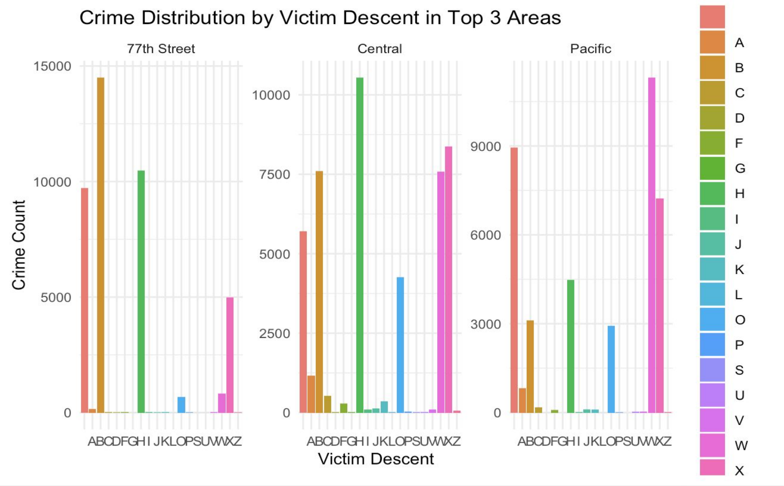
Top 3 Crime Areas Analysis:-

- **Pacific, Central, and 77th Street** reported the highest crime counts.
- **Trends**: Crime rates show consistent increases in Pacific and Central.

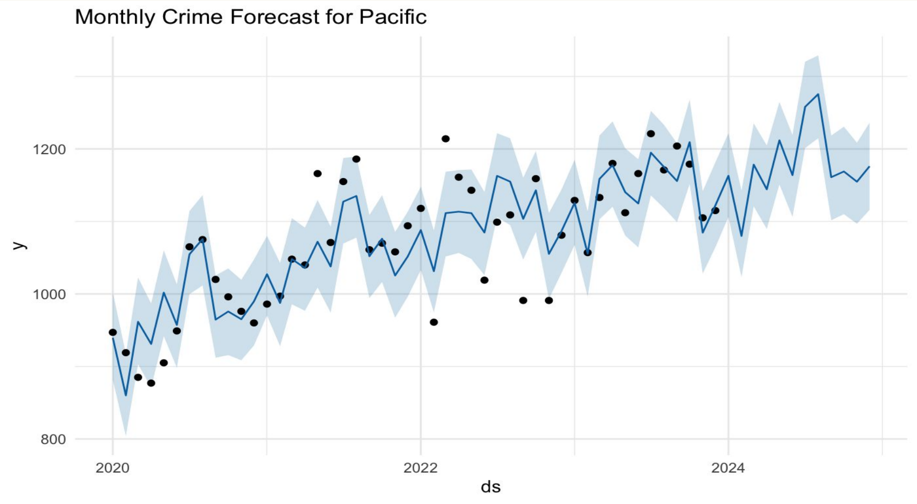


Victim Demographics:-

- **Sex**: Female victims dominate 77th Street; Males are higher in Central and Pacific.
- **Descent**: Hispanic (H) and White (W) victims are most common across all areas.



- **Top Crime Area**: The **Pacific area** is one of the top three neighborhoods with the highest crime rates in Los Angeles.
- **Rising Trend**: Crime has shown a steady upward trend from **2020 to 2024**, with projections indicating over **1200 monthly incidents** by late 2024.



Forecast Accuracy: $R^2 = 0.749$

Victim Descent and Victim Sex Models:-

- **Victim Descent Model**:
  - Achieved **65.8% accuracy**.
  - Highest sensitivity for class "H" (Hispanic).
- **Victim Sex Model**:
  - Achieved **68.8% accuracy**.
  - Strong performance for class "X" (Unknown sex)

SMART QUESTION 3

How effectively can crime prediction models utilize factors like victim descent, spatial data, and temporal patterns the likelihood and type of criminal incidents in different areas of Los Angeles?

Additional Pre-processing Steps & Model Selection

Crime Level Categories

- Violent Crimes.
- Theft Burglary.
- Vehicle Related.
- Other Crimes.

Spatial Zones

- Cell Division Aggregation.
- Spatial density calculations.
- Location-based clustering.

Temporal Features

- Morning (5AM-12PM)
- Afternoon (12PM-5PM)
- Evening (5PM-10PM)
- Night (10PM-5AM)

Model Selection

- Considered to be Optimal for categories defined.
- Features highlights the importance of ranking.
- High Dimensional Data can be handled in optimal level.

Best Model Accuracy & Comparison.

60.2%

Ensemble Method

62.7%

Random Forest

71.3%

XGBoost

Random Forest: - **Victim Descent** and **Victim Sex** are significant across most categories, particularly in **Vehicle Related Crimes**.
Spatial Zone and **Time Period** play critical roles in determining crime categories.
Area shows high importance in identifying **Theft/Burglary**.
Victim Age is important but has mixed effects across categories.

XGBoost: - Validating the Target Categories
Balanced Multi-Class Classification: XGBoost effectively classifies crimes into four categories using engineered spatial and temporal features, ensuring a balanced representation of all classes.
Key Metrics: Achieved high performance for frequent classes like Theft/Burglary and Vehicle Related Crimes,
Critical predictors include time period, spatial zones, and demographic variables like Victim Descent and Victim Sex.

Model & Contingency Analysis.

1. **Strong Performance for Common Crimes**:
   Both models effectively predict frequent categories like **Vehicle Related Crimes**, leveraging spatial and temporal features.

2. **Temporal Patterns**:
   Features like **time periods**, **high-risk hours**, and **peak times** significantly enhance the models' ability to predict crimes with distinct temporal patterns.

3. **Spatial Zones Improve Localization**:
   Incorporating spatial data, such as latitude/longitude zones and distance from city centers, helps differentiate crime likelihood across geographic regions.

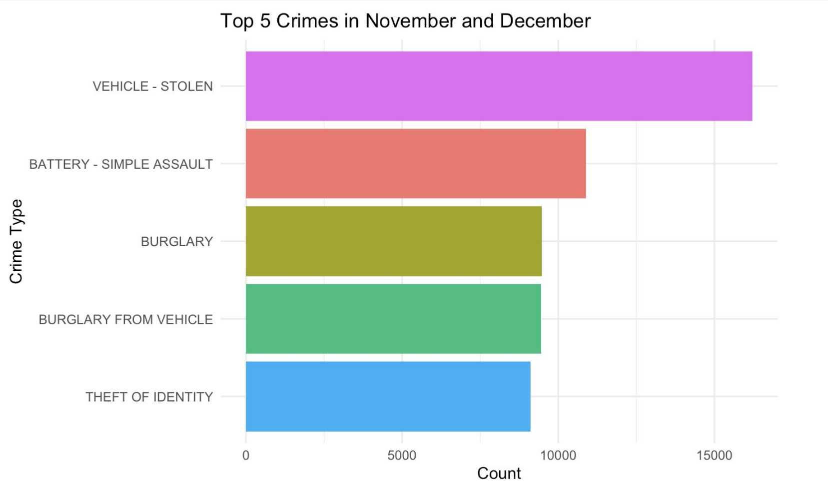
4. **Demographic Variables Are Key Predictors**:
   Factors like **Victim Descent**, **Victim Age**, and **Victim Sex** contribute meaningfully to classifying crimes, especially for **Theft/Burglary** and **Other Crimes**.

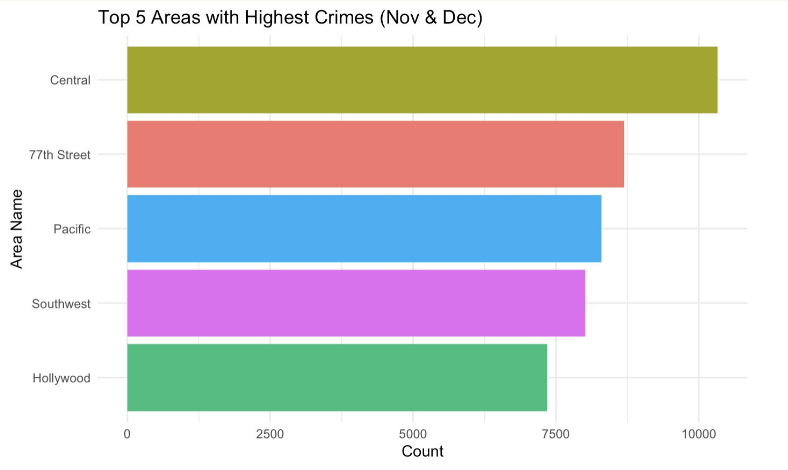
SMART QUESTION 4

How effectively can past data on holiday-season crimes in Los Angeles be utilized to predict the likelihood and occurrence of crimes during November and December?

Crime Analysis for Nov & Dec

- Total Records Analyzed: 144,212
- Vehicle theft is the most common crime during the holiday season.
- Crimes like burglary and assault also show significant activity.
- The Central district reported the highest crime activity during November and December.
- 77th Street, Pacific, Southwest, and Hollywood follow closely, collectively representing areas with significant criminal activity requiring targeted prevention efforts.
- Understanding these patterns can help law enforcement focus resources effectively.





Data Preprocessing

- Categorical variables (e.g., Victim Sex, Descent) were one-hot encoded.
- Age categorized into meaningful groups (e.g., Minor, Adult, Senior).
- New variables created: is_violent_crime, time_period, has_weapon.
- Dataset split: Training (70%) and Testing (30%).

age_group <fctr>	Count <int>
Adult	45946
NA	37822
Young Adult	30428
Senior	21639
Minor	4209
Elderly	4168

6 rows

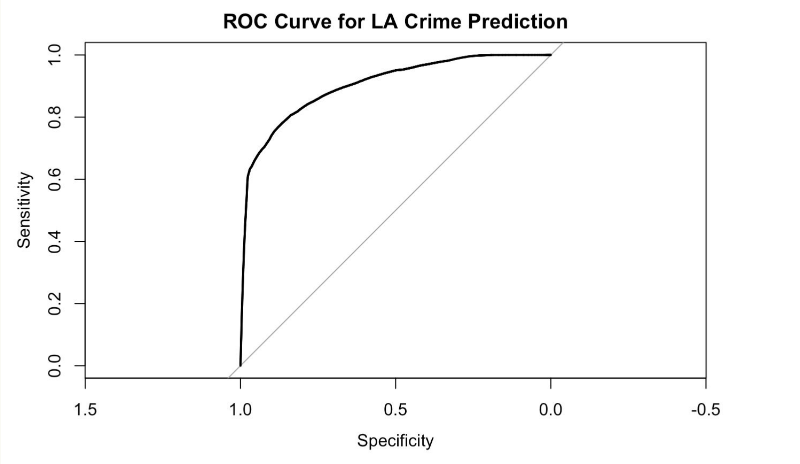
	Feature <chr>	Importance <dbl>
Vict.Sex	Vict.Sex	4.40
has_weapon	has_weapon	3.89
Vict.DescentC	Vict.DescentC	2.58
Vict.DescentV	Vict.DescentV	2.08
Vict.DescentJ	Vict.DescentJ	1.96

5 rows

Logistic Regression Model

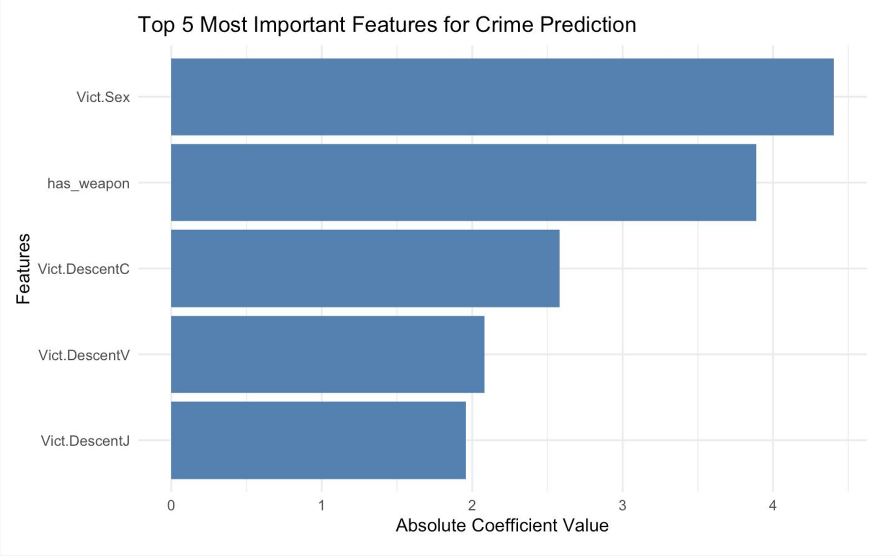
- Logistic regression with LASSO regularization used to prevent overfitting, with a model accuracy of **88.7%**.
- Features include victim age, sex, descent, crime time, and weapon involvement.
- Target Variable: Binary classification (Violent vs. Non-Violent Crimes).
- Evaluated on metrics: Accuracy, Precision, Recall, F1 Score, and ROC Curve.
- AUC Score : **0.903**

```
[1] "Confusion Matrix:"
      Predicted
Actual      0      1
     0 215406   6328
     1  26674  44710

Model Performance Metrics:
Accuracy: 0.887
Precision: 0.876
Recall: 0.626
F1 Score: 0.73
```



Feature Importance

- The victim's gender turned out to be the most important factor, showing that it strongly influences the likelihood of certain crimes.
- Crimes involving weapons are often more severe and follow distinct patterns, making weapon involvement a key predictor.
- Certain ethnic groups, like Chinese (C), Vietnamese (V), and Japanese (J), show trends in crime patterns, revealing potential vulnerabilities.



SMART QUESTION 5

How can we use factors like area, victim demographics, and weapons types, to predict whether a crime is violent or non-violent, and forecast the trend of violent crime incidents over the next two years?

DATA PRE-PROCESSING

Feature Selection

Columns such as Mocodes, Crm.Cd.2, and others were removed from the dataset due to their irrelevance to the smart question analysis and the high volume of null values."

Empty Strings

Rows containing empty strings or missing values, which were causing disruptions to the model, were removed."

Data Type Conversion

"Features such as 'Date Occurred' and 'Date Reported' were originally of character type and have been converted to the correct date format."

New Feature (Crime Type)

A new feature has been named “Crime Type” added to the dataset that classifies crimes as either violent or non-violent."

PART - 1

A. Logistic Regression

Precision: 0.781
Recall: 0.981
F1-Score: 0.87

```
Call:
glm(formula = Crime_Type ~ Area_Name + Victim_Age + Victim_Sex +
    Weapons_Used, family = binomial(), data = train_data_reduced)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.963569   0.123461   24.00  < 2e-16 ***
Area_NameCentral      0.068002   0.114821    0.59  0.55369
Area_NameDevonshire   0.132764   0.151873    0.87  0.38202
Area_NameFoothill    -0.142881   0.144976   -0.99  0.32435
Area_NameHarbor       0.548255   0.154100    3.56  0.00037 ***
Area_NameHollenbeck   0.050457   0.141758    0.36  0.72189
Area_NameHollywood   -0.266539   0.114432   -2.33  0.01985 *
Area_NameMission      0.325263   0.151905    2.14  0.03226 *
Area_NameN Hollywood  0.097721   0.140388    0.70  0.48638
Area_NameNewton       0.072060   0.126721    0.57  0.56959
Area_NameNortheast   -0.071381   0.143203   -0.50  0.61816
Area_NameOlympic      0.162297   0.127110    1.28  0.20166
Area_NamePacific     -0.557306   0.120816   -4.61  4.0e-06 ***
Area_NameRampart      0.077603   0.126746    0.61  0.54036
Area_NameSoutheast   -0.002390   0.117591   -0.02  0.98378
Area_NameSouthwest   -0.080335   0.117498   -0.68  0.49416
Area_NameTopanga      0.090270   0.149248    0.60  0.54529
Area_NameVan Nuys     0.096341   0.150091    0.64  0.52095
Area_NameWest LA     -0.694564   0.137848   -5.04  4.7e-07 ***
Area_NameWest Valley -0.173507   0.138397   -1.25  0.20995
Area_NameWilshire     0.319080   0.147473    2.16  0.03049 *
Victim_Age           -0.009391   0.001425   -6.59  4.4e-11 ***
Victim_SexM          -0.174624   0.048481   -3.60  0.00032 ***
Victim_SexX          -1.657937   0.101763  -16.29  < 2e-16 ***
Weapons_Used         -0.003275   0.000213  -15.36  < 2e-16 ***
```

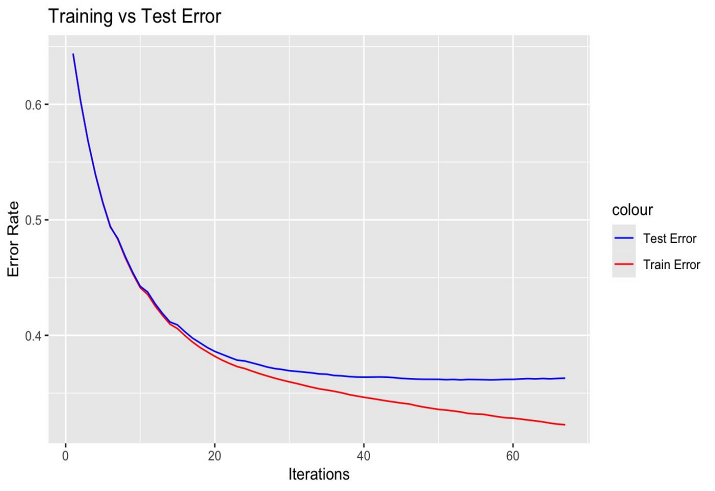
	VIOLENT	NON-VIOLENT
VIOLENT	1901	1396
NON-VIOLENT	20692	73889

PART - 1

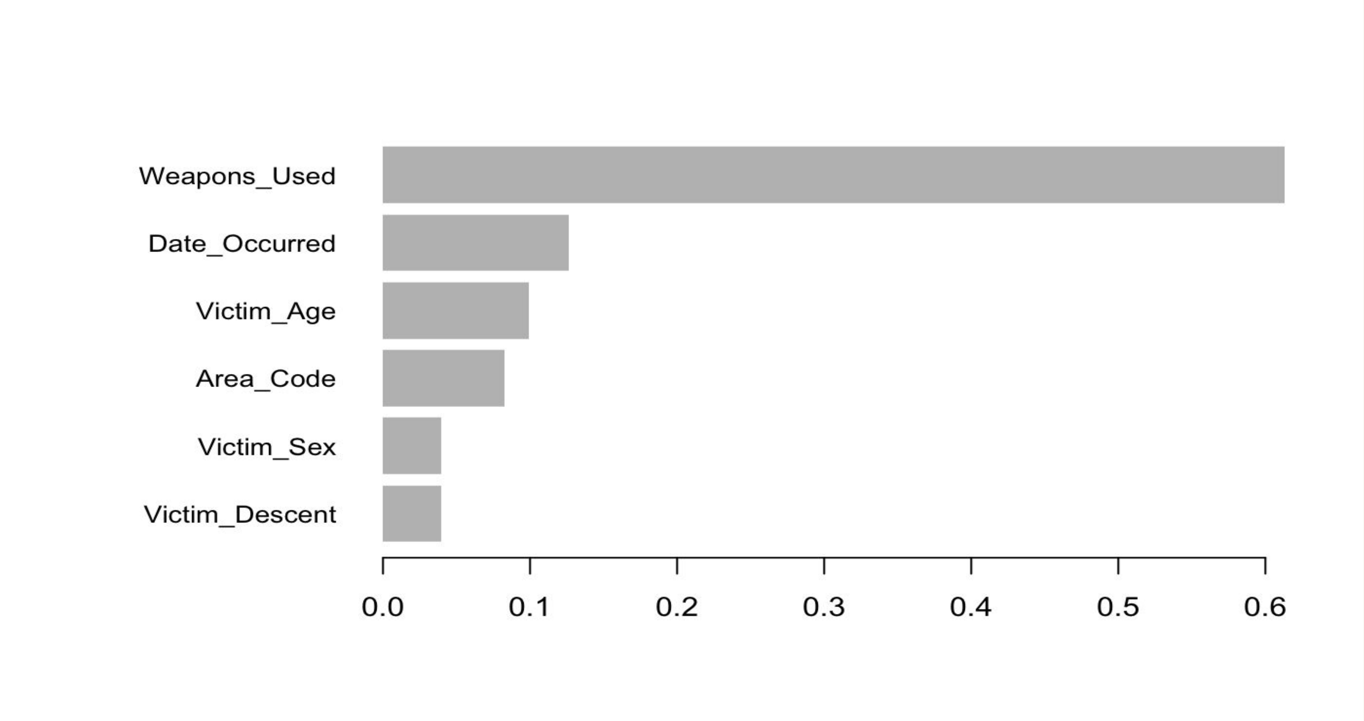
B. XGB MODEL

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2521  460
         1   84  360

               Accuracy : 0.841
                 95% CI : (0.828, 0.853)
    No Information Rate : 0.761
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.483
```

```
Precision: 0.846
Recall: 0.968
F1-Score: 0.903
```



FEATURE IMPORTANCE

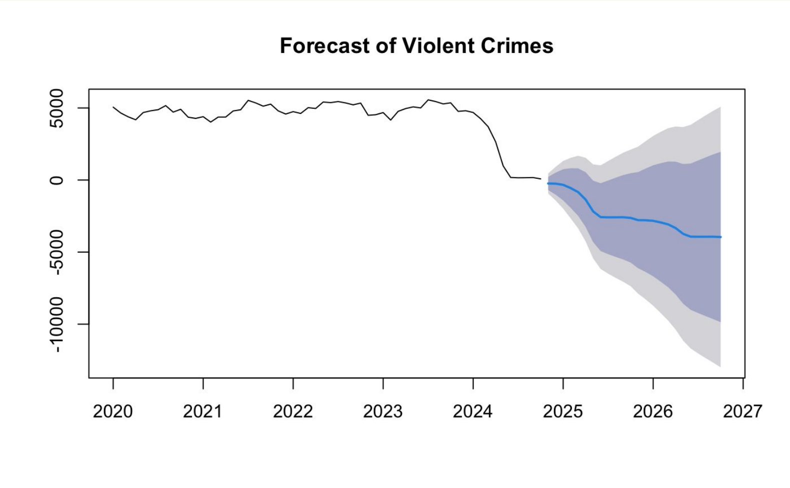


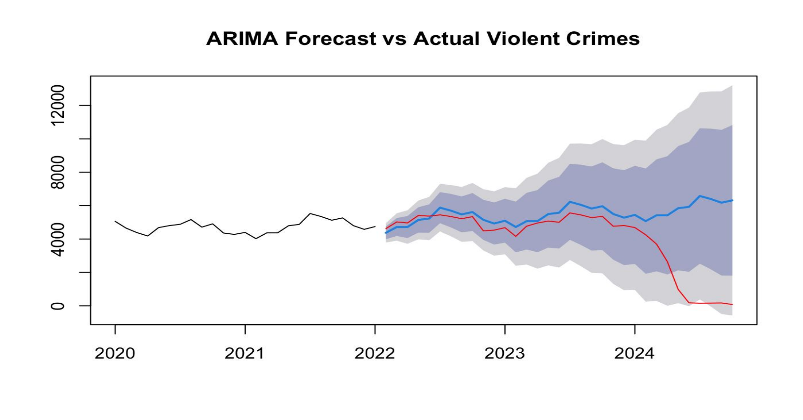
PART - 2

A. ARIMA MODEL

```
Coefficients:
         ar1    ar2   sar1
       0.353  0.202  0.490
s.e.   0.132  0.130  0.134

sigma^2 = 125944:  log likelihood = -416
AIC=840   AICc=840   BIC=848

Training set error measures:
              ME RMSE MAE  MPE MAPE  MASE     ACF1
Training set -36  342 253 2.19 19.6 0.266 -0.00634
```

The small MAPE and ACF1 values indicate that, overall, the model has some predictive power and has removed most of the correlation in the residuals.



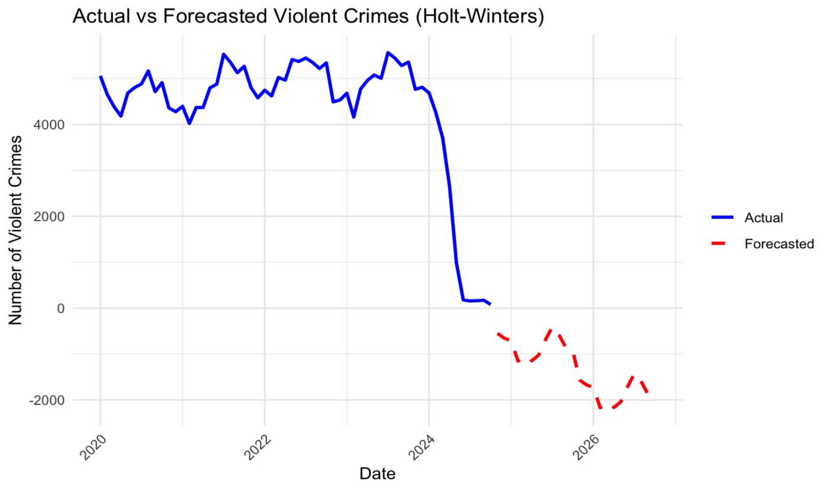


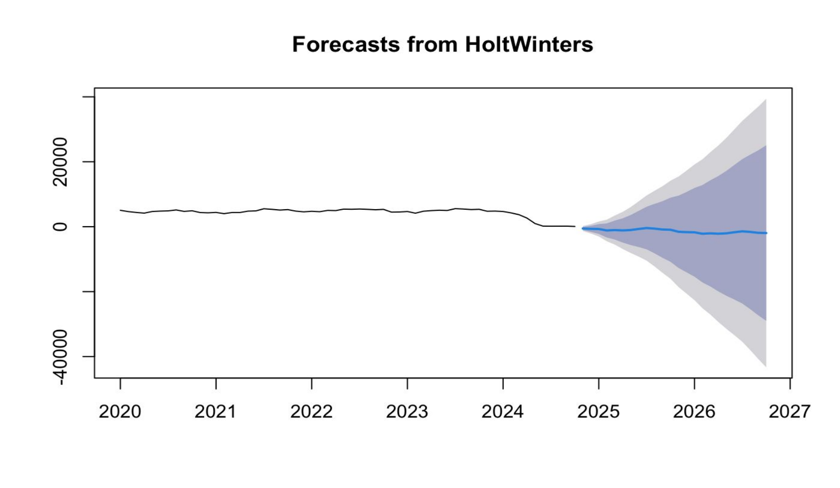
PART - 2

B. HOLT-WINTER MODEL

The Holt-Winters model has performed reasonably well, as indicated by the MASE (0.271) and ACF1 (0.0376). The model appears to perform better than a naive forecasting approach (which would predict no change from the previous time period).

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-2.83	382	258	21.3	50	0.271	0.0376





References

References for Crime Prediction Models

1. **Chen, T., & Guestrin, C. (2016)**: *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
   Describes the gradient boosting algorithm used for structured data analysis.

2. **Geospatial and Temporal Crime Analysis**:
   Chainey, S., & Ratcliffe, J. (2005). *GIS and Crime Mapping*. John Wiley & Sons.
   Highlights the importance of spatial and temporal data in understanding crime patterns.

3. **Evaluating Predictive Models**:
   Powers, D. M. (2011). *Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation*. Journal of Machine Learning Technologies, 2(1), 37–63.
   Discusses metrics like F2 score and their relevance in evaluating classification models.