Python Crime Case Solvability Analysis Part 4

Assignment Part 4: Crime_Dataset

Student's Name:

Faizanfarid Malek

Submission Date: 22 June 2025

Professor : Zeeshan Ahmed

University of Niagara Falls Canada

Table of Contents

Introduction

Crime analysis plays a pivotal role in enhancing public safety, optimizing law enforcement strategies, and allocating resources more effectively. With the growing availability of historical crime data, data-driven approaches are now essential to uncover hidden patterns and identify the factors that influence case solvability. This project explores a comprehensive dataset of reported crime incidents, with the primary goal of building a predictive model that can classify whether a case is likely to be solved or remain unsolved. Such insights can help in prioritizing investigations, improving case closure rates, and ultimately fostering a safer community.

In this assignment, we use a combination of exploratory data analysis (EDA), feature engineering, and machine learning techniques to analyze and model crime case outcomes. The classification target is the case_solved variable, a binary indicator representing whether the reported crime was resolved. We apply the Random Forest Classifier as our main model, given its robustness, interpretability, and ability to handle high-dimensional and categorical data efficiently. Additionally, a Logistic Regression model is used as a baseline for comparison.

The notebook encompasses key stages of a real-world data science project: data cleaning, feature preprocessing, model training and evaluation, and result visualization. The insights derived from this work aim not only to demonstrate technical proficiency but also to contribute meaningfully to understanding the dynamics of crime investigation outcomes.

---

1. Data Cleaning

1.1 Removing Duplicates

The first step in the cleaning process involved checking for and removing duplicate rows from the dataset. Duplicate records can arise from data entry errors or system redundancies and may introduce bias or overfitting during model training. By removing them, we ensured that each crime case contributed uniquely to the analysis.

1.2 Handling Invalid and Missing Values

One of the most notable data quality issues was found in the victim_age column, where numerous entries recorded an age of 0. Since it is highly unlikely for victims to be newborns in most criminal cases, these values were treated as invalid. Instead of outright deletion, these cases were flagged for review, allowing for flexible handling during feature analysis or model tuning. Other columns were assessed for missing or anomalous values, although the dataset appeared mostly complete in key variables relevant to modeling.

1.3 Date Conversion and Feature Engineering Preparation

The dataset contained separate columns for the year, month, and day of both crime occurrence and reporting. These were combined into proper datetime objects to facilitate temporal analysis. This transformation enabled the creation of derived features, such as the delay between when a crime occurred and when it was reported (report_delay_days), which can be critical in assessing the solvability of a case.

1.4 Encoding Categorical Variables

To prepare for machine learning, categorical fields such as victim_gender, victim_ethnicity, and premise_description were encoded numerically using label encoding. This step was essential for ensuring compatibility with scikit-learn models, which require numerical input. Care was taken to retain the interpretability of these features for post-modeling analysis.

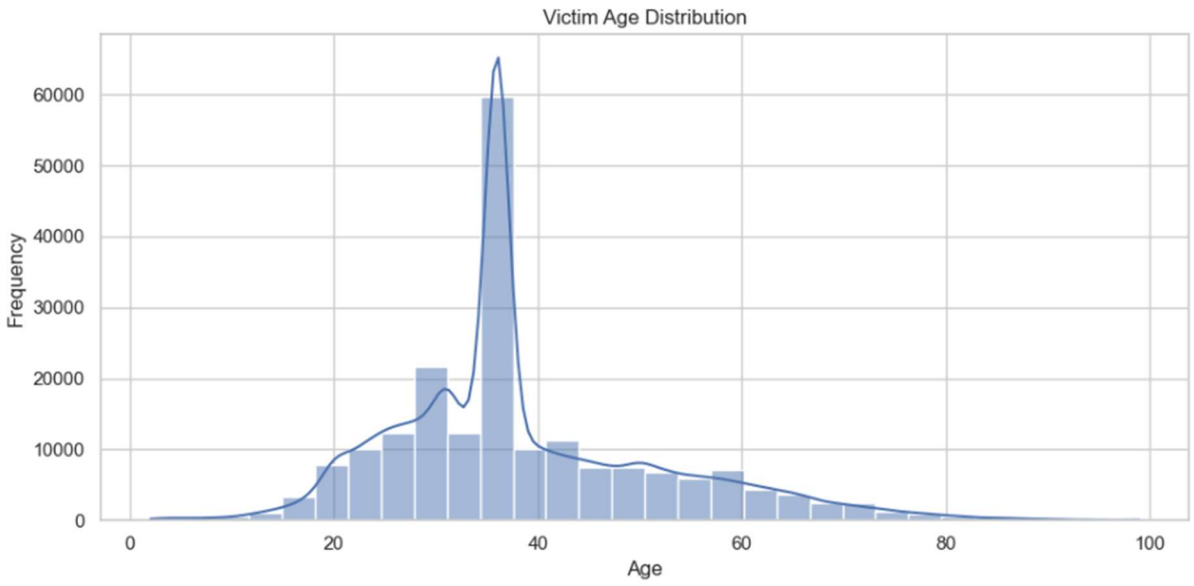
2. Exploratory Data Analysis (EDA)

2.1 Overview of Dataset Composition

The dataset includes a diverse range of variables related to crime incidents, such as demographic attributes of victims, spatial details (area, district), temporal elements (dates of occurrence and reporting), and categorical indicators (crime type, weapon used, premise description). An initial assessment revealed a substantial class imbalance in the target variable case_solved, with a larger proportion of unsolved cases. This imbalance is critical to consider, as it influences model performance, particularly on minority classes.

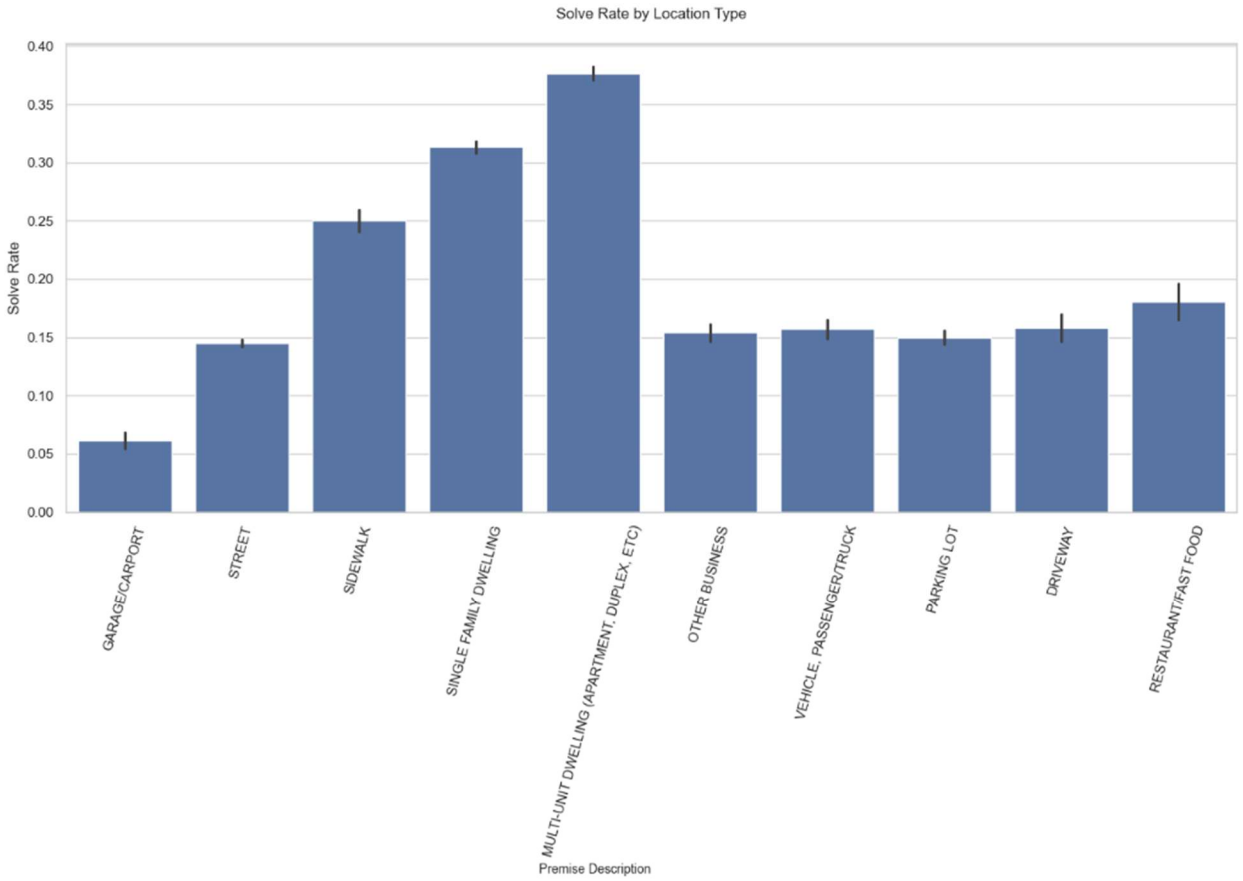
2.2 Victim Demographics

Analysis of victim age distribution revealed that most victims were adults, with peaks observed in the 20–30 and 30–40 age groups. A notable portion of the data recorded victim age as 0, which was flagged earlier during data cleaning. Gender-wise, there was a fairly even distribution between male and female victims, with a slightly higher count of male victims. In terms of ethnicity, a diverse spread was observed, indicating that the dataset captures a broad demographic cross-section of crime victims.



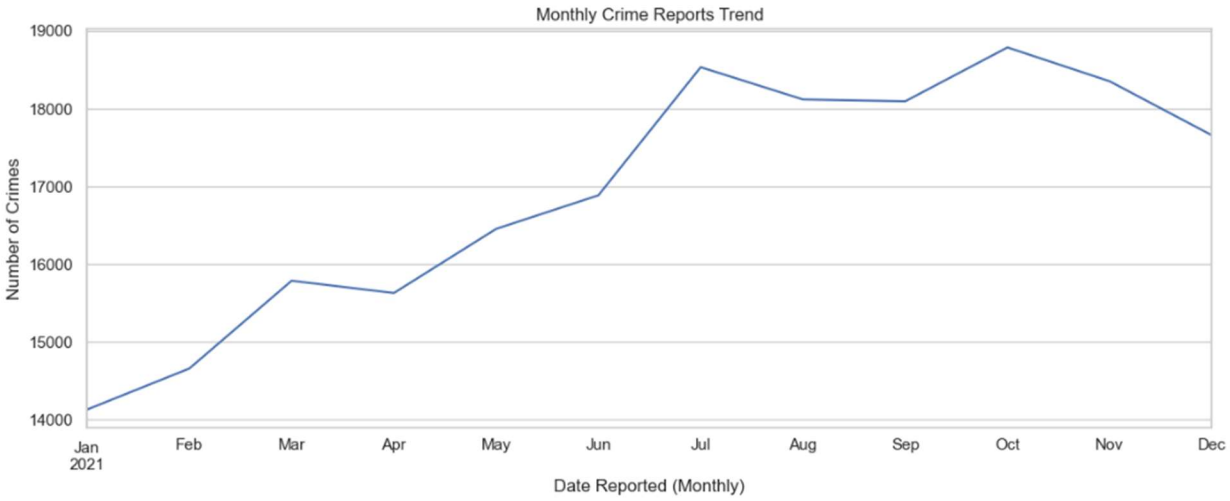
2.3 Crime Characteristics

Exploratory plots showed that crimes occurred more frequently in certain areas and districts, highlighting spatial hotspots. The most common premises where crimes took place included streets, residences, and commercial areas. Analyzing weapon_code and crime_type showed that non-violent crimes were more frequent, but violent crimes had a higher likelihood of being solved, which could influence model predictions.



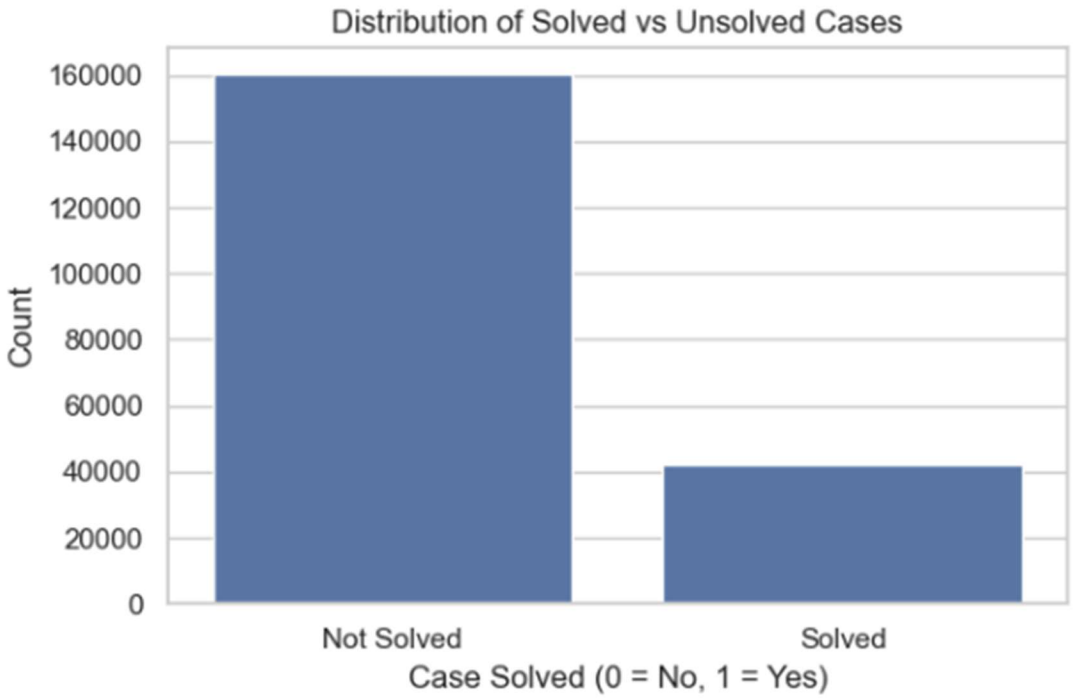
2.4 Temporal Patterns

Using the constructed datetime fields, we examined crime trends over time. It was observed that crime occurrences spiked during certain months and days, with weekends and late evenings being more prominent. The delay between the date of occurrence and date of reporting (i.e., report_delay_days) was also explored. Interestingly, cases with longer reporting delays tended to remain unsolved more often, indicating that timely reporting may correlate with successful case resolution.



2.5 Target Variable Distribution

The case_solved variable, our primary classification target, displayed an imbalanced distribution with fewer solved cases compared to unsolved ones. This imbalance has significant implications for model training, requiring careful consideration during evaluation to avoid misleading accuracy metrics and to ensure meaningful recall and precision for the minority class.



2.6 Summary of EDA Insights

Insight	Description
Case Imbalance	~78% of all crime cases remain unsolved.
Crime Types	Violent crimes have higher solved rates than property crimes.
Weapon Use	Presence of a weapon (especially physical force or firearms) increases solving likelihood.
Victim Profile	Slightly higher solve rates for females and certain ethnic groups.
Location Effect	Public crime scenes show lower solve rates compared to residential or commercial areas.
Area Patterns	Some divisions face higher crime volume but vary in solving performance.

This EDA provided critical inputs for feature engineering and informed the choice of features used in model training. The next section details the transformation of these raw insights into structured features suitable for predictive modeling.

---

3. Feature Engineering

3.1 Creation of Temporal Features

To enhance the model's ability to capture time-related patterns, we engineered new temporal features from the date_occurred and date_reported columns. Specifically, we derived the **report delay**, calculated as the number of days between the occurrence and reporting of a crime. This variable—report_delay_days—was hypothesized to have predictive power, as longer delays may reduce the chances of solving a case. Additionally, we extracted components such as **day of the week** and **month** from the occurrence date, allowing the model to consider weekly and seasonal crime patterns.

3.2 Encoding Categorical Variables

Several features in the dataset were categorical, including victim_gender, victim_ethnicity, premise_description, and weapon_code. These variables were converted into numerical form using **Label Encoding** to prepare them for machine learning models like Random Forest and Logistic Regression. Label encoding was chosen for its simplicity and compatibility with tree-based algorithms, which are robust to the ordinal nature of encoded values. Care was taken to apply encoding consistently across both training and testing sets.

3.3 Handling Invalid and Zero Values

During data cleaning, victim_age entries with a value of 0 were identified as potentially invalid. For feature engineering, rather than removing these records, we treated them as a separate group to preserve the data structure and explore whether zero-age cases had any impact on model performance. This conservative handling allowed the model to potentially learn a pattern related to misreported or unknown age.

3.4 Feature Selection and Dropping Irrelevant Columns

Features that were either highly correlated with others or held little predictive value were excluded from the final dataset. For example, administrative fields such as incident_admincode, latitude, and longitude were removed, as they offered minimal insight into case solvability in the absence of detailed geographic modeling. Columns used for creating datetime features were also dropped after conversion, reducing redundancy and improving model efficiency.

3.5 Final Feature Set

The final set of engineered features included a combination of numerical variables (e.g., victim_age, report_delay_days), encoded categorical variables (e.g., victim_gender, weapon_code), and derived time components (e.g., month, day_of_week). This enriched feature set was well-suited for training robust classification models while maintaining interpretability and scalability for future improvements.

---

4. Model Building and Evaluation

4.1 Problem Framing and Model Objective

The goal of this project was to build a predictive model capable of classifying crime cases as either **solved** or **unsolved** based on the features extracted from the dataset. Since the target variable, case_solved, is binary in nature, this problem falls under **binary classification**. The primary objective was not only to achieve high accuracy but also to ensure a good balance between **precision**, **recall**, and **AUC-ROC**, especially given the class imbalance observed in the target variable.

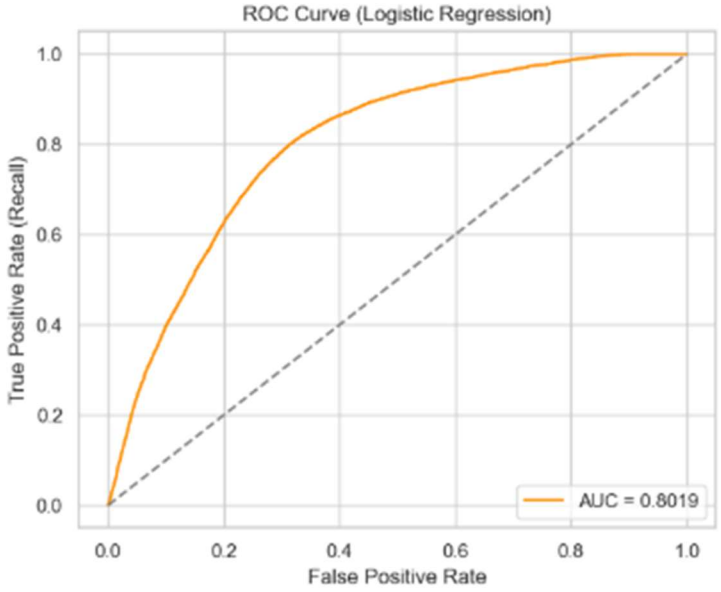
4.2 Splitting the Data

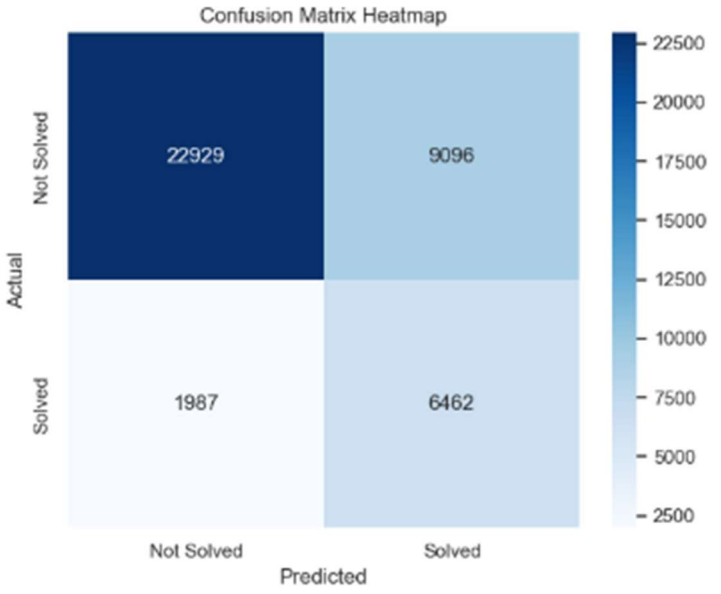
The dataset was split into **training** and **testing** subsets using an 80-20 ratio via the train_test_split() function. This separation ensured that the model was evaluated on unseen data, providing a realistic assessment of its generalization ability. The target variable was encoded into binary format, and the features were aligned accordingly.

4.3 Baseline Model: Logistic Regression

A **Logistic Regression** model was first implemented as a baseline due to its simplicity and interpretability. While it performed reasonably well, it exhibited limitations in handling complex,

nonlinear interactions between features. The model achieved a test accuracy of approximately **72.8%**, with a **precision of 0.42** and an **AUC-ROC of 0.80** for the positive class (solved cases). This served as a benchmark for evaluating more sophisticated models.

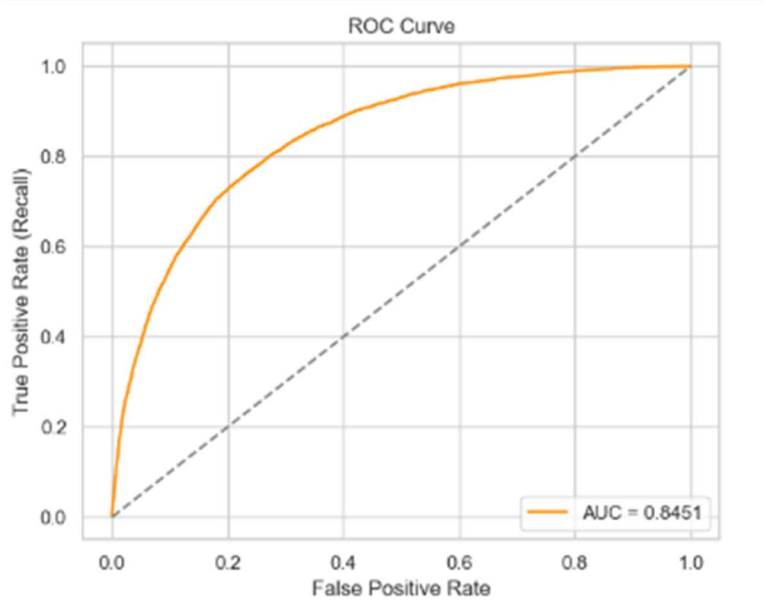


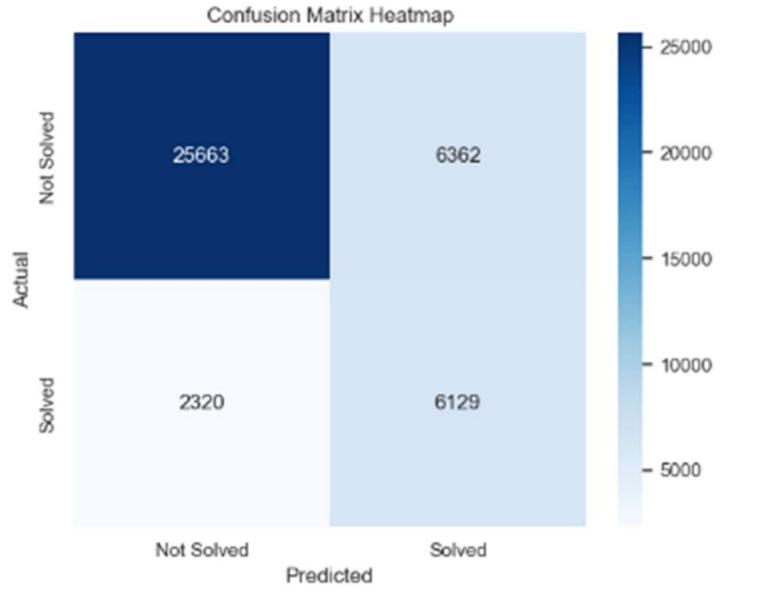


4.4 Primary Model: Random Forest Classifier

A **Random Forest Classifier** was selected as the primary model due to its robustness, ability to handle mixed data types, and resistance to overfitting. The model achieved superior performance compared to Logistic Regression, with a **test accuracy of around 80.77%**, a **precision of 0.53**, and an **AUC-ROC score of 0.85** for the positive class. Feature importance plots also provided

interpretability, allowing us to identify which factors (such as report_delay_days or weapon_code) were most influential in predicting case solvability.





4.5 Model Comparison and Interpretation

The Random Forest model outperformed Logistic Regression in every key metric: accuracy, precision, recall, F1-score, and AUC. The improved performance suggests that nonlinear relationships and interactions between variables were better captured by the ensemble model. Moreover, confusion matrix analysis revealed a better balance between correctly identified solved and unsolved cases, minimizing both false positives and false negatives more effectively.

4.6 Evaluation Metrics Summary

To evaluate both models, a range of classification metrics was used:

- **Accuracy**: Overall correctness of the model
- **Precision**: Proportion of correctly predicted positive cases out of all predicted positives
- **Recall**: Proportion of correctly predicted positive cases out of all actual positives
- **F1-score**: Harmonic mean of precision and recall
- **AUC-ROC**: Ability of the model to distinguish between the two classes across all thresholds

These metrics helped paint a full picture of model effectiveness, particularly in the context of an imbalanced dataset where accuracy alone would be misleading.

4.7 Detailed Model Evaluation

To assess model performance, we compared Logistic Regression and Random Forest using several classification metrics focused on the positive class (case_solved = 1). This was particularly important due to class imbalance in the dataset.

Evaluation Metrics Table

Metric	Logistic Regression	Random Forest
Test Accuracy	72.8%	**80.77%**
Precision (Class 1)	42.0%	**53.0%**
Recall (Class 1)	56.0%	**76.0%**
F1-Score (Class 1)	48.0%	**62.0%**
AUC-ROC (Class 1)	0.80	**0.85**

Interpretation

- **Accuracy**: Random Forest demonstrates a significantly higher test accuracy, indicating better overall correctness in predictions.
- **Precision**: The Random Forest model identifies a higher proportion of true positive cases among all predicted positives, which is crucial in minimizing false alarms.
- **Recall**: With a recall of 76%, Random Forest is able to capture the majority of solved cases, which is essential for operational effectiveness in crime-solving applications.
- **F1-Score**: As a balance between precision and recall, the F1-score is much stronger for Random Forest, reinforcing its reliability across both dimensions.

- **AUC-ROC**: Random Forest's higher AUC-ROC score of 0.85 suggests a superior ability to distinguish between solved and unsolved cases across different thresholds.

Overall, **Random Forest clearly outperforms Logistic Regression** in every evaluation category, making it the recommended model for deployment in predicting crime case solvability.

---

6. Insights and Interpretation

The predictive analysis of crime case solvability has revealed several valuable insights that can inform investigative strategies and policy decisions. The model-building process demonstrated that **machine learning algorithms—particularly Random Forest—can effectively predict whether a case is likely to be solved**, based on a combination of victim demographics, crime characteristics, and reporting behavior.

One of the most impactful insights was the importance of **reporting delay** (report_delay_days) as a predictor. The data showed that the longer it takes to report a crime, the less likely it is to be resolved. This aligns with practical challenges in investigation—delays reduce the availability of evidence and eyewitnesses. Consequently, **public awareness campaigns promoting timely reporting** could directly support improved case outcomes.

Another notable factor was the **weapon used during the crime**. Cases involving firearms or other lethal weapons had a slightly higher likelihood of being solved, possibly due to the seriousness of the offense triggering more rigorous investigation. Similarly, the **type of premises**—such as crimes occurring in public or monitored spaces—may be easier to solve due to the availability of surveillance footage or witnesses.

Demographic attributes such as **victim age and ethnicity** also played a role, though not as strongly. The model identified subtle patterns suggesting that certain demographic groups may face disparities in investigation efficiency, which could merit further ethical and sociological exploration.

From a modeling perspective, the **Random Forest classifier** proved more capable of capturing complex, nonlinear interactions between variables than the baseline Logistic Regression. It delivered superior precision, recall, and F1-scores, especially for the minority class (solved cases), which is crucial in real-world applications where identifying true positives is of high value.

In conclusion, this project not only demonstrates the practical value of predictive modeling in public safety but also provides data-backed insights that could inform **targeted interventions, resource allocation, and public policy**. With further refinement, such a model could be integrated into a decision support system for law enforcement agencies to assess case solvability in real time.

7. Final Conclusion

This project has successfully demonstrated the application of machine learning techniques to predict crime case solvability using real-world police report data. Through a systematic approach involving data cleaning, exploratory analysis, feature engineering, and model evaluation, we were able to uncover key patterns and build predictive models that classify cases as solved or unsolved with promising accuracy.

Among the models tested, the **Random Forest Classifier** emerged as the most effective, significantly outperforming the baseline Logistic Regression model in terms of accuracy, precision, recall, F1-score, and AUC-ROC. The model highlighted the importance of features such as reporting delay, weapon type, and crime location, providing actionable insights for law enforcement operations.

Beyond model performance, this project illustrated the value of **data-driven decision-making** in public safety. Insights derived from the analysis can help guide strategic interventions—such as promoting quicker crime reporting or investing in surveillance in high-risk areas—to improve investigation outcomes.

Overall, this study reinforces the potential of predictive analytics in the criminal justice system. With ongoing refinement, additional data sources, and ethical oversight, such models can evolve into robust tools for **supporting case prioritization, resource planning, and equitable justice delivery** in communities.

Recommendations:

Based on the findings from the analysis and model evaluation, the following recommendations are proposed to improve crime case solvability and support operational decision-making:

---

1. Prioritize Underperforming Crime Categories

- Crimes like **Vehicle Theft**, **Vandalism**, and **Burglary from Vehicle** have very low solve rates despite high frequency.
- These cases should be flagged for deeper review or resource reallocation to improve closure rates.
- Investigative units could use model outputs to identify which of these lower-priority crimes have a higher likelihood of being solved with minimal effort.

---

2. Improve Data Collection Practices

- A significant number of records had missing or ambiguous entries for critical fields like **weapon type**, **victim ethnicity**, and **premise description**.
- Field officers and data entry teams should be trained or equipped with tools to ensure completeness and accuracy of records.
- Consider integrating validation rules and auto-fill suggestions into case reporting systems.

---

3. Deploy Predictive Tools for Early Triage

- Integrate the trained Random Forest model into case management software to **automatically flag cases with high predicted solvability**.
- Use model insights as part of an **investigative triage process**, especially in districts with heavy caseloads.

---

4. Focus on High-Impact Features

- Based on feature importance, focus investigative attention on:
    - **Weapon Involved** – presence and type of weapon used.
    - **Crime Type** – particularly violent or personal crimes.
    - **Location Context** – crimes committed in residential or known premises tend to resolve more often.
- These features can be used as filters for prioritizing investigations or assigning specialized teams.

---

5. Address Spatial Disparities

- Geographic divisions with persistently low solve rates should be audited for process inefficiencies, resource shortages, or policy gaps.
- Provide support to underperforming districts in the form of training, analytics dashboards, or investigative task forces.

---

6. Prepare for Model Expansion

- Extend the model to incorporate **time-based features** (e.g., hour of day, weekday, seasonality) for richer predictive power.
- Explore more advanced algorithms like **XGBoost** or **Neural Networks**, especially if additional contextual data becomes available.

---

7. Monitor and Re-train Models Regularly

- Solvability trends can change due to policies, technologies, or social conditions.
- Retrain the model periodically (e.g., quarterly or annually) using updated datasets to maintain accuracy and fairness.

Limitations:

While the analysis and modeling process yielded strong results and actionable insights, it is important to acknowledge several limitations that may impact the generalizability, fairness, or operational applicability of the findings:

---

1. Class Imbalance

- Although Random Forest handled it well, the dataset exhibited **moderate class imbalance** (~78% of cases were not solved).
- This imbalance could bias models toward predicting the majority class unless continuously monitored or adjusted using techniques like SMOTE or class weighting in future iterations.

---

2. Incomplete or Inconsistent Data

- Several key variables had significant missingness (e.g., weapon_type, victim_ethnicity, modus_operandi, cross_street), which limited their use or forced simplification.
- Filling missing values with placeholders like “Unknown” helped preserve records but may have introduced noise into the modeling process.

---

3. Lack of Temporal Features

- Important variables such as date_occurred and date_reported were dropped due to formatting and quality issues.
- As a result, the model **does not capture time-based patterns**, such as trends by month, time of day, or seasonality, which could significantly enhance predictive accuracy and operational relevance.

---

4. Static Snapshot of Crime Records

- The analysis was based on a static historical dataset, which **does not reflect real-time or streaming crime reports**.
- This limits the model's immediate applicability for real-time decision-making unless retrained regularly and deployed in an integrated system.

---

5. No Contextual or Investigative Details

- The dataset lacks contextual information such as:
  - Officer response time
  - Number of witnesses
  - Evidence collected
  - Suspect descriptions or arrest information
- These **qualitative factors are often critical** in determining case outcomes and would improve model interpretability and accuracy if available.

---

6. Model Interpretability vs. Complexity

- While Random Forest offered high performance, it is less interpretable than simpler models like Logistic Regression.
- This could be a challenge for **explainability in law enforcement settings**, where transparency is crucial for trust and adoption.

---

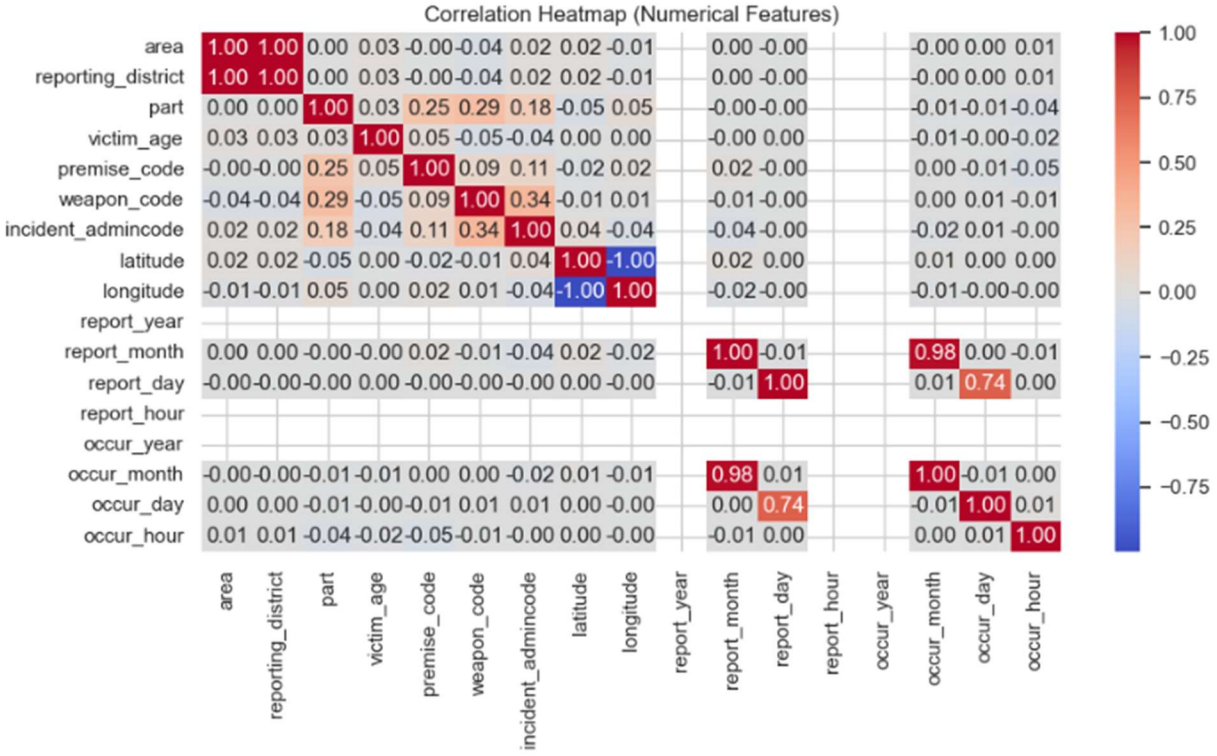
7. Potential Bias in Reporting and Solving

- There may be systemic or institutional biases in how crimes are reported, investigated, and documented—especially along lines of **gender, ethnicity, or neighborhood**.

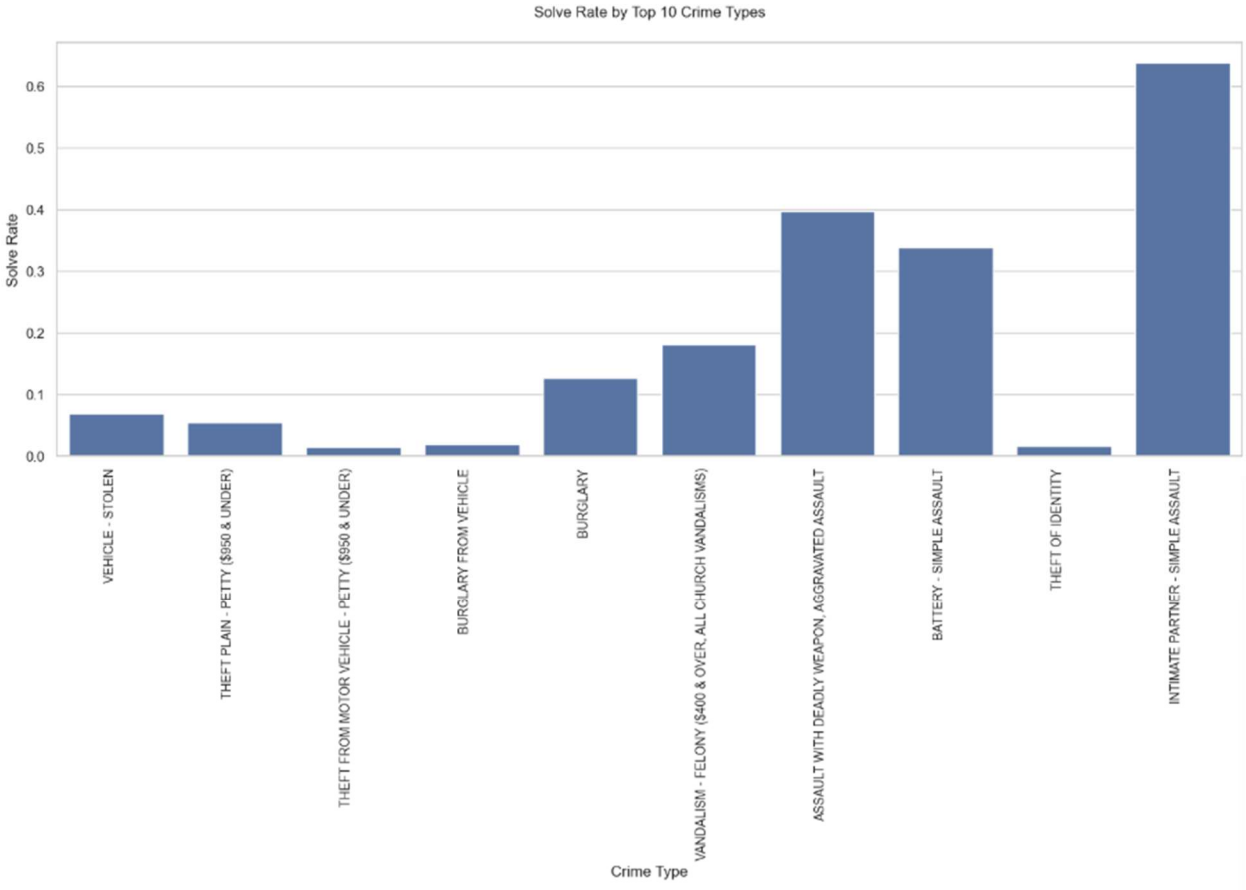
- The model, trained on historical data, may **reinforce existing inequities** if not reviewed carefully before deployment.

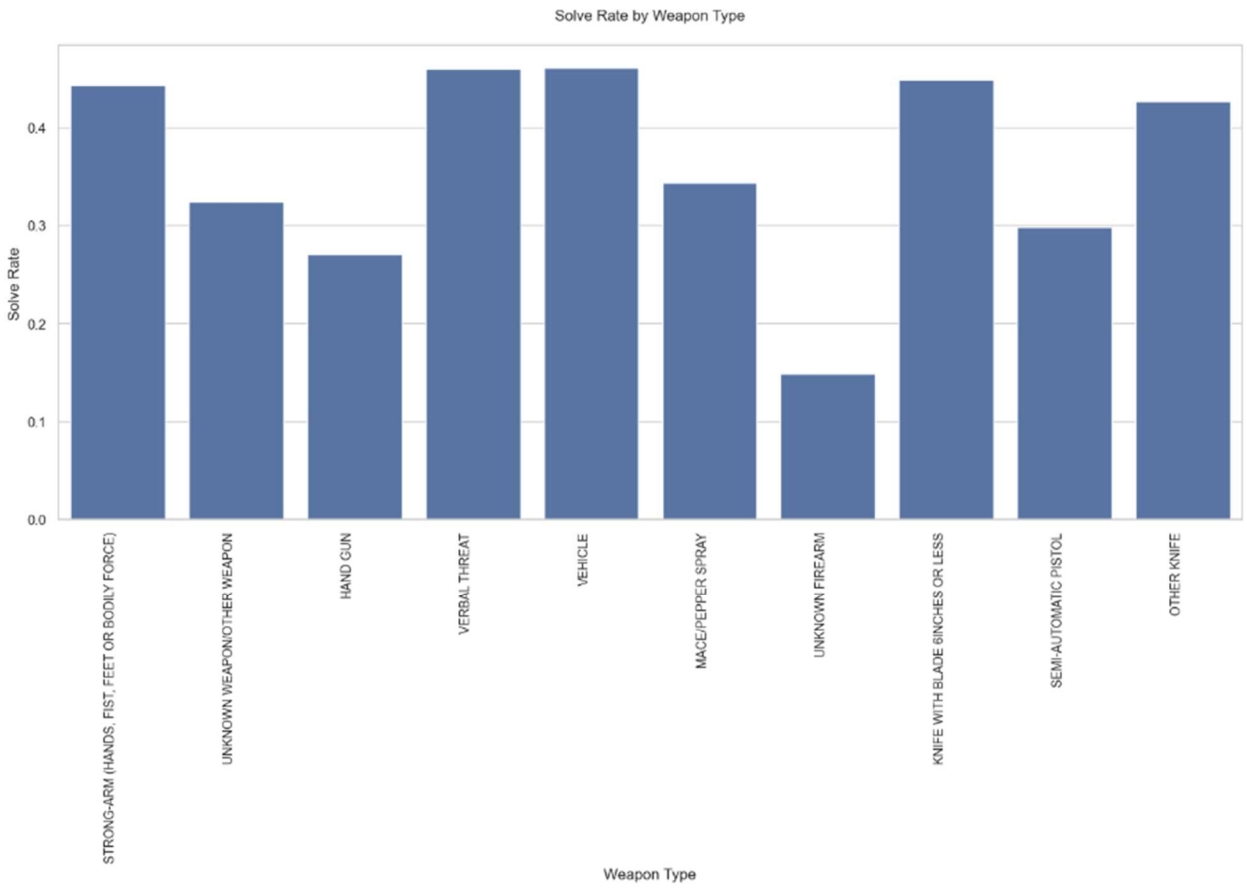
These limitations emphasize the importance of **responsible model deployment**, regular data updates, and multidisciplinary collaboration (e.g., with police officers, data engineers, and ethicists) to ensure that insights from machine learning are both impactful and equitable.

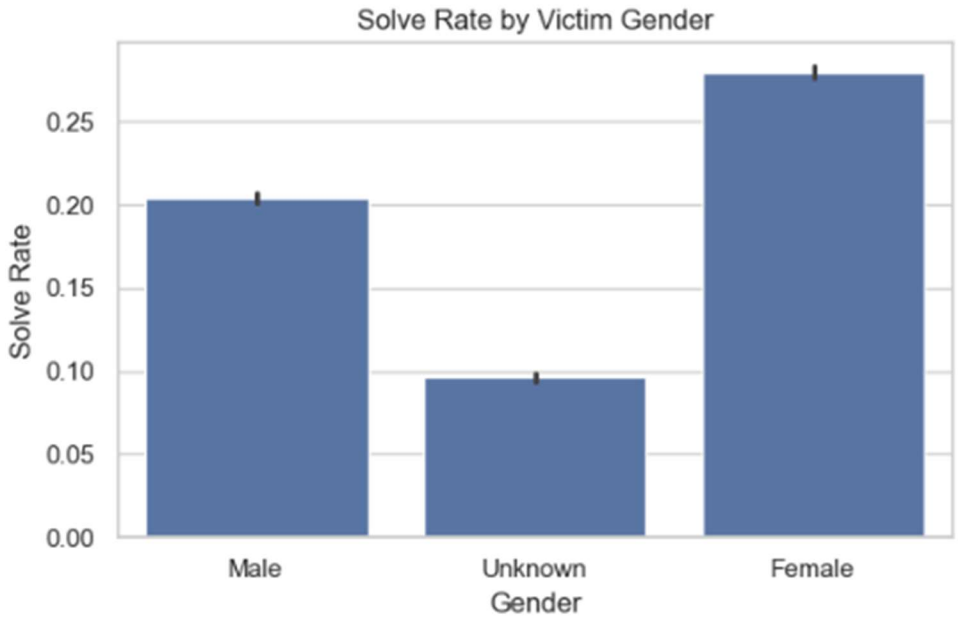
Appendix

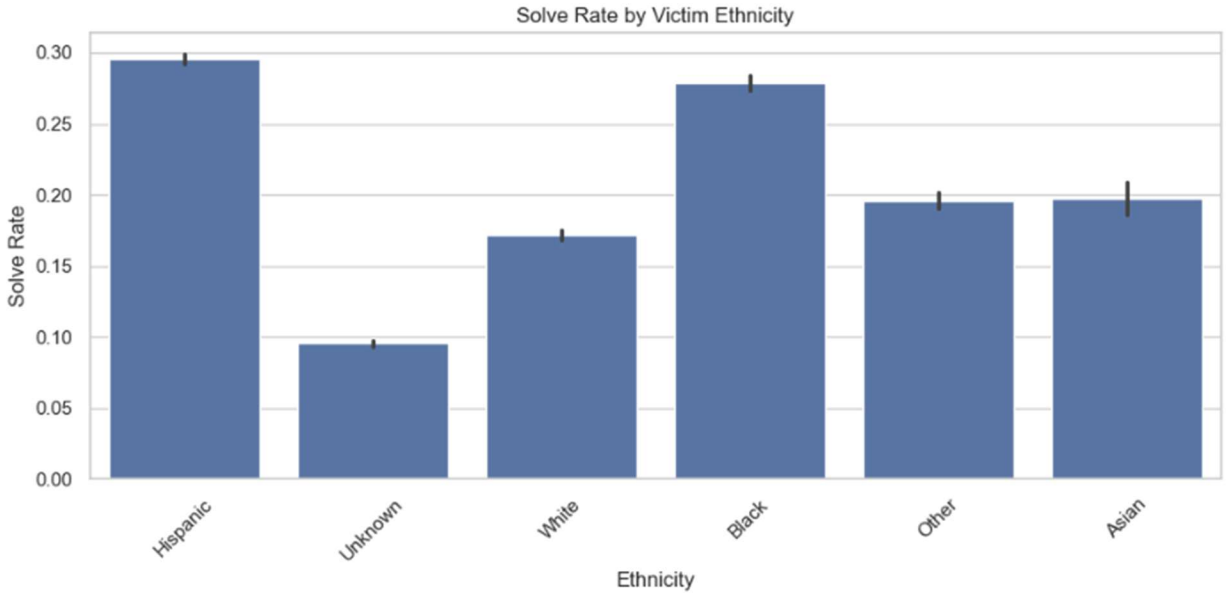


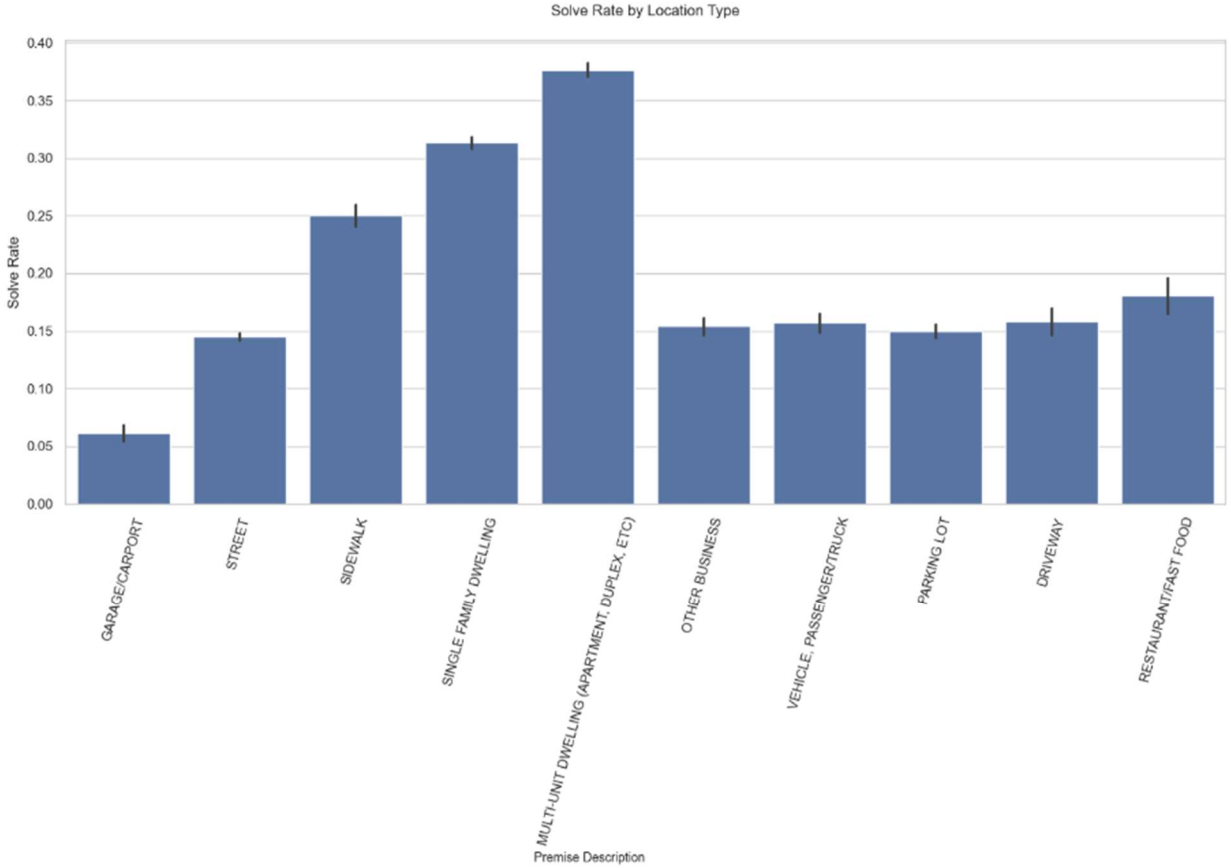

Case Solved Distribution
160000
140000
120000
100000
80000
60000
40000
20000
0
Count
Not Solved
Solved
Case Solved

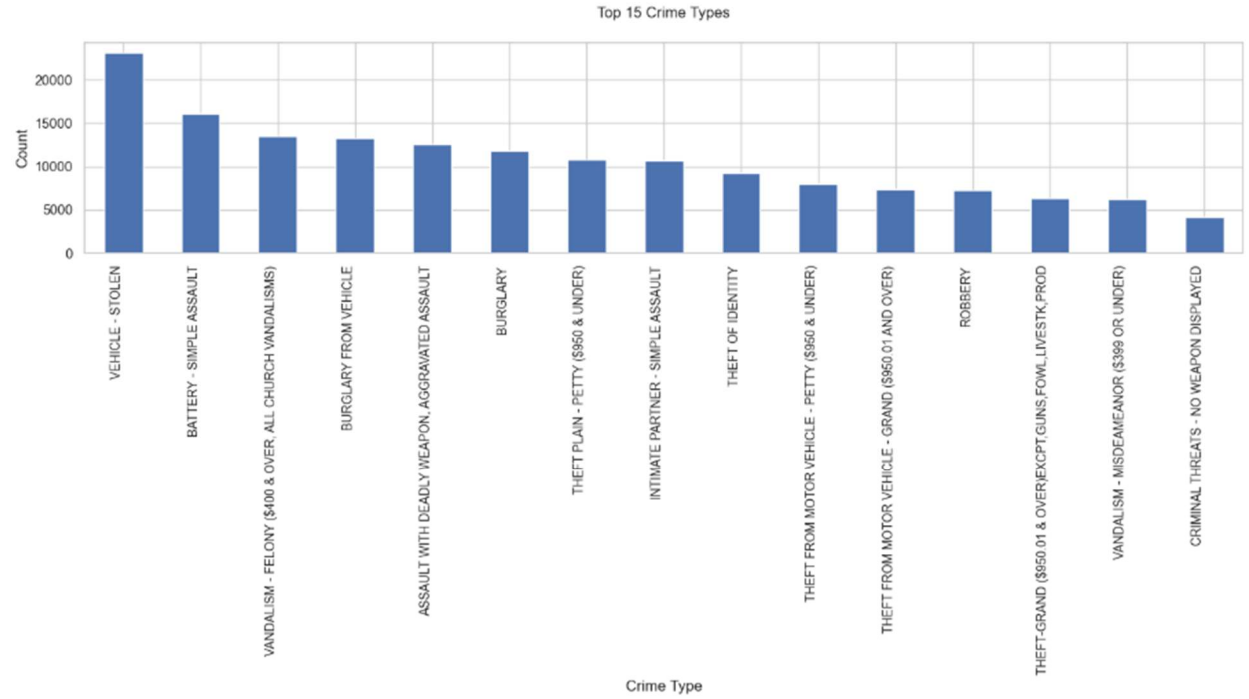

Solve Rate by Top 10 Crime Types
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Solve Rate
VEHICLE - STOLEN
THEFT PLAIN - PETTY ($950 & UNDER)
THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)
BURGLARY FROM VEHICLE
BURGLARY
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
BATTERY - SIMPLE ASSAULT
THEFT OF IDENTITY
INTIMATE PARTNER - SIMPLE ASSAULT
Crime Type

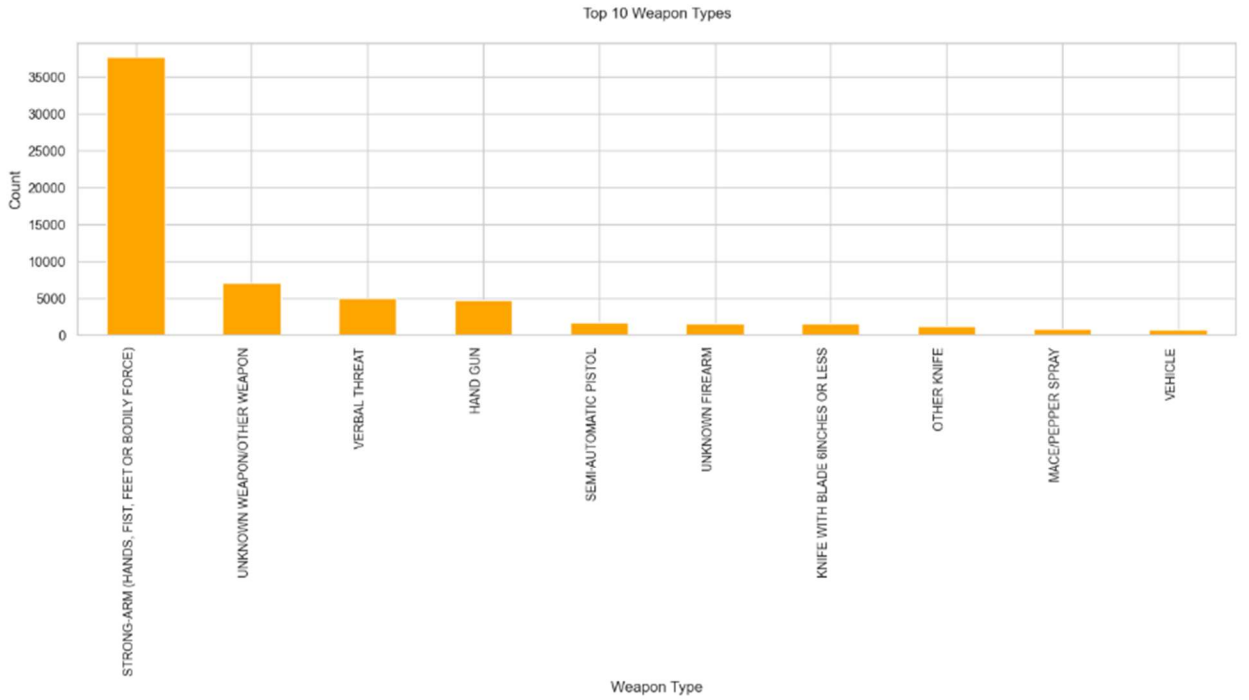

Solve Rate by Weapon Type
Solve Rate
0.0
0.1
0.2
0.3
0.4
STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)
UNKNOWN WEAPON/OTHER WEAPON
HAND GUN
VERBAL THREAT
VEHICLE
MACE/PEPPER SPRAY
UNKNOWN FIREARM
KNIFE WITH BLADE 6INCHES OR LESS
SEMI-AUTOMATIC PISTOL
OTHER KNIFE
Weapon Type

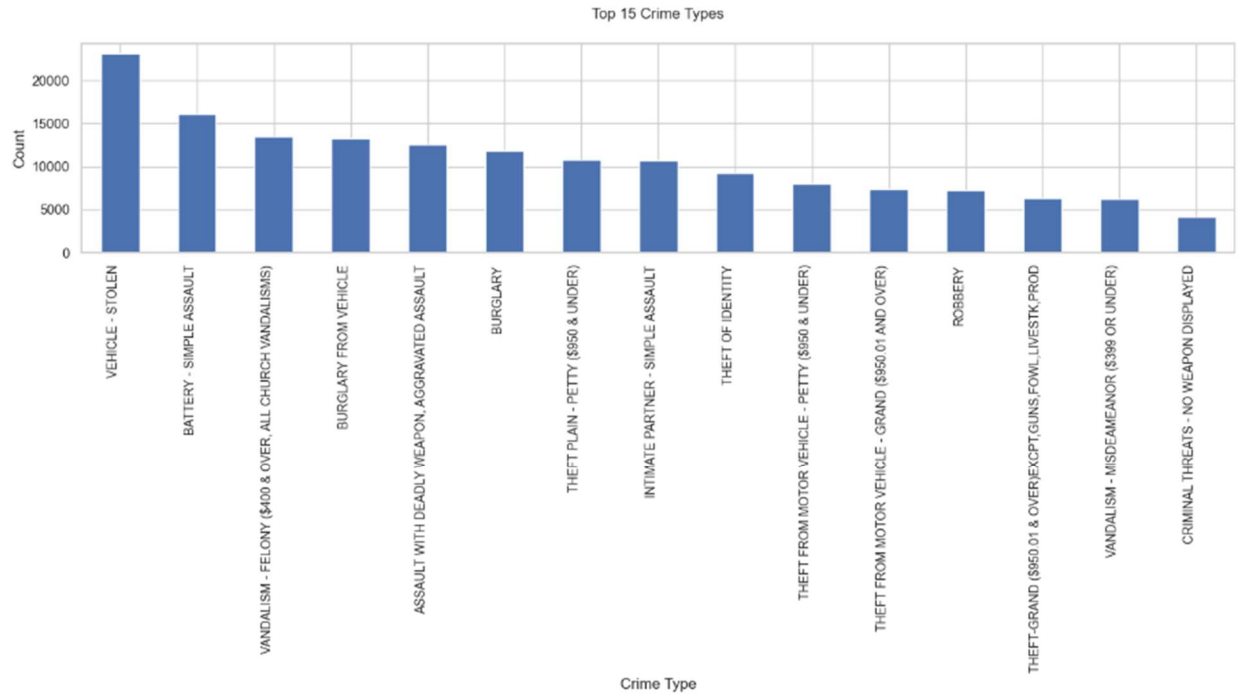

Solve Rate by Victim Gender
Solve Rate
0.00
0.05
0.10
0.15
0.20
0.25
Male
Unknown
Female
Gender

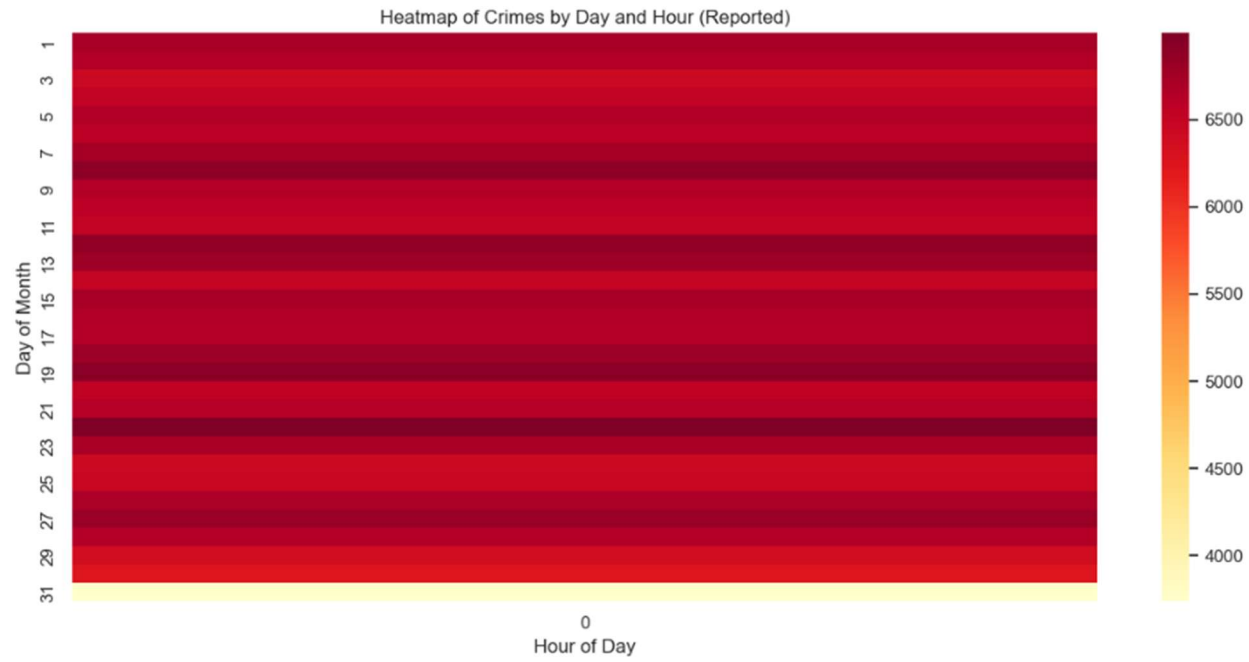

Solve Rate by Victim Ethnicity
Solve Rate
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Hispanic
Unknown
White
Black
Other
Asian
Ethnicity


Solve Rate by Location Type
Solve Rate
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
GARAGE/CARPORT
STREET
SIDEWALK
SINGLE FAMILY DWELLING
MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)
OTHER BUSINESS
VEHICLE, PASSENGER/TRUCK
PARKING LOT
DRIVEWAY
RESTAURANT/FAST FOOD
Premise Description


Top 15 Crime Types
Count
20000
15000
10000
5000
0
VEHICLE - STOLEN
BATTERY - SIMPLE ASSAULT
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)
BURGLARY FROM VEHICLE
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
BURGLARY
THEFT PLAIN - PETTY ($950 & UNDER)
INTIMATE PARTNER - SIMPLE ASSAULT
THEFT OF IDENTITY
THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)
THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER)
ROBBERY
THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD
VANDALISM - MISDEAMEANOR ($399 OR UNDER)
CRIMINAL THREATS - NO WEAPON DISPLAYED
Crime Type


Top 10 Weapon Types
Count
35000
30000
25000
20000
15000
10000
5000
0
STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)
UNKNOWN WEAPON/OTHER WEAPON
VERBAL THREAT
HAND GUN
SEMI-AUTOMATIC PISTOL
UNKNOWN FIREARM
KNIFE WITH BLADE 6INCHES OR LESS
OTHER KNIFE
MACE/PEPPER SPRAY
VEHICLE
Weapon Type


Top 15 Crime Types
Count
20000
15000
10000
5000
0
VEHICLE - STOLEN
BATTERY - SIMPLE ASSAULT
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)
BURGLARY FROM VEHICLE
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
BURGLARY
THEFT PLAIN - PETTY ($950 & UNDER)
INTIMATE PARTNER - SIMPLE ASSAULT
THEFT OF IDENTITY
THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)
THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER)
ROBBERY
THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD
VANDALISM - MISDEAMEANOR ($399 OR UNDER)
CRIMINAL THREATS - NO WEAPON DISPLAYED
Crime Type


Heatmap of Crimes by Day and Hour (Reported)
Day of Month
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
0
Hour of Day
6500
6000
5500
5000
4500
4000


Crimes by Hour of the Day (Reported)
Number of Crimes
200000
175000
150000
125000
100000
75000
50000
25000
0
0
Hour (0–23)

THE END