Understanding Crime in Los Angeles, California

Josiah Andrew Randleman

Northwest Missouri State University, Maryville MO 64468, USA
jrandl516@gmail.com and s567522@nwmissouri.edu

Abstract. Crime analysis can help us understand patterns and trends that can then be used to help improve safety and can help law enforcement properly allocate their resources to high risk areas. This analysis is aimed at Los Angeles, California and understanding high risk areas and crime trends from public crime data from 2020 to 2025. My methodologies will include Geo-spatial Crime Mapping, Temporal Crime Trend Analysis, Crime Type Categorization and Patterns. The insights generated from my analysis may be used to support law enforcement agencies in optimizing resource allocation and inform urban planning strategies aimed at improving public safety.

Keywords: Crime Analysis · Crime Hotspots · Law Enforcement Resource Allocation · Geo-spatial Mapping · Predictive Modeling · Temporal Analysis

1 Introduction

1.1 Domain

For this analysis, I am choosing the Crime Analysis Domain. For me, I work as a computer programmer / digital forensic investigator on fraud and embezzlement lawsuits. I would like to take the tools and skills that I have learned from this program and I would like to focus on crime trends, patterns, and predictive modeling and perform this on crime data from the City of Los Angeles California.

1.2 Data Problem and Significance

The data problem that I want to analyze is "What Locations and Areas in Los Angeles Are Most Prone to Specific Types of Crime?". Every area is completely different. Some areas have very low crime and then some areas have very high rates of crime. I want to analyze the various types of crimes and where they take place at. The reason why this is important and interesting is because then urban planners can invest and can improve lighting, surveillance, and can add different safety measures to areas that have more crime rates than others. This can also help aid law enforcement because they can allocate more resources to high-risk areas.

1.3 Project Implementation

The steps or phases I will perform will be Planning and Strategizing, Data Collection and Preprocessing, Exploratory Data Analysis, Predictive Modeling and Machine Learning, and then Results and Analysis.

1.4 Key Components and Limitations

My approach will be analyzing high-risk areas in Los Angeles, and my key components will include Geo-spatial Crime Mapping, Temporal Crime Trend Analysis, Crime Type Categorization and Patterns. Some limitations may include data reporting bias and time frame constraints.

2 Data Collection

2.1 Data Source

For this analysis, I will be using official data sourced from the United States Government website [3]. The title of this dataset is call "Crime Data from 2020 to Present" and is maintained by LAPD OpenData and includes crime data of Los Angeles, California. This dataset was last updated on March 19, 2025. This dataset contains 1,005,149 records of crime data from January 2020 thru March 2025 that pertains to Los Angeles, California.

2.2 Dataset Format

The Data.gov Los Angeles, California dataset is in the format of Comma Separated Values File (CSV).

2.3 Data Collection Methodology

The data was already formatted and structured as a CSV file. I was able to download it from the government website and I did not need to perform any data scraping techniques on it. The dataset was easy to access and easy to download.

2.4 Major Data Attributes

For my project, the major data attributes that I am planning to use would be DATE OCC (Date of Occurrence), TIME OCC (Time of Occurrence), AREA NAME, Crm Cd Desc (Crime Code Description), Vict Age, Vict Sex, Vict Descent, Weapon Desc, Status Desc, LAT and LON, LOCATION and Cross Street.

2.5 Data Extraction Details

I will create a crime severity categorization column. Such as grouping assaults, robberies, etc. into a violent crimes category. Crimes such as theft, burglary, vehicle theft would be in the property crimes and crimes such as vandalism and drug offenses would be in the public order crimes category. I will also look into victim profiling. I have these columns in my dataset: Vict Age, Vict Sex, Vict Descent. I can create an age group bins to group victims by age. I can also look into trends by victim sex and descent. Looking into these features can help provide valuable insights.

3 Data Curation and Cleaning

3.1 Original Dataset Overview

Before I performed any cleaning or preprocessing, the original dataset [3] consisted of **1,005,149 rows** and **28 columns**, each representing a specific aspect of a reported crime incident in Los Angeles from 2020 to the present. Each record corresponds to a unique crime report documented by the Los Angeles Police Department (LAPD).

The original columns are as follows:

- `DR_NO` – Division of Records Number, a unique identifier composed of year, area, and a sequence number.
- `Date Rptd` – The date when the crime was reported (MM/DD/YYYY).
- `DATE OCC` – The actual date the crime occurred (MM/DD/YYYY).
- `TIME OCC` – Time of occurrence in military (24-hour) format.
- `AREA` – A numerical code (1–21) representing one of LAPD's 21 geographical areas.
- `AREA NAME` – The name of the area or patrol division (e.g., 77th Street).
- `Rpt Dist No` – A 4-digit reporting district number indicating sub-regions within each area.
- `Part 1-2` – Classification of the crime as Part I or Part II offense under UCR standards.
- `Crm Cd` – Crime code, indicating the offense committed.
- `Crm Cd Desc` – Text description of the crime code.
- `Mocodes` – Modus Operandi codes indicating suspect behavior (e.g., forced entry, threatened with weapon).
- `Vict Age` – Victim's age (numerical), represented as a two-digit string.
- `Vict Sex` – Victim's gender: M (Male), F (Female), or X (Unknown).
- `Vict Descent` – Victim's race or ethnicity (e.g., W: White, H: Hispanic, B: Black, etc.).
- `Premis Cd` – Code representing the type of location where the crime occurred.
- `Premis Desc` – Text description of the premise code.
- `Weapon Used Cd` – Code representing the type of weapon used in the crime.

- `Weapon Desc` – Text description of the weapon code.
- `Status` – Case status (e.g., IC for Initial Case).
- `Status Desc` – Text description of the status code.
- `Crm Cd 1` – Primary crime code, indicating the most serious offense.
- `Crm Cd 2` – Additional, less serious offense (optional).
- `Crm Cd 3` – Additional offense (optional).
- `Crm Cd 4` – Additional offense (optional).
- `LOCATION` – Approximate address of the crime (rounded for privacy).
- `Cross Street` – Nearest cross street to the crime location.
- `LAT` – Latitude coordinate of the crime.
- `LON` – Longitude coordinate of the crime.

3.2 Data Curation Process

For this investigation, I developed a pipeline based in Python to curate and clean this data. My primary dataset [3] contains 1,005,149 records across 28 columns and comes from the Los Angeles Police Department (LAPD). Additionally, a secondary Modus Operandi (MO) reference file [2] containing 777 behavioral code descriptions was used to enhance the data with suspect behavioral patterns.

Below are my steps that I performed in my Python pipeline to curate and clean this dataset:

Step 1: Load Datasets
The first step I did was to load the LAPD crime dataset [3] and then the MO Code dictionary [2].

Step 2: Handle Missing and Blank Values
After performing an inspection of my dataset, it revealed that this dataset was missing around 5.5 million values across the 28 columns.

- All string-type columns were scanned for empty strings or whitespace using regular expressions and replaced with the placeholder `‘‘Unknown’’`.
- All numeric columns with missing values were cast to string type and similarly filled with `Unknown`, ensuring uniform formatting across the dataset.

Step 3: Remove Duplicates

Step 4: Format Dates and Extract Features

Step 5: Convert Military Time

Step 6: Standardize Victim Demographics In order to perform meaningful demographic profiling, the victim-specific fields—`Vict Age`, `Vict Sex`, and `Vict Descent`—were carefully cleaned and standardized.

Step 7: Expand MO Codes
The `Mocodes` field in the dataset represents Modus Operandi (MO) codes used by the Los Angeles Police Department to catalog specific behaviors or methods employed by suspects during the commission of a crime. These codes are numeric

and often space-separated, indicating multiple behavioral patterns associated with a single crime report.

To draw meaningful insight from this field, I downloaded the MO Crime dictionary [3] and I then imported that into my Python script. From there I created a new descriptive column that translated the MO Codes into meaningful descriptive columns.

Examples:

- A crime report with `Mocodes = "1822 1402 0344"` was expanded into:
  - `MO_Desc_1: Stranger`
  - `MO_Desc_2: Gang Involvement`
  - `MO_Desc_3: Removes Vict Property`

Step 8: Crime Category Mapping
To support high-level trend analysis, I took the `Crm Cd Desc` field and I grouped the records into six macro-level categories.

1. **Violent Crime** – Crimes involving physical harm or threat of harm to individuals.
2. **Property Crime** – Crimes involving theft or destruction of property without bodily harm.
3. **Public Order Crime** – Crimes that disrupt public peace, order, or morality.
4. **Sexual Offense** – Crimes involving sexual misconduct or assault.
5. **White Collar Crime** – Non-violent, financially motivated offenses, typically committed by individuals in professional settings.
6. **Other** – Crimes that do not fit neatly into any of the above categories or are administrative/legal in nature.

3.3 Tools and Techniques Used

For this pipeline that I created, it was done entirely in Python. Inside of this pipeline, I used a variety of libraries to curate and clean this data. The libraries that I used are as follows:

- `pandas` — This was the primary data manipulation library.
- `os` — This standard Python module handled directory creation and data export automation.

In this pipeline, I used a variety of techniques to process the data. I first load the crime dataset [3], then I load the MO code dictionary [2]. From there I then replace empty cells with "Unknown" to prevent NA. Then I check the column `DR_NO` and I drop and remove any duplicates. Examining my dataset before and after, no rows were dropped, thus leading me to conclude that all of the rows are then unique. From there, I take the dates and I parse them to date time objects. I then convert military time to standard 12-hour AM/PM format. I then create meaningful labels for my victim demographics. From there I expand the MO code field into more description columns. I then took the crime description and mapped it to a higher-level category.

3.4 Missing Value Handling Strategy

The original Los Angeles crime dataset contained a total of **5,592,494 missing values** distributed across numerous key columns.

- **Step 1: Detection of Hidden Blanks**

- **Step 2: Imputation for Categorical Fields**
  For object-type fields (e.g., `Vict Sex`, `Weapon Desc`), I filled missing cells with the constant `"Unknown"`.
- **Step 3: Imputation for Numeric Fields**
  Numeric columns like `Vict Age` were cast to strings and then I filled missing cells with the constant `"Unknown"`.
- **Step 4: Justification and Integrity**
  I chose a Placeholder-based imputation (`"Unknown"`) over methods like mean/mode to:
  - Avoid introducing artificial patterns,
  - Preserve categorical class diversity,
  - Enable later filtering or model exclusion.

Original Results:

Column Name	Missing Values
`Mocodes`	151,741
`Vict Sex`	144,765
`Vict Descent`	144,777
`Premis Cd`	16
`Premis Desc`	588
`Weapon Used Cd`	677,885
`Weapon Desc`	677,885
`Status`	1
`Crm Cd 1`	11
`Crm Cd 2`	935,996
`Crm Cd 3`	1,002,835
`Crm Cd 4`	1,005,085
`Cross Street`	850,909
Total	**5,592,494**

Table 1. Summary of Missing Values by Column

Final Results:

- **Remaining Missing Values:** 0
- **Columns Cleaned:** 13 with missing values
- **Placeholder Strategy:** Used `"Unknown"` for all categories
- **Export Compatibility:** Dataset is fully clean and model-ready

3.5 Final Dataset Dimensions

Initial Dataset Overview:

- **Records:** 1,005,149 crime reports
- **Attributes (Columns):** 28 original columns

Post-Cleaning and Feature Engineering Overview:

- **Final Record Count: 1,005,149**
  No rows were removed during the cleaning process.
- **Final Attribute Count: 42 columns**
  The attribute count increased from 28 to 42 due to additional derived fields that were created to enrich analytical possibilities. These included:
  - **Date/Time Features:**
    - `DayOfWeek` – Extracted weekday name from `DATE OCC`
    - `Month` – Extracted full month name from `DATE OCC`
    - Reformatted versions of `DATE OCC`, `Date Rptd`, and `TIME OCC`
  - **Demographic Enhancements:**
    - `Vict Age Group` – Categorical age brackets (e.g., Teen, Senior, Elderly)
    - Cleaned `Vict Sex` and expanded `Vict Descent`
  - **Modus Operandi (MO) Descriptions:**
    - `MO_Desc_1` through `MO_Desc_10` – Up to 10 behavioral descriptors per incident extracted from `Mocodes`
  - **Crime Categorization:**
    - `Crime_Category` – Macro-level classification into: Violent Crime, Property Crime, Public Order Crime, Sexual Offense, White Collar Crime, and Other

3.6 Key Data Attribute Definitions

- **DR_NO (Division of Records Number)** – A unique identifier for each crime report, combining the year, LAPD area code, and a sequential incident number.
- **DATE OCC (Date of Occurrence)** – The exact date the crime occurred.
- **TIME OCC (Time of Occurrence)** – Representing the time the crime occurred.
- **Date Rptd (Date Reported)** – The date the crime was officially reported to law enforcement.
- **AREA NAME** – A string identifier for the geographical patrol division where the crime took place (e.g., `77th Street`, `Hollywood`).
- **Crm Cd Desc (Crime Description)** – A textual descriptor of the specific crime committed (e.g., `BURGLARY FROM VEHICLE`, `RAPE, FORCIBLE`).
- **Crime_Category** – A derived field classifying crimes into six categories: Violent Crime, Property Crime, Public Order Crime, Sexual Offense, White Collar Crime, and Other.

- **Vict Age** – Age of the crime victim in years.
- **Vict Age Group** – A binned, categorical version of victim age.
- **Vict Sex** – The gender of the victim.
- **Vict Descent** – LAPD-coded race/ethnicity of the victim.
- **Mocodes** – Numeric codes describing suspect behavior.
- **LAT / LON** – Latitude and longitude coordinates of the reported crime location.

3.7 Dependent and Independent Variables

Dependent Variable:

- `Crime_Category` — It is a macro-level categorical label that assigns each crime incident to one of six high-level categories:
  1. Violent Crime
  2. Property Crime
  3. Public Order Crime
  4. Sexual Offense
  5. White Collar Crime
  6. Other

Independent Variables:

- **Temporal Features**
  - `TIME OCC` — The 12-hour formatted time when the crime occurred.
  - `DayOfWeek` — The day of the week on which the crime occurred (e.g., Monday, Tuesday).
  - `Month` — The month of the crime incident extracted from the date.
- **Geospatial Features**
  - `AREA NAME` — The LAPD division where the crime was reported.
- **Demographic Features**
  - `Vict Age` — Victim's age in years.
  - `Vict Sex` — Victim's biological sex (e.g., Male, Female, Unknown).
  - `Vict Descent` — Victim's racial or ethnic descent (e.g., Hispanic, White, Black, Other).
- **Situational and Contextual Features**
  - `Premis Desc` — Description of the premises where the crime took place (e.g., Parking Lot, Residence).
  - `Weapon Desc` — Description of the weapon used in the crime (e.g., Knife, Handgun, None).
- **Behavioral Features**
  - `MO_Desc_1` through `MO_Desc_10` — Behavioral descriptions associated with the Modus Operandi (MO) of the suspect (e.g., Armed Robbery, Gang Involvement, Entry by Force).

4 Exploratory Data Analysis (EDA)

4.1 What is Exploratory Data Analysis?

Exploratory Data Analysis is an approach where the goal is to analyze the data and then to summarize the main characteristics. This approach also helps us with finding patterns inside of our data.

4.2 Why is EDA Essential?

Exploratory Data Analysis is an essential step because it helps us to understand our data better before we make any assumptions. This approach can help us with finding where we might have outliners in our data and how our variables relate to one another.

4.3 EDA Techniques and Techniques Used

With regards to exploratory data analysis, there are a variety of techniques that may be employed. These include data inspection, summary statistics, handling missing values, data type conversion, data visualization, outliner detection and feature engineering to name a few. For me in my project, I employed a few of these techniques. I did a data inspection to check the number of rows and the types of my columns. I then did summary statistics to analyze my columns to see key insights. I also analyzed any missing values, and I replaced the empty values with 'Unknown'. From there, I performed some data visualization to analyze my distributions of some of my key attributes through histograms and bar plots. Finally, I performed feature engineering by creating new columns, such as victim age groups and crime severity categories, to add more analytical value to the dataset.

4.4 Details of Techniques and Results of EDA

In my exploratory data analysis, I performed these techniques:

- **Data Inspection** –The dataset contains 1,005,149 rows and 42 columns. The columns are mostly of the object type.

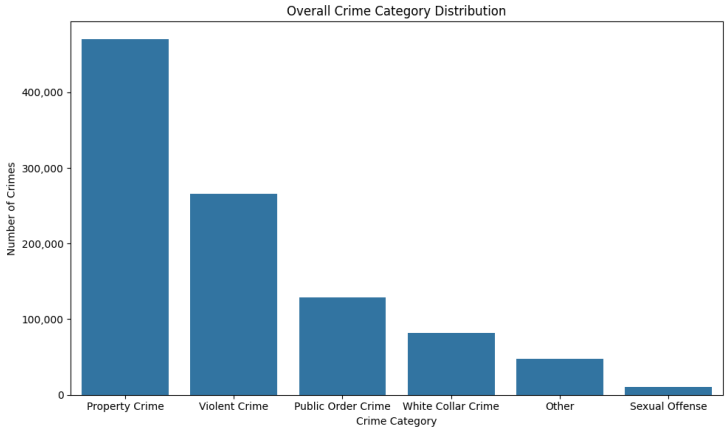
- **Summary Statistics** – In my pipeline, for each column, I print the top 5 values and the data type associated with the column.

- **Handling Missing Values** – In my dataset, I have 5,592,494 values that are missing. So, I iterate through my dataset and for each cell that is blank or missing, I fill it in with 'Unknown'.

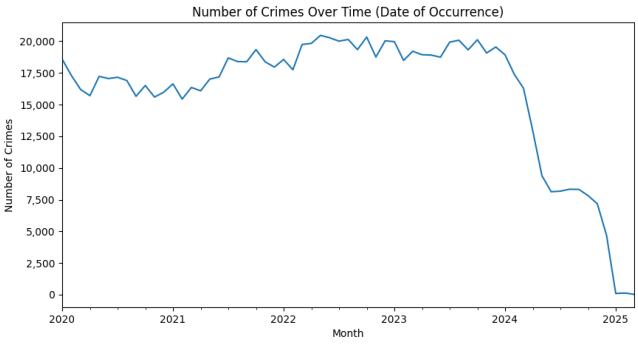
- **Feature Engineering** – From the Date OCC column, I create the columns DayOfWeek and Month. I then convert all military time to standard time. From the Vict Age column, I create the Vict Age Group by binding the ages into groups. I then take the Mocodes and I create description columns that adds the behavior of the suspect. I then create the Crime Category Column from the Crime Description. I binned these descriptor into 6 category groupings.
- **Data Visualization** - I then create various visuals to explain and show my data.

The results of the exploratory data analysis revealed several key insights:

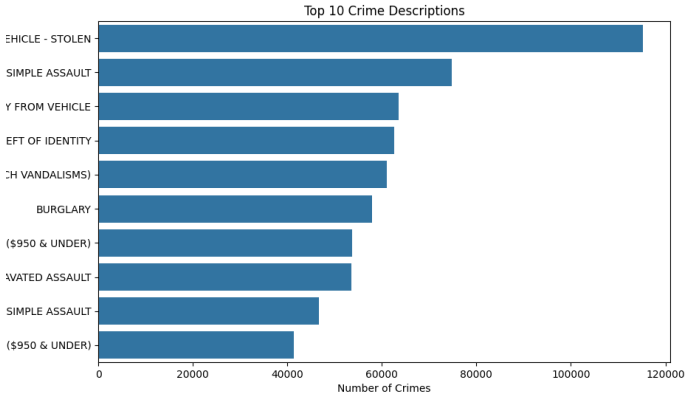
- **Overall Crime Distribution** - Property Crime was the most common crime category, followed by Violent Crime.



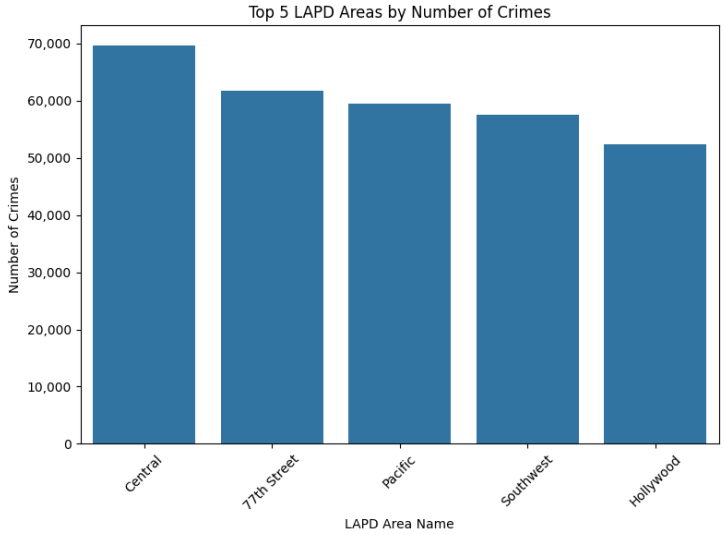
- **Crime Over Time** – Crime counts were relatively steady and consistent between 2020 and 2023, but noticeably declined in 2024 and 2025. According to the Mayor of Los Angeles, "These improvements are a direct result of strategic policing, targeted enforcement, and the invaluable collaboration with community organizations dedicated to violence prevention." [1]



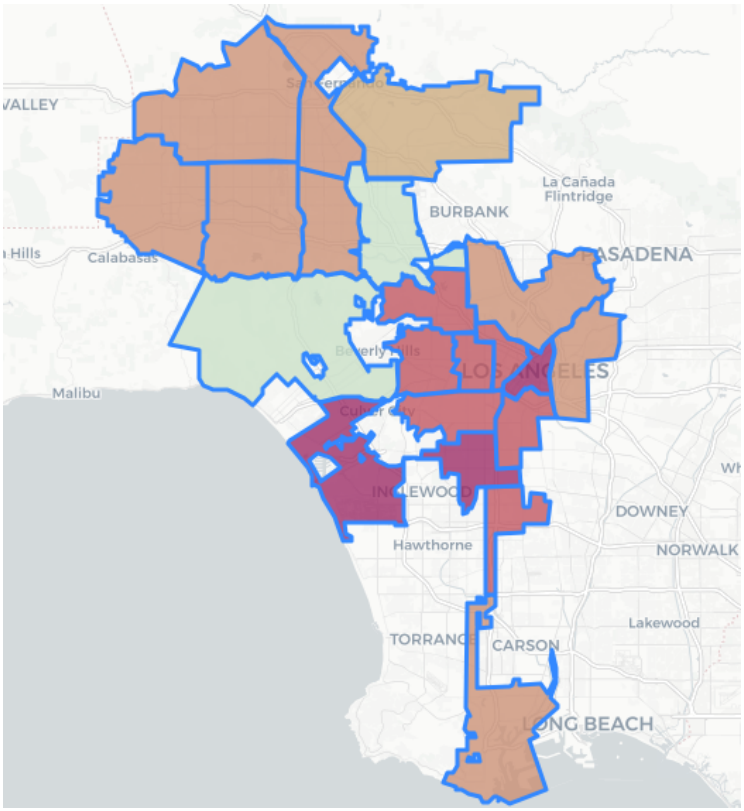
– **Top Crime Types** - "Vehicle - Stolen" and "Battery - Simple Assault" were among the most frequent crimes.



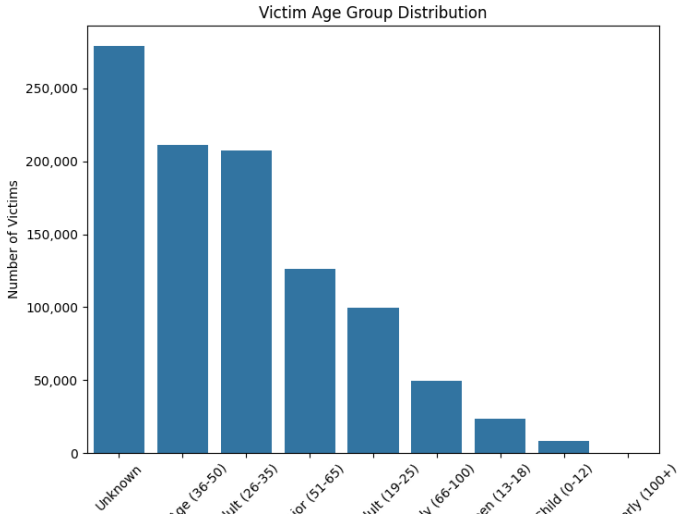
– **Top Crime Areas** - The highest number of crimes were reported in the "Central" and "77th Street" LAPD divisions.



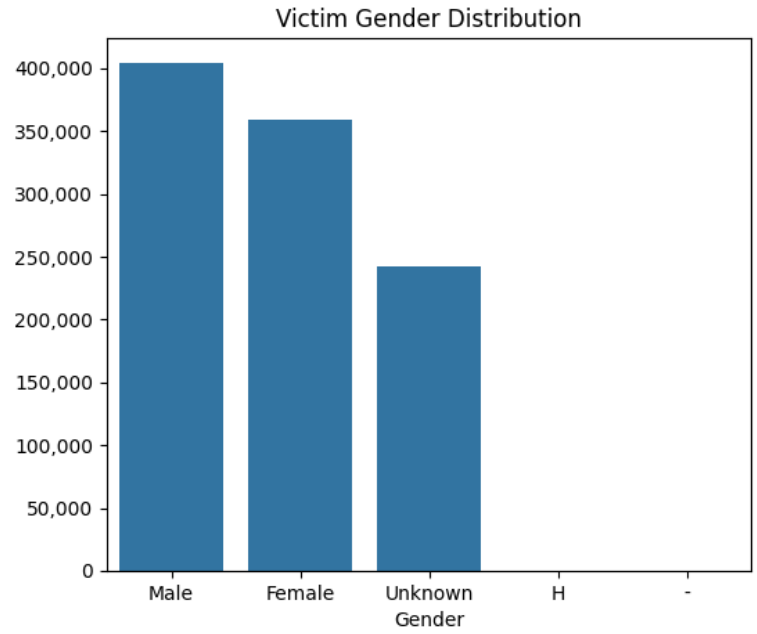
– **Crime Distribution** - This map shows the crime counts among the 21 LAPD areas. The darker shade it is the higher total crime counts there are. As we can see, the Central area has the darkest shade of red.



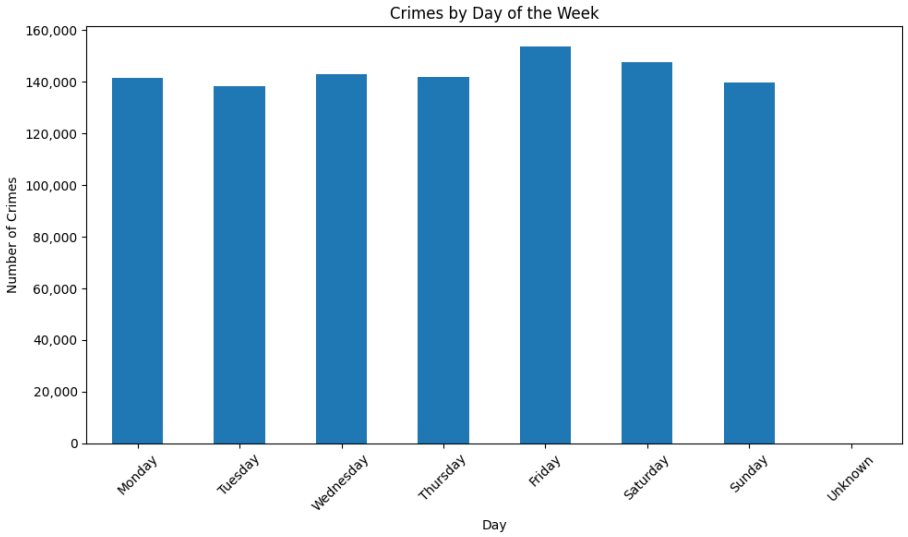
– **Victim Age** - Most victims fell into the "Adult (26–35)" and "Middle Age (36–50)" groups. However, a large portion of victim ages were recorded as "Unknown".



– **Victim Gender** - Males were slightly more frequent victims than females, but a significant number of records had unknown gender.

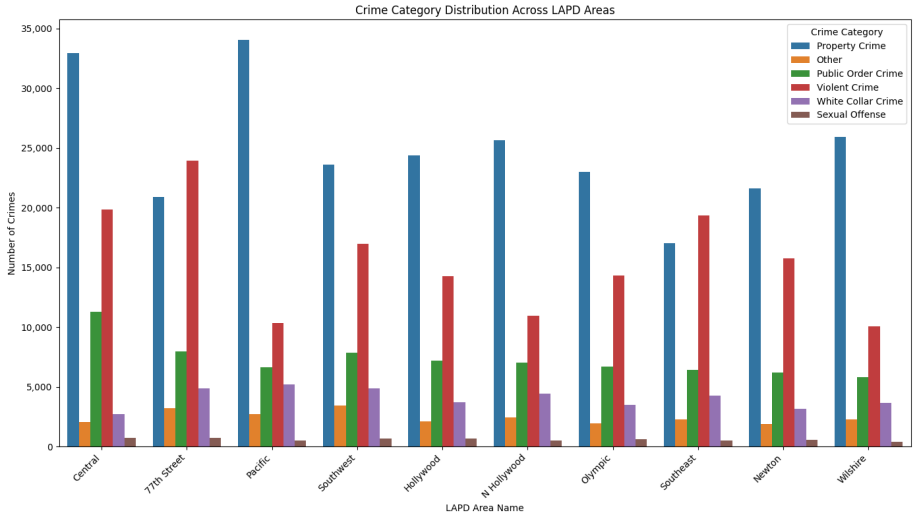


– **Crime by Day of Week** - Crimes occurred fairly evenly across weekdays, with a slight increase on Fridays and Saturdays.



– **Crime Category by Areas** - Central reported the highest overall number of crimes across all categories. Pacific had the greatest number of

Property Crimes, while 77th Street recorded the highest number of Violent Crimes. Additionally, Central led in Public Order Crimes as well.



4.5 What I Learned from the EDA Phase

For me, EDA helped me understand my data better and allowed me to draw meaningful insights from the dataset. Through the EDA process, I learned how important it is to first inspect and clean the data. I handled missing and blank values, and ensured that the date and time fields were properly formatted.

I also discovered patterns and trends within the data. For example, Property Crime was the most common crime type overall which was then followed by Violent Crime, there was a noticeable decline in reported crimes during 2024 and 2025, the top crime reported was the theft of vehicles which was then followed by simple assault. The Central division had the highest number of reported crimes. Additionally, I saw that victim demographics such as age and gender were often incomplete or unknown. The majority of victims are between the ages of 26-50. Males were also slightly more frequent victims than females. Overall, I learned a lot more about my data and how it was structured.

5 Predictive Analysis

5.1 Overview

Inside of my predictive analysis pipeline, I have designed a few different mechanisms to help me accomplish my goal. The goal for this is to design a model, that will accurately predict what the crime category will be based on various features that are collect by the police officers that pertain to the incident.

I designed these mechanisms for my predictive analysis pipeline:

- **Data Loading** – Loaded my cleaned dataset from a pickle file.
- **Data Sampling** – Randomly sampled 100,000 rows to help reduce the time complexity of my modeling.
- **Feature and Target Selection** – Selected `Crime_Category` as my dependent (target) variable. Then selected the independent variables listed in Section 4.7 as my features for my model.
- **Feature Encoding** – Applied one-hot encoding to categorical features.
- **Train-Test Split** – Then filtered 70% of my dataset into a training dataset and 30% into my testing dataset.
- **Model Training** – Then I trained a Random Forest Classifier using 100 estimators on my training dataset to predict what the `Crime_Category` will be.
- **Model Prediction** – Then used this trained model to predict what the `Crime_Category` will be in my test dataset.
- **Model Evaluation** – Then evaluated my model's performance and analyzed the accuracy metrics.
- **Model Saving** – I then save this trained Random Forest model as a pickle file.

5.2 Machine Learning Algorithm Used

For my predictive analysis, I used a Random Forest Classifier as my machine learning model to analyze my data. This model was trained on 100 estimators and the random state was set to 42.

5.3 Training and Testing Process

In the dataset, there are 1,005,149 rows. To run a Random Forest Classifier on that amount of data, would take a very long time. To save on time complexity, I randomly sample 100,000 rows from this dataset. For my training and testing datasets, I split them by 70% for my training and 30% for my testing.

5.4 Implementation and Evaluation Process

The implementation of the predictive analysis pipeline consisted of the following steps:

- **Data Loading and Sampling** — The cleaned dataset, containing 1,005,149 rows and 42 columns, was loaded from a pickle file. To optimize computational efficiency, a random sample of 100,000 rows was selected.

- **Feature and Target Preparation** — The independent variables listed in Section 4.7 were selected as features, and `Crime_Category` was selected as the dependent (target) variable. Categorical variables were encoded using one-hot encoding, resulting in a feature matrix with 5,529 features.
- **Train-Test Split** — The data was split into 70% training data and 30% testing data to ensure proper model evaluation on unseen data.
- **Model Training** — A Random Forest Classifier with 100 estimators was trained on the training dataset to learn patterns associated with different crime categories.
- **Model Prediction** — The trained model was used to predict the crime categories on the testing dataset.

The evaluation of the model's performance involved the following:

- **Accuracy Score** — The model achieved an overall accuracy of 90.93% on the testing dataset.
- **Classification Report** — Precision, recall, and F1-scores were computed for each crime category. The model performed particularly well on *Property Crime* and *Violent Crime*, while performance on *Sexual Offense* was lower, with a recall of 16%.
- **Confusion Matrix** — A confusion matrix was generated to visualize the correct and incorrect predictions across all crime categories. A normalized confusion matrix was also created to better understand the proportion of predictions relative to the true labels.
- **Visualizations** — The confusion matrix, normalized confusion matrix, and a heatmap of the classification report were generated and saved as PNG files.

The trained model was saved as a pickle file (`random_forest_model.pkl`) for future deployment and analysis.

The source code for this pipeline is available at [4].

5.5 Results of Predictive Analysis

The Random Forest Classifier achieved a strong, accurate performance on the crime category prediction task. My goal for my model, was to design a model that could predict what the crime category would be based on the various incident-related features collected by police officers.

Classification Metrics:

- Overall model accuracy: **90.93%**
- High performance was observed in categories such as:
  - **Property Crime** — Precision: 0.93, Recall: 0.97, F1-Score: 0.95
  - **Violent Crime** — Precision: 0.88, Recall: 0.99, F1-Score: 0.93
- Lower performance was observed in:
  - **Sexual Offense** — Precision: 0.93, Recall: 0.16, F1-Score: 0.27
- Macro Average Metrics:
  - Precision: 0.90, Recall: 0.72, F1-Score: 0.76

Visualizations and Interpretation of Model Performance:

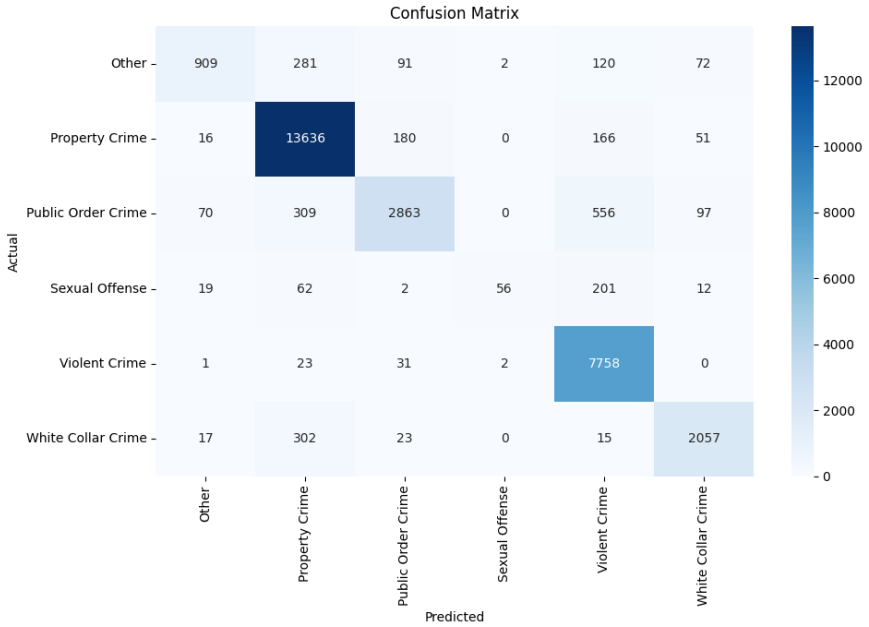


Fig. 1. Confusion Matrix of Predicted vs Actual Crime Categories

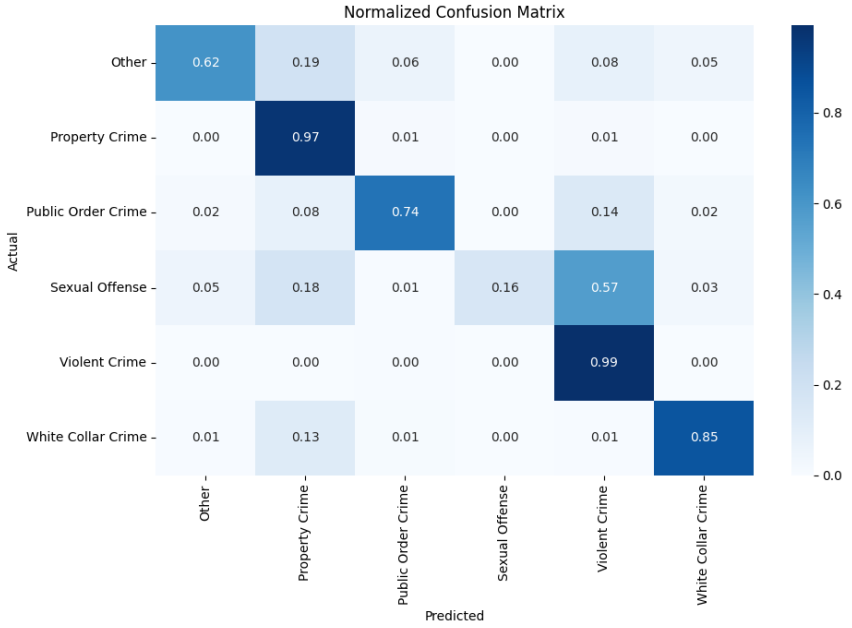


Fig. 2. Normalized Confusion Matrix

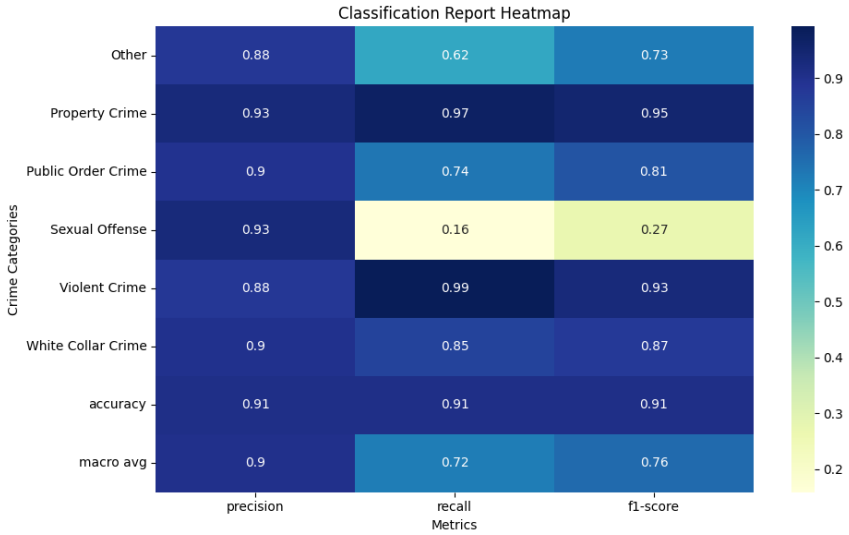


Fig. 3. Classification Report Heatmap (Precision, Recall, F1-Score)

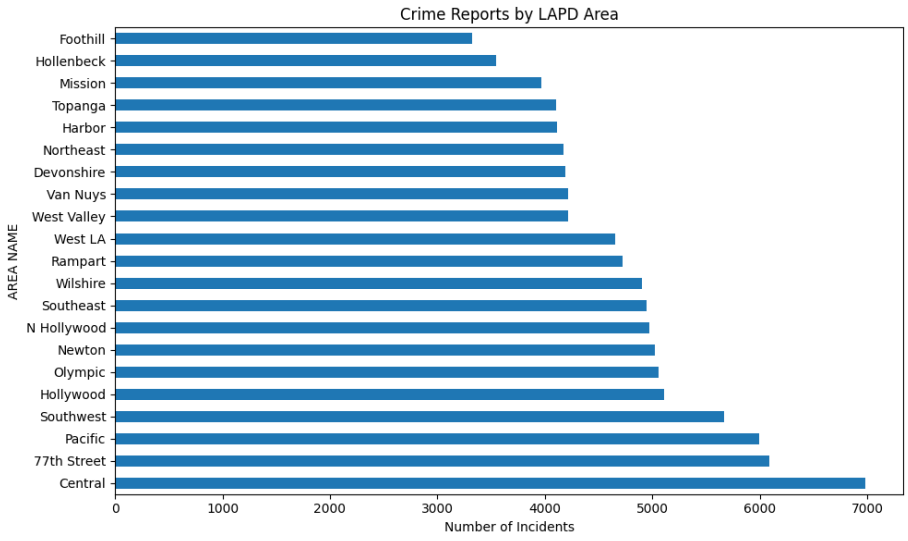


Fig. 4. Geographic Distribution of Crime Reports Across LAPD Areas

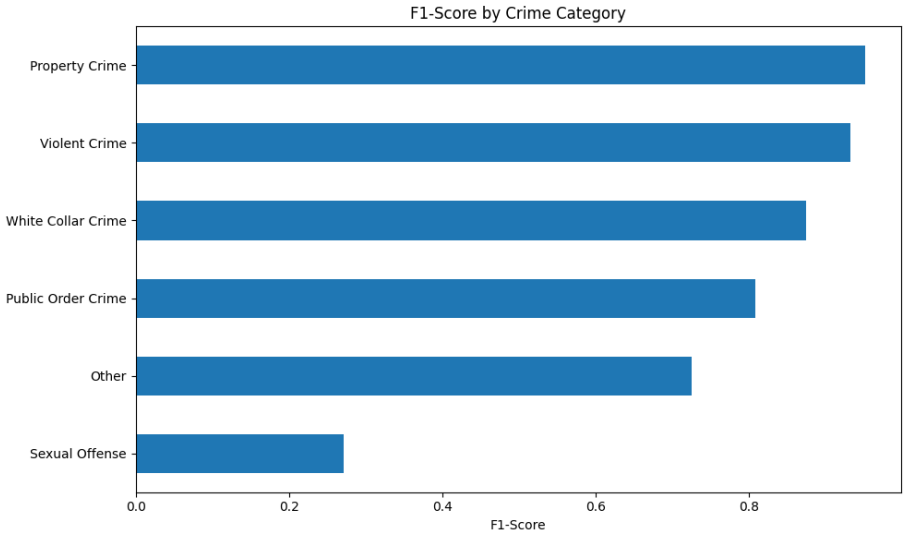


Fig. 5. F1-Score by Crime Category for Model Predictions

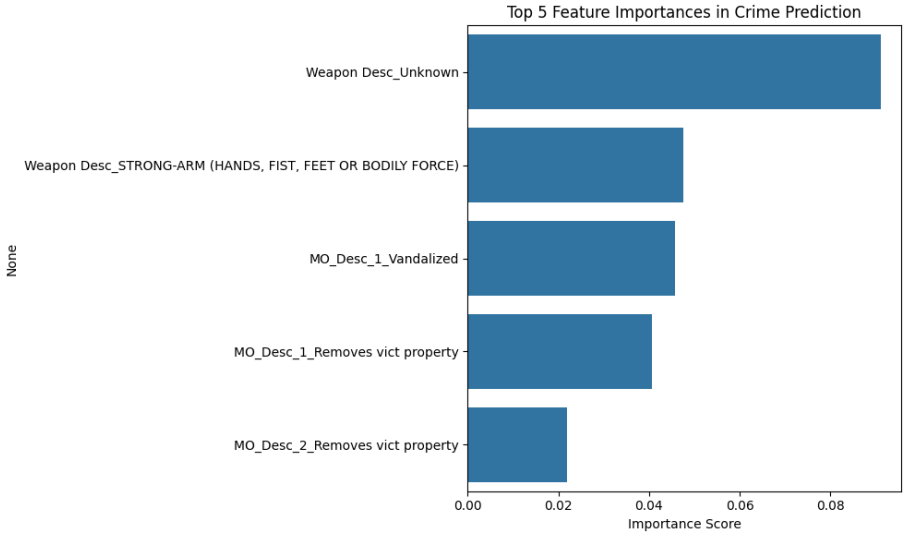


Fig. 6. Top 5 Most Important Features in Random Forest Model

5.6 Charts Used to Highlight Results

- **Confusion Matrix (Fig. 1)** – This confusion matrix chart displays the raw count of correct and incorrect predictions by crime category.
- **Normalized Confusion Matrix (Fig. 2)** – This is a normalized version of the confusion matrix that shows the percentage accuracy within each category.
- **Classification Report Heatmap (Fig. 3)** – This heatmap visualizes three key metrics (precision, recall, F1-score) across all crime categories.
- **Crime Distribution by LAPD Area (Fig. 4)** – This bar chart shows the total count of crimes reported by each area that is randomly selected into our training and test dataset.
- **F1-Score by Crime Category (Fig. 5)** – This bar chart isolates the F1-score for each crime category.
- **Top 5 Most Important Features (Fig. 6)** – This bar chart shows the top predictors used by the Random Forest classifier.

5.7 Explanation of Chart Results

The confusion matrix (Fig. 1) shows a high number of correct predictions along the diagonal, especially for categories such as Property Crime and Violent Crime. However, there are off-diagonal misclassifications, most notably with categories like Sexual Offense, which were frequently confused with Public Order Crime or Other.

The normalized confusion matrix (Fig. 2) further clarifies these findings. For categories such as Property Crime and Violent Crime, the normalized values approach 1.0 along the diagonal, indicating very strong class-specific accuracy. In contrast, Sexual Offense and White Collar Crime have lower diagonal values and higher off-diagonal spillover, suggesting poor recall for those classes despite relatively high precision.

The classification report heatmap (Fig. 3) quantifies these patterns. The model achieves high F1-scores for Property Crime (0.95) and Violent Crime (0.93), supported by strong precision and recall. However, Sexual Offense has a recall of only 0.16, despite a precision of 0.93, leading to a low F1-score of 0.27. This imbalance implies that while the model is conservative in predicting Sexual Offense, it often fails to detect it when it occurs.

The crime distribution by LAPD area (Fig. 4) contextualizes these performance differences by showing that Central and 77th Street divisions account for the majority of crime incidents in the sampled data.

The F1-score by crime category (Fig. 5) directly visualizes the disparity in model performance across crime types. It highlights that Property and Violent Crimes are predicted reliably, while Sexual Offense, White Collar Crime, and the Other category exhibit notably lower F1-scores.

The feature importance plot (Fig. 6) reveals that the most predictive features include Weapon Description, Modus Operandi (MO) descriptors, and AREA NAME.

5.8 Inferences from the Data

The model performed exceptionally well in predicting high-frequency crime categories such as Property Crime and Violent Crime. This suggests that these types of crimes have strong, consistent patterns within the dataset.

Crime categories with lower overall frequency or more ambiguous classification — such as Sexual Offense, White Collar Crime — were harder for the model to predict accurately.

The strong predictive power of features like Weapon Description, Area Name, and specific MO codes indicates that the model is heavily reliant on contextual and behavioral variables rather than victim demographics alone. This supports the hypothesis that where a crime happens and how it is committed (e.g., armed robbery, forced entry) is often more predictive than who the victim is.

The inferences drawn suggest that predictive modeling can complement traditional crime analysis, especially for frequent, well-defined offenses.

5.9 Statistical Conclusions

The Random Forest classifier achieved an overall accuracy of **90.93%**, indicating a strong ability to correctly classify the majority of crime incidents into their appropriate categories.

The macro-averaged performance metrics further validate the model's effectiveness:

- **Precision:** 0.90 – The model was correct 90% of the time when it predicted a given crime category.
- **Recall:** 0.72 – On average, the model correctly identified 72% of the true instances across all categories.
- **F1-Score:** 0.76 – The balance between precision and recall demonstrates moderate to strong classification performance.

High individual category performance was observed for:

- **Property Crime:** F1-score of 0.95 (Precision: 0.93, Recall: 0.97)
- **Violent Crime:** F1-score of 0.93 (Precision: 0.88, Recall: 0.99)

These results indicate that the model is especially effective at identifying and classifying more common and well-defined crime types.

In contrast, lower performance was noted in underrepresented or ambiguous categories:

- **Sexual Offense:** F1-score of 0.27 (Precision: 0.93, Recall: 0.16)

This disparity points to the model's difficulty in detecting less frequent crimes, which may require additional feature engineering to improve.

In conclusion, the statistical evidence suggests that the model is highly effective for frequent and structured crime categories, but additional refinement is needed to ensure equitable performance across all classes.

5.10 General Observations and Insights

Overall, the predictive model performed well and revealed several useful insights. Areas such as Central and 77th Street had the highest number of crime reports, which helped the model learn strong patterns from those divisions.

The features that most influenced the model's predictions were weapon descriptions, MO codes, and area names. This shows that how and where a crime happens is often more useful in predicting crime type than who the victim is.

While the model did better on common crimes like Property and Violent Crime, it had some difficulty with less frequent categories like Sexual Offense. This is likely due to fewer examples in the training data. Still, the high precision in those categories shows the model was careful about its predictions.

In general, this analysis showed that predictive modeling can be a strong tool for understanding crime trends and helping law enforcement focus on high-risk areas. With more balanced data and continued refinement, the model could become even more useful in real-world applications.

6 Results

The analysis revealed several important insights into crime trends, patterns, and predictability in Los Angeles between 2020 and 2025.

1. Crime Category Distribution: The most frequently occurring crime categories were *Property Crime* and *Violent Crime*, accounting for the majority of reported incidents. Specifically, theft-related crimes such as "Vehicle - Stolen" and assault-related crimes such as "Battery - Simple Assault" were the most common individual offenses. Together, these two categories represented a large proportion of total incidents across all LAPD divisions.

2. Geographic Distribution of Crime: Crime was not evenly distributed across the city. The *Central* and *77th Street* LAPD divisions reported the highest number of crimes overall. Central led in *Public Order Crimes*, while 77th Street had the highest counts of *Violent Crimes*. The *Pacific* division had a notable concentration of *Property Crimes*. These trends were visualized using choropleth maps and bar charts showing total incidents per division.

3. Temporal Patterns: The frequency of reported crimes remained relatively stable from 2020 through 2023, followed by a noticeable decline in 2024 and early 2025. Crime counts were slightly elevated on Fridays and Saturdays, suggesting a potential correlation with social or leisure activity.

4. Victim Demographics: Victim profiling showed that adults between the ages of 26 and 50 were the most common victims. Males were slightly more represented than females in reported incidents, although over 140,000 records lacked gender identification.

5. Predictive Model Performance: A Random Forest Classifier trained on a 100,000-row sample achieved an accuracy of **90.93%**. The model's best performance was on:

- **Property Crime** – Precision: 0.93, Recall: 0.97, F1-Score: 0.95

– **Violent Crime** – Precision: 0.88, Recall: 0.99, F1-Score: 0.93

The model struggled with underrepresented classes:

– **Sexual Offense** – Precision: 0.93, Recall: 0.16, F1-Score: 0.27

This suggests that while the model was cautious in predicting sensitive categories (high precision), it often failed to identify them correctly (low recall), likely due to the limited number of training samples for those categories.

6. Feature Importance: The most predictive features identified by the Random Forest model were:

– *Weapon Description*
– *Modus Operandi (MO) Descriptors*
– *LAPD Area Name*

This confirms that the nature of the crime (how it was committed) and its location were more indicative of crime category than victim demographics alone.

7. Model Interpretability: Visual diagnostics such as confusion matrices and classification heatmaps helped identify performance disparities across categories. While the model achieved high diagonal dominance for common classes, spillover into the *Other* and *Public Order Crime* categories was observed for harder-to-classify crimes.

8. Data Visualization: Over a dozen charts were created to support interpretation, including:

– F1-Score by Crime Category
– Crime Reports by Division
– Crime Distribution by Day of Week and Month
– Choropleth Maps and Heatmaps of Crime Incidents

These visualizations collectively reinforced the findings and provided individuals with intuitive tools for exploring high-risk areas and crime dynamics.

In summary, the results validate the feasibility of using data-driven approaches to understand and predict crime in large metropolitan areas. High-frequency crimes are predictable with strong accuracy, while low-frequency or sensitive crimes require further data enrichment or alternative modeling strategies.

7 Limitations

Several limitations were present in the dataset and the modeling process:

– **Missing Data:** Over 5.5 million missing values were identified, particularly in key fields such as Weapon Description and Victim Demographics. These were filled using a placeholder strategy ("Unknown"), which may have masked meaningful patterns.

- **Class Imbalance:** Certain crime categories like Sexual Offense and White Collar Crime were significantly underrepresented, which negatively affected the model's recall and generalizability for these classes.
- **Reporting Bias:** The dataset reflects only reported crimes. Many crimes, especially sensitive ones like sexual offenses or domestic violence, may go unreported, leading to incomplete insights.
- **Time Constraints:** For model efficiency, only a sample of 100,000 records (out of over one million) was used for training, which could limit the full potential of the model.
- **Simplified Categorization:** Grouping over 100 unique crime descriptions into six macro-level categories provided clarity but may have oversimplified nuanced distinctions between crime types.

8 Conclusions

This project successfully demonstrated how public crime data can be transformed into actionable insights through the use of data engineering, geospatial analysis, and machine learning. By analyzing over one million crime records from Los Angeles, we identified high-risk areas, prevalent crime types, and key victim demographics.

The Random Forest Classifier achieved a high overall accuracy of 90.93%, proving especially effective at predicting common crime categories such as Property and Violent Crime. These findings highlight the potential of predictive modeling to support strategic decision-making in law enforcement, such as resource allocation and patrol planning.

While the model faced challenges with rare and under reported crimes, the approach still represents a strong foundation for more sophisticated, real-world applications. Future work could involve integrating real-time data streams, improving model fairness across all categories, and collaborating directly with law enforcement agencies to translate findings into policy recommendations.

In sum, this project provides a scalable and interpretable framework for using data science to better understand crime dynamics in urban environments.

9 GitHub Repository and Reproducibility

All scripts, links to datasets, model outputs, and visualizations used throughout this project have been organized and published in a public GitHub repository for full transparency and reproducibility. The repository is located at [4].

References

1. Bass, K.: Lapd releases 2024 end-of-year crime statistics for the city of los angeles (2024), https://mayor.lacity.gov/news/lapd-releases-2024-end-year-crime-statistics-city-los-angeles#:~:text=%E2%80%9CA%2014%25%20reduction%20in%20homicides,organizations%20dedicated%20to%20violence%20prevention.

2. City of Los Angeles: MO Codes Numerical Reference PDF. https://data.lacity.org/api/views/d5tf-ez2w/files/8957b3b1-771a-4686-8f19-281d23a11f1b?download=true&filename=MO_CODES_Numerical_20180627.pdf
3. Los Angeles Police Department: Crime Data from 2020 to Present. https://catalog.data.gov/dataset/crime-data-from-2020-to-present
4. Randleman, J.: https://github.com/jrandl/Capstone-Project