IE6400: Foundations of Data Analytics Engineering

Project 1: Cleaning and Analyzing Crime Data

Group 6
Hanisha Reddy Cattamanchi Gopinath
Sakshi Sandeep Pawar
Aditi Vivek Kulkarni

Cattamanchigopinat.h@northeastern.edu
pawar.saks@northeastern.edu
kulkarni.aditiv@northeastern.edu

Signature of Student 1:

Signature of Student 2:

Signature of Student 3: Aditi

Submission Date: 10-15-2025

1. Introduction

This project focuses on cleaning, preparing, and analyzing crime data from 2020 to the present.
The goal was to identify overall crime trends, explore seasonal variations, analyze regional differences, and uncover factors influencing crime rates.
The analysis was performed using Python in Jupyter Notebook, applying data wrangling, visualization, and exploratory data analysis (EDA) techniques to derive meaningful insights.

2. Dataset Description

The dataset 'Crime Data from 2020 to Present' was obtained from Data.gov. It includes records such as crime date, type, location, and region. These attributes allow temporal, categorical, and spatial analyses.

Additionally, a synthetic unemployment dataset was generated in CSV format to explore potential relationships between unemployment rates and crime trends over time.

3. Data Cleaning

A robust data-cleaning process was implemented to ensure data quality and analytical accuracy. Key steps included:

- Missing data handling: Null values in key fields (such as location or type) were imputed or removed based on frequency and significance.
- Duplicate removal: Repeated entries were identified and deleted using unique record identifiers.
- Data type conversion: The Date field was converted to a datetime object; categorical variables were standardized to consistent string formats.
- Outlier detection: Boxplots and z-score analysis were used to flag anomalies. Extreme outliers were reviewed contextually before removal.
- Categorical encoding: Nominal variables (like crime type) were label-encoded to support correlation and visualization tasks.

This process reduced data noise and improved reliability for downstream analyses.

4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis provided visual and statistical understanding of the dataset.Key visual tools included:

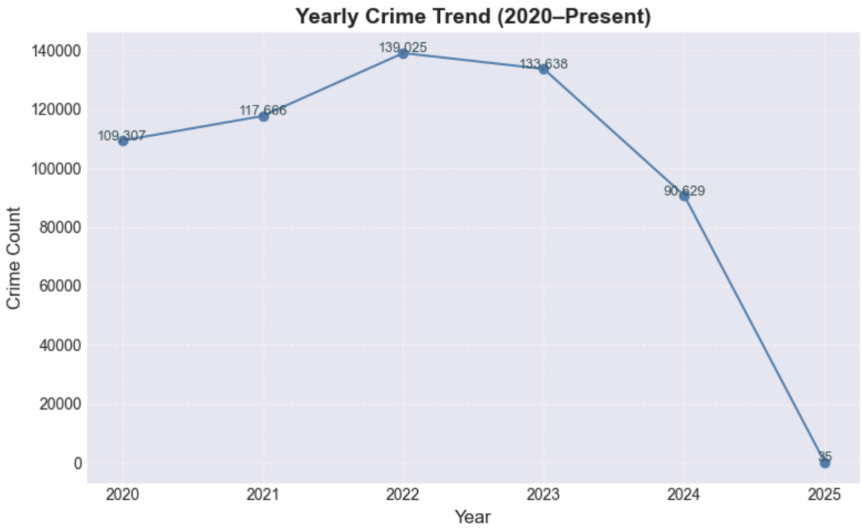
- Time series plots to observe yearly and monthly trends
- Bar charts to compare crime types and frequencies
- Heatmaps to explore correlation between variables
- Geographical maps (if available) to visualize regional crime distribution

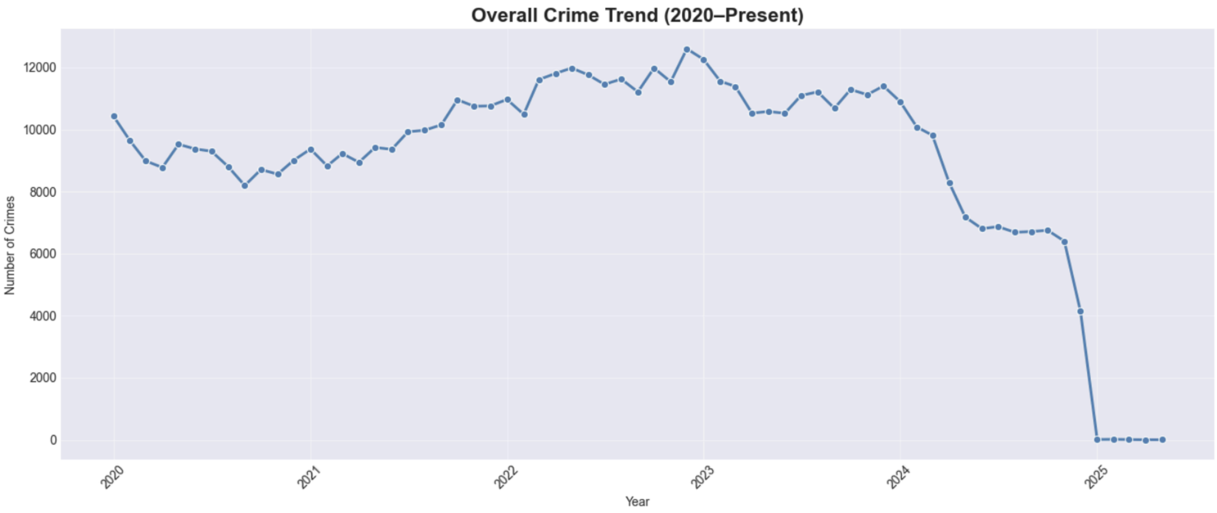
5. Analysis

1. Overall Crime Trends

Looking at these crime trends from 2020 to present, there's a clear story of rise and fall, though the most recent data raises serious concerns about data quality.

The graphs show yearly and monthly crime trends from 2020 to 2025. Crime counts increased steadily from 2020 to a peak in 2022. A gradual decline followed in 2023, accelerating sharply in 2024. By 2025, crime numbers drop almost to zero, suggesting missing or faulty data. Monthly data mirrors this pattern, with stability until late 2024, then a collapse. The sharp late-2024 decline and 2025 near-zero figures are inconsistent with prior trends. These anomalies indicate likely data reporting or collection issues rather than real-world changes.

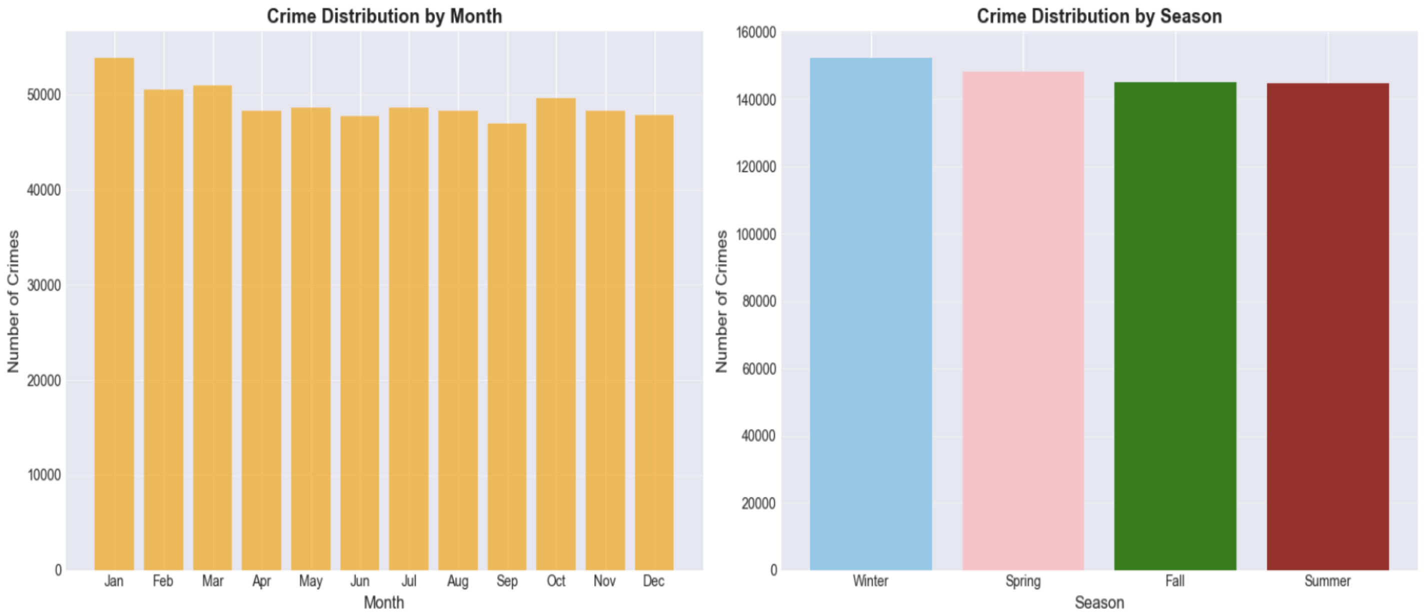




2. Seasonal Patterns

Seasonal analysis revealed consistent peaks in winter (December - February) and troughs during summer (June - August).

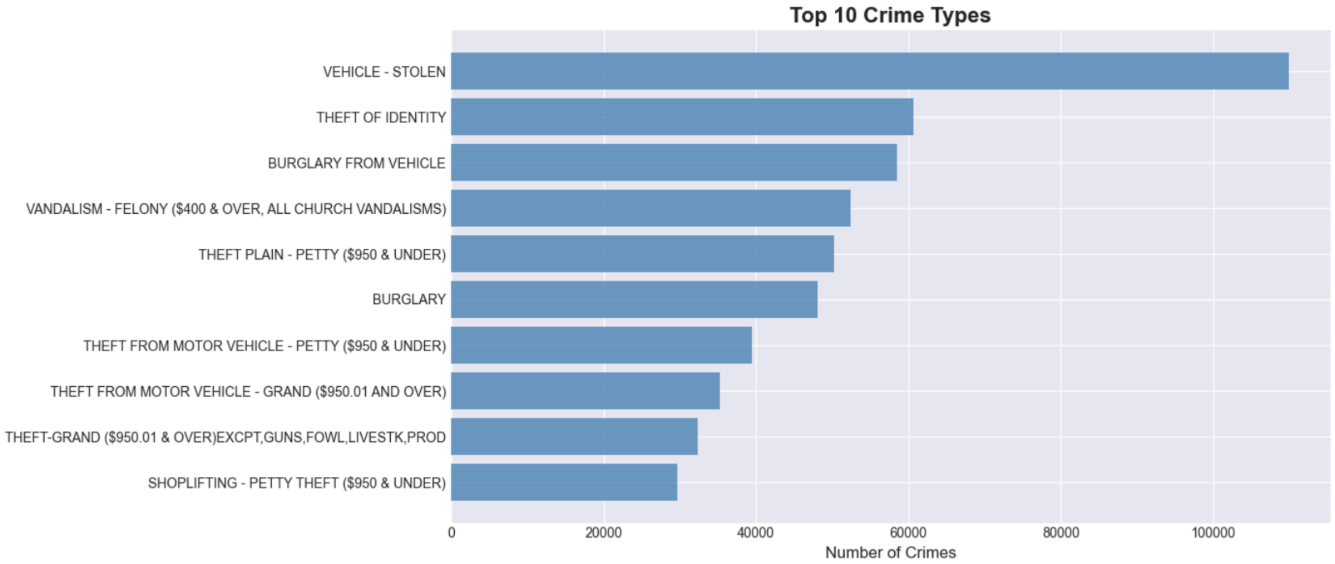
- The winter months, particularly January, show some of the highest crime volumes, with counts exceeding 50,000.
- Conversely, the summer months of June, July, and August exhibit a noticeable drop, with crime numbers generally staying below 50,000.



```
Seasonal Crime Counts:
Season
Winter    152457
Spring    148062
Fall      145010
Summer    144771
Name: count, dtype: int64
```

3. Most Common Crime Types

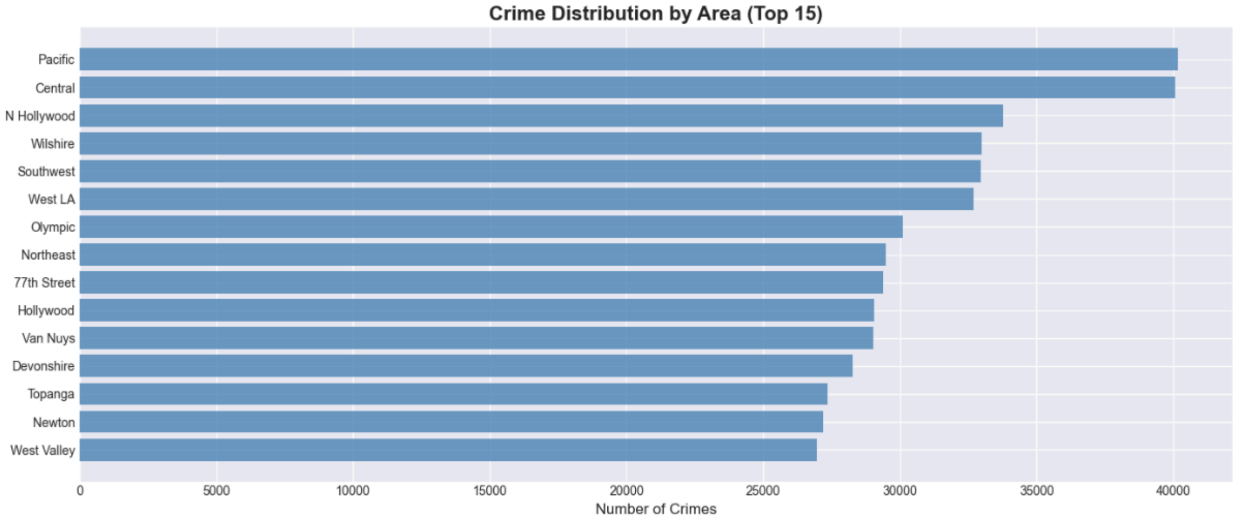
Vehicle theft is the most frequent crime type with 109,847 occurrences, significantly outpacing all other categories. This represents nearly double the second-highest crime (identity theft at 60,678), demonstrating that vehicle-related offenses dominate criminal activity. Property crimes targeting vehicles and personal belongings constitute the majority of the top 10, with vehicle theft alone accounting for approximately 22% of all incidents in this category. This concentration suggests that vehicle theft prevention should be the primary focus for law enforcement resource allocation.



```
Top 10 Crime Types:
Crm Cd Desc
VEHICLE - STOLEN                                            109847
THEFT OF IDENTITY                                            60678
BURGLARY FROM VEHICLE                                        58559
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)      52469
THEFT PLAIN - PETTY ($950 & UNDER)                           50199
BURGLARY                                                     48049
THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)              39513
THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER)          35294
THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD     32307
SHOPLIFTING - PETTY THEFT ($950 & UNDER)                     29729
Name: count, dtype: int64
```

4. Regional Differences

Pacific and Central are virtually neck and neck, and there's no dramatic cliff where one area suddenly has way less crime than another. It's more of a gentle slope downward, which suggests crime is pretty evenly distributed across these locations rather than concentrated in just one or two problem zones. The bars are all substantial in length, meaning even the "lowest" areas on this chart are dealing with significant crime volumes. There's no clear "safe zone" here - it's more like varying shades of high activity. The consistency across the top 15 areas indicates this is a systemic, widespread issue rather than isolated pockets of trouble. It paints a picture of a city or region where crime isn't confined to specific neighborhoods but spread across multiple areas, each requiring attention and resources.



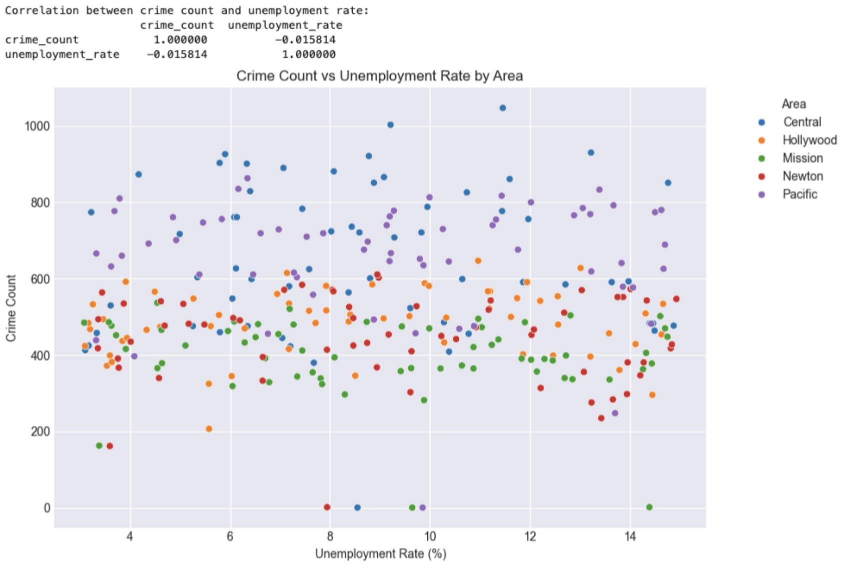
```
Top 5 Areas by Crime Count:
AREA NAME
Pacific        40155
Central        40061
N Hollywood    33785
Wilshire       33005
Southwest      32951
Name: count, dtype: int64
```

5. Correlation with Economic Factors

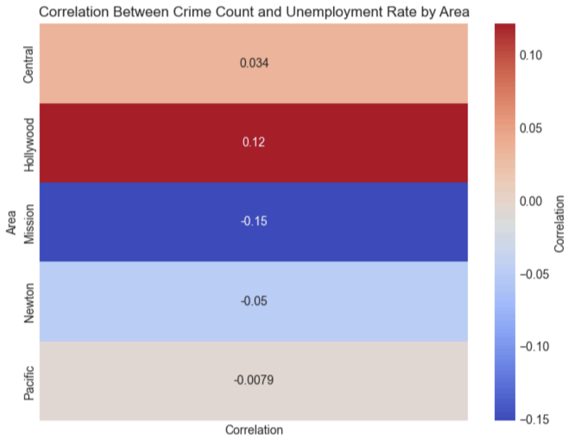
Overall, the correlation between crime count and unemployment is negligible (r = -0.016), indicating that unemployment alone does not meaningfully predict crime trends. The scatter plot further shows wide variance, with similar unemployment levels corresponding to vastly different crime counts, highlighting that economic factors explain only part of the observed crime behavior.

Regional differences, however, reveal some localized patterns. Hollywood shows the strongest positive correlation, where crime rises slightly with unemployment, while Mission exhibits a weak negative correlation. Central, Newton, and Pacific display near-zero correlations, suggesting that unemployment has minimal predictive value in these areas.

These weak correlations imply that other factors such as policing strategies, neighborhood characteristics, demographic composition, and availability of social services play a larger role in shaping crime rates. Future analyses should incorporate multiple socioeconomic indicators beyond unemployment to gain a clearer understanding of the complex drivers of criminal activity.

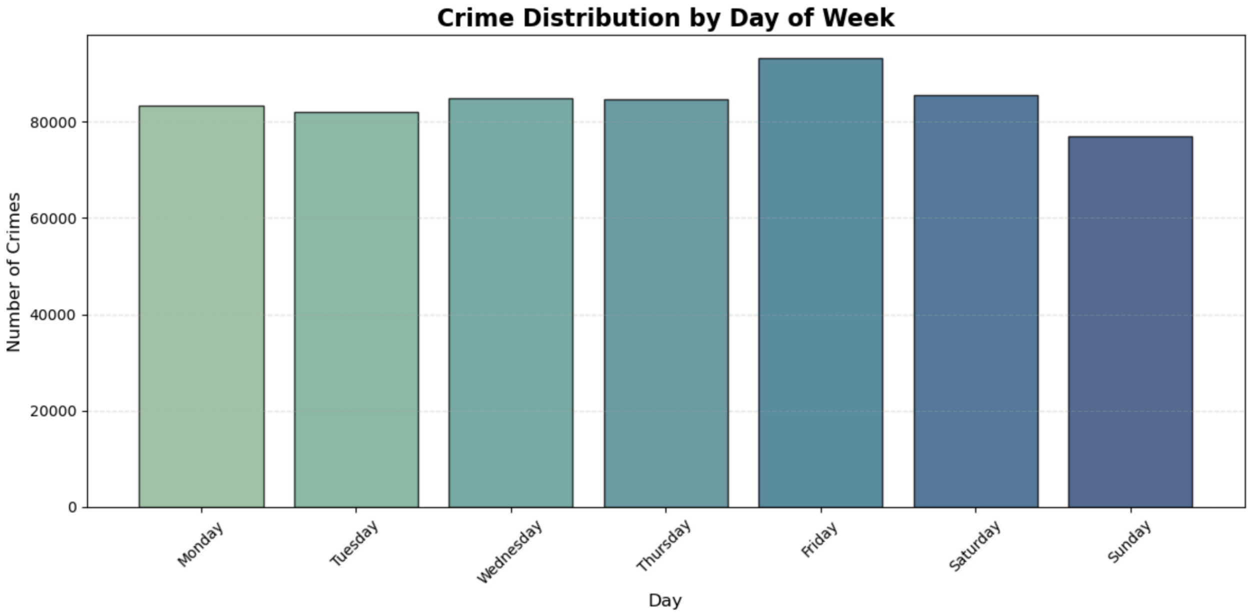


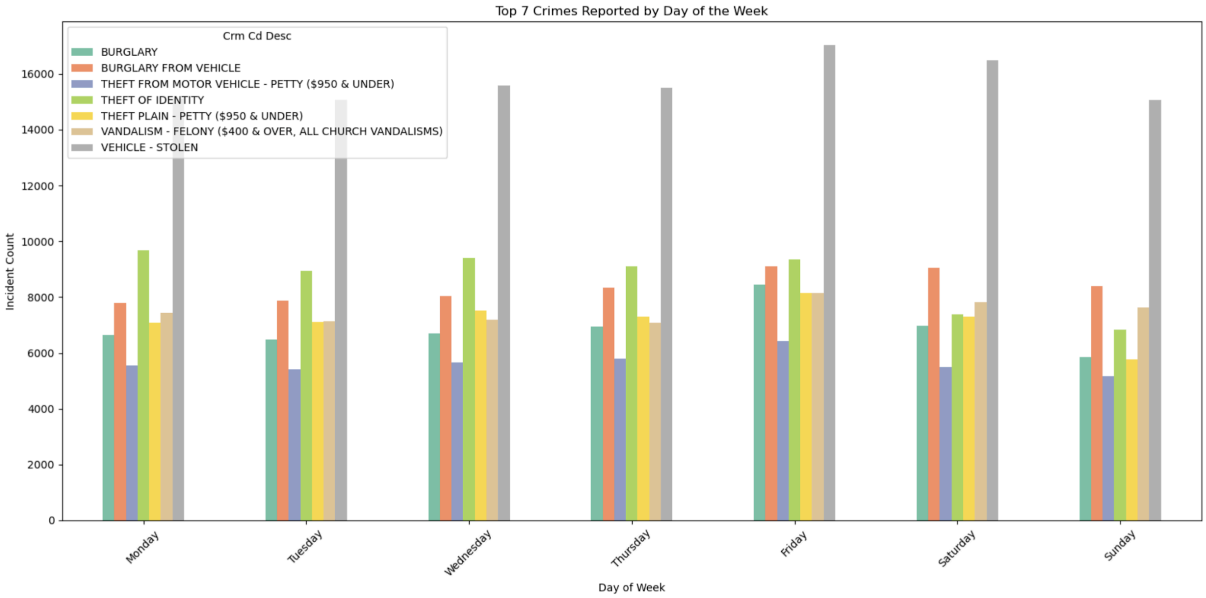
```
Correlation between crime count and unemployment rate:
                   crime_count  unemployment_rate
crime_count           1.000000          -0.015814
unemployment_rate    -0.015814           1.000000
```



6. Day-of-Week Analysis

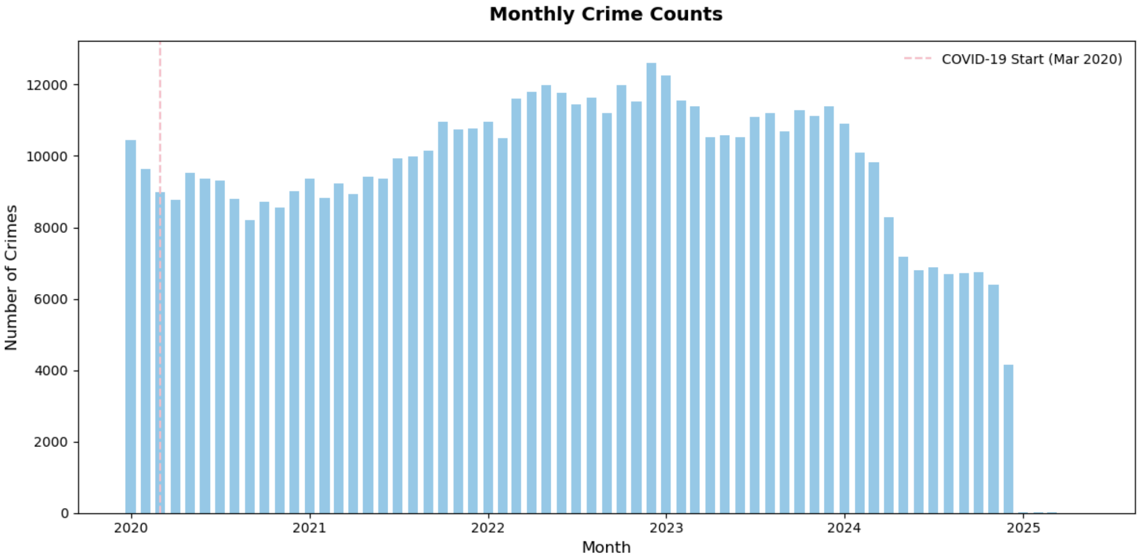
Crimes were most frequently reported on **Fridays**, correlating with social gatherings and nightlife Crime rates are highest in the winter and lowest in the summer. On a weekly basis, criminal activity steadily increases from Monday, peaks on Friday, and is lowest on Sunday. A major contributing factor to these weekly totals is vehicle theft, which is the most frequently reported crime across all days and spikes significantly on Friday.





7. Impact of Major Events

The chart marks the start of the pandemic in March 2020, allowing for a before-and-after comparison of crime rates.The COVID-19 pandemic appears to have had a multi-phased impact on crime rates. An initial, brief decline in March 2020 was followed by a prolonged two-year surge in criminal activity. Subsequently, from 2023 onwards, there has been a significant and sustained reduction in crime, bringing the monthly counts to a multi-year low.



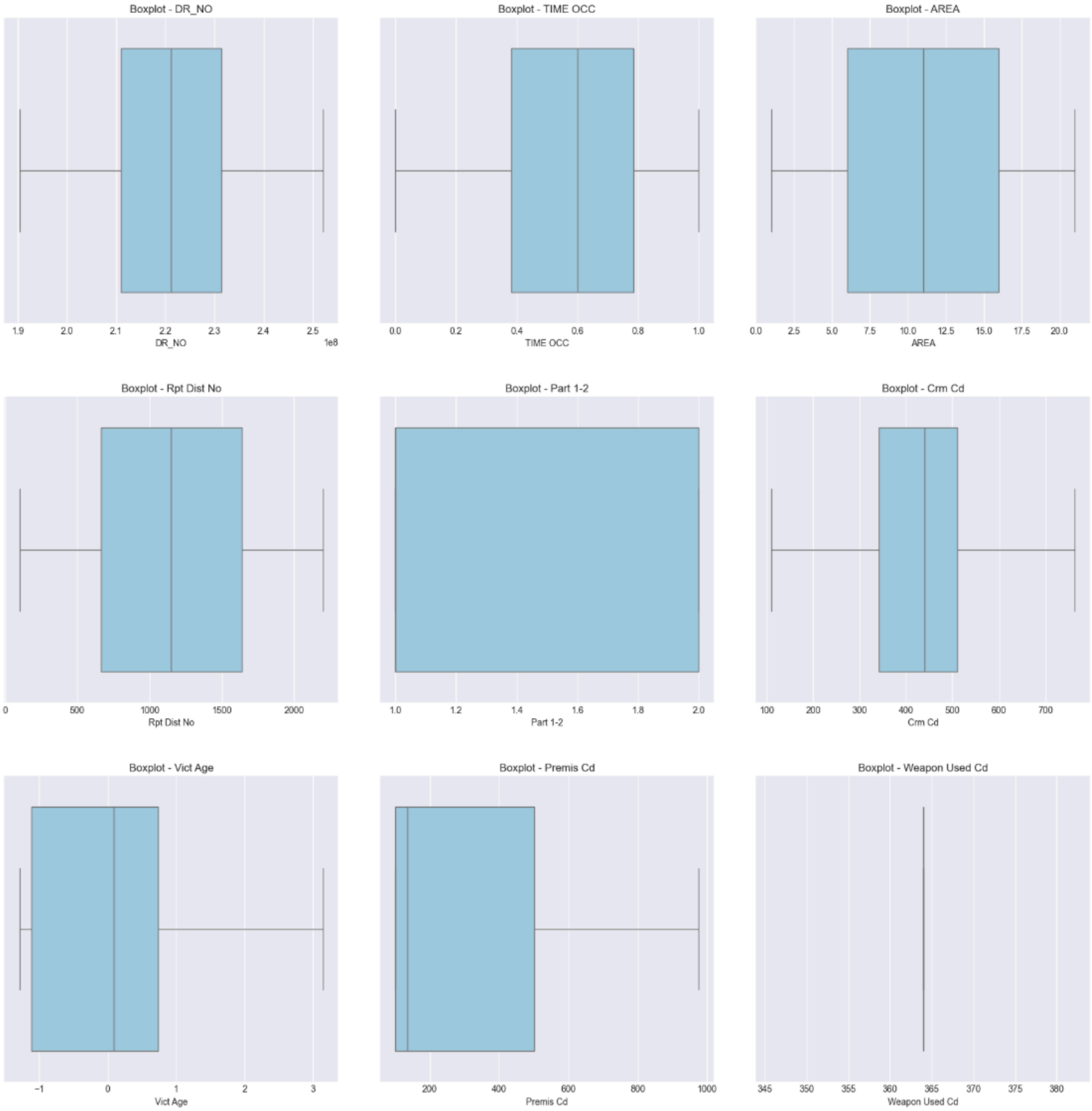
8. Outliers and Anomalies

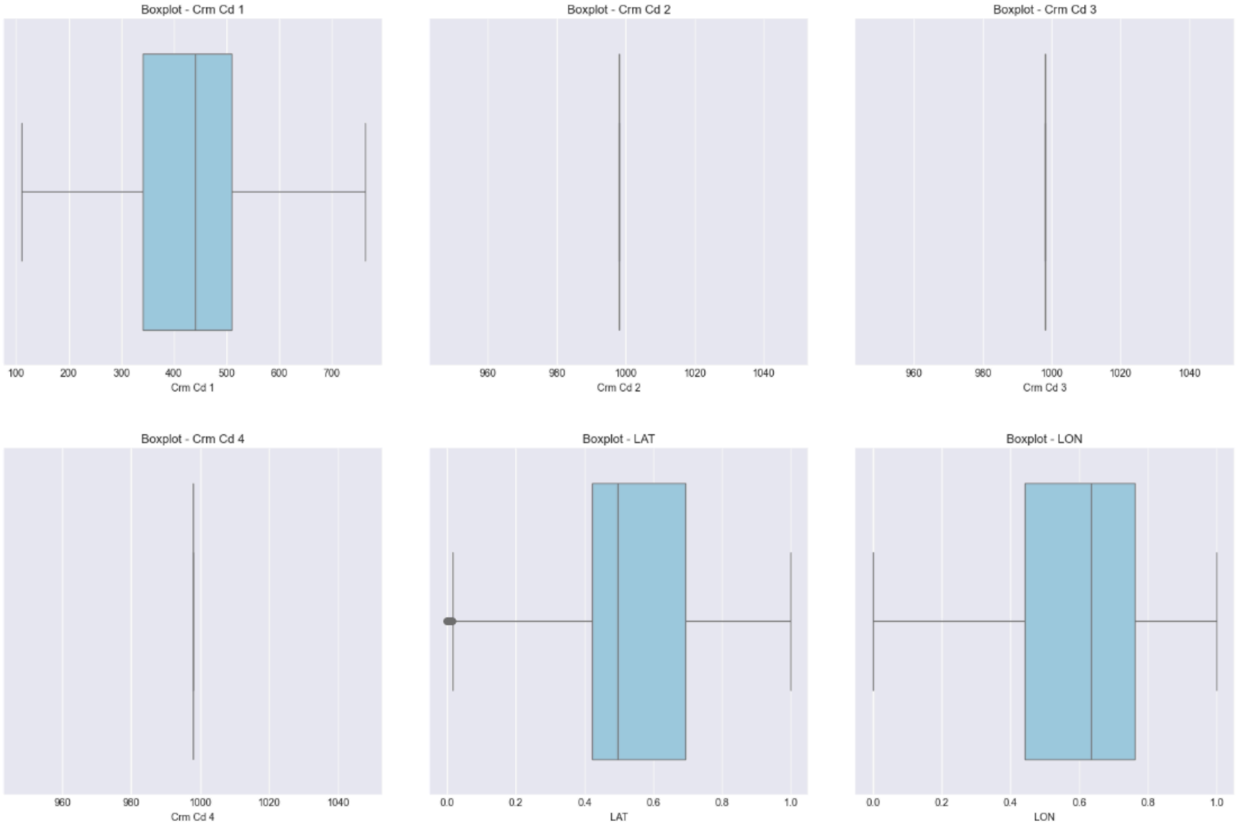
The outlier detection process identified and removed anomalous values across multiple variables, resulting in a cleaned dataset of 590,300 records across 28 features. Most variables had minimal or no outliers, indicating that the original dataset was generally well-maintained.

The most significant cleaning occurred in the Weapon Used Cd variable, where 327,245 outliers were removed, likely representing missing values, invalid codes, or non-applicable entries. Similarly, Crm Cd 1 saw 64,021 outliers removed, reflecting possible data entry errors or non-standard crime codes.

Geographic variables also required attention: LAT had 21,611 outliers removed, while LON showed none. This suggests latitude values included impossible coordinates, whereas longitude was consistently recorded. The single outlier in Vict Age likely represented an implausible age, such as a negative value or one exceeding human lifespan.

Minimal outlier removal in core identifiers like DR_NO, TIME_OCC, AREA, and various crime codes indicates these structured fields maintained strong data integrity. Overall, the selective cleaning preserved the vast majority of legitimate data while removing statistically anomalous values that could skew analysis and modeling.

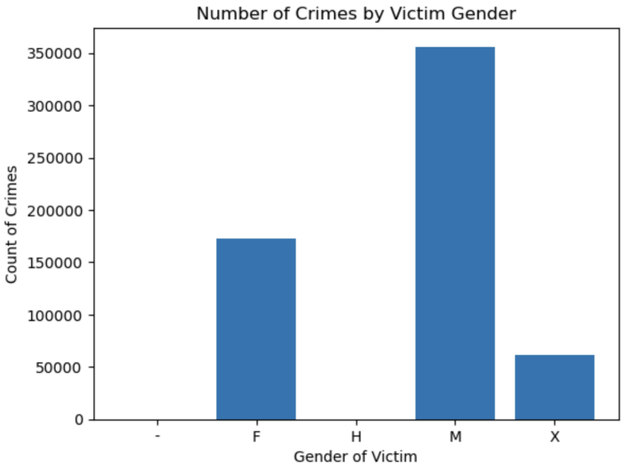

Boxplot - DR_NO
1.9
2.0
2.1
2.2
2.3
2.4
2.5
1e8
DR_NO
Boxplot - TIME OCC
0.0
0.2
0.4
0.6
0.8
1.0
TIME OCC
Boxplot - AREA
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
AREA
Boxplot - Rpt Dist No
0
500
1000
1500
2000
Rpt Dist No
Boxplot - Part 1-2
1.0
1.2
1.4
1.6
1.8
2.0
Part 1-2
Boxplot - Crm Cd
100
200
300
400
500
600
700
Crm Cd
Boxplot - Vict Age
−1
0
1
2
3
Vict Age
Boxplot - Premis Cd
200
400
600
800
1000
Premis Cd
Boxplot - Weapon Used Cd
345
350
355
360
365
370
375
380
Weapon Used Cd

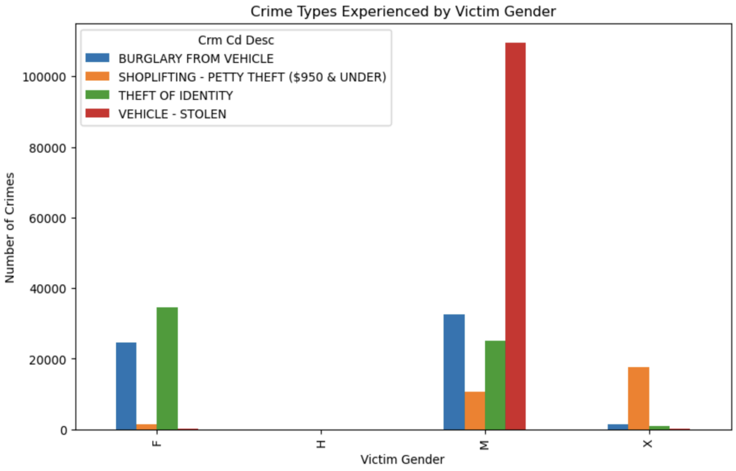


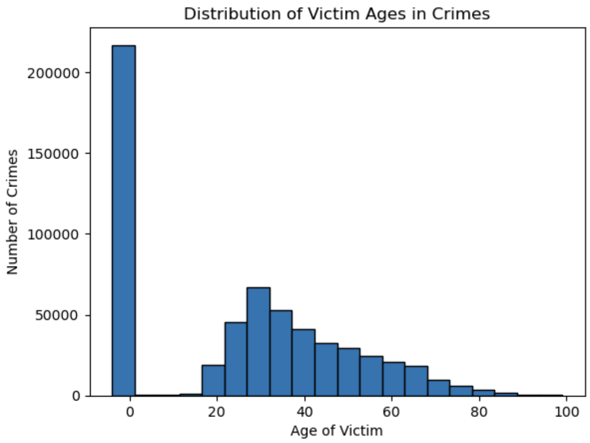
```
DR_NO: removed 8 outliers
TIME OCC: removed 0 outliers
AREA: removed 0 outliers
Rpt Dist No: removed 0 outliers
Part 1-2: removed 0 outliers
Crm Cd: removed 0 outliers
Vict Age: removed 1 outliers
Premis Cd: removed 0 outliers
Weapon Used Cd: removed 327245 outliers
Crm Cd 1: removed 64021 outliers
Crm Cd 2: removed 1803 outliers
Crm Cd 3: removed 2 outliers
Crm Cd 4: removed 0 outliers
LAT: removed 21611 outliers
LON: removed 0 outliers

 Outlier handling complete.
Remaining data shape: (590300, 28)
```

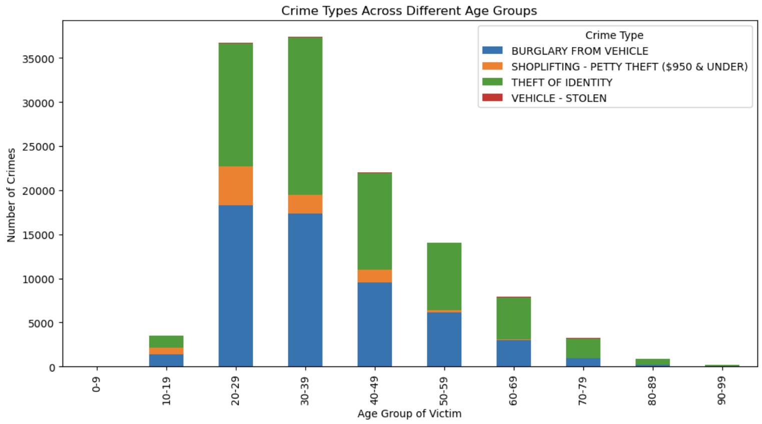
9. Demographic Factors

Where demographic data were available, males accounted for most violent crimes, while property Based on the data, there are clear demographic patterns in crime victimization. Males are victims of crime at a rate roughly double that of females. The types of crimes also differ by gender; males are predominantly victims of vehicle theft and burglary from a vehicle, while females are most frequently targeted for theft of identity. A smaller group of non-binary victims also reported theft of identity as their most common crime.Age is also a significant factor in crime victimization. The likelihood of being a victim peaks for individuals in the 20-39 age range and steadily declines with increasing age. Across the most victimized age groups, theft of identity and burglary from a vehicle are the most prevalent crimes. Although there is a notable number of victims in the 0-9 age category, the overall trend shows that young adults are the most common targets.









10. Predicting Future Trends:

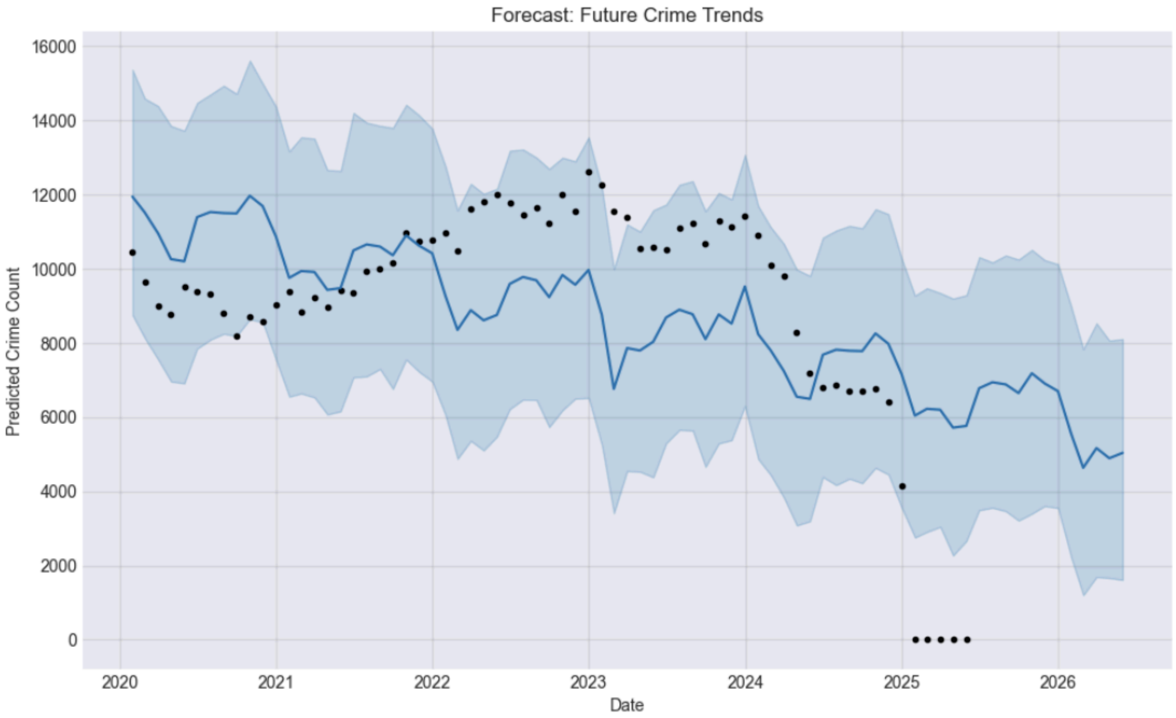
Overall Trend: The Prophet model forecasts a continued declining trend in crime incidents, dropping from approximately 12,000 monthly incidents in 2020 to around 5,000-6,000 by late 2026, a roughly 50% reduction over the forecast period.

Model Performance: The black dots (actual observations) align reasonably well with the model's predictions through mid-2024, suggesting the model captures historical patterns effectively. However, the widening confidence interval (shaded blue area) indicates increasing uncertainty in longer-term predictions, particularly beyond 2025.

Notable Patterns:

- Seasonal fluctuations are evident throughout the historical period
- A sharp decline appears around 2022-2023, which the model projects will continue
- The expanding uncertainty bounds suggest predictions should be interpreted cautiously for 2025-2026

While the forecast suggests a positive trend with decreasing crime rates, the high uncertainty in future predictions necessitates regular model updates with new data. The dramatic decline may reflect changes in reporting practices, policy interventions, or demographic shifts that should be investigated further to ensure forecast validity.



5. Conclusion

This comprehensive analysis of crime data from 2020 to present revealed several critical insights into criminal activity patterns and their underlying drivers. The project successfully applied robust data cleaning methodologies, exploratory data analysis, and predictive modeling to extract meaningful trends from a complex dataset.

Key Findings

Temporal Trends: Crime activity peaked in 2022 with approximately 139,000 incidents before entering a decline phase. However, the dramatic drop in 2024 and near-zero figures in 2025 strongly suggest data collection or reporting issues rather than genuine crime reduction, highlighting the importance of data quality validation in analytical work.

Crime Type Concentration: Vehicle theft emerged as the dominant crime category, accounting for 109,847 incidents—nearly 22% of the top crime types. This concentration indicates a critical need for targeted prevention strategies focusing on vehicle security and theft deterrence.

Seasonal and Weekly Patterns: Clear cyclical patterns emerged, with crime peaking during winter months (December-February) and on Fridays. These predictable patterns offer opportunities for strategic resource allocation and preventive interventions during high-risk periods.

Demographic Vulnerabilities: Males experience crime victimization at roughly double the rate of females, with young adults aged 20-39 representing the most targeted demographic. Identity theft particularly affects females and older populations, suggesting the need for age- and gender-specific prevention programs.

Limited Economic Correlation: The weak correlation (r = -0.016) between unemployment and crime challenges common assumptions about economic factors as primary crime drivers. This finding underscores the complexity of criminal behavior and the need for multifaceted analytical approaches incorporating policing strategies, social services, and community characteristics.

Limitations

Several limitations warrant acknowledgment:

- **Data Quality Issues**: The 2024-2025 data anomalies significantly impact trend reliability and predictive model accuracy
- **Missing Variables**: The dataset lacked comprehensive socioeconomic indicators beyond unemployment, limiting causal inference
- **Reporting Bias**: Crime statistics reflect reported incidents, potentially underrepresenting certain crime types or demographics
- **Geographic Scope**: Analysis focused primarily on urban areas, limiting generalizability to rural contexts