IE6400 Foundations of Data Analytics Engineering

Project 1

Cleaning and Analyzing Crime Data

Report

Student Name

Varaalakshime Vigneswara Pandiajan
Anita Janie Christdoss Chelladurai
Anoushka Karan Pakhare

Objective

The core objective of this project is to leverage the real-world crime dataset from 2020 to the present to conduct a comprehensive data pipeline. This encompassed rigorous data cleaning and preparation, detailed Exploratory Data Analysis (EDA) to discern underlying crime trends and patterns, and time-series forecasting to predict future crime rates. The derived insights are intended to support evidence-based decision making for law enforcement and public safety policy.

1. Data Acquisition and Inspection

1.1 Data Acquisition

The crime dataset was successfully acquired and loaded into the analysis environment. The initial dataset contained 532,824 records and 28 features.

1.2 Data Inspection

Initial inspection revealed that crucial time-series fields (Date Rptd, DATE OCC) were incorrectly classified as generic object (string) types, necessitating conversion. Furthermore, several descriptive columns, including Mocodes, Vict Sex, and Weapon Desc, exhibited a substantial number of missing values.

Looking at first ten rows:

Out[ ]:

	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd
0	211507896	04/11/2021 12:00:00 AM	11/07/2020 12:00:00 AM	845	15	N Hollywood	1502	2	354
1	201516622	10/21/2020 12:00:00 AM	10/18/2020 12:00:00 AM	1845	15	N Hollywood	1521	1	230
2	240913563	12/10/2024 12:00:00 AM	10/30/2020 12:00:00 AM	1240	9	Van Nuys	933	2	354
3	210704711	12/24/2020 12:00:00 AM	12/24/2020 12:00:00 AM	1310	7	Wilshire	782	1	331

tel:210704711

4	201418201	10/03/2020 12:00:00 AM	09/29/2020 12:00:00 AM	1830	14	Pacific	1454	1	420
5	240412063	12/11/2024 12:00:00 AM	11/11/2020 12:00:00 AM	1210	4	Hollenbeck	429	2	354
6	240317069	12/16/2024 12:00:00 AM	04/16/2020 12:00:00 AM	1350	3	Southwest	396	2	354
7	201115217	10/29/2020 12:00:00 AM	07/07/2020 12:00:00 AM	1400	11	Northeast	1133	2	812
8	241708596	04/20/2024 12:00:00 AM	03/02/2020 12:00:00 AM	1200	17	Devonshire	1729	2	354
9	242113813	12/18/2024 12:00:00 AM	09/01/2020 12:00:00 AM	900	21	Topanga	2196	2	354

10 rows × 28 columns

Checking data types :

```
DR_NO              int64
Date Rptd         object
DATE OCC          object
TIME OCC           int64
AREA               int64
AREA NAME         object
Rpt Dist No        int64
Part 1-2           int64
Crm Cd             int64
Crm Cd Desc       object
Mocodes           object
Vict Age           int64
Vict Sex          object
Vict Descent      object
Premis Cd        float64
Premis Desc       object
Weapon Used Cd   float64
Weapon Desc       object
Status            object
Status Desc       object
Crm Cd 1         float64
Crm Cd 2         float64
Crm Cd 3         float64
Crm Cd 4         float64
LOCATION          object
Cross Street      object
LAT              float64
LON              float64
dtype: object
```

Review column names and descriptions:

Out[ ]:

	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA
count	5.328240e+05	532824	532824	532824.000000	532824.000000
unique	NaN	1789	1096	NaN	NaN
top	NaN	11/01/2021 12:00:00 AM	01/01/2020 12:00:00 AM	NaN	NaN
freq	NaN	753	1164	NaN	NaN
mean	2.101192e+08	NaN	NaN	1338.927501	10.761542
std	8.055423e+06	NaN	NaN	652.943389	6.073469
min	8.170000e+02	NaN	NaN	1.000000	1.000000
25%	2.015060e+08	NaN	NaN	900.000000	6.000000
50%	2.108072e+08	NaN	NaN	1419.000000	11.000000
75%	2.121177e+08	NaN	NaN	1900.000000	16.000000
max	2.520042e+08	NaN	NaN	2359.000000	21.000000

11 rows × 28 columns

Out[ ]:

```
['DR_NO',
 'Date Rptd',
 'DATE OCC',
 'TIME OCC',
 'AREA',
 'AREA NAME',
 'Rpt Dist No',
 'Part 1-2',
 'Crm Cd',
 'Crm Cd Desc',
 'Mocodes',
 'Vict Age',
 'Vict Sex',
 'Vict Descent',
 'Premis Cd',
 'Premis Desc',
 'Weapon Used Cd',
 'Weapon Desc',
 'Status',
 'Status Desc',
 'Crm Cd 1',
 'Crm Cd 2',
 'Crm Cd 3',
 'Crm Cd 4',
 'LOCATION',
 'Cross Street',
 'LAT',
 'LON']
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 532824 entries, 0 to 532823
Data columns (total 28 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   DR_NO           532824 non-null  int64  
 1   Date Rptd       532824 non-null  object 
 2   DATE OCC        532824 non-null  object 
 3   TIME OCC        532824 non-null  int64  
 4   AREA            532824 non-null  int64  
 5   AREA NAME       532824 non-null  object 
 6   Rpt Dist No     532824 non-null  int64  
 7   Part 1-2        532824 non-null  int64  
 8   Crm Cd          532824 non-null  int64  
 9   Crm Cd Desc     532824 non-null  object 
 10  Mocodes         459598 non-null  object 
 11  Vict Age        532824 non-null  int64  
 12  Vict Sex        463186 non-null  object 
 13  Vict Descent    463182 non-null  object 
 14  Premis Cd       532817 non-null  float64
 15  Premis Desc     532586 non-null  object 
 16  Weapon Used Cd  187722 non-null  float64
 17  Weapon Desc     187722 non-null  object 
 18  Status          532824 non-null  object 
 19  Status Desc     532824 non-null  object 
 20  Crm Cd 1        532817 non-null  float64
 21  Crm Cd 2        41001 non-null   float64
 22  Crm Cd 3        1390 non-null    float64
 23  Crm Cd 4        43 non-null      float64
 24  LOCATION        532824 non-null  object 
 25  Cross Street    89270 non-null   object 
 26  LAT             532823 non-null  float64
 27  LON             532823 non-null  float64
dtypes: float64(8), int64(7), object(13)
memory usage: 113.8+ MB
```

2. Data Cleaning and Preparation

Meticulous data cleaning was executed to ensure the dataset's integrity and reliability for subsequent statistical and predictive modeling.

2.1 Handling Missing Data and Duplicates

No duplicate records were found in the dataset. Missing values were systematically addressed:

- Categorical Imputation: Missing values in fields like Vict Sex, Vict Descent, and Mocodes were imputed with the placeholder "Unknown" or "Not Specified."

```
DR_NO                  0
Date Rptd              0
DATE OCC               0
TIME OCC               0
AREA                   0
AREA NAME              0
Rpt Dist No            0
Part 1-2               0
Crm Cd                 0
Crm Cd Desc            0
Mocodes            73226
Vict Age               0
Vict Sex           69638
Vict Descent       69642
Premis Cd              7
Premis Desc          238
Weapon Used Cd    345102
Weapon Desc       345102
Status                 0
Status Desc            0
Crm Cd 1               7
Crm Cd 2          491823
Crm Cd 3          531434
Crm Cd 4          532781
LOCATION               0
Cross Street      443554
LAT                    1
LON                    1
dtype: int64
```

- Numerical Imputation: Null values in fields such as auxiliary crime codes and weapon codes were replaced with zero.

```
DR_NO             0
Date Rptd         0
DATE OCC          0
TIME OCC          0
AREA              0
AREA NAME         0
Rpt Dist No       0
Part 1-2          0
Crm Cd            0
Crm Cd Desc       0
Mocodes           0
Vict Age          0
Vict Sex          0
Vict Descent      0
Premis Cd         0
Premis Desc       0
Weapon Used Cd    0
Weapon Desc       0
Status            0
Status Desc       0
Crm Cd 1          0
Crm Cd 2          0
Crm Cd 3          0
Crm Cd 4          0
LOCATION          0
Cross Street      0
LAT               1
LON               1
dtype: int64
```

2.2 Data Type Conversion and Feature Engineering

All date columns were successfully converted to the correct datetime format. This enabled the engineering of new temporal features, including Year, Month, and DayName, which are essential for seasonal analysis.

Checking for Datatypes:

Out[ ]:

	0
DR_NO	int64
Date Rptd	datetime64[ns]
DATE OCC	datetime64[ns]
TIME OCC	int64
AREA	int64
AREA NAME	object
Rpt Dist No	int64
Part 1-2	int64
Crm Cd	int64
Crm Cd Desc	object
Mocodes	object
Vict Age	int64
Vict Sex	object
Vict Descent	object
Premis Cd	float64
Premis Desc	object
Weapon Used Cd	float64
Weapon Desc	object
Status	object
Status Desc	object
Crm Cd 1	float64
Crm Cd 2	float64
Crm Cd 3	float64
Crm Cd 4	float64
LOCATION	object
Cross Street	object
LAT	float64
LON	float64
Year	int32
Month	int32
Day	int32
DayOfWeek	int32
DayName	object
MonthName	object

2.3 Outlier Management

Outliers in key numerical features were managed to prevent statistical skew:

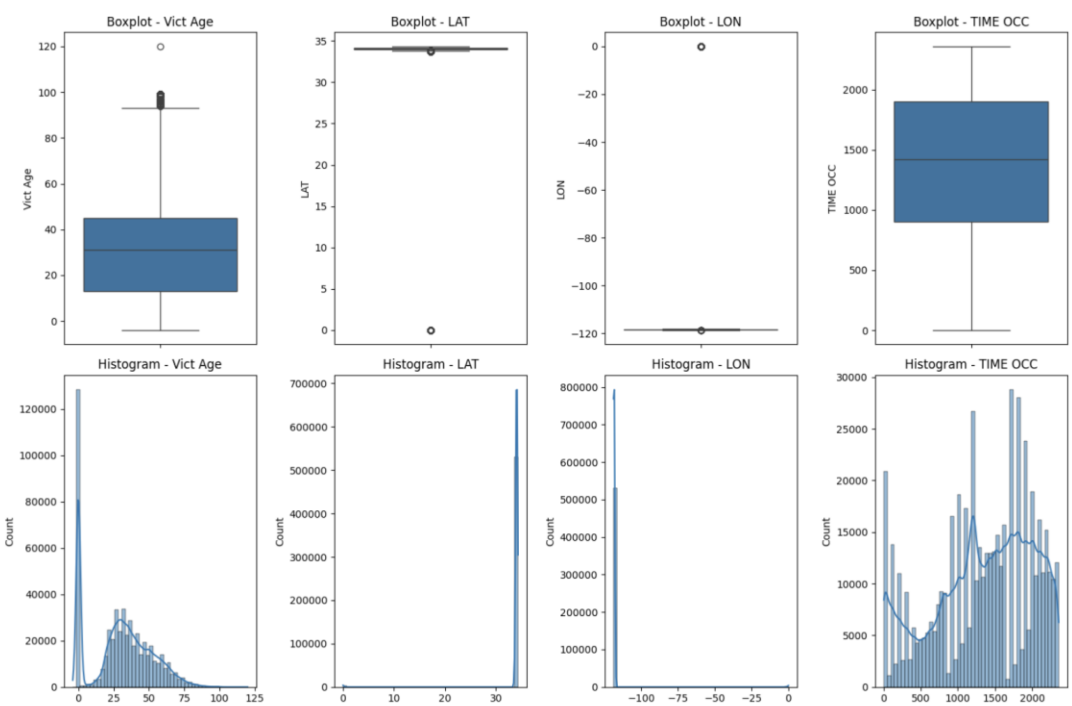
Out[ ]:

	DR_NO	Date Rptd	DATE OCC	TIME OCC
count	5.328240e+05	532824	532824	532824.000000
mean	2.101192e+08	2021-05-26 05:19:53.784063488	2021-05-10 15:46:02.145849600	1338.927501
min	8.170000e+02	2020-01-01 00:00:00	2020-01-01 00:00:00	1.000000
25%	2.015060e+08	2020-09-06 00:00:00	2020-08-26 00:00:00	900.000000
50%	2.108072e+08	2021-05-19 00:00:00	2021-05-04 00:00:00	1419.000000
75%	2.121177e+08	2021-12-29 00:00:00	2021-12-13 00:00:00	1900.000000
max	2.520042e+08	2025-03-28 00:00:00	2022-12-31 00:00:00	2359.000000
std	8.055423e+06	NaN	NaN	652.943389

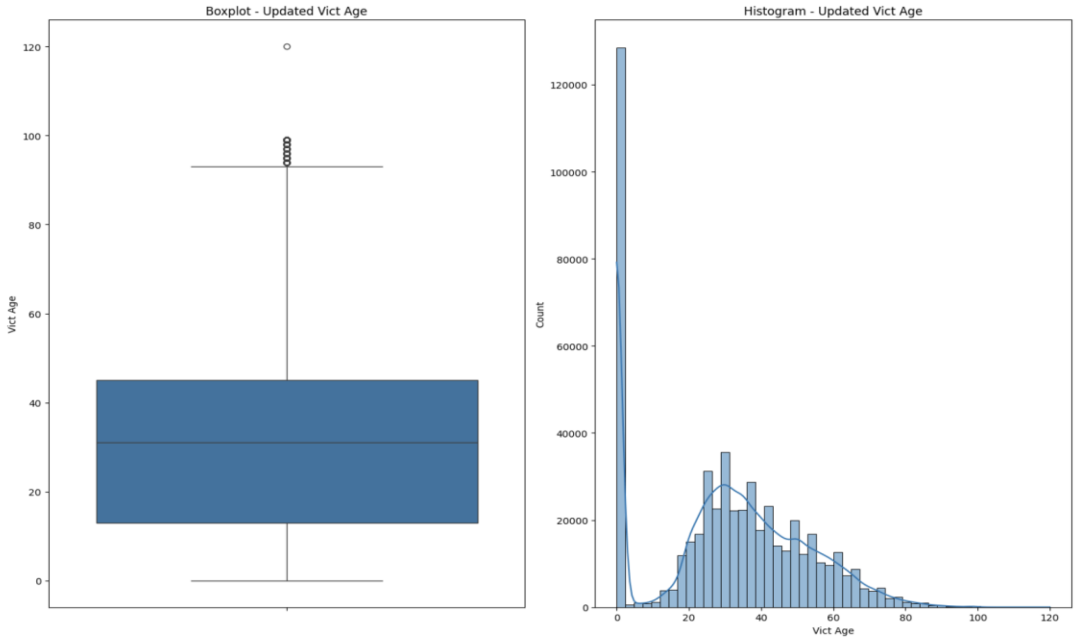
8 rows × 21 columns

- Victim Age: All negative victim ages were filtered out, establishing a valid range of 0 to 120.

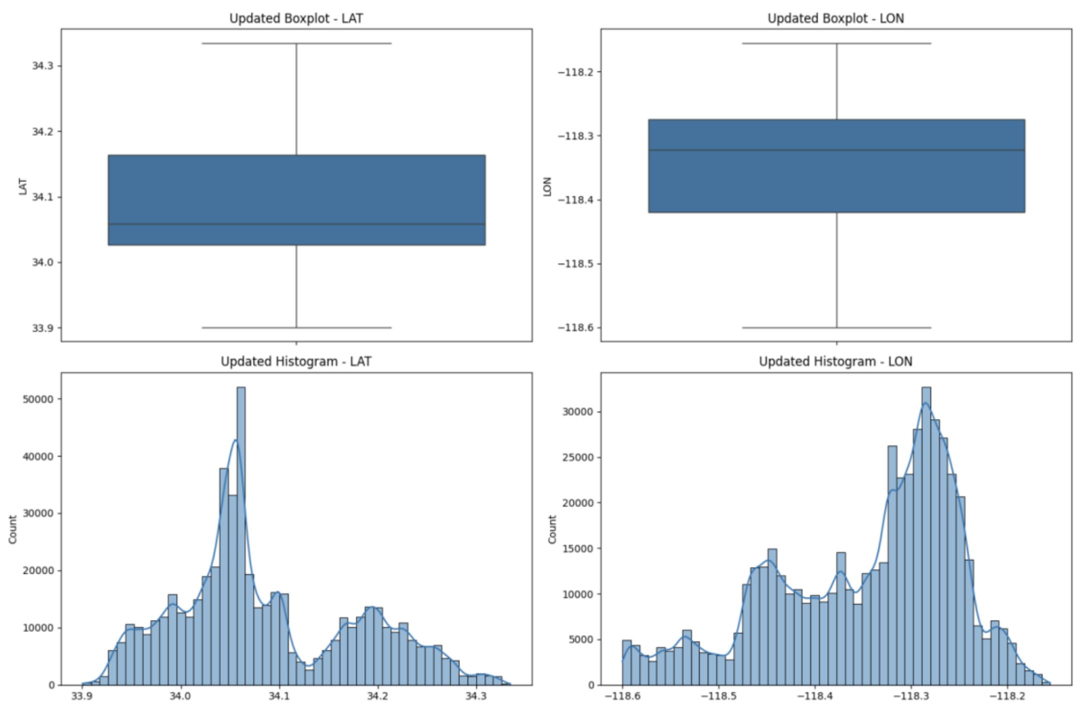
  Before dealing with outliers :



After dealing with outliers :



- Geographic Coordinates: Extreme outliers in LAT and LON were corrected by imputing the values with the column median, which is a robust measure against extreme values.



2.4 Standardization and Encoding

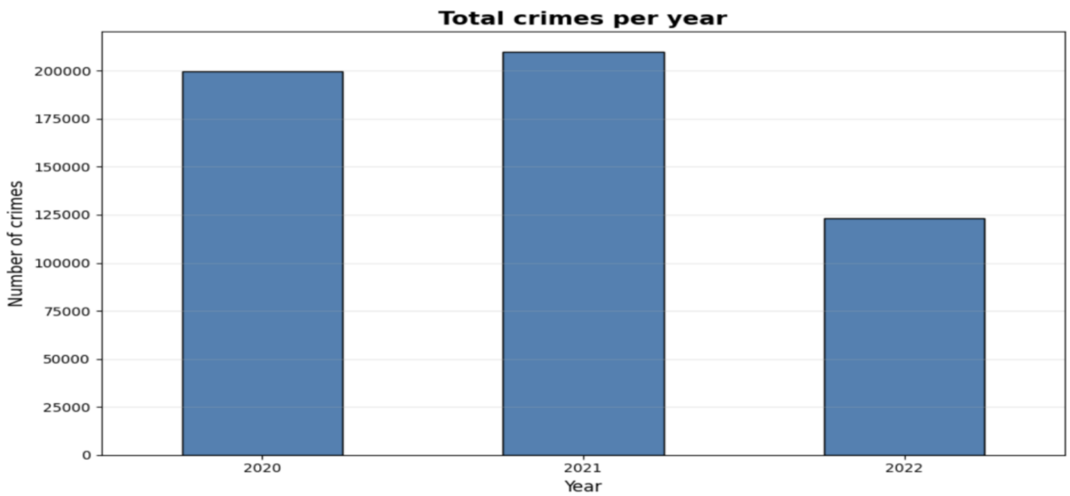
Final preparation steps included data transformation for modeling:

- Scaling: Vict Age was standardized using StandardScaler, while TIME OCC, LAT, and LON were normalized using MinMaxScaler.
- Encoding: Categorical variables were transformed using One-Hot Encoding (for low-to-medium cardinality features like AREA NAME) or Label Encoding (for high-cardinality features like Mocodes).

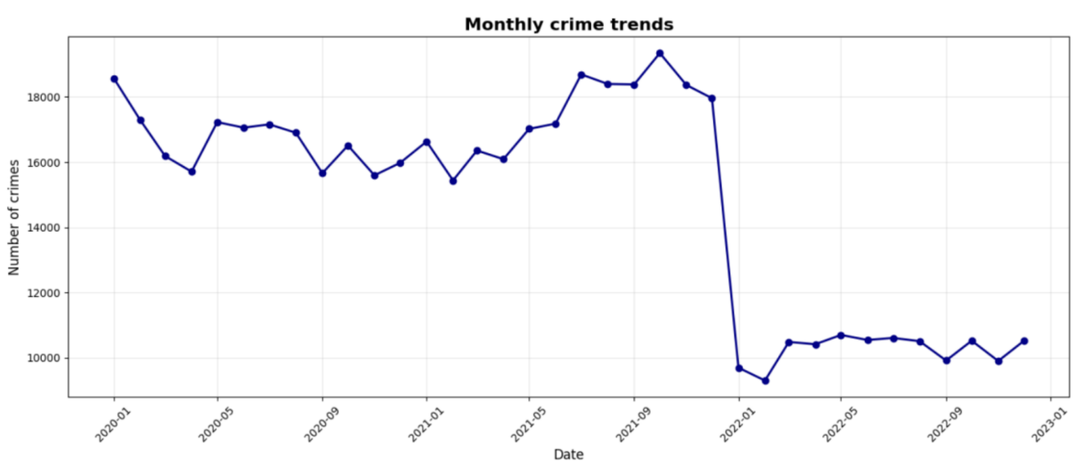
3. Exploratory Data Analysis (EDA)

3.1 Overall Crime Trends

Analysis of annual crime counts revealed significant volatility. Incidents increased marginally in 2021 (209,827) from 2020 (199,806), likely due to a return to normal activity post-initial pandemic lockdowns, before experiencing a sharp decline in 2022 (123,082).



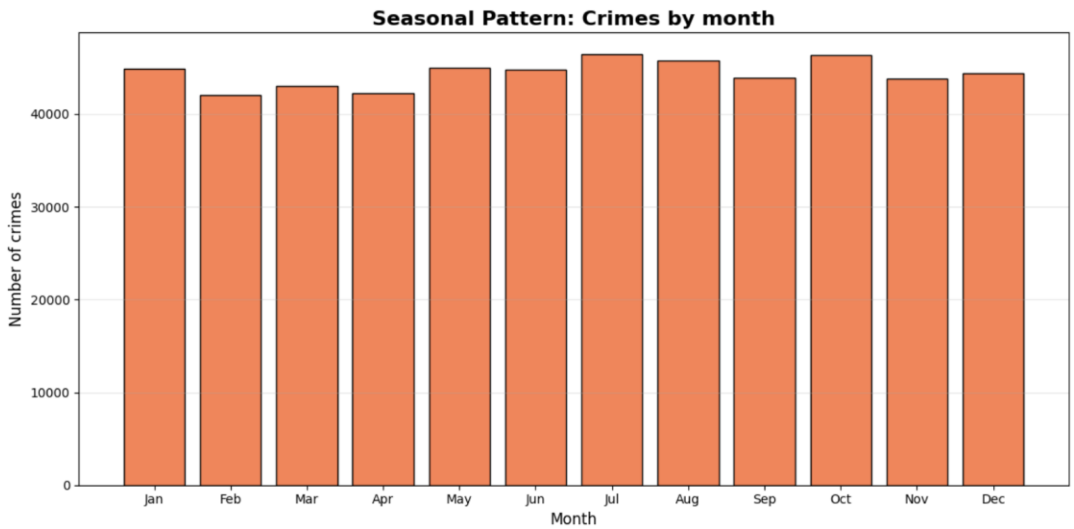
The monthly crime count showed a volatile, upward trend through 2021, followed by a dramatic, approximately 50% structural drop at the start of 2022, settling at a new, lower baseline.



3.2 Seasonal Patterns

A clear seasonal component was observed upon analyzing monthly crime rates.

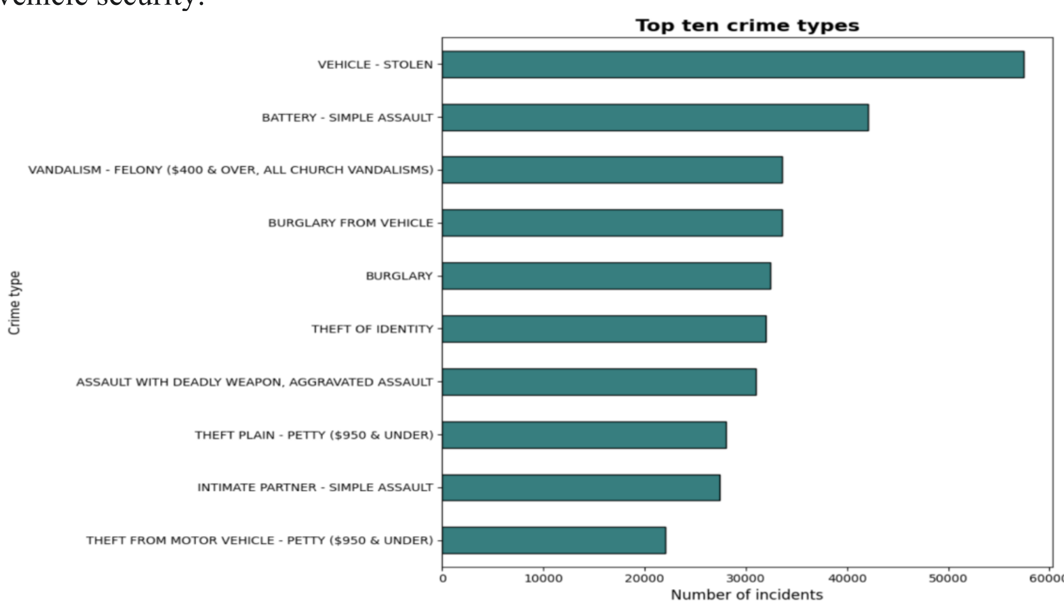
- Crime incidents are consistently highest during the summer and early fall specifically in July, August, and October suggesting a direct correlation with warmer weather and increased public mobility.

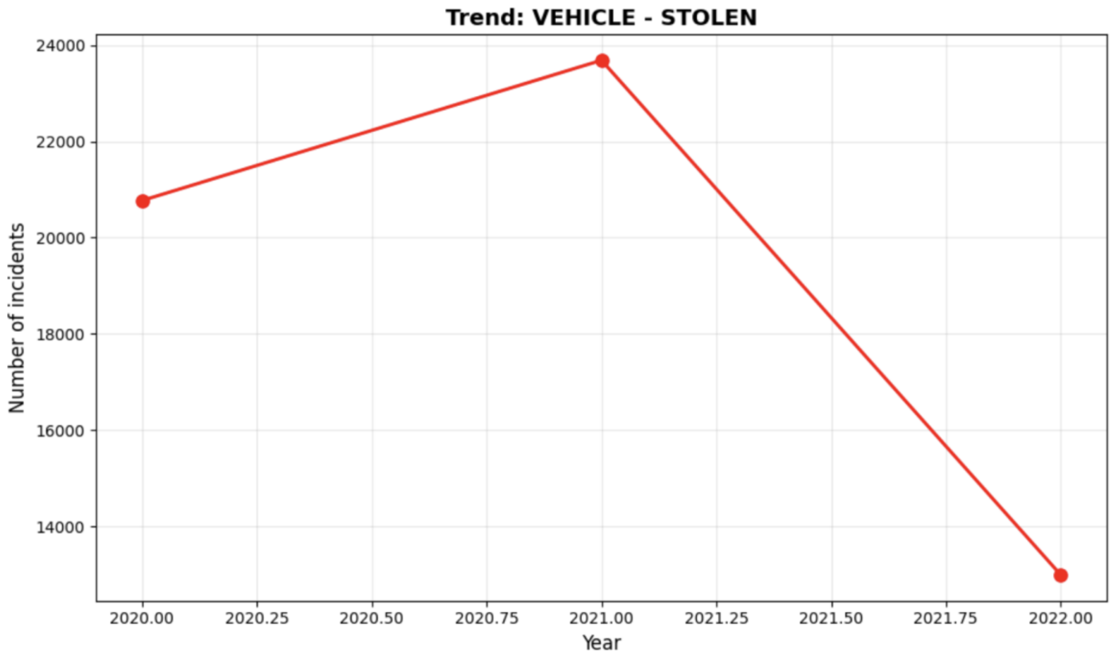


3.3 Most Common Crime Type

The dataset is heavily weighted toward specific offenses.

- Vehicle Theft is the dominant crime category, significantly outpacing others with 57,445 incidents. Simple Assault followed as the second most common offense. This trend emphasizes the need for specialized preventive efforts focused on vehicle security.

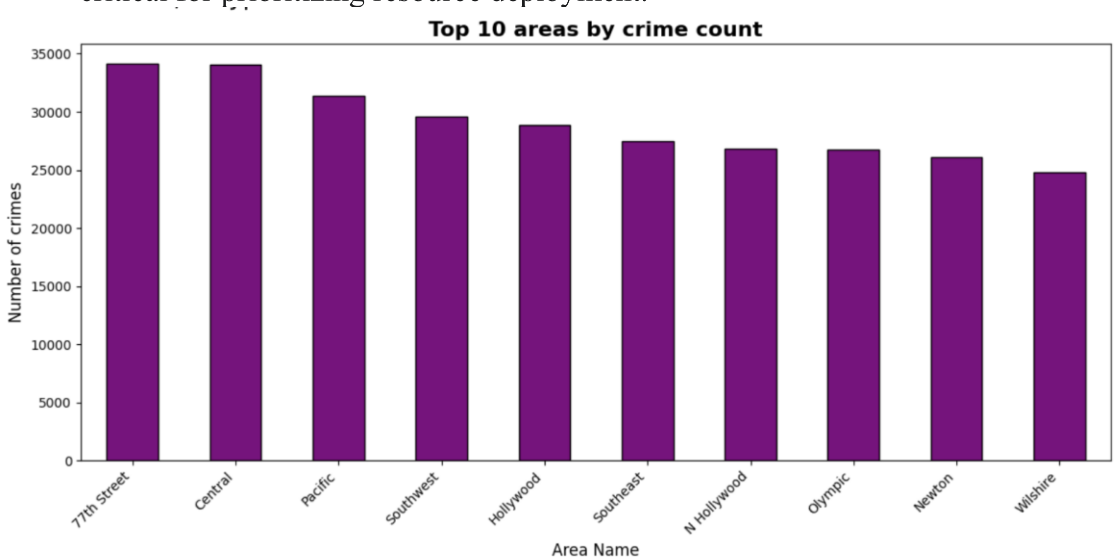




3.4 Regional Differences

Geographic analysis of crime distribution highlighted spatial disparities.

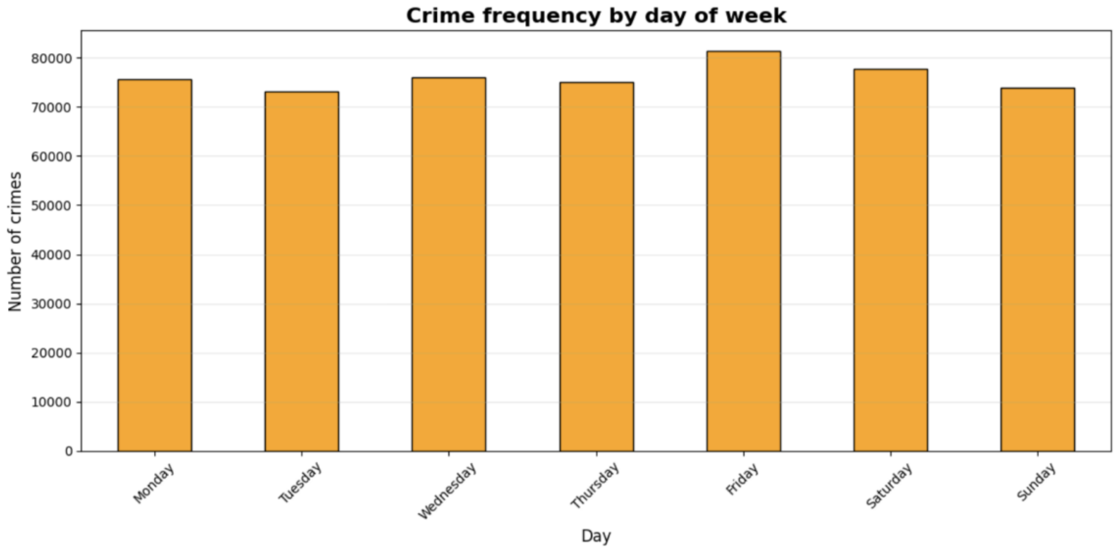
- Key Finding: The 77th Street and Central districts exhibit the highest crime concentrations, identifying them as persistent hotspots. This geographic insight is critical for prioritizing resource deployment.

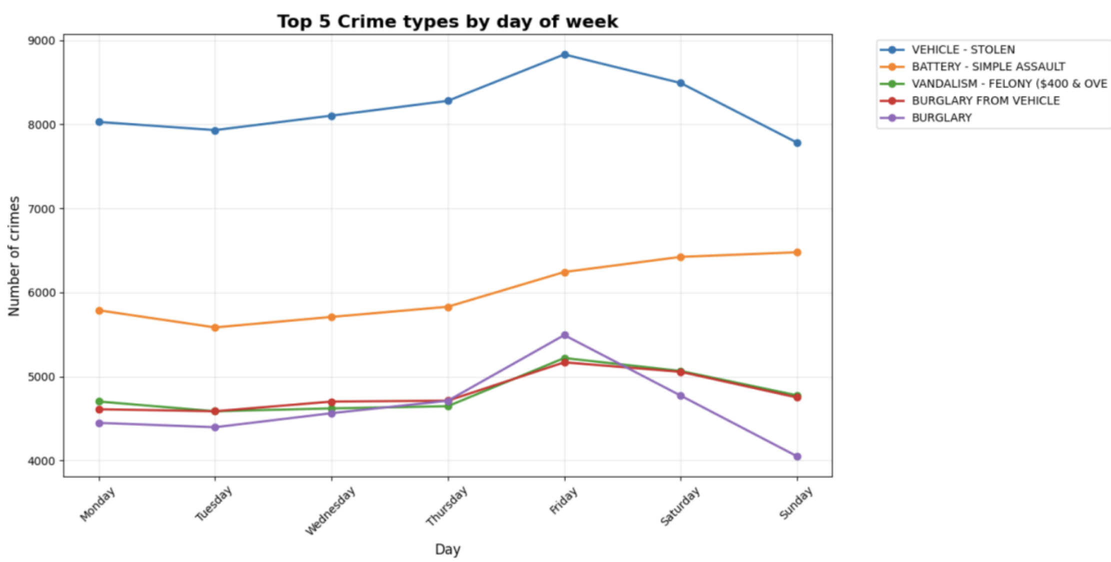


3.5 Day of the Week Analysis

Temporal analysis revealed a strong link between crime frequency and the weekly schedule.

- Key Finding: Crime frequency peaks on Fridays and Saturdays, suggesting a direct correlation with weekend leisure and social activities. This pattern holds true across the top crime types.





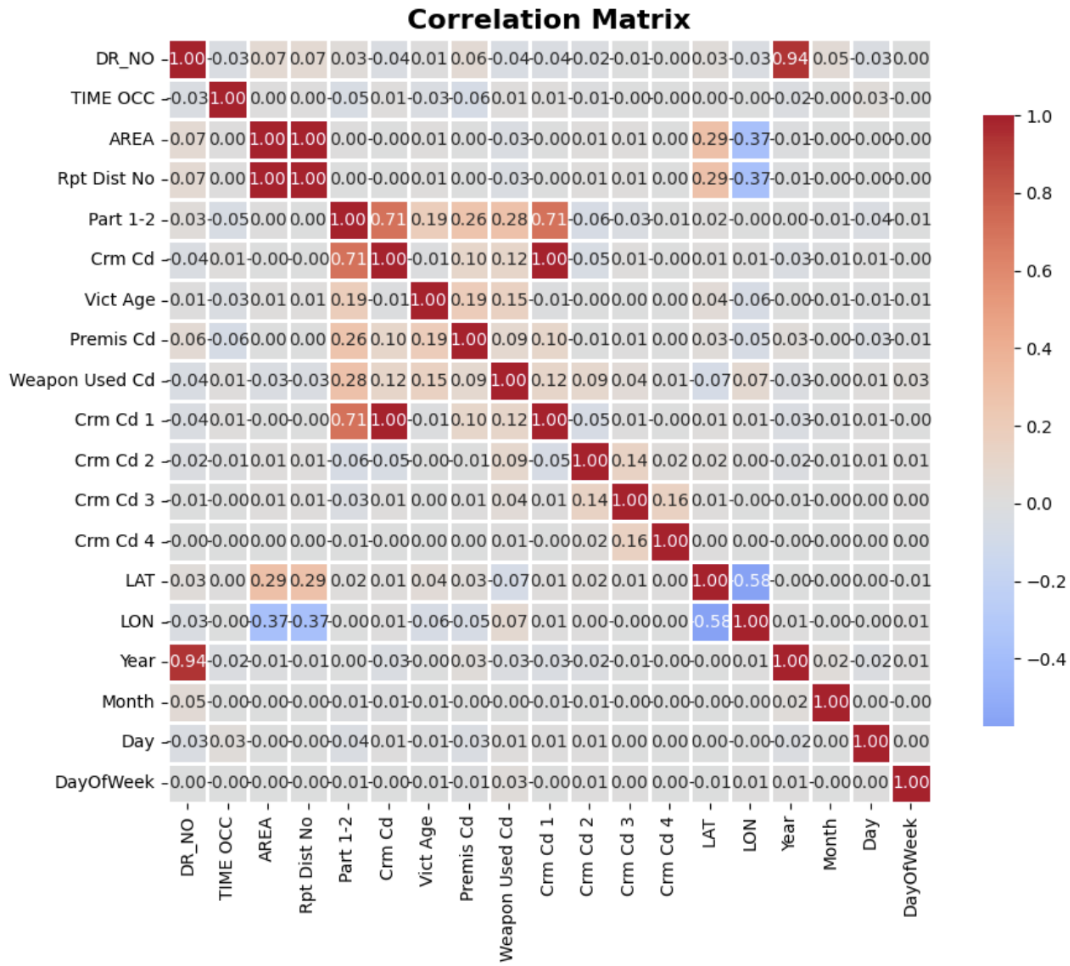
3.6 Demographic Factors

Analysis of demographic features, specifically victim age, showed specific patterns.

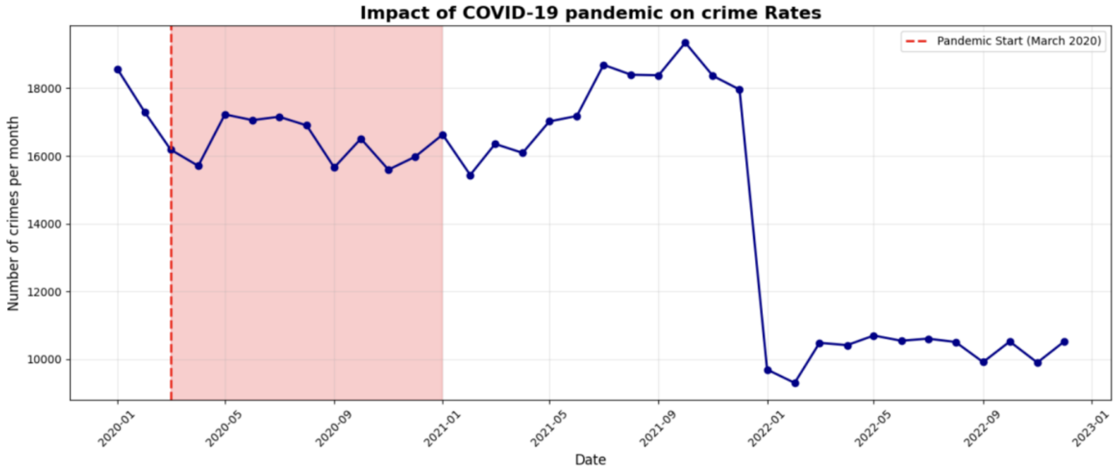
- Key Finding: While widely distributed, victims in the young adult and middle aged demographics (25-45) are consistently involved in the highest number of incidents, correlating with high-volume crimes like simple assault and vehicle-related offenses.

3.7 Correlation with Economic Factors and Impact of Major Events

- Economic Factors: Correlation analysis of internal numerical features showed weak to negligible correlations between features like age and time of occurrence. This limitation indicates that external, socio-economic data is required to fully explain variance in crime rates.



- Impact of Major Events: Analysis of the monthly trend clearly demonstrated the impact of the COVID-19 pandemic. Incident rates saw a noticeable dip during the initial severe lockdown phases of 2020, followed by a recovery and stabilization in 2021, affirming the influence of large-scale societal events on criminal opportunity.



4. Advanced Analysis: Predicting Future Trends

Time-series forecasting was conducted on the total monthly crime counts using the ARIMA(1,1,1) model to anticipate future public safety demands.

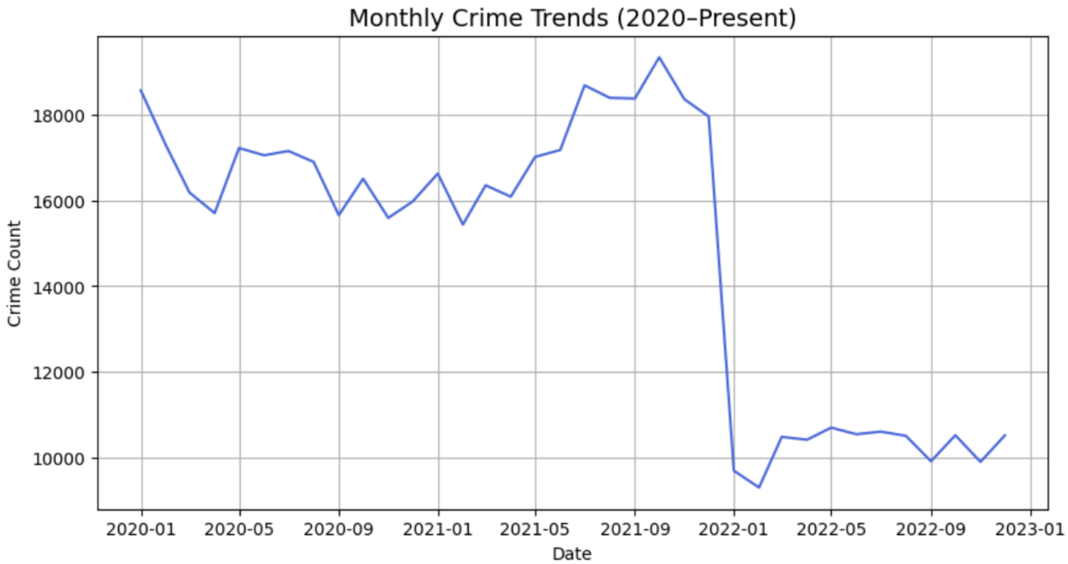
4.1 Stationarity Testing

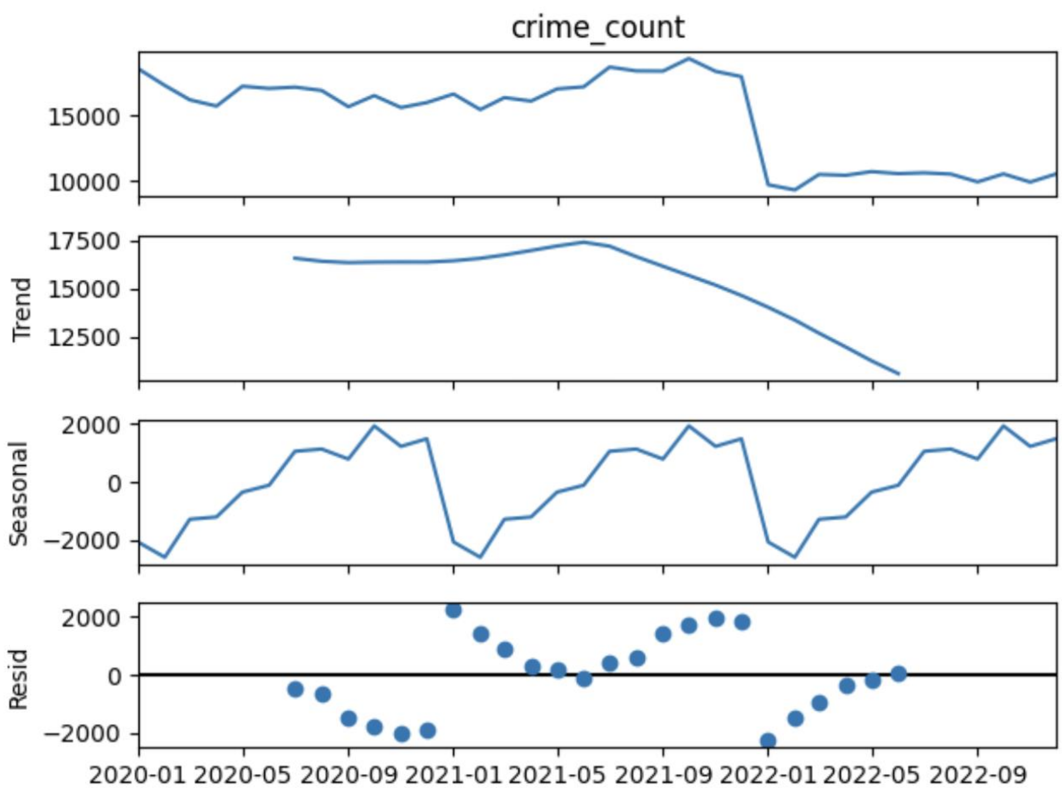
The Augmented Dickey-Fuller (ADF) test indicated the series was non-stationary. Therefore, a first-order differencing was applied to stabilize the data for modeling, which is a necessary preprocessing step for ARIMA analysis.

4.2 Forecast Results

The fitted model generated a forecast for the next 12 months, providing predicted values with a 95% confidence interval.

- Key Finding: The average monthly crime count over the last year stood at 10,256.83. The forecast projects a slight upward trend in crime frequency over the coming year. This incremental rise emphasizes the immediate need for law enforcement to adopt proactive planning and preventative strategies.






```
ADF Statistic: -1.331698229369143
p-value: 0.6145484904681282
⚠️ The series is not stationary (differencing will be applied).
```

5. Key Insights and Recommendations

1. Temporal Prioritization: Crime activity is strongly influenced by the calendar, peaking during summer/fall months and on weekend days (Friday/Saturday).
2. Geographic Focus: The 77th Street and Central districts are persistent high risk hotspots, demanding sustained resource concentration.
3. Specific Offense: The dominance of Vehicle Theft mandates specialized, targeted security and public awareness campaigns.
4. Future Outlook: The predictive modeling reinforces the conclusion that police forces must prepare for and manage a gradually rising crime rate.