Crime Severity Prediction -LA City

Presented By
Anvesh Reddy Pasula
Jaya Rupesh Maram
Dixith Kumar Bandari
Nikhila Cheela
Nasir Sohail Shaik

Introduction

Our project aims to analyze crime patterns in Los Angeles, identify high-risk neighborhoods, and predict future crime trends. By leveraging location and incident data, we developed models to uncover hot spots, forecast severity, and provide actionable insights. This analysis supports proactive safety measures and helps families make informed housing decisions.

Business Use Case:

- Law Enforcement & Public Safety
- Housing Advisory
- Urban Planning

Scope of data

- The Analysis is based on the January 1, 2020, to December 31, 2023.
- Focuses mainly on the historical data regarding crimes in the LA.

Data Acquisition

- The primary data source for this project will be the "LAPD Crime Data."
- All these records enable crime analysis.

Link - LAPD Crime Data - 2020 to 2023

Tools And Technologies :

- Programming Languages: Python serves as the primary language.

- Libraries and Frameworks: Utilize tools like Pandas, NumPy, sklearn, Matplotlib, Seaborn, and Folium etc.

Dataset

- The dataset contains details on LAPD crimes, including crime dates, crime descriptions, weapon descriptions, premise descriptions, etc . It also includes demographic information on age, gender, and race.

- Number of Rows : 892,934
- Number of Columns : 28

Few Column Names

- Date Rptd
- DATE OCC
- TIME OCC
- AREA
- AREA NAME
- Part 1-2
- Crm Cd
- Crm Cd Desc
- Vict Age
- Vict Sex
- Premis Cd
- Premis Desc
- Weapon Used Cd
- Weapon Desc
- Status Desc
- LOCATION
- LAT
- LON

Sample Data

DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	...	Status	Status Desc	Crm Cd 1	Crm Cd 2	Crm Cd 3	Crm Cd 4	LOCATION	Cross Street	LAT	LON
3/1/2020 0:00	2130	7	Wilshire	784	1	510	VEHICLE - STOLEN	...	AA	Adult Arrest	510.0	998.0	NaN	NaN	1900 S LONGWOOD AV	NaN	34.0375	-118.3506
2/8/2020 0:00	1800	1	Central	182	1	330	BURGLARY FROM VEHICLE	...	IC	Invest Cont	330.0	998.0	NaN	NaN	1000 S FLOWER ST	NaN	34.0444	-118.2628
11/4/2020 0:00	1700	3	Southwest	356	1	480	BIKE - STOLEN	...	IC	Invest Cont	480.0	NaN	NaN	NaN	1400 W 37TH ST	NaN	34.0210	-118.3002
3/10/2020 0:00	2037	9	Van Nuys	964	1	343	SHOPLIFTING-GRAND THEFT ($950.01 & OVER)	...	IC	Invest Cont	343.0	NaN	NaN	NaN	14000 RIVERSIDE DR	NaN	34.1576	-118.4387
8/17/2020 0:00	1200	6	Hollywood	666	2	354	THEFT OF IDENTITY	...	IC	Invest Cont	354.0	NaN	NaN	NaN	1900 TRANSIENT	NaN	34.0944	-118.3277

Handling Missing Data

- The initial data analysis identified significant missing values in some columns - 'Vict Sex', 'Weapon Used Cd', and 'Weapon Desc'.
- Handling Missing Values:
    - Premis Desc: Rows with missing values in the 'Premis Desc' column were dropped to ensure data completeness.
    - Weapon Description: Similarly, missing values in 'Weapon Desc' were also removed to maintain data integrity.
- Filtering Data: The dataset was further filtered to retain only the records where 'Vict Age' is greater than zero. This step ensured the analysis was based on realistic age values.

Categorization and Severity

- Due to the extensive variety of crime types, weapon types and premises in the dataset, they were categorized into broader categories with assigned severity levels based on their importance and impact:

Crime Types

Crime Category	Severity Level
Violent Crimes	10
Property Crimes	7
Drug/Alcohol Related	5
Sex Crimes	9
Theft/Fraud	6
Traffic Violations	4
Weapons Violation	8
Non-Violent Miscellaneous	3
Domestic Violence	7
Serious Violations	9

Weapon Types

Weapon Category	Severity Level
Firearms	10
Knives/Bladed Objects	8
Blunt Objects	6
Personal Weapons/Physical Force	4
Explosives/Flammable Objects	9
Chemical/Nontraditional Weapons	5
Threats	3
Miscellaneous Objects	2
Simulated/Toy Weapons	1
Unknown Weapon	2
Unknown or Miscellaneous	2

Premise Types

Premise Category	Severity Level
Residential	10
Business/Commercial	8
Transportation	6
Health Services	4
Educational	9
Government/Public Service	5
Recreational	3
Financial	2
Accommodation	1
Dim light areas	2

Categories and their Counts

```
df_1['Crime Description'].value_counts()
```

Property Crimes	545092
Violent Crimes	268115
Non-Violent Miscellaneous	32596
Theft/Fraud	23377
Sex Crimes	4261
Domestic Violence	3682
Serious Violations	2402
Traffic Violations	404
Drug/Alcohol Related	39
Weapons Violation	37

```
df_1['Weapon Description'].value_counts()
```

Weapon Description	
No Weapon	583043
Personal Weapons/Physical Force	166640
Firearms	36268
Unknown Weapons	33941
Threats	22724
Knives/Bladed Objects	19388
Miscellaneous Objects	18500
Blunt Objects	4970
Chemical/Nontraditional Weapons	3919
Explosives/Flammable Objects	1830
Simulated/Toy Weapons	1163

Name: count, dtype: int64

```
]: df_1['Premise Description'].value_coun
```

]:

Premise Description	
Transportation	383374
Residential	313237
Business/Commercial	135131
Dim light areas	19308
Government/Public Service	13845
Educational	8656
Accommodation	8201
Recreational	560

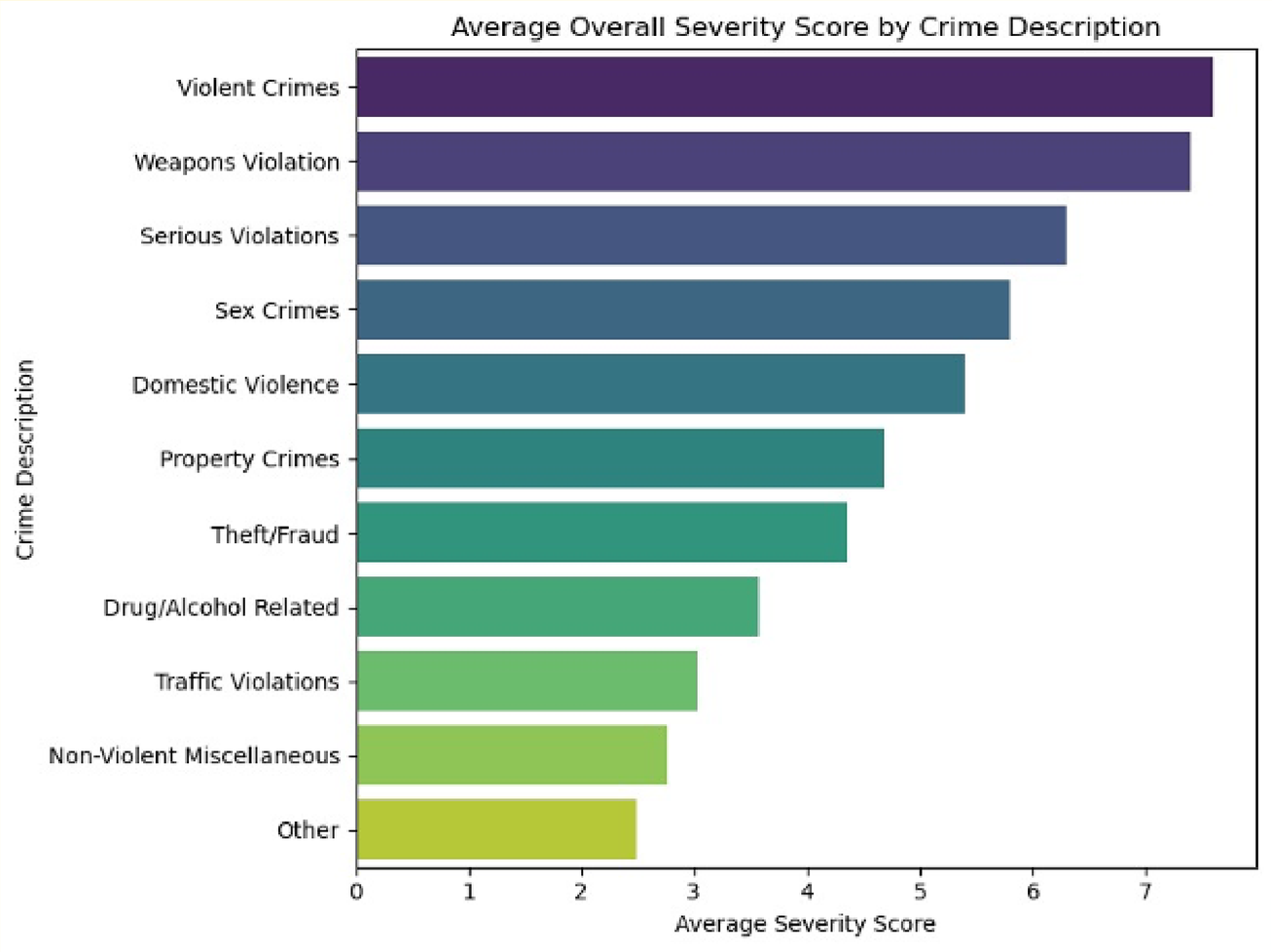
Name: count, dtype: int64

Severity Score allocation

- Incident severity computed using weighted technique:
  - Crime severity weighted highest at 50%.
  - Weapon participation weighted at 30%.
  - Premise location severity weighted at 20%.
- Comprehensive approach combines multiple factors into single severity measure.
- Facilitates prioritization of response activities and resource allocation.

Crime Description	Severity_Crime	Weapon Description	Severity_Weapon	Premise Description	Severity_Premise	Severity
Property Crimes	7	No Weapon	0.0	Transportation	5.0	4.5
Property Crimes	7	No Weapon	0.0	Transportation	5.0	4.5
Property Crimes	7	No Weapon	0.0	Residential	7.0	4.9
Property Crimes	7	No Weapon	0.0	Business/Commercial	6.0	4.7
Property Crimes	7	No Weapon	0.0	Transportation	5.0	4.5

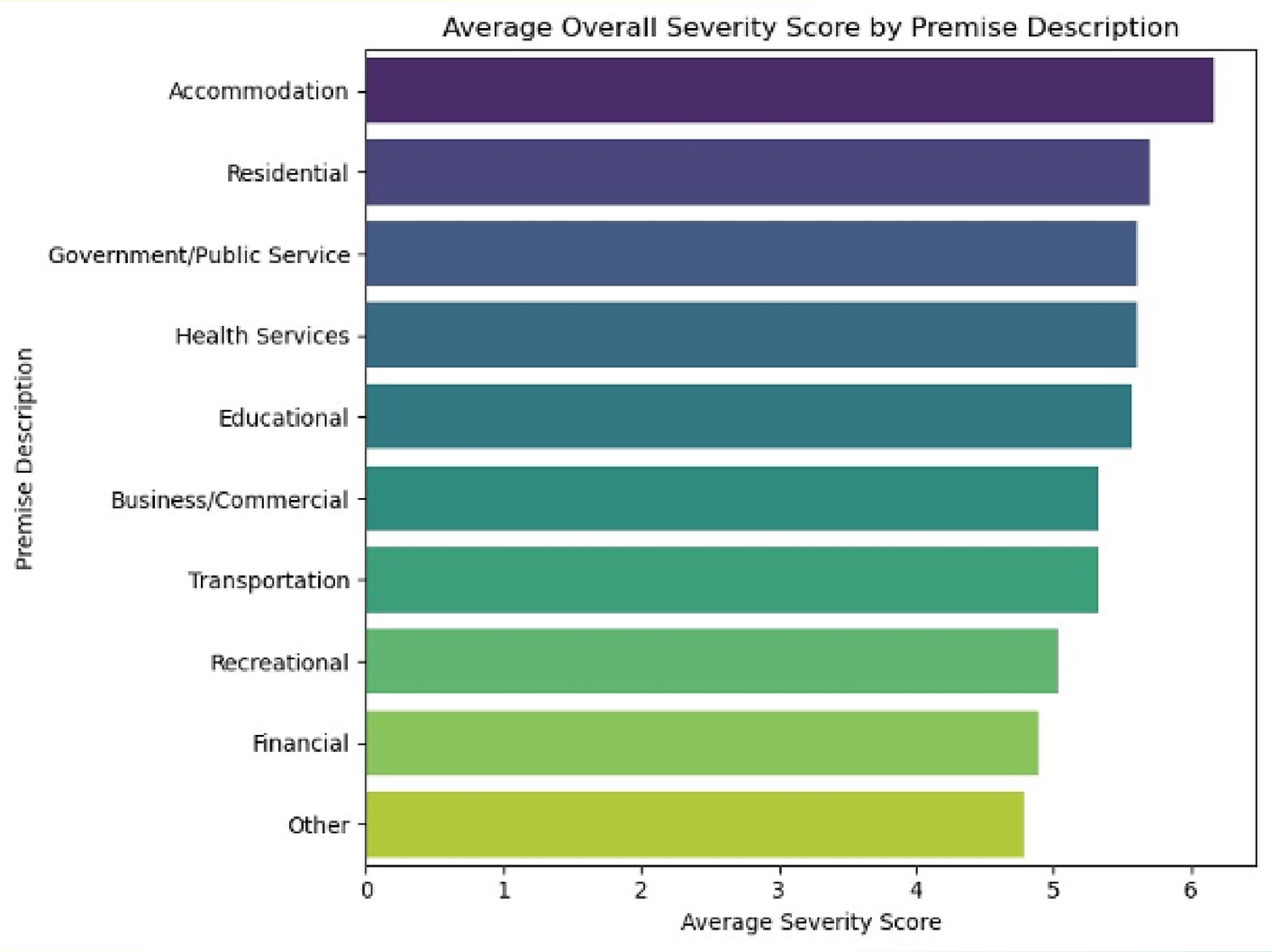
Average overall severity by crime description



Insights :

- Violent crimes and weapons violations have the highest severity scores.
- Non-violent crimes like traffic violations have the lowest scores.

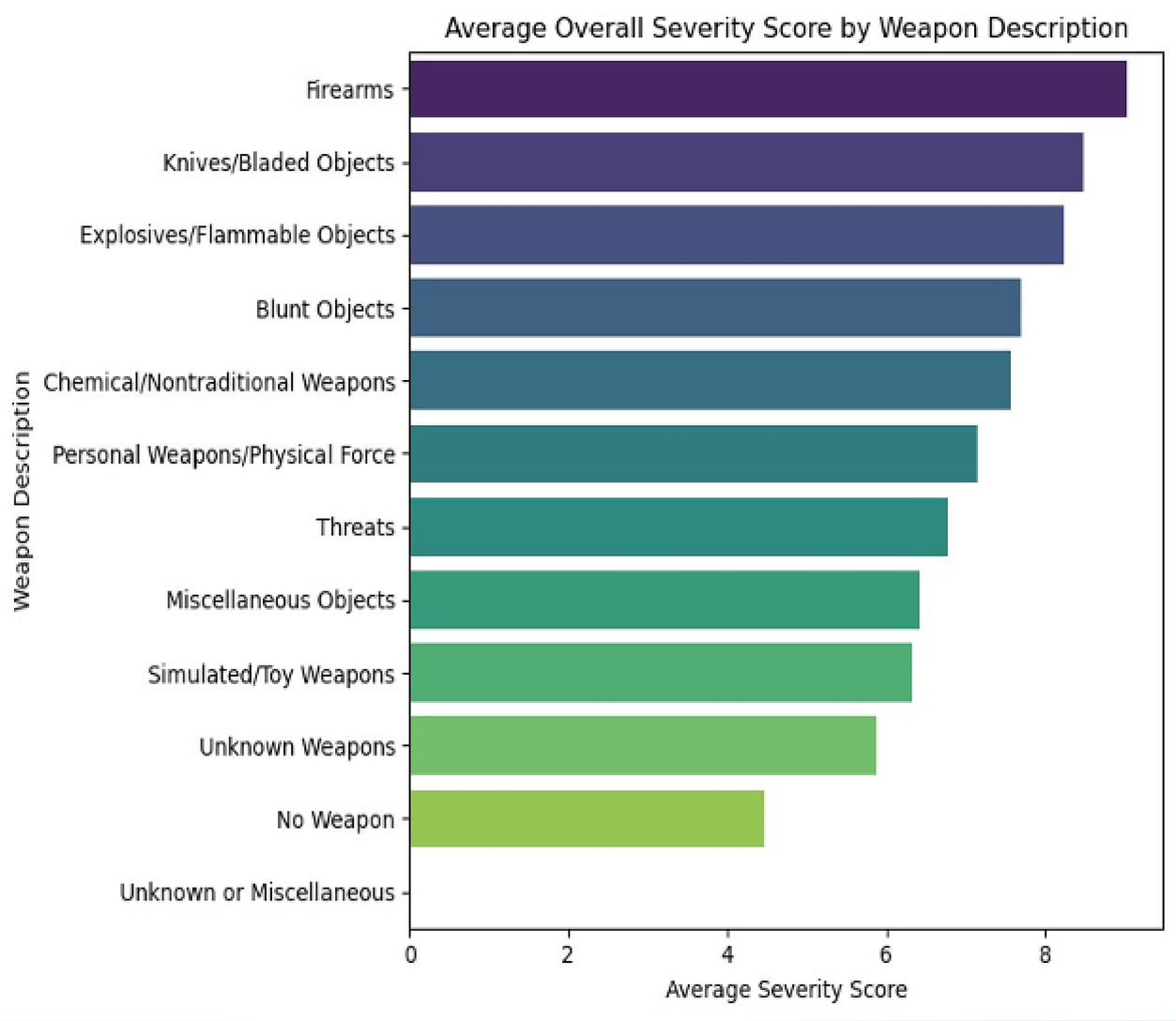
Average overall severity by premise description



Insights :

- Accommodation and residential premises have the highest average severity scores.
- Financial and "Other" premises have the lowest scores.

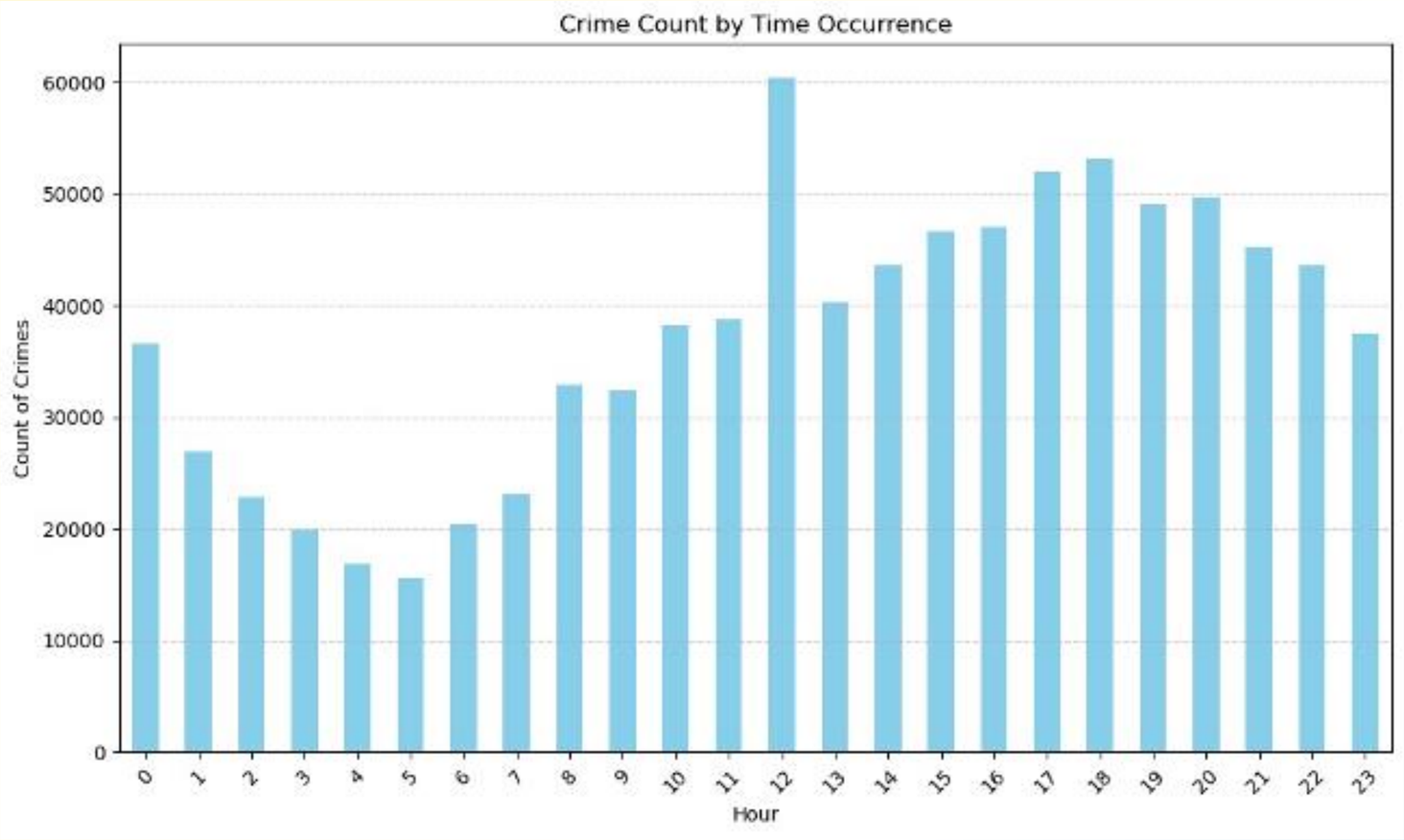
Average overall severity by weapon description



Insights :

- Firearms and knives have the highest average severity scores.
- Toy/simulated weapons and no weapons have the lowest scores.

Crime count by time occurence



Insights :

- Crime peaks around noon and early evening.
- Crime is lowest from 3 to 6 AM.

Clustering

Objective

- The main objective is to identify and distinguish crime hot spots in Los Angeles.
- Helps law enforcement focus efforts on high-risk areas for more effective crime prevention.
- Provides valuable housing location advice to families seeking safer neighborhoods.

Feature Selection

- Latitude/Longitude: Map crime incidents geographically to understand spatial distribution.
- Number of Crimes: Measure the overall frequency of criminal activities.

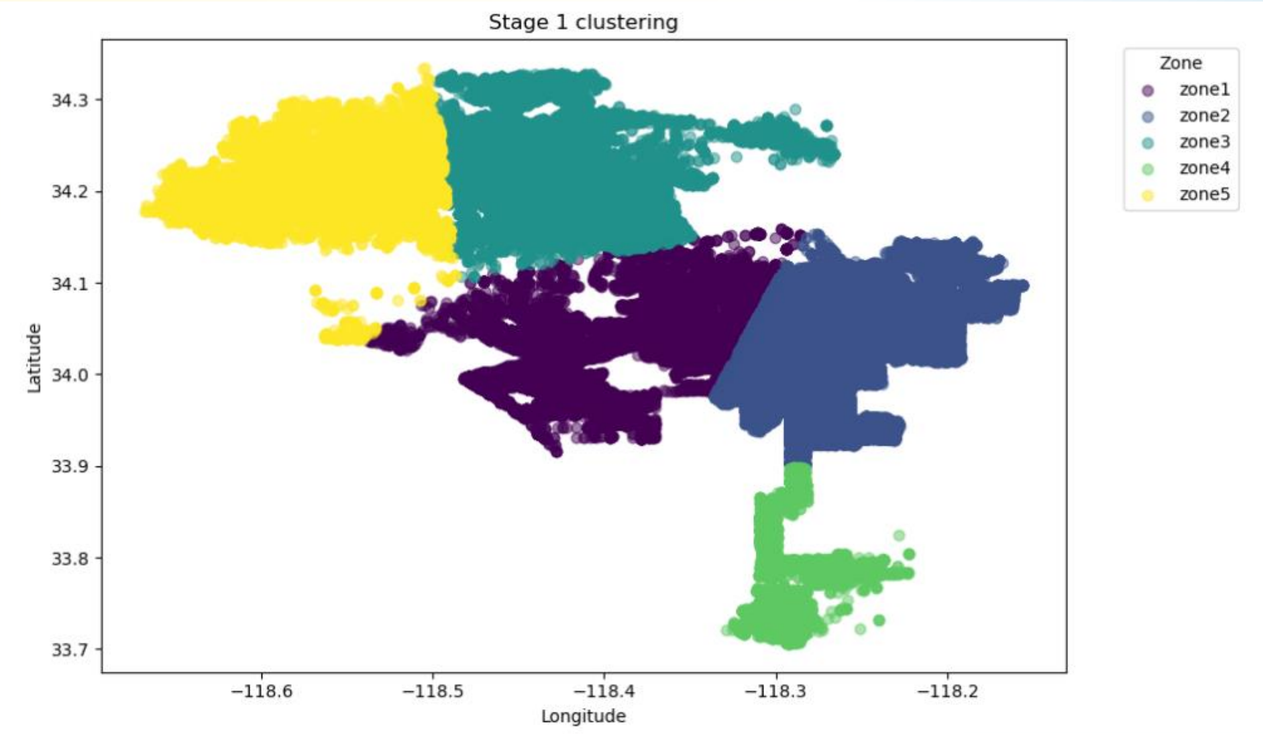
Clustering Approach:

- Algorithm: Two-stage hierarchical clustering for nested crime hot spots.
  - Stage 1: Divide LA into five main regions.
  - Stage 2: Subdivide each region into 20 sub-zones for granular insights.

Stage 1 Clustering

Key Insights

From the graph, we see that the clustering analysis successfully divided Los Angeles into five distinct regions using the K-means algorithm based on latitude, longitude, and the number of crimes. These regions are labeled as zones in our project and will be further subdivided in Stage 2 clustering.



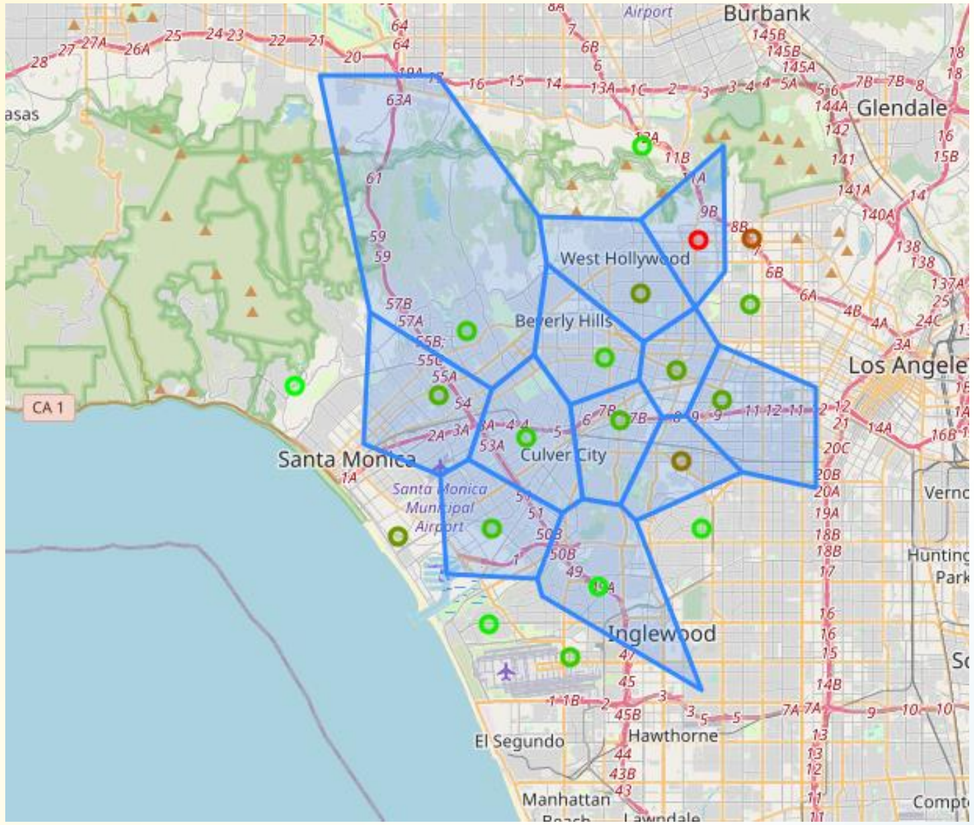
Stage 2 Clustering

In the second stage of clustering, we used the K-means algorithm again, this time to subdivide each of the five main regions into 20 sub-regions. Each sub-region was assigned a normalized score based on the number of crimes within its boundaries.

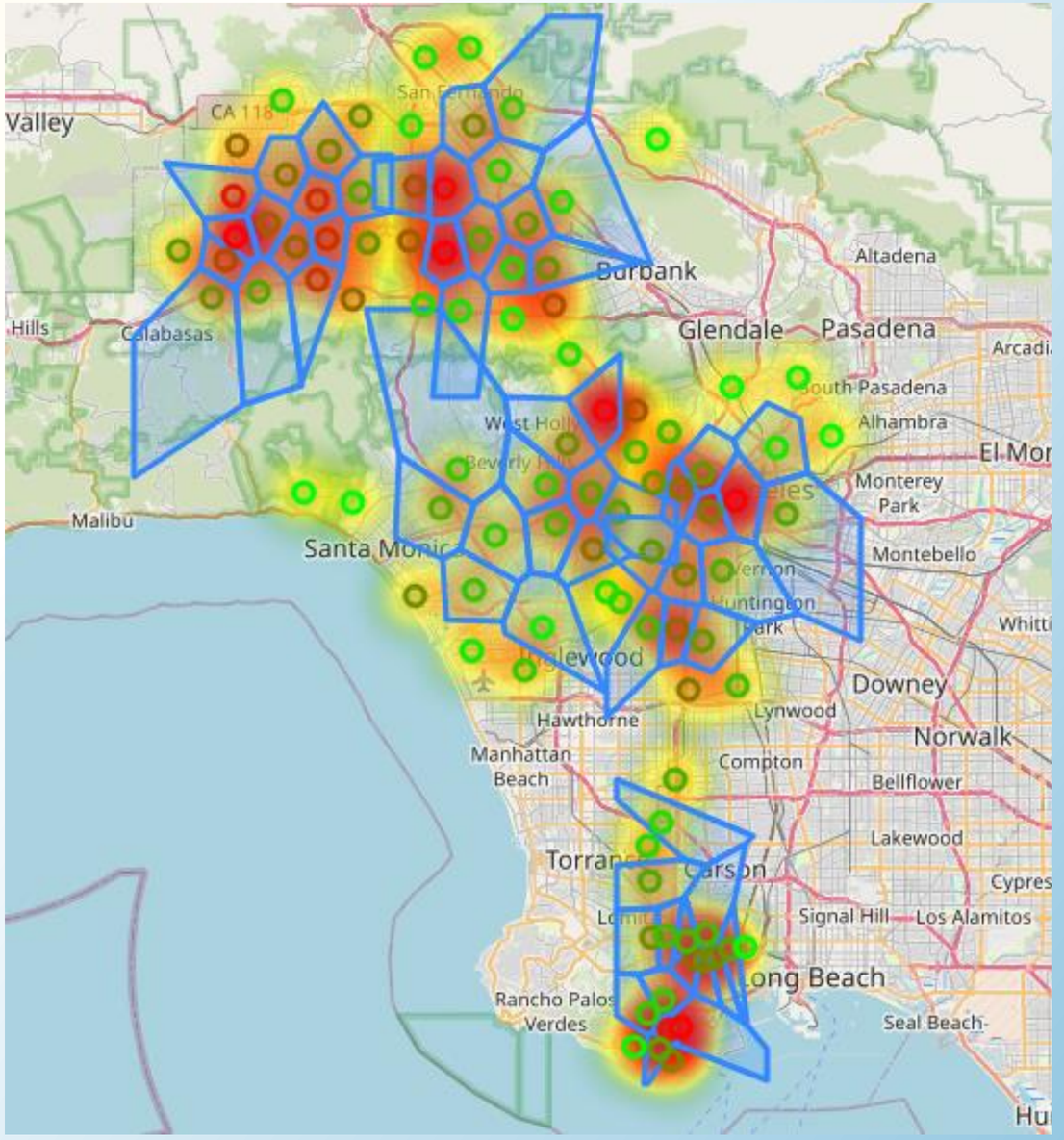
	Cluster_KMeans	LAT	LON	Zone	Crime_Count_Normalized
0	0	34.098753	-118.336987	8066	1.000000
1	1	33.991134	-118.469206	3487	0.413099
2	2	34.018456	-118.344783	4544	0.548577
3	3	34.056085	-118.378298	1829	0.200590
4	4	33.947551	-118.393306	1849	0.203153
5	5	34.065682	-118.439034	1877	0.206742
6	6	34.099141	-118.313856	5498	0.670854
7	7	34.051412	-118.346531	2997	0.350295
8	8	33.994275	-118.428005	2081	0.232889
9	9	34.033448	-118.371220	2369	0.269803
10	10	34.075341	-118.314240	2363	0.269034

- The subdivision of main zones into 20 sub-regions provides a more granular view of crime hot spots.
- Normalized scores enable a consistent comparison across all sub-regions, highlighting those with the highest crime activity for further investigation.

Spatial Analysis



Zone 1 clustering



Combined Los angels Crime Hope Spot Map

Feature Selection and Importance Analysis

Selected Columns:

Categorical features :

'AREA', 'Vict Sex', 'Status Desc', 'Hour', 'Month', 'Day', 'Year', and 'Crime Description'.

Numeric features :

'Vict Age' and 'Weapons Used'.

Feature Importance Analysis:

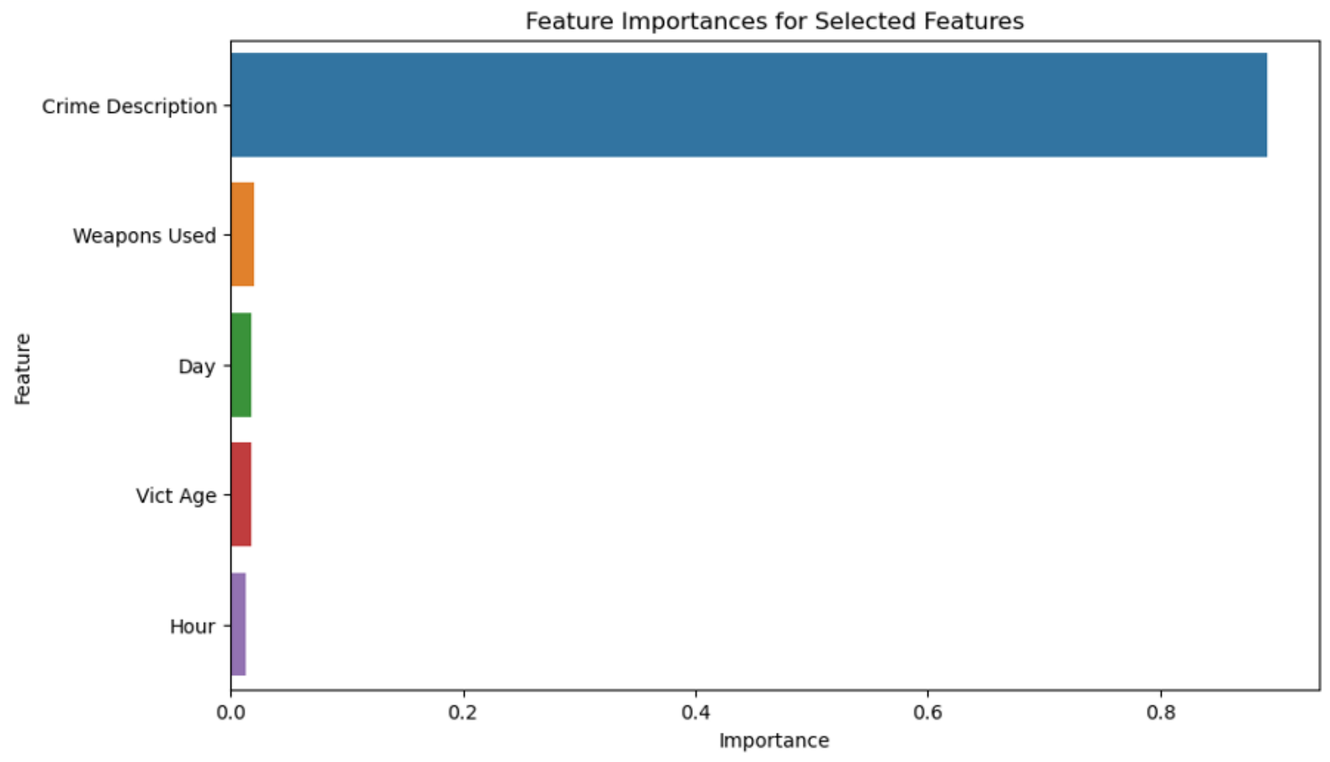


Fig - Feature Importance's for Selected Features [Using Random Forest]

Key Insights:

- Crime Description: Most significant predictor of crime severity.
- Weapons Used: Strongly correlates with increased crime severity.
- Day: Shows moderate importance, possibly reflecting specific crime trends on certain days.

Model Selection

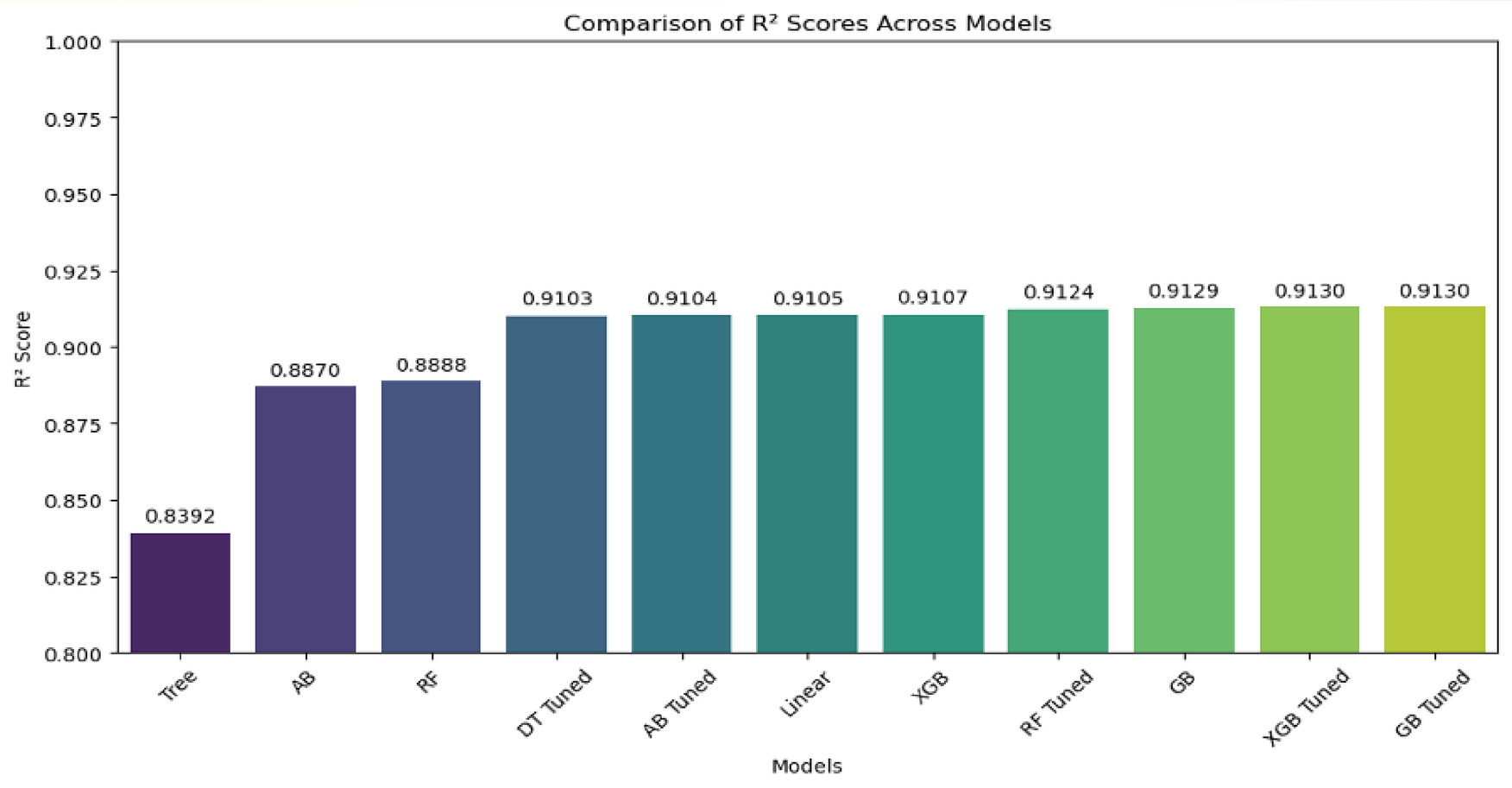
Predicting the Types of Crimes

Target = 'Severity'
Features=['Vict Age', 'Weapons Used', 'Hour', 'Day', and 'Crime Description']

Models Used :

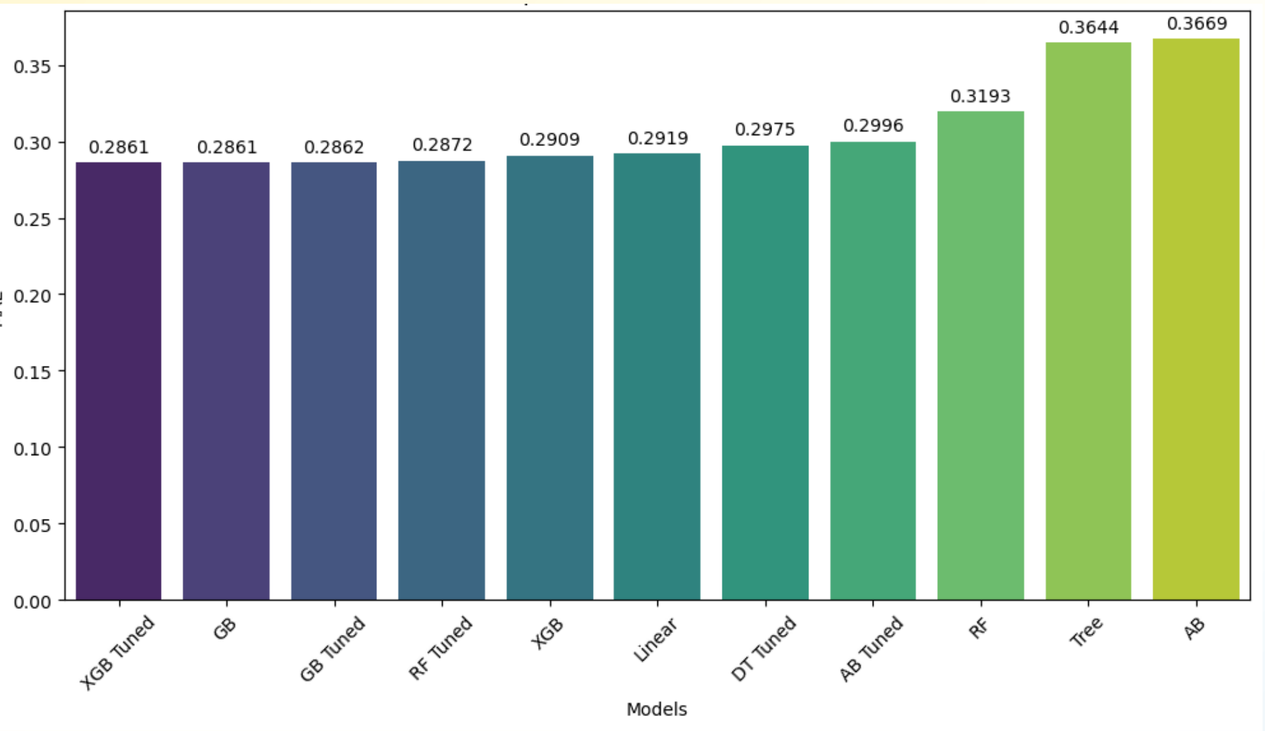
- Base Models
  - Linear Regression
- Ensemble Models
  - Random Forest Regressor
  - Decision Tree Regressor
- Boosting Models
  - AdaBoost Regressor
  - XGBoost Regressor
  - Gradient Boosting Regressor

Comparison of R^2 Scores Across Models

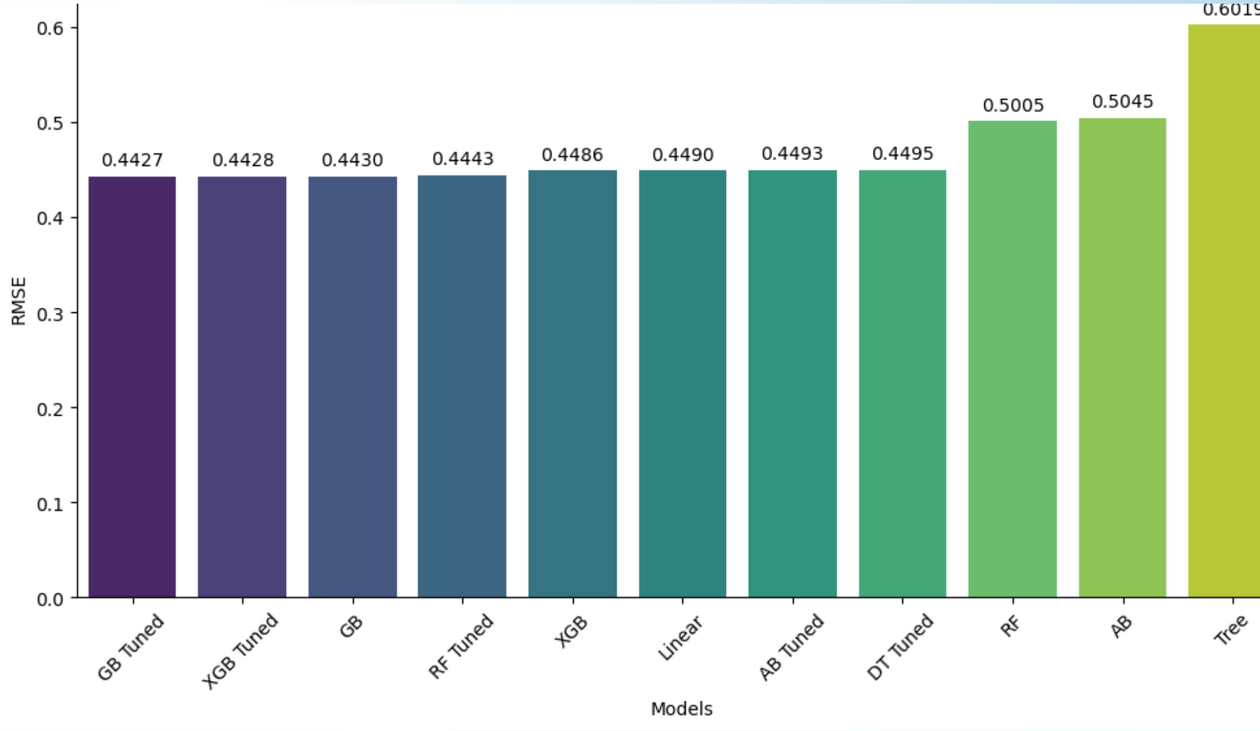


- Gradient Boosting and XGBoost proved to be the best models for this analysis due to their superior performance after tuning.
- Both models effectively captured the non-linear relationships present in the dataset and demonstrated their ability to adapt to the data structure.

Comparison of MAE Across Models



Comparison of RMSE Across Models



Gradient Boosting Regressor

Performance Metrics

Mean Squared Error (MSE): 0.196011

Root Mean Squared Error (RMSE): 0.442731

Mean Absolute Error (MAE): 0.286162

R^2 Score: 0.912987

Tuning: Grid search optimized parameters like learning rate, maximum depth, and the number of estimators, significantly improving the model's performance.

Insights: After being fine-tuned, the model showed remarkable predicted accuracy, indicating its capacity to adjust to intricate interactions within the dataset.

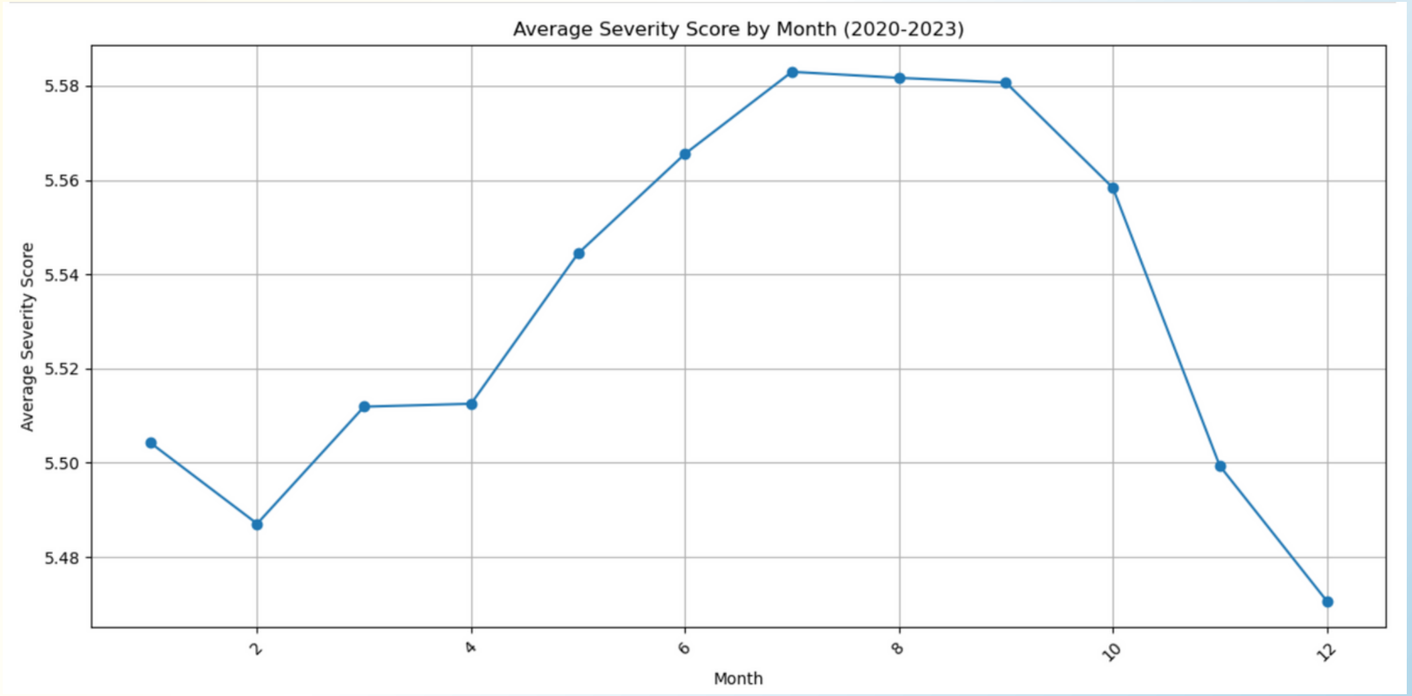
Time Series Analysis

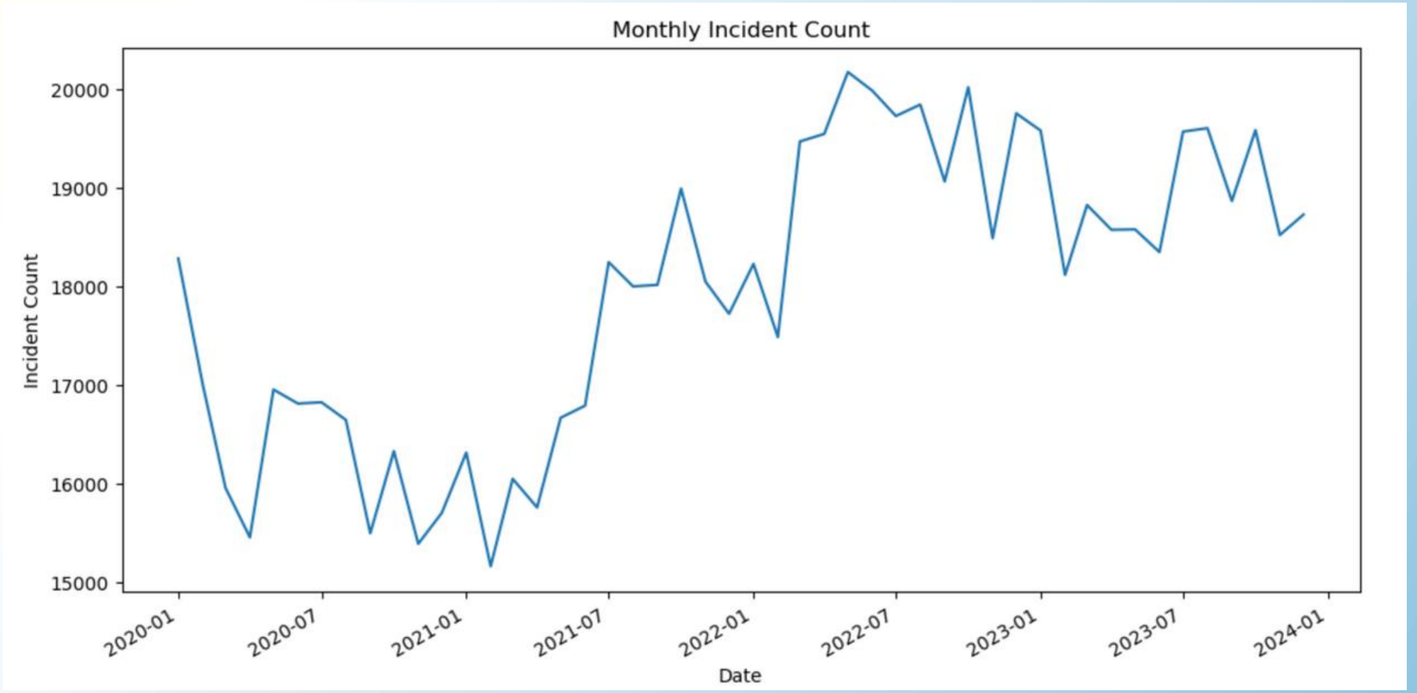
Time Series Analysis is instrumental in leveraging historical incident data to predict future trends, understand underlying patterns, and make informed decisions. By applying these methods, government can enhance operational efficiency, optimize resource allocation, and improve overall preparedness in managing incidents.

Use Case:

1. Forecasting Incident Counts: Time series analysis helps in predicting future incident counts based on historical data. This can be valuable for resource allocation, preparedness, and decision-making.
2. Identifying Trends and Seasonality: Time series methods can uncover underlying trends and seasonal patterns in incident data. This understanding is crucial for strategic planning and resource optimization.
3. Modeling Dependencies: Time series analysis can capture dependencies and correlations within the data, providing insights into how incidents evolve over time and identifying potential influencing factors.

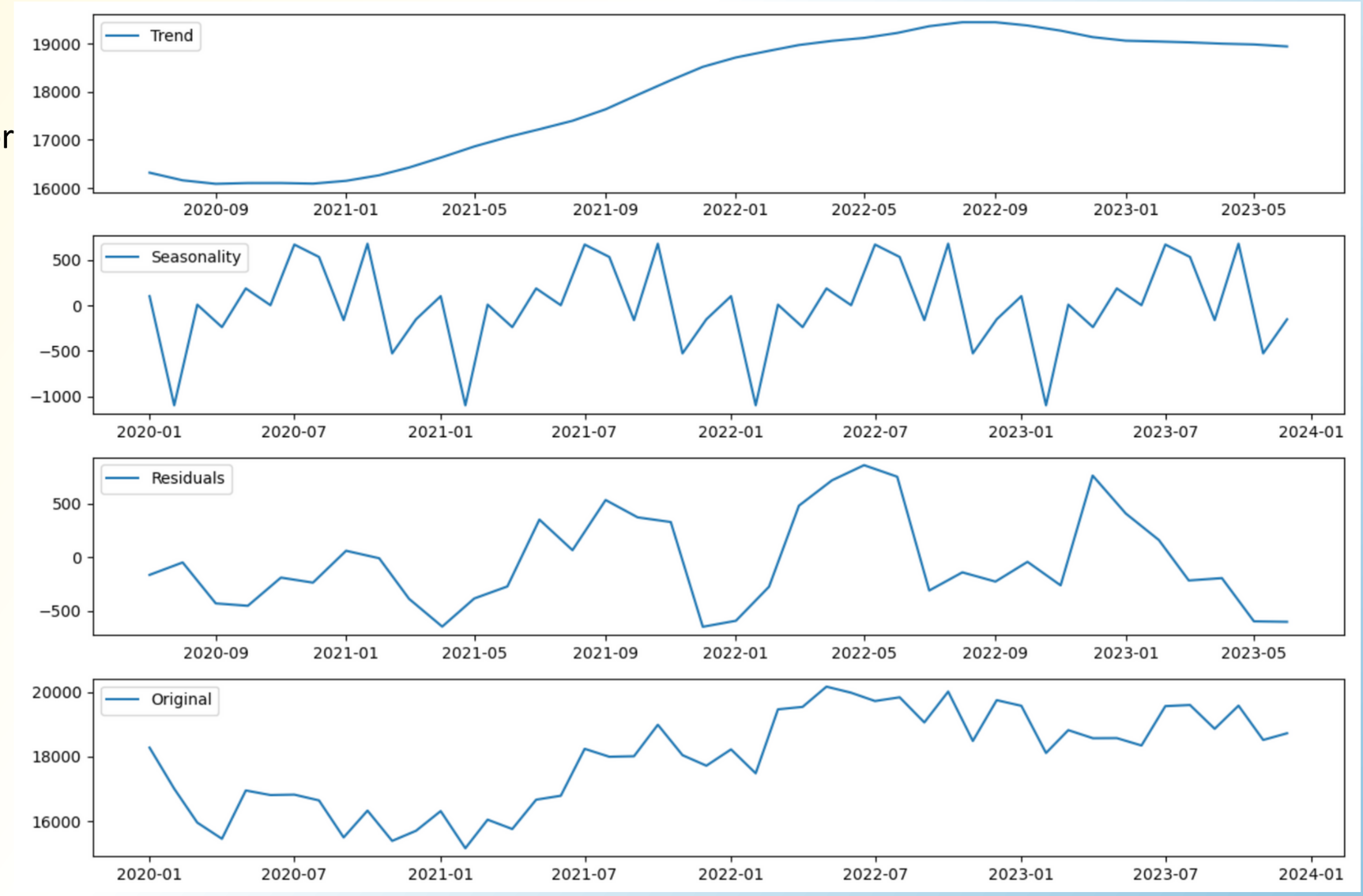
The "average severity score vs month" plot involves interpreting trends, identifying patterns, and translating these insights into actionable recommendations for incident management and decision-making. The "incident count vs date" plot, gives valuable insights into temporal variations in incident frequency and use this information to support data-driven decision-making and emergency response strategies.





The trend, seasonality, residuals, and original time series graphs are obtained from a seasonal decomposition analysis.

1. Trend: Identify long-term changes or movements in the data.
2. Seasonality: Highlight recurring patterns or cycles within the data.
3. Residuals: Assess the randomness or unexplained variability in the time series.
4. Interpretation: Use insights from trend, seasonality, and residuals to understand the data's behavior comprehensively.

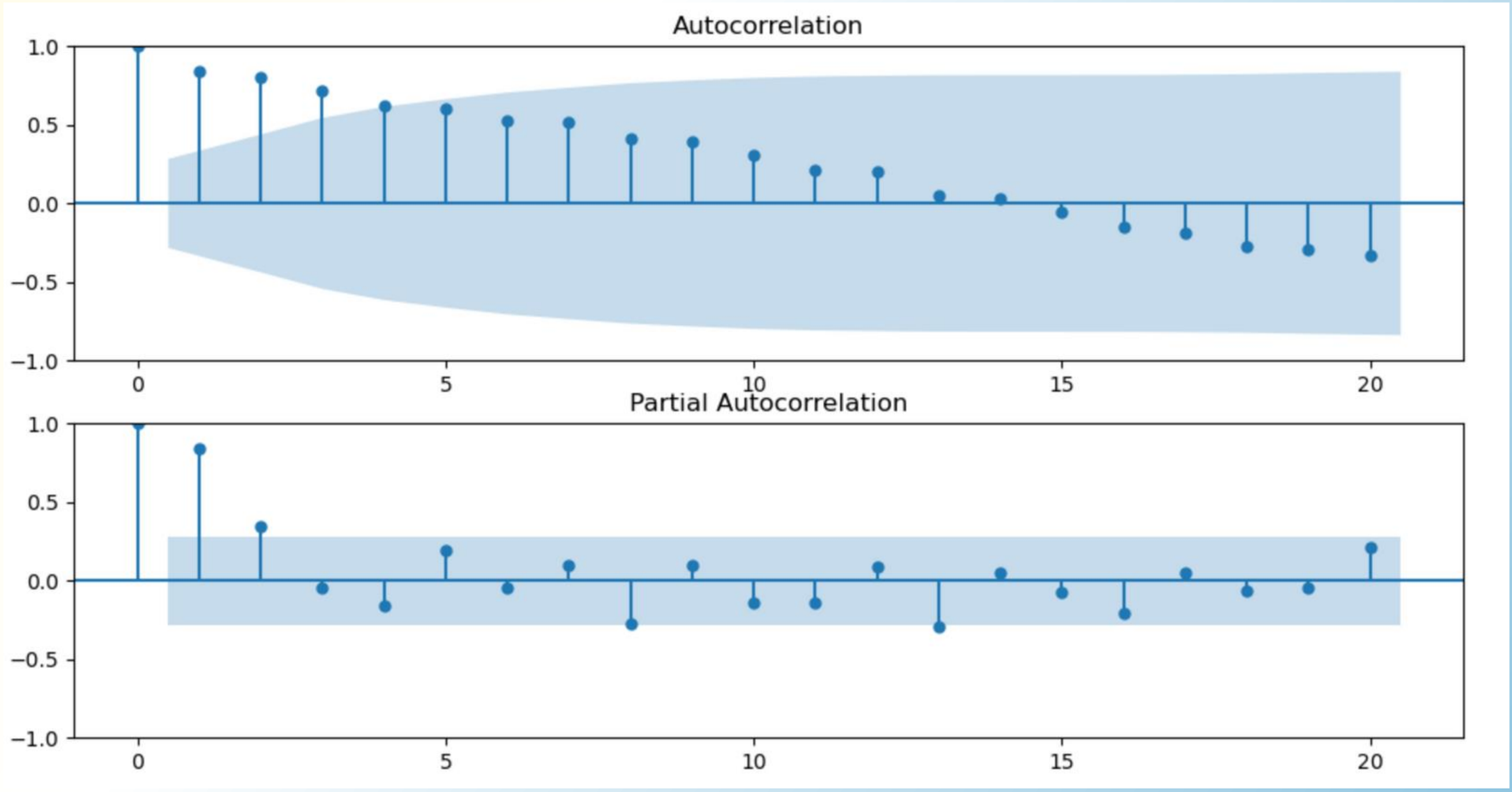


Autocorrelation Function (ACF) Plot:

The ACF plot shows the correlation coefficients between the time series and its lagged values up to a specified number of lags. We identify lag values where autocorrelation is significant, indicating potential seasonal patterns or autoregressive relationships.

Partial Autocorrelation Function (PACF) Plot:

The PACF plot shows the correlation coefficients between the time series and its lagged values, while removing the effect of intermediate lags. We can determine the order of autoregressive terms (p) and moving average terms (q) based on significant spikes or cutoffs in the plot.

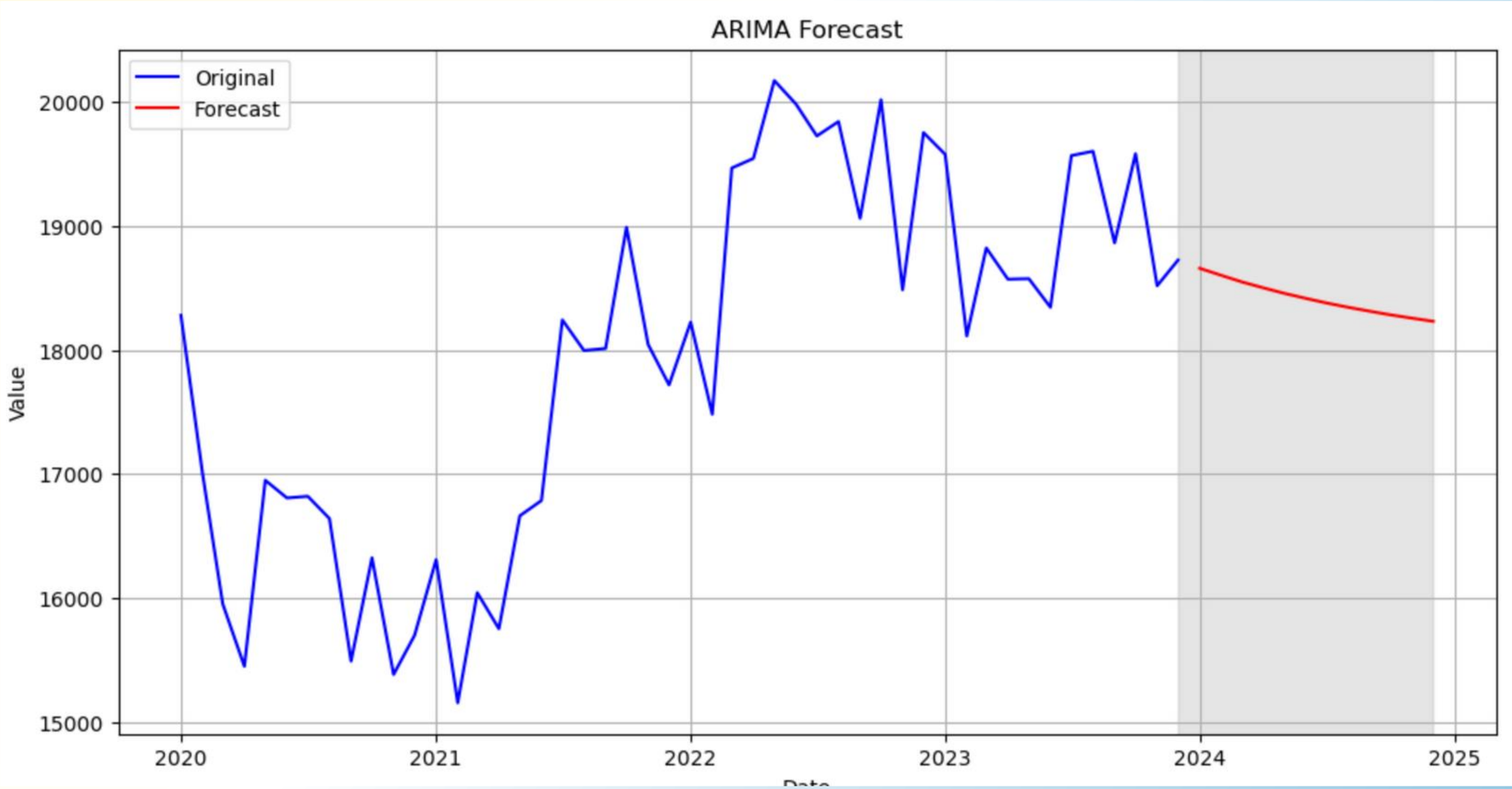


In our case, the p,d,q values are 1,0 , 1 respectively.

ARIMA MODEL

The ARIMA model was used to forecast severity scores across different months of 2024.

The forecasted values showed a consistent trend with small variations across the months, suggesting a stable pattern in the severity scores over time.

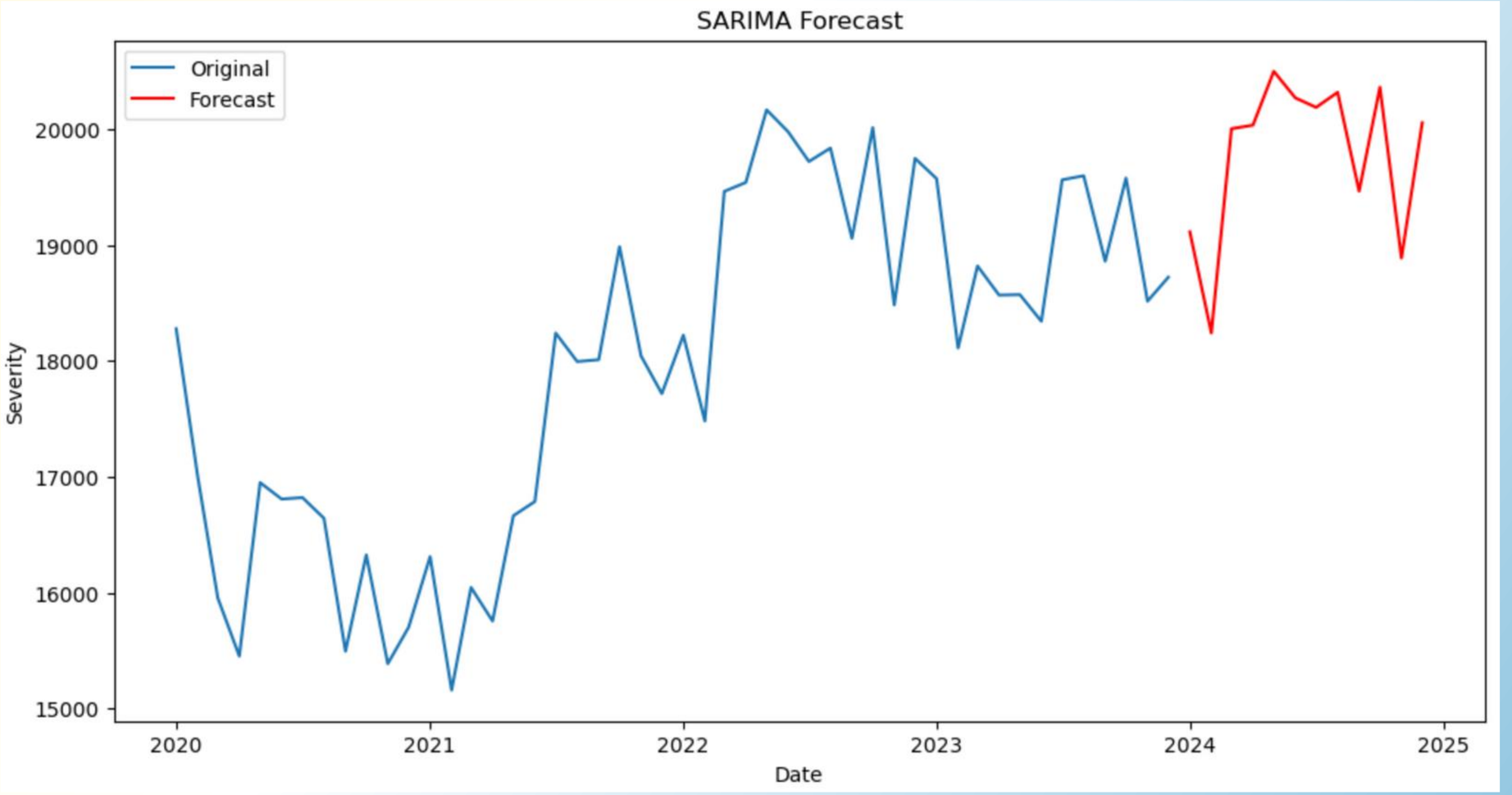


The model seemed to provide reasonable forecasts for the next 12 months based on the historical data, assuming a monthly forecasting interval.

SARIMA MODEL

SARIMA (Seasonal ARIMA) was applied to capture both trend and seasonal components in the time series data.

SARIMA is beneficial when dealing with data exhibiting clear seasonal patterns.



The model's performance can be evaluated based on its ability to capture and predict the seasonal variations observed in the data more accurately compared to a basic ARIMA model.

Conclusion

Through our comprehensive analysis, we achieved a multi-faceted understanding of crime in Los Angeles:

- Clustering Analysis: Grouping crime incidents geographically, we identified and mapped high-severity crime zones, providing law enforcement with valuable insights to prioritize high-risk areas for efficient resource allocation. This data also helps families make informed housing decisions based on neighborhood safety.
- Machine Learning: Predicting the severity of crimes using key features like crime description and weapon used, this model offers actionable insights for law enforcement to anticipate high-risk incidents. Urban planners can leverage these insights to design targeted crime prevention strategies.
- Time Series Analysis: By forecasting future crime trends using historical data, we enabled policymakers and law enforcement to anticipate upcoming patterns for 2024. This understanding allows for proactive, long-term public safety measures.

Together, these analytical models offer a detailed view of crime patterns and trends, enabling the implementation of strategic, effective solutions for safer neighborhoods in 2024 and beyond.