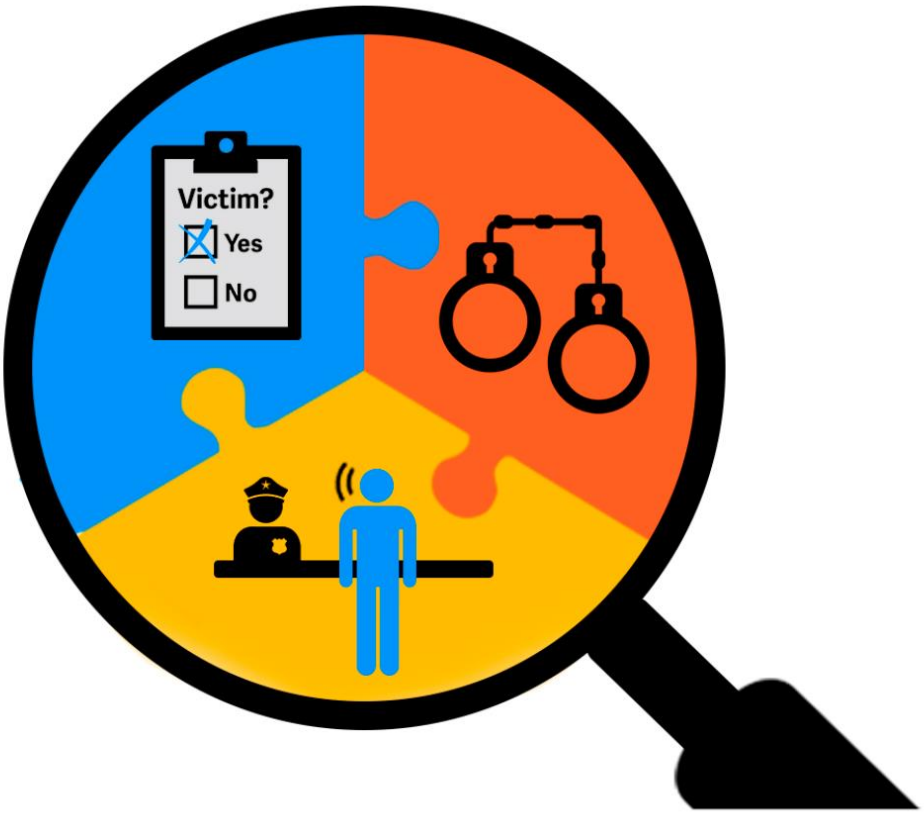


Cleaning and Analyzing Crime Data

IE6400 Foundations Data Analytics Engineering

Group Number 3

Jahnavi Mishra (002724552)

Sahita Bonthu (002823336)

Namrit Sheth (002244393)

Part 1: Introduction and Research Questions

The main objective of crime analysis is to aid the functioning of a police department, which encompasses tasks such as facilitating criminal investigations, apprehending and prosecuting offenders, enhancing patrol operations, implementing crime prevention and reduction tactics, and promoting problem-solving initiatives. Crime analysts provide essential insights to law enforcement agencies, improving their efficiency in crime reduction. They also aid law enforcement executives in making well-informed decisions concerning resource distribution, personnel placement, and strategic planning.

This investigative report aims to comprehensively analyze crime data. It begins by calculating and visualizing annual crime trends, followed by a monthly analysis of average crime occurrences over the years. The investigation identifies the most frequent crime type and compares crime rates between regions or cities using descriptive statistics and visualizations. Furthermore, it incorporates economic data to assess the correlation between economic factors and crime rates. Weekly crime frequencies are analyzed based on the day of the week, and significant events or policy changes during the dataset period are explored for their impact on crime rates. The report employs statistical methods and data visualization techniques to identify outliers and unusual patterns in the dataset. It also examines potential patterns and correlations between demographic factors and specific crime types. Finally, time series forecasting methods, such as ARIMA or Prophet, are utilized to predict future crime trends, with an emphasis on incorporating relevant external factors into the forecasting models.In this project, we have used Python and Jupyter notebook to clean the crime dataset from 2020 to present that we got from catalog.data.gov.

Part 2: Data Sources

The main portion of data used in this analysis was sourced from catalog.data.gov.

The table contains various columns providing information related to reported incidents. The apartment dataset preparation involved:

- DR_NO: This column represents the unique Report Number associated with each incident.
- Date Rptd: It denotes the date when the incident was reported.
- DATE OCC: This column records the date of occurrence of the incident.
- TIME OCC: It specifies the time when the incident occurred.
- AREA: A numeric code is used to represent the specific geographic area where the incident took place.
- AREA NAME: This column provides the name of the area, offering a more human-readable location reference.
- Rpt Dist No: Report District Number indicates the specific district or region where the incident was reported.
- Part 1-2: It categorizes the type of crime as either Part 1 or Part 2.
- Crm Cd: This is the Crime Code, a numerical identifier for the type of crime committed.
- Crm Cd Desc: It provides a description of the crime associated with the Crime Code.
- Mocodes: Modus Operandi describes the method used in committing the crime.
- Vict Age: This column records the age of the victim involved in the incident.
- Vict Sex: It specifies the gender of the victim.
- Vict Descent: This column records the descent or ethnicity of the victim, providing information about their background.
- Premis Cd: Premises Code is a numerical identifier for the type of location where the incident occurred.
- Premis Desc: It provides a description of the premises where the incident took place.
- Weapon Used Cd: If a weapon was used in the incident, this column contains a numerical code representing the type of weapon.
- Weapon Desc: This column offers a description of the weapon used, if applicable.
- Status: It indicates the current status of the case related to the incident.
- Status Desc: This column provides a detailed description of the status of the case.
- Crm Cd 1, Crm Cd 2, Crm Cd 3, Crm Cd 4: These columns may contain additional Crime Codes if multiple crimes were committed during the incident.
- LOCATION: This column specifies the exact location of the incident.
- Cross Street: It records the cross street or nearby location reference.
- LAT: Latitude coordinates of the incident.
- LON: Longitude coordinates of the incident.

Part 3: Results and Methods

The dataset we obtained for our study on crime contained substantial amounts of missing data and superfluous columns. In order to make the data suitable for visualization in our university project, we undertook various data cleaning steps. We removed irrelevant columns, addressed missing values, managed outliers, standardized data types, and readied the dataset for subsequent analysis. The resulting dataset, referred to as "proj1," now includes standardized numerical attributes and encoded categorical variables, rendering it well-prepared for the purposes of Exploratory Data Analysis in our project.

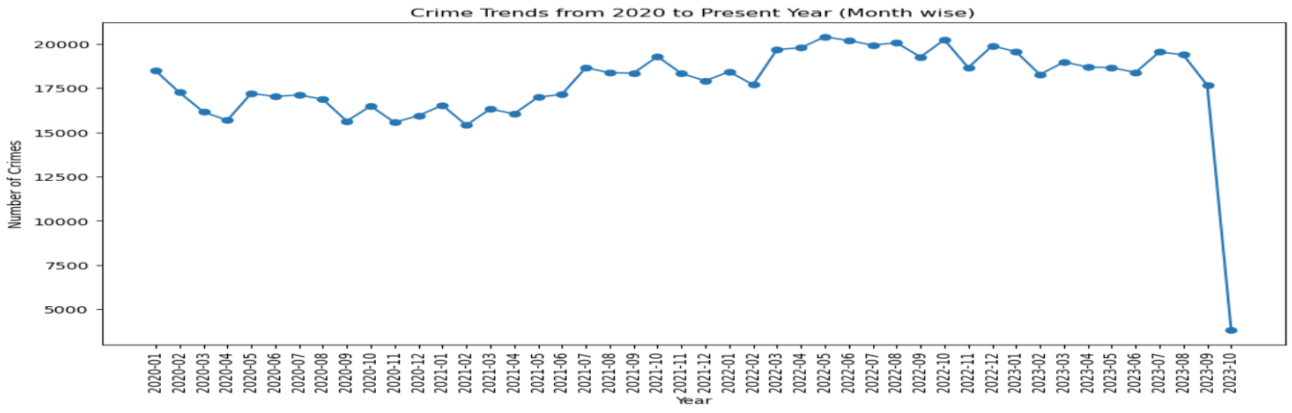
	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	...	LOCATION	Cross Street	LAT	LON
0	10304468	2020-01-08	2020-01-08	0.945293	0.1	Southwest	0.131554	2	0.607565	BATTERY - SIMPLE ASSAULT	...	1100 W 39TH PL	BROADWAY	0.990674	0.003116
1	190101086	2020-01-02	2020-01-01	0.139525	0.0	Central	0.029552	2	0.607565	BATTERY - SIMPLE ASSAULT	...	700 S HILL ST	BROADWAY	0.991600	0.003481
2	200110444	2020-04-14	2020-02-13	0.508482	0.0	Central	0.025739	2	0.868794	SEX OFFENDER REGISTRANT OUT OF COMPLIANCE	...	200 E 6TH ST	BROADWAY	0.991568	0.003541
3	191501505	2020-01-01	2020-01-01	0.733249	0.7	N Hollywood	0.687321	2	0.750591	VANDALISM - MISDEAMEANOR ($399 OR UNDER)	...	5400 CORTEEN PL	BROADWAY	0.995171	0.002239
4	191921269	2020-01-01	2020-01-01	0.175573	0.9	Mission	0.904194	2	0.744681	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VA...	...	14400 TITUS ST	BROADWAY	0.996665	0.001861

5 rows × 28 columns

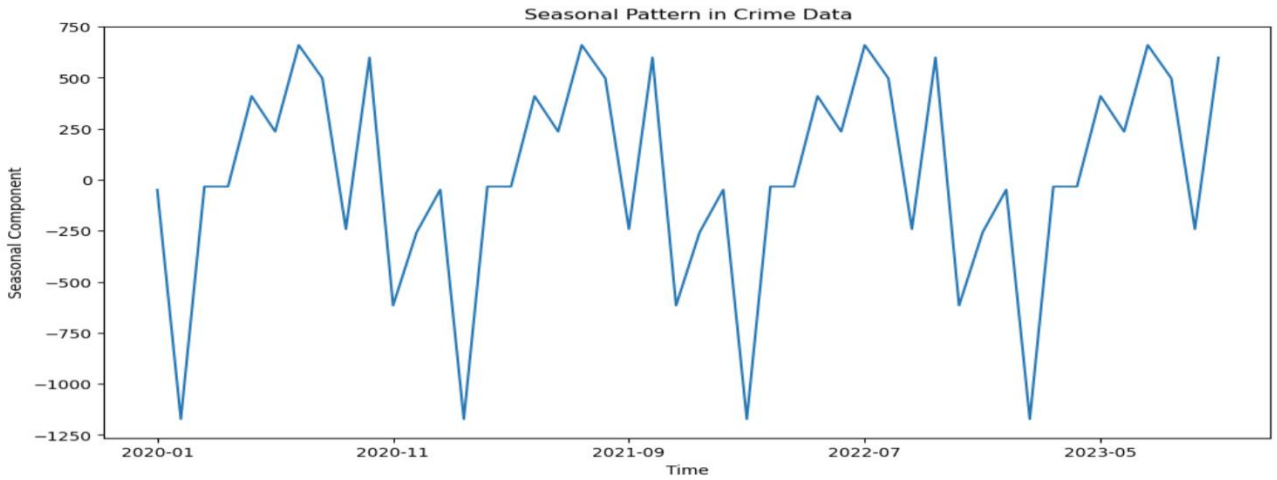
Vict Sex_encoded	Vict Descent_encoded	Premis Desc_encoded	Weapon Desc_encoded	Status_encoded	Status Desc_encoded
2	2	255	65	1	1
7	7	254	75	3	2
18	18	235	65	0	0
17	17	207	65	3	2
18	18	23	65	3	2

We initiate by extracting the year component from the 'DATE OCC' column within the 'proj1' dataset, resulting in the creation of a 'Year' column. Then we proceed to tally and arrange the occurrence of crimes on a yearly basis. Subsequently, the data is converted into a string format suitable for plotting, and a line graph is generated to depict the trends in crime over the years, starting from 2020 up to the current year. Following this, a 'Month' column is crafted by extracting both the month and year from the 'DATE OCC' column. We then tabulates and organizes crime occurrences on a monthly basis, which are also presented in a line graph. Seasonal decomposition is executed using the statsmodels library to dissect the monthly crime data into trend, seasonal, and residual elements, primarily aimed at identifying recurrent patterns. The code additionally pinpoints the most prevalent crime type in the dataset, filters the data to exclusively study this specific crime type on a monthly scale, and generates a line graph to visualize its trends across time. Ultimately, we computes and showcases crime rates for various cities, utilizing the 'AREA NAME' column, and these rates are graphically compared via a bar chart to reveal disparities in crime rates across different regions.

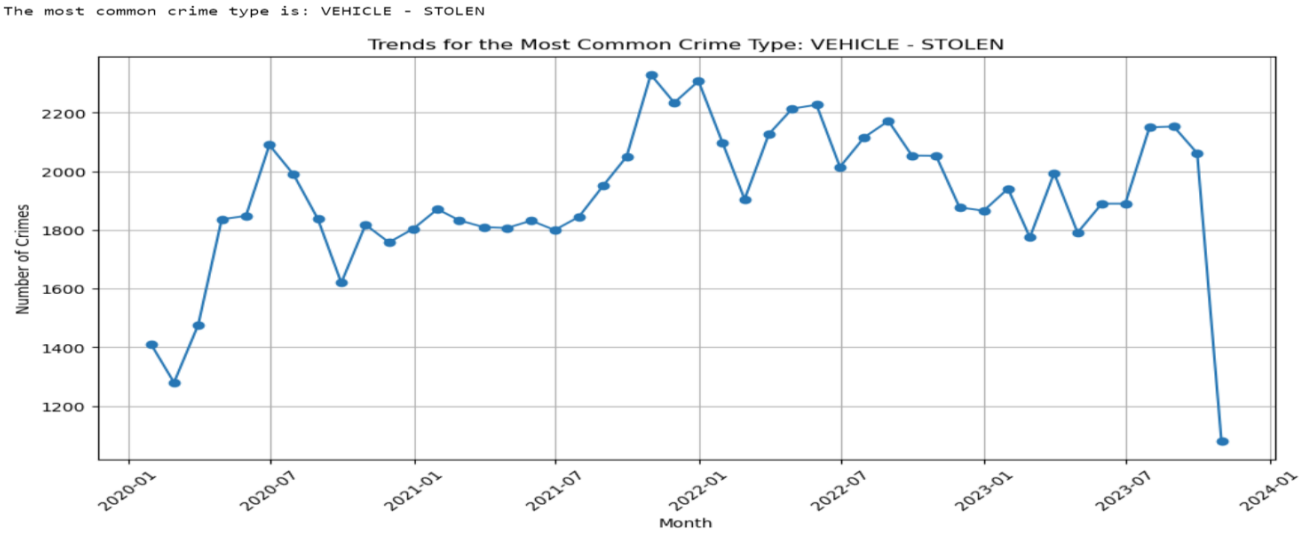
We have created a line plot that visualizes the yearly trends in crime data, helping to identify patterns and changes over time.



A line plot is generated that shows the trend in the number of crimes over the years, starting from 2020 and extending to the present year. The line plot visually displays whether crime rates have increased, decreased, or remained relatively stable over this time period.

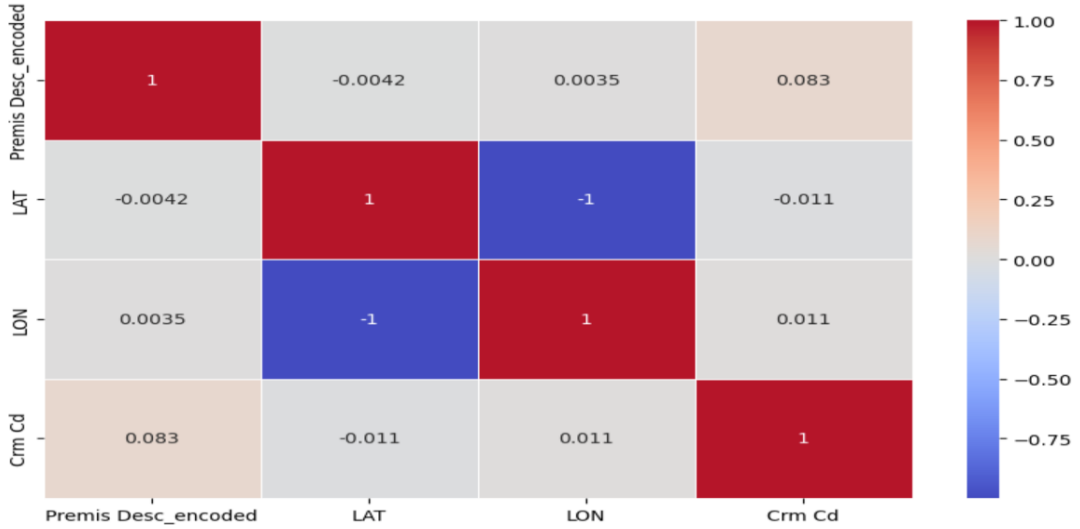


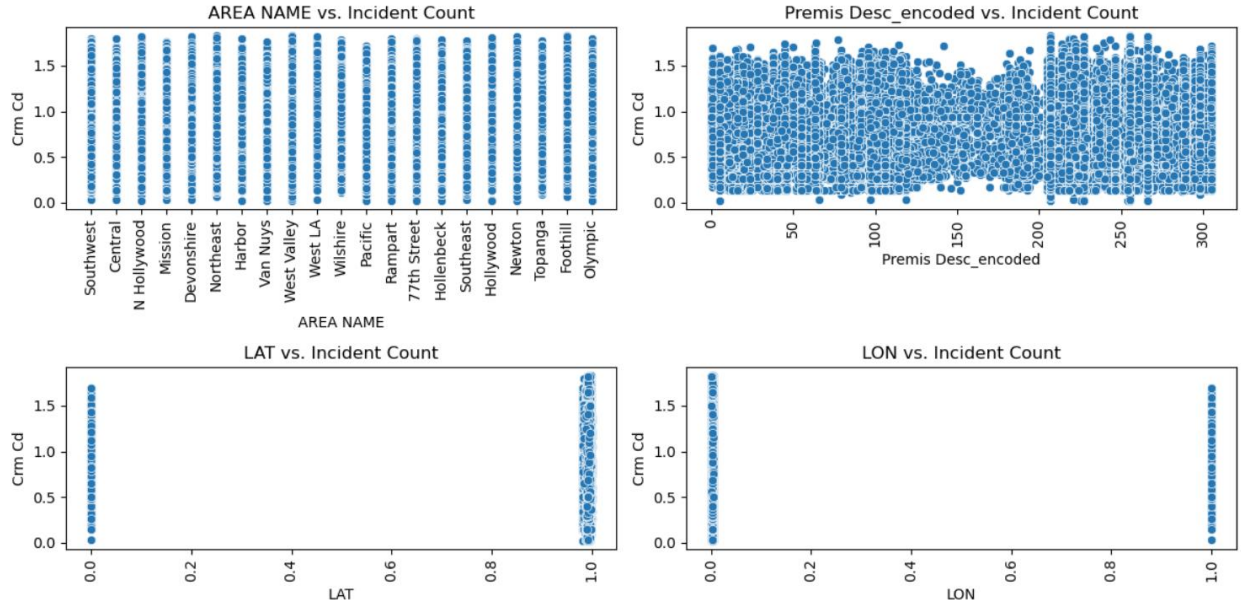
visualizing the monthly trends in this common crime type using a line plot, observing any patterns or changes in its occurrence over time.



bar chart that visually compares the crime rates in different cities or areas, providing a clear overview of which locations have higher or lower crime rates based on the data

We select a set of economic factors and aggregated different crime types to get total incident counts. It calculated the correlation coefficients between these factors and incident counts and visualizes the correlation matrix using a heatmap. We have visualized how these factors may be associated with crime rates.

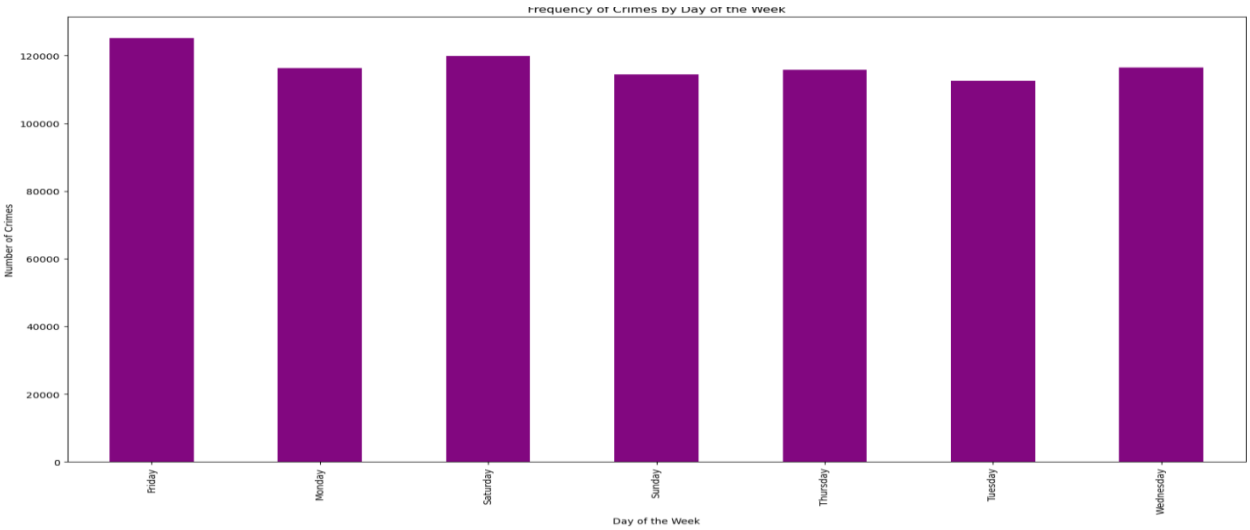


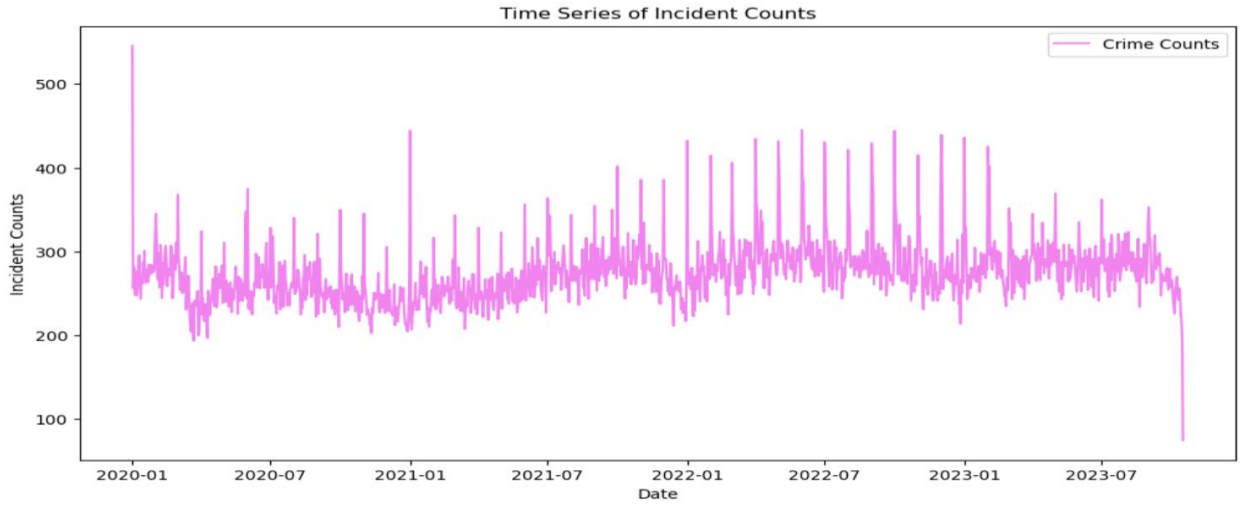


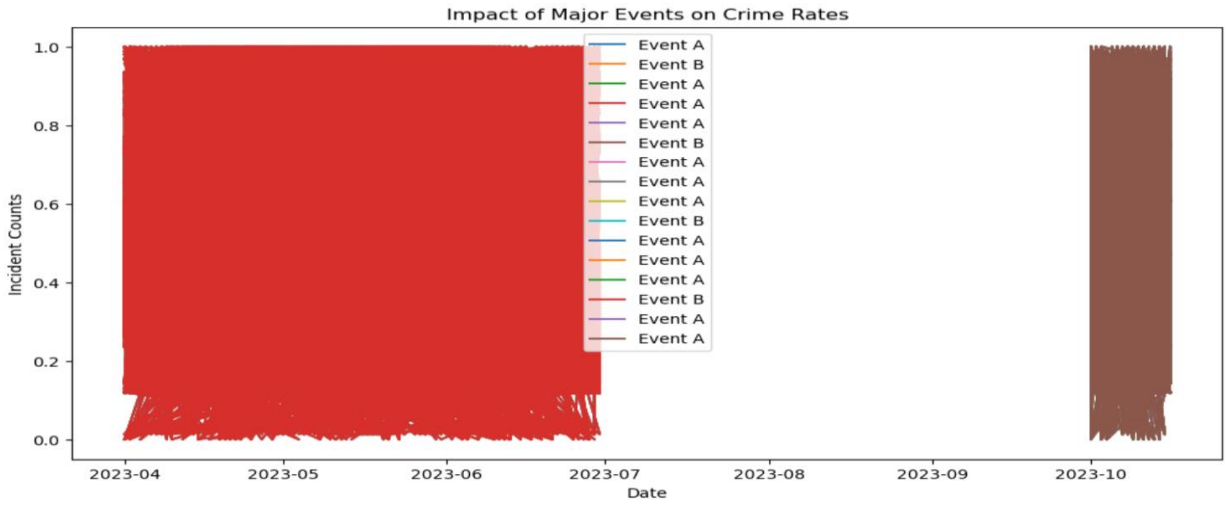
We then group the data by the day of the week and counts the occurrences of crimes on each day.

The date of occurrence is converted to a datetime format, defining major events or policy change dates, and creates binary indicator variables for these events. Then we created a time series of daily incident counts and visualized the time series to understand how crime rates change over time.To assess the impact of major events on crime rates, plot of the incident

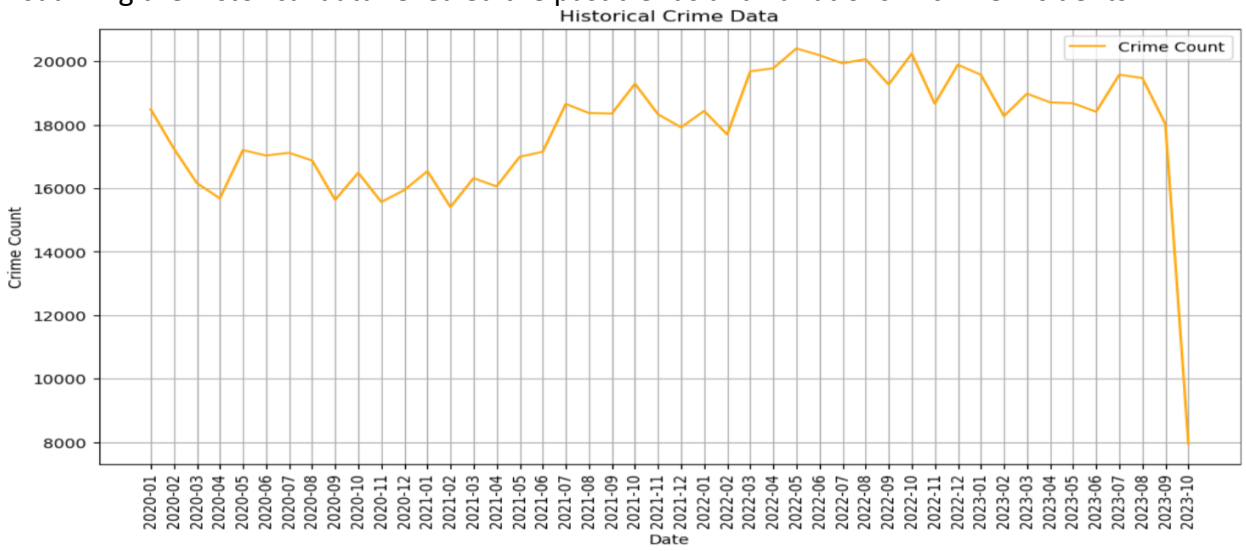
counts during these events are used by us, allowing for a visual examination of how specific events might influence crime trends.



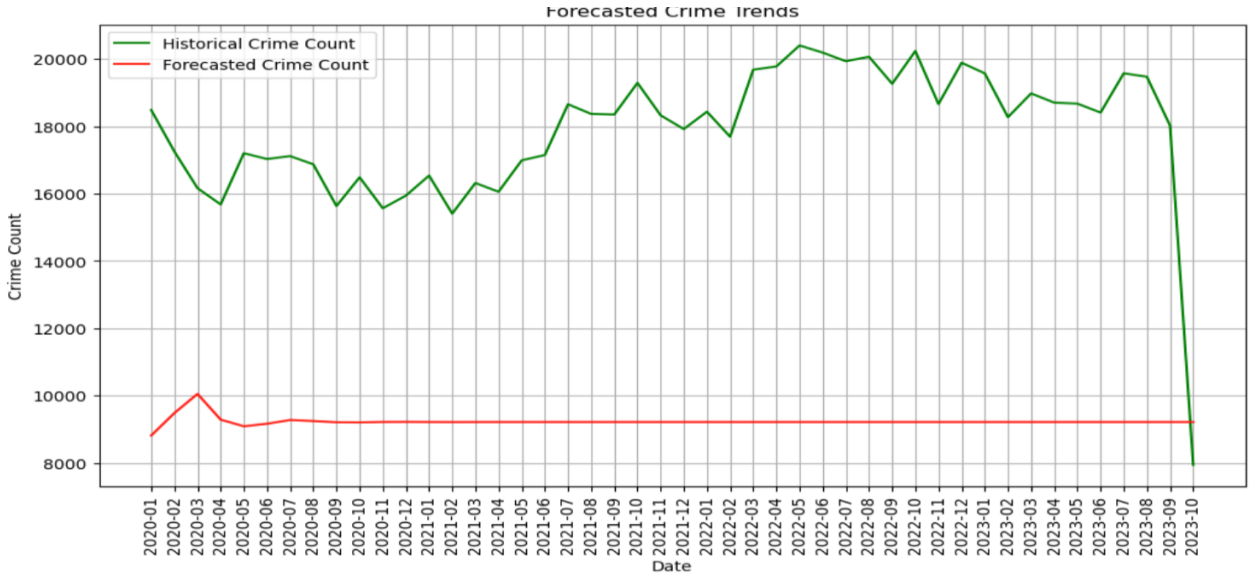




In this advanced analysis, we utilized predictive modeling techniques, specifically a time series forecasting method, to predict future crime trends based on historical data. We began by preparing the crime dataset, which included historical crime counts aggregated by month. Visualizing the historical data revealed the past trends and variations in crime incidents.

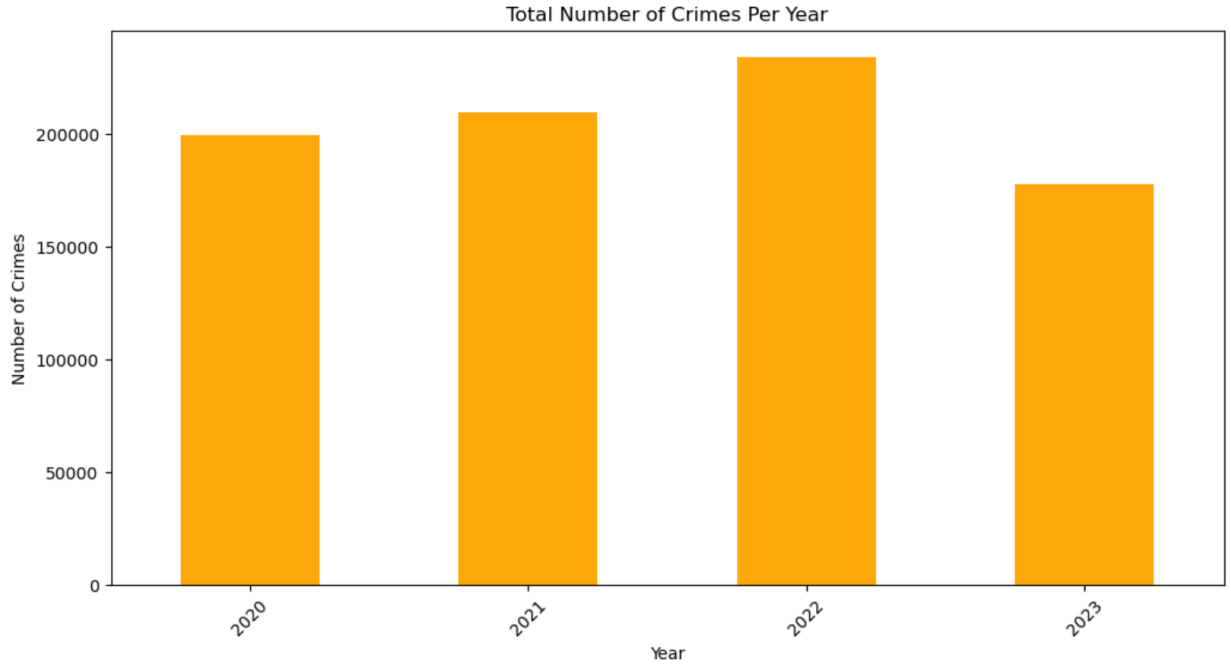


Next, we employed an ARIMA (AutoRegressive Integrated Moving Average) model to capture the time-dependent patterns in the crime data. We specified the model order as (5, 1, 0) to account for autoregressive, differencing, and moving average components. After fitting the ARIMA model to the historical data, we used it to forecast future crime trends.
The forecasted crime trends were then plotted alongside the historical data, allowing for a visual comparison between actual historical counts and the predicted counts for the forthcoming months. This analysis enables us to make informed predictions about potential changes in crime rates, contributing valuable insights for future planning and decision-making regarding law enforcement and public safety strategies.



Find the solutions to these questions:

1. Overall Crime Trends:
Calculate and plot the total number of crimes per year to visualize the trends.



2. Seasonal Patterns:
Group the data by month and analyze the average number of crimes per month over the years.

3. Most Common Crime Type:
Count the occurrences of each crime type and identify the one with the highest frequency.
Attaching the crimes with more than 2000 occurences
Crime Type: VEHICLE - STOLEN
Count: 87888

Crime Type: BATTERY - SIMPLE ASSAULT
Count: 65360

Crime Type: THEFT OF IDENTITY
Count: 51944

Crime Type: BURGLARY FROM VEHICLE
Count: 50275

Crime Type: VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)
Count: 50019

Crime Type: BURGLARY
Count: 49955

Crime Type: ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
Count: 47235

Crime Type: THEFT PLAIN - PETTY ($950 & UNDER)
Count: 41704

Crime Type: INTIMATE PARTNER - SIMPLE ASSAULT
Count: 41355

Crime Type: THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)
Count: 31696

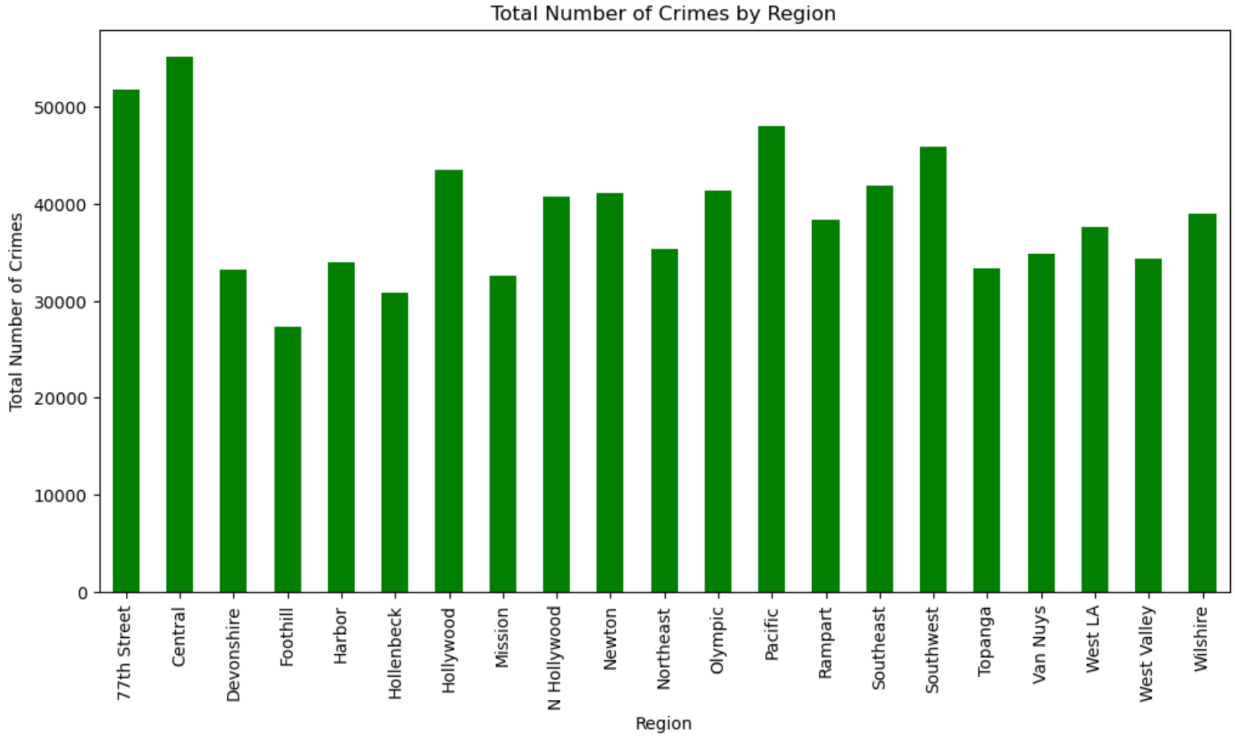
Crime Type: THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER)
Count: 29758

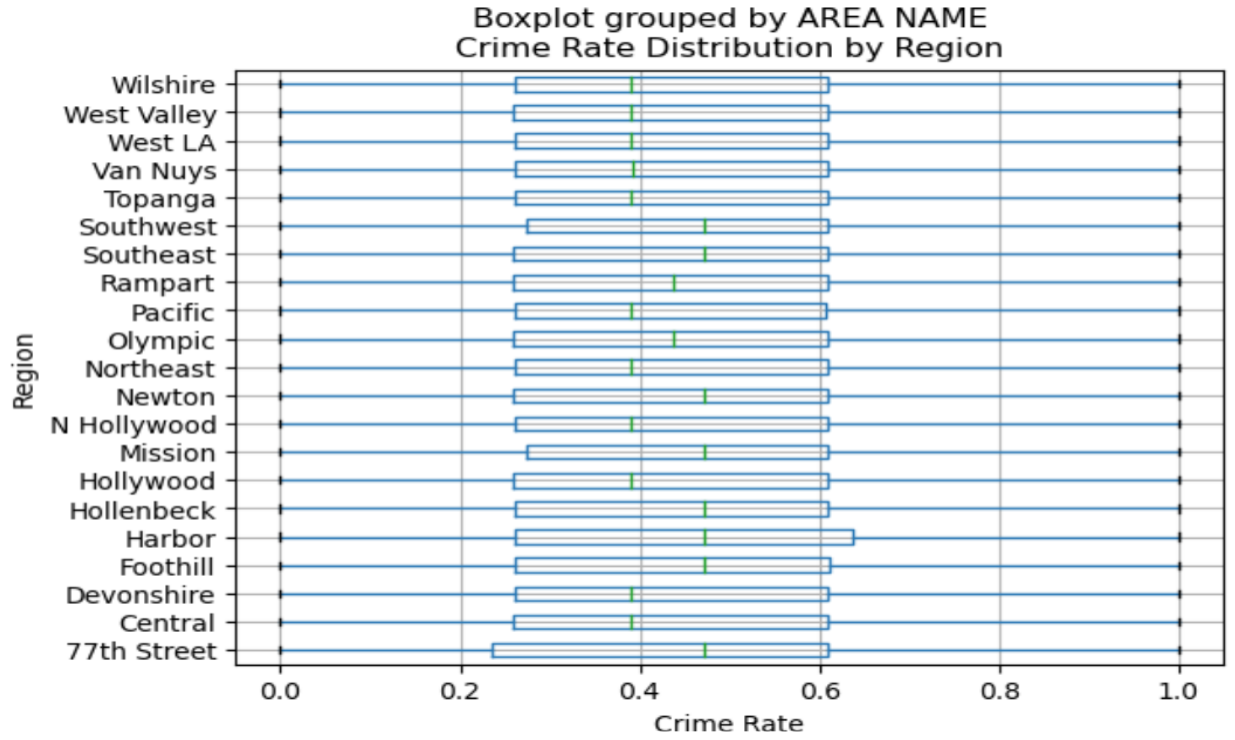
Crime Type: ROBBERY
Count: 28142

Crime Type: THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD
Count: 26694

4. Regional Differences:

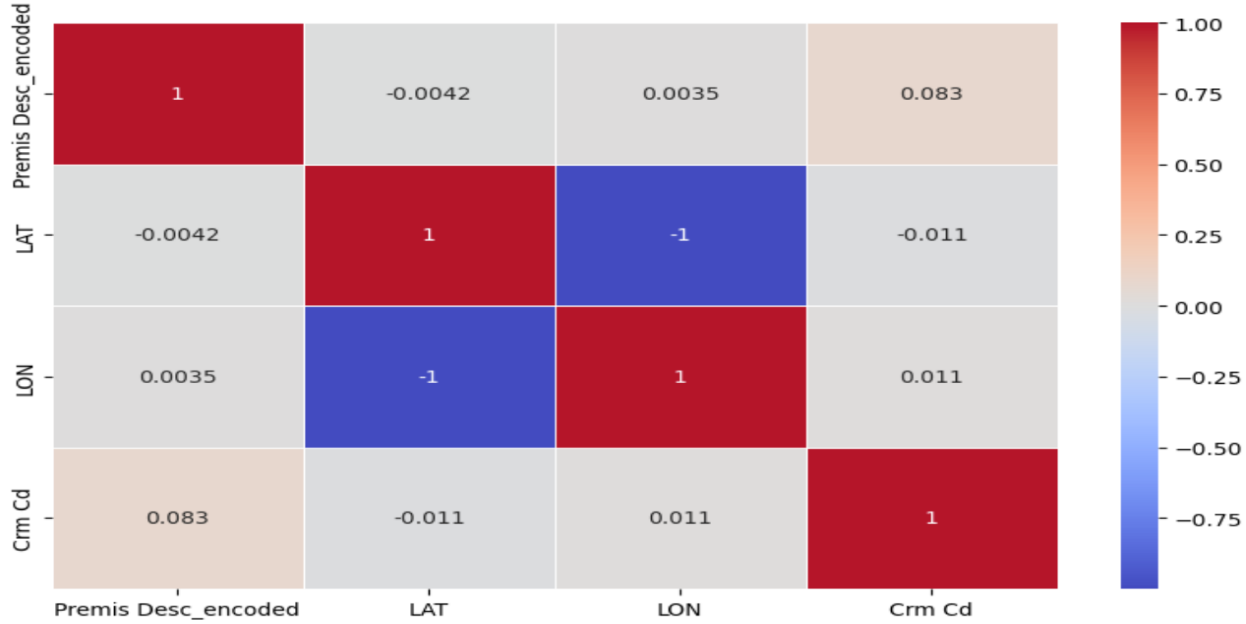
Group the data by region or city and compare crime rates between them using descriptive statistics or visualizations.





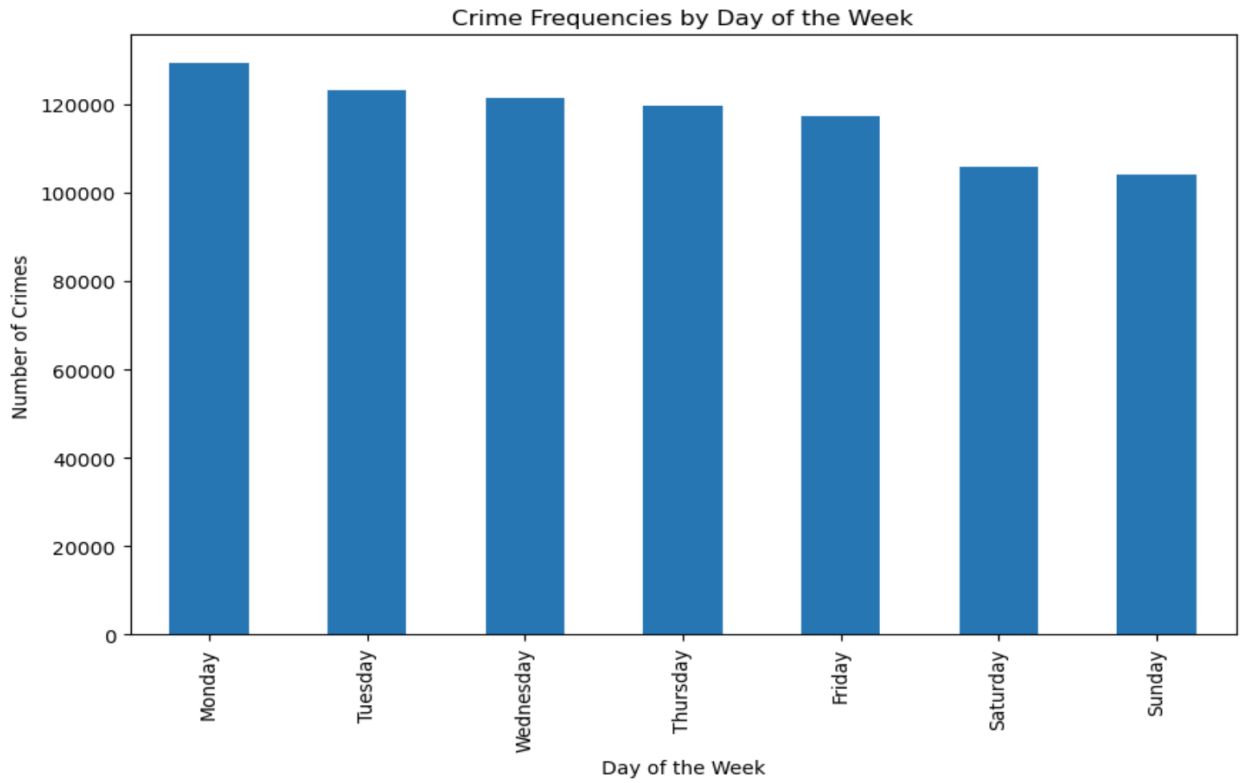
5. Correlation with Economic Factors:

o Collect economic data for the same time frame and use statistical methods like correlation analysis to assess the relationship between economic factors and crime rates.



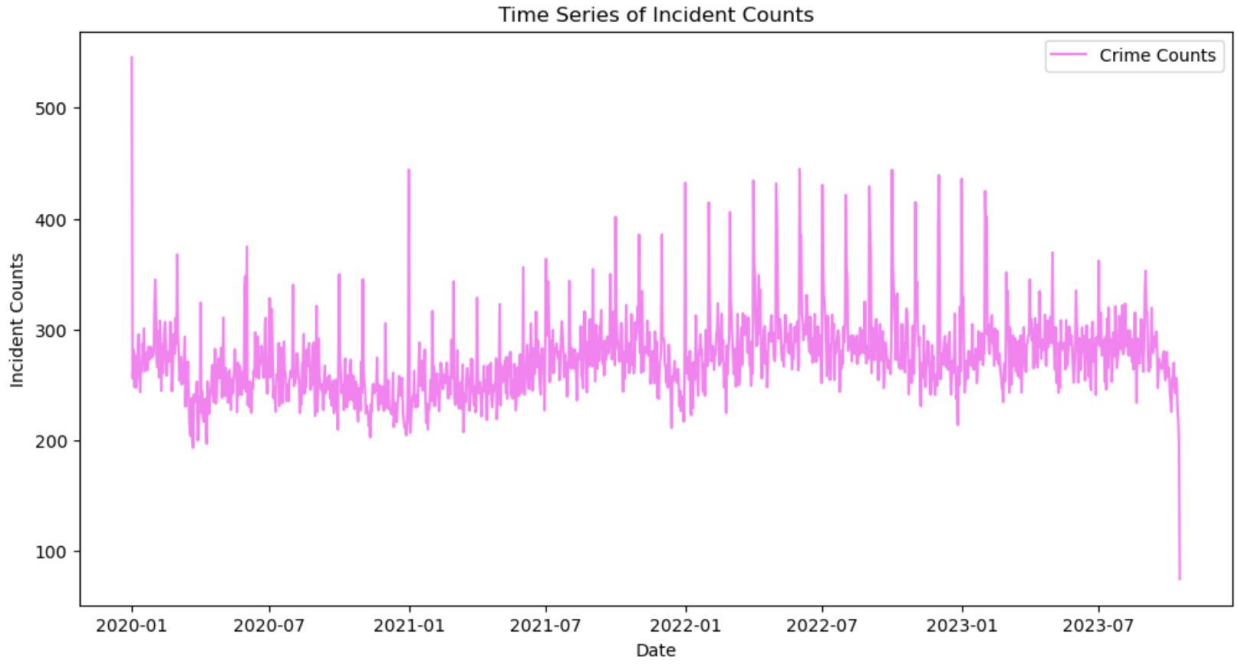
6. Day of the Week Analysis:

o Group the data by day of the week and analyze crime frequencies for each day.



7. Impact of Major Events:

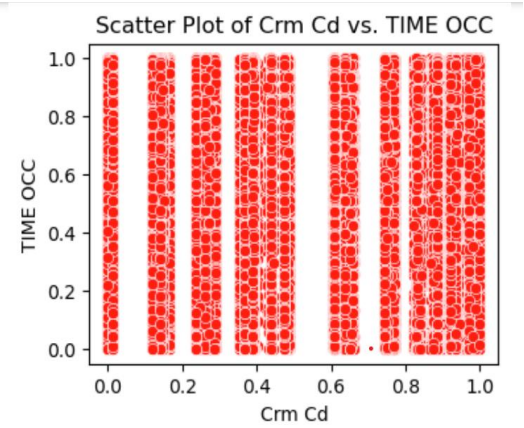
Identify major events or policy changes during the dataset period and analyze crime rate changes before and after these events.

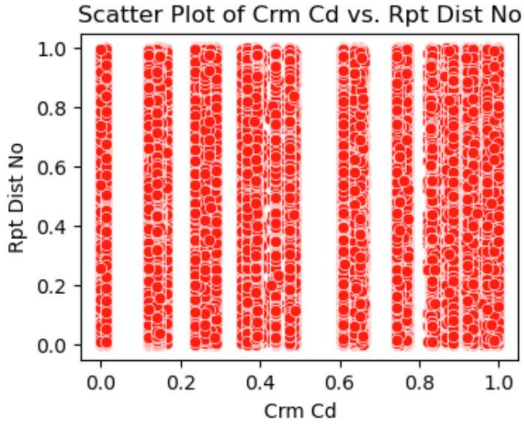


8. Outliers and Anomalies:

Use statistical methods or data visualization techniques to identify dataset outliers and investigate unusual patterns.

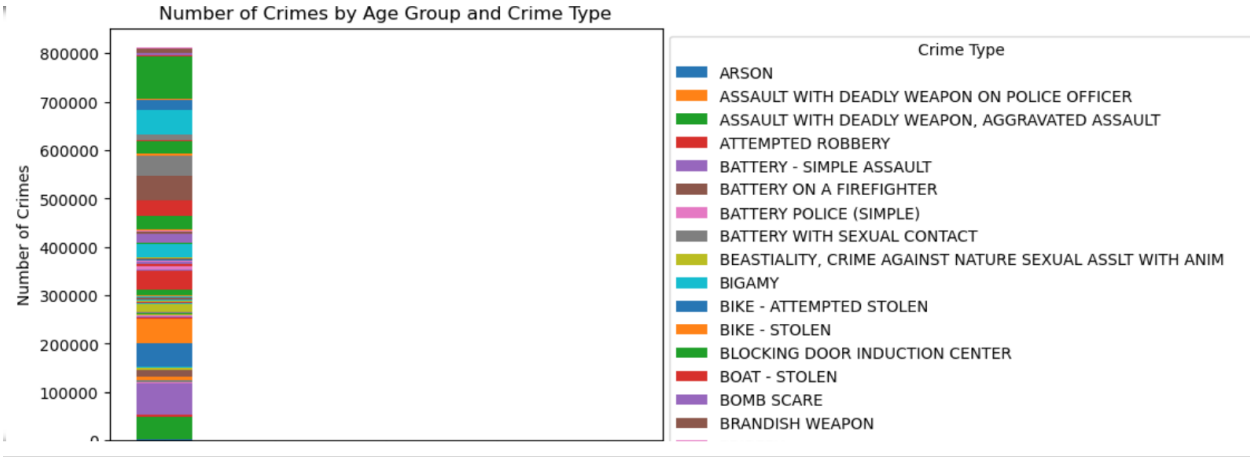
Outliers in DR_NO: 15
Outliers in TIME OCC: 0
Outliers in AREA: 0
Outliers in Rpt Dist No: 0
Outliers in Part 1-2: 0
Outliers in Crm Cd: 0
Outliers in Vict Age: 1
Outliers in Premis Cd: 0
Outliers in Crm Cd 1: 0
Outliers in Crm Cd 2: 60413
Outliers in Crm Cd 3: 2025
Outliers in Crm Cd 4: 60
Outliers in LAT: 31696
Outliers in LON: 11348

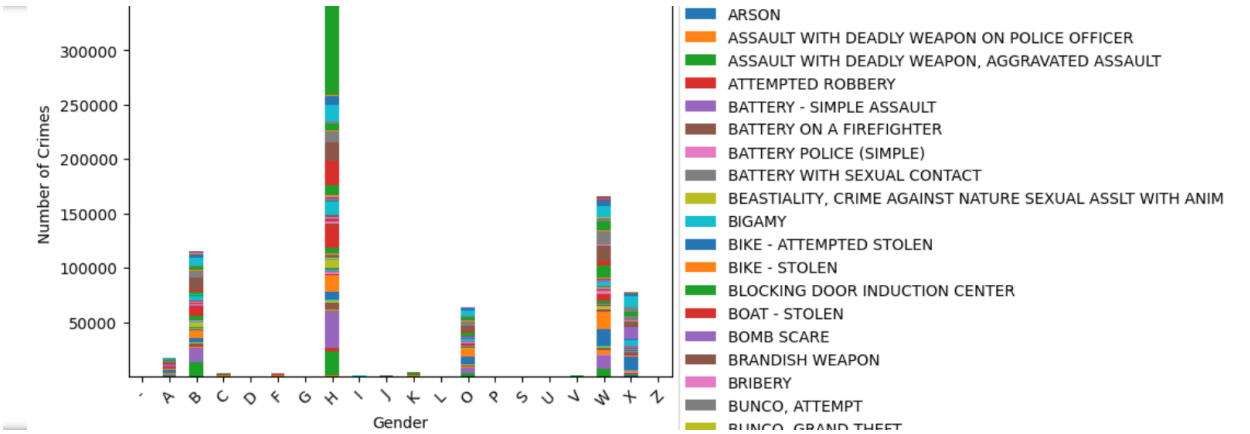




9. Demographic Factors:

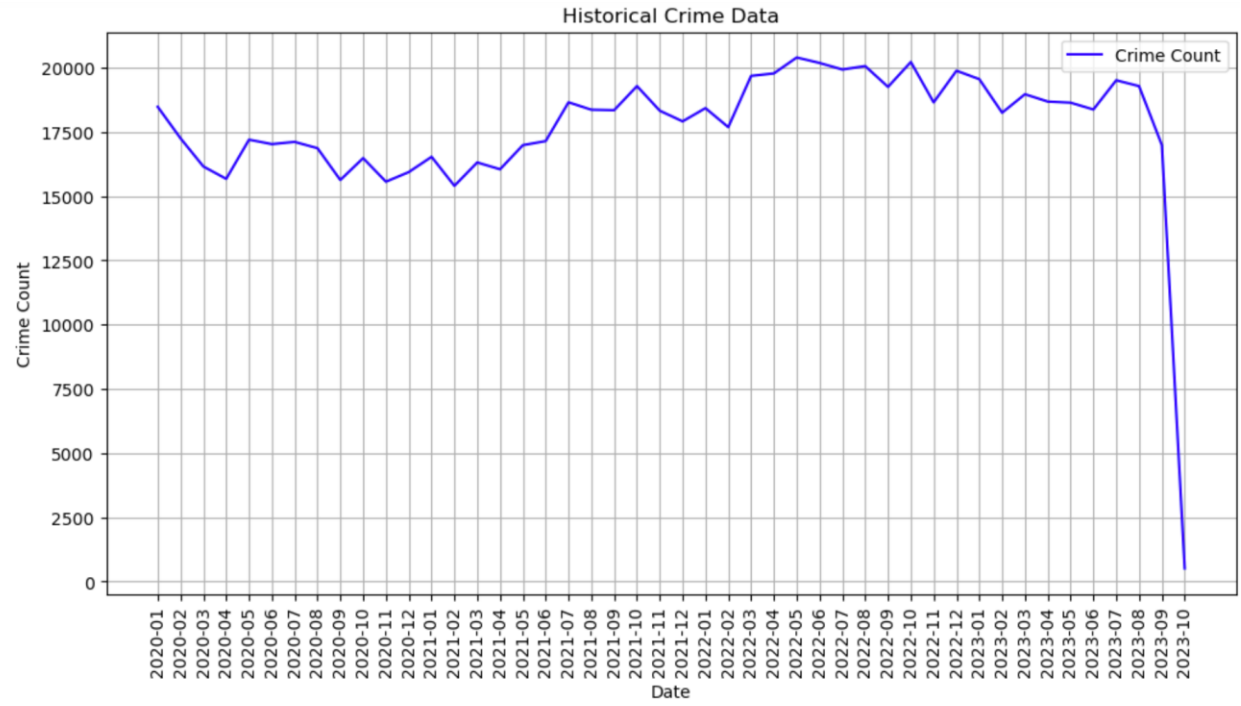
Analyze the dataset to identify any patterns or correlations between demographic factors (e.g., age, gender) and specific types of crimes.

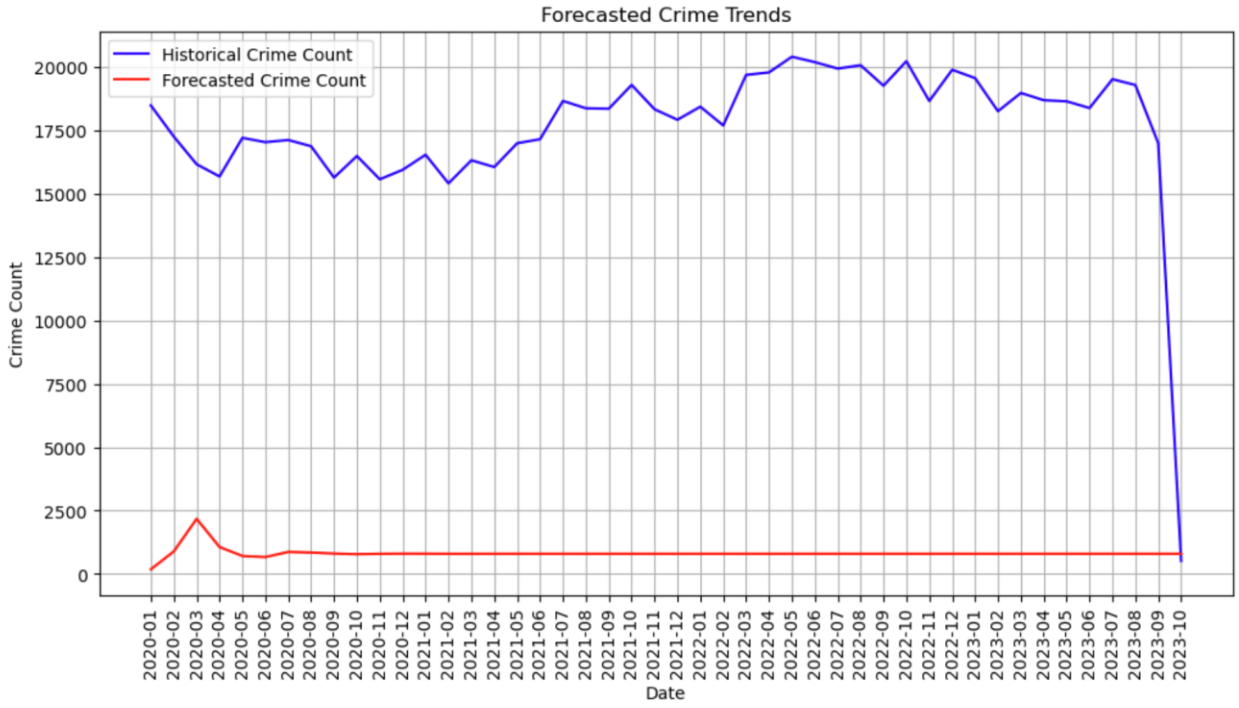




10. Predicting Future Trends:

o Employ time series forecasting methods, such as ARIMA or Prophet, to predict future crime trends based on historical data. Consider incorporating relevant external factors into your models.





Date	ForecastedCrimeCount	
2023-11-01	2020-01	181.913431
2023-12-01	2020-02	885.850696
2024-01-01	2020-03	2172.068050
2024-02-01	2020-04	1061.249245
2024-03-01	2020-05	707.275828
2024-04-01	2020-06	665.376388
2024-05-01	2020-07	867.160347
2024-06-01	2020-08	844.213804
2024-07-01	2020-09	803.251994
2024-08-01	2020-10	778.315463
2024-09-01	2020-11	794.147191
2024-10-01	2020-12	799.764198
2024-11-01	2021-01	798.496051
2024-12-01	2021-02	794.724051
2025-01-01	2021-03	794.954789
2025-02-01	2021-04	795.763247
2025-03-01	2021-05	796.084437
2025-04-01	2021-06	795.775296
2025-05-01	2021-07	795.657211
2025-06-01	2021-08	795.693945
2025-07-01	2021-09	795.756974
2025-08-01	2021-10	795.750513
2025-09-01	2021-11	795.733398
2025-10-01	2021-12	795.728780
2025-11-01	2022-01	795.734491
2025-12-01	2022-02	795.736589
2026-01-01	2022-03	795.735619
2026-02-01	2022-04	795.734512
2026-03-01	2022-05	795.734666
2026-04-01	2022-06	795.735000
2026-05-01	2022-07	795.735059
2026-06-01	2022-08	795.734950
2026-07-01	2022-09	795.734913
2026-08-01	2022-10	795.734935
2026-09-01	2022-11	795.734954
2026-10-01	2022-12	795.734951
2026-11-01	2023-01	795.734944
2026-12-01	2023-02	795.734944
2027-01-01	2023-03	795.734946
2027-02-01	2023-04	795.734946
2027-03-01	2023-05	795.734946
2027-04-01	2023-06	795.734946
2027-05-01	2023-07	795.734946

Part 5: Limitations and Future Work

In our project, it is crucial to explore potential avenues for future research and acknowledge the limitations of our analysis. In terms of future work, we can consider incorporating additional data sources, such as demographic information or data related to law enforcement resources, to achieve a more comprehensive understanding of crime trends. Advanced machine learning models, such as deep learning techniques or geospatial analysis, could be contemplated to enhance the accuracy of crime prediction. Real-time data analysis and monitoring may present an opportunity for a proactive approach to addressing emerging crime patterns. Spatial analysis, utilizing Geographic Information Systems (GIS), could assist in identifying specific crime "hotspots" and conducting more detailed spatial pattern analyses. Ethical considerations, particularly related to potential biases in the data or ethical concerns with predictive policing algorithms, should be an integral part of our future work.

On the other hand, it is essential to be mindful of the limitations inherent in our current analysis. These limitations may encompass issues related to data quality, such as the presence of missing or inaccurate data. Constraints regarding data availability might affect the inclusion of certain economic or demographic variables in our analysis. Generalization should be approached cautiously, as our findings are specific to the dataset and time frame used. Correlations between variables should not be automatically construed as indicating causation. Assumptions, such as the assumption of continued historical crime trends, may not always hold true. Privacy and ethical concerns need to be considered, particularly when dealing with sensitive crime data. If we have employed predictive models, it is important to acknowledge their limitations, including forecast accuracy and the potential for issues like overfitting and underfitting.