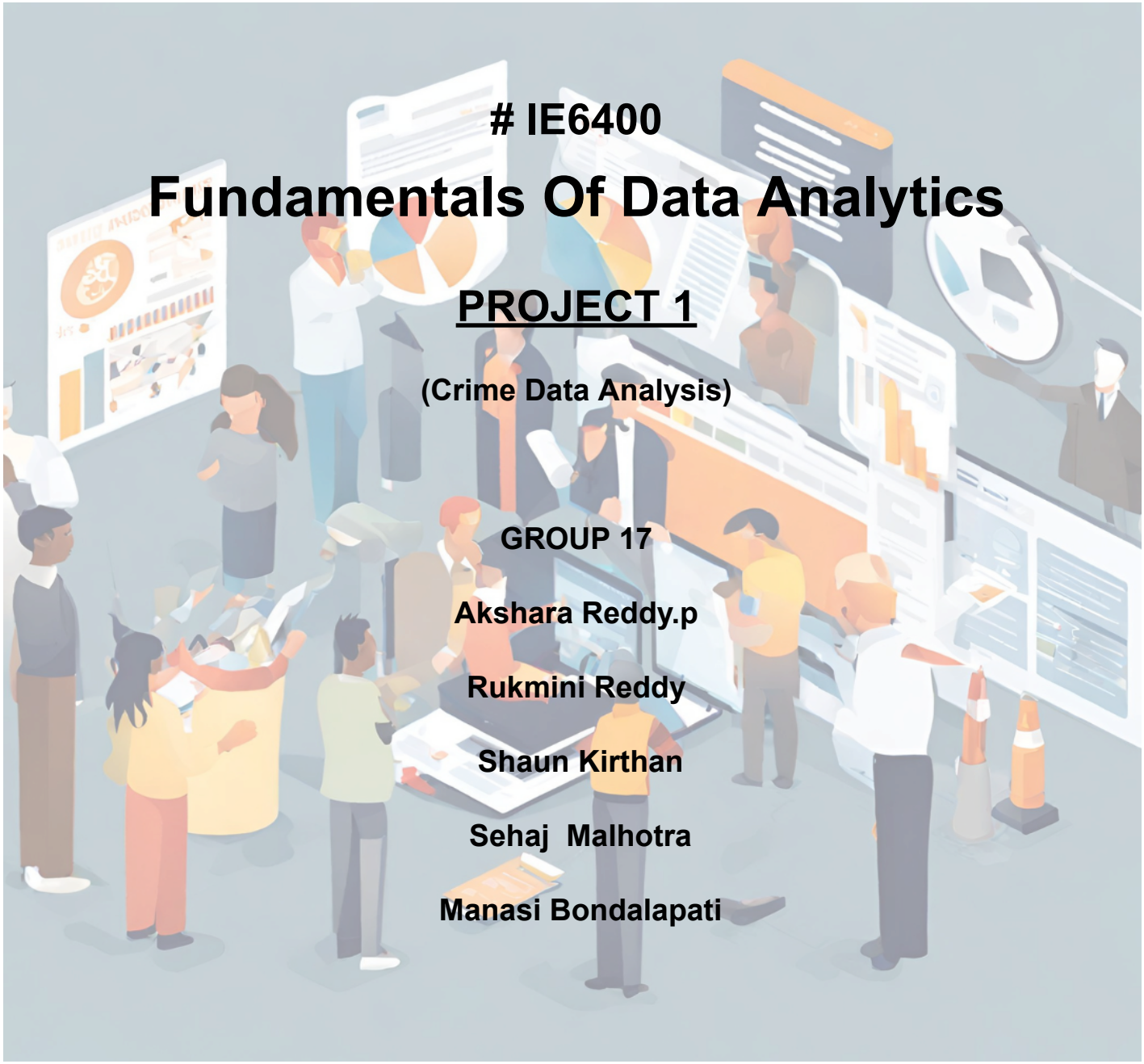
Northeastern University



IE6400

Fundamentals Of Data Analytics

PROJECT 1

(Crime Data Analysis)

GROUP 17

Akshara Reddy.p

Rukmini Reddy

Shaun Kirthan

Sehaj Malhotra

Manasi Bondalapati

College of Engineering, Northeastern University

Boston, MA 02115

Table Of Contents

Introduction

The aim of this project is to work with a real-world dataset containing crime data from 2020 to the present. The goal is to clean and prepare the dataset for analysis, perform exploratory data analysis, and answer specific questions related to crime trends, patterns, and factors influencing crime rates. The dataset used is the ‘Crime Data from 2020 to Present’.

The analysis commenced by importing the crime dataset into the designated data analysis tool, with a meticulous examination of the initial dataset structure, inclusive of the first few rows. A comprehensive review of the data types for each column was conducted, ensuring coherence and precision in the analytical process. Attention was paid to the column names and descriptions, facilitating a comprehensive understanding of the dataset's context.

The data cleaning process was initiated, involving a systematic identification of missing data and subsequent implementation of suitable handling strategies to maintain the integrity of the analysis. Rigorous checks were performed to detect any duplicate rows. To streamline the analysis, data types were appropriately adjusted, and measures were taken to manage outliers, mitigating their potential impact on the analytical outcomes. Standardization and normalization techniques were applied to numerical data, enabling effective comparisons, while categorical data were encoded for comprehensive analysis.

During the exploratory data analysis (EDA) phase, a range of visualization techniques were employed to decipher crime trends from 2020 to the present year. Emphasis was placed on the identification of seasonal patterns within the crime data, aiming to unveil any recurrent trends or patterns. Focused efforts were dedicated to uncovering the predominant crime types and their temporal variations over the study period, alongside an investigation into potential divergences in crime rates across different regions or cities. The analysis further delved into the potential correlations between available economic factors and crime rates, providing valuable insights for subsequent analysis. Moreover, the examination of the relationship between the day of the week and the frequency of specific crime types sought to uncover discernible patterns or trends that may underpin the data.

Data pre-processing,cleaning

In the initial phase of our project, we meticulously processed the Crime Data from 2020 to the present. This involved converting the dataset into a Pandas DataFrame, inspecting the data types of all the columns, and transforming the 'Date Rptd' and 'DATE OCC' columns into the datetime format for efficient time-based analysis. We conducted a comprehensive check for missing values, ascertaining the presence of null values across various columns. Fortunately, our dataset was devoid of any duplicate entries, ensuring data integrity.

Among the identified columns, 'Vict Sex' emerged with a significant count of 106,524 null values. To ensure the reliability of our analysis, we made a crucial decision to eliminate these null values. Consequently, we delved deeper into the 'Vict Sex' column to understand its categorical distribution. Leveraging the .value_counts() method, we gained valuable insights into the unique categories within 'Vict Sex,' including male, female, and others, refining our dataset for accurate gender-based analysis.

Similar scrutiny was applied to the 'Vict Age' column, where we encountered negative values that were promptly removed to uphold the integrity of our age-based analysis. Recognizing the importance of maintaining data consistency, we made the pivotal decision to remove outliers, denoted as 'h' and '-', ensuring that our dataset remained free from any anomalous data points that could skew our results.

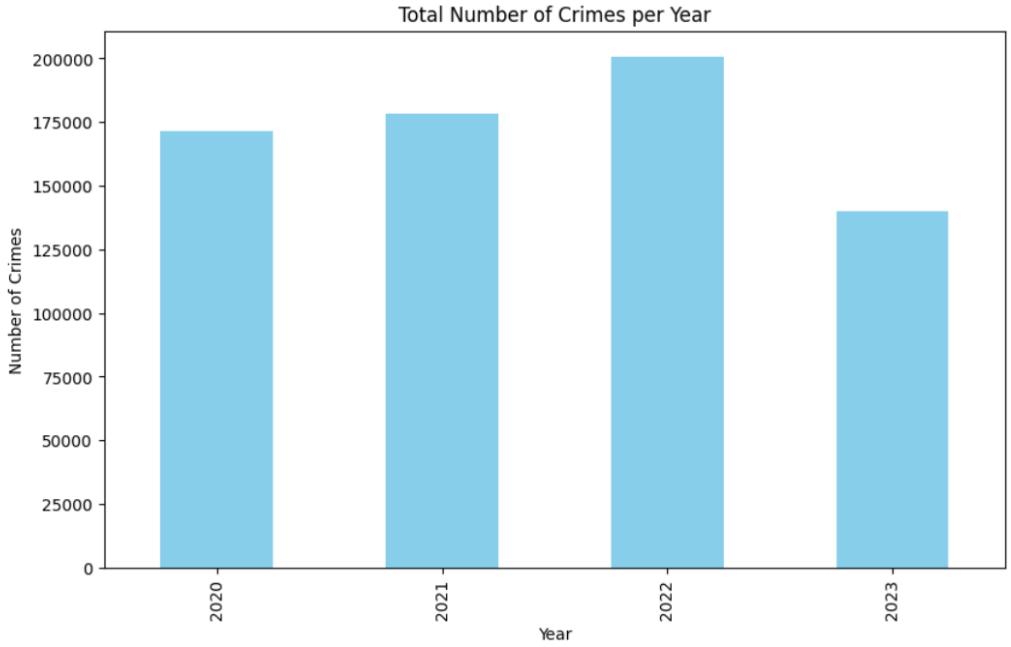
Further refinements were made to our dataset by addressing additional columns, such as 'Mocodes,' 'Vict Descent,' 'Premis Desc,' and 'Crm Cd 1,' all of which contained null rows that were considered non-essential for our current analysis. These rows were subsequently dropped, facilitating a more streamlined and concise dataset for our subsequent analysis.

We made the strategic decision to remove the 'Crm Cd 2,' 'Crm Cd 3,' and 'Crm Cd 4' columns from our DataFrame, citing redundancy and a need for increased processing efficiency. By meticulously conducting these cleaning and preprocessing operations, we were able to create a refined and robust dataset, representative of the essential information required for our analytical endeavors.

Our final step involved creating a new CSV file, housing the thoroughly cleaned and preprocessed data. This cleaned dataset serves as a solid foundation for our subsequent analysis and will pave the way for deriving meaningful insights and trends from the meticulously curated information. This data cleaning and preprocessing phase has laid a strong groundwork for our project, ensuring data reliability and enabling us to delve deeper into the intricate nuances of the crime data for informed decision-making and strategic planning.

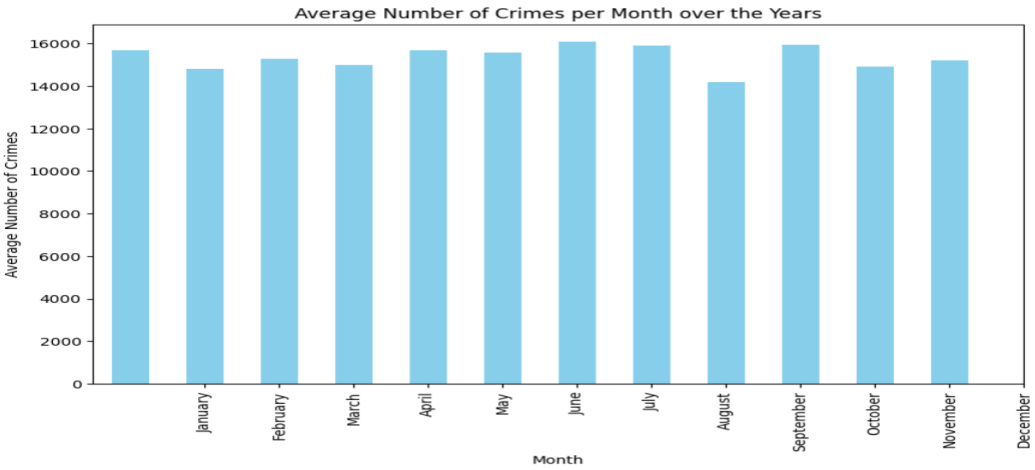
Analyzing trends

1. Overall Crime Trends:



Stable crime rate in 2020 and 2021- The number of crimes in 2020 and 2021 remained almost the same, around 175,000. Significant decrease in 2022 and 2023- There was a significant decrease in the number of crimes in 2022 and 2023. Lowest crime rate in 2023- The number of crimes in 2023 was less than half of the number of crimes in 2020 and 2021. This data suggests a positive trend in crime reduction over the last two years.This maybe due to the global pandemic.

2.Seasonal Patterns



The number of crimes is highest in the summer months, particularly in July and August. The number of crimes is lowest in the winter months, with February having the lowest average

number of crimes.This pattern could be due to more people being outside and active during the summer months, leading to more opportunities for crime.

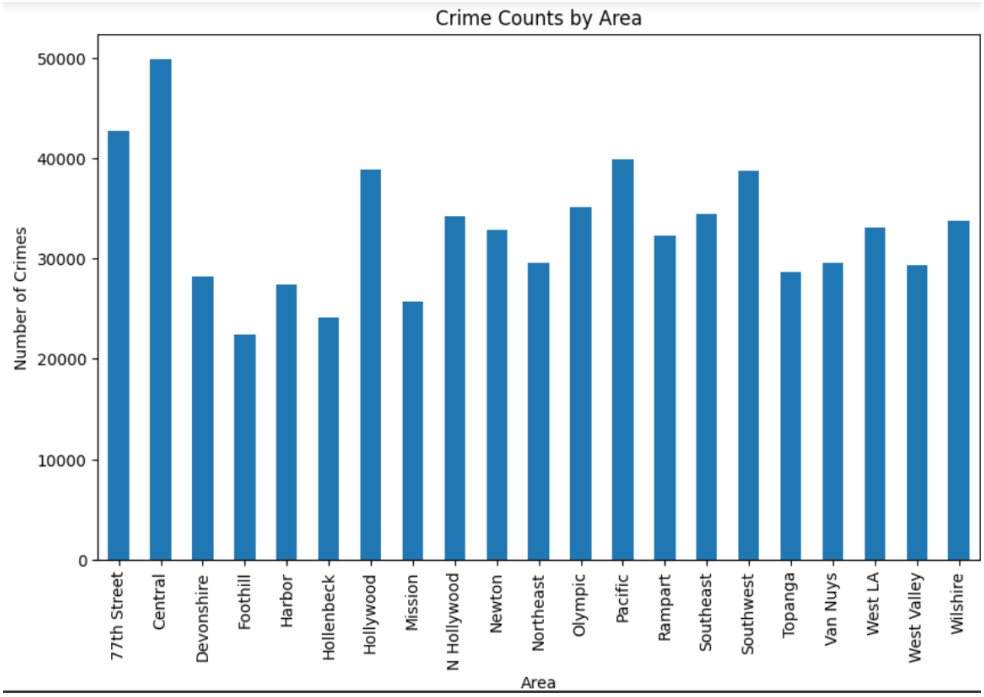
3.Most Common Crime Type

Crm Cd Desc	
ARSON	1844
ASSAULT WITH DEADLY WEAPON ON POLICE OFFICER	984
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	46166
ATTEMPTED ROBBERY	4173
BATTERY - SIMPLE ASSAULT	63648
	...
VEHICLE, STOLEN - OTHER (MOTORIZED SCOOTERS, BIKES, ETC)	761
VIOLATION OF COURT ORDER	5499
VIOLATION OF RESTRAINING ORDER	10119
VIOLATION OF TEMPORARY RESTRAINING ORDER	779
WEAPONS POSSESSION/BOMBING	29

Top 5 most common crimes:

'BATTERY - SIMPLE ASSAULT': 63648 occurrences
'BURGLARY FROM VEHICLE': 49342 occurrences
'VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)': 49000 occurrences
'THEFT OF IDENTITY': 48929 occurrences
'BURGLARY': 48046 occurrences

4.Regional Differences

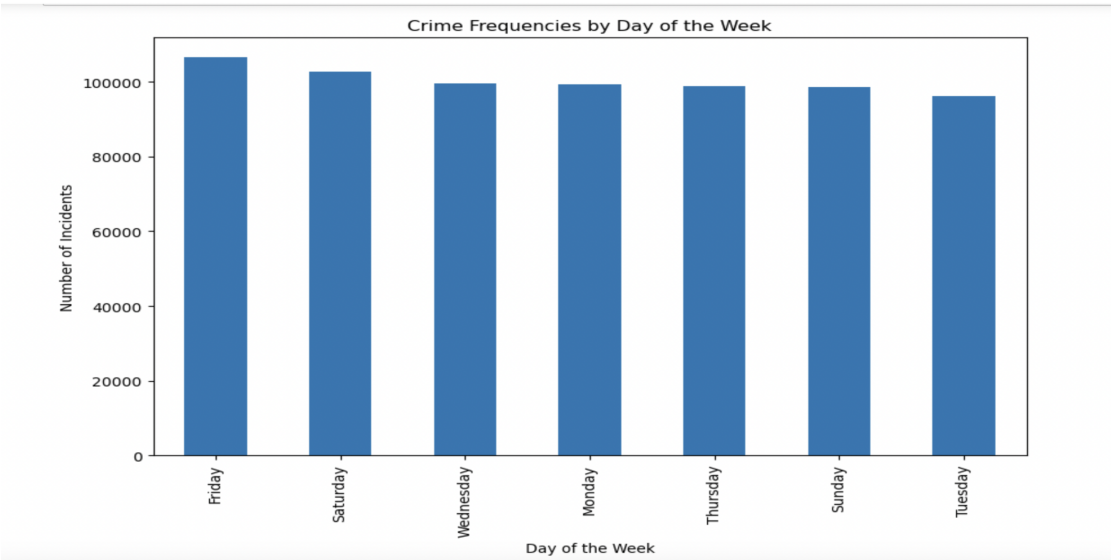


The average number of crime among all the different type of crime is approx 29522.Area with highest number of was ‘77th Street’, the most common in region was ‘ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT’ the count of was 38469.

5.Correlation with Economic Factors

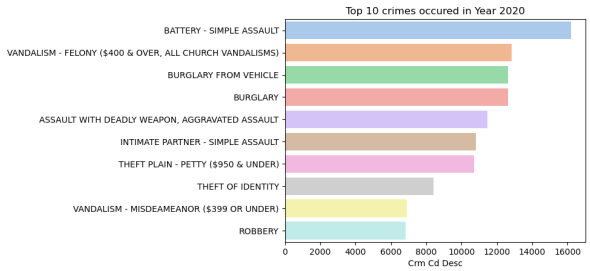
Based on the analysis conducted, it is evident that there exists a mutual influence between crime rates and various economic factors.The unemployment rate demonstrates the most substantial impact on crime rates, as indicated by its positive Pearson correlation coefficient. This positive coefficient signifies that an escalation in the unemployment rate is associated with a corresponding increase in crime trends.The correlation between crime rates and median household income is negative, indicating that an augmentation in median household income is linked to a reduction in crime rates. This suggests that regions with higher median household incomes tend to experience lower levels of criminal activity.

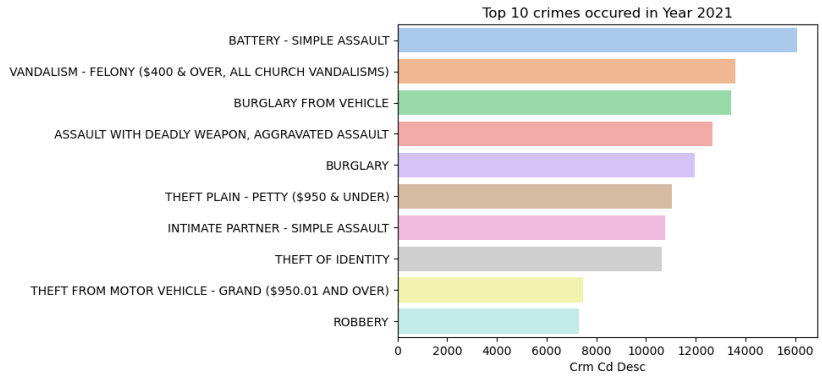
6.Day of the Week Analysis

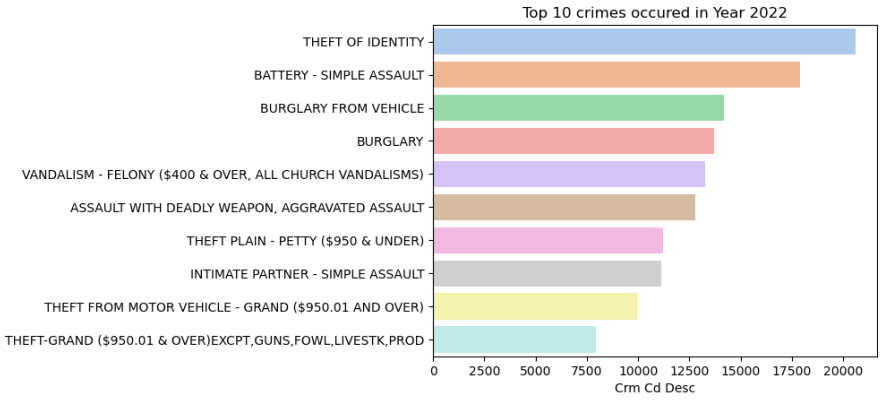


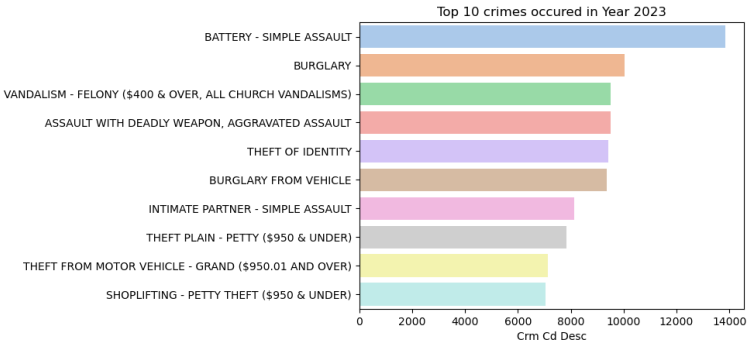
In this data analysis, we examined the relationship between crime rates and the day of the week. We extracted the day of the week from the "DATE OCC" column and counted the number of crimes that occurred on each specific day. The results indicate that the highest recorded crime rate is observed on Fridays, while the lowest crime rate is reported on Tuesdays.

7. Impact of Major Events:









The convergence of the COVID-19 pandemic, stringent ghost gun laws, and inflation has intricately shaped crime rates in distinct ways. The pandemic's digitalization surge inadvertently expanded the scope for cybercriminals, leading to a significant rise in identity theft cases in 2022. California's comprehensive gun legislation curtailed untraceable firearms, notably reducing crime rates related to these weapons.

Research revealed a notable link between inflation and increased criminal activities, driven by economic hardships faced by individuals during times of high inflation.

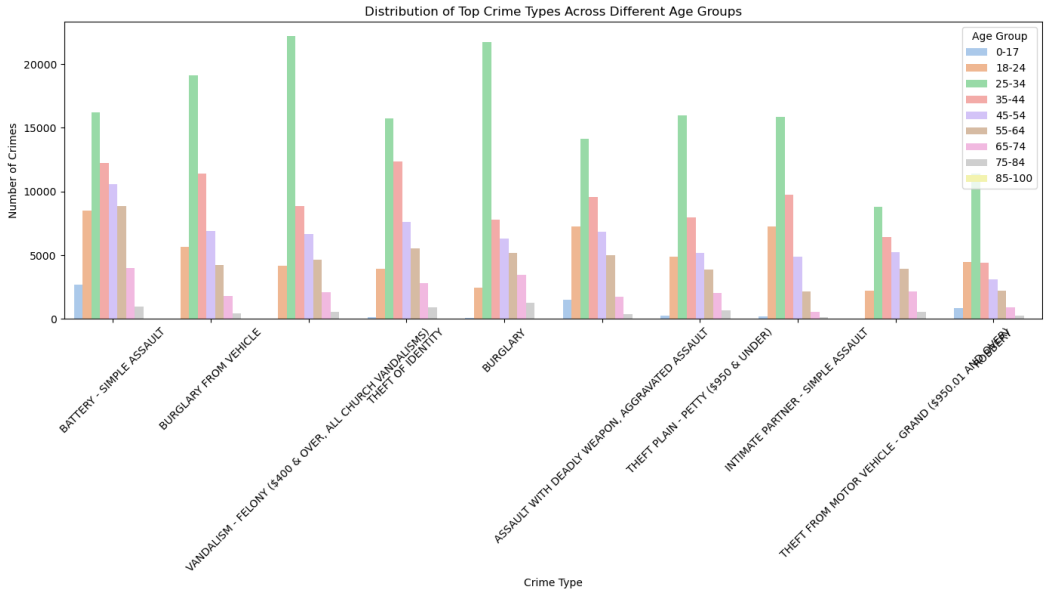
8. Outliers and Anomalies

	DR_NO	TIME OCC	AREA	Rpt Dist No	Part 1-2	Crm Cd	Vict Age	Year	YEAR	MONTH
0	10304468	2230	3	377	2	624	36	2020	2020	1
1	190101086	330	1	163	2	624	25	2020	2020	1
3	191501505	1730	15	1543	2	745	76	2020	2020	1
4	191921269	415	19	1998	2	740	31	2020	2020	1
67093	190101087	510	1	156	2	626	53	2020	2020	1
...	...	...	...	...	...	...	...	...	...	...
811576	230505717	1900	5	557	1	210	78	2023	2023	2
811599	230609390	1300	6	669	2	627	8	2023	2023	5
811627	230804108	1130	8	834	2	956	77	2023	2023	1
811629	232007343	1640	20	2038	1	341	77	2023	2023	3
811639	231608412	2130	16	1663	2	624	8	2023	2023	5

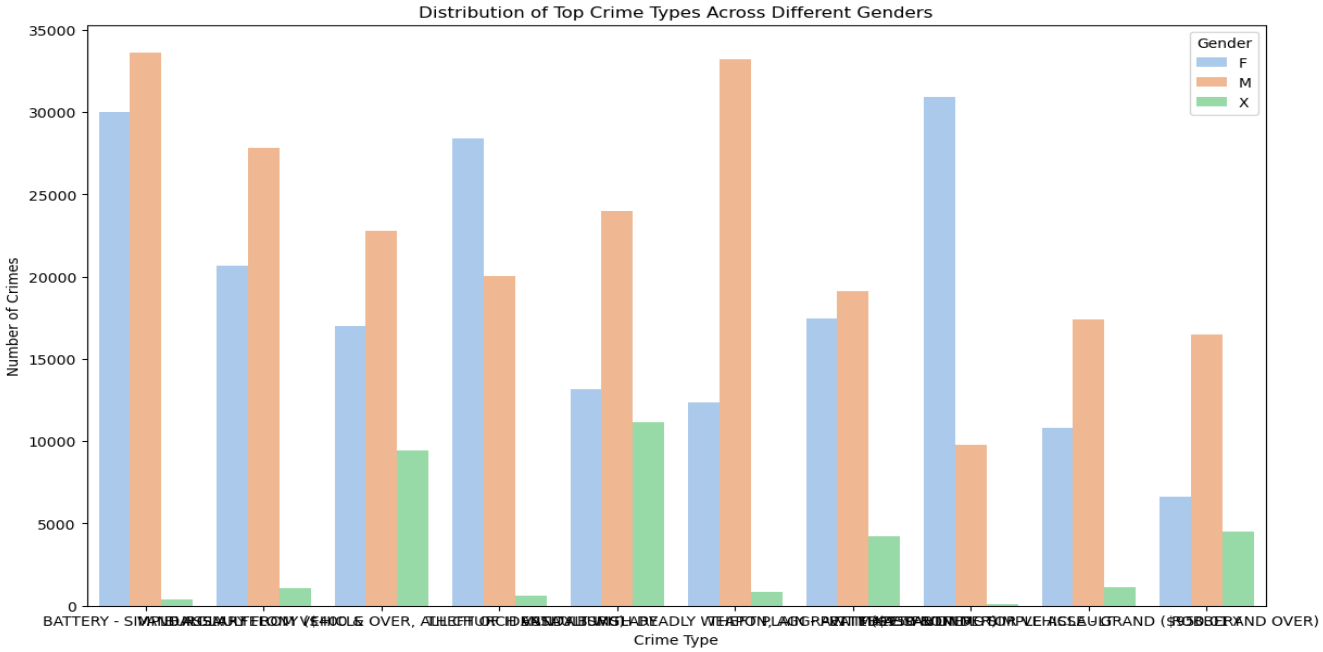
65844 rows × 10 columns

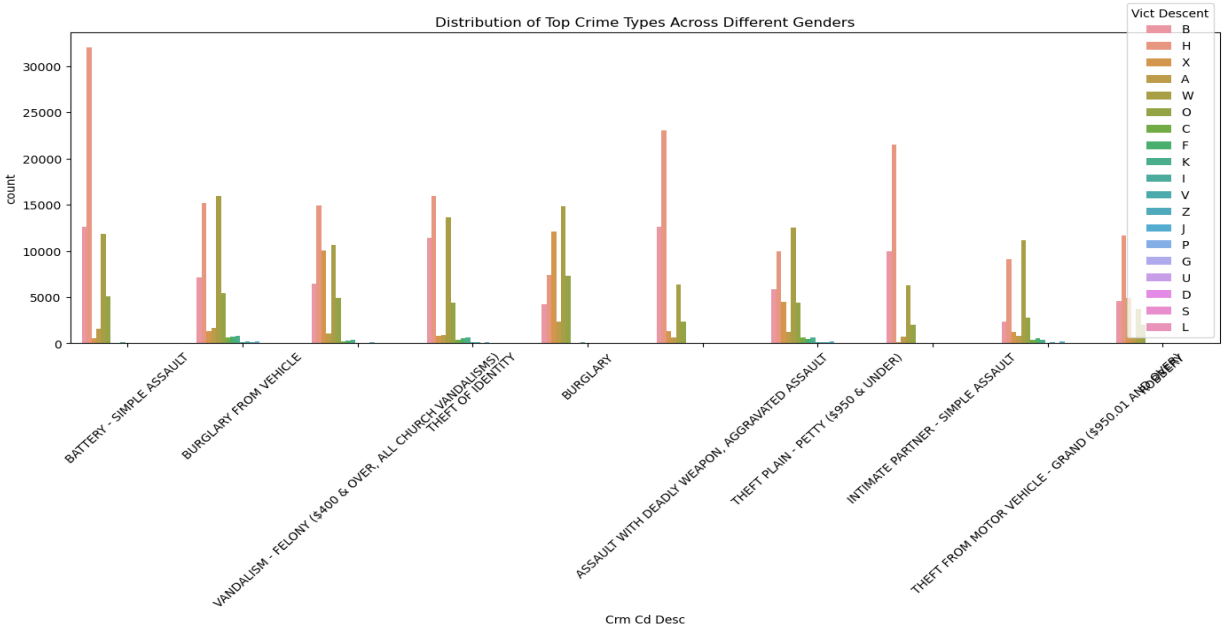
A Comprehensive view of how the top 10 crime types are distributed across different demographics, including age groups, genders, and racial or ethnic backgrounds. This can be valuable for identifying patterns and trends in crime data.This can be useful for studying disparities in crime victimization.

9.Demographic Factors



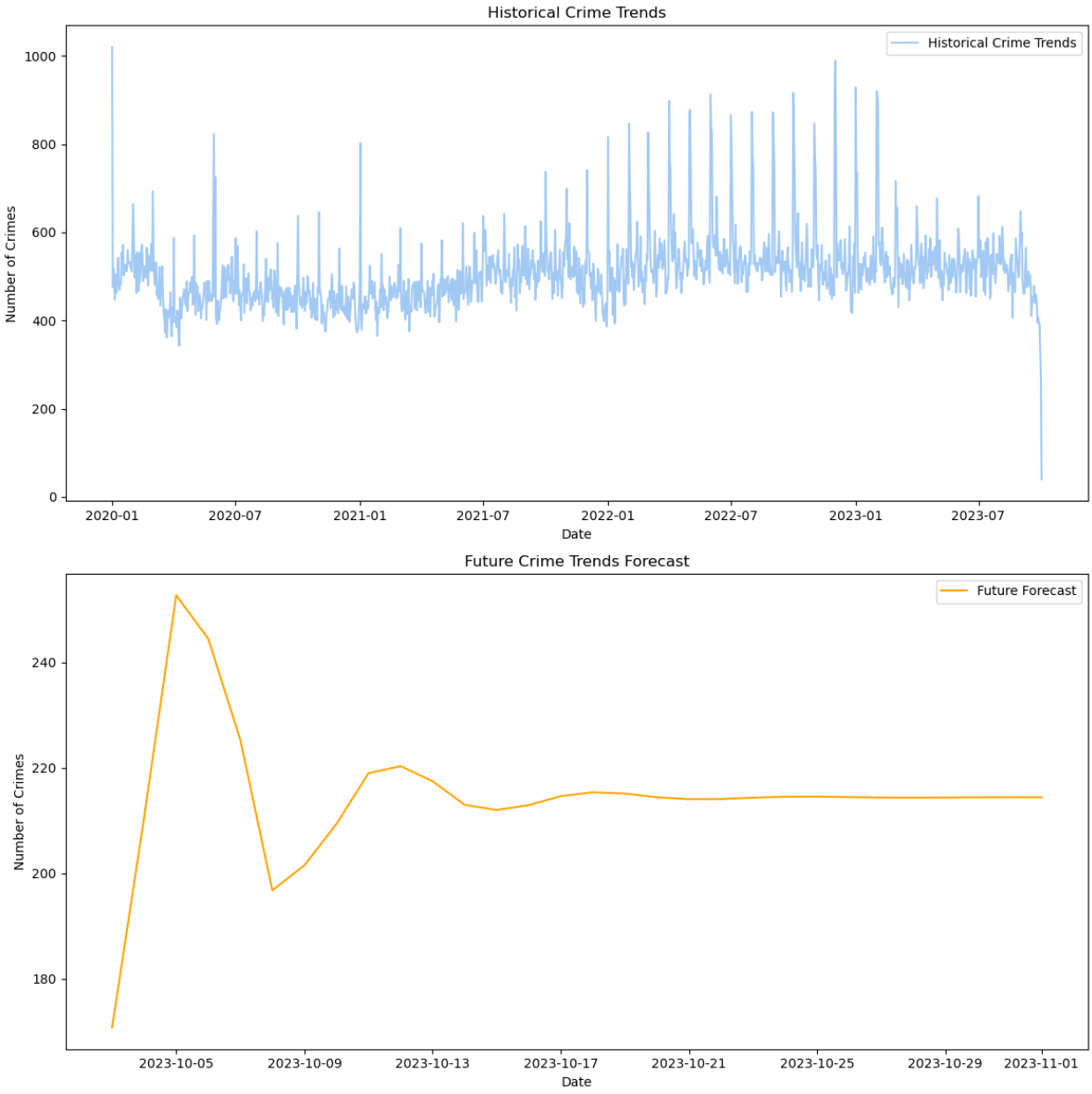
For the distribution of crime types across different age groups:The charts reveal how various crime types are distributed across different age groups, which can be useful for understanding the demographics of crime victims. For example, one can observe if certain crimes are more prevalent among specific age groups.





The charts illustrate how crime types are distributed among different genders. This information can help in analyzing whether certain types of crimes disproportionately affect specific genders.

10.Predicting Future Trends



The first graph shows historical crime trends in a specific area. The graph shows a fluctuating trend, indicating that the crime rate in this area is not stable and varies over time.: There is a noticeable peak in the middle of the graph. This could suggest a period of increased criminal activity.

The second graph gives the future predictions of the crime rates.The trends seem to be more stable after a while. This might be because of proper law enforcements and also the environmental factors such as the winter that is approaching which would cause people to stay in shelter rather than be outside.

Conclusion

The analysis of the crime data over the past few years reveals some intriguing patterns and trends. Notably, the stable crime rates observed in 2020 and 2021, with around 175,000 reported crimes, contrast sharply with the significant decrease in the following years, particularly in 2022 and 2023. Surprisingly, 2023 recorded the lowest crime rate, indicating a potential positive shift in crime reduction. This decline might be attributed to the impact of the global pandemic, altering societal dynamics and activities.

An intriguing seasonal pattern emerges, illustrating a surge in crime during the summer months, notably in July and August, and a contrasting decrease during the winter, with February marking the lowest average crime rate. This seasonal variation could be attributed to increased outdoor activities during the warmer months, creating more opportunities for criminal activities. The identification of the top five most common crimes sheds light on the prevalent types of criminal activities, ranging from 'BATTERY - SIMPLE ASSAULT' to 'BURGLARY,' providing valuable insights into the nature of the prevalent criminal activities within the analyzed dataset.

The analysis suggests a significant interplay between crime rates and economic factors, notably demonstrating the influential role of the unemployment rate and median household income. An increase in the unemployment rate appears to be associated with a corresponding rise in crime, while higher median household incomes exhibit an inverse relationship, indicating lower crime rates in areas with elevated household incomes.

The analysis of historical crime trends reveals a dynamic and fluctuating pattern within the specific area, characterized by notable variations in crime rates over time. The observed peak in the middle of the graph suggests a potential period of heightened criminal activity, emphasizing the need for a comprehensive understanding of the underlying socio-economic and environmental factors contributing to this trend.

The future predictions of crime rates indicate a more stabilized trend, potentially influenced by the implementation of robust law enforcement measures and community engagement initiatives. This stability may also be attributed to environmental factors, such as the approaching winter season, which typically limits outdoor activities and reduces opportunities for criminal incidents. The forecasted stability in crime rates underscores the importance of sustained efforts in proactive law enforcement.

References

1. Crime Data from 2020 to Present - Catalog
2. US Macro-Economic Factors data from 2002-2022 (kaggle.com)
3. What is Exploratory Data Analysis ? - GeeksforGeeks
4. Time Series Forecasting — ARIMA vs Prophet | by Krish Hariharan | Analytics Vidhya | Medium
5. What is Correlation Analysis? A Definition and Explanation
6. Difference between Anomaly and Outlier - Cross Validated
7. Descriptive Statistics: Definition, Overview, Types, Example
8. South, Scott J., and Steven F. Messner. “Crime and Demography: Multiple Linkages, Reciprocal Relations.” *Annual Review of Sociology*, vol. 26, 2000, pp. 83–106. *JSTOR*, http://www.jstor.org/stable/223438.
9. Predictive Modeling - Time-Series Regression, Linear Regression Models - MATLAB & Simulink
10. Class pdfs