IE6400 Foundations Data Analytics Engineering

Fall Semester 2023

Cleaning and Analyzing Crime Data

Final Report

Group 6

Giridhar Babu

Hemant Manohar Deshmukh

Mallika Gaikwad

Sharvari Pravin Deshpande

Medhavi Uday Pande

Index

ABSTRACT

In today's time, crime rate is exponentially increasing with a variety of crimes and violence taking place in the United States. Crime has been recorded in the United States since its founding and has fluctuated significantly over time, with a sharp rise after 1900 and reaching a broad bulging peak between the 1970s and early 1990s. After 1992, crime rates have generally trended downwards each year, with the exceptions of a slight increase in property crimes in 2001 and increases in violent crimes in 2005-2006, 2014-2016 and 2020-2021.

While official federal crime data beginning in 2021 has a wide margin of error due to the incomplete adoption of the National Incident-Based Reporting System by government agencies, federal data for 2020-2021 and limited data from select U.S. cities collected by the nonpartisan Council on Criminal Justice showed significantly elevated rates of homicide and motor vehicle theft in 2020-2022. Although overall crime rates have fallen far below the peak of crime scene in the United States during the late 1980s and early 1990s, the homicide rate in the U.S. has remained high, relative to other developed nations, with eight major U.S. cities ranked among the 50 cities with the highest homicide rate in the world in 2022. The aggregate cost of crime in the United States is significant, with an estimated value of $4.9 trillion reported in 2021.

With the statistics proving the importance of crime prevention, there is an urgent need to showcase analytics to prevent these kinds of crimes. Categorizing crimes according to their types and cities and evaluating the cause of the crimes taking place in the country. To address the issue, we need to develop an analysis report and predict future crime rates to try and curb crimes.

The primary objective of this project is to provide a comprehensive overview of the latest advancements in the field and contribute to the prospects of future research. The field of crime prediction has grown in importance, requiring creative, data-driven solutions to improve on current detection systems. The scope for researching and applying machine learning and deep learning approaches to crime detection has greatly increased with the increasing availability of crime data and the ongoing advancement of technology. In order to examine how these cutting-edge technologies aid in the identification, predicting, and prevention of criminal activity, this data can be examined through an examination of current developments in machine learning and deep learning applications for crime prediction. The main goals are to contribute to the prospects of future research and offer a thorough overview of the most recent developments in the field.

INTRODUCTION

Predicting future crimes becomes extremely important when addressing the problem of reducing crime. The task of crime prediction is intricate and demands sophisticated analytical instruments to bridge the gaps in current detection methods. Researchers now have a unique opportunity to examine and research crime detection utilizing machine learning and deep learning approaches thanks to the rising availability of crime data and the progress of existing technologies. This article will examine current developments in machine learning and deep learning for crime prediction and highlight how these cutting-edge technologies are being used to detect criminal activity, predict crime patterns, and prevent crime. It is based on the latest advancements in this field. Our main objective is to offer a thorough summary of current developments in this area and support future research efforts.

In recent years, researchers have focused a great deal of attention on predicting crime using exploratory data analysis and prediction approaches, with an emphasis on finding patterns and trends in the incidence of crimes. Access to the datasets used by researchers to forecast crime is made possible by this study, which also examines popular methods for predicting crime using deep learning and machine learning algorithms. These methods offer insights into many patterns and variables associated with criminal activity. The study also identifies future directions and possible gaps that could improve the precision of crime prediction. Lastly, scholars in this subject will find great use for the thorough summary of studies on machine learning and deep learning approaches to crime prediction provided in this publication. Through acquiring a more profound comprehension of crime prediction methodologies, law enforcement agencies.

With the help of this project, we will be able to identify which areas in the United States are safer than others by analyzing various trends related to the crime rates that occur in various regions and creating distinct trends based on the types of crimes that occur in various cities and regions. We forecast and create a time series model forecasting the future crime timeline based on historical data in order to obtain additional insights into local and timely crimes. We discovered the underlying cause of crimes and the economic elements influencing them. To evaluate the correlations between the economic variables and crime rates, we examined correlational events and statistical techniques. Understanding how large incidents like thefts and violence affect specific people reveals insights and useful trends.

To sum up, this study sets out to explore the current state of crime analysis and forecasts future rates. It emphasizes how important machine learning and deep learning are to improving our understanding of crime trends and strengthening our group's overall efforts to deter crime more successfully. The initiative also highlights how law enforcement organizations may use these technologies to battle crime in our changing society in a more proactive and data-driven manner.

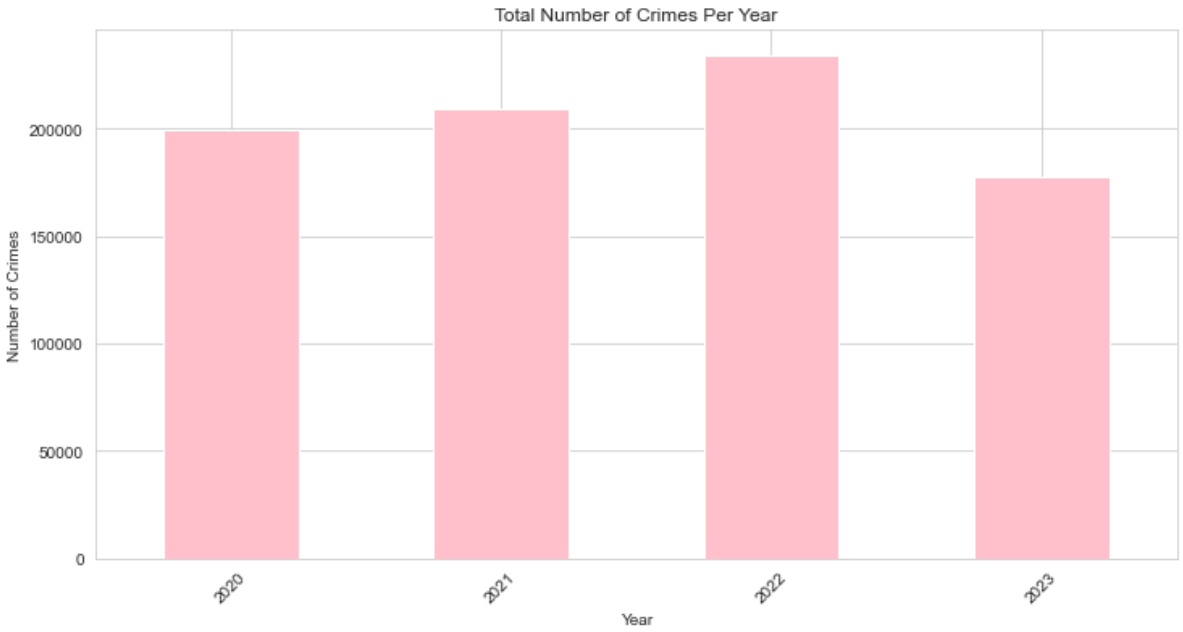
DATA SOURCES

The primary dataset utilized in this analysis was sourced from an official government website, containing incident records of crimes within the City of Los Angeles dating back to 2020. It's important to acknowledge that this dataset is a transcription of original crime reports, which were initially documented in paper format. Therefore, there may be some potential inaccuracies in the data due to the manual transcription process. Instances where specific location data was missing are denoted as (0°, 0°). For privacy protection, address information is provided only up to the nearest hundred block, with exact addresses withheld. It's crucial to recognize that the data's accuracy depends on the quality of the original records in the database. Any questions or concerns regarding data quality or specific data points are duly acknowledged and can be addressed through comments or inquiries.

Furthermore, in preparing the crime dataset, crucial steps were taken, encompassing data acquisition, thorough inspection, meticulous cleaning, and exploratory data analysis (EDA). Within this process, any redundant or irrelevant columns, such as URL references and date records, were pinpointed and subsequently eliminated to optimize the dataset. Additionally, the feature extraction technique was implemented to extract valuable insights from the available data. This rigorous data preparation is vital in guaranteeing the dataset's trustworthiness and pertinence for the comprehensive analysis carried out in this research. It serves as the fundamental framework upon which the subsequent findings and insights are built.

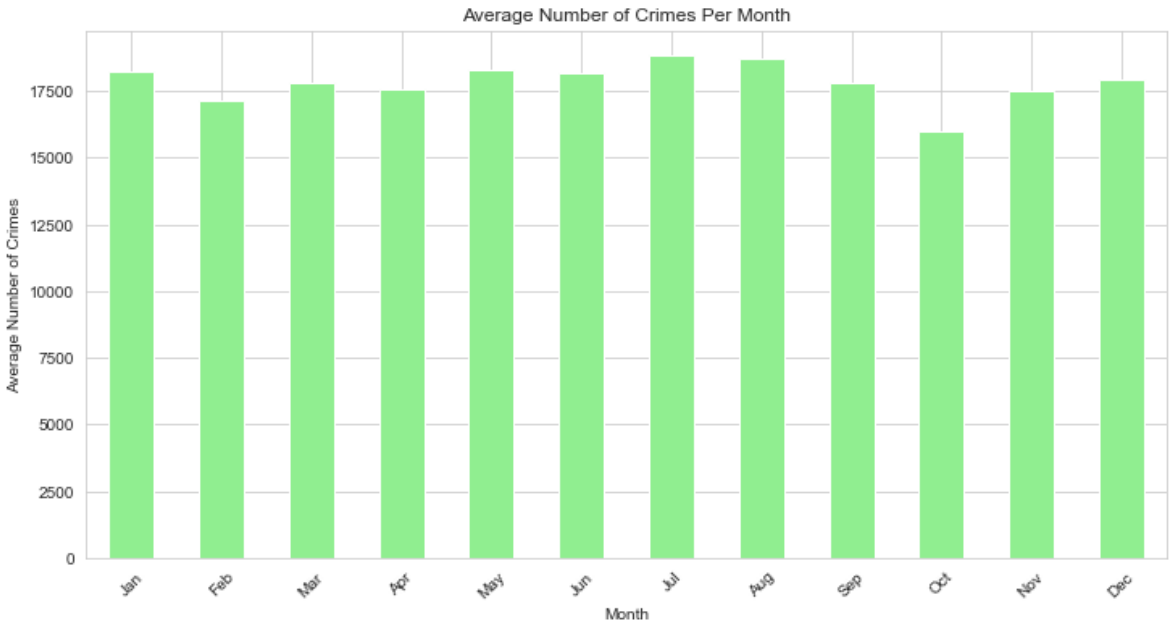
RESULT AND METHODS

1. Overall crime trends



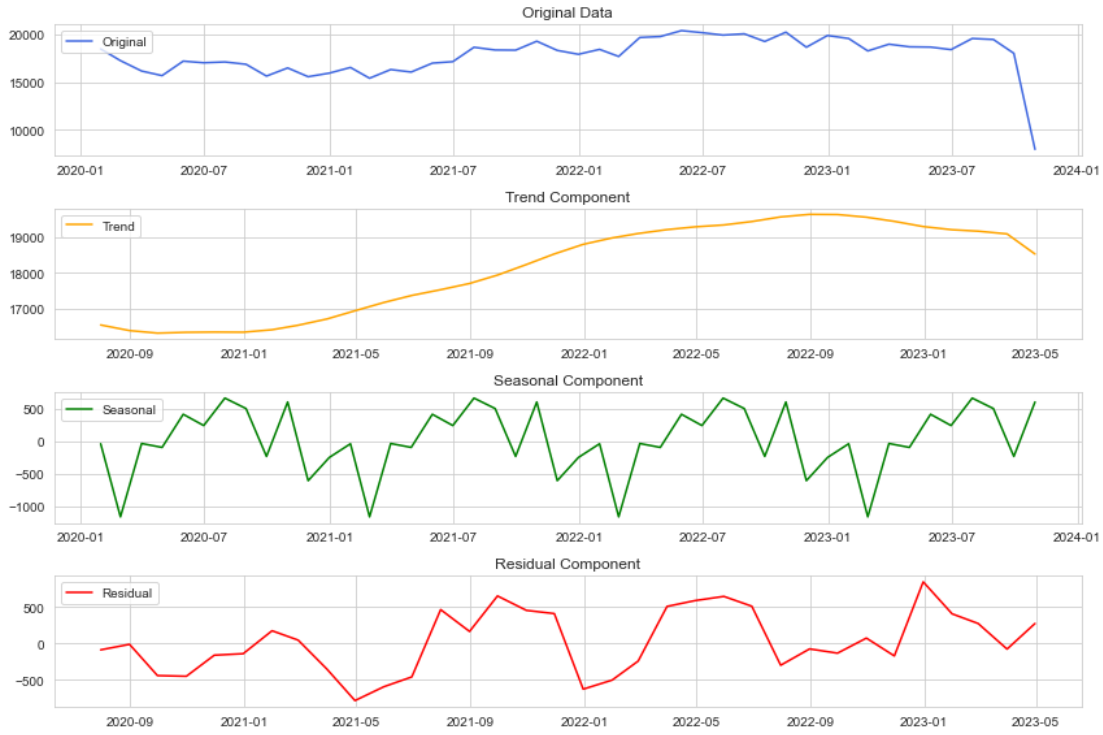
The bar graph visually illustrates the long-term trends in criminal activity. One can observe whether crime rates have been increasing, decreasing, or remaining relatively stable over time. 2022 had the highest number of crimes followed by 2021. 2023 had the lowest number of crimes. Some areas may experience cyclical patterns in crime rates, which could be related to economic conditions, population shifts, or other factors. Changes in law enforcement strategies or policies may be reflected in the trends. For example, the implementation of community policing initiatives or targeted crime reduction programs may lead to observable shifts in crime rates.

2. Seasonal Patterns



The analysis reveals how the number of reported crimes fluctuates throughout the year. Certain months may consistently exhibit higher or lower average crime rates. July has the highest average of crimes reported whereas October has the lowest average of crimes. Identifying seasonal trends in criminal activity can be crucial for law enforcement agencies and policymakers. This analysis can inform long-term planning for crime prevention strategies. By understanding seasonal patterns, policymakers can develop targeted interventions to address specific challenges during certain times of the year.

3. Analysis and visualization of seasonal patterns in crime data.



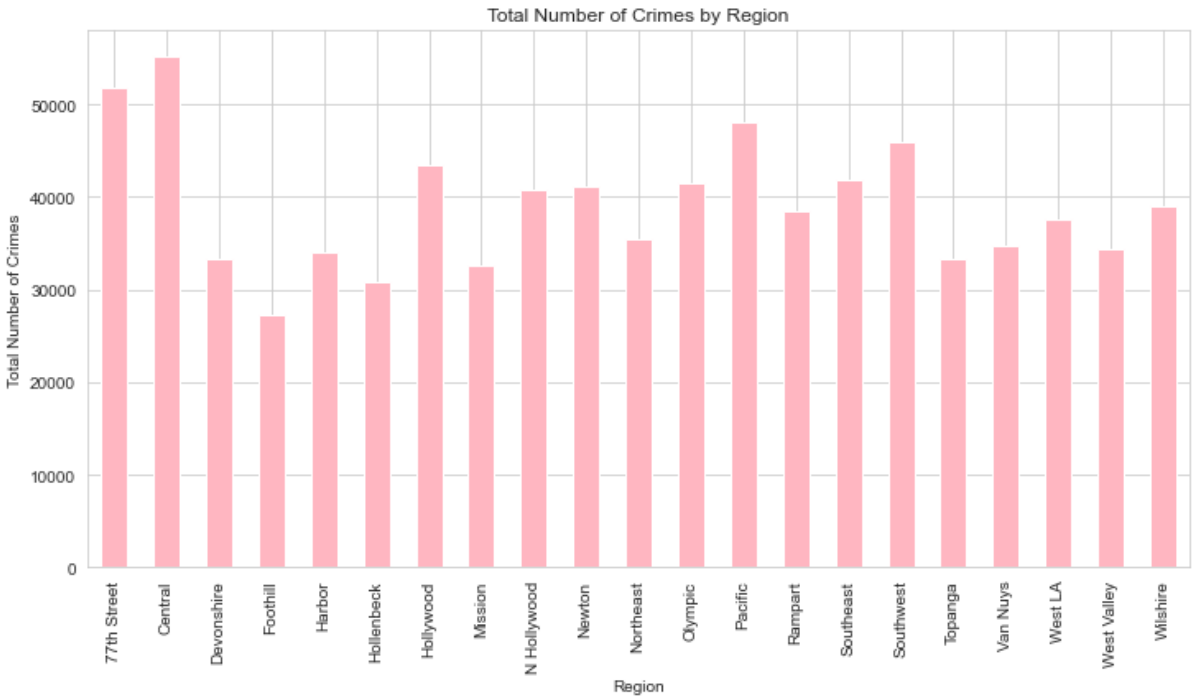
The graphs show the data by month, counting how many crime incidents occurred each month. This step creates a time series of monthly crime counts. These graphs are then separated in three parts:

- Trend Component: This graph represents the long-term pattern in the data, like whether the crime is generally increasing or decreasing over time.
- Seasonal Component: This graph captures recurring patterns or cycles in the data. For example, if crime tends to increase during certain months of the year, that would be a seasonal pattern.
- Residual Component: This graph is the part of the data that cannot be explained by the trend and the seasonal patterns. It represents random or irregular fluctuations.

4. Analysis of the relationship between the day of the week and the frequency of certain types of crimes

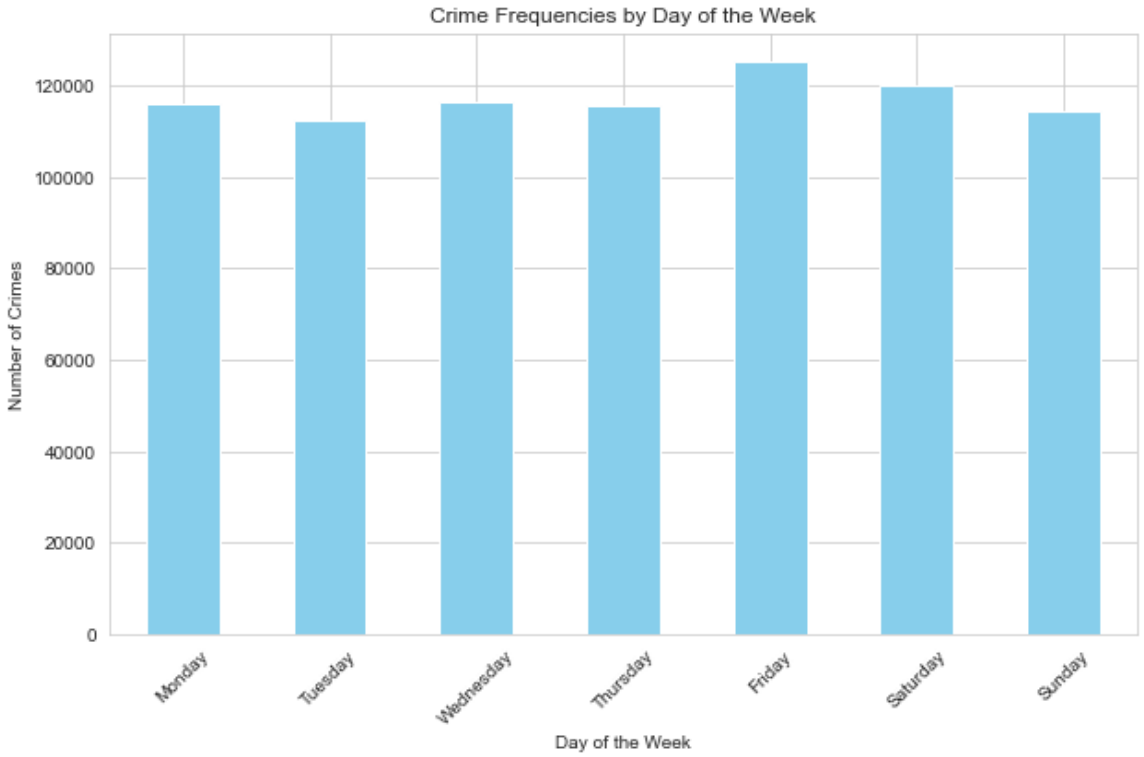
The bar graph visualizes the relationship between the day of the week and the frequency of certain types of crimes. Certain types of crimes may exhibit day-specific patterns. For example, Arson crime type is more prevalent on all days possibly influenced by factors such as heightened activity levels or specific opportunities that arise. Economic conditions and employment patterns may influence the types of crimes observed on specific days. Communities can use this information to implement targeted safety measures. For example, they may schedule neighborhood watch programs or increase security measures on days when certain types of crimes are more likely to occur.

5. Regional differences



The visual comparison of crime rates between different regions or cities highlights the disparities in criminal activity. Some regions may have significantly higher or lower crime rates compared to others. The “Central” region has a higher number of crimes (greater than 50000), and “foothill” has the lowest number of crimes (below 30000). The analysis helps identify regions with higher crime rates, often referred to as "hotspots." Conversely, regions with lower crime rates can be considered relatively safer zones. Law enforcement agencies and policymakers can use this information to allocate resources strategically. Higher crime regions may require increased police presence and targeted intervention programs.

6. Day of the week analysis

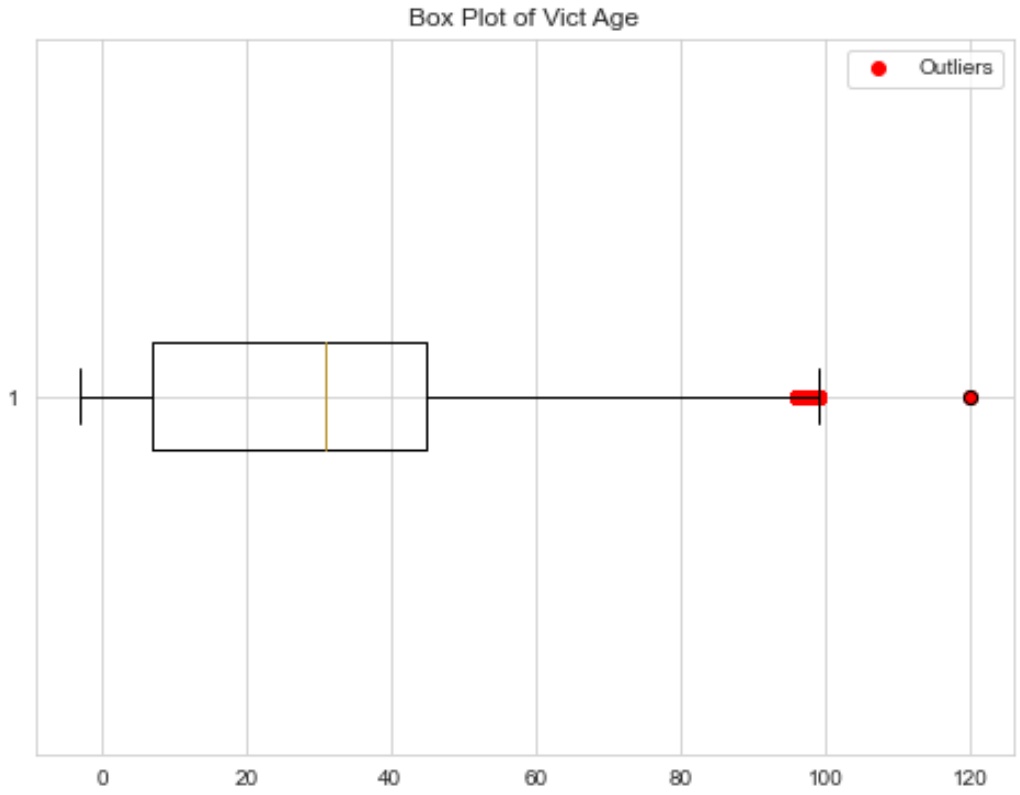


The resulting bar chart displays the number of crimes reported for each day of the week. The chart shows higher crime frequencies on weekends (Friday, Saturday and Sunday) compared to weekdays (Monday to Thursday). This is a common trend because weekends tend to have increased social activities, larger gatherings, and more people out and about, providing more opportunities for criminal activity. Economic considerations may play a role. For instance, some types of crimes, like theft or robbery, may be more prevalent when individuals are facing financial difficulties or looking for opportunities to steal. Within each day, crime rates may vary depending on the time of day. For example, nighttime hours may have higher crime rates due to reduced visibility and less public activity.

7. Demographic factors

The stacked bar graph represents the distribution of the crime rates across different age groups. It also shows the bifurcation of the crime types. This statistic is significant for understanding the age distribution of criminal activities in the dataset. We can conclude that fairly major part of the criminals belongs to the age range of 31-45. The criminal activities are the least between the age group of 0-18. However, this is considerable since minor crime rates should be as good as negligible. Hence, government can utilize the analysis output to organize necessary training sessions for minors.

8. Outliers and anomalies



Upon analyzing the data, we have spotted a few outliers. We have created boxplots and mapped the outliers, which indicate the values which are falling outside the usual range. The age of victims majorly lies between 5 to 42. However, we have observed from outliers around 100 and 120, indicating that there could be improvement in safety services being given to the elderly.

SUMMARY

In a nutshell, the analysis of crime data spanning from 2020 to 2023 unveiled several noteworthy trends. While crime rates exhibited a steady rise from 2020 to 2022, there was a substantial decline in 2023, potentially attributed to incomplete data. Interestingly, most months in 2023 witnessed a decrease in crime rates, except for January and February, which may have been influenced by policy changes. Seasonal examination indicated that the highest number of crimes occurred during the middle months of the year.

During the COVID-19 lockdown, there was a surge in identity theft, with crimes like credit card scams and loan fraud peaking in 2022. Analyzing the correlation between victim age and crime, it was apparent that individuals aged 0 to 20 experienced the highest number of crimes, with fluctuations in other age groups. Sundays consistently registered the highest crime rates.

In terms of geography, specific areas including 77th Street, Central, Hollywood, Pacific, and Southwest exhibited the highest crime rates. Gender-wise, crimes were evenly distributed between males and females, while other gender categories were less prevalent. Grasping the age and gender dynamics of crime holds significant importance for law enforcement and policy development. The diurnal pattern of crime frequency underscores the necessity for timely and data-driven strategies to tackle public safety challenges.

LIMITATIONS

The results of the exploratory data analysis are compiled to several factors.

1. The data being used could be lacking the dimension of completeness of data quality. For example, the crime rates seem to have deteriorated considerably in 2023. The conclusion of this could be considerable reduction in criminal instances. However, this interpretation could be misleading and could be a result of unavailability of data for that particular year.
2. Upon analysis, we have figured that the influence of policy changes has their impact on the crime rates. Further analysis would be required to confirm these findings. We may require data which is currently not available in the dataset. For instance, the access to legal documentation and contextual information can contribute to deeper analysis.
3. The dataset does not provide categorization of identity theft crimes. Hence, its classification and detailed analysis is restricted.
4. Subjective judgement in data and biased information can lead to problem of objectivity, reducing the data quality. There could be some other discrepancies and errors which could lead to false conclusions. Hence, the accuracy and reliability of the data sources can be major factors defining the relevance of the findings.

FUTURE SCOPE

In depth research in several verticals can be done to address the above-mentioned limitations and add to the better understanding of the crime trends.

1. We can incorporate directional steps and efforts to improve data quality and completeness. One of the ways to do this could be to include data from various sources for visualizing the crime trends more accurately.
2. Efforts can be made to study identity theft, its bifurcation and variance in its patterns over time. This will give analytical inputs for new laws creation and enforcement.
3. There is scope understanding the influence policy changes on the crime statistics. Induction of descriptive analysis of policy changes can help us do this. This analysis can be utilized to develop law enforcement statergies.
4. The analysis can be extended by including more specifics in the geographical analysis. There be inclusion of hotspots where the crimes are more often can be added. The neighborhoods of such hotspots can be made more aware and more force be allocated to in such areas.
5. Future crime trends can be calculated using predictive modelling. This will again contribute to the law enforcement strategies. Time series forecasting methods can also be used to construct prediction for future crime trends. This could add to strategic resource allocation as well.
6. Criminal activities can be linked to sociological factors. Inclusion of some personal background of the victims as well as criminals involved in the crimes can be helpful in figuring such factors.
7. Correlation analysis between crime and economic factors like unemployment rates, income ranges can get us to the root causes that cause the criminals to choose the path they do. This can help in taking appropriate measures in training the potential criminals.

All the above suggested analysis can lay the foundation of working on law enforcement, resource allocation, and putting efforts in shaping the criminal mindset.