CHICAGO : CRIME DATA ANALYSIS

PROJECT REPORT

of

IE 6600: COMPUTIZATION AND VISUALIZATION FOR ANALYTICS

BY

GROUP NUMBER 8:

ARCHIT SINGH (002813253)

GAURAV ASHVINBHAI RAMOLIYA (002837484)

DEPARTMENT OF COLLEGE OF ENGINEERING
NORTHEASTERN UNIVERSITY
BOSTON, MASSACHUSETTS – 022115
NOVEMBER 2023

ABSTRACT

This project involves a comprehensive analysis of crime data spanning from 2001 to the present, with a primary focus on data preparation and exploratory data analysis. Utilizing the Python programming language, we acquire the dataset and meticulously clean it by addressing issues such as missing data, duplicates, data type conversions, outliers, and categorical encoding.

The core of the project lies in Exploratory Data Analysis (EDA), where we visualize crime trends over time, identify seasonal patterns, and assess the most prevalent crime types. We also investigate regional and municipal variations in crime rates, explore correlations between economic indicators and crime rates, and study the relationship between the day of the week and specific crimes.

Moreover, we scrutinize the potential impact of major events and policy changes on crime rates. To facilitate these analyses, we ensure the presence of essential Python libraries. This project is expected to provide valuable insights into crime data, contributing to a deeper understanding of public safety dynamics.

ACKNOWLEDGEMTS

We wish to express our deep gratitude to those individuals who played essential roles in the successful completion of this data analysis report. Our collaborative efforts, commitment, and teamwork were the driving forces behind this project. We are thankful for the guidance and support provided by the following people:

Professor Sivarit (Tony) Sultornsanee
Associate Teaching Professor of Mechanical and Industrial Engineering

Professor Sivarit Sultornsanee's expertise and mentorship were instrumental in shaping the direction of our analysis. We are appreciative of the valuable insights and guidance he provided during the project.

Teacher Assistant – Fenil Savani, Mrudula challagonda

Team Members
Archit Singh, Gaurav Ashvinbhai Ramoliya

Our exceptional team members deserve our profound thanks for their unwavering commitment and collaboration. Together, we addressed various aspects of this project, including data sourcing, data cleaning, data analysis, and reporting. The project's quality and success would not have been achievable without their hard work and dedication.

This report stands as evidence of the exceptional teamwork and camaraderie that characterized our project. We take pride in working with such talented and cohesive team members. Our heartfelt appreciation goes out to everyone involved in this endeavour for their invaluable contributions.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.0 INTRODUCTION

In this project, we shall undertake an extensive analysis of crime data spanning from 2001 to the present. Our primary objective is to meticulously clean and prepare the dataset for a comprehensive analysis. This will encompass an exploration of crime trends, patterns, and the underlying factors influencing crime rates.

We shall employ the Python programming language as our primary analytical tool. The project commences with the acquisition of the dataset from the designated source and its subsequent integration into our data analysis environment. Subsequently, we shall conduct a meticulous examination of the dataset, including an initial display of its early records, an evaluation of data types for each column, and a review of column names and descriptions.

The data cleaning process will address issues of missing data, duplicate records, data type conversions, outlier management, and normalization. Furthermore, categorical data shall be encoded where applicable. The heart of the project lies in the Exploratory Data Analysis (EDA) phase, which entails visualizing crime trends spanning the entire temporal spectrum, identifying seasonal patterns, determining the most prevalent crime types, and discerning regional and municipal disparities in crime rates.

A sophisticated examination will explore correlations between economic indicators, where available, and crime rates. We shall delve into the relationship between the day of the week and the prevalence of specific crimes. Additionally, we shall scrutinise the potential influence of major events and policy alterations on crime rates.

To effectively execute these tasks, it is imperative to ensure the presence of essential Python libraries, including Pandas, Matplotlib, NumPy and Jupyter Notebook within the working environment. This project promises to deliver profound insights into crime data, offering valuable contributions to the understanding of public safety dynamics.

CHAPTER 2: DATA SOURCING AND CLEANING

The primary dataset utilized for this analysis was sourced from a government website and covers incident records of crimes occurring within the City of Chicago, with records dating back to 2001.

It is important to remember that this dataset is a transcription of the original crime reports that were first recorded on paper. As a result, errors could occur in the data as a result of the manual transcription procedure. Instances when precise location data was absent are denoted with the symbols (0°, 0°).

Address information is purposefully shortened to the closest hundred blocks in order to preserve people's privacy; precise addresses are not revealed. It's critical to recognize that the correctness of the data depends on the caliber of the original records stored in the database. Any queries or comments can be used to address any worries or questions about certain data points or the quality of the data.

Furthermore, the preparation of the crime dataset involved several critical steps, encompassing data acquisition, inspection, cleaning, and exploratory data analysis (EDA).

In order to simplify the dataset, extraneous or unnecessary columns—like URL references and date records—were found throughout this procedure and eliminated. The feature extraction technique was also used to glean relevant insights from the existing data. This thorough data preparation is essential to guaranteeing the validity and applicability of the dataset for the extensive analysis that this study is conducting.

It forms the foundational basis upon which the subsequent findings and insights are constructed.

The data-cleaning process for the crime dataset involved a sequence of methodical actions:

Preliminary Cleaning:

1. Initial assessment to pinpoint the presence of missing values in the data.

Column Standardization:

1. Functions were implemented to exclude unnecessary columns.

Data Enhancement and Error Handling:

1. A calculation of victim gender distribution was conducted.
2. The dataset was augmented with a binary indicator to reflect the use of weapons in crimes.
3. Default values were assigned to missing entries in the 'Premises Description' and 'Premises Code' fields.

Date and Time Formatting:

1. Missing 'Cross Street' information was incorporated into the 'Location' column.
2. The 'Date Reported' field was converted to a standard datetime format.
3. Time of occurrence entries were converted into an 'hour: minute' format.

Integration and Final Touches:

1. The dataset was restructured by combining the date and time of crime occurrence into a single column.
2. Age data was refined to include only logical age ranges.
3. Superfluous columns were discarded to streamline the dataset.

Post-Cleaning Validation:

1. The cleaned dataset was then examined for any remaining missing values.
2. Additional time-related features such as year, month, and a concatenated year-month string were extracted from the occurrence datetime to aid in temporal analysis.

CHAPTER 3: VISUALIZATION AND ANALYSIS

In Chapter 3, we explore the project's data through a series of visual representations and the subsequent analysis of these visuals. The objective is to translate complex data sets into understandable formats that can inform our questions and objectives.

3.1 Overall Crime Trends:

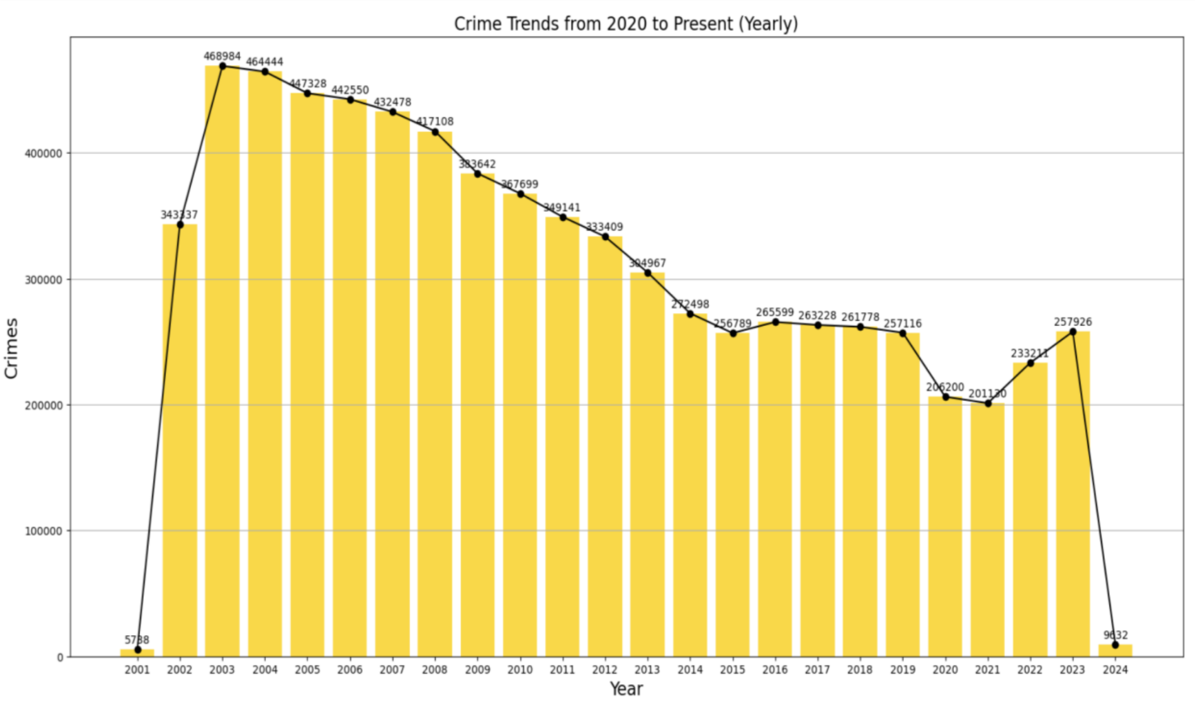


Figure 3.1 Yearly Average Crime Trends

The extensive examination of crime data spanning from 2001 to present has uncovered several noteworthy insights into the patterns of criminal activities. Firstly, it highlights a rising trend in crime rates, with the highest occurrences in 2003, closely trailed by 2021 and 2020. Interestingly, there was a significant surge in crime in 2002 to 2003.

We lack data for 2024 in our dataset, preventing us from making any conclusions or representing the sharp decline during that period in our graph.
Additionally, in a rapidly advancing world, crime has taken on new and diverse forms, contributing to its overall increase.
The histogram clearly illustrates that the year 2003 witnessed the highest frequency of criminal incidents compared to all the other years.

3.2 Seasonal Patterns:

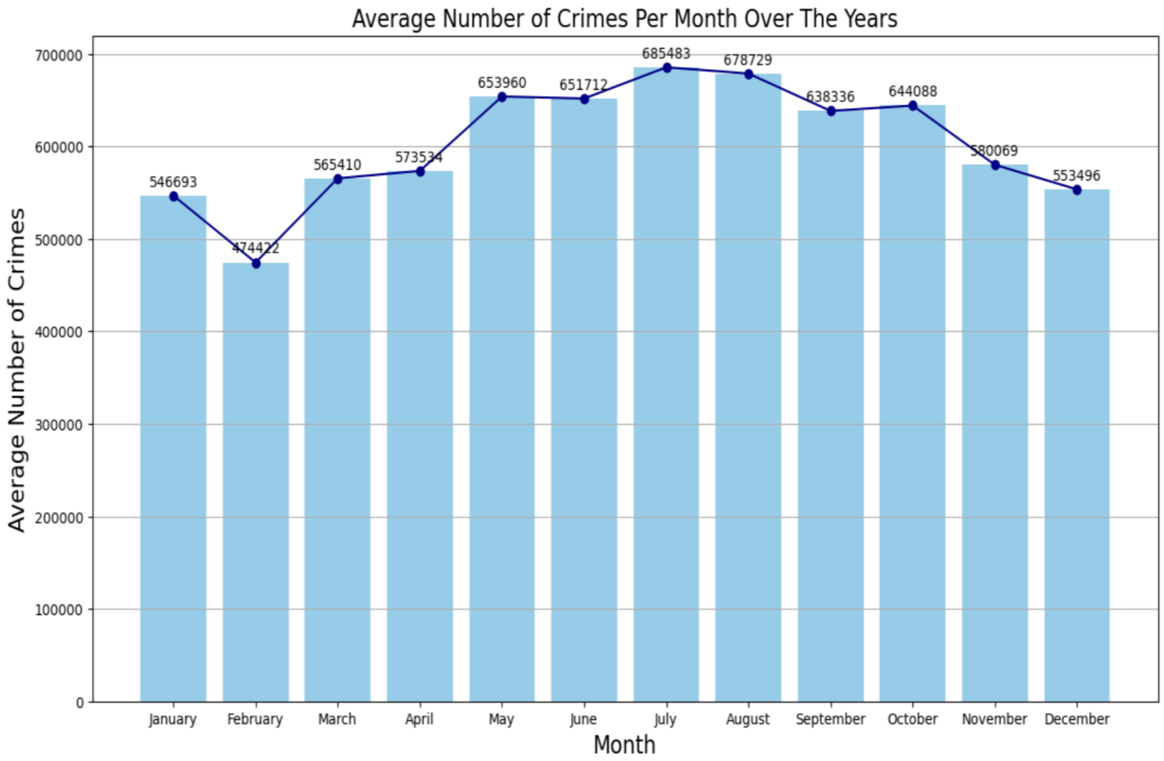


Figure 3.2 Monthly Crime Trends

Insights drawn from the graph reveal that the months of July, August, and September witness the highest crime rates, whereas November and December experience the lowest incidents. This pattern may be linked to warmer weather conditions, leading to larger gatherings and increased crime opportunities during the summer months. Conversely, during the holiday season in November and December, heightened law enforcement presence acts as a deterrent to criminal activity. Moreover, the extended daylight hours in summer provide additional windows for criminal incidents.

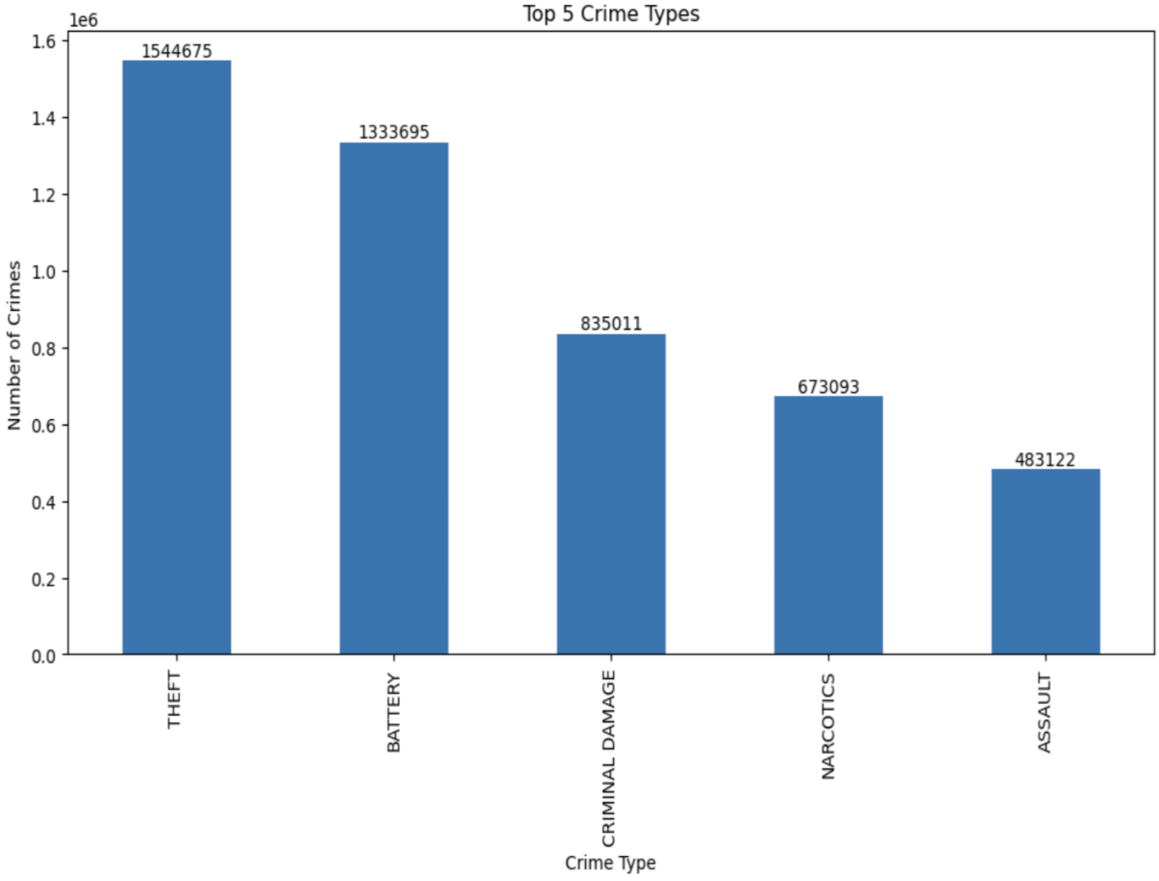
We now analyse the top 15 contributors to the crime data.

3.3 Most Common Crime Type:

It can be inferred that theft stands out as the most prevalent, followed by battery and criminal damage. These crimes are not only frequently committed but are also relatively easier to execute, possibly contributing to their higher incidence. Furthermore, the penalties for such offenses tend to be less severe, potentially encouraging their perpetration.

On the other hand, vandalism and shoplifting appear to be less common, possibly due to the Extensive presence of surveillance cameras on streets and within malls. This heightened surveillance leads to quicker identification of wrongdoers, which, in turn, results in more effective enforcement of punishments and serves as a potential deterrent against these crimes.

Figure 3.3 Top 5 Most Common Crimes

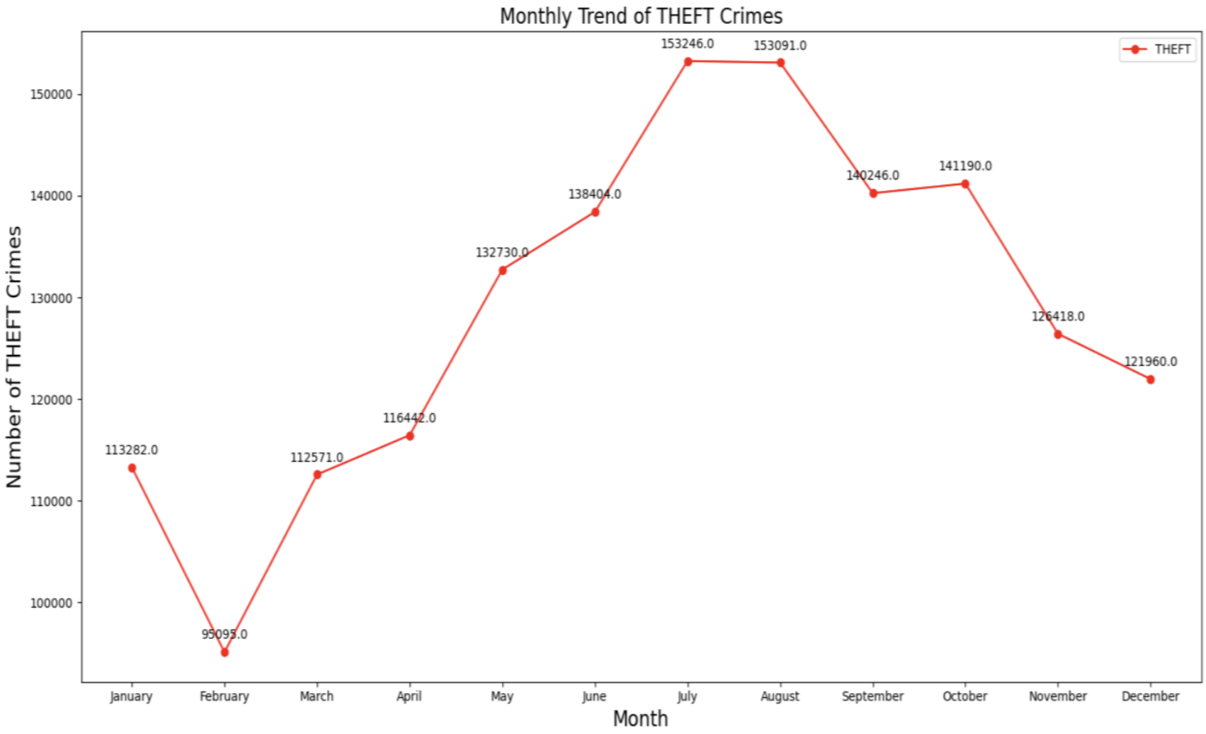


3.4 Monthly Analysis for THEFT crime:

Our analysis of theft crimes over a multi-year period aimed to identify significant seasonal and monthly trends. The graphical representation demonstrates the monthly volumes, with a visible peak in July which reaches levels markedly higher than other calendar cycles. Specifically, July experienced 153246 theft incidents per year - a concerning figure approximately. Mid-summer months see considerably elevated theft volumes, while December and January declines to monthly lows likely tied to holiday travel declines.

The peak observed in July warrant increased vigilance during this heavy theft season. Possible policy considerations connect seasonal staffing boosts in prevention and security roles to correlate with predictable incident spikes. While theft declines post-July, August theft numbers remain above Q1/Q4 lows - illustrating that one month alone of heightened focus will not address seasonal effects lasting several summer months. Sustained theft mitigation campaigns through August may help taper the July-centered peak.

Evaluating underlying drivers, summer and warm weather months typically see increased activity levels that likely enable more theft crimes. Vacations and travel also peak during summer, increasing vulnerabilities. Reviewing temperature, tourism and public events data could help confirm seasonal drivers. Please let me know if a report summary focusing on interpreting the monthly theft data trends and seasonal impacts would benefit from any other specific detail or focus areas. I can continue building out this analysis as needed.

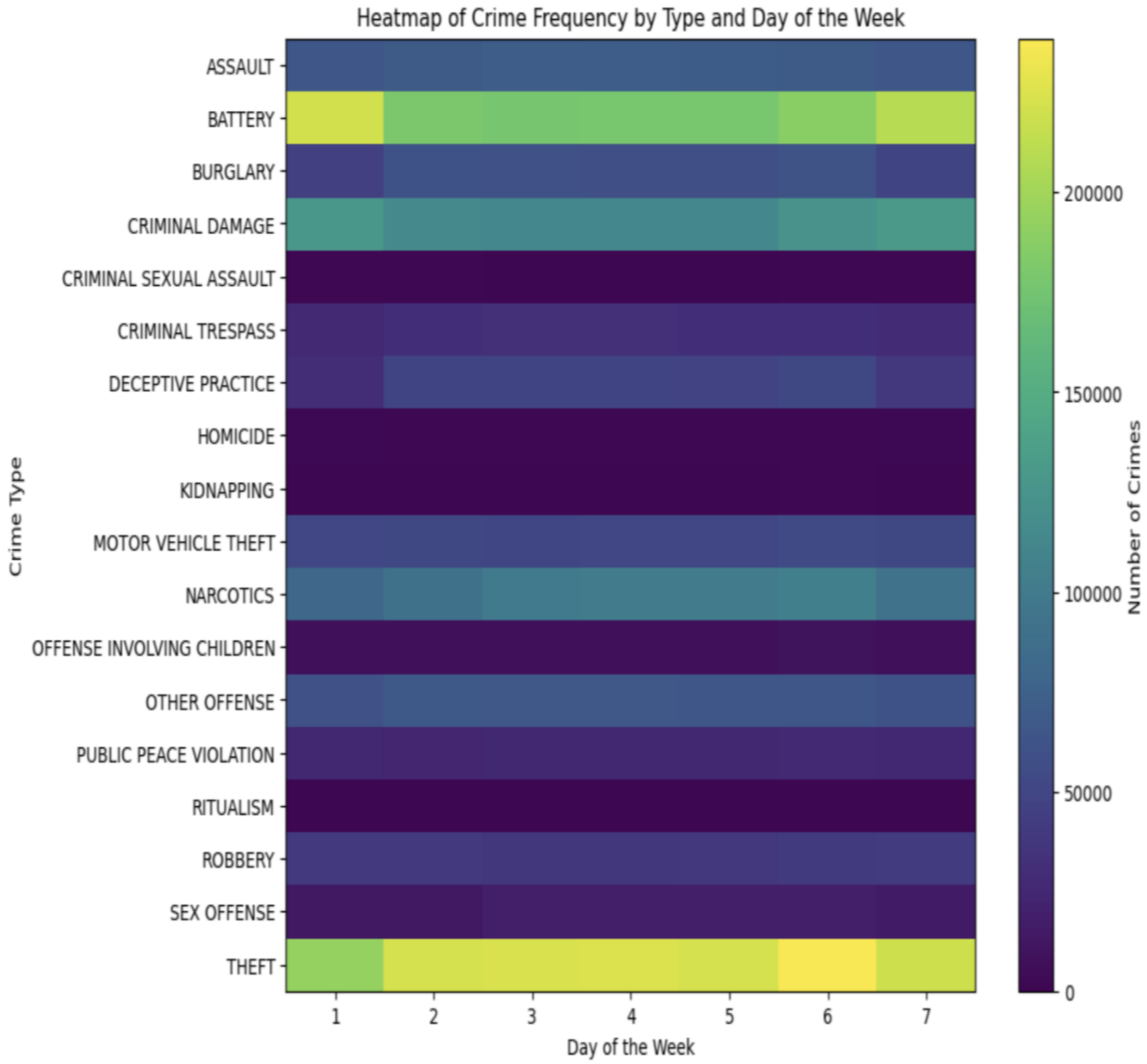


3.5 Heatmap Analysis:

The data indicates relative consistency. Theft and battery crimes maintain the highest frequencies every day of the week. On Mondays the average theft rate reaches more than 200000 incidents. Battery also hovers between 170000-200000 counts regardless of weekday. In contrast, more severe crimes like homicide and kidnapping remain less common occurrences. Even on peak weekdays, kidnappings and homicides emerge at rates considerably below higher-frequency crimes.

With exceptions for unpredictably severe crimes, steady weekday crime rates suggest resource allocation could align to typical daily volumes without significant weekday fluctuations. More analysis regarding weekend activity and overnight shifts when crimes may concentrate could reveal staffing optimization opportunities.

However, evaluating weekday consistency by more granular Crime Type, as well as Specific Location and Time of incidents, could expose more variability in crime rates. While daily fluctuations appear negligible in aggregate views, localized volatility may inform community policing tactics.



3.6 Crime analysis in district:

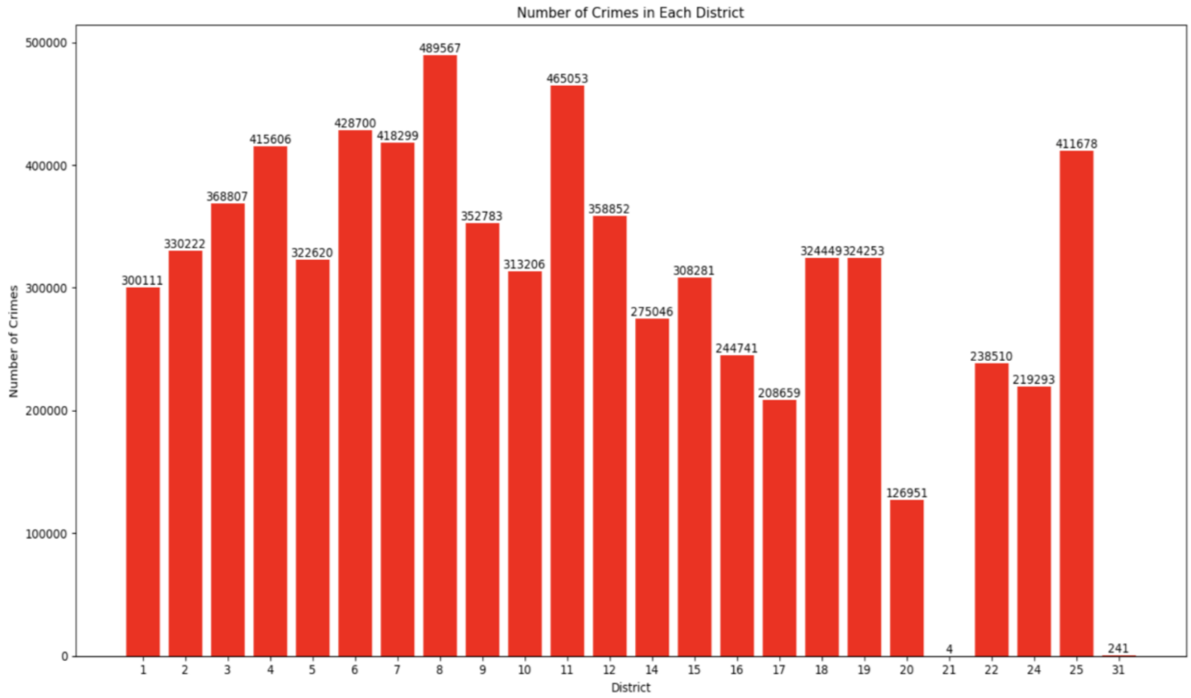


Figure 3.6 District with most number of crimes

In the above analysis, it can be observed that geographical analysis of crime rates by district provides actionable insights into where incidents concentrate. The data reveals sizable divergence across areas. Over a multi-year span, these districts each averaged over 400,000 reported crimes - figures that significantly exceed nearby districts.

Mapping yearly incident figures by district visually distinguishes the highest density areas. We can focus prevention resources, analytics, and law enforcement efforts directly into the three districts exhibiting such pronounced crime levels. Conversely, the graphical view also quickly highlights districts sustaining relatively low crime rates. Districts 17, 20 and 21 maintain the lowest occurrences, though we still should not rule out crime prevention activities in these areas.

Evaluating demographic, income and population density factors by district may reveal influences driving the concentration of crimes into specific divisions. Additional analysis can inform customized strategies, appropriate resourcing and tailored community outreach by district.

3.7 Analysis of District 8’s crime:

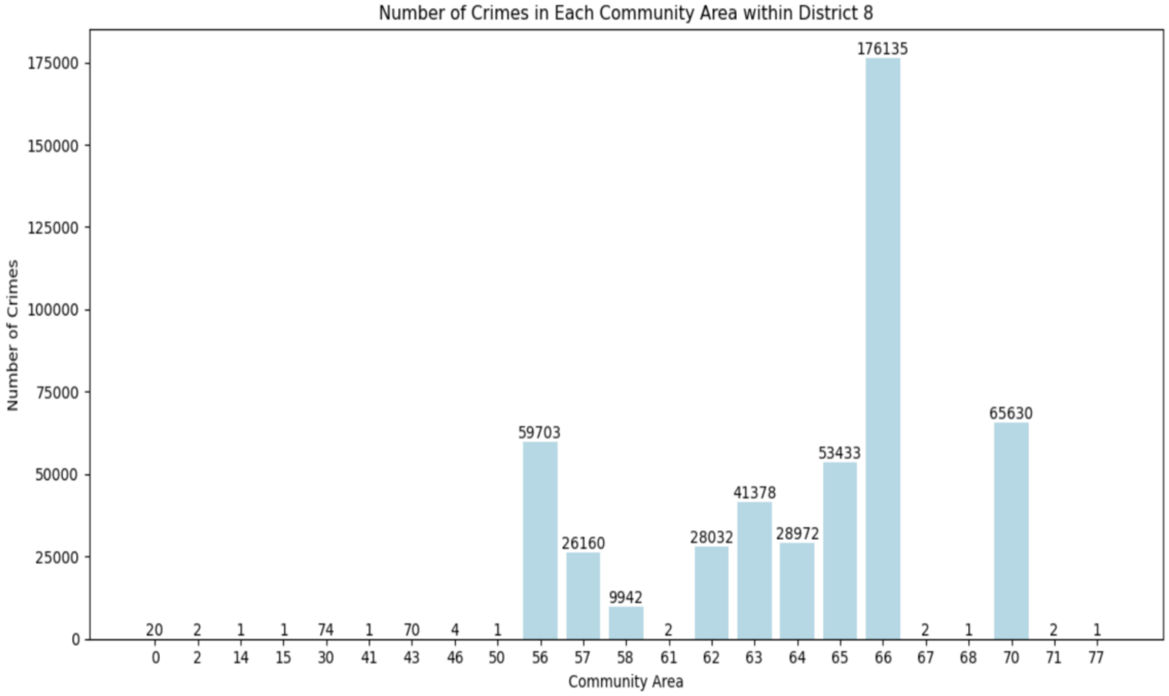


Figure 3.7 District 8 with most number of crimes in community area

District 8's sizable geographic span covers communities exhibiting wide divergence in crime volumes. The data indicates crime concentrates heavily into specific areas within the district. Most notably, the 66 community sustains markedly higher crime levels versus all other District 8 areas. 66 averaged over 175,000 total crime incidents during the analysis period - nearly double the next highest community.

Mapping granular crime statistics by community area spotlights community area 66 as an priority area for preventative measures. Additionally, Visualizing community-level crime data identifies the areas sustaining relatively low occurrences compared to community area 66. 57 and 65 communities see less than a fifth of community area 66 yearly crime totals, enabling resource prioritization. Evaluating income levels, resident demographics, law enforcement staffing and other factors differentiated across District 8 communities can reveal influences on disproportionate crime levels in community area 66. These analytical insights can inform effective, localized responses.

3.8 PROBABILITY OF EACH CRIME TYPE IN DISTRICT 8, COMMUNITY AREA 66:

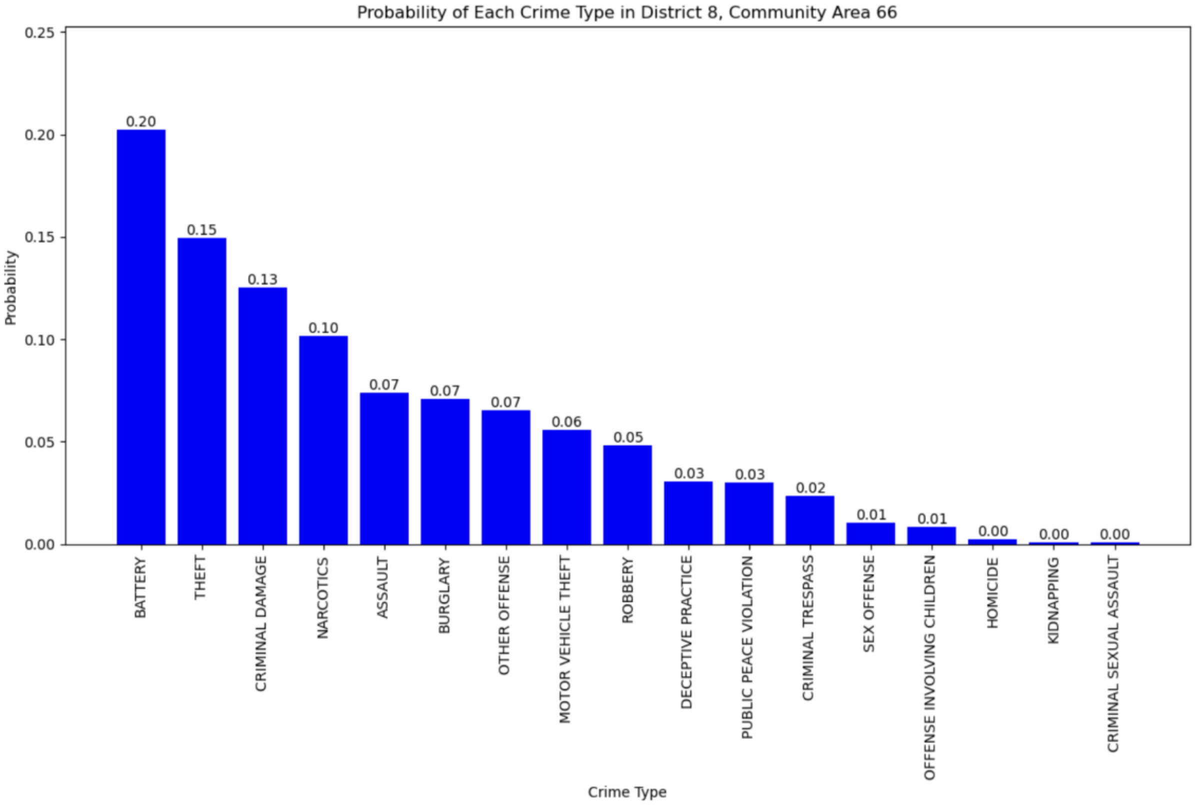


Figure 3.8 Probability of each crime type in community area 66

Crime probabilities across different incident types reveals consistent trends over the 5-year period reviewed. Battery and theft categorically maintain the highest likelihood of occurrence year-over-year.

Specifically, nearly one in every 10 crimes reported falls into the battery classification - the most common violation by a significant margin. With over 200000 annual battery incidents, mitigating this prevalent crime type could provide widespread community impact. Similarly, theft emerges in approximately 8% of incidents, also considerably exceeding other violation probabilities. As common property crimes with overlapping root causes, tactical approaches tailored to reducing battery and theft simultaneously could optimize prevention resources.

Conversely, the least probable incidents by magnitude relate to homicide, arson and kidnapping events. Despite their severity, these collective crime categories each show less than a 1% chance of occurrence. Thus, these outliers may warrant specialized investigative teams rather than extensive community policing concentrates. Comparing crime probabilities guides strategic resourcing recommendations based on violation frequency versus severity. We can deploy community safety resources commensurate to incident likelihood, while escalating acute investigation assets to less-common homicides.

3.9: PROBABILITY OF ARREST IN DISTRICT 8, COMMUNITY 66 VS WHOLE DATASET

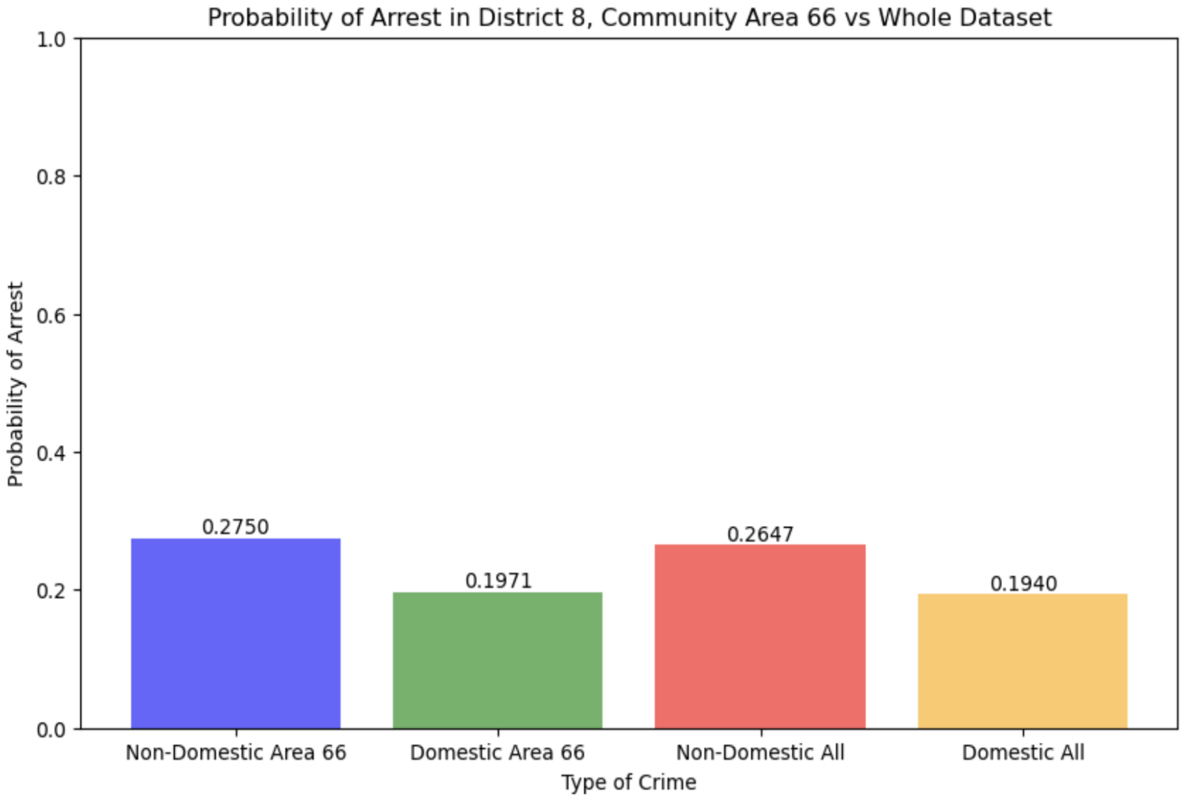


Figure 3.9 Probability of Arrest

Our analysis of arrest probability differences across domestic and non-domestic crimes aimed to identify any variance for Community Area 66 specifically, as well as citywide patterns. The data reveals sizable divergence.

In Area 66, non-domestic crimes, such as assaults and batteries between strangers or acquaintances, show nearly double the arrest likelihood versus comparable domestic disputes between family members. This data in arrest probability underscores a reluctance of victims to pursue charges against family members, despite violence similarities in domestic situations. It suggests a need for additional community engagement and law enforcement training to appropriately handle domestic issues. These views can inform appropriate response procedures.

3.7 Impact of Major Events:

a) George Floyd Protest

b) US Elections 2020

c) Minimum Wage for selected immigrant workers

d) Covid-19

CHAPTER 4 : LIMITATIONS

It is important to consider the limitations of the analysis. First off, the notable decline in crime rates in 2024 suggests that there might be incomplete data. It's critical to understand that this drop may not actually reflect a decline in criminal occurrences but rather the result of incomplete data for that particular year.

Second, the veracity and correctness of the crime data sources that were consulted for the analysis can greatly influence how credible the results are. Inaccuracies, biases, or contradictions in the data may cause conclusions to be drawn that are not accurate.

Moreover, although the research indicates that policy changes can affect crime rates, a deeper comprehension of these changes and their consequences would require access to legislative documents and contextual data, which might not be included in the dataset.

Finally, the dataset's inherent constraints and the way these crimes are classified may limit the classification and in-depth study of identity theft crimes.

CHAPTER 5: FUTURE WORK

Future study could go in a number of directions to get around these restrictions and enhance our understanding of crime trends. To ensure a more accurate portrayal of crime patterns, it is imperative to prioritize improving the quality and completeness of the data, potentially through the integration of data from numerous sources. Future studies could concentrate on a thorough examination of policy modifications in order to determine their impact on crime rates and investigate the outcomes of law enforcement tactics.

Expert research on identity theft, methods, and how these trends evolve over time may provide crucial information for cybersecurity and law enforcement initiatives. Additionally, carrying out a more thorough geographic study that identifies crime hotspots at the neighborhood level could offer better direction for allocating law enforcement resources.

Proactive law enforcement tactics may be made possible by the development of predictive models to foresee future trends in crime. Subsequent investigation of the characteristics of criminals and victims may reveal underlying societal causes influencing crime rates. In order to allocate resources and prevent crime, law enforcement organizations could use time-series forecasting techniques to anticipate future trends in crime.

Ultimately, a more comprehensive understanding of the elements that motivate criminal conduct would emerge from examining the relationships between crime rates and numerous social and economic variables, including income levels and unemployment rates. Through tackling these constraints and conducting additional research in these domains, scholars and law enforcement organizations can improve their comprehension of criminal trends, bolster deterrent tactics, and adjust to the evolving terrain of illicit activity and societal transformations.

In conclusion, SARIMAX (Seasonal Autoregressive Integrated Moving Average) yielded better forecasting accuracy compared to the Prophet model.

CHAPTER 6: CONCLUSION

The analysis of crime data spanning over two decades has provided valuable insights into crime trends, patterns, and influencing factors. Through meticulous data preparation and exploratory data analysis, we have uncovered significant findings that contribute to a deeper understanding of public safety dynamics in Chicago.

Key observations include the rising trend in crime rates over the years, seasonal patterns with higher crime rates during summer months, and the prevalence of certain crime types such as theft and battery. Geographic analysis has highlighted areas with high crime concentrations, enabling targeted preventive measures and resource allocation. Furthermore, the examination of arrest probabilities and the impact of major events on crime rates have underscored the complex dynamics at play.

However, it's important to acknowledge the limitations of the analysis, including potential data incompleteness and accuracy issues. Future research should focus on improving data quality, examining policy changes' impact on crime rates, and delving deeper into specific crime categories such as identity theft. Additionally, the development of predictive models and further exploration of socio-economic variables' influence on crime rates could enhance proactive law enforcement strategies.

Overall, this project has laid a foundation for continued research into crime dynamics, with implications for law enforcement, policymaking, and community safety initiatives.