EDA & PRE-PROCESSING:

In our data preparation journey, we initiated the process by loading a dataset of nearly 8 million records, each containing a wealth of information across multiple dimensions of crime incidents in the Chicago region. This included crime type, location, crime description as well the time. This dataset was then methodically cleaned to enhance its quality and usability.

The process involved converting date-related entries to a standardized datetime format to facilitate temporal analyses. Further refinement was achieved by segmenting the date information into discrete date and time elements, thereby simplifying the analysis of criminal events. Additionally, to ensure the integrity and quality of the dataset, all null values present within the columns were appropriately removed.

To ensure a focused and relevant dataset, we opted to exclude columns that were less relevant to our analysis, such as 'Ward' and 'Community Area'. To derive meaningful insights, we introduced a new categorization, 'Time Category', enabling us to classify crime incidents into distinct time segments of the day, thus painting a clearer picture of crime trends over time.

As we look ahead, our analysis will delve deeper into data cleaning and exploratory analysis. We will engage in advanced feature engineering to uncover hidden patterns and prepare the groundwork for predictive modeling.

ANALYSIS PLAN:

a. Understand the nature and frequency of crimes committed in Chicago from 2001 to 2024.
b. Identify any patterns or trends in the crime data using various machine learning methodologies
c. Learn about crime association in 2024.
d. Use geo-spatial clustering for identifying crime hotspots in Chicago.
e. Use classification reports to predict which crime type occurs the most.

METHODOLOGY:

The plan is to use various statistical and machine learning techniques to analyze the dataset and gain insights. The chosen methods include:

a. We will implement K-means clustering to identify patterns and hotspots in crime occurrences throughout Chicago. This method will allow us to understand the spatial distribution of crimes by partitioning crime locations into clusters based on their geographical coordinates. Through this approach, we aim to highlight areas with high crime density and provide a clear picture of the spatial distribution of crimes across the city.

b. We plan to use Hierarchical Clustering to gain deeper insights into the structure of the crime data. This clustering technique, which does not require pre-specification of the number of clusters, will offer a nuanced view of how crime incidents group together based on various attributes, such as time of day or crime type. We expect this method to complement our K-means analysis by offering a different perspective on the clustering of crime incidents.

c. Our analysis will include the use of Term Frequency-Inverse Document Frequency (TF-IDF) and Word Clouds to analyze the textual data, such as crime descriptions, within our dataset. By employing TF-IDF, we aim to understand the significance of various terms in the dataset, particularly those that are frequent in specific types of crimes but not common overall.

d. Word Clouds will be utilized to provide a visual representation of these significant terms, offering an intuitive understanding of the nature of the crimes. Moreover, we enhanced our TF-IDF model to create a classification report for tokenized text using the "Text" attribute in the "Cleaned_Chicago_df". We will also use the TF-IDF Vectorizer to create distribution of predicted crime labels.

e. We will utilize Association Rule Mining to uncover relationships between different types of crimes and other attributes in our dataset, such as time or location. This process will involve identifying frequent itemsets and then deriving strong association rules from these sets. Our goal is to discover patterns, such as specific crimes occurring more frequently in certain areas or times, which could provide valuable insights for crime prevention strategies.

PRELIMINARY RESULTS:

A. Clustering:

a. Used Hierarchical Clustering to analyze crime data based on crime type and time of day. This method, employing the 'complete' linkage approach, revealed three significant clusters in a dendrogram, indicating distinct crime patterns at different times.

b. K-means Clustering was also applied, where the optimal number of clusters (k=3) was determined using the Elbow Method. Both techniques provided valuable insights into the criminal activities in Chicago.

B. TF-IDF:

Using TF-IDF technique on Chicago crime data, we trained a model to spot 'Homicide' instances. It reached 99% accuracy, which, while impressive, may hint at overfitting or data inconsistencies that need further investigation.

After creating a classification report, we found out that the crime type "Arson" had the highest precision. "Assault" had the highest recall.

C. Association Rules:

Using the apriori algorithm on Chicago's 2024 crime data, we uncovered frequent combinations like battery and assault, with narcotics often preceding various crimes.

This analysis highlights the linkage between drug-related offenses and violent outcomes, such as assault and criminal damage. Property crimes like theft and criminal damage are common follow-ups, suggesting an escalation pattern from initial offenses to more severe property-related crimes.

Our time-of-day analysis revealed consistent crime patterns, underscoring the potential for law enforcement interventions at early stages to disrupt these sequences and prevent subsequent crimes.

NEXT STEPS:

Since the initial part of this project is completed, after facing so many challenges we have decided to take some important steps in the future.

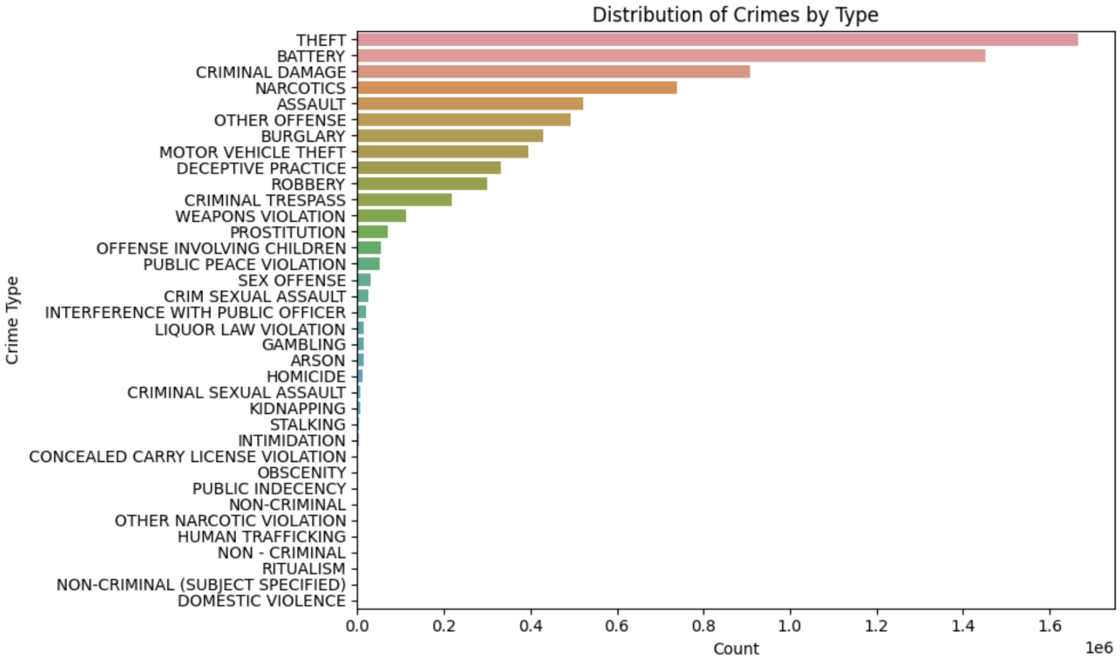
- We will utilize Big Data tools like Hadoop to handle big data since this project is working with a very huge dataset. We will also plan to use pySpark.
- Furthermore, we will expand more on sentiment analysis even though this Chicago dataset is completely about crime (negative sentiment analysis only).
- Additionally, for this dataset, we had to remove a lot of columns and rows because they contained insufficient information, hence for future analysis to ensure 100% model accuracy, we will use geo-encoding to encode longitude and latitude data from addresses.

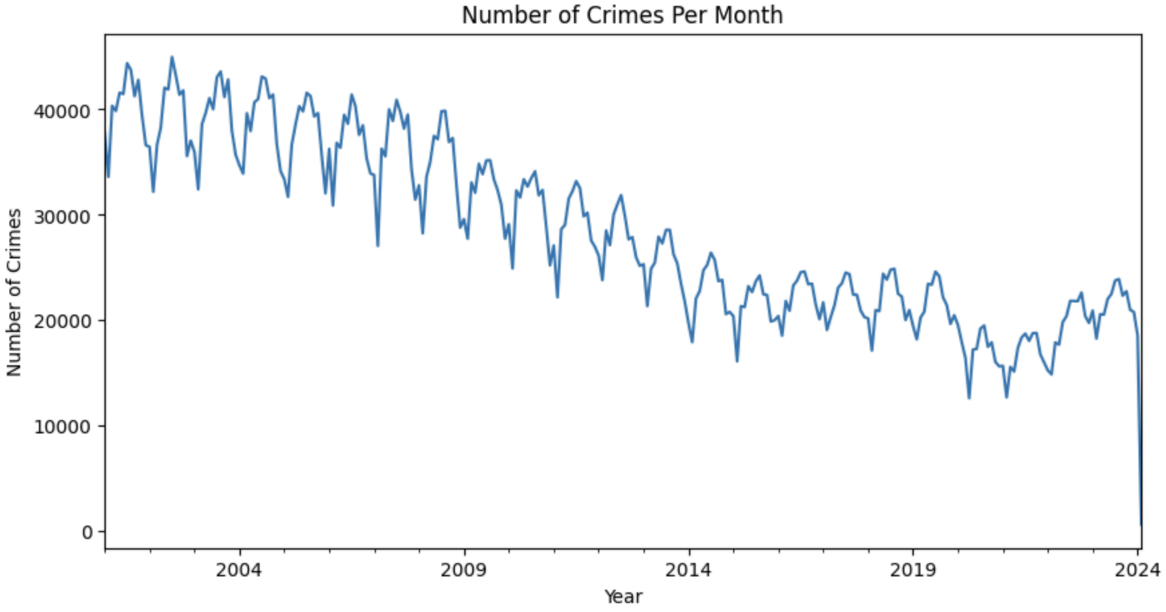
Member	**Shubham**	**Shayan**	**Ali**	**Sulaiman**
Contribution	**30%**	**30%**	**20%**	**20%**
Work Done by Subtopics	Proposal Documentation Slides K Means TF-IDF EDA Charts Data Cleaning Notebook Overview	Proposal Documentation Slides EDA & Github Upload Data Cleaning Notebook Overview Association Rule Mining	EDA Charts Hierarchical Clustering Proposal Documentation Slides	EDA Organization Data Dictionary Data Cleaning Hierarchical Clustering K-Means Proposal Documentation Slides

APPENDIX:

- Dataset Link:
  https://data.cityofchicago.org/widgets/ijzp-q8t2?mobile_redirect=true
- Github Link:
  - Repository:
    https://github.com/Shayanhasan99/BA_820
  - Kanban Board:
    https://github.com/users/Shayanhasan99/projects/1

- EDA Visualization:


Distribution of Crimes by Type
Crime Type
THEFT
BATTERY
CRIMINAL DAMAGE
NARCOTICS
ASSAULT
OTHER OFFENSE
BURGLARY
MOTOR VEHICLE THEFT
DECEPTIVE PRACTICE
ROBBERY
CRIMINAL TRESPASS
WEAPONS VIOLATION
PROSTITUTION
OFFENSE INVOLVING CHILDREN
PUBLIC PEACE VIOLATION
SEX OFFENSE
CRIM SEXUAL ASSAULT
INTERFERENCE WITH PUBLIC OFFICER
LIQUOR LAW VIOLATION
GAMBLING
ARSON
HOMICIDE
CRIMINAL SEXUAL ASSAULT
KIDNAPPING
STALKING
INTIMIDATION
CONCEALED CARRY LICENSE VIOLATION
OBSCENITY
PUBLIC INDECENCY
NON-CRIMINAL
OTHER NARCOTIC VIOLATION
HUMAN TRAFFICKING
NON - CRIMINAL
RITUALISM
NON-CRIMINAL (SUBJECT SPECIFIED)
DOMESTIC VIOLENCE
Count
1e6


Number of Crimes Per Month
Number of Crimes
Year