ANALYSIS ON CHICAGO CRIMES

By: Shubham, Shayan, Ali, Sulaiman

Project Overview

Motivation

Harnessing data to optimize crime management in Chicago for proactive public safety measures

Problem

Chicago's diverse crime landscape requires advanced data analytics to uncover hidden patterns and relationships for more effective resource allocation and crime prevention strategies

Business Relevance

The analysis benefits law enforcement, businesses, insurance companies, and families by enabling predictive policing, informed decision-making, and safer living environments through data-driven insights

Dataset Description

- ~ 8 Million rows
- 26 columns
- 36 Unique Crime Types
- 5% Missing Data for Location attributes

Column Name	Data Type	Description
ID	int	Unique identifier for the record.
Case Number	string	The Chicago Police Department RD Number (Records Division Number).
Date	datetime	Date when the incident occurred. This is sometimes a best estimate.
Block	string	The partially redacted address where the incident occurred, placed in the middle of the street block.
IUCR	string	The Illinois Uniform Crime Reporting code. This is directly linked to the type of crime that occurred.
Primary Type	string	The primary description of the IUCR code.
Description	string	The secondary description of the IUCR code, a subcategory of the primary type.
Location Description	string	Description of the location where the incident occurred.
Arrest	bool	Indicates whether an arrest was made.
Domestic	bool	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
Beat	string	Indicates the beat where the incident occurred. A beat is the smallest police geographic area.
District	int	Indicates the police district where the incident occurred.
Ward	int	The ward (City Council district) where the incident occurred.
Community Area	string	Indicates the community area where the incident occurred.
FBI Code	string	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System.
X Coordinate	float	The X coordinate of the location where the incident occurred.
Y Coordinate	float	The Y coordinate of the location where the incident occurred.
Year	int	Year the incident occurred.
Updated On	datetime	Date and time the record was last updated.
Latitude	float	The latitude of the location where the incident occurred.
Longitude	float	The longitude of the location where the incident occurred.
Location	string	The location where the incident occurred in a format that allows for mapping and data analysis.

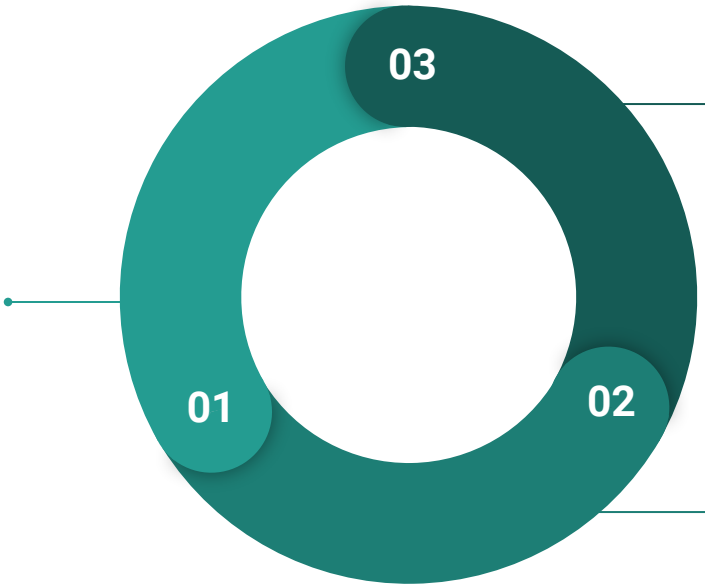
Data Processing

Methodologies used

Clustering

Hierarchical Clustering

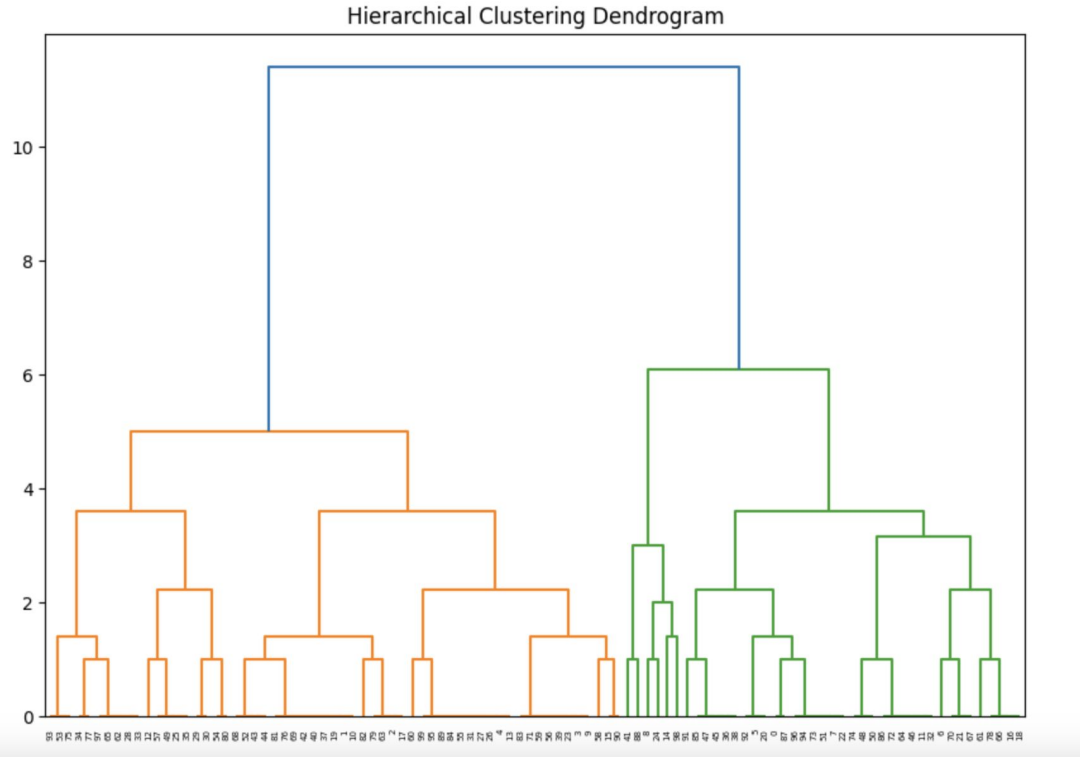
K-means Clustering



TF-IDF

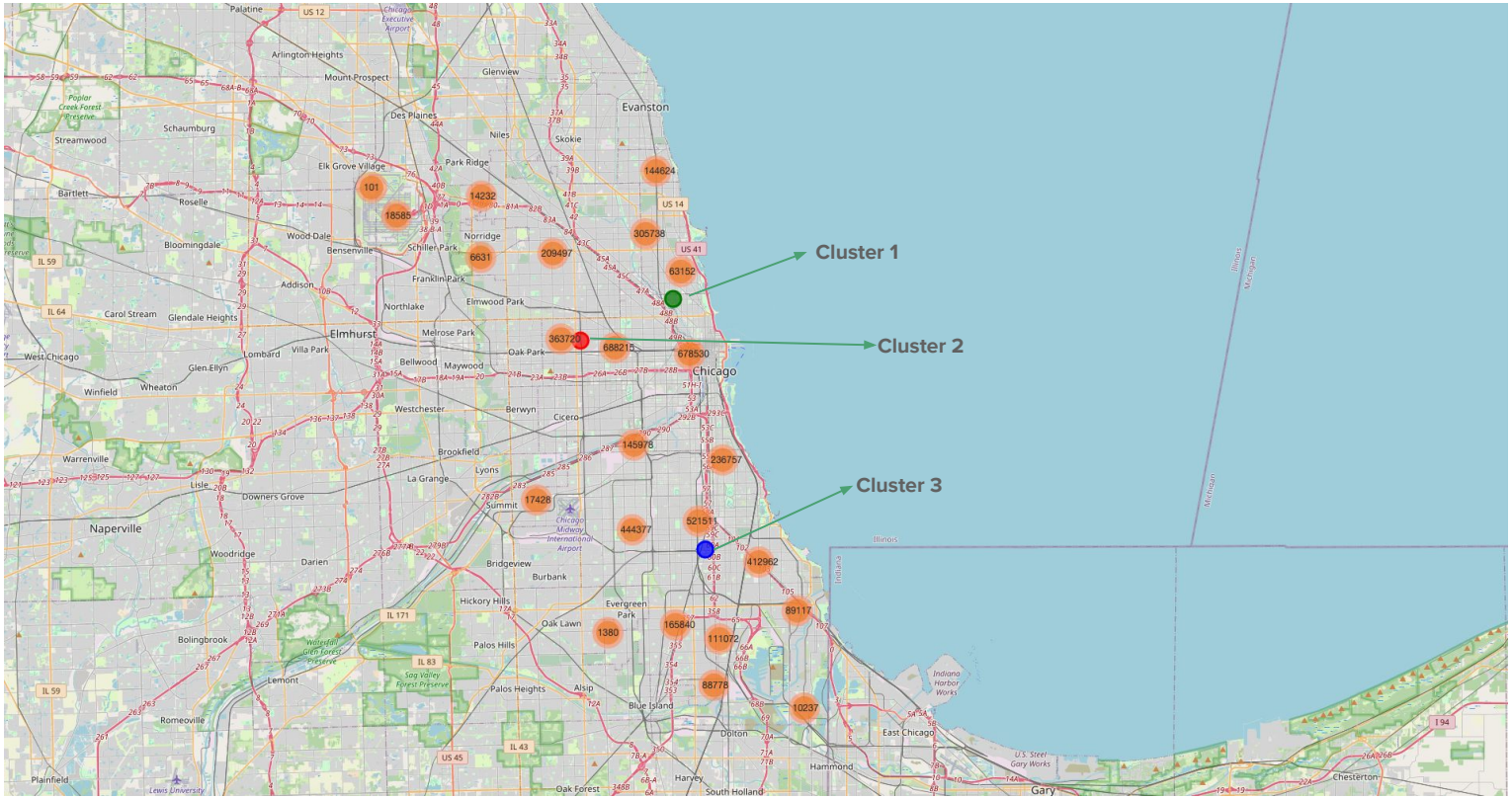
Association Rules

Clustering: (Hierarchical)



- Chosen for its ability to reveal hierarchical relationships in data.
- Employed **'Complete'** linkage method.
- Visualized results using a **dendrogram**.
- Revealed **Three distinct clusters**, offering insights into crime patterns in Chicago.

Clustering: (K-Means)

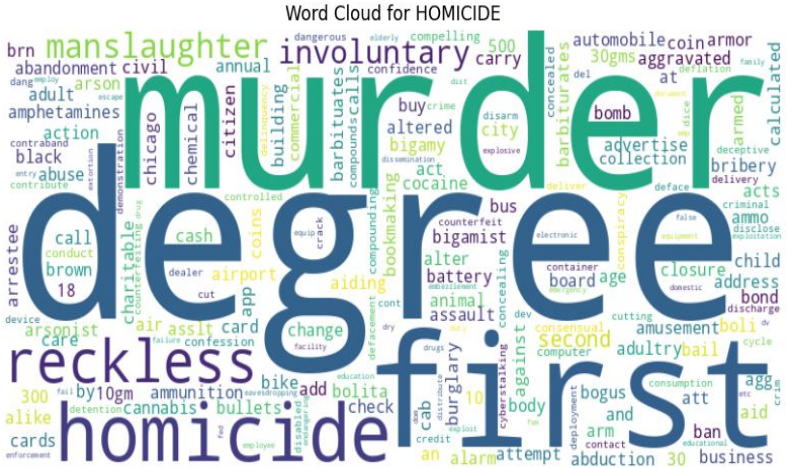


TF-IDF

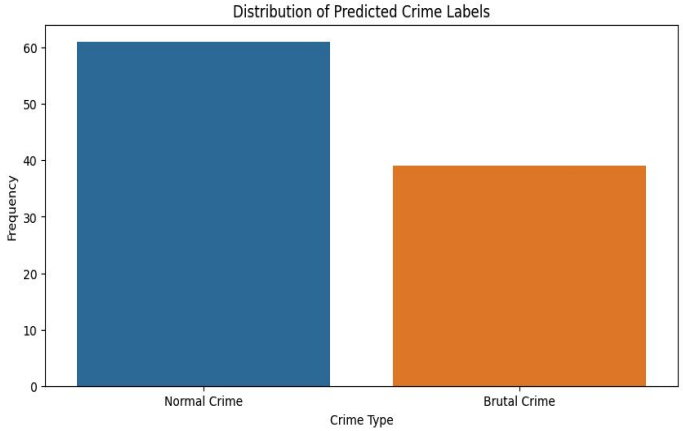
Extracting word frequencies

```
simple: 1554924
to: 1097361
500: 1065912
under: 674195
domestic: 672207
battery: 672113
and: 659602
poss: 611600
over: 513086
vehicle: 496834
property: 452097
theft: 426567
entry: 421989
aggravated: 421072
automobile: 358257
```

Word Clouds for 36 different crime types



Using TFIDF Vectorizer to Create Distribution of Predicted Crime Labels



Association Rule Mining

- Analysis of Crimes for the past two years

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
50466	(MOTOR VEHICLE THEFT, THEFT, CRIMINAL DAMAGE, ...	(BATTERY)	0.20	0.67	0.20	1.0	1.492537
107266	(BURGLARY, THEFT, OTHER OFFENSE, ASSAULT, DECE...	(CRIMINAL DAMAGE)	0.18	0.59	0.18	1.0	1.694915
47063	(BATTERY, BURGLARY, DECEPTIVE PRACTICE, CRIMIN...	(CRIMINAL DAMAGE, THEFT)	0.24	0.57	0.24	1.0	1.754386
107268	(BURGLARY, OTHER OFFENSE, CRIMINAL DAMAGE, ASS...	(THEFT)	0.18	0.75	0.18	1.0	1.333333
27058	(OTHER OFFENSE, DECEPTIVE PRACTICE, CRIMINAL T...	(THEFT)	0.23	0.75	0.23	1.0	1.333333
118021	(WEAPONS VIOLATION, BATTERY, OTHER OFFENSE, NA...	(CRIMINAL TRESPASS, CRIMINAL DAMAGE, THEFT)	0.15	0.36	0.15	1.0	2.777778

Challenges Faced

- Huge Dataset, had to create subsets to work with ML
- Map HTML File too big to open on chrome
- Association Rule Mining could not be done on large datasets.

Future Work

- Utilize Big Data Tools such as Hadoop and pySpark for efficient handling of large dataset.
- Expand sentiment analysis to extract negative sentiments.
- Enhance data quality by implementing geo-encoding techniques for accurate geographical information.
- Apply this analysis on another city.

Thank You!

Appendix

- Dataset Link: https://data.cityofchicago.org/widgets/ijzp-q8t2?mobile_redirect=true