***Analyzing Crime, Insights from the Windy City***

**Vasant Dave: 041154429**

**Hasibullah Yosufi: 041012318**

**Shelly Tyagi: 041180437**

**Md Irteza Chowdhury: 041126343**

**Algonquin College** of **Applied Arts and Technology ©2025**

**Start Date: 10th March,2024**

Contents

[1. Introduction 2](#_Toc53114217)

[2. Business Understanding 3](#_Toc531358132)

[2.1 Determine business objectives 3](#_Toc1146106274)

[2.2 Assess situation 4](#_Toc607620759)

[2.3 Determine goals 5](#_Toc1680846800)

[2.4 Produce project plan 5](#_Toc646554150)

[3. Data Understanding 7](#_Toc254219208)

[3.1. Collect initial data 7](#_Toc2082682527)

[3.2. Describe data 8](#_Toc1507185210)

[3.3. Explore data 13](#_Toc272244546)

[3.4. Verify data quality 23](#_Toc186577803)

[4. Classification by kNN 25](#_Toc840943773)

[4.1. Data Preparation 25](#_Toc54906156)

[4.1.1. Select data 26](#_Toc1728345221)

[4.1.2. Clean data 26](#_Toc454276200)

[4.1.3. Construct data 26](#_Toc1966452567)

[4.1.4. Integrate data: 29](#_Toc1502509407)

[4.1.5. Format data 29](#_Toc2119730695)

[4.2. Modelling 30](#_Toc1707369709)

[4.2.1. Select modeling techniques 30](#_Toc829288719)

[4.2.2. Generate test design 30](#_Toc1470942313)

[4.2.3. Build model 30](#_Toc216779867)

[4.2.4. Assess model 30](#_Toc111834323)

[4.3. Evaluation 30](#_Toc510774115)

[4.3.1. Evaluate results 30](#_Toc70308662)

[4.3.2. Interpret resultsy 30](#_Toc1533725583)

[4.3.3. Review of process 30](#_Toc2140350589)

[4.3.4. Determine next steps 30](#_Toc565035103)

[5. Classification by Decision Tree 30](#_Toc691433713)

[5.1. Data Preparation 30](#_Toc1649431160)

[5.1.1. Select data: 30](#_Toc117167322)

[5.1.2. Clean data 31](#_Toc2048670572)

[5.1.3. Construct data 31](#_Toc2041825172)

[5.1.4. Integrate data 35](#_Toc1190959069)

[5.1.5. Format data 35](#_Toc2062815099)

[5.2. Modelling 35](#_Toc272876477)

[5.2.1. Select modeling techniques 35](#_Toc1116848088)

[5.2.2. Generate test design 35](#_Toc1408470956)

[5.2.3. Build model 35](#_Toc1933142803)

[5.2.4. Assess model 35](#_Toc1683022839)

[5.3. Evaluation 35](#_Toc695085855)

[5.3.1. Evaluate results 35](#_Toc1325851318)

[5.3.2. Interpret results 35](#_Toc1948785171)

[5.3.3. Review of process 36](#_Toc563494564)

[5.3.4. Determine next steps 36](#_Toc993601221)

[6. Clustering by kMeans (clustering and finding outliers) 36](#_Toc703669370)

[6.1. Data Preparation 36](#_Toc261535139)

[6.1.1. Select data 36](#_Toc1311104558)

[6.1.2. Clean data 38](#_Toc1058983387)

[6.1.3. Construct data 38](#_Toc905376899)

[6.1.4. Integrate data 38](#_Toc800822471)

[6.1.5. Format data 38](#_Toc514389065)

[6.2. Modelling 39](#_Toc400232048)

[6.2.1. Select modeling techniques 39](#_Toc1048537319)

[6.2.2. Generate test design 39](#_Toc1088661187)

[6.2.3. Build model 39](#_Toc1647349796)

[6.2.4. Assess model 39](#_Toc798956358)

[6.3. Evaluation 40](#_Toc1246547312)

[6.3.1. Evaluate results 40](#_Toc125347076)

[6.3.2. Interpret results 40](#_Toc408300387)

[6.3.3. Review of process 40](#_Toc1330686332)

[6.3.4. Determine next steps 40](#_Toc1397133167)

[7. Outlier Detection by LOF and ISF (and common outliers) 40](#_Toc774546196)

[7.1. Data Preparation 40](#_Toc105013231)

[7.1.1. Select data 40](#_Toc1354471062)

[7.1.2. Clean data 41](#_Toc1328891717)

[7.1.3. Construct data 41](#_Toc324241874)

[7.1.4. Integrate data 43](#_Toc632189739)

[7.1.5. Format data 43](#_Toc594006027)

[7.2. Modelling 44](#_Toc1316533372)

[7.2.1. Select modeling techniques 44](#_Toc1108323120)

[7.2.2. Generate test design 44](#_Toc151497002)

[7.2.3. Build model 44](#_Toc2089109881)

[7.2.4. Assess model 44](#_Toc1537916079)

[7.3. Evaluation 44](#_Toc62182786)

[7.3.1. Evaluate results 44](#_Toc767749676)

[7.3.2. Interpret results 44](#_Toc1451010562)

[7.3.3. Review of process 44](#_Toc157274926)

[7.3.4. Determine next steps 44](#_Toc114676462)

[8. Conclusion 44](#_Toc1465734716)

# Introduction

This report analysis was undertaken as a part of academic project. The report is formatted in CRISP-DM format. The CRoss Industry Standard Process for Data Mining (CRISP-DM) is a process model with six phases that describes the data science life cycle. Six phases in chronological order are

* Business Understanding
* Data Understanding
* Data preparation
* Modelling
* Evaluation
* Deployment

The CRISP-DM framework is selected for its organized yet adaptable methodology, rendering it suitable for examining extensive crime data. Its iterative nature guarantees a good comprehension of crime patterns, facilitates data-informed insights and better accuracy for prediction models, while adjusting to new discoveries during the analytical process. Furthermore, we use RapidMiner as our tool for this analysis and machine learning. RapidMiner is a data science platform used for data preparation, machine learning, text mining, and predictive analytics, offering a user-friendly interface and a wide range of features that can benefit both technical and non-technical users, streamlining data analysis and model development.

The Dataset utilized is [City of Chicago's crime dataset](https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data), spanning from 2001 to the present, size of approximately 8 million rows, offers a comprehensive view of criminal activity within the windy city (metaphorical name for Chicago because of its unique meteorological conditions and urban structure of the boastful, and long-winded nature of its politicians and boosters, who were known for promoting the city aggressively). This dataset, available on the City of Chicago's Data Portal, provides detailed information on various types of crimes, including their locations, dates, whether an arrest was made and other relevant details. The dataset along with approximate [dashboard visualizations](https://data.cityofchicago.org/stories/s/Crimes-2001-to-present-Dashboard/5cd6-ry5g) is a valuable resource for understanding crime patterns, trends, and distributions across different neighborhoods and time periods. The primary goal of this report is to analyze and gain insights into crime and various factors associated with crime, in Chicago. The analysis will potentially try to answer questions related to crime types, location, arrest, with their associated correlated factors. We will employ different classification machine learning models to attempt to answer our framed questions or business objectives that align with our primary goal.

**Disclaimer:** These crimes may be based upon preliminary information supplied to the Police Department by the reporting parties that have not been verified. The preliminary crime classifications may be changed later based upon additional investigation and there is always the possibility of mechanical or human error. Therefore, the Chicago Police Department does not guarantee (either expressed or implied) the accuracy, completeness, timeliness, or correct sequencing of the information and the information should not be used for comparison purposes over time. The Chicago Police Department will not be responsible for any error or omission, or for the use of, or the results obtained from the use of this information. All data visualizations on maps should be considered approximate and attempts to derive specific addresses are strictly prohibited.

# Business Understanding

## 2.1 Determine business objectives

The dataset comes from the Chicago Police Department and contains records of arrests in Chicago with more than 1.8 million records of individuals aged 18 and above. In this report, we will analyze the dataset to uncover valuable insights into the patterns, trends, and factors and the demographic contributing to crime in the city of Chicago. Upon the initial assessment of the data and a thorough investigation of the metadata, our team has decided on one key research question that provide insights into the crime dataset, and we utilized two different classification models for the objective. The Questions can be answered by applying various machine learning techniques like KNN, Decision Tree, Clustering and Outlier Detection. Respectively, our key research question is

***What factors contribute to the likelihood of an arrest in each crime case?***

Classification Models selected: -

1. **KNN Analysis**
2. **Decision Tree and Random Forest**

Outlier detection models selected: -

1. **Clustering Analysis:** How do different types of crimes cluster across Chicago Dataset?
2. **Outlier Detection:** Are there any unusual crime patterns or anomalies in the factors influencing arrests (KNN) and crime types (Decision tree)?

## 2.2 Assess situation

1. **Source and Data Quality**

* The dataset is obtained from the Chicago Police Department (CPD) which covers arrests from 2001 up to 2025.
* The dataset is in CSV (Comma Separated Value) format, with a total size of 1.8 GB.
* The dataset incorporates over 8 million records of individuals aged 18 and above.
* The dataset has a total of 22 features.

1. **Contingency and Challenges**

* The dataset comprises more than 8 million rows, which may pose several computational challenges, including memory usage, processing speed, and model training methods such as K-NN and clustering, both of which require extensive numerical computation.
* The Dataset can have potential bias towards certain demographics, races, or even dates, for example, pre- pre-covid-19 and post-covid-19.

1. **Tools and Technologies**

* For this analysis, we will use Altair-RapidMiner for all modeling and analysis.
* All preprocessing tasks, such as filling in missing values, sampling, and formatting data, will be performed in RapidMiner.
* We will utilize a correlation matrix to identify patterns and relationships between features.

## 2.3 Determine goals

The goal of this analysis is to uncover hidden patterns and key factors contributing to different types of crime in the city of Chicago. By identifying and understanding these patterns and trends, this study aims to help the City of Chicago, and the Chicago Police Department (CPD) develop more effective strategies and policies to ensure a safer and more prosperous environment for its residents.

We utilize two main techniques to achieve feature selection, Association rule mining and Correlation matrix analysis. Both methods are covered in detail in the Explore data step in the Data understanding section of the report.

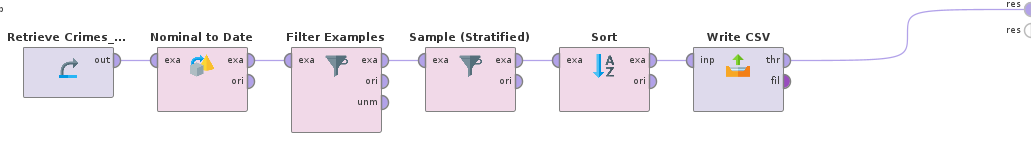
## 2.4 Produce project plan

|  |  |  |
| --- | --- | --- |
| **Steps** | **Sub-task** | **Name of the student** |
| Introduction |  | Vasant |
| Business Understanding |  | Hasib |
| Data Understanding  Each group must collectively understand all attributes. Individually, each member is responsible for understanding at least one-fourth of the attributes, assuming the group consists of four members. Once assigned, document the attribute names along with the names of the members who have understood them. | 1)Vasant: - Block, Beat, Community Area, Ward, District, X and Y coordinate  2)Hasib: - ID, Latitude, Longitude, Location, Updated\_ On, Year, IUCR  3)Irteza: - Primary Type, FBI Code, Date, Location Description  4)Shelly: - Arrest, Crime description, Case number, Domestic | Vasant |
| Task 1 – KNN |  | Shelly |
| Task 2 – Decision Tree & Random Forest |  | Irteza |
| Task 3 – Clustering & identify outliers from clusters |  | Vasant |
| Task 4 - Outlier Detection  LOF & ISF |  | Hasib |
| Compare Model results |  | Shelly |
| Conclusion |  | Irteza |

# Data Understanding

# Collect initial data

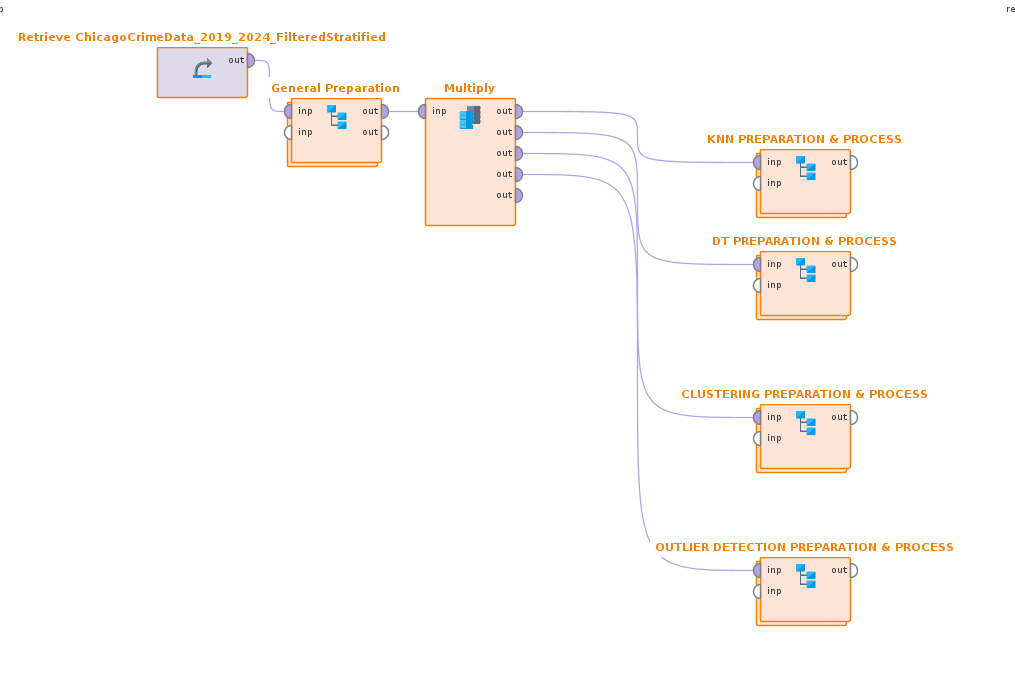
Data was collected from the [Chicago data portal](https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data_preview) which provides data from 2001 to 2025 (present minus the most recent seven days). It was fetched through the API endpoint available, and downloaded as a CSV Format. The Dataset contained 8 million rows of data. For our research and project purposes, we compressed the data down from 8 million rows to 10,000 instances (rows) and a specific time range that we saw appropriately captures some of the temporal trends in the dataset.



First, we loaded the 8 million rows dataset into the rapidminer. Upon loading the “Date” column got auto detected as type “nominal”. Therefore, we had to parse it into “Datetime” for us to filter them based on a year range. Our selected year range is 2019-2024. Some of the reasons for us selecting this time-period is due to its representation of specific events in human history, like the Onset and Decline of COVID-19(2020-2024). Lockdowns leading to decreased street crime but potential increases in domestic violence. Economic downturns, inflation job losses, and social distress (2022-2024) contributing to different types of crimes. Social unrest and protest (after the killing of George Floyd in 2020). Gun violence surge and Changes in law enforcement policies (2020-2022).

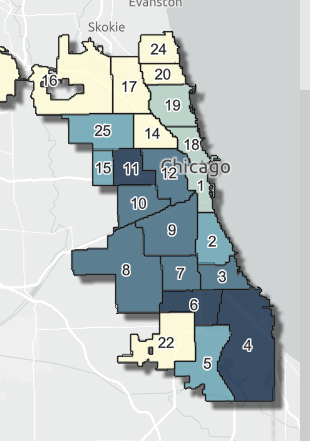
Next, we used Stratified sampling method with absolute value to get 10,000 instances from the original 8 million. Stratified sampling ensures that each subgroup (stratum) in a dataset is proportionally represented, making it ideal for analyzing crime data where categories like crime type, location, and time periods vary significantly. Formula for Stratified sampling

where nh is the sample size for stratum h, Nh is the population size of stratum h, N is the total population size, and n is the total sample size. We just sort in ascending order per date for initial interpretation and then export as csv file which we then finally use in our main RapidMiner process. Below is an image of our main Process where implementation of classification models and further steps for CRISP-DM are carried out.



# Describe data

Chicago crime data is spread across the 22 main police districts ([Reference](https://www.chicagopolice.org/police-districts/))

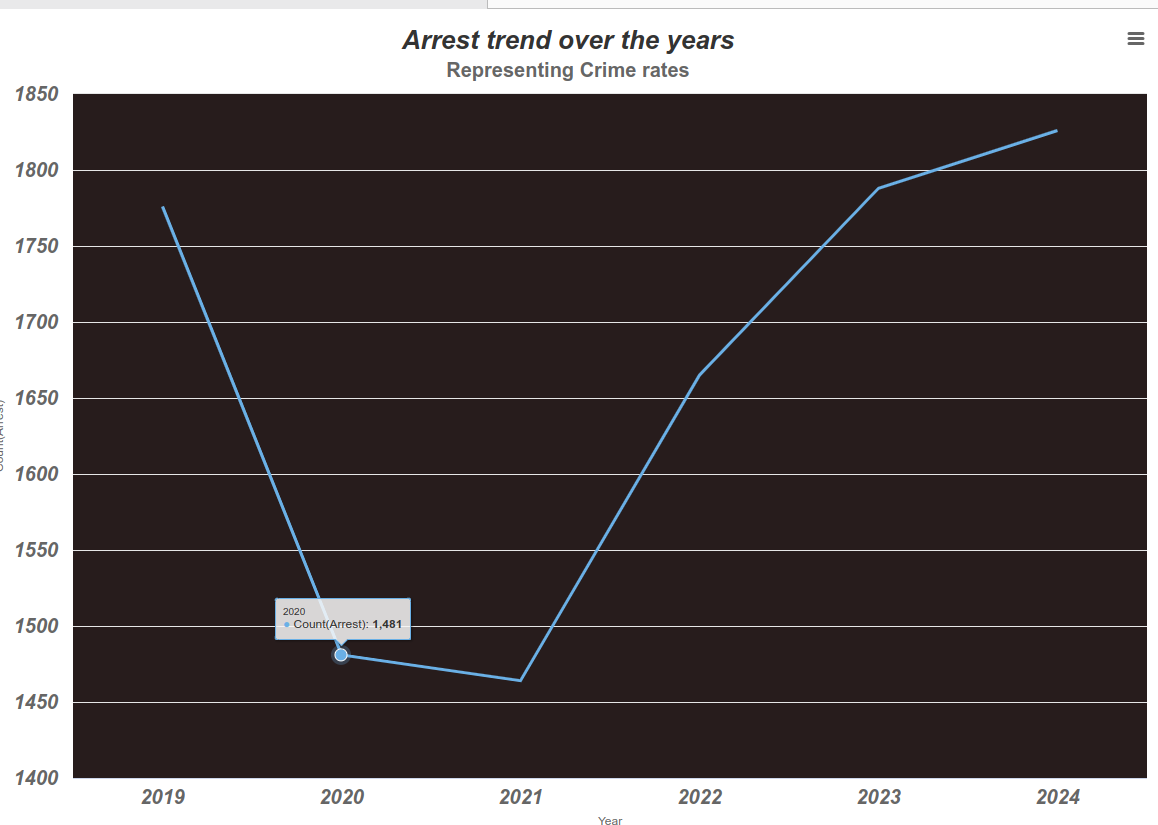


There are exactly in the original dataset 8,275,244 rows and 22 columns in the dataset. Each row represents a **reported crime.** Time period is 2001 to present (minus seven days). Data includes a mix of categorical, numerical, and timestamp fields that describe crime incidents, their locations, police involvement, and classification codes. Our New Dataset is like the original dataset below in all cases except we have compressed rows to 10000 with previously mentioned techniques.

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Description** | **API Field Name** | **Data Type** |
| **ID** | Unique identifier for the record. | id | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| **Case Number** | Chicago Police Department RD Number, unique to the incident. | case\_number | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| **Date** | Date and time when the incident occurred (sometimes an estimate). | date | [Floating Timestamp](https://dev.socrata.com/docs/datatypes/floating_timestamp.html) |
| **Block** | Partially redacted address where the incident occurred. | block | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| **IUCR** | Illinois Uniform Crime Reporting code, linked to the Primary Type. | iucr | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| **Primary Type** | The primary description of the IUCR code. | primary\_type | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| **Description** | Secondary description (subcategory) of the IUCR code. | description | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| **Location Description** | Description of the location where the incident occurred. | location\_description | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| **Arrest** | Indicates whether an arrest was made. | arrest | [Checkbox](https://dev.socrata.com/docs/datatypes/checkbox.html) |
| **Domestic** | Indicates if the incident was domestic-related. | domestic | [Checkbox](https://dev.socrata.com/docs/datatypes/checkbox.html) |
| **Beat** | Smallest police geographic area; each beat has a dedicated police car. | beat | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| **District** | Indicates the police district where the incident occurred. | district | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| **Ward** | The ward (City Council district) where the incident occurred. | ward | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| **Community Area** | Indicates the community area where the incident occurred (77 total). | community\_area | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| **FBI Code** | Crime classification based on the FBI’s National Incident-Based Reporting System. | fbi\_code | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| **X Coordinate** | The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block. | x\_coordinate | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| **Y Coordinate** | The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block. | y\_coordinate | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| **Year** | Year the incident occurred. | year | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| **Updated On** | Date and time the record was last updated. | updated\_on | [Floating Timestamp](https://dev.socrata.com/docs/datatypes/floating_timestamp.html) |
| **Latitude** | The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block. | latitude | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| **Longitude** | The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block. | longitude | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| **Location** | The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block. | location | [Location](https://dev.socrata.com/docs/datatypes/location.html) |

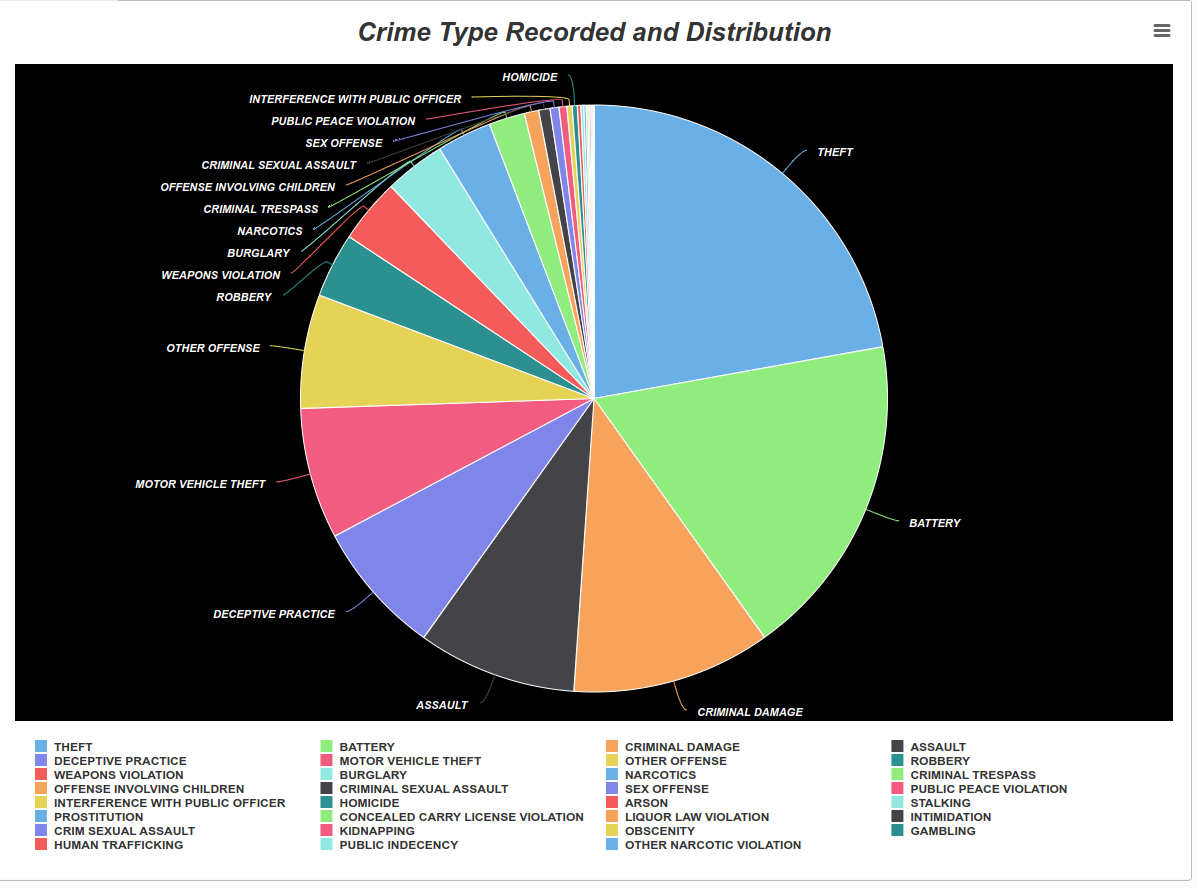
# Explore data

Our primary dataset for analysis will be the stratified and filtered subset, consisting of 10,000 instances, retaining all original columns from the full dataset. With our initial assumptions from the news and meta-data about the data, we started to visualize the dataset. We tried to see the temporal trend over the years in the rate of crimes that lead to an arrest (shown by line graph below). Our assumption hypothesis was proven right as we can see, it reveals a dramatic decrease in arrests from 2019 to 2021 (pre-covid and onset of covid), followed by a substantial surge peaking in 2023. In the context of Chicago, these fluctuations could reflect various factors, including but not limited to aforementioned factors like changes in policing strategies, shifts in crime patterns, or even socio-economic influences. It's important to note that arrests do not perfectly correlate with crime rates, but they offer a potential window into law enforcement activity within the city.

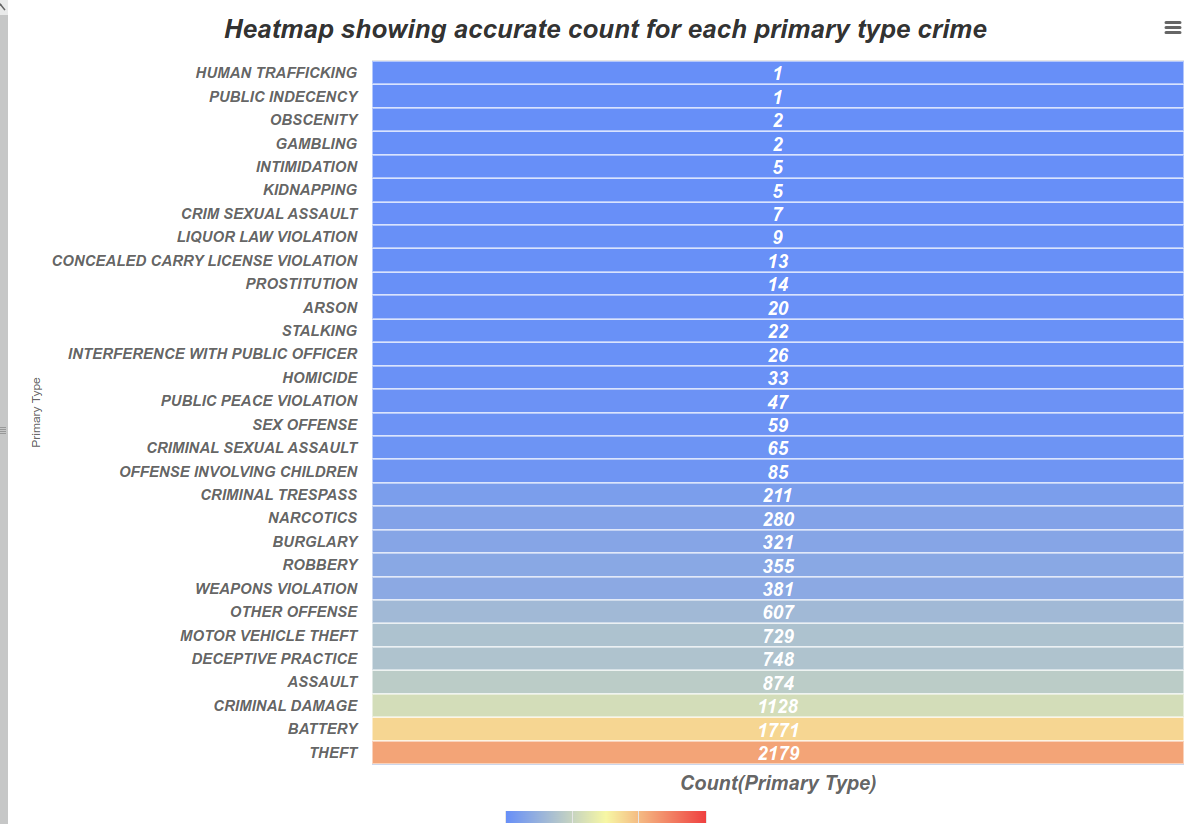


|  |
| --- |
|  |
|  |

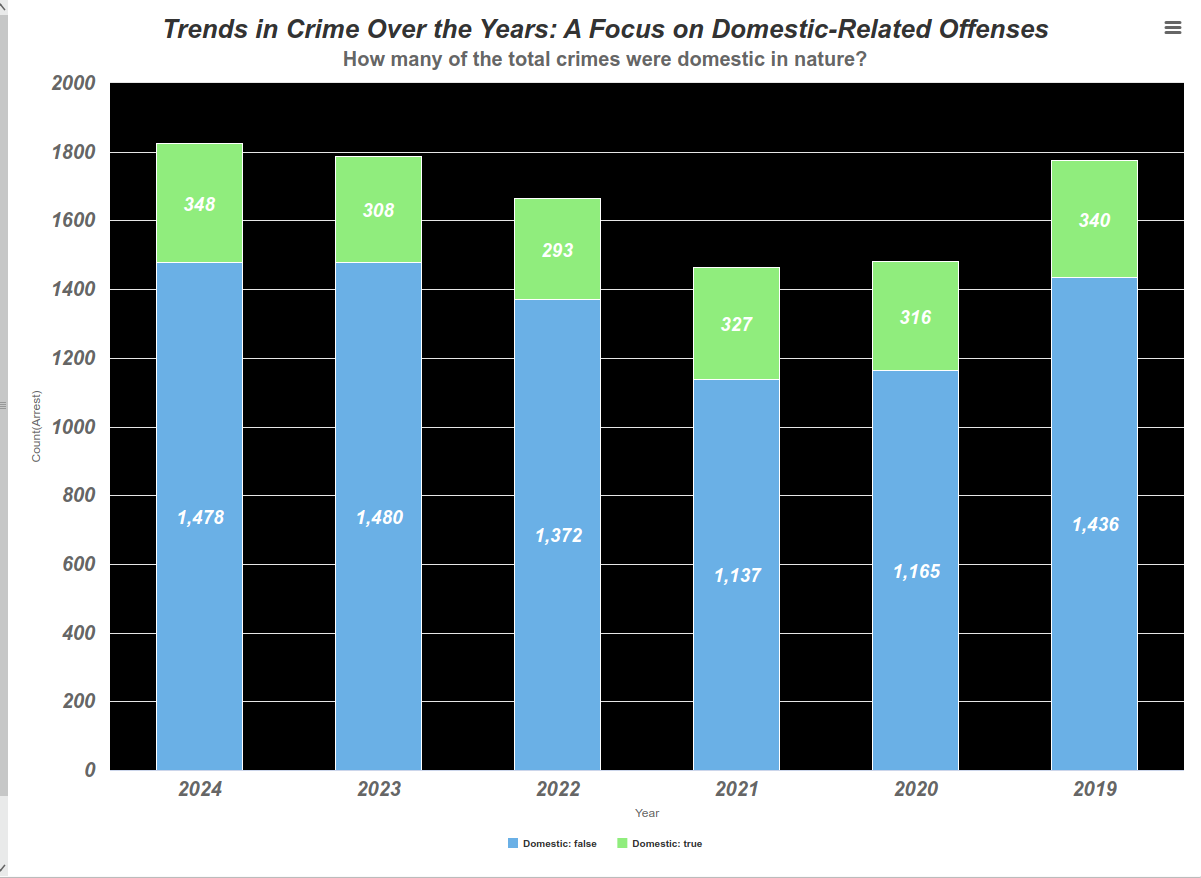
This side-by-side maps of Chicago (upper image from [Live Mapping of Chicago crime data for reference](https://data.cityofchicago.org/Public-Safety/Crimes-Map/dfnk-7re6) for territorial visualization and below visualized from RapidMiner based on longitude and latitude crimes) visually represents the geographical distribution of reported crimes, differentiated by whether an arrest was made. The concentration of both arrest (green) and non-arrest (blue) points is heavily focused within the city limits, particularly along the lakefront, indicating that crime is primarily an urban phenomenon in this area. The significant presence of blue points, representing crimes where no arrest was made, suggests a notable disparity between reported crimes and successful arrests, potentially indicating challenges in law enforcement or variations in crime types. Overall, there is a contrast between the dense cluster of crime data-points within Chicago and the relatively sparse distribution in the surrounding suburban regions.



To understand different types of crime recorded in the dataset (31 categories) and their frequency, we used pie chart. The size of each slice proportionally represents the frequency of each crime type, allowing for easy visual comparison. It immediately highlights that "THEFT" and "BATTERY" constitute the largest portions of reported crimes, suggesting these are the most prevalent offenses. The chart also showcases a wide variety of other crime categories, indicating a diverse range of criminal activity. Notably, less frequent crimes such as "HOMICIDE" and "HUMAN TRAFFICKING" are represented by very small segments, emphasizing their relative rarity compared to the dominant crime types. Overall, the chart offers a valuable categorical snapshot of the composition of recorded crimes.



A heatmap to show the count for each crime type in a sorted form. Provides similar information to the pie chart but with point precision numbers. Gradient color to show the binning of high value counts vs low and medium value counts.

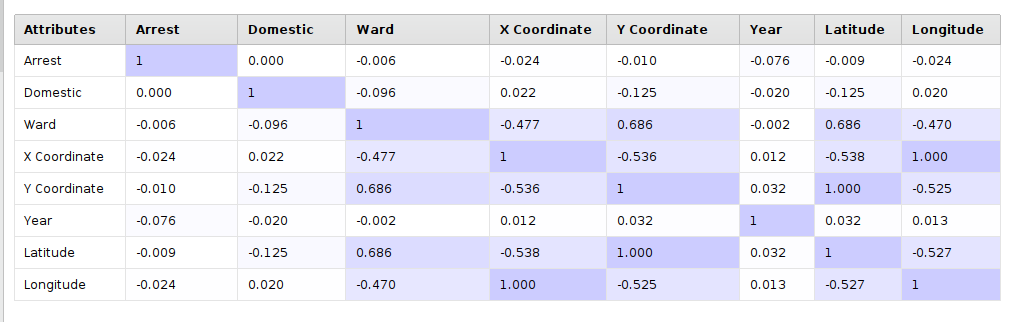
This stacked bar chart illustrates crime trends in Chicago from 2019 to 2024, with a specific focus on domestic-related offenses. Domestic crimes involve violence or abuse by someone with a close personal or familial relationship to the victim. These include spousal or intimate partner violence, child abuse (physical, sexual, or emotional abuse or neglect by a parent or caregiver), elder abuse (abuse or neglect by a family member or caregiver), and other family violence (violence between family members like siblings or parents and adult children).

The chart clearly highlights those non-domestic crimes, represented by the larger blue segments, dominate the city's crime landscape. In contrast, domestic crimes, shown in green, remain relatively stable over the years, but at a significantly lower volume. Notably, non-domestic crimes experienced a noticeable dip in 2020 and 2021, while domestic offenses held steady throughout the period. Overall, it suggests that while domestic crimes contribute to the city's crime rate, they are much less frequent than non-domestic offenses.

Next, as mentioned before in the “determine data mining goals” step, we use two techniques to identify feature relevance for our classification models. These methods help to uncover hidden patterns and

dependencies within the dataset, aiding in better feature selection and model performance. **Association Rule Mining** is a data mining technique that identifies relationships between variables (**categorical)** in large datasets by analyzing co-occurrence patterns. It uses metrics like support, confidence, and lift to evaluate the strength of these associations. **Correlation Matrix Analysis** is a statistical method used to measure the linear relationships between multiple continuous variables (**numerical)**, represented in a matrix format. It is based on the Pearson correlation coefficient:

Where **n** is the number of data points (i.e., **x, y** pairs), **∑XY** is the sum of the products of the **x** and **y** values, **∑X** and **∑Y** are the sums of the **x** and **y** values, respectively, and **∑X²** and **∑Y²** are the sums of the squares of the **x** and **y** values, respectively.

After Correlation matrix analysis in RapidMiner, it shows the correlation coefficients between pairs of attributes, with values ranging from -1 to 1. Higher positive correlation is represented by values closer to ”1” and stronger negative correlation is represented by values closer to -1 Notably, "Latitude" and "Ward" exhibit a strong positive correlation (0.686), suggesting that as the ward number increases, the latitude tends to increase as well, although this would not be relevant for us going further, as ward provides area code for city council district and does not provide us insights into On the other hand, "Latitude" and "Longitude" show a strong negative correlation, which is expected due to the nature of geographic coordinates. The correlation between "Domestic" and "Latitude" is moderately positive, hinting at a slight tendency for domestic crimes to be associated with higher latitude values. X and Y coordinate columns provide much more accurate correlations than Latitude and Longitude due to X and Y coordinates, when projected into a **Cartesian coordinate system** (such as UTM), provide **linear distances** that are consistent across the entire dataset, which is crucial for accurate correlation analysis. The matrix also reveals weak correlations between "Arrest" and other attributes, suggesting that arrests are not strongly associated with any single factor in this dataset. However, it is important to note that correlations don't imply causation. This doesn't mean that domestic crimes have no influence on arrests. It simply means that a simple linear relationship isn't evident in this data. It's possible that arrests for domestic crimes are influenced by a complex combination of factors beyond just whether a crime is domestic or not. Similarly, the weak correlation with "Year" doesn't mean time has no impact on arrests. The near-zero correlations with "Ward," "Latitude," and "Longitude" don't mean location is irrelevant to arrests. It could be that location plus some other factors together influences factors like police response times, which in turn affect arrests, but these indirect relationships aren't captured by simple linear correlations. Therefore, while the "Arrest" row reveals potential associations, it doesn't explain the causal mechanisms behind the outcomes.

|  |
| --- |
|  |
|  |

Before proceeding onto association rule mining, some preprocessing steps must be carried out. First the Categorical values are converted to binominal as the FP-Growth algorithm only accepts binominal values for rule mining (Nominal to binominal operator transforms nominal values (e.g., "short", "medium", "long") into **binary** attributes. For instance, each nominal category might become a separate binary attribute, where 1 indicates presence and 0 indicates absence.)

FP-Growth is an algorithm used to mine frequent itemset from the data. An itemset is a collection of items (in this case, binominal attributes), and a frequent itemset is a combination of items that appear together in the dataset more often than a specified minimum support threshold. Lastly create association rules based on confidence is applied. After Association rule mining in RapidMiner, table presents association rules, which are used to discover interesting relationships between categorical variables in large datasets.

Upon observing the table, a significant trend is the high confidence associated with "Arrest\_false" as a conclusion across multiple rules, indicating that a substantial portion of reported crimes do not result in arrests. Rules 9 through 10 demonstrate a strong correlation between "Primary Type\_BATTERY" and "FBI Code\_08B", suggesting a consistent association between battery-related offenses and specific FBI coding. Furthermore, rules reveal that both domestic and non-domestic incidents are linked to "Arrest\_false," with non-domestic crimes having a much higher support, implying their greater frequency. Some Rules highlight the interconnectedness of "Primary Type\_THEFT," "FBI Code\_06," "Arrest\_false," and "Domestic\_false," suggesting that theft-related incidents are often non-domestic and do not lead to arrests. Rules also indicate that crimes described as "SIMPLE" or occurring in "APARTMENTS" are also highly associated with "Arrest\_false." Overall, 50 plus rules are generated for interpretation in aspect to feature relevance.

Using both these methods, we select relevant columns that aid in our classification model’s feature selections and model performance.

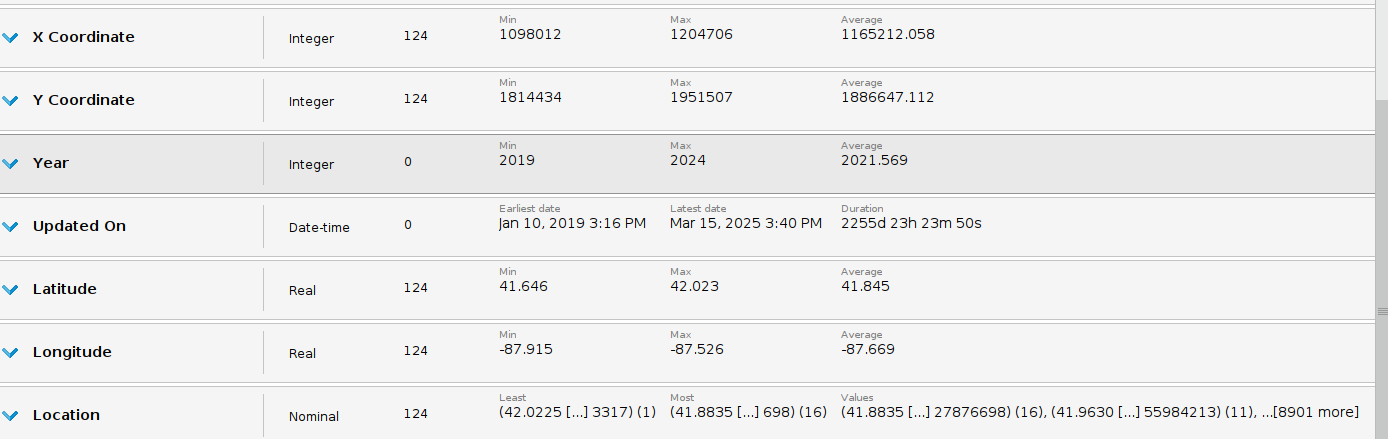
# Verify data quality

During the Initial loading of the data, it is noticed that some columns are interpreted as a different datatype by RapidMiner autodetection. Cross-checking with the metadata on the source portal. We found that Arrest, Domestic, Beat, district etc. Columns were coming as wrong types

|  |
| --- |
|  |
|  |

Therefore, we convert arrest domestic to binominal (even if we do not do this step, RapidMiner automatically interprets binominal as binary classification 0/1). Then we convert beat, district and community area to nominal datatypes to match the metadata declarations. We also set id column role as id for indexing or joining in later steps. We have very few missing values across the dataset





The techniques to impute these missing values are handled and documented in each classification subprocess Preparation steps.

# Classification by kNN

In this section of the report, we are trying to answer the question “what factors contribute to the likelihood of an arrest in a crime case?” using k-NN algorithm. K-Nearest Neighbors(k-NN) is a supervised machine learning method used for classification tasks. It works by identifying the k most similar instances to a new instance based on selected features. The majority class among these neighbors determines the predicted outcome for the new case. In this project k-NN will be used to predict the likelihood of an arrest in each crime case based on historical arrest data from the Chicago arrest dataset. By using features such as crime type, location, time and other relevant attributes the k-NN model will help identify patterns and key factors that are common among cases where arrests were made. This will allow us to analyze which factors are most strongly associated with arrests.

# Data Preparation

The process starts by bringing in the Chicago Arrest dataset. The Goal is to ensure that the dataset is free of inconsistencies, missing data and irrelevant attributes before building the kNN model. We have already stratified the data in our general prep so we can get started with the data.

# Select data

The first step is to select only the relevant features for modeling from the dataset. We filtered some common irrelevant columns in general prep and then each modelling process is selecting the attributes that are relevant for their model. The attributes we are keeping for k-NN are Primary Type, Date, ID, Location Description, Arrest (class column), Domestic, x Coordinate, y Coordinate. These attributes are most likely to influence the outcome of an instance based on the research we did in data understanding.

# Clean data

After selecting all relevant attributes, the next step is to clean the data. I used remove duplicates and replace missing values operator for this. In my data preparation I found that there are 3 attributes Location description, x coordinate, y coordinate with missing values. I decided to adopt different approaches to replace missing values for these attributes which I mentioned below.

**Remove duplicates** eliminates any duplicate rows that may bias the distance calculation in kNN. I used remove duplicates operator to filter any duplicates but found out there are no duplicates in this dataset.

**Replace missing values**: Handles any missing or null values, ensuring that the dataset is complete and ready for further processing. For location descriptions I replaced missing values with the value unknown so that I can create a separate category for it later in my construct data steps. For x and y coordinate I replaced them with the mean.

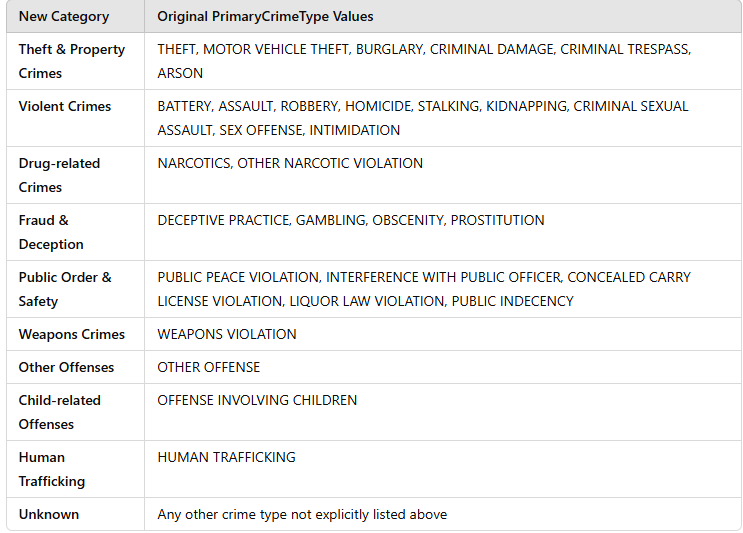
# Construct data

For this purpose I used Rename, Set Role and Generate attributes and date to numerical operators.

**Rename**: Renames attributes to more descriptive or meaningful names making downstream steps more manageable. I renamed prime type to PrimaryCrimeType, Description to CrimeDescription and location description to LocationDescription. Removing spaces from the attribute names makes it easier to use them later in generate attributes queries.

**Set Role:** Defines roles like setting target variables and unique identifiers. I set arrest as label for my dataset.

**DateProcessing(subprocess):** I created a subprocess called Date processing to extract useful information from date column. I used **date to numerical** operator to extract Hour, month and year from the date and renamed them appropriately. Since Year has a datatype of real after being extracted from the date column I changed it to nominal using **numerical to polynominal** operator. I kept hour and month numerical to normalize them later. In the end I used **select attribute** to move forward with newly created column and discard Date column since I don’t need it anymore.

**Generate attributes:** I used generate attributes to create new engineered attributes such as I grouped Primary crime type( 30 categories) attributes into 10 major crime type categories as below: 

And I grouped location description into location categories as below.



# Integrate data:

After all the cleaning and processing is done for the attributes, it is now time to gather the final data together to move ahead with. I am using select attribute operator for this purpose.

**Select Attributes**: After all the transformations I am only keeping the features relevant for classification. I am dropping all the previous columns for Primary Crime Type, Location description and crime description and only keeping the newly created categories columns.

# Format data

K-NN is a distance-based method, so we need all our attributes to be numerical. But our final attributes are either numerical or nominal data types. I will be using nominal to numerical and normalization operators to prepare the data for modelling.

**Normalize**: Scales all numeric features to the same range to ensure fair distance calculation. This operator will normalize all numeric attributes (Hour, Month, x coordinate, y coordinate).

**Nominal to numerical:** Converts categorical(nominal) attributes into numerical form using one-hot encoding. This operator will convert all the nominal attributes (CrimeCategory, Location Category, Domestic,Year) into 0 and 1 form.

# Modelling

# Select modeling techniques

In this part of project, the k-Nearest Neighbors (k-NN) classification algorithm was chosen to model and predict whether a crime case leads to an arrest. k-NN is a non-parametric algorithm that classifies data points based on the majority label of the k most similar instances, using a distance metric. This is particularly effective when patterns in the data are based on proximity or similarity.

In RapidMiner, the k-NN operator was used with the following settings:

**k: 7**

**Weighted Vote:** Enabled (closer neighbors have a higher influence)

**Distance Measure:** Mixed Euclidean Distance (combines both nominal and numerical distances)

**Nominal attributes:** Nominal Distance

**Numerical attributes**: Euclidean Distance

These settings ensure that both types of variables—such as CrimeCategory (nominal) and xCoordinate (numerical)—are incorporated appropriately in the model's similarity calculations.

# Generate test design

To train and evaluate the model effectively, a **Split Data** operator was used with the following configuration:

* **Split ratio**: 70% training, 30% testing
* **Stratified sampling**: Yes (based on the target variable Arrest)

This ensures that both classes (arrest and no arrest) are proportionally represented in both training and test sets. Cross-validation was not used in this process; instead, model performance was evaluated directly on the hold-out test set.

# Build model

After preprocessing, the dataset was normalized and all nominal variables were converted using one-hot encoding (Nominal to Numerical). The k-NN operator was applied to the training data using the settings above.

The model predicts the Arrest class label based on input features such as:

* **CrimeCategory**
* **LocationCategory**
* **Domestic**
* **Hour**
* **Month**
* **xCoordinate**
* **yCoordinate**
* **Year**

The model's output was then passed to the Apply Model and Performance (Classification) operators to generate predictions and evaluate accuracy, precision, recall, and other metrics.

# Assess model

After training the k-NN model, we assessed its performance by applying it to the test dataset using the **Apply Model** operator. This step allows us to compare the predicted outcomes with the actual labels (Arrest) in the test data.

The model's predictions were then passed to the **Performance (Classification)** operator, where key performance metrics such as accuracy, precision, and recall were computed. This assessment provides insights into how well the model generalizes to unseen data.

The assessment also highlighted the strengths and weaknesses of the k-NN algorithm in the context of this dataset:

* It performs well in correctly predicting non-arrest cases (high true negatives).
* It struggles more with correctly identifying actual arrests, which may be due to class imbalance or feature overlap in the data.

This stage helped us determine that while the model has good overall accuracy, further tuning or different modeling techniques may be needed to improve sensitivity to the minority class (true for Arrest).

# Evaluation

# Evaluate results

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Accuracy | 87.80 % |
| Precision(No arrest) | 88.53% |
| Precision (Arrest) | 74.68% |
| Recall(No arrest) | 98.44% |
| Recall(Arrest) | 27.58% |

Confusion Matrix:  
 - True Negatives (TN): 2,516  
 - False Positives (FP): 326  
 - False Negatives (FN): 40  
 - True Positives (TP): 118

# Interpret results

* The model is very good at detecting "No Arrest" cases (high recall: 98.44%)
* It is less effective at detecting actual arrests (recall: 26.58%)
* Precision for arrests is relatively decent at 74.68%, indicating good prediction confidence when it does classify a case as an arrest
* This performance imbalance likely reflects class imbalance in the dataset (more non-arrest cases than arrests)

The model identified several important predictors of arrest likelihood. The most influential factors were:

* **CrimeCategory**: More severe or violent crimes (e.g., assaults) were more likely to result in arrests.
* **LocationCategory**: Certain public locations and domestic settings had higher arrest rates.
* **Domestic flag**: Domestic cases often led to faster intervention and arrests.
* **Hour and Month**: Time-based patterns showed trends in arrest likelihood, possibly due to law enforcement resource allocation.

The use of weighted voting in k-NN gave greater importance to closer neighbors, which likely improved prediction performance, especially for ambiguous or borderline cases.

# Review of process

The process used in this project followed a clear, modular, and repeatable structure in RapidMiner, adhering to CRISP-DM methodology. Each step was implemented using distinct operator chains or subprocesses to ensure transparency and flexibility for iteration.

**1. Data Understanding & Cleaning**

The project began with importing the Chicago Arrest dataset, where initial filtering and stratified sampling were already performed to ensure representativeness across the Arrest label.

In the General Preparation subprocess, irrelevant or redundant columns such as Case Number, Location, and Updated On were removed using Select Attributes. Missing values were handled carefully:

LocationDescription was replaced with 'Unknown' (categorical handling)

xCoordinate and yCoordinate were imputed using the mean (numerical handling)

Duplicate records were checked using Remove Duplicates (none were found)

**2. Feature Engineering**

Meaningful new features were created using operators like Generate Attributes, Date to Numerical, and Rename:

Temporal features such as Hour, Month, and Year were extracted from the Date field

Crime types (originally 30+) were grouped into 10 higher-level CrimeCategory groups

Location descriptions were grouped into broader LocationCategory types

Attribute names were cleaned and standardized (e.g., removing spaces) for easier referencing

**3. Conversion to Modeling Format**

To prepare for k-NN (which is sensitive to data scaling and format):

Nominal to Numerical: Converted categorical variables using one-hot encoding

Normalize: Applied normalization to numerical attributes like Hour, Month, x Coordinate, and y Coordinate

This ensured all variables were on a comparable scale and ready for distance-based calculations.

**4. Model Training Using k-NN**

Using the k-NN operator:

k was set to 7 with weighted voting enabled

Mixed Euclidean Distance was used to account for both nominal and numerical attributes

The model was trained on the processed 70% training subset

**5. Model Assessment**

The Apply Model operator was used to generate predictions on the 30% test data. The predictions were passed into the Performance (Classification) operator, which output accuracy, precision, recall, and a confusion matrix.

# Determine next steps

* Address Class Imbalance - Try techniques like SMOTE or under-sampling to improve arrest prediction.
* Model Comparison - Compare k-NN with Decision Tree, Logistic Regression, or Random Forest.
* Parameter Tuning - Test different k values (e.g., 3, 5, 9) and see if prediction recall improves.
* Deploy Visualization - Consider mapping high-arrest-likelihood crime types or locations for deeper insights.

# Classification by Decision Tree

The primary goal of this classification process is to analyze crime patterns and determine key factors that influence crime categories. By using Decision Tree and Random Forest classification models, the analysis aims to identify patterns based on time of crime, arrest status, domestic involvement, geographic coordinates ( X and Y coordinates), and other descriptive crime-related attributes (community area, FBI code, beat, primary type, and location description). The insights from this classification can help in understanding crime trends and potentially aid in crime prevention strategies. The process involves filtering missing values, creating new attributes for better classification, setting roles for key variables, and discretizing time-based data to enhance predictive accuracy.

# Data Preparation

# Select data:

In the initial general data preparation, attribute selection was performed to ensure relevant features were included in the analysis. The dataset used for this classification task consists of the following attributes: Arrest, Beat, Community Area, Date, Description, Domestic, FBI Code, ID, Latitude, Location Description, Longitude, Primary Type, X Coordinate, and Y Coordinate. The goal is to determine key factors that influence the type of crime using Decision Tree and Random Forest classification models.

# Clean data

In this step, missing values were filtered out to ensure that only complete records were used in the analysis. I have used the The "Filter Examples" operator in RapidMiner was applied to remove incomplete or irrelevant data.

A screenshot of a computer

AI-generated content may be incorrect.

filters are applied to ensure that only records with complete values for certain attributes are used in the classification process. The selected attributes—X Coordinate, Y Coordinate, Community Area, and Location Description—must not be missing for a data record to be included in the analysis. This step helps improve the accuracy and reliability of the model by eliminating incomplete data entries that could otherwise impact classification performance

# Construct data

New Attributes Creation: Additional attributes were derived to enhance classification accuracy. Specifically, the "**Generate Attributes**" operator was used to generate a new attribute called Crime\_Category, which serves as the label class for classification. These attributes were based on date and primary type to improve understanding.

The newly created columns in the dataset serve to extract meaningful insights from existing attributes, helping improve the classification accuracy. These new attributes were generated using various functional expression:

* **Dayofweek**: Extracted from the "Date" column to determine the day of the week when a crime occurred, using date\_str(Date, "America/Newyork", "EEEE").
* **Month**: Extracted using date\_get(Date, DATE\_UNIT\_MONTH, "America/Newyork") + 1, which retrieves the month of occurrence.
* **Hour\_of\_crime**: Extracted from the timestamp using parse(date\_str(Date, "America/Newyork", "K")), helping analyze crimes based on time patterns.
* **Season**: Created using conditional logic to classify months into seasons (e.g., Spring, Summer, Fall, and Winter) based on the month value.
* **Crime\_Category**: Categorized crime types into broader groups for easier classification.

The **Crime\_Category** attribute was created by grouping different types of crimes into broader categories. I have created using **"Generate Attributes"** operator in RapidMiner, where the if conditions classify crimes based on their **Primary Type**.

Here’s how Crime\_Category was constructed:

1. Violent Crimes: If the crime type involves assault, homicide, kidnapping, battery, intimidation, or stalking, it is categorized as "Violent Crime."
2. Property Crimes: Crimes related to theft, burglary, robbery, arson, motor vehicle theft, and criminal damage/trespass are grouped as "Property Crime."
3. Sex & Child-Related Crimes: Offenses involving sexual assault, sex offenses, prostitution, public indecency, or crimes against children are labeled as "Sex & Child-Related Crime."
4. Drug & Alcohol Crimes: Crimes linked to narcotics and liquor law violations fall under "Drug & Alcohol."
5. Fraud & Deception: Includes gambling and deceptive practices.
6. Public Offenses: Includes weapons violations, public peace violations, interference with public officers, and concealed carry violations.
7. Other Offenses: If a crime does not fit into the above categories, it is classified as "OTHER OFFENSE**."**

* Year: Extracted from the "Date" column using date\_str(Date, "America/New\_York", "yyyy").
* Location\_Category: To enhance crime pattern analysis, crime locations were grouped into broader categories based on their nature. This categorization simplifies data interpretation and improves model efficiency in classification.

|  |  |
| --- | --- |
| * Category | * Locations Included |
| * Residential Area | * RESIDENCE, APARTMENT, RESIDENCE-GARAGE, RESIDENTIAL YARD (FRONT/BACK), RESIDENCE PORCH/HALLWAY, HOUSE, COLLEGE/UNIVERSITY RESIDENCE HALL, NURSING HOME/RETIREMENT HOME |
| * Transportation & Parking | * STREET, PARKING LOT/GARAGE, CTA BUS, CTA BUS STOP, CTA STATION, CTA PLATFORM, CTA TRAIN, CTA TRACKS - RIGHT OF WAY, AIRPORT EXTERIOR, AIRPORT TERMINAL, TAXICAB, VEHICLE NON-COMMERCIAL, VEHICLE - COMMERCIAL, VEHICLE - OTHER RIDE SHARE SERVICE (E.G., UBER, LYFT) |
| * Retail & Commercial | * GROCERY FOOD STORE, DEPARTMENT STORE, SMALL RETAIL STORE, CONVENIENCE STORE, RESTAURANT, BAR OR TAVERN, TAVERN/LIQUOR STORE, DRUG STORE, GAS STATION, APPLIANCE STORE, BARBERSHOP, BEAUTY SALON, MOVIE HOUSE/THEATER, COIN OPERATED MACHINE, CASINO/GAMBLING ESTABLISHMENT |
| * Public & Institutional | * SCHOOL - PUBLIC, SCHOOL - PRIVATE, LIBRARY, COLLEGE/UNIVERSITY, CHURCH/SYNAGOGUE/PLACE OF WORSHIP, GOVERNMENT BUILDING, FEDERAL BUILDING, HOSPITAL BUILDING/GROUNDS, POLICE FACILITY, MEDICAL/DENTAL OFFICE, FIRE STATION, DAY CARE CENTER, CREDIT UNION, CURRENCY EXCHANGE |
| * Industrial & Construction | * FACTORY/MANUFACTURING BUILDING, WAREHOUSE, CONSTRUCTION SITE, ABANDONED BUILDING, OTHER RAILROAD PROPERTY/TRAIN DEPOT, AIRPORT BUILDING, AIRPORT TRANSPORTATION SYSTEM (ATS) |
| * Outdoor & Recreational | * PARK PROPERTY, SPORTS ARENA/STADIUM, ATHLETIC CLUB, BEACH, LAKEFRONT/WATERFRONT/RIVERBANK, VACANT LOT, FOREST PRESERVE, PUBLIC GRAMMAR SCHOOL, CEMETERY, BRIDGE, ANIMAL HOSPITAL, AUTO/BOAT/RV DEALERSHIP, GANGWAY, BOWLING ALLEY |

These transformations allow for better pattern recognition in the dataset, making the classification process more efficient and easier.

**Attribute Role Assignment:** The "Set Role" operator was applied to designate "Crime\_Category" as the target label for the model. Assigning "Crime\_Category" as the target label aligns with Decision Tree and Random Forest models by enabling crime pattern recognition based on dataset features and supporting the prediction goal.

# Integrate data

Attribute Selection for Model Training: After preparing and transforming the dataset, the "Select Attributes" operator was used to ensure that only relevant features were included for classifications and the final date prep for tree-based classification has been done.

# Format data

Attribute Selection for Model Training: After preparing and transforming the dataset, the "Select Attributes" operator was used to ensure that only relevant features were included for classification.

Nominal to Numeric Transformation: Categorical attributes like "Arrest" and "Domestic" were converted into numerical values because machine learning models, including Decision Trees and Random Forests, require numerical input for processing. Converting these categorical variables allows the models to recognize patterns and relationships between these attributes and crime categories, improving classification accuracy.

Discretization by Time: The "Discretize by Frequency" operator was applied to group crime occurrences into specific time ranges (e.g., "Late Night" or "Morning"). This transformation helps capture temporal crime patterns, making it easier for the model to learn meaningful trends, such as increased crime rates at certain times of the day. It also reduces noise in raw time data, leading to better generalization and predictive performance

# Modelling

# Select modeling techniques

# Generate test design

# Build model

# Assess model

# Evaluation

# Evaluate results

# Interpret results

# Review of process

# Determine next steps

# Clustering by kMeans (clustering and finding outliers)

Clustering is an unsupervised machine learning method utilized to organize data points according to their similarities. In contrast to default classification, which involves predefined labels, clustering aids in uncovering inherent structures present in the data without any prior knowledge of categories. This characteristic renders it especially beneficial for exploratory data analysis and anomaly-detection.  
  
In the context of crime analysis, clustering can assist in spotting patterns in criminal activities, such as hotspots for certain crimes, seasonal variations, or relationships between various types of offenses. It is also useful for identifying outliers in unusual crime incidents that may signal reporting inaccuracies, emerging crime patterns, or significant events requiring further investigation.  
  
K-Means is a well-known clustering algorithm because of its effectiveness in managing large datasets. It divides data into K clusters by reducing the variance within each cluster. By implementing K-Means on Chicago’s crime dataset, we can reveal spatial and temporal patterns in criminal activities, categorize different types of offenses, and potentially pinpoint areas with rising or falling crime trends. Furthermore, identifying outliers in this dataset may assist law enforcement in concentrating on unique or infrequent crime occurrences.

# Data Preparation

Data preparation or pre-processing steps are required, the same as any other Machine learning model. Clustering is a distance-based method. Therefore, the aim of the distance-based relevant data-processing is to make the dataset noise free before applying clustering on it.

# Select data

After the General preparation, Relevant columns were taken into consideration for any of the classification machine learning tasks. Although this initial column selection is helpful. Many new attributes had to be generated in each specific task as seen before in KNN and Decision Tree. Feature engineering played a crucial role in refining the dataset. Derived attributes were created based on domain knowledge and statistical transformations. For example, categorical variables were encoded, temporal data was processed into meaningful time-based features, and location-based attributes were discretized for better interpretability.

Furthermore, in previous experiments with KNN and Decision Tree models, additional features significantly impacted classification accuracy. The inclusion of engineered attributes, such as crime type frequency, normalized coordinates, and contextual location groupings, improved the models’ ability to discern patterns. Therefore, In Clustering many new features were derived from the date column. Location Description was discretized into groupings of distinct categories. Different clustering techniques use Different set of attributes as their selection of attributes. A General set of attributes for all clustering methods are shown as in the image.

For all three clustering techniques, as per their reflective clustering approach selection, these many sets of columns were selected.

|  |  |  |
| --- | --- | --- |
| Spatial Clustering (Geographic) | Crime-Type Clustering | Final Set of Selected columns after “Clustering General Preparation.”  Ultimately, Same set used in the Clustering Outliers integrated approach |
|  |  |  |

All the relevant pre-processing steps before clustering are done after. Documented in the Format Data Step.

# Clean data

Only three columns are reported to have missing values approximating a total of 2.96% of the whole dataset. Any approach can be taken to handle these missing values like deletion (of missing value rows) or imputation or enhanced feature engineering before imputation. We Utilized the Replace missing values operator in RapidMiner (and used average/mean to impute). The reason for selection of this method was to not lose important data, which is likely to happen if methods like deletion are performed. We tried to remove duplicates on all columns as well, but in the end, we found no duplicates.

# Construct data

The primary reason for discretization for various attributes was the nature of high cardinality observed in many of the columns like location description, FBI code etc. As Clustering is a distance-based method sparse and high cardinality (many distinct values) fail to capture the relationships and nuance between attributes.

Date column was segregated, and new features were derived from the segregations. Year, Month, Day, Hour, were all separated from the date column.

|  |  |
| --- | --- |
|  |  |

Season (Six seasons) and Time of the day were generated respectively from Month and Hour. These new features aid in the clustering of the instances. Location type is the resulting discretization of the Location description column values. 100 plus distinct values of the location description column were broken down into 10 main features. The Primary type of crime (theft, robbery etc. ) was also discretized from 31 distinct features to half. District is discretized based on the four cardinal directions: north, south, east and west. Furthermore, FBI Code is also discretized based on the severity of the crime, and the official FBI group code index taken as the [source.](https://ucr.fbi.gov/nibrs/2011/resources/nibrs-offense-codes)

# Integrate data

Integration with other datasets wasn’t necessary for clustering and our purpose. Re-integration is achieved as the original column values were joined in all clustering chains/approaches to interpret the clustering results of the instances.

# Format data

Relevant processing steps for Distance based models are applied like One-hot-encoding for Categorical columns and Normalization/Standardization for numerical columns.

**Normalization (Min-Max Scaling)**: Rescales numerical values to a fixed range, typically 0 to 1, using the formula:

Columns like X and Y coordinates are normalized to a range of 0 to 1 to bring them within a specific scale, ensuring they do not disproportionately affect the clustering or modeling process.

**Standardization (Z-score Scaling)**: Transforms numerical data to have a mean of 0 and a standard deviation of 1, using:

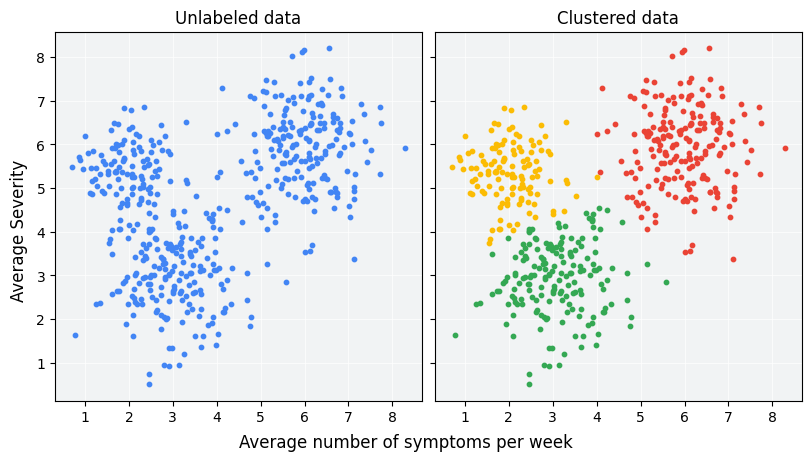
Columns like Year are standardized to have a mean of 0 and a standard deviation of 1. Standardizing the year preserves its temporal relationship, ensuring clustering models respect time-based trends. Normalizing the year distorts its sequence, treating it as an arbitrary number on the same scale as other variables. Standardization allows for better comparison with other features without losing time-based context.

**One-Hot Encoding**: Converts categorical variables into binary vectors, creating a separate column for each unique category where only one value is "1" and the rest are "0," allowing models to process categorical data numerically.

# Modelling

# Select modeling techniques

A typical example of clustering looks like this, image example from another dataset ([source](https://www.researchgate.net/figure/based-on-SSe-approach-a-Spatiotemporal-clusters-for-crime-dataset-b-Clusters-for_fig2_326403569))



In this clustering task, a **nested clustering** approach is utilized to gain unique insights from both spatial and crime-type data. Initially, two main clustering techniques are applied: **Spatial Clustering** and **Crime-Type Clustering**. These approaches are then integrated in a final clustering model to uncover deeper patterns and outliers.

1. **Spatial Clustering** groups crime incidents based on geographic proximity using location-based features. This helps identify crime hotspots and patterns across different areas, aiding in resource allocation and crime prevention strategies.
2. **Crime-Type Clustering** categorizes incidents based on crime characteristics such as primary type, this approach reveals patterns in the types of crimes occurring, helping to understand trends and correlations between different offenses.
3. To cluster and find outliers a final integrated Spatial and Crime-Type clustering is selected. This final clustering step combines both approaches to identify overlapping patterns, detect outliers, and ultimately generate clusters that can assist in predicting and addressing crime trends more effectively.

**The Elbow Method** will be used to assess the optimal number of clusters and the quality of the resulting groups. This ensures that the clustering approach is both efficient and provides actionable insights.

# Generate test design

In this phase, we created assessments to measure the performance of K-means clustering on the crime dataset. The assessment involved establishing the ideal number of clusters using methods such as the Elbow Method to evaluate the appropriateness of K-means for the specified data. We also chose to conduct clustering on a selection of features that denote both spatial (districts, geographical data) and temporal (time, year) characteristics. Critical variables were standardized to guarantee that all features contributed equally to the clustering activity. The results of these assessments will enable us to analyze the categorization of crime incidents based on trends and ascertain if the detected clusters provide significant insights.

# Build model

For the K-means clustering model, we processed the data by addressing missing values, standardizing numerical variables, and encoding categorical features (e.g., type of crime, area). The best number of clusters was identified using the Elbow Method, where the total of squared distances (inertia) is graphed against the number of clusters, and the "elbow" point signifies the preferred number of clusters. K-means clustering was subsequently utilized on the dataset with the determined number of clusters. We employed the K-means algorithm from RapidMiner, which repeatedly operates to reduce the within-cluster variance. The outcome was a compilation of cluster assignments for each crime record, aligning similar incidents based on spatial and temporal trends.

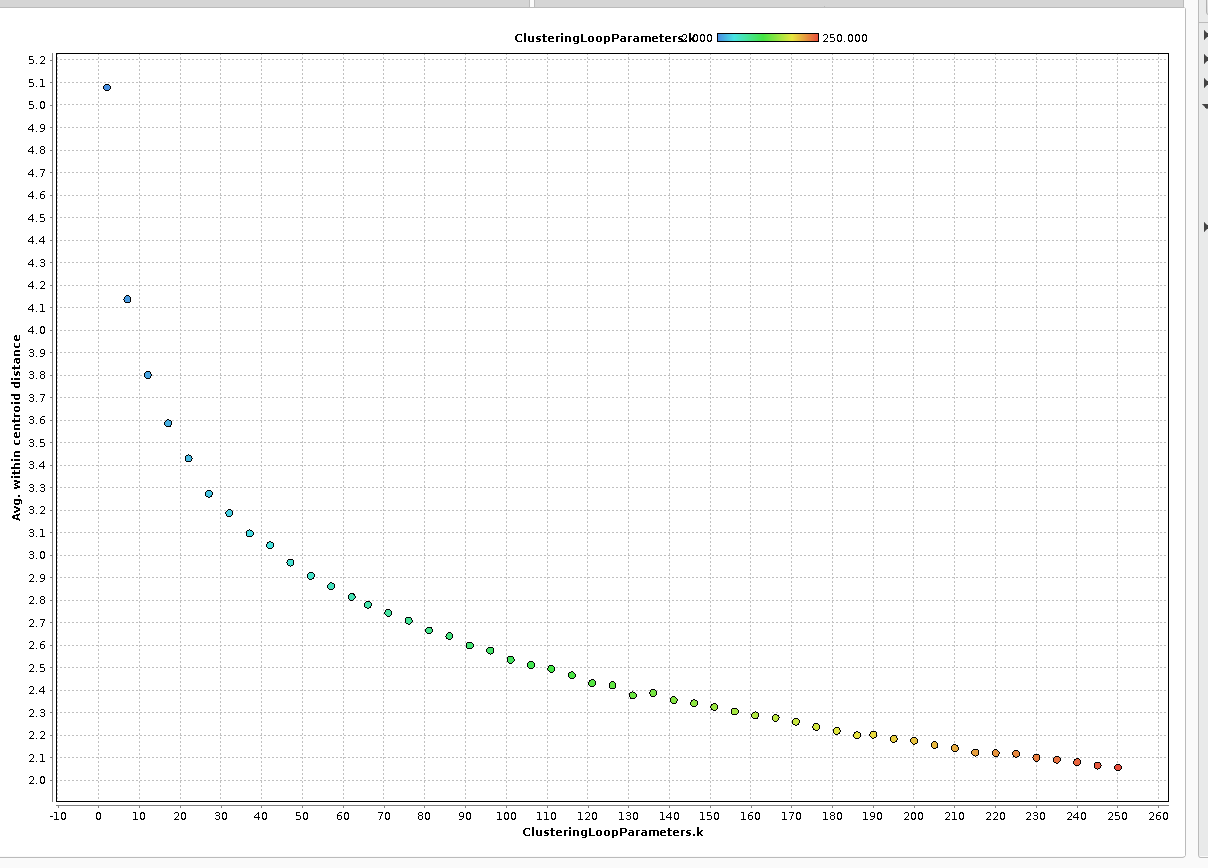
# Assess model

To assess the performance of the K-means clustering model, we evaluated the resulting clusters using both **internal** and **external** validation methods. Average distances within centroids were calculated to measure how well-separated the clusters are, with a higher score indicating better-defined clusters. We also analyzed the **cluster centroids** to understand the characteristics of each cluster, such as common crime types or locations. Additionally, we performed a qualitative assessment by visualizing the clusters on a map to see if the geographic distribution of crime incidents aligned with the expected outcomes. Finally, the clustering results were compared to existing classifications from decision tree to assess the model’s external validity. Based on these assessments, we concluded whether the clustering model provided meaningful insights into crime patterns in the dataset.

# Evaluation

# Evaluate results

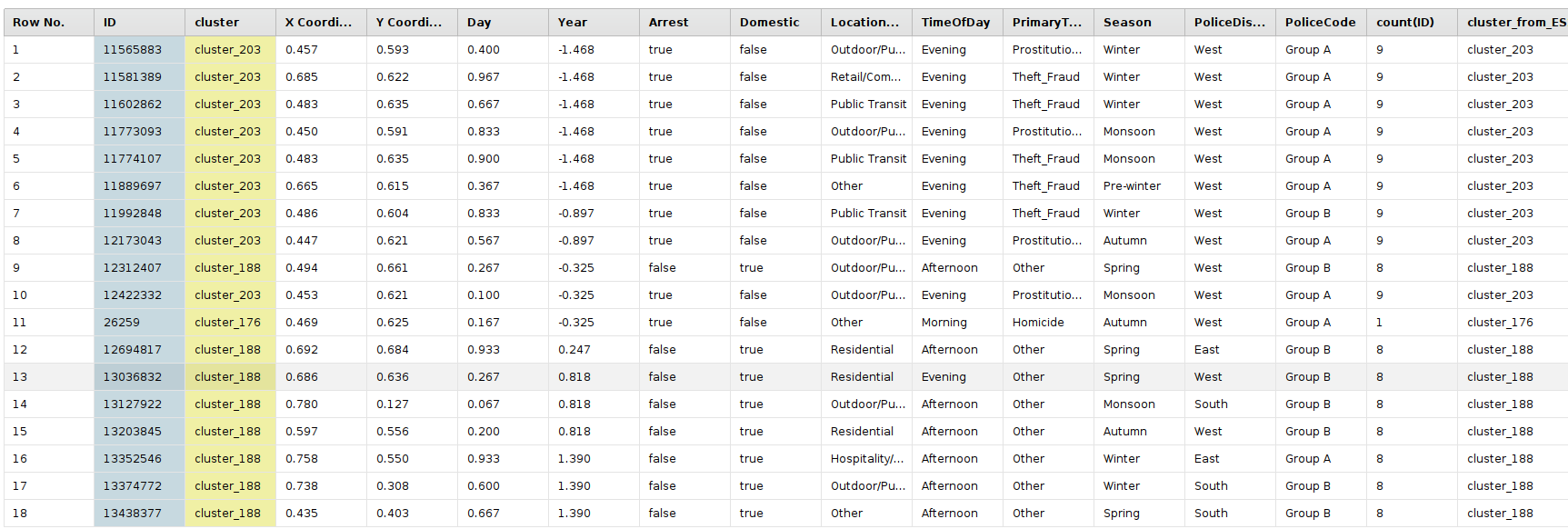
For all clustering techniques and the initial elbow plot, we tried to experiment with different hyperparameters like the min and max range for K clusters to loop, changing Steps and Scale functions. A final range of 2 to 250 clusters with 50 steps and linear scale function were selected.



From the provided chart, the sharpest bend appears around *k = 30 to 50.* Beyond this point, the decrease in the within-cluster distance slows down significantly, indicating diminishing returns in reducing clustering error. Given this observation, we are selecting as the optimal number of clusters for both clustering techniques, Spatial Clustering and Crime-Type Clustering. On the other hand, for integrated approach for clustering and outlier detection, a K cluster value of 260 was selected to seclude and detect outliers. This choice balances clustering accuracy and computational efficiency while ensuring meaningful groupings across different clustering methodologies.

# Interpret results

After Clustering the clusters below shared instances count of less than 10 were aggregated and stored.

After performing clustering on my Chicago crime dataset, clusters with fewer than 10 shared instances were specifically analyzed to determine if they followed any meaningful patterns. This was done as part of the clustering interpretation process, rather than treating them strictly as outliers. The assumption was that such small clusters might reveal underlying trends or unique crime patterns rather than being dismissed as noise. Upon analysis, these clusters did exhibit patterns, providing valuable insights into less frequent but potentially significant crime behaviors. No additional storage or aggregation was done beyond this interpretative step.

# Review of process

# Determine next steps

# Outlier Detection by LOF and ISF (and common outliers)

# Data Preparation

Our aim with outlier detection is to identify possible anomalies in the dataset. These anomalies may include unusual crime trends or suspicious crime types based on demographics, dates, or times. Outlier detection can also reveal potential biases in the dataset that may impact the final classification models and the analysis of crime factors.

While both ISF (Isolation Forest) and LOF (Local Outlier Factor) are effective methods for finding outliers, ISF is particularly suitable for identifying sudden crime trends or unusual crimes in specific locations and times. This is due to ISF's nature of randomly splitting datasets, which makes it effective at detecting anomalies occurring at unexpected or infrequent times. On the other hand, LOF is more suited for general outlier detection, uncovering biases in the dataset or addressing other potential data quality issues.

For this analysis, we stratified the data to include 10,000 instances spanning the years 2019 to 2024. Since RapidMiner only permits 10,000 records for non-subscribers, we further sampled the data down to 10,000 rows. This process also removed all instances with missing values.

# Select data

The dataset contains 22 attributes, plus 'ID,' making a total of 23. To achieve the best results in outlier detection, we selected all the features that play the most crucial roles in arrests, such as date and location. For this analysis, we used a total of 14 attributes: Arrest (nominal), Block (numeric), Date (date), Description (string), District (numeric), Domestic (nominal), Latitude (numeric), Location (string), Location Description (string), Longitude (numeric), Primary Type (string), Updated On (date), Year (numeric), and ID (ID).

# Clean data

To ensure data quality and avoid any biases, I used the ‘Remove Duplicate’ operator to remove all the duplicates that can increase the noise in our model and result.

# Construct data

I created eight new features, most of which come from the ‘Date’ column. These features:

|  |  |  |
| --- | --- | --- |
| Name | Description | Type |
| Dayofweek | Extract the day of the week, (Monday, Tuesday...) | String |
| Month | Extract month of the year | numeric |
| IsWeeknd | Is ‘Dayofweek’ is weekend or not. True for Saturday and Sunday | bool |
| Hour\_of\_crime | Extract the hour of crime from date in 24 hours format | time |
| Season | Group ‘Month’ into season if month equal to (3, 4, 5) then it is “spring”, if the month equals to (6,7,8) then Summer if it equal to (9,10,11) then Fall otherwise winter | String |
| Visibility | Group ‘Hour\_of\_crime’ into how much visibility would be at that time. If ‘Hour\_of\_crime’ is between 0 to 6 visibility is “Very Dark”, if between 6 and 8 or between 19 and 21 visibility is “Low Visibility” if between 8 and 19 then Bright otherwise Dark. | String |
| Year | Extract year from date | String |
| Crime\_Category | Group ‘Primary Type’ into 7 category they are:  **Violent Crime:** ASSULT, BATTERY, HOMICIDE, KIDNAPPING, INTIMIDATION, STALKING.  Property Crime: THEFT, BURGLARY, ROBBERY, MOTOR VEHICLE THEFT, CRIMINAL DAMAGE, ARSON, CRIMINAL TRESPASS.  **Sex & Child-Related:** CRIMINAL SEXUAL ASSULT, SEX OFFENSE, PROSTITUTION, OFFENSE INVOLVING CHILDREN, PUBLIC INDECENCY, OBSCENTITY.  **Drug & Alcohol:** NARCOTICS LIQUOR LAW VIOLATION  **Fraud & Deception:** GAMBLING, DECEPTIVE PRACTICE.  **Public Offense:** WEAPONS VIOLATION, PUBLIC PEACE VIOLATION, INTERFERENCE WITH PUBLIC OFFICER, CONCEALED CARRY LICENSE VIOLATION.  **OTHER OFFENSE:** for all other crimes assign | String |

# Integrate data

After creating new features and integrating them with the updated dataset, I selected the final list of attributes for the ISF and LOF outlier detection models. These attributes are Arrest, Crime\_Category, Domestic, Is Weekend, Season, Visibility, and Year

# Format data

All the attributes are of nominal type, with 'Crime\_Category' having the most distinct values, totaling seven.

One-hot encoding is required before feeding the data into the ISF and LOF models.

# Modelling

# Select modeling techniques

# Generate test design

# Build model

# Assess model

# Evaluation

# Evaluate results

# Interpret results

# Review of process

# Determine next steps

# Conclusion