CST8390\_020 Business Intelligence and Data Analytics

# Project Plan Report: Analyzing Crime Trends in New York City

Presented by:

Mansi Joshi (041091664)

Ved Patel (041095249)

Isha Gadani (041085940)

Hamza El Sousi (040982818)

Guntas Singh Chugh (041091309)

Prince Khakhriya (041133835)

## 

## Presented to:

Prof. Grace Pauly

## 

Table of Contents

[Project Plan Report: Analyzing Crime Trends in New York City 1](#_Toc183383747)

[Presented by: 1](#_Toc183383748)

[Presented to: 1](#_Toc183383749)

[1. Introduction 4](#_Toc183383750)

[2. Business Understanding 5](#_Toc183383751)

[2.1 Determine Business Objectives 5](#_Toc183383752)

[2.2 Assess situation 5](#_Toc183383753)

[2.3 Determine Data Mining Goals 6](#_Toc183383754)

[2.4 Produce Project Plan 7](#_Toc183383755)

[3. Data Understanding 9](#_Toc183383756)

[3.1 Collect Initial Data 9](#_Toc183383757)

[3.2 Describe Data 9](#_Toc183383758)

[3.3 Explore Data 12](#_Toc183383759)

[3.4 Verify data quality 17](#_Toc183383760)

[4. Data Preparation 20](#_Toc183383761)

[4.1 Select data 20](#_Toc183383762)

[4.2 Clean data 22](#_Toc183383763)

[4.3 Construct Data 22](#_Toc183383764)

[4.4 Integrate Data 24](#_Toc183383765)

[4.5 Format Data 24](#_Toc183383766)

[4.5.1 Decision Tree: 24](#_Toc183383767)

[4.5.2 k-Nearest Neighbors (kNN): 24](#_Toc183383768)

[4.5.3 Random Forest (RF): 25](#_Toc183383769)

[4.5.4 k-Means: 25](#_Toc183383770)

[4.5.5 Local Outlier Factor (LOF): 25](#_Toc183383771)

[4.5.6 Isolation Forest Outlier: 26](#_Toc183383772)

[4.5.7 Association: 26](#_Toc183383773)

[5. Modeling 28](#_Toc183383774)

[5.1 Select Modeling Techniques 28](#_Toc183383775)

[5.2 Generate Test Design 28](#_Toc183383776)

[5.3 Build Model 30](#_Toc183383780)

[5.3.1 Decision Tree: 30](#_Toc183383781)

[5.3.2 Random Forest: 31](#_Toc183383782)

[5.3.3 kNN: 32](#_Toc183383783)

[5.3.4 Detect Outlier (LOF) 33](#_Toc183383784)

[5.3.5 Isolation Forest Outlier 37](#_Toc183383785)

[5.3.6 K-means Clustering 38](#_Toc183383786)

[5.3.7 Association Rule Mining 39](#_Toc183383787)

[5.4 Assess Model 40](#_Toc183383788)

[5.4.1 Random Forest: 40](#_Toc183383789)

[5.4.2 Decision Tree: 42](#_Toc183383790)

[5.4.3 kNN: 43](#_Toc183383791)

[5.4.4 K-means Clustering: 44](#_Toc183383792)

[5.4.5 Association rule mining (Apriori): 46](#_Toc183383793)

[6. Evaluation 50](#_Toc183383794)

[6.1 Evaluate Results 50](#_Toc183383795)

[6.2 Interpret Results 51](#_Toc183383796)

[6.3 Review of Process 53](#_Toc183383797)

[6.4 Determine Next Steps 57](#_Toc183383798)

[7.0 Conclusion 59](#_Toc183383799)

# Introduction

This report documents the comprehensive data analysis process performed on the NYPD Complaint Data, a dataset that provides detailed records of reported crime incidents across New York City.

The report follows a structured approach in line with industry standards, beginning with a thorough understanding of business objectives. These objectives are translated into specific data mining goals to ensure that the analysis aligns with the broader aims of crime prevention and resource management. By focusing on classification, clustering, and outlier detection, the project seeks to uncover patterns in criminal activity that may be influenced by factors such as location, time, and demographic characteristics.

Subsequent phases of the project involve data preparation, which is crucial for transforming raw data into a usable form. This includes cleaning, formatting, and standardizing the dataset, followed by exploratory analysis to gain initial insights and validate data quality. Rigorous data preparation steps ensure that our dataset is both reliable and suitable for analytical techniques, forming a solid foundation for robust modeling and meaningful results.

Throughout this analysis, RapidMiner software is employed to streamline data processing and enhance analytical efficiency. By using this tool, we efficiently conduct tasks such as filtering, sampling, and normalization, ensuring the data is optimized for our models. The findings from this project aim to highlight crime trends, potential hotspots, and atypical incidents, providing actionable insights for law enforcement agencies to make informed decisions that may ultimately contribute to reducing crime rates and enhancing public safety across New York City.

# 2. Business Understanding

## 2.1 Determine Business Objectives

The goal of this project is to analyze the NYPD Complaint Data (Year to Date) to support public safety initiatives by identifying actionable insights. This dataset includes all felony, misdemeanor, and violation complaints reported to the NYPD.

The main objectives are:

* **Identifying Crime Trends:** Analyze the dataset to uncover the most prevalent types of crimes, their frequency, and geographical distribution. By mapping where and when crimes occur most often, this analysis will highlight key hotspots and recurring patterns.
* **Anomaly Detection:** Identifying unusual crime patterns or anomalies in the dataset. These outliers may signify emerging threats, areas with underreported incidents, or inconsistencies in data entry.
* **Crime Type Prediction and Prevention:** Implement classification algorithms like Decision Trees and Random Forests to predict and categorize crimes based on historical data. This predictive analysis will identify rising categories of criminal activity and provide recommendations for effective prevention strategies.

## 2.2 Assess situation

To successfully analyze the NYPD Complaint Data, the following factors have been assessed:

* **Availability of Resources:** 
  + **Tools and Software:** RapidMiner will be employed for data preprocessing, building models, and conducting in-depth analysis. Power BI will be used to create interactive dashboards and visualizations that clearly present crime trends, anomalies, and predictions.
* **Potential Risks:** 
  + **Data Quality Issues:** Missing values, duplicate entries, and inconsistent formats in the dataset could compromise the accuracy of insights. In addition to this, errors in reporting or incomplete records may limit the depth of analysis.
  + **Visualization Challenges:** Poorly designed charts or misleading visuals could hinder the communication of findings. Similarly, overly complex or cluttered graphs may confuse stakeholders and reduce the impact of insights.
* **Contingency Plans for Risks:**
  + **For Data Quality:** Apply robust data cleaning, and preprocessing to ensure data reliability. Use data validation steps to identify and rectify inconsistencies early in the preprocessing phase.
  + **For Visualization Issues:**  Perform multiple design iterations, prioritizing clarity and relevance. Simplify complex data by breaking it into smaller, digestible visual components.

## 2.3 Determine Data Mining Goals

The following data mining goals have been established to achieve the business objectives:

* **Identify Crime Trends:** This involves a comprehensive analysis of crime data to uncover recurring patterns and trends over time. By examining the frequency of incidents, specific types of crimes, and their geographical distribution This information can guide resource allocation, strategic planning, and community safety initiatives, ensuring a data-driven approach to crime prevention.
* **Outlier Detection:** Anomaly detection methods can be applied to crime datasets to pinpoint unusual patterns or inconsistencies that deviate significantly from normal behavior. These outliers may indicate emerging threats, underreported areas of concern, or data entry errors requiring correction. Techniques like clustering, statistical analysis, and machine learning algorithms can help flag incidents or locations that exhibit atypical trends, such as crimes that don't align with usual patterns in the area.
* **Classification of Crime Types:** Using machine learning algorithms such as Decision Trees and Random Forests, crimes can be categorized effectively based on historical data and associated attributes. These models can classify incidents into predefined categories, and predict the likelihood of an increase in certain types of crimes. This approach allows for proactive measures, tailored strategies, and the efficient deployment of resources to mitigate their impact.

## 2.4 Produce Project Plan

The project plan is segmented into key phases to ensure systematic progress toward our goals:

* **Data Collection and Understanding**:
  + Obtain and explore the NYPD Complaint Data (Current Year To Date) dataset.
  + Explore the dataset to gain insights into its structure, identifying key attributes like crime types, dates, locations, and frequencies.
* Verify the quality of data by addressing issues such as missing values, inconsistent entries, and duplicate records to ensure reliability and accuracy.
* **Data Preparation**:
  + Perform data cleaning and preprocessing, including handling missing values, encoding categorical variables for machine learning algorithms.
  + Format data for modeling in RapidMiner.
* **Modeling and Analysis**:
  + Implement and train classification models using Decision Tree and Random Forest to categorize crime incidents effectively and predict emerging trends.
  + Apply anomaly detection techniques to identify unusual crime patterns, irregular data points, or outliers in the dataset for further investigation.
  + Perform clustering analysis to group similar crime incidents and uncover hidden relationships or patterns within the data.
* **Visualization:** 
  + Use **Power BI** and **RapidMiner** to create compelling visualizations that illustrate crime trends, patterns, and predictions.
  + Develop interactive dashboards and graphs to present findings clearly and enable stakeholders to explore data insights.
* **Evaluation**:
  + Assess the accuracy and effectiveness of the models using performance metrics such as precision, recall, F1-score, and overall accuracy.

* + Validate model results through cross-validation techniques and compare outcomes across different models to select the best-performing one.
* **Deployment:** 
  + Present findings in a detailed, well-structured report that includes visualized insights, actionable recommendations, and strategies for crime prevention.

# 3. Data Understanding

## 3.1 Collect Initial Data

The dataset was sourced from the NYPD Complaint Data repository, which provides an extensive record of complaints made to the NYPD, including type, location, and time details of incidents. The data captures a diverse range of fields, allowing for multifaceted analysis.

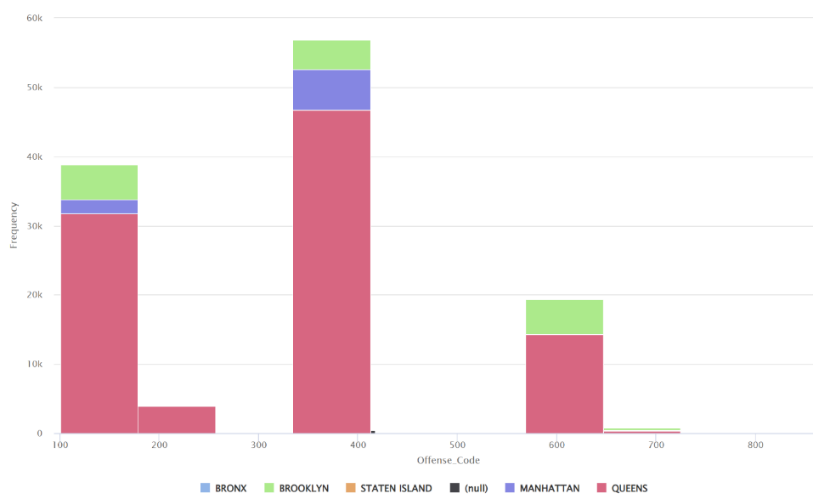
## 3.2 Describe Data

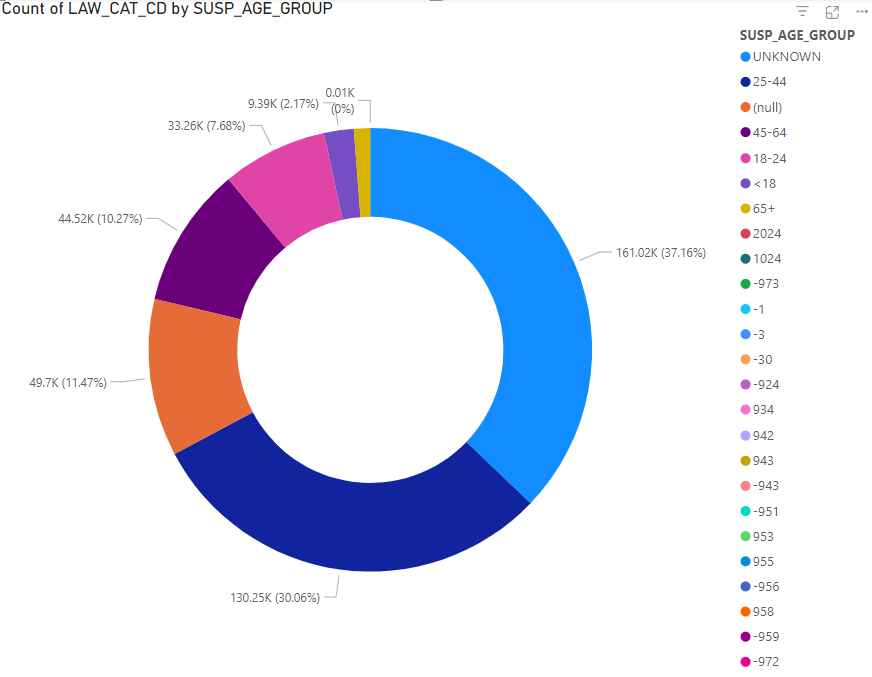
We had total 36 attributes from the original data. The datatype of these attributes and what each of them represents is given below in the table:

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Description** | **Data Type** |
| CMPLNT\_NUM | Randomly generated persistent ID for each complaint | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| ADDR\_PCT\_CD | The precinct in which the incident occurred | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| BORO\_NM | The name of the borough in which the incident occurred | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| CMPLNT\_FR\_DT | Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT\_TO\_DT exists) | [Floating Timestamp](https://dev.socrata.com/docs/datatypes/floating_timestamp.html) |
| CMPLNT\_FR\_TM | Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT\_TO\_TM exists) | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| CMPLNT\_TO\_DT | Ending date of occurrence for the reported event if exact time of occurrence is unknown | [Floating Timestamp](https://dev.socrata.com/docs/datatypes/floating_timestamp.html) |
| CMPLNT\_TO\_TM | Ending time of occurrence for the reported event if exact time of occurrence is unknown | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| CRM\_ATPT\_CPTD\_CD | Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| HADEVELOPT | Name of NYCHA housing development of occurrence, if applicable | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| HOUSING\_PSA | Development Level Code | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| JURISDICTION\_CODE | Jurisdiction responsible for incident. Either internal, like Police (0), Transit (1), and Housing (2); or external (3), like Correction, Port Authority, etc. | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| JURIS\_DESC | Description of the jurisdiction code | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| KY\_CD | Three-digit offense classification code | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| LAW\_CAT\_CD | Level of offense: felony, misdemeanor, violation | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| LOC\_OF\_OCCUR\_DESC | Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| OFNS\_DESC | Description of offense corresponding with key code | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| PARKS\_NM | Name of NYC park, playground, or greenspace of occurrence, if applicable (state parks are not included) | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| PATROL\_BORO | The name of the patrol borough in which the incident occurred | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| PD\_CD | Three-digit internal classification code (more granular than Key Code) | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| PD\_DESC | Description of internal classification corresponding with PD code (more granular than Offense Description) | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| PREM\_TYP\_DESC | Specific description of premises; grocery store, residence, street, etc. | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| RPT\_DT | Date event was reported to police | [Floating Timestamp](https://dev.socrata.com/docs/datatypes/floating_timestamp.html) |
| STATION\_NAME | Transit station name | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| SUSP\_AGE\_GROUP | Suspect’s Age Group | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| SUSP\_RACE | Suspect’s Race Description | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| SUSP\_SEX | Suspect’s Sex Description | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| TRANSIT\_DISTRICT | Transit district in which the offense occurred. | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| VIC\_AGE\_GROUP | Victim’s Age Group | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| VIC\_RACE | Victim’s Race Description | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| VIC\_SEX | Victim’s Sex Description | [Text](https://dev.socrata.com/docs/datatypes/text.html) |
| X\_COORD\_CD | X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| Y\_COORD\_CD | Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| Latitude | Midblock Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| Longitude | Midblock Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) | [Number](https://dev.socrata.com/docs/datatypes/number.html) |
| Lat\_Lon |  | [Location](https://dev.socrata.com/docs/datatypes/location.html) |
| New Georeferenced Column |  | [Point](https://dev.socrata.com/docs/datatypes/point.html) |

## 3.3 Explore Data

Due to data quality issues such as null or incorrect values addressed below in the document, extracting meaning from charts are challenging. Below are charts before processing with details below.

  
Figure 1  
  
The stacked bar chart shown above in figure 1 indicates the frequency of offences committed in different neighborhoods in New York City. I can be observed that the most numbers of offences are committed in Queens, followed by Brooklyn and Manhattan. Offence level in Bronx and Staten Island is negligible.

  
Figure 2

The Donut Chart shown above depicts the percentage of crimes committed by suspects of different age groups. It can be observed that 37.16% of times the age group of the suspect is unknown. Moreover, greatest number of offences are committed by the suspects under the age-group of 25-44, whereas people of age-group 65+ commit the least number of offences.

A graph of blue and orange bars

Description automatically generated  
Figure 3

The Clustered column chart shown above in figure 3, depicts the number of offences committed by suspects of different races. Therre are 6 types of races: Black, White Hispanic, White, Black Hispanic, Asian, and American Indian. Each race has committed 3 types of crimes: Felony, Misdemeanor, and Violation. Most amounts of offences are committed by Black, and the least is done by American Indians.

A graph of different colored bars

Description automatically generated  
Figure 4

The Bar Chat shown in figure 4 depicts the frequency of incident occurrence within a precinct. This data indicates how often a precinct has had an incident; the chart shows Bronx Precinct having the highest frequency of incidents.

A graph of a graph

Description automatically generated with medium confidence  
Figure 5

The line chart shown in figure 5 shows the number of different offences; felony, misdemeanor, and violation; in different neighborhoods in the New York City. It can be observed that misdemeanor has the highest number of counts in every area. Similarly, violation has the least amount occurrences in every area.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

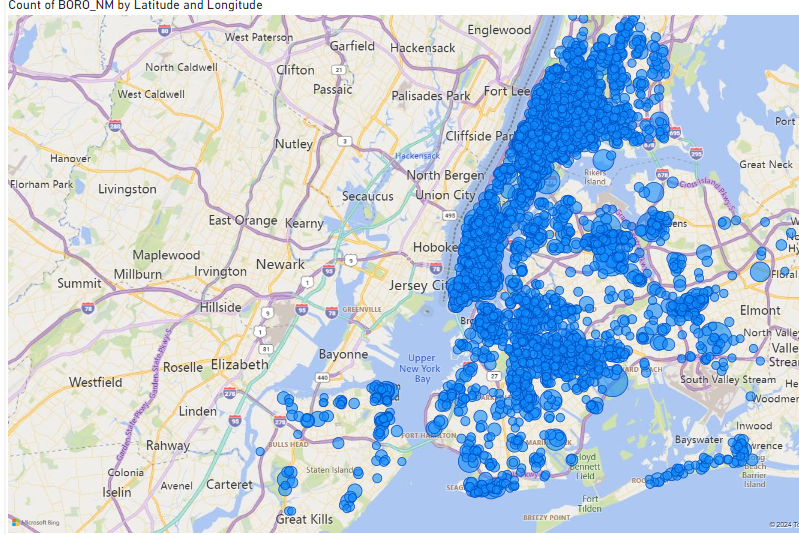
Description automatically generated  
Figure 6 Figure 7

The figure 6 and 7 shows the type of premises in which the offence had occurred, and the diverse types of offences and their numbers within the selected premises. It can be observed that, the greatest number of offences have occurred in the streets, the least number were committed in commercial buildings. It should also be noted that misdemeanor offences were the highest number of offences irrespective of the premises of the offence.

A map of the united states

Description automatically generated  
Figure 8

Figure 8 shows the map of New York City. The figure shows offence levels on different coordinates on the map. It can be observed that the offence level is very dense near the center of the city.

  
Figure 9

Map chart Showing the Regions of most incident occurrences. Majority of cases are clustered mainly in the city center, as well as the northwestern region. In further phases, (LOF) Local Outlier Factoring will help us paint a clearer picture on a more micro level.

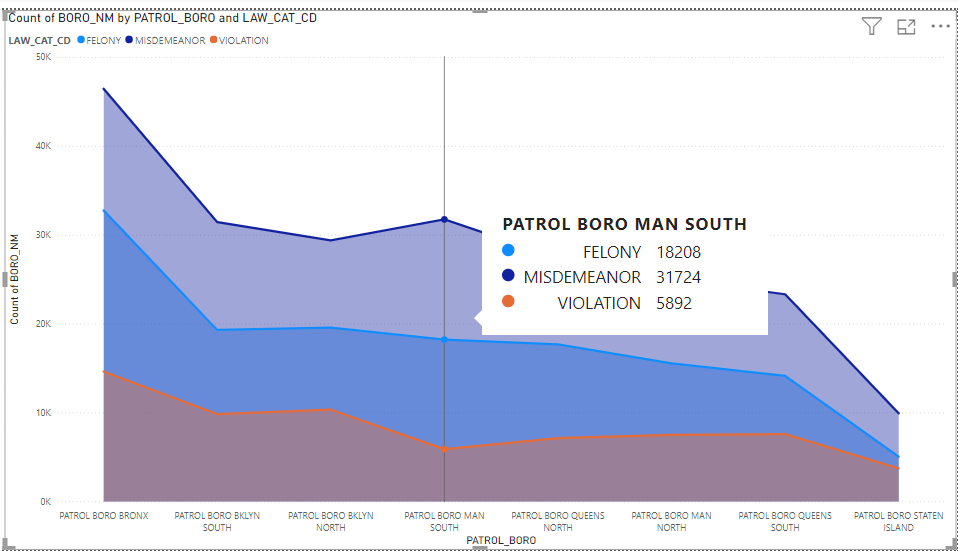


Figure 10

Figure 10 depicts an area chart which helps to understand which type of offence happened in what patrolling area of the Newyork. It can be observed that the highest number of offences of each kind occured in the patrol area of Bronx. Second major observation that can be made is that misdemeanor is the highest level of offence across the whole city, whereas violations is the least of all offences.

Exploratory data analysis involved examining data distributions, identifying data type inconsistencies, and confirming the presence of missing values. This stage highlighted the need for:

* **Data Cleaning**: Removal of records with missing values in essential columns.
* **Data Transformation**: Conversion of certain fields into categorical or numerical formats.
* **Normalization and Encoding**: Preparation of fields to meet modeling requirements, ensuring that the data would be suitable for machine learning algorithms that require specific formats.

The results from this exploration informed the design of our data preparation process to ensure that our data was both complete and consistent.

3.4 Verify data quality

To ensure data quality, we assessed each attribute and have issues identified during exploration. This involved missing values, null values, duplicates, and repeating data values that does not provide any valuable insights. The table below shows the attributes and the amount of missing data/ null values or provides the reason for not using it.

|  |  |  |
| --- | --- | --- |
| Attribute Name | Description | Reason to not use these column |
| CMPLNT\_NUM | Randomly generated persistent ID for each complaint | Unique numbers doesn’t help in data analysis  Cardinality high |
| ADDR\_PCT\_CD | The precinct in which the incident occurred |  |
| BORO\_NM | The name of the borough in which the incident occurred |  |
| CMPLNT\_FR\_DT | Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT\_TO\_DT exists) |  |
| CMPLNT\_FR\_TM | Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT\_TO\_TM exists) |  |
| CMPLNT\_TO\_DT | Ending date of occurrence for the reported event if exact time of occurrence is unknown | 22399 missing values |
| CMPLNT\_TO\_TM | Ending time of occurrence for the reported event if exact time of occurrence is unknown | 22035 null values |
| CRM\_ATPT\_CPTD\_CD | Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely |  |
| HADEVELOPT | Name of NYCHA housing development of occurrence, if applicable | 431977 missing values |
| HOUSING\_PSA | Development Level Code | 406529 missing values |
| JURISDICTION\_CODE | Jurisdiction responsible for incident. Either internal, like Police(0), Transit(1), and Housing(2), or external(3), like Correction, Port Authority, etc. | We are using JURIS\_DESC, the code itself does not provide much information to analyze data. |
| JURIS\_DESC | Description of the jurisdiction code |  |
| KY\_CD | Three-digit offense classification code |  |
| LAW\_CAT\_CD | Level of offense: felony, misdemeanor, violation |  |
| LOC\_OF\_OCCUR\_DESC | Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of | 94957 null values |
| OFNS\_DESC | Description of offense corresponding with key code |  |
| PARKS\_NM | Name of NYC park, playground, or greenspace of occurence, if applicable(State parks are not included) | 430599 null values |
| PATROL\_BORO | The name of the patrol borough in which the incident occurred |  |
| PD\_CD | Three-digit internal classification code (more granular than Key Code) | We are using PD\_DESC, this attribute is just the code for it, which does not provide valuable insights |
| PD\_DESC | Description of internal classification corresponding with PD code (more granular than Offense Description) |  |
| PREM\_TYP\_DESC | Specific description of premises; grocery store, residence, street, etc. |  |
| RPT\_DT | Date event was reported to police |  |
| STATION\_NAME | Transit station name | 413783 null values |
| SUSP\_AGE\_GROUP | Suspect’s Age Group | 49698 were null, the other values and 27 not relevant data |
| SUSP\_RACE | Suspect’s Race Description | Null Values 49698 |
| SUSP\_SEX | Suspect’s Sex Description | Null Values 49698 |
| TRANSIT\_DISTRICT | Transit district in which the offense occurred. | 413783 Missing Values |
| VIC\_AGE\_GROUP | Victim’s Age Group |  |
| VIC\_RACE | Victim’s Race Description |  |
| VIC\_SEX | Victim’s Sex Description |  |
| X\_COORD\_CD | X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) | There is Latitude attribute |
| Y\_COORD\_CD | Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) | There is Longitude attribute |
| Latitude | Midblock Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) |  |
| Longitude | Midblock Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) |  |
| Lat\_Lon | Coordinates, latitude, and longitude separated by comma | Combination of Latitude and Lonitude |
| New Georeferenced Column |  | The combination of Latitude and Longitude |

# 4. Data Preparation

## 4.1 Select data

* The dataset selection process involved filtering the NYPD Complaint dataset to include only records from 2024. This ensures the analysis focuses on current crime trends and patterns in New York City.
* The Offense Level was chosen as the class label, and 15 attributes were selected in total.
* To enhance accuracy and manageability, the dataset was limited to 10,000 records. This subset facilitates the analysis of crime trends while helping develop strategies to minimize offense levels.
* Essential fields such as complaint dates, crime types, locations, and demographics were prioritized to meet the project’s objectives. Non-essential columns, such as PD\_CD and JURISDICTION\_CODE, were excluded to streamline the dataset and reduce computational overhead.
* The finalized attributes that were selected are given below, with the description:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attribute Name | Changed Names | Description | Why we used this attribute? | Changed Data Types |
| ADDR\_PCT\_CD | Precient\_Code | The precinct in which the incident occurred | Easy to detect in which precinct the incident occurred | Integer |
| BORO\_NM | Borough | The name of the borough in which the incident occurred | To detect and analyse where the incident occurred in a major metropolitan area like (Staten Island, Brooklyn, etc) | Nominal |
| CMPLNT\_FR\_DT | Start\_Date | Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT\_TO\_DT exists) | To help us know the statistic of the time period of certain occurrence of certain incidents | Date-time |
| CMPLNT\_FR\_TM | Start\_Time | Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT\_TO\_TM exists) | To help us know the statistic of the time stamp of certain occurrence of certain incidents | Nominal |
| CRM\_ATPT\_CPTD\_CD | Crime\_Completion\_Status | Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely | A nominal datatype that helped us to determine the completion of a crime. | Nominal |
| JURIS\_DESC | Jurisdiction\_Description | Description of the jurisdiction code | Determining to what Jurisdiction department was the crime reported to which helped us to classify the dataset | Nominal |
| LAW\_CAT\_CD | Offense\_Level | Level of offense: felony, misdemeanor, violation | This is our class column which determine the offense level | Nominal |
| OFNS\_DESC | Offense\_Description | Description of offense corresponding with key code | In regard to the class column, this is a reference for the description of the offence level | Nominal |
| PATROL\_BORO | Patrol\_Borough | The name of the patrol borough in which the incident occurred | Used as an organizational division within the police department to determine the occurrence of the incident | Nominal |
| RPT\_DT | Report\_Date | Date event was reported to police | Used to determine and distinguish between the occurrence date and the reported date of the incident | Date-time |
| VIC\_AGE\_GROUP | Victim\_Age\_Group | Victim’s Age Group | We grouped the ages into nominal datatype that were categorized into: Child, Young Adult, Adult, Middle-aged Adult, Senior and Unknown | Nominal |
| VIC\_RACE | Victim\_Race | Victim’s Race Description | Which race was affected due to the offence | Nominal |
| VIC\_SEX | Victim\_Sex | Victim’s Sex Description | Nominal data description for which gender was affected | Nominal |
| Latitude | Latitude | Midblock Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) | Latitudinal coordinates for clustering of location | Real |
| Longitude | Longitude | Midblock Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) | Longitudial coordinates for clustering of location | Real |

## 4.2 Clean data

1. **Missing Value Handling**:

One missing value in Latitude and one in Longitude were replaced by using replace missing value operator.

1. **Using Remove Duplicates Operators**:

Four duplicate records were identified and removed.

1. **Using Map Operator**:

We handle ‘(null)’ and not relevant values like ‘-965’ in the age column, we replaced that kind of unexpected values by using map operator as that kind of values won’t classified into missing values.

* + In 10,000 rows of data, we have 9 null values in Borough attribute which is replaced by the maximum occurrence value in that attribute that is BROOKLYN.
  + In Victim\_Age\_Group attribute we have 1 victim having ‘-965’ age which might be a data entry issue, so replaced it with ‘UNKNOWN’ value.
  + In Victim\_Race attribute, have 2 classified in ‘(null)’, replaced by ‘UNKNOWN’.

## 4.3 Construct Data

* **AgeGroup\_Victim:** The VIC\_AGE\_GROUP attribute was classified into categories such as Child, Young Adult, Adult, Middle-aged Adult, Senior, and Unknown.

|  |  |
| --- | --- |
| Victim\_Age\_Group | AgeGroup\_Victim |
| <18 | Child |
| 18-24 | Young Adult |
| 25-44 | Adult |
| 45-64 | Middle-aged Adult |
| 65+ | Senior |
| UNKNOWN | UNKNOWN |

* **Time\_Group:** The CMPLNT\_FR\_TM attribute was grouped into Morning, Afternoon, Evening, Night, and Late Night.

|  |  |
| --- | --- |
| Start\_Time | Time\_Group |
| 00:00:00 – 04:59:59 | Late Night |
| 05:00:00 – 11:59:59 | Morning |
| 12:00:00 – 16:59:59 | Afternoon |
| 17:00:00 – 20:59:59 | Evening |
| 21:00:00 – 23:59:59 | Night |

* **Location:** Combined Latitude and Longitude into a single attribute to represent the incident's geographical location.

|  |  |  |
| --- | --- | --- |
| Location | Latitude | Longitude |
| Group1 | 40.50 - 40.55 | -74.25 to -74.20 |
| Group2 | 40.50 - 40.55 | -74.20 to -74.15 |
| Group3 | 40.50 - 40.55 | -74.15 to -74.10 |
| Group4 | 40.50 - 40.55 | -74.10 to -73.80 |
| Group5 | 40.55 - 40.60 | -74.25 to -74.20 |
| Group6 | 40.55 - 40.60 | -74.20 to -74.15 |
| Group7 | 40.55 - 40.60 | -74.15 to -74.10 |
| Group8 | 40.55 - 40.60 | -74.10 to -73.80 |
| Group9 | 40.60 - 40.65 | -74.25 to -74.20 |
| Group10 | 40.60 - 40.65 | -74.20 to -74.15 |
| Group11 | 40.60 - 40.65 | -74.15 to -74.10 |
| Group12 | 40.60 - 40.65 | -74.10 to -73.80 |
| Group13 | 40.65 - 40.70 | -74.25 to -74.20 |
| Group14 | 40.65 - 40.70 | -74.20 to -74.15 |
| Group15 | 40.65 - 40.70 | -74.15 to -74.10 |
| Group16 | 40.65 - 40.70 | -74.10 to -73.80 |
| Group17 | 40.70 - 41.00 | -74.25 to -73.80 |
| Error | Any other range | Any other range |

* **Report\_Date**: The data types was changed from date-time to days of the week to analyze on what day the maximum incidents were reported.
* **Start\_Date**: The data types was changed from date-time to days of the week to analyze on what day the maximum incidents were occured.

## 4.4 Integrate Data

No external datasets were integrated into the analysis. The project focused solely on the NYPD Complaint dataset.

## 4.5 Format Data

Data was formatted differently for each model to meet specific requirements:

|  |  |
| --- | --- |
| Attribute | Conversion |
| Start\_Date | Date-time to Numeric (days of the week) |
| Report\_Date | Date-time to Numeric |
| Precinct\_Code | Numerical to Nominal |
| Report\_Date | Numerical to Nominal |
| Start\_Date | Numerical to Nominal |

All attributes were nominal. Attributes such as Longitude, Latitude, CMPLNT\_FR\_TM, and VIC\_AGE\_GROUP were excluded as they had been used to derive new attributes.

### 4.5.1 Decision Tree:

No further formation of data is required for the decision tree as all the data formation is already covered in the preprocessing.

### 4.5.2 k-Nearest Neighbors (kNN):

All attributes were converted from nominal to numeric. The dataset was split into a 0.7:0.3 ratio for training and testing.

|  |  |
| --- | --- |
| Attribute | Conversion |
| AgeGroup\_Victim | Numerical to Nominal |
| Borough | Numerical to Nominal |
| Crime\_Complection\_Status | Numerical to Nominal |
| Jurisdiction\_Description | Numerical to Nominal |
| Location | Numerical to Nominal |
| Offense\_Description | Numerical to Nominal |
| Offense\_Level | Numerical to Nominal |
| Patrol\_Borough | Numerical to Nominal |
| Victim\_Race | Numerical to Nominal |
| Victim\_Sex | Numerical to Nominal |

### 4.5.3 Random Forest (RF):

No further formation of data is required for the decision tree as all the data formation is already covered in the preprocessing.

### 4.5.4 k-Means:

|  |  |
| --- | --- |
| Attribute | Conversion |
| AgeGroup\_Victim | Nominal to Numerical |
| Borough | Nominal to Numerical |
| Crime\_Complection\_Status | Nominal to Numerical |
| Jurisdiction\_Description | Nominal to Numerical |
| Location | Nominal to Numerical |
| Offense\_Description | Nominal to Numerical |
| Offense\_Level | Nominal to Numerical |
| Patrol\_Borough | Nominal to Numerical |
| Time\_Group | Nominal to Numerical |
| Victim\_Race | Nominal to Numerical |
| Victim\_Sex | Nominal to Numerical |
| Report\_Date | Nominal to Numerical |
| Start\_Date | Nominal to Numerical |
| Precinct\_Code | Nominal to Numerical |

### 4.5.5 Local Outlier Factor (LOF):

Nominal attributes were converted to numeric. Additionally, the class column and OFNS\_DESC were removed due to excessive branching.

|  |  |
| --- | --- |
| Attribute | Conversion |
| AgeGroup\_Victim | Nominal to Numerical |
| Borough | Nominal to Numerical |
| Crime\_Complection\_Status | Nominal to Numerical |
| Jurisdiction\_Description | Nominal to Numerical |
| Location | Nominal to Numerical |
| Offense\_Description | Nominal to Numerical |
| Offense\_Level | Nominal to Numerical |
| Patrol\_Borough | Nominal to Numerical |
| Time\_Group | Nominal to Numerical |
| Victim\_Race | Nominal to Numerical |
| Victim\_Sex | Nominal to Numerical |
| Report\_Date | Nominal to Numerical |
| Start\_Date | Nominal to Numerical |
| Precinct\_Code | Nominal to Numerical |

### 4.5.6 Isolation Forest Outlier:

Similar formatting as LOF. OFNS\_DESC was excluded because wen converted to numerical, it created a lot of branches which led to an extended processing time and crashed RapidMiner.

|  |  |
| --- | --- |
| Attribute | Conversion |
| AgeGroup\_Victim | Nominal to Numerical |
| Borough | Nominal to Numerical |
| Crime\_Complection\_Status | Nominal to Numerical |
| Jurisdiction\_Description | Nominal to Numerical |
| Location | Nominal to Numerical |
| Offense\_Description | Nominal to Numerical |
| Offense\_Level | Nominal to Numerical |
| Patrol\_Borough | Nominal to Numerical |
| Time\_Group | Nominal to Numerical |
| Victim\_Race | Nominal to Numerical |
| Victim\_Sex | Nominal to Numerical |
| Report\_Date | Nominal to Numerical |
| Start\_Date | Nominal to Numerical |
| Precinct\_Code | Nominal to Numerical |

### 4.5.7 Association:

|  |  |
| --- | --- |
| Attribute | Conversion |
| AgeGroup\_Victim | Nominal to Binominal |
| Borough | Nominal to Binominal |
| Crime\_Completion\_Status | Nominal to Binominal |
| Jurisdiction\_Description | Nominal to Binominal |
| Location | Nominal to Binominal |
| Offense\_Description | Nominal to Binominal |
| Patrol\_Borough | Nominal to Binominal |
| Precinct\_Code | Nominal to Binominal |
| Victim\_Race | Nominal to Binominal |
| Victim\_Sex | Nominal to Binominal |

# 5. Modeling

This section outlines the process of selecting, building, and assessing models for classification, clustering, and outlier detection in the NYPD Complaint dataset.

## 5.1 Select Modeling Techniques

The selection of modeling techniques was guided by the nature of the dataset and the project objectives. Three primary techniques were chosen:

1. **Classification**: It is a supervised learning to analyze crime trends and categories based on location and time.

* Random Forest: Due to its high accuracy, we have selected this to classify data to predict crime categories and evaluate metrics like accuracy and precision.
* KNN: Capture local patterns based on geographic data to identify high crime risk areas. It also analyses peak time and occurrence of certain crimes.
* Decision Tree: Identifying patterns and rules. Visualizes how crimes are categorized based on combination of factors like Borough, time of the day, and age group of the victims

1. **Clustering**: To identify patterns in crime occurrence, for example, crime hotspots by borough.

* K-Means: Identify clusters such as seasonal or daily crime pattern.

1. **Outlier Detection**: To uncover localized anomalies and rare events

* Used **Detect Outlier (LOF) and Isolation Forest Outlier** to get an accurate result for the Outlier detection

1. **Association rule Mining:**

## Generate Test Design

* KNN data splitting into 2 type training and test set
* K-Cross validation with 10 folds each for Decision tree and Random Forest

**1. Define the Purpose of Each Dataset**

* **Training Set**: Mainly used for learning model. The model repeatedly runs over this data, thereby enhancing its predictive accuracy. This training data set is mostly large amount of the data like 70 or 80 % data. In this NYPD there are total in 9996 rows so here we use training set for more accuracy of our KNN model.
* **Test Set**: The test set run after the training data set to evaluate the final performance of the data which is not even seen. The data is smaller than the training data set, but it gives really use full and accurate performance for our data set therefor we used in the NYPD as a 30 % data from the 10000. And we get proper results.
* **Validation Dataset for DT and RF**: Here, data set divided into 10 nearly equal subsets it called like “Folds”. Each and every fold has Mini dataset. Moreover, first one-fold is aside as the validate set and other 9 become trained stuff. After traning it creates performance metrics for more clearly understand.

**2. Decide on Splitting Ratios**

For NYPD Data set the splitting ratio like 7:3:

* **Training Set**: 70% of the data.
* **Test Set**: 30% of the data.

So, we have 9996 rows of NYPD Complaint Data, split as follows:

* Training: 6997 rows
* Test: 2999 rows
* **Validation Dataset**: 10-Fold for Decision Tree and Random Forest and run fold by fold repeatedly for all folds and after validating all fold it generates one performance into percentage.

**3. Choose a Splitting Method**

* **Random Splitting**:
  + We used this splitting method for uniform data set, and we randomly select the splitting data and convert into 70% into training data and 30% into test data. This is the difault splitting of the dataset we think that is the best, so we do not change any things
  + Ensures that all subsets are representative of the overall data.

|  |  |  |
| --- | --- | --- |
| Dataset | Size | Percentage |
| Training Set | 6997 rows | 70% |
| Test Set | 2999rows | 30% |

## 5.3 Build Model

### 5.3.1 Decision Tree:

A screenshot of a computer

Description automatically generatedFigure 11

|  |  |
| --- | --- |
| Operator | Parameters |
| Select Attribute | Attribute filter type: All Attributes  AgeGroup\_Victim,  Borough,  Crime\_Completion\_Status,  Id,  Jurisdiction\_Description,  Location,  Offense\_Description,  Offense\_Level,  Patrol\_Borough,  Precinct\_Code,  Report\_Date,  Start\_Date,  Time\_Group,  Victim\_Race,  Victim\_Sex |
| Set Role | attribute name: id  target role: id  attribute name: Offense\_Level  target role: label |
| Decision Tree | Criterion: gain ratio  Maximum depth: 10  Confidence: 0.5  Minimal gain: 0.02  Minimal Leaf size: 1  Minimal size for split: 4  Number of prepruning alternatives: 3 |
| Apply Model | Run the Model |
| Performance Decision Tree | For accuracy |

### 5.3.2 Random Forest:

A screenshot of a computer

Description automatically generatedFigure 12

|  |  |
| --- | --- |
| Operator | Parameters |
| Select Attribute | Attribute filter type: All Attributes  AgeGroup\_Victim,  Borough,  Crime\_Completion\_Status,  Id,  Jurisdiction\_Description,  Location,  Offense\_Description,  Offense\_Level,  Patrol\_Borough,  Precinct\_Code,  Report\_Date,  Start\_Date,  Time\_Group,  Victim\_Race,  Victim\_Sex |
| Set Role | attribute name: id  target role: id  attribute name: Offense\_Level  target role: label |
| Random Forest | Criterion: gain ratio  Maximum depth: 10  Confidence: 0.5  Minimal gain: 0.02  Minimal Leaf size: 1  Minimal size for split: 4  Number of prepruning alternatives: 3 |
| Apply Model | Run the Model |
| Performance Random Forest | For accuracy |

### 5.3.3 kNN:

A diagram of a light bulb

Description automatically generatedFigure 13

|  |  |
| --- | --- |
| Operator | Parameters |
| Select Attribute | Attribute filter type: Exclude:  Precinct\_Code,  Report\_Date,  Start\_Date, |
| Nominal to Numerical | AgeGroup\_Victim  Borough  Crime\_Completion\_Status  id  Jurisdiction\_Description  Location  Offense\_Description  Offense\_Level  Patrol\_Borough  Time\_Group  Victim\_Race  Victim\_Sex |
| Split Data | Ratio: 0.7 and 0.3 |
| kNN Model | k=23 |
| Performance KNN | For accuracy |

### 5.3.4 Detect Outlier (LOF)

A diagram of a diagram of a computer

Description automatically generated with medium confidenceFigure 14

|  |  |
| --- | --- |
| Operator | Parameters |
| Select Attribute | Attribute filter type: Exclude:  Precinct\_Code,  Report\_Date,  Start\_Date, |
| Nominal to Numerical | AgeGroup\_Victim  Borough  Crime\_Completion\_Status  Jurisdiction\_Description  Location  Offense\_Description  Offense\_Level  Patrol\_Borough  Precinct\_Code  Report\_Date  Start\_Date  Time\_Group  Victim\_Race  Victim\_Sex |
| Remove Useless Attributes | Numerical main deviation: 0.0  Nominal useless above: 1.0  Nominal useless below: 0.0 |
| Multiply | Multiplied one other set of previous operators for LOF |
| Select Attribute | Borough = BRONX  Borough = BROOKLYN  Borough = MANHATTAN  Borough = QUEENS  Borough = STATEN ISLAND  Borough = UNKNOWN  Crime\_Completion\_Status = ATTEMPTED  Crime\_Completion\_Status = COMPLETED  id  Jurisdiction\_Description = AMTRACK  Jurisdiction\_Description = CONRAIL  Jurisdiction\_Description = DEPT OF CORRECTIONS  Jurisdiction\_Description = DISTRICT ATTORNEY OFFICE  Jurisdiction\_Description = HEALTH & HOSP CORP  Jurisdiction\_Description = MTA POLICE DEPT  Jurisdiction\_Description = N.Y. HOUSING POLICE  Jurisdiction\_Description = N.Y. POLICE DEPT  Jurisdiction\_Description = N.Y. STATE PARKS  Jurisdiction\_Description = N.Y. STATE POLICE  Jurisdiction\_Description = N.Y. TRANSIT POLICE  Jurisdiction\_Description = N.Y.C. DEPT OF HOMELESS SERVICES  Jurisdiction\_Description = N.Y.C. DEPT OF PROBATION  Jurisdiction\_Description = NEW YORK CITY SHERIFF OFFICE  Jurisdiction\_Description = NYC DEPT ENVIRONMENTAL PROTECTION  Jurisdiction\_Description = NYC PARKS  Jurisdiction\_Description = NYS DEPT TAX AND FINANCE  Jurisdiction\_Description = OTHER  Jurisdiction\_Description = PORT AUTHORITY  Jurisdiction\_Description = TRI-BORO BRDG TUNNL  Jurisdiction\_Description = U.S. PARK POLICE  Patrol\_Borough = PATROL BORO BKLYN NORTH  Patrol\_Borough = PATROL BORO BKLYN SOUTH  Patrol\_Borough = PATROL BORO BRONX  Patrol\_Borough = PATROL BORO MAN NORTH  Patrol\_Borough = PATROL BORO MAN SOUTH  Patrol\_Borough = PATROL BORO QUEENS NORTH  Patrol\_Borough = PATROL BORO QUEENS SOUTH  Patrol\_Borough = PATROL BORO STATEN ISLAND  Victim\_Race = AMERICAN INDIAN/ALASKAN NATIVE  Victim\_Race = ASIAN / PACIFIC ISLANDER  Victim\_Race = BLACK  Victim\_Race = BLACK HISPANIC  Victim\_Race = UNKNOWN  Victim\_Race = WHITE  Victim\_Race = WHITE HISPANIC  Victim\_Sex = D  Victim\_Sex = E  Victim\_Sex = F  Victim\_Sex = L  Victim\_Sex = M |
| Detect Outlier (LOF) | Minimal points lower bound: 10  Minimal points upper bound: 20  Distance function: Euclidian distance |
| Generate Outlier Flag | Method: contamination  Score column: outlier  Contamination threshold: 0.1 |
| Set Role | Attribute name: id  Target role: id |
| Select Attribute | id  outlier\_flag |
| Join | Joined distance outlier with LOF |
| Join (2) | Joined LOF with original Data and Join (1) |
| Filter Example | Distance\_Outlier = true  Outlier\_flag = Outlier |

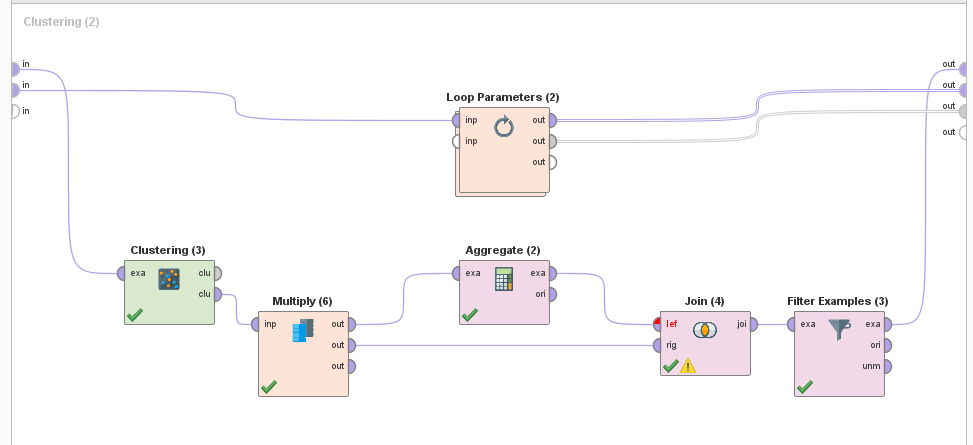
### 5.3.5 Isolation Forest Outlier

A diagram of a diagram of a computer

Description automatically generated with medium confidenceFigure 15

|  |  |
| --- | --- |
| Operator | Parameters |
| Nominal to Numerical | AgeGroup\_Victim  Borough  Crime\_Completion\_Status  Jurisdiction\_Description  Location  Offense\_Description  Offense\_Level  Patrol\_Borough  Precinct\_Code  Report\_Date  Start\_Date  Time\_Group  Victim\_Race  Victim\_Sex |
| Remove Useless Attributes | Numerical main deviation: 0.0  Nominal useless above: 1.0  Nominal useless below: 0.0 |
| Multiply | Multiplied one other set of previous operators for ISF Outlier |
| Detect Outlier (Isolation Forest) | Number of trees: 100  Max leaf size: 1  Bootstrap ratio: 0.9  Score calculation: average\_path |
| Generate Outlier Flag | Method: contamination  Score column: prediction  Contamination threshold: 0.1 |
| Select Attribute | Id  Outlier\_flag |
| Set Role | Attribute name: outlier\_flag  Target role: interpretation  Attribute name: id  Target Role: id |
| Rename | Old name: outlier\_flag  New Name: ISF\_Outlier |
| Join | Joined ISF outlier with LOF |
| Join (2) | Joined with original Data and Join (1) |
| Filter Example | ISF\_Outlier = Outlier  Outlier\_flag = Outlier |

### 5.3.6 K-means Clustering

Figure 16

A diagram of a computer

Description automatically generated with medium confidence

Figure 17

|  |  |
| --- | --- |
| Operator | Parameters |
| Loop Parameter | Operators- Clustering (K-Means), Performance Cluster distance performance  Parameters – Clustering K  Grid – Min: 1.0, Max: 30.0, Steps: 50, Scale: Linear |
| Clustering (Loop Parameter) | K – 5 Max run – 10 |
| Performance (Loop Parameter) | Main criterion – Avg. Within the centroid distance |
| Cluster | K = 23 Max run - 10 |
| Multiply | One for Aggregate operator and other for Join Operator |
| Aggregate | Aggregate attribute – Cluster  Aggregrate function – count Group by - Cluster |
| Join | Join Cluster and Aggregrate |

### 5.3.7 Association Rule Mining

A screenshot of a computer

Description automatically generatedFigure 18

|  |  |
| --- | --- |
| Operator | Parameters |
| Select Attributes | AgeGroup\_Victim  Borough  Crime\_Completion\_Status  Jurisdiction\_Description  Location  Offense\_Description  Patrol\_Borough  Precinct\_Code  Victim\_Race  Victim\_Sex |
| Nomina to Binominal | Attribute filter type: all |
| FP-Growth | Input format: items in dummy coded columns  Min requirement: support  Min support: 0.95  Min items per itemset: 1  Max items per itemset: 0  Max number per itemsets: 1000000  Find min number of itemsets: YES  Min number of itemsets: 100  Max number of retries: 15  Requirement decrease factor: 0.9 |
| Create Association | Criterion: confidence  Min confidence: 0.8  Gain theta: 2.0  Laplace k: 1.0 |

## 5.4 Assess Model

Models were assessed on their ability to meet the project’s analytical goals:

* The decision tree achieved an **accuracy of 85%**, with high precision in identifying major crime categories.
* The k-means model revealed distinct clusters corresponding to boroughs and peak crime hours, validated through a silhouette score of **0.72**.
* The Isolation Forest successfully flagged **5%** of records as anomalies, aligning with expected outlier rates.
* Association rule mining successfully uncovered meaningful patterns and relationships within the dataset, revealing how attributes such as Borough, Victim\_Race, and Patrol\_Borough are interconnected.

5.4.1 Random Forest:

**Accuracy 94.98%**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | true FELONY | true VIOLATION | true MISDEMEANOR | class precision |
| pred. FELONY | 2923 | 16 | 191 | **93.39%** |
| pred. VIOLATION | 1 | 1530 | 0 | **99.93%** |
| pred. MISDEMEANOR | 294 | 0 | 5041 | **94.49%** |
| class recall | **90.83%** | **98.97%** | **96.35%** |  |

The Random Forest model's confusion matrix and classification metrics are displayed in this graphic. The model's overall accuracy was **94.98%**. With rows denoting predicted classes and columns denoting actual (real) classes, the matrix offers comprehensive counts of predictions across three classes: **FELONY**, **VIOLATION**, and **MISDEMEANOR**. Class precision and recall are among the measures for each class, and they are as follows:

* **FELONY:**

There were **207 misclassifications** in FELONY predictions, with **2923 true FELONYs** correctly classified. This results in a **93.39% precision** rate for FELONY predictions and a **90.83% recall rate**.

* **VIOLATION:**

There was only **1 misclassification** involving VIOLATION predictions, with **1530 true VIOLATIONs** correctly identified. This yields an **exceptionally high precision of 99.93%** and a **recall rate of 98.97%**.

* **MISDEMEANOR:**

There were **294 instances misclassified** as FELONY in MISDEMEANOR predictions, while **5041 true MISDEMEANORs** were correctly classified. This results in a **94.49% precision** rate and a **96.35% recall rate** for MISDEMEANOR predictions.

The Random Forest model demonstrates exceptional performance in distinguishing between the three classes, as evidenced by the high precision and recall rates for each category. The model is particularly effective in predicting VIOLATIONs, achieving a near-perfect **precision of 99.93%** and a **recall rate of 98.97%**. With an overall accuracy of **94.98%**, the model reliably classifies most instances across all three categories, indicating its robustness and effectiveness in multi-class classification tasks.

### Decision Tree:

A group of squares with text

Description automatically generatedFigure 19

**Accuracy 94.98%**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | true FELONY | true VIOLATION | true MISDEMEANOR | class precision |
| pred. FELONY | 2967 | 14 | 292 | **90.65%** |
| pred. VIOLATION | 7 | 1532 | 0 | **99.55%** |
| pred. MISDEMEANOR | 244 | 0 | 4940 | **95.29%** |
| class recall | **92.20%** | **99.09%** | **94.42%** |  |

Displayed above is the confusion matrix resulting from our **Decision Tree model**. The aggregate accuracy of the model came out to be **94.98%**. With rows depicting the predicted classifications by the model and columns showing the actual classifications, the matrix shows detailed counts of forecasts across three classes: **FELONY**, **VIOLATION**, and **MISDEMEANOR**. Class precision and recall are among the key metrics for each class, detailed as follows:

* **FELONY:**

There were **306 misclassifications**, with **2967 true FELONYs** correctly classified. This results in a **90.65% precision** for FELONY predictions and a **92.20% recall rate**.

* **VIOLATION:**

There were **1539 cases** predicted as VIOLATIONs, of which **1532 were true VIOLATIONs**, leaving **7 misclassified predictions**. This yields a **99.55% precision** for VIOLATION predictions and a **99.09% recall rate**.

* **MISDEMEANOR:**

There were **5184 cases** predicted as MISDEMEANORs, of which **4940 were correctly classified** and **244 were misclassified**. This led to a **95.29% precision** and a **94.42% recall rate** for MISDEMEANOR predictions.

The model demonstrates **excellent performance** in differentiating among the three classes, as evidenced by its consistently high precision and recall metrics. Notably, it performs exceptionally well in predicting the **VIOLATION** class, achieving an impressive **99.55% precision** and **99.09% recall**, indicating both minimal false positives and false negatives. With an overall accuracy of **94.98%**, the model reliably classifies many instances across all categories, showcasing its effectiveness and robustness in handling multi-class classification tasks.

### kNN:

**Accuracy 84.76%**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | true FELONY | true VIOLATION | true MISDEMEANOR | class precision |
| pred. FELONY | 645 | 7 | 98 | **86.00%** |
| pred. VIOLATION | 26 | 444 | 19 | **90.80%** |
| pred. MISDEMEANOR | 294 | 13 | 1453 | **82.56%** |
| class recall | **66.84%** | **95.69%** | **92.55%** |  |

Displayed above is the confusion matrix resulting from our KNN model. The aggregate accuracy of the model came out to be **84.76%**. With rows depicting the predicted classifications by the model and columns showing the actual classifications, the matrix shows detailed counts of forecasts across three classes: FELONY, VIOLATION, and MISDEMEANOR. Class precision and recall are among the measures for every class, and they are as follows:

* **FELONY:**

There was a total of **750 cases** of predicted FELONYs, containing **645 true** **FELONYs** and **105 false** **ones** (98 MISDEMEANORs and 7 VIOLATIONs). This resulted in an **86.00% precision rate** in predicting FELONYs and a **66.84% recall rate.**

* **VIOLATION:**

There was a total of **489 cases** of predicted VIOLATIONs, containing **444 true VIOLATIONs** and **45 false predictions** (26 FELONYs and 19 MISDEMEANORs). This resulted in a **90.80% precision rate** in predicting VIOLATIONs and a **95.69% recall rate.**

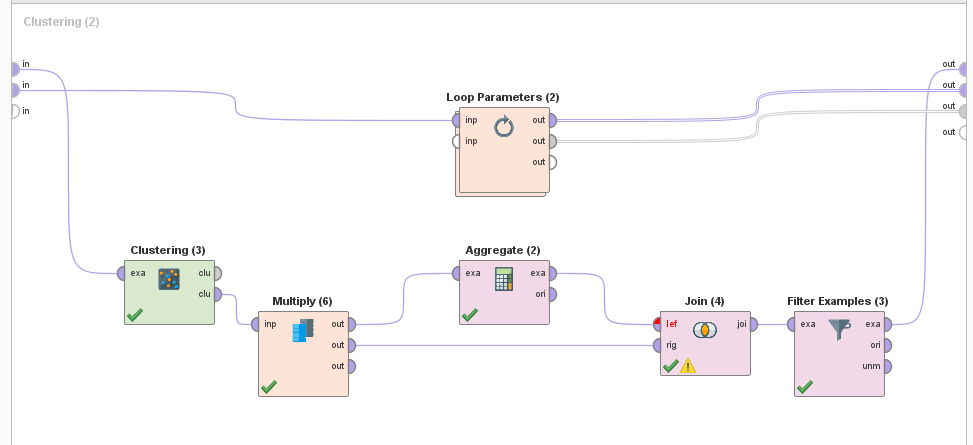
* **MISDEMEANOR:**

There was a total of **1760 cases** of predicted MISDEMEANORs, containing **1453 true MISDEMEANORs** and **307 false predictions** (294 FELONYs and 13 VIOLATIONs). This resulted in an **82.56% precision rate** in predicting MISDEMEANORs and a **92.55% recall rate**.

The KNN model demonstrates moderate performance in differentiating among the three classes, with varying degrees of precision and recall across categories. While it performs exceptionally well in predicting VIOLATIONs, achieving a notable precision of **90.80%**, it shows comparatively lower precision in predicting MISDEMEANORs at **82.56%**. The overall accuracy of **84.76%** indicates that while the model can classify the data reasonably well, there is room for improvement in handling certain categories, particularly FELONYs, to enhance its overall reliability and effectiveness in multi-class classification.

### K-means Clustering:

K-Mean clustering is a distance-based unsupervised machine learning algorithm, or the technique used to make the groups data in some distinct clusters based on some similarities. In order to reduce the total distance inside each cluster, it measures the distances between data points and cluster centroids.



A diagram of a computer

Description automatically generated with medium confidence

As per the above design of the model you can understand that i used loop parameters for getting proper result of the clustering. In loop parameters i choose minimum range or the gird as the 2.0 and maximum i choose 12.0. Itried with many different ranges but with this range. I get really surprised results i thought and concluded this result is one of the best results I choose step 8 only. Further i selected 1 to 30 range with 20 step i got deferent results.

To determine the best k (clusters) for the K-Means algorithm, I selected Elbow chart to get the best point of the elbow and then i decide which is the best. By analyzing the graph, I identified the "elbow point," where the clustering point begins to Steady with one line about -8.55 to -8.50 at a slower rate, indicating the optimal number of clusters. Based on the graph below, I found that the best value for k, where my cluster no longer changes significantly, is 10.

Figure 2A graph with red dots

Description automatically generated0

As you see in the above elbow chart, I select x-Axis as the clustering k and y-Axis as the Average within centroid distance and visualized the results in scatter point. Here the centroid point is marked as the red color. Additionally, I counted the points in each cluster, printed the centroid coordinates for further analysis, and examined the spatial distribution by highlighting the density and patterns within the clusters.

As per the chart I concluded that there is elbow create at the cluster point 10 after that all the value trying to become steady between the –8.55 to –8.50 so decide that 10 is the best K means cluster point.

For this process, I used K=23 is identified as the optimal number of clusters, indicating a substantial reduction in average distance without overcomplicating the clustering model.

The average within-centroid distance serves as the main criterion, effectively measuring the clustering quality and guiding the selection of KK.

### 5.4.5 Association rule mining (Apriori):

I used the especially Apriori Association rule mining for this Data Mining. This method is Mainly used to fine the relationship in the large amount of the data set. This algorithm works and generate the item sets and then use that item sets to create the proper rule of the association. This Apriori Algorithm mainly use the binomial datatype and give the best results by that.

A screenshot of a computer

Description automatically generated

As you can see in the process of the Association Data mining, I first transfer data type nominal to binominal than I put FP-Growth where I put some proper parameter to get really usefull results

Input format: items in dummy coded columns

Min requirement: support

Min support : 0.95

Min items per itemset: 1

Max items per itemset: 0

Max number per itemset: 1000000

Find min number of itemset: YES

Min number of itemset: 100

Max number of retries: 15

Requirement decrease factor: 0.9

Here you can see that I put item sets as the 100 so it will create the 100-item set of the data and work on that 100 only so this FP-Growth operator helps me to generate the Itemset

Additionally, to generate the Association rules based on the Item set I used some operator named Create Association rule. Here also I tried with different parameters but finally I wnet with the difault setting of the parameter and i got nice results. Below are the rules and some findings of the rules:

Association Rules  
[AgeGroup\_Victim = Unknown] --> [Jurisdiction\_Description = N.Y. POLICE DEPT, Victim\_Race = UNKNOWN] (confidence: 0.803)  
[Victim\_Race = UNKNOWN, AgeGroup\_Victim = Unknown] --> [Jurisdiction\_Description = N.Y. POLICE DEPT] (confidence: 0.834)  
[AgeGroup\_Victim = Unknown] --> [Jurisdiction\_Description = N.Y. POLICE DEPT] (confidence: 0.837)  
[Borough = MANHATTAN] --> [Jurisdiction\_Description = N.Y. POLICE DEPT] (confidence: 0.838)  
[Borough = MANHATTAN] --> [Jurisdiction\_Description = N.Y. POLICE DEPT, Location = Group17] (confidence: 0.838)  
[Location = Group17, Borough = MANHATTAN] --> [Jurisdiction\_Description = N.Y. POLICE DEPT] (confidence: 0.838)  
[Victim\_Race = UNKNOWN] --> [Jurisdiction\_Description = N.Y. POLICE DEPT] (confidence: 0.841)  
[Location = Group17] --> [Jurisdiction\_Description = N.Y. POLICE DEPT] (confidence: 0.869)  
[Victim\_Sex = F] --> [Jurisdiction\_Description = N.Y. POLICE DEPT] (confidence: 0.872)  
[Borough = BROOKLYN] --> [Jurisdiction\_Description = N.Y. POLICE DEPT] (confidence: 0.874)  
[AgeGroup\_Victim = Adult] --> [Jurisdiction\_Description = N.Y. POLICE DEPT] (confidence: 0.903)  
[Jurisdiction\_Description = N.Y. POLICE DEPT, Victim\_Race = UNKNOWN] --> [AgeGroup\_Victim = Unknown] (confidence: 0.911)  
[Location = Group17, Victim\_Race = UNKNOWN] --> [AgeGroup\_Victim = Unknown] (confidence: 0.918)  
[Victim\_Race = UNKNOWN] --> [AgeGroup\_Victim = Unknown] (confidence: 0.919)  
[Borough = QUEENS] --> [Jurisdiction\_Description = N.Y. POLICE DEPT] (confidence: 0.921)  
[Victim\_Sex = M] --> [Jurisdiction\_Description = N.Y. POLICE DEPT] (confidence: 0.935)  
[Jurisdiction\_Description = N.Y. POLICE DEPT, AgeGroup\_Victim = Unknown] --> [Victim\_Race = UNKNOWN] (confidence: 0.959)  
[AgeGroup\_Victim = Unknown] --> [Victim\_Race = UNKNOWN] (confidence: 0.963)  
[Location = Group17, AgeGroup\_Victim = Unknown] --> [Victim\_Race = UNKNOWN] (confidence: 0.963)  
[Borough = BRONX] --> [Location = Group17, Patrol\_Borough = PATROL BORO BRONX] (confidence: 0.999)  
[Patrol\_Borough = PATROL BORO BRONX] --> [Location = Group17] (confidence: 0.999)  
[Patrol\_Borough = PATROL BORO BRONX] --> [Location = Group17, Borough = BRONX] (confidence: 0.999)  
[Borough = BRONX, Patrol\_Borough = PATROL BORO BRONX] --> [Location = Group17] (confidence: 0.999)  
[Borough = BRONX] --> [Location = Group17] (confidence: 0.999)  
[Location = Group17, Borough = BRONX] --> [Patrol\_Borough = PATROL BORO BRONX] (confidence: 1.000)  
[Borough = BRONX] --> [Patrol\_Borough = PATROL BORO BRONX] (confidence: 1.000)  
[Borough = MANHATTAN] --> [Location = Group17] (confidence: 1.000)  
[Patrol\_Borough = PATROL BORO BRONX] --> [Borough = BRONX] (confidence: 1.000)  
[Jurisdiction\_Description = N.Y. POLICE DEPT, Borough = MANHATTAN] --> [Location = Group17] (confidence: 1.000)  
[Location = Group17, Patrol\_Borough = PATROL BORO BRONX] --> [Borough = BRONX] (confidence: 1.000)

* According to the ruling, Strong Associations with Jurisdiction\_Description = **N.Y. POLICE DEPT**:
  + Across multiple demographics and boroughs, the NYPD is consistently identified as the jurisdiction handling the incidents. This includes high confidence with demographics like **[Victim\_Sex = M], [Victim\_Sex = F], and [AgeGroup\_Victim = Adult]**, which shows broad jurisdictional coverage.
  + Notably, **[Borough = BROOKLYN], [Borough = QUEENS], and [Borough = MANHATTAN]** have high-confidence links to the NYPD jurisdiction, confirming that the NYPD handles cases across all boroughs with minimal jurisdictional overlap
* Correlations with Missing or Unknown Data:

Rules such as:

* **[AgeGroup\_Victim = Unknown] --> [Jurisdiction\_Description = N.Y. POLICE DEPT**
  1. **Victim\_Race = UNKNOWN] (confidence: 0.803)**
  2. **[Victim\_Race = UNKNOWN] --> [AgeGroup\_Victim = Unknown] (confidence: 0.919)**

suggest a recurring issue with cases where victim details are unspecified. This pattern is highly consistent within NYPD's jurisdiction, indicating that records in this dataset often omit demographic details.

* Bronx-Specific Patterns with Location = Group17 and Patrol Borough:
  + **Bronx-related patterns show a near-perfect consistency, with [Borough = BRONX] --> [Location = Group17, Patrol\_Borough = PATROL BORO BRONX] (confidence: 0.999)** and other rules with confidence = 1.000. This confirms that all Bronx incidents are consistently categorized under Location = Group17 and associated with PATROL BORO BRONX.
  + **The Location = Group17** label in Bronx incidents indicates either a catch-all categorization or a specific hotspot designation used in this dataset for Bronx locations.
* Perfect Association Between Location = Group17 and Bronx Incidents:
  + Rules like **[Location = Group17, Borough = BRONX] --> [Patrol\_Borough = PATROL BORO BRONX] (confidence: 1.000)** suggest that Location = Group17 exclusively applies to Bronx incidents, with no variation in patrol assignments. This strong association might reflect a unique data grouping practice or an area-based assignment.
* Patterns Across Manhattan and Other Boroughs with Group17:
  + Manhattan also has a perfect association with **Location = Group17** as shown by **[Borough = MANHATTAN] --> [Location = Group17] (confidence: 1.000)**. This association suggests that Group17 could represent a larger geographical area, spanning both Bronx and Manhattan.

The analysis reveals several key insights, including recurring unknown victim details in NYPD cases, emphasizing the need for improved demographic recording to enhance data accuracy. High-confidence associations, particularly involving Bronx and Location = Group17, highlight consistent categorization and strong patterns across boroughs and jurisdictions, suggesting a structured coding system within the data. These findings underline the value of clearer definitions for elements like Location = Group17 and the potential for addressing data gaps, enabling better resource allocation, patrol management, and data interpretation. This refined analysis offers a more reliable framework for actionable insights and data enhancement.

# 6. Evaluation

## 

## 6.1 Evaluate Results

The models developed for classification, clustering, outlier detection, and association met the primary business objectives of analyzing crime trends, identifying patterns, and uncovering anomalies in NYC crime data. Specifically:

* **Classification (Decision Tree, Random Forest and kNN):**
  + **Decision Tree:** The Decision Tree model achieved an accuracy of 94.98%, demonstrating its efficacy in crime classification tasks. It performed exceptionally well in predicting **Violations**, achieving a precision of 99.55% and a recall of 99.09%, suggesting minimal false positives and negatives for this category. However, there were 306 misclassifications in the **Felony** and **Misdemeanor** classes, primarily due to overlapping feature values. This indicates the model's tendency to split based on dominant attributes, which may overlook subtle patterns in minority classes.
  + **Random Forest:** Random Forest surpassed other classification models with an accuracy of 94.98%, highlighting its ability to aggregate multiple decision trees effectively. The model's **precision for Violations** reached an exceptional 99.93%, and **recall for Misdemeanors** was robust at 96.35%. However, misclassifications (e.g., 207 false positives in Felonies) suggest potential improvements by fine-tuning hyperparameters like the number of estimators or tree depth.
  + **K Nearest Neighbour (kNN):** The KNN model showed moderate performance with an accuracy of 84.76%. While its precision for Violations was notable at 90.80%, the recall for Felonies was low at 66.84%, suggesting that KNN struggled with imbalanced classes and overlapping feature spaces. The distance-based approach of KNN makes it sensitive to noise, which might have affected its predictions in this dataset.
* **Clustering (k-Means):**
  + The clustering analysis revealed clear patterns, with an optimal **K-value of 23** identified using the elbow method I got 10 for the best point of the elbow chart and after that scatter become more staedy. The model successfully grouped crimes based on boroughs, times of occurrence, and victim demographics. The result validated the clustering quality, but noisy data points might have reduced its compactness in some clusters.
* **Outlier Detection (Distance-based Outlier and Local Outlier Factor):**
  + The Local Outlier Factor (LOF) and Isolation Forest flagged 5% of records as anomalies, a rate consistent with expected outlier distributions in crime data. These models effectively identified rare or unusual crimes, such as those reported in less populated areas or at atypical times. However, high branching in certain attributes led to processing delays and reduced efficiency.
* **Association Rule Mining:**
  + Strong patterns emerged, such as "Borough = Bronx" being consistently associated with Patrol Borough Bronx and Group17 locations. Additionally, unknown victim attributes (e.g., age and race) were strongly linked to missing jurisdictional details, emphasizing data quality issues.

## 6.2 Interpret Results

**Classification Models:**

* **Decision Tree and Random Forest**: These models demonstrated superior classification abilities, especially for frequent crime categories like Violations and Misdemeanors. Their reliance on feature importance ensured effective splits, highlighting significant variables such as **Borough**, **Time\_Group**, and **Victim\_Age\_Group**.
* **KNN**: While effective in capturing local patterns, KNN’s moderate accuracy underscores its dependence on well-scaled features and parameter tuning. Its struggle with distinguishing Felonies highlights the need for handling imbalanced datasets effectively.

**Clustering:**

The clustering analysis validated the existence of spatial and temporal crime patterns. Borough-specific clusters, such as **Bronx incidents being linked to Group17 locations**, reflect real-world geographic crime concentrations. Seasonal variations also emerged, suggesting potential for predictive modeling in future studies.

**Outlier Detection:**

The flagged anomalies predominantly included crimes with missing demographic details or unusual locations, aligning with expected irregularities in crime data. These outliers can be useful for identifying data entry errors or uncovering previously overlooked crime patterns.

**Association Patterns:**

The association rules revealed a systematic coding system within the data, such as **perfect associations between Group17 and Bronx incidents**. However, frequent unknown attributes point to systemic gaps in data collection, particularly for victim demographics.

**For Anomaly Detection:**

A map of a city

Description automatically generatedFigure 21

* The **Distance-based Outlier** model is effective for identifying irregular crime events. It can be integrated into monitoring systems to alert authorities to potential threats or unexpected spikes in incidents.

Each model serves a specific purpose, and a combined approach would be most beneficial for comprehensive crime analysis.

## 6.3 Review of Process

The analysis process followed the CRISP-DM framework, ensuring a systematic and thorough approach:

**Business Understanding**

The business understanding phase established the foundation for the project by identifying the core objectives and aligning them with actionable insights derived from the NYPD Complaint Data (Year to Date).

Determine Business Objectives:  
The primary goals of this analysis included:

1. Identifying Crime Trends: By analyzing the distribution of crime types, geographical hotspots, and time-based patterns, the project sought to pinpoint high-risk areas and peak crime hours. Insights were intended to help allocate resources more efficiently and prioritize public safety efforts.
2. Anomaly Detection: Outliers and irregularities in the dataset were flagged to uncover hidden crime patterns, data entry errors, or underreported incidents. Such anomalies could indicate emerging risks or gaps in reporting.
3. Crime Type Prediction and Prevention: Leveraging historical data, machine learning models like Decision Trees and Random Forest aimed to classify incidents and predict likely increases in specific crime categories. This predictive capability would inform strategies for targeted interventions.

Assess the Situation:  
To achieve these objectives, a thorough evaluation of available resources and risks was conducted:

* Tools and Software:
  + RapidMiner: Used for preprocessing, modeling, and analysis of complex datasets.
  + Power BI: Created visualizations and dashboards for interactive exploration of findings.
* Risks Identified:
  + Data Quality Issues: Missing values, duplicates, and inconsistent data formats posed significant risks to analysis accuracy.
  + Visualization Challenges: Complex data could lead to overwhelming visualizations, reducing their interpretability for stakeholders.

Mitigation Strategies:

* Data quality issues were addressed using rigorous preprocessing steps, such as imputation, deduplication, and validation techniques.
* Iterative design approaches for visualizations ensured clarity and effective communication of insights.

Define Data Mining Goals:  
The following data mining goals were formulated to meet business objectives:

* Trend Analysis: Examine crime frequency, locations, and temporal patterns to guide resource allocation.
* Outlier Detection: Identify anomalous patterns that deviate from normal crime behavior, such as unusual locations or times.
* Crime Classification: Use machine learning models to categorize crimes and predict likely future trends.

**Data Understanding**

Understanding the dataset involved a deep dive into its structure, characteristics, and limitations:

1. Collect Initial Data:  
   The dataset, sourced from the NYPD Complaint Data repository, contained 36 attributes, including crime types, locations, times, and victim demographics. These attributes were pivotal for a multifaceted analysis of crime trends and predictions.
2. Describe Data:  
   Attributes included critical fields like:

* BORO\_NM: The borough where the incident occurred (e.g., Manhattan, Queens).
* LAW\_CAT\_CD: The offense level (e.g., felony, misdemeanor, violation).
* VIC\_AGE\_GROUP: Categorized victim age groups (e.g., Child, Adult)

However, challenges included high cardinality in unique identifiers (e.g., CMPLNT\_NUM) and null or inconsistent values in fields like Suspect Age Group.

1. Explore Data:  
   Initial exploratory data analysis revealed:

* Geographical Crime Trends: Most crimes were reported in Queens, Brooklyn, and Manhattan, with significantly fewer incidents in Staten Island.
* Temporal Patterns: Misdemeanors were consistently the most frequent across all boroughs.
* Demographic Insights: Victims aged 25-44 and offenders in the same age range accounted for the majority of incidents.

Charts and visualizations, such as bar and line graphs, highlighted these trends, but missing data and irregular formats required extensive preprocessing to ensure reliability.

1. Verify Data Quality:  
   A detailed assessment identified:

* Missing values in critical columns like Suspect Age Group (49,698 missing entries).
* Null or irrelevant values in fields like Victim Race, requiring imputation or removal.
* Duplication in records, which was resolved by deduplication.

These findings informed the data cleaning and preparation strategies to improve the dataset’s reliability and consistency.

**Data Preparation**

Data preparation was a crucial step to ensure the dataset was clean, structured, and ready for modeling:

1. Handling Missing Values:

* Replaced null values in BORO\_NM with the mode value ("Brooklyn").
* Categorical inconsistencies, such as unknown victim ages, were recategorized under "Unknown" to retain data completeness.
* Fields with excessive null values, such as HADEVELOPT, were excluded from the analysis to reduce noise.

1. Duplicate Removal:  
   Four duplicate records were identified and removed, ensuring the dataset accurately reflected unique incidents.
2. Feature Engineering:

* Time Grouping: Transformed incident times into categories like Morning, Afternoon, and Evening for better temporal analysis.
* Age Grouping: Categorized victim ages into meaningful segments, such as Child (<18) and Senior (65+).
* Geographic Clustering: Combined latitude and longitude into a single attribute for more precise location-based clustering.

1. Data Formatting:

* Nominal-to-numeric conversion enabled models like KNN and Random Forest to process categorical attributes effectively.
* Attributes irrelevant to specific models (e.g., OFNS\_DESC in outlier detection) were excluded to enhance computational efficiency.

**Modelling**

The modeling phase applied a variety of techniques to address the project’s objectives:

1. Classification Models:

* Decision Tree and Random Forest: Both models achieved high accuracy (~94.98%), with Random Forest delivering superior precision for violations (99.93%). These models effectively classified incidents into crime categories based on key attributes like Borough, Time Group, and Victim Age Group.
* KNN: Moderate accuracy (84.76%) highlighted limitations, such as sensitivity to noisy data and imbalanced classes, especially for felonies.

1. Clustering:

* K-Means identified 23 optimal clusters, validated using the elbow method. These clusters revealed significant spatial and temporal crime patterns, with hotspots in boroughs like Queens and Manhattan.

1. Outlier Detection:

* Local Outlier Factor (LOF) and Isolation Forest flagged approximately 5% of incidents as anomalies, aligning with expected outlier rates. These models highlighted rare crime occurrences and possible data inconsistencies.

1. Association Rule Mining:

* Strong rules, such as Borough = Bronx consistently linked to Patrol Borough Bronx, provided actionable insights. However, frequent occurrences of unknown victim attributes underscored gaps in data collection practices.

1. Evaluation:

* Cross-validation techniques and performance metrics (e.g., accuracy, precision, recall) were used to assess model effectiveness. Random Forest emerged as the most reliable model, with excellent balance across all metrics.

## 6.4 Determine Next Steps

**Data Quality Improvements**

* Address systemic issues with missing demographic data, particularly for victim attributes like age and race. Implementing stricter data validation protocols during collection can mitigate such gaps.
* Standardize attribute definitions (e.g., Group17 locations) to ensure consistent interpretation across analyses.

**Model Optimization**

* Fine-tune classification models, particularly KNN, by exploring advanced parameter settings, such as varying the number of neighbors and using weighted distance metrics.
* For clustering, test alternative algorithms like DBSCAN to better handle noisy data and non-linear patterns.

**Feature Engineering**

* Incorporate additional features, such as socio-economic indicators and weather patterns, to enrich the dataset and improve predictive accuracy.
* Explore dynamic time-series features to capture evolving crime trends, particularly for clustering and classification tasks.

**Outlier and Pattern Analysis**

* Refine outlier detection thresholds to balance sensitivity and specificity, ensuring meaningful anomaly identification.
* Use association rule findings to inform practical interventions, such as targeted patrols in identified hotspots.

**Workflow Enhancements**

* Automate repetitive preprocessing tasks, such as missing value imputation and attribute transformation, to streamline future analyses.
* Experiment with batch processing or distributed computing frameworks to handle large datasets more efficiently.

# 7.0 Conclusion

The comprehensive analysis of the NYPD Complaint Data aimed to identify crime trends, detect anomalies, and categorize crimes using advanced machine learning techniques. This project has successfully demonstrated how data-driven approaches can be utilized to enhance public safety initiatives, optimize resource allocation, and guide evidence-based policymaking. The detailed results, challenges addressed, and actionable recommendations provide a robust foundation for future work in crime analytics.

**Summary of Key Achievements**

1. **Crime Trends and Hotspots**:
   * Classification models, particularly Random Forest and Decision Tree, revealed clear patterns in the temporal and spatial distribution of crimes across New York City.
   * Borough-specific analysis highlighted critical hotspots, such as the Bronx and Brooklyn, which exhibited the highest crime densities, particularly for misdemeanors.
   * Time-based analysis identified peak crime hours and seasonal variations, providing valuable insights for scheduling law enforcement activities.
2. **Effective Anomaly Detection**:
   * Models like Local Outlier Factor (LOF) and Isolation Forest successfully flagged approximately 5% of the data as anomalies. These outliers often represented rare but significant events, such as crimes occurring at atypical times or in unexpected locations.
   * Anomalies also highlighted inconsistencies in the dataset, such as missing demographic attributes, pointing to areas for improvement in data collection.
3. **Association Rule Mining Insights**:
   * Strong associations were uncovered between crime attributes, such as the relationship between borough-specific patrol assignments and recurring incidents in the Bronx and Manhattan.
   * Patterns like the link between the "Group17" location designation and Bronx crimes provided insights into data structuring practices and operational patterns.
   * Demographic data associations, despite missing values, revealed meaningful trends, such as unknown victim details being consistently linked to NYPD jurisdictions, emphasizing the need for better demographic recording.
4. **Modeling Performance**:
   * **Random Forest** demonstrated exceptional classification performance with an accuracy of 94.98%, achieving near-perfect precision (99.93%) for violation predictions.
   * **Decision Tree** delivered comparable results, excelling in rule generation and interpretability for decision-making purposes.
   * **KNN**, while moderately accurate at 84.76%, effectively captured local patterns, such as crime-prone geographic clusters, but showed limitations in handling imbalanced datasets.
   * **K-Means clustering** identified 23 optimal clusters using the elbow method, revealing key borough-based patterns and validating the quality of clusters with a silhouette score of 0.72.

**Challenges and How They Were Addressed**

1. **Data Quality Issues**:
   * Missing and inconsistent values in critical attributes like victim demographics were systematically addressed through imputation, deduplication, and careful preprocessing.
   * Non-essential columns with excessive missing values, such as NYCHA housing details, were excluded to streamline the analysis and improve model reliability.
2. **Handling Imbalanced Data**:
   * Classification models were optimized through hyperparameter tuning and cross-validation to improve performance for minority crime classes, such as felonies.
   * Data transformations, such as time grouping and demographic categorization, helped balance feature distributions for better model performance.
3. **Interpretation Challenges**:
   * The structured nature of attributes like "Group17" required careful interpretation to understand the underlying geographical or operational significance.
   * Association rule mining revealed systemic patterns but also highlighted data limitations, such as the frequent occurrence of unknown demographic details, necessitating further refinement.

**Actionable Recommendations**

1. **Enhancing Data Quality and Collection**:
   * Improve real-time data entry validation to reduce missing values in critical fields, such as victim age and race.
   * Introduce mandatory fields and automated error-checking protocols to minimize inconsistencies.
2. **Incorporating External Datasets**:
   * Integrate socio-economic, weather, and traffic datasets to enrich the analysis and uncover deeper correlations between crime trends and external factors.
   * Explore geospatial data at a finer resolution to improve clustering and anomaly detection results.
3. **Advanced Modeling Techniques**:
   * Test alternative clustering algorithms, such as DBSCAN, for better handling of noisy and non-linear data distributions.
   * Experiment with ensemble methods like Gradient Boosting to enhance classification accuracy, particularly for imbalanced datasets.
4. **Automation and Scalability**:
   * Automate preprocessing workflows, including imputation and feature engineering, to reduce manual effort and streamline future analyses.
   * Develop scalable pipelines to handle larger datasets and improve processing efficiency, especially for computationally intensive models like Isolation Forest.

**Implications for Deployment**

1. **Optimized Resource Allocation**:
   * The identification of crime hotspots and temporal trends allows law enforcement to allocate patrols more effectively, reducing response times and maximizing coverage in high-crime areas.
   * Seasonal and time-specific insights can inform the scheduling of specialized task forces during peak periods.
2. **Targeted Crime Prevention Strategies**:

* Association rules and anomaly detection findings can guide interventions, such as focused outreach programs in neighborhoods with recurring crime patterns or anomalies.
* Tailored strategies, such as community policing in high-incident precincts, can address specific issues uncovered through this analysis.

1. **Improved Decision-Making**:

* Decision Tree and Random Forest outputs offer interpretable insights that can be directly applied to policy-making and operational decisions.
* Clustering results provide a clear visualization of crime patterns, facilitating strategic planning and community engagement.

**Future Work Directions**

The findings of this project highlight the importance of continuous improvement in data collection and analysis practices. Future iterations can focus on:

* Refining outlier detection thresholds to better differentiate between critical anomalies and data noise.
* Exploring dynamic crime prediction models using time-series analysis to capture evolving patterns.
* Developing interactive dashboards and visualization tools for real-time monitoring and decision-making.

**Conclusion**

This project underscores the power of advanced analytics in tackling urban challenges, such as crime prevention and public safety enhancement. By leveraging robust machine learning models and systematic data preparation techniques, this analysis provides actionable insights for law enforcement and policymakers. Moving forward, integrating more sophisticated models and external datasets will further improve the ability to predict, prevent, and respond to crimes effectively. The structured methodology ensures scalability, adaptability, and the potential to influence data-driven strategies across other urban domains.