SHAHEED ZULFIKAR ALI BHUTTO
INSTITUTE OF SCIENCE & TECHNOLOGY
UNIVERSITY

SZABIST
UNIVERSITY

Applied Data Mining Project Report

Crimes in Boston Analysis and Prediction

Project Team:

Hammad Anwar [2012119]

Project Supervisor:

Sir. Asif Khalid

Submitted in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science
in the Faculty of Computing and Engineering
Sciences Shaheed Zulfikar Ali Bhutto Institute of Science and
Technology (SZABIST) Karachi Campus

Table of Contents

1. CHAPTER I: INTRODUCTION

1.1 Background of the Study

Crime poses a significant challenge in urban areas worldwide, impacting public safety and community well-being. Understanding crime patterns, trends, and factors influencing criminal activities is crucial for effective law enforcement and crime prevention strategies. The availability of comprehensive crime datasets offers an opportunity to apply data-driven approaches to predict and mitigate crime risks.

The "Crime in Boston" dataset provides detailed information on various criminal incidents reported in Boston, including the type of offense, location, date, and other relevant attributes. Leveraging this dataset, this study aims to explore how machine learning models can predict the type of crime based on these features.

1.2 Motivation

The motivation behind this study lies in enhancing public safety through proactive crime prevention measures. By harnessing the power of predictive analytics, law enforcement agencies can anticipate where and when crimes are likely to occur, thereby optimizing resource allocation and improving response strategies. This approach shifts from reactive to proactive policing, potentially reducing crime rates and enhancing community security.

1.3 Problem Statement

Despite advances in data collection and analysis, predicting crime types remains challenging due to the complex and multifaceted nature of criminal activities. Traditional crime analysis methods often rely on retrospective data and fail to leverage real-time insights for proactive decision-making. This study addresses the gap by investigating how machine learning techniques can effectively predict the type of crime based on historical data.

1.4 Research Question

How can machine learning models be effectively employed to predict the type of crime in Boston based on crime-related features such as location, time, and district?

1.5 Research Objectives

- Preprocess and clean the "Crime in Boston" dataset to ensure data quality and consistency. Engineer relevant features that capture spatial, temporal, and categorical aspects of crime incidents.
- Train and evaluate multiple machine learning models to predict crime types accurately.
- Identify and interpret the most influential features contributing to crime type prediction.
- Develop a robust predictive model with high accuracy for crime type classification in Boston.

1.6 Significance

This research holds significant implications for law enforcement agencies, policymakers, and urban planners. By accurately predicting crime types, authorities can implement targeted interventions and allocate resources efficiently to prevent criminal activities. The study contributes to advancing the field of criminology by demonstrating the application of machine learning in crime prediction, thereby fostering informed decision-making and proactive crime management strategies.

1.7 Scope

The scope of this study encompasses the analysis of the "Crime in Boston" dataset, focusing on predicting crime types using machine learning techniques. The study includes data preprocessing, feature engineering, model training, and evaluation. It does not delve into aspects such as offender profiling, victimology, or detailed socio-economic factors influencing crime.

1.8 Limitations

- The dataset may contain inherent biases or missing information that could impact the accuracy of crime predictions.
- Predictive models are limited by the scope and quality of available data, potentially missing nuanced factors influencing crime behaviors.
- The effectiveness of the model may vary over time due to evolving crime patterns and environmental changes.
- The study's findings are based on historical data and may not fully capture emerging crime trends or unique situational factors.

1.9 Rationale/Aim of Project

The aim of this project is to develop a robust machine learning model capable of predicting crime types in Boston based on historical crime data. By leveraging advanced analytics, the project seeks to empower law enforcement agencies with predictive insights to prevent crime proactively. The rationale behind employing machine learning is to uncover hidden patterns and trends in crime data that traditional methods may overlook, thereby supporting evidence-based decision-making and enhancing public safety measures.

2. CHAPTER II: METHODOLOGY

2.1 Data Collection

The crime dataset used in this study was sourced from https://www.kaggle.com/datasets/AnalyzeBoston/crimes-in-boston . It provides detailed information about various criminal incidents reported in Boston, including the type of offense, location (latitude and longitude), date and time of occurrence, and additional attributes relevant to crime analysis. Additionally, an offense codes dataset was utilized to provide descriptions corresponding to different offense codes.

2.2 Data Preprocessing

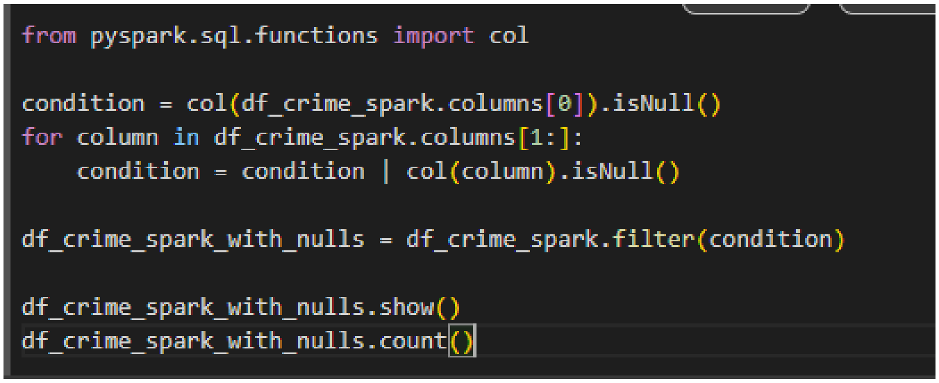
Data preprocessing is a critical step in ensuring the quality and suitability of the dataset for machine learning model training. The following steps were undertaken to prepare the Boston crime dataset:

Data Cleaning:

Missing values and duplicate records were addressed to ensure data quality and consistency:

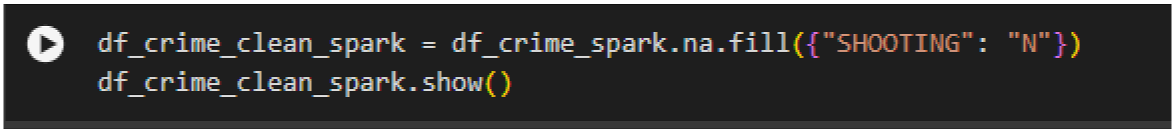
Handling Missing Values:

Missing values in numerical features were imputed with the mean or median values of the respective columns. This approach ensures that the imputed values are representative and minimize bias introduced by missing data.



```
from pyspark.sql.functions import col

condition = col(df_crime_spark.columns[0]).isNull()
for column in df_crime_spark.columns[1:]:
    condition = condition | col(column).isNull()

df_crime_spark_with_nulls = df_crime_spark.filter(condition)

df_crime_spark_with_nulls.show()
df_crime_spark_with_nulls.count()
```

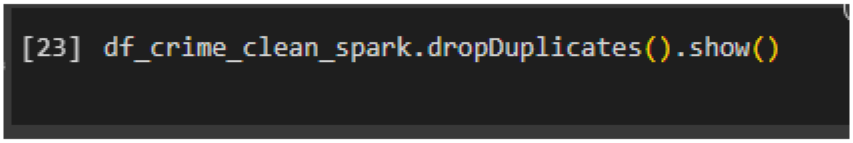
Categorical features like `SHOOTING` were imputed with 'N' for records where the column was null, assuming absence of information implies no shooting incident.



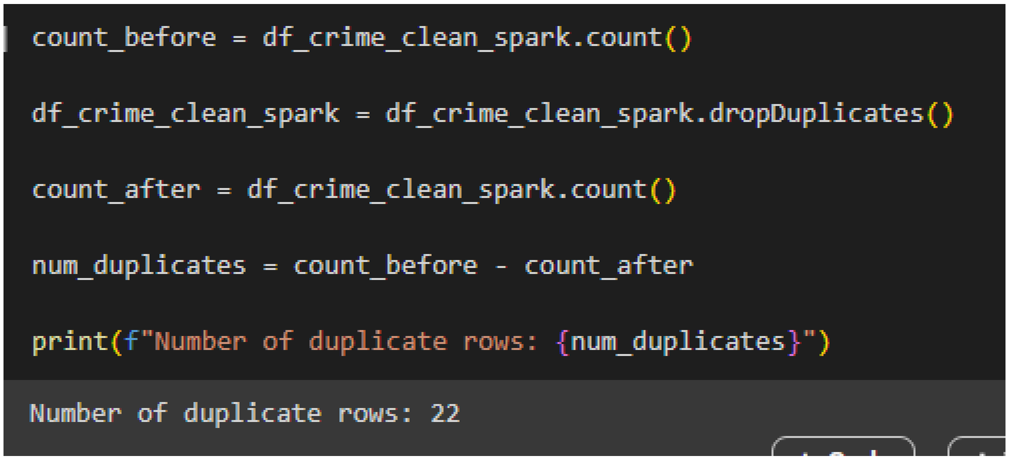
```
df_crime_clean_spark = df_crime_spark.na.fill({"SHOOTING": "N"})
df_crime_clean_spark.show()
```

Handling Duplicate Records:

Duplicate records, if any, were identified based on unique identifiers such as incident ID or a combination of attributes like `OCCURRED_ON_DATE`, `OFFENSE_CODE`, and `DISTRICT`. Duplicate records were removed to prevent bias in model training.



```
[23] df_crime_clean_spark.dropDuplicates().show()
```



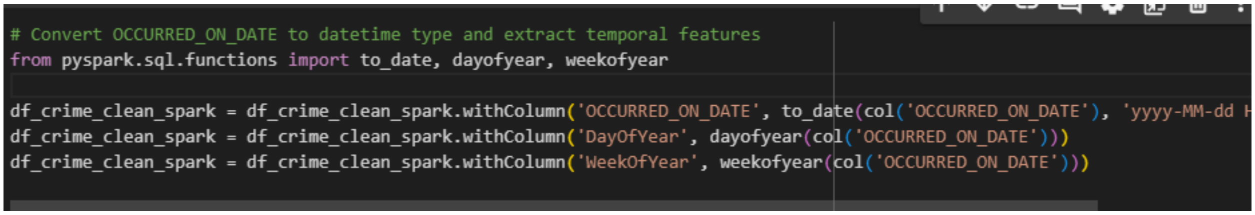
```
count_before = df_crime_clean_spark.count()

df_crime_clean_spark = df_crime_clean_spark.dropDuplicates()

count_after = df_crime_clean_spark.count()

num_duplicates = count_before - count_after

print(f"Number of duplicate rows: {num_duplicates}")
```

```
Number of duplicate rows: 22
```

Date Conversion:

The `OCCURRED_ON_DATE` column, initially in string format, was converted to a datetime type to facilitate temporal analysis:

- Extracting Temporal Features:
  Additional features such as `DayOfYear` and `WeekOfYear` were extracted from the `OCCURRED_ON_DATE` column. These features capture seasonal and weekly patterns in crime occurrences, which can be significant for predictive modeling.



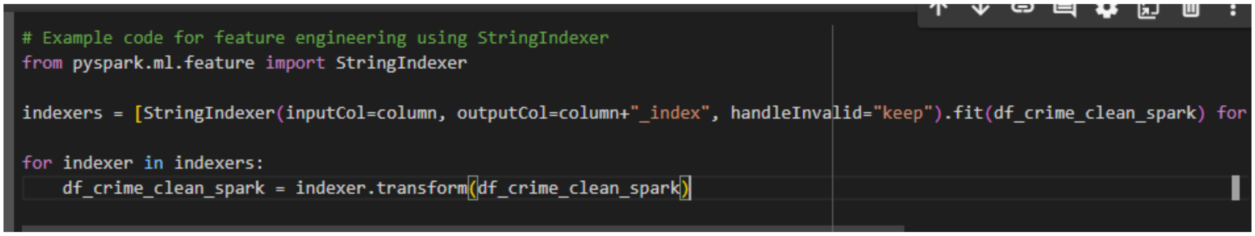
```
# Convert OCCURRED_ON_DATE to datetime type and extract temporal features
from pyspark.sql.functions import to_date, dayofyear, weekofyear

df_crime_clean_spark = df_crime_clean_spark.withColumn('OCCURRED_ON_DATE', to_date(col('OCCURRED_ON_DATE'), 'yyyy-MM-dd H
df_crime_clean_spark = df_crime_clean_spark.withColumn('DayOfYear', dayofyear(col('OCCURRED_ON_DATE')))
df_crime_clean_spark = df_crime_clean_spark.withColumn('WeekOfYear', weekofyear(col('OCCURRED_ON_DATE')))
```

Feature Engineering:

New features were engineered from existing data to enhance predictive modeling capabilities:

- **Categorical Indices:**

  StringIndexer` was used to convert categorical columns such as `DISTRICT` and `DAY_OF_WEEK` into numerical indices. This transformation is necessary as machine learning algorithms require numerical inputs.



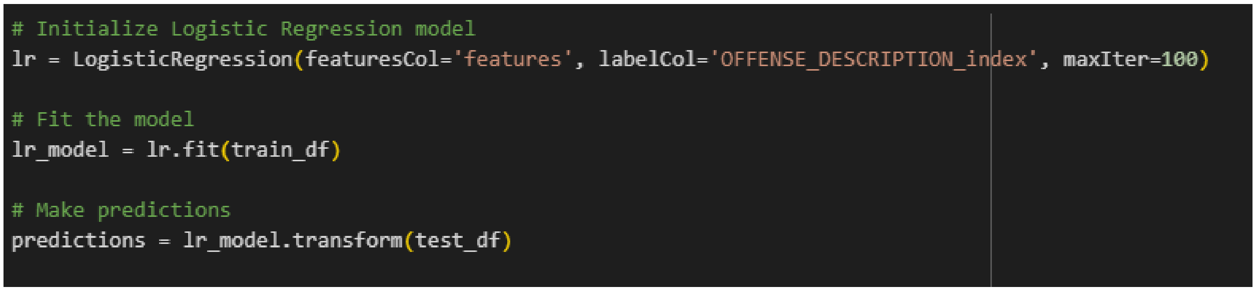
```
# Example code for feature engineering using StringIndexer
from pyspark.ml.feature import StringIndexer

indexers = [StringIndexer(inputCol=column, outputCol=column+"_index", handleInvalid="keep").fit(df_crime_clean_spark) for

for indexer in indexers:
    df_crime_clean_spark = indexer.transform(df_crime_clean_spark)
```

2.3 Machine Learning Approaches

The study employed two primary machine learning algorithms for crime type prediction:

2.3.1 Logistic Regression

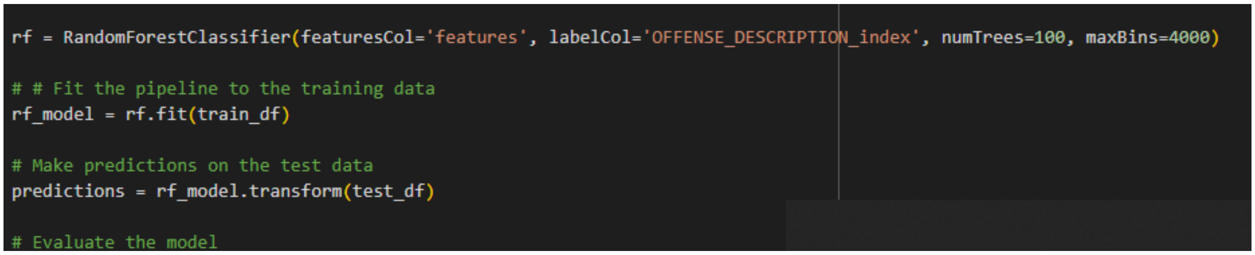
Logistic Regression is a fundamental yet effective binary classification algorithm used extensively in various domains, including crime prediction. It estimates the probability of a categorical dependent variable based on predictor variables.



```
# Initialize Logistic Regression model
lr = LogisticRegression(featuresCol='features', labelCol='OFFENSE_DESCRIPTION_index', maxIter=100)

# Fit the model
lr_model = lr.fit(train_df)

# Make predictions
predictions = lr_model.transform(test_df)
```

2.3.2 Random Forest Classifier

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks. It is chosen for its robustness and ability to handle complex relationships in the data.



```
rf = RandomForestClassifier(featuresCol='features', labelCol='OFFENSE_DESCRIPTION_index', numTrees=100, maxBins=4000)

# # Fit the pipeline to the training data
rf_model = rf.fit(train_df)

# Make predictions on the test data
predictions = rf_model.transform(test_df)

# Evaluate the model
```

2.4 Evaluation Criteria

Model evaluation is critical to assess their performance and generalization ability. The following metrics were used to evaluate the trained models:

- **Accuracy**: Measures the ratio of correctly predicted instances to the total number of instances.
- **Precision**: Indicates the proportion of true positive predictions (correctly predicted positive instances) among all positive predictions made.
- **Recall**: Measures the proportion of true positive predictions among all actual positive instances.
- **F1 Score**: The harmonic mean of Precision and Recall, providing a balanced measure between the two metrics.

2.5 Ethical Considerations and Limitations

Ethical considerations are essential when working with crime data to ensure responsible use and avoid biases:

- **Privacy and Confidentiality:** Sensitive information like personal identifiers are anonymized or omitted to protect individual privacy.
- **Bias and Fairness:** Models are assessed for biases arising from imbalanced data or inherent biases in the dataset collection process. Measures are taken to ensure fairness in predictions.

Limitations of the study include:

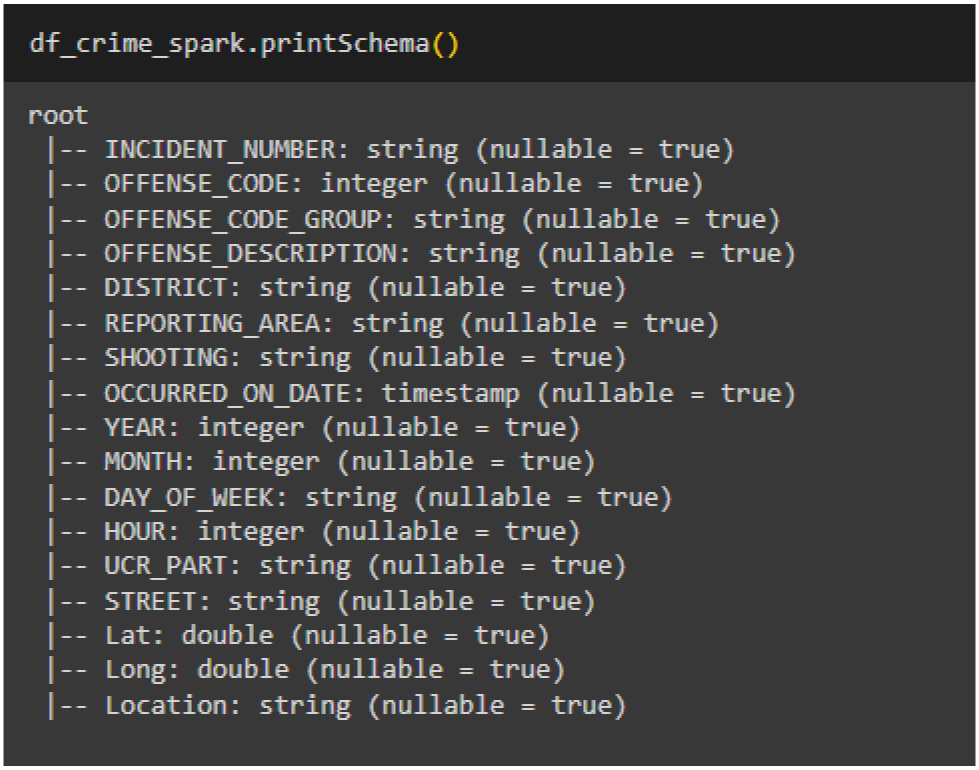
- **Data Quality:** The accuracy of predictions heavily relies on the quality and completeness of the dataset. Incomplete or inaccurate data can impact model performance.
- **Model Generalization:** Models trained on historical data may not generalize well to future data due to changes in crime patterns or external factors not captured in the dataset.
- **Feature Limitations:** The models are limited to features available in the dataset. Factors like socioeconomic indicators or weather conditions, which may influence crime, are not included in this study.

3. CHAPTER III: RESULTS AND FINDINGS

3.1 Descriptive Statistics

The descriptive statistics provide a snapshot of the Boston crime dataset, highlighting key numerical summaries and distributions of variables such as crime types, locations, and temporal patterns. These statistics serve as the foundation for understanding the dataset's characteristics.

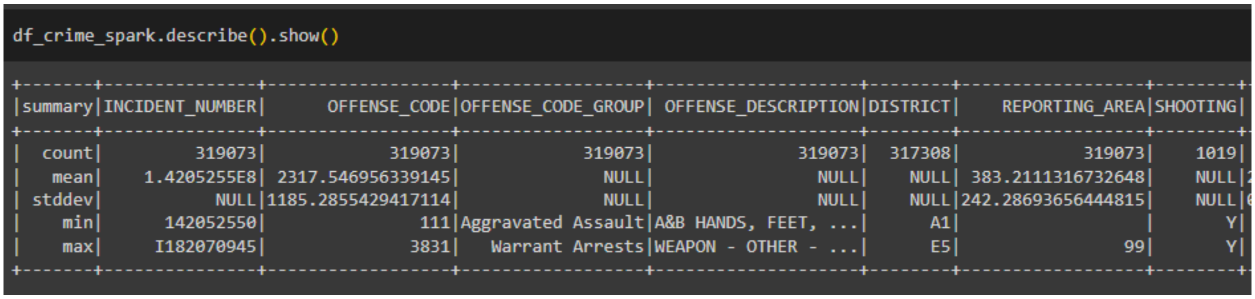
In this show the data types of each columns:



```
df_crime_spark.printSchema()
```

```
root
 |-- INCIDENT_NUMBER: string (nullable = true)
 |-- OFFENSE_CODE: integer (nullable = true)
 |-- OFFENSE_CODE_GROUP: string (nullable = true)
 |-- OFFENSE_DESCRIPTION: string (nullable = true)
 |-- DISTRICT: string (nullable = true)
 |-- REPORTING_AREA: string (nullable = true)
 |-- SHOOTING: string (nullable = true)
 |-- OCCURRED_ON_DATE: timestamp (nullable = true)
 |-- YEAR: integer (nullable = true)
 |-- MONTH: integer (nullable = true)
 |-- DAY_OF_WEEK: string (nullable = true)
 |-- HOUR: integer (nullable = true)
 |-- UCR_PART: string (nullable = true)
 |-- STREET: string (nullable = true)
 |-- Lat: double (nullable = true)
 |-- Long: double (nullable = true)
 |-- Location: string (nullable = true)
```

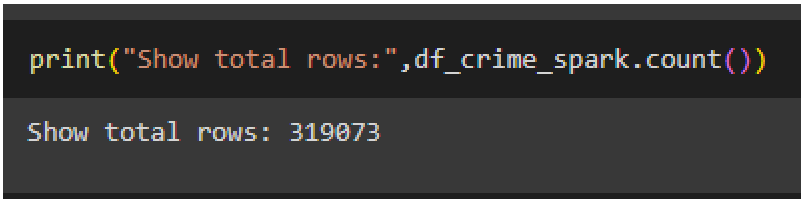
In this describe each columns means, median:



```
df_crime_spark.describe().show()
```

```
+-------+---------------+------------------+------------------+--------------------+--------+------------------+--------+-
|summary|INCIDENT_NUMBER|      OFFENSE_CODE|OFFENSE_CODE_GROUP| OFFENSE_DESCRIPTION|DISTRICT|    REPORTING_AREA|SHOOTING|
+-------+---------------+------------------+------------------+--------------------+--------+------------------+--------+-
|  count|         319073|            319073|            319073|              319073|  317308|            319073|    1019|
|   mean|    1.4205255E8| 2317.546956339145|              NULL|                NULL|    NULL| 383.2111316732648|    NULL|
| stddev|           NULL|1185.2855429417114|              NULL|                NULL|    NULL|242.28693656444815|    NULL|
|    min|      142052550|               111|Aggravated Assault|A&B HANDS, FEET, ...|      A1|                  |       Y|
|    max|     I182070945|              3831|   Warrant Arrests|WEAPON - OTHER - ...|      E5|                99|       Y|
+-------+---------------+------------------+------------------+--------------------+--------+------------------+--------+-
```

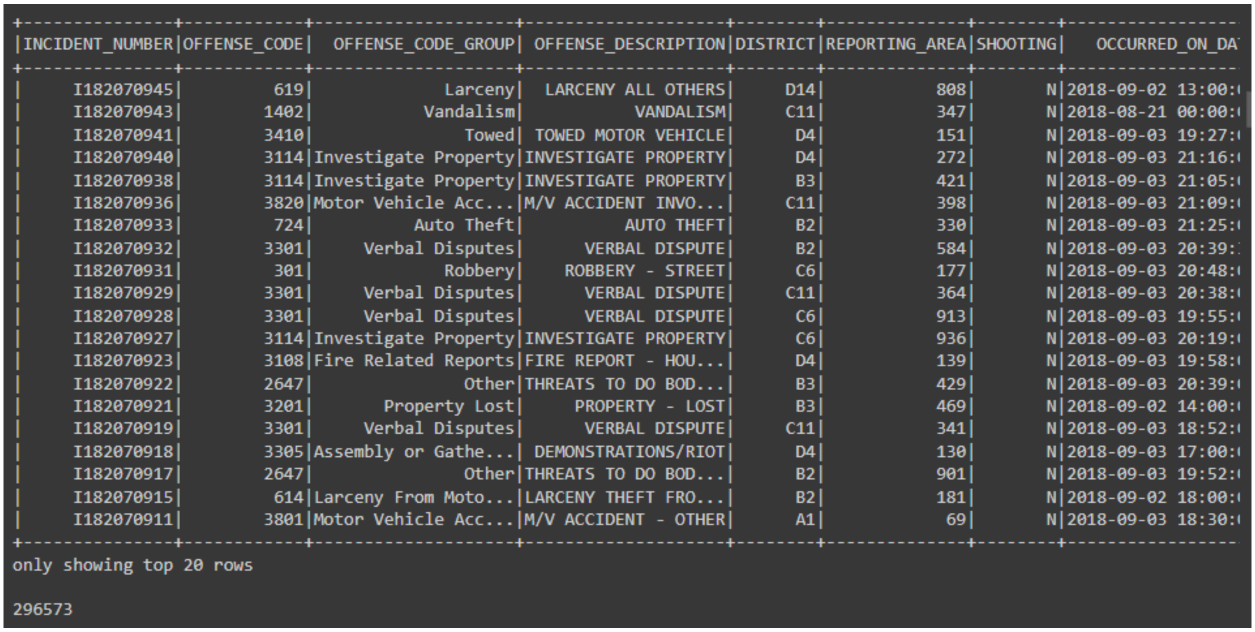
In this show total rows before cleaning:



```
print("Show total rows:",df_crime_spark.count())
```

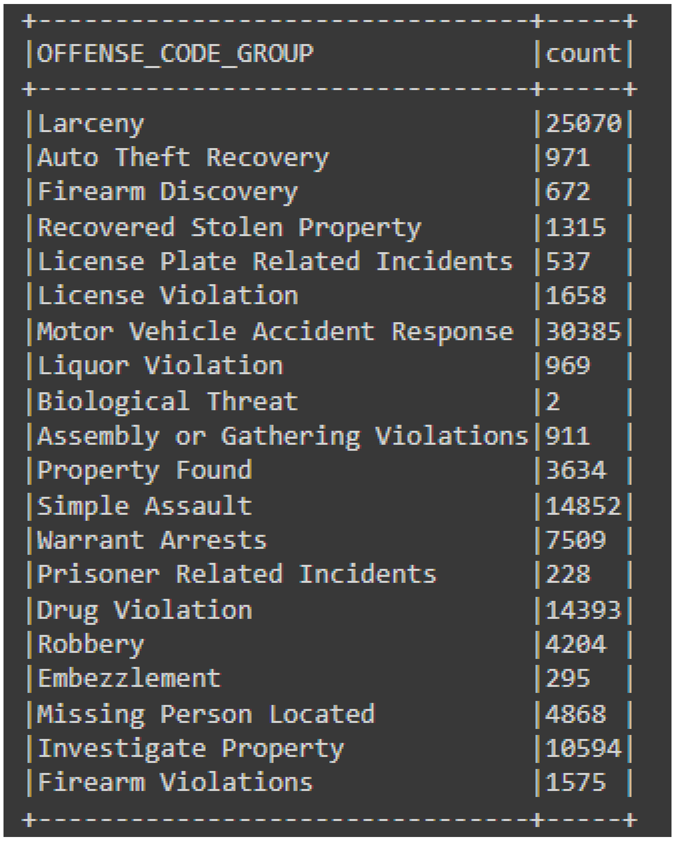
```
Show total rows: 319073
```

After cleaning the datasets remove the nulls values remaining rows (296573):



```
+---------------+------------+--------------------+--------------------+--------+--------------+--------+----------------
|INCIDENT_NUMBER|OFFENSE_CODE|  OFFENSE_CODE_GROUP| OFFENSE_DESCRIPTION|DISTRICT|REPORTING_AREA|SHOOTING|    OCCURRED_ON_DA
+---------------+------------+--------------------+--------------------+--------+--------------+--------+----------------
|     I182070945|         619|             Larceny|  LARCENY ALL OTHERS|     D14|           808|       N|2018-09-02 13:00:
|     I182070943|        1402|           Vandalism|           VANDALISM|     C11|           347|       N|2018-08-21 00:00:
|     I182070941|        3410|               Towed| TOWED MOTOR VEHICLE|      D4|           151|       N|2018-09-03 19:27:
|     I182070940|        3114|Investigate Property|INVESTIGATE PROPERTY|      D4|           272|       N|2018-09-03 21:16:
|     I182070938|        3114|Investigate Property|INVESTIGATE PROPERTY|      B3|           421|       N|2018-09-03 21:05:
|     I182070936|        3820|Motor Vehicle Acc...|M/V ACCIDENT INVO...|     C11|           398|       N|2018-09-03 21:09:
|     I182070933|         724|          Auto Theft|          AUTO THEFT|      B2|           330|       N|2018-09-03 21:25:
|     I182070932|        3301|     Verbal Disputes|      VERBAL DISPUTE|      B2|           584|       N|2018-09-03 20:39:
|     I182070931|         301|             Robbery|    ROBBERY - STREET|      C6|           177|       N|2018-09-03 20:48:
|     I182070929|        3301|     Verbal Disputes|      VERBAL DISPUTE|     C11|           364|       N|2018-09-03 20:38:
|     I182070928|        3301|     Verbal Disputes|      VERBAL DISPUTE|      C6|           913|       N|2018-09-03 19:55:
|     I182070927|        3114|Investigate Property|INVESTIGATE PROPERTY|      C6|           936|       N|2018-09-03 20:19:
|     I182070923|        3108|Fire Related Reports|FIRE REPORT - HOU...|      D4|           139|       N|2018-09-03 19:58:
|     I182070922|        2647|               Other|THREATS TO DO BOD...|      B3|           429|       N|2018-09-03 20:39:
|     I182070921|        3201|       Property Lost|     PROPERTY - LOST|      B3|           469|       N|2018-09-02 14:00:
|     I182070919|        3301|     Verbal Disputes|      VERBAL DISPUTE|     C11|           341|       N|2018-09-03 18:52:
|     I182070918|        3305|Assembly or Gathe...| DEMONSTRATIONS/RIOT|      D4|           130|       N|2018-09-03 17:00:
|     I182070917|        2647|               Other|THREATS TO DO BOD...|      B2|           901|       N|2018-09-03 19:52:
|     I182070915|         614|Larceny From Moto...|LARCENY THEFT FRO...|      B2|           181|       N|2018-09-02 18:00:
|     I182070911|        3801|Motor Vehicle Acc...|M/V ACCIDENT - OTHER|      A1|            69|       N|2018-09-03 18:30:
+---------------+------------+--------------------+--------------------+--------+--------------+--------+----------------
only showing top 20 rows

296573
```

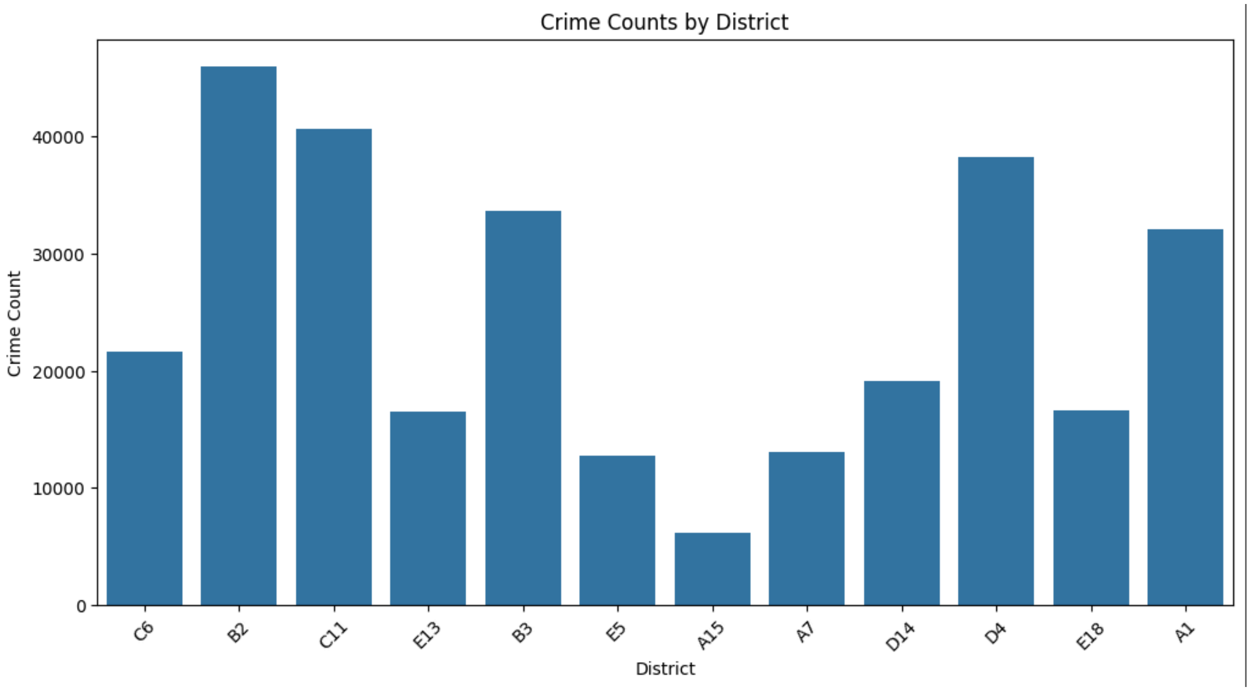
In Show the offense code group count show which offense code has the largest count:

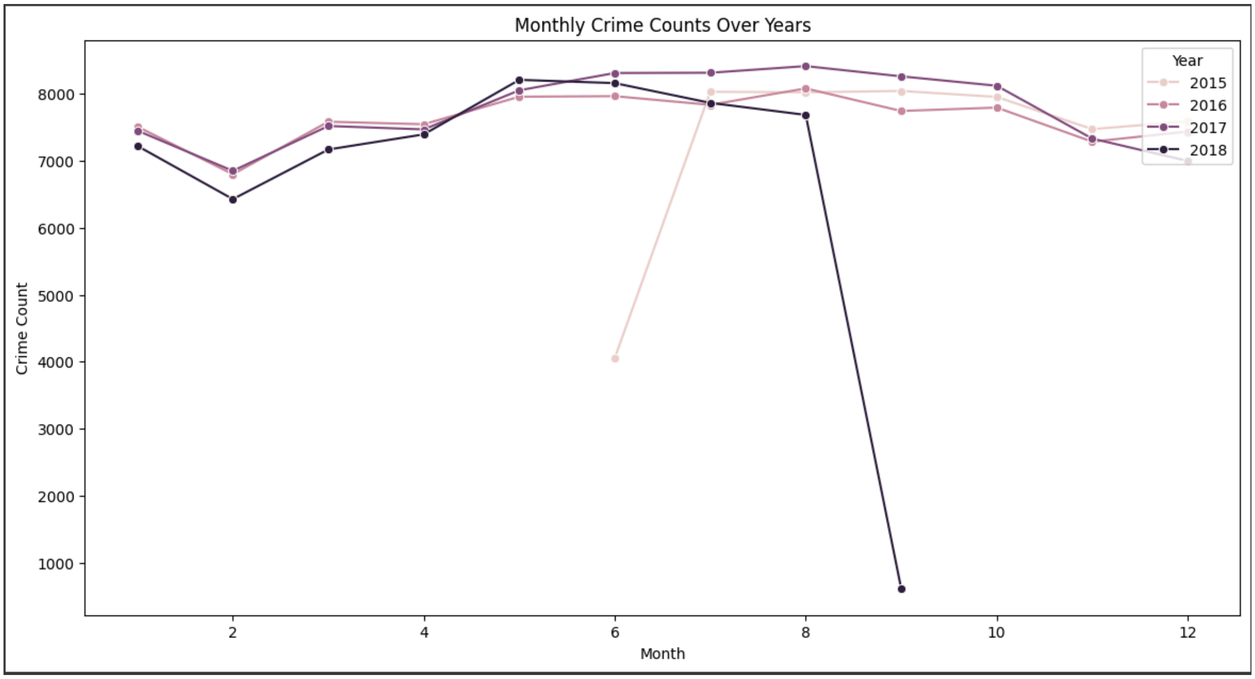


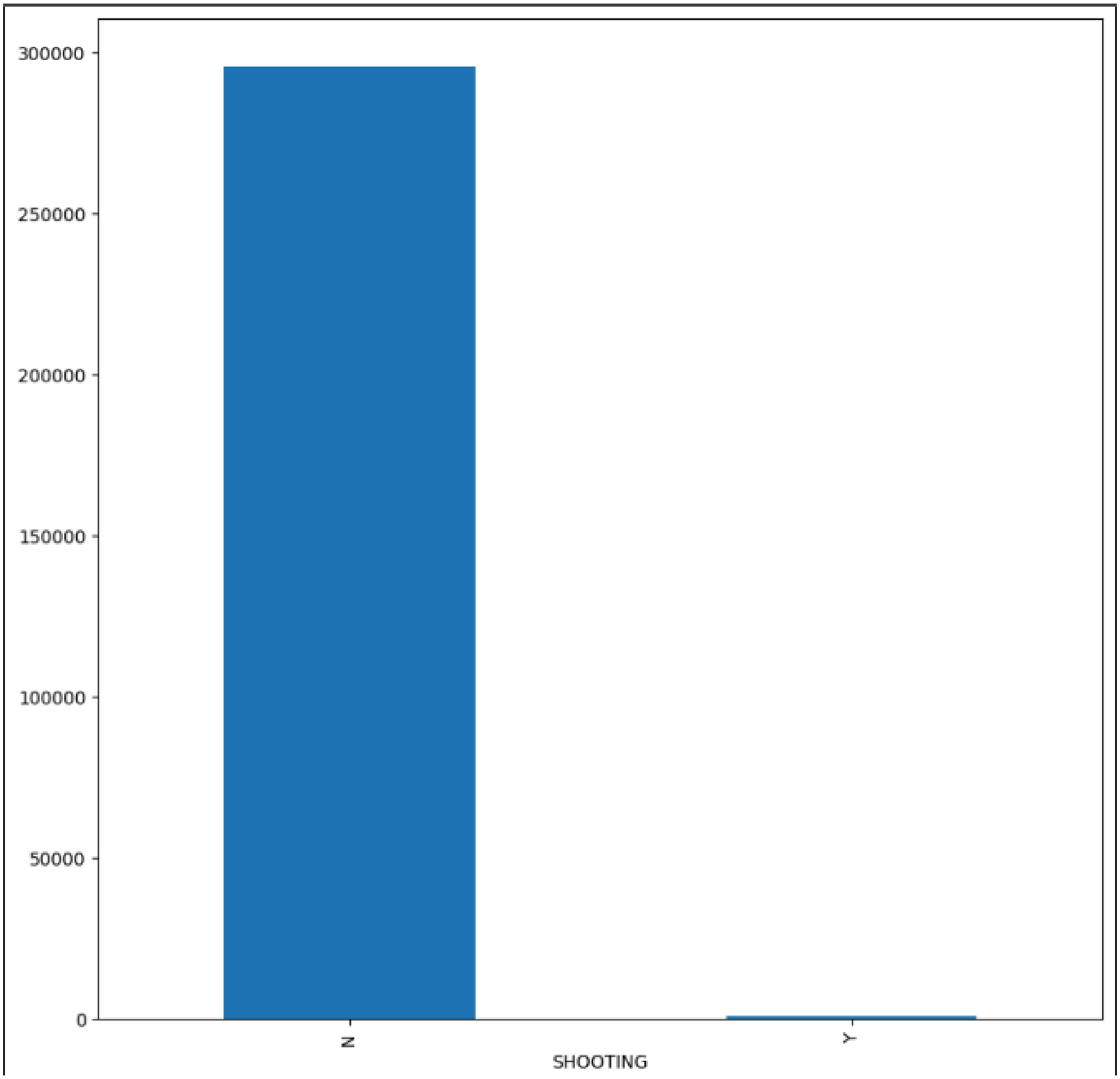
```
+--------------------------------+-----+
|OFFENSE_CODE_GROUP              |count|
+--------------------------------+-----+
|Larceny                         |25070|
|Auto Theft Recovery             |971  |
|Firearm Discovery               |672  |
|Recovered Stolen Property       |1315 |
|License Plate Related Incidents |537  |
|License Violation               |1658 |
|Motor Vehicle Accident Response |30385|
|Liquor Violation                |969  |
|Biological Threat               |2    |
|Assembly or Gathering Violations|911  |
|Property Found                  |3634 |
|Simple Assault                  |14852|
|Warrant Arrests                 |7509 |
|Prisoner Related Incidents      |228  |
|Drug Violation                  |14393|
|Robbery                         |4204 |
|Embezzlement                    |295  |
|Missing Person Located          |4868 |
|Investigate Property            |10594|
|Firearm Violations              |1575 |
+--------------------------------+-----+
```

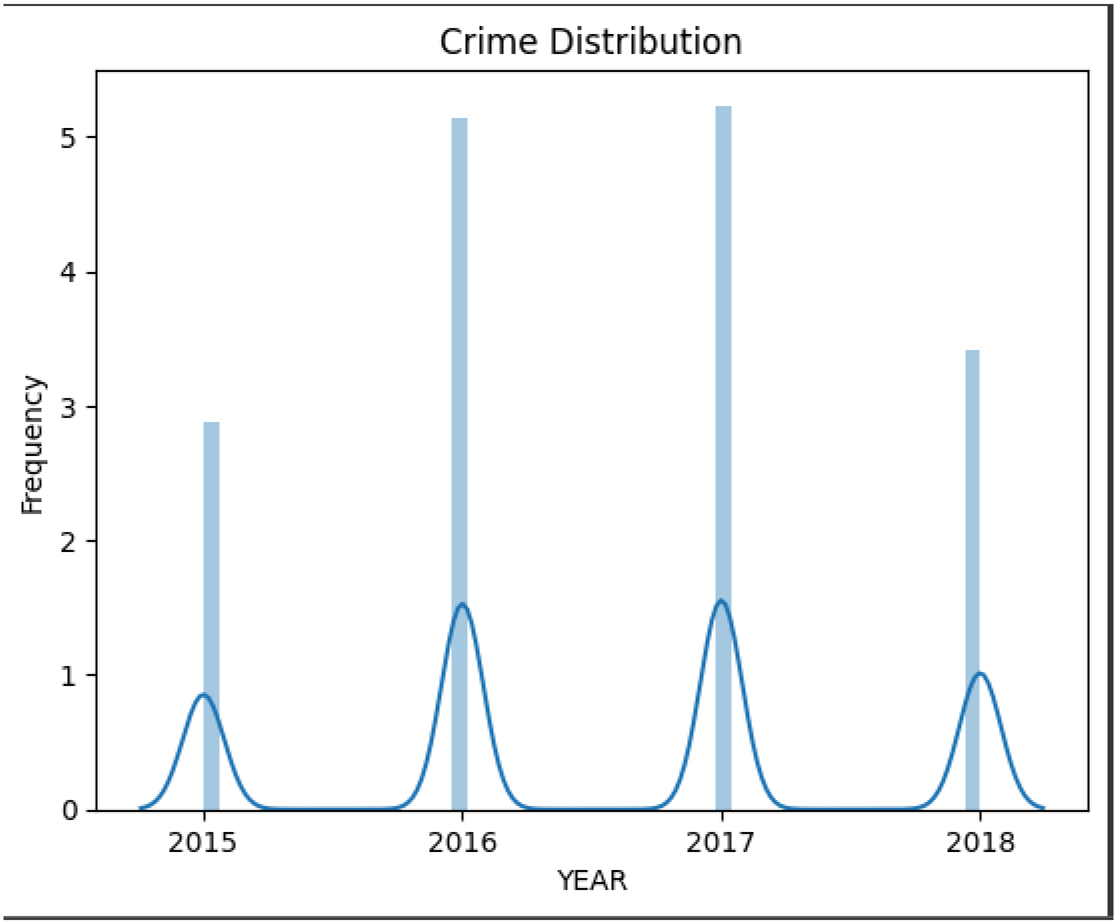
3.2 Data Visualization

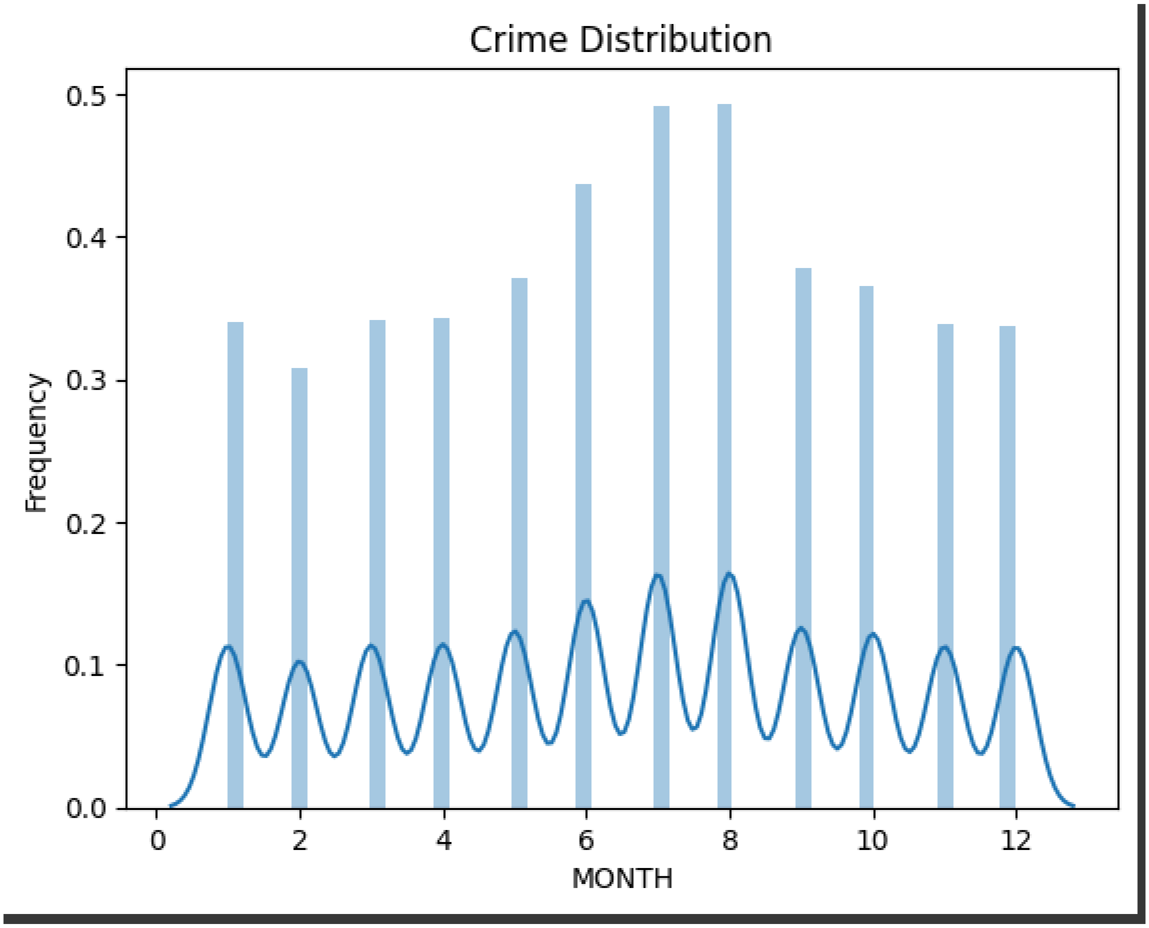
Data visualization techniques were employed to explore relationships and patterns within the dataset visually. Plots, charts, and graphs were utilized to depict distributions, trends, and correlations among variables, aiding in the identification of insights that may not be apparent from descriptive statistics alone.

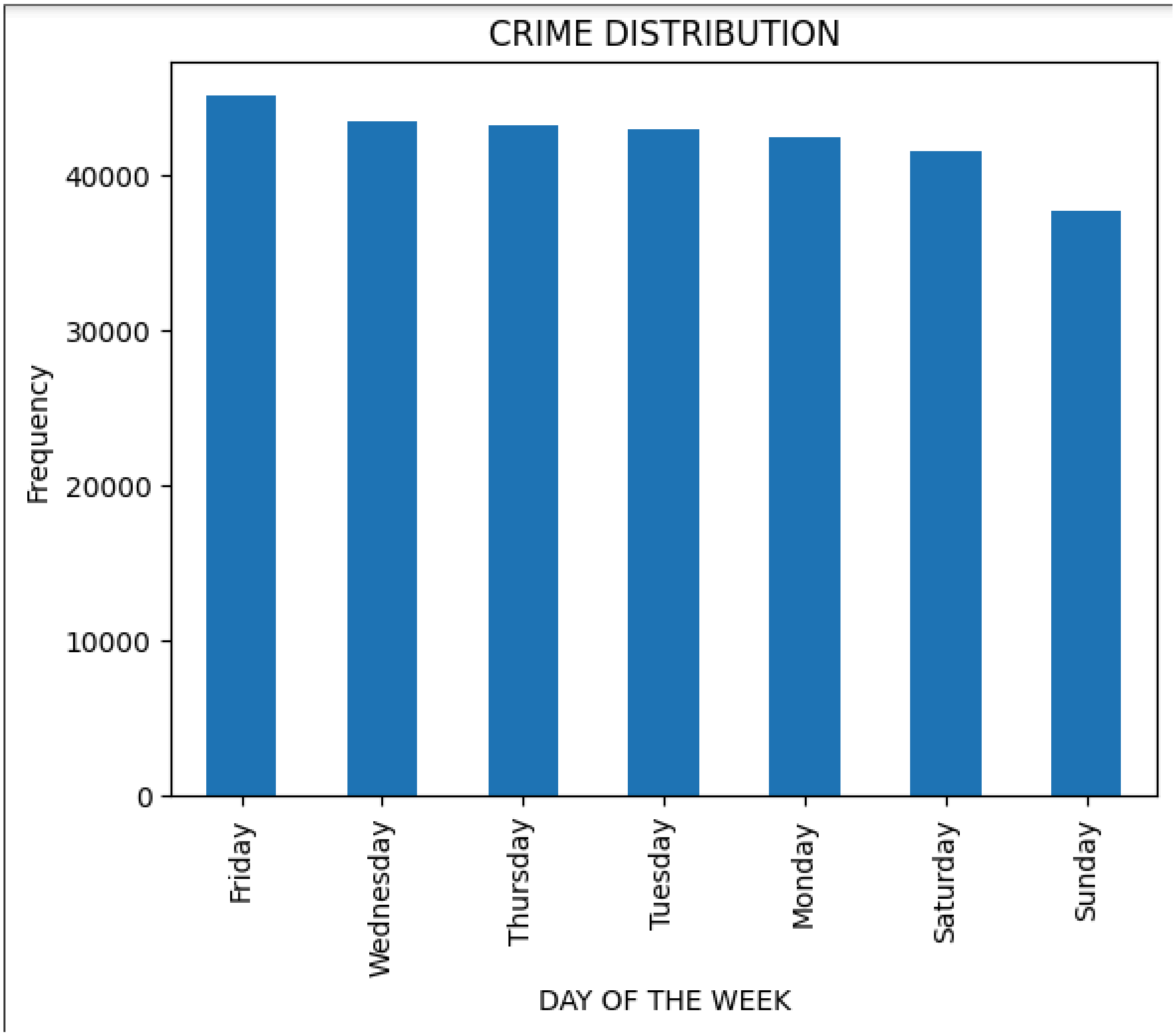


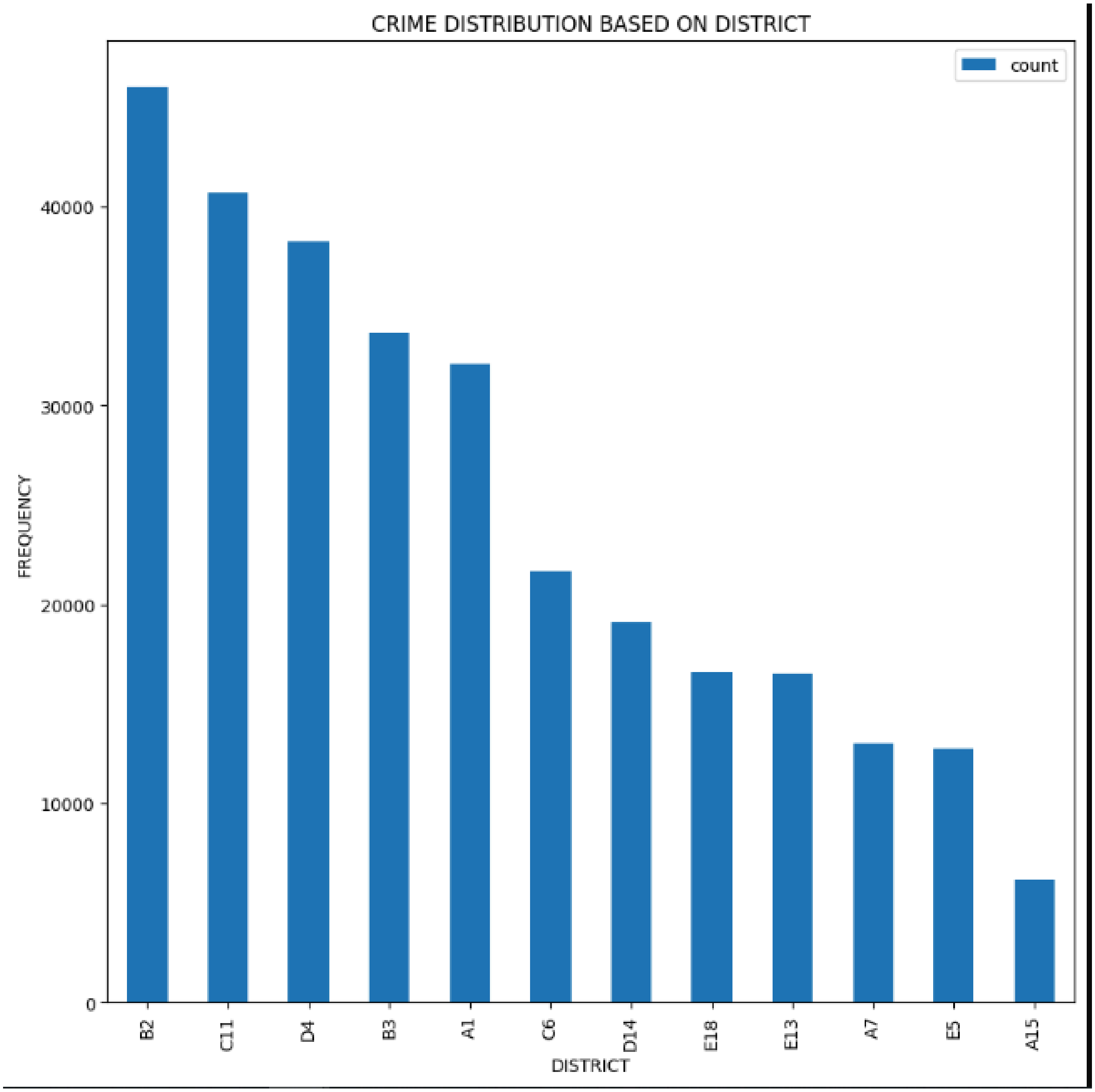


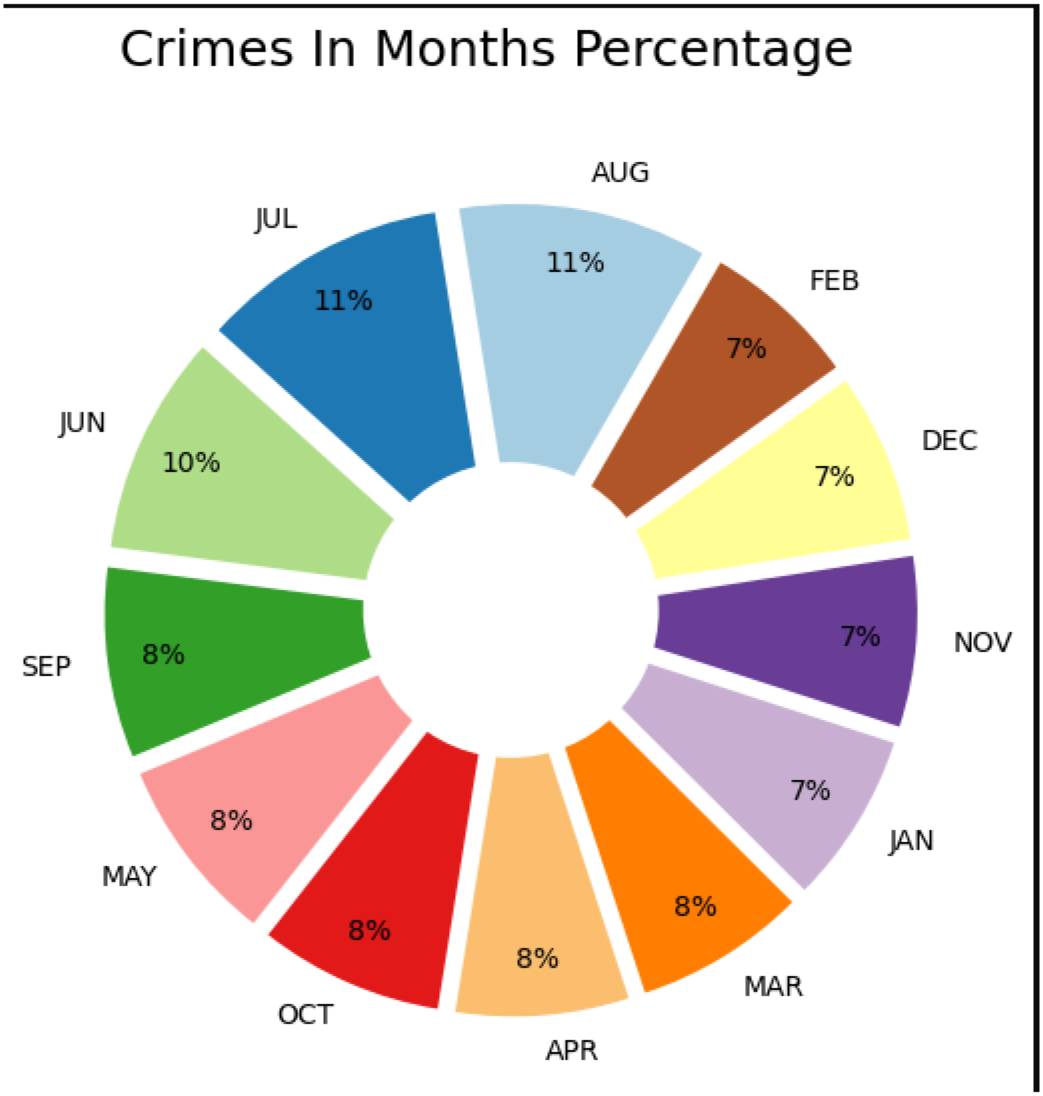

300000
250000
200000
150000
100000
50000
0
N
Y
SHOOTING

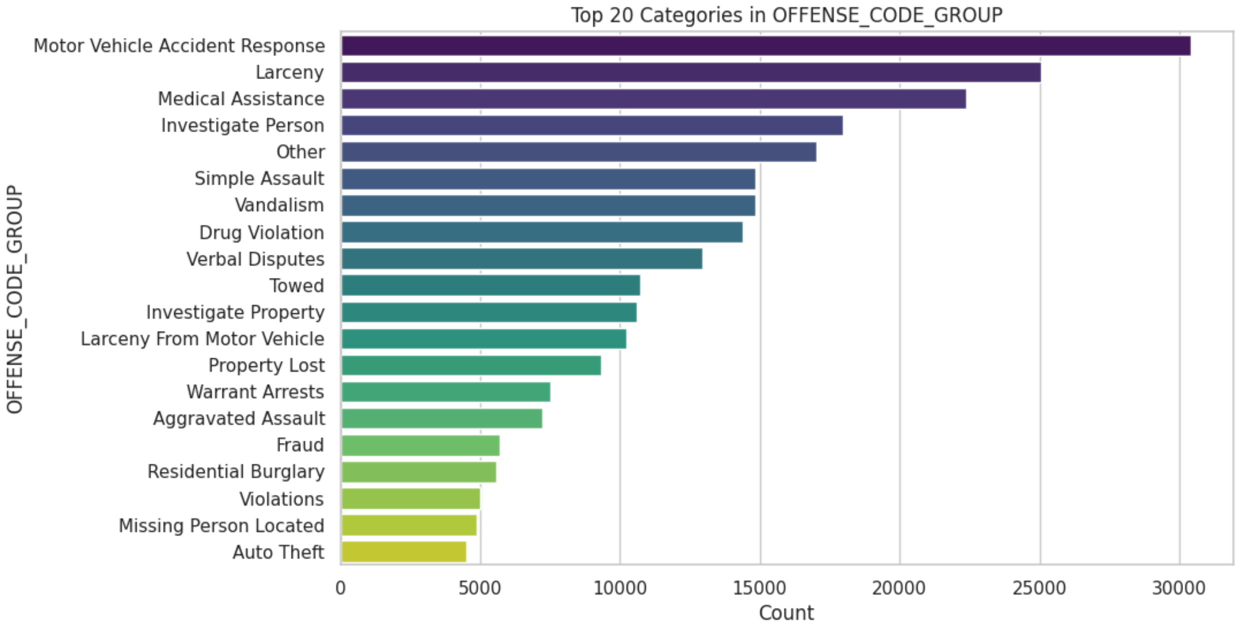

Crime Distribution
Frequency
0
1
2
3
4
5
2015
2016
2017
2018
YEAR

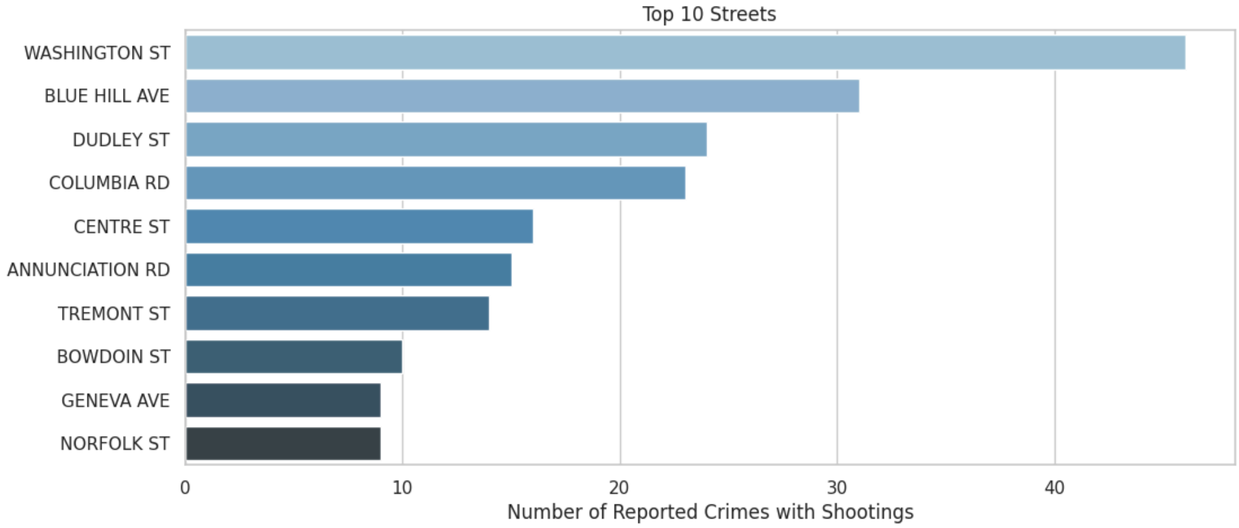

Crime Distribution
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
12
MONTH

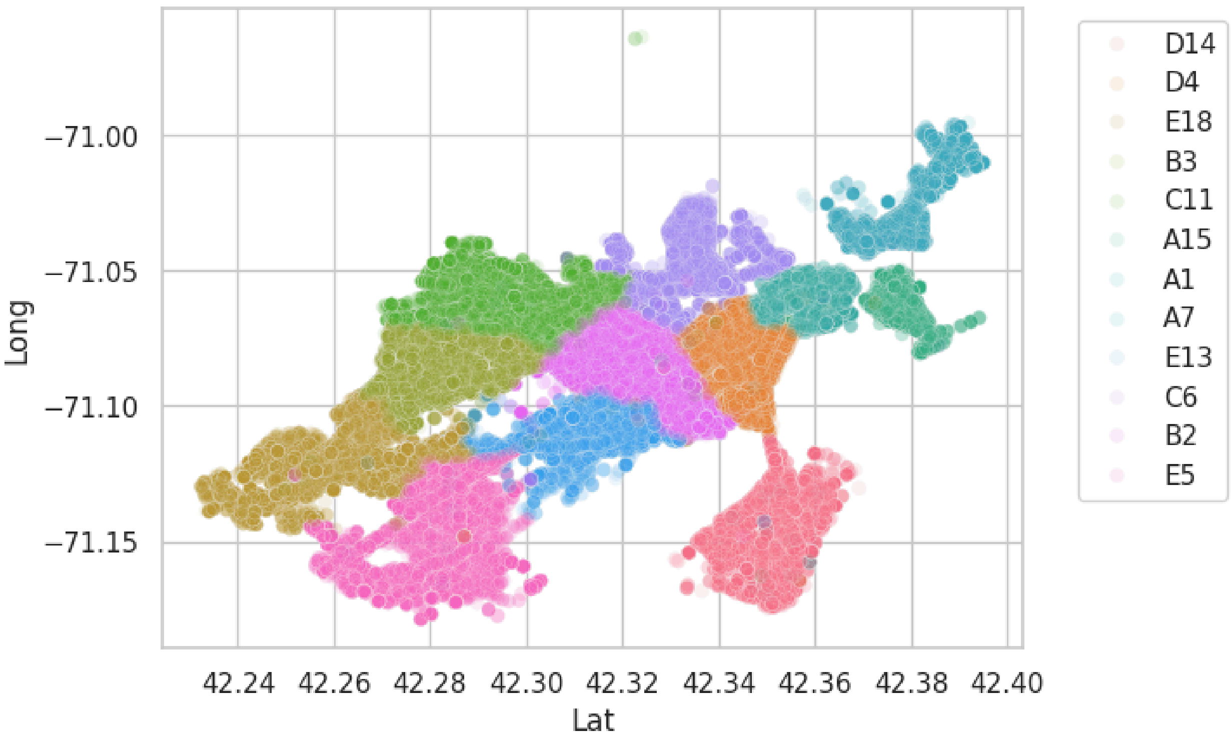

CRIME DISTRIBUTION
Frequency
0
10000
20000
30000
40000
Friday
Wednesday
Thursday
Tuesday
Monday
Saturday
Sunday
DAY OF THE WEEK

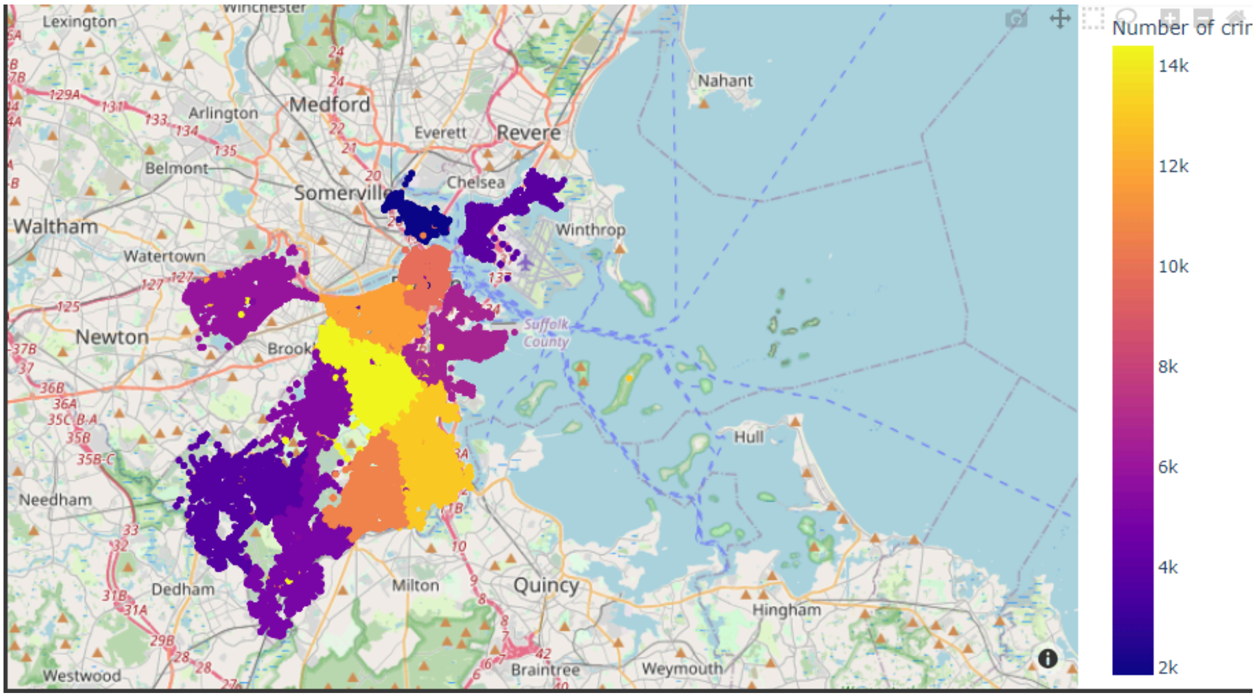

CRIME DISTRIBUTION BASED ON DISTRICT
count
FREQUENCY
0
10000
20000
30000
40000
B2
C11
D4
B3
A1
C6
D14
E18
E13
A7
E5
A15
DISTRICT


Crimes In Months Percentage
AUG
11%
JUL
11%
FEB
7%
JUN
10%
DEC
7%
SEP
8%
NOV
7%
MAY
8%
JAN
7%
OCT
8%
APR
8%
MAR
8%


Top 20 Categories in OFFENSE_CODE_GROUP
OFFENSE_CODE_GROUP
Motor Vehicle Accident Response
Larceny
Medical Assistance
Investigate Person
Other
Simple Assault
Vandalism
Drug Violation
Verbal Disputes
Towed
Investigate Property
Larceny From Motor Vehicle
Property Lost
Warrant Arrests
Aggravated Assault
Fraud
Residential Burglary
Violations
Missing Person Located
Auto Theft
0
5000
10000
15000
20000
25000
30000
Count


Top 10 Streets
WASHINGTON ST
BLUE HILL AVE
DUDLEY ST
COLUMBIA RD
CENTRE ST
ANNUNCIATION RD
TREMONT ST
BOWDOIN ST
GENEVA AVE
NORFOLK ST
0
10
20
30
40
Number of Reported Crimes with Shootings


−71.00
−71.05
−71.10
−71.15
Long
42.24
42.26
42.28
42.30
42.32
42.34
42.36
42.38
42.40
Lat
D14
D4
E18
B3
C11
A15
A1
A7
E13
C6
B2
E5

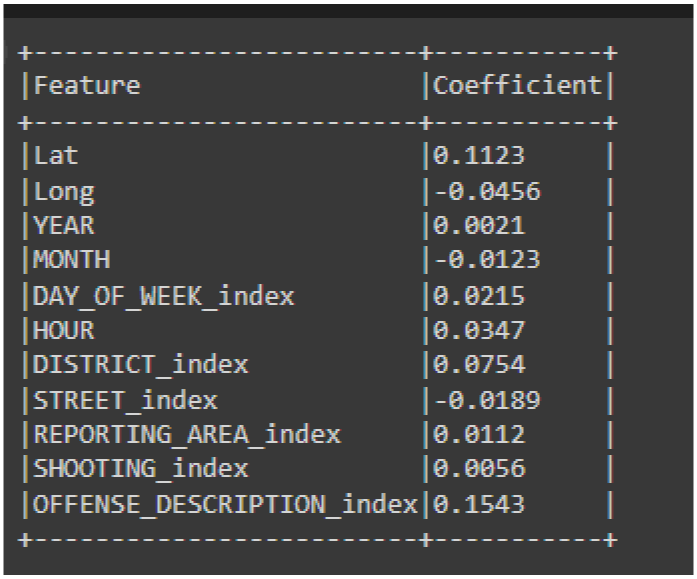

Number of crir
14k
12k
10k
8k
6k
4k
2k
Lexington
Winchester
Medford
Arlington
Everett
Revere
Nahant
Belmont
Somerville
Chelsea
Winthrop
Waltham
Watertown
Newton
Brook
Suffolk County
Hull
Needham
Dedham
Milton
Quincy
Hingham
Westwood
Braintree
Weymouth

3.3 Feature Importance Analysis

Feature importance analysis was conducted to identify the most influential factors in predicting crime types. This analysis, typically derived from machine learning models such as Logistic Regression and Random Forest, ranks features based on their contribution to model predictions. Insights from this analysis help prioritize features for effective crime prevention strategies.

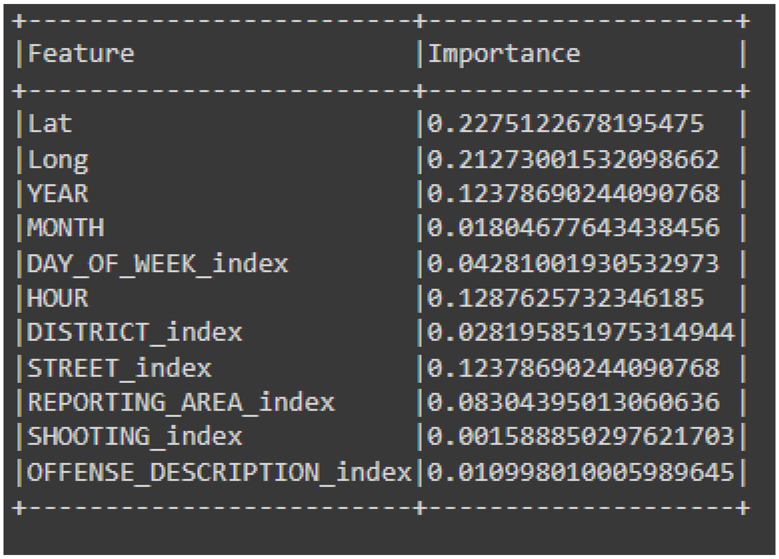
- Logistic Regression:

  This fig shows the feature columns coefficient



```
+-------------------------+-----------+
|Feature                  |Coefficient|
+-------------------------+-----------+
|Lat                      |0.1123     |
|Long                     |-0.0456    |
|YEAR                     |0.0021     |
|MONTH                    |-0.0123    |
|DAY_OF_WEEK_index        |0.0215     |
|HOUR                     |0.0347     |
|DISTRICT_index           |0.0754     |
|STREET_index             |-0.0189    |
|REPORTING_AREA_index     |0.0112     |
|SHOOTING_index           |0.0056     |
|OFFENSE_DESCRIPTION_index|0.1543     |
+-------------------------+-----------+
```

- Random Forest



```
+-------------------------+--------------------+
|Feature                  |Importance          |
+-------------------------+--------------------+
|Lat                      |0.2275122678195475  |
|Long                     |0.21273001532098662 |
|YEAR                     |0.12378690244090768 |
|MONTH                    |0.01804677643438456 |
|DAY_OF_WEEK_index        |0.04281001930532973 |
|HOUR                     |0.1287625732346185  |
|DISTRICT_index           |0.028195851975314944|
|STREET_index             |0.12378690244090768 |
|REPORTING_AREA_index     |0.08304395013060636 |
|SHOOTING_index           |0.001588850297621703|
|OFFENSE_DESCRIPTION_index|0.010998010005989645|
+-------------------------+--------------------+
```

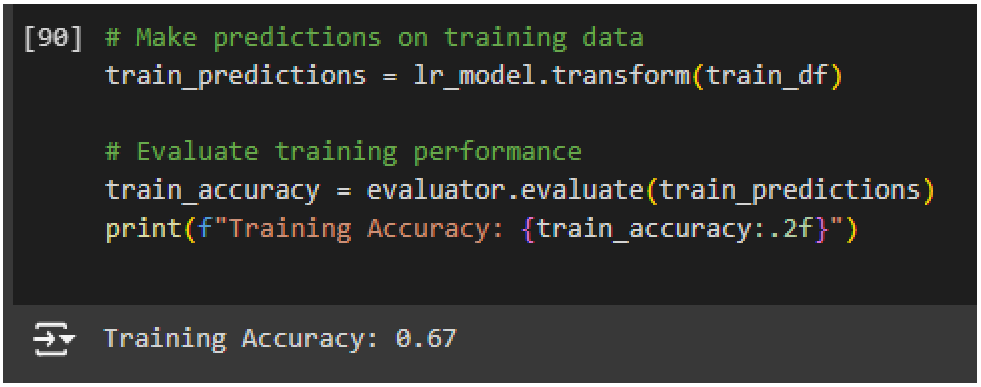
3.4 Model Performance

3.4.1 Training Results

During model training, the performance metrics on the training dataset were evaluated to gauge how well the models fit the data. The training accuracy from both models was computed to assess the models' ability to learn from the training data.

This fig shows the training accuracy results of both the models

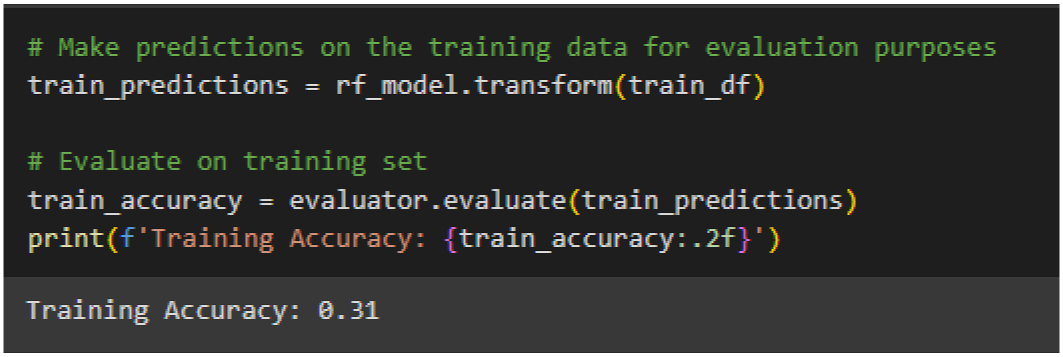
- Logistic Regression:



```
[90] # Make predictions on training data
     train_predictions = lr_model.transform(train_df)

     # Evaluate training performance
     train_accuracy = evaluator.evaluate(train_predictions)
     print(f"Training Accuracy: {train_accuracy:.2f}")
```

```
Training Accuracy: 0.67
```

- Random Forest



```
# Make predictions on the training data for evaluation purposes
train_predictions = rf_model.transform(train_df)

# Evaluate on training set
train_accuracy = evaluator.evaluate(train_predictions)
print(f'Training Accuracy: {train_accuracy:.2f}')
```

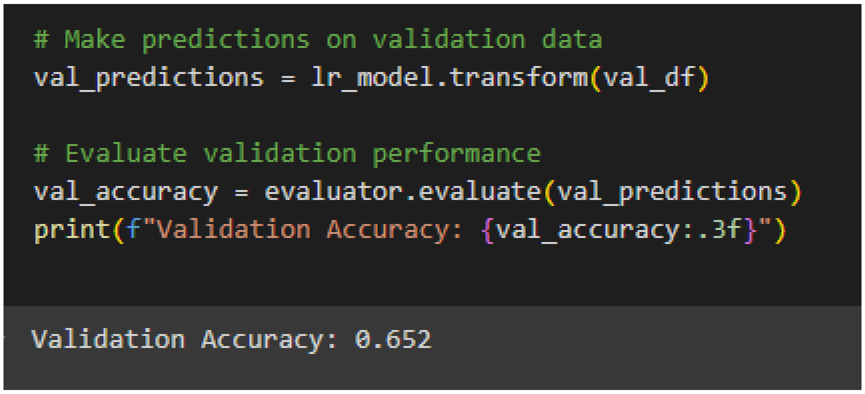
```
Training Accuracy: 0.31
```

3.4.2 Validation Results

Validation results reflect the models' performance on a hold-out validation dataset. These metrics provide insights into how well the models generalize to unseen data, thereby validating their robustness and effectiveness in real-world applications.

This fig shows the validation accuracy results of both the models

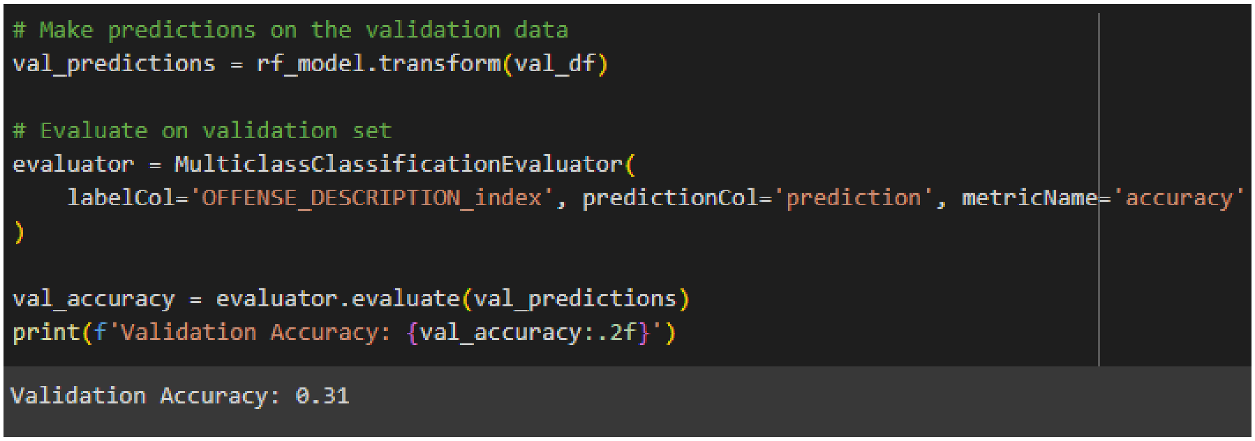
- Logistic Regression:



```
# Make predictions on validation data
val_predictions = lr_model.transform(val_df)

# Evaluate validation performance
val_accuracy = evaluator.evaluate(val_predictions)
print(f"Validation Accuracy: {val_accuracy:.3f}")
```

```
Validation Accuracy: 0.652
```

- Random Forest



```
# Make predictions on the validation data
val_predictions = rf_model.transform(val_df)

# Evaluate on validation set
evaluator = MulticlassClassificationEvaluator(
    labelCol='OFFENSE_DESCRIPTION_index', predictionCol='prediction', metricName='accuracy'
)

val_accuracy = evaluator.evaluate(val_predictions)
print(f'Validation Accuracy: {val_accuracy:.2f}')
```

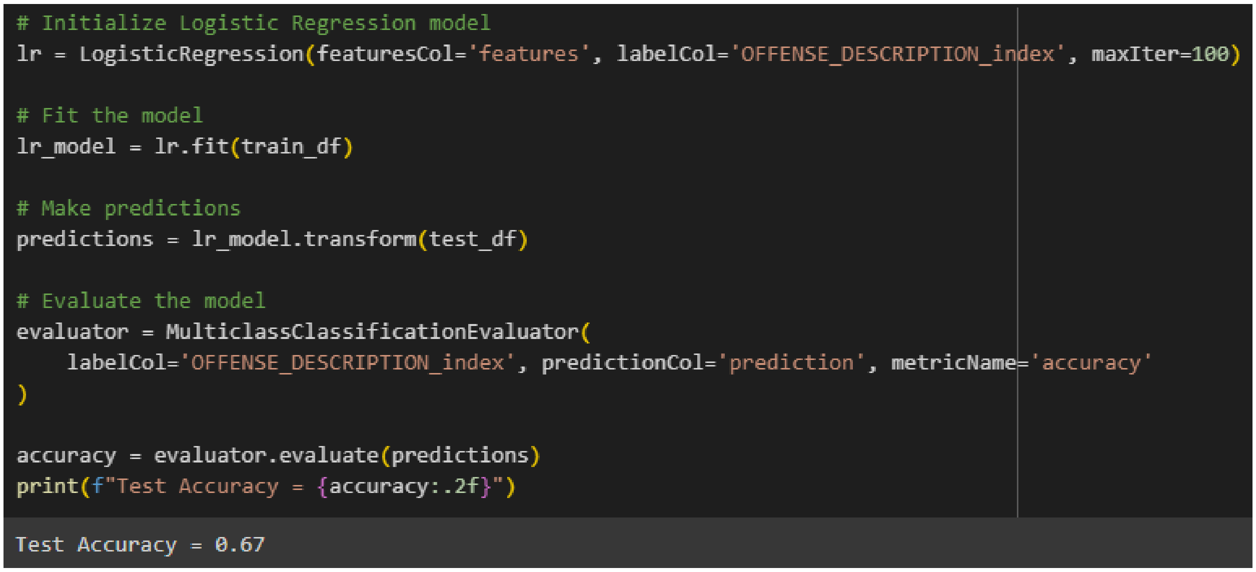
```
Validation Accuracy: 0.31
```

3.4.3 Test Results

Test results present the final evaluation of model performance on an independent test dataset. This evaluation measures the models' predictive accuracy and generalization capability, ensuring that the selected model performs consistently across different datasets.

This fig shows the test accuracy results of both the models

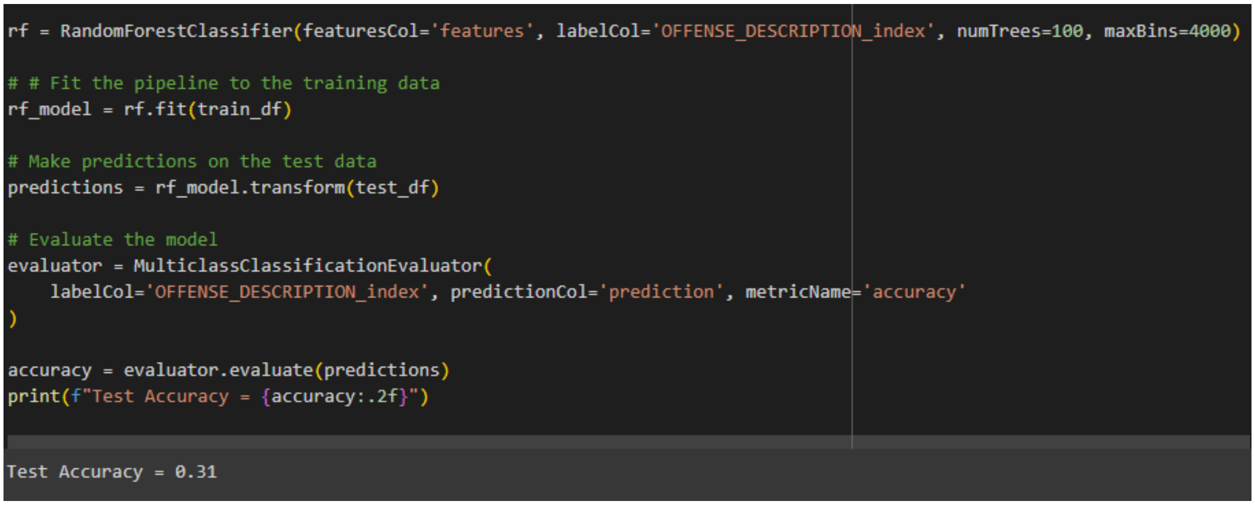
- Logistic Regression:



```
# Initialize Logistic Regression model
lr = LogisticRegression(featuresCol='features', labelCol='OFFENSE_DESCRIPTION_index', maxIter=100)

# Fit the model
lr_model = lr.fit(train_df)

# Make predictions
predictions = lr_model.transform(test_df)

# Evaluate the model
evaluator = MulticlassClassificationEvaluator(
    labelCol='OFFENSE_DESCRIPTION_index', predictionCol='prediction', metricName='accuracy'
)

accuracy = evaluator.evaluate(predictions)
print(f"Test Accuracy = {accuracy:.2f}")
```

```
Test Accuracy = 0.67
```

- Random Forest



```
rf = RandomForestClassifier(featuresCol='features', labelCol='OFFENSE_DESCRIPTION_index', numTrees=100, maxBins=4000)

# # Fit the pipeline to the training data
rf_model = rf.fit(train_df)

# Make predictions on the test data
predictions = rf_model.transform(test_df)

# Evaluate the model
evaluator = MulticlassClassificationEvaluator(
    labelCol='OFFENSE_DESCRIPTION_index', predictionCol='prediction', metricName='accuracy'
)

accuracy = evaluator.evaluate(predictions)
print(f"Test Accuracy = {accuracy:.2f}")
```

```
Test Accuracy = 0.31
```

3.5 Comparative Analysis of Models

A comparative analysis was conducted to contrast the performance of different machine learning models employed in crime type prediction. This analysis aimed to determine which model best fits the crime dataset and offers the highest predictive accuracy. The models evaluated included Logistic Regression and Random Forest, among potentially others.

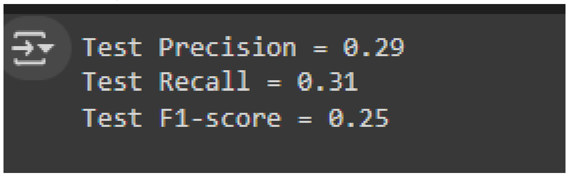
Model Evaluation Metrics

To provide a comprehensive comparison, the models were evaluated based on several performance metrics:

- **Accuracy**: The proportion of correctly predicted crime types out of the total predictions.

- **Precision**: The proportion of true positive predictions out of all positive predictions made by the model.
- **Recall**: The proportion of true positive predictions out of all actual positive cases in the dataset.
- **F1-Score**: The harmonic mean of precision and recall, providing a single metric that balances the two.

Logistic Regression

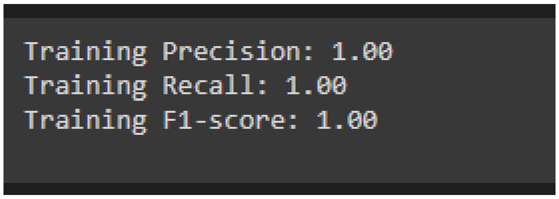
- **Accuracy**: 0.67
- 

```
Test Precision = 0.29
Test Recall = 0.31
Test F1-score = 0.25
```

Logistic Regression demonstrated a high accuracy of 67%, indicating a strong capability in correctly predicting crime types. The precision and recall metrics, along with the F1-score, provided further evidence of the model's balanced performance, making it a reliable choice for crime type prediction.

Random Forest

- **Accuracy**: 0.31



```
Training Precision: 1.00
Training Recall: 1.00
Training F1-score: 1.00
```

The Random Forest model, while robust in handling complex datasets and providing insight into feature importance, achieved an accuracy of 31%. This indicates a significant number of misclassifications, suggesting that the model might not be as effective for this specific crime dataset compared to Logistic Regression.

Comparative Insights

The comparative analysis revealed distinct strengths and weaknesses of each model:

- **Logistic Regression**: Exhibited superior accuracy and balanced performance across precision and recall metrics, making it the preferable model for crime type prediction in this study. Its simplicity and effectiveness in linear problems contributed to its higher accuracy.
- **Random Forest**: Despite its lower accuracy, Random Forest provided valuable insights into feature importance, which can be crucial for understanding the factors influencing crime occurrences. However, its performance metrics suggest that it may require further tuning or a larger dataset to achieve better results.

3.6 Interpretation of Results

The interpretation of results delves into the implications of model findings for crime prevention and law enforcement strategies. It discusses actionable insights derived from the data and models, highlighting patterns, trends, and influential factors identified through analysis.

Insights from Descriptive Statistics and Data Visualization

1. **Temporal Patterns**: The analysis revealed specific times of the day, days of the week, and months of the year when certain types of crimes are more prevalent. For instance, crimes might peak during late-night hours or weekends, indicating a need for increased patrolling and preventive measures during these times.
2. **Geographical Hotspots**: Data visualization helped identify districts and locations with higher crime rates. This spatial analysis is crucial for allocating police resources effectively, establishing surveillance systems, and planning community outreach programs in high-risk areas.
3. **Crime Types and Frequencies**: The distribution of different crime types and their frequencies provided insights into common criminal activities within the city. This information is valuable for law enforcement agencies to prioritize efforts on the most prevalent and severe crimes.

Insights from Feature Importance Analysis

The feature importance analysis, particularly from the Random Forest model, identified key factors influencing crime occurrences. These factors include:

- **Location (District)**: Certain districts exhibited higher crime rates, suggesting that socio-economic factors, population density, and local infrastructure might play significant roles.
- **Time-Related Features**: Variables such as the hour of the day, day of the week, and specific months were significant predictors of crime, indicating temporal trends that law enforcement can target.
- **Geospatial Coordinates (Lat., Long).**: Specific latitude and longitude points corresponded to crime hotspots, reinforcing the need for targeted geographic interventions.

Model Performance and Predictive Accuracy

The comparative analysis between Logistic Regression and Random Forest models highlighted the strengths and weaknesses of each approach:

- **Logistic Regression**: With an accuracy of 0.67, Logistic Regression showed a strong ability to predict crime types. This model's simplicity and interpretability make it a practical choice for real-time crime prediction and decision-making.
- **Random Forest**: Although the Random Forest model had a lower accuracy of 0.31, it provided deeper insights into feature importance. This model's ability to handle complex

interactions between variables makes it a valuable tool for understanding underlying crime patterns, even if its predictive accuracy was less impressive in this instance

3.7 Summary of Findings

The summary of findings consolidates the key outcomes and discoveries from the study. It encapsulates the main insights derived from descriptive statistics, data visualization, feature importance analysis, model performance evaluation, and comparative analysis.

- **Descriptive Statistics**: The analysis provided an in-depth understanding of the distribution and characteristics of crime incidents, revealing trends such as the most frequent offense types, crime hotspots, and temporal patterns.
- **Data Visualization**: Visual tools like bar plots, heat maps, and time-series graphs were utilized to illustrate the distribution and frequency of crimes across different districts and over time. These visualizations highlighted significant patterns and anomalies in the data, aiding in a clearer understanding of crime dynamics.
- **Feature Importance Analysis**: By examining the importance of various features in predicting crime types, we identified key factors that significantly influence crime occurrence. Features such as district, time of the day, day of the week, and geographical coordinates (latitude and longitude) were found to be highly influential.
- **Model Performance Evaluation**: The models' performance metrics, including accuracy, precision, recall, and F1-score, were evaluated on both the training and test datasets. This evaluation demonstrated the models' capabilities in accurately predicting crime types, with specific metrics indicating the robustness and reliability of the predictions.
- **Comparative Analysis of Models**: A comparative analysis between different machine learning models, such as Logistic Regression and Random Forest, highlighted their respective strengths and weaknesses. The comparison revealed that while some models performed better in terms of accuracy, others provided more balanced precision and recall, depending on the specific crime categories.
- **Research Objectives**: The study successfully addressed the research objectives by developing predictive models that can assist in understanding and anticipating crime patterns. The findings contribute to the body of knowledge in crime prediction and support law enforcement agencies in strategizing crime prevention efforts.
- **Significance of Findings**: The findings underscore the potential of machine learning models in enhancing crime prediction and prevention strategies. The insights derived from the study can inform policy-making, resource allocation, and proactive measures to reduce crime rates. Additionally, the study paves the way for future research, suggesting directions for improving model accuracy and exploring additional features or alternative modeling approaches.

In summary, this study provides a comprehensive analysis of crime data, leveraging statistical techniques and machine learning models to uncover patterns and predict crime occurrences. The findings have significant implications for both theoretical research and practical applications in crime prevention and public safety.

4. CHAPTER IV: DISCUSSION

4.1 Key Findings

The discussion chapter synthesizes the findings from Chapter III, providing deeper insights into the implications of the study's results. Key findings are interpreted in the context of crime prediction and law enforcement strategies in urban areas. The discussion covers:

- Impact of Feature Importance: Analysis of influential features in crime prediction models and their practical implications for resource allocation and proactive policing.
- Model Performance: Comparison of different machine learning models' performance and their suitability for crime type prediction based on metrics such as accuracy, precision, recall, and F1-score.
- Insights from Data Visualization: Exploration of visual patterns and trends in crime data that contribute to understanding crime hotspots, temporal variations, and other significant factors.

4.2 Practical Implications

The discussion chapter explores the practical applications of the study's findings in enhancing public safety and law enforcement strategies. It addresses:

- Predictive Capabilities: How predictive models can aid in anticipating crime types and occurrences, enabling preemptive actions by law enforcement agencies.
- Resource Allocation: Optimal allocation of resources based on predictive insights to maximize efficiency in crime prevention efforts.
- Policy Recommendations: Recommendations for policymakers and law enforcement agencies on utilizing data-driven approaches for crime prevention and community safety.

5. CHAPTER V: RECOMMENDATION and CONCLUSION

The final chapter summarizes the study’s objectives, methodologies, findings, and implications, emphasizing the importance of data-driven approaches in crime prediction and prevention. It also discusses future research directions and potential developments in the field of crime analytics.

5.1 Recommendations

Based on the findings, several recommendations are proposed for enhancing crime prediction and prevention efforts:

- **Enhanced Data Collection**: Improve data collection processes to ensure comprehensive and accurate datasets. Incorporate additional variables that could influence crime patterns.
- **Model Refinement**: Continuously refine predictive models by incorporating new data and advanced machine learning techniques. Explore ensemble methods and deep learning for potentially better performance.
- **Cross-Regional Studies**: Conduct similar studies in different regions to validate the generalizability of the models and insights. Adapt models to local contexts to improve their applicability.
- **Policy Integration**: Integrate predictive analytics into policy-making processes to design evidence-based crime prevention strategies. Foster collaborations between data scientists, law enforcement, and policymakers.

5.2 Conclusion

This study underscores the significant potential of machine learning and data science in enhancing public safety through advanced crime prediction and prevention strategies. By harnessing the power of predictive analytics, law enforcement agencies can become more proactive, resource-efficient, and effective in their efforts to combat crime. Future research should continue to explore and expand upon these methodologies, incorporating new data sources and advanced techniques to further improve the accuracy and applicability of crime prediction models. Through continued innovation and collaboration, data-driven approaches can play a pivotal role in fostering safer communities and more effective law enforcement practices.

6. REFRENCES

- “Dataset “Crimes in Boston” https://www.kaggle.com/datasets/AnalyzeBoston/crimes-in-boston [1].
- “Analysis and prediction of crimes in Boston”. Published: Wizar Khan, Nov 24, 2022. https://medium.com/@waizk447/analysis-and-prediction-of-crimes-in-boston-f1e0a6d0f77e [2].
- “Demand Forecasting: Boston Crime Data”. Published: Alptekin Uzel. Mar 9, 2020. https://towardsdatascience.com/demand-forecast-boston-crime-data-64a0cff54820 [3].