KLE Society's

KLE Technological University



**Exploratory Data Analysis**

**(22ECAC210)**

**Course Project Report on**

**“Boston Crime Analysis”**

*Submitted in partial fulfilment of the requirement for the award of*

**Degree of Bachelor of Engineering**

**in**

**Computer Science and Engineering**

*Submitted By*

**Mohd. Arman Khan – 01fe21bci037**

**Siddharth J K- 01fe21bci052**

**Sameer Nadar – 01fe21bci352**

**Mahendra Addi I - 01fe21bci351**

**Ayesha Mulla- 01fe22bci405**

*Submitted To*

**SCHOOL OF COMPUTER SCIENCE & ENGINEERING,**

KLE TECHNOLOGICAL UNIVERSITY

HUBLI – 580 031 (India).

Abstract

The city of Boston has long been a focal point for crime analysis due to its dynamic urban landscape and diverse population. This report delves into a comprehensive study of Boston's crime patterns, aiming to uncover key insights that can aid law enforcement agencies, policymakers, and communities in developing effective strategies to address crime and enhance public safety. Through an extensive examination of crime data spanning several years, this analysis identifies prevalent crime types and their distribution across various neighbourhoods. The report investigates the temporal trends of criminal activities, exploring whether certain crimes exhibit seasonality or fluctuations over time.Furthermore, the study explores potential socio-economic and demographic factors that may influence crime rates in different areas of the city. By investigating correlations between crime and factors such as poverty, education levels, and unemployment rates, we gain valuable knowledge to inform targeted intervention programs. An essential aspect of the analysis involves mapping crime hotspots using geospatial tools to visualize the concentration of criminal incidents in specific regions. This spatial understanding helps identify areas that require immediate attention and resource allocation to prevent and combat criminal activities effectively. Moreover, this report assesses the impact of law enforcement strategies and community-oriented policing initiatives on crime reduction. Evaluating the effectiveness of various intervention programs provides valuable insights into the most successful approaches for crime prevention and control.

The findings of this research offer significant implications for Boston's law enforcement agencies and policymakers. By understanding the city's crime patterns and underlying factors, stakeholders can implement evidence-based policies and initiatives that prioritize crime prevention and community safety. However, it is essential to acknowledge the limitations of this analysis, such as potential data gaps and methodological constraints. Nevertheless, this study serves as a crucial foundation for further research and the development of targeted strategies to combat crime in Boston effectively. Ultimately, by fostering a collaborative approach between law enforcement, government agencies, and local communities, Boston can aspire to become a safer and more secure city for all its residents.

Table of Content:

1. Introduction 4

1.1 Overview of the project 4

1.2 Importance of EDA in data analysis 4

1.3 Objectives of the course project 5

2. Data Collection 5

3. Data Exploration 5

4. Data Cleaning 7

4.1 Steps in data cleaning 7

4.2 Data cleaning in dataset 8

5. Feature Selection 8

5.1 Feature selection for prediction of district of crimes 8

5.2 Feature selection for prediction of number of crimes…..............................................9

6. Data Analysis

6.1 Steps for data analysis……………………………………………...………...………9

6.2 Model Training……………………………………………………………………... 9

7. Conclusions……...…...…………………………………………………………………..10

8. References……………………………………….……………………………………….11

# **1.Introduction**

## **1.1** **Overview of the project**

The Boston Crime Dataset is a comprehensive collection of reported criminal incidents have occurred within the city of Boston. It includes various types of crimes, such as thefts, assaults, burglaries, robberies, drug-related offenses, and more. The dataset typically comprises information from multiple years, allowing for longitudinal analyses and the detection of long-term trends.

## **1.2 Importance of EDA in data analysis**

Exploratory Data Analysis (EDA) is essential in data analysis as it helps understand data patterns, identify missing values, and detect outliers. It aids in making data-driven decisions, selecting appropriate models, and improving feature engineering for enhanced predictive performance. EDA uncovers insights that guide hypothesis testing and highlights potential biases, enabling data professionals to communicate findings effectively through data visualization. It serves as a critical initial step, providing a solid foundation for successful data analysis and informed decision-making. It is an iterative process that lays the ground work for successful and meaningful data analysis

## **Objectives of the course project**

The objectives of crime analysis in the Boston Dataset are multi-faceted and aimed at gaining insights into criminal activities within the city. By conducting crime analysis on this dataset, law enforcement agencies, researchers, and policymakers can achieve the following objectives:

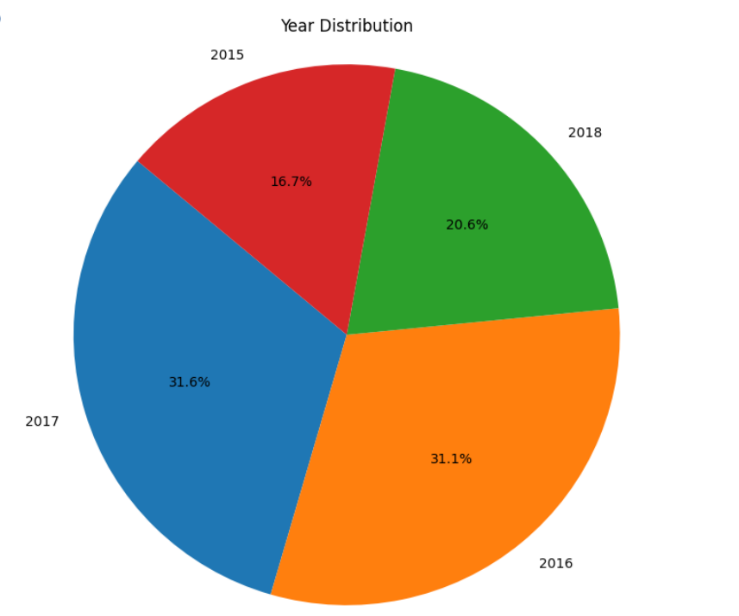
* Crime Pattern Identification: The primary objective of crime analysis is to identify patterns and trends in criminal activities within different areas of Boston. By analysing the spatial and temporal distribution of crimes, analysts can identify crime hotspots and understand where and when certain types of crimes are more prevalent.
* Hotspot Identification and Resource Allocation: Crime analysis helps in identifying high-crime areas or hotspots. Law enforcement agencies can then allocate their resources strategically to these areas, focusing on crime prevention and reducing criminal incidents.
* Crime Type Analysis: Analysing different types of crimes in the dataset allows researchers to understand the prevalence and severity of specific offenses. This insight can guide the allocation of resources and the development of specialized crime prevention strategies for different crime categories.
* Identification of High-Priority Offenders: Data analysis can help in identifying repeat offenders or individuals associated with multiple criminal incidents. This information aids in prioritizing investigative efforts and targeting resources towards apprehending high-priority offenders.
* Support for Policy Decisions: Crime analysis provides evidence-based insights that support policy and decision-making processes related to public safety, law enforcement strategies.
* Public Safety Communication: Crime analysis findings can be used to inform the public about crime trends, safety concerns, and precautionary measures. Timely and accurate information enhances public awareness and encourages community involvement in crime prevention efforts.
* Predictive Policing: By applying predictive modelling techniques to historical crime data, crime analysis can help law enforcement predict potential crime hotspots or emerging crime patterns. This proactive approach allows for the prevention of criminal activities before they occur.

# **2.Data Collection**

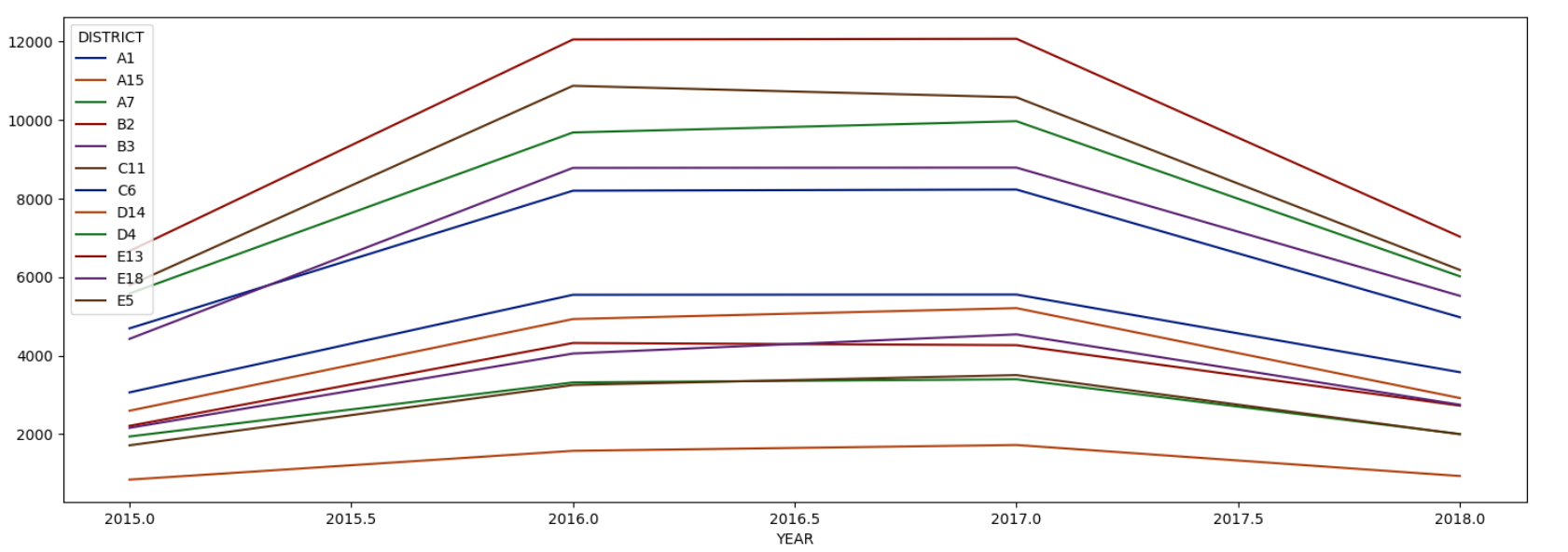
In this study on Boston crime analysis and crime prediction using machine learning and exploratory data analysis (EDA), the primary data source utilized is a structured collection of crime-related data points obtained from Kaggle, a reputable online platform for sharing datasets and data science resources. The dataset serves as the foundation for conducting thorough EDA to extract meaningful insights and patterns that can aid in understanding crime dynamics within the city and subsequently inform effective crime prevention and law enforcement strategies. The dataset obtained from Kaggle comprises diverse observations or samples representing various criminal incidents reported in Boston over a significant period. Each data point includes essential attributes, such as crime type, location coordinates, date and time of occurrence, and other relevant information. By accessing this comprehensive dataset, we gain access to a wealth of information that allows us to analyse crime patterns, identify potential hotspots, analyse temporal trends, and develop predictive models.

**3. Data Exploration**

The visual analysis of the data presented in the graph reveals several notable insights. Firstly, the correlation between the "Month" and "Year" variables shows a negative trend, suggesting that as the years progress, crime rates tend to decrease month by month. On the other hand, the "Offense Code" and "Year" variables exhibit a moderately positive correlation, indicating a potential increase in reported offenses over the years. Additionally, the "Latitude" and "Longitude" variables demonstrate a moderately negative correlation, implying that crimes tend to cluster in specific geographical regions.Moving on to the boxplots, they are all uniformly blue and display similar size and shape across the 16 different districts, implying comparable distributions of offenses in each district. The tree map visualizes various crime types, represented by differently colour rectangles, with their sizes proportional to the number of crimes committed. The most frequent crime is "Motor Vehicle Accident Response" with 30,264 incidents, closely followed by "Larceny" with 24,900 incidents. Other crime types, such as vandalism, simple assault, drug violations, investigate person, property lost, and fraud, are also present in the dataset.Regarding the time of day, the graph indicates that the highest crime count occurs at 17:00 (5:00 PM), while the lowest count is observed at 5:00 (5:00 AM).



The pie chart further illustrates the distribution of crimes across four years, with the largest portion representing 2017 at 31.6%, followed by 2016 at 31.1%. The years 2018 and 2015 account for 20.6% and 16.7% of the distribution, respectively, showcasing a general decrease in crime rates over the analysed years.



Lastly, the lines depicting the correlation values from "Offense Code" to "Longitude" exhibit a declining trend, reinforcing the notion that the variables become less correlated as we move from offense codes to geographical coordinates.These visual insights provide valuable cues for further analysis and model development, enabling a deeper understanding of crime patterns and trends in the city of Boston.

**4. Data Cleaning**

**4.1 Steps in Data cleaning**

Data cleaning is a critical step in the data analysis process that involves identifying and rectifying errors, inconsistencies, and inaccuracies in the dataset. Here are the essential steps for data cleaning:

* Identify and Handle Missing Values**:** Check for missing data in the dataset and decide how to handle them. You can choose to drop rows with missing values, impute missing values with mean/median/mode, or use advanced imputation methods depending on the dataset and the missing value patterns.
* Remove Duplicate Entries: Look for duplicate rows or observations in the dataset and eliminate them to avoid biased analysis or model training.
* Correct Data Types: Ensure that each attribute has the correct data type. For example, categorical variables should be represented as factors, dates should be in datetime format, and numerical variables should be represented as integers or floats.
* Check for Outliers: Identify outliers that might impact the analysis or modelling process. Decide whether to remove them, transform the data, or handle them using robust statistical methods.
* Normalize Data: If required, normalize numerical data to bring them to a similar scale, especially when using algorithms sensitive to the magnitude of attributes.
* Handle Inconsistent or Erroneous Values: Identify any inconsistent or erroneous values that do not make logical sense and rectify them. This may involve cross-referencing with external data sources or using domain knowledge.
* Validate and Standardize Data: Validate categorical variables to ensure they contain only valid categories. Standardize text data by converting it to a consistent case (e.g., lowercase) for easier analysis.
* Encode Categorical Variables: Convert categorical variables into numerical representations using one-hot encoding, label encoding, or other suitable techniques for analysis or model training.
* Remove Irrelevant or Redundant Features: Eliminate attributes that do not contribute significantly to the analysis or modelling process to simplify the dataset and improve computational efficiency.
* Address Data Integrity Issues: Detect and correct any data integrity problems, such as referential integrity issues in relational databases.
* Feature Engineering: Create new features or transform existing ones to extract more relevant information and improve the performance of predictive models.

**4.2 Data cleaning in Dataset**

In the data preprocessing phase of our analysis on the Boston crime dataset, we encountered several issues related to missing values and data format. Notably, the "District" column contained 1765 null values, representing less than 0.6% of the total rows in the dataset. As these null values cannot be replaced with any specific value, we made the decision to drop these rows from the dataset to maintain data integrity and ensure accurate analysis. To enhance the temporal analysis of crime incidents, we addressed the data type of the "OCCURRED\_ON\_DATE" column. Originally stored as an object, we converted it to the datetime format, allowing for more precise and meaningful temporal exploration. Another challenge we encounteredwas the presence of duplicate incident numbers, as multiple offense types could be associated with the same crime incident. To accurately identify and analyse unique incidents, we created a copy of the dataset and removed duplicate rows based on incident numbers, ensuring that each incident was represented only once in the dataset. One significant issue we faced was the substantial number of null values in the "Shooting" column, accounting for approximately 99% of the data. However, due to the column's importance for our predictive modelling, we made the strategic decision not to drop it from the dataset. Instead, we replaced the null values in the "Shooting" column with the appropriate value "NO" to preserve the data's utility while addressing the missing information.Through these data preprocessing steps, we aimed to enhance the quality and usability of the dataset, enabling more accurate and insightful crime analysis and prediction. By carefully handling missing values, data types, and duplicate entries, we ensure the dataset's reliability, laying a robust foundation for the subsequent stages of our analysis and modelling efforts.

**5. Feature selection**

Feature selection plays a crucial role in the predictive modelling process, as it involves identifying and selecting the most relevant attributes from the dataset that significantly contribute to the prediction of the target variable, which in this case is the district where a crime occurred. The goal is to create an efficient and accurate predictive model while reducing the risk of overfitting and improving model interpretability.

**5.1. Feature selection for predicting district of crimes**

For our crime dataset, we are attempting to predict the district based on several attributes. The x\_train attributes used for training our predictive model include "Offense Code," "Offense Code Group," "Shooting," "Year," "Month," "Day," "Hour," "Street," "Latitude," and "Longitude."The process of feature selection entails several methods, such as univariate feature selection, recursive feature elimination, and feature importance techniques provided by machine learning algorithms like Random Forest or Gradient Boosting models. These methods aid in identifying the most influential attributes that contribute to predicting the target variable, "District," effectively.Moreover, domain knowledge and insights gained from exploratory data analysis (EDA) can also guide feature selection. Attributes that have a significant impact on crime occurrences and district patterns, as identified during EDA, can be given priority in the selection process.By carefully choosing relevant features, we aim to create a predictive model that accurately associates the various attributes with the correct district where a crime took place. Effective feature selection ensures a streamlined and optimized model, improving its performance and applicability in real-world scenarios.As we proceed with the feature selection process, we will assess the importance and relevance of each attribute in predicting the district and retain only those attributes that add significant value to our predictive model. This streamlined set of attributes will form the foundation of our model, allowing us to develop a reliable crime prediction system that can aid law enforcement agencies in allocating resources and implementing targeted crime prevention strategies in the city of Boston.

**5.2. Feature selection for predicting number of crimes**

We are predicting the number of crimes using a time series approach with the attribute "Occurred\_On\_Date." By leveraging time series analysis, we can detect and analyse patterns, trends, and seasonality in crime incidents over time. This enables us to model and forecast the future number of crimes in each district, helping law enforcement agencies allocate resources more effectively and implement targeted crime prevention strategies.Through careful feature selection and time series analysis, we aim to develop a robust predictive model that can accurately associate the attributes with the correct district and provide valuable insights into crime patterns and trends in the city of Boston. This comprehensive approach to feature selection and time series modelling will contribute to enhancing public safety and supporting law enforcement efforts in the proactive management of crime in the region.

**6. Data Analysis**

**6.1 Steps for Data analysis**

* Choose a Model: Select an appropriate machine learning algorithm that suits the nature of the problem (classification, regression, etc.) and the characteristics of the dataset.
* Model Initialization: Initialize the selected model with appropriate hyperparameters. Hyperparameters are tuning parameters that control the learning process of the model.
* Model Training: Train the model on the training dataset. The model learns the patterns and relationships between the input features and the target variable.
* Model Evaluation: Evaluate the trained model using the testing dataset. Various evaluation metrics such as accuracy, precision, recall, F1-score, and mean squared error (MSE) are used based on the type of problem (classification or regression).
* Hyperparameter Tuning: Optimize the model's hyperparameters to improve its performance. This can be done using techniques like grid search, random search, or Bayesian optimization.
* Cross-Validation: Perform cross-validation to assess the model's generalization performance and reduce the risk of overfitting. Common methods include k-fold cross-validation and stratified cross-validation.
* Finalize the Model: After achieving satisfactory performance on the testing set, finalize the model for deployment. Re-train the model on the entire dataset if necessary.

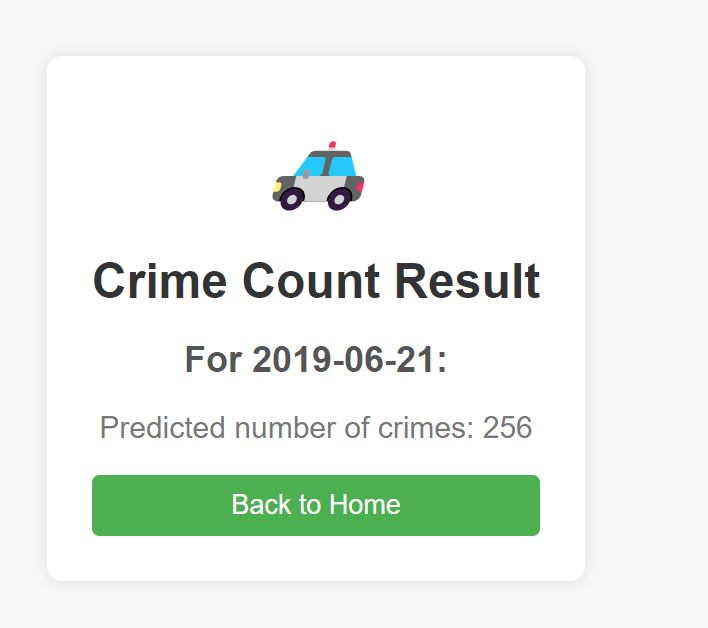
**6.2 Model Training**

In our crime data analysis, we adopted a comprehensive approach to build predictive models for district prediction and crime forecasting using time series analysis. The dataset was divided into a 70:30 split, with 70% used for model training and the remaining 30% for evaluation. For district prediction, we explored three different machine learning algorithms: K-Nearest Neighbours (KNN), Decision Tree, and Random Forest Classifier. Each algorithm was trained on the training dataset, utilizing features such as "Offense Code," "Offense Code Group," "Shooting," "Year," "Month," "Day," "Hour," "Street," "Latitude," and "Longitude" to predict the district where crimes occurred. We evaluated the models using the testing dataset, comparing performance metrics such as accuracy, precision, recall, and F1-score. The goal was to identify the best-performing model that could accurately predict the districts associated with different types of crimes.In addition to district prediction, we also incorporated time series analysis to forecast the number of crimes over a period. To achieve this, we employed Facebook's Prophet model, a powerful tool for time series forecasting. Leveraging the "Occurred\_On\_Date" attribute, we analysed patterns and trends in crime occurrences over time. The Prophet model learned from historical crime data and generated forecasts for future crime counts, offering valuable insights to law enforcement agencies for resource allocation and proactive crime prevention efforts.By combining traditional machine learning algorithms for district prediction with cutting-edge time series analysis using Prophet, we aimed to create a comprehensive crime analysis framework. This approach allows us to gain a holistic understanding of crime patterns, trends, and associations with specific districts, enabling law enforcement agencies and policymakers to make data-driven decisions for enhancing public safety and improving crime prevention strategies in the city of Boston.

Following accuracies were obtained while predicting district using these models:

|  |  |
| --- | --- |
| Models | Accuracy |
| KNN classifier | 80% |
| Decision Tree classifier | 98% |
| Random forest classifier | 96% |

Predicting number of crimes on a given date using Facebook’s Prophet model:

**7.Conclusions**

Feature selection played a crucial role in the prediction of districts for crime incidents. By carefully selecting relevant attributes, such as "Offense Code," "Shooting," "Year," "Month," and others, we were able to train effective predictive models using machine learning algorithms like KNN, Decision Tree, and Random Forest Classifier. These models accurately predicted the districts associated with crimes, enabling law enforcement agencies to allocate resources more efficiently and implement targeted crime prevention strategies.In addition to district prediction, we employed time series analysis using Facebook's Prophet model to forecast the number of crimes over time. Leveraging the "Occurred\_On\_Date" attribute, we detected patterns, trends, and seasonality in crime incidents. The Prophet model provided valuable insights into future crime counts, empowering law enforcement agencies with proactive crime management strategies.

Our data-driven approach to crime analysis presented a comprehensive understanding of crime patterns and trends in Boston. By combining traditional machine learning techniques with advanced time series analysis, we have laid a strong foundation for data-informed decision-making in crime prevention and law enforcement efforts. This analysis serves as a valuable resource for city officials and law enforcement agencies to devise effective strategies, allocate resources efficiently, and create a safer and more secure environment for the residents of Boston.

**8.References**