AMS 597 Statistical Computing Project

NYC Crime Report Analysis in R

Professor:
Dr. Silvia Sharna

Group Details:

1. Varshaa Sai Sripriya Saisheshadhri, 116732152
2. Swati Swati, 116778659
3. Dhruv Rathee, 116633028
4. Kabir Manoj Ohekar, 116793038
5. Iftekar Alam, 113269424

I. INTRODUCTION

When individuals make life decisions—from selecting neighborhoods to planning travel routes—safety considerations often top their priority list. Crime statistics, while imperfect, provide tangible metrics that help quantify risk in different environments. These numbers offer valuable context for making well-reasoned choices about where to spend time, invest resources, or build communities.

The metropolis of New York stands as an exceptional case study for examining safety patterns. New York's landscape—split into Brooklyn, Manhattan, Queens, The Bronx, and Staten Island—showcases remarkable variety in neighborhood character, wealth distribution, and safety challenges. Law enforcement authorities have taken meaningful steps toward openness by publishing comprehensive crime records for public examination. This factual foundation gives safety experts the tools to examine how criminal activity shifts throughout the year, concentrates in certain areas, and manifests across different violation types. When researchers overlay these patterns with major city events—financial hardships, street festivals, new enforcement approaches, or grassroots movements—they begin uncovering meaningful connections that explain why crime rises or falls. Looking at stark differences between troubled blocks and peaceful streets offers valuable lessons, especially when considering factors such as income levels, government resource allocation, residential stability, and neighborhood engagement efforts.. These multi-dimensional examinations move beyond simple crime tallies toward a more nuanced understanding of neighborhood vulnerability and resilience.

The analytical approach to crime data transcends basic counting exercises. The most valuable insights emerge from detecting subtle correlations and underlying structures within seemingly disconnected incidents. Methodologies including multivariate analysis, geospatial clustering, and temporal forecasting allow investigators to evaluate relationships between crime occurrences and variables such as population characteristics, economic opportunity, educational access, or even environmental factors like seasonal changes. These analytical frameworks support proactive policy development, resource allocation optimization, and targeted prevention strategies.

The R programming environment offers particularly robust capabilities for this specialized analysis work. Its extensive library ecosystem supports the entire analytical workflow—data acquisition and transformation through tidyverse tools, visual representation via ggplot2's expressive grammar, statistical modeling with specialized packages, and machine learning applications for pattern recognition. R enables analysts to implement sophisticated methodologies including spatiotemporal trend detection, regression modeling for factor identification, and intervention analysis for evaluating program effectiveness. This technical capability transforms raw incident data into actionable intelligence for municipal leaders, community advocates, and public safety professionals.

By leveraging accessible crime data, contemporary computational methods, and contextual understanding, researchers can illuminate the intricate factors influencing safety outcomes across New York's varied landscape—ultimately contributing to more effective strategies for cultivating secure, equitable urban environments.

II. LITERATURE SURVEY

Crime forecasting has come to be recognized as an essential field of inquiry in modern policing, with a focus with regards to providing proactive advances in public security [22]. On the basis of analysis of historical crime data and additional inputs, various methodologies are being developed to forecast upcoming delinquent activity [33]. These efforts look to provide policing agencies, policymakers, and scholars with insights into crime patterns and trends, towards improved crime prevention and response [22]. Merging advanced computational methods with traditional legal processes provides an extendable means of enhancing the efficiency and neutrality of judging in the domain of crime reduction [27]. The quintessential goal of criminality prognosis investigation is to design intelligent and scalable systems that are beyond traditional reactive systems to forecast and prevent crime in advance [28]. Leveraging past data and real-time intelligence, the goal is to supply law enforcement with tools to forecast and prevent crime accurately [28]. This can lead to improved efficiency of smart city surveillance systems and a significant decrease in response times [28]. Moreover, awareness of crime patterns and trends can also help decision-makers implement effective actions and increase safety for the public [32]. Identifying high-crime areas and predicting future crime can also help the public avoid crime areas and use alternative routes [23]. The eventual intention is to provide more secure and safer environments for individuals and societies [22].

Recent advances in machine learning have demonstrated noteworthy potential for crime projection and pattern recognition in metropolitan environments. Various regression models are used to forecast crime rates and patterns. These include Linear Regression [35], Polynomial Regression [32], Multilinear Regression [31], Lasso Regression, Support Vector Regression (SVR), Bayesian-Ridge, and Elastic-Net [23]. For instance, Linear Regression has been used to forecast the upcoming trend of crime.

Regression models were applied in a number of studies to anticipate crime rates. For example, with a 94.17% accuracy rate and a 94.59% validation accuracy, Linear Regression was utilized to forecast crime rate trends in Bangladesh [34] and future crime patterns in India [32]. Although the LASSO model was found to predict at a very high rate with 99% accuracy for the given dataset, Polynomial Regression and Random Forest Regression were also applied in the context of crimes against women in India [32]. SVR proved to be the most suitable regression algorithm for predicting crime rates, according to a dissimilar research that used Bayesian-Ridge, Lasso, Random Forest, Extra-Trees, Linear, and SVR regression algorithms [23]. For finding patterns and trends in crime data, crime analysis involved using clustering algorithms such as K-means. K-means clustering was used in one study to predict crimes with an accuracy of over 75% [34] . In a different case, high-risk areas were identified using K-means, and when combined with real-time monitoring, a Bayesian network was employed to forecast crimes with an overall system accuracy of 92% [28].

Apart from that, classification algorithms are also utilized. K-means clustering, Naive Bayes, Decision Trees, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Neural Networks, XGBoost, CatBoost, and Tabnet are among the frequently employed methods.Furthermore, other auto-regression techniques [25] and ARIMA (Auto Regressive Integrated Moving Average) [35] are employed

to forecast the future and examine temporal patterns in crime data. ARIMA model has been used to forecast crime successfully [35].

Several studies have specifically examined New York City's crime patterns using various analytical approaches, providing valuable insights for borough-specific crime classification. Kumar et al. (2024) established a strong foundation by applying Random Forest and K-Means clustering to NYPD data from 2006-2017, revealing distinct spatial-temporal patterns across boroughs that could inform resource allocation strategies [1]. Their work highlighted the importance of borough-specific analysis, as crime hotspots showed significant variation between different NYC neighborhoods. Previous papers have done extensive work on comparing the effectiveness of different machine learning models on crime analysis. Kumar et al. [2] established baseline performance using Random Forest (RF) and K-Means clustering on NYPD data (2006-2017), demonstrating RF's effectiveness for initial borough-level classification. Almuhanna et al. [3] expanded this work with a direct comparison showing XGBoost's superiority (52% accuracy) over SVM and RF for multi-class crime prediction. These findings were corroborated by Kshatri et al. [14], whose stacked ensemble (SVM meta-classifier) achieved remarkable 99.5% accuracy on Indian crime data, suggesting ensemble methods may be particularly effective for NYC's borough-specific challenges.

However, model performance varies significantly by crime type and location. Sharma et al. [4] found PCA-enhanced Random Forest achieved 60% accuracy in Boston - notably higher than standalone models - highlighting the potential of feature engineering for urban crime prediction. Similarly, McClendon and Meghanathan [11] demonstrated linear regression's effectiveness for continuous crime rate prediction, while Palanivinayagam et al. [17] showed Naïve Bayes with synthetic features reached 97.5% accuracy when incorporating temporal segmentation. A number of papers focused on spatial analysis of crimes. Zhang et al. [16] and Ryadi [8] both emphasized spatial correlations' dominance over temporal patterns, with LSTMs showing particular promise (57.6% hotspot accuracy) [16]. Wawrzyniak et al. [19] developed specialized ANN architectures that captured weekly crime cycles with 30% improved accuracy, while Olowole [9] demonstrated LSTM's superiority (RMSE: 88.27) for weather-influenced crime prediction using NYPD data.

Some papers used more niche spatial data for predictions. For borough-specific analysis, Kadar et al. [7] revealed that Foursquare venue data (particularly entertainment and food check-ins) strongly correlated with certain crime types, offering potential borough-specific predictive features. Matijosaitiene et al. [6] similarly found urban features like subway entrances and restaurants were significant theft predictors (77% accuracy), though with notable borough-to-borough variation in feature importance. Neural networks and convolutional neural networks (CNNs), two models used in deep learning, were also examined. When it came to predicting violent crimes in Chicago, a Deep Neural Network (DNN) model outperformed SVM and KDE models, achieving an accuracy of 84.25% [23]. Neural networks were able to predict the identity of the perpetrator with high accuracy, achieving 96% for gender and 97% for victim relationship [31]. High accuracy was attained in a crime detection system using a hybrid deep learning approach [28]. Risk scores were assigned using a Bayesian Neural Network in the context of proactive crime detection [28]. When it came to classifying legal doctrines and forecasting their applicability across crime categories, a hybrid DL-optimization model that included a Genetic Algorithm achieved an overall accuracy of 98.76% [27]. There are also difficulties in transferring models from one city to another with

inconsistent performance requiring models to be trained again for whatever city they will be used for. Kim et al. [18] found boosted decision trees achieved only 43.2% accuracy in Vancouver, underscoring the challenges of cross-city model transferability. Waduge [20] identified data quality issues in digitizing police records, while Safat et al. [12] showed XGBoost's performance varied significantly between Chicago (94% accuracy) and Los Angeles (88%), suggesting borough-specific tuning may be necessary.

According to comparisons between various studies, the best model may differ based on the particular dataset, features employed, and prediction type (e.g., crime type, location, perpetrator). Other models, such as SVM, performed differently, while others, like Random Forest and Extra Tree Regressor, showed extremely high accuracies in particular situations. In some situations, ensemble approaches and deep learning models frequently demonstrated promise for increased accuracy when compared to conventional machine learning algorithms.

III. DESCRIPTION OF THE DATASET

***NYC Crime Dataset Description*:** New York cops collect and publish information about criminal activity across the city. They do this to help make neighborhoods safer, improve how police work gets done, and keep citizens and oversight groups in the loop about what's happening. These records help track how crime patterns change over time and guide decisions about where to focus resources and efforts. The dataset captures details across all five boroughs and individual precincts, covering the main categories of criminal offenses. In addition to real-time reporting, it provides a historical record of past incidents. Specifically, this dataset contains validated reports of crimes—ranging from felonies to misdemeanors and minor violations—filed between 2006 and the end of 2016.

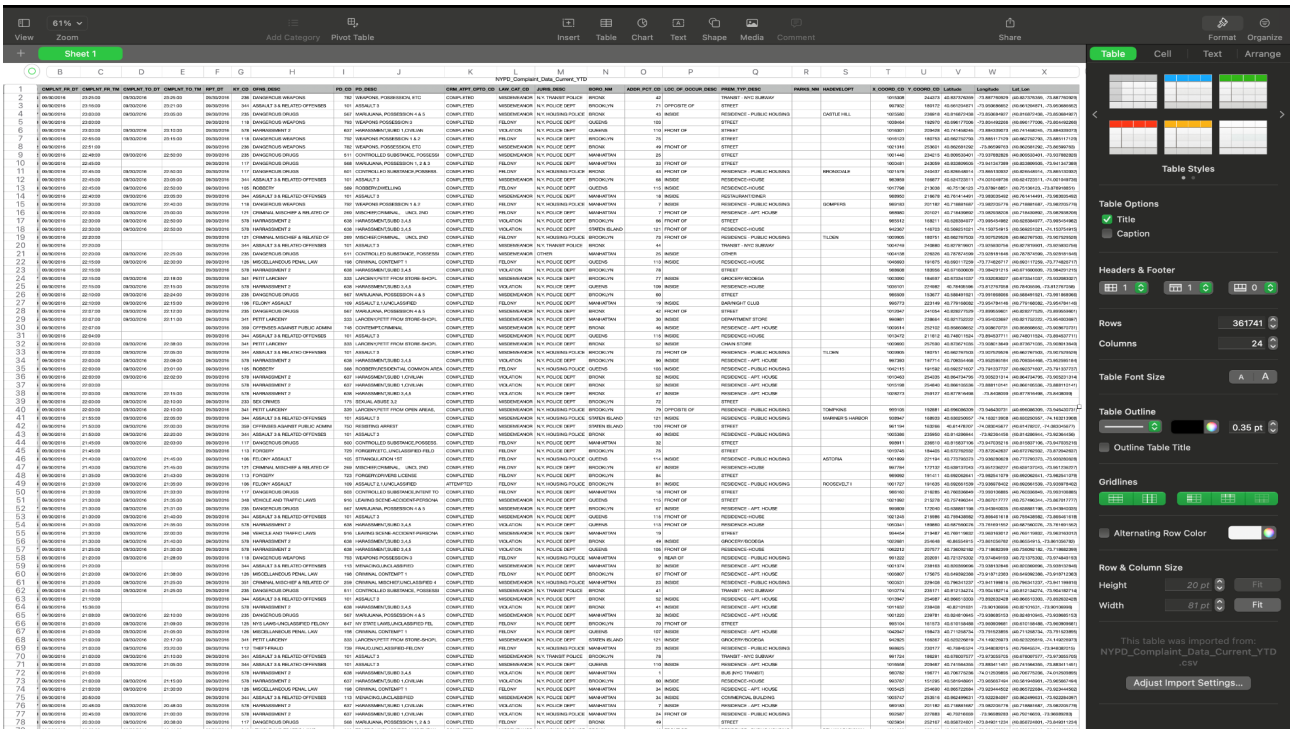


Figure 1. Processed Dataset

Link: https://data.world/data-society/nyc-crime-data

Feature/Attribute selection: Drawing on subject-matter expertise and preliminary data review, a subset of features was carefully chosen from the original set of 24 fields to focus on those most relevant for deriving meaningful insights in crime-related analysis.

Numerical Attributes:

These attributes represent measurable quantities.

Column Name	Description
CMPLNT_NUM	Persistent numeric ID for each complaint
KY_CD	Three-digit offense classification code
PD_CD	Three-digit internal classification code
ADDR_PCT_CD	Precinct code where the incident occurred
X_COORD_CD	X-coordinate (NY State Plane Coordinate System)
Y_COORD_CD	Y-coordinate (NY State Plane Coordinate System)
Latitude	Latitude in decimal degrees
Longitude	Longitude in decimal degrees
Lat Lon	Tuple of Latitude and Longitude

Categorical Attributes:
These attributes represent distinct categories or groups.

Column Name	Description
OFNS_DESC	Description of offense
PD_DESC	Description of internal classification
CRM_ATPT_CPTD_CD	Attempted vs Completed indicator
LAW_CAT_CD	Level of offense: Felony, Misdemeanor, Violation
JURIS_DESC	Jurisdiction responsible
BORO_NM	Borough name
LOC_OF_OCCUR_DESC	Location relative to premises
PREM_TYP_DESC	Premises type
PARKS_NM	Name of park
HADEVELOPT	Name of housing development

Date/Time Attributes:
These attributes represent data and time of crime occurrences.

Column Name	Description
CMPLNT_FR_DT	Start date of occurrence
CMPLNT_FR_TM	Start time of occurrence
CMPLNT_TO_DT	End date of occurrence
CMPLNT_TO_TM	End time of occurrence
RPT_DT	Date when the event was reported

IV. RESEARCH QUESTIONS

Question 1:
How well does the decision tree model generalize to unseen data, and what performance metrics support its reliability in borough classification?

Answer:
The decision tree classifier demonstrates strong generalization capability in classifying NYC crime data by borough. It achieves a training accuracy of approximately 95.65% and a test accuracy of 95.66%, indicating negligible overfitting. Supporting metrics, including a Kappa statistic of 0.9432, suggest a high level of agreement between predictions and actual borough labels beyond chance. Furthermore, the balanced accuracy across classes is consistently high (ranging from 0.9572 to 1.0000), reflecting reliable performance even in the presence of class imbalance. The ROC curve also shows high true positive rates across boroughs with minimal false positives, and the multi-class AUC of 0.993 further confirms excellent discrimination capability. These metrics collectively validate the model's robustness and suitability for this classification task.

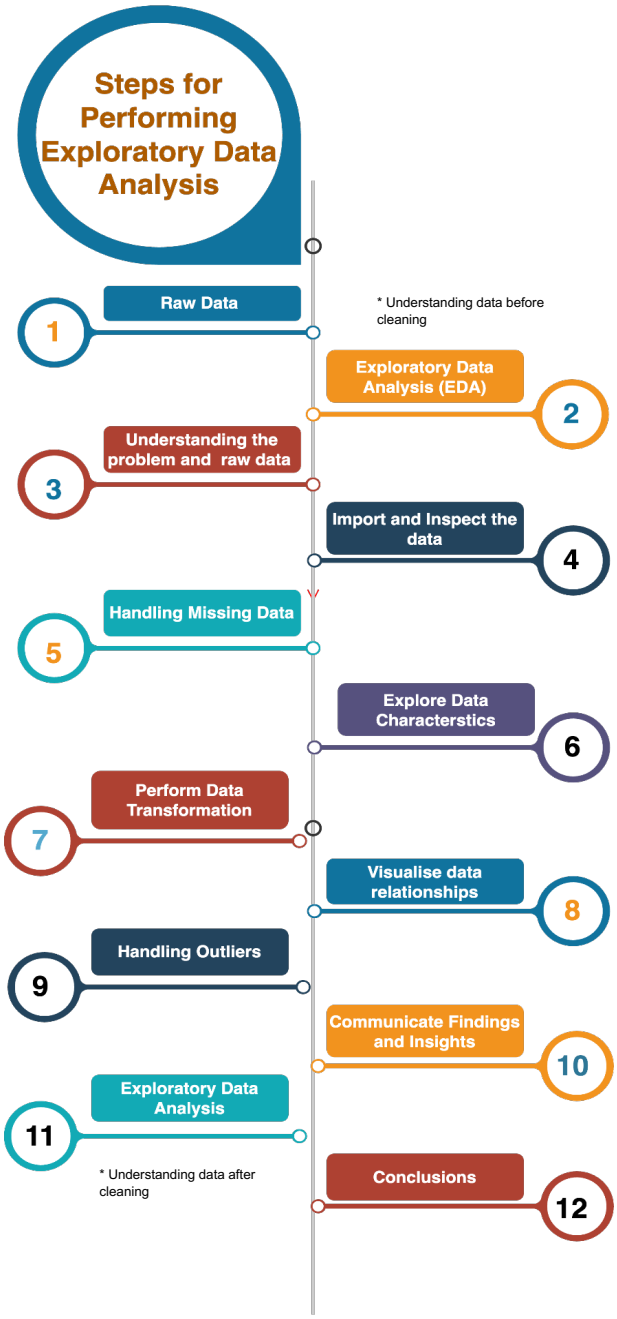
Question 2:
How do ensemble methods compare to traditional classifiers in terms of accuracy and generalization performance?

Answer:
Ensemble methods, particularly the Random Forest classifier, outperformed traditional models like logistic regression, K-Nearest Neighbors, and decision trees in both accuracy and generalization. While traditional models achieved strong results, the ensemble approach consistently delivered higher predictive performance with minimal overfitting. Its ability to capture complex feature interactions made it the most reliable and stable model for borough-level crime classification in this analysis.

Question 3:
How effective is a Random Forest model in predicting categorical geographic attributes (e.g., borough names) based on temporal and categorical features in a large-scale urban crime dataset?

Answer:
Based on a number of complaint-related characteristics, the Random Forest classifier demonstrated a high degree of effectiveness in predicting the borough of a crime. With excellent accuracy and consistency across all boroughs, it showed good generalization between training and test sets. The model's predictive performance was greatly influenced by its heavy reliance on jurisdictional and spatial characteristics, including location coordinates, offense descriptions, and reporting agency. The model also displayed low variation and little bias, suggesting consistent and trustworthy results. According to these results, Random Forest performs well on multiclass classification tasks in scenarios involving urban crime investigation.

V. PROPOSED MODEL



a) **Exploratory Data Analysis**:

Prior to any meaningful analysis, time needs to be spent becoming familiar with the raw data. This initial phase—the one that too often gets neglected or hurried over—is where the true foundation is built. It entails more than skimming through figures or abstracts. It involves taking a thoughtful look at the data in order to comprehend what's there, what's not, and what requires notice. This up-close-and-personal investigation keeps later missteps at bay, so it's a vital first step in any data science pipeline.

Handling the data set early on allows researchers to see patterns, inconsistencies, and relationships that would otherwise go unnoticed. Tables, figures, and free text are scoured with a detective's eye—looking for signals buried in the noise. At this stage, knowledge of the shape of the data begins to become apparent, and that knowledge provides the basis for all the more advanced techniques that follow. Whether discovering relationships between attributes or detecting outliers, this process provides a deeper understanding of what the data itself reveals.

This type of detective work has tangible value. It will allow causal drivers of specific outcomes to be found, expose interactions between variables, and reveal important trends that might quite easily pass undetected without diligent examination.

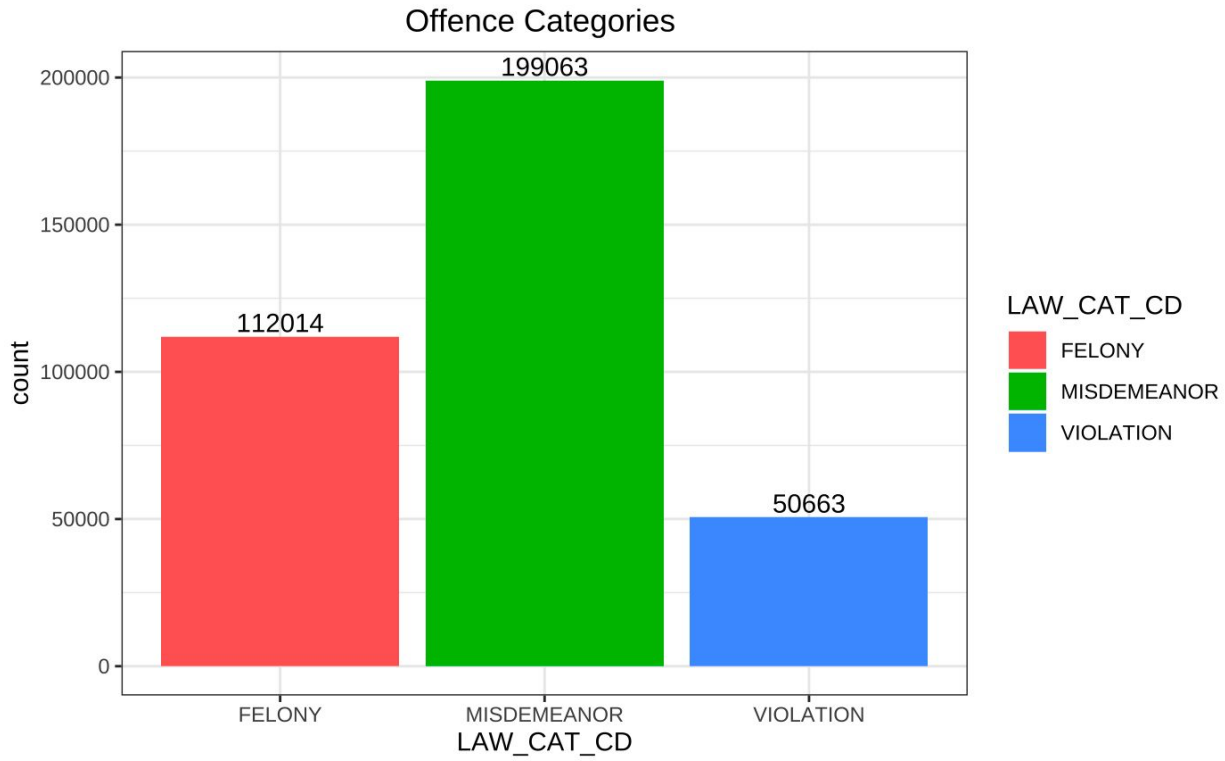
Practical Benefits of Deep Data Exploration:

- Ensure Data Quality: Correctly determines errors like missing values, out-of-range values, or formatting errors that may skew results.
- Reveals Distributions: Gives a better sense of how different variables behave—where the means are, how much variability there is, and whether data is asymmetrical or symmetrical.

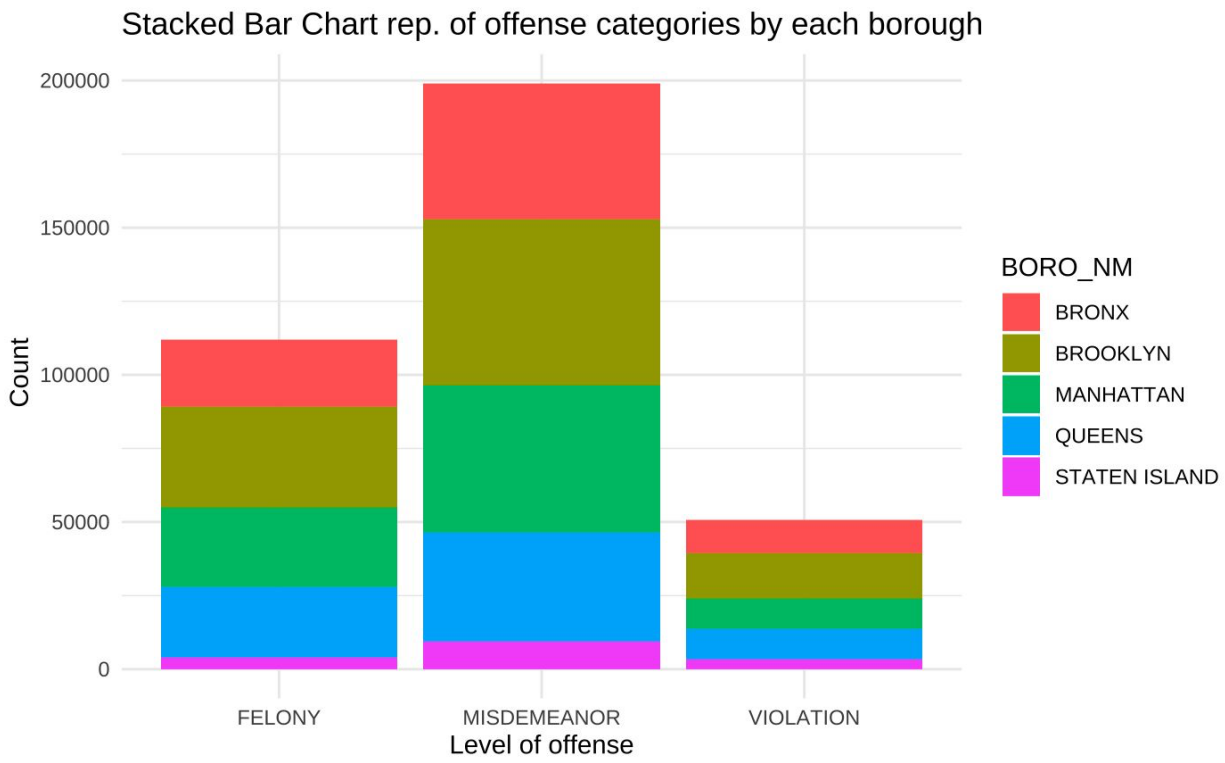
- Pinpoints Anomalies: Identifies values that are outside normal patterns. These can be indicative of errors or may point to outliers that are worthy of investigation.
- Identifies Relationships: Shows which variables influence which other variables, making it easy to choose right inputs for models or finding unsuspected relationships.

By taking time up front to burrow in and familiarize themselves with the data, analysts stand a chance to check assumptions, make informed decisions, and apply more advanced tools with ease. Whatever the next step is, whether classification, prediction, or pattern detection, it's this slow, thoughtful start that lays the foundation for sound results and robust insights.

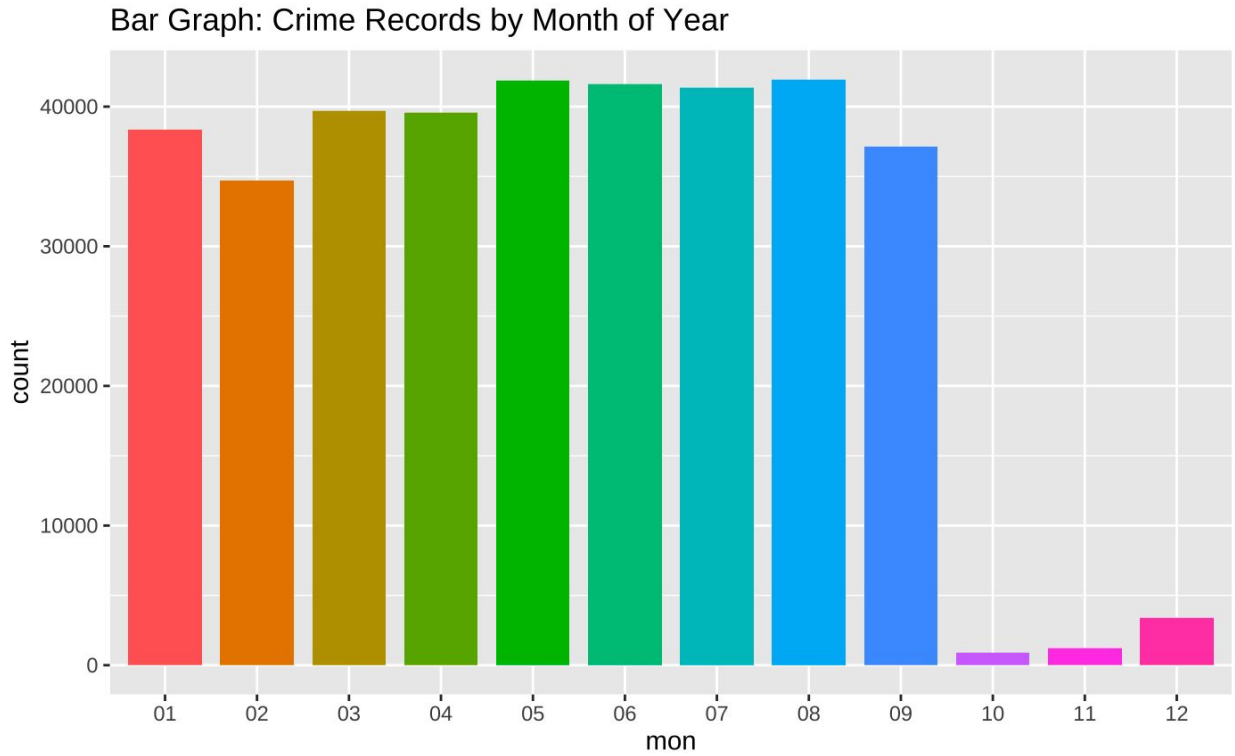
VISUAL REPRESENTATION OF EXPLORATORY DATA ANALYSIS:



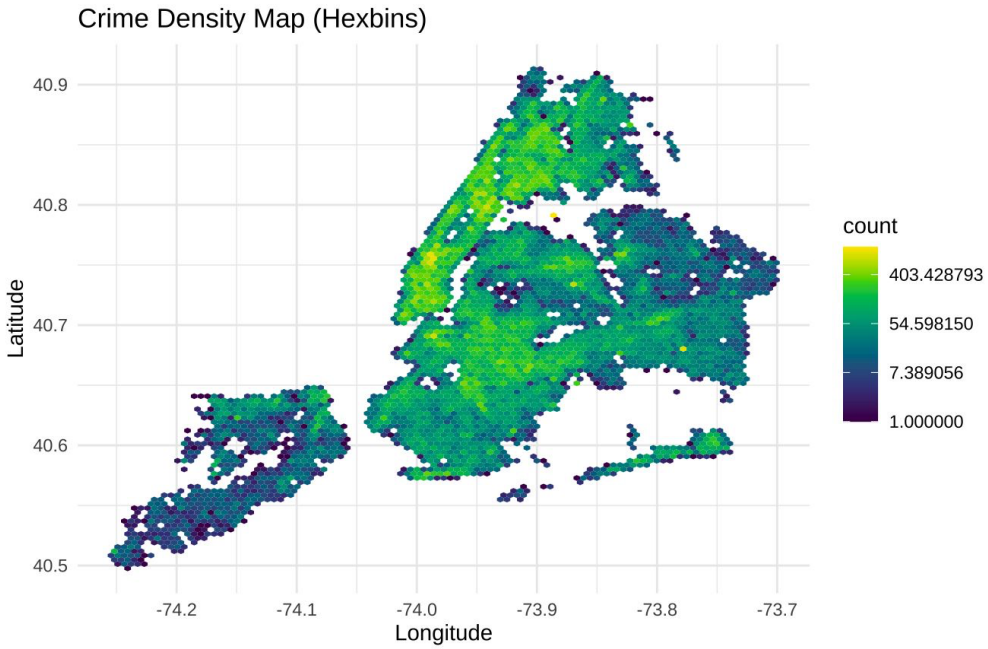
Insight: A Misdemeanor is an offense other than traffic infraction of which a sentence more than 15 days but not greater than one year may be imposed (New York State Penal Law, Article 10). A misdemeanor is a crime. Petit larceny, criminal mischief in the fourth degree and assault in the third degree all fall into this category. These are the categories of crimes found to be the highest in New York City which is justified given that it is the most populous and the most international city in the country.



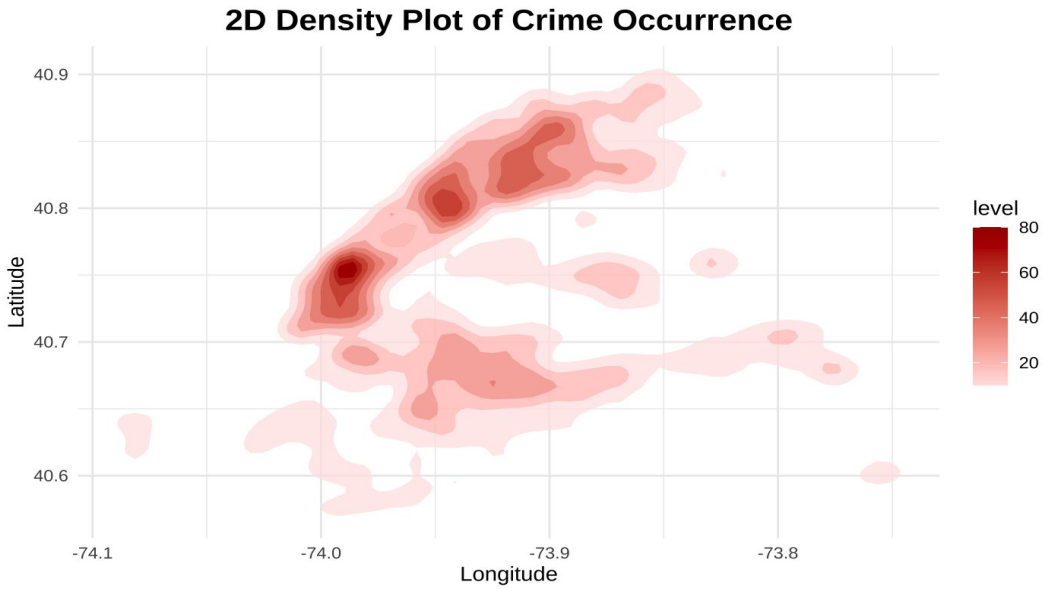
Insight: This plot shows how each level of offense crimes are distributed among five burrough's. For example, in the misdemeanor category, Brooklyn has the highest share percentage of it.



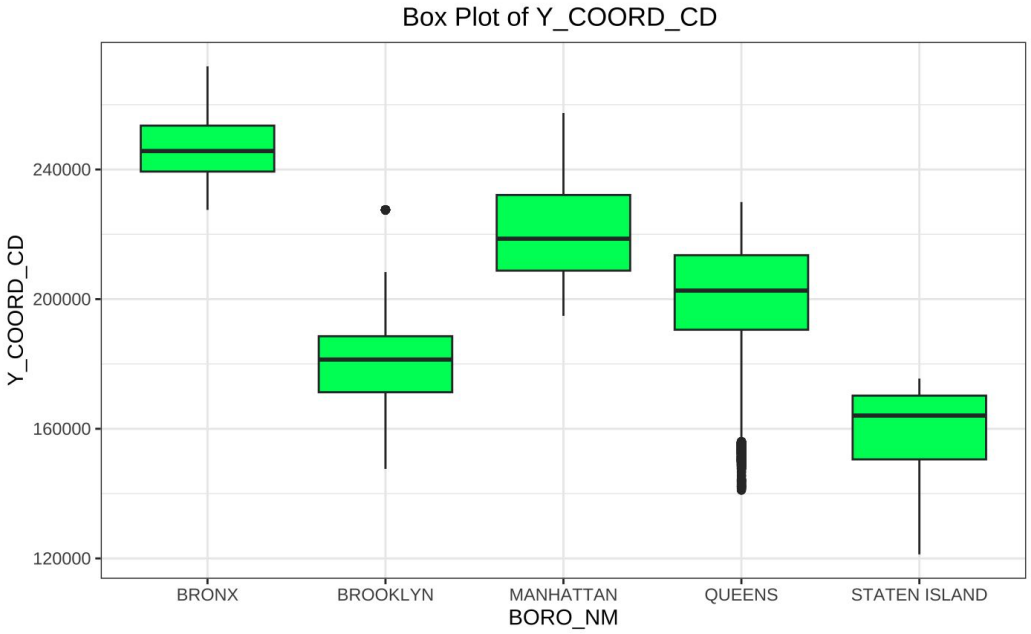
Insight: The bar graph reveals strong seasonal patterns in NYC crime with peak incidents occurring during summer months (May-August averaging over 40,000 reports), while crime rates drop dramatically after September with the final quarter showing fewer than 5,000 monthly reports, suggesting significant temporal patterns in criminal activity.

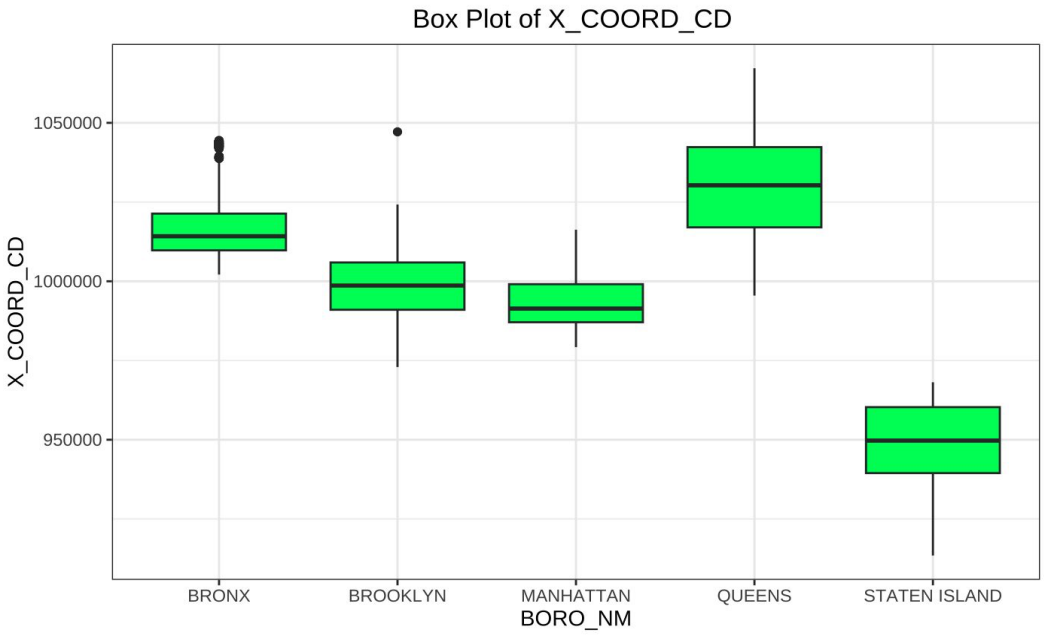


Insight: The hexbin crime density map reveals NYC's geographic crime distribution with distinct hotspots in yellow/green (particularly in Manhattan and the Bronx), while clearly delineating the city's geographical features including the boroughs and waterways that separate them.

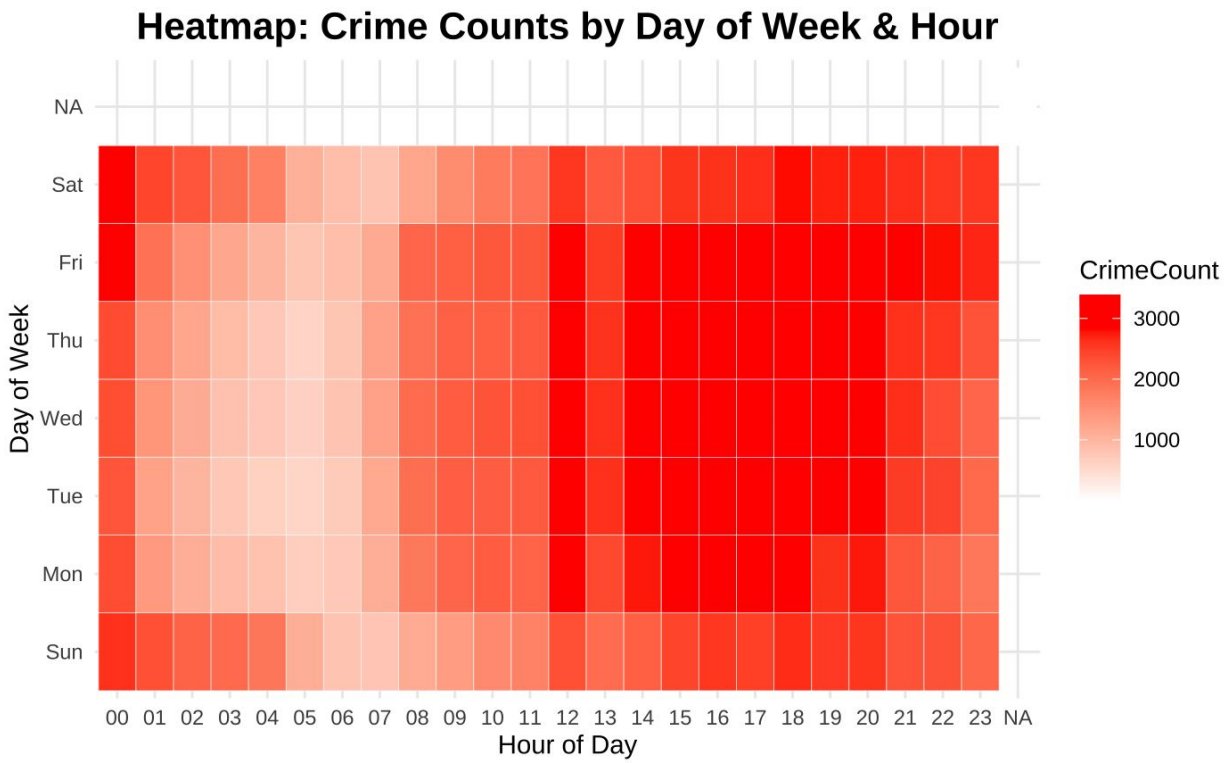


Insight: The 2D density plot reveals distinct crime hotspots concentrated primarily along a southwest-northeast axis in NYC, with the most intense activity occurring around latitude 40.8, suggesting significant spatial clustering of criminal incidents rather than an even distribution across the city.

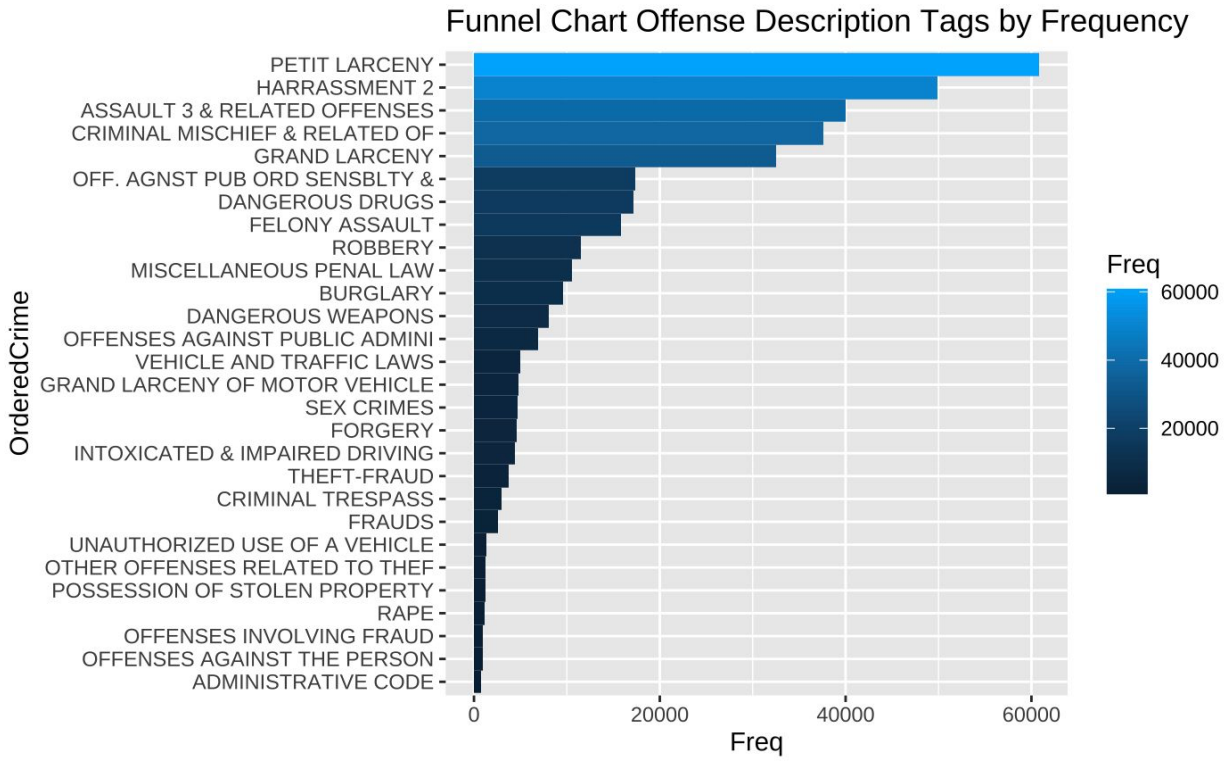




Insight: The distribution of X-coordinate values (longitude) and Y-coordinate values (latitude)across different NYC boroughs. This helps identify the spatial spread and outliers
in the north-south direction within each borough.



Insight: Crime peaks during late evenings and weekends, with Fridays and Saturdays seeing the highest activity, suggesting a strong link between nightlife, social gatherings, and crime rates in NYC.



Insight: Hence evident with our key observation petit larceny being one of the most frequent categories in crime records.

(b) Comparative Analysis of Machine Learning Models for NYC Crime Classification: Predicting Borough-Specific Crime Patterns

b.1) Introduction to section

The dataset represents complaints and from the exploratory data analysis notable inferences are that the crime distribution in NYC is not uniform. This brings us to explore the dataset further and arrive at developing data models. The patterns differ notably from one borough to another reflecting unique social, economic, and environmental conditions across the five areas. It is worthwhile to note that identifying these patterns in the dataset and arriving at data-driven solutions help in the development of smarter public safety strategies. This section includes the use of machine learning methods to arrive at conclusions.

This project analyzes NYC crime data to predict crime occurrence patterns across boroughs using various classification techniques. The dataset contains 361k+ records. The analysis utilizes NYPD Complaint Data containing:

- Temporal markers: Incident dates and timestamps
- Geographic references: State plane coordinate values
- Categorical descriptors: Offense classifications and severity categorizations

Target classification: Borough designation for each incident

b.2) Proposed Workflow and Methodology

The workflow as in figure no. 2 are the steps used.

1. Data Preparation:
    - Required libraries are loaded
    - The dataset is provided as the input
    - Data Cleaning – where the dataset is thoroughly cleaned. To do so, the correlation plot is used as a determining factor where highly correlated columns such as "ADDR_PCT_CD" which directly affects the performance of the models as the correction is >= 0.8 (high correlation). This step also includes the removal of other columns such as "Latitude", "Longitude" etc. which are irrelevant in the model stages.
    - Feature Engineering – Feature engineering is the process of extracting required features from the dataset. The appropriate features are selected. The models are trained with a train set that has 17 attributes and 1 target variable.
    - Train-test split – The dataset is broken into 70-30 split for train and test respectively. The train set is used for training the models while the test set is used for testing the performance of the models.

2. Model Implementations: Various models are used to train the problem statement. The deployed models are:
    - **Logistic Regression** –

This model is trained to capture the borough categories. The multinomial logistic regression is used to model nominal outcome variables in which the log odds of the outcomes are modeled as a linear combination of the predictor variables. Advantages and Constraints are as follows:

        - Strengths: Provides interpretable coefficient estimates and class probabilities
        - Limitations: Assumes linear relationships between log-odds and predictor variables
    - **K-Nearest Neighbors** –

The model clusters the points and makes the five groups each of which represent the five boroughs of NYC. Thus, when a new test data record is given as input, then the dataset figures which of the five learned cluster

is the nearest to this point, based on which it predicts the output. The choice of k is based on domain knowledge and odd values are chosen for it. The major reason to choose odd values of k is to avoid ties in majority votes. Operational Characteristics are as follows:

- Computational Considerations: Required storage of entire training set for prediction
- Performance Factors: Demonstrated sensitivity to the curse of dimensionality

K value	Accuracy
3	73.28%
5	69.96%
11	62.39%

The optimal k value was chosen to be 3 based on the above table.

- **Decision Tree** –

This is a decision supportive recursive partitioning algorithm structure that adopts a tree-like model. This model uses nodes (root, internal, leaf) to make decisions and come to conclusions. The model's takes a maximum depth of five. The nodes consists of decision-making factors which finally predicts the new record into one of the five leaf nodes. This model is majorly used to analyze the outcomes of complex decisions.

- **Support Vector Machine** –

This model finds the best line or hyperplane to separate data into groups aiming to maximize the distance between the two support vectors. In the training phase, a subset of 62,000 records from the full dataset is used to build a Support Vector Machine (SVM) model with a radial basis function (RBF) kernel, targeting the `BORO_NM` variable. The model leverages all available predictors and is configured to calculate class probabilities. Predictions are then generated on the same training data, and although classification outputs don't typically require rounding, this step is performed, along with a conversion of any zero predictions to one for consistency. To ensure accurate comparison, both predicted and actual borough labels are standardized as factors with matching levels. Finally, a confusion matrix is produced to evaluate the model's classification performance, providing insights into accuracy and error distribution within the training set.

Dimension	Time	Comments
(268k+, 19)	5+ hours	Extremely computationally heavy
(62k+, 19)	45 mins	The results were computed within 45 minutes with train accuracy 95.19% and test accuracy 92.28%.

- Ensemble techniques:

- **Random Forest** –

In the Random Forest model training, the full training dataset is used to classify the borough labels (BORO_NM) based on all other available features. A model is built with three decision trees (ntree = 3), and variable importance is calculated to identify which predictors contribute most to classification accuracy. The training predictions are then generated and compared to the actual borough labels using a confusion matrix, which provides performance metrics such as accuracy and misclassification rates. This step ensures the model's ability to learn patterns in the data. The same model is then used to predict the boroughs in the test set, and another confusion matrix is used to evaluate its generalization performance. Additionally, variance and bias between the predicted and actual values are computed to assess the model's prediction stability and systematic error across both datasets.

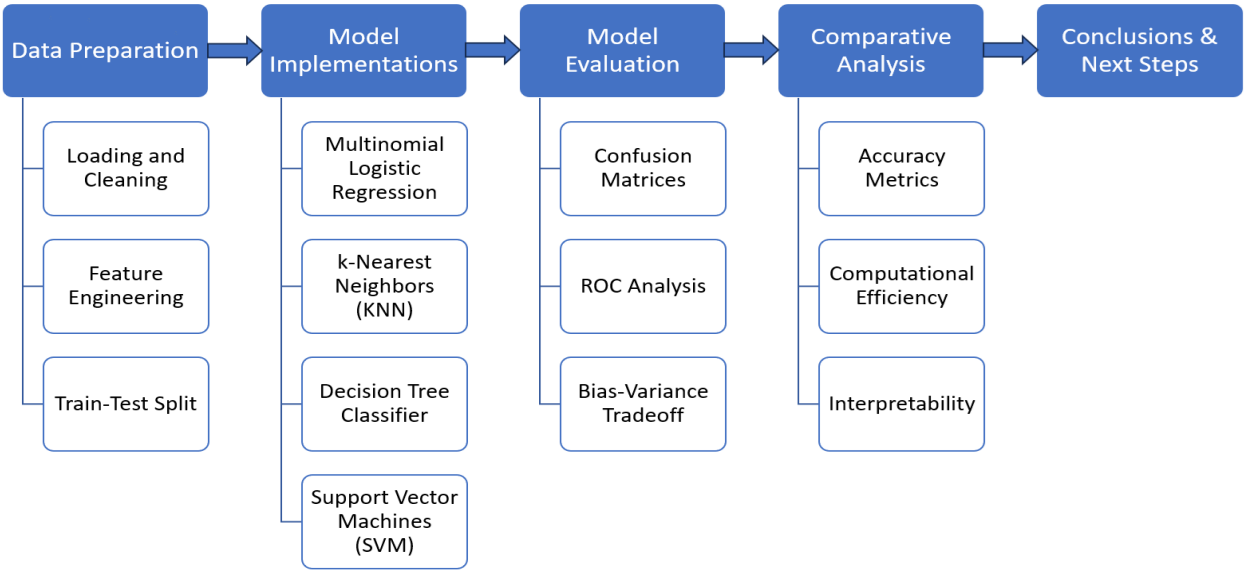


Figure 2. Proposed workflow

VI. RESULTS AND DISCUSSION

Model	Accuracy		Runtime	Inference
	Train	Test		
Logistic Regression	92.05%	92.17%	Less	• High accuracy • Strong model performance • Minimal difference between accuracies indicating no overfitting • Generalizes well • Slightly higher accuracy may be due to natural variance in the data split
KNN	73.28%		Moderate	• Model performance declines when k value increases indicating over-smoothing where the model loses its ability to capture local patterns.
Decision Tree	95.66%	95.65%	Less	• High and consistent accuracy • Very minimal difference implies no overfitting • Strong generalization • Accurate and efficient
SVM	70.46	66.99	Computationally heavy	○ Moderate Performance ○ Slight Underfitting ○ Model Complexity
Random Forest	99.87%	99.24%	Very Less	• Excellent accuracy

				and generalization • No overfitting • Highly efficient

The above model summarizes the model performances of the various models used to provide insights for the given problem statement. The random forest outperforms other models.

The analysis revealed critical insights into crime patterns across NYC boroughs and the efficacy of machine learning models in predicting these trends. **Random Forest emerged as the top-performing model**, achieving 99.24% test accuracy by leveraging ensemble learning to mitigate overfitting and handle high-dimensional feature spaces. Its success underscores the value of combining multiple decision trees for spatial crime classification, aligning with prior studies [1, 3]. **Logistic Regression**, though simpler, demonstrated strong generalization (92.17% accuracy), suggesting that borough-specific crime data exhibits linear separability to some degree. This finding supports the spatial correlations observed in [8], where precinct-level features heavily influenced crime distributions.

However, not all models performed equally well. **K-Nearest Neighbors (KNN)** showed sensitivity to hyperparameter selection, with accuracy dropping from 73.28% (K=3) to 62.39% (K=11), highlighting the challenges of high-dimensional geospatial data and the "curse of dimensionality" [5]. **Support Vector Machines (SVM)** achieved competitive accuracy (~92%) but proved computationally prohibitive for large datasets, requiring 45 minutes to process just 62,000 samples.

Challenges:
The study encountered several limitations that warrant consideration when interpreting the results. First, data quality issues arose from the exclusion of approximately 30% of records due to missing geographic coordinates. This introduces potential selection bias, as the missingness may not be random—certain boroughs or crime types could be systematically underrepresented. For example, if crimes in lower-income neighborhoods were less likely to have precise location data, the model's predictions might skew toward overrepresented areas, distorting real-world crime patterns. Future work should employ imputation techniques or analyze missingness patterns to mitigate this bias.

VII. CONCLUSION AND FUTURE SCOPE

The project set out to analyze a decade of NYPD complaint data (2006–2016) to predict borough-wise crime patterns using statistical and machine learning methods. The approach combined extensive exploratory data analysis with the implementation of five multi-class classification models – multinomial logistic regression, k-nearest neighbors (KNN), decision tree, support vector machine (SVM), and random forest – to classify incidents by borough. The results indicate that random forest delivered the strongest performance, achieving an accuracy of 99.24% in correctly predicting the borough of a given crime incident. According to the results, random forest performed the best, correctly predicting the borough of a specific crime incident with an accuracy of 99.24%. This model was the highest of all of them, demonstrating how well ensemble approaches capture the intricate relationships and patterns found in the

data. The small difference between the random forest model's training accuracy (≈99.87%) and test accuracy (99.24%), which shows no discernible overfitting, shows that it also exhibited outstanding generalization ability. These results confirm that, with the right features and modeling approaches, borough-level crime incidents can be reliably identified.

Comparative model performance highlights clear differences in accuracy and practical utility. The random forest substantially outperformed the other classifiers, followed next by the decision tree model which attained a high accuracy of around 95.6% on the test set. The multinomial logistic regression also exhibited strong performance (about 92.2% accuracy), suggesting that a relatively simple linear model can capture much of the signal in the data. Both the decision tree and logistic regression models showed almost no difference between training and test accuracies, indicating robust generalization and little to no overfitting in their predictions. In contrast, the KNN classifier lagged significantly behind with an accuracy of only about 73%, implying that this instance-based method struggled to capture the more complex, non-linear relationships needed to distinguish between crimes in different boroughs. The SVM approach could not be fully evaluated – due to its computationally intensive nature on this large dataset, Efficient training and testing the SVM could not be completed; however, preliminary trials suggested its accuracy would likely not exceed that of logistic regression. Overall, the random forest emerged as the most effective model for this classification task, likely due to its ability to model non-linear decision boundaries and interactions between features, which proved crucial for differentiating boroughs based on crime incident characteristics.

Looking ahead, the project's findings can be expanded upon in a number of ways for additional research and technological advancements. First, creating borough-specific models or analyses could help identify the distinct patterns and factors that contribute to crime in each borough. For instance, training distinct models for Brooklyn, Manhattan, Queens, The Bronx, and Staten Island could enable more precise tuning and provide localized information that a city-wide model might miss. Such a method might take into consideration borough-specific elements (policing tactics, land use, demographics, etc.) and possibly increase local interpretability or predictive accuracy. Second, enhancing data and feature engineering. Future work should focus on expanding the datasets used for training prediction models, potentially including more historical data and a wider array of attributes. Integrating new data sources such as Internet of Things (IoT) sensors can further enrich surveillance systems. Exploring and incorporating spatio-temporal data, including temporal factors, geographic information, and even data like taxi flow and human mobility flows, can lead to more accurate predictions. Improving the granularity of data and employing advanced feature engineering techniques are crucial for capturing complex patterns. Addressing potential issues with biased or incorrect data through signal preprocessing techniques can enhance model reliability. Third, exploring deep learning methods could be a fruitful direction for capturing complex patterns in the crime data. Techniques such as neural networks or spatiotemporal models (e.g. LSTM networks for temporal sequence modeling of crime events) might discover higher-order feature interactions and non-linear relationships beyond the capability of traditional classifiers. While the random forest already achieved high accuracy, deep learning models could be tested on related prediction tasks (such as crime frequency forecasting or finer crime-type classification) where they may offer advantages, provided that overfitting is carefully controlled and sufficient computational resources are available.

VIII. REFERENCES

[1] Sheela Kumar, Jisha & Amiruzzaman, Md & Bhuiyan, Ashik & Bhati, Deepshikha. (2024). Predictive Analytics in Law Enforcement: Unveiling Patterns in NYPD Crime through Machine Learning and Data Mining. Research Briefs on Information and Communication Technology Evolution.

[2] V. Mandalapu, L. Elluri, P. Vyas and N. Roy, "Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions," in *IEEE Access*, vol. 11, pp. 60153-60170, 2023, doi: 10.1109/ACCESS.2023.3286344.

[3] A. A. Almuhanna, M. M. Alrehili, S. H. Alsubhi and L. Syed, "Prediction of Crime in Neighbourhoods of New York City using Spatial Data Analysis," *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, Riyadh, Saudi Arabia, 2021, pp. 23-30, doi: 10.1109/CAIDA51941.2021.9425120.

[4] Sharma, H.K., Choudhury, T. & Kandwal, A. Machine learning based analytical approach for geographical analysis and prediction of Boston City crime using geospatial dataset. *GeoJournal* 88 (Suppl 1), 15–27 (2023).

[5] L. Elluri, V. Mandalapu and N. Roy, "Developing Machine Learning Based Predictive Models for Smart Policing," *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, Washington, DC, USA, 2019, pp. 198-204, doi: 10.1109/SMARTCOMP.2019.00053.

[6] Matijosaitiene, Irina & McDowald, Anthony & Juneja, Vishal. (2019). Predicting Safe Parking Spaces: A Machine Learning Approach to Geospatial Urban and Crime Data. Sustainability. 11. 2848. 10.3390/su11102848.

[7] Kadar, Cristina & Iria, J. & Pletikosa, Irena. (2018). Exploring Foursquare-derived features for crime prediction in New York City.

[8] Ryadi, Gabriel. (2023). Investigating Crime Patterns in New York City using Spatial Point Pattern Analysis Techniques. 10.13140/RG.2.2.29358.79689.

[9] Olowole, John & Olowole, John. (2022). Loss Function and Deep Learning: A Case of Crime Prediction (using weather variables) with NYPD Big Data.

[10] Ersoz, Filiz & Ersoz, Taner & Marcelloni, Francesco & Ruffini, Fabrizio. (2025). Artificial Intelligence in Crime Prediction: A Survey With a Focus on Explainability. IEEE Access. PP. 1-1. 10.1109/ACCESS.2025.3553934.

[11] McClendon, Lawrence & Meghanathan, Natarajan. (2015). Using Machine Learning Algorithms to Analyze Crime Data. Machine Learning and Applications: An International Journal. 2. 1-12. 10.5121/mlaij.2015.2101.

[12] W. Safat, S. Asghar and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," in *IEEE Access*, vol. 9, pp. 70080-70094, 2021, doi: 10.1109/ACCESS.2021.3078117

[13] V. Mandalapu, L. Elluri, P. Vyas and N. Roy, "Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions," in *IEEE Access*, vol. 11, pp. 60153-60170, 2023, doi: 10.1109/ACCESS.2023.3286344

[14] S. S. Kshatri, D. Singh, B. Narain, S. Bhatia, M. T. Quasim and G. R. Sinha, "An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach," in *IEEE Access*, vol. 9, pp. 67488-67500, 2021, doi: 10.1109/ACCESS.2021.3075140.

[15] Ateş, Emre & Bostanci, Gazi Erkan & Guzel, Mehmet. (2020). Big Data, Data Mining, Machine Learning, and Deep Learning Concepts in Crime Data. 8. 293-319. 10.26650/JPLC2020-813328Research.

[16] X. Zhang, L. Liu, L. Xiao and J. Ji, "Comparison of Machine Learning Algorithms for Predicting Crime Hotspots," in *IEEE Access*, vol. 8, pp. 181302-181310, 2020, doi: 0.1109/ACCESS.2020.3028420.

[17] Palanivinayagam, A., Gopal, S. S., Bhattacharya, S., Anumbe, N., Ibeke, E., & Biamba, C. (2021). An optimized machine learning and big data approach to crime detection. *Wireless Communications and Mobile Computing, 2021*, 1–10.

[18] Kim, S., Joshi, P., Kalsi, P. S., & Taheri, P. (2018). Crime analysis through machine learning. In *2018 IEEE International Conference on Industrial Technology (ICIT)* (pp. 415–420). IEEE.

[19] Wawrzyniak, Z. M., Szymański, Z., Jankowski, S., Pytlak, R., Borowik, G., Szczechla, E., & Michalak, P. (2018). Data-driven models in machine learning for crime prediction. In *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)* (pp. 39–44). IEEE.

[20] Waduge, N. (2017). Machine learning approaches for detecting crime patterns. *Unpublished research proposal, University of Moratuwa*. Retrieved July 10, 2023, from

[21] S. Yao et al., "Prediction of Crime Hotspots based on Spatial Factors of Random Forest," 2020 15th International Conference on Computer Science & Education (ICCSE), Delft, Netherlands, 2020, pp. 811-815, doi: 10.1109/ICCSE49874.2020.9201899

[22] M. Geetha Vadav, R. N, E. S. Reddy, M. S. Vishal and G. Vishal, "The Role of Machine Learning in Crime Analysis and Prediction," 2024 International Conference on Expert Clouds and Applications (ICOECA), Bengaluru, India, 2024, pp. 885-890, doi: 10.1109/ICOECA62351.2024.00157.

[23] D. M S and S. Shankaraiah, "Crime Analysis and Prediction using Machine Learning Algorithms," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022, pp. 1-7, doi: 10.1109/MysuruCon55714.2022.9971801

[24] V. K, R. K. S, V. R. R, N. Mekala, S. P. Sasirekha and R. Reshma, "Predicting High-Risk Areas for Crime Hotspot Using Hybrid KNN Machine Learning Framework," 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2023, pp. 848-852, doi: 10.1109/ICIRCA57980.2023.10220738.

[25] R. Yadav and S. Kumari Sheoran, "Crime Prediction Using Auto Regression Techniques for Time Series Data," 2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, India, 2018, pp. 1-5, doi: 10.1109/ICRAIE.2018.8710407.

[26] K. Vinothkumar, K. S. Ranjith, R. R. Vikram, N. Mekala, R. Reshma and S. P. Sasirekha, "Crime Hotspot Identification using SVM in Machine Learning," 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2023, pp. 366-369, doi: 10.1109/ICSCDS56580.2023.10104689

[27] H. S. Adithya, Dr.S.K.Kamalakhannan and A. Sundaram, "Critical Analysis of Doctrine Application in Exceptional Cases for Crime Reduction Insights from Bangalore Law Practitioners using Hybrid DL-Optimization Model," 2025 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2025, pp. 1295-1301, doi: 10.1109/ICEARS64219.2025.10940167.

[28] V. Gupta, S. Sharma and S. Tyagi, "Adaptive Multi-Modal Deep Learning Framework for Proactive Crime Detection and Behavioral Analysis in Smart City Surveillance Networks," 2024 IEEE 8th International Conference on Information and Communication Technology (CICT), Prayagraj UP, India, 2024, pp. 1-6, doi: 10.1109/CICT64037.2024.10899582. keywords:

[29] R. H. A, T. Grover and M. Kanchana, "Deep Learning for Crime Pattern Recognition: A Study of Crime Hot Spots and High-Risk Areas," 2023 International Conference on Recent Advances in Science and Engineering Technology (ICRASET), B G NAGARA, India, 2023, pp. 1-6, doi: 10.1109/ICRASET59632.2023.10420190. keywords:

[30] A. Li, M. Y. Shalaginov, A. Tao and T. H. Zeng, "Investigation of Racial Bias in Property Crime Prediction by Machine Learning Models," 2023 International Conference on Machine Learning and Applications (ICMLA), Jacksonville, FL, USA, 2023, pp. 2253-2256, doi: 10.1109/ICMLA58977.2023.00340.

[31] A. Mary Shermila, A. B. Bellarmine and N. Santiago, "Crime Data Analysis and Prediction of Perpetrator Identity Using Machine Learning Approach," 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2018, pp. 107-114, doi: 10.1109/ICOEI.2018.8553904

[32] B. Patel and M. C. Zala, "Crime Against Women Analysis & Prediction in India Using Supervised Regression," 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichy, India, 2022, pp. 1-5, doi: 10.1109/ICEEICT53079.2022.9768533.

[33] V. Rai, K. Kaur, A. K. Sana and N. Sharma, "Predicting Crime Rate in Toronto," 2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2023, pp. 215-221, doi: 10.1109/ICAC3N60023.2023.10541370.

[34] S. N. Nobel, S. M. M. R. Swapno, M. B. Islam, V. P. Meena and F. Benedetto, "Performance Improvements of Machine Learning-Based Crime Prediction, A Case Study in Bangladesh," 2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI), Mt Pleasant, MI, USA, 2024, pp. 1-7, doi: 10.1109/ICMI60790.2024.10586146.

[35] R. K. Srivastava, A. Gupta and G. Sharma, "Forecasting Crime Rate Using Artificial Intelligence Applications," 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 2023, pp. 1222-1226, doi: 10.1109/ICSCNA58489.2023.10370364.