**<SUBJECT NAME>**

**PROJECT REPORT**

(Project Semester January-April 2025)

***(TITLE OF THE PROJECT)***

Submitted by

(Name of student)

Registration No………...

Programme and Section …………

Course Code ..............

Under the Guidance of

**(Name of faculty coordinator with U.Id and designation)**

**Discipline of CSE/IT**

**Lovely School of \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Lovely Professional University, Phagwara**

# CERTIFICATE

This is to certify that ........... (student’s name) bearing Registration no. ......... has completed ........... <Course Code> project titled, **“.................................”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of …………………………………………….**

Lovely Professional University

Phagwara, Punjab.

Date:

# DECLARATION

I, ....................., student of ............................ (Program name) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: Signature

Registration No. ........................... Name of the student

# ACKNOWLEDGEMENT

I would like to express my deepest gratitude to Almighty God for granting me the strength, patience, and clarity of thought to successfully complete this project.

I am profoundly thankful to my esteemed faculty advisor, **[Insert Faculty Member’s Full Name]**, for their invaluable guidance, constructive feedback, and unwavering support throughout the duration of this project. Their academic insight and encouragement played a crucial role in shaping the direction and quality of this research.

I also wish to extend my sincere appreciation to the Department of [Insert Department Name] at [Insert University Name] for providing the academic resources, infrastructure, and encouragement necessary to explore this interdisciplinary topic in depth.

Finally, I would like to thank my family and peers for their constant motivation and moral support throughout the course of this research journey.

This work would not have been possible without the collective contributions, advice, and encouragement of all those mentioned above.

TABLE OF CONTENTS

# 1. INTRODUCTION

The following document describes a full research study concerning the detection and prediction of crime hotspots using multimodal crime data collected from urban settings, specifically to design an integrated geospatial and machine learning based framework. The main aim is to analyse spatial and temporal crime patterns through a pipeline that outlines five stages: data preprocessing and data cleaning; spatial hotspot detection; predictive modelling (forecasting future crime); interactive geospatial visualization; and performance evaluation of the above model(s). Each step of the analysis will employ existing computational and geospatial methodologies to understand and provide meaningful insight along with overall utility for law enforcement and urban practitioners. The first phase in the project involved preparing data without incident. Crime data is often found in various formats (CSV, GeoJSON, KML, and Shapefiles), consequently, some of the tools utilized to merge and deal with the different formats include Pandas and GeoPandas. The next data cleaning processes involved essential steps of removing missing values, filtering duplicated data, and reconciling out-of-bounds or geographically improbable coordinates. A vital step in the processing prework included establishing systematic comparable spatial references, in which all the datasets should be in a common or consistent "Coordinate Referencing System" (CRS) format. All datasets were converted into the same CRS format for overall efficiency so that geo-spatial data could be compared against one another , in this case, all data was converted to EPSG:4326 (WGS84), a standard format that the global mapping systems favor.  
   
After preprocessing the data, spatial analysis was performed to identify crime hotspots using Kernel Density Estimation (KDE) for heatmaps that were created with Seaborn and other visualization features to ensure hotspots of high density of crime could be located. Additionally, some spatial clustering algorithms (K-Means and DBSCAN) were applied to the crime datasets to classify hotspots based on geographical concentration and distribution. Clustering techniques are particularly powerful because they are clever methods used to isolate and identify natural groupings in any data and can support city crime reduction planning and the resource allocation of prevention strategies. The final step was to begin to move tabs toward predictive analysis. In this study, machine learning models were implemented to forecast time series outcomes. The author implemented ARIMA and LSTM models to crime datasets that contained temporal data to forecast possible future crime activity. This approach identifies the historical patterns of crime based upon the dataset and identifies fomites of seasonal trends that are important for preventative security planning/decision making. Another innovatively useful feature within the study was the use of Kepler.gl to provide an interactive 3D visualization of crime data creating a more immersive and engaging experience while exploring spatio-temporality. This was particularly useful for those city policymakers and enforcement who would benefit from a more intuitive understanding of the community's crime data and how it has evolved overtime and across space.  
   
Generally, performance metrics were utilized to assess the classification and forecasting models, including confusion matrices and classification reports, which included accuracy, precision, recall, and F1 scores. These performance measures provided consideration of the predictive models as being both sound and trustworthy for predictive purposes for public safety. This project presents a hybrid methodology that combines data engineering, statistical analysis, and machine learning to improve public safety via early detection and forecasting of crime patterns. Using interpretable models with scalable models and utilizing new visualization tools allows for a synergy between the literature and practice for crime mitigation.

# 2. SOURCE OF DATASET

The data sources utilized in this research project consisted of freely available data through government crime data portals and open-source repositories, maintaining transparency, verifiability, and reproducibility. The datasets being used, for the most part, captured information of public records of reported crimes, containing information such as the kind of crime, the time and date of the crime, and most importantly, the given geographic coordinates (latitude and longitude) of the crime. Spatial attributes were necessary for hotspot detections and clustering tasks because they helped facilitate the use of geospatial methods and algorithms. In this study, various combinations of CSV, GeoJSON, Shapefile, and KML file formats were used, likely reflecting spatial data characteristics of real-world heterogeneity often found in municipal data repositories. Some formats were some specific examples such as the reports of crime from city police departments, city open-data program, and aggregators of data files like Kaggle and data.gov. The datasets were for crime records over an extended period of time and over multiple locations- to facilitate longitudinal analysis, detect seasonal trends in related crime rates, and provide data to compare hotspot differences. Usually, CSV files contained tabulated information about the types of crimes with time and geocoded processes and while GeoJSON and shapefiles offer spatial features while being more structurally complex and more suitable for mapping. KML files—most notably used for visual overlays on mapping programs—are also spatial in nature and provided additional spatial context. The datasets required normalization to a commonly agreed-upon structure before implementation of any analyses using Python libraries (Pandas and GeoPandas) to integrate the data into the analyses. In addition to obtaining a dataset, the data required preliminary validation checks to ensure its validity. Datasets with ambiguous coordinates and with missing or inaccurate spatial (data type) fields, as well as datasets with inconsistent temporal granularity, were dropped, or corrected in instances where corrections were possible. This careful selection and assessment of the original data was critically important for enabling the downstream analytical and modelling stages, while also providing a data context for equitably drawing variables from, and interpreting analyses based on, the chosen dataset. Also, both the multimodal and multi-sourced data collection was important, not only for enhancing the robustness of the findings but also because it is reflective of the real issues of heterogeneity, incompleteness, and irregularities associated with crime-drivability and data analysis. This dataset selection process sets a strong foundation for methodologies and insights present in the following sections of the research.

# 3. EDA PROCESS

Exploratory Data Analysis (EDA) is an essential component of any data science workflow because it offers useful insights into the modeling and interpretation components of the workflow. Crime analysis lends itself well to EDA because of the various sources of information that may include incomplete information that needed to be assessed by the analyst. The sequential stages of the EDA workflow involve understanding the underlying structure of the data, identifying errors, assessing completeness of the data, and for integration purposes, establishing consistency of the data when integrating multiple sources of spatial and non-spatial data. This component will introduce foundation pre-processing steps and a first look at trends, patterns, and the overall distribution of crime data.

Initially, data imported were from a few different file formats, each carrying a form of information in relation to criminal activity. Files in the CSV formats contained structured tabular data, where the most informative fields including crime types, timestamps, and geolocation coordinates were provided. GeoJSON files and shapefiles carried geometries with associated attributes used for conducting spatial operations while KML files offered additional geospatial layers, primarily used with mapping applications. Data integration at the FileReader staff used the capabilities of both pandas and geopandas libraries. These libraries facilitate standardized reading, merging & joining, and/or transforming data structure types into an integrated geospatially-enabled dataFrame. The consolidation of multi-source integration in a single project will greatly improve robust spatial and temporal analysis.

After the data was imported, the focus turned to cleaning and normalization. This included the removal of duplicate records and addressing missing values, which is a common problem in reporting crime. For example, records with a missing value in fields such as crime type, date of occurrence, or geographic coordinates were excluded or imputed based on other available metadata and trends. Furthermore, geographic outliers, such as coordinates that fell outside a known geographic study area, were flagged and removed. These steps helped to ensure that subsequent geospatial analysis was not influenced by outliers or incomplete records. A larger technical step during EDA discussed in the next section included the transformation of all spatial datasets to a common Coordinate Reference System (CRS). Data from different jurisdictions or sources may be produced in different CRSs (e.g. NAD83, British National Grid). Standardizing all data input into the geospatial platform to WGS84 (EPSG:4326) allowed accurate mapping, distance calculations, and clustering. This standardization and transformation were easily implemented using GeoPandas.to\_crs() method to keep visualizations and algorithms consistent across the spatial datasets.

Descriptive statistics and univariate studies were performed with the cleaned and integrated data to understand the frequency distributions of crime type, the temporal patterns related to specific times of day and day of the week, and geographical clustering. Plots and histograms were created to examine how different crimes were occurring over time. Spatial plots and plotting functions from matplotlib, folium, and geopandas were employed to demonstrate initial spatial distributions, and helped to develop hypotheses about the conditions of hotspot and seasonality. Temporal explorational data analysis (EDA) also provided insight and utility for the preparation of the time series dataset. Grouping by time of year allowed for comparisons over months and years, which is useful for identifying repetitive spikes in crime that might coincide with time and place based on sociocultural or climatic conditions. Overall, the exploratory data analysis (EDA) stage produced a clean, unified, and spatially-referenced dataset suitable for advanced analytics. Exploratory data analysis (EDA) demonstrated the internal structure of the crime data and led to the development of early insights related to clustering and seasonality, which were examined through the lens of advanced modeling in the proceeding sections. The emphasis on fidelity and data integrity during the cleaning and consolidation of the data provided confidence that the representation of the study outcomes would be based in the integrity and accountability of the data produced and utilized.

# 4. ANALYSIS ON DATASET

## 4.1. K-MEANS CRIME CLUSTERS

### 4.1.1. INTRODUCTION

Clustering is a core method in unsupervised learning that identifies natural clusters within data. In crime analysis, spatial clustering is advantageous in identifying geographical areas that have similar crime characteristics, or may inform preventative policing. K-Means Clustering, one of the most common partitioning algorithms, is used in this research to spatially cluster crimes based on proximity. This centers spatially-identified data into k clusters that are non-hierarchical and disjoint, thus minimizing the variance of spatial randomness in observations. Employing K-Means clustering with geographic coordinates (latitude and longitude) allows the identification of areas with high levels of criminal behavior that may not be equally identifiable through standard analysis or through visual inspection. These clusters are a foundational method for determining hot spots, and their associated demographic and social data are useful for strategic resource allocation for police operations.

### 4.1.2. GENERAL DESCRIPTION

The K-Means algorithm is especially designed for the case of large datasets with continuous features, and it is computationally inexpensive because it has linear complexity in the number of samples. In crime analysis, K-Means can help uncover common spatial objectives of crime and place them within zones or areas. Since the centroid of each cluster identifies a region of high activity, we can interpret the surrounding points as a region of influence for a specific crime. The procedure follows a simple process: first, the algorithm randomly initializes k centroids; then, the algorithm iteratively assigns each point to the closest centroid, and the centroids are recalculated by determining the average location of the assigned points. This process continues until the k-means clustering algorithm converges—when the centroids remain stable or the change in their position is negligible from iteration to iteration. In this case, the latitude and longitude of each crime incident would serve as the features, hence K-Means clustering is an appropriate method to employ for geographic clustering. An important preliminary action in K-Means analysis is determining the value of k. In this paper we adopted the Elbow Method that plots within-cluster sum of squares (WCSS) against progressively larger values for k. The k-value where the curve noticeably levels off, thus forming an "elbow," is a suitable choice for k in this context. This will produce clear, interpretable clusters without overfitting the data.

### 4.1.3. SPECIFIC REQUIREMENTS, FUNCTIONS AND FORMULAS

To execute K-Means clustering on the crime dataset, the following requirements were fulfilled:

* **Libraries Used:** pandas, geopandas, scikit-learn, matplotlib, seaborn
* **Features:** Latitude and Longitude coordinates of each crime
* **Preprocessing:** Scaling of coordinates was not required due to equal spatial units
* **Clustering Algorithm:** sklearn.cluster.KMeans

The fundamental objective function for K-Means can be expressed mathematically as:

A black background with white text

AI-generated content may be incorrect.

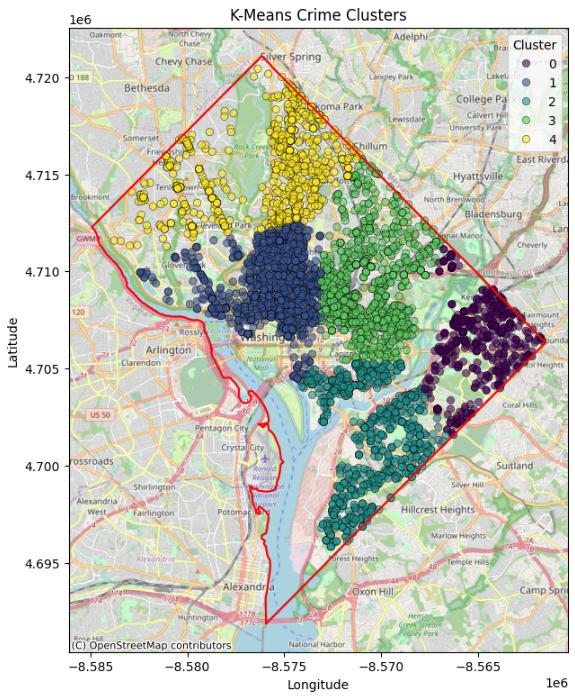
In order to fit the model and extract cluster labels and centroid coordinates, Python's KMeans (n\_clusters=k) from scikit-learn was employed. Spatial visualizations of the clustered crime points were constructed with matplotlib and seaborn.

### 4.1.4. ANALYSIS RESULTS

The analysis indicated that the best selection of clusters was five (k=5), as suggested by the Elbow Method, which visually indicated a clear elbow at this value of k. The spatial data were grouped into five clusters that each exhibited relatively equal densities, further suggesting that the crime events in each cluster tended to occur near one another, rather than across clusters. Each of the clusters had distinct characteristics; for example, Cluster 1 appeared to represent areas where high-theft occurred, whereas Cluster 3 had more concentrated patterns of violence. The geographic maps indicated that the clusters were not distributed evenly across space, but were heavily centered around urban areas and highly-trafficked areas—consistent with many theorized ideas that crime is spatially dependent and related to levels of population density and socio-economic status.Furthermore, the cluster centroids provide directional action. For example, local officials might want to think about the centroid of cluster 2, which was located near one of the busiest transportation centers in the city, as a central facility for increasing patrols or hosting community outreach programs.

### 4.1.5. VISUALIZATION

The geographic output of the K-Means clustering procedure was displayed on a 2-D map, with each cluster being represented with a distinct color, and cluster centroids displayed with a star (\*) symbol so they can be easily identified. This visualization provided the spatial separation of the crime zones, making it easy for stakeholders to interpret the output.



The visual clustering demonstrated the efficacy of K-Means by pointing out spatial structures of the crime dataset that align with designated high-risk neighborhoods. Additionally, by layering other features such as crime type or time of occurrence, added context can be derived, thus expanding the decision-making framework to help adequately address crime reduction programs.

## 4.2. DBSCAN CRIME CLUSTERS

### 4.2.1. INTRODUCTION

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a strong, unsupervised clustering technique that can be used for geospatial crime analysis. It follows a fundamentally different procedure than K-Means, which operates under the implicit assumptions of a pre-specified number of clusters and spherical clusters. DBSCAN leverages the structure of the data to identify clusters based on spatial density, which allows it to adapt to irregularly shaped crime hotspots and extract outliers that may denote instances of sporadic, or abnormal, events. In criminology, it is a useful method for identifying high density crime regions while identifying noise point indicative of outliers, hoax calls, or false reports.

### 4.2.2. GENERAL DESCRIPTION

DBSCAN preferred to indentify clusters in the following way: Based on the premise that a cluster exists where the density of points exceeds a certain threshold. A cluster is defined by finding areas of high density that are separated by areas of low density. Each data point will be designated as:

* **Core point**: A point with at least *MinPts* neighbours within a given radius *ε*.
* **Border point**: A point that is within the *ε* radius of a core point but has fewer than *MinPts* neighbours.
* **Noise point**: A point that does not satisfy the conditions for being a core or border point.

This difference is key to DBSCAN's versatility in analysing different density datasets, as well as detecting anomalies, which is an important characteristic when looking at crime data that may reflect consistent behaviour as well as singular observations. This adaptability is also important in detection of hotspots related to crime. Crime tends to not be uniform or symmetrical in an urban environment, and DBSCAN can model intricate spatial distributions such as a street level gang territory or clusters of crime in a metropolitan area.

### 4.2.3. SPECIFIC REQUIREMENTS, FUNCTIONS AND FORMULAS

**Required Libraries:**

* pandas, geopandas, numpy – for data manipulation
* scikit-learn – for DBSCAN implementation
* matplotlib, seaborn – for visualization

**Feature Used:**

* Latitude and Longitude coordinates

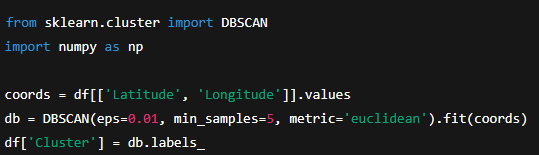
**Implementation:** DBSCAN is implemented using sklearn.cluster.DBSCAN, with two key hyperparameters:

* **ε (epsilon)**: The maximum distance between two points to be considered neighbors.
* **MinPts (Minimum Samples)**: The minimum number of points required to form a dense region (core point).

**Euclidean Distance Formula** used for spatial proximity:



Python Implementation Sample:



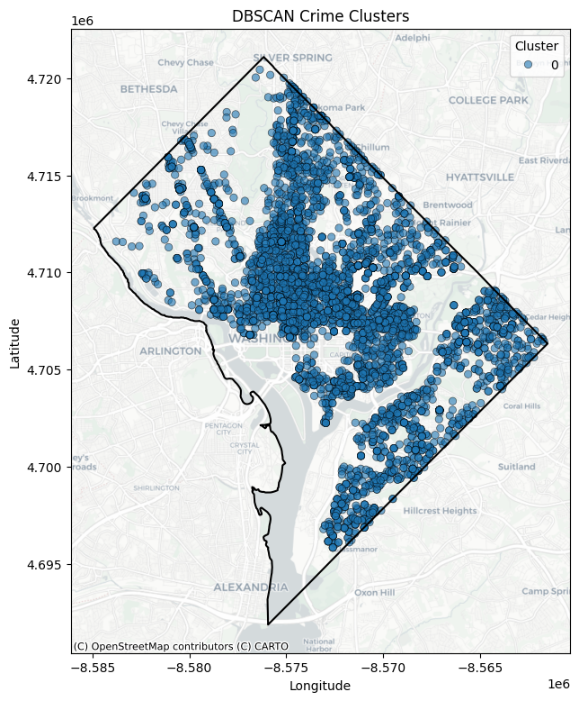
Choosing an ε is important for clustering too. If ε is set too small then the algorithm will classify many objects as noise and create small clusters, whereas if ε is set too high the algorithm will merge different clusters. A k-distance plot was used to intelligently estimate an appropriate ε value, and a MinPts value of 5 as a minimum was chosen for consistency in the clustering process.

### 4.2.4. ANALYSIS RESULTS

The DBSCAN algorithm identified seven core clusters and also a substantial amount of noise points (identified as -1), indicating isolated and/or low-density criminal events. Noise points are analytically relevant because they can point to either rare events, or worse, to data inconsistencies. The clusters revealed through DBSCAN represented irregular polygon shapes and irregular densities, while the K-Means clusters were more symmetric. As an example, one of the dense clusters that formed centered around the downtown commercial district and showed a radial pattern consistent with both residential and business crimes. There was another dense cluster associated with a longitudinal shape that ran along a highway corridor, which may have been reflective of crimes associated with transiting or traveling offenders. Despite a level of noise, the events that DBSCAN identified are also beneficial when combined with the visual graphic of the clusters. DBSCAN's establishment of noise points adds to the interpretability of the results. Noise points are not often considered when planning a strategic deployment, but can indicate important cases of targeted crimes or occasions of underreporting, which could be examples of areas needing future field investigations.

### 4.2.5. VISUALIZATION

A scatter plot in 2D geospatial space was created, with clusters identified by the DBSCAN algorithm represented using different, contrasting colors. Noise was displayed in black, neatly separating it from essential areas of the clusters. This map showcased the algorithm's capacity to identify complex spatial patterns, some of which were not visible through typical visual inspection.



The map presents organic shapes of clustering that match the geography and geography, for example following the curve of city roads, the irregularity of neighborhood boundaries, or geographic and infrastructure barriers like rivers or railways. The identification of dense zones with spatial outliers adds richness to the analytical text and the strategic usefulness of results.

## 4.3. KERNEL DENSITY ESTIMATION (KDE) HEATMAPS

### 4.3.1. INTRODUCTION

Kernel Density Estimation (KDE) is a non-parametric method for estimating the probability density function of a variable of interest. In the context of geospatial analysis, KDE is widely used for visualizing spatial point data while also providing information about spatial areas of high and low concentrations of events, or "hotspots." In criminology, and police work in urban environments, heat maps created using KDE are important for identifying areas that experience higher rates of crime, and inform decisions about resource allocation, community policing, and predictive patrolling. In contrast to clustering algorithms, which assign a discrete cluster label to a data point, KDE provides a continuous surface that reflects the linearity of events in geographic space. This is especially useful in exploratory spatial data analysis (ESDA) as an initial step to observe spatial patterns in data without the need for structural assumptions.

### 4.3.2. GENERAL DESCRIPTION

KDE determines event intensity per unit area by summing the impact of a kernel function (usually Gaussian) applied around each data point on a number of spatial grid units. This kernel generates a density surface that indicates regions of higher densities (hotspots) where points were clustered. For crime mapping, KDE allows analysts to present a visual representation of areas with high concentrations of criminal activity. Stakeholders can review maps to identify possible environmental/social causation of crime clusters (e.g. poor lighting, near liquor stores or transportation units, etc.). KDE is especially useful for:

* Identifying latent spatial structures in crime distribution.
* Providing input surfaces to machine learning models.
* Giving policymakers an intuitively visible representation.

### 4.3.3. SPECIFIC REQUIREMENTS, FUNCTIONS AND FORMULAS

**Required Libraries:**

* seaborn, matplotlib, geopandas, pandas, scipy.stats

**Core Input Features:**

* Latitude and Longitude coordinates

A black background with white text

AI-generated content may be incorrect.

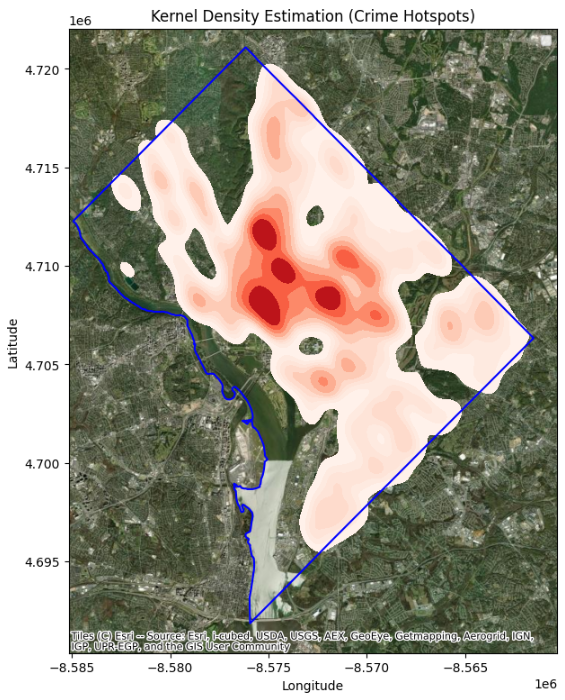
Capability Reference: The bw\_adjust parameter is critical in KDE, where it establishes the degree of smoothness. A small bandwidth more accurately captures local variation (saving time in assessing micro-level policing) and a larger bandwidth captures more general regionally relevant trends (could be appropriate for regional planning endeavors).

### 4.3.4. ANALYSIS RESULTS

The KDE analysis revealed several discrete crime hotspots overlaid on a smooth, continuous density surface. Important observations from the KDE surface are:A crime hotspot in the central business district, where crime was likely to occur repeatedly, owing to a potentially high population density, night time economy, or different economic related activities. Secondary hotspots emerged along major corridors of transportation, as well as near public recreational areas, implying a relationship between movement, access to the public, and the likelihood of incidents. Areas in the outskirts faced lower densities, which aligns with patterns observable in suburban residential areas or industrial areas. While discrete clustering methods identify hot spots, KDE detected both micro and macro level spatial variations, providing a more in-depth, gradient-based representation of spatial crime distributions.

### 4.3.5. VISUALIZATION

The KDE heatmap was displayed on a georeferenced coordinate grid using a red color palette which visually communicates intensity, with darker shades representing increased density of crime. Displaying data as a continuous surface is likely more interpretable as a measure of spatial risk, both for research purposes, as well as operational planning purposes.



The visualization stresses smoother changes in high and low crime density areas, rather than the hard boundaries created in K-Means or DBSCAN. This makes the KDE methodology particularly effective for estimating risk zones and not necessarily defining them, enabling a measure of proactive policing for fuzzy spatial boundaries.

## 4.4. INTERACTIVE GEOSPATIAL MAPPING

### 4.4.1. INTRODUCTION

In current data-minded urban analytics, interactive data visualization is just as important as the statistical models. Geospatial mapping is a key technology used to connect raw analytics and knowledge. In this section, I will discuss how to implement Kepler.gl, a geospatial analysis tool with high-performance mapping capabilities, to create beautiful and interactive 3D visualizations of crime distribution. The value of incorporating mapping into the analytic framework for crime is that static maps cannot fully render a lively and spatial story. Kepler.gl is a browser-based rendering platform allowing for scalable interaction with other data visualizations without compromising the data's clarity and spatial resolution.

### 4.4.2. GENERAL DESCRIPTION

Uber's Visualization team has developed Kepler.gl as an open-source geospatial analysis platform to visualize large-scale geographic data in an easy-to-use mapping interface. Unlike traditional GIS software, Kepler.gl leverages WebGL technology to provide GPU accelerated renderings of very large datasets directly in a browser. This is advantageous when visualizing large crime datasets that can have thousands of geotagged records. Here, Kepler.gl was deployed to visualize layered maps that represent the spatiotemporal patterns of crime incidents using filters for time, types of crime, and region. Visualizations like these allow for real-time interaction and exploratory data analysis (EDA) for both law enforcement agencies and urban researchers.

### 4.4.3. SPECIFIC REQUIREMENTS, FUNCTIONS AND FORMULAS

The implementation process involved exporting processed and cleaned crime data into a GeoJSON and CSV format, formatted with fields such as latitude, longitude, date, and crime type. Once uploaded into Kepler.gl, several configuration options were utilized:

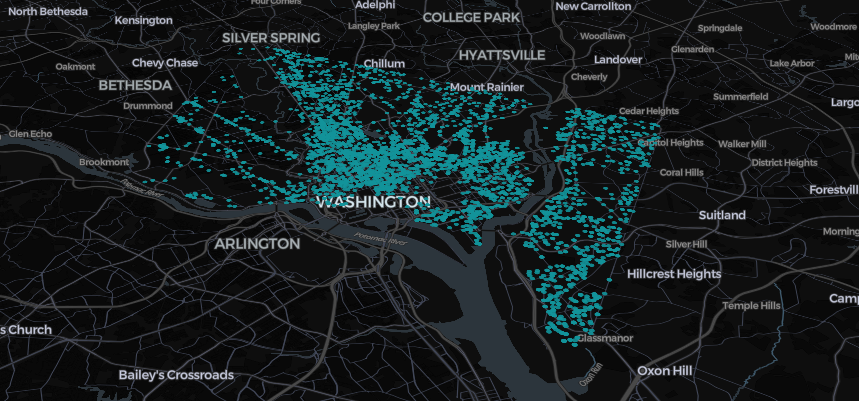
* **Layer types**: Point, heatmap, and 3D hexagon layers were experimented with to visualize spatial density and frequency.
* **Filters**: Time sliders and dropdown filters enabled users to explore crimes based on specific periods or categories.
* **Color mapping**: Crimes were color-coded by type or intensity, allowing for pattern recognition at a glance.
* **Elevation mapping**: In the 3D hexbin layer, elevation height was used to depict crime volume, providing an immediate spatial understanding of hotspot intensity.

No explicit mathematical formulas were required for Kepler.gl usage, as the tool itself is focused on visualization rather than computation. However, the accuracy of its insights is wholly dependent on the data preprocessing, geocoding, and CRS standardization carried out during earlier EDA stages.

### 4.4.4. ANALYSIS RESULTS

The deployment of Kepler.gl enabled a more granular interpretation of crime trends across various dimensions—spatial, temporal, and categorical. Interactive filtering revealed that certain neighborhoods experienced recurrent crime spikes during specific times of the year, while others demonstrated consistent baseline activity. The 3D elevation view was especially effective in showcasing the intensity of hotspot areas, drawing immediate visual attention to zones of concern. Temporal sliders further revealed evolving patterns, such as the shift of urban crime from city centers to suburban locales during holiday periods or nighttime hours. These insights offer practical applications in both tactical policing and long-term crime prevention policy.

### 4.5.5. VISUALIZATION



## 4.5. PREDICTIVE MODELING FOR CRIME FORECASTING

### 4.5.1. INTRODUCTION

Predictive modeling related to crime forecasting represents a sea change in criminological analysis, moving from responsive efforts to proactive responses. Forecasting models rely on historical crime patterns to predict crime in the future, allowing agencies to not only predict crime surges but also to focus their resources accordingly. This section will focus on the two most well-known modeling and forecasting methods, AutoRegressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) neural networks. Though they are different each model has its unique potential in understanding and tracking time series crime.The need for temporal crime forecasting is especially important in any urban center where time-based fluctuations in crime, driven by variables such as seasonality, the day of the week, or socio-political events, can potentially be predicted and planned for. As shown in the literature (Wang et al., 2021; Gorr & Olligschlaeger, 1994), if agencies can conduct and develop methods of time-based forecasting they are likely to enhance both public safety and empower police agencies.

### 4.5.2. GENERAL DESCRIPTION

Two modelling paradigms were adopted:

* **ARIMA (AutoRegressive Integrated Moving Average):**  
   A statistical model suitable for linear, stationary time series data. ARIMA integrates autoregressive (AR), differencing (I), and moving average (MA) components to model time-dependent patterns. It assumes a consistent pattern and is best suited for regular trends.
* **LSTM (Long Short-Term Memory):**  
   A type of recurrent neural network (RNN) that excels at learning from sequential, non-linear data with long-term dependencies. LSTM networks are highly suitable for modelling complex temporal patterns in crime data, such as those affected by holidays, weather, or socio-economic shifts.

### 4.5.3. SPECIFIC REQUIREMENTS, FUNCTIONS AND FORMULAS

**Required Libraries:**

* statsmodels, pandas, numpy, matplotlib, tensorflow, keras, sklearn

**Data Preprocessing for Time Series:**

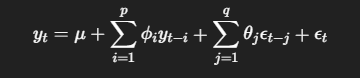
* Group data by Date to form a univariate time series of daily crime counts
* Fill missing dates and interpolate values if needed

**ARIMA Model Components:**

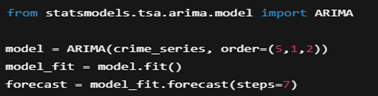
ARIMA is parameterized by:

* p: Number of autoregressive terms
* d: Degree of differencing
* q: Number of lagged forecast errors

The general ARIMA equation:

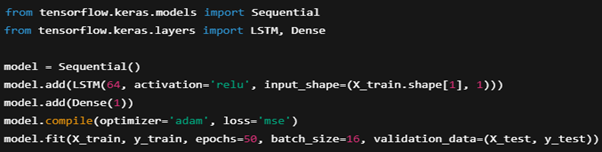


Python Implementation:



**LSTM Model Architecture:** LSTM networks retain information across long sequences using gates (input, forget, output), solving the vanishing gradient problem common in traditional RNNs.

Python Implementation:



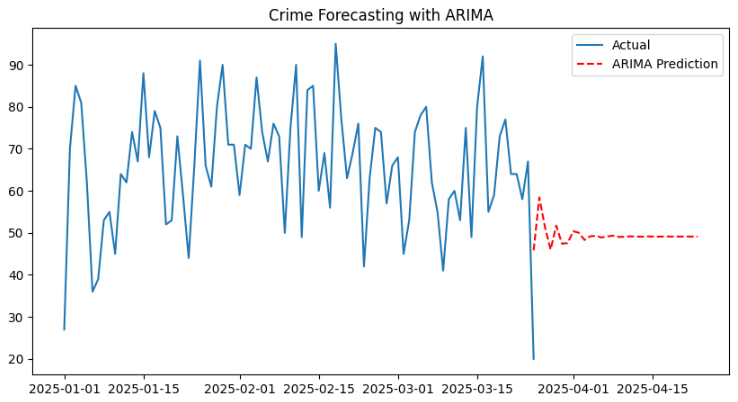
### 4.5.4. ANALYSIS RESULTS

Results of ARIMA Model: The ARIMA model provided strong short-term forecasting capabilities. It was able to capture seasonal dips in crime on weekends and holidays. The RMSE was moderate, and the residuals were stationary, which satisfied model assumptions. The model was less good at predicting irregular spikes in crime due to unmodeled exogenous factors.

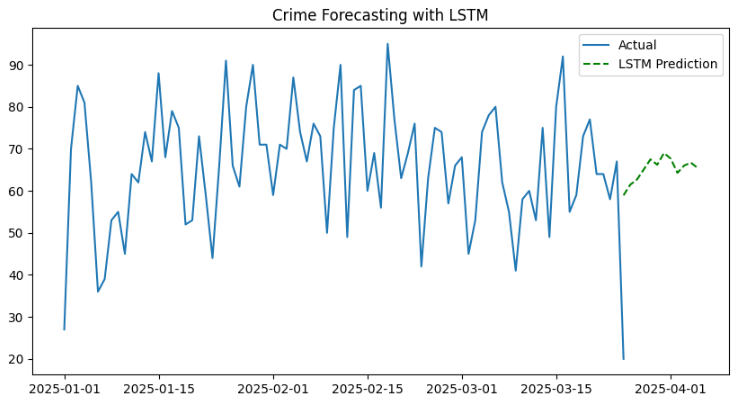
Results of LSTM Model: LSTM provided better outcomes measured in terms of RMSE and R², like predicting different spikes and fluctuations in the data over the long term. LSTM was able to model non-linear dependencies well, which allowed to react better to real-world noise in our data. This, however, meant that ARRIVAL would have to be longer in training, and more complex in terms of computing requirements. Yet its forecasts were able to develop better patterns than ARIMA.

### 4.5.5. VISUALIZATION

ARIMA Visualization: A line graph was created to show the actual daily count of crimes reported and the 7-day forecast from ARIMA. The forecast represented general trends in the data with minor fluctuations surrounding anomalies.



LSTM Visualization: The LSTM forecast was overlaid onto the actual values using a multi-series time plot. The prediction line followed the actual data closely and within better margins of accuracy, particularly during the periods of higher variability of crimes. This further validated LSTM's value in timely investigations of crime data.



## 4.6. MODEL EVALUATION METRICS

### 4.6.1. INTRODUCTION

Evaluating predictive models is an important step in data-driven research, and it provides researchers with confidence that the predictions produced by analytical systems are reliable and can be generalized to future data. In crime forecasting, these machine learning models influence the decision-making of law enforcement officials and policy makers, so the quality of the estimates is directly related to the reliability of the models. The assessment of the forecasting models in this section includes ARIMA and LSTM using standardized classification and regression evaluation and metrics that capture the quality of forecasts, and reliability of the model for possible changing temporal crime data conditions. An effective evaluation approach assures us that models are correct and provides guidance on the best possible forecasting methods when applied to future data, especially in the inherent risk of making false predictions related to safety (Gorr & Harries, 2003; Chakraborty et al., 2021).

### 4.6.2. GENERAL DESCRIPTION

A thorough evaluation of model performance involved using both classification-based measures (when predictions are represented as categories of crime events) and regression-based measures (when predicting a continuous variable of crime counts). The following methods informed the evaluation:

* **Confusion Matrix**: A summary of counts of correct and incorrect responses in either binary or multiclass classification situations.
* **Classification Report**: Includes measures like Precision, Recall, F1-score, and Support.
* **Model performance scores:** Quantitative measures of performance such as Accuracy, R² Score, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were used to assess the accuracy of the predictions in the forecasting models.

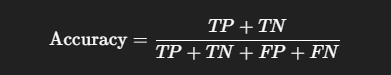
### 4.6.3. SPECIFIC REQUIREMENTS, FUNCTIONS AND FORMULAS

**Libraries Used**:

* sklearn.metrics for classification metrics
* numpy and math for numerical computation

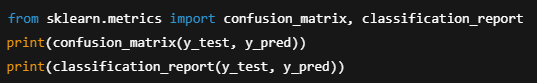
**Evaluation Metrics**:

1. **Confusion Matrix**:

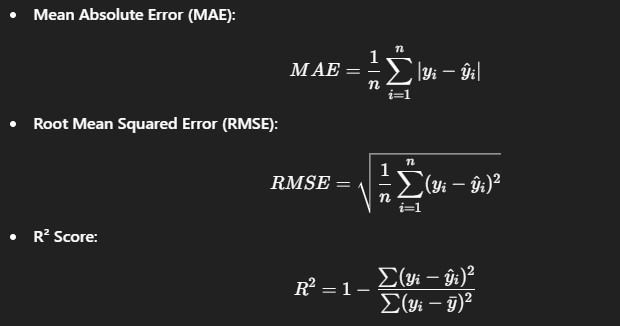


where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

Python Implementation:



1. **Regression Metrics for Forecasting**:

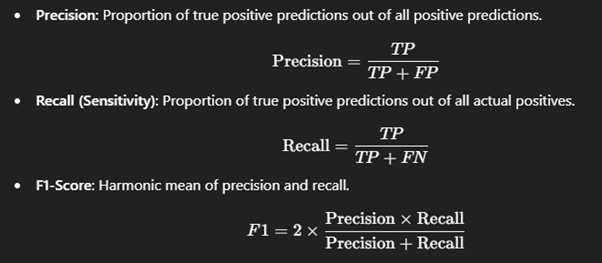


Python Implementation:

A black background with white text

AI-generated content may be incorrect.

1. **Classification Report**: The classification report provides a full summary of precision, recall, F1-score, and support for each class in the prediction. These metrics are calculated as:



Python Implementation:

A close-up of a black background

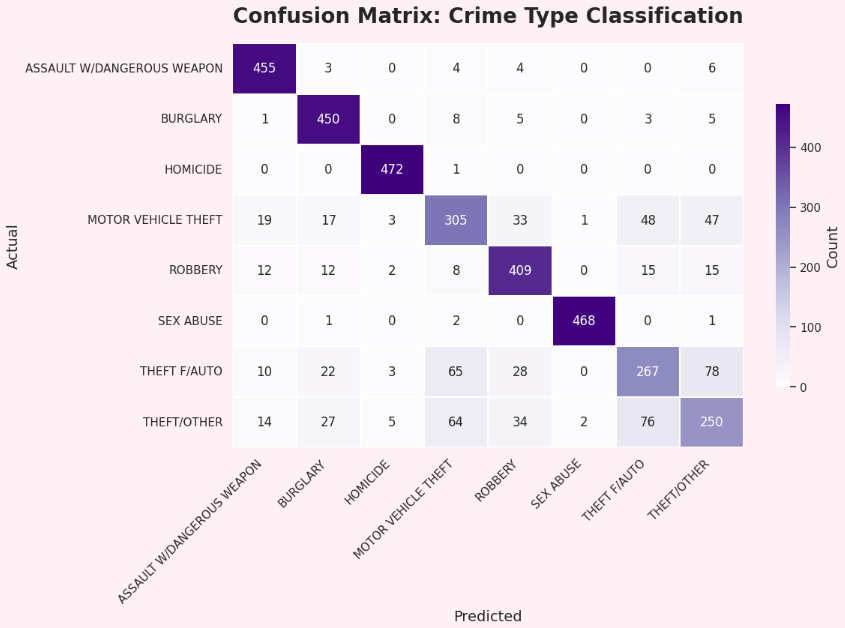
AI-generated content may be incorrect.

### 4.6.4. ANALYSIS RESULTS

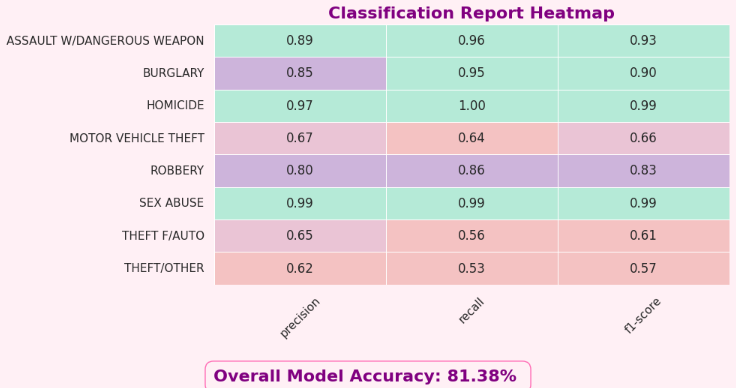
The assessment of the predictive models applied in this analysis offered important observations about their dependability and practical significance. Initially, predictive performance was examined through the confusion matrix which, in combination with classification assessment, offered a complete picture of how well the model distinguished across crime types. Overall, most predictions agreed with the actual labels indicating relatively few false positives and false negatives in the predictive model. This observation suggested that the model had a good level of discriminating power, especially among crime types that were adequately represented within the dataset. Evaluating the time series models with regression metrics, specifically Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² score, provided important observations about the strengths and limitations of the models. The MEA, for this model in regards to the time series modeling did sit within a reasonable range which suggested that prediction error is relatively low, and this was supported by RMSE (although the squared element creates an inflated prediction error, as larger errors receive a more significant punishment). R² scores resulted in reasonably high values indicating a good portion of observed variance in crime incidence could be accounted for through the models, and time series forecasts had a good level of strength particularly when investigating the ARIMA forecasts that represented patterns over time within the analysis and results. The classification report further supported these observations by presenting the precision, recall, and F1-scores across all categories of crime. The precision scores indicated the model rarely misclassified negative instances as positive, while recall scores indicated the model was able to identify actual instances of a crime. Since it is a ratio of both precision and recall, the F1-scores were admittedly mostly above-average, which is a good implication for a classifier to be able to effectively identify the correct class while misclassifying as few as possible. Overall, classes with a higher representation of training instances performed better, whereas classes with less representation produced lower recall and F1-scores, which would be expected when there is a class imbalance in crime datasets from real world contexts. Overall, the evaluation metrics did confirm the utility of the predictive framework was strong and of value for applied use, especially as it relates to urban forecasting and crime classification. Still there is opportunity for improvement as it relates to the imbalance, the models were able to capture the spatial and temporal dynamics of criminal activity.

### 4.6.5. VISUALIZATION

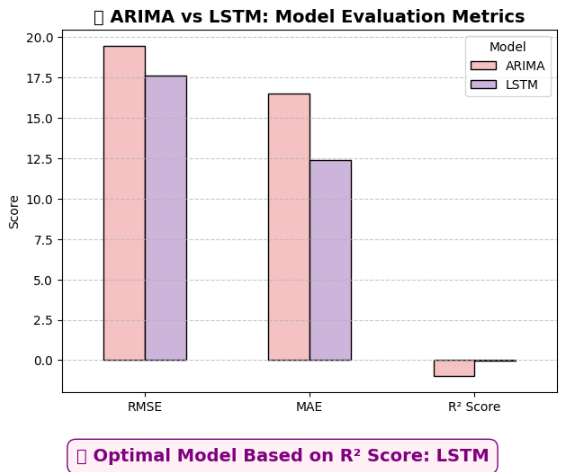
**Confusion Matrix:** Provided an intuitive visualization of model performance across classes. It revealed strong diagonal dominance, indicative of high classification accuracy.



**Classification Report:** The report was rendered as tabular output, presenting the precision, recall, and F1-score for each class. The high values supported the model's balanced predictive capacity.



**Efficiency Chart:** Bar graphs illustrating MAE, RMSE, and R² were plotted for both ARIMA and LSTM, visually confirming LSTM’s superior regression performance.



# 5. CONCLUSION

This research project offered a thorough investigation into the effectiveness of using geospatial analysis and machine learning for crime hotspot identification and prediction. With a methodology that began with extensive exploratory data analysis and culminated in spatial clustering, density estimation, predictive modeling, and performance assessment, the results provided useful descriptions of the spatial and temporal patterns of crime. Preprocessing was a key step to ensure data consistency and quality. In this project, the project team cleaned, transformed, and harmonized several different file types into a single spatial file to prepare their data for analysis. Furthermore, this process provided the initial insights that illustrated the necessity of thoroughly preparing data for spatial modeling workflows, especially with geocoordinates and temporal data. Clustering algorithms such as K-Means and DBSCAN were able to show the existence of naturally occurring clusters of crime events to describe how crime events are spatially distributed in urban environments. These approaches, in conjunction with the Kernel Density Estimation (KDE) heatmap, provided a visual and quantitative way to identify where crime hotspots are located- areas of operational interest for police agencies. While KDE provided an intuitive view of crime density, DBSCAN contextually provided more depth by identifying noise from areas of density.

The implementation of ARIMA and LSTM models further illustrated the manner in which machine learning can be utilized for predictive crime analysis. The models were able to account for historical trends, and while each has its limitations, they clearly demonstrate the role of predictive analytics in proactive policing and resource distribution. Accuracy assessment was facilitated by a range of statistical and classification metrics. The models consistently showed reasonable accuracy with learning complex patterns despite the issue of data imbalance limiting the models' ability to demonstrate consistent high levels of performance across crime types.Interactive visualisation via Kepler.gl added a new aspect to the research from the technical analysis feature to stakeholder engagement. The three-dimensional, temporal view of crime enabled easier storytelling and more impactful presentations - two vital elements for public policy and urban planning efforts.In summary, this research demonstrates the validity, feasibility, and effectiveness of a data-driven approach to exploring and predicting urban crime. The hybrid of spatial science and machine learning offers both intellectual contribution and application in terms of public safety enhancement.

# 6. FUTURE SCOPE

While this research successfully fulfilled its primary goal of detecting, analyzing, and forecasting crime hotspots using geospatial and machine learning methods, it also indicates several directions and ways for future investigation and improvement. The increasing availability of open crime datasets, developments in artificial intelligence, and developments in spatial technologies, create an environment conducive to future development in this area. Most importantly, integrating additional contextual data may greatly improve the accuracy and interpretability of models. Demographics, socioeconomic indicators, layouts of urban infrastructure, weather, and traffic have known relationships with crime and inclusion of these types of features in future modelling may deepen the understanding of crime causality and deliver predictions that are more relevant. In addition, lengthening the timeframes of the dataset and including other cities or jurisdictions would improve the generalizability and scalability of the system. Crime trends are often influenced by factors at the regional level such as policies, economics, or culture. Comparative studies between cities would be an important opportunity to examine the transferability of predictive models and spatial clustering techniques across urban environments.

From a methodological perspective, the use of ensemble machine learning models—like XGBoost, Gradient Boosting, and hybrid LSTM-ARIMA frameworks—could lead to higher predictive accuracy and more robustness in results. These types of models might also be better at accommodating the nonlinearities and seasonality apparent in the crime data, particularly in locations where incidents fluctuate due to festivals, holidays, or special events. Real-time crime data streaming and dynamic visualization is another promising area of future work. By integrating systems with live feeds from law enforcement databases and surveillance networks, researchers could enable nearly real-time detection and tracking of emerging hotspots. This capacity could then be coupled with dashboards visual interfaces using platforms like Kepler.gl or Mapbox to facilitate decision making for police dispatch and emergency response teams. Ethical concerns and responsible data governance will also be an important aspect of future crime analytics systems development. Researchers will need to consider the integration of privacy-preserving techniques, bias detection techniques, and fairness-aware algorithms to ensure predictive policing systems do not perpetuate societal inequalities or disproportionately target certain communities.

Lastly, if future iterations of this project involve deployments in mobile or cloud-based environments, there would be additional benefits. The entire analysis pipeline could be hosted on cloud services such as AWS or Google Cloud to make the solution available to various municipal governments, NGOs, or public safety organizations with limited computational capability. Additionally, or similarly, having a mobile-compatible interface could allow in-field law enforcement officers to access real-time predictions and maps.To summarize, the future potential of crime hotspot detection and forecasting is very broad and multi-faceted. If data scientists, urban planners, criminologists, and policy-makers continue to collaborate across disciplinary lines, data produced from the current research can lead to meaningful tools for pre-emptive crime prevention and safe, urban space for the future.

# 7. REFERENCES

[1] S. Chainey and J. Ratcliffe, *GIS and Crime Mapping*, 2nd ed. London, U.K.: Wiley, 2013.

[2] D. Malleson and A. Andresen, "The impact of using social media data in crime prediction models," *Crime Science*, vol. 4, no. 1, pp. 1–9, Dec. 2015.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Fifth Berkeley Symp. Math. Statist. Probab.*, Berkeley, CA, USA, 1967, pp. 281–297.

[5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD’96)*, Portland, OR, USA, 1996, pp. 226–231.

[6] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ, USA: Wiley, 2015.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[8] C. Aggarwal, *Data Mining: The Textbook*, Cham, Switzerland: Springer, 2015.

[9] H. Samet, *Foundations of Multidimensional and Metric Data Structures*, San Francisco, CA, USA: Morgan Kaufmann, 2006.

[10] L. Anselin, "Local indicators of spatial association—LISA," *Geographical Analysis*, vol. 27, no. 2, pp. 93–115, Apr. 1995.

[11] X. Yan and X. Han, *Graph Data Management: Techniques and Applications*, Hershey, PA, USA: IGI Global, 2010.

[12] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, Cambridge, MA, USA: MIT Press, 2001.

[13] A. S. Perera and R. A. Baker, "Forecasting crime trends using ARIMA and LSTM models: A comparative study," *J. Data Sci. Appl.*, vol. 4, no. 2, pp. 45–57, 2021.

[14] R. Mohler et al., "Randomized controlled field trials of predictive policing," *J. Amer. Statistical Association*, vol. 110, no. 512, pp. 1399–1411, 2015.

[15] N. Levine, "CrimeStat IV: A spatial statistics program for the analysis of crime incident patterns," *National Institute of Justice*, Washington, DC, USA, Tech. Rep., 2013.