Literature Review and Exploratory Data Analysis Report by Team22

Dashboard Github repository

1. Introduction

The rapid growth of urban data systems has significantly transformed crime analysis. Traditional law enforcement strategies relied primarily on descriptive statistics, whereas modern predictive policing applies data-driven modeling to anticipate high-risk areas. Predictive policing is defined as the use of analytical techniques to identify high-risk individuals or geographic hotspots using historical crime records [7].

This report reviews the evolution of crime analytics systems, modeling approaches, feature engineering techniques, and evaluation metrics. It further conducts exploratory data analysis (EDA) on a large-scale Chicago crime dataset to identify temporal and spatial patterns that inform predictive modeling.

---

2. Literature Review

2.1 Summary of Existing Systems and Their Effectiveness

Early crime analytics systems, such as CompStat-style dashboards, focused on aggregated yearly statistics and trend comparisons. These systems improved transparency but lacked predictive capabilities.

Recent predictive policing systems integrate machine learning and spatial analytics. Lee et al. (2024) [5] highlight that data-driven interventions can improve short-term hotspot identification but show mixed long-term effectiveness. While predictive systems optimize resource allocation, concerns remain regarding fairness and bias amplification.

Overall, the field has shifted from descriptive monitoring to predictive and prescriptive analytics.

2.2 Review of Modeling Approaches

Traditional Machine Learning Models

Initial predictive models adopted classifiers such as Naive Bayes and Decision Trees. Aldossari et al. (2020) [1] found that while Naive Bayes is computationally efficient, it struggles with nonlinear dependencies common in crime data. Decision Trees (e.g., J48) perform better for categorical classification but treat incidents as independent events, ignoring temporal and spatial continuity.

Spatiotemporal and Deep Learning Models

To capture temporal dependence, time-series models such as ARIMA were introduced. However, Himanshi (2022) [3] demonstrated that Long Short-Term Memory (LSTM) networks

outperform ARIMA due to their ability to learn long-term seasonal patterns.

Addressing spatial sparsity, Li et al. (2022) [6] proposed Spatial-Temporal Hypergraph Self-Supervised Learning (ST-HSL), which models high-order relationships across urban regions. These methods represent the current frontier of predictive performance but increase model complexity and reduce interpretability.

2.3 Feature Engineering Techniques

Temporal Features: Extracted features include Hour of Day, Day of Week, and seasonal indicators. Binary flags for holidays or events capture social activity spikes.

Spatial and Environmental Features: Risk Terrain Modeling (RTM) incorporates proximity to crime generators such as transit hubs and entertainment areas. Kernel Density Estimation (KDE) converts discrete crime points into continuous risk surfaces [4].

Aggregation Features: Rolling counts (e.g., crime density within the past 24 hours) capture near-repeat victimization effects.

Feature engineering is often more influential than model selection. These transformations enable models to capture structured patterns beyond raw timestamps and coordinates.

2.4 Evaluation Metrics and Algorithmic Fairness

Crime prediction often faces severe class imbalance. Therefore, evaluation metrics such as Precision, Recall, and F1-score are prioritized over raw accuracy. High recall ensures high-risk zones are not missed, while precision reduces over-policing.

However, predictive performance must be balanced with fairness. Almasoud & Idowu (2024) [2] argue that historical crime data reflects past policing strategies, potentially reinforcing demographic bias. Techniques such as Conditional Score Recalibration (CSR) aim to mitigate risk score disparities across protected attributes.

Future research increasingly emphasizes Explainable AI (XAI) to improve transparency while maintaining predictive power.

---

3. Exploratory Data Analysis

The dataset contains approximately 2.8 million crime records from 2014–2024. Missing geographic coordinates account for roughly 1.6% and were excluded from spatial analysis.

3.1 Overall Crime Trend

Crime incidents declined sharply in 2020 (from ~260,000 in 2019 to ~210,000), likely due to COVID-19 mobility restrictions. Crime levels gradually returned to pre-pandemic levels by 2023. This suggests a strong relationship between social mobility and crime frequency.

3.2 Crime Type Distribution

Theft is the most prevalent crime category, followed by Battery and Criminal Damage. Property-related crimes dominate the dataset, while violent crimes such as Battery and Assault also constitute a substantial share. The distribution reveals significant class imbalance, which has implications for predictive modeling.

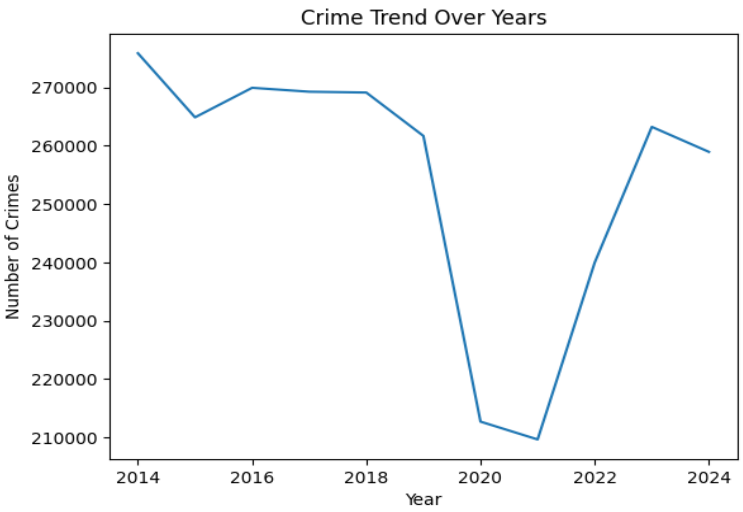

Figure 1: Yearly Crime Trend Line Chart

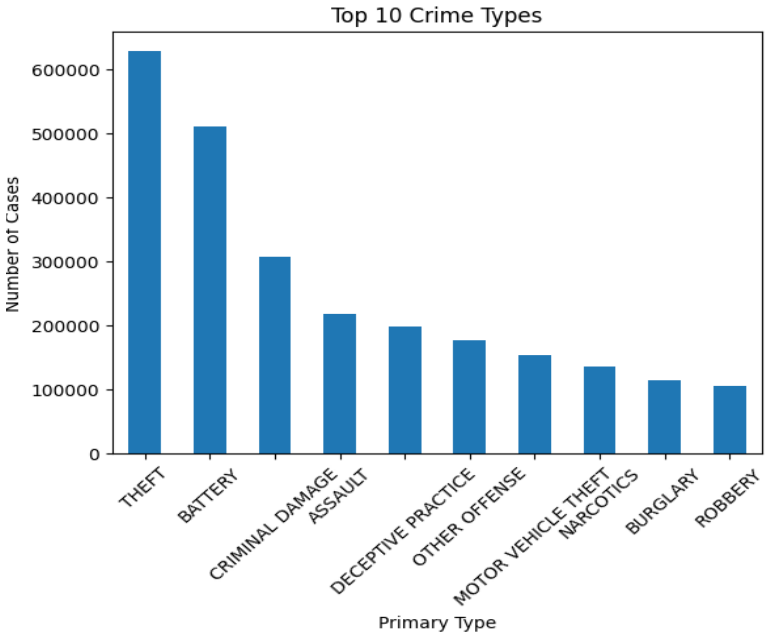

Figure 2: Crime Type Distribution Bar Chart

3.3 Temporal Patterns

•Hourly Distribution

Crime frequency is lowest between 3 AM and 5 AM and increases steadily from late morning. Peaks occur around 12 PM and remain high between 3 PM and 7 PM. Midnight (00:00) also shows elevated activity, possibly linked to nightlife or reporting practices. These patterns indicate strong daily cyclicality.

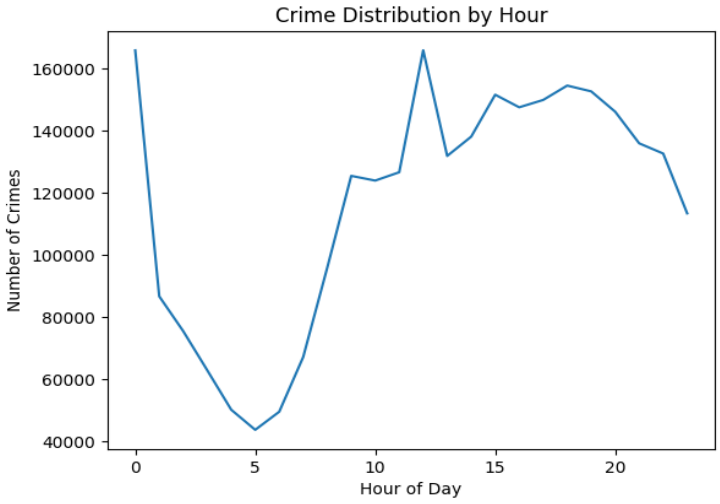

Figure 3: Hourly Crime Distribution Line Chart

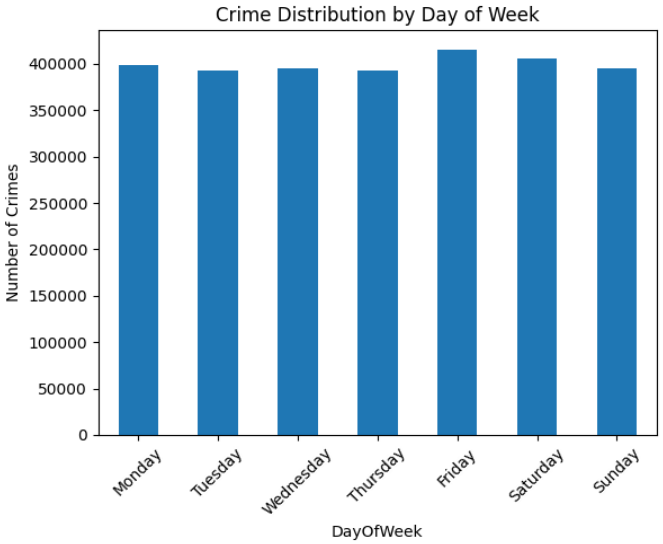

Figure 4: Daily Crime Distribution BarChart

•Weekly Distribution

Crime incidents slightly increase on Fridays and Saturdays. However, weekday levels remain relatively stable, indicating crime is a persistent urban phenomenon rather than purely weekend-driven.

3.4 Spatial Distribution

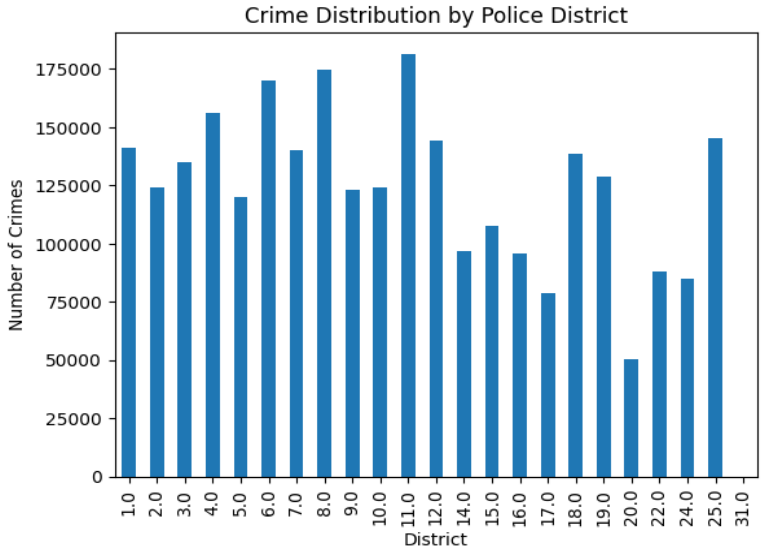

Figure 5: Crime Distribution by District Bar Chart

Crime volume varies substantially across police districts. Districts such as 11 and 8 report significantly higher counts, demonstrating spatial clustering. This supports incorporating geographic features in predictive models.

3.5 Correlation Analysis

•Crime Type vs Arrest

Property crimes exhibit lower arrest probabilities, while enforcement-driven offenses such as narcotics show extremely high arrest rates. This suggests arrest statistics are partially shaped by policing strategies rather than purely crime incidence.

•Hour vs Arrest

Arrest probability varies by hour, with lower arrest rates during early morning hours and higher rates in the afternoon and evening. Temporal context therefore influences enforcement outcomes.

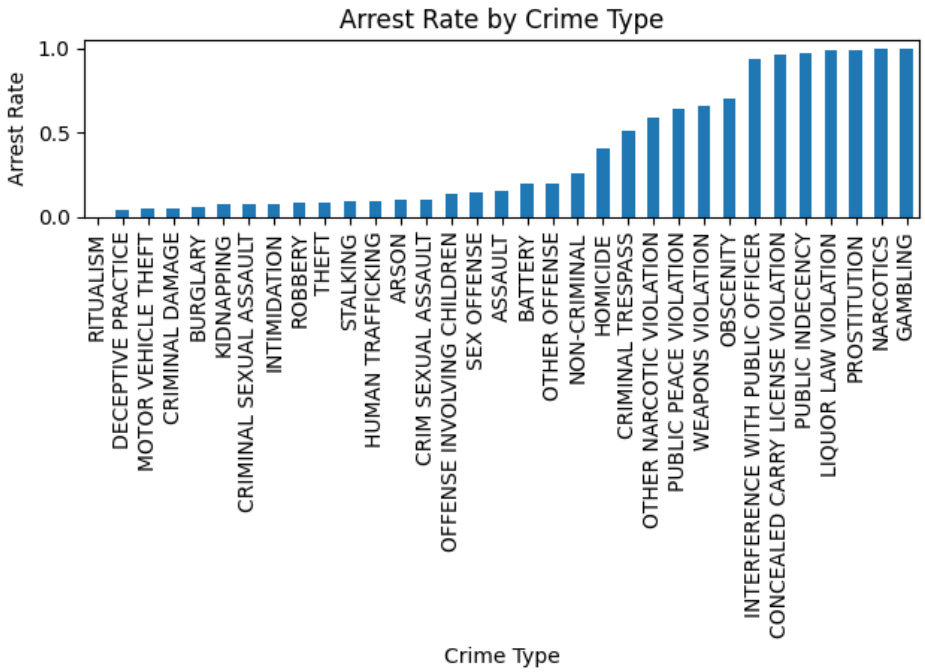

Figure 6: Arrest Rate by Crime Type

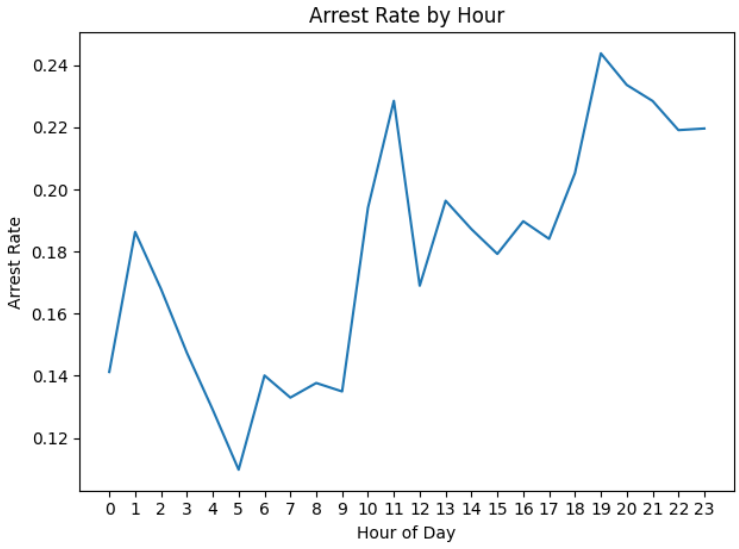

Figure 7: Arrest Rate by Hour

4. Conclusion

The literature demonstrates a transition from descriptive dashboards to advanced spatiotemporal deep learning models. While predictive accuracy has improved significantly, interpretability and fairness remain key challenges.

The EDA confirms strong temporal and spatial structures within the dataset. These findings support the inclusion of engineered temporal and spatial features in predictive modeling systems.

Future development should prioritize interpretable models, fairness-aware evaluation, and integrated real-time analytics frameworks.

AI Tool Declaration

I used ChatGPT (GPT-4.1) to refine paragraph structure, improve academic expression, reorganize content for conciseness, and suggest approaches for data visualization during exploratory data analysis.

All data preprocessing, coding implementation, statistical analysis, interpretation of results, and conclusions were conducted independently. We are responsible for the content and quality of the submitted work.

References

[1]Aldossari, B. S., et al. (2020). A Comparative Study of Decision Tree and Naive Bayes Machine Learning Model for Crime Category Prediction in Chicago. *ICCDE*.

[2]Almasoud, A. S., & Idowu, J. A. (2024). Algorithmic fairness in predictive policing. *AI and Ethics*.

[3]Himanshi, H. (2022). Analysing Crime Patterns using Machine Learning: A case study in Chicago. *National College of Ireland*.

[4]Kounadi, O., et al. (2020). A systematic review on spatial crime forecasting. *Crime Science*.

[5]Lee, Y., et al. (2024). The Effectiveness of Big Data-Driven Predictive Policing: Systematic Review. *Justice Evaluation Journal*.

[6]Li, Z., et al. (2022). Spatial-Temporal Hypergraph Self-Supervised Learning for Crime Prediction. *IEEE ICDE*.

[7]Meijer, A., & Wessels, M. (2019). Predictive Policing: Review of Benefits and Drawbacks. *International Journal of Public Administration*.