NATIONAL UNIVERSITY OF SINGAPORE

Group: 6 Feb 15, 2026

Phase 1: Foundation - Literature Review and Exploratory Data Analysis

by

Tian Yi Li Xinyao Wen Yuxin Wang Yikai Lei Jiajia

Abstract: This study examines place-based spatiotemporal patterns in Chicago crime data from 2015–2025 to inform practical predictive policing modelling. Using 2015–2024 for model development and holding out 2025 for testing, we preprocess incident records to construct core temporal features and ensure spatial consistency. Exploratory analysis shows clear temporal regularities, strong spatial concentration of incidents, and heterogeneity in arrest outcomes across offence types and neighbourhoods, motivating hotspot-oriented forecasting and evaluation aligned with operational coverage constraints.

Keywords: predictive policing; spatiotemporal hotspot forecasting; Chicago crime data; exploratory data analysis

1 Literature Review

1.1 Predictive Policing Systems and Effectiveness

Predictive policing uses historical crime data, sometimes augmented with contextual information, to forecast where and when crime is likely to occur in cities, with the goal of informing patrol allocation and tactical deployment [13, 8]. In practice, most operational systems adopt place-based spatiotemporal hotspot forecasting—exemplified by CompStat and commercial tools such as PredPol—partly for technical feasibility and partly to avoid the ethical risks of person-based prediction; however, deployments vary widely across jurisdictions in transparency, institutional integration, and legal oversight [2, 6].

Effectiveness is typically assessed via predictive accuracy against simple baselines and operational outcomes such as efficiency or crime reduction. While reported accuracy gains are common, systematic evidence for sustained crime reduction remains limited, in part because many evaluations rely on retrospective analyses rather than prospective field trials [10]. Moreover, organizational context, data quality, feedback loops, and bias can confound both measured performance and real-world impact, underscoring the need for careful model design, feature construction, and evaluation strategies in applied work [6].

1.2 Modeling Approaches for Spatiotemporal Crime Prediction

Building on the preceding discussion, the variation in reported effectiveness between studies is closely tied to modeling choices, which determine how spatial structure and temporal dynamics are represented and how predictions are operationalized. Place-based spatiotemporal forecasting dominates the literature because it aligns directly with hotspot policing and patrol planning[3, 7, 1].

Hotspot-oriented statistical baselines, such as kernel density estimation, remain common operational benchmarks because they exploit crime concentration and near-repeat patterns and produce interpretable risk surfaces; however, their sensitivity to spatial partitioning and temporal aggregation limits performance at very fine resolutions [15].

Feature-based machine learning models treat prediction as supervised learning in spatial units (e.g. grid cells) and are frequently implemented using ensemble

methods such as random forests, which integrate heterogeneous covariates and capture nonlinear effects [5].Structured spatiotemporal models, including point-process formulations, explicitly encode clustering dynamics and have been evaluated in controlled field trials[12].Deep learning approaches further extend modeling capacity but raise concerns about evaluation rigor, interpretability, and sensitivity to discretizations [14].

Overall, these approaches differ primarily in how they encode urban structure and balance flexibility against interpretability, highlighting the need for careful feature design and evaluation alignment in applied studies.

1.3 Feature Engineering for Spatiotemporal Crime Prediction

Feature engineering is central to predictive policing because it translates raw incident data into representations of temporal rhythms, spatial exposure, and contextual variation. Beyond historical crime counts, studies commonly incorporate demographic indicators, land-use proxies, mobility or activity signals, weather, and calendar effects to better explain shifting hotspot patterns [11].

Temporal dependencies are typically encoded through lagged-history variables that capture short-term persistence, weekly cycles, and seasonal trends . For instance, one district-scale framework decomposes temporal history into three components: short-term “closeness” lags, weekly periodic signals, and longer-term seasonal trends. [4]. This decomposition is practically important for temporal pattern analysis because different crime categories and neighborhoods can be dominated by different rhythms (routine activities vs. seasonal variation), and the engineered representation makes that separability available to multiple downstream model classes (regression/trees, sequence models, or hybrid deep architectures).

Calendar and weather variables are frequently used to explain deviations from baseline periodicity, sometimes through composite exposure indices rather than raw meteorological inputs. Importantly, feature design is inseparable from concerns about data quality and bias, as recorded incidents reflect reporting and enforcement practices as well as underlying victimization. Reviews therefore emphasise careful handling of missingness, stability checks over time, and analysis of how feature choices influence spatial disparities in predicted risk [11].

1.4 Evaluation Metrics and Operational Appropriateness

Evaluation protocols should reflect operational goals—namely identifying where and when limited patrol resources should be allocated—rather than focusing solely on event-level classification. Consequently, hotspot-oriented measures such as the Predictive Accuracy Index (PAI), which normalises captured incidents by covered area:

$$\mathrm{PAI} = \frac{n/N}{a/A},$$

where n is the number of crimes falling inside the predicted hotspot area a, and N is the total number of crimes in the study area A [16].These are often paired with hit-rate metrics to reflect deployment trade-offs [9].

The literature nevertheless reports diverse metric choices, including classification scores (precision, recall, AUC) and regression errors (MAE, RMSE). In sparse and imbalanced spatiotemporal data, such metrics can be misleading unless paired with forward-chaining validation schemes and sensitivity analysis over hotspot coverage levels [9].Recent work further argues for incorporating fairness and data-quality auditing into evaluation frameworks, since strong predictive scores do not necessarily imply real-world effectiveness [10]. Taken together, defensible evaluation therefore requires combining hotspot utility, probabilistic discrimination, temporal robustness, and societal constraints rather than relying on a single headline score.

1.5 Conclusion

Urban predictive policing research has largely converged on place-based spatiotemporal forecasting. Reported gains are often driven by data representation and feature design rather than model complexity, and they do not reliably translate into sustained crime reduction. Therefore, evaluation should align with deployment choices (e.g., hotspot utility under realistic coverage), be robust across spatiotemporal resolutions, and include temporal validation, transparency, and fairness checks to mitigate biased data and inequitable impacts. Key gaps remain in shared benchmarks, cross-city generalisation, and evaluation frameworks that link forecasts to measurable and equitable public-safety outcomes.

2 Exploratory Data Analysis

Motivated by the literature's emphasis on place-based spatiotemporal prediction—where performance depends on temporal rhythms, spatial concentration, and heterogeneity across offence types and neighbourhoods—this section explores Chicago crime data (2015–2025) to test whether these empirical patterns hold in the study area and to translate them into practical modelling considerations; the analysis examines temporal regularities, spatial concentration, and variation in arrest outcomes across offence types and locations, with all figures reported in the Appendix and the accompanying interactive dashboard.

2.1 Dataset Overview and Pre-processing

We restrict the original incident dataset (2001 onward) to 2015–2025 for relevance and tractability, using 2015–2024 for model development and holding out 2025 for testing; timestamps are parsed to derive basic temporal features (year, month, weekday, hour), records with invalid coordinates or missing district identifiers are removed to ensure spatial consistency, low-coverage fields (e.g., ward) and redundant/leakage-prone identifiers are dropped, and remaining missingness is handled via explicit indicators (e.g., community area) or an `UNKNOWN` category (location description), yielding a cleaned dataset of several million observations for the downstream analyses.

2.2 Temporal Patterns

Aggregation by year shows a gradual decline in total reported incidents across the late 2010s followed by renewed growth after 2021, indicating that crime dynamics are non-stationary over the study period(Fig. 1). Monthly seasonality is pronounced, with higher volumes during summer months and lower activity in winter(Fig. 2), while weekday–weekend contrasts and hourly profiles reveal stable diurnal cycles that peak in late afternoon and evening periods(Fig. 3).

Heatmaps stratified by year–month and year–hour demonstrate that although overall volumes fluctuate, core temporal rhythms remain persistent across years, suggesting structural rather than transient effects(Figs. 4 and 5). Weekly aggregation highlights systematic weekday–weekend contrasts (Fig. 6), while joint weekday–hour heatmaps reveal recurring high-risk windows concentrated on weekend evenings (Fig. 7).

2.3 Spatial Distribution

Spatial aggregation by police district and community area reveals pronounced geographic heterogeneity, with a limited subset of districts accounting for a disproportionate share of incidents (Figs. 8 and 9). Examination of location descriptions indicates that street environments, residential premises and commercial facilities dominate reported incidents (Fig. 10).

Geocoded scatter visualisations based on a large random subsample illustrate dense spatial clusters in central and historically high-crime neighbourhoods alongside lower-intensity peripheral regions (Fig. 11). This persistent concentration is consistent with the hotspot-oriented paradigms reviewed earlier and justifies framing the prediction task at the level of spatial units rather than individual offenders.

2.4 Crime–Arrest Relationships

Arrests generally move with overall crime volume, but the gap between incidents and arrests shifts across years. This likely reflects changes in enforcement outcomes, reporting behaviour, or case processing (Figs. 12 and 13). Arrest likelihood also varies by offence category. Some primary types show consistently higher arrest rates than others, so treating all offences as one uniform target would hide important differences (Fig. 14).

Arrest patterns are not uniform over time or space. The arrest proportion changes across the day, suggesting time-of-day effects on enforcement outcomes (Fig. 16). District comparisons show further geographic variation (Fig. 17). Crime mix also differs by district: some districts skew toward property offences, while others show a higher share of violent offences (Fig. 18). These patterns point to strong heterogeneity by offence type, time window, and local context.

2.5 Implications for Subsequent Modelling

The EDA supports several modelling choices. Strong temporal cycles motivate calendar features and lagged histories, while persistent spatial concentration favours hotspot-based prediction over districts or grid cells. Variation in arrest outcomes and offence frequencies requires imbalance-aware evaluation beyond accuracy, and non-stationary trends call for time-ordered validation such as rolling or forward-chaining schemes.

References

[1] U. M. Butt, S. Letchmunan, F. H. Hassan, et al. Spatio-temporal crime hotspot detection and prediction: A systematic literature review. *IEEE Access*, 8:166553–166570, 2020.

[2] W. Hardyns and A. Rummens. Predictive policing: A theoretical framework. *Policing: An International Journal*, 41(4):481–495, 2018.

[3] R. He, Y. Lu, C. Jiang, Y. Deng, X. Li, and D. Shi. Progress in research and practice of spatial-temporal crime prediction over the past decade. 2023. PDF provided in course materials.

[4] M. Hou, X. Hu, J. Cai, X. Han, and S. Yuan. An integrated graph model for spatial–temporal urban crime prediction based on attention mechanism. *ISPRS International Journal of Geo-Information*, 11(5):294, 2022.

[5] R. Jain, A. Chilambuchelvan, M. S. Kumar, et al. Predictive policing in urban environments using random forest framework for safer smart cities. In *15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024.

[6] S. Kim. Toward a more contextual approach to predictive policing: A systematic literature review. *Sociology Compass*, 2025. Advance online publication.

[7] O. Kounadi, A. Ristea, A. Araujo, and M. Leitner. A systematic review on spatial crime forecasting. *Crime Science*, 9(1):7, 2020.

[8] O. Kounadi, A. Ristea, M. Leitner, M. Langford, and P. J. Brantingham. A systematic review of crime prediction methods. *ISPRS International Journal of Geo-Information*, 9(6):348, 2020.

[9] O. Kounadi, A. M. Ristea, and M. Leitner. Progress in research and practice of spatial-temporal crime prediction over the past decade. *Crime Science*, 9(1):1–19, 2020.

[10] Y. Lee, B. Bradford, and K. Posch. The effectiveness of big data-driven predictive policing: A systematic review. *Justice Evaluation Journal*, 7(2):127–160, 2024.

[11] Y. Lee et al. The effectiveness of big data-driven predictive policing: A systematic review. 2024. Systematic review (full bibliographic details as in the provided PDF).

[12] G. O. Mohler, M. B. Short, S. Malinowski, M. Johnson, G. E. Tita, A. L. Bertozzi, and P. J. Brantingham. Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, 110(512):1399–1411, 2015.

[13] J. H. Ratcliffe. Predictive policing. *Crime Science*, 4(1):1–6, 2015.

[14] A. Rummens and W. Hardyns. Comparison of near-repeat, machine learning, and risk terrain modeling for making spatiotemporal predictions of crime. *Applied Spatial Analysis and Policy*, 13(4):1035–1053, 2020.

[15] A. Rummens and W. Hardyns. The effect of spatiotemporal resolution on predictive policing model performance. *International Journal of Forecasting*, 37(1):125–133, 2021.

[16] A. Rummens, W. Hardyns, L. Pauwels, and T. Vander Beken. The Penalized Predictive Accuracy Index: A new metric for evaluating geographic crime predictions. *ISPRS International Journal of Geo-Information*, 10(9):597, 2021.

A An appendix

Exploratory Data Analysis Figures

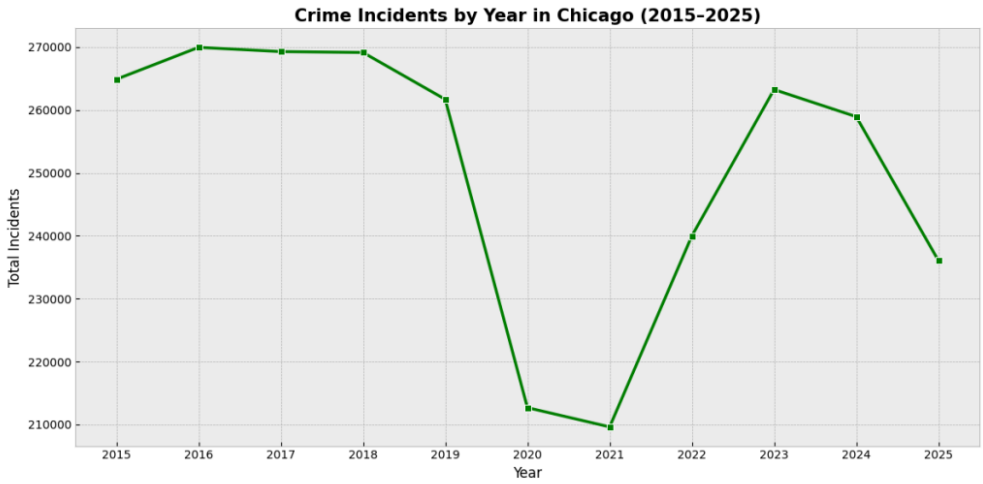

Figure 1: Long-term trend of total reported crime incidents in Chicago (2015–2025).

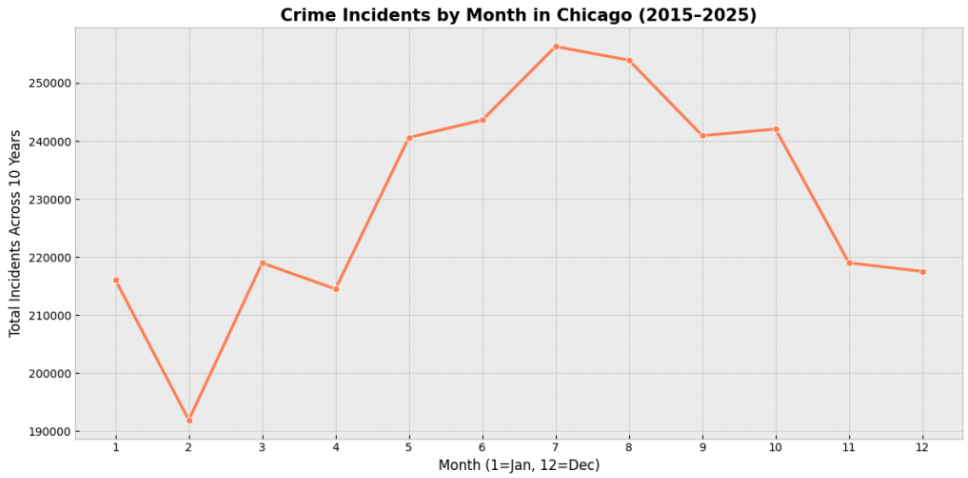

Figure 2: Diurnal pattern of crime incidents aggregated by hour of day (2015–2025).

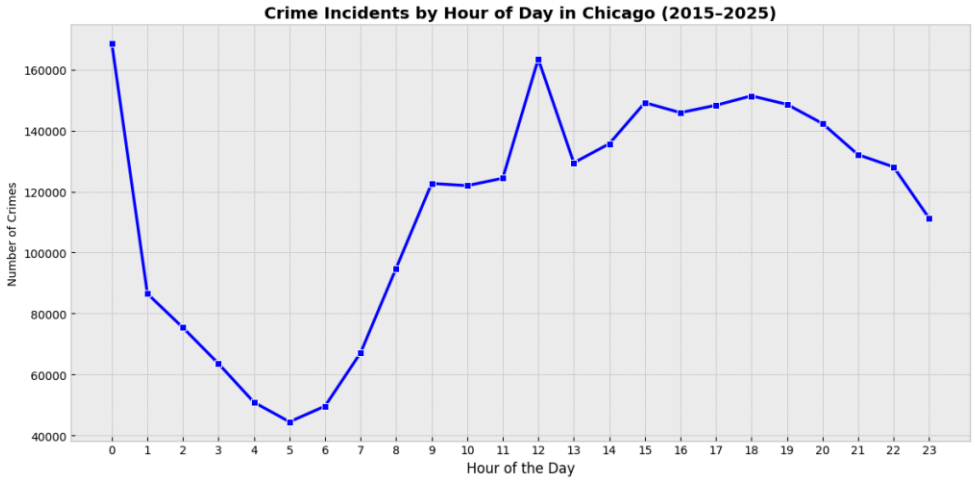

Figure 3: Diurnal pattern of crime incidents aggregated by hour of day (2015–2025).

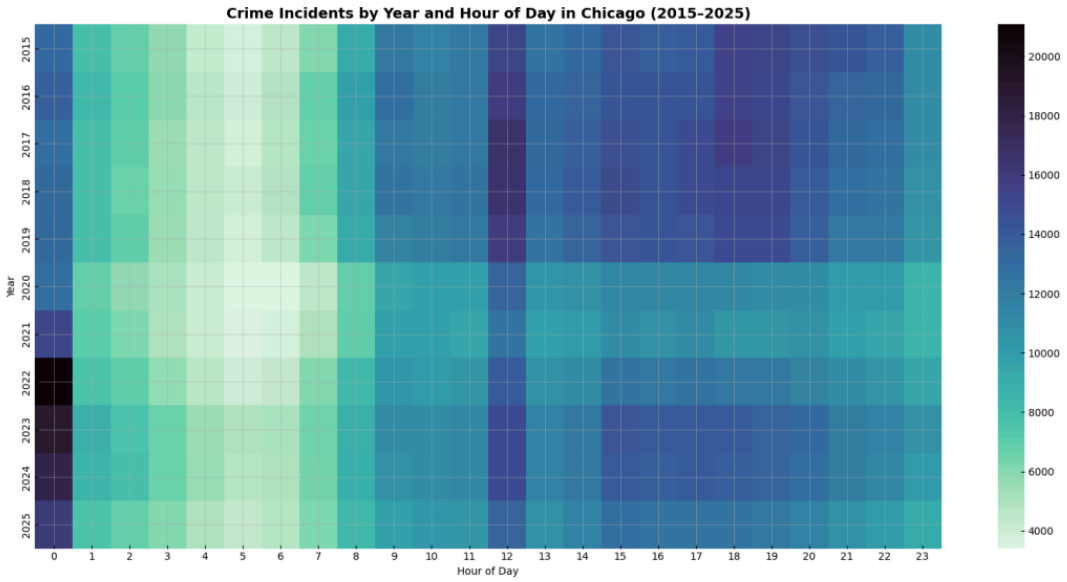

Figure 4: Heatmap of crime incidents by year and hour of day (2015–2025).

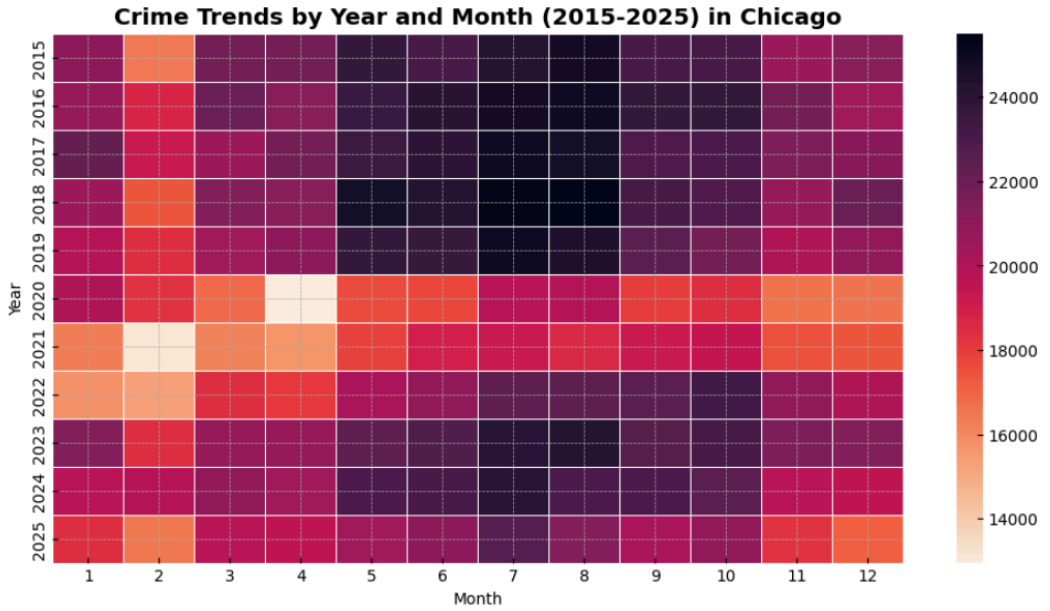

Figure 5: Heatmap of crime incidents by year and month (2015–2025).

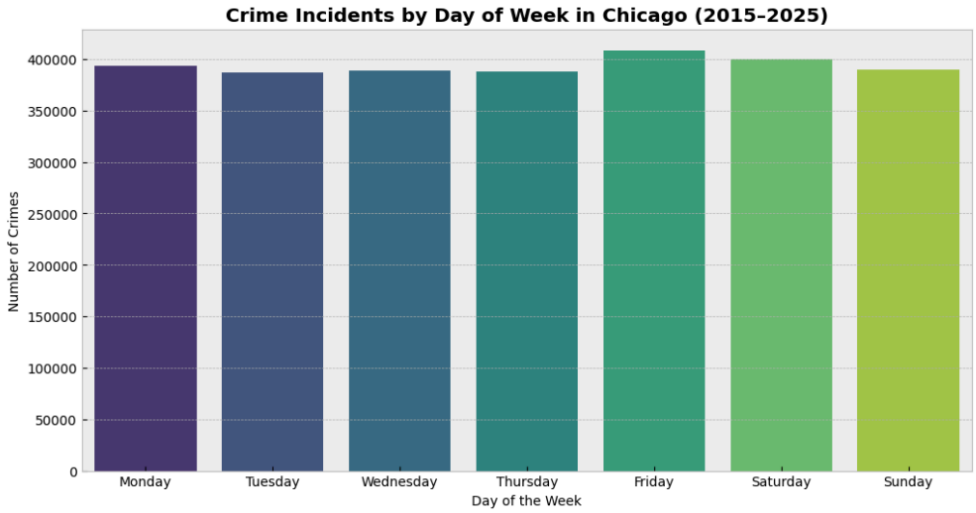

Figure 6: Crime incidents aggregated by day of week (2015–2025).

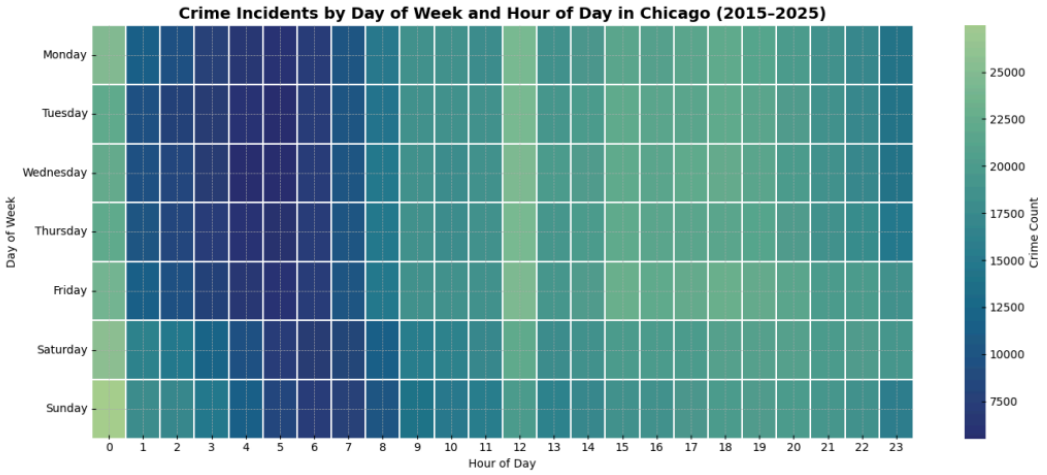

Figure 7: Heatmap of crime incidents by day of week and hour of day (2015–2025).

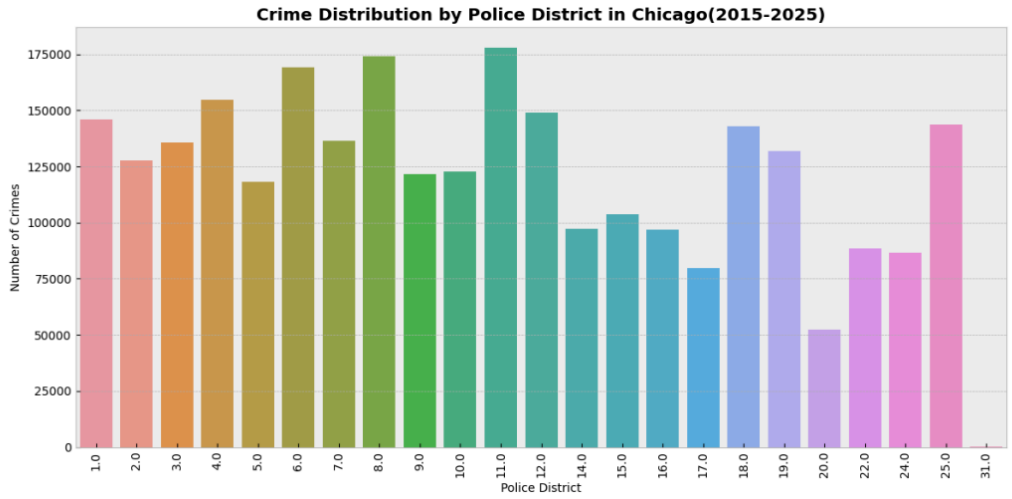

Figure 8: Distribution of crime incidents by police district (2015–2025).

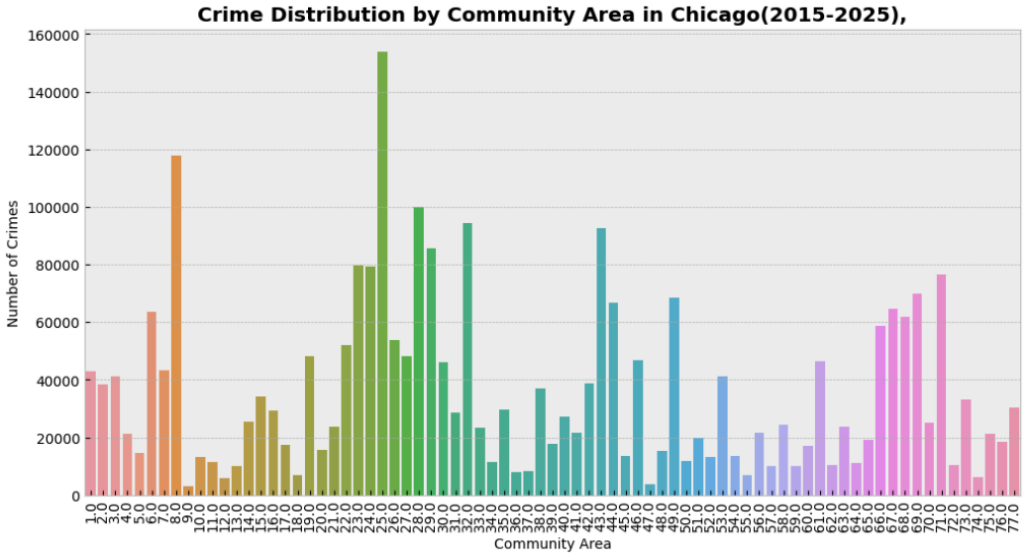

Figure 9: Distribution of crime incidents by community area (2015–2025).

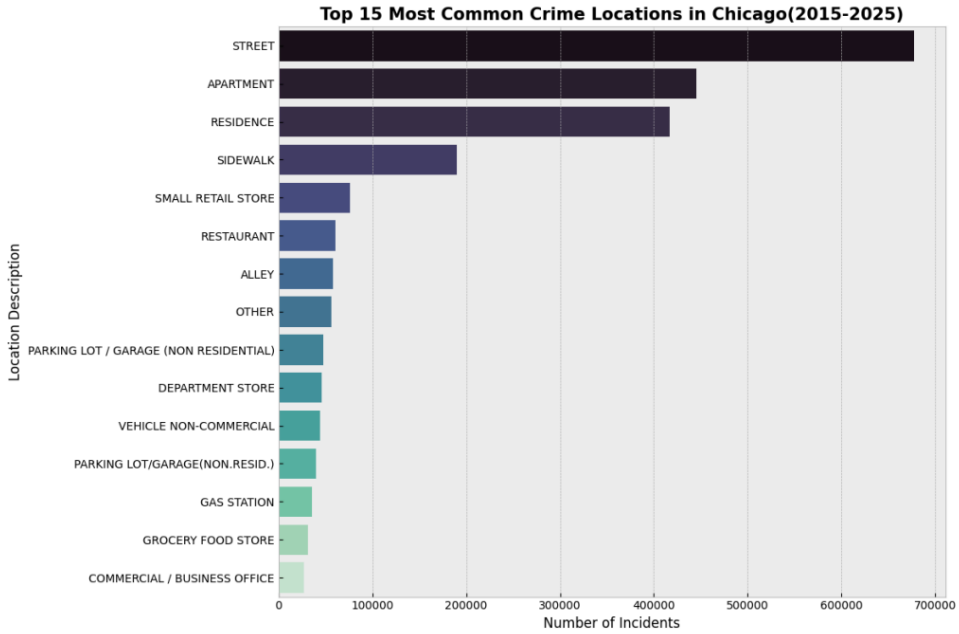

Figure 10: Top 15 most common location descriptions for reported crime incidents (2015–2025).

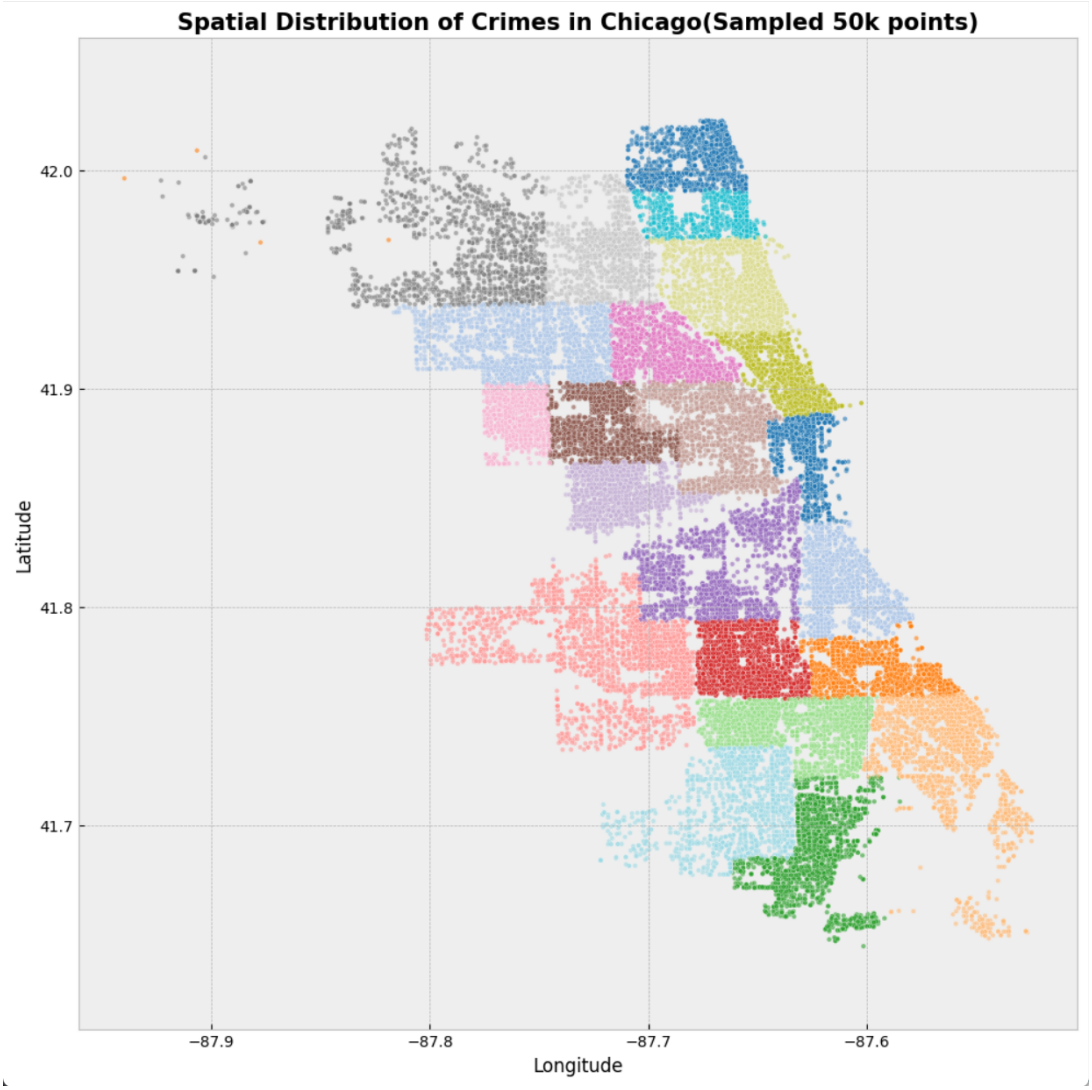

Figure 11: Spatial distribution of crime incidents in Chicago (random sample of 50k records), coloured by police district.

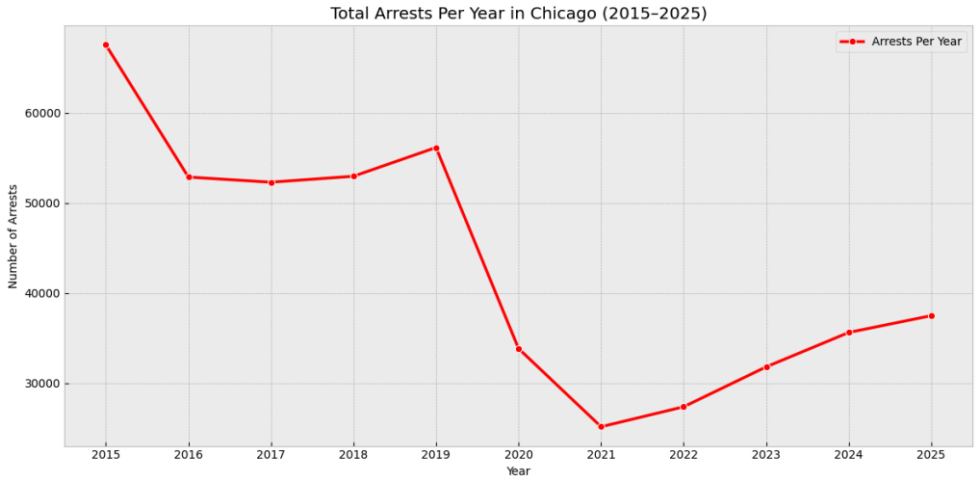

Figure 12: Annual trend of incidents that resulted in an arrest (2015–2025).

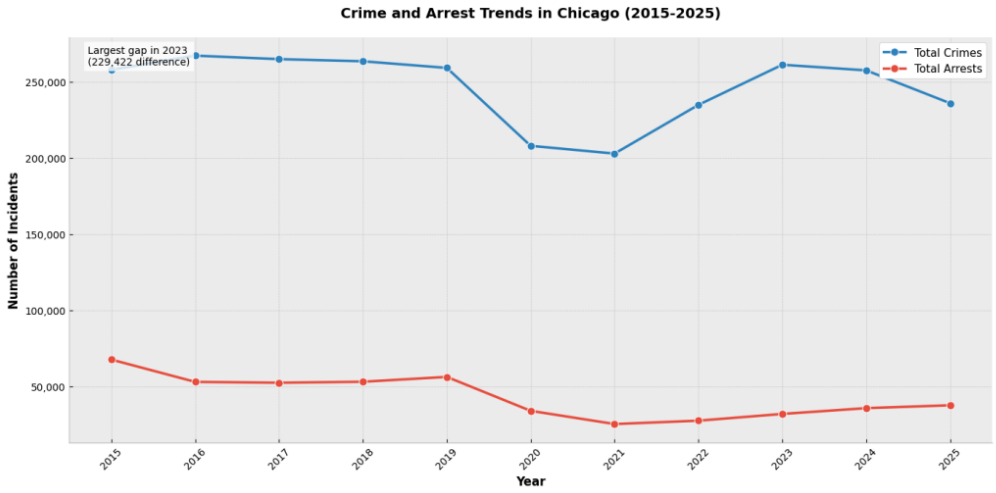

Figure 13: Comparison of annual total crime incidents and arrest incidents (2015–2025).

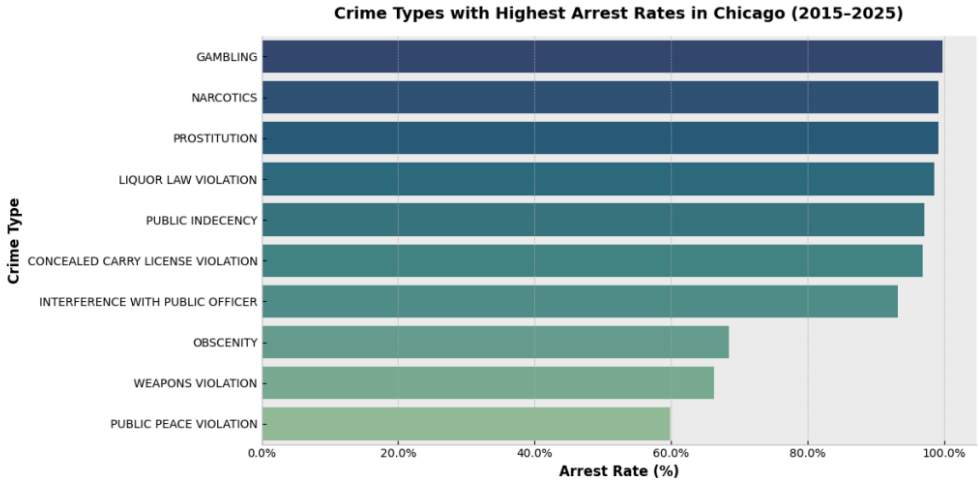

Figure 14: Crime categories with the highest arrest rates, computed as arrests divided by total incidents by primary type (2015–2025).

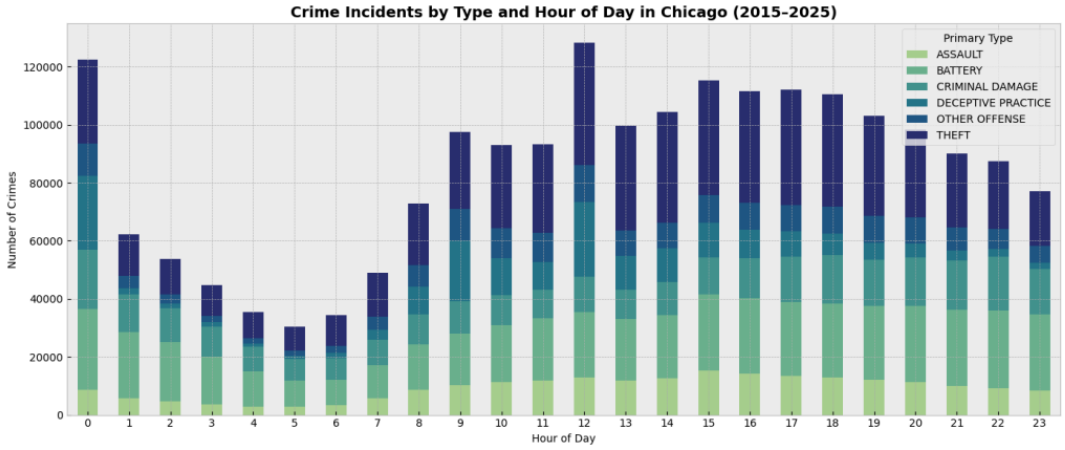

Figure 15: Incidents by hour of day for the top primary crime categories (stacked counts; 2015–2025).

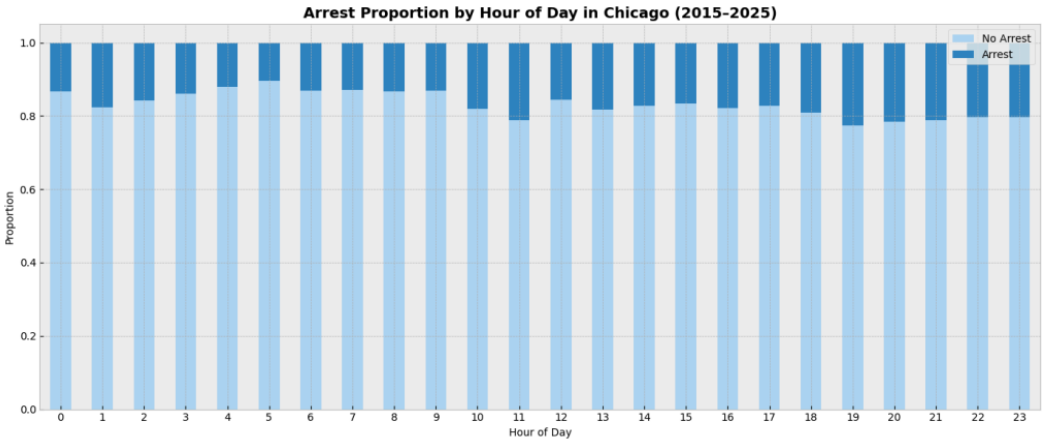

Figure 16: Proportion of incidents resulting in arrest by hour of day (2015–2025).

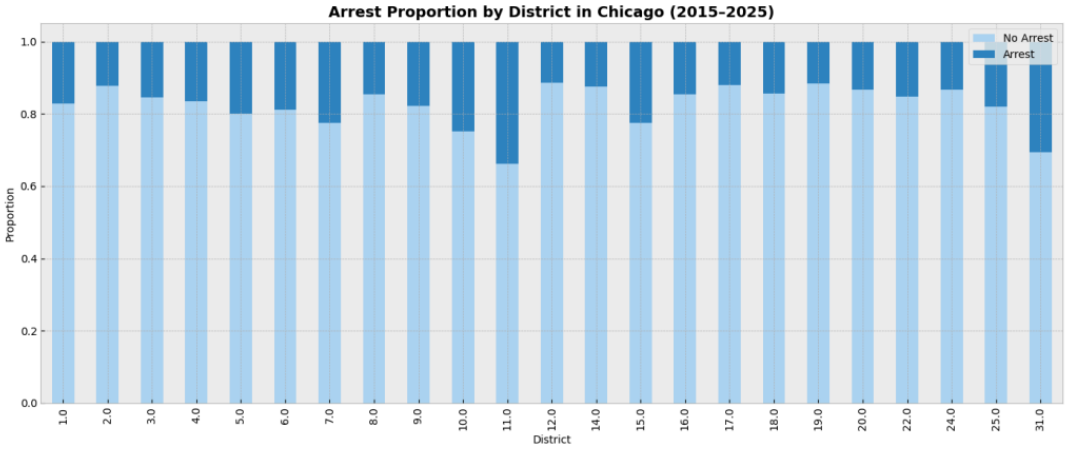

Figure 17: Proportion of incidents resulting in arrest by police district (2015–2025).

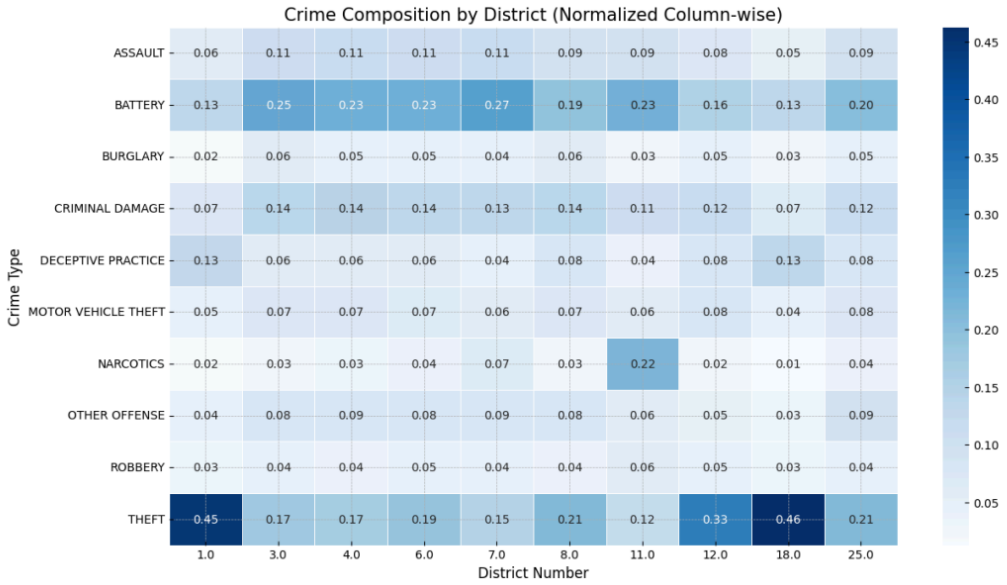

Figure 18: Crime composition by district for the top districts and top crime categories; values are column-normalised to show within-district composition (2015–2025).