Crime Analysis in Chicago Using Data-Driven Techniques

Mahita Maddipati
Virginia Polytechnic Institute and State University
Falls Church, Virginia, USA
mmahita@vt.edu

Prerna Sehrawat
Virginia Polytechnic Institute and State University
Falls Church, Virginia, USA
prernasehrawat23@vt.edu

Siddhi Kasera
Virginia Polytechnic Institute and State University
Falls Church, Virginia, USA
siddhikasera@vt.edu

Abstract

This research investigates Chicago crime patterns through modern data-driven methodologies, offering new insights into urban safety challenges. Using clustering algorithms like DBSCAN and predictive models like Random Forest, we identified high-crime hotspots and explored temporal variations in crime occurrences. The analysis revealed significant patterns, including the concentration of crime around transit hubs and underserved neighborhoods, alongside peaks during summer and late-night hours. Additionally, graph theory was employed to evaluate Chicago's transit network, pinpointing critical nodes where strategic police presence could enhance safety and response times. The study further highlights the predictability of specific crime types, with theft emerging as the most predictable class, achieving near-perfect accuracy across models. By combining predictive analytics, spatial analysis, and network modeling, this research provides actionable strategies for urban planners and law enforcement agencies. These findings offer a pathway to more targeted interventions, improved resource allocation, and the development of safer, more resilient urban environments.

Keywords

Crime pattern analysis, redictive modeling, spatial clustering, DBSCAN algorithm, Random Forest classification, crime hotspots, temporal crime trends, seasonal crime variation, graph theory, transit network analysis, high-risk crime nodes, urban safety, data-driven urban planning, community safety, law enforcement optimization

ACM Reference Format:
Mahita Maddipati, Prerna Sehrawat, and Siddhi Kasera. 2018. Crime Analysis in Chicago Using Data-Driven Techniques. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Urban crime is a multifaceted issue that impacts public safety, community well-being, and the efficient allocation of resources. For cities like Chicago, where crime rates remain a significant concern, understanding the dynamics of crime is critical for effective prevention and intervention strategies. Crime is not evenly distributed across time and space; instead, it follows discernible patterns influenced by socioeconomic factors, infrastructure design, and community engagement. This complexity necessitates innovative, data-driven approaches to identify these patterns and design actionable solutions.

This research integrates advanced analytical tools, including clustering algorithms, predictive modeling, and graph theory, to uncover meaningful insights into Chicago's crime landscape. By leveraging historical crime data, we identified hotspots where criminal activities are most concentrated. DBSCAN clustering, in particular, effectively isolates irregularly shaped clusters of high-crime activity, often located near transit hubs and under-policed areas. These findings highlight the spatial distribution of crime and its relationship with infrastructure and community design.

Temporal analysis revealed additional layers of complexity. Seasonal trends show a noticeable increase in crime rates during summer, coinciding with heightened outdoor activity and social interactions. Late-night hours also emerged as high-risk periods for crimes such as battery, robbery, and theft, underscoring the need for targeted interventions during these times. Predictive modeling provided another dimension to the analysis, demonstrating that certain crimes, like theft, follow consistent patterns and can be accurately predicted using Random Forests. Other crimes, such as vandalism, showed moderate predictability, indicating that interventions may need to focus on a broader set of factors for effective prevention.

The study also employed graph theory to analyze Chicago's transit network, identifying critical nodes such as high-traffic train stations where increased police presence could have the most significant impact. These findings offer actionable insights for optimizing resource allocation, enhancing surveillance, and improving overall safety in the city.

By integrating insights from spatial analysis, temporal trends, and predictive modeling, this research provides a comprehensive framework for addressing urban crime. It highlights high-risk areas and peak times and offers practical recommendations for law enforcement and urban planners. This study underscores the importance of using data-driven strategies to create safer, more equitable, and resilient urban environments.

2 Related Research

"The Trend of Crime in Chicago" by Harry Willbach (1941) analyzed 21 years of arrest data (1919–1939) for males aged 16 and

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

older. Crimes were categorized into Crimes Against the Person (e.g., murder, assault, rape) and Crimes Against Property (e.g., larceny, burglary, robbery). The study investigated whether Chicago's crime trends paralleled the downward trajectory observed in New York City during the same period. Descriptive statistics and linear regression models were employed to identify trends and smooth fluctuations in annual data. While Willbach's approach provided valuable historical insights, our research leverages modern machine-learning techniques to classify crime types and predict patterns, enabling law enforcement to allocate resources more efficiently and implement proactive crime prevention strategies.

"Crime Pattern Detection Using Data Mining" by Shyam Varan Nath proposed using clustering algorithms to detect crime patterns and expedite crime-solving. The study applied K-means clustering to sheriff's office crime data and validated its results, identifying key attributes using expert-based semi-supervised learning. A dynamic weighting scheme for attributes was introduced to better categorize crime types. However, the study lacked predictive capabilities for identifying crime hotspots over time. Building on this, our research employs advanced techniques, such as DBSCAN clustering, to handle irregular spatial clusters and noise in crime data.

In addition, we integrate network analysis to uncover relationships between crime locations and types, offering more profound insights into urban crime dynamics. Our approach also incorporates temporal analysis, examining time-of-day and seasonal crime trends to provide actionable insights, complemented by interactive visualization tools like Folium and heatmaps. These enhancements address the limitations of Nath's methodology and adapt it to current crime data for improved scalability and applicability.

The association between weather and the number of daily shootings in Chicago (2012–2016) by Paul M. Reeping and David Hemenway examined the influence of weather conditions, such as temperature, humidity, and precipitation, on daily shootings. The study highlighted how high temperatures and deviations from historical averages increased shooting incidents, accounting for factors such as weekends and holidays. Although their work focused on daily weather variations and shootings, our research broadened the scope by analyzing seasonal trends across all types of crime. This temporal analysis offers a more comprehensive understanding of how environmental factors impact crime patterns and provides insight into crime prevention strategies tailored to seasonal variations. Building on the and findings of these previous studies, our research bridges critical gaps, combining spatial, temporal, and predictive analyses to deliver a more dynamic, scalable, and actionable framework for understanding urban crime.

3 Methodology

To analyze crime patterns in Chicago, we employed a comprehensive methodology combining data preprocessing, spatial and temporal analysis, clustering techniques, and network modeling.

3.1 Data Preprocessing

The dataset was processed using Pandas, Matplotlib, and Seaborn libraries. Initial steps involved inspecting the dataset structure, checking data types, handling missing values, normalizing formats, and removing duplicates. Latitude and longitude coordinates were validated for accurate spatial analysis. Temporal attributes were converted into datetime formats to derive features like hour, day of the week, and seasonality, enabling meaningful time-based analysis. Cleaned and preprocessed data were saved into new files for further analysis.

3.2 Spatial Visualization and Analysis

We used Folium for interactive geographic visualizations and overlaid spatial clusters and heatmaps to identify crime hotspots. GeoPandas facilitated geospatial data management, while Shapely was used for geometry construction and manipulation, such as analyzing point distributions. Contextily enriched our maps with geographic context, and shapefiles representing Chicago's wards and neighborhoods were integrated to provide granular spatial insights.

We utilized DBSCAN (Density-Based Spatial Clustering of Applications with Noise) for hotspot detection, a clustering algorithm well-suited for identifying irregularly shaped clusters in noisy datasets. DBSCAN parameters— ϵ (neighborhood radius) and MinPts (minimum points to form a cluster)—were fine-tuned through iterations and visual validation using heatmaps and geographic overlays. Using GeoPandas, the resulting clusters were overlaid onto Chicago's map, highlighting areas with high crime density. Transit routes and stops were added to identify infrastructure within these clusters. Clusters containing more than 5,000 nodes were classified as "crime hotspots," further analyzed for their proximity to police stations.

3.3 Temporal Analysis

Temporal patterns were explored by grouping data by seasons (Winter, Spring, Summer, and Fall) and deriving insights into seasonal trends and time-of-day variations. Aggregated data for crime counts by hour and day of the week were visualized using line plots, bar plots, and heatmaps generated with Matplotlib and Seaborn. These visualizations revealed temporal hotspots and provided actionable insights for resource allocation.

3.4 Predictive Modeling

Robust validation techniques such as Stratified K-Fold cross-validation were employed to address class imbalances. Models, including Gaussian Naive Bayes (GNB), Random Forest, Decision Tree, and Logistic Regression, were trained to predict crime types. Performance metrics like accuracy, precision, recall, F1-score, and ROC-AUC ensured thorough evaluations. Gaussian Naive Bayes, leveraging Bayes' theorem, performed well on certain datasets, efficiently handling continuous features. Predictive modeling enabled the identification of crime patterns and the classification of crime types.

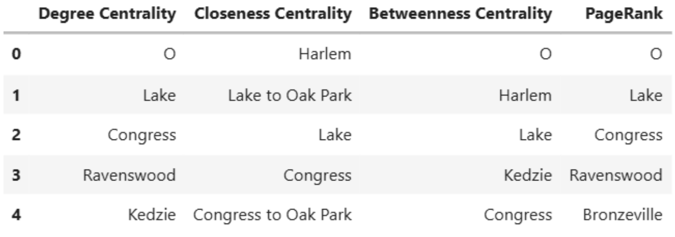
3.5 Network Modeling

The transit system and its connections were modeled as a graph G=(V,E)G=(V,E), where nodes V represented transit stops and edges

E denoted routes. Using NetworkX, centrality measures were applied:

• **Degree Centrality:** Highlighted nodes with the highest direct connections, identifying major transit hubs.
• **Betweenness Centrality:** Identified nodes frequently traversed in shortest paths, marking critical areas for crime flow.
• **Closeness Centrality**: Measured nodes' proximity to others, identifying highly connected locations.

Clustering coefficients assessed local connectivity, and the Giant Connected Component (GCC) was analyzed for its role in overall network efficiency. Recommendations were provided to enhance security at critical transit nodes based on their centrality and vulnerability.



	Degree Centrality	Closeness Centrality	Betweenness Centrality	PageRank
0	O	Harlem	O	O
1	Lake	Lake to Oak Park	Harlem	Lake
2	Congress	Lake	Lake	Congress
3	Ravenswood	Congress	Kedzie	Ravenswood
4	Kedzie	Congress to Oak Park	Congress	Bronzeville

Figure 1: Centrality of critical nodes

3.6 Integrated Analysis

Combining temporal trends, DBSCAN clustering, and network theory enabled a holistic analysis of crime patterns. Crime hotspots were identified, and gaps in police station coverage were mapped to inform resource allocation. Insights from transit network modeling supported strategic interventions, such as reinforcing security at critical hubs and improving overall network resilience.

4 Experiment

4.1 Data Description

The datasets used in this study provided a detailed understanding of crime patterns and their connection to transit infrastructure in Chicago. The main dataset, sourced from the Chicago Data Portal, included 256,020 crime incidents with information such as the date and time of occurrence, geographic coordinates, detailed crime descriptions, and administrative data like police beats and ward numbers. This dataset allowed for an in-depth analysis of how crimes are distributed across the city and how they vary over time. To complement this, data from the Chicago Transit Authority (CTA) was incorporated, which included records on transit lines, detailing routes for L rail lines, and transit stops, providing spatial information for 302 stations.

Additionally, data on police station locations was used to assess their proximity to high-crime areas. Spatial visualizations were enhanced using shapefiles to create accurate overlays of transit routes, stops, and police stations. Together, these datasets formed a strong foundation for analyzing crime trends and their relationship to urban infrastructure, offering valuable insights to improve safety and optimize resource allocation.

To complement this, data from the Chicago Transit Authority (CTA) was incorporated, which included records on transit lines, detailing routes for L rail lines, and transit stops, providing spatial information for 302 stations. Additionally, data on police station locations was used to assess their proximity to high-crime areas. Spatial visualizations were enhanced using shapefiles to create accurate overlays of transit routes, stops, and police stations. Together, these datasets formed a strong foundation for analyzing crime trends and their relationship to urban infrastructure, offering valuable insights to improve safety and optimize resource allocation.

4.2 Our analysis

An analysis of the frequency of crimes revealed significant patterns among different types of crimes.

4.2.1 *Description of Crime Types.*

- **Top Crime Types:** Theft, Assault, and Battery emerged as the most prevalent categories, underscoring their prominence in the dataset.
- **Visualization:** A bar chart was generated to illustrate the count of each crime type, providing a clear representation of their distribution.
- **Key Insight:** Theft accounted for the highest proportion of crimes, highlighting the need for targeted interventions to address this category effectively.An analysis of crime frequencies revealed significant patterns among different crime types.

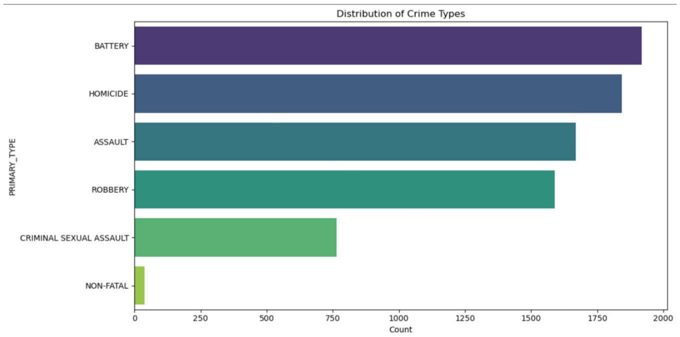


Figure 2: Distribution of Crime Types

4.2.2 ***Temporal Analysis.*** Temporal trends were examined by analyzing year-over-year and quarterly variations in crime occurrences.

- **Yearly Trends:** Yearly crime counts exhibited fluctuations, with noticeable spikes during specific years. These anomalies may be attributed to policy shifts, socio-economic factors, or changes in reporting practices.

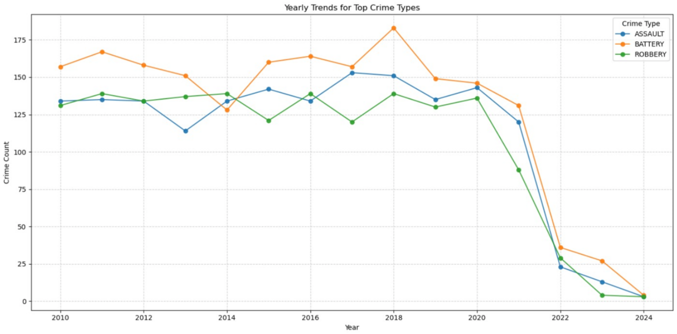


Figure 3: Yearly Trends for Top Crime Types

- **Quarterly Trends:** Seasonal variations were evident, with crime rates peaking in Q3 (summer months). This aligns with established patterns linking increased outdoor activity to higher crime rates.

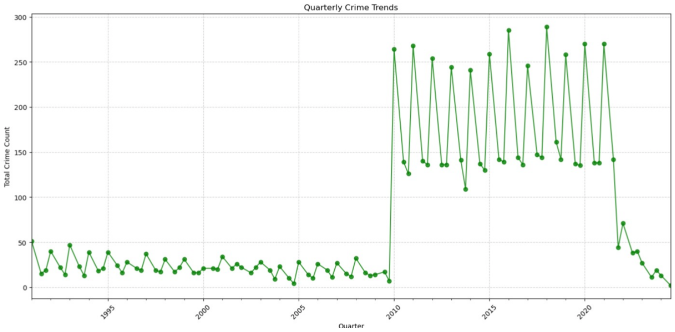


Figure 4: Quarterly Crime Trends

4.3 Models

To predict crimes in Chicago, we evaluated four machine learning models: Logistic Regression, Random Forest, Decision Tree, and Gaussian Naive Bayes (GNB). Among these, Random Forest consistently emerged as the best-performing model, delivering high precision, recall, and F1 scores across most crime classes.

4.3.1 Model Performance Overview.

- **Random Forest:** Demonstrated exceptional performance, particularly for Theft and Fraud, achieving near-perfect recall (1.0) and F1 scores (0.98). It consistently provided robust predictions across all classes.
- **Decision Tree:** Showed excellent recall for Fraud (0.95) and Murder (0.98) but struggled with low precision for other classes, leading to higher false-positive rates.
- **Logistic Regression:** Underperformed in most classes, particularly for Arson, Assault, and Vandalism, with limited predictability.
- **Decision Tree:** Performed strongly, closely matching Random Forest in predicting Theft and Fraud.

Algorithm	Class	Recall	Precision	F1-Score
Random Forest	Theft	1.00	0.98	0.98
Random Forest	Fraud	1.00	0.98	0.98
Decision Tree	Theft	0.98	0.97	0.97
Decision Tree	Fraud	0.98	0.97	0.97
Gaussian Naive Bayes	Fraud	0.95	0.65	0.77
Gaussian Naive Bayes	Murder	0.98	0.70	0.82
Logistic Regression	Arson	0.50	0.30	0.38
Logistic Regression	Assault	0.55	0.40	0.46
Logistic Regression	Vandalism	0.60	0.45	0.51

Table 1: Performance Metrics of Machine Learning Algorithms

4.3.2 ***Model Performance Overview.*** The models were evaluated on six crime categories: Arson, Assault, Fraud, Murder, Theft, and Vandalism. The following insights were derived:

- **Theft:** Performed strongly, closely matching Random Forest in predicting Theft and Fraud. The most predictable class, with Random Forest and Decision Tree achieving near-perfect precision, recall (1.0), and F1 scores (0.98). Logistic Regression also performed well, with an F1 score of 0.92.
- **Fraud:** Random Forest excelled with an F1 score of 0.75 and recall of 0.78. GNB delivered the highest recall (0.95) but suffered from low precision (0.30), resulting in false positives.
- **Murder:** All models performed effectively. Random Forest offered the best balance with an F1 score of 0.90 and recall of 0.91, while GNB achieved the highest recall (0.98), showcasing its strength in identifying Murder cases.
- **Arson:** Random Forest and Decision Tree provided moderate performance with F1 scores (0.43) and recall (0.42). Logistic Regression and GNB struggled to capture the scattered nature of Arson crimes.
- **Assault:** Decision Tree and Random Forest outperformed other models, though overall performance was modest, with F1 scores (0.43). GNB exhibited poor recall (0.05).
- **Vandalism:** Random Forest and Decision Tree performed moderately well, with F1 scores (0.70) and recall (0.70). GNB proved ineffective, achieving a recall of only 0.05.

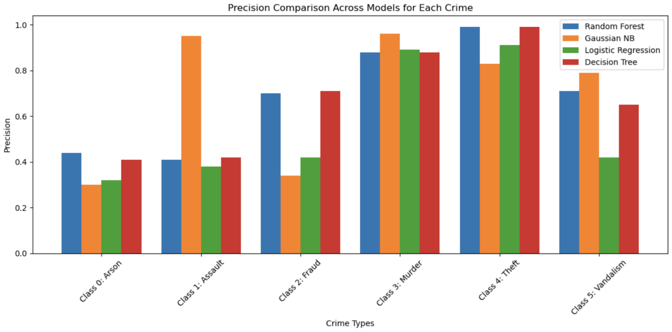


Figure 5: Precision Comparison across Model of each crime

4.3.3 *Validation and Insights.*

To ensure robust and unbiased performance evaluation, we applied Stratified K-Fold Cross-Validation. This technique accounted for imbalanced class distributions and provided consistent metrics across datasets.

- **Random Forest:** Exhibited minimal sensitivity to data splits, with consistently high accuracy and stability across folds.
- **Gaussian Naive Bayes:** Delivered consistent results, emphasizing its robustness despite lower precision for certain classes.
- **Conclusion:** Random Forest emerged as the most reliable and versatile model, making it ideal for building an effective crime prediction system.

4.3.4 *Urban Recommendations.*

Based on the predictive models and findings, we propose targeted strategies to mitigate urban crime:

- **Infrastructure Improvements:** Enhance public safety by installing better street lighting and CCTV cameras in high-crime areas, particularly to deter theft, vandalism, and assault. Emergency call boxes and improved public spaces such as parks and transport hubs can increase community safety.
- **Hotspot Policing:** Deploy law enforcement resources to high-crime areas identified through predictive models, such as Ward 20. Specific measures include increased surveillance of abandoned buildings to prevent arson and targeted patrols to address violent crimes like murder.
- **Socio-Economic Initiatives:** Introduce education and vocational training programs in high-crime wards to address socio-economic contributors to crime. For instance, areas with high illiteracy rates can benefit from youth engagement programs that offer constructive alternatives to vandalism and assault.
- **Urban Design Enhancements:** Improve visibility and accessibility in neglected spaces to reduce crime opportunities. Legal graffiti zones can be established to mitigate vandalism.
- **Technology Integration:** Leverage crime mapping and predictive analytics to dynamically allocate resources and identify emerging hotspots. These tools enhance the efficiency of urban safety systems and enable proactive interventions.

4.4 Seasonal Analysis

We examined crime trends by season, using a line graph that categorized data into Fall, Spring, Summer, and Winter.

Key Observations:

- **1990–2009:** Crime counts remained consistently low across all seasons, fluctuating between 10 and 30 incidents annually. There were no significant seasonal variations during this period.
- **2010–2019:** Crime counts spiked sharply in 2010, rising to 130–140 incidents annually across all seasons. This change likely reflects improvements in data collection rather than a real-world surge in crime. Crime rates remained high throughout this decade, with Summer consistently showing slightly higher counts than other seasons.
- **2020–2024:** A significant decline in crime counts was observed, reaching near-zero levels by 2024. This reduction may be attributed to societal disruptions during the COVID-19 pandemic, improved law enforcement strategies, or changes in reporting practices.

4.4.1 *Season-Wise Distribution of Crime Types.*

We analyzed crime types across four seasons—Fall, Spring, Summer, and Winter—highlighting key seasonal trends:

- **Summer:** Consistently reported higher crime counts for most types, including Assault, Battery, Criminal Sexual Assault, Homicide, and Robbery. This increase is likely driven by greater outdoor activity and social interactions during warmer months.
- **Winter:** Recorded the lowest crime counts, possibly due to reduced public activity in colder weather.
- **Crime Type Insights:**
  - **Assault and Battery:** Higher counts during Spring and Summer, reflecting increased social interactions.
  - **Criminal Sexual Assault:** Highest counts in Summer, attributed to increased public events and outdoor activities.
  - **Homicide:** Consistently high in all seasons, with a notable peak in summer.
  - **Robbery:** More prevalent in Summer and Spring, linked to increased foot traffic and outdoor gatherings.

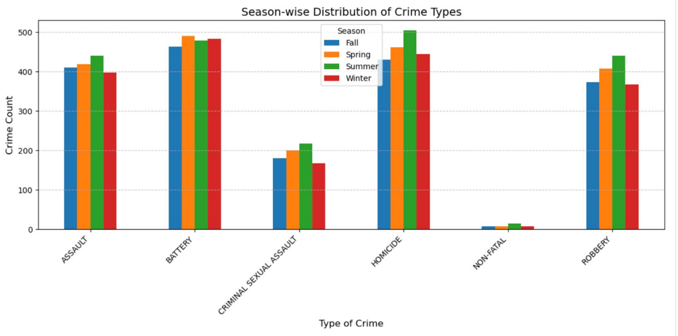


Figure 6: Season-wise distribution of Crime Types

These findings underscore the need for targeted police and surveillance during the warmer months to effectively address seasonal spikes in crime.

4.4.2 *Seasonal Crime Trends at CTA Train Locations.*

We analyzed crime trends at Chicago Transit Authority (CTA) train locations across seasons:

- **Spring and Summer:** Reported the highest crime counts, likely due to increased foot traffic and activity at transit hubs during warmer months.
- **Fall and Winter:** Recorded lower crime counts, reflecting decreased public activity and reduced opportunities for crime.

In summary, Spring and Summer exhibit higher crime rates at CTA train locations due to increased usage and public activity

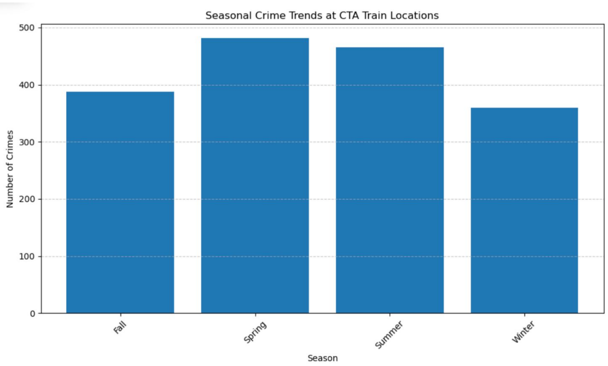


Figure 7: Seasonal Crime Trends at CTA Train Locations

during the warmer months. Fall and winter see a notable decline in crime counts, reflecting reduced foot traffic and activity.

4.4.3 Top 5 Crimes on CTA Trains by Time of Day . We categorized crimes into Morning, Afternoon, and Late Night, analyzing Assault, Battery, Criminal Damage, Robbery, and Theft:

- **Late Night:** Consistently recorded the highest crime counts in all categories, reflecting reduced public presence and lower surveillance during these hours.
- **Afternoon:** Showed notable crime counts for Theft and Battery, linked to increased passenger traffic.
- **Morning:** Reported the lowest crime counts, reflecting limited criminal activity during early hours.

The predominance of late-night crimes highlights a lack of monitoring during these hours, making trains and stations vulnerable to vandalism and property damage. For Robbery, crime counts are highest during Late Night, followed by the Afternoon. The Morning period sees minimal incidents. The concentration of robberies during Late Night suggests opportunistic crimes when fewer passengers and reduced security measures make stations and trains easier targets. Enhanced security measures, including surveillance and increased patrols, are essential during Late Night and Afternoon hours to mitigate risks and improve passenger safety.

4.5 Datasets Analysis

This study utilized a primary dataset comprising 256,020 crime incidents in Chicago. The dataset included key temporal, spatial, and descriptive attributes such as:

- **Date and Time of Occurrence:** Specific timestamps of crimes.
- **Latitude and Longitude:** Spatial coordinates for geospatial analysis.
- **Crime Descriptions:** Detailed categorizations of incidents.
- **Administrative Data:** Information such as police beat and ward numbers.

The data was sourced from the Chicago Data Portal. Additionally, for analyzing transit infrastructure, datasets from the Chicago Transit Authority (CTA) Data Portal were incorporated. These included:

- **CrimeTransit Lines Dataset:** Containing 153 entries detailing L rail lines, such as the Red and Blue Lines, with route geometries.
- **Transit Stops Data set:** Comprising 302 entries with station names and their latitude-longitude coordinates.
- **Police Stations Dataset:** Mapping the locations of police stations across Chicago, was vital for evaluating their proximity to identified crime hotspots.

To enrich spatial analysis, these datasets were supplemented with shapefiles for transit routes, stops, and police station overlays, ensuring accurate geospatial visualizations.

4.6 Experimental Setup

The analysis was conducted using Python 3.12 and employed the following libraries and tools:

Tool/Library	Purpose
Pandas and NumPy	For data manipulation and exploration.
Matplotlib and Seaborn	For generating heatmaps, bar plots, and other visualizations.
Folium	For interactive spatial visualizations.
GeoPandas and Contextily	For overlaying spatial data on basemaps.
Shapely.geometry	For geometric analysis of spatial relationships.
NetworkX	For modeling transit networks and calculating centrality measures.
DBSCAN	For density-based clustering of crime incidents.

Table 2: Tools and Libraries Used for Data Analysis and Visualization

DBSCAN was used to identify spatial crime clusters, while K-means clustering was also evaluated as a benchmark. NetworkX enabled modeling of transit stations and routes, allowing the identification of critical nodes based on centrality measures to optimize resource allocation.

4.7 Trends

4.7.1 ***Temporal Analysis****.* We categorized crimes into Morning, Afternoon, and Late Night, analyzing Assault, Battery, Criminal Damage, Robbery, and Theft:

- **Crime Timing:**
  - Incidents peaked at midnight and between 12 PM and 11 PM, marking these as high-risk periods requiring enhanced law enforcement.
  - The lowest crime rates were observed between 4 AM and 7 AM, suggesting reduced criminal activity during early morning hours.
- **Monthly Trends:**
  - Crime rates were significantly higher during summer months (June to August), with notable drops in December and January.
  - These trends align with increased outdoor activity during warmer months, necessitating proactive measures during these periods.

4.7.2 ***Categorical Analysis****.* An in-depth examination of crime categories revealed theft as the most prevalent type, followed by battery, criminal damage, and assault. Specific findings included:

- **Theft and Battery:** Predominantly occurred late at night or early in the morning.

- **Criminal Damage:** Displayed a more uniform distribution throughout the day.
- **Seasonal Trends:** A heatmap showed theft, battery, and weapons violations were most frequent during summer, consistent with the broader temporal analysis.

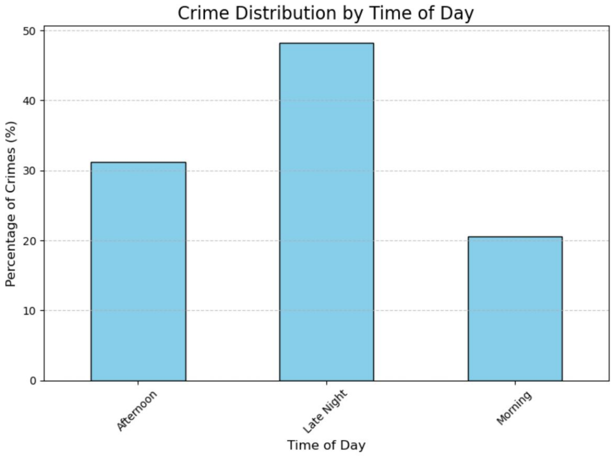


Figure 8: Crime Distribution by Time of Day

Arrest likelihood varied across crime categories, with higher rates observed for concealed carry violations and liquor law infractions, whereas theft showed a notably lower arrest likelihood.

4.8 DBSCAN Clustering Results

Using the K-distance plot, an optimal epsilon (ϵ) value was determined for DBSCAN clustering. The refined parameters identified multiple clusters:

- **Cluster 0:** Contained 94,938 incidents, making it the densest cluster.
- **Clusters 1 and 2:** Followed with 31,788 and 21,685 incidents, respectively.

These clusters correlated with high-crime areas in Chicago, primarily in the northern, central, and southern regions.

- **Transit Correlation:** Cluster 0 intersected with 68 transit stations, highlighting a strong relationship between crime density and transit activity.
- **Police Station Proximity:** Several high-crime clusters, particularly in southern and central regions, lacked adequate police station coverage, revealing critical gaps in law enforcement resources.

4.9 Police Station and Graph Theory Analysis

By leveraging NetworkX, we evaluated transit routes to identify critical nodes for optimizing police station placement:

- **Key Nodes:** Stations like Harlem, Lake, and Kedzie emerged as high-betweenness centrality nodes, marking them as strategic locations for new police stations.

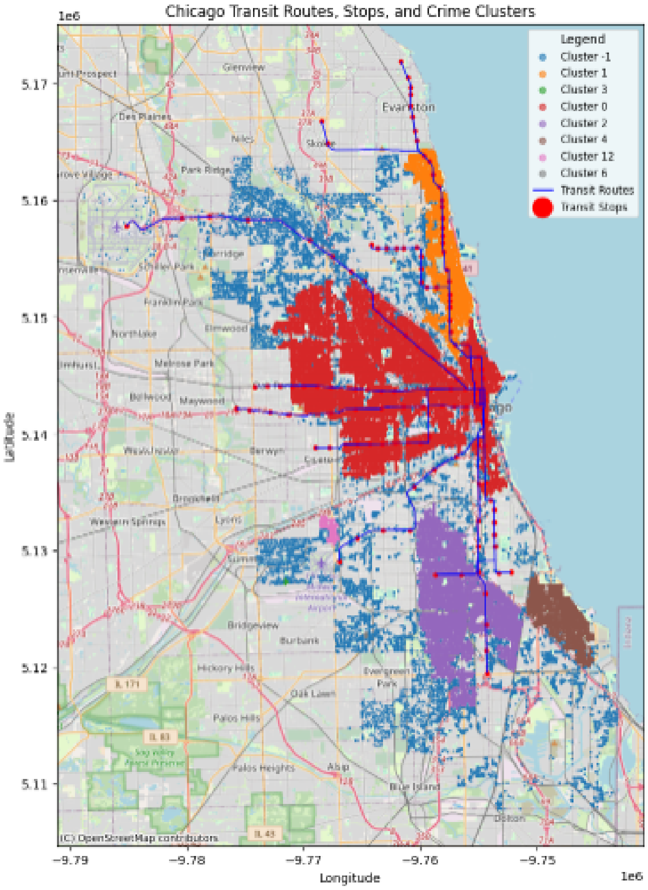


Figure 9: Chicago transit routes and crime clusters

- **Bottlenecks** Critical connections, such as Harlem-O and Kedzie-Midway, were identified as vulnerable points that could disrupt network flow if compromised.

Simulations of graph densification demonstrated improved connectivity by reducing the diameter of the Giant Connected Component (GCC) to 17, enhancing network resilience. However, the low clustering coefficient indicated sparse overall connectivity.

4.10 Insights and Recommendations

The DBSCAN clustering and graph theory analyses yielded actionable insights for improving crime prevention strategies:

- **Prioritize High-Crime Clusters:** Transit hubs with high crime density, such as those in Cluster 0, require increased police presence and surveillance.
- **Strategic Law Enforcement Placement:** Stations with high centrality, such as Harlem and Lake, should be prioritized for resource allocation to improve coverage and response times.
- **Enhance Transit Connectivity:** Adding routes or connections to low-density areas can reduce crime opportunities and improve transit safety.

- **Temporal Crime Management:** Tailoring law enforcement efforts to peak crime hours, especially during late-night and summer months, can optimize resource utilization.

Incorporating these strategies, along with real-time crime mapping and AI-powered surveillance, offers the potential to significantly enhance urban safety and foster resilient communities

4.11 Result

Our analysis of Chicago crime data revealed significant temporal, spatial, and categorical patterns. Theft emerged as the most prevalent and predictable crime type, with Random Forest outperforming other models in precision, recall, and F1 scores across all categories. Seasonal trends highlighted increased crime rates during summer months and late-night hours, emphasizing the need for targeted interventions during these periods. Spatial clustering using DBSCAN identified high-crime clusters near transit hubs, underscoring gaps in police station coverage in key areas. These insights suggest that combining predictive modeling, enhanced infrastructure, and strategic law enforcement placement can effectively reduce crime and improve urban safety.

5 Discussion

The findings of this study offer valuable insights into crime patterns in Chicago, with significant implications for urban planning, law enforcement, and public policy. The prevalence of theft as the most frequent crime type underscores the need for targeted preventive measures, including enhanced surveillance, public awareness campaigns, and community engagement initiatives. Seasonal variations, with higher crime rates during summer months, call for season-specific strategies such as increased patrols and resource allocation during warmer periods. Temporal trends revealing yearly fluctuations suggest the influence of external factors, such as socio-economic changes or policy shifts, which could guide adaptive planning.

The superior performance of the Random Forest model in predicting crimes, particularly theft and fraud, highlights its potential as a decision-making tool for dynamic resource allocation and hotspot identification. High-crime clusters near transit hubs further support the strategic deployment of law enforcement and surveillance. However, ethical considerations must guide the implementation of these insights to ensure privacy, fairness, and equity, avoiding systemic biases while fostering community trust.

In addition to enforcement measures, socio-economic programs addressing root causes of crime, such as education and employment initiatives, should be prioritized for long-term impact. Together, these findings and recommendations provide a comprehensive framework for improving urban safety and resilience while ensuring ethical and effective resource utilization.

6 Conclusion

This study utilized data mining techniques, including DBSCAN and graph theory analysis, to uncover crime patterns and optimize resource allocation in Chicago. By analyzing historical datasets and spatial data, we identified key crime clusters, such as Cluster 0, which contained over 94,000 incidents concentrated near transit hubs, and highlighted critical gaps in police station coverage. Graph theory metrics pinpointed transit stations like Harlem, Lake, and Kedzie as strategic nodes for enhanced law enforcement placement, ensuring better coverage in high-crime areas.

Our machine learning evaluation of Random Forest, Logistic Regression, Decision Tree, and Gaussian Naive Bayes revealed Random Forest as the most reliable model, delivering consistently high precision, recall, and F1 scores across crime categories like Theft and Fraud. Temporal and seasonal analysis identified Late Night and Summer as high-risk periods, with crimes such as Theft and Battery peaking during these times, emphasizing the importance of tailored interventions during these periods.

Recommendations include deploying enhanced surveillance, optimized police presence, and AI-powered real-time monitoring at critical nodes to address spatial and temporal crime trends. Additionally, improving transit network connectivity through graph densification can mitigate crime in isolated areas. These data-driven insights empower urban planners and law enforcement to implement targeted strategies that enhance public safety and foster resilient urban environments.

7 Acknowledgments

We would like to express our sincere gratitude to our professor, Dr. Naren Ramakrishnan, for his invaluable time spent reviewing our research and providing insightful feedback. We also extend our thanks to the Data.gov portal for making crime datasets accessible and to the Chicago Transit Authority (CTA) for providing their transit dataset, which significantly enriched our research.

8 Author Contributions

The project began with all three authors collaboratively identifying and reviewing relevant research papers. After a thorough discussion, we selected key references and formulated a plan of action. Prerna took the lead on exploratory data analysis (EDA), focusing on data cleaning and initial analyses. Siddhi was responsible for running machine learning models to derive insights from the data. Mahita contributed by analyzing transit routes, performing clustering and evaluating their urban relevance. Throughout the project, we held regular meetings to discuss our findings and strategize on how to expand our research. After finalizing our analyses, all authors collaborated on writing and refining the research paper to ensure clarity and coherence.

9 GitHub Repository

The code and supplementary materials for this research are available on GitHub: Urban-Computing-Final-Project.

References

[1] Abdullah Alqahtani. 2019. Crime analysis in Chicago city. In *2019 10th International Conference on Information and Communication Systems (ICICS)*. IEEE, 185–190.

[2] Author(s). 2007. Crime pattern detection using data mining. In *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on*. IEEE.

[3] Anthony A. Braga, Andrew V. Papachristos, and David M. Hureau. 2014. The effects of hot spots policing on crime: An updated systematic review and meta-analysis. *Justice Quarterly* 31, 4 (2014), 633–663.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise (DBSCAN). In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD ’96)*. AAAI Press, 226–231.

[5] Kai Hou, Lei Zhang, Xiaotong Xu, Fan Yang, Bing Chen, Wei Hu, and Rong Shu. 2022. High ambient temperatures are associated with urban crime risk in Chicago. *Science of the Total Environment* 843 (2022), 158846.

[6] Andrew V. Papachristos, David M. Hureau, and Anthony A. Braga. 2013. The Corner and the Crew: The Influence of Geography and Social Networks on Gang Violence. *American Sociological Review* (2013).

[7] Robert J. Sampson and William J. Wilson. 1995. Toward a theory of race, crime, and urban inequality. In *Crime and Inequality*, John Hagan and Ruth D. Peterson (Eds.). Stanford University Press, XX–XX.

[8] William R. Smith, Stephen G. Frazee, and Elizabeth L. Davison. 2000. Furthering the integration of routine activity and social disorganization theories: Small units of analysis and the study of street robbery as a diffusion process. *Criminology* 38, 2 (2000), 489–523.

[9] Ralph B. Taylor and Adele V. Harrell. 1996. *Physical environment and crime*. Technical Report. U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

[10] David Weisburd and John E. Eck. 2004. What can police do to reduce crime, disorder, and fear? *The Annals of the American Academy of Political and Social Science* 593, 1 (2004), 42–65.

[11] H. Willbach. 1941. The trend of crime in Chicago. *Journal of Criminal Law and Criminology* 31, 6 (1941), 748–757.

[6] [1] [5] [11] [2] [7] [3] [9] [8] [10] [4]