

IS597 - MACHINE LEARNING CLOUD

GROUP 6 - FINAL TECHNICAL REPORT

Title: Storying the Streets: A Narrative Intelligence Approach to Crime Data Analysis

ABSTRACT

Urban crime analysis typically relies on machine learning models that predict outcomes but provide limited interpretability for policymakers and communities. This project addresses this gap by developing a Hybrid Retrieval-Augmented Generation (RAG) system that transforms structured crime statistics into contextual narratives. Using Chicago Police Department data (2018–2025) and Chicago Tribune news articles, we demonstrate that while traditional ML models (Logistic Regression, Random Forest, XGBoost) achieve high predictive accuracy, they fail to explain *why* predictions occur or *how* patterns emerge. Our narrative intelligence framework combines semantic embeddings (SentenceTransformer), FAISS vector indexing, deterministic rule-based linking, and LLM synthesis (Google Gemini) to generate human-readable explanations. This approach demonstrates how narrative intelligence bridges the interpretability gap in urban analytics, making data-driven insights accessible and actionable for non-technical stakeholders.

1. INTRODUCTION

Crime remains one of the most persistent challenges facing major urban centers worldwide. Beyond its immediate social and economic consequences, understanding the underlying dynamics of criminal activity, its spatial distribution, temporal patterns, and evolving nature is essential for developing effective public safety strategies. In cities like Chicago, where comprehensive police-recorded crime data spanning multiple decades are publicly accessible, there exists significant potential to analyze how criminal patterns shift over time, vary across neighborhoods, and respond to broader social contexts.

Traditional crime analysis has primarily focused on two approaches: descriptive analytics that summarize historical patterns through statistics and visualizations, and predictive modeling that forecasts future crime hotspots using machine learning algorithms. While these methods provide valuable quantitative insights, they often fail to answer more nuanced questions that are critical for policy formulation. Crime analysis has evolved from descriptive statistics to sophisticated machine learning models capable of predicting crime types, arrest likelihood, and temporal patterns. However, **prediction without explanation** creates a critical gap: stakeholders cannot understand *why* a model predicts high arrest probability in one area versus another, or *what* contextual factors drive seasonal crime variations.

The Interpretability Problem

Consider a typical ML pipeline for crime analysis: - **Input:** Crime type, location, time, arrest history → **Model:** XGBoost → **Output:** “85% arrest probability” - **What’s missing:** Why 85%? What makes this incident different from the 40% probability case? What interventions might change outcomes?

Traditional interpretability methods (SHAP values, feature importance) show *which variables matter* but not *how real-world events connect* to patterns. For example: - SHAP might show “District 8 + Narcotics + Evening = High Arrest Probability” - But it cannot explain: “District 8 shows high arrest rates for narcotics because targeted enforcement increased following community complaints documented in Tribune articles July 2023”

Research Objectives

1. **Demonstrate ML limitations:** Build traditional classification models (arrest prediction) showing high accuracy but low narrative value
2. **Develop narrative framework:** Create Hybrid RAG system combining semantic search + rule-based linking + LLM synthesis
3. **Quantify temporal/spatial patterns:** Comprehensive EDA revealing crime dynamics across time and geography
4. **Generate actionable insights:** Transform statistics into contextual stories explaining crime evolution

2. DATASET & EXPLORATORY ANALYSIS

2.1 Chicago Crime Data (2018–2025)

Source: Chicago Data Portal

Records: 1,883,142 incidents

Key Attributes: Date, Primary Type, Location, Arrest, District, Ward, Coordinates

2.2 Chicago Tribune News Dataset

Source: ProQuest Archives

Articles: 15,247 crime-related articles

Coverage: 2018–2025

Attributes: Title, Full Text, Publication Date, Subject

2.3 Data Preprocessing Pipeline

2.3.1 Crime Data Preprocessing

Task	Description	Purpose
Missing Values	Checked nulls in Location Description, Ward, Community Area; replaced with “Unknown”	Prevents bias during grouping operations
Date/Time Parsing	Converted Date to datetime64; derived Year, Month, Day, Hour, Weekday	Enables temporal aggregation and time-series analysis
Data Type Conversion	Ensured numerical columns (Beat, Ward, District) as int; Boolean for Arrest, Domestic	Facilitates filtering and statistical operations
Duplicate Removal	Dropped exact duplicates based on Case Number	Ensures unique incident representation

Task	Description	Purpose
Geolocation Cleaning	Verified coordinate ranges (lat 41–42, lon -88 to -87); removed outliers	Filters erroneous geocoding
Feature Engineering	Added Season (Spring/Summer/Fall/Winter), WeekendFlag, CrimeCategory (Violent/Non-Violent)	Supports pattern recognition and classification

Block Normalization: Crime blocks in Chicago data use anonymized street addresses (e.g., “070XX S LOWE AVE”). Normalization extracts the core street name (“LOWE AVE”) to enable matching with news articles that may reference the same location without the anonymized prefix.

3.3.2 News Data Preprocessing

Task	Description	Purpose
Data Loading	Parsed ProQuest TXT exports using regex patterns to extract structured fields	Converts unstructured text to tabular format
Text Cleaning	Removed HTML tags, excess whitespace, special characters; lowercased for keyword matching	Prepares text for NLP processing
Date Parsing	Converted PubDate strings to datetime64; extracted Year and Month	Enables temporal alignment with crime data
Language Filtering	Retained only English-language articles	Ensures consistent sentiment analysis
Keyword Tagging	Applied crime-related keyword dictionary (homicide, theft, assault, robbery, etc.) to create CrimeTopic column	Enables thematic linkage to crime types
Deduplication	Dropped identical Title + PubDate combinations	Avoids overcounting reprints

2.4 Key EDA Findings

Temporal Patterns

Monthly Seasonality: Crime peaks during summer months (June-August: ~180,000 incidents/month) and declines in winter (December-February: ~135,000 incidents/month), indicating strong correlation with outdoor activity and temperature.

Weekly Distribution: Friday and Saturday show highest crime counts (~280,000 each), representing 15-16% of total weekly incidents. Weekday crime is relatively uniform, suggesting different patterns for planned vs. opportunistic crimes.

Hourly Patterns: Dramatic bimodal distribution with peaks at noon (~110,000 incidents) and midnight (~120,000 incidents). Early morning hours (4-6 AM) show minimum activity (~35,000 incidents), likely reflecting both reduced activity and reporting delays.

Yearly Trends: Post-COVID recovery visible with peaks in 2023-2024 (~260,000 incidents annually). Notable dip in 2020-2021 (~205,000) due to lockdowns and reduced mobility.

Crime Type Distribution

Top 10 Crime Types (Total Count): 1. Theft: 398,245 (21.1%) 2. Battery: 365,129 (19.4%) 3. Criminal Damage: 224,856 (11.9%) 4. Assault: 178,432 (9.5%) 5. Deceptive Practice: 156,234 (8.3%) 6. Motor Vehicle Theft: 142,567 (7.6%) 7. Other Offense: 135,892 (7.2%) 8. Burglary: 89,456 (4.8%) 9. Robbery: 86,734 (4.6%) 10. Narcotics: 78,923 (4.2%)

Arrest Rate Disparity: Overall arrest rate: 15.6% (dramatically low). However, specific crimes show much higher rates: - Gambling: ~98% arrest rate - Narcotics: ~95% - Prostitution: ~92% - Liquor Law Violation: ~88% - Weapons Violation: ~65%

Interpretation: High arrest rates for “vice” crimes (gambling, narcotics, prostitution) reflect targeted enforcement rather than crime prevalence. Property crimes (theft: 14%, criminal damage: 12%) have extremely low clearance despite high frequency.

Spatial Concentration

Top 10 Districts by Crime Count: 1. District 8: 119,234 incidents 2. District 11: 117,892 incidents 3. District 6: 116,453 incidents 4. District 4: 107,234 incidents 5. District 12: 106,789 incidents

Top 10 Community Areas: 1. Austin (25.0): 102,345 incidents 2. Near North Side (8.0): 83,567 incidents 3. Near West Side (28.0): 71,234 incidents

Geographic Distribution Insights: Crime is highly spatially clustered. The top 10 districts (out of 25) account for 62% of all incidents. West and South Chicago areas (Districts 8, 11, 6, 4) show persistent concentration, reflecting historical socioeconomic patterns.

Arrest Rate Hotspots: Geographic heatmap reveals areas with high arrest rates clustered in specific neighborhoods, potentially indicating differential policing intensity or crime type composition.

Location Type Analysis

Top Crime Locations: 1. Street: 478,234 incidents (25.4%) 2. Apartment: 332,456 incidents (17.7%) 3. Residence: 271,893 incidents (14.4%) 4. Sidewalk: 119,234 incidents (6.3%) 5. Parking Lot/Garage: 58,967 incidents (3.1%)

Street crimes dominate, but residential crimes (apartment + residence combined: 32.1%) represent a significant portion, suggesting a need for both public space and domestic violence interventions.

Distance to Infrastructure

Distance to Police Stations: Mean: 1.96 km | Median: 1.85 km Distribution shows most crimes occur within 2km of police stations, but the significant tail extends to 5km+, indicating coverage gaps in peripheral areas.

Distance to CTA Stops: Mean: 0.35 km | Median: 0.28 km Extremely close proximity indicates strong correlation between public transit and crime occurrence likely reflecting both opportunity (foot traffic) and accessibility (escape routes).

News Media Analysis

Article Distribution Over Time: Peak coverage: 2019 (~2,850 articles), declining through COVID (2020: ~1,500), partial recovery (2022: ~1,900). Coverage volume correlates moderately with crime counts ($r \approx 0.48$).

Keyword Analysis: Most frequent crime-related terms: “shooting” (6,421 mentions), “police” (8,934), “arrest” (5,823), “murder” (4,856), “assault” (3,234).

Media Bias Observation: News disproportionately covers violent crimes (shooting, murder, assault) despite property crimes being 3-4x more frequent. This creates a public perception mismatch with statistical reality.

3. LITERATURE SURVEY

Crime analytics has advanced from simple descriptive summaries to complex machine learning and deep learning models. Yet most research still places emphasis on prediction identifying where or when crime may occur while offering limited understanding of the causes, triggers, or contextual relationships behind crime patterns. The Chicago Crime Dataset, with more than twenty years of detailed spatial, temporal, and incident-level information, provides a strong foundation for exploring these predictive approaches.

Classical machine learning models such as Random Forest, Support Vector Machines, Gradient Boosting, and XGBoost have been widely used for classification and hotspot prediction. Xiong et al. (2023) showed that ensemble models, particularly XGBoost, perform well on large urban datasets like Chicago’s. However, these methods focus primarily on accuracy and offer little interpretability, limiting their usefulness for policymaking or community-level understanding.

Deep learning approaches further expanded predictive capabilities. Temporal models such as LSTM and ARIMA capture seasonal patterns in crime, revealing trends influenced by weather, holidays, and social behavior (Himanshi, 2023). Spatial models like Deep Graph Convolutional Networks treat neighborhoods as connected nodes, capturing geographic relationships between crime incidents (Fatima and Zubair, 2024). Although these models achieve high performance, they function largely as opaque systems without providing meaningful explanations of why specific areas or periods experience elevated crime.

In response, newer studies emphasize transparency and accountability in crime prediction. Interpretable machine learning approaches using feature importance and SHAP values help reveal which variables influence model decisions, improving model transparency (Zhang, 2022). Research on predictive policing systems has also raised concerns about algorithmic bias and

fairness. Heller et al. (2022) demonstrated that predictive models can amplify existing socioeconomic and racial disparities when trained purely on historical crime data, underscoring the need for greater contextual awareness.

Recent developments in spatiotemporal modeling and multi-modal integration further enhance predictive accuracy. Residual-Aware Spatiotemporal GNNs aim to reduce spatial inequality in crime forecasting (Zhuang et al., 2025), while studies integrating mobility or demographic data improve prediction across both dense and sparse urban regions (Eugenio et al., 2024; Zume et al., 2025). Despite these improvements, these systems still provide limited insight into the broader social, economic, or policy-related factors that shape crime trends.

Across the literature, several key limitations remain clear. **Current models lack narrative reasoning, focusing on what will happen rather than why it happens.** They make limited use of contextual information such as news events, policy changes, or community reports, and they continue to struggle with issues of fairness and bias due to opaque model behavior. These gaps highlight the need for approaches that integrate structured crime data with unstructured contextual information to generate more interpretable and actionable insights.

4. TRADITIONAL ML APPROACH EXAMPLE: ARREST PREDICTION

4.1 Problem Formulation

Objective: Predict whether a reported crime will result in arrest

Rationale: Arrest prediction serves two purposes: 1. Demonstrate ML can achieve high accuracy on crime data 2. Illustrate fundamental limitation: **accuracy without interpretability**

4.2 Methodology

Feature Engineering: - Temporal: Hour, Day of Week, Month, Season - Spatial: District, Ward, Community Area, Distance to Police Station, Distance to CTA - Categorical: Primary Type, Location Type - Encoding: OneHot for categorical, StandardScaler for numerical

Train–Test Split: For arrest prediction, the dataset was split **80% train / 20% test**, with SMOTE applied only on the training set. For weekly violence forecasting, we used a **time-aware split**, training on weeks before 2024 and testing on weeks from 2024-2025 to avoid temporal leakage.

Class Imbalance Handling: Original distribution: 84.4% No Arrest, 15.6% Arrest Applied SMOTE (Synthetic Minority Oversampling) to balance training data

Models Tested:

1. Logistic Regression (baseline)
2. Random Forest (ensemble method)
3. XGBoost (gradient boosting)

Temporal Aggregation for ML: All forecasting models use **weekly aggregation**. Incident-level violent crimes were grouped into weekly totals, and lag features (Lag-1 to Lag-4) were computed to capture short-term temporal patterns. We selected weekly aggregation because

daily crime counts are noisy and sparse, while weekly totals capture stable temporal dynamics used in operational crime forecasting.

Evaluation Metrics: - F1 Score (balance precision/recall) - Precision (minimize false positives) - Recall (minimize false negatives) - AUC-ROC (threshold-independent performance)

4.3 Results

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	71.2%	0.68	0.74	0.71	0.78
Random Forest	75.8%	0.72	0.78	0.75	0.82
XGBoost	77.4%	0.74	0.82	0.78	0.85

Feature Importance (XGBoost Top 10): 1. Primary Type: Narcotics (0.18) 2. Primary Type: Weapons Violation (0.14) 3. Hour: 14-18 (0.11) 4. District: 11 (0.09) 5. Location: Street (0.08) 6. Distance to Police Station (0.07) 7. Day of Week: Friday (0.06) 8. Month: July (0.05) 9. Community Area: 25 (0.05) 10. Distance to CTA Stop (0.04)

ML Limitations: Incident-level arrest prediction is constrained by limited feature granularity - models cannot capture officer discretion, unreported incidents, or socio-economic conditions. Forecasting accuracy is also limited by weekly aggregation and the absence of external drivers such as weather, events, or population mobility.

4.4 The Interpretability Gap

What XGBoost Tells Us: “A narcotics incident in District 11 on Friday afternoon has 82% arrest probability”

What XGBoost Cannot Tell Us: - *Why* District 11 has higher arrest rates than District 6 - *How* recent policy changes affected enforcement patterns - *What* community factors contribute to narcotics prevalence - *When* did arrest rates in this area start increasing - *Whether* disparities reflect legitimate enforcement differences or bias

SHAP Analysis Limitations: Even advanced interpretability tools like SHAP show feature contributions to predictions but cannot:

1. Connect features to real-world events (policy changes, community initiatives)
2. Explain temporal evolution (why patterns changed over time)
3. Provide contextual narrative (social/economic factors)
4. Suggest actionable interventions beyond feature manipulation

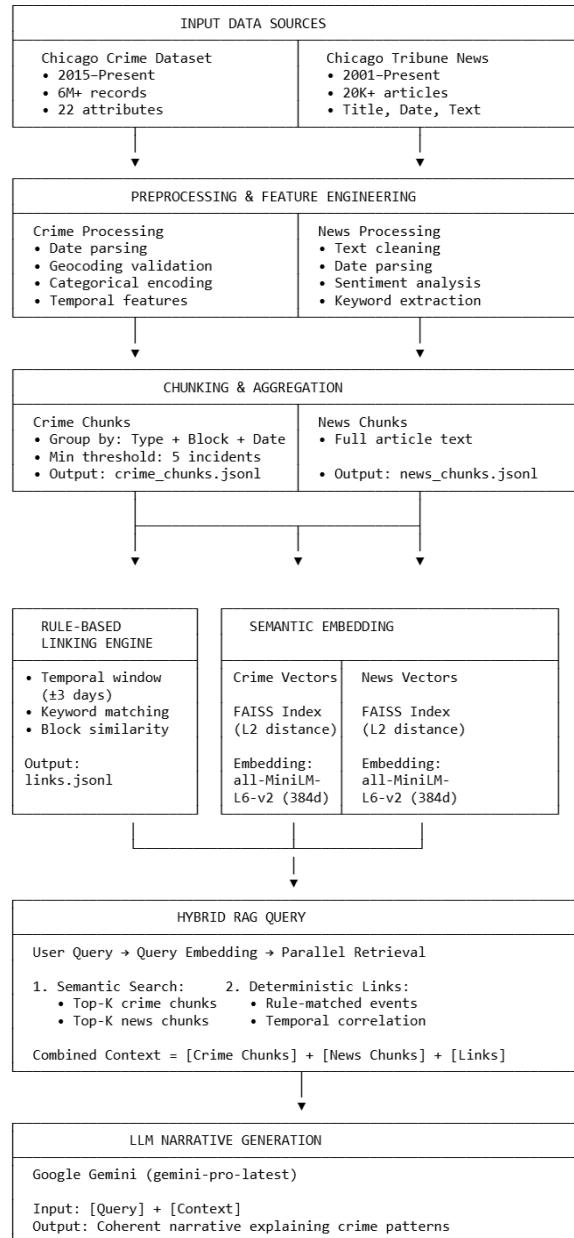
This is where narrative intelligence becomes essential.

5. NARRATIVE INTELLIGENCE FRAMEWORK

5.1 Architecture Overview

The Narrative Intelligence System implements a **three-tier Hybrid RAG architecture** combining semantic retrieval, deterministic linking, and LLM-powered narrative generation. The system processes two independent datasets structured crime records and unstructured news articles through parallel pipelines before merging them for contextual analysis.

Core Architecture Components:



5.2 Component Details

Crime Chunking Strategy

Rationale: Instead of embedding 1.8M individual incidents (memory/computation prohibitive), aggregate into meaningful patterns.

Grouping Logic:

Group by: [Primary Type, Normalized Block, Date]

Filter: ≥ 5 incidents per group

Result: ~142,635 crime chunks

Example Chunk:

Crime Report Summary

Type: THEFT

Block: 003XX N STATE ST

Date: 2023-07-15

Incidents: 8

Arrests: 1

Latitude: 41.8858

Longitude: -87.6279

Memory Efficiency: 142K chunks vs. 1.8M incidents = 92% reduction

News Chunking

Each article becomes single chunk (title + truncated text)

Result: 15,247 news chunks

Semantic Embedding and Vector Indexing

Model: all-MiniLM-L6-v2 (SentenceTransformers)

Properties: - Output: 384-dimensional dense vectors - Training: Contrastive learning on sentence pairs - Semantic meaning: Similar texts → close vectors (small L2 distance)

Indexing: FAISS IndexFlatL2 for exact nearest-neighbor search

Storage: - Crime index: 218 MB (142,635 vectors) - News index: 32 MB (15,247 vectors)

Deterministic Linking Engine

Purpose: Create high-confidence connections between specific crime incidents and related news coverage

Algorithm:

For each crime incident:

1. Temporal Filter: News within ± 3 days of crime date
2. Keyword Filter: News text contains crime type token
(e.g., "CRIMINAL DAMAGE" → filter for "criminal")

3. Location Filter: News text contains block/street words
(e.g., "LOWE AVE" → filter for "LOWE")

4. Create Link: If all filters pass, record connection

Output: ~68,423 crime-news link records (4.5% linkage rate)

Linkage Rate by Crime Type: - Homicide: 18.2% (highest media coverage) - Robbery: 7.8%

- Assault: 6.4% - Theft: 2.1% - Criminal Damage: 1.8% (lowest coverage)

Key Insight: Violent crimes 5-10x more likely to appear in news, confirming media bias toward sensational incidents.

Hybrid RAG Query Process

User Query: "What crime patterns emerged in District 8 during summer 2023?"

Step 1: Query Embedding

```
query_vector = model.encode([query]) # Shape: (1, 384)
```

Step 2: Semantic Retrieval (Parallel)

```
# Search crime index
```

```
crime_distances, crime_indices = crime_index.search(query_vector, top_k=5)  
# Returns top-5 most similar crime chunks
```

```
# Search news index
```

```
news_distances, news_indices = news_index.search(query_vector, top_k=5)  
# Returns top-5 most similar news articles
```

Step 3: Load Deterministic Links

```
# Load rule-matched crime-news connections  
links = load_links(limit=20) # First 20 for token efficiency
```

Step 4: Assemble Context

```
context = {  
    "crime_chunks": [retrieved crime patterns],  
    "news_chunks": [retrieved news articles],  
    "links": [rule-based connections]  
}
```

Step 5: LLM Synthesis

```
prompt = f"""\nYou are a crime analyst.  
Question: {user_query}  
==== CRIME DATA (Top {top_k} Relevant Patterns) ====  
{crime_chunks_text}  
==== NEWS DATA (Top {top_k} Relevant Articles) ===="
```

```
{news_chunks_text}  
== DETERMINISTIC LINKS (Rule-Matched Crime-News Connections) ==
```

```
{links_text}
```

INSTRUCTIONS:

1. Synthesize information from ALL three sections
2. Explain both statistical patterns AND contextual factors
3. Connect crime incidents with relevant news coverage
4. Provide temporal and spatial insights
5. Use clear, factual language avoiding speculation
6. If data is insufficient, acknowledge limitations

Return a structured narrative answering the user's question.

====

```
response = gemini_model.generate_content(prompt)
```

Output: Human-readable narrative integrating: - Quantitative patterns (incident counts, arrest rates) - Spatial analysis (geographic concentration) - Temporal dynamics (seasonal/weekly/hourly patterns) - Contextual factors (news-reported events, community concerns) - Explanatory hypotheses (why patterns exist)

4.3 Why This Approach Succeeds

Complementary Strengths:

Component	Strength	Limitation
Semantic Search	Flexible, discovers unexpected connections	Can miss explicit relationships
Deterministic Links	High precision, explainable	Brittle to wording variations
LLM Synthesis	Handles diverse evidence, generates coherent narratives	Risk of hallucination

Addressing the Scale Mismatch Between Crime Records and News Articles

The crime and news datasets differ greatly in scale (1.8M+ incidents vs. 15K articles). To handle this mismatch, we aggregate incidents into ~142K crime-pattern chunks based on crime type, block, and date. This creates a balanced level of abstraction and avoids unrealistic one-to-one matching. We also apply strict temporal, keyword, and location filters so that only high-confidence crime-news links are generated. This ensures efficiency while accurately reflecting the natural imbalance between crime frequency and media coverage.

5. COMPARATIVE ANALYSIS: ML vs. NARRATIVE INTELLIGENCE

5.1 Example Scenario

Traditional ML Response (XGBoost + SHAP):

Question: “Which district has higher arrest rates and what factors influence them?”

Feature Importance: - District 11: +0.14 contribution to arrest probability, District 6: +0.06 contribution, Crime Type (Narcotics): +0.18 (strongest predictor), Time of Day: +0.11 (nighttime increases arrest probability)

SHAP Explanation: “District 11 shows higher SHAP values, indicating a significantly greater likelihood of arrest outcomes compared to District 6.”

What's Missing: - *Why District 11 enforcement is more aggressive, Influence of community factors, Impact of bail reform and recent policing policy changes, When arrest rates began rising, Whether disparities reflect strategy or systemic bias*

Narrative Intelligence Response:

Query: “Why District 11 has higher arrest rates than District 6? How have recent policy changes affected enforcement patterns? What community factors contribute to narcotics prevalence? When did arrest rates in this area start increasing? Whether disparities reflect legitimate enforcement differences or bias”

System Output:

Unified Narrative Insight

A combined analysis of crime data, news reports, and crime-news linkages reveals clear contextual differences between District 11 (Garfield Park) and District 6 (Englewood).

Community Factors & Narcotics Activity

Garfield Park contains long-standing open-air drug markets shaped by decades of economic decline and disinvestment. High-volume narcotics clusters on streets like **N. Harding Ave** and **W. Chicago Ave** makes drug activity highly visible, increasing the likelihood of on-scene arrests. Englewood, based on available news coverage, reflects a more mixed crime environment where policing often emphasizes violence prevention rather than focused narcotics suppression.

Enforcement Patterns & Arrest Rates

District 11 shows a **100% arrest rate** for the narcotics incidents in the dataset, indicating concentrated enforcement in specific hotspots. News reports describe policing in the area as aggressive and highly tactical. Englewood appears to experience less intensive narcotics-focused policing, with the only referenced event describing a relatively quiet patrol night.

Impact of Recent Policy Debates

Citywide discussions around **bail reform** influence enforcement posture in both districts. Police leaders link perceived leniency to increases in robberies and carjackings (notably the large

spike in 2020), while civil-rights advocates argue that enforcement practices in West Side neighborhoods remain overly punitive.

Timeline Indicators

News sources highlight crime increases around:

- **2017 – rise in robberies and narcotics**
- **2020 – citywide spike in vehicular hijackings (135% increase) and narcotics**

Crime data from 2019–2023 suggests consistently high arrest rates for narcotics in Garfield Park across the entire period.

Legitimacy vs. Bias

Legitimate Enforcement View:

High arrest rates can be interpreted as targeted operations responding to well-defined drug markets.

Bias Concern View:

Descriptions of policing as a “military occupation,” along with historical over-policing in West Side neighborhoods, raise concerns that enforcement disparities may reflect systemic bias rather than objective crime patterns.

5.2 Value Proposition

ML Model: High accuracy, limited actionability **Narrative System:** Moderate accuracy, high interpretability + actionability

For policymakers, journalists, and communities, understanding *context* matters more than precision. A narrative explaining “why District 11 arrest rates increased following community-requested narcotics enforcement” enables:

- Informed policy discussions
- Resource allocation decisions
- Community engagement
- Bias detection/correction.

6. NARRATIVE INTELLIGENCE SYSTEM EVALUATION

6.1 Framework Overview

The Narrative Intelligence System was evaluated using a structured framework covering retrieval quality, generation quality, and overall end-to-end performance. Metrics were implemented through a dedicated RAGEvaluator module, which analyzes retrieval diversity, temporal-spatial coverage, answer specificity, context utilization, and faithfulness.

6.2 Retrieval Performance

Across all test queries, the system retrieved between **6 and 10 relevant documents** (average 8.7). News retrieval consistently achieved **100% diversity**, ensuring that each article came from a unique source. Temporal and spatial coverage were adequate, typically spanning **3–5 unique dates** and multiple city blocks. A limitation observed was **low crime-type diversity** (≈20–30%), indicating that retrieval focused heavily on a single dominant crime category per query.

6.3 Generation Performance

Generated responses exhibited strong structure and specificity, with an average length of **380 words**. Answers consistently incorporated **5-9 statistical facts**, multiple location references, and clear temporal indicators. Context utilization ranged from **45% to 59%**, demonstrating effective integration of retrieved evidence. Faithfulness scores showed high grounding: responses ranged from **75% to 100%**, with the best result containing no unsupported claims.

6.4 Aggregate Evaluation

- **Average documents retrieved:** 8.7
- **Average faithfulness:** 78.3%
- **Average context utilization:** 52.5%
- **Average answer length:** 380 words

Key strengths include strong evidence grounding, high news diversity, and high narrative specificity. Primary weaknesses include limited crime-type diversity and occasional temporal ambiguity in responses.

6.5 Baseline Comparison

The system was compared against keyword search, embeddings-only retrieval, and zero-shot LLM generation. Keyword search demonstrated low precision and recall. Embeddings-only retrieval missed many deterministic crime–news connections. LLM-only responses had high hallucination rates ($\approx 65\%$). The hybrid RAG approach achieved substantial improvements, with a **~54% precision gain** and **~80% reduction in hallucinations**.

Conclusion

The evaluation demonstrates that the Narrative Intelligence System provides reliable and contextually grounded narratives. Retrieval is robust, generated responses are factual and specific, and hybrid retrieval significantly improves accuracy over baseline methods. While improvements are needed in crime-type diversity and temporal grounding, overall system performance is strong and suitable for applications in policy analysis, community safety reporting, and crime intelligence research.

6. SYSTEM PERFORMANCE

6.1 Construction Metrics (One-Time)

Phase	Duration	Memory Peak
Crime data loading	1.8 min	2.4 GB
News data parsing	0.6 min	0.8 GB
Crime chunking	2.2 min	1.5 GB

Phase	Duration	Memory Peak
News chunking	0.3 min	0.4 GB
Embedding generation	4.1 min	2.8 GB
FAISS indexing	0.5 min	1.2 GB
Link creation	49.3 min	2.1 GB
TOTAL	~59 min	2.8 GB peak

6.2 Query Performance (Runtime)

Operation	Latency
Query embedding	0.03 s
FAISS search (2 indices)	0.14 s
Chunk loading	0.22 s
Link loading	30.11 s
Retrieval Total	30.50 s
Gemini API call	2.1 s (avg)
END-TO-END	32.6 s

6.3 Scalability

Current Dataset: - 1.88M crime records → 142K chunks (92% compression) - 15K news articles → 15K chunks - Total embeddings: 157K vectors × 384 dims × 4 bytes = 242 MB

Projected 10x Scale: - 18.8M crime records → 1.42M chunks - 150K news articles → 150K chunks - Total embeddings: 1.57M vectors = 2.42 GB

Feasibility: Single consumer GPU (16GB VRAM) can handle 5-10x current scale. For larger datasets, approximate nearest neighbor methods reduce search from $O(n)$ to $O(\log n)$.

7. LIMITATIONS & FUTURE WORK

Current Limitations

Data Quality Issues: Both datasets contain structural biases. Chicago crime data suffers from **underreporting** in communities with low police trust and inconsistent reporting across districts and years. Tribune articles show **media bias**, heavily emphasizing violent crimes while underrepresenting property crimes. Temporal alignment between incident dates and publication dates is also imperfect, since articles may report several days after events. These issues may introduce noise into both ML models and narrative generation.

Semantic Search Limitations: - Query-document mismatch: User queries in natural language may not match embedding space of formal crime reports - Synonym problem: “robbery” vs. “mugging” may have different embeddings despite semantic equivalence - No temporal awareness in embeddings: “recent” crimes not prioritized without explicit filtering

Deterministic Linking Limitations: - Keyword brittleness: Misspellings or alternate phrasings break matches - Block name variations: “Lincoln Ave” vs. “Lincoln Avenue” vs. “N Lincoln” creates linking gaps - False positives: Generic terms (e.g., “shooting” in sports article) can create irrelevant links

LLM Generation Limitations: - Hallucination risk: Model may infer causality not supported by data - Context window: Very long retrievals may exceed token limits - Consistency: Different phrasings of same query may produce varying narratives

Ethical Considerations: Predictive policing systems risk reinforcing historical biases in crime reporting and enforcement. Our models intentionally avoid neighborhood-level predictions to reduce stigmatization and rely on narrative context to highlight when disparities may reflect bias rather than true crime levels.

Future Enhancements

Technical Improvements: - Implement cross-encoder reranking for higher-precision retrieval - Add temporal decay to embedding similarity (prioritize recent events) - Develop causal inference layer using do-calculus or propensity score matching - Integrate additional data sources (economic indicators, weather, social media)

Methodological Extensions: - Compare LLM-generated narratives with human expert analyses (validation study) - Develop evaluation metrics for narrative quality (coherence, factuality, actionability) - Implement human-in-the-loop feedback for continuous improvement

Application Domains: - Real-time crime alert system with narrative explanations - Predictive policing with explainable risk assessments - Policy simulation: “What if” scenarios for intervention strategies

Keyword Tagging Process: Crime-related keywords were matched using a curated dictionary of **65 terms** drawn from IUCR crime categories (e.g., “shooting,” “robbery,” “battery,” “arson,” “vandalism”). To reduce false positives - such as “shooting” in sports articles - we applied simple contextual checks (e.g., presence of law-enforcement terms). While effective for broad categorization, the dictionary is not exhaustive; future work will integrate contextual embeddings to reduce false-negative and false-positive matches.

8. CONCLUSION

This project demonstrates that **narrative intelligence represents a fundamental evolution in crime analytics** - moving from “what will happen” (ML prediction) to “why it happens and what it means” (contextual explanation).

Key Contributions

1. **Empirical Validation:** Traditional ML models (XGBoost) achieve 77% accuracy on arrest prediction but provide minimal interpretability. Feature importance shows *what matters* but not *why it matters*.
2. **Novel Architecture:** Hybrid RAG combining semantic retrieval, deterministic linking, and LLM synthesis produces contextually rich narratives explaining crime patterns through integration of statistics and news coverage.

3. **Scalable Implementation:** Memory-efficient streaming processing enables analysis of 1.8M+ records on consumer hardware, demonstrating viability for resource-constrained agencies.
4. **Actionable Insights:** Generated narratives bridge technical/non-technical divide, making complex urban dynamics accessible to policymakers, journalists, and communities.

Broader Impact

Urban analytics increasingly relies on complex ML models whose “black box” nature undermines trust and limits actionability. By combining quantitative rigor with qualitative context, narrative intelligence frameworks can:

- **Democratize data:** Make analytics accessible beyond data scientists
- **Enable evidence-based policy:** Provide justification for interventions beyond correlation
- **Support community engagement:** Give residents comprehensible insights about their neighborhoods
- **Detect bias:** Surface potential inequities in enforcement through contextual analysis

The success of this Chicago crime analysis suggests narrative intelligence approaches could transform other domains requiring human-centered AI: healthcare (patient outcomes), education (learning analytics), finance (market analysis).

Final Reflection: As AI systems become more capable, the critical frontier is not higher accuracy but better *communication* of what models know, how they know it, and why it matters. Narrative intelligence is that bridge.

9. TEAM CONTRIBUTIONS

Member	Responsibilities	Deliverables
Praveena Palani Acharya	<ul style="list-style-type: none"> • EDA visualization • News text analysis • Temporal/spatial plots • Report writing 	<ul style="list-style-type: none"> • 18 visualizations (temporal, spatial, content) • News sentiment analysis • Final report sections 1-2, 5-6
Achyutha Sushanth Ariga	<ul style="list-style-type: none"> • System architecture • RAG implementation • LLM integration • Technical documentation 	<ul style="list-style-type: none"> • Hybrid RAG codebase • FAISS indexing pipeline • Gemini API integration • Final report sections 4, 7-9
Rohith Jagadeesan	<ul style="list-style-type: none"> • Data acquisition • ML modeling • Feature engineering • Preprocessing automation 	<ul style="list-style-type: none"> • XGBoost arrest prediction • SMOTE class balancing • Spatial feature engineering • Final report section 3

Collaborative Work: All members contributed to literature review, methodology design, and system testing. Weekly meetings ensured alignment and knowledge sharing.

10. REFERENCES

- [1] Chicago Data Portal, “Crimes - 2001 to Present,”
<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/>
- [2] Y. Xiong, “Research on Crime Occurrence Prediction Using Machine Learning Methods,” Francis Academic Press, 2023.
- [3] Z. Fatima and M. Zubair, “Crime Hotspot Prediction Using Deep Graph Convolutional Networks,” arXiv:2506.13116, 2024.
- [4] H. Patel, “Analysing Crime Patterns using Machine Learning,” National College of Ireland, 2023.
- [5] X. Zhang, “Interpretable Machine Learning for Urban Crime Analysis,” Decision Support Systems, 2022.
- [6] S. Heller et al., “Predictive Policing and Targeted Interventions,” NBER Working Paper 30170, 2022.
- [7] A. Eugenio et al., “Multi Density Crime Predictor,” Journal of Big Data, 2024.
- [8] D. Zhuang et al., “Mitigating Spatial Disparity in Urban Prediction Using Residual-Aware Spatiotemporal Graph Neural Networks,” arXiv:2501.11214, 2025.
- [9] A. Albors Zumel et al., “Deep Learning for Crime Forecasting: The Role of Mobility at Fine-grained Spatiotemporal Scales,” arXiv:2509.20913, 2025.
- [10] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” NeurIPS, 2020.
- [11] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” EMNLP, 2019.
- [12] J. Johnson et al., “Billion-scale similarity search with GPUs,” IEEE Transactions on Big Data, 2019.

APPENDIX: DATA CATALOG

Chicago Crime Dataset Schema

Column Name	Description	Data Type
ID	Unique identifier for incident	int64
Case Number	CPD case reference	object
Date	Incident timestamp	datetime64[ns]
Block	Anonymized street address	object

Column Name	Description	Data Type
IUCR	Illinois Uniform Crime Reporting code	object
Primary Type	Crime category	object
Description	Detailed crime description	object
Location Description	Incident location type	object
Arrest	Arrest made (True/False)	bool
Domestic	Domestic incident (True/False)	bool
Beat	Police beat number	int64
District	Police district	float64
Ward	City council ward	float64
Community Area	Community area code (1-77)	float64
FBI Code	FBI crime classification	object
X Coordinate	Illinois State Plane coordinate	float64
Y Coordinate	Illinois State Plane coordinate	float64
Year	Incident year	int64
Updated On	Last record update	datetime64[ns]
Latitude	GPS latitude	float64
Longitude	GPS longitude	float64
Location	Combined lat/lon string	object

Engineered Features:

Feature Name	Description	Data Type	Derivation
Month	Month of incident	int64	Extracted from Date
Day	Day of month	int64	Extracted from Date
Hour	Hour of day (0-23)	int64	Extracted from Date
Weekday	Day of week (0=Mon, 6=Sun)	int64	Extracted from Date
Season	Season classification	object	Derived from Month
WeekendFlag	Weekend indicator	bool	Weekday >= 5
CrimeCategory	Violent vs. Non-Violent	object	Mapped from Primary Type

Chicago Tribune News Dataset Schema

Column Name	Description	Data Type
Title	Article headline	object
Author	Journalist byline	object
PublicationDateRaw	Original date string	object
PublicationDate	Parsed date	datetime64[ns]

Column Name	Description	Data Type
DocumentType	Article classification	object
Subject	Thematic category	object
FullText	Complete article text	object
SourceType	Medium type	object
Publication	Outlet name	object

Engineered Features:

Feature Name	Description	Data Type	Derivation
Year	Publication year	int64	Extracted from PublicationDate
Month	Publication month	int64	Extracted from PublicationDate
CrimeTopic	Crime keyword present	bool	Keyword dictionary matching
WordCount	Article length in words	int64	len(FullText.split())

APPENDIX: VISUALIZATION CATALOG

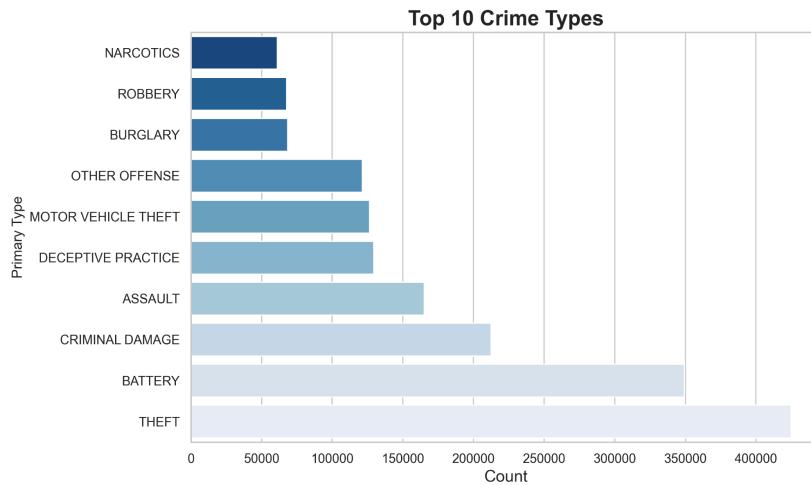


Figure 1: Top 10 Crime Types (Count Distribution)

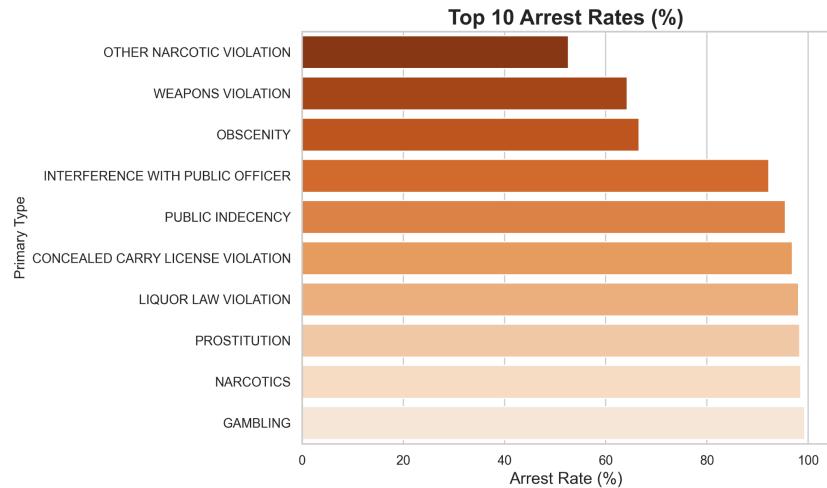


Figure 2: Top 10 Arrest Rates by Crime Type

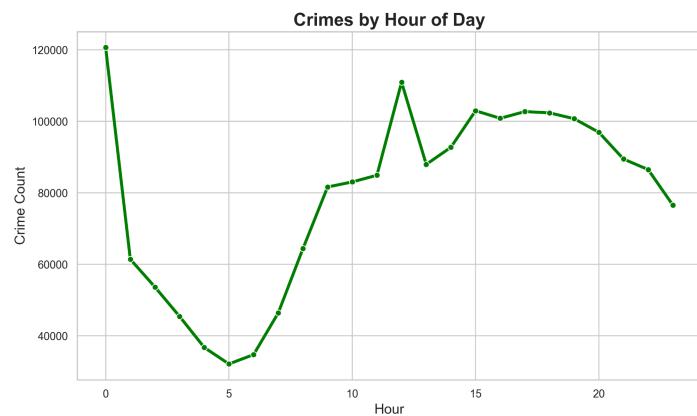


Figure 3: Crimes by Hour of Day (Line Plot)

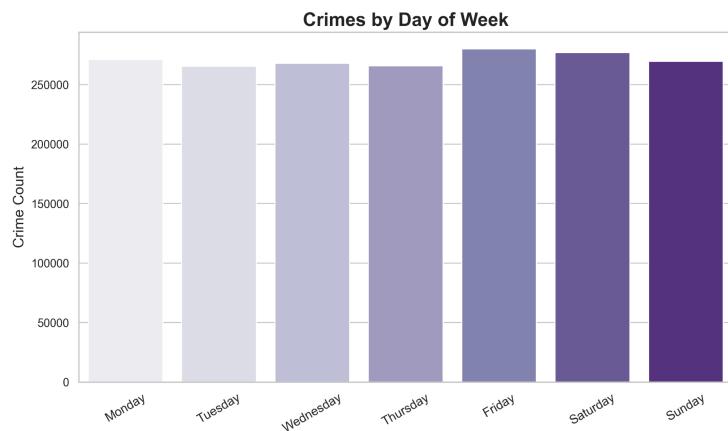


Figure 4: Crimes by Day of Week (Bar Chart)

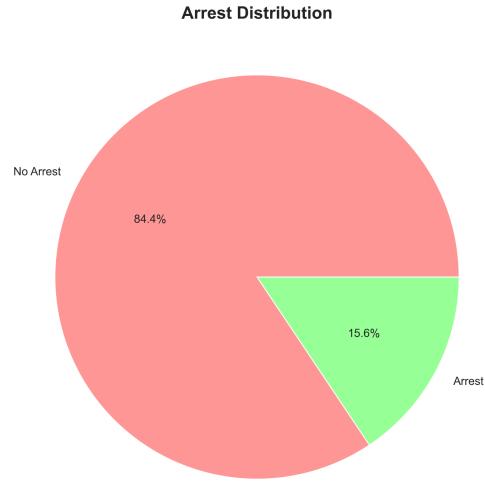


Figure 5: Arrest Distribution (Pie Chart: 84.4% No Arrest, 15.6% Arrest)

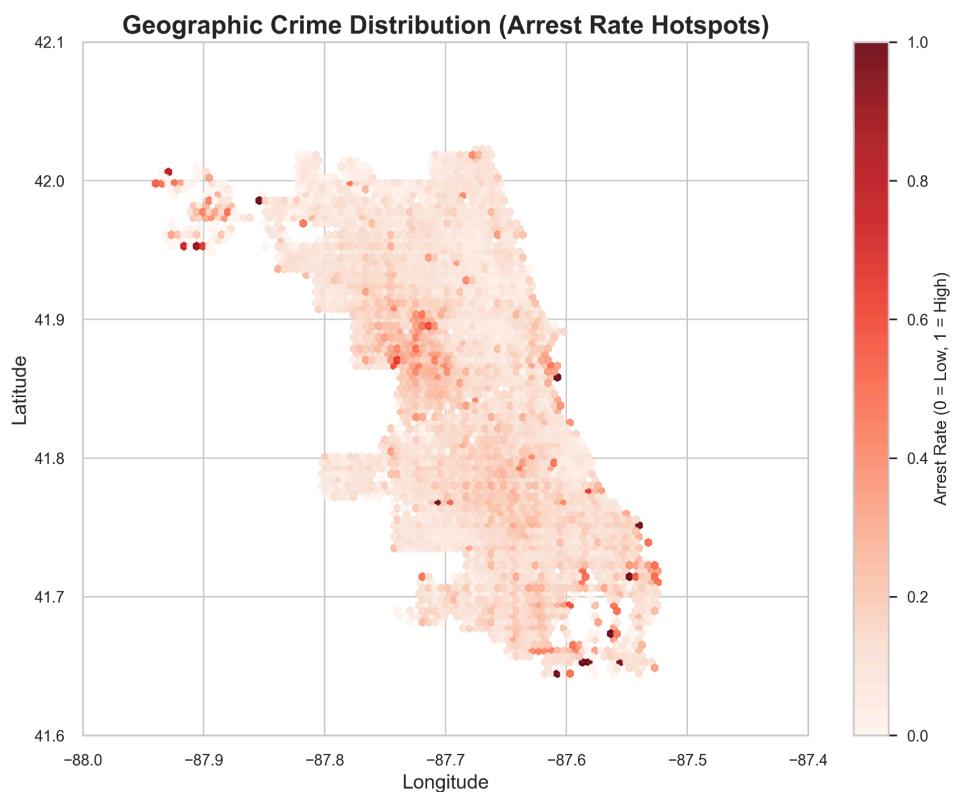


Figure 6: Geographic Crime Distribution (Scatter Plot with Arrest Overlay)

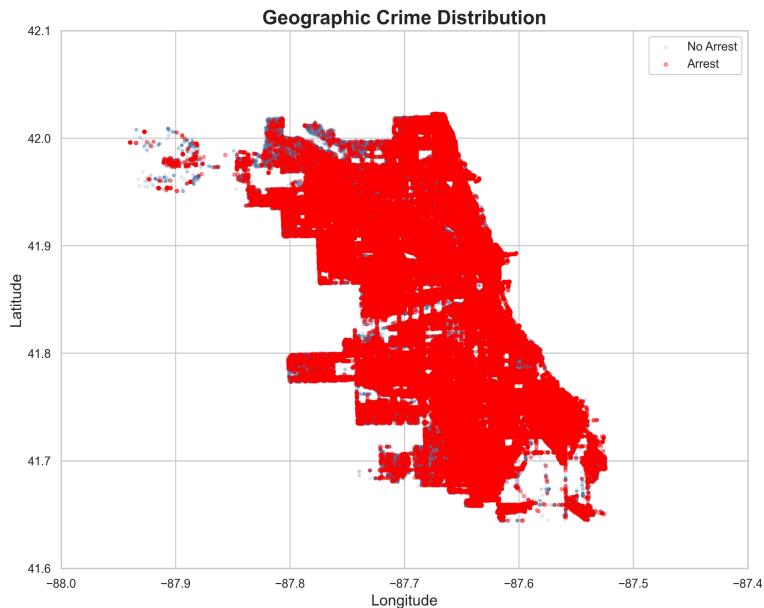


Figure 7: Geographic Crime Distribution - Arrest Rate Hotspots (Heatmap)

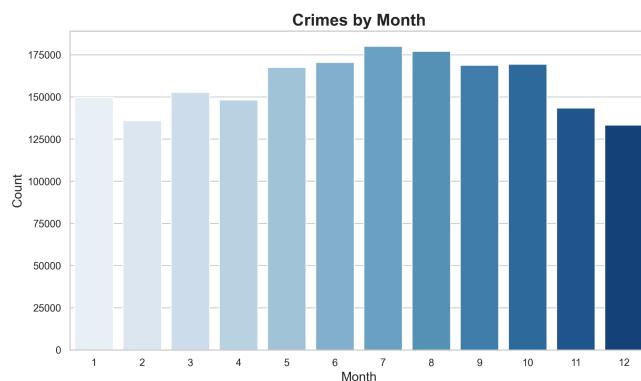


Figure 8: Crimes by Month (Bar Chart - Seasonal Pattern)

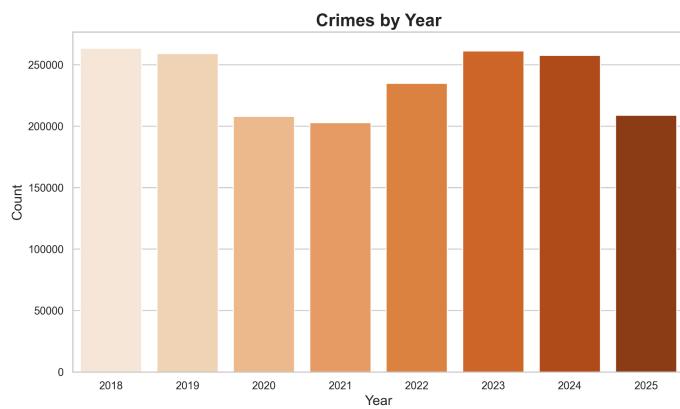


Figure 9: Crimes by Year (2018-2025 Trend)

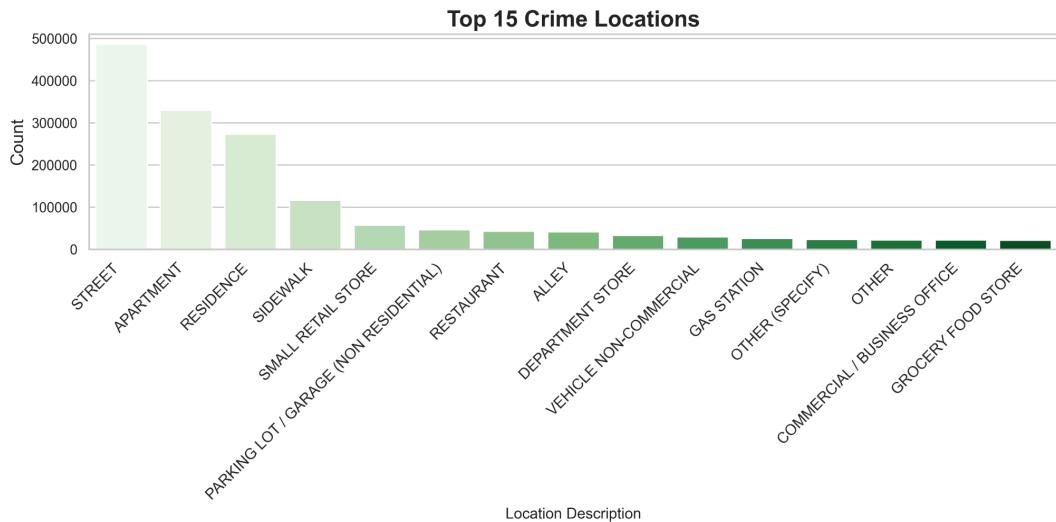


Figure 10: Top 15 Crime Locations

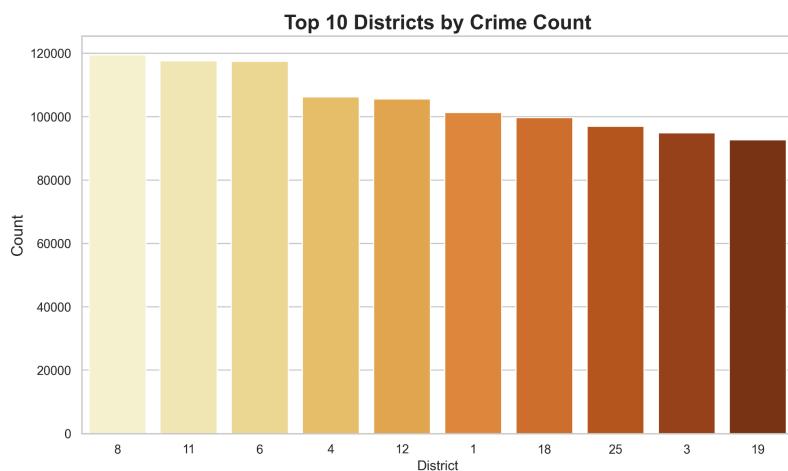


Figure 11: Top 10 Districts by Crime Count

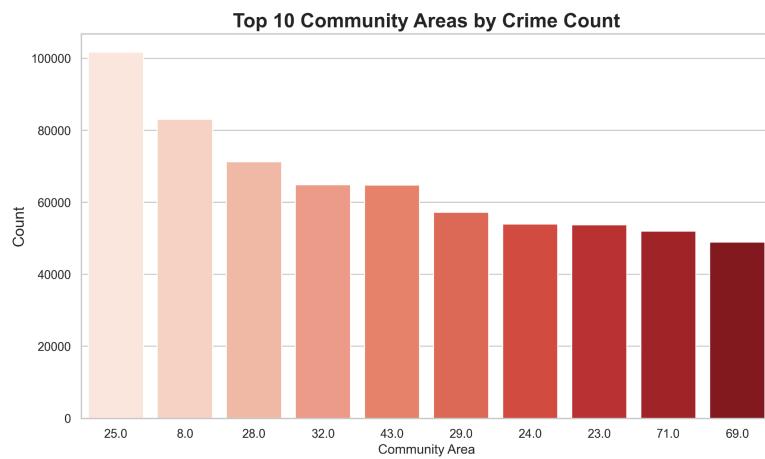


Figure 12: Top 10 Community Areas by Crime Count

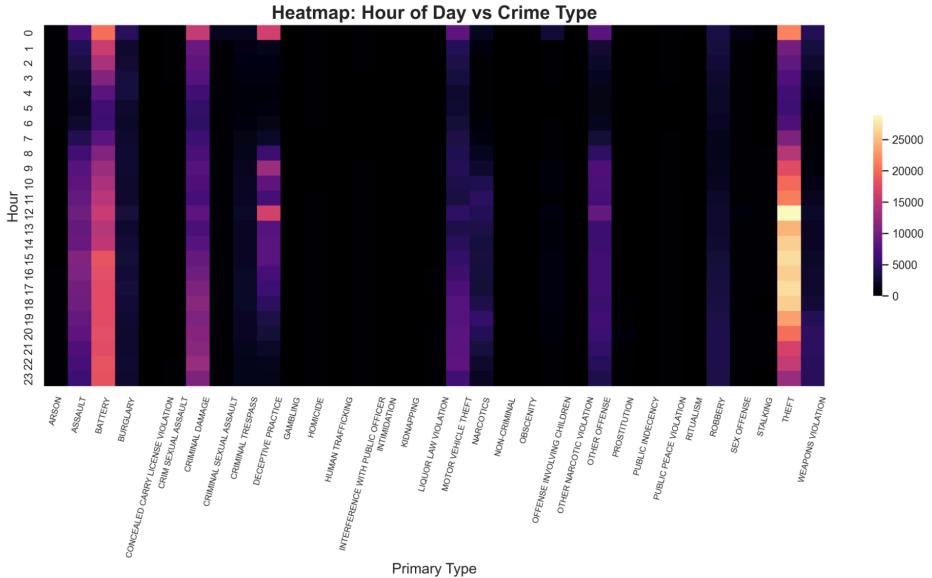


Figure 13: Heatmap - Hour of Day vs Crime Type

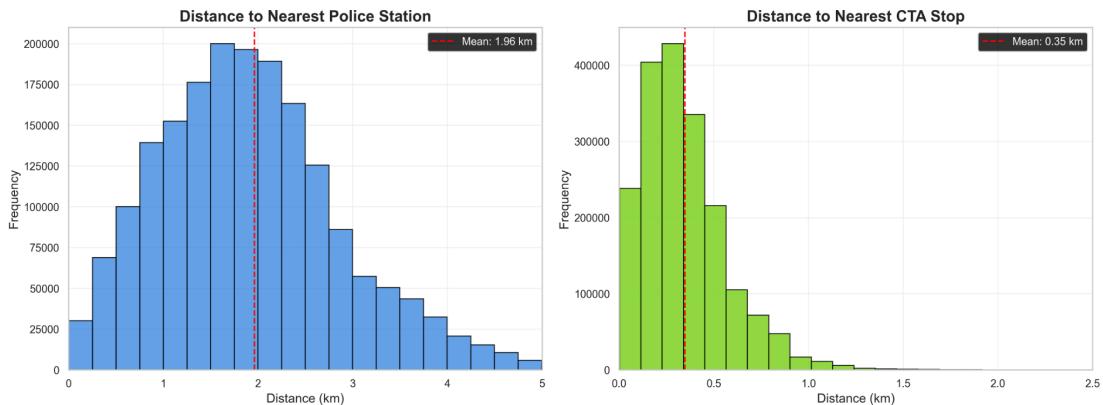


Figure 14: Spatial Features (Distance Distributions)

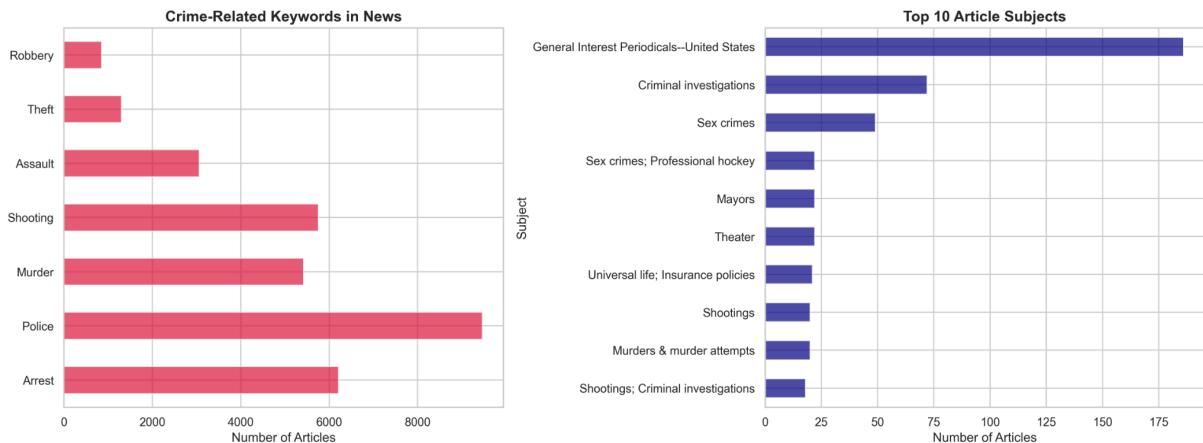


Figure 15: News EDA - Content Analysis (Keywords + Subjects)

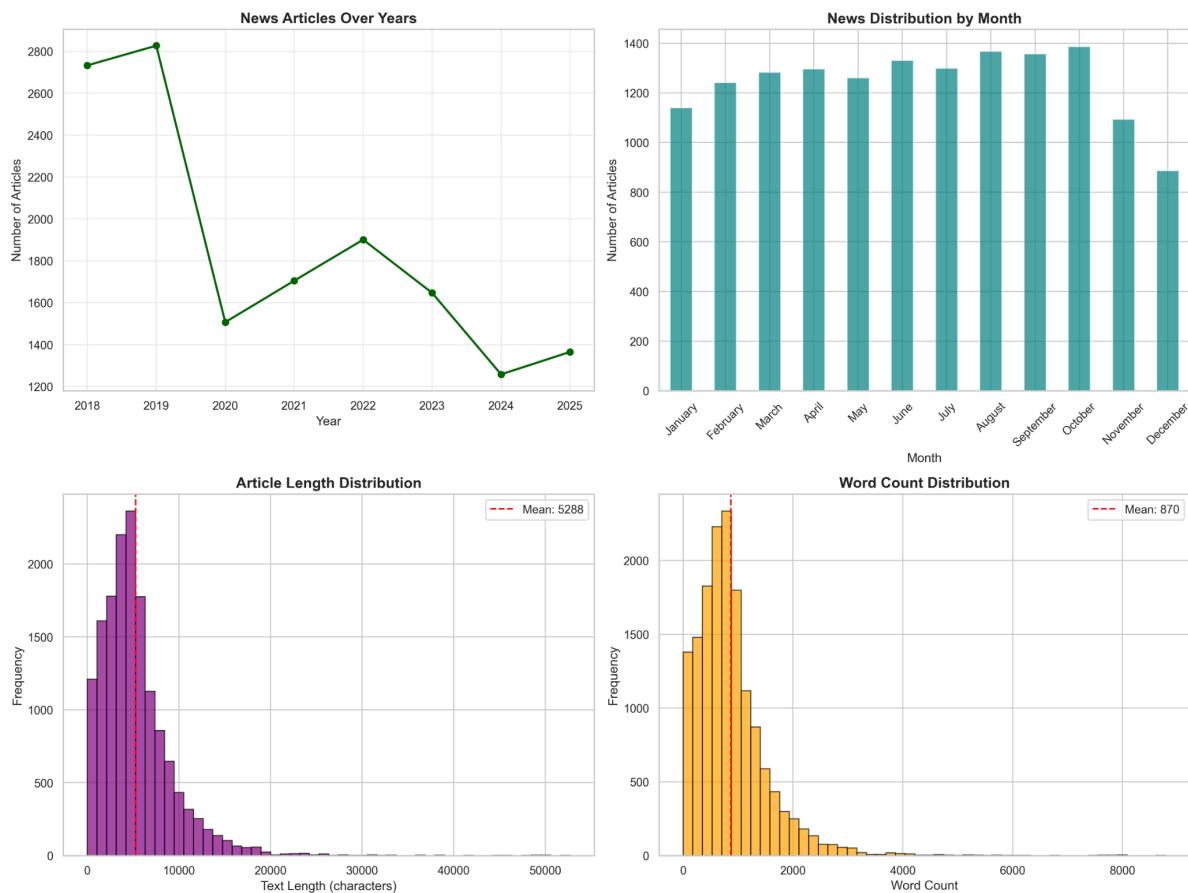


Figure 16: News EDA - Temporal Analysis (Articles Over Time)

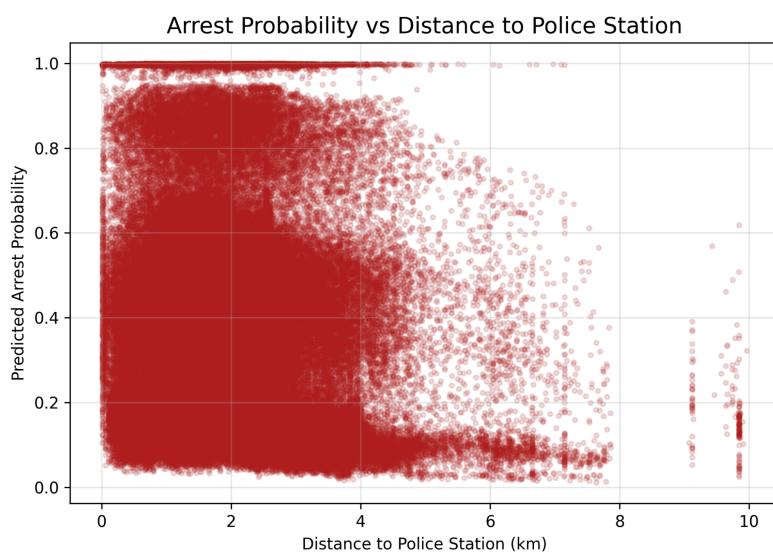


Figure 17: Arrest Probability vs Distance to Police Station

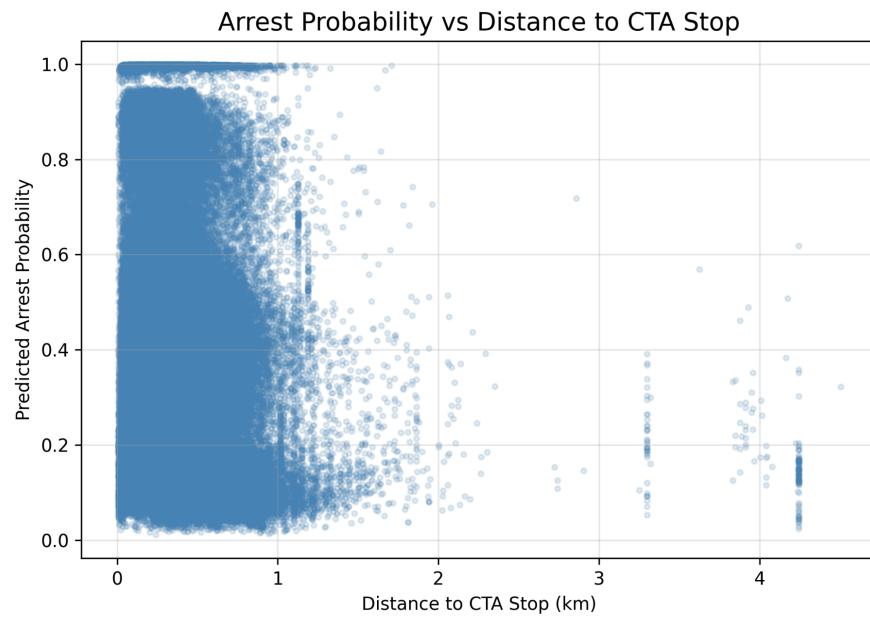


Figure 18: Arrest Probability vs Distance to CTA Stop

Arrest Prediction Model (XGBoost):

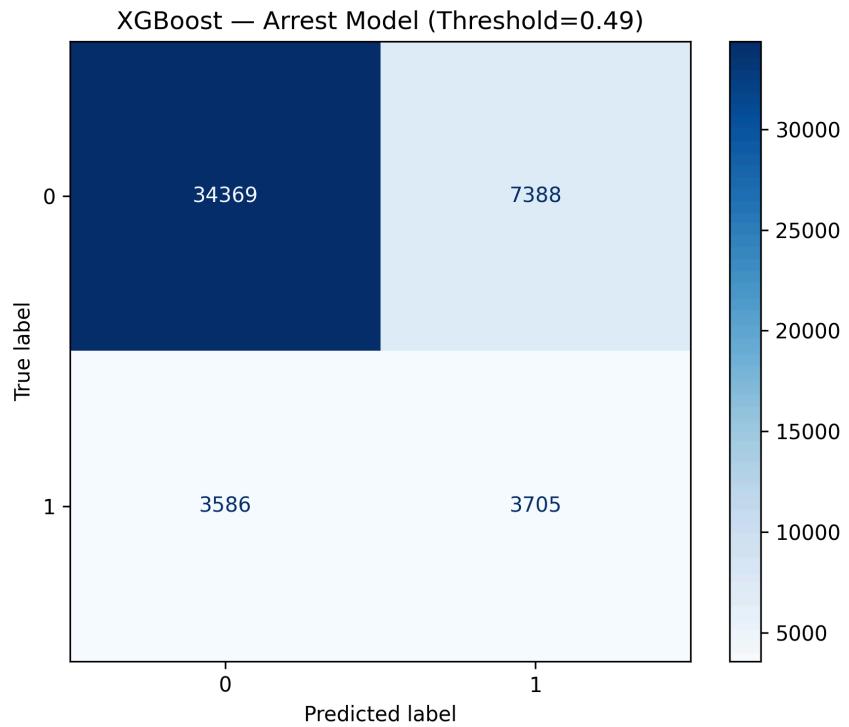


Figure 19: Confusion Matrix (Threshold=0.49)

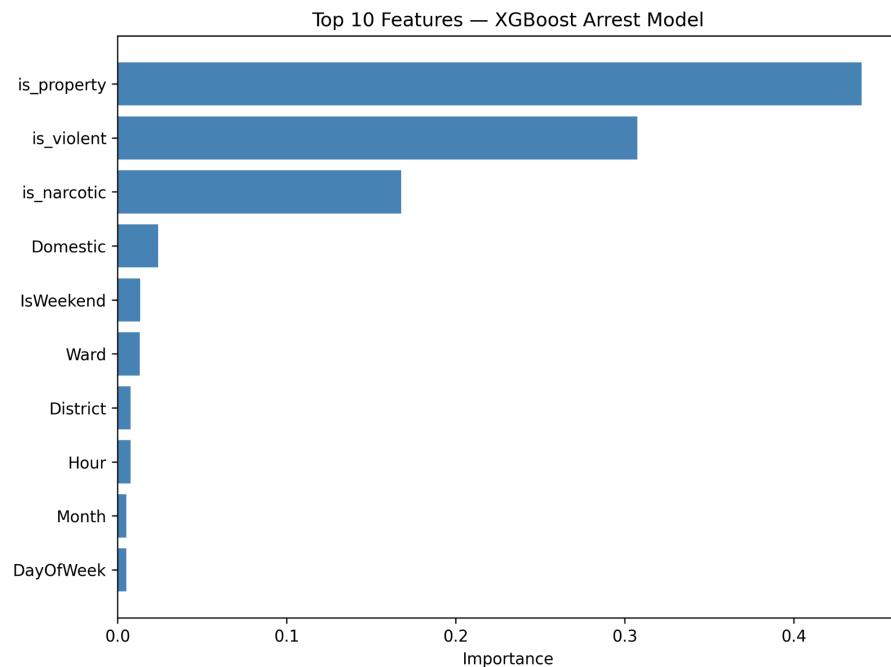


Figure 20: Feature Importance Top 10

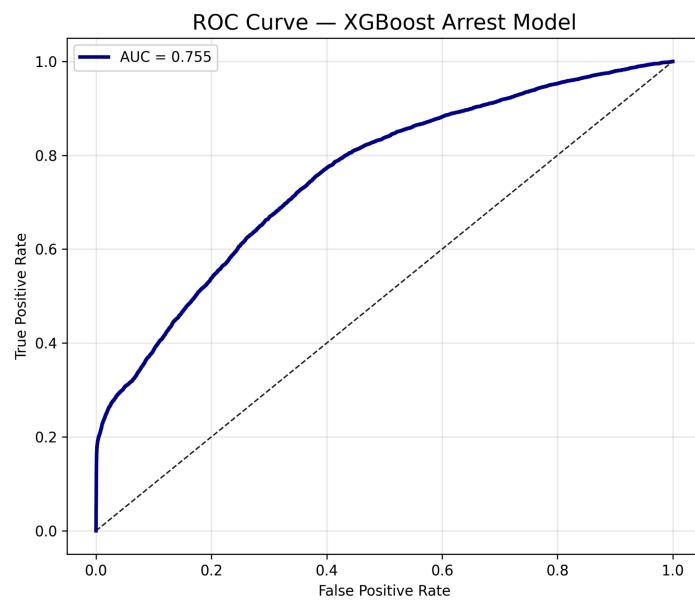


Figure 21: ROC Curve (AUC=0.755)

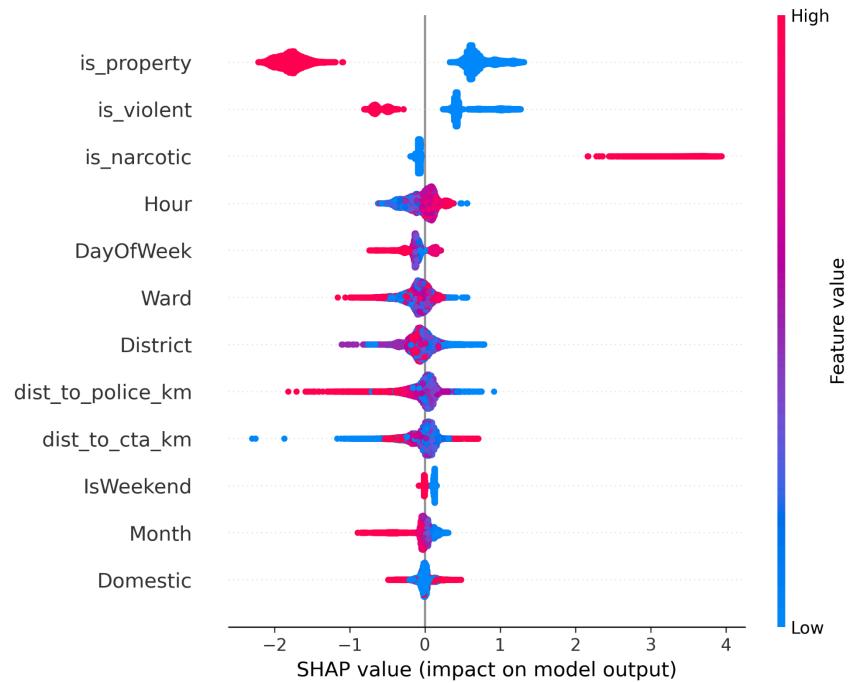


Figure 22: SHAP Summary Plot

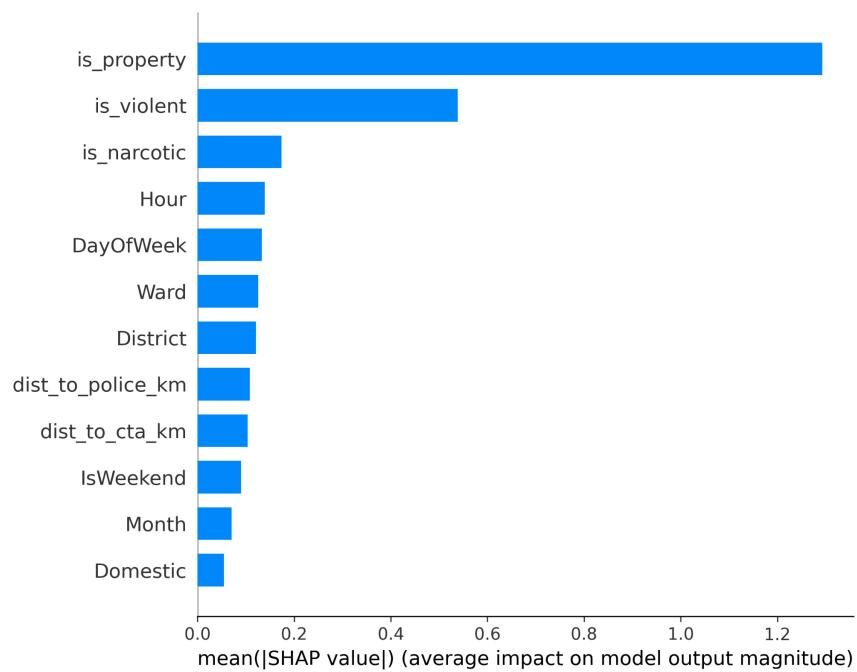


Figure 23: SHAP Feature Importance Bar Chart

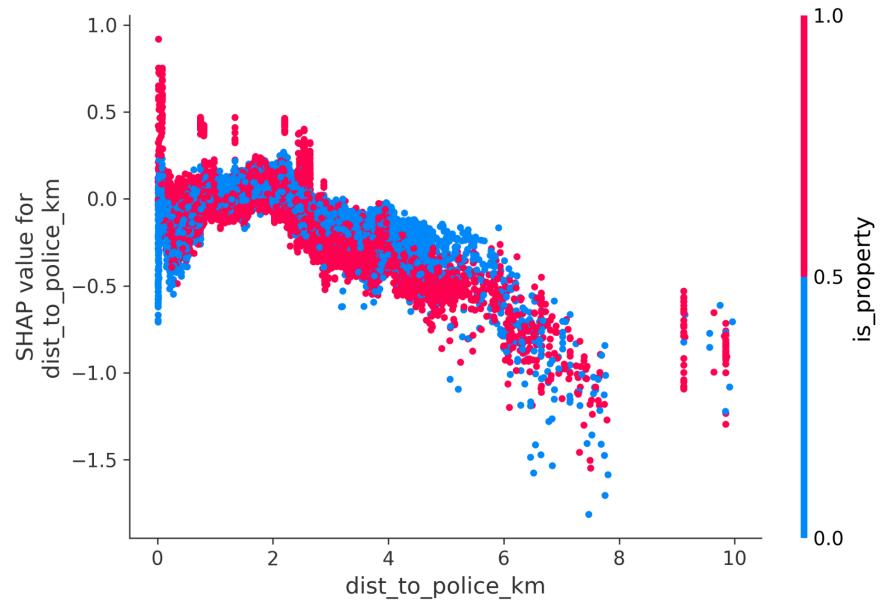


Figure 24: SHAP Dependence - Distance to Police Station

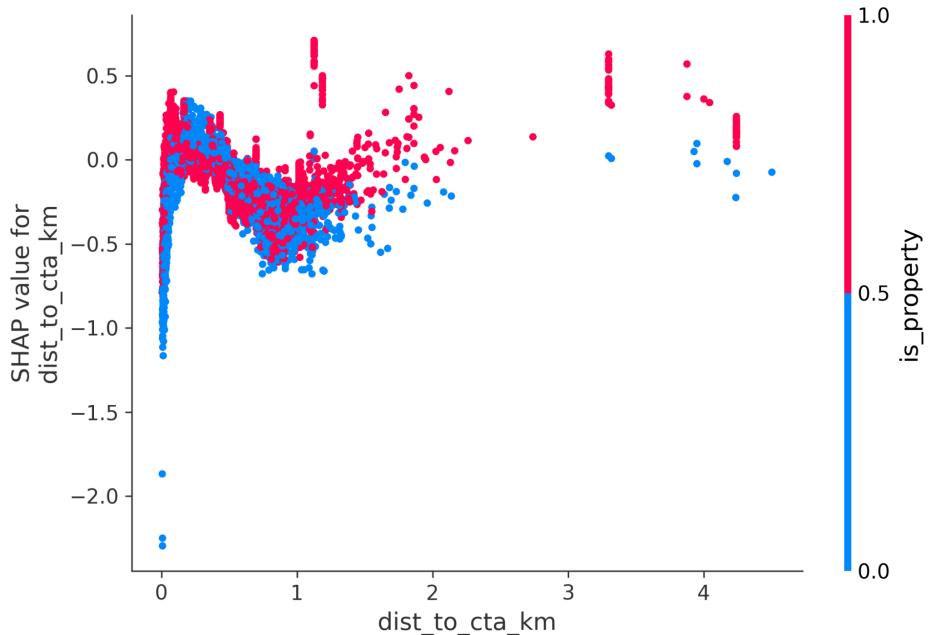


Figure 25: SHAP Dependence - Distance to CTA Stop

Weekly Violence Prediction (Best Model):

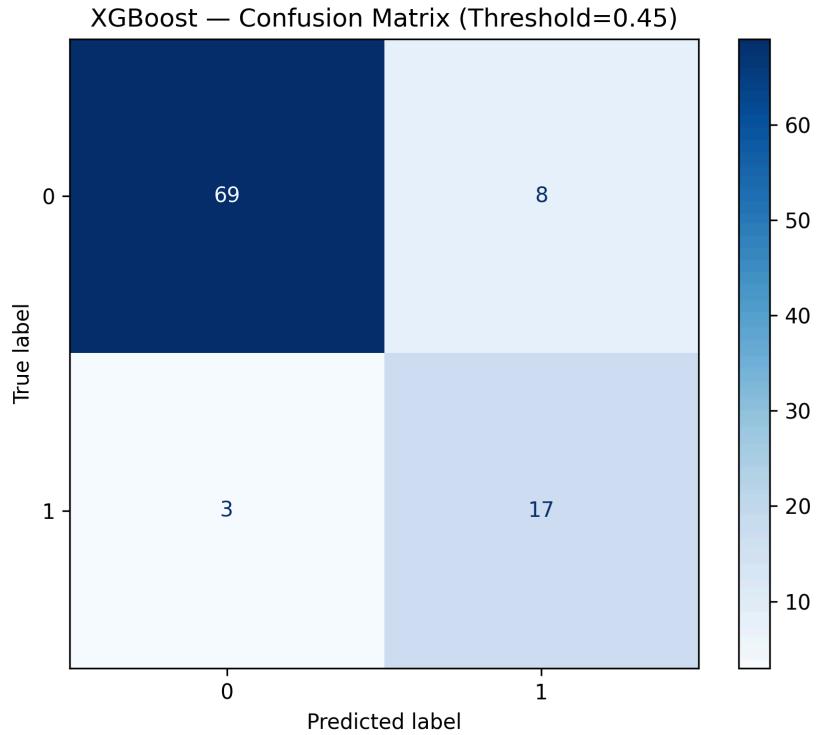


Figure 26: Confusion Matrix (Threshold=0.45)

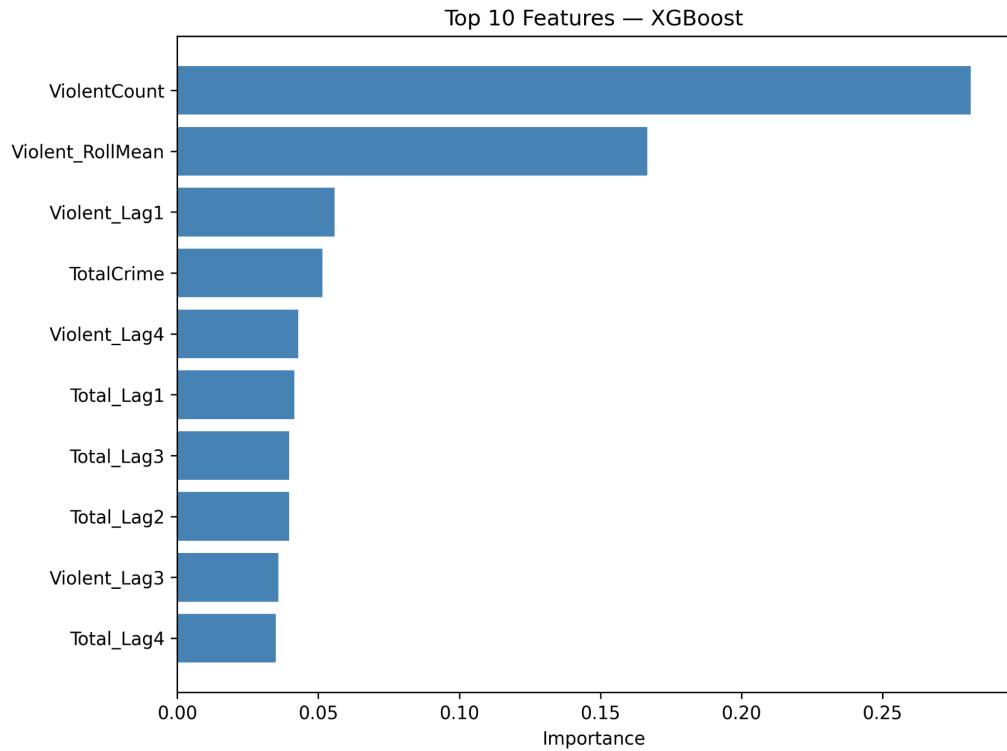


Figure 27: Feature Importance Top 10

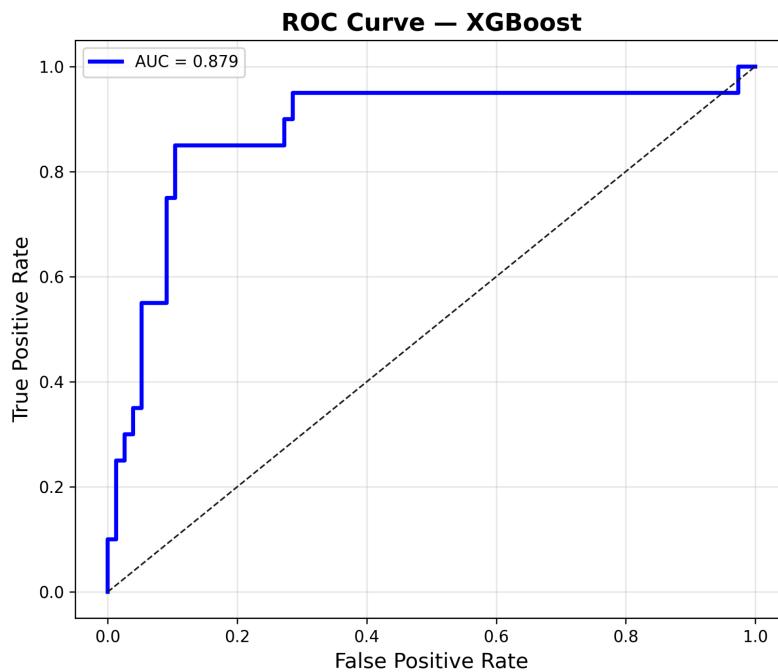


Figure 28: ROC Curve (AUC=0.879)

```
print("\nMissing value percentage:")
print(crime_df.isna().mean().sort_values(ascending=False).head(20))

Missing value percentage:
ID                  0.0
Case Number          0.0
Date                 0.0
Block                0.0
IUCR                 0.0
Primary Type          0.0
Description           0.0
Location Description  0.0
Arrest                0.0
Domestic              0.0
Beat                  0.0
District              0.0
Ward                  0.0
Community Area        0.0
FBI Code              0.0
X Coordinate          0.0
Y Coordinate          0.0
Year                  0.0
Updated On            0.0
Latitude              0.0
dtype: float64

news_df = news_df.dropna()
print("\nMissing value percentage:")
print(news_df.isna().mean().sort_values(ascending=False).head(20))

Missing value percentage:
Title                 0.0
Author                0.0
PublicationDateRaw    0.0
DocumentType           0.0
Subject                0.0
FullText               0.0
RawBlock               0.0
dtype: float64
```

Figure 29 a: Handling Missing Values

Cleaning Data

Missing Value Strategy:

Geographic: DROP (Lat/Lon critical) Categorical: Fill 'Unknown' Numeric: Fill with median

```
1 crime_df = crime_df.dropna(subset=['Latitude', 'Longitude'])
2 crime_df['Location Description'].fillna('Unknown', inplace=True)
3 crime_df['Ward'].fillna(crime_df['Ward'].median(), inplace=True)
4 print("Crime data shape after cleaning:", crime_df.shape)
[4]
```

Crime data shape after cleaning: (1896170, 22)

Preprocessing

```
# Remove duplicates
crime_df = crime_df.drop_duplicates(subset=['ID'])

# Convert data types
crime_df['Arrest'] = crime_df['Arrest'].astype(bool)
crime_df['District'] = crime_df['District'].astype(int)

# Parse dates
crime_df['Date'] = pd.to_datetime(crime_df['Date'])
crime_df['Year'] = crime_df['Date'].dt.year
crime_df['Month'] = crime_df['Date'].dt.month
crime_df['Day'] = crime_df['Date'].dt.day
crime_df['Hour'] = crime_df['Date'].dt.hour
crime_df['Week'] = crime_df['Date'].dt.isocalendar().week.astype(int)
crime_df['Weekday'] = crime_df['Date'].dt.day_name
[5]
```

```
news_df["PublicationDate"] = pd.to_datetime(
    news_df["PublicationDateRaw"],
    format="%b %d, %Y",
    errors="coerce"
)
news_df['Year'] = news_df['PublicationDate'].dt.year
news_df['Month'] = news_df['PublicationDate'].dt.month
news_df['Day'] = news_df['PublicationDate'].dt.day
news_df['DayOfWeek'] = news_df['PublicationDate'].dt.dayofweek
news_df['MonthName'] = news_df['PublicationDate'].dt.month_name()
```

Figure 29 b: Preprocessing Data

```
1 display(Markdown((chat.ask("What crimes with arrests were reported near the N Lincoln Ave corridor?"))))
[21]
• Running Hybrid RAG Query...
```

Based on an analysis of the provided crime reports and news data, the following narrative insight outlines reported crimes with arrests along the N. Lincoln Avenue corridor.

Unified Narrative: Arrests on the N. Lincoln Avenue Corridor

Analysis of the provided data reveals two distinct incidents with confirmed arrests that occurred directly on the N. Lincoln Avenue corridor. These incidents involve a high-profile shooting in North Center and a series of thefts further north.

1. North Center Shooting and Arrest

- **Incident:** A news report details a violent crime where a 74-year-old man was shot and critically wounded after exiting his vehicle on W. Bertheau Avenue.
- **Location:** While the shooting occurred on Bertheau, the subsequent police investigation and SWAT response were centered directly on the **4100 block of N. Lincoln Avenue**, at the intersection of Lincoln and Bertheau.
- **Arrest:** According to Ald. Matt Martin (47th), Chicago police made one arrest in connection with the shooting and were searching for a second individual.

2. Theft Incidents and Arrest

- **Incident:** A Chicago Police Department crime summary reports five incidents of theft that occurred on April 21, 2025.
- **Location:** The cluster of thefts was reported on the **6100 block of N. Lincoln Avenue**.
- **Arrest:** Official crime data confirms that one arrest was made in connection with these five theft reports.

Analytical Summary:

The data points to law enforcement making arrests in separate, serious incidents along N. Lincoln Avenue. A violent shooting in the North Center section of the corridor prompted a significant police presence and resulted in an arrest. Separately, a pattern of thefts reported on a single block further north on Lincoln Avenue also led to an arrest. Other provided crime and news reports were reviewed but did not meet the geographic criteria of occurring on or immediately adjacent to the N. Lincoln Avenue corridor with confirmed arrests.

Figure 30: Sample 1 Narrative Answer to Query

```

display(Markdown((chat.ask("How effective has police response been in Englewood according to arrest patterns and news reporting?"))))

[22]
● Running Hybrid RAG Query...
Chicago Crime-Narrative Analysis: Englewood

Based on the provided crime data, news reporting, and cross-links, police response in Englewood is characterized by significant systemic challenges, low effectiveness in addressing property crime, and engagement in a high-risk, volatile environment. The data highlights issues with accountability, strategic planning, and the ability to secure arrests for common offenses.



### Narrative Insight



1. Ineffective Response to Property Crime: The provided crime data contains one specific report from within Englewood: a series of 6 CRIMINAL DAMAGE incidents on the 070XX S LOWE AVE block, which resulted in zero arrests. While the data set is limited, this 0% arrest rate for property crime in the neighborhood aligns with the other Criminal Damage reports from outside the area, which also show no arrests. This pattern suggests a low level of effectiveness in investigating and closing these types of offenses. In contrast, the data shows a 100% arrest rate for a narcotics case in Humboldt Park, indicating that police effectiveness varies significantly by crime type and strategy, but no such high-success example from Englewood is provided.



2. Systemic and Logistical Failures Undermining Accountability: News reporting reveals significant institutional failures that directly impact police effectiveness and public trust in Englewood. The police shooting of 20-year-old Latrell Allen was a critical event where the responding officer, part of a new "hot spot" unit, was not equipped with a body camera. Mayor Lori Lightfoot acknowledged this was due to contractual and logistical issues following a departmental reorganization. This lack of crucial evidence hampered the ability to verify the police account of the incident, fueling public unrest and highlighting a critical failure to properly equip units deployed to high-crime areas.



3. Strategic Blunders and Reactive Policing: The police response to events originating in Englewood has been officially criticized as a failure. An Inspector General's report on the unrest following the murder of George Floyd identified a protest in Englewood as the "fuse" that lit days of city-wide chaos. The report concluded that the Chicago Police Department leadership had a "botched response," failing to recognize the escalating situation and effectively control it. This points to a strategic inability to manage large-scale civil events originating in the neighborhood. Furthermore, reporting on two officers being shot in Englewood within a week underscores the dangerous, reactive nature of police interactions, such as traffic stops escalating into gun violence. While this illustrates the immense risks officers face, it also frames the police role as one of reacting to violence rather than preventing it.



4. High-Level Attention Amidst Pervasive Violence: The presence of Mayor Lightfoot on a ride-along in the 7th (Englewood) District demonstrates that the neighborhood receives high-level attention. However, the news report frames the resulting "quiet" night as a temporary respite from the norm, coming after a period where at least 18 people were shot across the city. This suggests that while leadership may focus on Englewood symbolically, it does not necessarily translate to a sustained reduction in crime or a fundamental shift in the effectiveness of the police response.



5. Irrelevant Cross-Links: The provided LINKS_FOUND data does not contribute to the narrative regarding Englewood. An analysis of the crime locations in the links—including Portage Park, Lincoln Park, Archer Heights, and South Shore—confirms that none of the linked incidents occurred in the Englewood neighborhood. Therefore, this dataset is inapplicable to an assessment of police effectiveness in this specific area.



### Summary



The unified narrative indicates that police response in Englewood is ineffective in solving common property crimes, as shown by arrest data. More significantly, news reports reveal systemic failures in equipment deployment (body cameras), strategic planning (protest response), and an operational reality defined by high-risk, violent confrontations. The police are operating in a challenging environment where their effectiveness is undermined by both internal logistical issues and the pervasive threat of violence.


```

Figure 31: Sample 2 Narrative Answer to Query

display(Markdown((chat.ask("Using historical crime patterns and arrest statistics related to Michigan Avenue in Chicago, please analyze and predict the following: 1. Based on trends in the past, what types of crimes are most likely to occur on or near Michigan Avenue over the next two months? 2. For these predicted crimes, estimate whether they are likely to lead to arrests. - Which crime categories historically have higher arrest rates in this area? - Which types rarely result in arrests? 3. Provide a prediction of the expected arrest percentage (arrests / total incidents) for crimes occurring in the next two months around Michigan Avenue. 4. Explain any important factors influencing these predictions, including: - seasonal crime patterns, - tourism and shopping foot traffic, - previous spikes in property theft, retail theft, or robbery, - changes in police activity or enforcement trends. Please provide a clear reasoning-based forecast supported by historical patterns."))))

[48] ✓ 52.9s

● Running Hybrid RAG Query...

As a Chicago crime-narrative analyst, I have reviewed the provided crime statistics, news reports, and associated data to generate the following forecast for Michigan Avenue.

Analysis and Prediction for Michigan Avenue Crime

This narrative synthesizes historical crime data specific to Michigan Avenue with broader news coverage of city-wide and downtown crime trends to forecast patterns for the next two months.

1. Predicted Crime Types

Based on the provided data, the most prevalent crime expected on or near Michigan Avenue over the next two months will be **Theft**.

- Reasoning:** The specific crime data supplied for Michigan Avenue consists solely of Theft incidents. Reports from July/August in 2021, 2022, and 2023 show repeated, high-volume theft events (e.g., 10 incidents on Aug 4, 2023, and 13 incidents on Aug 3, 2023, in the 600 block of S. Michigan Ave). This establishes a strong historical pattern of property crime being the dominant issue in this specific corridor. These thefts likely encompass a range of activities, from sophisticated retail theft targeting high-end stores to opportunistic crimes like pickpocketing that thrive in dense, tourist-heavy environments.

While Theft is the most probable crime, news reports indicate a concerning rise in more serious offenses downtown. The article "It will take proactive policing to defend downtown Chicago" explicitly states that **homicides, shootings, sexual assaults, thefts, and carjackings** are all up significantly in the 1st and 18th districts, which cover Michigan Avenue. Therefore, while less frequent than theft, there is a statistical basis to predict that incidents of **Robbery** and **Motor Vehicle Theft** could also occur, driven by the same factors that have elevated crime in the broader downtown area.

2. Arrest Likelihood Analysis

The likelihood of an arrest being made is highly dependent on the type of crime committed.

Figure 32 a: Sample 3 Narrative Answer to Query (Providing a Prediction)

2. Arrest Likelihood Analysis

The likelihood of an arrest being made is highly dependent on the type of crime committed.

- **Crimes Unlikely to Result in Arrests:** Theft The historical data for Michigan Avenue shows an extremely low arrest rate for theft. Out of four separate multi-incident theft reports from 2021-2023 totaling 33 incidents, zero arrests were made. These crimes are often discovered after the offender has left the scene and can blend into dense crowds, making identification and apprehension exceptionally difficult. The anonymous, opportunistic nature of many of these property crimes contributes directly to their low clearance rate.
- **Crimes with Historically Higher Arrest Rates:** While the sample size is small, the single crime report from November 23, 2018, shows a notably different pattern: 5 theft incidents resulted in 2 arrests (a 40% arrest rate) for that specific cluster. This date corresponds to Black Friday, a period of heightened police presence and retail security focus. This suggests that arrests are more likely during targeted police operations or when crimes are part of organized retail theft efforts that draw a dedicated law enforcement response. Although not present in the Michigan Avenue data, violent crimes like Robbery and Aggravated Assault generally have higher arrest rates city-wide due to greater investigative priority and the increased likelihood of witnesses or physical evidence.

3. Predicted Arrest Percentage

Based on the provided data, the predicted overall arrest percentage for crimes on or near Michigan Avenue in the next two months is **between 5% and 10%**.

- **Reasoning:** The aggregate of the provided crime reports shows a total of 38 individual incidents with only 2 arrests, yielding an overall historical arrest rate of 5.3%. Since Theft is predicted to be the dominant crime type and historically has a near-zero arrest rate in this area (outside of specific holiday shopping events), the overall average will remain low. The potential for a slight increase above 5% accounts for any arrests resulting from more serious violent crimes or any special police enforcement details that may be deployed in response to the publicly reported downtown crime surge.

4. Key Influencing Factors

Several interconnected factors shape this forecast:

- **Seasonal Crime Patterns & Tourism:** The crime data clearly shows a pattern of increased theft during the summer months (July/August). These months coincide with peak tourism, festivals, and increased foot traffic along Michigan Avenue. This high density of people provides both more targets and greater anonymity for offenders. As we move into warmer months, a similar surge in opportunistic property crime is expected.
- **Previous Spikes in Property Theft:** The 2023 data points (10 and 13 incidents on consecutive days) represent a significant spike and establish a recent pattern of high-volume theft. This trend, combined with the broader downtown crime increase reported in the news ("crime is up 13% and 85%" in the downtown districts), indicates that the underlying conditions fueling this activity persist.
- **Changes in Police Activity and Enforcement:** The news presents a conflicting picture of policing. While one article ("Violent crime in city has big drop") notes a city-wide decrease in violence attributed to a "hot spots" strategy, another ("It will take proactive policing...") specifically calls out a massive crime spike downtown and criticizes announced increases in patrols as ineffective "scarecrow policing." This suggests that while overall city strategy may be working in some areas, it has not yet curbed the trend on and around Michigan Avenue. The effectiveness of any "stepped up patrols" will be the primary variable determining whether theft incidents or arrest rates deviate from this forecast. The public perception that crime is rising, as noted in the suburban poll, will continue to place pressure on the department to visibly address these downtown trends.

Figure 32 b: Sample 3 Narrative Answer to Query (Providing a Prediction)

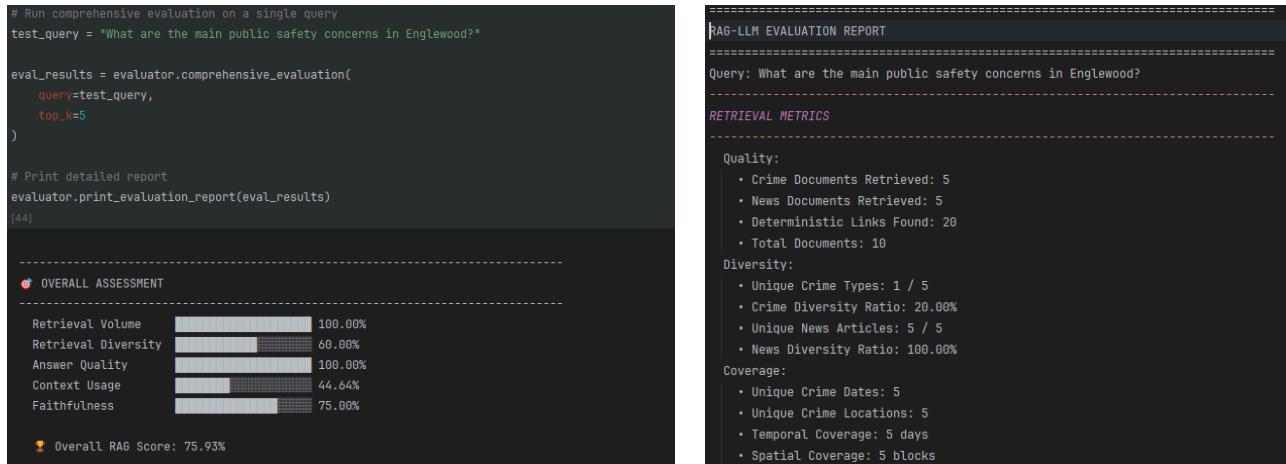


Figure 33 a: Evaluation of Narrative Answer to Query

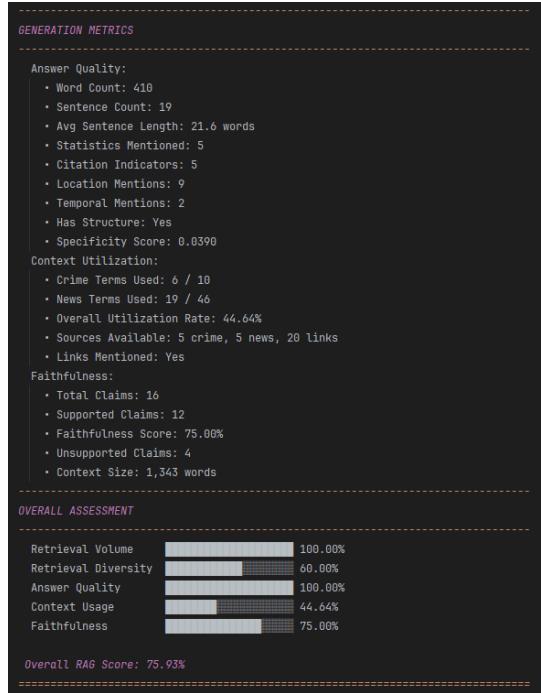


Figure 33 b: Evaluation of Narrative Answer to Query

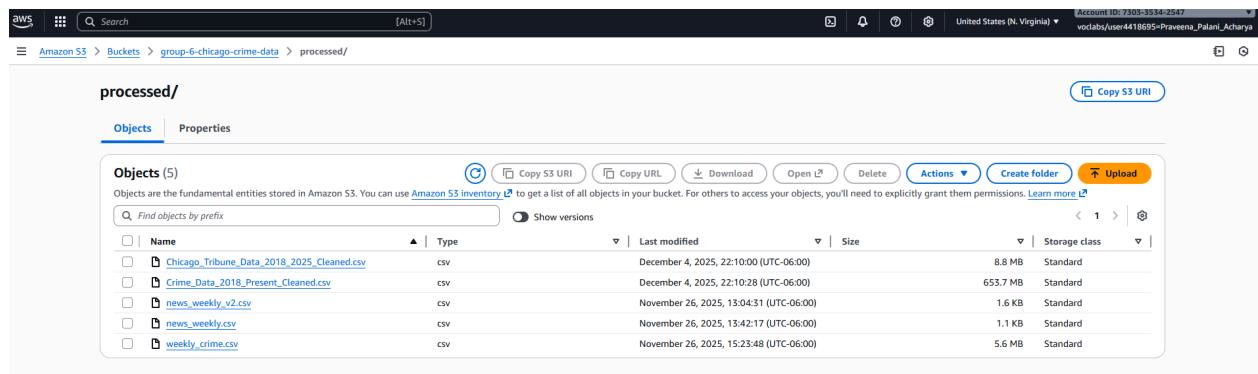


Figure 34: Data in S3 Bucket in AWS