

Project Title and Summary: Syracuse Open Data Initiative – Crime Data Insights

Crime Data Insights in Syracuse: A 2024 Spatial–Temporal Transparency Tool

This project will build a clear, publicly accessible analytical tool that helps Syracuse residents and civic stakeholders understand *when* and *where* reported crime incidents occurred in 2024. Using official Syracuse Open Data crime records, the project will transform raw incident-level data into structured spatial and temporal views maps, trends, and distributions without predictive claims or causal assumptions. The goal is transparency and situational understanding, not enforcement strategy. By presenting validated, responsibly framed insights, the project supports informed community awareness and responsible public discussion around city safety patterns.

Problem Statement:

Residents, journalists, and community organizations in Syracuse frequently encounter crime statistics without sufficient context. Raw counts, headlines, or isolated incidents often dominate public understanding, leading to incomplete or misleading narratives about safety. While the City of Syracuse publishes detailed crime data, the format and complexity of the datasets make them inaccessible to most non-technical users.

The central question this project answers is: **How are reported crime incidents in Syracuse distributed across space and time in 2024?**

This question is being asked implicitly by residents trying to understand their neighborhoods, by journalists contextualizing incidents, and by civic organizations planning outreach or programming. The answer matters because transparent, well-framed descriptive analysis reduces speculation, supports informed civic dialogue, and aligns with the city's goal of making open data usable, not just available. This project explicitly avoids causal explanations or predictive claims, focusing instead on defensible descriptive patterns grounded in official data.

Data Sources:

Primary datasets (Syracuse Open Data Portal):

- **Crime Data 2024 – Part 1 Offenses**
 - ❖ Incident-level records with offense category, date/time, and latitude/longitude
 - ❖ Strengths: geocoded, time-stamped, standardized reporting
 - ❖ Limitations: reporting bias, underreporting, no demographic or outcome data

- **Crime Data 2024 – Part 2 Offenses**

- ❖ Supplementary offense categories published separately by the city
- ❖ Strengths: expands coverage of non–Part 1 incidents
- ❖ Limitations: analytical comparability must be justified; may remain separate

No external datasets will be merged in the initial scope to avoid introducing unverifiable assumptions or demographic inference risks.

Technical Approach:

The analysis will follow a reproducible pipeline: raw data ingestion, structured cleaning, exploratory statistical analysis, and presentation-layer visualization. Initial work will focus on data quality assessment missing coordinates, temporal gaps, and spatial coverage followed by descriptive aggregation across time (hour of day, day of week, month) and space (point-level mapping and neighborhood-level aggregation where appropriate).

LLM augmentation will be used selectively and transparently. Large language models will assist in:

- Generating candidate exploratory questions after baseline statistics exist
- Drafting plain-language narrative summaries for non-technical audiences
- Supporting documentation clarity (README, methodology explanations)

All LLM-generated outputs will be validated against ground-truth calculations derived from Python/Pandas analysis, following techniques from prior tasks on LLM validation and bias detection. Any unsupported claims will be rejected and documented. No LLM output will be treated as analytical evidence.

Deliverable Description:

The final deliverable will be an interactive exploratory dashboard accompanied by a written analytical report. The dashboard will allow users to filter crime incidents by offense type and time window while visualizing spatial distribution on a city map and temporal patterns through charts. The report will document methodology, limitations, and validated findings. Together, these artifacts will function as a standalone civic transparency tool suitable for public sharing.

Success Criteria:

- Users can identify spatial concentrations of reported incidents without interpretation bias.
- Temporal patterns (hour, weekday, month) are clearly visible and validated.
- All visualizations are traceable to documented data transformations.
- LLM-assisted narratives contain zero unvalidated numerical claims.
- Documentation enables a third party to reproduce the analysis end to end.
- The project avoids predictive, causal, or stigmatizing claims.

Timeline:

- **Week 1–2:** Dataset review, stakeholder framing, proposal finalization
- **Week 3–4:** Data acquisition, data dictionary creation, quality audit, baseline EDA
- **Week 5–6:** Pipeline implementation, architecture documentation
- **Week 7–8:** Dashboard prototype, initial visualizations
- **Week 9–10:** Feature completion, validation, refinement
- **Week 11–12:** Documentation, usability polish, final testing
- **Week 13:** Presentation preparation and demo

Risks and Mitigations:

- **Data quality issues (missing or inaccurate coordinates)**
Mitigation: exclude invalid records transparently and document impact
- **Over-interpretation by users**
Mitigation: explicit limitations, neutral language, no rankings or predictions
- **LLM narrative bias**
Mitigation: strict validation workflow and rejection logging
- **Scope creep**
Mitigation: single-year focus and descriptive-only mandate