Phase 1 Project Proposal: Syracuse Road Resilience

Project Title and Summary

Project Title: Syracuse Road Resilience: Pavement Health and Emergency Route Vulnerability

Summary: This project will integrate the 2020 Pavement Ratings with the designated Emergency Snow Routes to create a prioritized, map-based analysis of critical infrastructure vulnerability. The objective is to identify and visualize every segment of an emergency route currently rated as "Poor" or "Fair" pavement quality. The final deliverable will be an interactive web dashboard and an executive report designed to guide the City of Syracuse's Department of Public Works (DPW) and Office of Accountability, Performance and Innovation (API) in strategically allocating maintenance funds to maximize public safety and civic resilience, especially during winter emergencies.

Problem Statement

The integrity of critical transportation arteries is essential for public safety and the efficient delivery of emergency services, particularly in a high-snow city like Syracuse. The Emergency Snow Routes, as defined in the City Charter, represent the most critical components of the road network during a major event. However, the current status of the road network is uneven. We lack a clear, data-driven synthesis that answers the simple but crucial question: **Which specific, high-priority emergency road segments are currently struggling with degraded pavement?**

This project directly addresses the needs of City officials who must make challenging resource allocation decisions. By cross-referencing pavement condition with infrastructure criticality, the analysis moves beyond a general assessment of road quality to a targeted, resilience-focused prioritization model. Answering this question will allow the city to preemptively mitigate risks, reduce wear and tear on emergency vehicles, and ensure critical access remains viable for residents and first responders, thereby translating raw data into actionable civic policy.

Data Sources

The project relies on two primary datasets from the Syracuse Open Data portal:

- **Dataset 1: Pavement Ratings (2020)**
  - **Fields Used:** BPID, Miles, Street Name, From Street, To Street, Ward, Rating_2020, Rating Category.
  - **Quality Notes:** Provides a detailed, segment-by-segment assessment using the

NYSDOT 1-10 scale, which offers high resolution for analysis. Includes excellent spatial metadata (Ward, cross streets).
    - **Limitations Identified:** The data is from **2020**, meaning it represents conditions four years prior. This limits the ability to make real-time assertions about current road status. The data is explicitly marked as being for "planning purposes only."
- **Dataset 2: Emergency Snow Routes**
    - **Fields Used:** CompleteSt, NYSStreetI (for potential joining), LeftFromAd, LeftToAddr (for spatial context).
    - **Quality Notes:** Comprehensive list of official emergency routes as defined by the City Charter. The large number of records (3,685) suggests good coverage of the city's network.
    - **Limitations Identified:** Data freshness is questionable, last updated in **2017**. Field names are abbreviated and require initial validation to confirm meaning (e.g., PreModifie, PreDirecti).
- **External Data Sources:** No external data is strictly required, but the project may optionally incorporate a **GeoJSON file of City Ward boundaries** to improve map visualization and allow for more robust, non-string-based spatial joins.

Technical Approach

The analysis will follow a reproducible, three-stage pipeline leveraging Python for data science and LLMs for insight generation.

1. **Data Integration and Transformation (Pandas/Python):** The initial step involves spatially joining the two datasets. We will attempt a deterministic join using string matching (Street Name, From Street, To Street, and Block) as a primary key. Failing this, we will use a geographic information system (GIS) approach to overlay the two road networks, tagging each rated segment with an is_emergency_route Boolean flag. Data will be normalized (e.g., standardizing street suffix abbreviations) to maximize join accuracy. Critical metrics, such as the total mileage of "Poor" pavement on emergency routes, will be calculated and grouped by Ward and Functional Classification.
2. **LLM-Augmented Analysis and Hypothesis Generation:** LLM augmentation will be used selectively in the exploratory phase (Phase 2). After performing standard statistical grouping (pure Pandas/Python), the LLM will be prompted to generate **interpretive narratives** and **causal hypotheses**. For example, we might present the LLM with the finding that "Ward 3 has 40% of the city's Poor-rated emergency route mileage but only 15% of the total Poor-rated mileage" and ask it to speculate on reasons related to city planning, material, or traffic density. This aligns with the pattern-finding and ethical considerations explored in previous tasks.
3. **Validation Strategy:** All quantitative LLM outputs and statistical claims will be rigorously validated. Numerical outputs from LLM prompts will be checked against the ground-truth calculations performed independently using standard Python/Pandas functions (Task 5 techniques). Furthermore, any LLM-generated narrative or hypothesis suggesting potential bias in resource allocation will be explicitly checked against the dataset's

available demographic and geographic features (Task 8 techniques) to prevent the unintentional reinforcement of systemic bias.

Deliverable Description

The core deliverable will be a **single-file, interactive web dashboard** built using a Python framework (e.g., Streamlit or Plotly Dash). This dashboard will feature a full-screen, responsive map of Syracuse using a library like Folium or Leaflet. The map will display all Emergency Snow Routes, color-coded based on their Pavement Rating (1-10 scale). Key interactive elements will include filters by **Ward**, **Rating Category** (Poor/Fair), and **Functional Classification**. The dashboard will also feature dynamic charts summarizing the total miles of vulnerable (Poor/Fair) emergency routes per Ward. An accompanying executive summary report (Markdown/PDF) will be produced, featuring the top 10 most critical road segments prioritized by poor pavement rating and high functional classification.

Success Criteria

- **Integration & Data Quality:** Successfully achieve greater than 90% match rate when joining the Pavement Ratings and Emergency Snow Routes datasets based on street segments.
- **Actionable Insight:** The analysis must successfully identify, quantify (by mileage), and rank the top 10 most vulnerable street segments (Emergency Route + Poor/Fair Rating) for repair prioritization.
- **Civic Utility:** The final dashboard must be deemed easily understandable and functionally useful by a non-technical end-user.
- **LLM Validation:** All LLM-generated claims, narratives, or hypotheses included in the final report must be validated against ground-truth statistical calculations, with documentation of the validation process included in the METHODOLOGY.md.

Timeline

Week	Phase	Activity
Week 1	Discovery & Proposal	Write and finalize project proposal. Conduct initial data surface-level inspection.
Week 2	Discovery & Proposal	Refine data acquisition plan; Begin securing/downloading raw data files.

Week 3	Acquisition & Exploration	Data Acquisition; Create Data Dictionary; Thorough Data Cleaning (handling nulls, string standardization).
Week 4	Acquisition & Exploration	Execute spatial/string merge of datasets; Perform initial EDA (statistics, preliminary visualizations); LLM-assisted hypothesis generation; Finalize **Exploration Report**.
Week 5	Development	Setup project repository; Implement the final, clean data processing pipeline script.
Week 6	Development	**Architecture Review**; Develop the final analytical model and prioritization logic; Unit testing for core calculations.
Week 7	Development	Begin development of the interactive web dashboard (map integration, chart components).
Week 8	Development	**Working Prototype** milestone: Core map visualization and filtering functionality operational.
Week 9	Development	Refine UI/UX; Implement dynamic data summaries; Develop the PDF report generation script.
Week 10	Development	**Feature Complete** milestone: All planned functionality implemented;

		Draft core documentation (README, TECHNICAL).
Week 11	Polish & Documentation	Full testing and QA (bug fixes, performance); Finalize **TECHNICAL.md** and **METHODOLOGY.md**.
Week 12	Polish & Documentation	Refine visual design; Complete presentation materials and the one-page project summary.

Risks and Mitigations

- **Risk: Inaccurate Spatial Join (Data Integration)**
  - **Description:** String-based joins relying on Street Name and From/To Street often fail due to inconsistencies in formatting (e.g., "ST" vs. "Street", directionals).
  - **Mitigation:** The primary plan is to use fuzzy matching techniques (e.g., Levenshtein distance) or normalization functions to clean strings prior to joining. The ultimate fallback is to use the shared geographic identifiers (Ward, Block) or external GeoJSON data to enforce a geographic, rather than string, match.
- **Risk: Data Staleness Leading to Invalid Recommendations**
  - **Description:** The Pavement Ratings are from 2020. City staff may dismiss the findings if significant repairs have occurred since then.
  - **Mitigation:** This limitation will be **prominently noted** in the dashboard and report. The narrative will frame the results as identifying **historic vulnerabilities** and prioritizing areas where **current pavement health is highly likely to be poor if no maintenance has occurred**. The report will include a strong recommendation for the city to conduct an updated survey of these high-priority segments.
- **Risk: LLM Over-Reliance**
  - **Description:** Relying on the LLM to perform calculations or advanced statistical analysis leading to inaccuracies.
  - **Mitigation:** The LLM will **only** be used for natural language tasks (hypothesis generation, narrative drafting, code scaffolding) and **never** for core calculations. All derived statistics will be calculated exclusively using standard Python libraries, with LLM outputs serving only as a cross-check for simple, easily verifiable metrics (as per Task 5 validation techniques).