A Data-Driven View of Housing Quality and Stability in Syracuse Neighborhoods

Summary:

This project aims to analyze Syracuse Open Data related to housing conditions to identify patterns in property quality, complaints, and potential risk factors across neighborhoods. By combining multiple housing-related datasets, the project will provide residents, community organizations, and city officials with a clearer, data-driven understanding of where housing challenges are concentrated and how they vary geographically and over time. The final deliverable will translate complex municipal data into accessible insights that support informed decision-making and community awareness.

Problem Statement

The Problem:

Housing quality and stability are critical issues affecting residents' health, safety, and economic well-being in Syracuse. While the city collects extensive data on housing conditions, inspections, and complaints, this information is often difficult for residents and community stakeholders to interpret or connect across datasets. As a result, it is challenging to understand which neighborhoods face the greatest housing risks and what patterns may indicate systemic issues.

Why It Matters:

This project seeks to answer the question: *Where are housing-related risks and challenges most concentrated in Syracuse, and how do these patterns differ across neighborhoods?* This question matters to residents who want transparency about their neighborhoods, to city officials responsible for allocating inspection and remediation resources, and to community organizations advocating for safe and equitable housing. Making this data understandable and actionable can support better policy decisions and targeted interventions.

Data Sources:

Example:

- **Syracuse Property Assessment Data**
  Use: Property characteristics and valuation
  Notes: Appears comprehensive; may require cleaning for missing values

- **Housing Code Violations**
  Use: Indicators of housing quality and enforcement activity
  Notes: Potential reporting bias; depends on complaint behavior

- **Vacant Property Registry**
  Use: Identify neighborhood-level vacancy patterns
  Notes: May lag real-time conditions

- **Lead Risk / Inspection Data**
  Use: Health-related housing risks
  Notes: Coverage may be limited to inspected properties

- **External Data (Possible – TBD)**
  Examples: Census or ACS neighborhood demographics
  Notes: To be evaluated in Phase 2 for alignment and bias

Technical Approach:

The project will follow a reproducible data analysis pipeline, beginning with raw data acquisition from the Syracuse Open Data portal and progressing through data cleaning, transformation, and exploratory analysis. Summary statistics and visualizations will be used to identify patterns related to housing quality, complaints, and risk indicators across neighborhoods and time.

Large Language Models (LLMs) will be used to assist with exploratory hypothesis generation and narrative interpretation, particularly for identifying potential relationships and summarizing trends. All LLM-generated insights will be validated against ground-truth statistical calculations to ensure accuracy and prevent hallucination. Bias detection techniques learned in prior tasks will be applied to ensure that narratives do not unfairly attribute blame or misrepresent communities.

Deliverable Description:

The final deliverable will be an interactive analytical report or dashboard that visualizes housing conditions and risk indicators across Syracuse neighborhoods. The tool will allow users to explore patterns geographically and temporally, supported by clear explanations and documented limitations. The deliverable will be designed for accessibility to non-technical audiences while maintaining analytical rigor.

Success Criteria:

Example measurable outcomes:

- Clear identification of neighborhood-level housing patterns

- At least 3 validated insights supported by data and visualization

- Transparent documentation of data limitations and bias risks

- A deliverable that can be understood by non-technical users

- Reproducible analysis pipeline with clear documentation

Timeline (Week-by-Week – High Level)

- **Weeks 1–2:** Data discovery, proposal finalization

- **Weeks 3–4:** Data acquisition, cleaning, exploratory analysis

- **Weeks 5–8:** Core analysis and prototype development

- **Weeks 9–10:** Feature completion and validation

- **Weeks 11–12:** Documentation, polish, refinement

- **Week 13+:** Presentation and showcase preparation

Risks and Mitigations (Bulleted List)

Example:

- **Risk:** Incomplete or outdated housing data
  Mitigation: Clearly document limitations and supplement with external data if appropriate

- **Risk:** Reporting bias in complaint-based datasets
  Mitigation: Avoid causal claims and contextualize findings carefully

- **Risk:** Overly broad project scope
  Mitigation: Focus on a small number of well-defined housing questions