Exploration Report

Syracuse Unfit Properties Dataset

---

1. Dataset Overview & Acquisition

Dataset Description

The **Unfit Properties** dataset contains records of residential properties in the City of Syracuse that have been officially declared *unfit for habitation* by city code enforcement authorities. These designations typically indicate serious housing condition issues that may pose health or safety risks.

Data Source

- **Source:** City of Syracuse Open Data Portal
- **Dataset:** Unfit Properties
- **Portal:** https://data.syr.gov

Acquisition Details

- **Acquisition Date:** *(recorded in docs/acquisition_metadata.json)*
- **Acquisition Method:** CSV download from the Open Data portal
- **API Endpoint:** Not finalized / static download used
- **Raw Data Storage:**
    - data_raw/Unfit_Properties_raw.csv
- **Processed Data Storage:**
    - data_processed/Unfit_Properties_clean.csv

Raw data was preserved exactly as downloaded. All transformations were performed on a separate processed copy to ensure reproducibility and auditability.

---

2. Data Dictionary

A data dictionary was created programmatically due to the absence of a fully documented schema in the source portal.

Location:

- docs/data_dictionary.csv

For each column, the dictionary includes:

- Column name
- Data type
- Missing value count and percentage
- Example values
- Placeholder for semantic description

This dictionary serves as a living document and will be expanded during later project phases.

---

3. Data Quality Assessment

3.1 Missing Values

A column-level missingness analysis was performed.

Key Observations:

- Several columns exhibit **high missing percentages**, indicating optional or inconsistently captured fields.
- Geographic attributes (e.g., coordinates or neighborhood fields) show partial completeness.
- Missing values are **not uniformly distributed**, suggesting structural rather than random absence.

Impact:

- High-missing columns are unsuitable for aggregate analysis without imputation or filtering.
- Some analyses (e.g., neighborhood comparisons) must explicitly account for incomplete geographic coverage.

Artifact:

- outputs/missing_values_summary.csv

3.2 Inconsistencies (Formatting & Categorical Values)

Text columns were evaluated for:

- Leading/trailing whitespace
- Case inconsistencies
- Inflated cardinality caused by formatting differences

Key Observations:

- Multiple categorical fields show reduced unique counts after lower-casing, indicating inconsistent label formatting.
- Several fields contain leading/trailing whitespace that would affect grouping operations.

Mitigation Applied:

- Whitespace trimming was applied uniformly in the processed dataset.
- Case normalization will be applied selectively in later phases where grouping is required.

Artifact:

- outputs/text_inconsistencies.csv

3.3 Temporal Coverage

Date-like fields were parsed and analyzed to assess time range and completeness.

Key Observations:

- The dataset spans multiple years, indicating longitudinal coverage.
- Some date fields contain non-parseable values or missing entries.
- The data reflects **event-based records** rather than continuous time series updates.

Implications:

- Temporal trend analysis is feasible at annual or coarse time resolution.
- Finer-grained seasonal analysis may be unreliable due to gaps.

Artifact:

- outputs/temporal_coverage.csv

3.4 Geographic Coverage

Geographic completeness was assessed using available coordinate and location fields.

Key Observations:

- Latitude/longitude data is present for a majority, but not all, records.
- Geographic bounds align with expected Syracuse city extents.
- Some records lack sufficient location detail for neighborhood-level aggregation.

Implications:

- Citywide spatial patterns can be analyzed.
- Neighborhood-level conclusions must note partial coverage and potential bias.

---

4. Processed Dataset Creation

A lightly cleaned dataset was created with:

- Whitespace trimming for all text fields
- No value imputation
- No categorical recoding

This ensures analytical integrity while avoiding premature assumptions.

Processed Dataset:

- data_processed/Unfit_Properties_clean.csv

---

5. Summary Statistics

Numeric Attributes

- Standard descriptive statistics (count, mean, std, min, max) were computed for all numeric fields.
- Numeric fields primarily represent identifiers or spatial coordinates rather than continuous measurements.

Full Dataset Summary

- Categorical dominance is evident, reflecting the administrative nature of the dataset.
- Several fields exhibit high cardinality, requiring careful grouping strategies.

---

6. Exploratory Visualizations

A total of **five visualizations** were produced to understand structure, coverage, and distributions.

Visualization 1: Missingness by Column

Insight:
A small subset of columns accounts for the majority of missing data, suggesting optional metadata fields rather than systemic data loss.

Visualization 2: Categorical Value Distribution

Insight:
One or more categorical fields show strong skew, indicating that certain property statuses or classifications dominate the dataset.

Visualization 3: Records Over Time

Insight:
The number of unfit property records varies by year, suggesting possible changes in enforcement intensity, reporting practices, or housing conditions over time.

Visualization 4: Geographic Scatter Plot

Insight:
Unfit properties are geographically clustered rather than uniformly distributed, indicating spatial concentration of housing challenges.

Visualization 5: Row Completeness Distribution

Insight:
Most records are largely complete, but a non-trivial minority lack multiple fields, reinforcing the need for cautious filtering.

---

7. Key Findings & Hypotheses

Ground-Truth Findings (Validated)

1. The dataset contains a substantial number of records, supporting city-level and neighborhood-level analysis.
2. Missing data is concentrated in specific metadata and geographic fields rather than core identifiers.

3. Spatial clustering suggests localized housing condition issues rather than citywide uniformity.

Ground-truth statistics were saved and used as the sole numeric reference:

- outputs/ground_truth_findings.json

LLM-Assisted Hypotheses (Not Yet Confirmed)

LLM tools were used **only to generate hypotheses**, not conclusions.

Hypothesis 1:
Unfit property designations are concentrated in specific neighborhoods rather than evenly distributed.

Hypothesis 2:
Periods of increased unfit designations align with heightened inspection or enforcement efforts rather than sudden housing deterioration.

Hypothesis 3:
Properties lacking complete geographic data may systematically differ from fully geocoded records, indicating reporting bias.

---

8. LLM Validation & Bias Checks

Validation Strategy (Task 5)

- All numeric claims were validated against Pandas-generated statistics.
- LLM outputs were restricted to hypothesis generation.
- No LLM-generated figures were accepted without independent verification.

Bias Checks (Task 8)

- Prompts were reviewed to avoid framing neighborhoods as “bad” or “problematic.”
- Findings are framed as **recorded designations**, not inherent neighborhood conditions.
- Enforcement and reporting biases are explicitly acknowledged.

---

9. Dataset Limitations

What This Data Cannot Answer

- It does not capture all substandard housing—only properties officially declared unfit.
- It cannot explain *why* properties became unfit.
- It does not include demographic, economic, or ownership context.

Potential Biases

- Complaint-driven and enforcement-driven reporting bias
- Inconsistent geographic completeness
- Temporal variation in inspection practices

Impact on Conclusions

- Results must be interpreted as **administrative records**, not comprehensive measures of housing quality.
- Neighborhood comparisons must be contextualized carefully to avoid stigmatization.

---

10. Conclusion

This exploratory analysis establishes a solid foundation for deeper investigation into housing condition patterns in Syracuse. The dataset demonstrates sufficient coverage and structure to support spatial and temporal analysis, provided its limitations are clearly acknowledged. Future phases will integrate additional datasets and validated analyses to contextualize and extend these findings.