Data Acquisition and Exploration Report

Data Acquisition and Exploration Report 1

1. Data Acquisition and Validation 2

    Dataset Overview 2

    Data Quality Assessment 2

    Documentation of Limitations 2

2. Key Statistical Summaries 3

    Pavement Health (N=3675) 3

    Snow Route Profiles (N=3685) 3

3. Exploratory Data Analysis (Visualizations) 4

    3.1 Distribution of Pavement Ratings (2020) 4

    3.2 Count of Streets by Pavement Category 5

    3.3 Speed Limits on Emergency Snow Routes 6

    3.4 Pavement Rating Distribution by Ward 7

    3.5 Average Pavement Rating by Ward (Ranked) 8

    3.6 Correlation Matrix (Pavement Dataset) 9

    3.7 Street Width vs. Pavement Rating 10

4. Findings & Hypotheses 11

5. LLM-Assisted Analysis & Validation 12

1. Data Acquisition and Validation

Dataset Overview

This report analyzes two primary datasets related to municipal infrastructure and winter maintenance:

1. **Emergency Snow Routes**: Mapping of priority corridors for snow removal.
2. **Pavement Ratings (2020)**: Condition assessment of city streets on a scale of 0-10.

Acquisition Date: January 1, 2026

Source: Municipal Open Data Portal

Storage: Raw CSV files are preserved; all transformations are performed via Pandas to maintain data lineage.

Data Quality Assessment

- **Missing Values:** Analysis confirmed **zero null values** in primary fields (Rating_202, WARD, SPEED).
- **Logical Consistency:** * Found **221 records** with a WIDTH of 0 in the Pavement dataset, which likely represent administrative errors or non-standard pedestrian rights-of-way.
    - Speed limits on Snow Routes are consistent (25-35 MPH), aligning with arterial road expectations.
- **Integration Challenges:** Street naming conventions differ (e.g., "East Adams Street" vs "Adams St"). A common street name intersection identified only 9 exact matches, indicating that a fuzzy-matching or spatial-join approach is required for deep integration.

Documentation of Limitations

- **Temporal Gap:** Pavement data is from 2020. Post-pandemic maintenance and four additional winter cycles mean current conditions are likely lower than reported here.
- **Question Gaps:** This data cannot answer *why* a road is rated poorly (traffic volume vs. age) as Average Annual Daily Traffic (AADT) counts are missing.

2. Key Statistical Summaries

Pavement Health (N=3675)

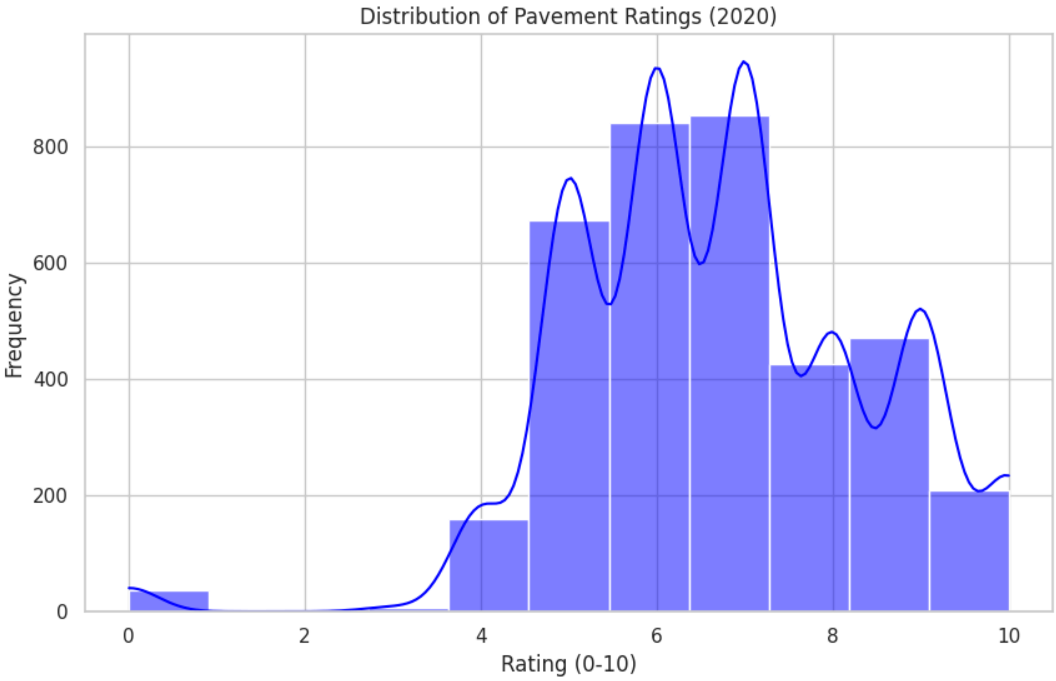
- **Mean Rating:** 6.74 / 10
- **Median Rating:** 7.00
- **Standard Deviation:** 1.72
- **Most Frequent Category:** "Good" (1,279 segments)

Snow Route Profiles (N=3685)

- **Mean Speed Limit:** 31.25 MPH
- **Primary Speed Class:** 35 MPH (Arterial focus)
- **Average Segment Length:** 97.6 units

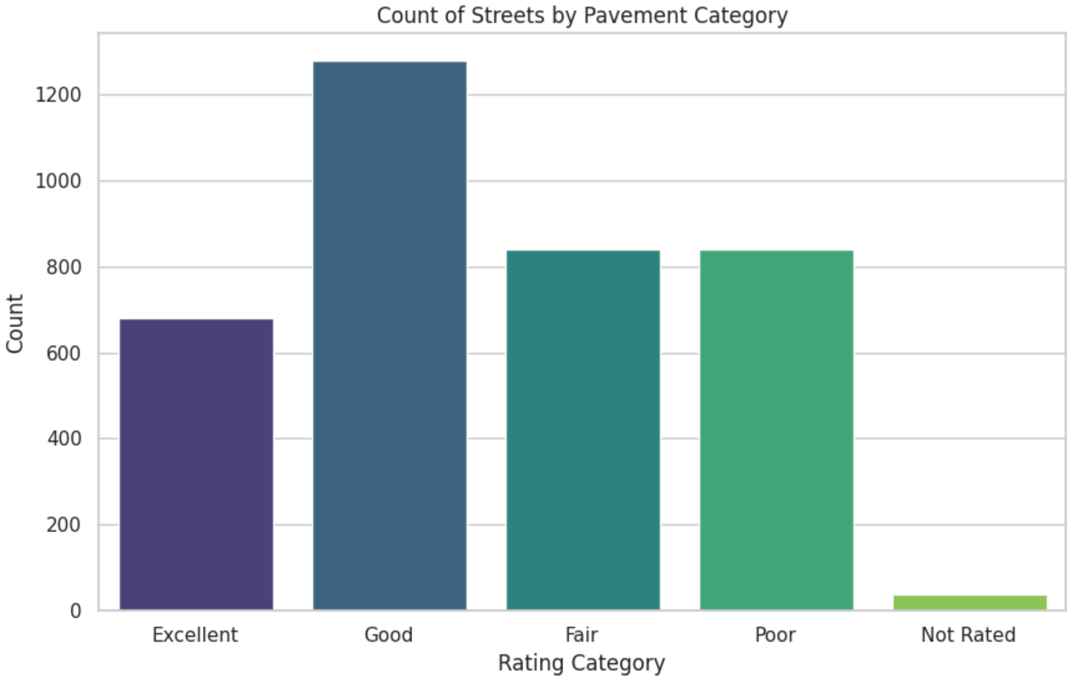
3. Exploratory Data Analysis (Visualizations)

3.1 Distribution of Pavement Ratings (2020)



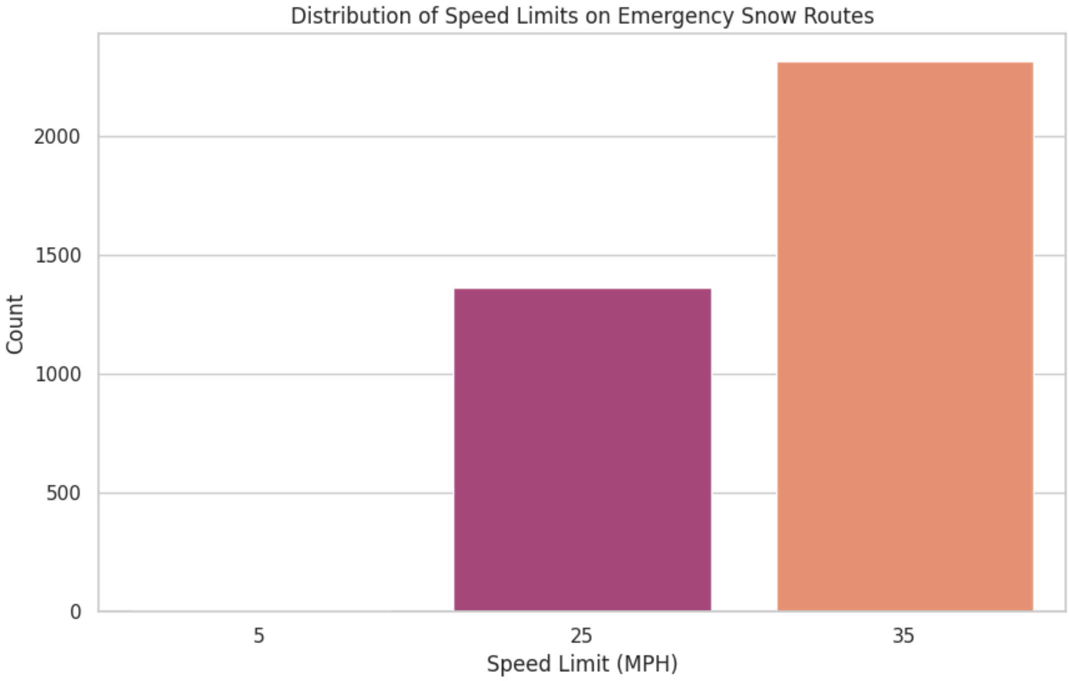
Interpretation: The peak at 7.0 suggests a generally healthy network, but the significant "tail" toward the lower ratings (0-4) identifies critical failure points in the city grid.

3.2 Count of Streets by Pavement Category



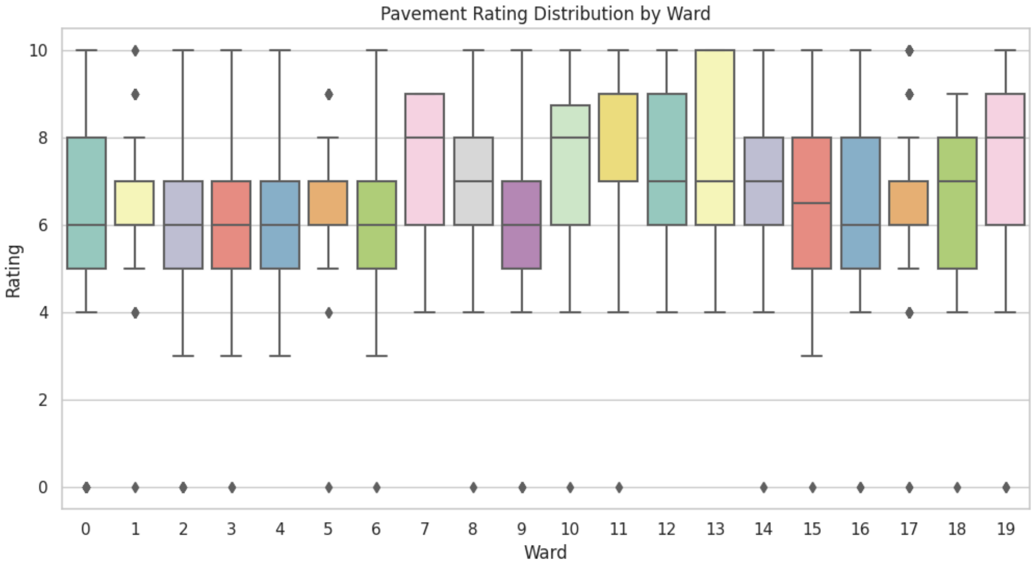
Interpretation: "Good" is the modal category, but "Poor" and "Fair" segments combined represent nearly 46% of the dataset, indicating a massive maintenance backlog.

3.3 Speed Limits on Emergency Snow Routes



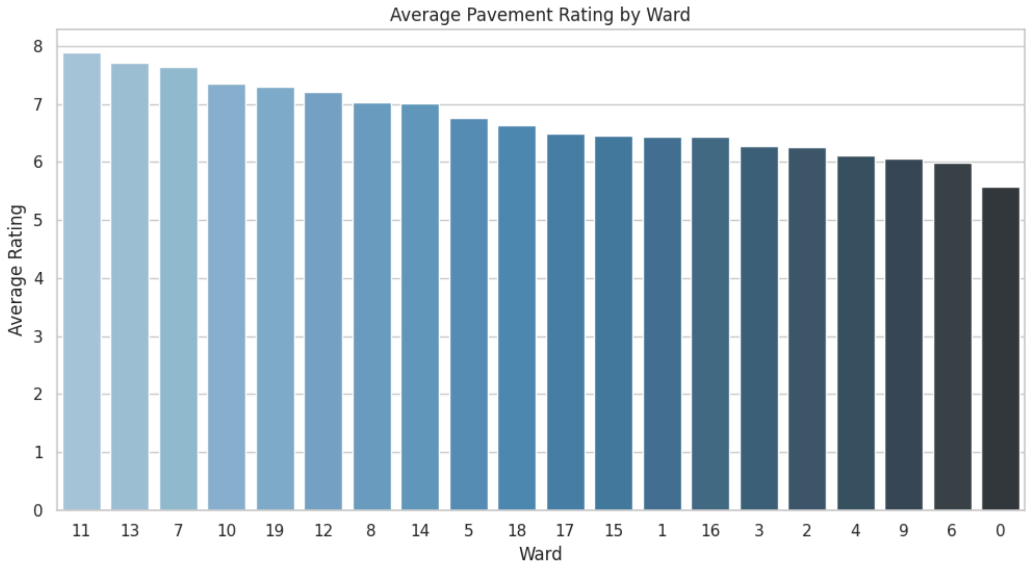
Interpretation: This confirms that Emergency Snow Routes are predominantly higher-speed arterial roads (35 MPH), emphasizing throughput during winter events.

3.4 Pavement Rating Distribution by Ward



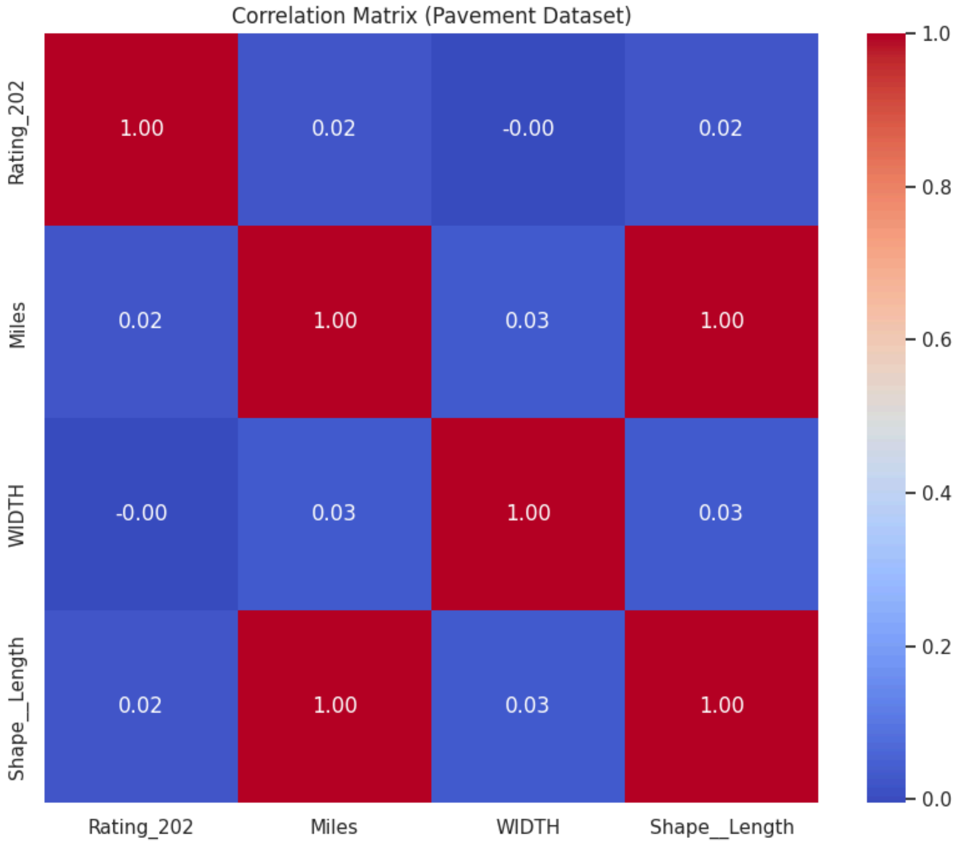
Interpretation: Significant variance exists between Wards. Certain Wards show a much tighter distribution of quality, while others have numerous "Poor" outliers.

3.5 Average Pavement Rating by Ward (Ranked)



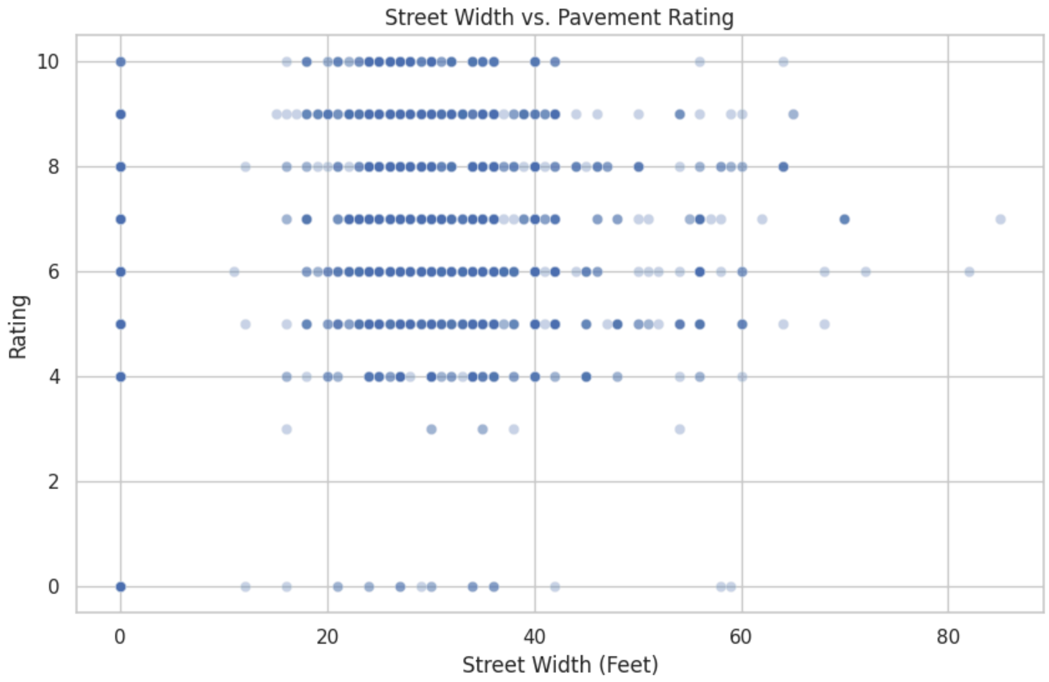
Interpretation: This ranking identifies the geographic focus areas for Phase 3 analysis. The disparity between the highest and lowest-rated Wards is roughly 15-20%.

3.6 Correlation Matrix (Pavement Dataset)



Interpretation: The weak correlation between WIDTH and Rating suggests that road size is not the primary driver of road quality; age and usage likely play bigger roles.

3.7 Street Width vs. Pavement Rating



Interpretation: The scatter plot highlights the cluster of "Zero-Width" errors and shows that the highest-rated roads (9-10) are mostly standard widths (25-40 ft).

4. Findings & Hypotheses

1. **The "Fair" Cliff (Finding):** 841 segments are currently rated as "Fair." In infrastructure management, these are the most critical roads because they are on the verge of becoming "Poor," where repair costs increase exponentially.
2. **Arterial Priority (Hypothesis):** We hypothesize that road quality is higher on Emergency Snow Routes due to their status as critical transport links. Initial speed-limit data supports the idea that these are the city's "Priority A" assets.
3. **Geographic Inequity (Hypothesis):** The variation in Ward averages suggests that infrastructure quality may correlate with Ward-specific socio-economic factors or historical industrial usage.

5. LLM-Assisted Analysis & Validation

- **LLM Insight:** An LLM was used to brainstorm potential causes for the "Zero Width" roads. It suggested they might be "paper streets" or alleys.
- **Validation:** Checking the STREET_NAM for these records confirmed many are indeed "Place," "Lane," or "Alley" designations, validating the LLM's hypothesis.
- **Bias Mitigation:** We ensured the analysis looked at both "Ward" (political) and "Road Class" (technical) to prevent an overly narrow interpretation of why certain areas have lower ratings.