AI Explanations & Insights Report

Generated at: 2026-02-08T21:46:04.541941

Data Quality Summaries

• ingested_Activity Logs.csv: The activitydate column is critically incomplete, with 66.76% of values missing (only 33.2% complete for a required column). Additionally, fertilizer_amount is 14.92% missing and irrigationhours is 10.13% missing.
• ingested_Weather.csv: The observationdate column is significantly incomplete, with 49.84% of values missing (only 50.2% complete for a required column). Other substantial missing data includes rain_unit (19.87%), rainfall (9.9%), and temperature (4.98%). This dataset also contains 29008 duplicate rows based on stationid and observationdate.
• ingested_Station Region.csv: This dataset has 680 duplicate rows based on the stationcode key.
• ingested_Reference Units.csv: The standard_unit and conversion_factor columns both have 25.0% missing values.
• Pipeline Stability: Several pipeline errors occurred, including 2 instances of "Input X contains NaN" (likely related to missing data) and 2 instances of feature mismatch errors, indicating potential downstream processing failures due to data quality.

ingested_Weather.csv

Rows: 50000, Cols: 6

Column	Missing %
observationdate	49.84
rain_unit	19.87
rainfall	9.9
temperature	4.98
stationid	0.0

ingested_Station Region.csv

Rows: 800, Cols: 3

Column	Missing %
stationcode	0.0
region	0.0
region_type	0.0

ingested_Activity Logs.csv

Rows: 30000, Cols: 5

Column	Missing %
activitydate	66.76
fertilizer_amount	14.92
irrigationhours	10.13
region	0.0
croptype	0.0

ingested_Reference Units.csv

Rows: 4, Cols: 3

Column	Missing %
standard_unit	25.0
conversion_factor	25.0
unit	0.0

Anomaly Explanations (Representative)

Here are explanations for the representative anomalies in each dataset, based on missing/invalid values and extreme magnitudes.

Weather Dataset

The weather dataset shows anomalies primarily related to extreme temperature values and missing data.

1. Station S177 on 2023-01-16 (Anomaly Score: 0.932)
• Reason: The temperature value is explicitly NaN (Not a Number), indicating a missing or invalid data point for a critical measurement. This makes the record anomalous as a data quality issue.

2. Station S346 on 2023-01-30 (Anomaly Score: 1.000)
• Reason: The temperature of 44.27 is an extreme magnitude. Assuming typical temperature scales (e.g., Celsius), this is an exceptionally high temperature, far outside the expected range for January in most regions, making it a statistical outlier. If Fahrenheit, it's still notably high compared to typical daily averages, suggesting an unusual peak.

3. Station S279 on 2023-01-03 (Anomaly Score: 0.980)
• Reason: Similar to S346, the temperature of 44.63 is another instance of an extreme high magnitude. This record stands out due to this unusually elevated temperature reading.

4. Station S096 on 2023-01-29 (Anomaly Score: 0.915)
• Reason: The temperature of 29.04 represents an extreme low magnitude. While not as strikingly high as the 44+ degree readings, this temperature could be unusually cold, especially if the dataset typically records warmer climates or if this is outside the expected range for the station's location in late January.

Activity Dataset

Anomalies in the activity dataset are largely characterized by missing key information, often alongside an 'unknown' region.

1. Region 'unknown' on 2023-05-19 (Anomaly Score: 1.000)
• Reason: Both irrigationhours and fertilizer_amount are NaN, meaning critical activity metrics are entirely missing for this record. The region being "unknown" further highlights a potential data completeness or categorization issue, making this a significant data quality anomaly.

2. Region 'unknown' with activitydate: NaN (Anomaly Score: 0.914)
• Reason: The activitydate is NaN, indicating the specific date of the activity is missing. Additionally, fertilizer_amount is also NaN. Missing the date for an activity record is a fundamental data quality problem, making the record difficult to interpret or use for time-series analysis.

3. Region 'West ' on 2023-04-13 (Anomaly Score: 0.839)
• Reason: Similar to the top anomaly, both irrigationhours and fertilizer_amount are NaN. Even though the region is known ("West "), the complete absence of data for these two crucial activity metrics makes this record highly anomalous due to severe data incompleteness.

4. Region 'unknown' with activitydate: NaN (Anomaly Score: 0.751)
• Reason: The activitydate is NaN, and irrigationhours is also NaN. While fertilizer_amount is present, the absence of an activity date and one of the two key metrics severely impacts the record's utility and indicates significant data quality issues.

5. Region 'North' on 2023-06-14 (Anomaly Score: 0.730)
• Reason: In this case, both irrigationhours (5.98) and fertilizer_amount (49.89) are present. The anomaly is due to these values

being extreme magnitudes within their local neighborhood or compared to typical values for the 'North' region at that time of year, making it a statistical outlier despite no missing data.

Station Dataset

All listed anomalies in the station dataset point to a single station with an unusual volume of observations.

1. Station S297 in ' north' region (Anomaly Score: 0.732)
• Reason: This record is anomalous because stationid S297 has an "Unusual observation count" of 920. This count is significantly higher than the regional average of 318 for the ' north' region, indicating an extreme magnitude in data volume compared to its peers.

2. Station S297 in 'west' region (Anomaly Score: 0.726)
• Reason: Again, stationid S297 shows an "Unusual observation count" of 920, which is an extreme magnitude compared to the regional average of 314 for the 'west' region. This consistently high observation count for a single station suggests it might be over-reporting or collecting data at a different frequency than other stations.

3. Station S297 in 'unknown' region (Anomaly Score: 0.670)
• Reason: The stationid S297 is again flagged for an "Unusual observation count" of 920, which is an extreme magnitude against a regional average of 357. The 'unknown' region for this record, combined with the extreme observation count, could indicate both a data categorization issue and an operational anomaly for this particular station.

Dataset: weather

Anomalies: 2500 / 50000

• Station S346, date 2023-01-30, temp 44.27433119318408, rain 0.2107320531999024, score 1.0: Statistical outlier (score: 1.000)
• Station S371, date 2023-01-29, temp 43.39121230910064, rain 0.0377894947559187, score 0.9944085134645024: Statistical outlier (score: 0.994)
• Station S279, date 2023-01-03, temp 44.6328849454904, rain 0.320070240060416, score 0.9801960979743932: Statistical outlier (score: 0.980)
• Station S230, date 2023-01-13, temp 44.25379105105102, rain 0.6265978202537825, score 0.956282083533104: Statistical outlier (score: 0.956)
• Station S048, date 2023-01-04, temp 41.39940204267012, rain 0.3525379908654646, score 0.9464122842012416: Statistical outlier (score: 0.946)
• Station S029, date 2023-01-23, temp 36.66343583144092, rain 0.19959658228893, score 0.9453200163105692: Statistical outlier (score: 0.945)
• Station S177, date 2023-01-16, temp nan, rain 0.012155707875, score 0.9318224538344: Statistical outlier (score: 0.932)
• Station S178, date 2023-03-09, temp 41.82752106964189, rain 0.1268993806264884, score 0.9242496912563948: Statistical outlier (score: 0.924)
• Station S079, date 2023-01-26, temp 43.07715441014209, rain 0.5780416693747348, score 0.9226296780688134: Statistical outlier (score: 0.923)
• Station S065, date 2023-02-12, temp 44.61761693794386, rain 0.5665099700109477, score 0.9216290132797812: Statistical outlier (score: 0.922)

Dataset: activity

Anomalies: 1374 / 30000

• Region unknown, date 2023-05-19, crop , irrigation nan, fertilizer nan, score 1.0: Statistical outlier in local neighborhood (score: 1.000)
• Region unknown, date nan, crop , irrigation 5.617074151757526, fertilizer nan, score 0.9144864161082896: Missing activity date
• Region unknown, date nan, crop , irrigation 5.494239874267086, fertilizer nan, score 0.8850069927411767: Missing activity date
• Region West , date 2023-04-13, crop , irrigation nan, fertilizer nan, score 0.8394434125384006: Statistical outlier in local neighborhood (score: 0.839)
• Region unknown, date nan, crop , irrigation nan, fertilizer 54.819635296717074, score 0.7507359262456135: Missing activity date
• Region North, date 2023-06-14, crop , irrigation 5.985662848513651, fertilizer 49.89587826498891, score 0.7298198099896092: Statistical outlier in local neighborhood (score: 0.730)
• Region unknown, date nan, crop , irrigation nan, fertilizer 47.205706233540326, score 0.7140518567761128: Missing activity date
• Region unknown, date nan, crop , irrigation nan, fertilizer 47.56264329923287, score 0.7031532912542942: Missing activity date
• Region unknown, date nan, crop , irrigation nan, fertilizer 47.59708376216747, score 0.6979947089280887: Missing activity date

• Region unknown, date nan, crop , irrigation nan, fertilizer 53.84434558466699, score 0.6884072933593138: Missing activity date

Dataset: station

Anomalies: 8 / 800

stationcode	region	temp_count	temp_mean	rain_mean	anomaly_score	anomaly_reason
S297	north				0.7320895059031696	Unusual observation count (920 vs regional avg 318)
S297	west				0.725955803350194	Unusual observation count (920 vs regional avg 314)
S297	west				0.725955803350194	Unusual observation count (920 vs regional avg 314)
S297	unknown				0.6702197647320405	Unusual observation count (920 vs regional avg 357)
S297	unknown				0.6702197647320405	Unusual observation count (920 vs regional avg 357)
S297	SOUTH				0.6008538600591692	Unusual observation count (920 vs regional avg 352)
S297	SOUTH				0.6008538600591692	Unusual observation count (920 vs regional avg 352)
S297	SOUTH				0.6008538600591692	Unusual observation count (920 vs regional avg 352)

Decision Support

Data Quality Actions
• Prioritize addressing the extremely high missing percentages for 'observationdate' (49.84%) in ingested_Weather.csv and 'activitydate' (66.76%) in ingested_Activity Logs.csv to ensure data completeness for time-series analysis.
• Implement robust missing value handling for critical data points like 'rain_unit' (19.87%), 'rainfall' (9.9%), and 'temperature' (4.98%) in ingested_Weather.csv, and 'fertilizer_amount' (14.92%) and 'irrigationhours' (10.13%) in ingested_Activity Logs.csv, as these likely contribute to "Input X contains NaN" pipeline errors.
• Investigate and correct the source of "Pipeline failed: X has 8 features, but StandardScaler is expecting 5 features" errors, which indicates a fundamental mismatch in data schema or processing steps.
• Address the "Failed to validate Weather.csv: Object of type Series is not JSON serializable" error to ensure Weather.csv can be correctly parsed and used in the data pipeline.
• Rectify the 25.0% missing data for both 'standard_unit' and 'conversion_factor' in ingested_Reference Units.csv to ensure accurate unit standardization across datasets.

Station Maintenance
• Investigate station 'S297' immediately due to repeated high anomaly scores and consistently high observation counts (920) significantly exceeding regional averages (e.g., 318, 314, 357).
• Analyze the data collection mechanism for station 'S297' to determine if the "Unusual observation count" is due to sensor malfunction, misconfiguration, or genuine data influx.
• Review the total of 8 anomalies detected across 800 station records (1% anomaly rate) to identify any other stations exhibiting unusual behavior beyond S297.
• Cross-reference 'stationid' 'S297' with its associated regions (' north', 'west', 'unknown') to clarify regional assignments and ensure data consistency.

Suspicious Activity Patterns
• Conduct a deeper investigation into the 2500 weather anomalies (5% of total weather records) to identify specific patterns, locations, or timeframes for unusual weather events.
• Analyze the 1374 activity anomalies (approximately 4.58% of total activity records) to understand if these relate to specific crop types, regions, or unusual operational practices.
• Correlate weather anomalies with activity anomalies to determine if extreme weather conditions are influencing agricultural activities or vice-versa.
• Review data collection processes for both weather and activity logs, especially considering the high missing percentages for 'observationdate' (49.84%) and 'activitydate' (66.76%), as incomplete data can mimic or obscure true anomalies.
• Investigate if the missing 'rain_unit' (19.87%) and 'rainfall' (9.9%) in weather data contribute to or obscure anomalies related to

precipitation.

Governance / Audit Notes

Audit events: 68 | Errors: 16