AI Explanations & Insights Report

Generated at: 2026-02-07T10:59:02.028979

Data Quality Summaries

- Critical Missing Date Data: The ingested_Weather.csv dataset has 24920 (49.84%) missing values in its required observationdate column, and ingested_Activity Logs.csv has 20029 (66.76%) missing values in its required activitydate column.
- High Duplicate Records: ingested_Weather.csv contains 29008 duplicate rows on stationid and observationdate, while ingested_Station Region.csv has 680 duplicate rows out of 800 based on stationcode.
- Additional Missing Values:
- ingested_Weather.csv also suffers from 19.87% missing rain_unit, 9.9% missing rainfall, and 4.98% missing temperature.
- ingested_Activity Logs.csv has 14.92% missing fertilizer_amount and 10.13% missing irrigationhours.
- ingested_Reference Units.csv shows 25.0% missing values for both standard_unit and conversion_factor.
- Data Processing Errors: Pipeline logs indicate errors during validation and filtering for ingested_Weather.csv and ingested_Station Region.csv.

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, grouped by dataset:

### Weather Dataset Anomalies

1. Extreme High Temperature (e.g., Station S346, 2023-01-30, temperature: 44.27):
This record is anomalous due to an extreme magnitude in the temperature reading. A temperature of 44.27 degrees (assuming Celsius, which is typical for weather data unless specified otherwise) in January is exceptionally high for most regions globally, indicating a potential sensor malfunction, an extreme heat event, or a data entry error.

2. Missing Temperature Data (e.g., Station S177, 2023-01-16, temperature: NaN):
This anomaly is characterized by a missing value (NaN) for the temperature field. This directly indicates a data quality issue where temperature data was not recorded or is invalid for this specific station and date, making the record incomplete.

3. Extreme Low Temperature (e.g., Station S096, 2023-01-29, temperature: 29.04):
This record highlights an extreme magnitude in temperature on the lower end, at 29.04 degrees. While not as high as the first example, if the typical climate for the station's location in January is significantly different, this could represent an unusually cold day or a possible sensor error, flagging it as an outlier.

### Activity Dataset Anomalies

1. Multiple Missing Key Metrics (e.g., region 'unknown', 2023-05-19, score 1.0):
This record is highly anomalous because both critical activity metrics, irrigationhours and fertilizer_amount, are missing (NaN). This represents a severe data completeness issue, suggesting that no quantifiable activity data was recorded for this date, despite the record existing. The region being "unknown" further indicates poor data quality for this entry.

2. Missing Activity Date (e.g., region 'unknown', activitydate: NaN, score 0.914):
Several anomalies show a missing activitydate (NaN), such as this example where irrigationhours (5.617) is present but the activitydate is not. The absence of a date makes it impossible to correctly timestamp and contextualize the recorded activity, rendering the data partially invalid due to crucial temporal information being absent.

3. Missing Activity Date and Irrigation Hours (e.g., region 'unknown', activitydate: NaN, irrigationhours: NaN, score 0.751):
This anomaly demonstrates missing values for both activitydate and irrigationhours (NaN), while fertilizer_amount (54.819) is present. This indicates an incomplete record where key activity details and their temporal context are unavailable, making the fertilizer_amount data difficult to interpret or use effectively.

4. Statistical Outlier with All Values Present (e.g., region 'North', 2023-06-14, score 0.730):
This record from the "North" region on 2023-06-14 contains explicit values for irrigationhours (5.98) and fertilizer_amount (49.89), with no missing data. However, it is flagged as a "Statistical outlier." This means that the combination or specific magnitudes of these values are highly unusual when compared to typical patterns for activities in the "North" region or for that time of year, even without apparent data quality issues like NaN or obvious extreme magnitudes.

### Station Dataset Anomalies

1. Unusually High Observation Count for Station S297 (e.g., all top anomalies):
All the listed anomalies for the 'station' dataset point to stationid: "S297". The consistent anomaly_reason across these entries is "Unusual observation count (920 vs regional avg X)". This clearly indicates that Station S297 has an extremely high number of recorded observations (920) compared to the regional averages (which range from 314 to 357). This extreme magnitude suggests either a highly active station, a misconfiguration in data collection, or a data aggregation error, consistently marking S297 as an outlier based on its volume of observations.

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 investigation into the 'observationdate' column in 'Weather.csv' due to a critical 49.84% missing data rate, severely impacting data completeness.
- Address the high percentage of missing 'activitydate' (66.76%) in 'Activity Logs.csv' to ensure proper tracking and analysis of agricultural activities.
- Resolve pipeline errors preventing validation of 'Weather.csv' and 'Station Region.csv' due to "Object of type Series is not JSON serializable," hindering data ingestion.
- Improve data collection for 'rain_unit' (19.87% missing) and 'rainfall' (9.9% missing) in 'Weather.csv' to ensure complete environmental data.
- Review missing 'fertilizer_amount' (14.92%) and 'irrigationhours' (10.13%) in 'Activity Logs.csv' to better understand and optimize resource usage.
- Rectify the 25.0% missing 'standard_unit' and 'conversion_factor' in 'Reference Units.csv' to ensure accurate unit transformations.

Station Maintenance

- Immediately investigate station S297 due to its consistently "Unusual observation count (920)" compared to regional averages (e.g., 318 for 'north', 314 for 'west', 357 for 'unknown').
- Verify the physical status and sensor calibration of station S297 to rule out equipment malfunction or misconfiguration, given it's listed across multiple regions with the same anomaly.
- Cross-reference data from neighboring stations to determine if the high observation count for S297 is an isolated issue or part of a wider regional data anomaly.
- Conduct a historical data audit for station S297 to identify the onset and duration of these unusual observation patterns.
- Review data transmission protocols and regional assignments for station S297 given its repeated high anomaly scores.

Suspicious Activity Patterns

- Investigate the 5% anomaly rate in weather data (2500 anomalies out of 50000 rows) for potential sensor tampering or unusual environmental events.
- Analyze the 4.58% anomaly rate in activity logs (1374 anomalies out of 30000 rows) to identify any unauthorized or irregular agricultural practices.
- Correlate high missing data percentages for 'activitydate' (66.76%), 'fertilizer_amount' (14.92%), and 'irrigationhours' (10.13%) in 'Activity Logs.csv' with the activity anomaly rate.
- Examine if significant missing 'observationdate' (49.84%), 'rain_unit' (19.87%), and 'rainfall' (9.9%) in 'Weather.csv' are linked to periods of weather data anomalies or deliberate data suppression.
- Assess if the unusual observation count at station S297 (920 vs regional averages) contributes to the overall weather anomaly rate, potentially indicating data injection or faulty readings.

Governance / Audit Notes

Audit events: 112 | Errors: 64