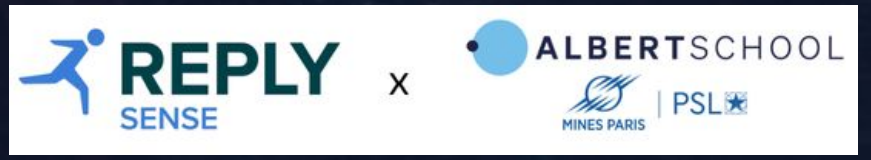
Data Preparation for Anomaly Detection

Optimizing Substation Energy Consumption Analysis

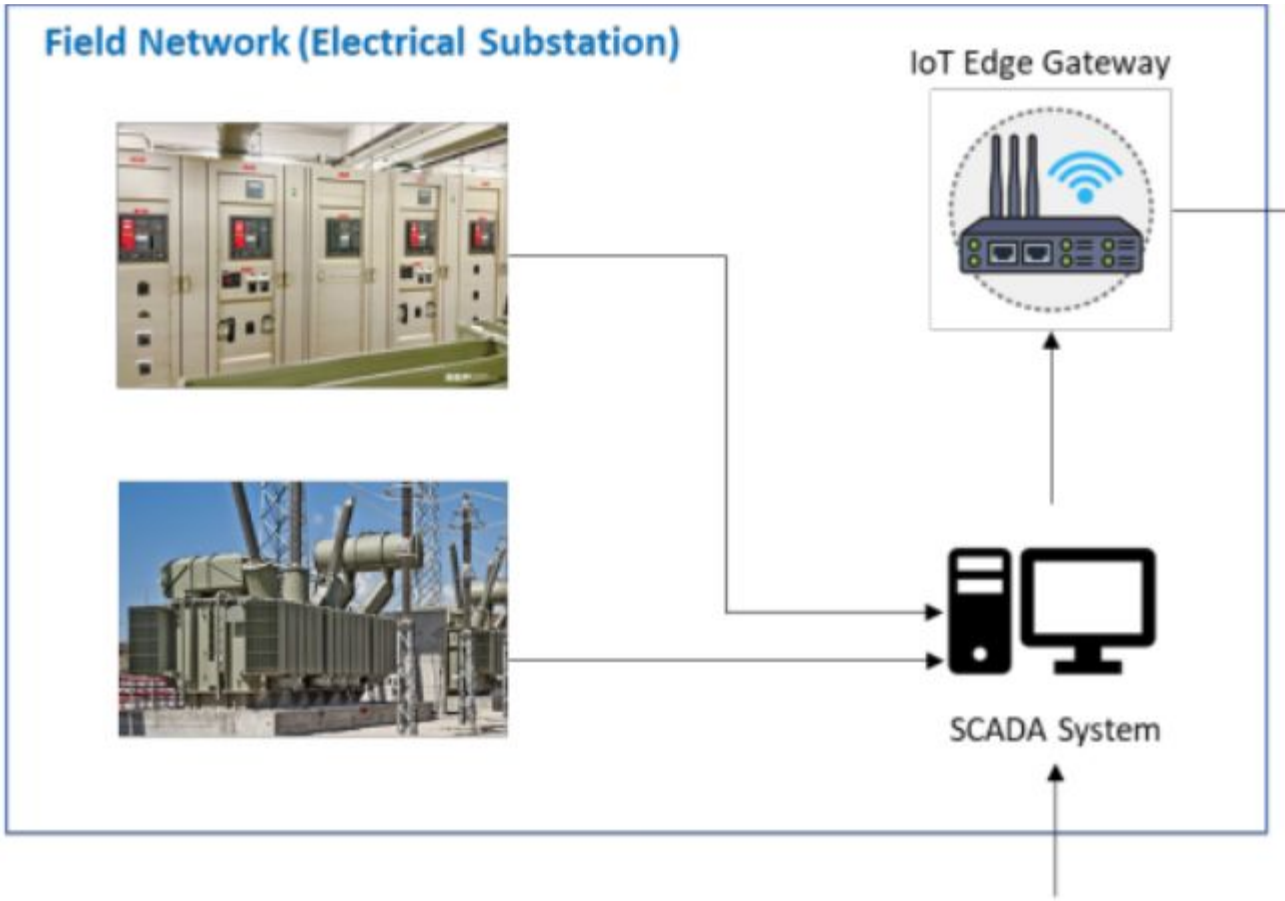
Group #1



The Pipeline: Where Quality Issues Emerge

Our analysis operates within a complex data pipeline spanning from the physical edge to the cloud.

- **Field Network:** Raw signals from electrical substations.
- **IoT Edge Gateway:** First point of aggregation.
- **Cloud Platform & Data Lake:** Central repository where data quality issues often become visible.



REPLY SENSE x ALBERTSCHOOL MINES PARIS | PSL

Establish Critical Infrastructure & Data Needs

Critical Infrastructure

- Substations = critical systems needing constant monitoring.
- IoT → huge time-series data volumes.
- Reliable data is essential for secure power delivery.

Data Quality Barrier

- Raw data = noise, gaps, inconsistencies, duplicates.
- Data pipeline introduces quality loss.
- Poor data → unreliable ML & anomaly detection.

REPLY SENSE x ALBERTSCHOOL MINES PARIS | PSL

Audit Reveals Critical Data Integrity Issues

Significant integrity issues were identified in the provided dataset before processing, which would inevitably alter any predictive modeling efforts.

4,189

MISSING VALUES (8.41%)

4,108

ERRONEOUS ZEROS (8.25%)

1,493

DUPLICATE ROWS

REPLY SENSE x ALBERTSCHOOL MINES PARIS | PSL

Two-Phase Data Cleaning Strategy

1. Structural Cleaning

✔️ Removed duplicate rows

✔️ Resampling to 10 min intervals

2. Value Imputation

✔️ Handling missing values

✔️ Handling zeros

✔️ Linear interpolation to fill gaps

Issue	Before	After	Result
Missing Values	4189	0	100% resolved
Erroneous Zeros	4108	0	100% resolved
Duplicate Timestamps	1,493	0	100% removed
Total Rows	51,290	52,417	Data consolidated

REPLY SENSE x ALBERTSCHOOL MINES PARIS | PSL

Transforming Data for Model Use

Feature Name	Description / Logic
TIMEVALUE	Difference between current and previous row timestamp (Delta T).
DAY_OF_WEEK	Extracted integer (0=Monday ... 6=Sunday) to capture weekly cycles.
AVG_CONSUMPTION	Average of Zone 1, 2, and 3 consumption at that timestamp.
EXPECTED_MEAN	Expected Mean Consumption value at a specific Hour / Day, from the pivot table.
EXPECTED_STD	Expected Standard Deviation value at a specific Hour / Day, from the pivot table.

REPLY SENSE x ALBERTSCHOOL MINES PARIS | PSL

Data Cleaning Success: A Reliable Baseline

The rigorous preprocessing pipeline successfully consolidated the data, creating a clean baseline for the model.

100%
MISSING VALUES RESOLVED

100%
ZEROS CORRECTED

52,417
TOTAL CLEAN ROWS

REPLY SENSE x ALBERTSCHOOL MINES PARIS | PSL

Z-Score Anomaly Detection

The Logic

We implemented a statistical approach to identify anomalies. By calculating the Z-score, we measure how many standard deviations a data point is from the expected mean for that specific time and day.

Threshold: Any data point where the absolute Anomaly Score exceeded 2.5 was flagged as an anomaly.

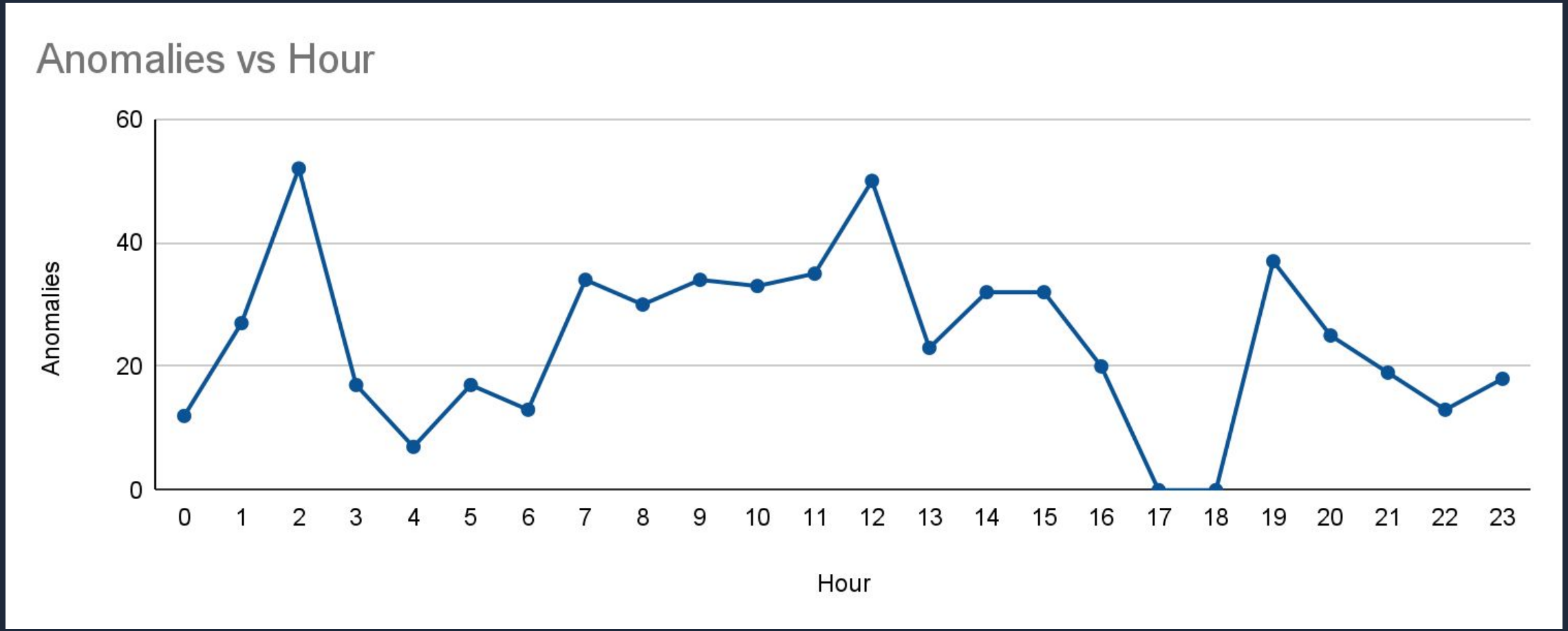
The Formula

$$\text{Anomaly\_Score} = \frac{\text{Avg\_consumption - Expected\_Mean}}{\text{Expected\_StdDev}}$$

Is_Anomaly = 1 IF ABS(Anomaly_Score) >= 2.5

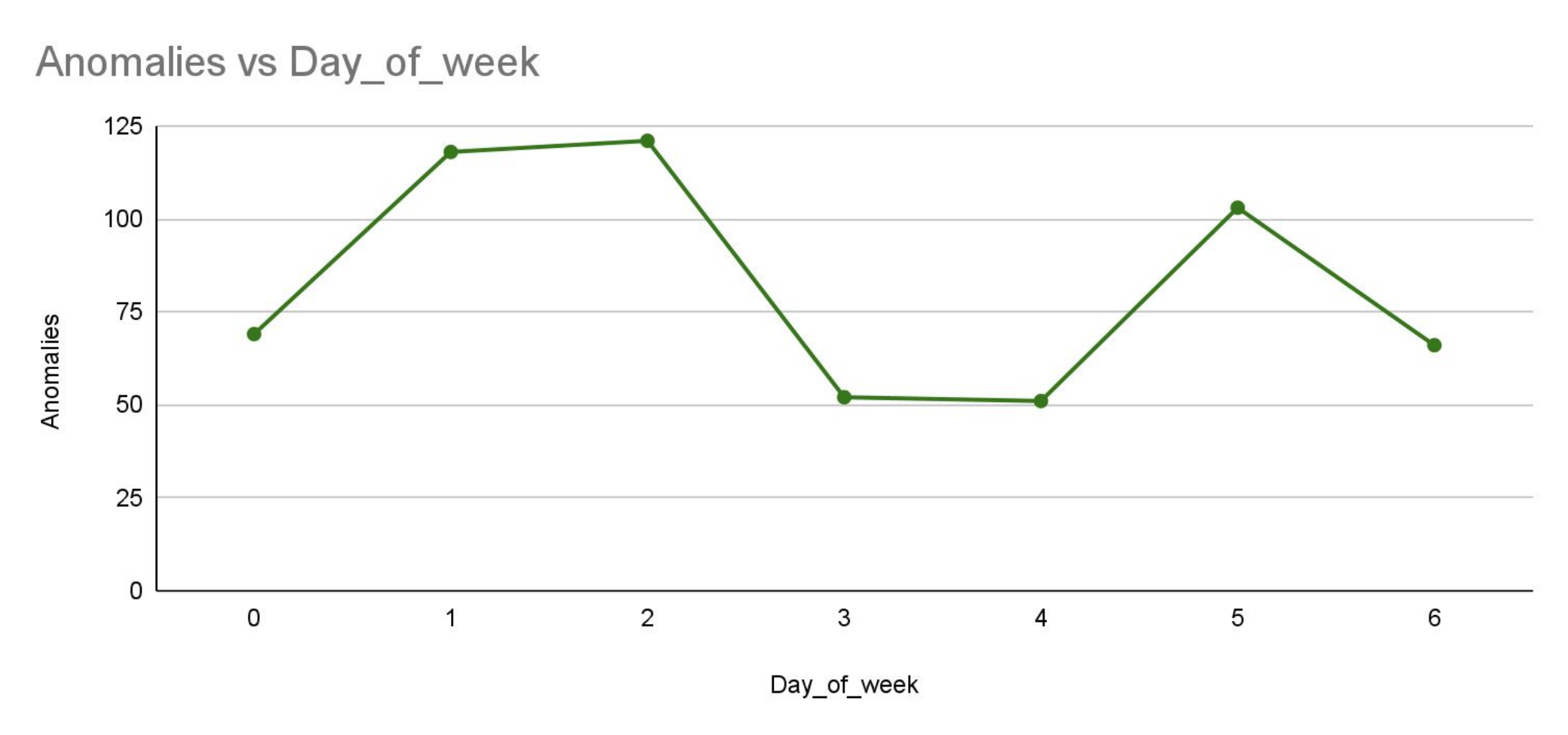
REPLY SENSE x ALBERTSCHOOL MINES PARIS | PSL

Visualizing Detected Anomalies



REPLY SENSE x ALBERTSCHOOL MINES PARIS | PSL

Visualizing Detected Anomalies



REPLY SENSE x ALBERTSCHOOL MINES PARIS | PSL

Impact on Anomaly Detection

	Before	After
Anomalies	**971**	**580**

The cleaning process reduced detected anomalies from 971 to 580 (40.27% reduction), filtering out false positives caused by data quality issues.

40%

REDUCTION

REPLY SENSE x ALBERTSCHOOL MINES PARIS | PSL

Architecture Recommendations to Prevent Data Integrity Issues

Edge Validation

Implement basic validation logic (Range checks, Null checks) directly at the IoT Gateway level to prevent transmission of obviously erroneous data.

Timestamp Synchronization

Ensure SCADA and Edge Gateway clocks are synchronized via the Network Time Protocol to eliminate the 1,493 duplicate timestamp errors observed in the dataset.

REPLY SENSE x ALBERTSCHOOL MINES PARIS | PSL