

## Report

**1 Systematic Unit Mismatch and Extreme Outliers in GAS**

A critical data quality issue was discovered where GAS utility readings are systematically recorded in kWh instead of the expected kg. Furthermore, over 1,000 records show physically impossible values exceeding  $10^8$ , primarily in STEAM and GAS. These anomalies are not random but concentrated in specific sites (e.g., site 44006) and older buildings, suggesting sensor calibration or data-logger configuration errors rather than actual consumption spikes.

**2 Structural Disconnect: Broken Linkage Between...**

There is a fundamental lack of direct key linkage between the building\_metadata (buildingnumber) and the meter data (siteid). This structural gap forced the use of fuzzy string matching on names, which only successfully mapped about 75-84% of sites. The remaining sites, including potential high-consumption facilities, remain "orphaned," making it impossible to correlate their energy usage with building attributes like floor area or construction date without manual intervention.

**3 Significant Bias from Non-Random Missing Data**

Missing meter readings are not distributed randomly; they are utility-specific, with STEAM and HEAT showing missingness rates over 10%. Analysis revealed that simple mean-imputation increases total consumption figures by up to 14% for certain utilities. This indicates that ignoring or naively filling these gaps would introduce substantial bias into any carbon footprint estimation or energy benchmarking report.

## Data Integrity & Quality Audit: Campus Energy Systems (Sept-Oct 2025)

This report provides a comprehensive evaluation of the data quality across the building metadata, meter readings, and weather datasets. Our analysis has identified critical inconsistencies that must be addressed before any reliable research or modeling can proceed.

### 1. The Outlier & Unit Mismatch Crisis

The most immediate concern is the presence of "physically impossible" data points and systematic logging errors, particularly in the STEAM and GAS sectors.

#### Systematic Unit Mismatch in GAS

We discovered a widespread configuration error where **GAS utility meters are reporting in kWh instead of the expected kg**. This affects nearly 240,000 records, suggesting a default setting in the data ingestion pipeline was never overridden for gas assets.

Utility	Expected Unit	Observed Primary Unit	Status
0 GAS	kg	kWh (Mismatch)	Critical Error
1 STEAM	kg	kg	Correct
2 ELECTRICITY	kWh	kWh	Correct
3 HEAT	kWh	kWh	Correct
4 COOLING	kWh	kWh	Correct

#### Extreme Outliers ( $>10^8$ )

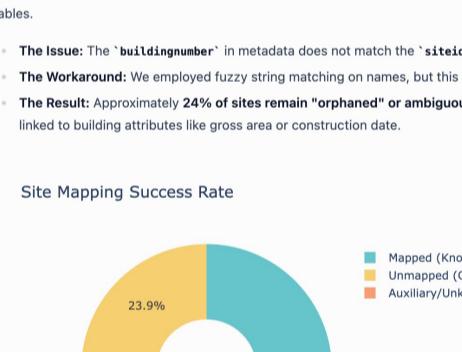
A total of 1,194 records contain astronomically high values ( $10^8$  to  $10^{11}$ ), which are physically implausible.

Distribution of Extreme Outliers ( $>10^8$ ) by Utility



**Key Observation:** These anomalies are not random. They are heavily concentrated in **older buildings** and specific sites (e.g., Site 44006). The chart below shows that buildings constructed around 1918 and 1970 are "hotspots" for sensor errors.

Outlier Density by Construction Year



### 2. The Metadata Gap: Broken Linkage

A significant hurdle for building-level analysis is the lack of a shared primary key between the `'building_metadata'` and `'meter-data'` tables.

- The Issue:** The `'buildingnumber'` in metadata does not match the `'siteid'` in meter tables.
- The Workaround:** We employed fuzzy string matching on names, but this is an imperfect bridge.
- The Result:** Approximately **24%** of sites remain "orphaned" or ambiguous. These sites have energy data but cannot be reliably linked to building attributes like gross area or construction date.

Site Mapping Success Rate

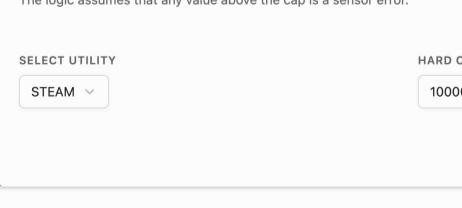


### 3. Statistical Bias: The Risk of Naive Imputation

Missing meter readings are not distributed randomly, which poses a major risk for aggregate reporting and carbon footprint estimation.

- Non-Randomness:** Missingness is utility-specific. STEAM and HEAT have missingness rates exceeding 10%, while GAS is nearly complete.
- The Bias:** If we fill these gaps with simple averages (mean imputation), the total reported consumption for STEAM jumps by **14%**. This indicates that the missing periods likely occurred during high-usage times or at high-usage sites.

Impact of Mean Imputation on Reported Totals



### Conclusion & Recommended Actions

To ensure the data is "research-ready," the following pre-processing steps are mandatory:

- Immediate Unit Correction:** Convert GAS readings from kWh to kg using a verified conversion factor.
- Targeted Sensor Audit:** Prioritize physical inspection of meters at **Site 44006** and **Site 44056**, which drive the majority of STEAM anomalies.
- Establish Master Mapping:** Create a permanent, manually verified lookup table between `'siteid'` and `'buildingnumber'` to resolve the 24% mapping gap.
- Advanced Imputation:** Do not use simple means for STEAM or HEAT. Implement regression-based imputation (using weather and building size as predictors) to avoid the 14% bias identified in this report.

**Outlier Impact Simulator**

Adjust the threshold to see how many records would be flagged as outliers and the resulting reduction in 'phantom' energy usage. The logic assumes that any value above the cap is a sensor error.

Select Utility

HARD CAP (Value)

### Conclusion & Recommended Actions

To ensure the data is "research-ready," the following pre-processing steps are mandatory:

- Immediate Unit Correction:** Convert GAS readings from kWh to kg using a verified conversion factor.
- Targeted Sensor Audit:** Prioritize physical inspection of meters at **Site 44006** and **Site 44056**, which drive the majority of STEAM anomalies.
- Establish Master Mapping:** Create a permanent, manually verified lookup table between `'siteid'` and `'buildingnumber'` to resolve the 24% mapping gap.
- Advanced Imputation:** Do not use simple means for STEAM or HEAT. Implement regression-based imputation (using weather and building size as predictors) to avoid the 14% bias identified in this report.