

1.2.1 Statistical Analysis Report

1. Introduction

This report presents the statistical analysis of the unified and validated water consumption dataset described in *Validation Report 1.1.1.b*. The objective is to quantify the key relationships between meter characteristics and consumption behavior, identify anomaly patterns, and extract insights that will inform feature selection and thresholding for subsequent modeling tasks under WP 1.2.1.b.

The analysis focuses on several core aspects: the statistical and distributional exploration of consumption, the examination of temporal and categorical variability across meter brands, models and diameters, and the identification and characterization of anomalous readings. The report concludes with the implications of these findings for future variable selection and predictive modeling.

2. Dataset Overview

The dataset comprises millions of daily water consumption readings collected from a wide variety of meters differing in brand, model and nominal diameter. The primary variables include CONSUMO_REAL (the measured daily water consumption), MARCA_COMP, CODI_MODEL, and DIAM_COMP (categorical descriptors of the meter), as well as FECHA (the daily timestamp). Weather variables—tavg, tmin, tmax, and prcp—were also joined to each record to allow cross-analysis with environmental factors.

Following the data cleaning and validation procedures outlined in WP 1.1.1.b, the dataset was verified for internal consistency, duplicate removal and standardized data types. The analyses presented here were conducted using the cleaned *df_plot* version of the dataset, which is optimized for visualization and grouped computations.

3. Exploratory Statistical Analysis

3.1 Consumption Distribution

Initial examination of CONSUMO_REAL revealed a heavily right-skewed distribution dominated by numerous zero readings, particularly associated with one widely deployed meter brand (5557SZ47QZAZ56EQ). To better understand the underlying variability, a $\log(x+1)$ transformation was applied. This transformation produced a distribution much closer to normality and highlighted subtle differences within specific meter groups.

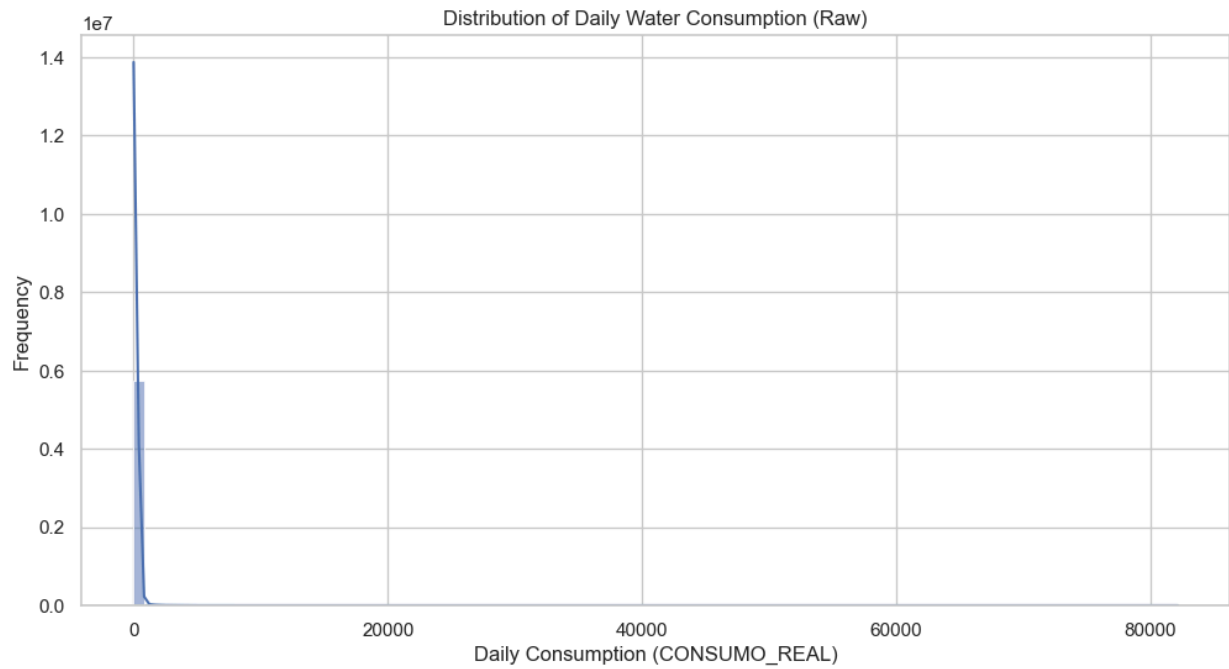


Figure 1: Distribution of Consumption (Raw)

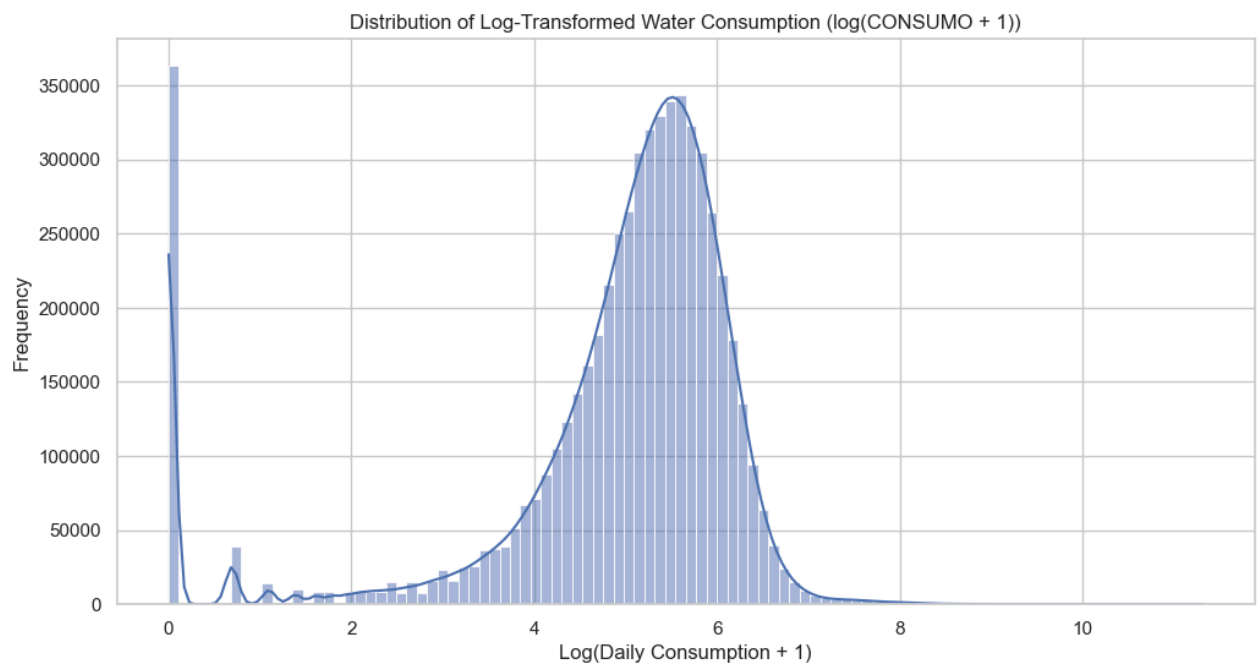


Figure 2: Distribution of Consumption (Log-Transformed)

Overall, the raw data show strong zero inflation and right skewness, while the transformed data display a more symmetric distribution. This transformation supports the later application of robust anomaly detection techniques, such as the Median Absolute Deviation (MAD) and the Interquartile Range (IQR) methods.

3.2 Consumption by Brand and Diameter

When comparing consumption levels across meter brands and diameters, significant differences emerged. Meters with larger diameters (20 mm and 30 mm, both represented by the 5557SZ47QZAZ56EQ brand) exhibited higher and more dispersed consumption values, which aligns with industrial or multi-user contexts. In contrast, 15 mm meters showed lower, more stable consumption typical of residential use.

Across the 15 mm group, brands behaved similarly, with only minor differences in spread. These findings suggest that diameter is the primary driver of consumption variability, while brand effects are secondary and mainly relevant within smaller residential meters.

The following boxplot compares CONSUMO_REAL across brands and nominal diameters on a logarithmic scale.

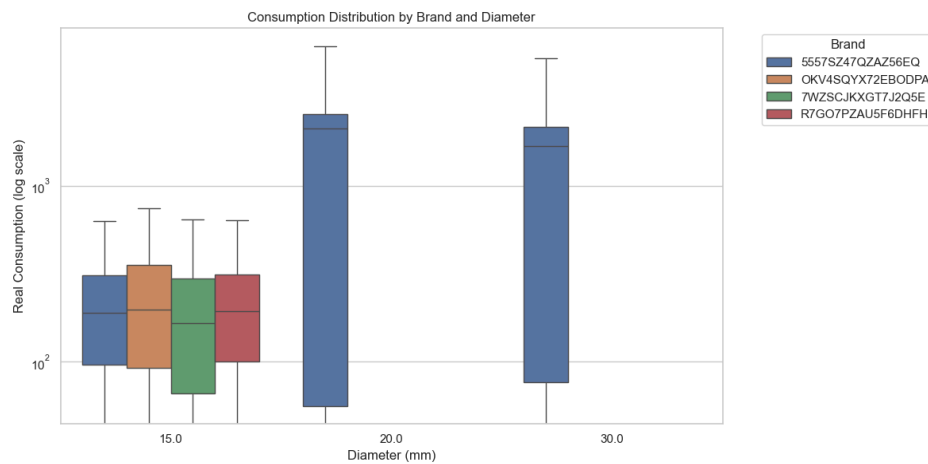


Figure 3: Consumption Distribution by Brand and Diameter

3.3 Temporal Consumption Patterns

An analysis of average weekly water consumption revealed clear periodic fluctuations for all meter diameters, suggesting consistent weekly and seasonal usage cycles. Consumption tends to increase during the summer months, especially for larger meters. The 20–30 mm meters demonstrated much higher and more volatile consumption levels, while the 15 mm group maintained low, stable readings throughout the year.

These findings indicate that consumption behavior is systematically influenced by both meter capacity and temporal factors, confirming the presence of seasonal demand patterns that should be incorporated into anomaly detection and forecasting models.

The following plot shows the average weekly consumption across the full observation period, grouped by meter diameter.

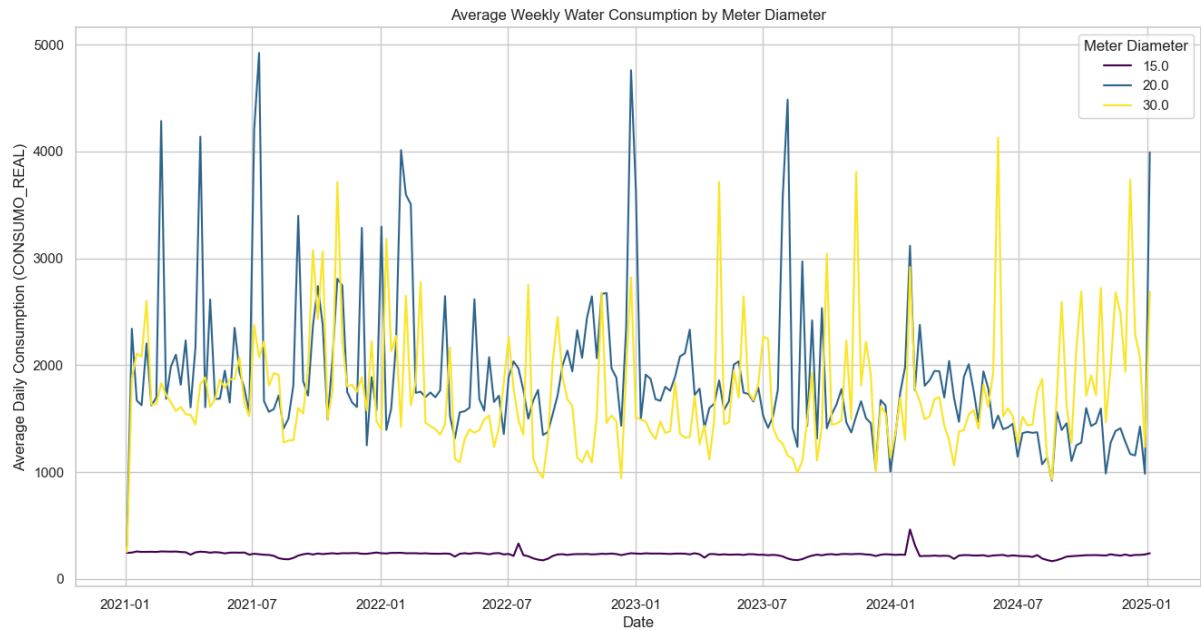


Figure 4: Average Weekly Water Consumption by Meter Diameter

3.4 Correlation with Weather

A correlation matrix between consumption and daily weather variables (temperature and precipitation) was computed.

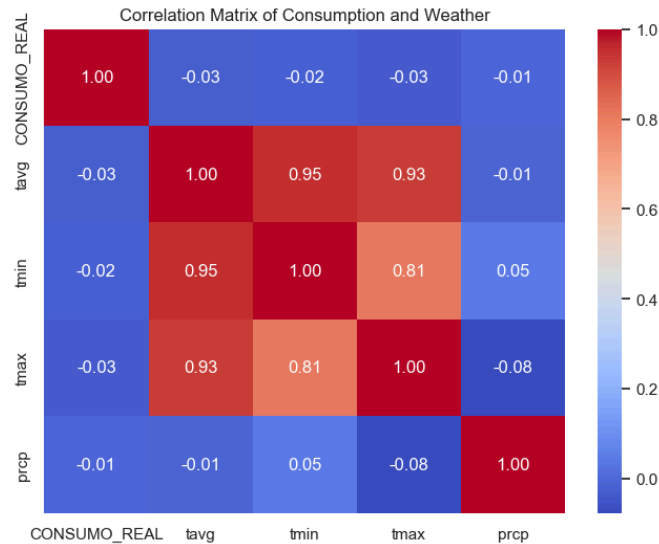


Figure 5: Correlation Matrix of Consumption and Weather Variables

The correlation analysis revealed negligible linear relationships. Neither temperature nor rainfall showed a strong direct correlation with daily water consumption. While weather may indirectly influence consumption through seasonal patterns, it does not explain short-term anomalies.

4. Anomaly Detection and Characterization

4.1 Detection Methods

Anomalies were identified separately for each combination of MARCA_COMP, CODI_MODEL and DIAM_COMP using three complementary statistical criteria:

1. **Standard deviation rule:** $|x - \mu| > 3\sigma$
2. **Interquartile range (IQR) rule:** $x \notin [Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR]$
3. **Modified Z-score (MAD-based):** $|z| > 3$

A data point was labeled as anomalous if any of these methods flagged it. Anomalies were further categorized into high anomalies (unusually large consumption) and low anomalies (near-zero or unexpectedly small consumption).

The following figure compares anomaly rates across meter groups (brand-diameter) using the previous detection criteria introduced: standard deviation, interquartile range (IQR) and modified z-score.

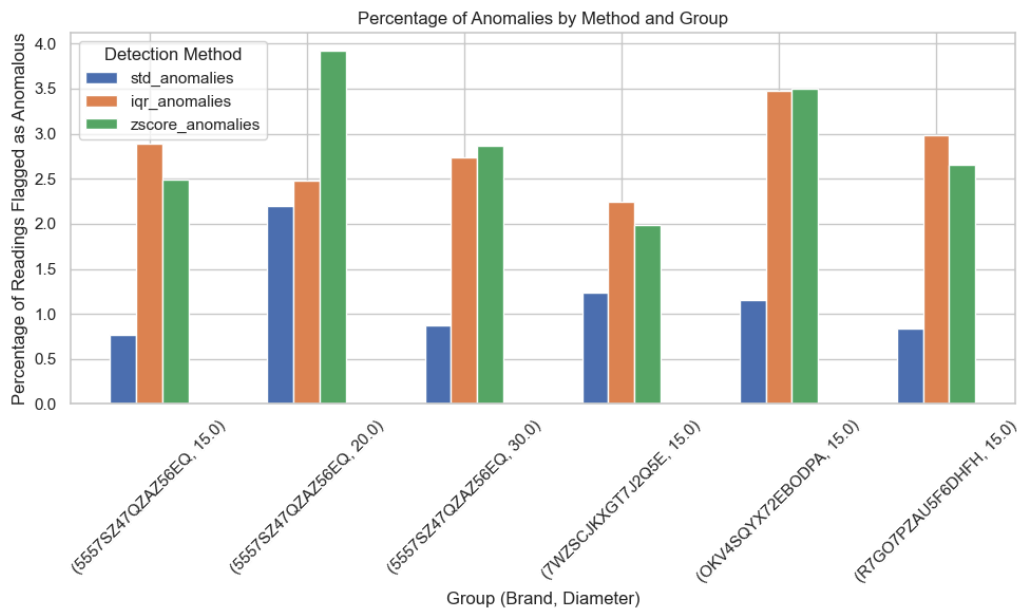


Figure 6: Percentage of Anomalies by Detection Method and Group

Results show that the IQR and MAD-based z-score methods are more sensitive, identifying approximately 2–4% of readings as anomalous, whereas the standard deviation approach is more conservative, flagging less than 1% in most cases. The **5557SZ47QZAZ56EQ (20 mm)** meter group showed the highest anomaly rates across all methods, confirming its systematically deviant behavior. The consistency of results across detection methods supports the robustness of the approach and indicates that the anomalies reflect genuine irregularities rather than methodological bias.

4.2 Seasonal Distribution of Anomalies

Analysis of monthly anomaly percentages showed that rates remain relatively stable throughout the year for most meter types, suggesting minimal seasonal influence. However, the **5557SZ47QZAZ56EQ** brand, particularly the 20 mm variant, consistently exhibited higher anomaly rates exceeding 3–4%, marking it as a recurrent outlier.

Other brands maintained uniformly low anomaly rates, typically below 1.5%, though a mild summer increase was observed in some groups, likely due to greater consumption variability. Overall, these findings confirm that **meter characteristics**, rather than seasonality, are the dominant factor influencing anomaly frequency.

The following heatmap illustrates the monthly percentage of anomalous readings across brand-diameter combinations.

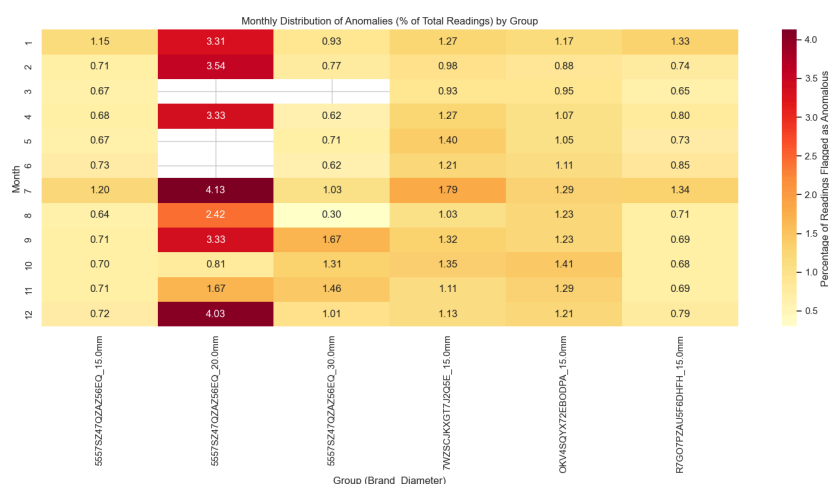


Figure 7: Monthly Distribution of Anomalies (% of Total Readings) by Group

4.3 Anomaly Rates by Brand and Diameter

Normalized anomaly rates were computed to account for different data volumes per brand and diameter.

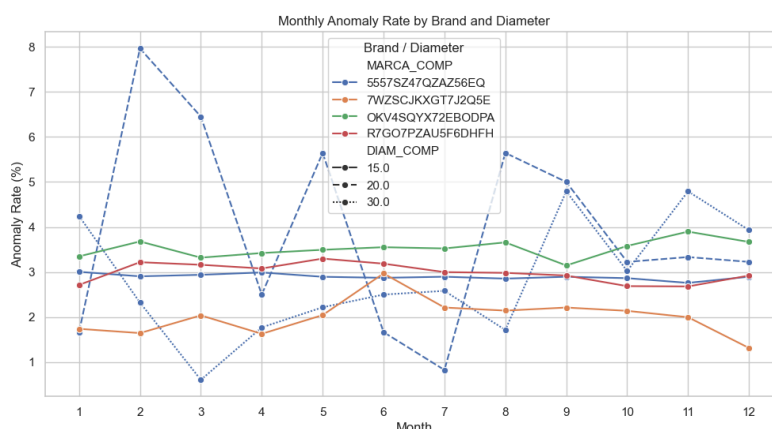


Figure 8: Monthly Anomaly Rate by Brand and Diameter

After normalization for differing group sizes, the results revealed that 15 mm meters tend to show higher relative anomaly rates and greater seasonal variation, while 20–30 mm meters are more stable over time. Among brands, **5557SZ47QZAZ56EQ (15 mm)** consistently demonstrated the highest anomaly rate, reinforcing its reputation as the most variable group.

4.4 Anomaly Rates by Model and Diameter

To capture finer product-level differences, anomaly rates were also aggregated by model within each diameter.

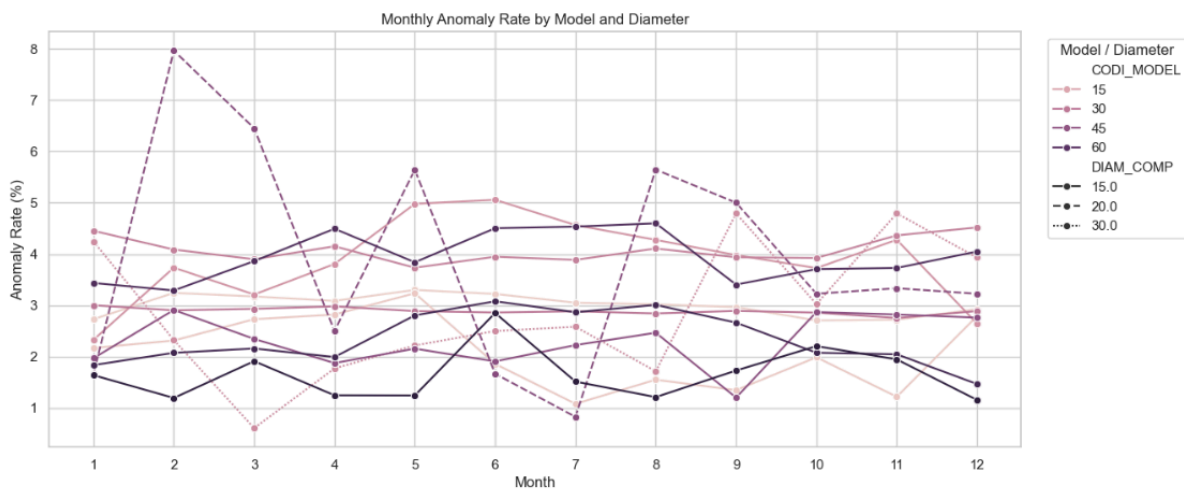


Figure 9: Monthly Anomaly Rate by Model and Diameter

When analyzing anomalies at the model level within each diameter group, additional variability emerged. Certain models within the same diameter class displayed distinct patterns in both average anomaly rate and seasonal behavior. These differences may stem from manufacturing, calibration or firmware characteristics. Clustering the models by diameter revealed blocks of similar behavior. For instance, 15 mm models commonly exhibited higher summer anomaly rates, indicating that model-specific effects are non-negligible.

4.5 Temporal Patterns

Monthly trends indicate that anomalies are not randomly distributed throughout the year. There are slight increases during the summer months (June to August) and occasionally in January, reflecting seasonal shifts in water usage such as irrigation or temperature-related effects. Nonetheless, the overall anomaly trend remains stable, suggesting that the deviations observed are systematic rather than stochastic noise.

5. Interpretation and Feature Implications

The findings can be summarized in terms of feature relevance for modeling:

- **Meter characteristics** (**DIAM_COMP**, **MARCA_COMP**, **CODI_MODEL**) exert a strong influence on both consumption levels and anomaly rates, making them critical inputs for predictive modeling.
- **Temporal features** (month, seasonal aggregates) demonstrate moderate relevance, capturing periodic consumption and anomaly fluctuations.
- **Derived statistical metrics** (rolling mean, IQR, MAD, percentage of zero readings) effectively summarize meter-specific behavior and support robust anomaly thresholds.
- **External weather variables** (tavg, prcp) show weak short-term correlations and should either be excluded or aggregated at coarser temporal resolutions (e.g., weekly or monthly averages).

Future feature engineering should therefore prioritize **meter characteristics**, **temporal variables** and **robust statistical metrics**, while considering weather data only as a potential long-term contextual factor.

6. Conclusions

Overall, the analysis confirms that consumption data are highly right-skewed and zero-inflated, primarily due to the dominance of one brand. Applying a log transformation enhances interpretability and reveals meaningful underlying structure. Diameter has a major influence on consumption behavior—15 mm meters show greater variability and anomaly rates, while 20–30 mm meters remain more stable.

Brand and model heterogeneity is significant; certain models consistently display higher anomaly frequencies even under comparable usage conditions. Most anomalies correspond to high-consumption spikes, with relatively few low-value outliers aside from zero-heavy meters.

Seasonal fluctuations are evident but modest and weather variables have minimal direct impact on short-term consumption anomalies. Consequently, predictive modeling should emphasize meter attributes, temporal context and robust statistical indicators rather than external environmental variables.