1.1.1.b Validation Report

1. Introduction

This report summarizes the quality assessment and validation of the dataset used in our project *"Incidències en Comptadors Intel·ligents"*, conducted for **Aigües de Barcelona**. The goal of this validation exercise was to ensure that the data used for subsequent analysis and modeling is accurate, consistent and complete. Reliable data quality is essential for building meaningful statistical insights and trustworthy machine learning models capable of detecting anomalies and patterns in smart water meter consumption.

2. Dataset Description

The validated dataset originates from the internal systems of Aigües de Barcelona, containing anonymized consumption records from smart water meters. The raw data was received in *Parquet* format under the filename *Dades_Comptadors_anonymized.parquet*. It comprises several million individual observations, each representing the daily consumption associated with a unique water supply contract.

Each record corresponds to a specific customer-meter pair (POLIZA_SUMINISTRO) and date (FECHA). The dataset includes both numeric and categorical attributes describing consumption levels, customer location and operational metadata. Table 1 summarizes the principal fields.

POLIZA_SUMINISTRO	The unique supply contract or service code identifying the customer's water connection.
FECHA	The date of the consumption record or measurement.
CONSUMO_REAL	The actual water consumption (typically in liters or cubic meters) recorded for that date.
SECCIO_CENSAL	The census section or geographic code representing the customer's location.
US_AIGUA_GEST	The water usage management code indicating the type of user (e.g., domestic, commercial, industrial).
NUM_MUN_SGAB	A municipal identifier code related to the local water management authority.
NUM_DTE_MUNI	The department or district number within the municipality.
NUM_COMPLET	The complete meter or installation identifier code for the connection.
DATA_INST_COMP	The date the water meter or measuring device was installed.
MARCA_COMP	The brand or manufacturer of the installed water meter.
CODI_MODEL	The model code of the water meter.
DIAM_COMP	The diameter of the water meter (in millimeters), indicating its capacity or flow range.[1]

[1] ***Table 1: Table displaying the fields and meaning of input data object***

3. Data Loading

The validation process began with the loading of the smart water meter consumption data (meter_data). The dataset contained 12 columns and approximately 17 million records, confirming a large-scale dataset suitable for statistical analysis and machine learning. A schema check confirmed that all expected columns were present and correctly named according to the data specification, with no unexpected or missing fields.

During this step, the data types of each column were carefully reviewed. The initial inspection revealed that both date variables — FECHA (consumption date) and DATA_INST_COMP (installation date) — were stored as object types. These were correctly converted to datetime64 format to enable proper temporal operations such as filtering, time-based grouping, and joining with external weather data. This conversion also ensured that chronological integrity could be verified throughout the dataset.

The categorical identifiers (POLIZA_SUMINISTRO, US_AIGUA_GEST, NUM_COMPLET, MARCA_COMP) were stored as *object* types and were inspected for structural uniformity. No irregular encodings, extraneous spaces, or invalid characters were detected, confirming that these fields were formatted consistently. The numeric attributes (CONSUMO_REAL, SECCIO_CENSAL) were correctly typed as *int64* or *float64*, suitable for analytical and modeling operations. However, while the variables (NUM_MUN_SGAB, NUM_DTE_MUNI, CODI_MODEL, DIAM_COMP) are stored as numeric fields, their interpretation here is inherently categorical. Although represented as numbers, these values denote discrete classes rather than measurable quantities along a scale. For instance, diameter values such as 15 mm or 100 mm correspond to distinct meter categories, each associated with specific operational characteristics, rather than reflecting proportional magnitude differences. Therefore, they were changed from numeric to categoric.

Finally, the target variable CONSUMO_REAL (actual water consumption) was verified to contain valid numeric entries with no non-numeric contamination or null values.

4. Data Cleaning

The second stage of validation addressed the completeness and integrity of the dataset. Missingness analysis revealed that three columns — POLIZA_SUMINISTRO, FECHA, and CONSUMO_REAL — contained no missing values, while the remaining nine columns (SECCIO_CENSAL, US_AIGUA_GEST, NUM_MUN_SGAB, NUM_DTE_MUNI, NUM_COMPLET, DATA_INST_COMP, MARCA_COMP, CODI_MODEL, and DIAM_COMP) exhibited substantial missingness. The pattern of missing values was highly consistent across these variables, suggesting that the incomplete records corresponded to instances where the meter had not yet been installed.

To validate this hypothesis, the cleaning process applied a temporal filter ensuring that only readings with a measurement date (FECHA) greater than or equal to the installation date (DATA_INST_COMP) were retained. This operation effectively removed the majority of missing values, resulting in a dataset with complete operational data for all retained records.

Additional integrity checks were performed to ensure there were no duplicate records per contract and date. Where multiple records existed for the same (POLIZA_SUMINISTRO) and (FECHA) , attributes were found to be identical. These were therefore classified as redundant record replications, and duplicates were removed under the assumption that they did not correspond to distinct consumption measurements. Similarly, the dataset was screened for invalid consumption values, such as negative readings in CONSUMO_REAL; none were found.

After these steps, the cleaned dataset exhibited a consistent structure with minimized missingness, accurate data types, and unique entries per contract and date combination. The result was a dataset that accurately reflected real operational readings and was ready for integration with external data sources.

5. Data Integration

Following cleaning and verification, the next phase involved validating the integration of the smart meter data with the weather datasets. Four separate weather files were loaded, each corresponding to a different municipality. The integration process required matching both data sources using two join keys: FECHA (date) and NUM_MUN_SGAB (municipality identifier).

Before the merge, the FECHA column in both datasets was standardized to a consistent datetime format to ensure proper alignment. A dedicated municipality code column was added to the weather data to match the format used in the meter dataset. Once aligned, the datasets were successfully merged, producing a combined file containing both consumption and weather-related variables such as temperature and precipitation.

A validation of join-key consistency confirmed that all merge operations succeeded without schema conflicts or loss of key integrity. The temporal range between datasets was also verified to ensure that both sources covered the same date intervals. Missing precipitation data were treated as valid “no rain” cases, and therefore imputed with zeros to maintain analytical coherence.

Post-integration checks confirmed that all merged records retained complete structural consistency, with no duplicated or orphaned keys. The resulting dataset offered full coverage of water consumption and corresponding weather conditions across the included municipalities, ensuring that subsequent analyses could rely on both accurate and contextually enriched information.