Data Cleaning Report

Anomalies, Missing Values, Inconsistencies:

Missing Values

- **Energy Usage Datasets**:
  - **Electricity, Gas, Hotwater, Irrigation, Solar, Steam, Water**:
    - ➢ A portion of meter reading entries were missing. We filled these using linear interpolation within each building_id group based on surrounding timestamps, followed by forward and backward filling to handle any remaining missing values.
    - ➢ This method was chosen to maintain the trends of the meter readings with realistic implied values, as opposed to filling them with 0s or removing them entirely.
- **Metadata**:
  - We filled missing values in this dataset with "Unknown" to ensure all buildings could be categorized.
- **Weather**:
  - We filled these missing values with the mean airTemperature across the dataset.

Anomalies

- Replaced negative meter readings with 0s
- Capped extreme outliers at 5 standard deviations above mean
- Invalid values in numeric columns (e.g., yearbuilt, sqm) were coerced to NaN during conversion to numeric types and then filled with the median value for that column.

Inconsistencies

We checked for inconsistencies that could cause errors during analysis.

- **Timestamp Formats**:
  - Timestamps in the energy datasets were standardized to UTC format using a consistent format (%Y-%m-%d %H:%M:%S).
  - Timestamps in the weather and metadata datasets were similarly standardized to UTC format.

Raw Data Issues Summary Report

This report summarizes the percentage of missing and negative values for energy datasets (meter readings) and missing values for non-energy datasets (metadata and weather). For energy datasets, metrics are calculated across building columns (excluding 'timestamp').

Summary Table

Dataset	Total Readings	Missing Values (%)	Negative Values (%)
chilledwater.csv	9,736,920	6.95	0.0
electricity.csv	27,684,432	4.74	0.0
gas.csv	3,105,288	3.34	0.0
hotwater.csv	3,245,640	6.20	0.0
solar.csv	87,720	20.13	0.0
irrigation.csv	649,128	10.70	0.0
steam.csv	6,491,280	11.26	0.0
water.csv	2,561,424	5.78	0.0
metadata.csv	52,352	53.81	N/A
weather.csv	3,311,660	19.71	N/A

Notes:

- **Energy Datasets**: Metrics are calculated across building columns (excluding 'timestamp').

- **Non-Energy Datasets**: Only missing values are reported, as negative values are not applicable.

- Total Readings: Number of meter readings (rows building columns) for energy datasets, or total cells for non-energy datasets.

- Missing Values: Percentage of NaN values.

- Negative Values: Percentage of readings < 0.

- Datasets were processed using Dask for large files and pandas for smaller ones.

Cleaned Data Issues Summary Report

This report summarizes the vastly improved percentage of missing and negative values for cleaned energy datasets (melted format: timestamp, building_id, value) and missing values for cleaned non-energy datasets (metadata and weather).

Summary Table

Dataset	Total Readings	Missing Values (%)	Negative Values (%)
cleaned_electricity.csv	27,684,432	0.13	0.0
cleaned_gas.csv	3,105,288	0.0	0.0
cleaned_hotwater.csv	3,245,640	0.0	0.0
cleaned_solar.csv	87,720	20.0	0.0
cleaned_irrigation.csv	649,128	0.0	0.0
cleaned_steam.csv	6,491,280	5.14	0.0

cleaned_water.csv	2,561,424	1.37	0.0
cleaned_metadata.csv	29,448	0.76	N/A
cleaned_weather.csv	3,311,660	0.0	N/A

Notes

- **Energy Datasets**: Metrics are calculated for the 'value' column in melted format (timestamp, building_id, value). Electricity data combines cleaned_electricity-0.csv and cleaned_electricity-1.csv.

- **Non-Energy Datasets**: Only missing values are reported, as negative values are not applicable.

- Total Readings: Number of rows in the 'value' column for energy datasets, or total cells for non-energy datasets.

- Missing Values: Percentage of NaN values in the 'value' column.

- Negative Values: Percentage of 'value' < 0.

- Datasets were processed using Dask for large files and pandas for smaller ones.

Data Preparation Summary

- **Folder Setup**: Verified the existence and write permissions of the "building data" folder in the Downloads directory.

- **Utility Data Cleaning** (electricity, gas, hotwater, irrigation, solar, steam, water):
    - Loaded large datasets using **Dask** for efficient processing.
    - **Standardized timestamps** to UTC.
    - **Reshaped** datasets to long format (building ID, timestamp, value).
    - **Handled anomalies**:

        - Replaced negative values with 0.
        - Capped extreme outliers at 5 standard deviations from the building-specific mean.
    - **Interpolated missing values** within each building ID group using linear interpolation.
    - **Saved** cleaned datasets in multiple CSV files prefixed with cleaned_
- **Weather Data Cleaning**:
    - Loaded the weather dataset using **Pandas**.
    - Standardized timestamps to UTC.
    - Imputed missing values in numeric columns with their **mean**.
    - Converted site IDs to string format.
    - Saved the cleaned weather dataset as `cleaned_weather.csv`.
- **Metadata Cleaning**:
    - Loaded the metadata file.
    - Converted specific columns to numeric, coercing errors to `NaN`.
    - Imputed missing values:
        - **Categorical columns**: filled with `"Unknown"`.
        - **Numeric columns**: filled with the **median**.
    - Dropped irrelevant or redundant columns.
    - Converted building ID and site ID to string format.
    - Saved the cleaned metadata dataset as `cleaned_metadata.csv`.
- **Final Step**:
    - Cleaned datasets were saved in the same folder with filenames prefixed by `cleaned_`.
    - Cleaned datasets were then utilized in energy_analysis.py to aggregate energy usage by different groupings (building, building type, hour, month) and look for significant overusers/irregular trends