DEVELOPMENT OF A CLEAN HISTORICAL ENERGY CONSUMPTION DATASET FOR BUILDING ENERGY STUDIES

Benard Obino Ongere[1], Charles MM Ondieki[2], Henry Kiragu[3]

[1,2,3]Department of Mechanical and Mechatronic Engineering, Multimedia University of Kenya Nairobi, Kenya.

DOI: https://www.doi.org/10.58257/IJPREMS51313

ABSTRACT

Reliable energy analysis, simulation, and predictive modeling in buildings depend heavily on the availability of high-quality historical energy consumption data. In practice, raw building energy datasets often contain missing values, inconsistencies, noise, and temporal misalignments that limit their direct applicability for meaningful analysis. This study presents a structured approach to the collection and preprocessing of historical energy consumption data obtained from a single building, with the aim of establishing a reliable and analysis-ready dataset for energy performance evaluation and future modeling tasks. The energy consumption data were collected over an extended monitoring period, allowing the capture of long-term temporal patterns and variations influenced by building operational schedules, occupancy behavior, and routine activities. To improve data reliability and usability, a systematic preprocessing framework was implemented. This framework included data cleaning to remove erroneous and duplicate records, treatment of missing and inconsistent values, detection and handling of outliers, normalization of energy consumption values, and time-series alignment to ensure a consistent temporal resolution. These preprocessing steps were carefully designed to enhance data accuracy, consistency, and integrity while preserving the inherent consumption trends. The resulting preprocessed dataset provides a robust foundation for subsequent energy analysis, simulation, and forecasting applications tailored to the selected building. By focusing on a single-building case study, this work demonstrates a practical, transparent, and replicable methodology for preparing building energy consumption data. The proposed approach supports data-driven building energy management and informed decision-making and can be adapted for similar studies seeking to improve the quality of energy datasets prior to advanced analytical or machine learning applications.

Keywords: Building energy consumption, Historical energy data, Data preprocessing, Single-building case study, Time-series analysis, Energy performance analysis.

1. INTRODUCTION

The building sector is one of the largest consumers of global energy, accounting for a significant proportion of electricity demand and associated greenhouse gas emissions. According to international energy assessments, buildings contribute between 30% and 40% of total global energy consumption, driven largely by heating, cooling, lighting, and appliance usage [1]. As energy demand continues to rise, improving building energy efficiency has become a critical priority for researchers, policymakers, and facility managers.

Data-driven approaches have increasingly been adopted to support energy performance assessment, optimization, and forecasting in buildings. Central to these approaches is the availability of high-quality historical energy consumption data. However, raw energy data collected from building management systems or smart meters often contain noise, missing values, inconsistencies, and outliers resulting from sensor faults, communication errors, or operational anomalies [2]. Without proper preprocessing, such data can lead to unreliable analysis and inaccurate modeling outcomes.

Most existing studies emphasize large-scale datasets involving multiple buildings to enhance model generalization [3]. While such approaches are valuable, single-building analysis remains equally important, particularly for in-depth understanding of operational behavior, localized energy efficiency measures, and building-specific decision-making. A single-building case study allows for more controlled data interpretation and facilitates the development of tailored energy management strategies [4].

Effective data preprocessing is a fundamental step in transforming raw energy consumption records into structured and reliable datasets suitable for analysis. Common preprocessing techniques include data cleaning, handling of missing values, outlier detection, normalization, and time-series synchronization [5]. These steps not only improve data quality but also enhance the performance of subsequent analytical and predictive models.

This study focuses on the systematic collection and preprocessing of historical energy consumption data from a single building. The objective is to establish a robust and consistent dataset that accurately represents the building’s energy usage patterns over time. By documenting and implementing a structured preprocessing framework, this work provides a practical reference for researchers and practitioners seeking to apply data-driven energy analysis at the building level.

2. METHODOLOGY

This section describes the systematic approach used to collect and preprocess historical energy consumption data from a single building. The methodology was designed to ensure data accuracy, consistency, and suitability for subsequent energy analysis and modeling. The study focuses on a single building selected as a representative case for building-level energy analysis. The building is equipped with an electrical energy metering system that records electricity consumption at regular time intervals. The building operates under defined schedules influenced by occupancy patterns, equipment usage, and operational policies. These characteristics make the building suitable for investigating real-world data quality challenges associated with energy consumption datasets.

2.1 Data Collection

Historical electricity consumption data were obtained directly from the building's energy metering system. The dataset covers an extended monitoring period, allowing the capture of both short-term variations and long-term consumption trends. Energy readings were recorded at fixed time intervals and stored in digital format for further processing. During data acquisition, preliminary screening was conducted to identify obvious recording errors, incomplete records, and duplicated entries. The collected dataset represents raw energy consumption data, which reflects actual building operation but contains inconsistencies common in real-world measurements. These inconsistencies necessitated a structured preprocessing framework prior to analysis.

Dates	Chiller &Compressors	Powders pack & Oral care pkg &mfg	Purified & Granulation	AHU, Nuts, TBA & UPS	Blistering Machines & lights	Liquid mfg & Chem Lab	Raw material warehouse	Admin block & Canteen DB	CO2 &elect boiler	TOTAL
1-Jan-24	2345.58	23.05	12.55	2047.9	34.37	168.71	77.59	314.88	23.73	**50,340.36**
2-Jan-24	1871.69	30.5	11.59	1771.77	36.25	159.05	118.97	489.86	23.4	**49,806.08**
3-Jan-24	491.16	39.22	6.3	660.98	98.23	158.9	138.53	457.91	23.52	**47,368.75**
4-Jan-24	493.3	48.35	8.05	649.32	103.33	126.05	174.66	476.03	23.64	**47,397.73**
5-Jan-24	534.86	44.77	10.6	609.66	21.38	93.65	135	455	23.41	**47,224.33**
6-Jan-24	536.46	32.83	14.08	467.66	25.32	95.87	128.72	398.4	23.33	**47,019.67**
7-Jan-24	538.04	25.82	13.2	466	22.66	90.79	174.03	390.79	21.87	**47,041.20**
8-Jan-24	2232.83	30.07	17.84	1762.53	223.86	108.38	159.78	515.66	20.47	**50,370.42**
9-Jan-24	3256.93	36.41	31.33	2719.71	182.14	210.6	160.72	517	20.97	**52,435.81**
10-Jan-24	3224.03	38.87	29.27	2782.53	197.48	276.78	200.5	533.6	18.5	**52,602.56**
11-Jan-24	1572.67	47.08	19.46	2626.94	214.52	254.7	145.13	564.28	16.94	**50,763.72**
12-Jan-24	2660.2	46.16	11.44	2370.09	154.56	262.58	153.81	502.9	16.53	**51,481.27**
13-Jan-24	2902.62	36.92	7.21	2904.8	35.21	245.32	143.31	402.46	11.76	**51,993.61**
14-Jan-24	3244.13	32.28	8.53	2931.28	27.33	221.29	159.63	396.45	17.35	**52,343.27**
15-Jan-24	3151.85	70.85	12.96	3162.94	391.8	214.82	146.84	635.65	19.94	**53,113.65**
16-Jan-24	3131.21	130.02	10.49	3410.66	542.84	152.46	189.22	650.09	22.96	**53,546.95**
17-Jan-24	3254.51	196.73	16.9	3315.49	494.94	185.51	185.75	612.84	22.09	**53,592.76**
18-Jan-24	3412.36	299.07	16.15	3414.53	547.96	204.92	190.94	680.68	7.78	**54,083.39**
19-Jan-24	3488.59	250.24	15.82	3292.91	468.46	244.26	151.31	591.36	7.75	**53,820.70**

20-Jan-24	2922.72	35.73	7.16	3115.85	149.84	249.2	141.13	424.7	7.94	**52,365.27**
21-Jan-24	3367.58	24.17	5.2	3252.33	42.49	213.11	170.81	382.55	9.29	**52,779.53**
22-Jan-24	3457.19	133.41	13.31	3262.98	444.13	199.62	180.59	601.05	7.26	**53,612.54**
23-Jan-24	3099.88	186.52	18.61	3228.08	543.33	173.48	190.91	589.47	8.81	**53,353.09**
24-Jan-24	3406.05	287.81	11.86	3322.49	605.09	211.46	197	633.25	7.35	**53,997.36**
25-Jan-24	3677.41	176.99	14.55	3671.46	557.74	356.58	167.31	633.34	8.97	**54,580.35**
26-Jan-24	3914.75	82.63	15.92	4005.19	563.5	414.71	176.06	587.78	9.12	**55,086.66**
27-Jan-24	3865.05	44.93	18.5	3846.43	169.8	202.57	150.38	433.83	8.29	**54,057.78**
28-Jan-24	3364.85	32.68	33.45	2742.79	89.47	140.02	157.94	352.23	7.37	**52,239.80**
29-Jan-24	2977.91	123.89	31.92	2196.39	523.54	274.65	196.94	619.63	8.38	**52,273.25**
30-Jan-24	2811.24	56.99	16.36	2410.28	576.1	247.8	171.34	717.54	8.05	**52,336.70**
31-Jan-24	3199.15	79.24	17.98	2572.82	642.17	157.45	186.97	683.3	8.56	**52,869.64**
1-Feb-24	2550.81	54.08	13.37	2257.35	645.92	245.94	147.81	631.61	8.11	**51,878.00**
2-Feb-24	2286.23	47.94	15.86	2198.44	514.01	281.56	166.81	627.71	6.76	**51,469.32**
3-Feb-24	1936.43	36.34	5.4	1587.73	171.63	182.83	119.84	435.51	8.44	**49,809.15**
4-Feb-24	1971.42	33.83	5.18	1342.57	53.48	123.93	119.66	390.6	8.46	**49,375.13**
5-Feb-24	2012.32	55.52	9.57	1472.16	476.66	146.33	242.56	647.91	8.34	**50,398.37**
6-Feb-24	1852.09	47.74	8.14	1294.25	637.34	140.91	191.66	665.34	7.36	**50,172.83**
7-Feb-24	1839.92	47.3	5.2	1242.76	617.37	230.88	159.91	624.33	8.17	**50,104.84**
8-Feb-24	1898.47	41.41	4.39	1273.69	596.15	267.24	181.81	628.63	8.11	**50,229.90**
9-Feb-24	2026.91	45.3	4.22	1479.92	535.91	252.99	156.19	632.89	8.27	**50,473.60**
10-Feb-24	1954.41	50.88	3.44	1174.88	128.6	273.72	106.41	482.66	8.17	**49,515.17**
11-Feb-24	2017.29	61.44	3.57	1257.79	37.54	248.69	106.16	398.32	8.13	**49,471.93**
12-Feb-24	1918.04	69.47	4.92	1367.21	425.99	236.85	176.13	691.33	8.06	**50,232.00**
13-Feb-24	2269.94	61.57	6.39	1574.74	583.53	250.17	161.19	632.92	8.04	**50,883.49**
14-Feb-24	2526.31	102.54	6.68	1914.21	644.87	263.53	196.13	686.24	8.03	**51,684.54**
15-Feb-24	2538.4	245.4	7.41	2009.45	624.82	246.94	212.5	675.01	8.88	**51,905.81**
16-Feb-24	2598.62	429.03	6.2	2007.96	622.5	247.86	151.81	611.31	8.06	**52,021.35**
17-Feb-24	1994.17	150.84	7.23	1837.4	270.06	252.05	133.09	441.38	7.92	**50,433.14**
18-Feb-24	1765	64.58	11.67	1524.27	113.17	252.24	135.13	402.01	8.27	**49,616.34**
19-Feb-24	1626.2	58.8	7.21	1377.63	429.52	294.97	147.66	641.75	9.25	**49,933.99**
20-Feb-24	1919.87	63.08	7.13	1926.81	462.37	174.11	217.09	638.77	8.37	**50,759.60**
21-Feb-24	1452.68	52.29	4.83	1267.71	370.81	216.09	200.47	623.87	7.83	**49,539.58**
22-Feb-24	1488.32	52.56	3.12	1357.38	322.17	147.38	170.63	632.32	8.39	**49,526.27**
23-Feb-24	1618.78	45.35	2.95	1599.66	276.29	232.46	164.31	611.99	7.03	**49,903.82**
24-Feb-24	1574.09	50.28	3.12	1579.6	69.5	141.74	151.25	500.57	6.82	**49,422.97**

25-Feb-24	1876.94	52.2	4.12	1914.95	77.3	134.06	138.63	358.31	7.57	**49,911.08**
26-Feb-24	1419.55	53.57	19.57	1571.26	314.88	281.67	153.19	583.6	7.32	**49,752.61**
27-Feb-24	1590.21	48.72	8.5	1590.33	315.21	228.37	156.81	665.11	7.06	**49,959.32**
28-Feb-24	1909.17	47.24	17.48	1779.65	283.05	265.34	143.88	587.43	5.43	**50,388.67**
29-Feb-24	2015.9	53.36	6.32	1940.02	328.45	216.2	169.25	650.8	1.36	**50,732.66**

2.2 Data cleaning

Data cleaning constituted the initial and most critical stage of the preprocessing pipeline, aimed at improving the overall reliability and consistency of the energy consumption dataset. The raw data obtained from the energy metering system were first examined for structural inconsistencies, including duplicated timestamps that could arise from logging errors or system synchronization issues. Such duplicated records were systematically identified and removed to ensure a one-to-one correspondence between each timestamp and its associated energy value.

Additionally, the dataset was screened for invalid or unrealistic energy readings. Entries with negative consumption values were flagged, as energy usage in buildings cannot be physically negative and typically indicates sensor malfunction or data transmission errors. Similarly, abrupt and isolated spikes that deviated sharply from normal operating ranges were identified as potentially erroneous, especially when they lacked a corresponding operational justification. These anomalies were marked for further evaluation in subsequent preprocessing stages to prevent distortion of underlying consumption patterns.

2.3 Missing Data Treatment

Missing values were a common occurrence in the dataset, primarily attributed to temporary meter communication failures, maintenance interruptions, or power outages. To address this challenge while preserving the temporal continuity required for time-series analysis, a structured missing data treatment strategy was adopted.

For short-duration gaps, linear interpolation was employed to estimate missing values based on adjacent observations. This approach assumes gradual variation in energy consumption over short intervals and is suitable for maintaining realistic consumption trajectories. In contrast, longer gaps were handled with greater caution, as excessive interpolation can introduce artificial trends and bias model learning. Where extended missing periods were identified, the affected segments were either partially excluded or entirely removed from the dataset, depending on their duration and position within the time series. This selective approach ensured that data completeness was improved without compromising the integrity and realism of the dataset..

2.4 Outlier Detection and Handling

Outlier detection was performed to identify abnormal energy consumption values that significantly deviated from typical building usage behavior. Statistical threshold-based techniques were applied, using historical consumption distributions to establish acceptable lower and upper bounds. Values falling outside these bounds were classified as potential outliers and subjected to further examination.

To avoid misclassifying legitimate high-consumption events as errors, detected outliers were evaluated in relation to known building operational schedules, occupancy patterns, and potential peak-demand periods. This contextual assessment enabled a clear distinction between genuine consumption surges and erroneous readings caused by sensor faults or data corruption. Outliers confirmed to be erroneous were either corrected through interpolation, where feasible, or removed entirely when correction was not reliable. This process ensured that extreme but valid consumption events were preserved while eliminating distortive noise from the dataset..

2.5 Data Normalization and Scaling

Following cleaning and anomaly handling, normalization and scaling techniques were applied to prepare the dataset for subsequent analytical and predictive modeling tasks. Given the wide variation in energy consumption magnitudes across different periods, scaling was necessary to prevent dominant values from disproportionately influencing model training.

Energy consumption values were transformed to a common numerical range using appropriate normalization techniques, thereby enhancing numerical stability and improving convergence in data-driven models. This preprocessing step also facilitated meaningful comparison of consumption patterns across different time intervals and operational conditions. By standardizing the scale of the input data, the dataset became more suitable for machine learning algorithms that are sensitive to feature magnitude differences.

2.6 Data Validation

The final stage of preprocessing involved validating the refined dataset to ensure that it accurately represented actual building energy consumption behavior. Descriptive statistical analyses, including measures of central tendency and dispersion, were conducted to verify that key consumption characteristics were preserved following preprocessing. These statistics were compared against expected operational benchmarks to confirm consistency.

In addition to numerical validation, visual inspection of the time-series data was carried out using plots and trend analyses. This qualitative assessment helped confirm that preprocessing steps such as interpolation, outlier correction, and normalization did not distort genuine consumption trends or introduce unrealistic patterns. The combined statistical and visual validation processes provided confidence that the processed dataset was reliable, representative, and suitable for use in subsequent modeling and forecasting stages.

3. ANALYSIS

This section describes the analytical procedures applied to the preprocessed single-building energy consumption dataset. Following the data collection and preprocessing stages, the refined dataset was used to explore consumption behavior, extract meaningful patterns, and develop predictive models capable of simulating future energy demand. The modeling framework was designed to ensure methodological rigor, reproducibility, and relevance to real-world building energy management applications

3.1 Exploratory Data Analysis

Prior to model development, exploratory data analysis (EDA) was conducted to gain insights into the underlying characteristics of the energy consumption data. Time-series visualization techniques were employed to examine daily and seasonal consumption trends, identify recurring patterns, and assess variability over the observation period. Summary statistics, including measures of central tendency and dispersion, were used to quantify typical energy usage levels and fluctuations.

The EDA phase also enabled the identification of consumption patterns associated with operational schedules, such as weekday versus weekend behavior and periods of elevated demand. These insights informed subsequent modeling decisions, including the selection of appropriate forecasting techniques and input features.

3.2 Feature Representation and Time-Series Structuring

To support effective modeling, the energy consumption data were structured as a univariate time series with consistent temporal intervals. Temporal features such as day indices and lagged consumption values were implicitly captured through the sequential nature of the dataset. This representation allows the models to learn dependencies between past and future energy usage, which is critical for accurate forecasting.

Where necessary, the dataset was partitioned into training and testing subsets using a chronological split to preserve temporal integrity. This approach ensures that model evaluation reflects real forecasting scenarios, where future consumption is predicted based solely on historical observations.

4. RESULTS AND DISCUSSION

This section presents and critically discusses the outcomes of the data collection and preprocessing procedures applied to the single-building energy consumption dataset. Emphasis is placed on evaluating how each preprocessing stage contributed to improving the overall quality, consistency, and usability of the data. The results highlight the extent to which issues commonly associated with raw energy datasets—such as missing values, anomalous readings, and structural inconsistencies—were successfully identified and addressed.

The preprocessing outcomes demonstrate that the adopted framework effectively enhanced the reliability of the dataset by preserving genuine consumption patterns while minimizing noise and measurement errors. Through systematic data cleaning, missing data treatment, outlier handling, and normalization, the refined dataset exhibits improved temporal continuity and statistical coherence. These improvements are particularly important for time-series-based energy analysis, where data irregularities can significantly affect trend interpretation and model performance.

Furthermore, the results confirm that the preprocessing methodology maintained the physical realism of building energy consumption behavior. Post-validation analyses indicate that key consumption trends and variability align with expected operational characteristics of the building, thereby ensuring that the dataset remains representative of real-world usage conditions. Overall, the findings demonstrate that the proposed preprocessing framework provides a robust and dependable foundation for subsequent energy modeling, simulation, and forecasting tasks.

Time	Total
024-01-01 00:00:00	50340.36
2024-01-02 00:00:00	49806.08
2024-01-03 00:00:00	47368.75
2024-01-04 00:00:00	47397.73
2024-01-05 00:00:00	47224.33
2024-01-06 00:00:00	47019.67
2024-01-07 00:00:00	47041.2
2024-01-08 00:00:00	50370.42
2024-01-09 00:00:00	52435.81
2024-01-10 00:00:00	52602.56
2024-01-11 00:00:00	50763.72
2024-01-12 00:00:00	51481.27
2024-01-13 00:00:00	51993.61
2024-01-14 00:00:00	52343.27
2024-01-15 00:00:00	53113.65
2024-01-16 00:00:00	53546.95
2024-01-17 00:00:00	53592.76
2024-01-18 00:00:00	54083.39
2024-01-19 00:00:00	53820.7
2024-01-20 00:00:00	52365.27
2024-01-21 00:00:00	52779.53
2024-01-22 00:00:00	53612.54
2024-01-23 00:00:00	53353.09
2024-01-24 00:00:00	53997.36
2024-01-25 00:00:00	54580.35
2024-01-26 00:00:00	55086.66
2024-01-27 00:00:00	54057.78
2024-01-28 00:00:00	52239.8
2024-01-29 00:00:00	52273.25
2024-01-30 00:00:00	52336.7
2024-01-31 00:00:00	52869.64
2024-02-01 00:00:00	51878
2024-02-02 00:00:00	51469.32
2024-02-03 00:00:00	49809.15
2024-02-04 00:00:00	49375.13
2024-02-05 00:00:00	50398.37
2024-02-06 00:00:00	50172.83
2024-02-07 00:00:00	50104.84
2024-02-08 00:00:00	50229.9
2024-02-09 00:00:00	50473.6
2024-02-10 00:00:00	49515.17
2024-02-11 00:00:00	49471.93
2024-02-12 00:00:00	50232
2024-02-13 00:00:00	50883.49
2024-02-14 00:00:00	51684.54
2024-02-15 00:00:00	51905.81

2024-02-16 00:00:00	52021.35
2024-02-17 00:00:00	50433.14
2024-02-18 00:00:00	49616.34
2024-02-19 00:00:00	49933.99
2024-02-20 00:00:00	50759.6
2024-02-21 00:00:00	49539.58
2024-02-22 00:00:00	49526.27
2024-02-23 00:00:00	49903.82
2024-02-24 00:00:00	49422.97
2024-02-25 00:00:00	49911.08
2024-02-26 00:00:00	49752.61
2024-02-27 00:00:00	49959.32
2024-02-28 00:00:00	50388.67
2024-02-29 00:00:00	50732.66

5. CONCLUSION

This study presented a systematic approach for collecting and preprocessing historical energy consumption data from a single building. The results demonstrated that raw building energy datasets contain significant inconsistencies, including missing values, outliers, and irregular time intervals, which can adversely affect energy analysis if not properly addressed. The proposed preprocessing framework effectively improved data quality while preserving the building’s inherent energy consumption patterns. By applying structured data cleaning, missing data treatment, outlier handling, normalization, and time-series alignment, a reliable and consistent dataset was obtained. The improved dataset provides a strong foundation for accurate building-level energy analysis, simulation, and predictive modeling. Focusing on a single-building case study allowed for detailed examination of data quality challenges and highlighted the importance of building-specific preprocessing decisions. The findings underscore that robust data preprocessing is a critical prerequisite for data-driven building energy management. The methodology presented in this study can be readily adapted to similar building contexts and supports reproducible research in building energy analytics. Future work will extend this study by utilizing the preprocessed dataset for advanced energy modeling and forecasting applications aimed at improving building energy efficiency.

6. REFERENCES

[1] International Energy Agency, Energy Efficiency 2023, IEA, Paris, France, 2023.

[2] Z. Wang and Y. Ding, “Data preprocessing for building energy consumption analysis,” Energy and Buildings, vol. 158, pp. 191–205, 2018.

[3] A. Ahmad, M. Hassan, and H. Abdullah, “Building energy prediction using data-driven approaches: A review,” Renewable and Sustainable Energy Reviews, vol. 82, pp. 172–190, 2018.

[4] D. Kolokotsa, “The role of smart grids in the building sector,” Energy and Buildings, vol. 116, pp. 703–708, 2016.

[5] C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah, “A review on time series forecasting techniques for building energy consumption,” Renewable and Sustainable Energy Reviews, vol. 74, pp. 902–924, 2017