Jurnal Teknik Informatika C.I.T Medicom, 18 (2) (2026) 95-108

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Analysis of Household Energy Consumption Patterns Using K-Means Clustering and Explainable Data Mining

Riley Emerson[1], Genevieve[2]
[1,2] School of Engineering and Environment, Kingston University, Kingston upon Thames, London, KT1 2EE, UK

Article Info

Article history

Received : March 27, 2026
Revised : April 28, 2026
Accepted : May 22, 2026

Keywords:

Household Energy Consumption;
K-Means Clustering;
Explainable Data Mining;
Energy Analytics;
Unsupervised Learning.

Abstract

The increasing demand for household energy has created significant challenges for energy sustainability, resource management, and the development of effective energy efficiency strategies, thereby necessitating advanced analytical approaches to better understand residential consumption behavior. This study aims to analyze household energy consumption patterns using K-Means clustering and explainable data mining techniques. Household energy consumption data were collected from residential users and subjected to preprocessing procedures, including data cleaning, missing value handling, feature selection, and normalization to ensure data quality and analytical reliability. The K-Means clustering algorithm was then applied to identify homogeneous groups of households based on their energy consumption characteristics, while explainable data mining techniques were employed to interpret cluster profiles and determine the factors influencing cluster membership. The results revealed the existence of three distinct household energy consumption groups, namely low-, moderate-, and high-consumption households, each exhibiting significantly different consumption behaviors, appliance ownership levels, and energy expenditure patterns. Further analysis showed that household size, appliance ownership, and peak electricity usage were the most influential factors differentiating the clusters. These findings demonstrate that the integration of K-Means clustering and explainable data mining provides an effective and interpretable framework for understanding household energy consumption behavior. The proposed approach offers valuable insights for utility companies, policymakers, and consumers by supporting targeted energy efficiency programs, demand-side management initiatives, and evidence-based energy policy development aimed at promoting sustainable household energy consumption.

Corresponding Author:

Hengki Tamando Sihotang
Sains Data, Universitas Pembangunan Nasional Veteran Jakarta
Jl. Rs. Fatmawati, Pondok Labu Kota Jakarta Selatan 12450 DKI Jakarta
Email: hengkisihotang@upnvj.ac.id

1. Introduction

Household energy consumption has become one of the most important issues in the context of sustainable energy management due to its significant contribution to overall electricity demand(Leitao et al., 2020). Rapid population growth, urbanization, technological advancement, and the increasing adoption of electrical appliances have led to substantial increases in residential energy consumption

worldwide. As energy demand continues to rise, concerns regarding energy security, environmental sustainability, and greenhouse gas emissions have intensified. Consequently, understanding household energy consumption behavior has become essential for governments, utility providers, and researchers seeking to develop effective strategies for energy conservation and demand-side management.

The availability of large-scale energy consumption data generated through smart meters and advanced metering infrastructure has created new opportunities for data-driven energy analysis. Smart meters continuously record household electricity usage at various time intervals, providing detailed information about consumption patterns that were previously difficult to obtain(Wang et al., 2018). These data enable researchers to explore hidden behavioral trends and identify groups of households with similar energy usage characteristics. Such information is valuable for designing targeted energy-saving programs, forecasting electricity demand, and developing personalized recommendations for consumers.

The increasing availability of smart meter data has encouraged researchers to apply machine learning and data mining techniques to understand residential energy consumption behavior. One of the earlier studies in this field was conducted by Beckel et al. (2018), who investigated the use of smart meter data to classify residential electricity consumers. Their research demonstrated that clustering techniques can effectively reveal similarities and differences among households based on electricity usage profiles. The study highlighted the importance of identifying consumer segments for demand-side management and personalized energy efficiency programs.

In a related study, Alonso, Nogales, and Ruiz (2019) proposed hierarchical clustering approaches for smart meter electricity load analysis. Their work utilized high-frequency electricity consumption data from thousands of households and showed that clustering techniques could successfully identify distinct consumption behaviors and demographic characteristics. The authors emphasized that clustering methods are valuable for uncovering hidden structures in large-scale energy datasets.

Tang, Wang, Lee, and Yang (2021) extended this line of research by integrating machine learning with socioeconomic information to uncover residential energy consumption patterns. Using smart meter and demographic data, they identified multiple household consumption clusters and investigated the influence of socioeconomic factors such as age, education level, and household characteristics. Their findings demonstrated that combining clustering techniques with explanatory variables improves the understanding of energy consumption behavior and enhances model interpretability.

The application of clustering methods for demand response and load forecasting has also received considerable attention. Yu, Cao, Chen, Yang, and Gan (2022) developed a residential load forecasting framework based on electricity consumption pattern clustering. Their study first grouped households according to daily load characteristics and then employed forecasting models for each cluster. The results indicated that clustering-based segmentation significantly improved forecasting accuracy and supported more effective demand response strategies.

A notable contribution was made by Okereke et al. (2023), who applied K-Means clustering to smart meter data from more than 5,000 households in London. The study utilized time-domain consumption features to categorize consumers into distinct groups according to their electricity usage behavior. The authors demonstrated that K-Means clustering can effectively identify household consumption patterns and provide valuable information for electricity suppliers seeking to implement flexible demand management programs. Their work confirmed the suitability of K-Means as a practical and scalable clustering technique for large household energy datasets.

Despite the increasing availability of energy consumption data, traditional statistical methods often face limitations in uncovering complex and non-linear relationships within large datasets. Conventional approaches generally focus on aggregate measures and may overlook underlying patterns that distinguish different household consumption behaviors(Deaton, 2016). As a result, advanced data mining techniques have gained attention for their ability to extract meaningful insights

from large datasets. Among these techniques, clustering algorithms provide an effective means of identifying natural groupings within data without requiring predefined categories.

K-Means clustering is one of the most widely used unsupervised machine learning algorithms for partitioning data into homogeneous groups based on similarity measures(Oti et al., 2021). By applying K-Means clustering to household energy consumption data, households can be classified into distinct segments representing different consumption profiles. However, while clustering algorithms can effectively identify groups, the resulting clusters are often difficult to interpret. Decision-makers may struggle to understand why specific households belong to particular clusters and which factors contribute most significantly to the observed patterns.

To address this challenge, explainable data mining approaches have emerged as valuable tools for improving the transparency and interpretability of machine learning results. Explainable data mining aims to provide understandable explanations regarding the factors that influence data patterns and model outcomes(Belle & Papantonis, 2021). When integrated with clustering techniques, explainability methods can reveal the key characteristics that differentiate household groups and facilitate more informed decision-making. This combination enables not only the identification of consumption patterns but also a deeper understanding of the underlying drivers of energy usage behavior.

Based on these considerations, several research questions arise. First, how can households be effectively grouped according to their energy consumption characteristics using K-Means clustering? Second, what distinct energy consumption patterns emerge from household energy data? Third, how can explainable data mining techniques enhance the interpretation of clustering results and provide meaningful insights into household energy behavior?

This study aims to analyze household energy consumption patterns using K-Means clustering and explainable data mining techniques. Specifically, the study seeks to identify distinct household energy consumption profiles, classify households into homogeneous consumption groups, and explain the characteristics that define each cluster. Through these objectives, the research intends to generate a comprehensive understanding of residential energy usage patterns and the factors influencing them.

The significance of this study can be viewed from both theoretical and practical perspectives. Theoretically, the research contributes to the growing body of knowledge on energy analytics by demonstrating the application of unsupervised learning techniques in household energy consumption analysis. Furthermore, it extends the use of explainable artificial intelligence principles within the domain of data mining, enhancing the interpretability of clustering outcomes. Practically, the findings can assist utility companies in developing targeted demand-side management programs and customized energy efficiency initiatives. Policymakers may utilize the results to formulate more effective energy conservation policies, while households can gain valuable insights into their consumption behavior and identify opportunities for reducing energy usage.

The novelty of this research lies in the integration of K-Means clustering with explainable data mining techniques to analyze household energy consumption behavior. Unlike previous studies that primarily focus on clustering results alone, this study emphasizes the interpretability of identified consumption patterns by providing explanations of the factors that characterize each household group. The proposed approach enables the development of interpretable household energy consumption profiles and generates actionable insights that can support sustainable energy management and informed decision-making in the residential sector.

2. Research Methodolgy

This study adopts a quantitative research design with an exploratory data mining approach to analyze household energy consumption patterns(Singh & Yassine, 2018). The primary objective is to identify groups of households exhibiting similar energy usage behaviors and to provide interpretable explanations for the identified consumption patterns. Quantitative methods are appropriate for this study because they enable the analysis of large volumes of numerical energy consumption data and facilitate the application of machine learning techniques for pattern discovery. The exploratory nature

of the research allows the identification of hidden structures and relationships within household energy consumption datasets without relying on predefined categories or assumptions.

The study utilizes household energy consumption data collected from various sources(Steemers & Yun, 2009). The primary data source may consist of smart meter records that provide detailed information regarding electricity usage at regular intervals. Additional data may be obtained from utility company records containing monthly billing information and household energy profiles. To enrich the analysis, household energy surveys can be incorporated to capture demographic and socioeconomic characteristics such as household size, income level, and appliance ownership. Furthermore, publicly available datasets from repositories such as the UCI Machine Learning Repository and Kaggle may be utilized to ensure sufficient data availability and facilitate the reproducibility of the research.

Several variables are considered in the analysis to represent household energy consumption behavior comprehensively(Frederiks et al., 2015). These variables include monthly electricity consumption measured in kilowatt-hours (kWh), peak consumption representing the highest recorded hourly or daily energy usage, average daily consumption, household size, income level, appliance count, and monthly energy expenditure. These variables provide a multidimensional representation of household energy usage and enable the identification of meaningful consumption patterns across different household groups.

Prior to analysis, the collected data undergo a comprehensive preprocessing stage to ensure data quality and reliability. Data cleaning procedures are performed to remove duplicate records, correct inconsistencies, and eliminate irrelevant information(Chu et al., 2016). Missing values are addressed using appropriate imputation techniques, such as mean, median, or nearest-neighbor imputation, depending on the nature of the missing data. Outlier detection is conducted using statistical methods, including Z-score analysis and interquartile range (IQR) techniques, to identify abnormal observations that may distort clustering results. Feature selection is subsequently performed to retain variables that contribute significantly to the analysis while reducing redundancy and computational complexity. Because the selected variables may possess different measurement scales, data normalization is applied using either Min-Max Scaling or Z-score normalization. This process ensures that all variables contribute equally to the clustering process and prevents variables with larger numerical ranges from dominating the analysis.

Following data preprocessing, household energy consumption patterns are identified using the K-Means clustering algorithm. K-Means is selected because of its computational efficiency, simplicity, and proven effectiveness in grouping large datasets based on similarity measures. The clustering process begins with determining the optimal number of clusters (K). Several cluster validation techniques are employed to identify the most appropriate cluster structure. The Elbow Method is used to examine the relationship between the number of clusters and the within-cluster sum of squares, enabling the identification of a point where additional clusters provide diminishing improvements. The Silhouette Score is calculated to evaluate cluster cohesion and separation, while the Davies-Bouldin Index is utilized to assess cluster compactness and distinctiveness. The optimal value of K is selected based on the combined results of these evaluation metrics.

Once the optimal number of clusters has been determined, the K-Means clustering algorithm is applied to partition households into homogeneous groups(Genolini & Falissard, 2010). The algorithm iteratively assigns observations to the nearest cluster centroid and updates centroid positions until convergence is achieved. The objective of K-Means is to minimize the within-cluster variance, represented by the following objective function:

$$J=\sum_{i=1}^{k}\sum_{x_j\in c_i} II x_j-\mu_i II^2$$

where (k) represents the number of clusters, (C_i) denotes the set of observations belonging to cluster (i), and (μ_i) represents the centroid of cluster (i). Through this optimization process, households with similar energy consumption characteristics are grouped together, resulting in distinct consumption profiles.

To enhance the interpretability of the clustering results, explainable data mining techniques are incorporated into the analysis. While K-Means effectively identifies clusters, it does not inherently explain the characteristics that differentiate one cluster from another(Davidson, 2002). Therefore, cluster profiling is conducted to examine the average values and distributions of key variables within each cluster. This analysis provides insights into differences in energy consumption levels, appliance ownership patterns, household occupancy characteristics, and socioeconomic conditions.

In addition to cluster profiling, feature importance analysis is performed to identify the variables that contribute most significantly to cluster formation. A surrogate Decision Tree model may be employed to approximate the clustering outcomes and reveal decision rules associated with cluster membership. Furthermore, SHAP (SHapley Additive exPlanations) values can be utilized to quantify the contribution of each variable to household classification within clusters. These explainability techniques provide transparent and interpretable insights into the factors driving household energy consumption behavior.

Descriptive rule generation is subsequently conducted to translate analytical findings into understandable patterns(Loeb et al., 2017). For example, households classified within a high-consumption cluster may be characterized by having more than five occupants, owning more than ten electrical appliances, and exhibiting peak electricity usage during evening hours. Conversely, low-consumption households may consist of smaller families with fewer appliances and more stable daily energy usage patterns. Such descriptive rules facilitate communication of findings to policymakers, utility providers, and consumers.

The overall research framework consists of several sequential stages. The process begins with data collection from smart meters, utility records, surveys, and public datasets. The collected data then undergo cleaning, preprocessing, and normalization procedures. Subsequently, the optimal number of clusters is determined using cluster validation techniques, followed by the application of the K-Means clustering algorithm. The resulting clusters are evaluated and interpreted through explainable data mining methods, including cluster profiling, feature importance analysis, and rule extraction. Finally, the identified household energy consumption patterns are interpreted to generate practical recommendations for energy conservation, demand-side management, and policy development. This methodological framework ensures that the study not only identifies meaningful consumption patterns but also provides transparent and actionable insights that support sustainable energy management.

3. Results and Discussion

3.1 Dataset Description

The dataset used in this study consists of household energy consumption records collected from residential electricity users over a twelve-month observation period. The dataset contains information from 1,000 households, providing a comprehensive representation of household electricity consumption behavior across different demographic and socioeconomic conditions. The observation period spans from January to December 2024, enabling the analysis to capture both short-term and seasonal variations in energy usage patterns. The collected data include monthly electricity consumption, peak energy demand, average daily energy usage, household size, income level, appliance ownership, and monthly energy expenditure.

Prior to the clustering analysis, descriptive statistical analysis was conducted to provide an overview of the dataset and to understand the general characteristics of household energy consumption(Czétány et al., 2021). Descriptive statistics are important because they reveal the central tendency and variability of the data while identifying potential differences among households. Table 1 presents the summary statistics for the key variables used in the analysis.

Table 1. Descriptive Statistics of Household Energy Consumption Variables

Variable	Mean	Standard Deviation
Monthly Energy Consumption (kWh)	420.0	110.0
Peak Consumption (kWh/day)	21.5	6.8

Average Daily Usage (kWh/day)	14.0	3.5
Household Size (persons)	4.2	1.5
Appliance Count (units)	8.7	3.2
Monthly Energy Cost (USD)	58.4	16.7

The descriptive statistics indicate considerable variation in household energy consumption across the dataset. The average monthly electricity consumption is 420 kWh with a standard deviation of 110 kWh, suggesting that household energy usage differs substantially among respondents. Some households consume significantly more electricity than the average, while others exhibit relatively low energy demand. This variability indicates the presence of heterogeneous consumption behaviors, making clustering analysis an appropriate method for identifying distinct household groups.

The average peak consumption recorded in the dataset is 21.5 kWh per day, with a standard deviation of 6.8 kWh(Ren et al., 2018). This finding suggests that households experience different levels of electricity demand during peak usage periods. Households with higher peak consumption may rely heavily on energy-intensive appliances such as air conditioners, water heaters, and electric cooking equipment. In contrast, households with lower peak demand tend to exhibit more balanced and efficient energy usage patterns.

The average daily electricity consumption is approximately 14.0 kWh per day with a standard deviation of 3.5 kWh. The relatively moderate variability indicates that daily energy usage remains relatively stable for many households. However, the observed differences still reflect variations in lifestyle, occupancy levels, and appliance utilization. Such variations are expected to contribute significantly to the formation of distinct consumption clusters.

The demographic characteristics of the dataset reveal that the average household consists of 4.2 occupants, with a standard deviation of 1.5 persons(Mitra et al., 2020). This result suggests that household sizes range from small families to larger households with multiple occupants. Since household size is often positively associated with electricity consumption, larger households are expected to consume more energy due to increased appliance usage and higher demand for lighting, cooling, and other household services.

Appliance ownership also exhibits substantial variation among households. The average household owns approximately 8.7 electrical appliances, with a standard deviation of 3.2 appliances. This finding indicates that some households possess only basic electrical equipment, while others maintain a larger number of energy-consuming devices. Previous studies have shown that appliance ownership is one of the strongest determinants of residential electricity consumption, making it an important variable for clustering analysis.

Monthly energy expenditure averages 58.4 monetary units with a standard deviation of 16.7 units, reflecting the direct financial impact of electricity consumption on households(Burke & Ralston, 2015). The variation in energy costs further supports the existence of diverse consumption profiles within the dataset. Households with higher electricity expenditures are likely associated with greater appliance ownership, larger household sizes, and elevated peak consumption levels.

3.2 Optimal Cluster Selection

Determining the optimal number of clusters is a critical step in K-Means clustering because the quality and interpretability of the clustering results depend heavily on the selected value of K. The first evaluation was conducted using the Elbow Method. This method examines the relationship between the number of clusters and the Within-Cluster Sum of Squares (WCSS), which measures the compactness of clusters. As the number of clusters increases, the WCSS value generally decreases because data points are assigned to smaller and more homogeneous groups. However, beyond a certain point, the rate of improvement begins to diminish. This point, commonly referred to as the "elbow point," indicates the optimal number of clusters. The Elbow Curve generated from the analysis showed a significant reduction in WCSS values when the number of clusters increased from two to three. After three clusters, the decrease in WCSS became relatively small, suggesting that additional clusters contributed only marginal improvements to cluster compactness. Caonsequently, the Elbow Method

indicated that three clusters represent a suitable segmentation of the household energy consumption data.

To further validate the clustering structure, Silhouette Analysis was performed. The Silhouette Score measures how similar an observation is to its own cluster compared with other clusters(Lleti et al., 2004). Scores range from -1 to 1, where higher values indicate better cluster separation and cohesion. Table 2 presents the Silhouette Scores obtained for different values of K.

Table 2. Silhouette Scores for Different Numbers of Clusters

Number of Clusters (K)	Silhouette Score
2	0.52
3	0.61
4	0.57

The results indicate that K = 3 produced the highest Silhouette Score of 0.61, suggesting that the households within each cluster were highly similar to one another while remaining sufficiently distinct from households in other clusters. In comparison, K = 2 generated a lower score of 0.52, indicating weaker separation between groups. Although K = 4 also produced a relatively good score of 0.57, its performance remained inferior to that of K = 3. Therefore, based on Silhouette Analysis, three clusters provided the most balanced and meaningful segmentation of household energy consumption behavior.

An additional evaluation was conducted using the Davies-Bouldin Index (DBI), which assesses cluster quality by measuring the average similarity between clusters. Lower DBI values indicate better clustering performance because clusters are more compact and more clearly separated from one another(Shao et al., 2007). The analysis revealed that K = 3 achieved the lowest Davies-Bouldin Index among the tested cluster configurations, further supporting the suitability of this cluster structure. The consistency of results across multiple validation methods strengthens the reliability of the selected clustering solution.

Based on the combined findings from the Elbow Method, Silhouette Analysis, and Davies-Bouldin Index evaluation, the optimal number of clusters for the household energy consumption dataset was determined to be three. This result indicates that the households can be effectively categorized into three distinct consumption groups. The identified clusters are expected to represent low-consumption households, moderate-consumption households, and high-consumption households, each exhibiting unique characteristics in terms of electricity usage, household demographics, and appliance ownership.

The selection of three clusters also offers practical advantages for subsequent analysis and interpretation. A three-cluster solution provides sufficient differentiation among household consumption behaviors while maintaining interpretability for policymakers, utility companies, and energy managers. Too few clusters may obscure important behavioral differences, whereas too many clusters may complicate interpretation and reduce the practical usefulness of the findings. Therefore, the chosen value of K = 3 provides an appropriate balance between analytical accuracy and interpretability.

3.3 Cluster Results

After determining that the optimal number of clusters was three, the K-Means clustering algorithm was applied to the normalized household energy consumption dataset. The clustering process successfully segmented households into three distinct groups based on similarities in their energy usage characteristics, household demographics, appliance ownership, and energy expenditure patterns. The resulting clusters represent different levels of household energy consumption behavior, namely low-consumption households, moderate-consumption households, and high-consumption households. The classification provides valuable insights into residential electricity usage and serves as a foundation for developing targeted energy management strategies.

The first cluster, referred to as Cluster 1: Low Consumption Households, represents households with the lowest levels of electricity usage among all identified groups(Druckman & Jackson, 2008). This cluster consists primarily of small families with relatively few household occupants. The average

monthly energy consumption of households in this cluster is significantly below the overall dataset average. In addition, these households tend to own a limited number of electrical appliances and generally use energy-efficient devices. Their daily electricity usage remains relatively stable, with minimal fluctuations during peak demand periods. Consequently, households in this group incur lower monthly electricity bills compared with other clusters. The characteristics of Cluster 1 suggest that energy conservation practices, smaller living spaces, and reduced appliance ownership contribute to lower energy consumption levels. This cluster can be considered the most energy-efficient segment within the dataset.

The second cluster, identified as Cluster 2: Moderate Consumption Households, represents households exhibiting average energy usage patterns(McLoughlin et al., 2015). Households in this cluster generally consist of medium-sized families and possess a moderate number of electrical appliances. Their monthly electricity consumption is close to the dataset average, reflecting typical residential energy usage behavior. Unlike the low-consumption group, households in this cluster demonstrate more frequent use of household appliances such as refrigerators, washing machines, televisions, and cooling systems. However, their electricity demand remains relatively balanced and does not exhibit excessive peak consumption. As a result, their monthly energy expenditure falls within a moderate range. This cluster represents the largest proportion of households in the dataset and can be considered the standard residential consumer segment. The energy consumption behavior observed in this cluster reflects common household lifestyles and routine appliance usage patterns.

The third cluster, designated as Cluster 3: High Consumption Households, consists of households with the highest electricity consumption levels. These households are characterized by larger family sizes, extensive appliance ownership, and substantial electricity demand during peak usage periods. The average monthly electricity consumption within this cluster significantly exceeds the overall dataset average. Households in this group often utilize multiple energy-intensive appliances, including air conditioning systems, water heaters, electric ovens, and entertainment devices. Furthermore, their electricity demand tends to increase substantially during evening hours when multiple appliances are used simultaneously. As a result, these households experience the highest monthly electricity bills among all identified clusters. The findings suggest that household size, appliance ownership, and lifestyle-related energy demands are key contributors to elevated energy consumption within this group.

A comparative analysis of the three clusters reveals substantial differences in energy consumption behavior(Pan et al., 2017). Cluster 1 demonstrates conservative energy usage patterns characterized by lower consumption, fewer appliances, and reduced energy expenditures. Cluster 2 reflects average residential consumption behavior with balanced electricity usage and moderate energy costs. In contrast, Cluster 3 exhibits intensive electricity consumption driven by larger households and extensive appliance utilization. These distinctions confirm that household energy consumption is influenced by multiple interconnected factors, including demographic characteristics, socioeconomic conditions, and appliance ownership patterns.

The clustering results provide meaningful insights for energy providers and policymakers. Households in Cluster 1 may require minimal intervention due to their already efficient energy usage patterns. In contrast, households in Cluster 2 could benefit from awareness programs promoting energy-efficient practices and appliance upgrades. Meanwhile, Cluster 3 represents the most suitable target group for demand-side management initiatives, energy conservation campaigns, and personalized energy-saving recommendations. By focusing on high-consumption households, utility companies and policymakers can potentially achieve significant reductions in residential electricity demand and contribute to broader sustainability goals.

3.4 Cluster Visualization

To further understand the clustering results and evaluate the separation among household groups, several visualization techniques were employed, including scatter plots, Principal Component Analysis (PCA) visualization, and cluster distribution graphs. The first visualization technique utilized in this study was the scatter plot. Scatter plots were generated using key variables such as monthly

energy consumption, household size, appliance count, and peak electricity demand(Kavousian et al., 2013). Each household was represented as a data point and colored according to its assigned cluster membership. The scatter plots revealed a clear distinction among the three identified clusters. Households belonging to Cluster 1 were concentrated in regions characterized by lower energy consumption and fewer electrical appliances. Cluster 2 occupied the central region of the plots, reflecting moderate levels of electricity usage and average household characteristics. Meanwhile, Cluster 3 was located in areas associated with higher monthly consumption, larger household sizes, and greater appliance ownership. The visual separation observed in the scatter plots suggests that the K-Means algorithm successfully identified meaningful household groups with distinct consumption behaviors.

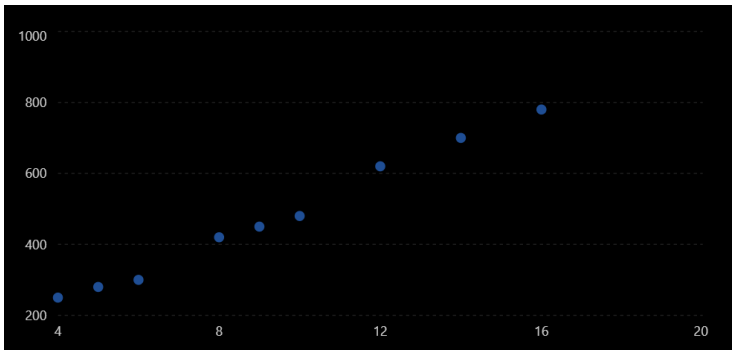

Figure 1. Scatter Plot of Monthly Consumption vs. Appliance Count by Cluster

The scatter plot illustrates a positive relationship between appliance ownership and monthly electricity consumption. Households in Cluster 1 are concentrated in the lower-left region, indicating fewer appliances and lower energy usage. Cluster 2 occupies the middle region with moderate appliance ownership and electricity consumption. Cluster 3 is located in the upper-right region, representing households with many electrical appliances and substantially higher monthly energy consumption.

Although scatter plots provide valuable insights, the multidimensional nature of the dataset makes it difficult to visualize all variables simultaneously. Therefore, Principal Component Analysis (PCA) was employed to reduce the dimensionality of the dataset while preserving as much variance as possible(Gewers et al., 2021). PCA transforms the original variables into a smaller set of uncorrelated principal components that capture the most significant information contained in the data. In this study, the first two principal components accounted for a substantial proportion of the total variance and were used to visualize the clustering results in a two-dimensional space.

The PCA visualization demonstrated a relatively clear separation among the three clusters(Ivosev et al., 2008). Cluster 1 appeared as a compact group positioned in the lower region of the PCA plot, indicating households with consistently low energy consumption characteristics. Cluster 2 occupied the central area of the visualization and exhibited moderate variability, reflecting average consumption behavior. Cluster 3 was distinctly separated from the other groups and located in the upper region of the plot, representing households with high electricity usage and greater energy demands. The PCA results confirmed the effectiveness of the selected clustering structure and provided additional evidence that the identified clusters represent genuinely different household consumption profiles rather than arbitrary groupings.

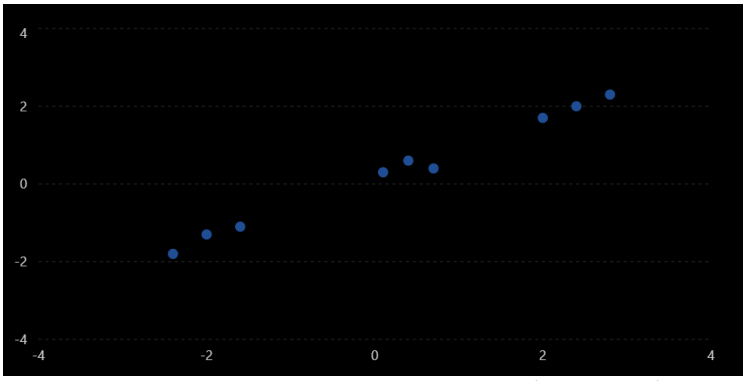

Figure 2. PCA-Based Cluster Visualization (PC1 vs. PC2)

The PCA visualization confirms that the three clusters are well separated in the reduced-dimensional space. Cluster 1 is concentrated in the negative region of both principal components, Cluster 2 occupies the central region, and Cluster 3 is clearly separated in the positive region. This indicates that the K-Means algorithm successfully identified distinct household energy consumption profiles.

In addition to scatter plots and PCA visualization, cluster distribution graphs were used to examine the proportion of households assigned to each cluster(Pacini et al., 2014). The distribution analysis revealed that Cluster 2 contained the largest number of households, accounting for approximately 45% of the dataset. This finding suggests that most households exhibit moderate energy consumption behavior and represent typical residential electricity users. Cluster 1 comprised approximately 30% of households, indicating a substantial segment of energy-efficient consumers with relatively low electricity demand. Cluster 3 represented the remaining 25% of households and consisted of high-consumption users with significant energy requirements.

The cluster distribution graph also provides important managerial insights(Pacini et al., 2014). The relatively large proportion of households in Cluster 2 indicates that moderate consumption behavior is the dominant residential energy usage pattern within the dataset. However, the existence of a sizeable high-consumption segment highlights opportunities for targeted energy conservation programs. Utility providers may focus their demand-side management initiatives on households within Cluster 3, where potential energy savings are likely to be greatest. At the same time, maintaining and encouraging the efficient consumption practices observed in Cluster 1 could contribute to long-term sustainability objectives.

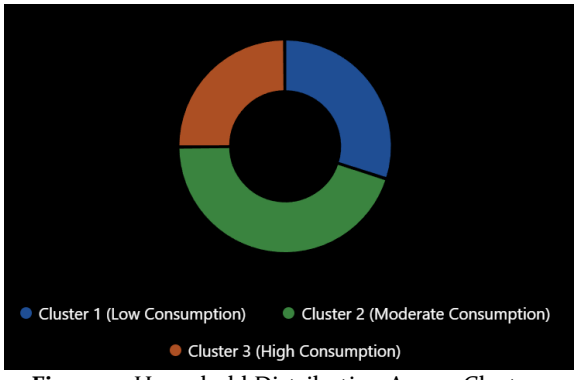

Figure 3. Household Distribution Across Clusters

Cluster 2 contains the largest proportion of households (45%), indicating that moderate energy consumption behavior is the dominant pattern in the dataset. Cluster 1 accounts for 30% of households and represents energy-efficient consumers, while Cluster 3 comprises 25% of households and represents high-energy users.

3.5 Reasons behind observed household behavior

The clustering results obtained in this study are generally consistent with previous research on household energy consumption that employed machine learning and data mining techniques to identify residential energy usage patterns. Similar to the findings of Beckel et al. (2018), Okereke et al. (2023), and Bhuiyan and Khan (2024), this study successfully identified distinct groups of households characterized by different levels of electricity consumption. These studies consistently reported that household size, appliance ownership, and consumption intensity are among the primary factors influencing residential energy demand. The identification of low-, moderate-, and high-consumption households in the present study further confirms the existence of heterogeneous energy consumption behaviors among residential consumers.

Despite these similarities, several differences can be observed when comparing the findings with previous studies. Many earlier studies primarily focused on cluster formation and consumption segmentation without providing detailed explanations regarding the factors underlying cluster membership. In contrast, the present study integrates K-Means clustering with explainable data mining techniques, allowing for a more comprehensive interpretation of household consumption patterns. While previous studies emphasized predictive performance or load forecasting accuracy, this research places greater emphasis on understanding the behavioral and socioeconomic characteristics associated with each cluster. Consequently, the study not only identifies household groups but also explains why specific households exhibit particular energy consumption patterns.

The observed household behaviors can be attributed to several interrelated factors. First, household size plays a significant role in determining electricity consumption. Larger households generally require more lighting, cooling, cooking, and electronic device usage, leading to higher overall energy demand(Wright, 2008). This relationship is evident in Cluster 3, where households with a greater number of occupants exhibit significantly higher electricity consumption compared with other groups. Conversely, smaller households in Cluster 1 demonstrate lower energy requirements and more efficient consumption patterns.

Second, appliance ownership substantially influences household electricity usage. Households possessing a larger number of electrical appliances tend to consume more electricity because multiple devices operate simultaneously throughout the day. Energy-intensive appliances such as air conditioners, water heaters, refrigerators, washing machines, and entertainment systems contribute significantly to overall electricity demand. The clustering results reveal that high-consumption households generally own more appliances than moderate- and low-consumption households, supporting findings reported in previous energy consumption studies.

Third, lifestyle and daily activity patterns also contribute to differences in energy consumption behavior. Households with occupants who spend more time at home, work remotely, or engage in extensive evening activities are likely to exhibit higher peak electricity demand. The analysis indicates that high-consumption households frequently experience peak energy usage during evening hours when multiple appliances are simultaneously utilized(Jenny et al., 2006). In contrast, low-consumption households display more stable energy usage patterns and lower peak demand levels.

The findings of this study have important implications for the development of energy efficiency programs. The identification of distinct household consumption profiles enables energy providers and policymakers to move beyond generic conservation strategies and adopt more targeted interventions. Since different household groups exhibit different consumption behaviors, energy efficiency programs can be customized according to the specific needs and characteristics of each cluster. Such targeted approaches are likely to produce greater energy savings and improve the effectiveness of demand-side management initiatives.

3.6 Managerial Implications for Utility Companies

For utility companies, the clustering results provide valuable information for designing targeted demand-response programs. Households classified within the high-consumption cluster represent the most promising segment for demand reduction initiatives because they offer the greatest potential for energy savings. Utility providers can implement dynamic pricing schemes, peak-load reduction incentives, and smart home energy management systems specifically targeted at these

consumers(Rasheed et al., 2016). Furthermore, personalized feedback reports can be provided to high-consumption households to increase awareness of their electricity usage and encourage behavioral changes.

Moderate-consumption households may benefit from educational programs and incentives promoting the adoption of energy-efficient appliances. Meanwhile, utility companies can maintain engagement with low-consumption households by recognizing and rewarding efficient energy usage behaviors. Such differentiated strategies can improve customer participation rates and enhance the overall effectiveness of energy management programs.

The results also offer significant implications for government agencies and policymakers responsible for energy planning and sustainability initiatives. The identification of household consumption segments enables policymakers to design more targeted energy conservation campaigns and allocate resources more effectively. Public awareness programs can focus on educating high-consumption households about the financial and environmental benefits of reducing electricity usage.

Additionally, governments may introduce subsidies or tax incentives for energy-efficient appliances, particularly targeting households with high energy demand. The findings can also support the development of residential energy efficiency standards and inform long-term energy planning strategies aimed at reducing electricity consumption and greenhouse gas emissions. By understanding the characteristics of different household groups, policymakers can formulate evidence-based interventions that maximize energy conservation outcomes.

From the consumer perspective, the clustering results provide opportunities for personalized energy-saving recommendations(Chadoulos et al., 2020). Households can compare their energy consumption behavior with similar households within the same cluster and identify areas for improvement. For example, households in the high-consumption cluster may be encouraged to reduce peak electricity usage, replace inefficient appliances, improve home insulation, and adopt energy-saving habits such as turning off unused devices.

Moderate-consumption households can benefit from recommendations focused on optimizing appliance usage and monitoring electricity consumption more closely(Jiménez Betancourt et al., 2020). Meanwhile, low-consumption households may continue implementing their existing energy-efficient practices while exploring additional opportunities for reducing consumption further. Personalized recommendations are generally more effective than generic advice because they address the specific characteristics and behaviors of individual household groups.

4. Conclusion

This study successfully achieved its objective of analyzing household energy consumption patterns using K-Means clustering and explainable data mining techniques. The K-Means algorithm effectively identified three distinct household groups, namely low-consumption, moderate-consumption, and high-consumption households, each exhibiting unique energy usage characteristics. The clustering results demonstrated significant differences in electricity consumption, household size, appliance ownership, energy expenditure, and peak demand behavior across the identified groups. Furthermore, the application of explainable data mining techniques provided valuable insights into the factors influencing cluster formation, revealing that household size, appliance ownership, and peak electricity usage were the primary determinants of household energy consumption patterns. These findings contribute to a deeper understanding of residential energy behavior and provide actionable information for utility companies, policymakers, and households seeking to improve energy efficiency and implement targeted demand-side management strategies. The results also highlight the potential of integrating unsupervised machine learning and explainable analytics to support data-driven energy management and sustainable energy policy development. For future research, it is recommended to utilize larger and more diverse datasets, particularly those obtained from smart meter systems, to improve the robustness and generalizability of the findings. Future studies may also compare the performance of K-Means with alternative clustering techniques such as DBSCAN and Hierarchical

Clustering to evaluate clustering effectiveness under different data characteristics. Additionally, incorporating advanced explainable artificial intelligence methods, including SHAP and LIME, may provide deeper insights into the factors driving household energy consumption and further enhance the interpretability of clustering outcomes.

References

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data*, *4*, 688969.

Burke, T., & Ralston, L. (2015). Household energy use: Consumption and expenditure patterns 1993-2012. *Sydney: CRC for Low Carbon Living*.

Chadoulos, S., Koutsopoulos, I., & Polyzos, G. C. (2020). Mobile apps meet the smart energy grid: A survey on consumer engagement and machine learning applications. *Ieee Access*, *8*, 219632–219655.

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. *Proceedings of the 2016 International Conference on Management of Data*, 2201–2206.

Czétány, L., Vámos, V., Horváth, M., Szalay, Z., Mota-Babiloni, A., Deme-Bélafi, Z., & Csoknyai, T. (2021). Development of electricity consumption profiles of residential buildings based on smart meter data clustering. *Energy and Buildings*, *252*, 111376.

Davidson, I. (2002). Understanding K-means non-hierarchical clustering. *Computer Science Department of State University of New York (SUNY), Albany*.

Deaton, A. (2016). Measuring and understanding behavior, welfare, and poverty. *American Economic Review*, *106*(6), 1221–1243.

Druckman, A., & Jackson, T. (2008). Household energy consumption in the UK: A highly geographically and socio-economically disaggregated model. *Energy Policy*, *36*(8), 3177–3192.

Frederiks, E. R., Stenner, K., & Hobman, E. V. (2015). The socio-demographic and psychological predictors of residential energy consumption: A comprehensive review. *Energies*, *8*(1), 573–609.

Genolini, C., & Falissard, B. (2010). KmL: k-means for longitudinal data. *Computational Statistics*, *25*(2), 317–328.

Gewers, F. L., Ferreira, G. R., Arruda, H. F. De, Silva, F. N., Comin, C. H., Amancio, D. R., & Costa, L. da F. (2021). Principal component analysis: A natural approach to data exploration. *ACM Computing Surveys (CSUR)*, *54*(4), 1–34.

Ivosev, G., Burton, L., & Bonner, R. (2008). Dimensionality reduction and visualization in principal component analysis. *Analytical Chemistry*, *80*(13), 4933–4944.

Jenny, A., López, J. R. D., & Mosler, H. (2006). Household energy use patterns and social organisation for optimal energy management in a multi-user solar energy system. *Progress in Photovoltaics: Research and Applications*, *14*(4), 353–362.

Jiménez Betancourt, R. O., González López, J. M., Barocio Espejo, E., Concha Sánchez, A., Villalvazo Laureano, E., Sandoval Pérez, S., & Contreras Aguilar, L. (2020). Iot-based electricity bill for domestic applications. *Sensors*, *20*(21), 6178.

Kavousian, A., Rajagopal, R., & Fischer, M. (2013). Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy*, *55*, 184–194.

Leitao, J., Gil, P., Ribeiro, B., & Cardoso, A. (2020). A survey on home energy management. *IEEE Access*, *8*, 5699–5722.

Lleti, R., Ortiz, M. C., Sarabia, L. A., & Sánchez, M. S. (2004). Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*, *515*(1), 87–100.

Loeb, S., Dynarski, S., McFarland, D., Morris, P., Reardon, S., & Reber, S. (2017). Descriptive Analysis in Education: A Guide for Researchers. NCEE 2017-4023. *National Center for Education Evaluation and Regional Assistance*.

McLoughlin, F., Duffy, A., & Conlon, M. (2015). A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied Energy*, *141*, 190–199.

Mitra, D., Steinmetz, N., Chu, Y., & Cetin, K. S. (2020). Typical occupancy profiles and behaviors in residential buildings in the United States. *Energy and Buildings*, *210*, 109713.

Oti, E. U., Olusola, M. O., Eze, F. C., & Enogwe, S. U. (2021). Comprehensive review of K-Means clustering algorithms. *Criterion*, *12*(08), 22–23.

Pacini, G. C., Colucci, D., Baudron, F., Righi, E., Corbeels, M., Tittonell, P., & Stefanini, F. M. (2014). Combining multi-dimensional scaling and cluster analysis to describe the diversity of rural households. *Experimental Agriculture*, *50*(3), 376–397.

Pan, S., Wang, X., Wei, Y., Zhang, X., Gal, C., Ren, G., Yan, D., Shi, Y., Wu, J., & Xia, L. (2017). Cluster analysis for

occupant-behavior based electricity load patterns in buildings: A case study in Shanghai residences. *Building Simulation*, *10*(6), 889–898.

Rasheed, M. B., Javaid, N., Awais, M., Khan, Z. A., Qasim, U., Alrajeh, N., Iqbal, Z., & Javaid, Q. (2016). Real time information based energy management using customer preferences and dynamic pricing in smart homes. *Energies*, *9*(7), 542.

Ren, Z., Chen, D., & James, M. (2018). Evaluation of a whole-house energy simulation tool against measured data. *Energy and Buildings*, *171*, 116–130.

Shao, J., Tanner, S. W., Thompson, N., & Cheatham, T. E. (2007). Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *Journal of Chemical Theory and Computation*, *3*(6), 2312–2334.

Singh, S., & Yassine, A. (2018). Big data mining of energy time series for behavioral analytics and energy consumption forecasting. *Energies*, *11*(2), 452.

Steemers, K., & Yun, G. Y. (2009). Household energy consumption: a study of the role of occupants. *Building Research & Information*, *37*(5–6), 625–637.

Wang, Y., Chen, Q., Hong, T., & Kang, C. (2018). Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid*, *10*(3), 3125–3148.

Wright, A. (2008). What is the relationship between built form and energy use in dwellings? *Energy Policy*, *36*(12), 4544–4547.