**Unveiling Hidden Patterns in Household Power Consumption:**

**A Clustering Analysis Approach**

(Case Study of a Residential Building in Sceaux, France)

# I. **Introduction**

As global attention on sustainable energy practices continues to intensify, it has become even more critical to understand, on a granular level, the electricity consumption patterns within individual households and the user behaviour driving these patterns.

Electricity demand all over the world is at an all-time high, and whilst new ways to bolster its production are being explored, the question remains whether it will be enough to meet future demand.

According to Connaissance des Énergies (2023), worldwide electricity consumption reached an all-time high of 27,520.5 TWh in 2021, with just ten countries accounting for almost 70% of that figure, and the trio of China, the United States and India alone consuming more than half of that.

As some studies like Zhifeng G.A, Kaile Z, et al (2018) have revealed before, the most influencing factor driving electrical consumption is user behaviour, and it is this context that prompted our investigation into the hidden consumption patterns, allowing us to comprehend the diverse drivers behind these usage behaviours.

The dataset under scrutiny comprises 2,075,259 measurements recorded at a one-minute sampling rate, spanning almost four years in a residential building in Sceaux, France. This dataset, focusing on various electrical quantities and appliances, serves as the foundation for uncovering consumption drivers and subsequently informing effective sustainability measures.

# II. **Demystifying the Problem:**

As previously mentioned, the dataset we will be working with is an energy usage dataset that records electricity consumption values in watts, obtained through a smart meter. This dataset covers a period of four years and represents the energy consumption of a single household, where energy production is based on solar panels.

The main objective of this study is to track electricity usage patterns by utilising two unsupervised machine learning techniques: K-Means and Agglomerative Hierarchical Clustering. The results obtained from both approaches will be thoroughly examined and discussed in the context of this case study.

## **Research Questions**

The study aims to answer three main research questions:

1. User Behavior Analysis: What user behaviours are observable in the dataset, and what clusters can be discerned from these patterns?
2. Temporal Patterns: How do electricity consumption patterns fluctuate over the four years, and can clusters be identified to represent distinct temporal behaviours?
3. Appliance-Specific Clusters: What patterns are associated with specific appliances, shedding light on their contributions to overall consumption?

## **Research Objectives**

The objectives of this research lie in applying advanced techniques to conduct:

1. Apply K-Means and Agglomerative Hierarchical Clustering to unveil hidden patterns in electricity consumption.
2. Explore and compare the results obtained from both clustering techniques.
3. Correlate identified clusters with temporal trends, appliance-specific behaviours, and user patterns.
4. Provide insights into the potential impact of the identified consumption patterns on sustainability practices.

## **About the Dataset**

The dataset comprises a number of key consumption variables and is delineated at a snapshot examination below

| **Variable** | **Description** | **Data Type** |
| --- | --- | --- |
| date | Represents the calendar date of each measurement, providing a temporal dimension to the dataset.  *(dd/mm/yyyy)* | Date |
| time | Represents the time of day for each measurement, allowing for the analysis of duration patterns. *(hh:mm:ss)* | Time |
| global\_active\_power | Global minute-averaged active power *(in kilowatt).*  Indicates the overall active power consumed. | Numeric |
| global\_reactive\_power | Global minute-averaged reactive power *(in kilowatt).*  Represents the reactive power consumed by the household, contributing to the overall power demand. | Numeric |
| voltage | Minute-averaged voltage *(in volts)*  Provides information about the voltage level, a crucial factor for electrical devices. | Numeric |
| global\_intensity | Global minute-averaged current intensity *(in ampere)*  Indicates the current intensity, reflecting the strength of the electrical current. | Numeric |
| sub\_metering\_1 | Sub metering No. 1  *(in watt-hour of active energy)*  Corresponds to the kitchen, measuring energy consumption from appliances like dishwasher, oven, and microwave. | Numeric |
| sub\_metering\_2 | Sub metering No. 2  *(in watt-hour of active energy)*  Corresponds to the laundry room, measuring energy consumption from appliances like washing machine, tumble-drier, refrigerator, and light | Numeric |
| sub\_metering\_3 | Sub-metering No. 3  *(in watt-hour of active energy)*  Corresponds to an electric water heater and an air conditioner, measuring energy consumption in these areas. | Numeric |

**Additional Notes:**

* A derived variable, ‘Unaccounted\_Use’ was created with the formula *`(global\_active\_power \* 1000 / 60 - sub\_metering\_1 - sub\_metering\_2 - sub\_metering\_3)`*calculating the active energy consumed every minute by electrical equipment not measured in sub-meterings 1, 2, and 3.
* For clarity, the submetering readings were renamed to *Kitchen\_Use, Laundry\_Use and Heating\_Use.*

The dataset was obtained from the UCI Machine Learning Repository. It is publicly accessible and was originally analysed in the work of [Moro et al., 2014]. The link to the dataset has been added to the Appendix section.

# III. **Literature Review**

In the exploration of the intricate relationship between household dynamics and electricity consumption, a spectrum of research has unfolded, revealing nuanced insights into the factors that influence energy use. Jones, Fuertes, and Lomas (2015) illuminated a correlation between household income and energy consumption, unveiling a tendency for higher income to be associated with increased energy usage. In supporting this, the study by Cayla, Maizi, and Marchand (2011) suggested that households with lower incomes face constraints, impacting their energy consumption patterns.

A pivotal factor in this intricate web of influences is the number of occupants in a household, as articulated by Gram-Hanssen (2004). His research posited that the number of occupants emerged as the most significant predictor of electricity consumption, intertwining with household income and apartment size’ to collectively explain most of the variance in electricity consumption. Supporting this notion, Zhou and Teng (2013) uncovered a quantifiable relationship, indicating an 8% increase in household electricity consumption in China for each additional occupant, a trend similarly observed in Indian households by Tiwari (2000).

The advent of smart metres marked a transformative shift in the landscape of energy consumption research. Preceding this technology, researchers heavily relied on subjective data obtained through traditional metres, household interviews, and surveys. However, the introduction of smart metres ushered in a new era, enabling the recording of half-hourly consumption data and automatic metre readings. The UK Department for Energy Security and Net Zero (2021) underscores the profound impact of smart metres, empowering consumers to meticulously track and manage their energy use, fostering both cost savings and emissions reduction.

With this evolution, researchers gained access to objective data that became fertile ground for various data mining and unsupervised machine learning techniques, including clustering analysis.

**Clustering with K-Means Algorithm**

Clustering, according to Han, J. and Kamber, M (2012) is a process of partitioning data objects into groups, or clusters based on their unique semantics and characteristics, so that the objects within a cluster are similar to one another and dissimilar from the objects in other clusters.

Among various clustering techniques, K-means, developed by MacQueen, is the most widely used, and this is mostly because it can quickly and efficiently cluster large amounts of data, including outliers.

The work by Flath, Nicolay, Conte, and van Dinther (2012) stands as a testament to the power of K Means. Employing K-means, they dissected consumption patterns by day and week, unveiling nine distinctive clusters with unique load profiles. Notably, their findings uncovered the high residential electricity usage during winter, whilst also revealing variations in consumption patterns during weekdays versus weekends.

K-means in its approach, uses Euclidean distance formula to find the correlation between two objects:



where *pi* and *ci* are the attributes of a given object, and i varies from 1 to n.

The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided and the data points are clustered based on feature similarity (Jessica Temporal, 2019).

The algorithm has three steps, which starts by giving the algorithm a set of data points and a value of K as a number of clusters, then randomly selects K centres of clusters, and eventually calculates the distance between each data point and the centres of the clusters K by using one of the distance measurements. It then finally assigns each data point to the closest centre and loops iteratively until convergence.

The k-Means algorithm was chosen because it is not affected by the order of objects. When using K-Means to cluster numeric data, it will be easier to find outliers and noise. This advantage is well in accordance with the characteristics of this electricity dataset, in which the data is numerical and the size of the dataset is quite big.

**Global Active Energy vs Global Reactive Energy**

According to Samaneh Zamanloo et al (2021), Global Active Power represents the actual energy consumed by electrical devices and is usually measured in kilowatts (kW) or MW. It is the power manifested in various physical forms such as electromagnetic, mechanical or acoustic waves and is symbolised by the letter “P”.

Reactive power, on the other hand, is simply the amount of power that continuously oscillates between the electrical grid (source) and the building ( load), basically representing the power which cannot be used. Reactive power is mostly used to overcome and control fluctuations in voltage levels. It is denoted by the “Q”.



# IV. **Methodology and Implementation**



As illustrated in the above “Fig. 1”, the proposed approach and implementation for this study is cascaded over 8 distinct phases as follows:

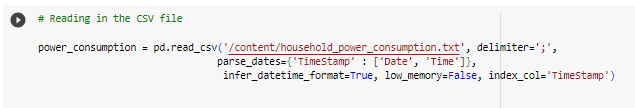
1. Data Cleaning and Preprocessing
2. Outlier detection and removal
3. Exploratory analysis and visualisation
4. Standardisation and feature selection
5. Algorithm 1: K Means Clustering
6. Algorithm 2: Hierarchical Clustering
7. Evaluation and Comparison
8. Interpretation and Recommendation

## Importation of Required Libraries

To get started, we import the necessary libraries in Python required to conduct a comprehensive clustering analysis.

## Dataset Preprocessing

The first phase involved loading the dataset and performing a number of important cleaning commands to preprocess for use.

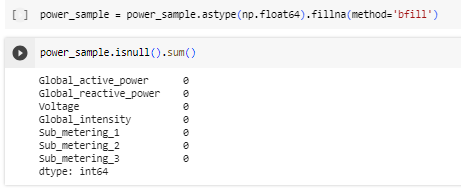
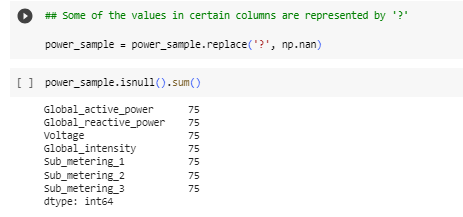


The original dataset containing 2,075, 259 rows was imported, with the 'Date' and 'Time' columns merged into a single DateTimeIndex named 'TimeStamp', serving as the temporal anchor for our analysis. Subsequently, 10,000 records were randomly sampled from the main dataset, enhancing computational efficiency and focusing the analysis on a manageable subset.



### **Handling Missing Values**:

Upon inspecting the sampled data, it was discovered that a certain number of rows (75) contained missing values represented by the character *'?'*. These characters were subsequently transformed and replaced with NaN values. In order to avoid the impact of missing values on the clustering results, the values were then filled using a forward fill method('bfill') and the data type changed from *object* to *float64*.



### Feature Engineering

The first step entailed renaming the sub-metering columns to provide a more intuitive representation of energy usage and interpretability of the dataset.

Subsequently, the challenge was to fashion additional informative features that could demonstrate the seasonality in the data as well as the unaccounted electricity consumed within the house.



Below are the newly derived columns:

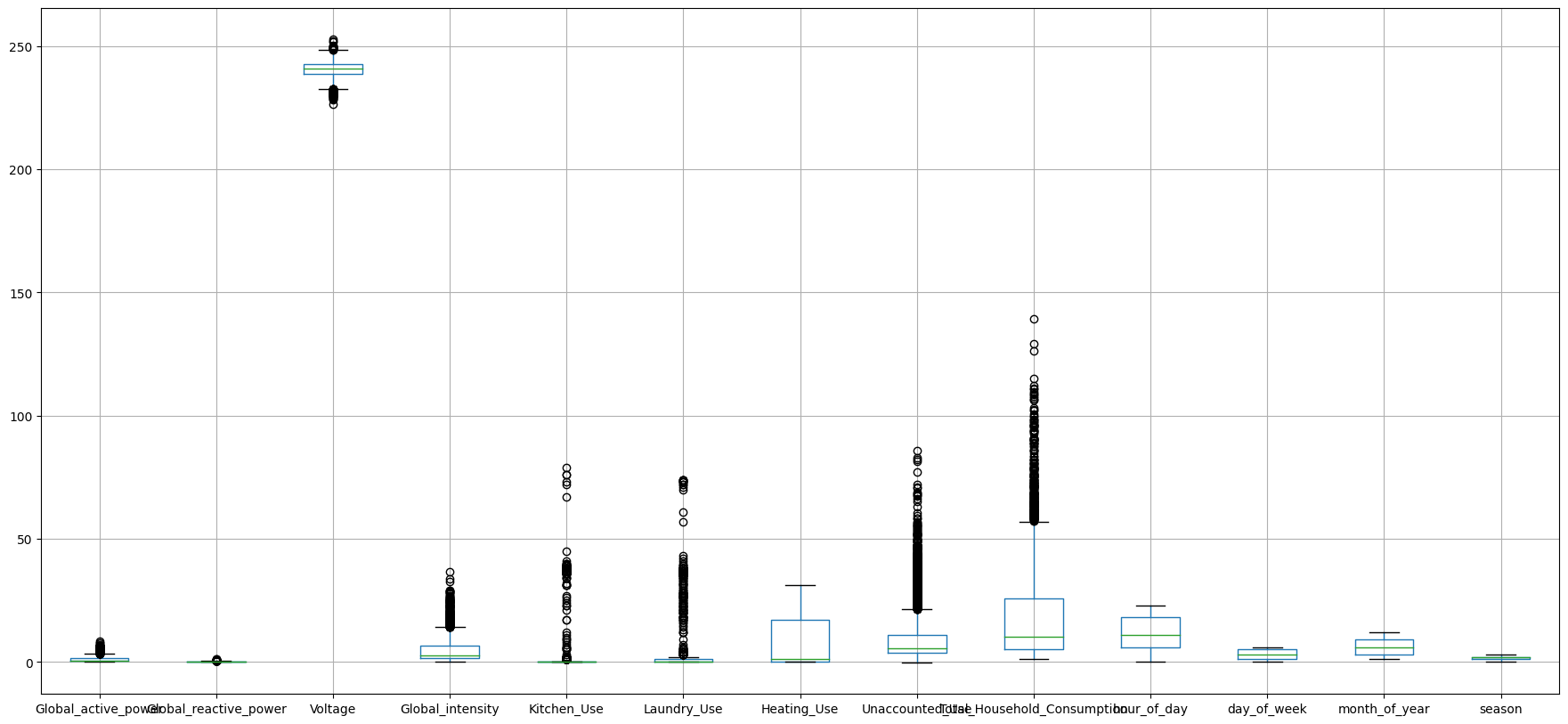
***Unaccounted\_Use:***Representing the active energy consumption not measured by sub-meterings 1, 2, and 3. This variable captures energy usage from unidentified sources within the house..

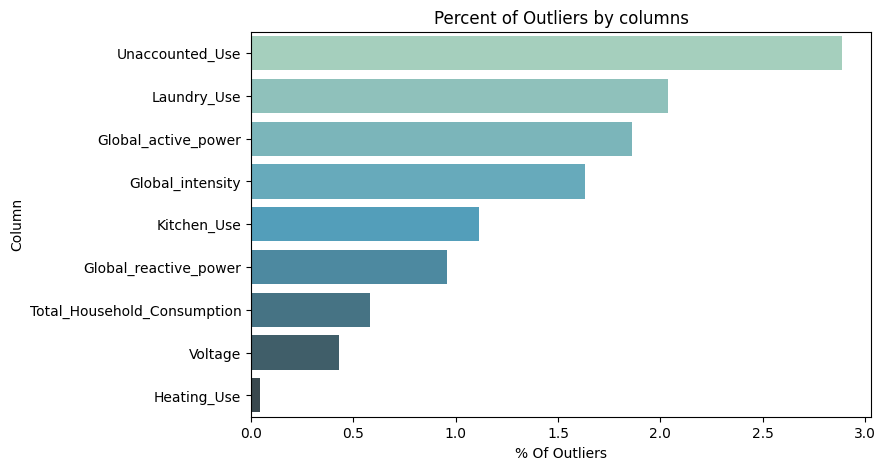
***Total\_Household\_Consumption*:** This combined all four energy readings to provide a holistic view of the household's total energy consumption. This variable sums up kitchen, laundry, heating, and unaccounted energy usage.

**Temporal variables:** New variables like h*our\_of\_day, day\_of\_week, month\_of\_year, and season* were also created. These variables capture temporal patterns by extracting the hour of the day, day of the week, month of the year, and season from the timestamp index. They enable the analysis of energy consumption variations across different time intervals.

## Outlier detection and removal

Recognizing the damaging impact of outliers, we ran a series of commands to identify the presence of outliers in the data using boxplots and their respective locations. Outliers are basically extreme values within each variable which could prove damaging if not handled (Hodge V H and Austin Jim, 2004).



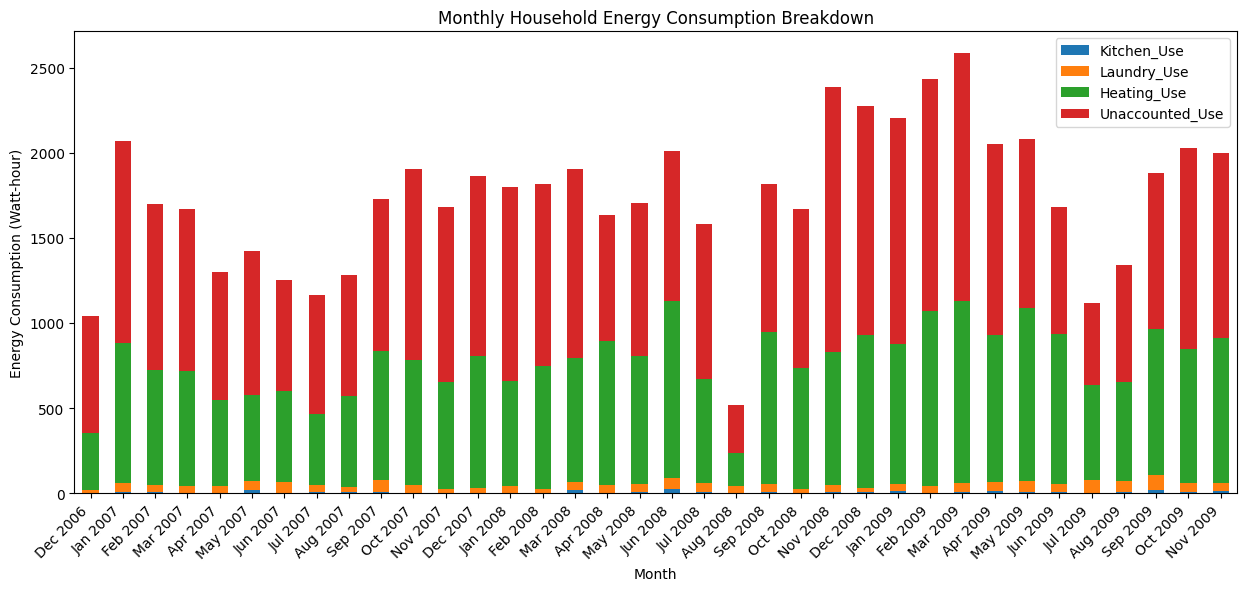


The Python commands iterated over selected numerical features, calculating their mean and standard deviation, and establishing a threshold for identifying outliers. The observations beyond three standard deviations from the mean were then considered outliers and subsequently removed.



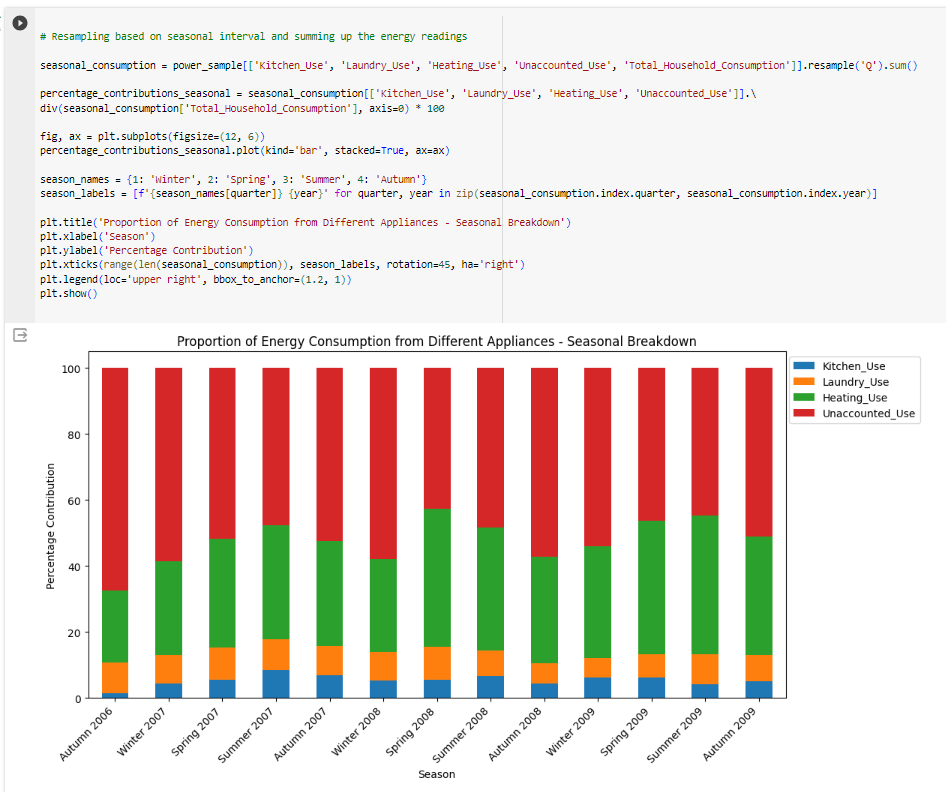
## Exploratory Data Analysis (EDA)

Considered one of the most critical phases in this study, the EDA is central to conducting a clear decomposition of the key components in the data using visualization techniques.



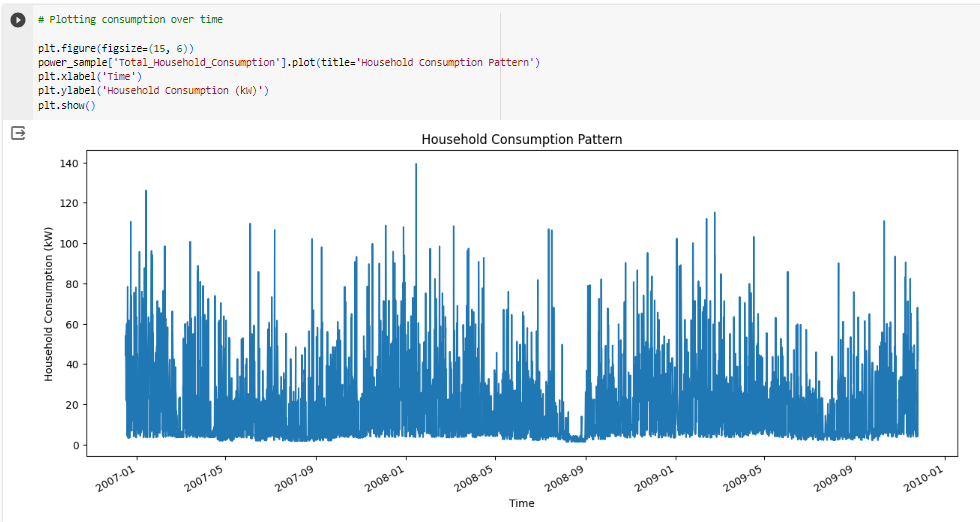
For instance, the figure above aided a clear representation of how different energy readings contribute to the total household energy consumption on a monthly basis. This can be considered a pattern in the forming, as it clearly shows the variations in energy usage across the specified categories.

To further explore the seasonal dynamics around consumption, the dataset was resampled at a quarterly frequency and the percentage contributions of the different appliances to the total household consumption are visualized using a stacked bar plot.

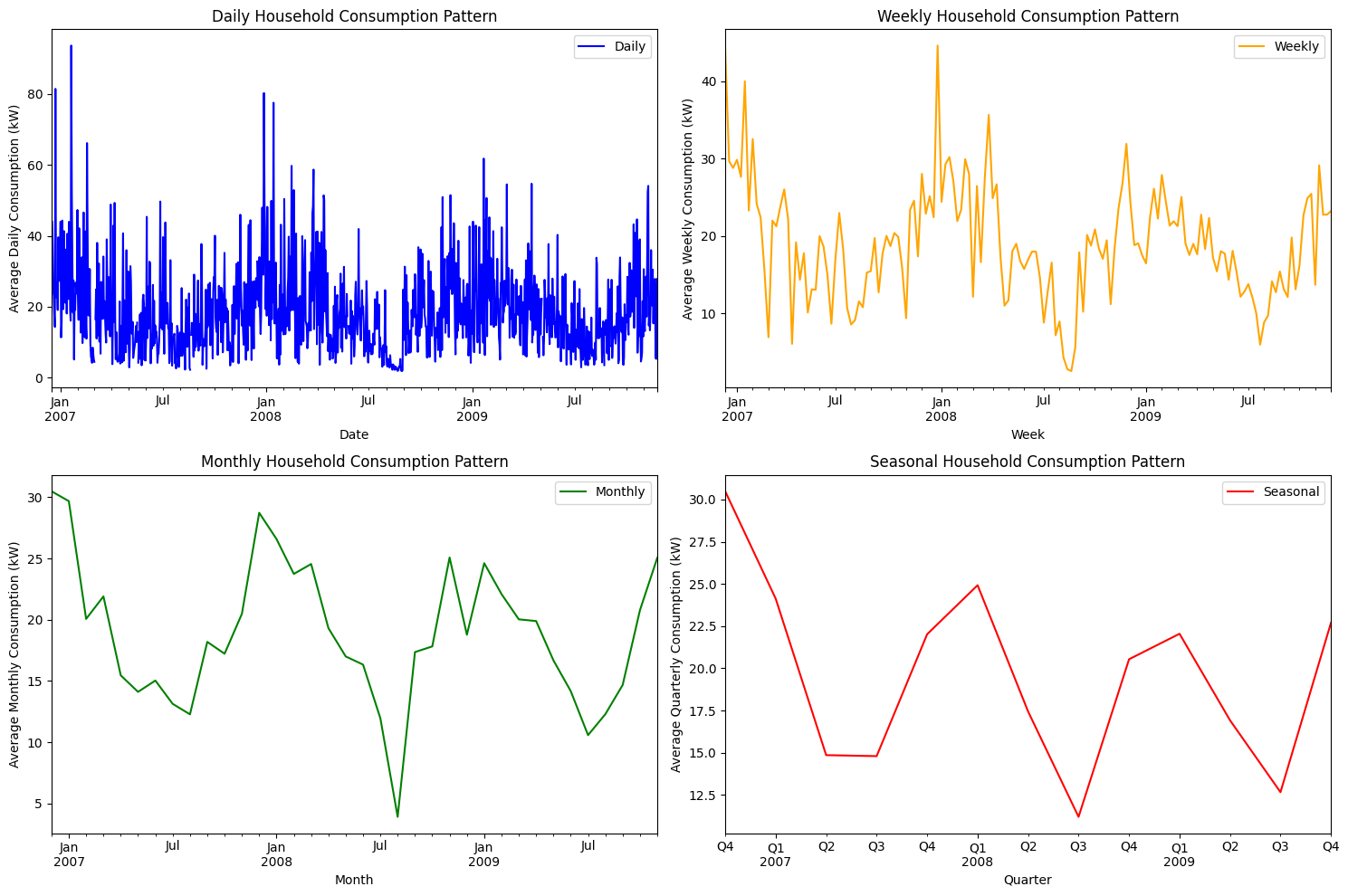


Clearly, we can observe a strong pattern, indicating a large volume of energy being used for unaccounted and yet unknown appliances.

We also further decomposed our dataset to show patterns at different time intervals. Visualized using a line plot, we can observe the overall pattern of household energy consumption over time.



The dataset was then resampled to calculate the mean consumption for different periods: daily, weekly, monthly, and seasonal (quarterly). The resulting patterns are visualized together for a comprehensive understanding.



The daily consumption pattern exhibits clear fluctuations, reflecting variations in energy usage throughout a typical day, with peaks and troughs aligning with typical daily routines, suggesting higher energy demand during specific hours.

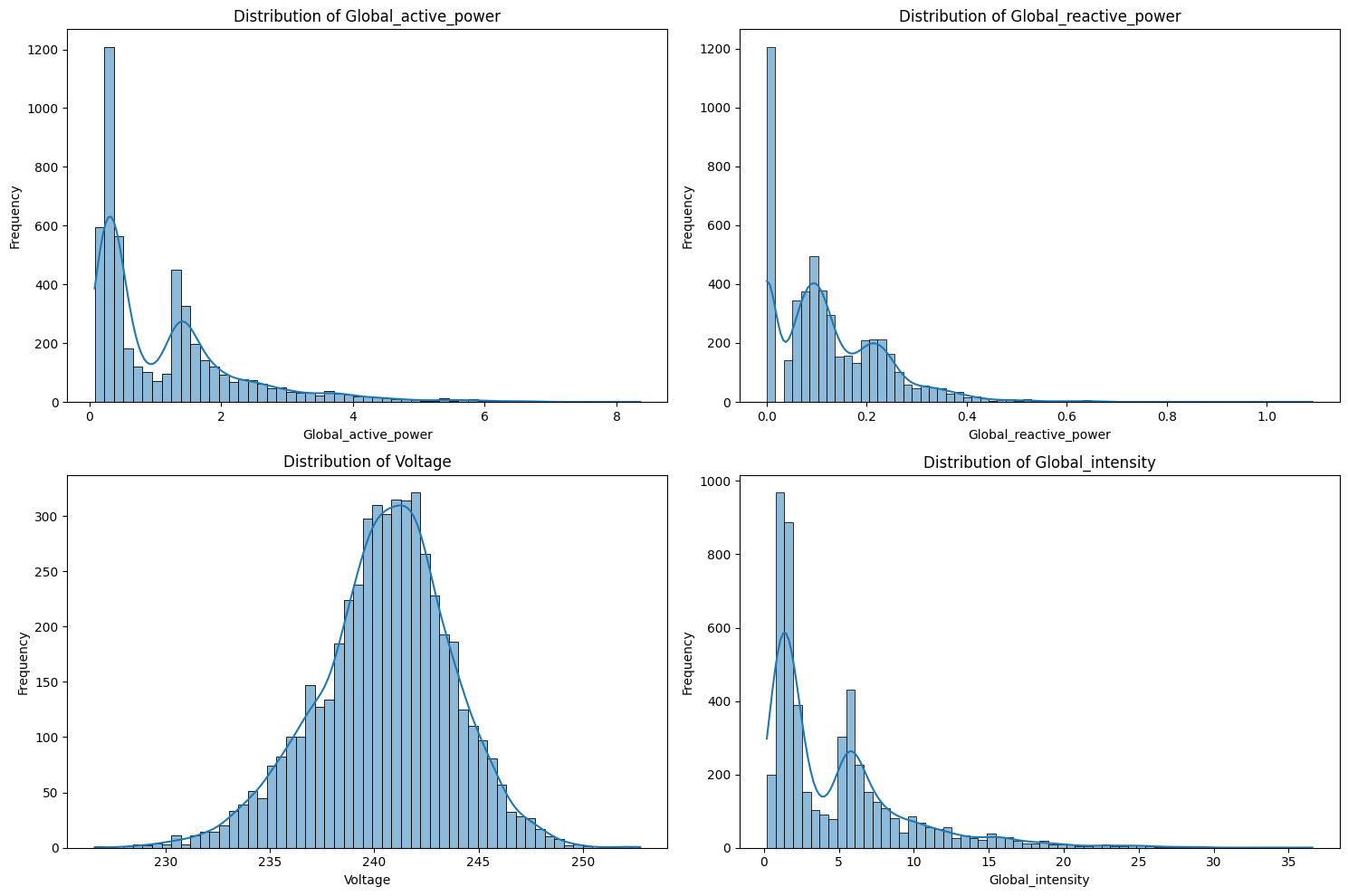
The weekly patterns highlighted variations in energy consumption over different days of the week, with the weekends showing distinctive consumption patterns compared to weekdays, indicating potential lifestyle-related differences.

The seasonal pattern emphasizes the impact of changing weather conditions on energy consumption. Notably, higher energy usage during winter quarters aligns with expectations, as heating systems and additional lighting may contribute to increased demand.

Overall, the visualizations reflected a clear seasonal pattern, with more pronounced energy usage observed during winter periods. We can also infer that resampling over larger time intervals, such as monthly or seasonal, diminishes the periodicity in the dataset.

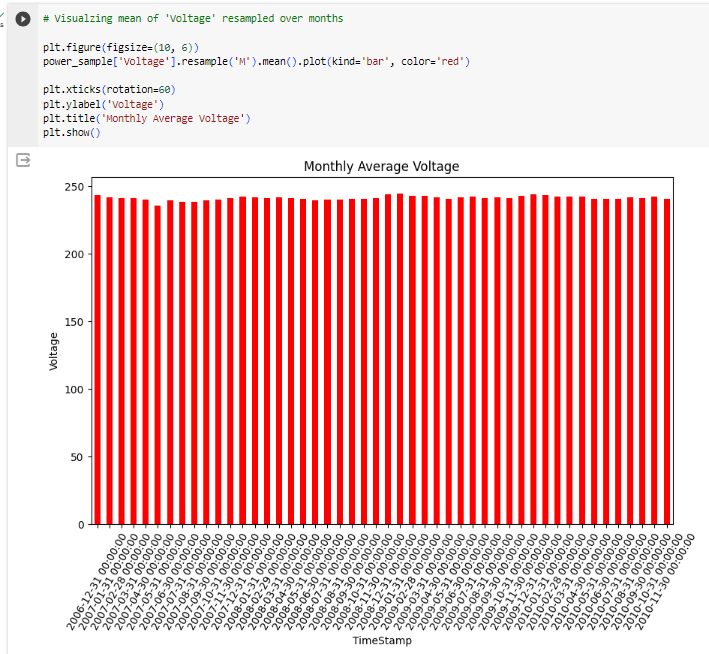
**Variable Distribution**

To further gather insights into the distribution of our other key variables, four histograms with kernel density estimates were plotted.



It is observed that the distribution Global\_active\_power: distribution of global active power appears to be right-skewed, indicating that higher power consumption values are more frequent. The presence of a long tail suggests the existence of periods with lower energy consumption. Similarly, the distribution of Global\_reactive\_power and Global\_intensity also exhibits right-skewness, indicating that higher reactive power values are more common.

Contrastingly, the distribution of voltage appears to be relatively symmetric, with a central peak, indicating a more balanced distribution and the the lack of pronounced skewness suggests that voltage values are distributed more evenly. A further visualization of Voltage also showed a relatively constant level over monthly periods. This is an important discovery, as it shows that Voltage has an almost little effect on consumption decisions.

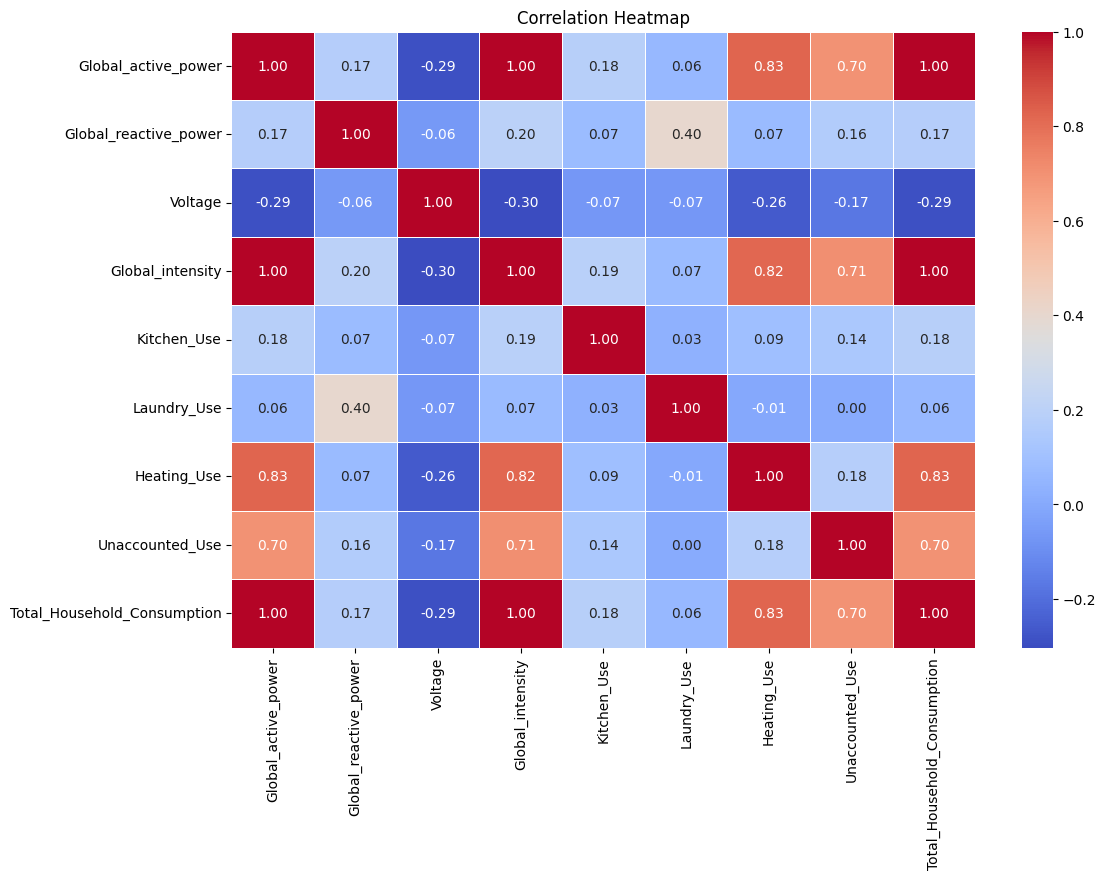


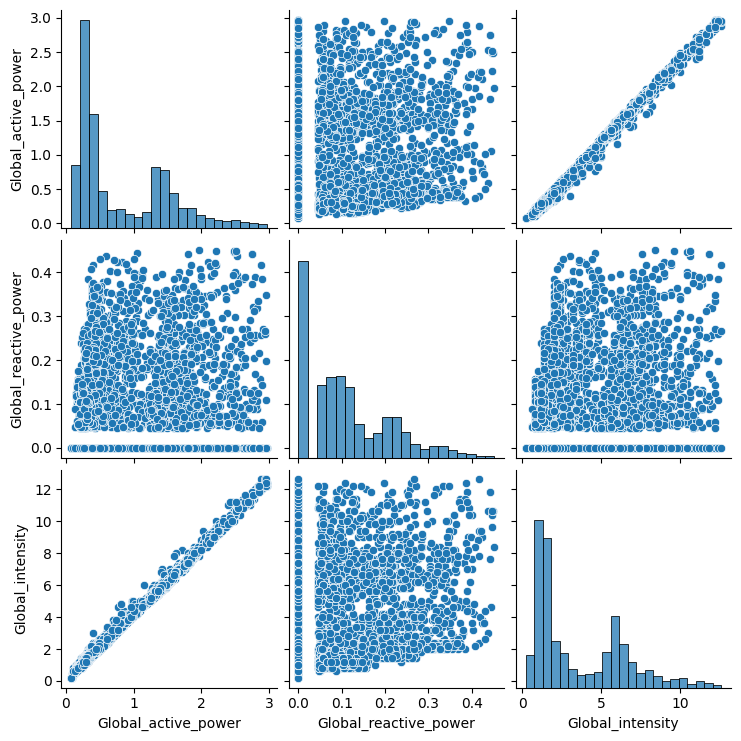
These observations lay the groundwork for further exploration, especially now that we understand the factors contributing to increased energy usage.

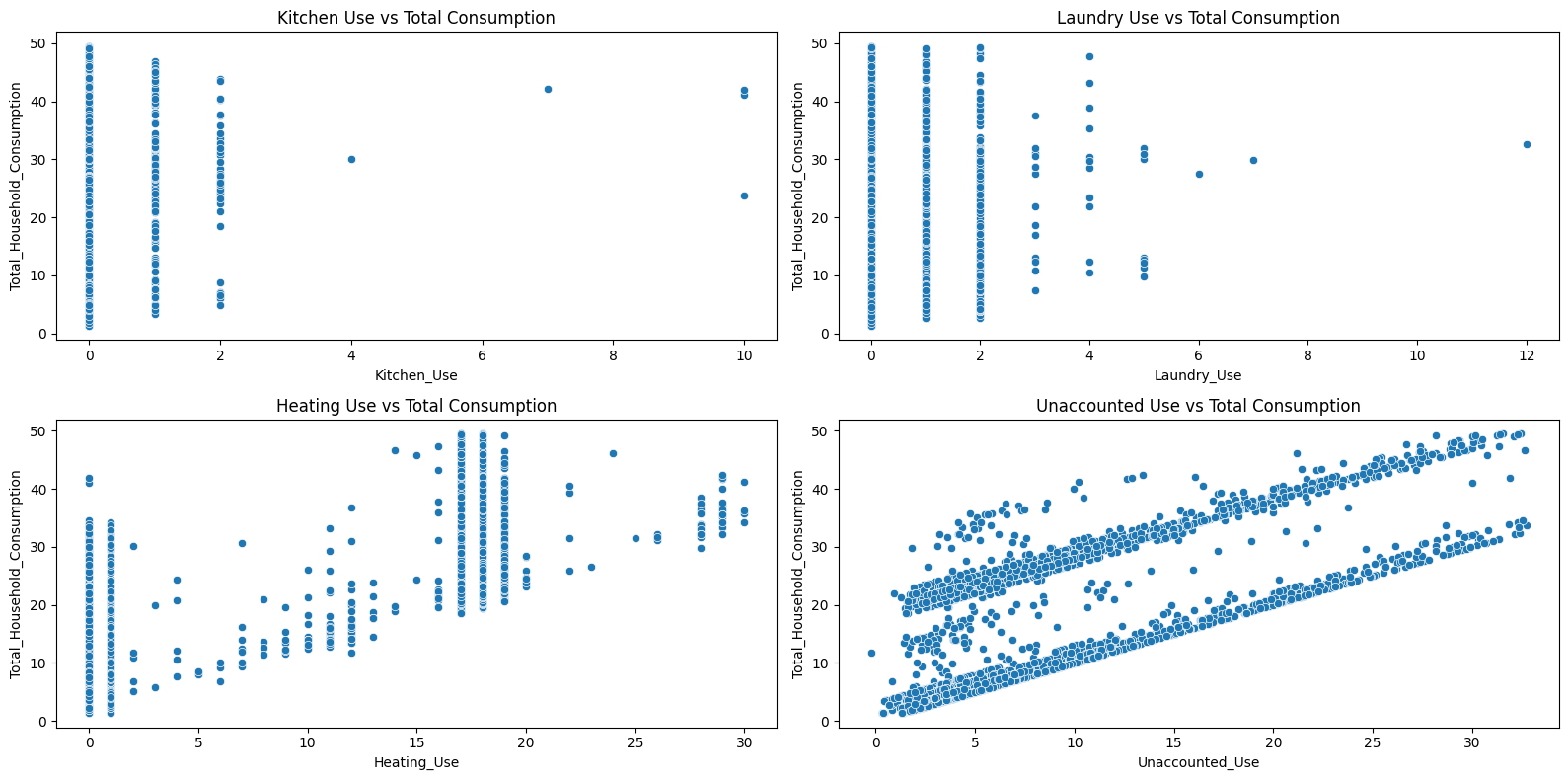
## Correlation Analysis

Employing a correlation heatmap, we visually explored the pairwise correlations between all the variables in the dataset and as expected, there is a perfect positive correlation of 1.00 between *global\_active\_power* and *total\_household\_consumption*, affirming their strong relationship. Similarly, there is a strong and perfect natural relationship between *Global\_intensity* and *Global\_active\_power*.

The sub-metering variables (*Kitchen\_Use*, *Laundry\_Use*, *Heating\_Use and Unaccounted\_use*) also show positive relationships with global active power, especially Heating and *Unaccounted\_use* , which both show a correlation of 0.83 and 0.70 respectively. This reaffirms our earlier observations that a significant portion of household demand is directly from these two categories.







## Clustering Analysis

### **Feature Selection & Scaling**

In preparation for the clustering analysis, the significant features were selected, and the data was scaled using the Standard Scaler.

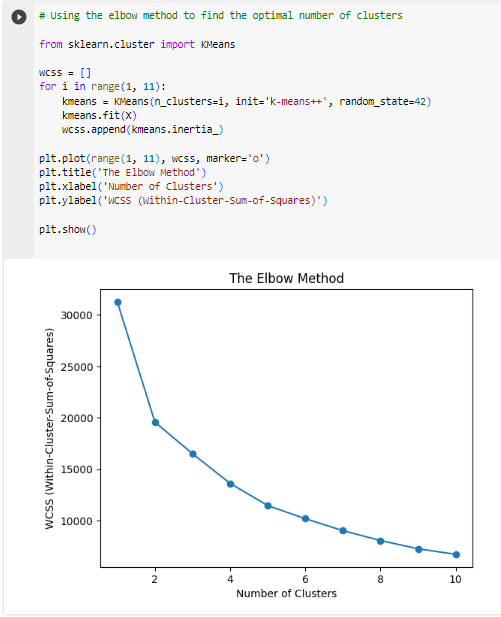


*'Voltage'* was excluded based on earlier observations from the EDA and correlation analysis. Scaling the variables is crucial to ensuring that all features contribute equally to the clustering algorithm, preventing any particular feature from dominating the analysis due to differences in scale.

### **Algorithm 1: K Means Clustering**

To initiate the process of finding our clusters using Kmeans, we first instantiate the Elbow Method. Recall that, the basic idea behind k-means clustering is to define clusters such that the total intra-cluster variation (also known as total within-cluster sum of square) is minimized. When plotted, the optimal number of clusters is typically located at the formation of an ‘elbow’, where the reduction in WCSS typically slows down.

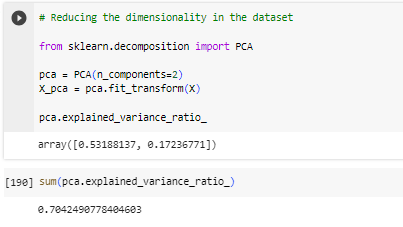
Illustrated below is our implementation of the Elbow technique, where number 5 represents the appropriate number of clusters for our analysis.



#### Dimensionality Reduction with PCA

Given that we selected about 7 significant features, the first logical step to undertake was to transform the data in a manner that would enable Kmeans to efficiently determine which features exhibited influence in terms of electric consumption behaviour.

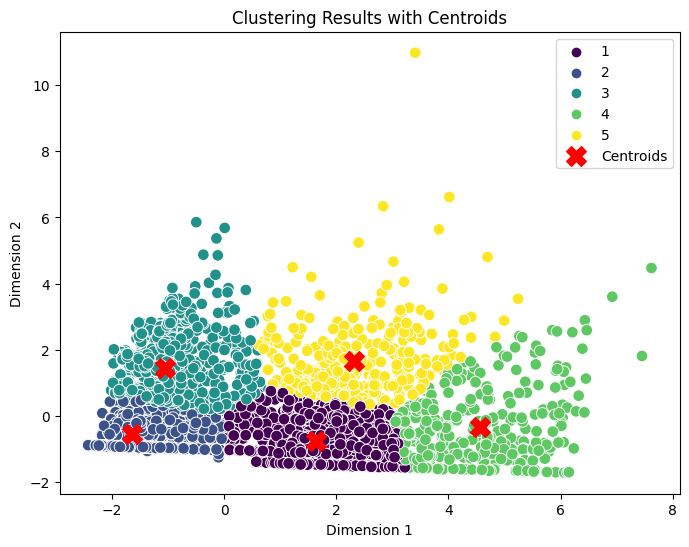
The Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while retaining most of the variance. The transformed dataset X\_pca now contains two principal components that capture a significant portion of the original variance.



The first principal component explained 53.19% of the variance, while the second component contributed 17.24%. Combined, the two components explain approximately 70.42% of the total variance in the dataset.

#### Cluster Identification

On fitting the reduced dimensions onto the K-Means algorithm with a pre-determined number of 5 clusters, the scatter plot below illustrates the clustering results with red markers representing the centroids of each cluster.

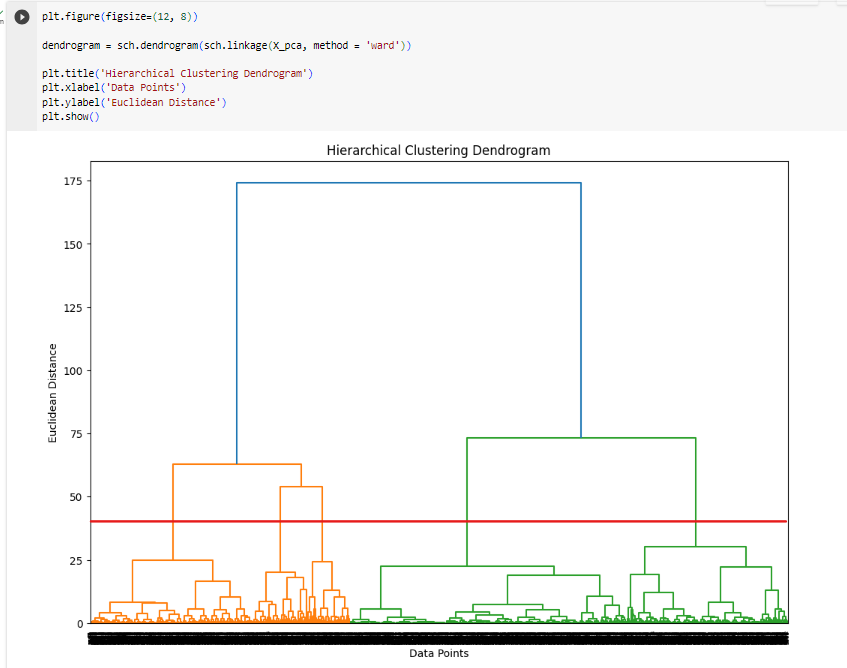


### **Algorithm 2: Hierarchical Clustering**

Hierarchical clustering is a method that builds a hierarchy of clusters, where data points are successively grouped together based on their similarity, forming a tree-like structure known as a dendrogram. (Fionn Murtagh, 2021).

The vertical lines in the dendrogram represent the clusters' formation, while the horizontal lines indicate the distance at which the clusters merge.

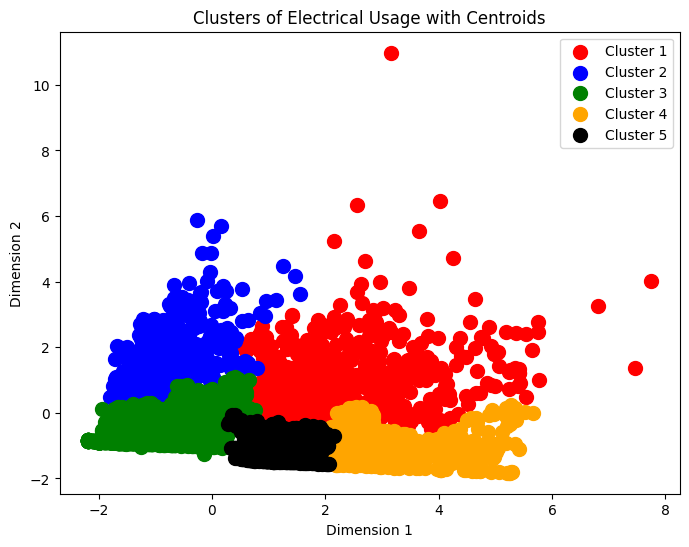
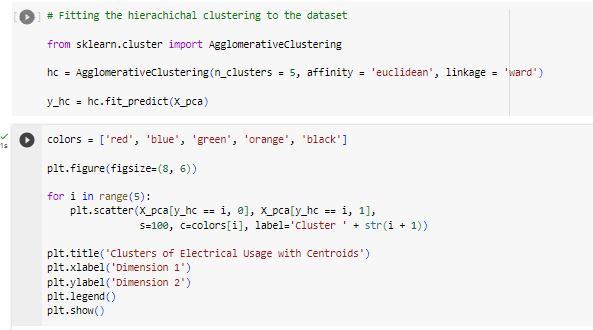
Similar to the elbow method, the dendrogram was plotted, and the ideal number of clusters appeared to be 5. This determination was made by drawing a horizontal redline across the trees, to determine the point resulting in the least loss of information.



#### Cluster Identification

Subsequently, the Agglomerative algorithm was instantiated and applied to the PCA-reduced data, grouping the data points into clusters based on their Euclidean distances in a way that minimized the variance within each cluster.

The resulting clusters were then visualized, with each cluster formed based on the proximity of data points



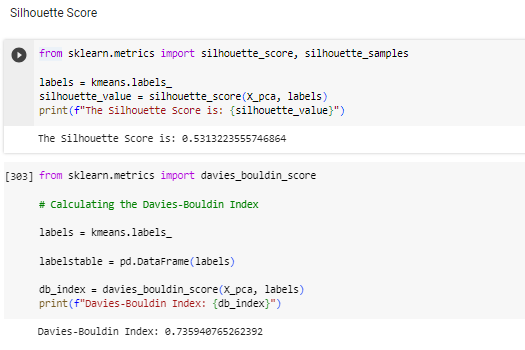
## Cluster Evaluation and Comparison

To evaluate the result of both algorithms, two common evaluation metrics, i.e. the Silhouette Score and Davies-Bouldin Index, were computed.

The Davies-Bouldin Index typically falls in the range of 0 to infinity, with values close to 0, suggesting ideal and compact clusters, whilst higher values, anything above 1, generally indicate that the clustering may not be optimal (Slobodan Petrovic, 2019).

The Silhouette Score, on the other hand, calculates how well-separated the clusters are. The general range Silhouette Scores are from -1 to 1, with a score near 1 indicating how far the data point is away from neighbouring clusters and well matched to its own cluster. A score less than 0 indicates that the data points might have been assigned to the wrong cluster. (Educative, 2023).

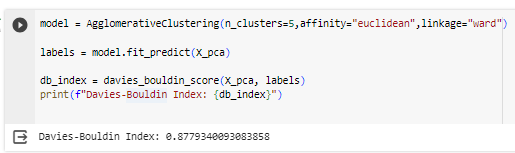
**Kmeans**:



The Silhouette Score of approximately 0.53, showed a moderate level of separation among the clusters, whilst the Davies-Bouldin Index of 0.73 indicated a good level of cluster compactness and separation.

**Hierarchical:**

When computed, the Davies-Bouldin Index was revealed to be 0.88, indicating a lesser discernment in identifying the right clusters.

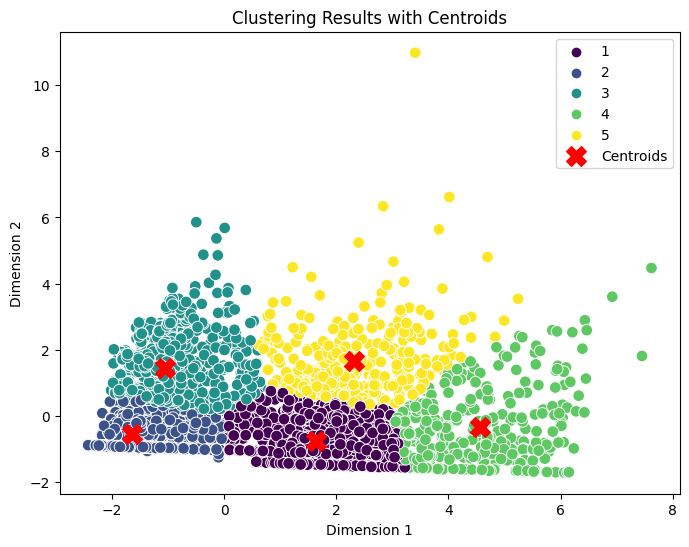


## Cluster Interpretation

Based on the result from our analysis, we identified five clusters using both algorithms and the figure plotted below shows the distribution of the identified clusters.

To further break it down, a heatmap detailing an overview of the cluster profiles was also constructed, with each cluster containing the mean value of the various features relevant to them.





**Cluster 1 (Teal): Balanced Energy Consumers**

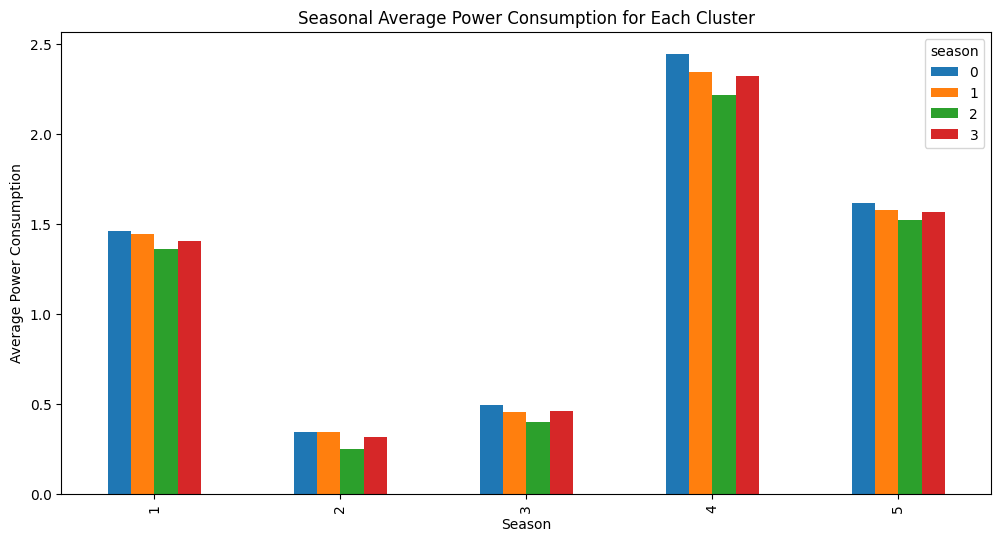
This cluster depicts high values across the key metrics, including a global active power of 1.43, voltage of 240.49, and global intensity of 5.90. Notably, we can observe a substantial use of heating, averaging 14.15 kW. The unaccounted use and total household consumption are also comparatively high, with averages of 9.45 and 23.77, respectively. This household has a balanced usage of various appliances.

**Cluster 2 (Blue): Low Consumers**

This household are the eco-friendly Minimalists and are characterised by low values across all variables, with a global active power of 0.31, voltage of 242.03, and global intensity of 1.33. The usage for Kitchen and laundry is also minimal, contributing to a low average total household consumption of 5.16. This cluster represents a household with minimal consumption, possibly one with energy-efficient practices.

**Cluster 3 (Purple): Moderate Consumers**

This cluster shows a moderate consumption of 7.30, with a global active power of 0.44, voltage of 241.74, and global intensity of 2.01. Within this cluster, the use of electricity for laundry purposes is significant, suggesting a young family scenario with children requiring frequent laundry activities.



**Cluster 4 (Yellow):** **High Consumers** ( High Heating Usage)

This cluster stands out with the highest energy consumption of 39.42. Within the household, heating and unaccounted usage is exceptionally high, 18.00 and 20.88, respectively. This might represent the household’s intense energy usage, during winter season, with high heating requirements.

**Cluster 5 (Green):** **High Consumers** ( Significant Kitchen and Laundry Usage)

Cluster 5 has the second highest consumption of 26.14, with notably high usage patterns registered in the kitchen and laundry. The global active power of 1.57, voltage of 240.17, and global intensity of 6.60 also indicates a high usage. This cluster is likely a large family with more appliances used for kitchen and laundry activities.

# **References**

Ember - <https://www.connaissancedesenergies.org/lelectricite-dans-le-monde-en-2021-annee-de-records-220330>.

[Summary of the IPCC's Sixth Report | Knowledge of energies (connaissancedesenergies.org)](https://www.connaissancedesenergies.org/synthese-du-sixieme-rapport-du-giec-230320)

[Residential electricity consumption behavior: Influencing factors, related theories and intervention strategies - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S1364032117311164)

Jones, R., Fuertes, A., & Lomas, K. (2015). The socio-economic, dwelling and appliance related factors affecting electricity consumption in domestic buildings. Renewable & Sustainable Energy Reviews, 43, 901–917. doi:10.1016/j.rser.2014.11.08

Cayla, J., Maizi, N., & Marchand, C. (2011). The role of income in energy consumption behaviour: Evidence from French household data. Energy Policy, 39(12), 7874–7883. doi:10.1016/j.enpol.2011.09.036

Gram-Hanssen, K. (2013). New needs for a better understanding of household’s energy consumption, lifestyle or practices? Architectural Engineering and Design Management, 10(1-2), 91–107. doi:10.1080/174 52007.2013.83725

Tiwari, P. (2000). Architectural, Demographic, and Economic Causes of Electricity Consumption in Bombay. Journal of Policy Modeling, 22(1), 81–98. doi:10.1016/S0161-8938(98)00003-9

Han, J.; Kamber, M.; Pie, J. Data Mining Concepts and Techniques; Academic Press; Morgan Kaufmann Publisher: Waltham, MA 02451, USA, 2012.

Jessica Temporal, How to define the optimal number of clusters for KMenas, https://jtemporal.com/kmeans-and-elbow-method/, (2019).

Hodge V H and Austin Jim 2004 A survey of outlier detection methodologies J. Artificial Intelligence Review 22 85-126

<https://www.cse.unsw.edu.au/~cs2521/19T3/assigns/ass2/OverviewHacLanceWilliams.pdf>

[Smart meters: a guide for households - GOV.UK (](https://www.gov.uk/guidance/smart-meters-how-they-work)[www.gov.uk](http://www.gov.uk)[)](https://www.gov.uk/guidance/smart-meters-how-they-work)

[Optimal two-level active and reactive energy management of residential appliances in smart homes - ScienceDirect (oclc.org)](https://www-sciencedirect-com.salford.idm.oclc.org/science/article/pii/S2210670721002584?via%3Dihub)

<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b2db00f73fc6b97ebe12e97cfdaefbb2fefc253b>

<https://www.educative.io/answers/what-is-silhouette-score>

Important

[Open Access proceedings Journal of Physics: Conference series (iop.org)](https://iopscience.iop.org/article/10.1088/1757-899X/105/1/012020/pdf)

[Household-Electricity-Consumption/Houshold Electricity Consumption .ipynb at main · royalbaswan/Household-Electricity-Consumption · GitHub](https://github.com/royalbaswan/Household-Electricity-Consumption/blob/main/Houshold%20Electricity%20Consumption%20.ipynb)

[household-electricity-consumption/house electric consumption (2).ipynb at master · Oumayma-mbarek/household-electricity-consumption · GitHub](https://github.com/Oumayma-mbarek/household-electricity-consumption/blob/master/house%20electric%20consumption%20(2).ipynb)

[Frontiers | Real-time recommendations for energy-efficient appliance usage in households (frontiersin.org)](https://www.frontiersin.org/articles/10.3389/fdata.2022.972206/full)

[Time Series Analysis of Household Electric Consumption with XGBoost Model | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/9946913)

[Processes | Free Full-Text | Energy Consumption Patterns and Load Forecasting with Profiled CNN-LSTM Networks (mdpi.com)](https://www.mdpi.com/2227-9717/9/11/1870/htm)