Exploring Residential Energy Behaviors: Identifying Key Factors and Perspectives for Optimization

Kévin Tang

Mohamed-Essaid Khanouche

Parisa Ghodous

August 2024

Claude Bernard University Lyon 1



**Abstract.** Household energy consumption represents a significant share of overall energy consumption. To deepen our understanding of surplus energy production, management, and storage, it is crucial to elucidate the implicit patterns of consumer behaviors and identify factors influencing their performance. This paper mainly aims to descriptively analyze the household energy consumption pattern using RECS2020 data, with a focus on the impact of behavior on energy consumption. To this end, we focus on selecting the most relevant feature subset from a large dataset, which allows for better insights, reduced computation time, and improved predictive performance. The results of this study can help policymakers examine household behaviors at various levels of society to ensure that their quality of life is adequate or needs improvement.

**Mots-clés**: Feature selection - Smart home - Household energy consumption - Correlation analysis

# **Summary**

[**Summary 3**](#_xgf0xldnyqvp)

[**1. Introduction 4**](#_hxsr4p7b4f5r)

[**2. Related Work 5**](#_8obrolxg0qqi)

[**3. Proposed methodology 5**](#_urcet1nvncnz)

[3.1 Dataset Overview 6](#_hn0b8h4g5rs2)

[3.2 Preprocessing 6](#_vhvdqzbv6h9y)

[3.2.1 Removal of variables concerning energies other than electricity 6](#_h9c292i5q4o4)

[3.2.2 Removal of imputation and calibration indicator variables 6](#_ip3ji9y7ul7k)

[3.2.3 Missing Values ​​Management 6](#_arr6os2u099l)

[3.2.4 Encoding Categorical Variables 7](#_t97pc73dozzx)

[3.2.5 Reduction of target variables 7](#_mpqsrqk59ldw)

[3.2.6 Reducing the number of variables 7](#_ptga0o6sl9ur)

[3.2.7 Reducing the number of samples 8](#_7oknh6azls2g)

[3.2.8 Pre-processing Conclusion 8](#_bkkom789404n)

[3.3 Variable selection 9](#_ndiid1z50cuk)

[3.3.1 Correlation across the dataset 9](#_oa6nvlf7npi)

[3.3.2 Correlation on categorized data 9](#_hqjdismvl78)

[3.3.3 Information Gain on the entire data set 11](#_8tjo4s2pzwyb)

[3.3.4 Information Gain on categorized data 12](#_tdi2gu8y5mkj)

[3.3.5 Intersection of the 4 methods 13](#_4mw8sel8knjb)

[3.4 Descriptive analysis of selected variables 13](#_ywgwnirghde5)

[3.4.1 Determining the optimal number of clusters 13](#_qzloabjmokzq)

[3.4.2 K-Means 14](#_hb82l8u6z04e)

[3.4.3 Kernel K-Means 15](#_mvvk7bxqdua)

[3.4.4 DBSCAN 17](#_u1mineb8g1m2)

[3.4.3 Cluster analysis 18](#_c5nxoqv80xgf)

[3.5 Descriptive analysis of selected behavioral variables 19](#_ttwzph31klbt)

[3.5.1 Determining the optimal number of clusters 20](#_2zq1vu93icns)

[3.5.2 Two clusters 20](#_2oug6ctsl8yg)

[3.5.3 Seven clusters 21](#_1ad7t07cg0xn)

[3.5.4 Thirteen clusters 22](#_l7zixlt1ikvk)

[3.5.5 DBSCAN 23](#_o08ul9yxrtt6)

[**4. Discussion 23**](#_r355w4su6hmv)

[**5. Conclusion 24**](#_4bdilksimxm9)

[**Appendix 26**](#_4n8xawru4qiq)

[**References 27**](#_pa4t1a18xbx4)

# **1. Introduction**

The world population reached 8 billion inhabitants in November 2022, one billion more than in 2010, and is expected to cross the 10 billion mark before the end of the century according to UN forecasts [2,3]. This rapid growth, combined with urbanization and increasing energy needs, is leading to a significant increase in household energy consumption. Currently, households account for 36% of electricity consumption in metropolitan France, making this sector the most energy-intensive in proportion, ahead of the tertiary and industrial sectors [4].

Today, energy suppliers and companies specializing in connected objects offer energy consumption monitoring systems for individuals. These devices allow consumers to better understand and manage their energy use by providing them with detailed and real-time data on their consumption. However, to further improve energy management and maximize savings, it is crucial to develop a more detailed understanding of household consumption patterns. This includes identifying factors that influence energy consumption, such as socio-demographic characteristics, lifestyle habits, and appliance efficiency. A thorough analysis of these behaviors and factors can not only help reduce energy consumption, but also tailor energy efficiency strategies to consumers’ specific needs.

By better understanding consumption behaviors, policymakers can develop more targeted and effective policies. For example, they could promote sustainable energy practices, encourage the adoption of more efficient appliances, or offer financial incentives for energy-responsible behaviors. Moreover, consumers themselves, armed with better information, can make more informed decisions and adjust their habits to achieve substantial energy savings.

In this paper, we analyze the RECS2020 dataset, which contains information on residential energy consumption of US households in 2020. Using feature selection techniques, we identify the most influential variables with respect to annual energy consumption and the cost paid to energy suppliers. Once the relevant features are selected, we apply clustering algorithms to partition households into meaningful groups. These groups are then analyzed to derive useful information about household energy consumption patterns and their relationship with variables such as income and living space size.

In addition, we pay special attention to the behavioral characteristics of households among those selected. This analysis allows us to better understand how household habits and behaviors influence their energy consumption, thus providing valuable information for the development of more efficient energy strategies adapted to different consumer profiles.

# **2. Related Work**

In this section, we review related work that identifies the defining characteristics of energy consumption and discuss the methods used to select these characteristics.

The study of [5] focuses on identifying factors influencing the energy consumption of US households using the RECS 2015 dataset, a version prior to the RECS 2020. Although their research and methodology are similar to those of our study, they do not place a particular emphasis on household behaviors. Their partitioning results are limited to differentiating households based on their wealth, thus separating rich households from poor households.

The work of [7] in 2011 identified five factors reflecting household lifestyle, associated with behaviors such as air conditioning use, laundry, personal computer use, climate zone of residence, and television use. These factors were estimated from RECS data from 2001 and 2005.

Like our study, the objective of [6] is to build behavioral archetypes to better understand energy consumption habits in the residential sector. The data used come from the ENERGIHAB project of the French National Research Agency and include a set of 35 variables covering various aspects such as hygiene, food, heating, lighting, leisure practices and housing occupation. These data were collected from 1363 households in Île-de-France. Thanks to these data and an original method, the study succeeded in identifying and building seven distinct behavioral archetypes. However, despite its contributions, this research has some limitations. The number of variables is relatively small, which can limit the depth of the analysis. In addition, the geographical scope of the study is limited to the Île-de-France region, which may reduce the generalizability of the results to the whole of France or to other geographical contexts. Finally, the data date from 2010 and were collected by telephone, which could lead to bias, errors or a lack of up-to-date information.

# **3. Proposed methodology**

The objective of our research is to identify the behaviors that have the greatest impact on household energy consumption, based on the 2020 RECS dataset. We have therefore selected three target variables: the total amount of electricity used in kilowatt-hours (kWh), the total amount of electricity used in BTU (BTUEL), and the total cost of electricity consumed in dollars (DOLLAREL).

In this section, we will first describe the dataset used. Then, we will detail the different preprocessings applied to this dataset, as well as the feature selection methods used to identify the most significant behaviors for our research. Finally, we will analyze the results obtained after applying the clustering algorithms.

## **3.1 Dataset Overview**

The Residential Energy Consumption Survey (RECS) is a comprehensive survey conducted in the United States to collect information on the residential energy consumption of American households. Conducted by the Energy Information Administration (EIA), this survey is one of the most comprehensive and reliable sources of data on energy use in the residential sector.

The 2020 edition of the RECS collects data on various aspects of household energy consumption, including the types and quantities of fuels used, dwelling characteristics, household appliances, heating and cooling systems, as well as recent behaviors such as the use of solar energy or the location of electric vehicle charging. The survey thus covers a wide range of 799 variables, allowing for an in-depth and multidimensional analysis of energy consumption habits.

The RECS uses a representative sample of 18,496 American households to ensure that the results can be generalized to the entire population. Detailed surveys are sent to selected households and can be administered as paper questionnaires, online, by telephone interviews or during on-site visits.

For more information, please visit the EIA website [1].

## **3.2 Preprocessing**

To ensure the quality and relevance of the analyses, several preprocessing steps were applied to the RECS 2020 data. These steps are crucial to clean and prepare the data before applying feature selection techniques and clustering algorithms.

### 3.2.1 Removal of variables concerning energies other than electricity

Since our research aims to identify behaviors influencing electricity consumption, we removed 117 variables relating to other types of energy, such as natural gas, propane, fuel oil and wood. By excluding these irrelevant variables, we were able to focus our analysis specifically on the factors influencing electricity consumption. This approach reduces the complexity of the dataset and facilitates the interpretation of the results.

### 3.2.2 Removal of imputation and calibration indicator variables

Imputation indicator variables flag missing or estimated values ​​for other variables, while calibration variables adjust response weights to reflect a representative population. To refine our analysis and ensure data relevance, we removed 407 of these variables used by the EIA. While these variables are crucial to the EIA’s internal processes, they are not necessary for our study, which focuses specifically on electricity consumption behaviors.

### 3.2.3 Missing Values ​​Management

To ensure the quality and completeness of our dataset, we adopted a rigorous method to handle missing values ​​(NaNs).

The first step is to identify all rows in the dataset containing one or more missing values. This identification is crucial to target the data requiring imputation. To determine the most appropriate values ​​to replace NaNs, we calculated the Euclidean distances between each row containing NaNs and the complete rows in the dataset. To do this, we used the ‘nan\_euclidean\_distances’ function from the ‘sklearn.metrics.pairwise’ library. This function calculates distances taking into account missing values, ignoring NaN positions when calculating the distance.

Once the distances are calculated, each NaN value is replaced by the corresponding value of the closest row, i.e. the one with the smallest Euclidean distance. This similarity-based imputation method ensures that missing values ​​are replaced consistently with existing data, thereby minimizing potential bias and errors.

### 3.2.4 Encoding Categorical Variables

To prepare our dataset for further analysis, we converted the categorical variables into a form that machine learning algorithms can understand. We used the OneHotEncoder method from the scikit-learn library. This technique transforms each categorical value into a new binary column (0 or 1), indicating the presence or absence of that value.

In our dataset, we identified 7 categorical variables that required encoding. Applying one-hot encoding to these variables resulted in the creation of 144 new columns. This process allowed us to convert the categorical data into a numerical representation while preserving the integrity and meaning of the original information.

### 3.2.5 Reduction of target variables

At the beginning of our study, we identified three main target variables to assess electricity consumption: the total amount of electricity used in kilowatt hours (kWh), the total amount of electricity used in BTUs (BTUEL), and the total cost of electricity consumed in dollars (DOLLAREL). However, after performing a correlation analysis between these variables, we found that some of them were highly correlated. In particular, the BTUEL variable was highly correlated with kWh, because these two measures are interconvertible (1 kWh is equivalent to approximately 3,412 BTUs).

As a result, we decided to reduce the number of target variables to two: kWh and DOLLAREL. This reduction simplifies the analysis while retaining the essential measures to assess electricity consumption.

### 3.2.6 Reducing the number of variables

As part of data preprocessing, we identified highly correlated pairs of variables in our dataset to reduce redundancy. This analysis led to the removal of 105 variables.

For example, all variables associated with state zip codes were highly correlated with those representing state names. We therefore retained only the variables relating to state names. Similarly, the variable concerning subarctic climate was highly correlated with the IECC climate code for Alaska [8]. We retained one of these two variables to avoid duplicating climate information.

Another example is the variable indicating the presence of air conditioners, which was highly correlated with the variable showing the use of electricity for cooling. This suggested that all air conditioners use electricity as an energy source, justifying the removal of the variable on the use of electricity for cooling.

This approach simplified the dataset while retaining the most relevant information for our analysis.

### 3.2.7 Reducing the number of samples

Isolated data, often considered as outliers, can introduce significant biases in machine learning models and alter the results of statistical analyses. To ensure the accuracy and reliability of analyses, it is crucial to identify and remove these outliers.

In our study, we used the K-nearest neighbors (KNN) algorithm to detect isolated data. This method allowed us to identify samples that were significantly different from others due to their low density or distance from neighboring points. A total of 93 samples were classified as isolated.

Removing these outliers led to a reduction in the total number of households in our dataset from 18,496 to 18,403. By removing these isolated data, we improved the quality of the dataset, which allows for more accurate and reliable analyses.

### 3.2.8 Pre-processing Conclusion

Following the rigorous pre-processing steps, we significantly refined our dataset to make it more relevant and usable for our analysis.

The number of samples was reduced from 18,496 to 18,403, following the removal of 93 outliers identified as isolated. This reduction allowed us to maintain a more coherent and representative database.

Regarding the variables, we simplified our dataset by reducing the total number of variables from 799 to 310. This reduction was achieved by eliminating variables related to other types of energy, imputation and calibration indicator variables, as well as those that were highly correlated. This approach reduced redundancies and focused our analysis on the most relevant factors.

These pre-processing steps not only optimized data quality, but also facilitated more precise and meaningful analyses. Ultimately, these adjustments help to obtain more reliable results and deepen our understanding of electricity consumption behaviors.

## **3.3 Variable selection**

Variable selection is a crucial step in data preprocessing, aiming to improve the quality of predictive models by identifying the most relevant variables. In our study, we applied several methods to refine our choice of variables.

### 3.3.1 Correlation across the dataset

The first method of variable selection is based on the analysis of the correlation between all the variables in the dataset and the target variables.

We first calculated the correlation between each variable and the target variable KWH to identify those with a significant relationship with electricity consumption. The variables whose correlation with KWH was higher than the average of the calculated correlations were selected.

Similarly, we evaluated the correlation between each variable and the second target variable, DOLLAREL, to identify the variables influencing the cost of electricity consumption. The variables whose correlation with DOLLAREL was also higher than the average of the calculated correlations were retained.

Finally, the relevant variables for each of the two targets were crossed to obtain their intersection, which made it possible to select 106 important variables for the two target variables, guaranteeing their relevance in both analysis contexts.

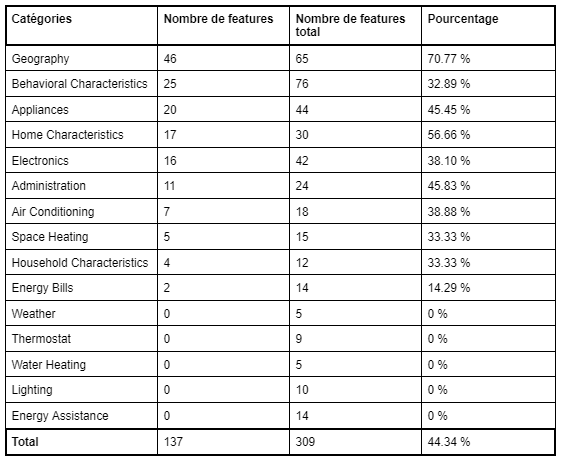
### 3.3.2 Correlation on categorized data

The second method of variable selection is based on the categorization of the variables in the dataset and the analysis of the correlation within these categories. This approach allows to structure the selection process by treating the variables according to their specific categories.

The EIA partitioned the variables into 14 distinct categories. We added an additional category specifically dedicated to behavioral variables, comprising 88 manually identified variables. The 15 categories are as follows:

* Administration: This category groups administrative and contextual variables that provide essential information for the identification, climatic classification and urban type of respondents. It concerns 4 variables, encoded in 24 columns.
* Geography: This category groups geographic variables that provide essential information on the location and geographical characteristics of respondents such as the state of residence. It concerns 5 variables, encoded in 65 columns.
* Weather: This category groups meteorological variables that provide essential information on the climatic conditions and meteorological characteristics of the areas inhabited by respondents such as outdoor, soil or light bulb temperatures. It concerns 5 variables.
* Home characteristics: This category groups variables that provide detailed information on the physical and structural characteristics of respondents' dwellings such as the size and type of dwelling. It concerns 30 variables.
* Household characteristics: This category groups together variables that provide detailed information on the socio-demographic and economic characteristics of respondents’ households, such as the number of inhabitants, their level of education or household income. It concerns 12 variables.
* Appliances: This category groups together variables that provide detailed information on the ownership, use and characteristics of household appliances present in respondents' homes. It concerns 44 variables.
* Electronics: This category groups together variables that provide detailed information on the ownership, use and characteristics of electronic devices present in respondents' homes. It concerns 42 variables.
* Lighting: This category groups together variables that provide detailed information on the types, use and characteristics of lighting systems present in respondents' homes. It concerns 10 variables.
* Air conditioning: This category groups together variables that provide detailed information on air conditioning systems, including their presence, use and characteristics in respondents' homes. It concerns 18 variables.
* Water heating: This category groups variables that provide detailed information on water heating systems, including their type, use and characteristics in respondents' homes. It concerns 5 variables.
* Space Heating: This category groups together variables that provide detailed information on heating systems, including their type, use and characteristics in respondents' homes. It concerns 15 variables.
* Thermostat: This category groups variables that provide information on temperature control devices, including their type, use and characteristics in respondents' homes. It concerns 9 variables.
* Energy bills: This category groups variables that provide detailed information on the responsibility for paying energy bills, the presence of a smart meter or the location of the electric vehicle charging. It concerns 14 variables.
* Energy assistance: This category groups together variables that provide information on aid and subsidies received for energy consumption. It concerns 14 variables.
* Behavioural characteristics: This category groups together variables from all other categories that provide information on households' energy consumption habits and behaviours, including daily routines or usage preferences. It concerns 76 variables.

For each category, we calculated the correlation between the variables and the two target variables, then selected those whose correlation was higher than the average within each category. This approach allowed us to retain a total of 137 variables.

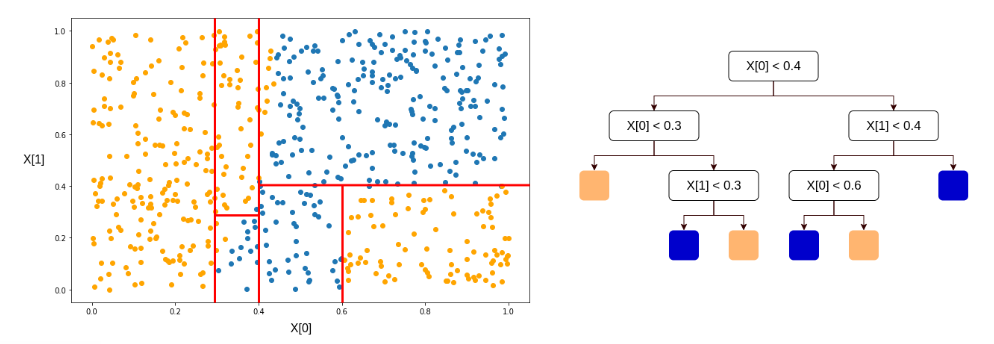


*Figure 1: Table of results of method 2*

### 3.3.3 Information Gain on the entire data set

The **DecisionTreeRegressor** is a regression algorithm that builds a decision tree to predict continuous values, such as energy consumption (KWH) or cost (DOLLAREL).

**How the algorithm works:** The decision tree divides the data by asking successive questions about the characteristics of the training data. For each new observation, the tree follows the branches corresponding to the characteristics of the observation and predicts a value based on the average of the data values ​​that arrive at the terminal leaf.



*Figure 2: Illustration of how DecisionTreeRegressor works*

**Feature importance assessment:** The algorithm assesses the importance of each feature by measuring the reduction in impurity, quantified by the squared error, brought by each division of the tree. At each node, it selects the feature that minimizes this impurity, thus optimizing the predictions.

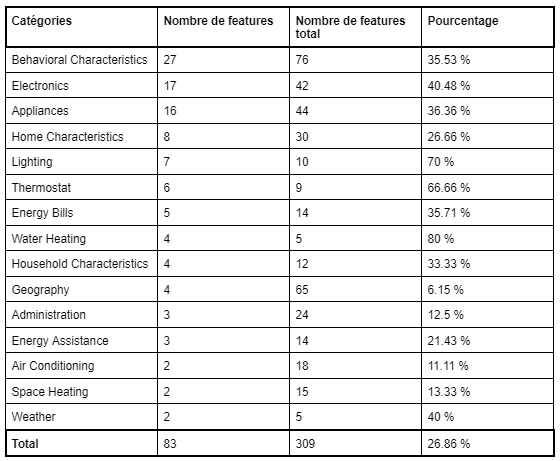
**Feature Selection:** After determining the importance of each feature, we rank these features based on their contribution to reducing the overall squared error. A cumulative curve is then plotted to visualize the contribution of each feature. By identifying the point where the curve levels off, we can determine the optimal number of features to retain, thus balancing accuracy and model simplicity.

This approach allows to select the most relevant variables and to build an efficient regression model, capable of making accurate predictions while avoiding overfitting. Applied to the entire preprocessed data, this method allowed to select 63 variables.

### 3.3.4 Information Gain on categorized data

As mentioned earlier, the original data are organized into different categories. For each category, we applied the elbow method to identify the optimal number of variables to keep. This approach consists of plotting the cumulative impurity reduction as a function of the number of retained variables and determining the threshold beyond which the addition of additional variables only marginally improves the impurity reduction.

By applying this method individually to each category, we were able to select 83 relevant variables. This approach ensures that the chosen variables contribute significantly to the prediction while maintaining the simplicity and efficiency of the model.



*Figure 3: Table of results of method 4*

### 3.3.5 Intersection of the 4 methods

The intersection of the results obtained by the four selection methods allowed us to identify 28 variables among the initial 309 with the greatest impact on energy consumption. This joint selection highlights the most significant variables by combining the perspectives of the different methods. Among these 28 variables, 12 come from the 76 behavioral variables, highlighting their particular importance in the analysis of energy consumption.

The complete list of the 28 selected variables can be found in Appendix A.

## **3.4 Descriptive analysis of selected variables**

After preprocessing, we selected 28 features with the highest influence on the target attributes. Since the data is unlabeled, we use unsupervised clustering methods to analyze this data.

### 3.4.1 Determining the optimal number of clusters

To determine the optimal number of clusters in our analysis, we used four complementary methods: the elbow method with inertia, the silhouette score, the Davies-Bouldin score, and the Calinski-Harabasz criterion. Each method provides a different perspective on the quality of clustering and helps identify the most appropriate number of clusters to best represent the structure of the data.

**Elbow Method with Inertia:** The elbow method is a classic approach to determining the optimal number of clusters. It relies on inertia, also called the intra-cluster sum of squares. Inertia measures the spread of data points within each cluster: the lower the inertia, the closer the points are to the center of the cluster. Plotting inertia against the number of clusters yields a curve that typically shows a rapid reduction at first, followed by a leveling off. The "elbow" in this curve, where the reduction in inertia begins to decrease significantly, indicates the optimal number of clusters. This point represents a balance between cluster compactness and model complexity.

**Silhouette score:** The silhouette score evaluates the quality of clustering by measuring the similarity of points within their own cluster relative to the similarity with points in other clusters. The score ranges from -1 to 1: a score close to 1 indicates that points are well clustered in their own cluster and far from other clusters, while a score close to -1 suggests that points might be misclassified. For different numbers of clusters, we calculated the average silhouette score. The number of clusters that maximizes this score is considered optimal, as it indicates the best separation and cohesion between clusters.

**Davies-Bouldin Score:** The Davies-Bouldin score is another measure of clustering quality, which assesses the compactness and separation of clusters. It is calculated by measuring the average of the similarity ratios between each cluster and the most similar cluster. A lower score indicates better separation between clusters, suggesting that the clusters are well-distinct and compact. Comparing the scores for different numbers of clusters, the number of clusters that minimizes the Davies-Bouldin score is chosen as optimal, as it reflects the best separation of the groups.

**Calinski-Harabasz criterion:** The Calinski-Harabasz criterion, also known as the "variance ratio criterion", evaluates the quality of clustering by comparing the intra-cluster variance to the inter-cluster variance. The higher this ratio, the better the separation between clusters and the more compact the clusters are. We calculated this criterion for different numbers of clusters and selected the number that maximizes the Calinski-Harabasz criterion. This indicates that the clusters are well-defined with clear separation and high internal cohesion.

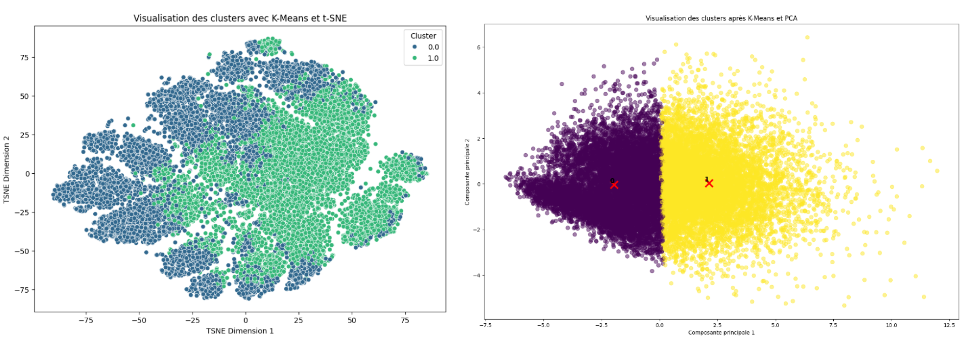
The application of the four methods for selecting the optimal number of clusters led to a convergent conclusion: all methods suggest that the optimal number of clusters is 2. The agreement between these different methods reinforces the conclusion that two clusters are the most appropriate choice to capture the underlying structure of the data. This multi-method consensus validates the robustness and reliability of this solution, thus ensuring an adequate representation of the groups within our analysis.

### 3.4.2 K-Means

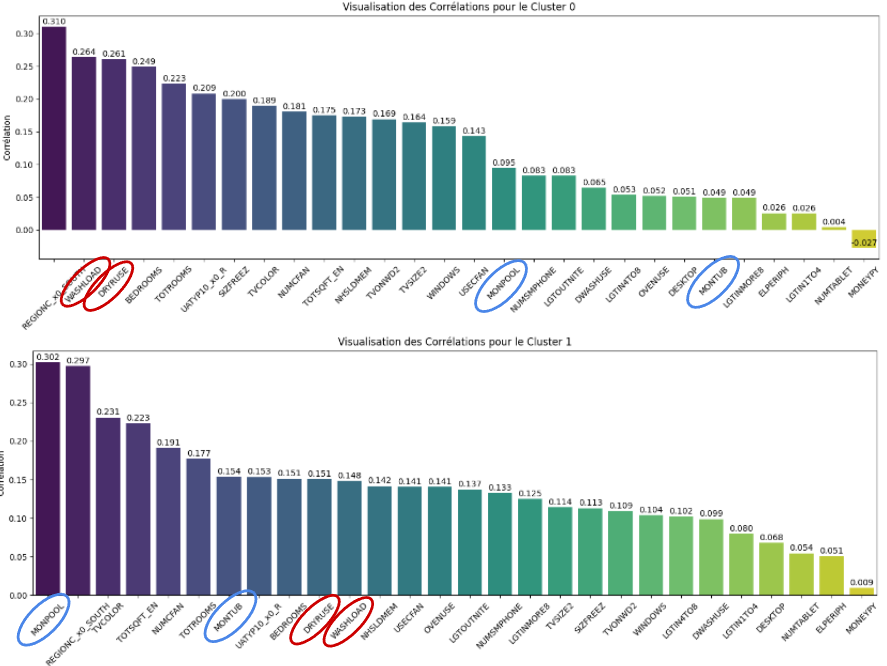
K-means is an unsupervised clustering method that aims to divide a data set into a fixed number of groups, called clusters. The goal is to minimize the variance within clusters and maximize the variance between clusters.

The K-means algorithm works in several steps. The algorithm begins by initializing the cluster centers. These centers can be determined randomly or by using specific techniques such as K-means++ to improve the initial selection of centers. Each data point is assigned to the cluster whose center is closest. This proximity is usually measured by the Euclidean distance between the data points and the cluster centers. Once all the points have been assigned to clusters, the cluster centers are recalculated by taking the average of the points belonging to each cluster. The steps of assigning points and updating the centers are repeated until the cluster centers converge, i.e. the changes become negligible, or until a stopping criterion is reached (such as a maximum number of iterations). The algorithm therefore stops when the cluster centers no longer vary significantly, which indicates that the clusters have reached a form of stability.

In this research, the K-means algorithm was applied using the optimal number of clusters determined by previous methods, set at two clusters.



*Figure 4 : Visualization of the clustering of the selected variables with K-Means*



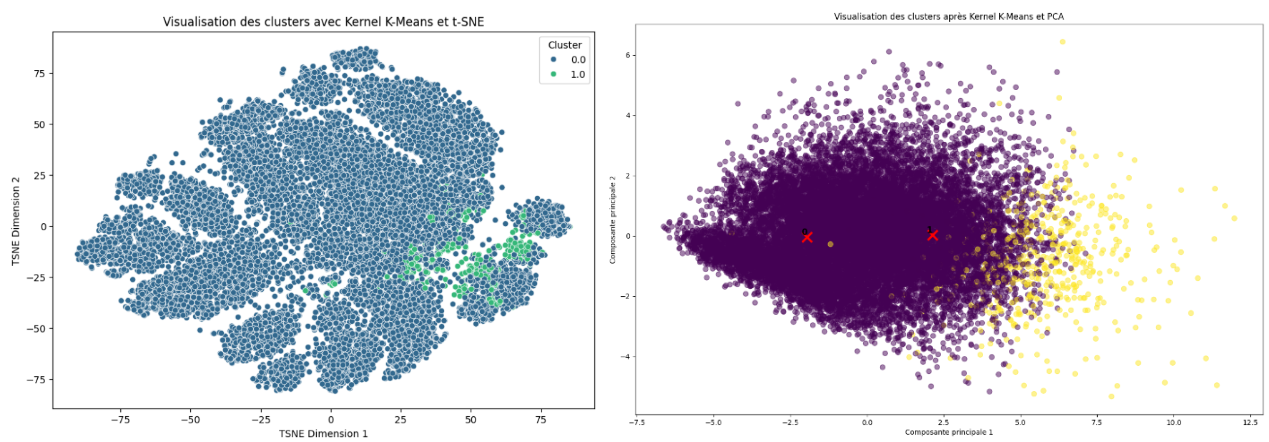
*Figure 5 : Cluster Correlations with K-Means (2 Clusters)*

### 3.4.3 Kernel K-Means

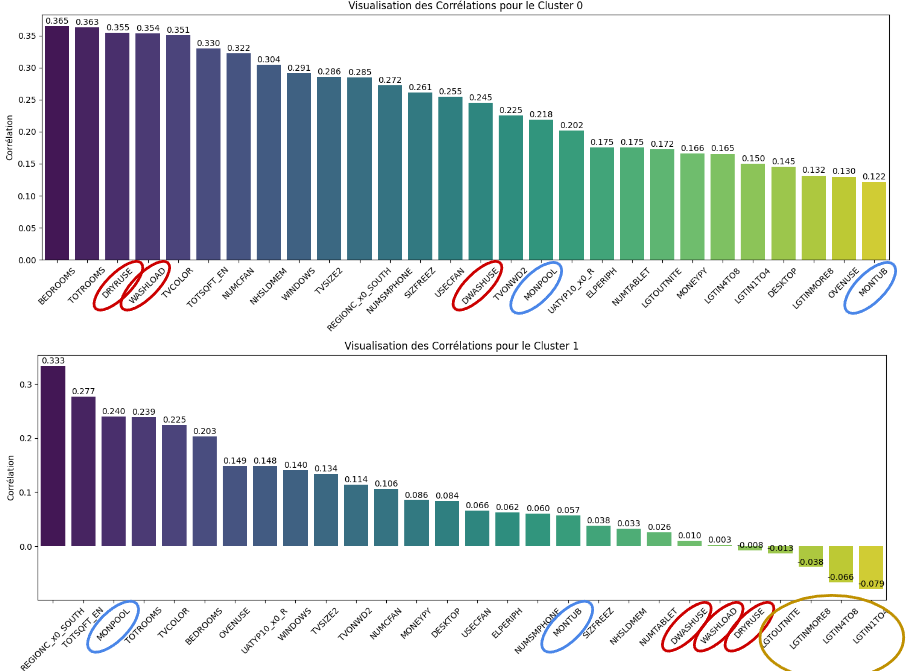
Kernel K-means is an extension of the classical K-means algorithm that uses a kernel function to map the data into a higher-dimensional space, thereby capturing nonlinear structures in the data. By transforming the data using a kernel function, Kernel K-means can identify clusters that would not be detectable by classical K-means because they lie in a space where the boundaries between clusters are nonlinear.

The algorithm follows similar steps to K-means: initializing cluster centers, assigning points to clusters based on distances computed with the kernel matrix, updating cluster centers, and iterating until convergence. However, instead of using traditional Euclidean distance, Kernel K-means uses distances computed in the transformed space by the kernel function. This approach helps to better capture complex relationships between data points.

In our study, after applying K-means, we used Kernel K-means to check whether more complex structures existed in the data. The results confirmed that the identified clusters were robust and allowed us to better understand household energy consumption behaviors.



*Figure 6 : Visualization of clustering of selected variables with Kernel K-Means*

**

*Figure 7 : Cluster Correlations with Kernel K-Means (2 Clusters)*

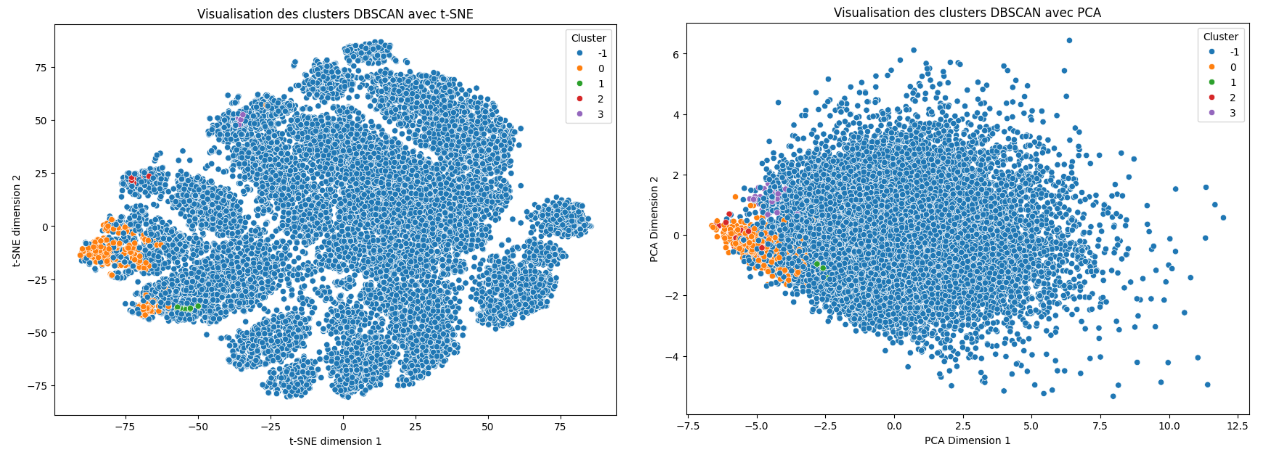
### 3.4.4 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a widely used unsupervised clustering algorithm to identify clusters in datasets based on the density of data points. Unlike other clustering methods such as K-Means, which require defining the number of clusters in advance, DBSCAN can automatically identify the number of clusters based on the local density of points. The algorithm relies on two main parameters: ε (epsilon), which represents the maximum distance between two points for them to be considered neighbors, and minPts, the minimum number of points in the ε neighborhood for a point to be considered a core point.

DBSCAN starts by selecting an unvisited point in the dataset and identifies all points in its neighborhood ε. If the number of neighboring points is greater than or equal to minPts, the point is considered a core point and a new cluster is created. If the number of neighboring points is less than minPts, the point is marked as noise, although it can be included in a cluster later if it is a neighbor of a core point. For each core point, the algorithm explores all points in its neighborhood and, if these points are also core points, their neighborhood is also explored. This process continues until the cluster can no longer be expanded. The algorithm repeats these steps until all points are visited.

DBSCAN has several advantages. It can identify clusters of arbitrary shapes, unlike K-Means which assumes spherical clusters, and explicitly handles noise points, which improves the quality of the formed clusters. In addition, it automatically determines the number of clusters based on the density of the points, which is advantageous when the number of clusters is not known in advance. It is also effective for databases containing large volumes of data and works well with databases of varying density.

Applying the DBSCAN algorithm to our dataset, with ε parameters set to 2 and minPts set to 29, revealed the formation of four distinct clusters. However, it is important to note that these clusters are relatively small in size. The majority of data points were classified as anomalies, reflecting the strict nature of the algorithm in defining cluster points based on local density. This feature of DBSCAN highlights its ability to identify and isolate outliers in the datasets, but it may also indicate that the ε and minPts parameters may need adjustment to better capture potential cluster structures present in the data.



*Figure 8 : Visualization of clustering of selected variables with DBSCAN*

### 3.4.3 Cluster analysis

Both clustering methods, K-means and Kernel K-means, segmented households into two distinct groups, each highlighting specific energy consumption characteristics.

Households in **Cluster 0** are characterized by smaller dwellings, lower income, and lower ownership and use of household appliances. The main drivers of energy consumption in this cluster are related to the essential needs of a dwelling, mainly the use of household appliances for daily tasks such as cooking, refrigeration, and washing.

Households in **Cluster 1**, on the other hand, are distinguished by larger dwellings, higher income, and much more intensive use of household appliances and electronics. In this cluster, the main drivers of energy consumption include leisure-related amenities, such as swimming pools and hot tubs, which are largely responsible for the high electricity consumption.

**K-Means** identified two distinct clusters, but more broadly, including a wide range of high-income households in Cluster 1. The findings show a clear distinction between low-income households with basic needs and high-income households with leisure amenities.

**Kernel K-Means** identified a specific subgroup of very wealthy households in Cluster 1, reducing the total number of samples in this cluster compared to K-means. This means that Kernel K-means was able to distinguish a group of extremely wealthy households, whose energy consumption habits are even more marked by leisure amenities and luxury consumption behaviors.

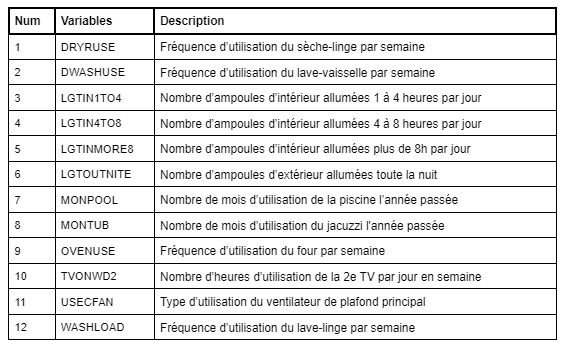
The differences between the two clusters highlight the various aspects of household energy consumption. For Cluster 0, energy consumption is mainly due to essential needs, reflecting households living in modest housing with lower incomes. Strategies to reduce energy consumption in this cluster could focus on the energy efficiency of household appliances and daily practices. For Cluster 1, high energy consumption is mainly attributed to recreational amenities such as swimming pools and hot tubs, characteristic of high-income households living in large dwellings. Efforts to reduce energy consumption in this group could include measures to make these amenities more efficient or promote less energy-intensive alternatives.

To complete this analysis, we also applied the DBSCAN algorithm which revealed the formation of four distinct clusters, but it is important to note that these clusters are relatively small in size. The majority of the data points were classified as anomalies. Unlike the results obtained with previous algorithms, these four clusters identified by DBSCAN group together households with lower income than those in the previous clusters. The analysis reveals that, despite differences in income within these clusters and in housing size, these households share similar behaviors in their low use of household equipment and appliances. Households in these clusters are generally poorly equipped, without a swimming pool or ceiling fan. Some have only one television and live alone or in pairs. Their only leisure activity is often this television. Some households do not even have a refrigerator or oven.

In summary, this analysis reveals that energy consumption behaviors differ greatly by income and housing size, directly influencing the potential energy-saving strategies to be adopted for each group. These results highlight the importance of targeting specific energy efficiency measures adapted to the characteristics and needs of different segments of the population, in order to maximize energy savings and improve environmental sustainability.

## **3.5 Descriptive analysis of selected behavioral variables**

After segmenting households into two distinct clusters and analyzing energy consumption characteristics, we focused our attention on behavioral variables among those that were selected. Recall that the initial objective of our study was to analyze the influence of behavior on energy consumption. Among the 28 variables resulting from the selection of variables, 12 are specifically related to household behaviors. By isolating these behavioral variables, we can examine in more detail how household habits and practices impact their energy consumption. This step is crucial to understand the specific behavioral factors that contribute to the variation in energy consumption between different households.



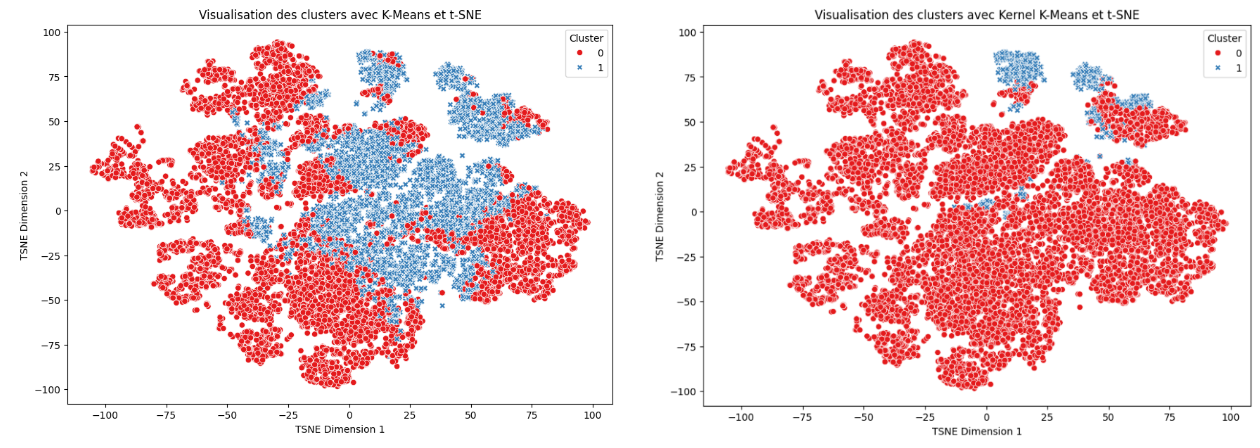
*Figure 9 : Table listing selected behavioral variables*

### 3.5.1 Determining the optimal number of clusters

As before, we determined the optimal number of clusters using multiple assessment methods. Both the elbow method and the Calinski-Harabasz criterion indicated that the optimal number of clusters is 2. In contrast, the silhouette score suggested 7 clusters, while the Davies-Bouldin index recommended 13 clusters. These discrepancies highlight the importance of considering multiple criteria for a robust assessment of the number of clusters, as each method uses different criteria and may interpret the data differently.

### 3.5.2 Two clusters

The analysis revealed two main clusters, as previously observed. The results obtained with the K-means and Kernel K-means methods show a similarity with the clusters identified in previous analyses. Both methods confirmed the separation between low- and high-income households. Kernel K-means once again highlights a specific subgroup among the wealthiest households.



*Figure 10 : Visualization of the clustering of behavioral variables into 2 clusters*

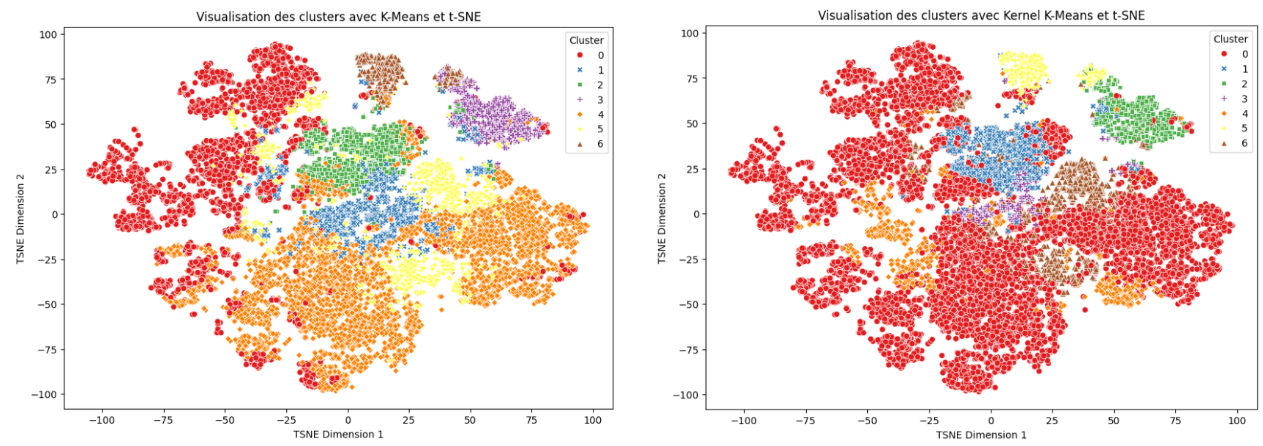
### 3.5.3 Seven clusters

By applying K-means and Kernel K-means algorithms to segment the data into 7 clusters, we further analyzed household energy consumption behaviors. The results show that both clustering methods produce similar profiles in terms of segmentation.

The clusters reveal distinct energy consumption patterns. For example, Cluster 0 groups households that use all appliances moderately. Cluster 1 is characterized by high use of lights, while Cluster 2 consists of households that use kitchen appliances intensively. Cluster 3 is distinguished by high consumption related to the swimming pool and jacuzzi. Cluster 4 includes households that use the swimming pool, jacuzzi, and kitchen appliances little, while Cluster 5 is characterized by high use of household appliances and television. Finally, Cluster 6 represents households with high consumption of the jacuzzi, swimming pool, and lights, but low use of television.

The analysis highlights several key findings regarding the drivers of energy consumption. The swimming pool emerges as the most important driver of energy consumption, particularly in clusters where its use is predominant. In comparison, the use of lights and the oven appear as less significant factors of energy consumption, although the use of the oven has a greater impact than that of lights. Household appliances are a more important factor in energy consumption compared to the oven. In contrast, television has little influence on energy consumption, as shown by its limited impact in the clusters.

These results highlight the diversity of energy consumption behaviors among households and highlight the importance of examining specific equipment to develop effective strategies to reduce energy consumption.



*Figure 11 : Visualization of the clustering of behavioral variables into 7 clusters*

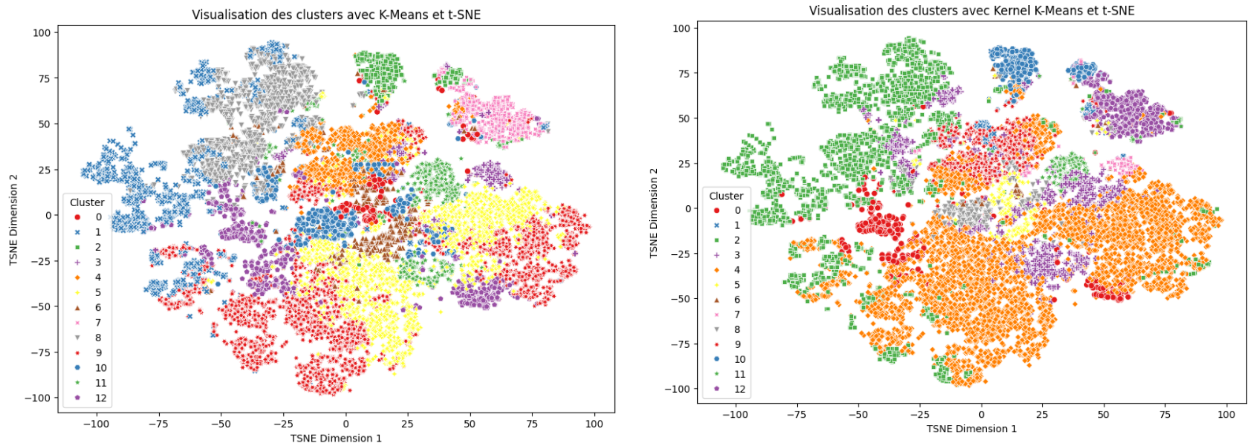
### 3.5.4 Thirteen clusters

Applying K-means and Kernel K-means algorithms to segment the data into 13 clusters provided a more detailed view of household energy consumption behaviors.

With K-means, the swimming pool was identified as the most significant energy consumption factor, highlighting its predominant impact. Household appliances were found to be crucial for energy consumption, surpassing lights in terms of importance. Using the oven was found to be more influential than television. On the other hand, leaving the lights on all night became a more important consumption factor than prolonged use of lights during the day. Conversely, leaving the lights on for a short period (less than 4 hours per day) was found to be a less significant consumption factor than television, which remains the factor with the least impact.

With Kernel K-means, similar findings were observed, but with additional nuances. Household appliances continue to play a prominent role compared to lights. Oven use is more significant than television use, but the importance of prolonged use of lights (less than 8 hours per day) was found to be greater than that of the oven. Leaving lights on all night remains a more significant consumption factor than prolonged use of lights during the day. In addition, the dishwasher was observed to be a less significant energy consumption factor compared to the washing machine and the dryer.

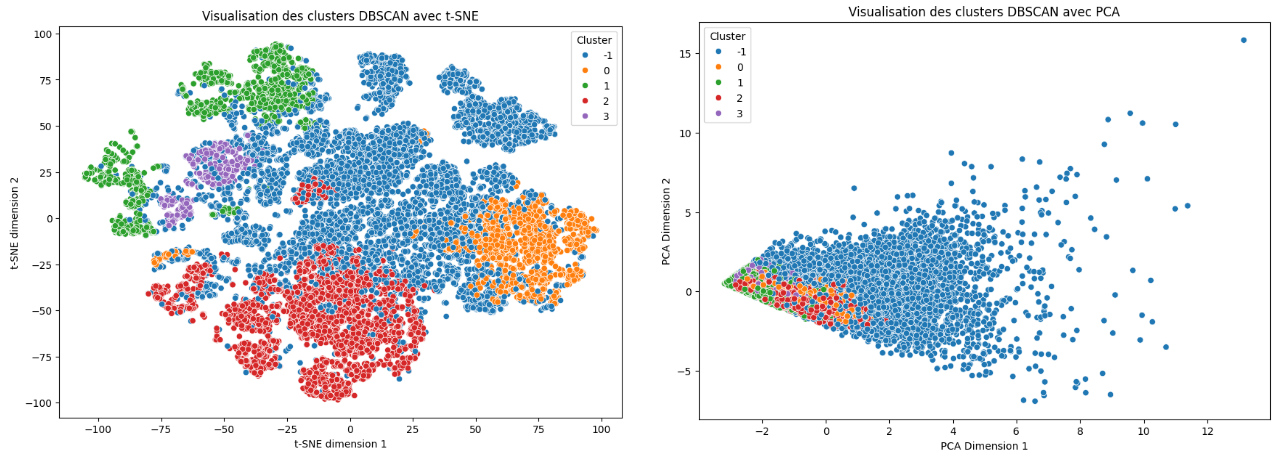
Compared to the previous analysis with 7 clusters, the segmentation into 13 groups allowed to distinguish even more specific behaviors. For example, the impact of the dishwasher compared to other kitchen appliances was better detailed, and the differences between various durations of use of lights became more apparent. These distinctions provide a more nuanced view of energy behaviors and highlight the importance of detailing consumption habits to better target energy efficiency interventions.



*Figure 12 : Visualization of the clustering of behavioral variables into 13 clusters*

### 3.5.5 DBSCAN

The clusters generated by the DBSCAN algorithm are distinguished mainly by the presence or absence of certain elements, such as the swimming pool, the jacuzzi, and the ceiling fan, as well as by the use or non-use of outdoor lights at night and the oven.



*Figure 13 : Visualization of behavioral variables clustering with DBSCAN*

# **4. Discussion**

In this study, we identified several key behaviors influencing residential energy consumption. This section highlights the main findings in relation to relevant data and trends:

1. Pool usage

The analysis revealed that the use of swimming pools is the most energy-intensive factor. This observation is reinforced by market data [9]: between 2019 and 2021, the sale of heat pumps for swimming pools increased by 101%. In the United States, the residential swimming pool market is significant, with 10.7 million swimming pools, of which 10.4 million are residential, and almost 59% are in-ground. In fact, the United States is the country with the most private swimming pools in the world. These figures highlight the importance of energy efficiency in swimming pool heating systems. Strategies such as optimizing heating systems or using thermal blankets could significantly reduce the energy consumption associated with swimming pools.

1. Appliances

Household appliances also account for a significant share of energy consumption. According to the International Energy Agency (IEA), modern, high-efficiency appliances can reduce energy consumption by up to 15% compared to older models [10]. Policies that encourage the purchase of more efficient appliances could have a significant impact on overall energy consumption.

1. Lighting

Prolonged lighting, including outdoor lights left on at night, accounts for 10–15% of average residential electricity consumption [11]. Switching to LED bulbs, which use up to 10 times less energy than incandescent bulbs and 6–8 times less than halogen bulbs [12], and implementing automated lighting control systems can result in significant energy savings.

1. Kitchen

Kitchen appliances, such as ovens and stoves, play a role in household energy consumption. Adopting more energy-efficient cooking practices, such as using convection ovens or induction cooktops, and optimizing the efficiency of these appliances, can reduce this consumption. However, the relatively small impact of cooking on the electricity consumption of American households is explained by their cooking habits. Americans spend on average half as much time at the table as the French (1h02 versus 2h13) [26], cook little [27], and prefer junk food and fast food, especially in low-income families. Fresh produce is expensive and less accessible than processed products and fast food menus, which limits the opportunities to cook [25].

1. Television and interior lighting

Although the impact of television and indoor lighting for short periods is relatively small, these behaviors, when widespread, contribute to energy consumption. Promoting practices such as automatically switching electronic devices to standby could provide modest but significant savings.

# **5. Conclusion**

This study conducted an in-depth analysis of residential energy consumption using preprocessing, feature selection and clustering methods to understand the influence of household behaviors and identify distinct energy consumption groups.

We began with rigorous data preprocessing, reducing the number of samples from 18,496 to 18,403 and the variables from 799 to 310. Four feature selection methods reduced the number of relevant variables to 28, 12 of which were behavioral.

The application of clustering algorithms, such as K-Means and Kernel K-Means, revealed significant variations in energy behaviors. The behaviors were classified according to their energy impact, from the most to the least energy-consuming: use of the swimming pool, household appliances, prolonged lighting of outdoor lights, prolonged lighting of indoor lights, use of kitchen appliances, watching television, and finally, indoor lighting for short periods.

This prioritization highlights the importance of adopting differentiated strategies for energy consumption management. Leisure facilities, particularly swimming pools, require targeted interventions to improve their energy efficiency. Household appliances also need to be optimized to reduce their impact. Finally, although lighting is less energy-intensive than leisure facilities, measures such as the use of LED bulbs and automated control systems can offer substantial savings.

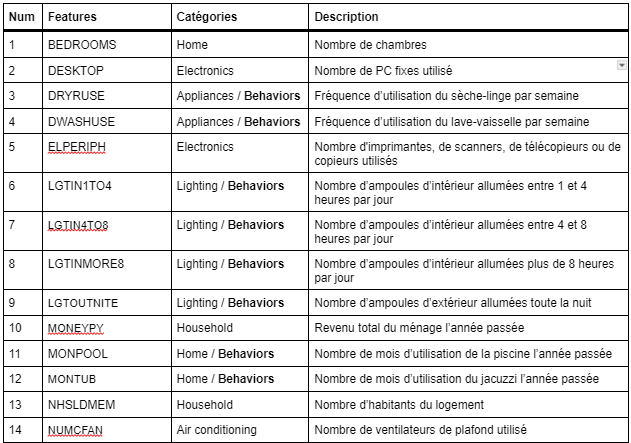
In the future, digital technologies, such as the rise of video streaming and cloud computing, which are very resource-intensive, could increase residential energy consumption. The impact of these technologies on electricity consumption deserves special attention in order to develop effective strategies to counter their effect on energy demand.

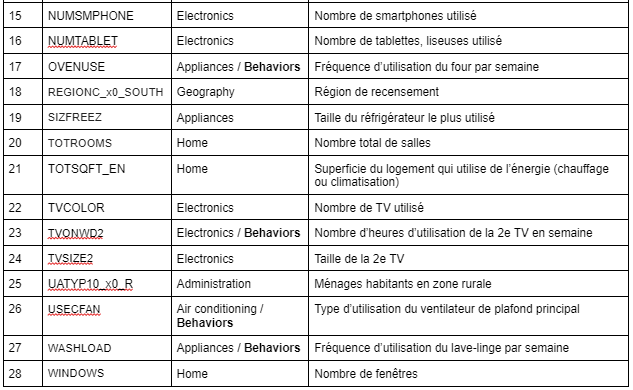
Finally, it is worth noting that electricity consumption prediction could benefit from the integration of emerging digital technologies. Emerging technologies are playing an increasingly crucial role in managing and reducing residential energy consumption. The Internet of Things (IoT) and connected objects enable real-time data collection on energy consumption, providing opportunities for more detailed analyses and more accurate forecasts. Home energy management systems (HEMS) can monitor and optimize energy use, while smart grids enable a more balanced and efficient distribution of electricity. The Internet of Behavior (IoB) could also play a crucial role by enabling a deep understanding of user habits and behaviors. These technologies could transform the way we anticipate energy needs and optimize consumption management, providing powerful tools to improve energy efficiency and reduce household carbon footprints.

Today, as we reach Earth Overshoot Day in 2024, this study highlights the critical environmental impact of our energy behaviors. Earth Overshoot Day marks the date when humanity has consumed all the resources that Earth can renew in one year. Adopting targeted strategies to optimize energy consumption at the household level is essential to foster greater awareness and more sustainable practices, helping to reduce our environmental footprint and preserve the planet’s resources.

# **Appendix**

Appendix A : Table listing selected variables





# **References**

1. U.S. Energy Information Administration, 2020 Residential Energy Consumption Survey. <https://www.eia.gov/consumption/residential/data/2020/>
2. Évolution du nombre d'habitants sur Terre entre 1950 et 2024 et projections jusqu'en 2100. Statista. <https://fr.statista.com/statistiques/564933/population-mondiale-jusqu-en-2080/>
3. Population. Nations Unies. <https://www.un.org/fr/global-issues/population>
4. D. Chareyron, H. Horsin-Molinaro, B. Multon. Concepts et chiffres de l'énergie : la consommation de l’électricité domestique en France. Culture Sciences Physiques, ENS Lyon. <https://culturesciencesphysique.ens-lyon.fr/ressource/chiffres-energie-electricite-domestique.xml>
5. H. M. Sani, S. O. Tehrani, B. Behkamal, H. Amintoosi : Extracting Effective Features for Descriptive Analysis of Household Energy Consumption Using Smart Home Data. 2019.
6. M. Heinrich, M. Ruellan, L. Oukhellou, A. Samé, J.-P Lévy : From energy behaviors to lifestyles: Contribution of behavioral archetypes to the description of energy consumption patterns in the residential sector. 2022.
7. T. F. Sanquist, H. Orr, B. Shui, A. C. Bittner : Lifestyle factors in U.S. residential electricity consumption. 2012.
8. IECC. 2021 International Energy Conservation Code. <https://codes.iccsafe.org/content/IECC2021P2>
9. Analyse du marché : Quelle est la taille du marché mondial de la piscine ?. Aquark. <https://www.aquark.com/fr/analyse-du-marche-mondial-de-la-piscine/>
10. U.S. Energy Information Administration. A call to action on efficient and smart appliances. <https://www.iea.org/articles/a-call-to-action-on-efficient-and-smart-appliances>
11. Engie. LED, halogène, connectée… quelle est la consommation d’une ampoule ?. <https://particuliers.engie.fr/economies-energie/conseils-economies-energie/conseils-calcul-consommation/consommation-ampoule.html>
12. Élisabeth Chesnais. Ampoules LED De belles économies. UFC Que Choisir. <https://www.quechoisir.org/guide-d-achat-ampoules-basse-consommation-led-n11547/>
13. M. Labiadh. Methodology for construction of adaptive models for the simulation of energy consumption in buildings. 2022.
14. H. Elayan, M. Aloqaily, F. Karray, M. Guizani. Decentralized IoB for Influencing IoT-based Systems Behavior. 2024.
15. H. Elayan, M. Aloqaily, F. Karray, M. Guizani. Internet of Behavior and Explainable AI Systems for Influencing IoT Behavior. 2023.
16. J. Wang, X. Chen, F. Zhang, F. Chen, Y. Xin. Building Load Forecasting Using Deep Neural Network with Efficient Feature Fusion. 2021.
17. J. W. Chan, C. K. Yeo. Electrical Power Consumption Forecasting with Transformers. 2022.
18. D. Hadjout, J.F. Torres, A. Troncoso, A. Sebaa, F. Martínez-Alvarez. Electricity consumption forecasting based on ensemble deep learning. 2021.
19. D. Hadjout, A. Sebaa, J. F. Torres, F. Martínez-Álvarez. Electricity consumption forecasting with outliers handling based on clustering and deep learning with application to the Algerian market. 2023.
20. D. Syed, H. Abu-Rub, A. Ghrayeb, S. S. Refaat. Household-Level Energy Forecasting in Smart Buildings Using a Novel Hybrid Deep Learning Model, 2021.
21. M. Alhussein , K. Aurangzeb, S. I. Haider. Hybrid CNN-LSTM Model for Short-Term Individual Household Load Forecasting. 2020.
22. X. Li, Y. Zhong, W. Shang, X. Zhang, B. Shan, X. Wang. Total electricity consumption forecasting based on Transformer time series models. 2022.
23. Y. Liu, D. Zhang, H. B. Gooi. Optimization Strategy Based on Deep Reinforcement Learning for Home Energy Management. 2020.
24. S.-H. Kim, C. Lee, C.-H. Youn. An Accelerated Edge Cloud System for Energy Data Stream Processing Based on Adaptive Incremental Deep Learning Scheme. 2020.
25. EuropUSA. Etats-Unis : les habitudes alimentaires. <https://www.europusa.com/vivre-aux-etats-unis/vie-personnelle-aux-usa/vie-quotidienne/etats-unis-les-habitudes-alimentaires/>
26. SudOuest. Les Français passent deux fois plus de temps à table que les Américains. <https://www.sudouest.fr/premium/art-de-vivre/les-francais-passent-deux-fois-plus-de-temps-a-table-que-les-americains-3136883.php>
27. Christina Gierse. Les Français et la nourriture aux Etats-Unis : un sujet « touchy ». Studyrama. <https://www.studyrama.com/pro/destination/les-francais-et-la-nourriture-aux-etats-unis-un-sujet-touchy-21891.html>
28. Blent. https://blent.ai/blog/a/arbres-de-decision-en-machine-learning