Exploration des Comportements Énergétiques Résidentiels : Identification des Facteurs Clés et Perspectives pour l’Optimisation

Kévin Tang

Mohamed-Essaid Khanouche

Parisa Ghodous

Août 2024

Université Claude Bernard Lyon 1



**Résumé**. La consommation énergétique des ménages représente une part significative de la consommation énergétique globale. Pour approfondir notre compréhension de la production, de la gestion et du stockage des excédents d’énergie, il est crucial d’élucider les schémas implicites des comportements des consommateurs et d’identifier les facteurs influençant leurs performances. Cet article vise principalement à analyser de manière descriptive le modèle de consommation énergétique des ménages en utilisant les données du RECS2020, en mettant l’accent sur l’impact du comportement sur la consommation d'énergie. À cette fin, nous nous concentrons sur la sélection du sous-ensemble de caractéristiques le plus pertinent à partir d’un large ensemble de données, ce qui permet une meilleure compréhension des informations, une réduction du temps de calcul et une amélioration des performances prédictives. Les résultats de cette étude peuvent aider les décideurs à examiner les comportements des familles à divers niveaux de la société pour s’assurer que leur qualité de vie est adéquate ou nécessite des améliorations.

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

# **Sommaire**

[**Sommaire 3**](#_sr7c106oktkx)

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

[**2. Travaux connexes 5**](#_8obrolxg0qqi)

[**3. Méthodologie proposée 6**](#_urcet1nvncnz)

[3.1 Présentation du dataset 6](#_hn0b8h4g5rs2)

[3.2 Prétraitements 7](#_vhvdqzbv6h9y)

[3.2.1 Suppression des variables concernant les énergies autres que l’électricité 7](#_h9c292i5q4o4)

[3.2.2 Suppression des variables indicatrices d’imputation et de calibration 7](#_ip3ji9y7ul7k)

[3.2.3 Gestion des valeurs manquantes 7](#_arr6os2u099l)

[3.2.4 Encodage des variables catégorielles 8](#_t97pc73dozzx)

[3.2.5 Réduction des variables cibles 8](#_mpqsrqk59ldw)

[3.2.6 Réduction du nombre de variables 8](#_ptga0o6sl9ur)

[3.2.7 Réduction du nombre d’échantillons 9](#_7oknh6azls2g)

[3.2.8 Conclusion du pré-traitement 9](#_bkkom789404n)

[3.3 Sélection des variables 10](#_ndiid1z50cuk)

[3.3.1 Corrélation sur l’ensemble des données 10](#_oa6nvlf7npi)

[3.3.2 Corrélation sur les données catégorisées 10](#_hqjdismvl78)

[3.3.3 Gain d’information sur l’ensemble des données 12](#_8tjo4s2pzwyb)

[3.3.4 Gain d’information sur les données catégorisées 13](#_tdi2gu8y5mkj)

[3.3.5 Intersection des 4 méthodes 14](#_4mw8sel8knjb)

[3.4 Analyse descriptive des variables sélectionnées 14](#_ywgwnirghde5)

[3.4.1 Détermination du nombre optimal de clusters 15](#_qzloabjmokzq)

[3.4.2 K-Means 16](#_hb82l8u6z04e)

[3.4.3 Kernel K-Means 17](#_mvvk7bxqdua)

[3.4.4 DBSCAN 18](#_u1mineb8g1m2)

[3.4.3 Analyse des clusters 20](#_c5nxoqv80xgf)

[3.5 Analyse descriptive des variables comportementales sélectionnées 21](#_ttwzph31klbt)

[3.5.1 Détermination du nombre optimal de clusters 22](#_2zq1vu93icns)

[3.5.2 Deux clusters 22](#_2oug6ctsl8yg)

[3.5.3 Sept clusters 22](#_1ad7t07cg0xn)

[3.5.4 Treize clusters 23](#_l7zixlt1ikvk)

[3.5.5 DBSCAN 24](#_o08ul9yxrtt6)

[**4. Discussion 25**](#_r355w4su6hmv)

[**5. Conclusion 26**](#_4bdilksimxm9)

[**Annexes 28**](#_56t5skyox9fm)

[**Références 29**](#_pa4t1a18xbx4)

# **1. Introduction**

La population mondiale a atteint 8 milliards d'habitants en novembre 2022, soit un milliard de plus qu'en 2010, et devrait franchir la barre des 10 milliards avant la fin du siècle selon les prévisions de l’ONU [2,3]. Cette croissance rapide, conjuguée à l'urbanisation et à l'augmentation des besoins en énergie, entraîne une hausse significative de la consommation énergétique des ménages. Actuellement, les ménages représentent 36 % de la consommation d’électricité en France métropolitaine, faisant de ce secteur le plus énergivore en proportion, devant le secteur tertiaire et le secteur industriel [4].

Aujourd'hui, les fournisseurs d’énergie et les entreprises spécialisées dans les objets connectés proposent des systèmes de suivi de la consommation énergétique pour les particuliers. Ces dispositifs permettent aux consommateurs de mieux comprendre et gérer leur usage de l'énergie en leur fournissant des données détaillées et en temps réel sur leur consommation. Toutefois, pour améliorer encore la gestion de l’énergie et optimiser les économies, il est crucial de développer une compréhension plus fine des schémas de consommation des ménages. Cela inclut l'identification des facteurs influençant la consommation d'énergie, tels que les caractéristiques sociodémographiques, les habitudes de vie et l’efficacité des appareils ménagers. Une analyse approfondie de ces comportements et facteurs peut non seulement aider à réduire la consommation d’énergie, mais aussi à adapter les stratégies d'efficacité énergétique aux besoins spécifiques des consommateurs.

En comprenant mieux les comportements de consommation, les décideurs peuvent élaborer des politiques plus ciblées et efficaces. Par exemple, ils pourraient promouvoir des pratiques énergétiques durables, encourager l'adoption d'appareils plus efficaces ou offrir des incitations financières pour des comportements énergétiquement responsables. De plus, les consommateurs eux-mêmes, munis de meilleures informations, peuvent prendre des décisions plus éclairées et ajuster leurs habitudes pour réaliser des économies d'énergie substantielles.

Dans cet article, nous analysons l'ensemble de données RECS2020, qui contient des informations sur la consommation énergétique résidentielle des ménages américains en 2020. En utilisant des techniques de sélection de caractéristiques, nous identifions les variables les plus influentes en ce qui concerne la consommation d'énergie annuelle et le coût payé aux fournisseurs d'énergie. Une fois les caractéristiques pertinentes sélectionnées, nous appliquons des algorithmes de clustering pour partitionner les ménages en groupes significatifs. Ces groupes sont ensuite analysés afin de déduire des informations utiles sur les modes de consommation d'énergie des ménages et leur relation avec des variables telles que les revenus et la taille des espaces de vie.

De plus, nous portons une attention particulière aux caractéristiques comportementales des ménages parmi celles qui ont été sélectionnées. Cette analyse permet de mieux comprendre comment les habitudes et comportements des ménages influencent leur consommation d'énergie, fournissant ainsi des informations précieuses pour le développement de stratégies énergétiques plus efficaces et adaptées aux différents profils de consommateurs.

# **2. Travaux connexes**

Dans cette section, nous examinons les travaux connexes qui identifient les caractéristiques déterminantes de la consommation d'énergie et discutons des méthodes utilisées pour sélectionner ces caractéristiques.

L’étude de [5] se concentre sur l’identification des facteurs influençant la consommation d’énergie des ménages américains à l’aide du jeu de données RECS de 2015, une version antérieure au RECS de 2020. Bien que leur recherche et la méthodologie employée soient similaires à celles de notre étude, ils ne mettent pas particulièrement l’accent sur les comportements des ménages. Leurs résultats de partitionnement se limitent à différencier les ménages en fonction de leur richesse, séparant ainsi les ménages riches des ménages pauvres.

Le travail de [7] a permis en 2011 d’identifier cinq facteurs reflétant le mode de vie des ménages, associés à des comportements tels que l’utilisation de la climatisation, la lessive, l’utilisation d’un ordinateur personnel, la zone climatique de résidence et l’utilisation de la télévision. Ces facteurs ont été estimés à partir des données RECS de 2001 et 2005.

Tout comme notre étude, l’objectif de [6] est de construire des archétypes comportementaux pour mieux comprendre les habitudes de consommation d’énergie dans le secteur résidentiel. Les données utilisées proviennent du projet ENERGIHAB de l’Agence Nationale de la Recherche française et comprennent un ensemble de 35 variables couvrant divers aspects tels que l’hygiène, l’alimentation, le chauffage, l’éclairage, les pratiques de loisirs et l’occupation du logement. Ces données ont été recueillies auprès de 1363 ménages en Île-de-France. Grâce à ces données et à une méthode originale, l’étude a réussi à identifier et construire sept archétypes comportementaux distincts. Cependant, malgré ses contributions, cette recherche présente certaines limites. Le nombre de variables est relativement restreint, ce qui peut limiter la profondeur de l’analyse. De plus, la portée géographique de l’étude est limitée à l’Île-de-France, ce qui peut réduire la généralisation des résultats à l’ensemble du territoire français ou à d’autres contextes géographiques. Enfin, les données datent de 2010 et ont été collectées par téléphone, ce qui pourrait entraîner des biais, des erreurs ou un manque d’informations actualisées.

# **3. Méthodologie proposée**

L’objectif de notre recherche est d'identifier les comportements ayant le plus d’impact sur la consommation énergétique dans les ménages, à partir de l’ensemble de données RECS de 2020. Nous avons donc sélectionné trois variables cibles : la quantité totale d’électricité utilisée en kilowattheures (kWh), la quantité totale d’électricité utilisée en BTU (BTUEL), et le coût total de l’électricité consommée en dollars (DOLLAREL).

Dans cette section, nous décrirons d'abord l’ensemble de données utilisé. Ensuite, nous détaillerons les différents prétraitements appliqués à ce dataset, ainsi que les méthodes de sélection des caractéristiques employées pour identifier les comportements les plus significatifs pour notre recherche. Enfin, nous analyserons les résultats obtenus après l’application des algorithmes de clustering.

## **3.1 Présentation du dataset**

Le RECS (Residential Energy Consumption Survey) est une enquête détaillée menée aux États-Unis, visant à recueillir des informations sur la consommation d'énergie résidentielle des ménages américains. Réalisée par l'Energy Information Administration (EIA), cette enquête est l'une des sources de données les plus complètes et fiables sur l'utilisation de l'énergie dans le secteur résidentiel.

L'édition 2020 du RECS collecte des données sur divers aspects de la consommation énergétique des foyers, y compris les types et quantités de combustibles utilisés, les caractéristiques des logements, les appareils électroménagers, les systèmes de chauffage et de climatisation, ainsi que des comportements récents tels que l’utilisation de l’énergie solaire ou le lieu de recharge des véhicules électriques. L'enquête couvre ainsi un large éventail de 799 variables, permettant une analyse approfondie et multidimensionnelle des habitudes de consommation d'énergie.

Le RECS utilise un échantillon représentatif de 18 496 ménages américains pour s'assurer que les résultats peuvent être généralisés à l'ensemble de la population. Les enquêtes détaillées sont envoyées aux ménages sélectionnés et peuvent être administrées sous forme de questionnaires papier, en ligne, par interviews téléphoniques ou lors de visites sur site.

Pour plus d’informations, vous pouvez consulter le site de l’EIA [1].

## **3.2 Prétraitements**

Pour assurer la qualité et la pertinence des analyses, plusieurs prétraitements ont été appliqués aux données du RECS 2020. Ces étapes sont cruciales pour nettoyer et préparer les données avant l’application des techniques de sélection de caractéristiques et des algorithmes de clustering.

### 3.2.1 Suppression des variables concernant les énergies autres que l’électricité

Étant donné que notre recherche vise à identifier les comportements influençant la consommation d’électricité, nous avons supprimé 117 variables relatives à d'autres types d'énergie, telles que le gaz naturel, le propane, le fioul et le bois. En excluant ces variables non pertinentes, nous avons pu concentrer notre analyse spécifiquement sur les facteurs influençant la consommation d'électricité. Cette démarche permet de réduire la complexité du dataset et facilite l'interprétation des résultats.

### 3.2.2 Suppression des variables indicatrices d’imputation et de calibration

Les variables indicatrices d’imputation signalent les valeurs manquantes ou estimées pour d'autres variables, tandis que les variables de calibration ajustent les poids des réponses pour refléter une population représentative. Afin d’affiner notre analyse et de garantir la pertinence des données, nous avons supprimé 407 de ces variables utilisées par l'EIA. Bien que ces variables soient cruciales pour les processus internes de l’EIA, elles ne sont pas nécessaires pour notre étude, qui se concentre spécifiquement sur les comportements de consommation électrique.

### 3.2.3 Gestion des valeurs manquantes

Pour assurer la qualité et la complétude de notre dataset, nous avons adopté une méthode rigoureuse pour traiter les valeurs manquantes (NaN).

La première étape consiste à identifier toutes les lignes du dataset contenant une ou plusieurs valeurs manquantes. Cette identification est cruciale pour cibler les données nécessitant une imputation. Pour déterminer les valeurs les plus appropriées pour remplacer les NaN, nous avons calculé les distances euclidiennes entre chaque ligne contenant des NaN et les lignes complètes du dataset. Pour ce faire, nous avons utilisé la fonction ‘nan\_euclidean\_distances’ de la bibliothèque ‘sklearn.metrics.pairwise’. Cette fonction calcule les distances en tenant compte des valeurs manquantes, en ignorant les positions de NaN lors du calcul de la distance.

Une fois les distances calculées, chaque valeur NaN est remplacée par la valeur correspondante de la ligne la plus proche, c’est-à-dire celle avec la plus petite distance euclidienne. Cette méthode d'imputation basée sur la similarité garantit que les valeurs manquantes sont remplacées de manière cohérente avec les données existantes, minimisant ainsi les biais et les erreurs potentiels.

### 3.2.4 Encodage des variables catégorielles

Pour préparer notre dataset à une analyse approfondie, nous avons converti les variables catégorielles en une forme que les algorithmes de machine learning peuvent comprendre. Nous avons utilisé la méthode OneHotEncoder de la bibliothèque scikit-learn. Cette technique transforme chaque valeur catégorielle en une nouvelle colonne binaire (0 ou 1), indiquant la présence ou l'absence de cette valeur.

Dans notre dataset, nous avons identifié 7 variables catégorielles nécessitant un encodage. L'application de l'encodage one-hot à ces variables a abouti à la création de 144 nouvelles colonnes. Ce processus a permis de convertir les données catégorielles en une représentation numérique tout en préservant l'intégrité et la signification des informations originales.

### 3.2.5 Réduction des variables cibles

Au début de notre étude, nous avions identifié trois variables cibles principales pour évaluer la consommation d'électricité : la quantité totale d’électricité utilisée en kilowattheures (kWh), la quantité totale d’électricité utilisée en BTU (BTUEL), et le coût total de l’électricité consommée en dollars (DOLLAREL). Cependant, après avoir réalisé une analyse de corrélation entre ces variables, nous avons constaté que certaines d'entre elles étaient fortement corrélées. En particulier, la variable BTUEL était fortement corrélée avec kWh, car ces deux mesures sont interconvertibles (1 kWh équivaut à environ 3 412 BTU).

En conséquence, nous avons décidé de réduire le nombre de variables cibles à deux : kWh et DOLLAREL. Cette réduction permet de simplifier l'analyse tout en conservant les mesures essentielles pour évaluer la consommation d’électricité.

### 3.2.6 Réduction du nombre de variables

Dans le cadre du prétraitement des données, nous avons identifié les paires de variables fortement corrélées dans notre dataset pour réduire la redondance. Cette analyse a conduit à la suppression de 105 variables.

Par exemple, toutes les variables associées aux codes postaux des états étaient fortement corrélées avec celles représentant les noms des états. Nous avons donc conservé uniquement les variables relatives aux noms des états. De même, la variable concernant le climat subarctique était étroitement corrélée avec le code climatique de l'IECC pour l'Alaska [8]. Nous avons retenu l'une de ces deux variables pour éviter la duplication des informations climatiques.

Un autre exemple concerne la variable indiquant la présence de climatiseurs, qui était fortement corrélée avec celle montrant l'utilisation d'électricité pour le refroidissement. Cela a suggéré que tous les climatiseurs utilisent l'électricité comme source d'énergie, justifiant ainsi la suppression de la variable sur l'utilisation de l'électricité pour le refroidissement.

Cette démarche a permis de simplifier le dataset tout en conservant les informations les plus pertinentes pour notre analyse.

### 3.2.7 Réduction du nombre d’échantillons

Les données isolées, souvent considérées comme des points aberrants, peuvent introduire des biais significatifs dans les modèles d'apprentissage automatique et altérer les résultats des analyses statistiques. Pour garantir la précision et la fiabilité des analyses, il est crucial d'identifier et de supprimer ces données aberrantes.

Dans notre étude, nous avons utilisé l'algorithme des K-nearest neighbors (KNN) pour détecter les données isolées. Cette méthode nous a permis d’identifier les échantillons se distinguant considérablement des autres en raison de leur faible densité ou de leur éloignement par rapport aux points voisins. Au total, 93 échantillons ont été classifiés comme isolés.

La suppression de ces points aberrants a conduit à une réduction du nombre total de ménages dans notre dataset, passant de 18 496 à 18 403. En éliminant ces données isolées, nous avons amélioré la qualité du dataset, ce qui permet d’obtenir des analyses plus précises et fiables.

### 3.2.8 Conclusion du pré-traitement

À l'issue des étapes rigoureuses de prétraitement, nous avons considérablement affiné notre dataset pour le rendre plus pertinent et exploitable pour notre analyse.

Le nombre d’échantillons a été réduit de 18 496 à 18 403, suite à la suppression de 93 points aberrants identifiés comme isolés. Cette réduction a permis de maintenir une base de données plus cohérente et représentative.

En ce qui concerne les variables, nous avons simplifié notre dataset en réduisant le nombre total de variables de 799 à 310. Cette diminution a été réalisée en éliminant les variables relatives à d'autres types d'énergie, les variables indicatrices d’imputation et de calibration, ainsi que celles fortement corrélées. Cette démarche a permis de réduire les redondances et de concentrer notre analyse sur les facteurs les plus pertinents.

Ces étapes de prétraitement ont non seulement optimisé la qualité des données, mais ont également facilité des analyses plus précises et significatives. En fin de compte, ces ajustements permettent d’obtenir des résultats plus fiables et d’approfondir notre compréhension des comportements de consommation électrique.

## **3.3 Sélection des variables**

La sélection des variables est une étape cruciale dans le prétraitement des données, visant à améliorer la qualité des modèles prédictifs en identifiant les variables les plus pertinentes. Dans notre étude, nous avons appliqué plusieurs méthodes pour affiner notre choix de variables.

### 3.3.1 Corrélation sur l’ensemble des données

La première méthode de sélection des variables repose sur l’analyse de la corrélation entre toutes les variables du dataset et les variables cibles.

Nous avons d'abord calculé la corrélation entre chaque variable et la variable cible KWH pour identifier celles ayant une relation significative avec la consommation électrique. Les variables dont la corrélation avec KWH était supérieure à la moyenne des corrélations calculées ont été sélectionnées.

De manière similaire, nous avons évalué la corrélation entre chaque variable et la deuxième variable cible, DOLLAREL, pour identifier les variables influençant le coût de la consommation d’électricité. Les variables dont la corrélation avec DOLLAREL était également supérieure à la moyenne des corrélations calculées ont été retenues.

Enfin, les variables pertinentes pour chacune des deux cibles ont été croisées pour obtenir leur intersection, ce qui a permis de sélectionner 106 variables importantes pour les deux variables cibles, garantissant leur pertinence dans les deux contextes d’analyse.

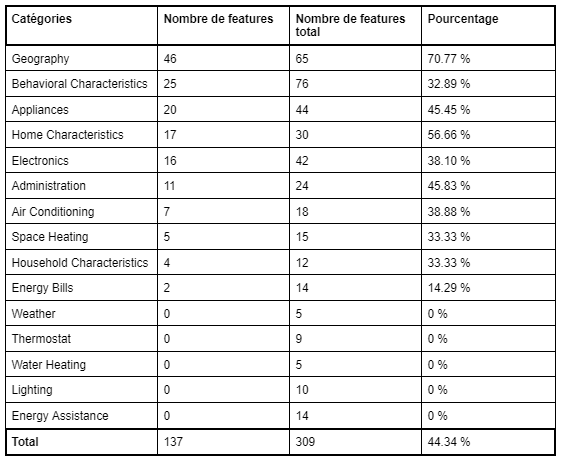
### 3.3.2 Corrélation sur les données catégorisées

La deuxième méthode de sélection des variables repose sur la catégorisation des variables du dataset et l’analyse de la corrélation au sein de ces catégories. Cette approche permet de structurer le processus de sélection en traitant les variables selon leurs catégories spécifiques.

L'EIA a partitionné les variables en 14 catégories distinctes. Nous avons ajouté une catégorie supplémentaire spécifiquement dédiée aux variables liées au comportement, comprenant 88 variables identifiées manuellement. Les 15 catégories sont les suivantes :

* Administration : Cette catégorie regroupe des variables administratives et contextuelles qui fournissent des informations essentielles pour l'identification, la classification climatique et le type urbain des répondants. Elle concerne 4 variables, encodées sur 24 colonnes.
* Géographie : Cette catégorie regroupe des variables géographiques qui fournissent des informations essentielles sur la localisation et les caractéristiques géographiques des répondants telles que l'État de résidence. Elle concerne 5 variables, encodées sur 65 colonnes.
* Météo : Cette catégorie regroupe des variables météorologiques qui fournissent des informations essentielles sur les conditions climatiques et les caractéristiques météorologiques des zones habitées par les répondants telles que la température extérieure, du sol ou des ampoules. Elle concerne 5 variables.
* Caractéristiques du logement : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur les caractéristiques physiques et structurelles des logements des répondants telles que la taille et le type de logement. Elle concerne 30 variables.
* Caractéristiques du ménage : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur les caractéristiques sociodémographiques et économiques des ménages des répondants telles que le nombre d’habitants, leur niveau d’éducation ou les revenus du ménage. Elle concerne 12 variables.
* Electroménager : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur la possession, l'utilisation et les caractéristiques des appareils électroménagers présents dans les logements des répondants. Elle concerne 44 variables.
* Électronique : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur la possession, l'utilisation et les caractéristiques des appareils électroniques présents dans les logements des répondants. Elle concerne 42 variables.
* Éclairage : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur les types, l'utilisation et les caractéristiques des systèmes d'éclairage présents dans les logements des répondants. Elle concerne 10 variables.
* Climatisation : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur les systèmes de climatisation, incluant leur présence, leur utilisation et leurs caractéristiques dans les logements des répondants. Elle concerne 18 variables.
* Chauffage de l’eau : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur les systèmes de chauffage de l'eau, y compris leur type, leur utilisation et leurs caractéristiques dans les logements des répondants. Elle concerne 5 variables.
* Chauffage de l’espace : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur les systèmes de chauffage, incluant leur type, leur utilisation et leurs caractéristiques dans les logements des répondants. Elle concerne 15 variables.
* Thermostat : Cette catégorie regroupe des variables qui fournissent des informations sur les dispositifs de contrôle de température, y compris leur type, leur utilisation et leurs caractéristiques dans les logements des répondants. Elle concerne 9 variables.
* Factures d’énergie : Cette catégorie regroupe des variables qui fournissent des informations détaillées sur la responsabilité du paiement des factures d’énergie, la présence d’un compteur intelligent ou le lieu de recharge du véhicule électrique. Elle concerne 14 variables.
* Aide à l’énergie : Cette catégorie regroupe des variables qui fournissent des informations sur les aides et subventions reçues pour la consommation d'énergie. Elle concerne 14 variables.
* Caractéristiques comportementales : Cette catégorie regroupe des variables de toutes les autres catégories qui fournissent des informations sur les habitudes et les comportements des ménages en matière de consommation d'énergie, y compris les routines quotidiennes ou les préférences d'utilisation. Elle concerne 76 variables.

Pour chaque catégorie, nous avons calculé la corrélation entre les variables et les deux variables cibles, puis sélectionné celles dont la corrélation était supérieure à la moyenne au sein de chaque catégorie. Cette approche a permis de conserver un total de 137 variables.

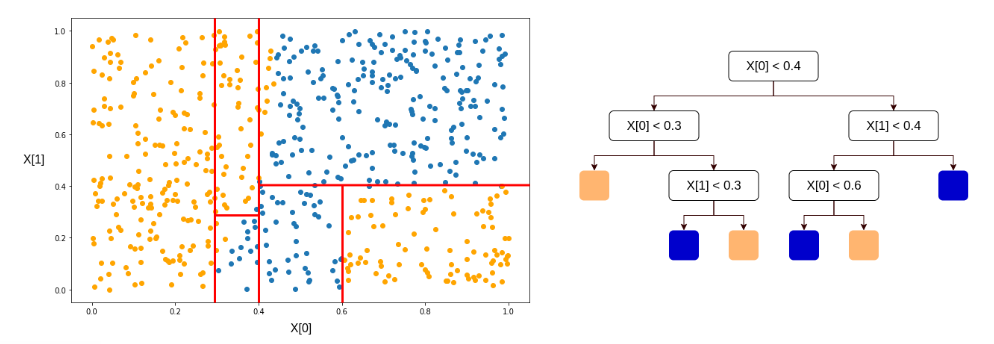


*Figure 1 : Tableau résultat de la méthode 2*

### 3.3.3 Gain d’information sur l’ensemble des données

Le **DecisionTreeRegressor** est un algorithme de régression qui construit un arbre de décision pour prédire des valeurs continues, telles que la consommation d’énergie (KWH) ou le coût (DOLLAREL).

**Fonctionnement de l'algorithme** : L’arbre de décision divise les données en posant des questions successives sur les caractéristiques des données d’entraînement. Pour chaque nouvelle observation, l’arbre suit les branches correspondant aux caractéristiques de l’observation et prédit une valeur basée sur la moyenne des valeurs des données qui arrivent à la feuille terminale.



*Figure 2 : Illustration du fonctionnement de DecisionTreeRegressor*

**Évaluation de l'importance des caractéristiques** : L'algorithme évalue l’importance de chaque caractéristique en mesurant la réduction de l’impureté, quantifiée par l’erreur quadratique apportée par chaque division de l’arbre. À chaque nœud, il sélectionne la caractéristique qui minimise cette impureté, optimisant ainsi les prédictions.

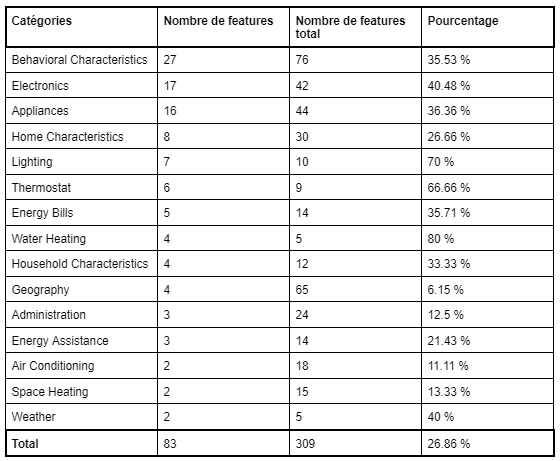
**Sélection des caractéristiques** : Après avoir déterminé l'importance de chaque caractéristique, nous classons ces caractéristiques en fonction de leur contribution à la réduction de l’erreur quadratique globale. Une courbe cumulative est ensuite tracée pour visualiser la contribution de chaque caractéristique. En identifiant le point où la courbe se stabilise, nous pouvons déterminer le nombre optimal de caractéristiques à retenir, équilibrant ainsi précision et simplicité du modèle.

Cette approche permet de sélectionner les variables les plus pertinentes et de construire un modèle de régression efficace, capable de faire des prédictions précises tout en évitant le surapprentissage. Appliquée à l'ensemble des données prétraitées, cette méthode a permis de sélectionner 63 variables.

### 3.3.4 Gain d’information sur les données catégorisées

Comme mentionné précédemment, les données d'origine sont organisées en différentes catégories. Pour chaque catégorie, nous avons appliqué la méthode du coude afin d'identifier le nombre optimal de variables à conserver. Cette approche consiste à tracer la réduction cumulative de l'impureté en fonction du nombre de variables retenues et à déterminer le seuil au-delà duquel l'ajout de variables supplémentaires n'améliore que marginalement la réduction d'impureté.

En appliquant cette méthode de manière individuelle à chaque catégorie, nous avons pu sélectionner 83 variables pertinentes. Cette approche permet d'assurer que les variables choisies contribuent de manière significative à la prédiction tout en maintenant la simplicité et l'efficacité du modèle.



*Figure 3 : Tableau résultat de la méthode 4*

### 3.3.5 Intersection des 4 méthodes

L'intersection des résultats obtenus par les quatre méthodes de sélection a permis d'identifier 28 variables parmi les 309 initiales ayant le plus grand impact sur la consommation d'énergie. Cette sélection conjointe met en lumière les variables les plus significatives en combinant les perspectives des différentes méthodes. Parmi ces 28 variables, 12 proviennent des 76 variables comportementales, soulignant leur importance particulière dans l'analyse de la consommation d'énergie.

La liste complète des 28 variables sélectionnées se trouve dans l'annexe A.

## **3.4 Analyse descriptive des variables sélectionnées**

À l'issue du prétraitement, nous avons sélectionné 28 caractéristiques présentant la plus forte influence sur les attributs cibles. Étant donné que les données ne sont pas étiquetées, nous avons recours à des méthodes de clustering non supervisées pour analyser ces données.

### 3.4.1 Détermination du nombre optimal de clusters

Pour déterminer le nombre optimal de clusters dans notre analyse, nous avons utilisé quatre méthodes complémentaires : la méthode du coude avec l'inertie, le silhouette score, le score de Davies-Bouldin, et le critère de Calinski-Harabasz. Chaque méthode fournit une perspective différente sur la qualité du clustering et aide à identifier le nombre de clusters le plus approprié pour représenter au mieux la structure des données.

**Méthode du Coude avec l'Inertie** : La méthode du coude est une approche classique pour déterminer le nombre optimal de clusters. Elle repose sur l'inertie, également appelée somme des carrés intra-cluster. L'inertie mesure la dispersion des points de données au sein de chaque cluster : plus l'inertie est faible, plus les points sont proches du centre du cluster. En traçant l'inertie en fonction du nombre de clusters, nous obtenons une courbe qui montre généralement une réduction rapide au début, suivie d'une stabilisation. Le "coude" de cette courbe, où la réduction de l'inertie commence à diminuer de manière significative, indique le nombre optimal de clusters. Ce point représente un équilibre entre la compacité des clusters et la complexité du modèle.

**Silhouette score** : Le silhouette score évalue la qualité du clustering en mesurant la similarité des points au sein de leur propre cluster par rapport à la similarité avec les points des autres clusters. Le score varie de -1 à 1 : un score proche de 1 indique que les points sont bien regroupés dans leur propre cluster et éloignés des autres clusters, tandis qu'un score proche de -1 suggère que les points pourraient être mal classifiés. Pour différents nombres de clusters, nous avons calculé le silhouette score moyen. Le nombre de clusters qui maximise ce score est considéré comme optimal, car il indique la meilleure séparation et la plus grande cohésion entre les clusters.

**Score de Davies-Bouldin** : Le score de Davies-Bouldin est une autre mesure de la qualité du clustering, qui évalue la compacité et la séparation des clusters. Il est calculé en mesurant la moyenne des ratios de similarité entre chaque cluster et le cluster le plus similaire. Un score plus bas indique une meilleure séparation entre les clusters, suggérant que les clusters sont bien distincts et compacts. En comparant les scores pour différents nombres de clusters, le nombre de clusters qui minimise le score de Davies-Bouldin est choisi comme optimal, car il reflète la meilleure séparation des groupes.

**Critère de Calinski-Harabasz** : Le critère de Calinski-Harabasz, également connu sous le nom de "variance ratio criterion", évalue la qualité du clustering en comparant la variance intra-cluster à la variance inter-cluster. Plus ce ratio est élevé, meilleure est la séparation entre les clusters et plus les clusters sont compacts. Nous avons calculé ce critère pour différents nombres de clusters et sélectionné le nombre qui maximise le critère de Calinski-Harabasz. Cela indique que les clusters sont bien définis avec une séparation claire et une cohésion interne élevée.

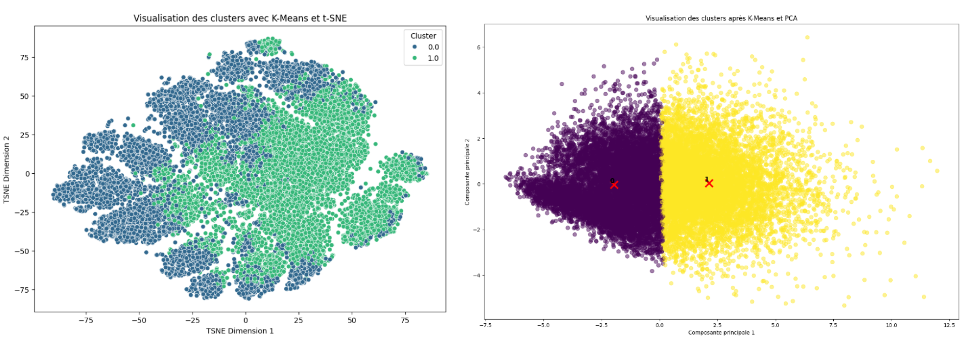
L'application des quatre méthodes de sélection du nombre optimal de clusters a conduit à une conclusion convergente : toutes les méthodes suggèrent que le nombre optimal de clusters est 2. L'accord entre ces différentes méthodes renforce la conclusion selon laquelle deux clusters sont le choix le plus approprié pour capturer la structure sous-jacente des données. Ce consensus multi-méthodes valide la robustesse et la fiabilité de cette solution, garantissant ainsi une représentation adéquate des groupes au sein de notre analyse.

### 3.4.2 K-Means

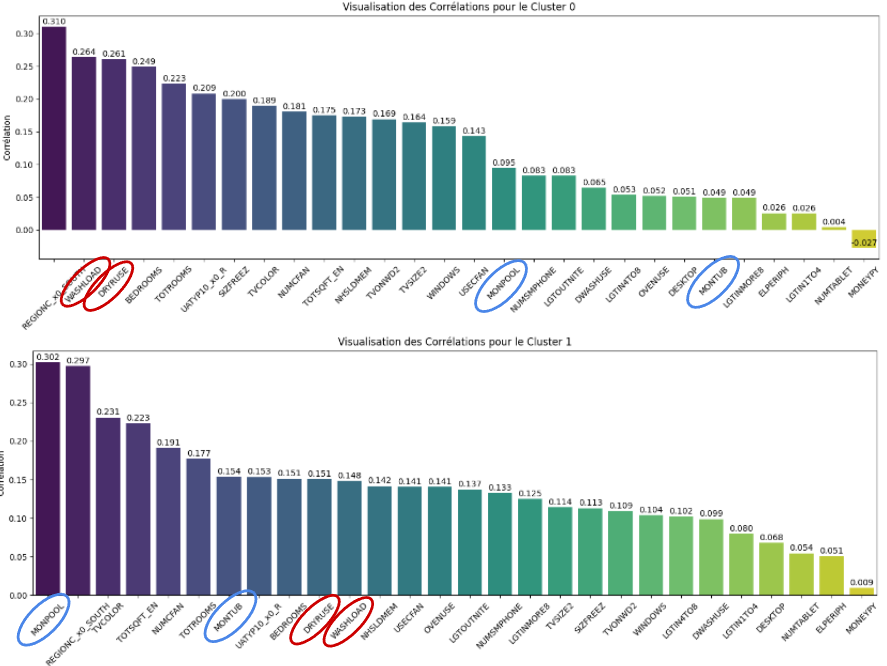
Le K-means est une méthode de clustering non supervisée qui vise à diviser un ensemble de données en un nombre fixe de groupes, appelés clusters. L'objectif est de minimiser la variance au sein des clusters et de maximiser la variance entre les clusters.

L'algorithme K-means fonctionne en plusieurs étapes. L'algorithme commence par l'initialisation des centres des clusters. Ces centres peuvent être déterminés aléatoirement ou en utilisant des techniques spécifiques comme K-means++ pour améliorer la sélection initiale des centres. Chaque point de données est assigné au cluster dont le centre est le plus proche. Cette proximité est généralement mesurée par la distance euclidienne entre les points de données et les centres des clusters. Une fois que tous les points ont été assignés à des clusters, les centres des clusters sont recalculés en prenant la moyenne des points appartenant à chaque cluster. Les étapes d'assignation des points et de mise à jour des centres sont répétées jusqu'à ce que les centres des clusters convergent, c’est-à-dire que les changements deviennent négligeables, ou jusqu’à ce qu’un critère d’arrêt soit atteint (tel qu’un nombre maximal d’itérations). L’algorithme s’arrête donc lorsque les centres des clusters ne varient plus de manière significative, ce qui indique que les clusters ont atteint une forme de stabilité.

Dans le cadre de cette recherche, l'algorithme K-means a été appliqué en utilisant le nombre optimal de clusters déterminé par les méthodes précédentes, établi à deux clusters.



*Figure 4 :* Visualisation du clustering des variables sélectionnées avec K-Means



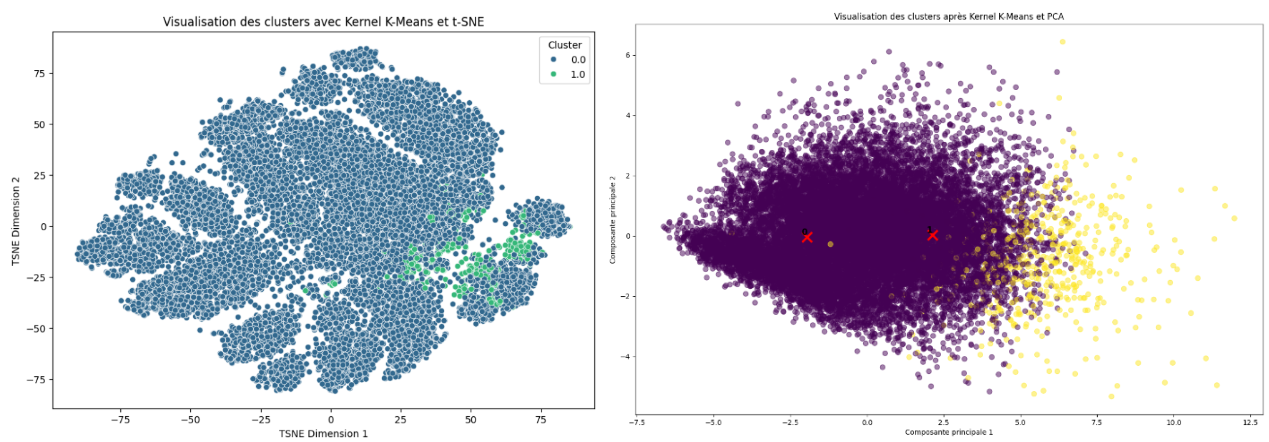
*Figure 5 : Corrélations des Clusters avec K-Means (2 Clusters)*

### 3.4.3 Kernel K-Means

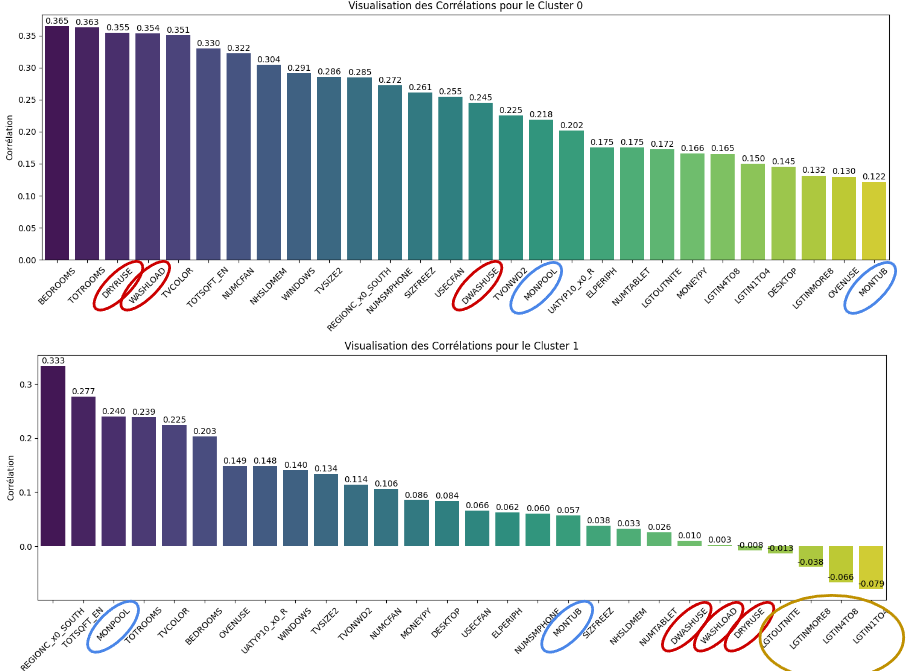
Kernel K-means est une extension de l'algorithme K-means classique qui utilise une fonction noyau pour mapper les données dans un espace de dimension plus élevée, permettant ainsi de capturer des structures non linéaires dans les données. En transformant les données à l'aide d'une fonction noyau, Kernel K-means peut identifier des clusters qui ne seraient pas détectables par K-means classique, car ils se trouvent dans un espace où les frontières entre les clusters ne sont pas linéaires.

L'algorithme suit des étapes similaires à K-means : initialisation des centres de clusters, assignation des points aux clusters en fonction des distances calculées avec la matrice noyau, mise à jour des centres de clusters, et itération jusqu'à convergence. Cependant, au lieu d'utiliser la distance euclidienne traditionnelle, Kernel K-means utilise les distances calculées dans l'espace transformé par la fonction noyau. Cette approche permet de mieux capturer des relations complexes entre les points de données.

Dans notre étude, après avoir appliqué K-means, nous avons utilisé Kernel K-means pour vérifier si des structures plus complexes existaient dans les données. Les résultats ont confirmé que les clusters identifiés étaient robustes et nous ont permis de mieux comprendre les comportements de consommation énergétique des ménages.



*Figure 6 :* Visualisation du clustering des variables sélectionnées avec Kernel K-Means

**

*Figure 7 : Corrélations des Clusters avec Kernel K-Means (2 Clusters)*

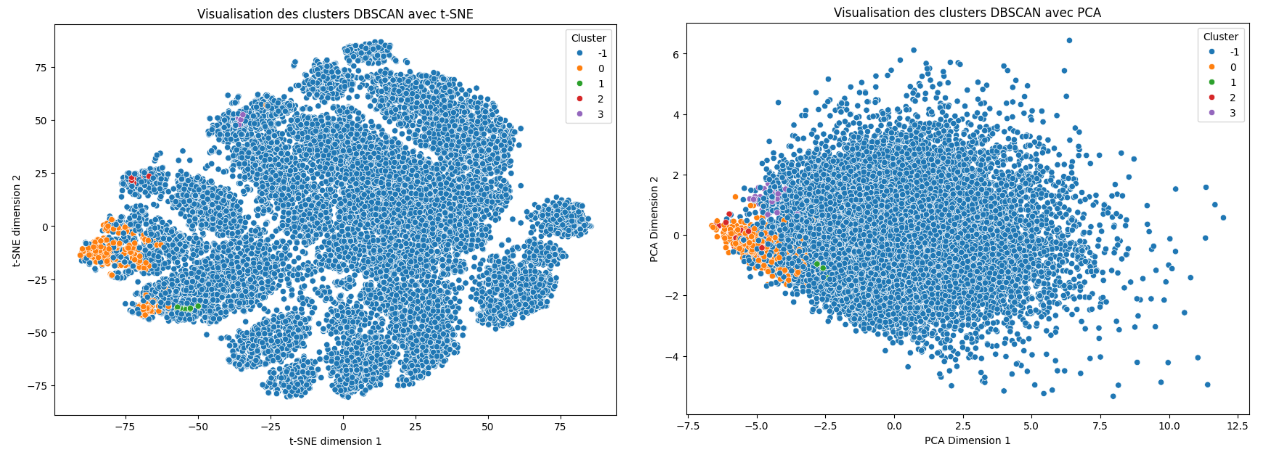
### 3.4.4 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est un algorithme de clustering non supervisé largement utilisé pour identifier des clusters dans des ensembles de données en se basant sur la densité des points de données. Contrairement à d'autres méthodes de clustering telles que K-Means, qui nécessitent de définir le nombre de clusters à l'avance, DBSCAN peut identifier automatiquement le nombre de clusters en fonction de la densité locale des points. L'algorithme repose sur deux principaux paramètres : ε (epsilon), qui représente la distance maximale entre deux points pour qu'ils soient considérés comme voisins, et minPts, le nombre minimum de points dans le voisinage ε pour qu'un point soit considéré comme un point central (core point).

DBSCAN commence par sélectionner un point non visité dans l'ensemble de données et identifie tous les points dans son voisinage ε. Si le nombre de points voisins est supérieur ou égal à minPts, le point est considéré comme un point central et un nouveau cluster est créé. Si le nombre de points voisins est inférieur à minPts, le point est marqué comme du bruit, bien qu'il puisse être inclus plus tard dans un cluster s'il est un voisin d'un point central. Pour chaque point central, l'algorithme explore tous les points dans son voisinage et, si ces points sont également des points centraux, leur voisinage est également exploré. Ce processus continue jusqu'à ce que le cluster ne puisse plus être étendu. L'algorithme répète ces étapes jusqu'à ce que tous les points soient visités.

DBSCAN présente plusieurs avantages. Il peut identifier des clusters de formes arbitraires, contrairement à K-Means qui suppose des clusters sphériques, et traite explicitement les points de bruit, ce qui améliore la qualité des clusters formés. De plus, il détermine automatiquement le nombre de clusters en fonction de la densité des points, ce qui est avantageux lorsque le nombre de clusters n'est pas connu à l'avance. Il est également efficace pour les bases de données contenant de grands volumes de données et fonctionne bien avec les bases de données de densité variable.

L'application de l'algorithme DBSCAN à notre ensemble de données, avec des paramètres ε définis à 2 et minPts fixé à 29, a révélé la formation de quatre clusters distincts. Cependant, il est important de noter que ces clusters sont relativement petits en taille. La majorité des points de données ont été classés comme des anomalies, reflétant la nature stricte de l'algorithme dans la définition des points de cluster en fonction de la densité locale. Cette caractéristique de DBSCAN souligne sa capacité à identifier et isoler les points aberrants dans les ensembles de données, mais elle peut également indiquer que les paramètres ε et minPts pourraient nécessiter un ajustement pour mieux capturer les structures de clusters potentielles présentes dans les données.



*Figure 8 : Visualisation du clustering des variables sélectionnées avec DBSCAN*

### 3.4.3 Analyse des clusters

Les deux méthodes de clustering, K-means et Kernel K-means, ont segmenté les ménages en deux groupes distincts, chacun mettant en évidence des caractéristiques spécifiques de consommation énergétique.

Les ménages appartenant au **Cluster 0** se caractérisent par des logements plus petits, un revenu plus faible, et une possession et utilisation moindre d'appareils électroménagers. Les principaux facteurs de consommation d'énergie dans ce cluster sont liés aux besoins essentiels d'un logement, principalement l'utilisation d'appareils électroménagers pour des tâches quotidiennes comme la cuisson, la réfrigération et le lavage.

Les ménages du **Cluster 1**, en revanche, se distinguent par des logements plus grands, un revenu élevé, et une utilisation beaucoup plus intensive des appareils électroménagers et électroniques. Dans ce cluster, les principaux facteurs de consommation d'énergie incluent des équipements liés aux loisirs, tels que les piscines et les jacuzzis, qui sont largement responsables de la consommation élevée d'électricité.

**K-Means** a identifié deux clusters distincts, mais de manière plus générale, incluant une large gamme de ménages à revenu élevé dans le Cluster 1. Les conclusions montrent une distinction claire entre les ménages à faible revenu avec des besoins essentiels et ceux à revenu élevé avec des équipements de loisir.

**Kernel K-Means** a identifié un sous-groupe spécifique de ménages très riches dans le Cluster 1, réduisant ainsi le nombre total d'échantillons dans ce cluster comparé à K-means. Cela signifie que Kernel K-means a pu distinguer un groupe de ménages extrêmement aisés, dont les habitudes de consommation énergétique sont encore plus marquées par des équipements de loisir et des comportements de consommation luxueux.

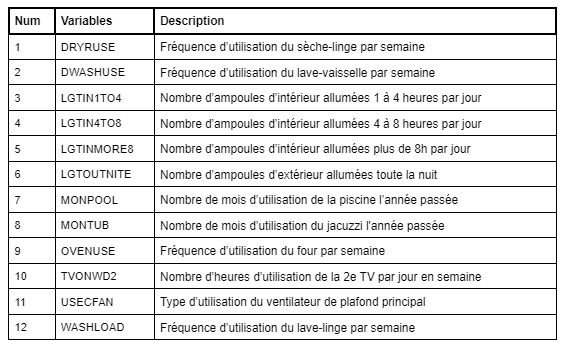
Les différences entre les deux clusters mettent en lumière les divers aspects de la consommation énergétique des ménages. Pour le Cluster 0, la consommation d'énergie est principalement due à des besoins essentiels, reflétant des ménages vivant dans des logements modestes avec des revenus plus faibles. Les stratégies de réduction de la consommation d'énergie dans ce cluster pourraient se concentrer sur l'efficacité énergétique des appareils électroménagers et des pratiques quotidiennes. Pour le Cluster 1, la consommation d'énergie élevée est majoritairement attribuée à des équipements de loisir tels que les piscines et les jacuzzis, caractéristiques de ménages à revenu élevé vivant dans de grands logements. Les efforts pour réduire la consommation d'énergie dans ce groupe pourraient inclure des mesures pour rendre ces équipements plus efficaces ou promouvoir des alternatives moins énergivores.

Pour compléter cette analyse, nous avons également appliqué l'algorithme DBSCAN qui a révélé la formation de quatre clusters distincts, mais il est important de noter que ces clusters sont relativement petits en taille. La majorité des points de données ont été classés comme des anomalies. Contrairement aux résultats obtenus avec les algorithmes précédents, ces quatre clusters identifiés par DBSCAN regroupent des ménages ayant un revenu plus faible que ceux des clusters précédents. L'analyse révèle que, malgré des différences de revenus au sein de ces clusters et de taille de logements, ces ménages partagent des comportements similaires dans leur faible utilisation des équipements et appareils domestiques. Les ménages dans ces clusters sont généralement peu équipés, sans piscine ou ventilateur de plafond. Certains n'ont qu'une seule télévision et vivent seuls ou à deux. Leur seul loisir est souvent cette télévision. Certains ménages ne possèdent même pas de réfrigérateur ou de four.

En somme, cette analyse révèle que les comportements de consommation énergétique diffèrent grandement selon le revenu et la taille du logement, influençant directement les stratégies potentielles d'économie d'énergie à adopter pour chaque groupe. Ces résultats mettent en lumière l'importance de cibler des mesures spécifiques d'efficacité énergétique adaptées aux caractéristiques et aux besoins des différents segments de la population, afin de maximiser les économies d'énergie et d'améliorer la durabilité environnementale.

## **3.5 Analyse descriptive des variables comportementales sélectionnées**

Après avoir segmenté les ménages en deux clusters distincts et analysé les caractéristiques de consommation énergétique, nous avons focalisé notre attention sur les variables comportementales parmi celles qui ont été sélectionnées. Rappelons que l'objectif initial de notre étude était d'analyser l'influence du comportement sur la consommation énergétique. Parmi les 28 variables issues de la sélection des variables, 12 sont spécifiquement liées aux comportements des ménages. En isolant ces variables comportementales, nous pouvons examiner plus en détail comment les habitudes et les pratiques des ménages impactent leur consommation d'énergie. Cette étape est cruciale pour comprendre les facteurs comportementaux spécifiques qui contribuent à la variation de la consommation d'énergie entre différents ménages.



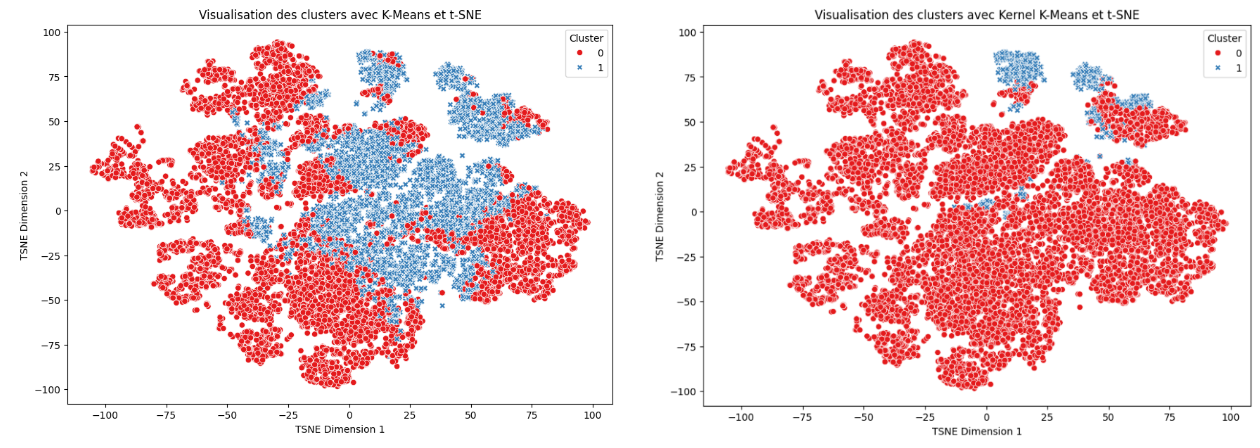
*Figure 9 : Tableau listant les variables comportementales*

### 3.5.1 Détermination du nombre optimal de clusters

Comme précédemment, nous avons déterminé le nombre optimal de clusters en utilisant plusieurs méthodes d’évaluation. La méthode du coude et le critère de Calinski-Harabasz ont toutes deux indiqué que le nombre optimal de clusters est 2. En revanche, le silhouette score a suggéré 7 clusters, tandis que le Davies-Bouldin index a recommandé 13 clusters. Ces divergences soulignent l’importance de considérer plusieurs critères pour une évaluation robuste du nombre de clusters, car chaque méthode utilise des critères différents et peut interpréter les données de manière distincte.

### 3.5.2 Deux clusters

L'analyse a révélé deux clusters principaux, comme observé précédemment. Les résultats obtenus avec les méthodes K-means et Kernel K-means montrent une similarité avec les clusters identifiés dans les analyses antérieures. Les deux méthodes ont confirmé la séparation entre les ménages à revenu faible et élevé. Kernel K-means met une nouvelle fois en évidence un sous-groupe spécifique parmi les ménages les plus aisés.



*Figure 10 : Visualisation du clustering des variables comportementales en 2 clusters*

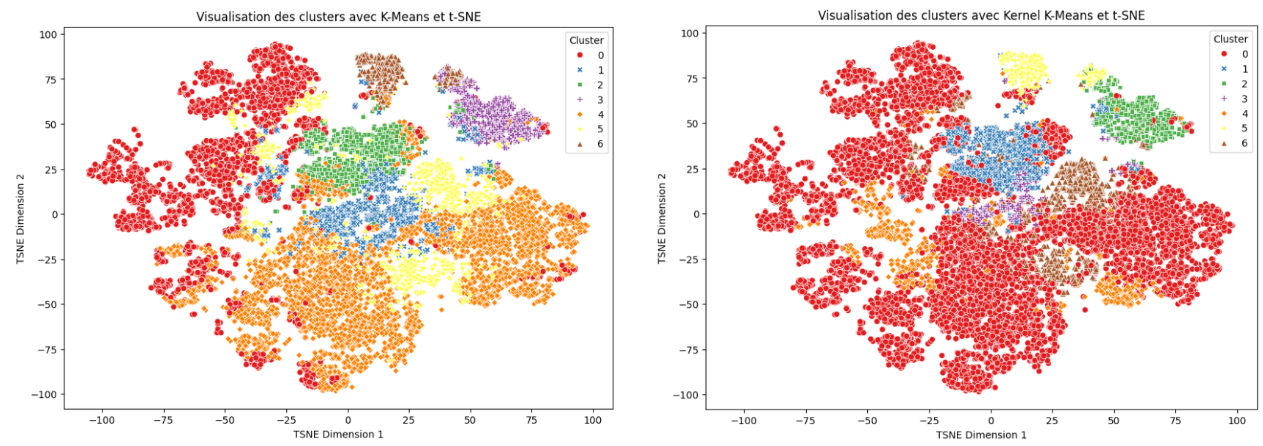
### 3.5.3 Sept clusters

En appliquant les algorithmes de K-means et de Kernel K-means pour segmenter les données en 7 clusters, nous avons approfondi l'analyse des comportements de consommation énergétique des ménages. Les résultats montrent que les deux méthodes de clustering produisent des profils similaires en termes de segmentation.

Les clusters révèlent des habitudes distinctes de consommation d'énergie. Par exemple, le Cluster 0 regroupe les ménages qui utilisent tous les équipements de manière modérée. Le Cluster 1 se caractérise par une utilisation élevée des lumières, tandis que le Cluster 2 est constitué de ménages utilisant intensivement les appareils de cuisine. Le Cluster 3 se distingue par une consommation importante liée à la piscine et au jacuzzi. Le Cluster 4 comprend les ménages utilisant peu la piscine, le jacuzzi, et les appareils de cuisine, tandis que le Cluster 5 se caractérise par une forte utilisation des appareils électroménagers et de la télévision. Enfin, le Cluster 6 représente les ménages avec une consommation élevée du jacuzzi, de la piscine, et des lumières, mais une faible utilisation de la télévision.

L’analyse met en évidence plusieurs conclusions clés concernant les facteurs de consommation d’énergie. La piscine se révèle être le facteur le plus déterminant de la consommation d'énergie, en particulier dans les clusters où son utilisation est prédominante. En comparaison, l’utilisation des lumières et du four apparaît comme des facteurs de consommation d'énergie moins significatifs, bien que l’utilisation du four ait un impact plus important que celle des lumières. Les appareils électroménagers sont un facteur plus important dans la consommation d'énergie par rapport au four. En revanche, la télévision influence peu la consommation d'énergie, comme le montre son impact limité dans les clusters.

Ces résultats soulignent la diversité des comportements de consommation énergétique parmi les ménages et mettent en lumière l'importance d'examiner les équipements spécifiques pour élaborer des stratégies efficaces de réduction de la consommation d'énergie.



*Figure 11 : Visualisation du clustering des variables comportementales en 7 clusters*

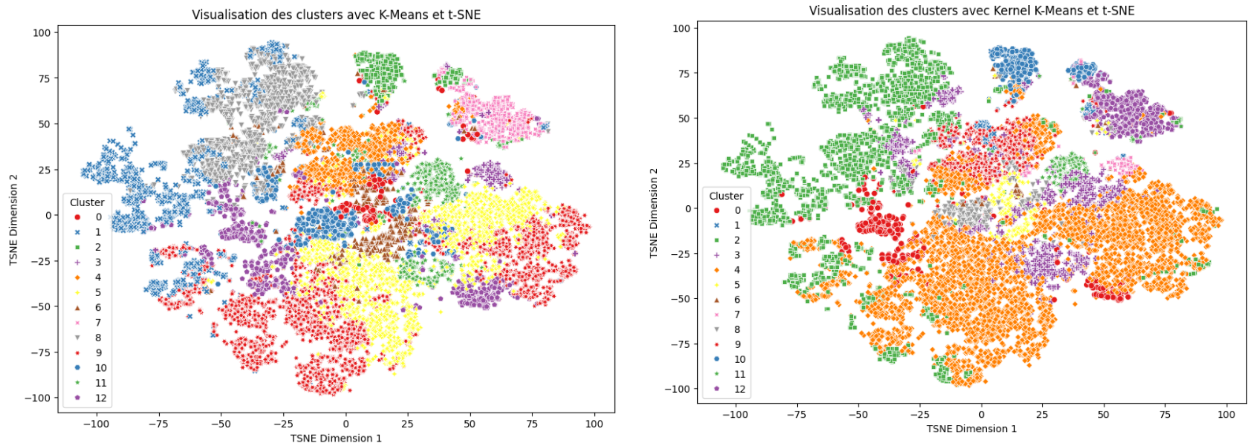
### 3.5.4 Treize clusters

L'application des algorithmes K-means et Kernel K-means pour segmenter les données en 13 clusters a permis d'obtenir une vue plus détaillée des comportements de consommation énergétique des ménages.

Avec K-means, la piscine a été identifiée comme le facteur le plus significatif de consommation d'énergie, soulignant son impact prédominant. Les appareils électroménagers se sont révélés cruciaux pour la consommation énergétique, surpassant les lumières en termes d'importance. L'utilisation du four a été jugée plus influente que la télévision. En revanche, laisser les lumières allumées toute la nuit est devenu un facteur de consommation plus important que l'utilisation prolongée des lumières pendant la journée. À l'inverse, laisser les lumières allumées pendant une courte période (moins de 4 heures par jour) est considéré comme un facteur de consommation moins significatif que la télévision, qui reste le facteur ayant le moins d’impact.

Avec Kernel K-means, des conclusions similaires ont été observées, mais avec des nuances additionnelles. Les appareils électroménagers continuent de jouer un rôle prééminent par rapport aux lumières. L'utilisation du four est plus significative que celle de la télévision, mais l'importance de l'utilisation prolongée des lumières (moins de 8 heures par jour) s'est révélée supérieure à celle du four. Laisser les lumières allumées toute la nuit reste un facteur de consommation plus significatif que l'utilisation prolongée des lumières durant la journée. De plus, il a été observé que le lave-vaisselle représente un facteur de consommation d'énergie moins important comparé au lave-linge et au sèche-linge.

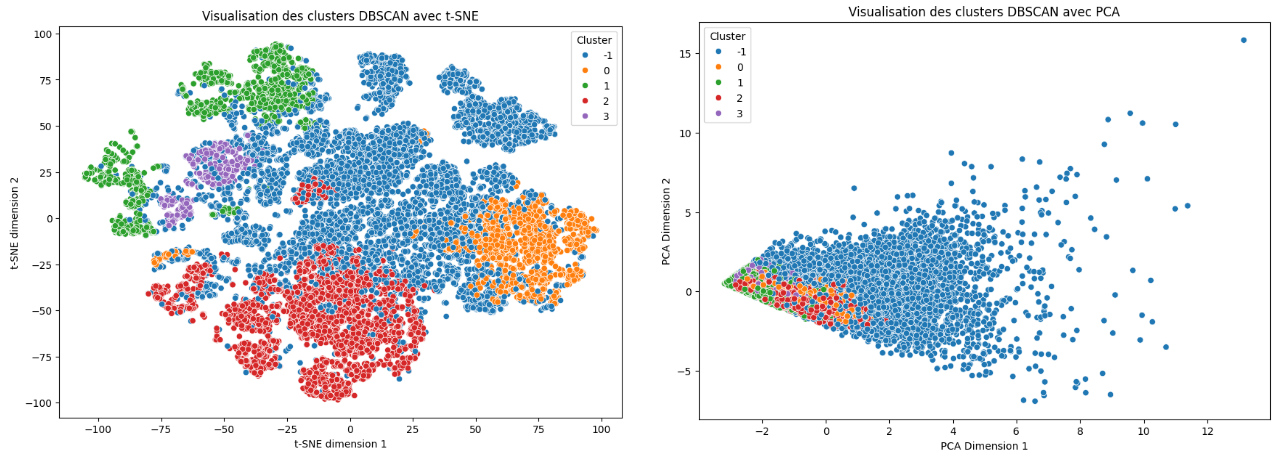
Comparativement à l'analyse précédente avec 7 clusters, la segmentation en 13 groupes a permis de distinguer des comportements encore plus spécifiques. Par exemple, l'impact du lave-vaisselle par rapport aux autres appareils de cuisine a été mieux détaillé, et les différences entre diverses durées d'utilisation des lumières sont devenues plus apparentes. Ces distinctions fournissent une vue plus nuancée des comportements énergétiques et soulignent l'importance de détailler les habitudes de consommation pour mieux cibler les interventions en matière d'efficacité énergétique.



*Figure 12 : Visualisation du clustering des variables comportementales en 13 clusters*

### 3.5.5 DBSCAN

Les clusters générés par l'algorithme DBSCAN se distinguent principalement par la présence ou l'absence de certains éléments, tels que la piscine, le jacuzzi, et le ventilateur de plafond, ainsi que par l'utilisation ou la non-utilisation des lumières extérieures la nuit et du four.



*Figure 13 : Visualisation du clustering des variables comportementales avec DBSCAN*

# **4. Discussion**

Dans cette étude, nous avons identifié plusieurs comportements clés influençant la consommation énergétique résidentielle. Cette section met en lumière les résultats principaux en lien avec des données et tendances pertinentes :

1. Utilisation de la piscine

L'analyse a révélé que l'utilisation des piscines est le facteur le plus énergivore. Cette observation est renforcée par des données du marché [9] : entre 2019 et 2021, la vente de pompes à chaleur pour piscines a augmenté de 101%. Aux États-Unis, le marché des piscines résidentielles est significatif, avec 10,7 millions de piscines, dont 10,4 millions sont résidentielles, et près de 59% sont enterrées. En effet, les États-Unis sont le pays avec le plus de piscines privées au monde. Ces chiffres soulignent l'importance de l'efficacité énergétique des systèmes de chauffage des piscines. Des stratégies telles que l'optimisation des systèmes de chauffage ou l'utilisation de couvertures thermiques pourraient réduire considérablement la consommation énergétique associée aux piscines.

1. Les appareils électroménagers

Les appareils électroménagers représentent également une part importante de la consommation d'énergie. Selon l’Agence Internationale de l’Énergie (AIE), les appareils modernes à haute efficacité peuvent réduire la consommation d'énergie jusqu'à 15% par rapport aux modèles plus anciens [10]. Les politiques encourageant l'achat d'appareils plus efficaces pourraient avoir un impact significatif sur la consommation globale d'énergie.

1. Eclairage

L'éclairage prolongé, notamment les lumières extérieures laissées allumées la nuit, représente entre 10 et 15 % de la consommation d'électricité résidentielle moyenne [11]. L'adoption d'ampoules LED, qui consomment jusqu'à 10 fois moins d'énergie que les ampoules incandescentes et 6 à 8 fois moins que les ampoules halogènes [12], ainsi que la mise en place de systèmes de contrôle d'éclairage automatisés, peuvent entraîner des économies d'énergie significatives.

1. Utilisation des appareils de cuisine

Les appareils de cuisine, tels que les fours et les cuisinières, jouent un rôle dans la consommation d'énergie domestique. Adopter des pratiques de cuisson plus économes en énergie, comme l'utilisation de fours à convection ou de cuisinières à induction, et optimiser l'efficacité de ces appareils, peut réduire cette consommation.

Cependant, l'impact relativement faible de la cuisine sur la consommation électrique des ménages américains s'explique par leurs habitudes culinaires. Les Américains passent en moyenne deux fois moins de temps à table que les Français (1h02 contre 2h13) [26], cuisinent peu [27], et préfèrent la malbouffe et les fast-foods, surtout dans les familles à revenu modeste. Les produits frais sont chers et moins accessibles que les produits industriels et les menus de fast food, ce qui limite les occasions de cuisiner [25].

1. Télévision et éclairage intérieur

Bien que l'impact de la télévision et de l'éclairage intérieur pendant de courtes périodes soit relativement faible, ces comportements, lorsqu'ils sont généralisés, contribuent à la consommation d'énergie. Promouvoir des pratiques telles que la mise en veille automatique des appareils électroniques pourrait permettre des économies modestes mais significatives.

# .**5. Conclusion**

Cette étude a réalisé une analyse approfondie de la consommation énergétique résidentielle en utilisant des méthodes de prétraitement, de sélection de caractéristiques et de clustering pour comprendre l'influence des comportements des ménages et identifier des groupes distincts de consommation énergétique.

Nous avons commencé par un prétraitement rigoureux des données, réduisant le nombre d'échantillons de 18 496 à 18 403 et les variables de 799 à 310. Quatre méthodes de sélection des caractéristiques ont permis de réduire le nombre de variables pertinentes à 28, dont 12 étaient comportementales.

L'application des algorithmes de clustering, tels que K-Means et Kernel K-Means, a révélé des variations significatives dans les comportements énergétiques. Les comportements ont été classés selon leur impact énergétique, du plus au moins énergivore : l'utilisation de la piscine, les appareils électroménagers, l'éclairage prolongé des lumières extérieures, l'éclairage prolongé des lumières intérieures, l'utilisation des appareils de cuisine, le visionnage de la télévision, et enfin, l'éclairage intérieur pendant de courtes périodes.

Cette hiérarchisation met en évidence l'importance d'adopter des stratégies différenciées pour la gestion de la consommation d'énergie. Les équipements de loisirs, en particulier les piscines, nécessitent des interventions ciblées pour améliorer leur efficacité énergétique. Les appareils électroménagers doivent également être optimisés pour réduire leur impact. Enfin, bien que l'éclairage soit moins énergivore que les équipements de loisirs, des mesures telles que l'utilisation d'ampoules LED et de systèmes de contrôle automatisés peuvent offrir des économies substantielles.

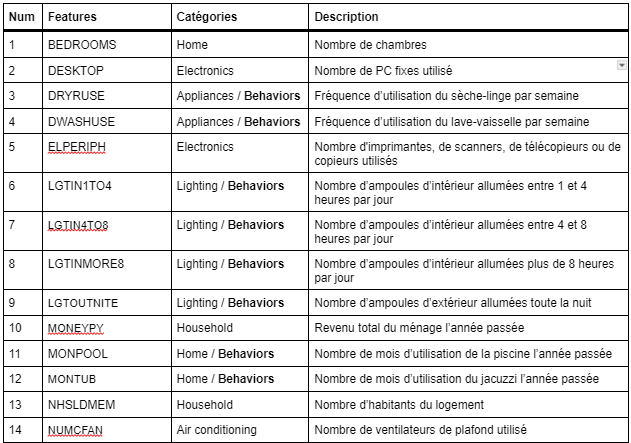
A l’avenir, les technologies numériques, telles que l'essor du streaming vidéo et du cloud computing, très gourmands en ressources, pourraient accroître la consommation énergétique résidentielle. L'impact de ces technologies sur la consommation d'électricité mérite une attention particulière afin d'élaborer des stratégies efficaces pour contrer leur effet sur la demande énergétique.

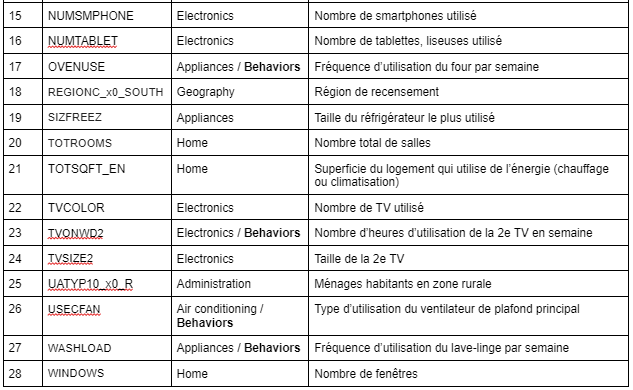
Pour finir, il est pertinent de noter que la prédiction de la consommation électrique pourrait bénéficier de l'intégration des technologies numériques émergentes. Les technologies émergentes jouent un rôle de plus en plus crucial dans la gestion et la réduction de la consommation d'énergie résidentielle. L'Internet des objets (IoT) et les objets connectés permettent une collecte de données en temps réel sur la consommation énergétique, offrant des opportunités pour des analyses plus fines et des prévisions plus précises. Les systèmes de gestion de l'énergie domestique (HEMS) peuvent surveiller et optimiser l'utilisation de l'énergie, tandis que les réseaux intelligents (smart grids) permettent une distribution plus équilibrée et efficace de l'électricité. L'Internet du comportement (IoB) pourrait également jouer un rôle crucial en permettant une compréhension approfondie des habitudes et des comportements des utilisateurs. Ces technologies pourraient transformer la manière dont nous anticipons les besoins énergétiques et optimisons la gestion de la consommation, en fournissant des outils puissants pour améliorer l'efficacité énergétique et réduire l'empreinte carbone des ménages.

Aujourd'hui, alors que nous atteignons le Jour du Dépassement de cette année 2024, cette étude souligne l'impact environnemental crucial de nos comportements énergétiques. Le Jour du Dépassement marque la date à laquelle l'humanité a consommé toutes les ressources que la Terre peut renouveler en une année. Adopter des stratégies ciblées pour optimiser la consommation d'énergie au niveau des ménages est essentiel pour favoriser une prise de conscience accrue et des pratiques plus durables, contribuant ainsi à réduire notre empreinte écologique et à préserver les ressources de la planète.

# **Annexes**

Annexe A : Tableau listant les variables sélectionnées





# **Références**

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