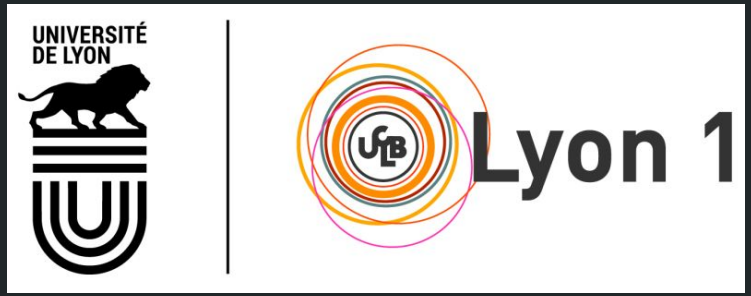
Soutenance de stage

Kévin TANG
M2 Data Science
Université Claude Bernard Lyon 1
2023-2024



Sommaire

1. Introduction
2. Description de la structure
3. Environnement de travail
4. Les missions
5. Les problèmes rencontrés
6. Bilan

Introduction

Contexte

Croissance démographique

- La population mondiale a atteint 8,2 milliards d'habitants en 2024
- Estimation de 10 milliards avant 2100 selon l'ONU
- Besoins énergétiques plus élevés

Internet of Things (IoT)

- Essor des objets connectés : 50 milliards en 2025, 100 milliards en 2030
- Collecte de données massives

Internet of Behaviors (IoB)

- Analyser les comportements, les habitudes des utilisateurs dans un but précis
- Objectif d'optimiser la consommation énergétique

Objectifs du stage

1. **Analyser l'influence des comportements sur la consommation d'énergie**

   Étudier comment les différents comportements affectent la consommation d'énergie résidentielle.

2. **Identifier des modèles de consommation distincts**

   Repérer des tendances et des profils de consommation d'énergie parmi les ménages.

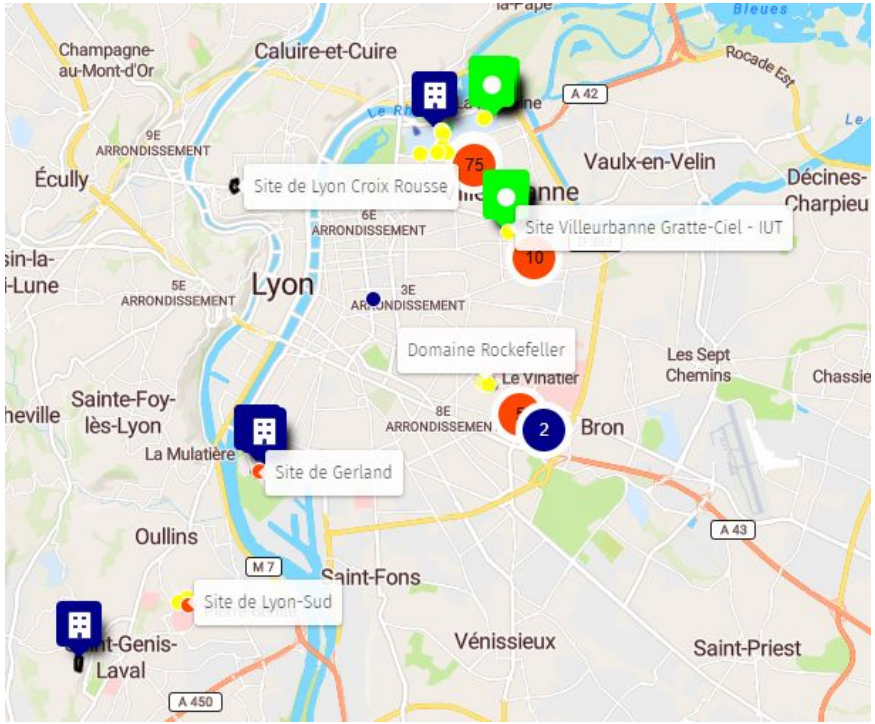
3. **Proposer des recommandations ciblées**

   Offrir des conseils pour réduire la consommation d'énergie en fonction des résultats obtenus.

Description de la structure

L’Université Claude Bernard Lyon 1

L’Université Claude Bernard Lyon 1

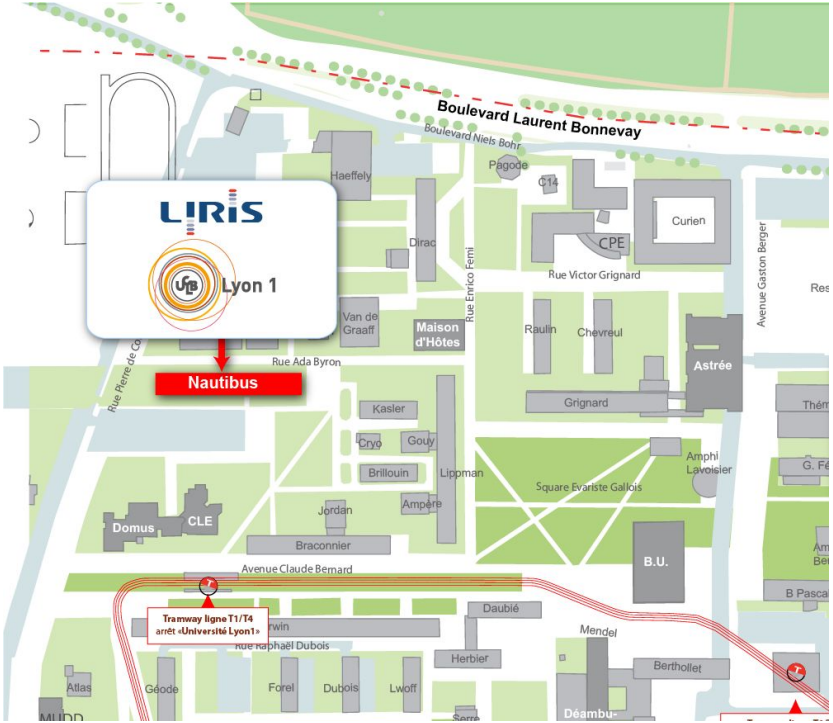


- Fondée en 1971
- 5 sites dans la métropole de Lyon (+3 hors Lyon)
- Plus de 47000 étudiants sur ces différents campus
- 85 laboratoires et plateformes de recherche

Campus LyonTech-LaDoua

- 22000 étudiants
- 1500 chercheurs et 1300 doctorants
- Établissements présents : Faculté des Sciences et Technologies, UFR STAPS, IUT Lyon 1, INSA Lyon, CPE Lyon, ENSSIB, PolyTech Lyon…
- Partenaires de recherche : CNRS, l’INRAE, et l’INRIA…

Laboratoire LIRIS



- Fondé en 2003 sous la tutelle du CNRS, de l'INSA Lyon, de l'Université Claude Bernard Lyon 1, de l'Université Lumière Lyon 2, et de l'École Centrale de Lyon.
- 330 membres
- Domaines de recherche : IA, Analyse de données massives, Vision par ordinateur, Cybersécurité...

Environnement de travail

Lieu et matériel

Lieu

- En présentiel dans les BOX du LIRIS
- En distanciel

Outils

- Contacts : Mails
- Réunions : Zoom
- Développement : Google Colab ou Jupiter Notebook
- Présentations : Google Slides ou PowerPoint
- Rédaction de l’article : Overleaf

Les missions

Etat de l’art

Internet of Behaviors (IoB)

- Concept de l’IoB par Gote Nyman
- Applications possibles (santé, commerce, marketing, énergie…)
- Enjeux éthiques

Extraction de features dans le domaine énergétique

- Identification de patterns
- Etude des styles de vie
- Construction d’archétypes comportementaux
- Clustering

Prédiction de la consommation électrique

- LSTM
- Hybride CNN-LSTM
- Transformers

Les moyens d’influences de l’Internet of Things

- Framework IoB décentralisé
- Explainable AI

Le dataset RECS2020

Qu'est-ce que le RECS 2020 ?

- Description : Le RECS (Residential Energy Consumption Survey) est une enquête menée par l'Energy Information Administration (EIA) aux États-Unis pour collecter des données sur la consommation d'énergie des ménages.

- Objectif : Comprendre les habitudes de consommation d'énergie et les caractéristiques des ménages américains.

- Période : Données collectées en 2020.

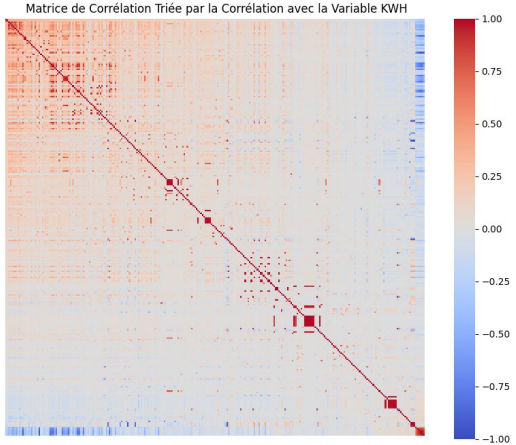
RECS 2020

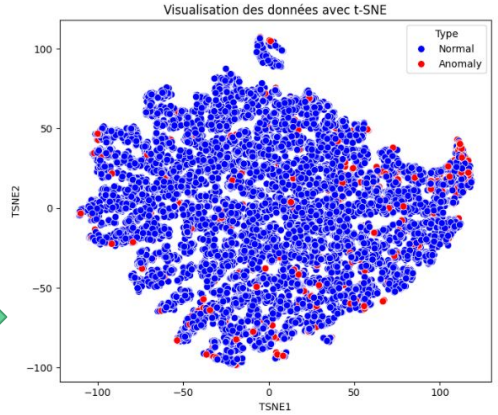
Structure du dataset

- Nombre de variables : 799
- Nombre de lignes : 18496
- Méthode de collecte : Le RECS utilise un échantillon représentatif des ménages américains pour s'assurer que les résultats peuvent être généralisés à l'ensemble de la population.
  Des enquêtes détaillées sont envoyées aux ménages sélectionnés. Ces enquêtes peuvent être administrées sous forme de questionnaires papier, en ligne, des interviews téléphoniques ou des visites sur site.

Les pré-traitements

01	Suppression des variables concernant les énergies autres que l'électricité	• Identification manuelle • Suppression de 117 variables
02	Suppression des variables indicatrices d'imputation et de calibration	• Identification manuelle • Suppression de 407 variables
03	Encodage des variables catégorielles	• LabelEncoder • Aucun changement en nombre de variables
04	Gestion des valeurs manquantes	• Identification des NaN • Suppression de 290 échantillons, soit 1.57% du dataset
05	Réduction des variables cibles	• Suppression d'une des variables cibles • 1 kWh vaut 3412.14 BTU
06	Réduction du nombre de variables	• Etude de corrélation • Suppression de 69 variables
07	Réduction du nombre d' échantillons	• IsolationForest • Suppression de 894 échantillons, soit environ 5% du dataset





Conclusion du pré-traitement

Taille du dataset après pré-traitement

- Nombre d’échantillons : 18496 -> 17312
- Nombre de variables : 799 -> 208

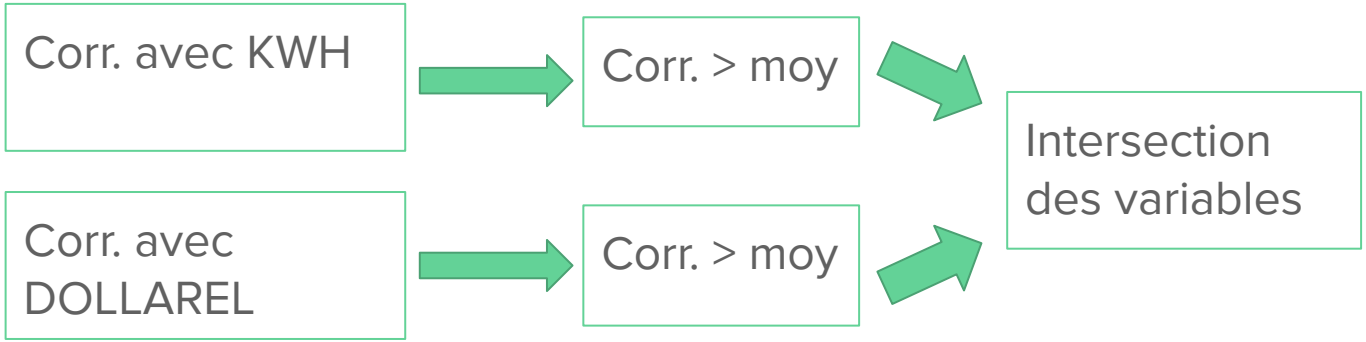
Sélection de features

Sélection de features avec 4 méthodes :

- Corrélation sur l’ensemble des données
- Corrélation sur les données catégorisées
- Gain d’information sur l’ensemble des données
- Gain d’information sur les données catégorisées

Méthode 1 : Corrélation sur l'ensemble des données

- Calcul de la corrélation entre toutes les variables et la variable target 'KWH'
- Calcul de la corrélation entre toutes les variables et la variable target 'DOLLAREL'
- Sélection des variables avec une corrélation supérieure à la moyenne
- Intersection des variables
- Résultat : 79 variables retenues sur les 208 variables



Méthode 2 : Corrélation sur les données catégorisées

- Catégorisation des variables :
  Dans le dataset d'origine, les variables sont déjà toutes catégorisées

2	Variable	Type	Description and Labels	Response Codes	Section
210	TEMPGONEAC	Num	Summer thermostat setting or temperature in home when no one is home during the day	50 - 90 -2 Not applicable	THERMOSTAT
211	TEMPNITEAC	Num	Summer thermostat setting or temperature in home at night	50 - 90 -2 Not applicable	THERMOSTAT
212	H2OAPT	Num	Water heating equipment serves multiple housing units in building	1 Yes 0 No -2 Not applicable	WATER HEATING
156	SSTV	Num	Use smart speakers to control TV or peripherals	1 Yes 0 No -2 Not applicable	ELECTRONICS
157	SSOTHER	Num	Use smart speakers to control some other device	1 Yes 0 No -2 Not applicable	ELECTRONICS
158	HEATHOME	Num	Space heating equipment used	1 Yes 0 No	SPACE HEATING

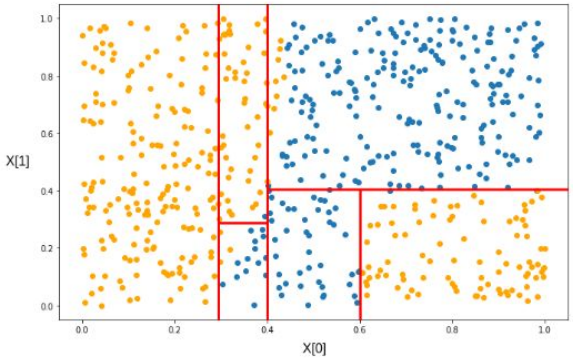
- Application de la méthode 1 dans chaque catégorie

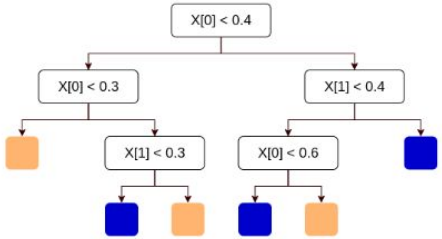
Méthode du coude appliquée au gain d’information

Définition du DecisionTreeRegressor : algorithme de machine learning qui utilise un arbre de décision pour prédire des valeurs continues, ici la valeur de KWH ou DOLLAREL.

Principe du DecisionTreeRegressor : Un arbre de décision est une structure arborescente qui pose des questions sur les caractéristiques des données d'entraînement. À chaque question, l'arbre divise les données en fonction des réponses.

Importance des variables : Calcul de l’erreur quadratique moyenne avant et après chaque division. L'importance des caractéristiques représente la contribution relative de chaque caractéristique à la réduction de l'erreur quadratique.

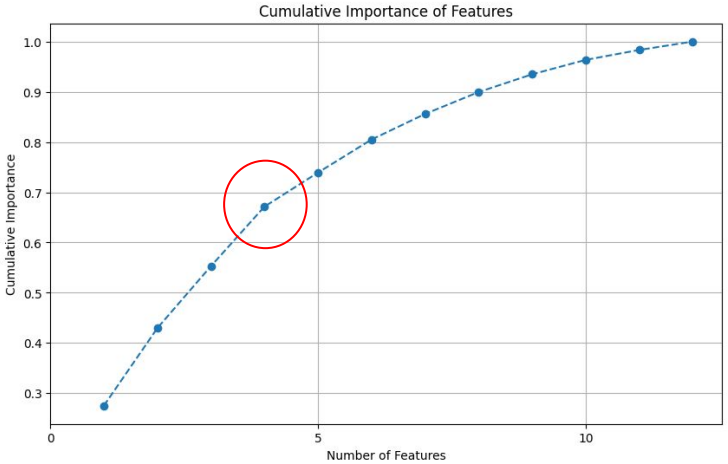


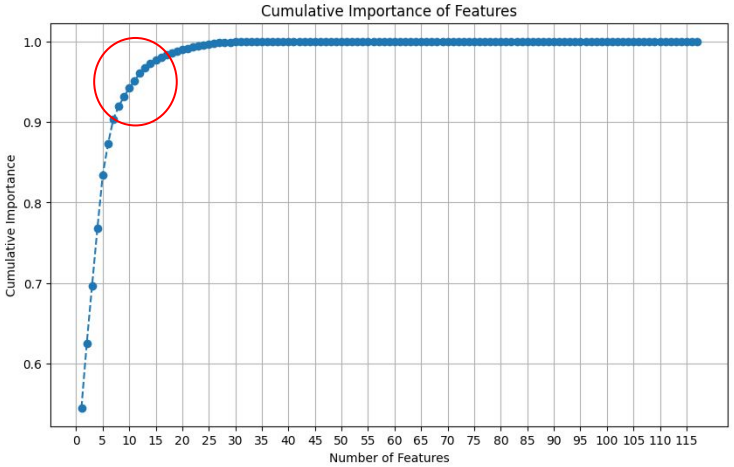


Méthode du coude appliquée au gain d’information

Méthode du coude

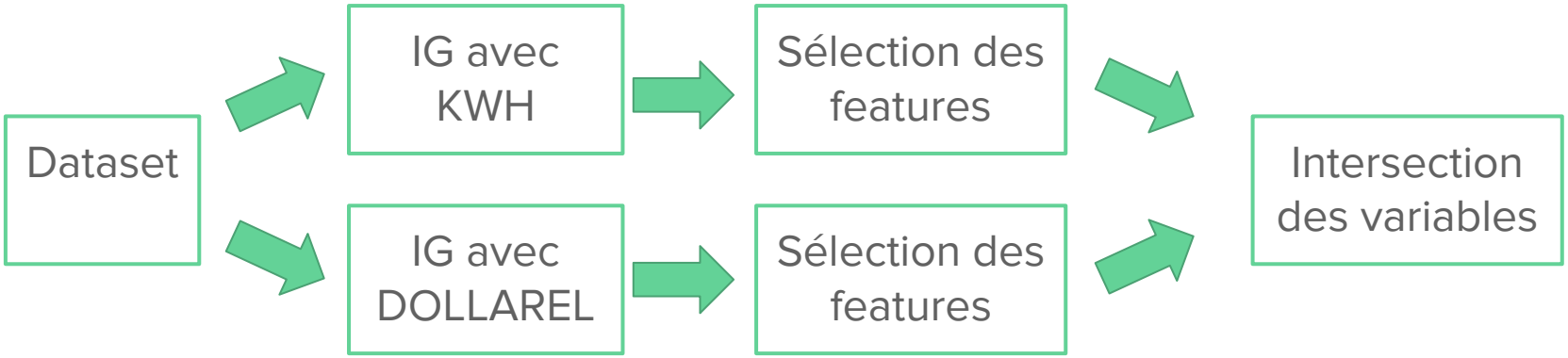
Après le calcul de l'importance des variables, celles-ci sont triées par ordre décroissant et leurs importances sont cumulées.





Méthode 3 : Gain d'information sur l'ensemble des données

La méthode précédemment décrite a été utilisée sur l'ensemble des données pré-traitées.
Résultat : 44 variables retenues sur les 309 variables.



Méthode 4 : Gain d’information sur les données catégorisées

Comme indiqué précédemment, les données d’origine sont catégorisées.

La méthode 3 a donc été appliqué dans chaque catégorie

Intersection des 4 méthodes

L'intersection des quatre méthodes a révélé **17 variables**, parmi les 206 variables d'origine, qui influencent le plus la consommation d'énergie.
Parmi celles-ci, **9 sont issues des 65 variables comportementales**.

Num	Features	Catégories	Description
1	BEDROOMS	Home	Nombre de chambres
2	COMBODVR	Electronics	Nombre de décodeurs câble ou satellite avec DVR utilisés
3	DRYRUSE	Appliances / **Behaviors**	Fréquence d'utilisation du sèche-linge par semaine
4	DWASHUSE	Appliances / **Behaviors**	Fréquence d'utilisation du lave-vaisselle par semaine
5	LGTIN1TO4	Lighting / **Behaviors**	Nombre d'ampoules d'intérieur allumées entre 1 et 4 heures par jour
6	LGTIN4TO8	Lighting / **Behaviors**	Nombre d'ampoules d'intérieur allumées entre 4 et 8 heures par jour
7	LGTINMORE8	Lighting / **Behaviors**	Nombre d'ampoules d'intérieur allumées plus de 8 heures par jour
8	MONEYPY	Household	Revenu total du ménage l'année passée
9	MONPOOL	Home / **Behaviors**	Nombre de mois d'utilisation de la piscine l'année passée
10	NHSLDMEM	Household	Nombre d'habitants du logement
11	SIZFREEZ	Appliances	Taille du réfrigérateur le plus utilisé
12	TOTROOMS	Home	Nombre total de salles
13	TOTSQFT_EN	Home	Superficie du logement qui utilise de l'énergie (chauffage ou climatisation)
14	TVCOLOR	Electronics	Nombre de TV utilisé
15	TVONWD2	Electronics / **Behaviors**	Nombre d'heures d'utilisation de la 2e TV en semaine
16	USECFAN	Air conditioning / **Behaviors**	Type d'utilisation du ventilateur de plafond principal
17	WASHLOAD	Appliances / **Behaviors**	Fréquence d'utilisation du lave-linge par semaine

Les variables comportementales sélectionnées

Num	Variables	Description
1	DRYRUSE	Fréquence d'utilisation du sèche-linge par semaine
2	DWASHUSE	Fréquence d'utilisation du lave-vaisselle par semaine
3	LGTIN1TO4	Nombre d'ampoules d'intérieur allumées 1 à 4 heures par jour
4	LGTIN4TO8	Nombre d'ampoules d'intérieur allumées 4 à 8 heures par jour
5	LGTINMORE8	Nombre d'ampoules d'intérieur allumées plus de 8h par jour
6	MONPOOL	Nombre de mois d'utilisation de la piscine l'année passée
7	TVONWD2	Nombre d'heures d'utilisation de la 2e TV par jour en semaine
8	USECFAN	Type d'utilisation du ventilateur de plafond principal
9	WASHLOAD	Fréquence d'utilisation du lave-linge par semaine

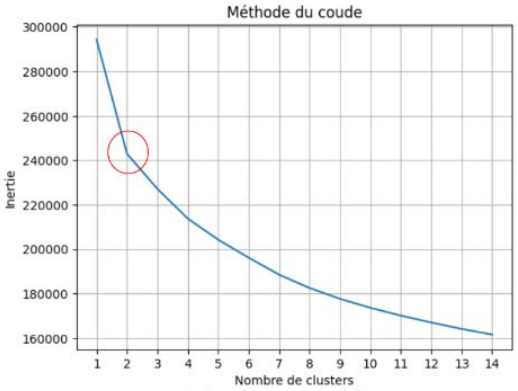
Analyse des données : les méthodes utilisées

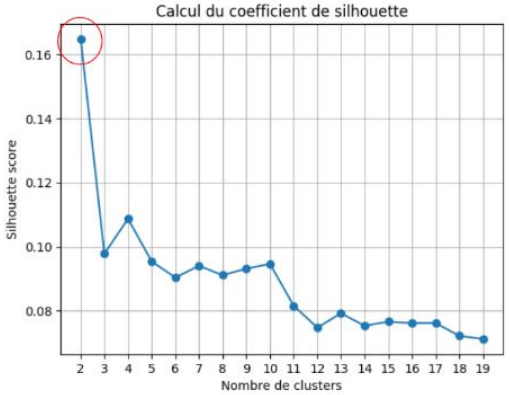
- Clustering : K-Means, Kernel K-Means et DBSCAN
+ Détermination du nombre de clusters

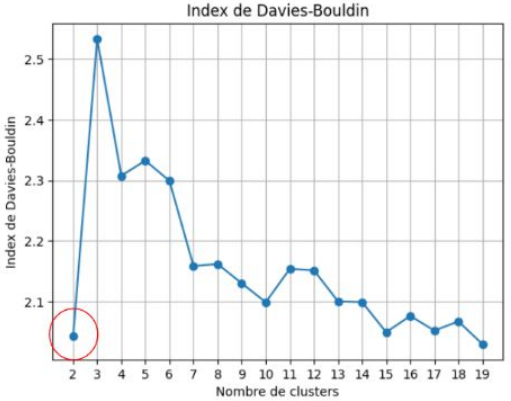
- Analyse statistique : distribution, proportions, moyennes...

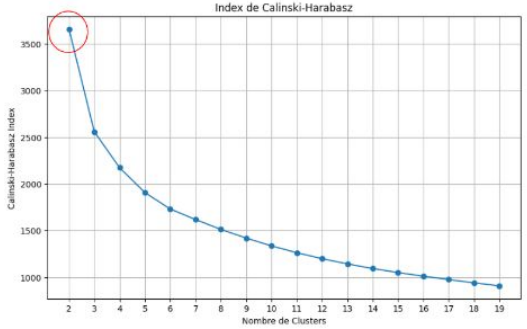
- Analyse des corrélations

Clustering : Détermination du nombre de clusters

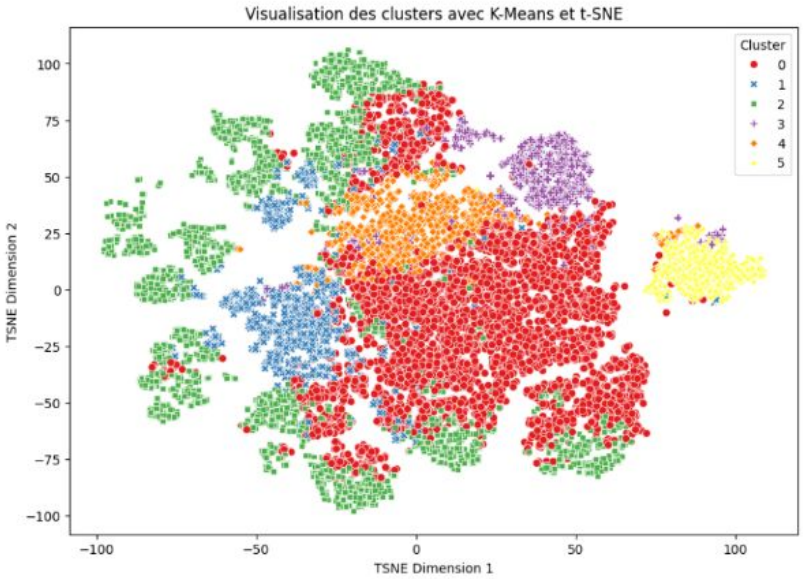


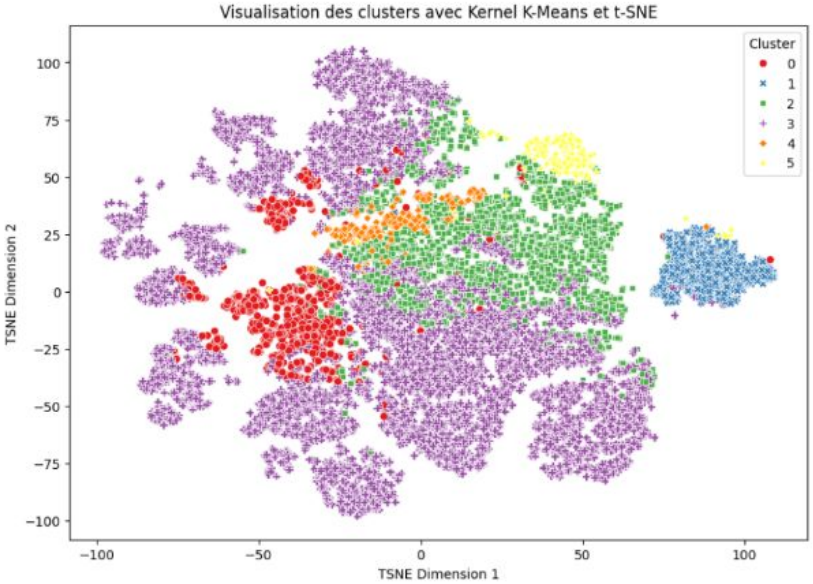






Clustering : K-Means, Kernel K-Means et DBSCAN

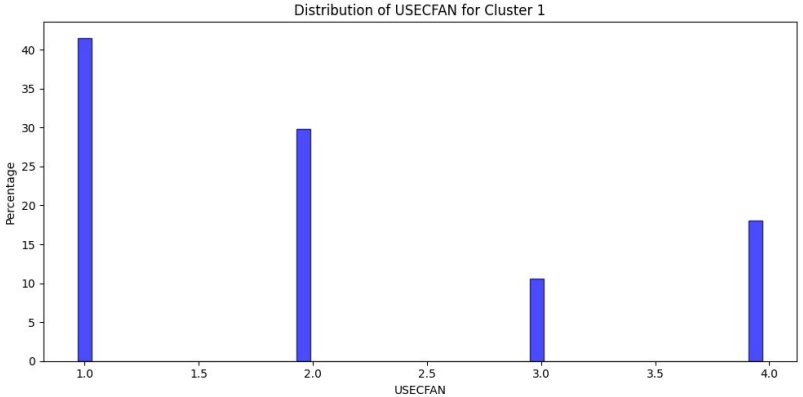




Analyse statistique : distribution, proportions, moyennes…

Catégories	Variables	Cluster 0	Cluster 1	Cluster 2
Home	MONPOOL	0.632653	0.388889	5.191740
Electronics	TVONWD2	3.198671	3.674543	3.071584
Appliances	DRYRUSE	2.499880	6.092226	4.739821
	DWASHUSE	1.908110	4.626524	3.710581
	WASHLOAD	2.499824	6.004934	4.718101
Lightning	LGTIN1TO4	3.657475	8.539834	6.761062
	LGTIN4TO8	2.303228	6.158655	4.573255
	LGTINMORE8	1.057168	3.154578	1.967552
Air conditioning	USECFAN	2.194599	2.674675	2.685466

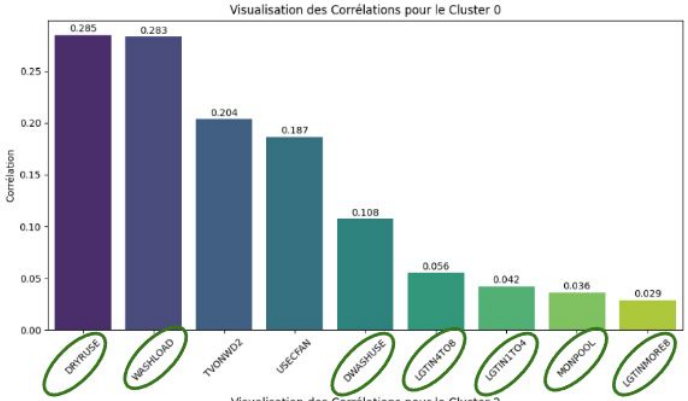
	Moy KWH	Moy DOLLAREL	Classement Énergivore
Cluster 0	8558.481171	1163.097435	3
Cluster 1	13027.244041	1648.425821	2
Cluster 2	17686.589371	2248.897148	1

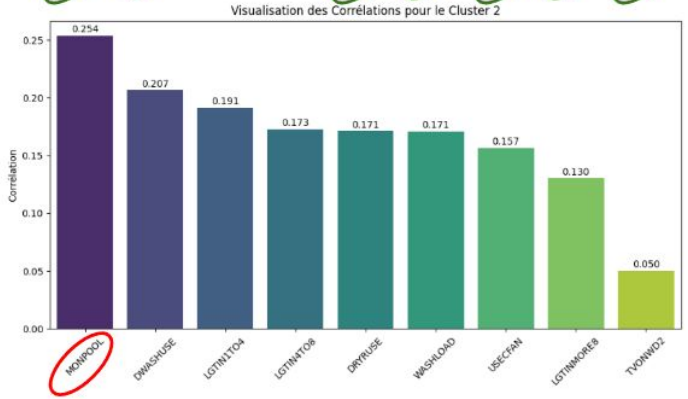


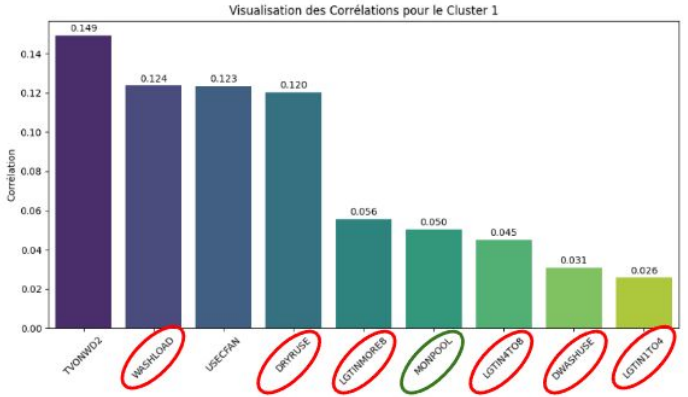
Cluster 1 :

	prediction	USECFAN	count	percentage
9	1	1	39	41.489362
10	1	2	28	29.787234
11	1	3	10	10.638298
12	1	4	17	18.085106

Analyse des corrélations







	Moy KWH	Moy DOLLAREL	Classement Énergivore
Cluster 0	8558.481171	1163.097435	3
Cluster 1	13027.244041	1648.425821	2
Cluster 2	17686.589371	2248.897148	1

Exemples de résultats

- **La piscine** : le facteur le plus influent sur la consommation électrique

Éléments de contexte : USA 1er marché mondial de la piscine avec 10.4 millions de piscines résidentielles (France 2e avec environ 3.2 millions de piscines privées) souvent équipées de pompe à chaleur.

Qui est concerné ? Plutôt dans les ménages aisés et ayant de grandes maisons, ce qui explique aussi le fait qu'ils ont aussi une utilisation intensive des lumières.

Exemples de résultats

- **Les lumières** : le type d’ampoule importe plus que son utilisation

Éléments de contexte :

- Environ 37% des ménages utilisent peu ou pas du tout d’ampoules LED durant la journée
- 14% des ménages qui laissent allumé des lumières toute la nuit n’utilisent pas d’ampoules LED

Exemples de résultats

- **Les appareils électroménagers et de cuisine** :

Éléments de contexte :
- “les Américains ont de très belles cuisines, mais ne cuisinent pas”
- Les Américains passent en moyenne deux fois moins de temps à table que les Français (1h02 contre 2h13)
- Les produits frais sont chers et moins accessibles que les produits industriels et les menus de fast food

Qui est concerné ? Les familles avec enfants utilisent jusqu’à 2 fois par jour les appareils électroménagers (lave-vaisselle, lave-linge, sèche-linge...) et ont une utilisation plus importante de la cuisine, contrairement au reste de la population américaine.

Exemples de résultats

- **La télévision** :

Éléments de contexte : moins de 10% des ménages possèdent désormais des téléviseurs à plasma ou tube cathodique.

Qui est concerné ?
- Les retraités ont tendance à rester jusqu'à 10h par jour devant leur télévision
- Les familles avec enfants utilisent plusieurs télévisions, dont souvent un est utilisé pour les jeux-vidéos

Présentation des résultats

- Réunion toutes les 1-2 semaines (par Zoom si distanciel)
- Présentation PowerPoint de ce que j'ai fait/modifié à chaque réunion
- Écriture de l'article final via Overleaf

Google Slides

Les problèmes rencontrés

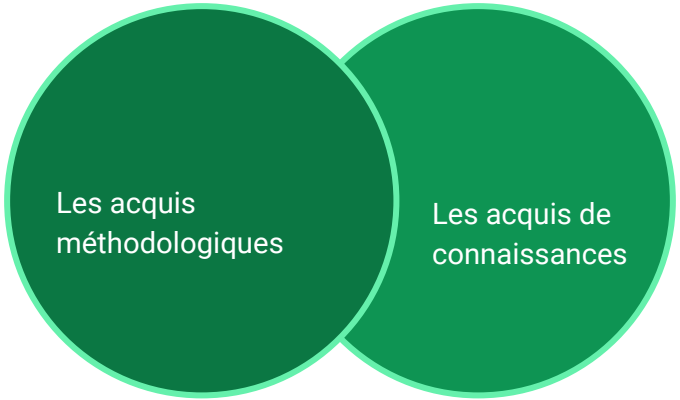
Problèmes rencontrés

01	Diversité et précision des données	• Peu de données provenant d'objets connectés dans RECS2020
02	Données non temporelles	• Impossibilité d'identifier des tendances saisonnières • Impossibilité de construire des modèles prédictifs précis
03	Pas d'accès aux données voulues	• Données d'EDF non disponible pour cause de sécurité et confidentialité

Bilan du stage

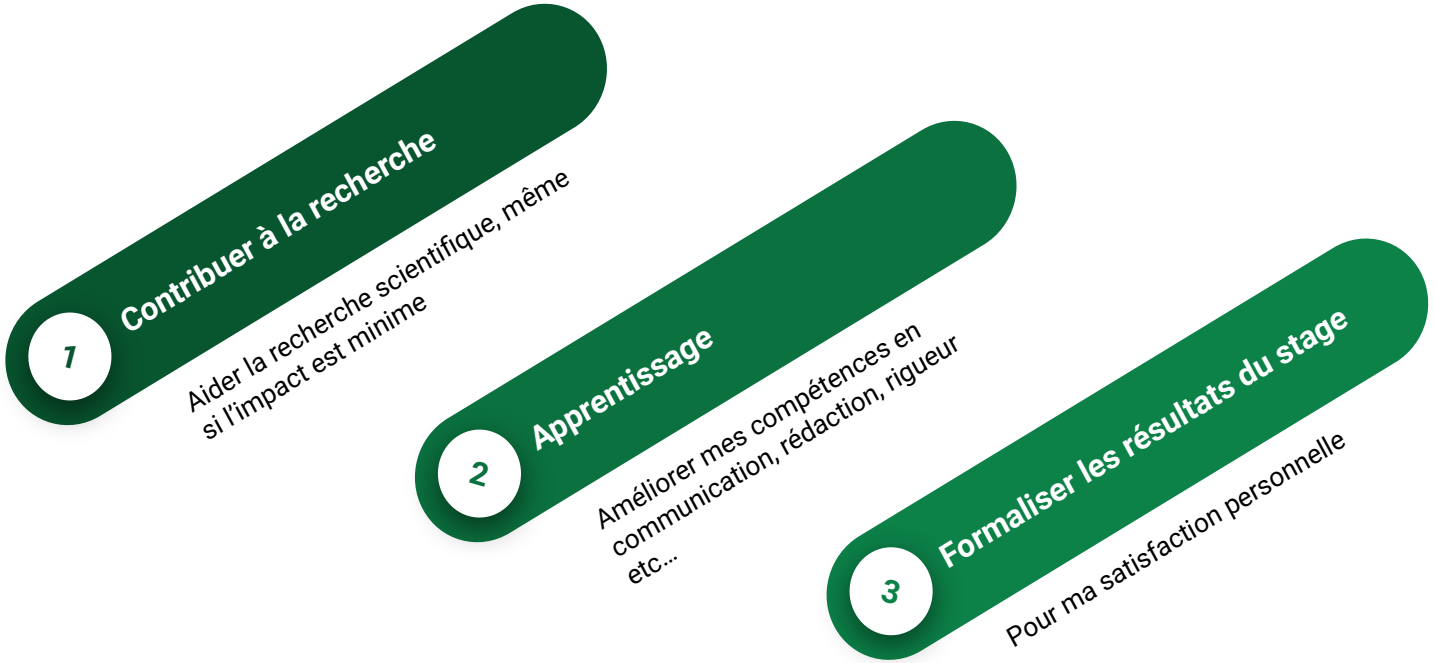
Les acquis

- Méthodologie de la recherche
- Rigueur dans la préparation des données
- Analyse des données



- Introduction au secteur de l’ énergie
- Introduction à l’utilisation des données des objets connectés (IoT & IoB)

Écriture d’un article



Remerciements

Enseignant référent :

M. Rémy CAZABET

Encadrants de stage :

Mme. Parisa GHODOUS

M. Mohamed-Essaid KHANOUCHE

Participants extérieurs :

M. Christian OBRECHT

M. Ahror BELAID

MERCI DE
VOTRE
ATTENTION