CONSOMMATION D'ELECTRICITE ANNUELLE PAR SECTEUR D'ACTIVITE AGREGEE ET PAR COMMUNE EN FRANCE
PROJET DATA MINING

Armand L'Huillier et Camil Zahi

1 Description des données

Nous avons 3 bases à notre disposition :

- base-pop-historiques-1876-2019.xlsx
- base-cc-tourisme-2022-geo2022-CSV/base-cc-tourisme-2022-geo2022.csv
- conso-elec-gaz-annuelle-par-secteur-dactivite-agregee-commune-france.csv

1.1 Séries historiques de population (1876 à 2019)

Il s'agit d'une base composée de **34 967 observations et de 37 variables** dont 4 variables géographiques qui serviront de clé de jointure et avec la population associée à plusieurs années comprises entre 1876 et 2019. Pour les colonnes du nombre d'habitants dans la population antérieure à 1962 ("PSDC62"), il y a quelques valeurs manquantes. Il n'y en a aucune pour les autres variables.

1.2 La capacité des communes en hébergement touristique

Il s'agit d'une base composée de **34 983 observations et de 52 variables** dont 1 variable géographique qui servira de clé de jointure ("CODGEO"). Les autres variables nous informent sur la capacité à recevoir des touristes pour la ville (nombre de chambres d'hôtel, de terrain de camping...) Il n'y a aucune valeur manquante.

1.3 Consommation annuelle d'électricité et de gaz par commune et par secteur d'activité

Il s'agit d'une base composée de **46 584 observations et de 36 variables** dont plusieurs variables géographiques qui serviront de clé de jointure. Les autres variables nous renseignent sur la consommation de gaz et/ou d'électricité, sur une année bien précisée, pour les communes. Il n'y a que quelques valeurs manquantes (sur la colonne "code_postal").

2 Analyse exploratoire des données

2.1 Construction de la base d'analyse

On remarque que les bases les plus importantes sont « consommation » et « population », car « consommation » sera la target et « population » sera la base la plus adéquate pour prédire la consommation d'électricité. Si nous faisons un merge left ou right pendant un merge sur ces deux bases, nous aurions plusieurs milliers (voire des dizaines de milliers de villes avec des valeurs manquantes d'un côté ou de l'autre (il y aurait donc soit des milliers de target manquantes, soit des milliers de variables majeures manquantes).

On préfère donc réaliser un **merge en 'inner'** pour garder les informations cruciales. Nous obtenons 'Base1'.

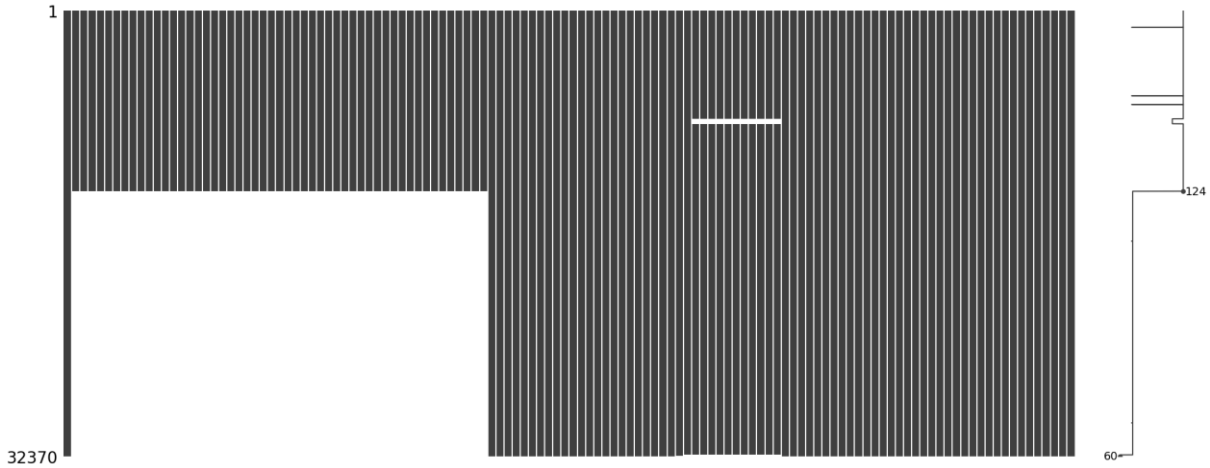
Ensuite, nous avons décidé de faire un **merge 'left'** entre le dataframe 'Base1' et la base capacité. Ceci nous permet de conserver toutes les villes avec les informations précieuses sur la consommation et les populations, tout en ajoutant de l'information sur la capacité de tourisme quand il y a les clés CODE-GEO qui matchent. Cette étape nous permet d'obtenir un dataframe 'Base' avec **32370 observations.**

Cependant, nous remarquons la présence de **doublons**. Des villes apparaissent plusieurs fois. Par exemple pour l'année 2021, une même ville peut avoir une observation qui renseigne ses caractéristiques en consommation d'électricité et une autre observation pour ses consommations en gaz.

En anticipant la partie machine learning, nous savons qu'on cherchera à prévoir la **consommation d'électricité des villes**. Or le gaz ne pourra pas nous servir : il n'est renseigné que pour la moitié des villes que nous avons dans le dataframe, donc si l'on voulait utiliser le gaz pour prédire l'électricité, nous devrions imputer pas moins de 50% de valeurs sur le gaz pour les villes. Nous considérons que cette information n'est pas suffisante, nous gardons alors seulement les observations qui communiquent une information sur la consommation d'électricité et non de gaz.

2.2 Imputation des valeurs manquantes et des outliers

Finalement, nous allons travailler sur une base **avec 22227 villes** ('base_elec'), c'est une base que nous allons chercher à nettoyer premièrement par une imputation des valeurs manquantes. Nous allons imputer le dataframe 'base_elec' en procédant par colonnes.



En noir : la présence des valeurs.

En blanc : l'absence des valeurs (valeurs manquantes)

Dans le coin inférieur gauche : La fusion a créé des valeurs manquantes puisque nous avons décidé de ne garder les informations des capacités de touristes des communes que lorsque nous avions des informations sur la consommation d'électricité et de gaz de ces communes. Par conséquent, la présence de ces valeurs manquantes est donc normale et nous allons les imputer par la régression.

Dans le centre droit vers le haut et tout en bas : le peu de valeurs manquantes est dû aux quelques codes postaux manquants et aux quelques données de population manquantes pour la période précédente à 1962. Nous les imputerons également.

D'abord nous regardons la présence de **valeurs manquantes des colonnes issues de la base conso**. Il n'y a qu'un **nombre infime de valeurs manquantes** sur les codes postaux. Nous décidons de **remplacer par le « code commune ».**

Ensuite, nous regardons les **valeurs manquantes des colonnes issues de la base pop**. Les colonnes concernées par des valeurs manquantes sont toutes numériques. Les valeurs manquantes sont peu nombreuses : 470 au maximum **(environ 1%**). Nous décidons de **remplacer par la médiane** car cette méthode ne va pas bouleverser le dataframe et les valeurs imputées seront réalistes.

En ce qui concerne les **colonnes issues de la base tourisme**, cela est dû au **merge choisi, nous nous retrouvons avec beaucoup de valeurs manquantes** (58% pour la colonne la plus atteinte par les valeurs manquantes). Donc nous ne voulons pas imputer avec une méthode simple (moyenne, médiane) car cela risquerait de changer drastiquement les colonnes issues de la base de tourisme. Donc nous optons pour une **imputation des variables issues de la base de tourisme par régression**. Nous utilisons toutes les colonnes numériques de 'Base_elec' pour prédire les valeurs manquantes des colonnes issues de la base de tourisme.

Concernant les valeurs extrêmes, Il est généralement recommandé de traiter les valeurs extrêmes lorsque nous allons modéliser une variable, comme la consommation d'électricité. En effet, les valeurs extrêmes peuvent avoir un impact significatif sur les résultats de votre modèle et peuvent distordre les résultats si elles ne sont pas prises en compte de manière adéquate.

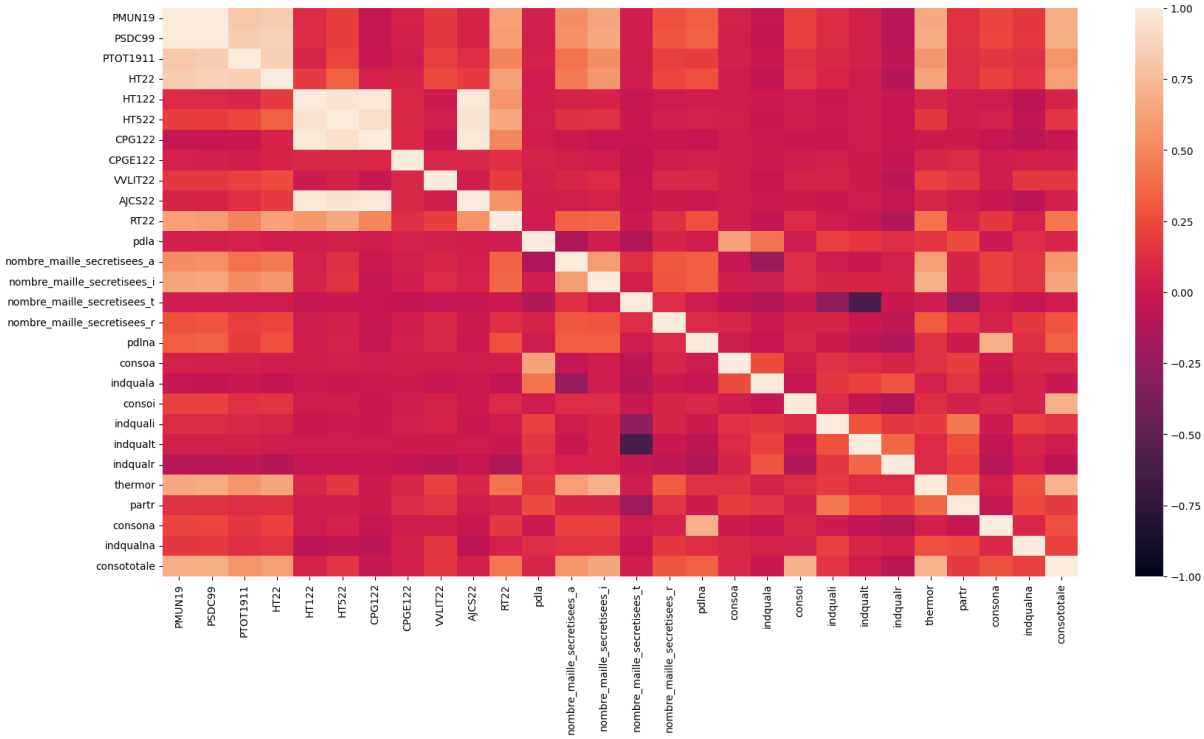
Il existe plusieurs façons de traiter les valeurs extrêmes. Il est important de noter que la façon de traiter les valeurs extrêmes dépendra de la situation spécifique et des objectifs de modélisation. Il est donc recommandé de bien comprendre les données et de faire des tests pour déterminer la meilleure approche à adopter dans votre cas particulier.

De cette manière, nous pensons qu'il serait acceptable de **ne rien faire pour les outliers dans ce cas**, puisque nous travaillons sur la **consommation d'électricité par commune**, or les **communes sont de tailles différentes et logiquement il y a beaucoup plus de petites villes que de grandes villes**. Ainsi traiter les valeurs extrêmes se révèlerait être contre-productif puisque cela constituerait une **perte d'information plus particulièrement pour les grandes villes qui sont moins nombreuses**. Ici, ils ne sont pas non plus le résultat d'une erreur de mesure donc nous jugeons préférable le fait de ne pas apporter de traitement spécifique sur les outliers.

2.3 Corrélation

Maintenant, pour les étapes des ACP et Clustering, nous avons besoin de regarder les corrélations entre les variables. Nous avons déjà remarqué que des variables sont très corrélées. Nous pouvons sans doute avoir des problèmes de combinaisons linéaires entre les variables : en sommant le nombre de chambres (1 à 5 étoiles) nous pouvons sans doute retrouver le total du nombre de chambres. L'étape de corrélation va aussi nous permettre de résoudre les **problèmes de multicolinéarité** dans cette étude.

On voit que **des variables sont très corrélées entre elles**. Par exemple, sur les variables numériques provenant du dataset de la population, toutes les variables sont corrélées à plus de 76%. Et leur corrélation peut atteindre 99%.



On a donc besoin de lancer des modèles ACP pour réduire la dimension et mieux comprendre notre dataset et la différenciation des villes.

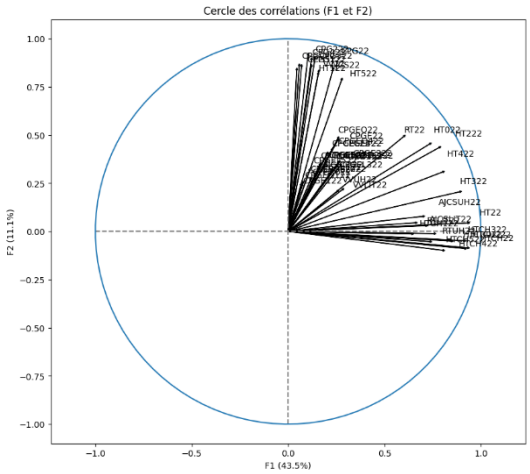
3 Analyse en Composantes Principales

L'analyse en Composantes Principales (ACP) est une technique de réduction de dimension qui permet de représenter graphiquement des données à plusieurs dimensions sous forme de projections sur un plan à deux dimensions. L'ACP consiste à trouver un nouvel ensemble de variables appelées "composantes principales", qui sont des combinaisons linéaires des variables d'origine, de sorte que la variance des données soit maximale sur chacune des composantes. Ces composantes sont ordonnées dans l'ordre croissant.

Dans l'ACP, nous utilisons toutes les variables, mais nous notons que **des variables sont extrêmement corrélées**, ce qui va se répercuter dans la visualisation de l'ACP avec les flèches : nous allons avoir des **flèches parfois très proches, qui se chevauchent**.

On obtient **deux premières composantes qui expliquent environ 55% de la variance expliquée**. Pour analyser les impacts des 54 variables, nous affichons le graphique des cercles des corrélations par groupes de variables :

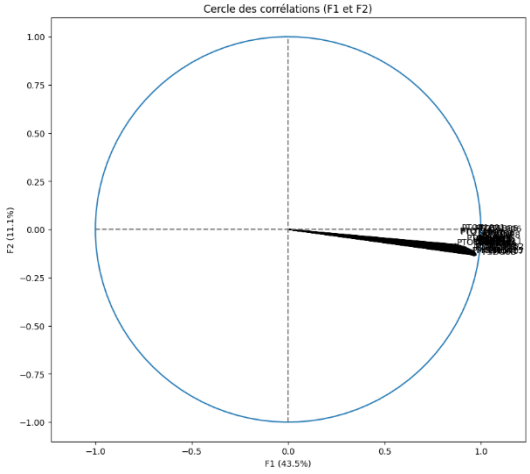
On trace un premier graphique avec toutes les variables numériques concernant le tourisme :



Sur la composante principale 1, nous remarquons que les variables déterminantes sont les chambres, les hôtels, les auberges de jeunesse. Or, les communes avec beaucoup d'hôtels et de chambres seront principalement des **grandes villes urbaines et touristiques** (Paris, Marseille...)

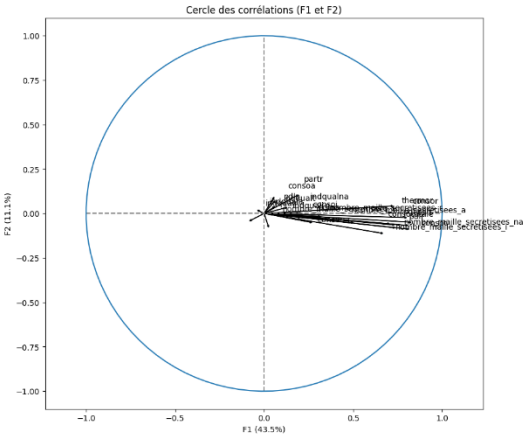
Sur la composante principale 2, nous remarquons que les variables déterminantes sont les variables liées au camping, elles sont un rôle positif par rapport à l'axe : donc les **communes rurales et balnéaires** qui sont touristiques ressortiront sur cet axe.

On trace un deuxième graphique avec la variable numérique concernant la population :

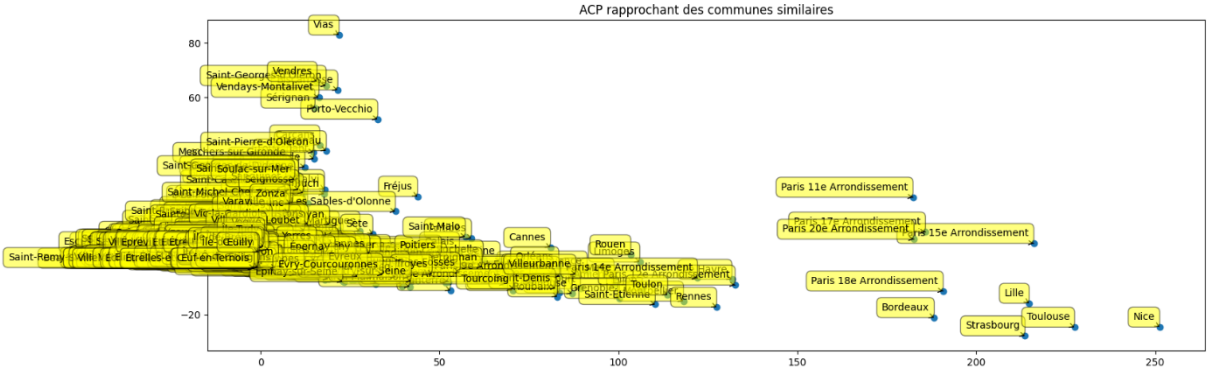


La population joue un rôle important, mais **toutes les variables de population sont finalement très corrélées entre elles** : les flèches se chevauchent toutes. En plus nous voyons que ces variables sont déterminantes pour construire la composante principale 1, mais pas beaucoup pour construire la composante principale 2.

Et le dernier graphique avec les variables de consommation :



La quantité d'énergie consommée par commune est déterminante pour expliquer la composante principale CP1, mais pas la CP2. **Plus la consommation est élevée, plus les communes se trouvent à droite dans le graphique.**



Interprétation du graphique :

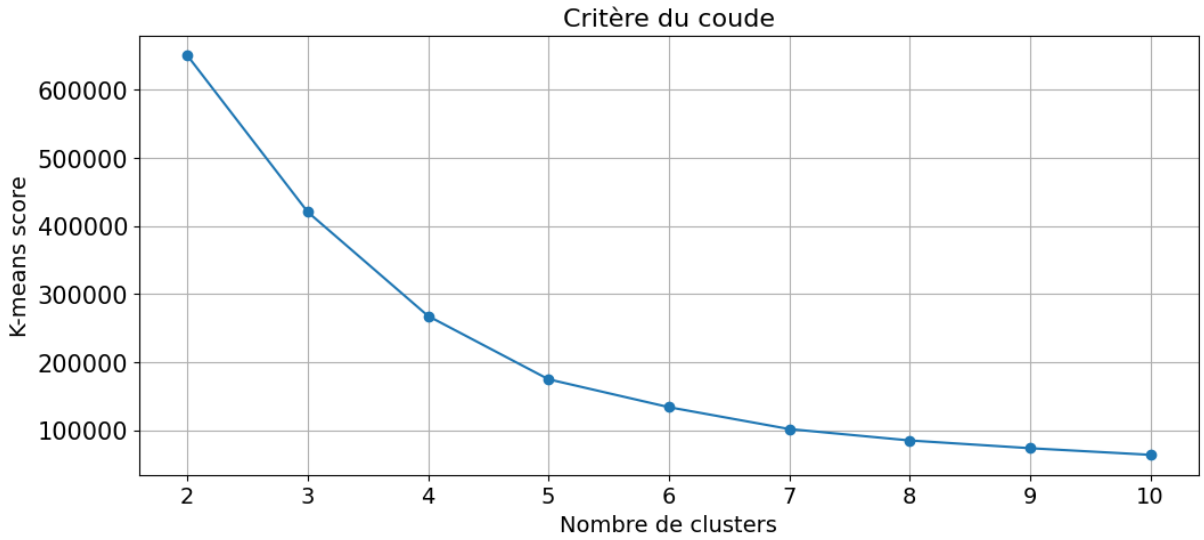
- **Les deux premiers axes représentent à peu près 55% de la variance expliquée.** Il y a deux groupes de variables (flèches vers le nord, flèches vers l'Est)
- La **composante principale 1 indique la taille de la ville** : population, nombre de lits dans les auberges de jeunesse, d'hôtel, consommation d'énergie.
    - **Les grandes villes se trouvent à l'Est** aux vues des variables qui influent positivement sur cet axe : les arrondissements parisiens, Nice, Lille...
- **Les flèches vers le Nord du graphique concernent des villages/petites villes dans les campagnes** : les variables qui influent positivement sur cet axe sont les variables sur les places dans les campings, les hébergements dans Villages vacances; et aussi la population, qui est négativement corrélée à l'axe 2.
    - **les petits villages touristiques sont dans le Nord du graphique** : Vias, St-Pierre d'Oléron...

PS : Nous avons une grande masse de communes qui sont proches de 0 dans le graphique. C'est parce qu'une **grande quantité de communes ont une population faible et qu'elles sont peu touristiques**, alors elles sont toutes au même endroit sur le graphique.

4 Clustering

L'algorithme KMeans est une méthode de clustering non-supervisée qui permet de regrouper un ensemble d'observations homogène "n" en "k" cluster de manière à minimiser la somme des distances intra-clusters.

Le critère du coude est une technique couramment utilisée pour choisir le nombre de clusters optimal à utiliser avec l'algorithme de clustering KMeans. Le principe de cette technique est de tracer un graphique de la distance inter-clusters moyenne en fonction du nombre de clusters, puis de choisir le nombre de clusters au niveau du "coude" du graphique, c'est-à-dire l'endroit où la distance inter-clusters moyenne commence à diminuer moins rapidement.

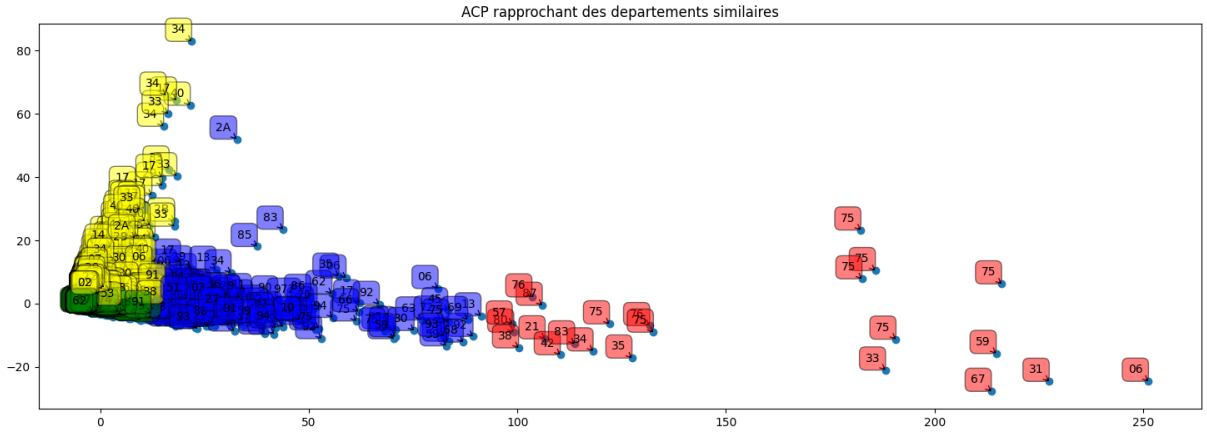


L'indice de silhouette est une mesure qui permet d'évaluer la qualité des clusters obtenus à l'aide d'un algorithme de clustering. Il se calcule en mesurant la distance moyenne entre les observations d'un même cluster et la distance moyenne entre les observations d'un cluster voisin. Plus le coefficient de silhouette est proche de 1, meilleurs sont les clusters obtenus.

Pour K = 2, le coefficient de silhouette est 0.957
Pour K = 3, le coefficient de silhouette est 0.748
Pour K = 4, le coefficient de silhouette est 0.777
Pour K = 5, le coefficient de silhouette est 0.790
Pour K = 6, le coefficient de silhouette est 0.797
Pour K = 7, le coefficient de silhouette est 0.811

Les coefficients de silhouette pour K = 4 et K = 5 semblent bons.

Ces deux critères nous permettent de choisir aisément un nombre de clusters égal à 4. En effet, pour 5 clusters, le coefficient de silhouette n'augmente que très peu mais l'interprétation sera difficilement compréhensible. Nous nous limitons donc à 4 clusters.



L'axe Ouest - Est s'explique par le niveau de la population de la commune : avec d'un côté les communes les plus peuplées (Paris, Nice...) et de l'autre les communes les moins peuplées.

L'axe Nord - Sud s'explique par le niveau de tourisme de la commune : avec d'un côté les communes les plus touristiques (Paris, Nice...) et de l'autre les communes les moins touristiques.
Une grande majorité des variables se trouve au milieu du graphique de l'ACP. Ce qui indique que l'ACP a du mal à vraiment discriminer : trop peu de variance expliquée. Mais nous pouvons quand même dire qu'il s'agit de communes qui attirent peu les touristes.

L'algorithme de clustering KMeans a regroupé 4 clusters :

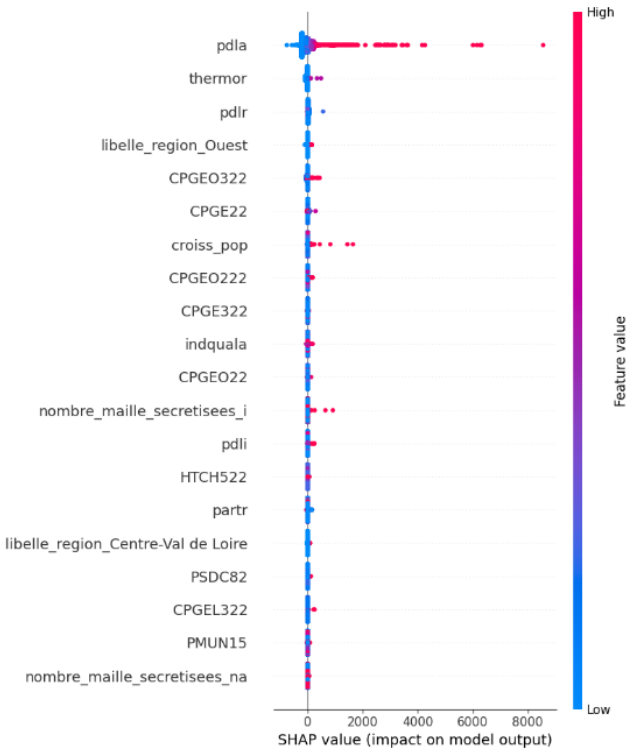
- Un cluster avec les **territoires où la population se révèle très forte et la capacité touristique également** (Paris-75, Strasbourg-67, Nice-06...)
- Un autre cluster regroupe des **villes moyennes**.
- Un cluster de **petites villes et villages avec une population faible mais rassemblant des touristes** (les villages balnéaires, les villages vacances...)
- Un cluster de **village peu peuplé et ne rassemblant pas ou très peu de touristes.**

5 Machine Learning

Pour **prédire la consommation électrique des communes dans l'agriculture, dans l'industrie et dans le résidentiel**, nous mettons en place un **Random Forest** que l'on a choisi car les RF donnent des bons résultats puisqu'ils prennent en compte les effets non linéaires et ne sont pas gênés par les corrélations entre les variables.

5.1 Consommation d'électricité dans l'agriculture des communes

R2 Value Train: 0.54 R2 Value Test: 0.43



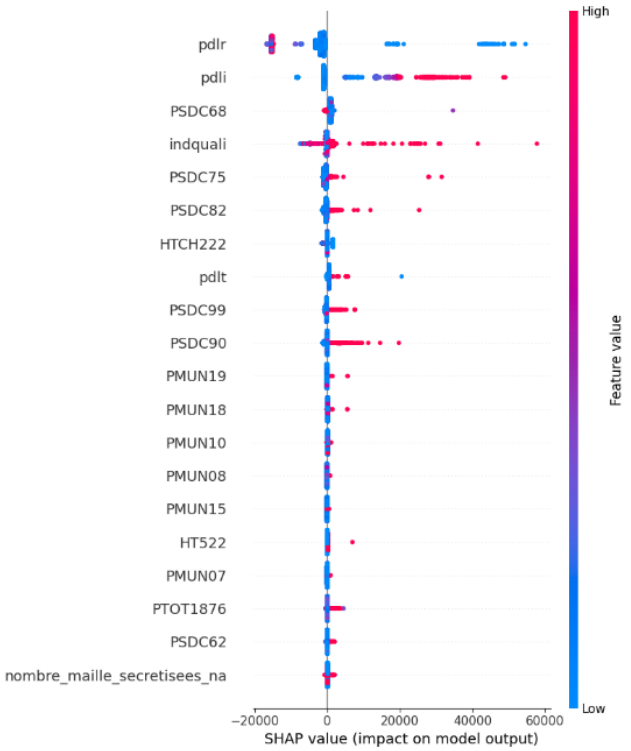
On garde un R2 faible sur le test, indiquant que notre modèle estime mal la consommation des villes en électricité pour l'agriculture.

Le graphique qui présente les valeurs de SHAP montre qu'une variable est essentielle pour prédire la consommation : c'est « plda ». Une variable est intéressante : la croissance de la population de la commune « croiss_pop ». Nous constatons que **plus la population de la commune a cru, moins elle consomme de l'électricité pour l'agriculture**. Elle traduit les changements économiques de la France au cours du 20eme siècle. L'économie se tourne vers l'industrie, puis vers les services, et au même moment la population augmente massivement. Dans les communes où la population a décru, il y a eu un exode rural, la commune est restée un village où les agriculteurs continuent de travailler. Voici l'explication de la variable « croiss_pop ».

Nous expliquons ce faible score dans l'agriculture dû à **une hétérogénéité du secteur de l'agriculture**. En effet, les grandes cultures ont besoin de pétrole principalement pour faire fonctionner leurs tracteurs et engin-automoteurs. Un agriculteur de grande culture aura moins besoin d'électricité que de pétrole tandis qu'un agriculteur d'élevage ne consommera peu de pétrole mais a besoin de chauffer ses bâtiments d'élevage et de locaux pour les animaux qui nécessitent de l'électricité. De plus, nous pouvons faire une différence en matière de taille d'exploitation. Une grande exploitation compte pour un seul « pdla » mais représente des quantités d'énergie différentes.

5.2 Consommation d'électricité dans l'industrie des communes

R2 Value Train: 0.72 R2 Value Test: 0.21

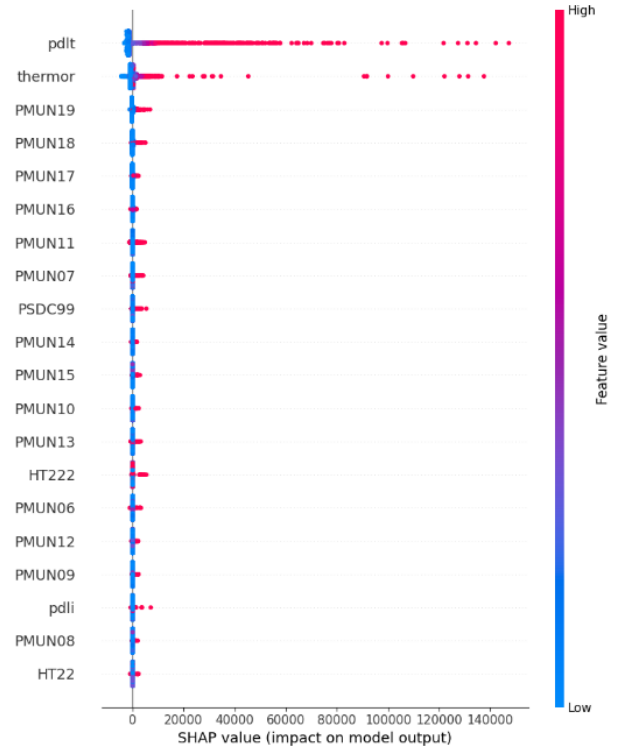


On voit que **ce modèle overfit énormément**, même avec un RF avec des arbres de profondeur égale à 3. Le modèle du RandomForest n'est pas adapté pour ce travail de prédiction. Mais l'analyse de SHAP indique que les variables les plus importantes sont « pdli », « pdlr », « indquali », « PSDC90 » ... : essentiellement des variables quantitatives. Ce qui veut dire qu'on pourrait modéliser un ElasticNet à la place du RF, ce qui permettrait de réduire drastiquement l'overfitting s'il est bien paramétré. Pour autant, la précision de la prédiction et du R2 sur le dataset Test n'augmenteront pas automatiquement.

5.3 Consommation d'électricité dans le résidentiel des communes

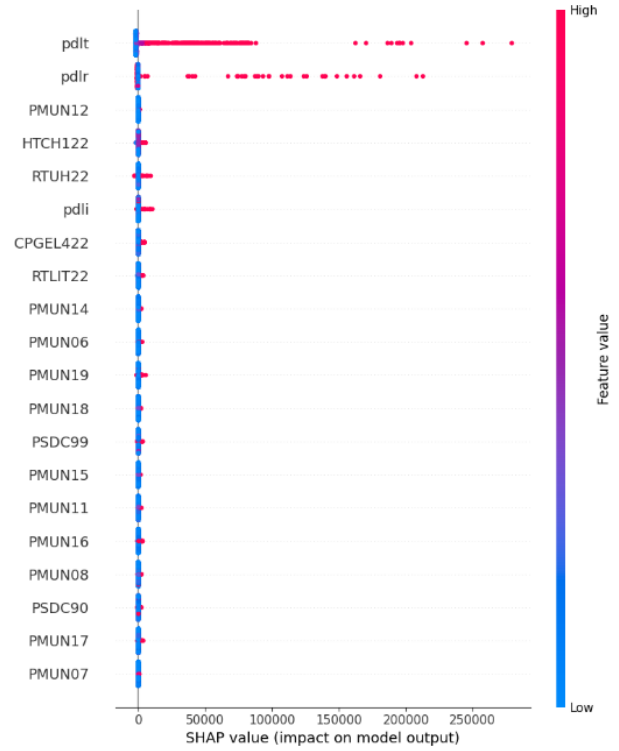
R2 Value Train: 0.97 R2 Value Test: 0.93

Ici, nous remarquons que grâce à deux variables, le RF est très bon pour prédire la consommation du secteur résidentiel.
Ici « pdlr » est très bon pour prédire la consommation, c'est parce que les ménages sont assez homogènes (plus que les agriculteurs en particulier). Et donc nous pouvons dire que **plus il y a de points de branchements à l'électricité, plus la consommation est élevée, dans le résidentiel**, ce phénomène est linéaire ce qui doit expliquer le très bon R2 sur le dataset Test.



5.4 Consommation d'électricité dans le secteur tertiaire des communes

R2 Value Train: 0.92 R2 Value Test: 0.87



Par curiosité, nous avons voulu lancer un modèle pour prédire la consommation dans le secteur tertiaire, le secteur des services. Et nous voyons que le modèle est performant.
Globalement, nous voyons voit que **quand la commune est très peuplée, alors elle doit avoir une économie très tournée vers les services**, et donc nous arrivons à bien prédire la consommation pour ce secteur.