Rapport MLOps
(Machine Learning Operations)

–

Étude de la consommation quotidienne d'électricité et de gaz

Héloïse Cammerman
Capucine Pollet
Thibaut Maton
Lamine Top

Professeur : M. Vasseur

Table des matières

I. Introduction

Ce projet vise à prédire la consommation brute quotidienne totale, qui combine à la fois la consommation d'électricité et celle de gaz. En nous appuyant sur les variables déjà disponibles et en enrichissant notre jeu de données avec des informations complémentaires, nous cherchons à estimer cette consommation de manière précise et robuste.

Notre objectif principal est d'améliorer progressivement les performances des modèles de prédiction. Pour cela, nous utiliserons des métriques adaptées à ce problème de régression afin d'évaluer rigoureusement chaque approche.

Notre méthodologie repose d'abord sur une préparation minutieuse des données : ajout de nouvelles variables, gestion des valeurs manquantes, analyse descriptive des distributions pour mieux comprendre les données et détecter d'éventuelles anomalies, ainsi qu'une étude des relations entre les variables explicatives et la cible. Ensuite, nous procéderons à l'entraînement et à l'évaluation de plusieurs modèles, en nous basant sur une spécification initiale commune. Enfin, nous sélectionnerons et justifierons le modèle retenu en fonction des résultats obtenus grâce aux métriques de qualité choisies.

II. Analyse des données

1) Description du fichier

Notre analyse repose sur le fichier « consommation-quotidienne-brute.csv », qui recense les données de consommation quotidienne d'électricité et de deux types de gaz (NaTran et Teréga) en France. Ce jeu de données, disponible sur data.gouv.fr, couvre une période allant du 31 décembre 2011 au 31 octobre 2025, avec pour chaque enregistrement la date et l'heure associées à la consommation.

À l'état brut, cette base comprend **12 colonnes** pour un total de **242 544 lignes**. Parmi ces colonnes, on retrouve 12 variables distinctes :

Variable	Type	Description
Date - Heure	Date Ex : 2025-10-31T22:00:00+00:00 → le 31/10/2025 à 22h	La date et l'heure correspondantes à la consommation
Date	Date Ex : 31/10/2025	La date correspondante à la consommation
Heure	Date Ex : 22:00	L'heure correspondante à la consommation
Consommation brute gaz (MW PCS 0°C) - NaTran	Variable quantitative	Consommation brute du gaz NaTran
Statut - NaTran	Variable qualitative	Statut du gaz. Soit « Définitif », soit « Meilleur statut »
Consommation brute gaz (MW PCS 0°C) - Teréga	Variable quantitative	Consommation brute du gaz Teréga
Statut - Teréga	Variable qualitative	Statut du gaz. Soit « Définitif », soit « Définitive »
Consommation brute gaz totale (MW PCS 0°C)	Variable quantitative	Somme des variables « Consommation brute gaz (MW PCS 0°C) - NaTran » et « Consommation brute gaz (MW PCS 0°C) - Teréga »
Consommation brute électricité (MW) - RTE	Variable quantitative	Consommation d'éléctricité
Statut - RTE	Variable qualitative	Statut du gaz. Soit « Définitif », soit « Consolidé »
Consommation brute totale (MW)	Variable quantitative	Somme des variables « Consommation brute gaz (MW PCS 0°C) - NaTran », « Consommation brute gaz (MW PCS 0°C) - Teréga et « Consommation brute électricité (MW) - RTE »
flag_ignore	Variable qualitative	Pour savoir si il faut ignorer la ligne correspondante. Avec uniquement la valeur « non »

2) Ajout de nouvelles variables

Pour améliorer la prédiction de la consommation, nous avons enrichi notre jeu de données en ajoutant des **variables calendaires** et des **variables météorologiques**.

Les **variables calendaires** ont été créées à partir des dates déjà présentes dans le fichier initial. Nous en avons extrait le jour, le mois, l'année, le jour de la semaine (en lettre et en numéro), ainsi qu'une indication précisant si le jour était un week-end ou non. Nous avons également intégré une variable indiquant si la date correspondait à une période de **vacances scolaires**, en tenant compte des trois zones de vacances en France.

Au total, ces ajouts ont permis d'incorporer **9 nouvelles variables** liées aux dates.

Variable	Type	Description
jour	Date. Ex : 31	Numéro du jour dans le mois
mois	Date. Ex : 10 → octobre	Numéro du mois dans l'année
annee	Date. Ex : 2025	L'année
jour_semaine	Date. Ex : lundi	Jour de la semaine en lettres
jour_semaine_num	Date. Ex : 4 → vendredi	Jour de la semaine en chiffre (avec lundi = 0)
week_end	Booléen	Si oui ou non la date correspond à un jour de week-end
vacances_zone_A	Booléen	Si oui ou non la date correspond à une période de vacances pour la zone A
vacances_zone_B	Booléen	Si oui ou non la date correspond à une période de vacances pour la zone B
vacances_zone_C	Booléen	Si oui ou non la date correspond à une période de vacances pour la zone C
Jour_ferie	Booléen	Si oui ou non la date correspond à un jour férié

Pour enrichir davantage notre jeu de données et améliorer la performance de nos modèles, nous avons également intégré des **données météorologiques**. Ces informations ont été récupérées via l'API Open-Meteo.

Les données météorologiques ajoutées se composent de **7 variables supplémentaires** :

Variable	Type	Description
timestamp	Date Ex : 2025-10-31 22:00:00 → le 31/10/2025 à 22h	Date de la météo que nous allons fusionner avec la date de notre jeu de données
temp	Variable quantitative	Température
temp_feels	Variable quantitative	Température ressentie
humidity	Variable quantitative	Taux d'humidité
wind_speed	Variable quantitative	Vitesse du vent
cloud_cover	Variable quantitative	Couverture nuageuse. Taux de remplissage du ciel par des nuages
pressure	Variable quantitative	Pression atmosphérique

3) Nettoyage des données

Après l'ajout des variables calendaires et météorologiques, notre jeu de données comptait initialement 27 variables. Plusieurs ajustements ont été nécessaires pour optimiser sa qualité et sa pertinence :

Nous avons supprimé la variable « timestamp », utilisée uniquement pour fusionner les données avec celles de la météo, car elle était redondante avec « Date - Heure ». Cette dernière, déjà décomposée en deux variables distinctes, n'apportait plus d'information utile et a donc également été retirée.

Les variables « Statut - Teréga » et « flag_ignore » ont été supprimées, car elles ne présentaient qu'une seule modalité et n'apportaient aucune information exploitable.

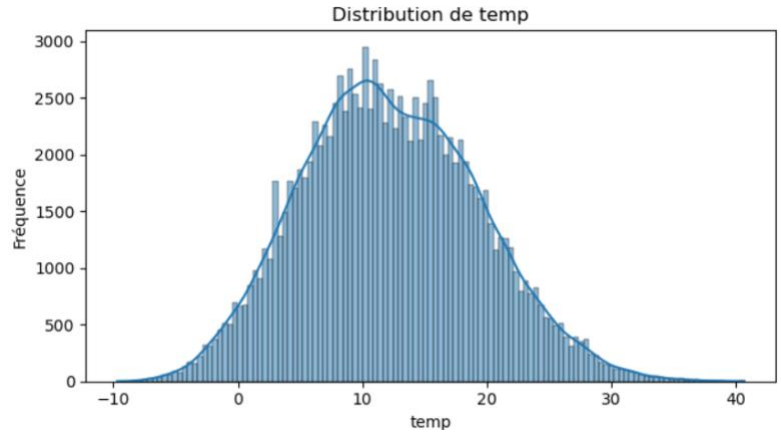
Par ailleurs, les variables « Consommation brute gaz totale (MW PCS 0°C) », « Consommation brute électricité (MW) – RTE », « Consommation brute gaz (MW PCS 0°C) – NaTran » et « Consommation brute gaz (MW PCS 0°C) – Teréga » n'ont pas été retenues pour l'entraînement des modèles. En effet, leur somme correspond exactement à la variable cible que nous cherchons à prédire. Notre objectif est de prédire la consommation totale sans connaître à l'avance les consommations individuelles en gaz (NaTran et Teréga) et en électricité.

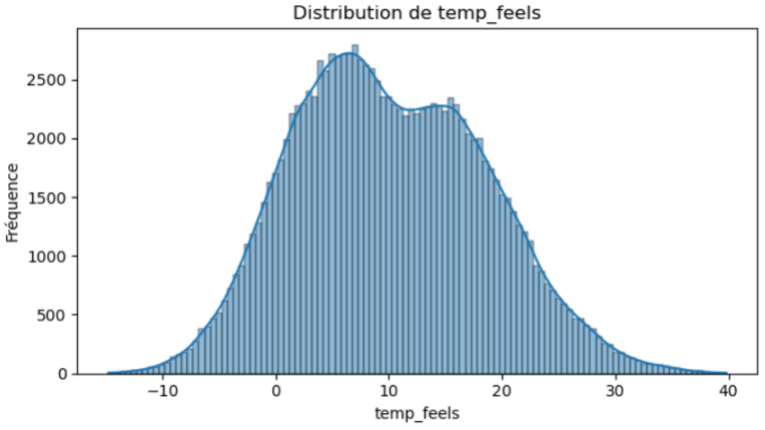
Nous avons également traité les valeurs nulles : la moitié des données contenaient des valeurs nulles, car les enregistrements étaient initialement disponibles toutes les demi-heures. Nous avons choisi de ne conserver que les données horaires, ce qui a réduit considérablement le nombre de valeurs nulles. Les 100 valeurs nulles restantes ont été supprimées. Enfin, après la fusion avec les données météorologiques, une ligne correspondant au 31/12/2011 présentait des valeurs nulles pour les variables météo et a donc été retirée.

À l'issue de ces traitements, notre jeu de données final comprend **25 colonnes** pour **121 171 lignes**.

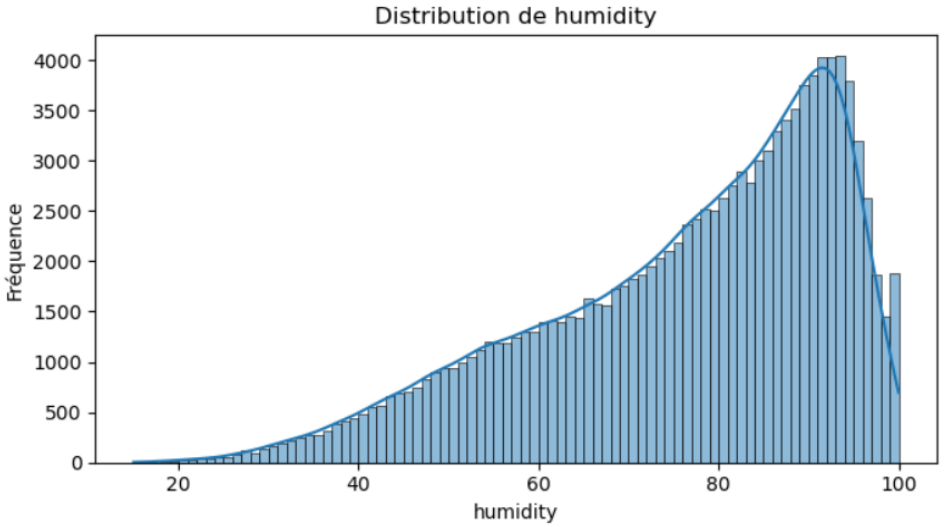
4) Distribution des données

Avant de procéder à l'entraînement des modèles, nous avons étudié la distribution de nos variables quantitatives afin de mieux les comprendre et d'identifier d'éventuelles particularités.

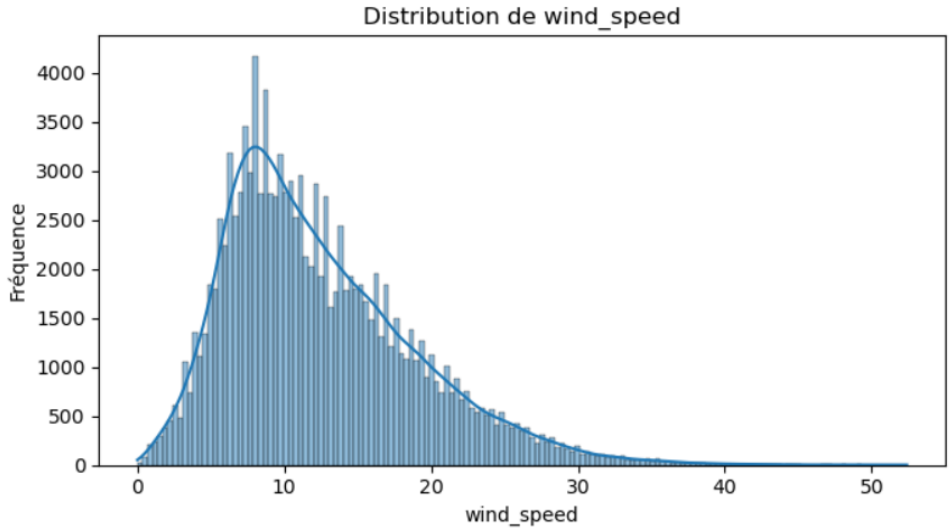




La distribution de la variable temp (tout comme celle de la variable temp_feels) est unimodale et globalement en cloche, ce qui suggère une forme proche d'une distribution normale. Le pic se situe autour de 10–12, indiquant la valeur la plus fréquente. On observe une légère asymétrie à droite (queue plus étendue vers les fortes valeurs), avec quelques valeurs élevées allant jusqu'à ~40, ce qui peut correspondre à des épisodes extrêmes. Les valeurs négatives existent mais restent rares. Dans l'ensemble, la variable est bien concentrée autour de sa moyenne, avec une dispersion modérée et peu d'anomalies marquées.

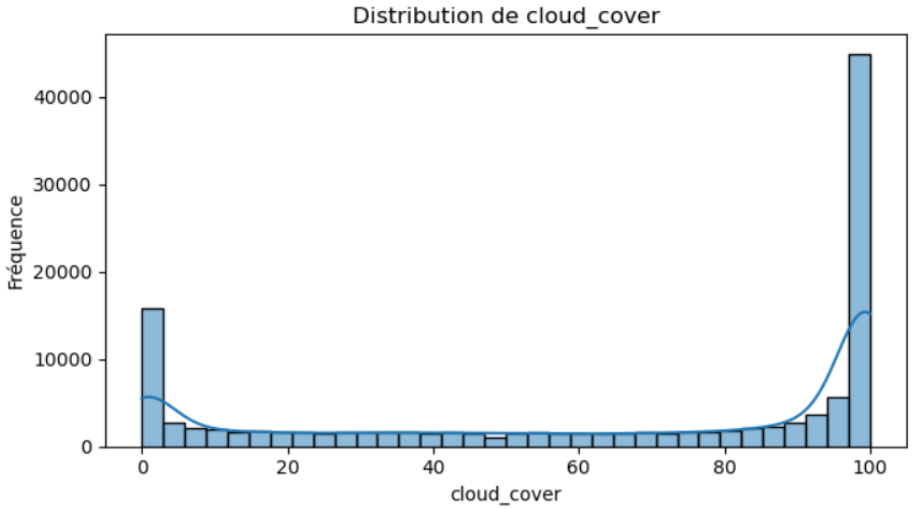


La distribution de humidity, quand à elle, est fortement asymétrique à gauche (queue étendue vers les faibles valeurs), avec une concentration marquée des observations entre 70 % et 95 %. Le pic se situe autour de 85–90 %, indiquant que des niveaux d'humidité élevés sont très fréquents. Les faibles humidités (< 40 %) sont rares, ce qui suggère un contexte climatique globalement humide. La borne supérieure proche de 100 % crée un effet de plafond, pouvant limiter la variabilité et influencer certains modèles statistiques.

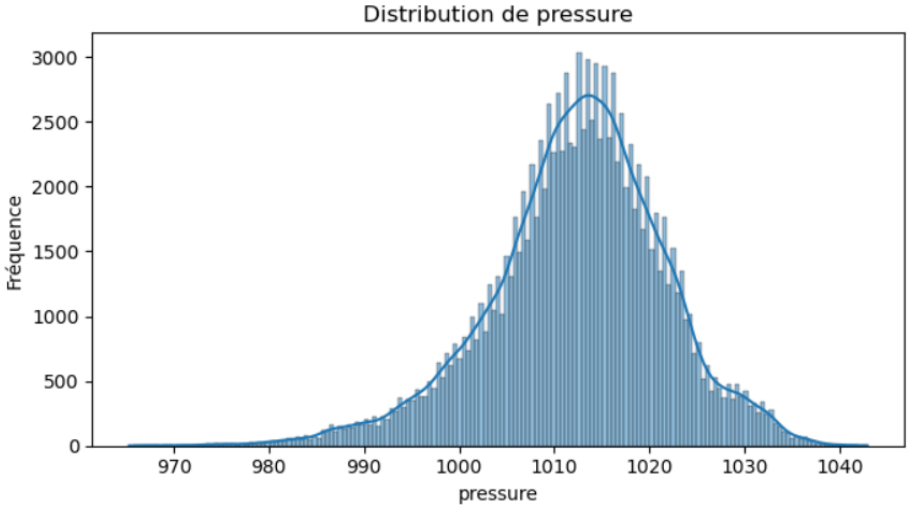


La distribution de la variable wind_speed, elle, est oposée à celle de la variable humidity car elle est fortement asymétrique à droite, avec une concentration des valeurs faibles à modérées (environ 5 à 15) et un pic autour de 8–10. Les vitesses élevées sont rares mais présentes, formant une longue queue jusqu'à des valeurs extrêmes (> 40).

Cette forme est typique des variables de vent et indique une variabilité élevée, avec des épisodes ponctuels de vents forts susceptibles d'influencer fortement les modèles sensibles aux outliers.

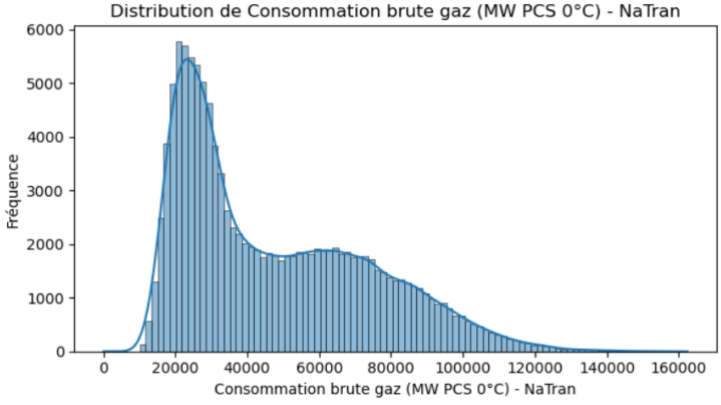


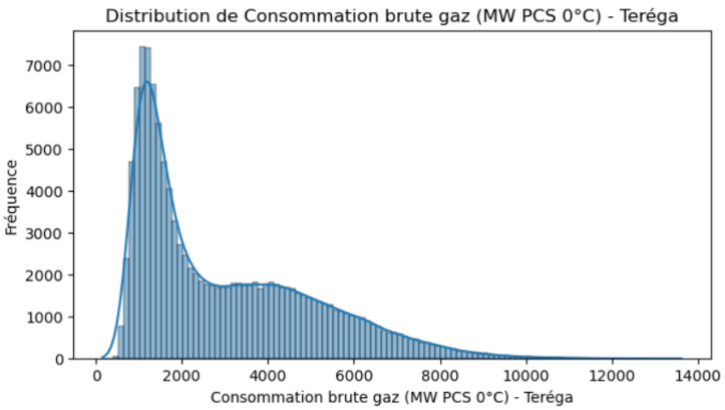
La distribution de cloud_cover elle est assez asymétrique avec une grande partie de ces valeurs en 0 et 100. Ce qui signifie un ciel souvent soit couvert de nuage soit entièrement dégagé.



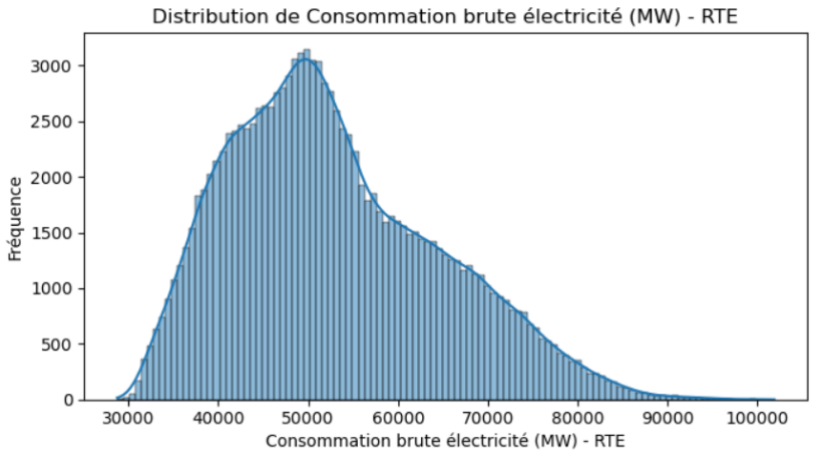
La distribution de la variable pressure est unimodale et globalement en cloche, ce qui suggère une forme proche d'une distribution normale. Le pic se situe autour de 1010–1020, indiquant la valeur la plus fréquente. On observe une légère asymétrie à gauche (queue plus étendue vers les fortes valeurs), avec quelques valeurs plus basses allant jusqu'à 965, ce qui peut correspondre à des épisodes extrêmes.

Dans l'ensemble, la variable est bien concentrée autour de sa moyenne, avec une dispersion modérée et peu d'anomalies marquées.

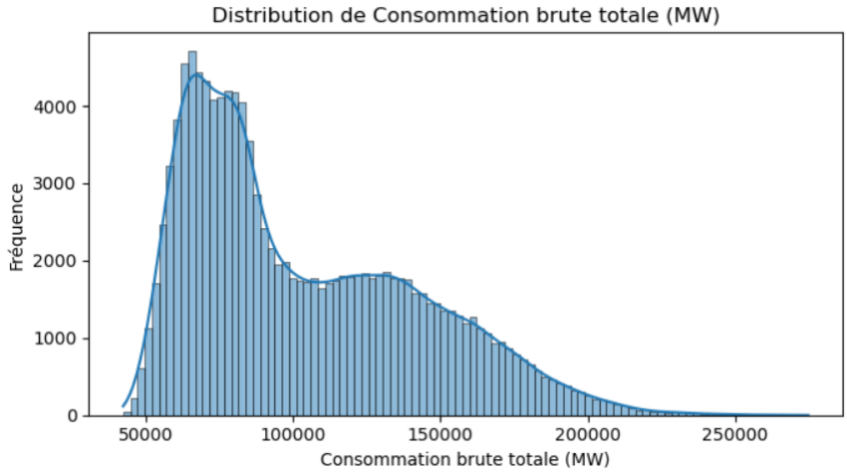




Cette distribution montre une consommation brute de gaz très concentrée entre 10 000 et 40 000 MW pour le NaTran et 1 000 et 2 000 pour Teréga, avec un pic autour de 20 000 MW pour NaTran et autour de 1 500 pour Teréga. La longue traîne vers la droite révèle des pics de consommation élevés, probablement liés à des périodes de froid intense. Cela indique une forte variabilité et la présence de valeurs extrêmes à prendre en compte dans les modèles.



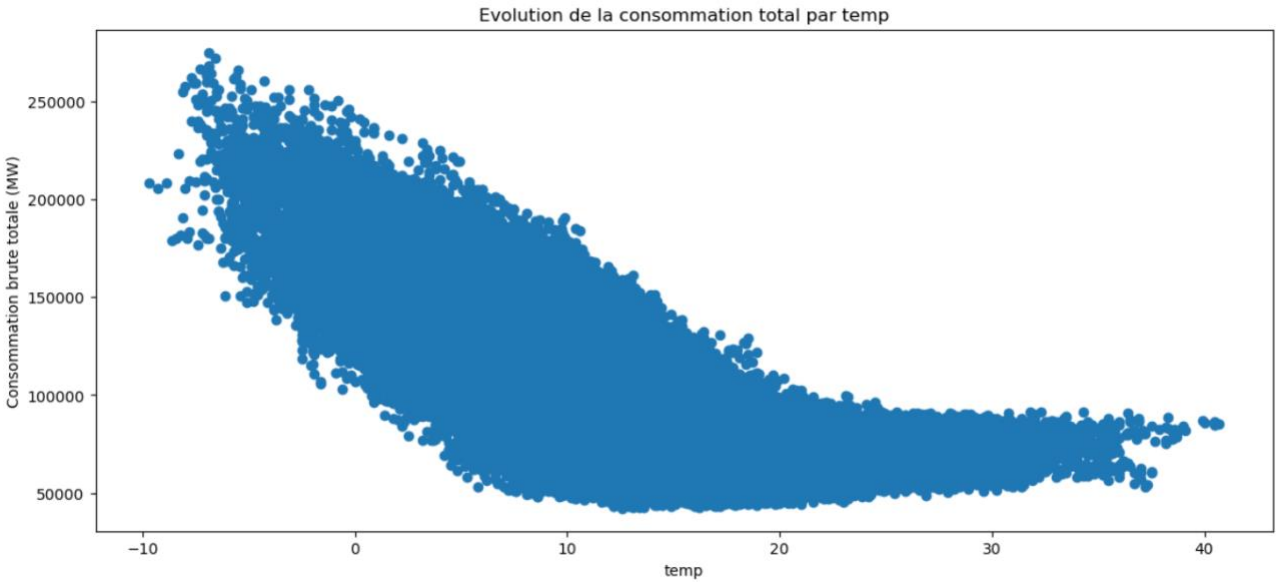
La distribution de la variable électricité est globalement en cloche, ce qui suggère une forme proche d'une distribution normale. Le pic se situe autour de 50 000, indiquant la valeur la plus fréquente. On observe une asymétrie à droite, avec quelques valeurs élevées allant jusqu'à 100 000.



La consommation brute total est notre variable cible. Cette distribution montre une consommation total assez concentrée entre 50 000 et 100 000 MW. La présence d'une longue queue à droite met en évidence des niveaux de consommation exceptionnellement élevés, vraisemblablement associés à des épisodes de froid marqué. Cette asymétrie traduit une variabilité importante ainsi que l'existence de valeurs extrêmes qu'il est nécessaire de considérer lors de la modélisation.

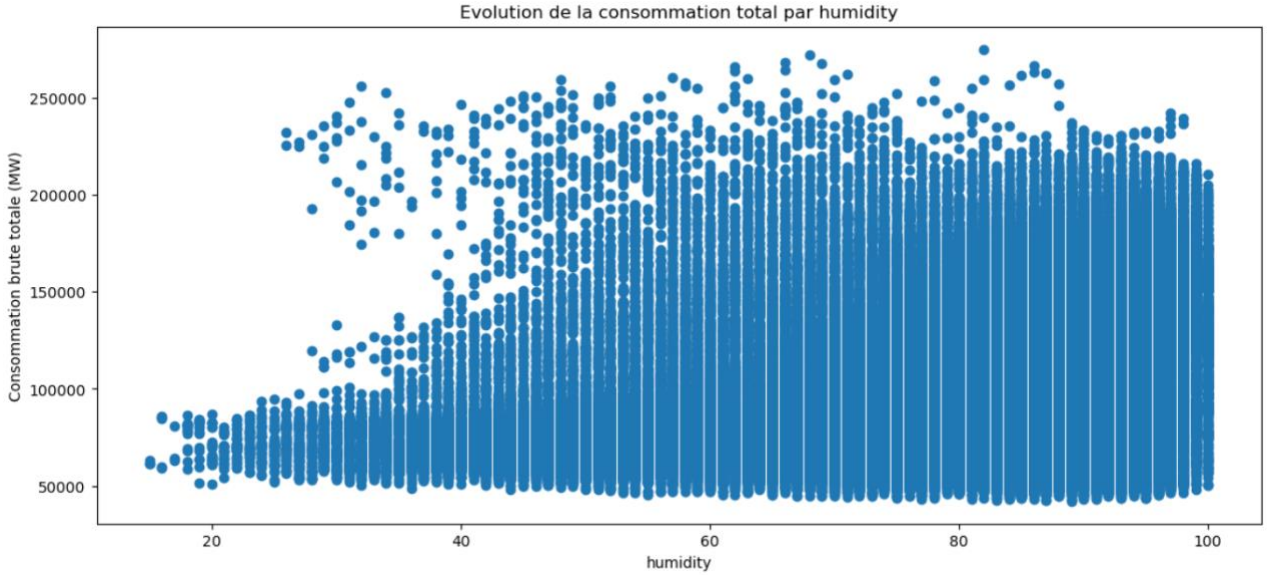
5) Relations avec la variable cible

Nous avons examiné l'évolution de la consommation brute totale en fonction des autres variables, afin d'évaluer et de mieux comprendre leur impact. Parmi celles-ci, nous présentons ici quatre variables dont l'influence nous a semblé la plus marquée. Les analyses complémentaires concernant les autres variables sont disponibles en annexe.

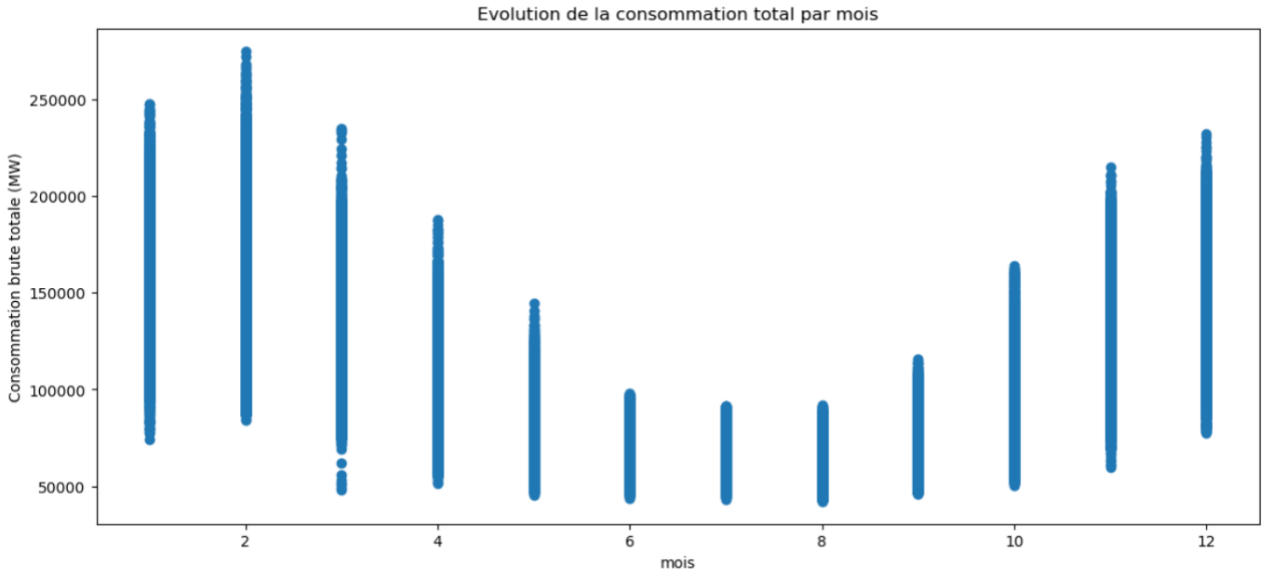


Ce premier graphique révèle une corrélation nette entre la consommation brute totale et la température extérieure. On observe en effet que les consommations augmentent significativement lorsque les températures descendent en dessous de 0°C, atteignant des niveaux environ deux fois plus élevés que ceux enregistrés lors de périodes plus douces.

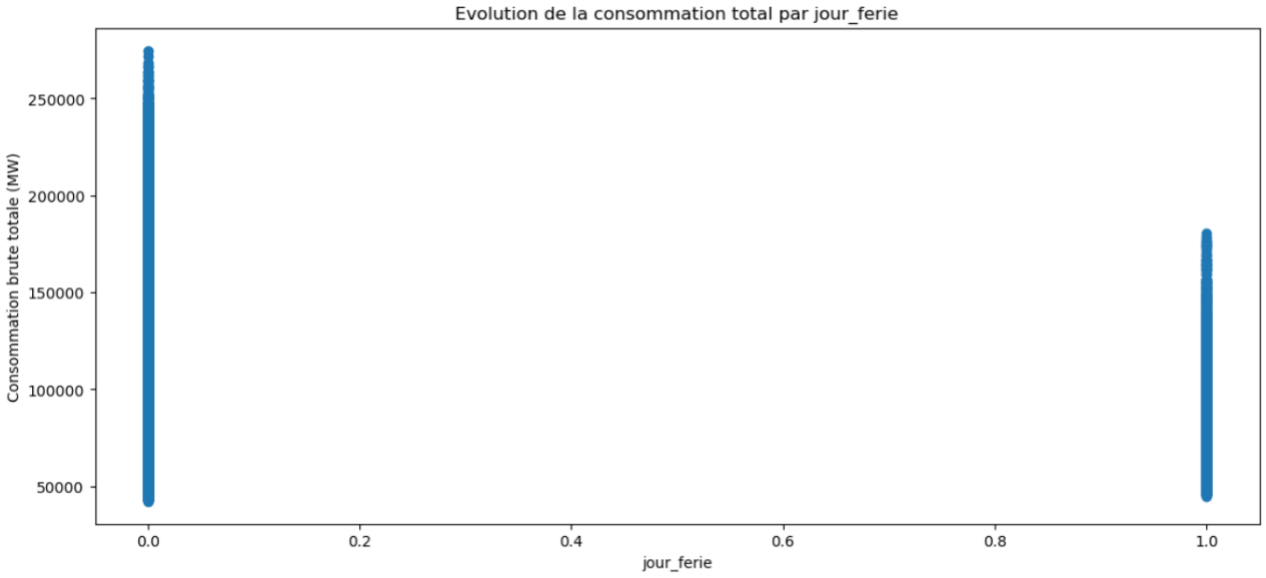
Ce deuxième graphique met en évidence une relation entre le taux d'humidité et la consommation



brute totale. On constate que les consommations tendent à augmenter lorsque le niveau d'humidité est plus élevé.



Ce graphique, qui représente la consommation par mois, montre clairement que les consommations de gaz et d'électricité sont significativement plus élevées pendant les saisons froides. Cette hausse peut s'expliquer par l'utilisation accrue du chauffage par les ménages français durant ces périodes.



Ce dernier graphique met en lumière l'influence des jours fériés sur la consommation. On observe que la consommation est plus élevée les jours non fériés (valeur = 0) que les jours fériés (valeur = 1).

L'ensemble de ces données, qu'elles soient calendaires ou météorologiques, s'avérera précieuse pour affiner nos modèles et prédire avec précision la consommation de gaz et d'électricité.

III. Modèles

1) Modélisation et métriques

Les données ont été séparées en un jeu d'entraînement (80 %) et un jeu de test (20 %). Ce découpage permet d'entraîner les modèles sur un volume de données suffisant tout en conservant un jeu de test indépendant pour l'évaluation des performances. La même séparation est utilisée pour l'ensemble des modèles afin de garantir la comparabilité des résultats.

Les performances des modèles sont évaluées à l'aide des métriques suivantes :

- MAE : Mean Absolute Error (erreur absolue moyenne). Cela permet de mesurer la moyenne des erreurs en valeur absolue entre les prédictions et les vraies valeurs.
- RMSE : Root Mean Squared Error (racine de l'erreur quadratique moyenne). Cela nous permet de mesurer la racine carrée de la moyenne des erreurs au carré.
- R^2 : coefficient de détermination. Cela nous permet de mesurer la proportion de la variance expliquée par le modèle.

Ces métriques sont analysées conjointement afin d'obtenir une évaluation complète et équilibrée des performances des modèles.

2) Modèles

1. Modèle baseline naïve (t-1)

Le premier modèle testé est un modèle baseline naïf, conçu comme une référence simple pour évaluer les performances des autres approches. Son principe est élémentaire : la valeur future est égale à la dernière valeur observée (principe de persistance). Si un modèle plus sophistiqué obtient des résultats inférieurs à cette baseline, il sera considéré comme non performant.

2. Régression linéaires

Premier modèle explicatif de notre étude, la régression linéaire suppose une relation linéaire entre les variables explicatives et la cible, avec des effets additifs et constants pour chaque variable. Ses principaux atouts résident dans sa simplicité, sa rapidité d'exécution et son interprétabilité, ce qui en fait un outil idéal pour une première analyse.

3. Random forest (foret aléatoire)

Ce modèle non linéaire repose sur un ensemble d'arbres de décision, lui permettant de capturer des interactions complexes et des relations non linéaires entre les variables. Robuste face au bruit et aux valeurs aberrantes, il offre une grande flexibilité tout en conservant une bonne capacité de généralisation.

4. Gradient Boosting

Le *Gradient Boosting* est une méthode itérative qui apprend progressivement à corriger ses erreurs, en capturant des structures fines dans les données. Cela lui permet souvent d'atteindre des performances supérieures aux autres modèles. Cependant, cette puissance s'accompagne de certains inconvénients : une sensibilité accrue à l'overfitting, une exécution plus lente, et une configuration plus délicate. Son utilisation doit donc être maîtrisée pour éviter les écueils.

3) Résultats

Les résultats que nous obtenons suite aux entraînements des modèles sont visibles dans le tableau ci-dessous.

Model	MAE	RMSE	R2
Baseline (t-1)	3217.87	4550.13	0.986
Gradient Boosting	10418.13	13903.77	0.8936
Random Forest	10611.51	14347.71	0.8867
Linear Regression	19988.74	24047.92	0.6818

Les résultats révèlent que le modèle baseline naïf (t-1) surpasse largement tous les autres modèles, avec les meilleures valeurs de MAE et RMSE, ainsi qu'un R^2 proche de 0,99. Cette performance exceptionnelle s'explique par une forte autocorrélation de la série temporelle : la valeur à l'instant t est en effet très proche de celle observée à t-1, un phénomène typique des séries temporelles.

Les modèles plus sophistiqués, comme le Random Forest et le Gradient Boosting, n'apportent aucun gain significatif. Cela suggère qu'ils ne parviennent pas à exploiter efficacement la dépendance temporelle dominante, ou que les variables explicatives ajoutées contiennent peu d'informations complémentaires utiles.

Quant à la régression linéaire, elle se révèle la moins performante, ce qui indique que les relations entre les variables ne sont ni simples ni suffisamment informatives pour expliquer la consommation.

En résumé, la dynamique temporelle semble expliquer l'essentiel du phénomène étudié, rendant le modèle *baseline* naïf particulièrement difficile à surpasser.

4) Choix du modèle

Le choix du modèle dépend avant tout de la disponibilité des données de la veille. Si la consommation réelle de la veille est connue, le modèle baseline naïf (t-1) s'impose comme la solution la plus fiable, tirant parti de la forte autocorrélation de la série temporelle pour offrir une prédiction précise.

En revanche, lorsque la consommation de la veille n'est pas disponible (par exemple pour des prédictions à plus long terme ou des scénarios prospectifs) le modèle Gradient Boosting devient le plus pertinent. Grâce à sa capacité à exploiter les variables météorologiques et calendaires, il permet d'estimer la consommation même en l'absence de données temporelles immédiates.

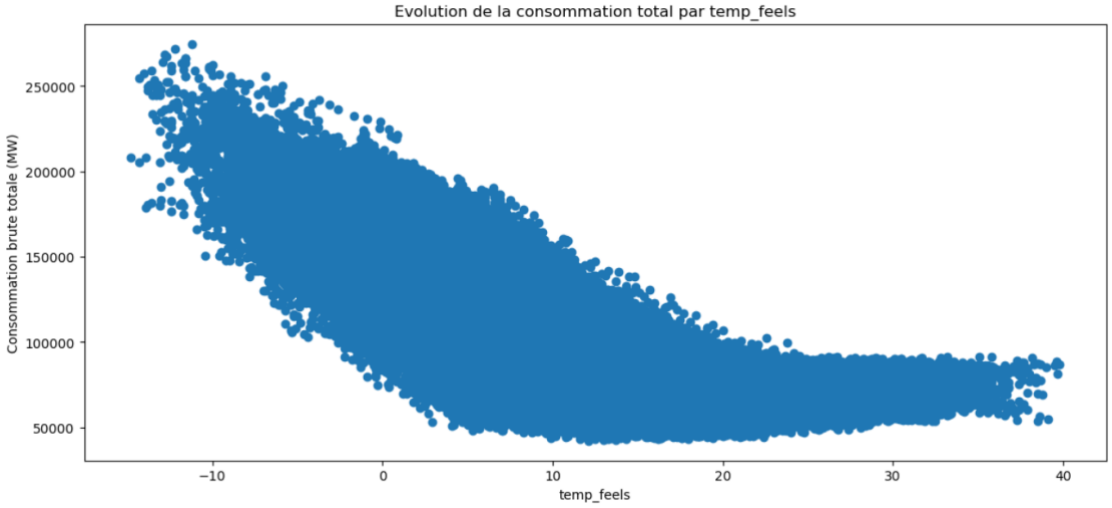
IV. Conclusion

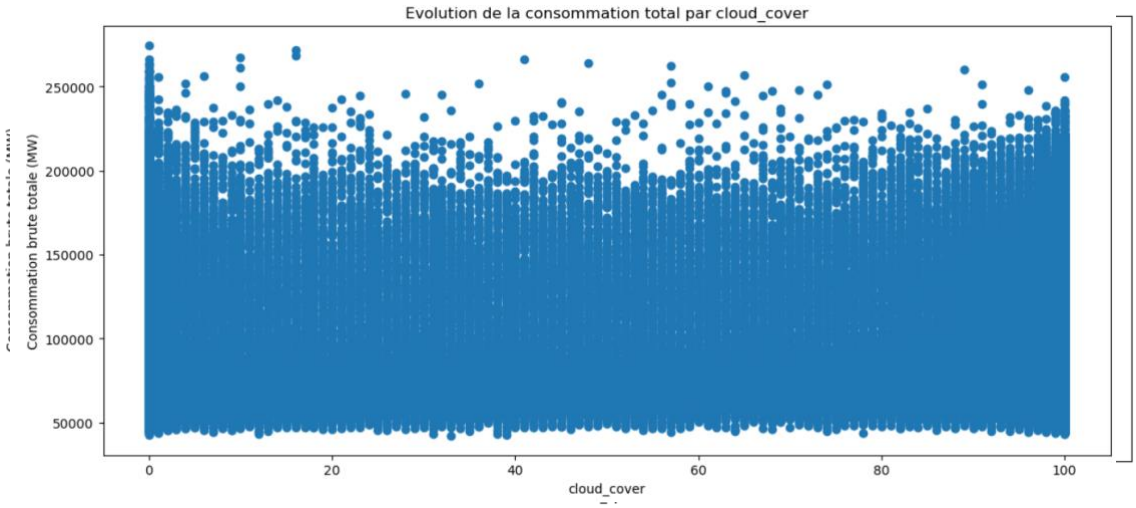
L'objectif de cette étude était de prédire la consommation brute totale d'énergie (regroupant à la fois l'électricité et le gaz) en France pour une date donnée. Pour y parvenir, nous avons enrichi notre jeu de données initial en intégrant des variables calendaires (comme le mois de l'année, les jours fériés, etc.) et des données météorologiques (température, humidité, etc.). Une analyse approfondie de la distribution de ces variables et de leurs liens avec la consommation a révélé des corrélations significatives : par exemple, une hausse de la consommation énergétique pendant les mois d'hiver ou lors des périodes de basses températures.

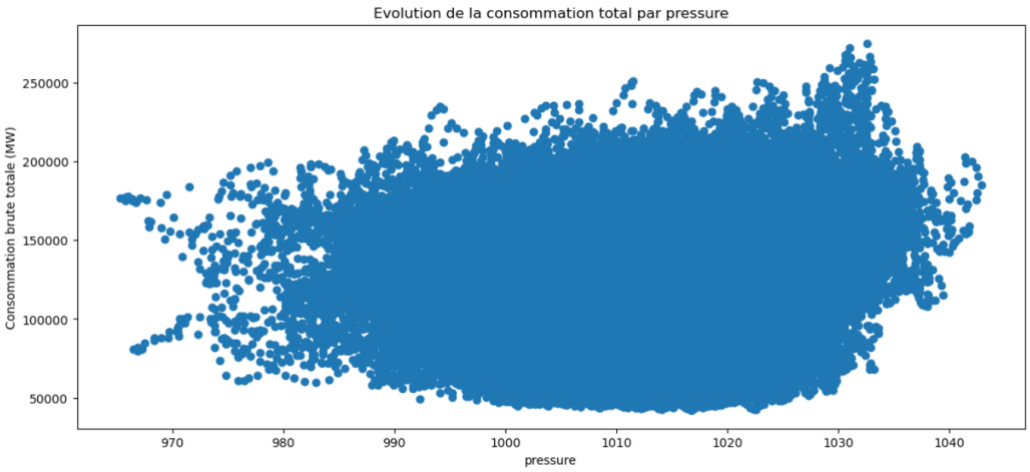
Plusieurs modèles ont ensuite été entraînés et comparés afin d'identifier celui offrant les meilleures performances. Contre toute attente, c'est le modèle baseline naïf (qui se base simplement sur la valeur de consommation de la veille) qui s'est avéré le plus efficace, grâce à la forte autocorrélation de la série temporelle. Cependant, dans les situations où la consommation de la veille n'est pas connue, le modèle Gradient Boosting se distingue comme la meilleure alternative, fournissant une estimation fiable de la consommation en exploitant les données météorologiques et calendaires.

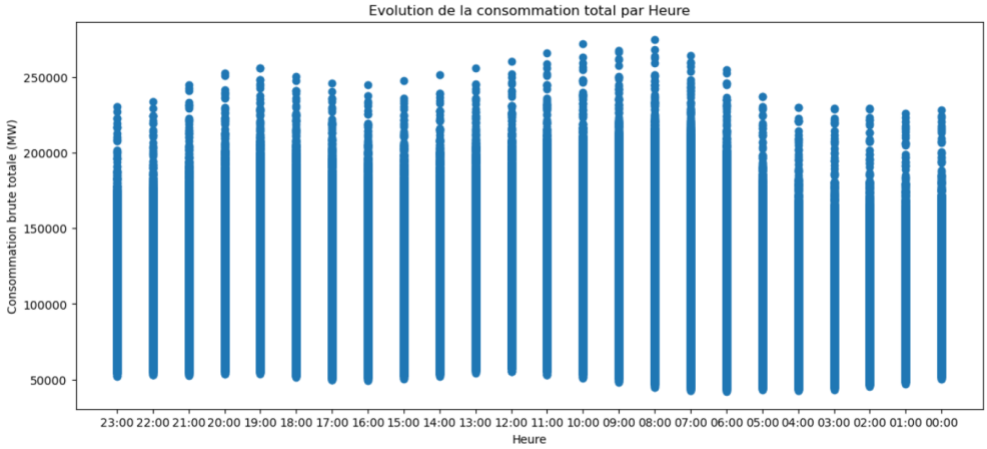
V. Annexes

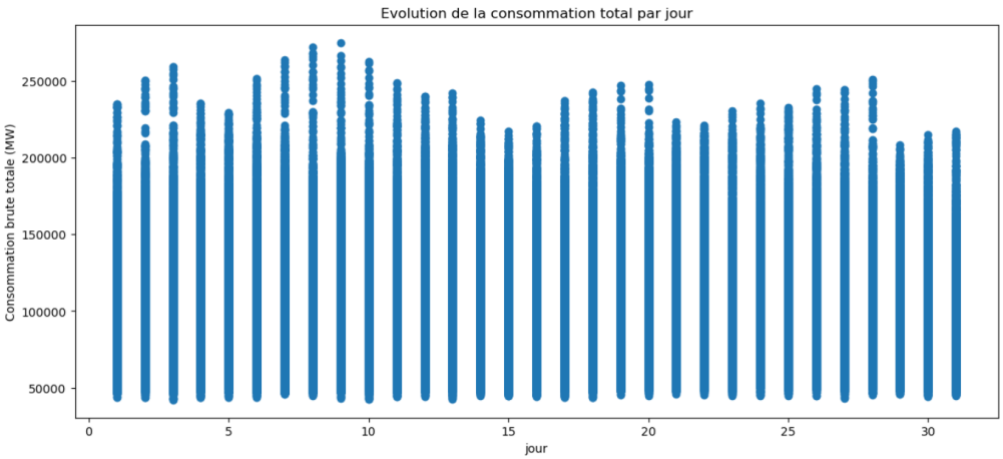
Evolution de la consommation par rapport à nos autres variables

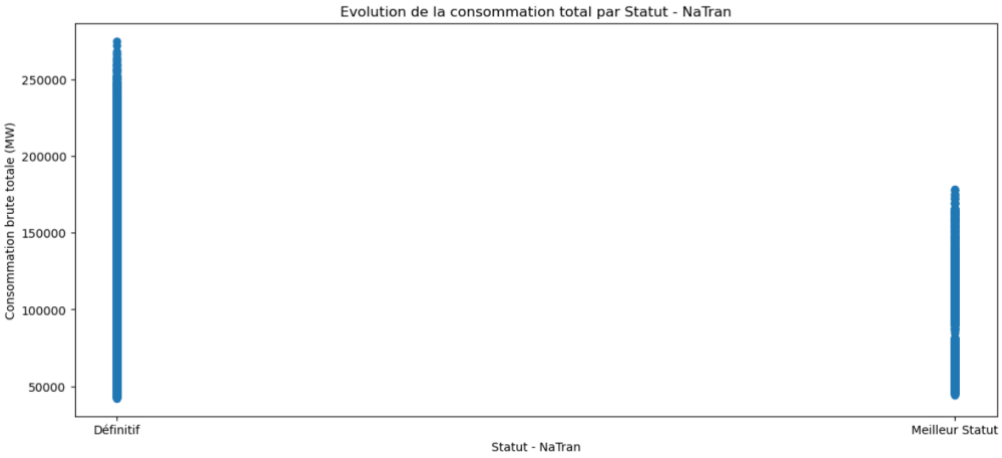


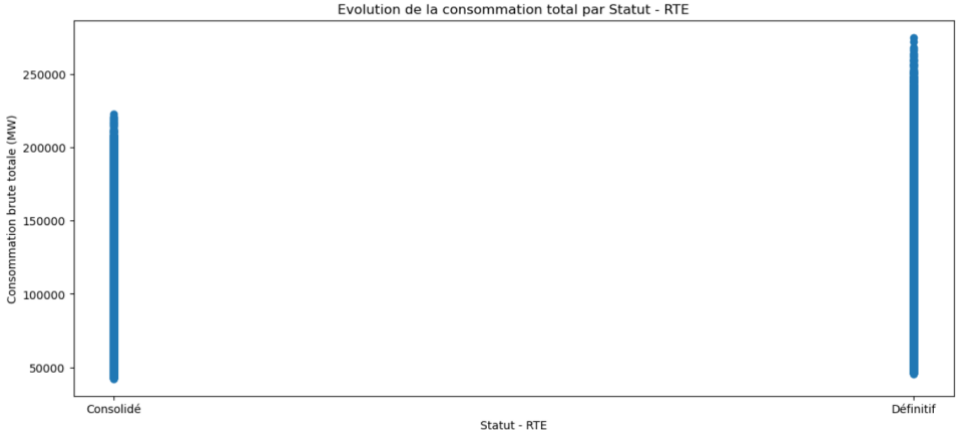


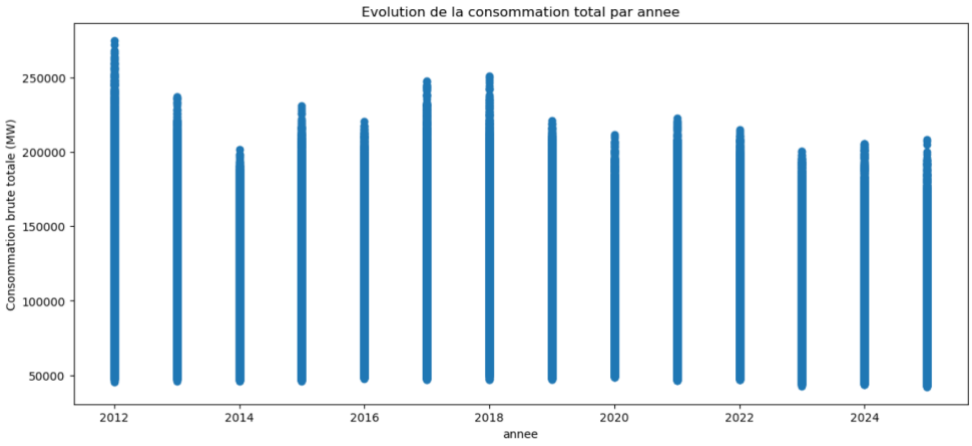

Evolution de la consommation total par pressure
Consommation brute totale (MW)
250000
200000
150000
100000
50000
970
980
990
1000
1010
1020
1030
1040
pressure

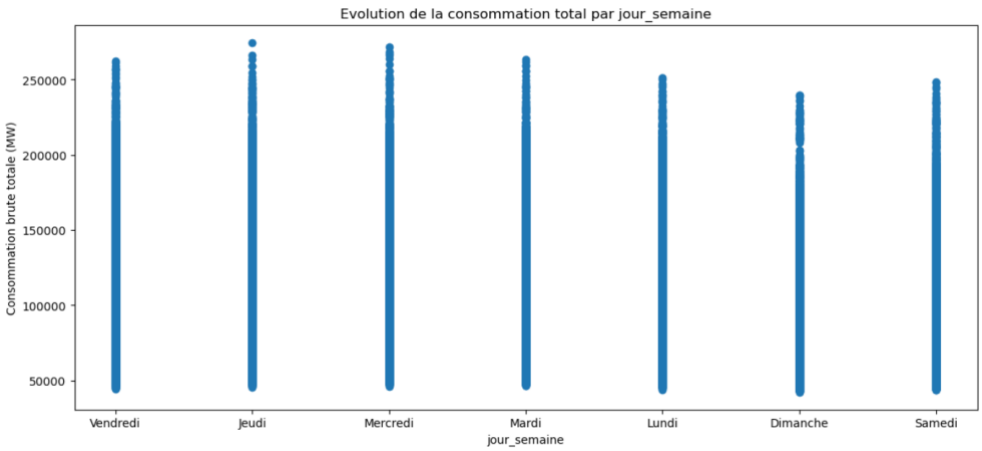

Evolution de la consommation total par Heure
Consommation brute totale (MW)
250000
200000
150000
100000
50000
23:00 22:00 21:00 20:00 19:00 18:00 17:00 16:00 15:00 14:00 13:00 12:00 11:00 10:00 09:00 08:00 07:00 06:00 05:00 04:00 03:00 02:00 01:00 00:00
Heure

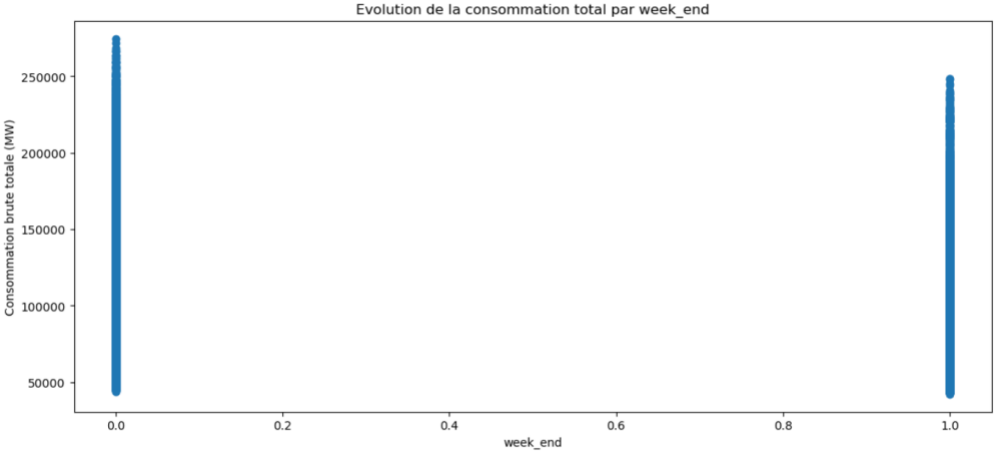

Evolution de la consommation total par jour
Consommation brute totale (MW)
250000
200000
150000
100000
50000
0
5
10
15
20
25
30
jour

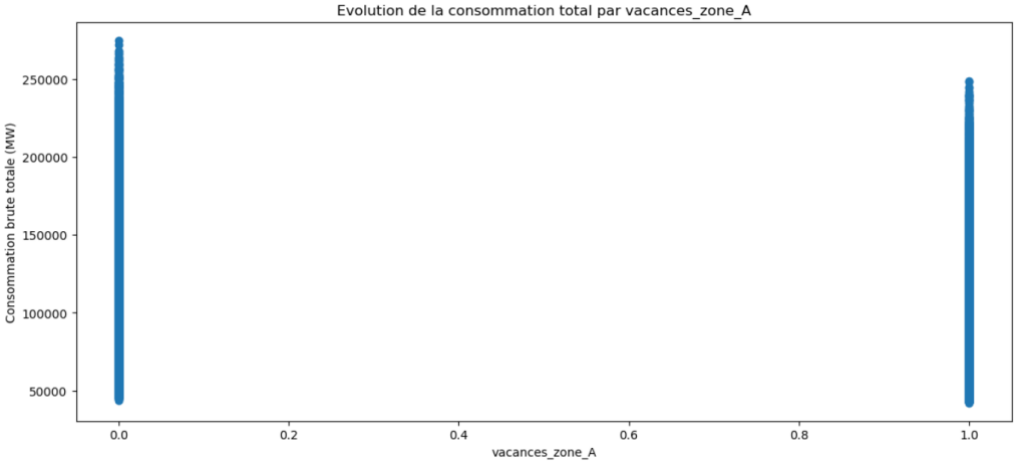

Evolution de la consommation total par Statut - NaTran
Consommation brute totale (MW)
250000
200000
150000
100000
50000
Définitif
Meilleur Statut
Statut - NaTran

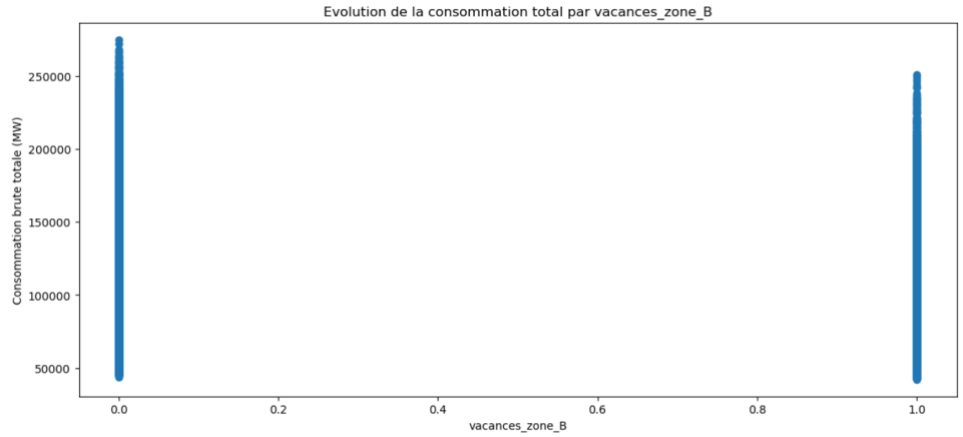

Evolution de la consommation total par Statut - RTE
Consommation brute totale (MW)
250000
200000
150000
100000
50000
Consolidé
Définitif
Statut - RTE

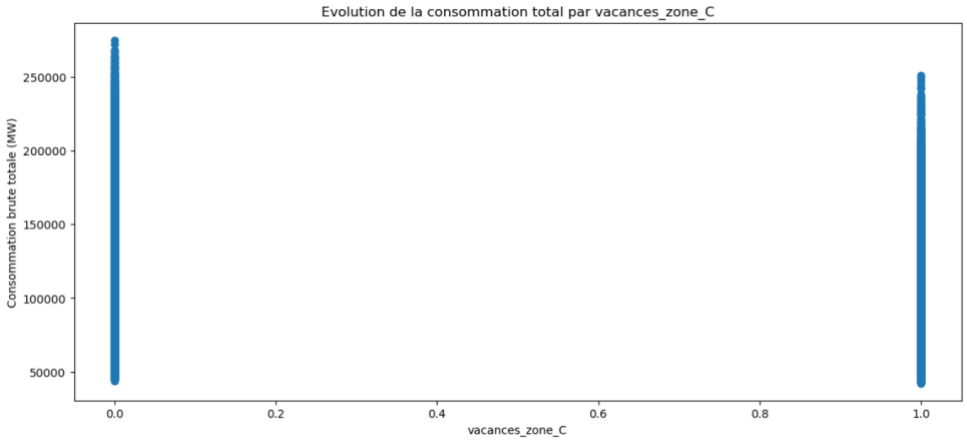

Evolution de la consommation total par annee
Consommation brute totale (MW)
250000
200000
150000
100000
50000
2012
2014
2016
2018
2020
2022
2024
annee


Evolution de la consommation total par jour_semaine
Consommation brute totale (MW)
250000
200000
150000
100000
50000
Vendredi
Jeudi
Mercredi
Mardi
Lundi
Dimanche
Samedi
jour_semaine


Evolution de la consommation total par week_end
Consommation brute totale (MW)
250000
200000
150000
100000
50000
0.0
0.2
0.4
0.6
0.8
1.0
week_end


Evolution de la consommation total par vacances_zone_A
Consommation brute totale (MW)
250000
200000
150000
100000
50000
0.0
0.2
0.4
0.6
0.8
1.0
vacances_zone_A


Evolution de la consommation total par vacances_zone_B
Consommation brute totale (MW)
250000
200000
150000
100000
50000
0.0
0.2
0.4
0.6
0.8
1.0
vacances_zone_B


Evolution de la consommation total par vacances_zone_C
Consommation brute totale (MW)
250000
200000
150000
100000
50000
0.0
0.2
0.4
0.6
0.8
1.0
vacances_zone_C