

A photograph of the Seattle skyline at sunset. The Space Needle is prominent on the left, illuminated with warm orange lights. The city buildings are silhouetted against a vibrant orange and yellow sky. In the distance, Mount Rainier is visible as a dark, hazy peak. A bridge with red lights is visible in the foreground.

# Prédiction des besoins en consommation des bâtiments

Emeline Tapin - 07/2023

# Prédiction des besoins en consommation des bâtiments

## Sommaire

---

- Rappel du contexte et des objectifs
- Analyse exploratoire & Feature engineering
- Prédiction de la consommation énergétique totale
- Prédiction des émissions de gaz à effets de serre
- Conclusion

# Rappel du contexte et des objectifs

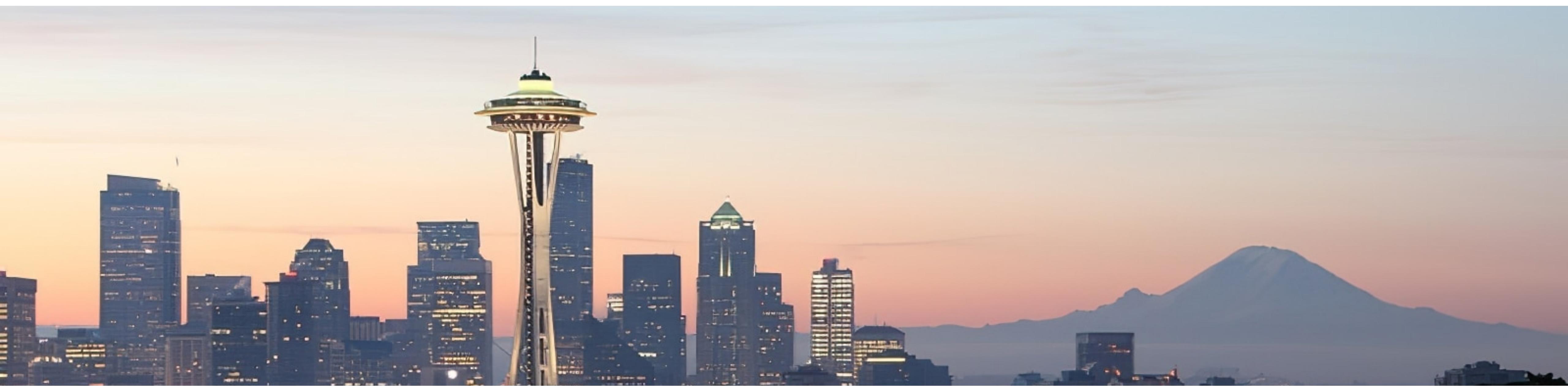
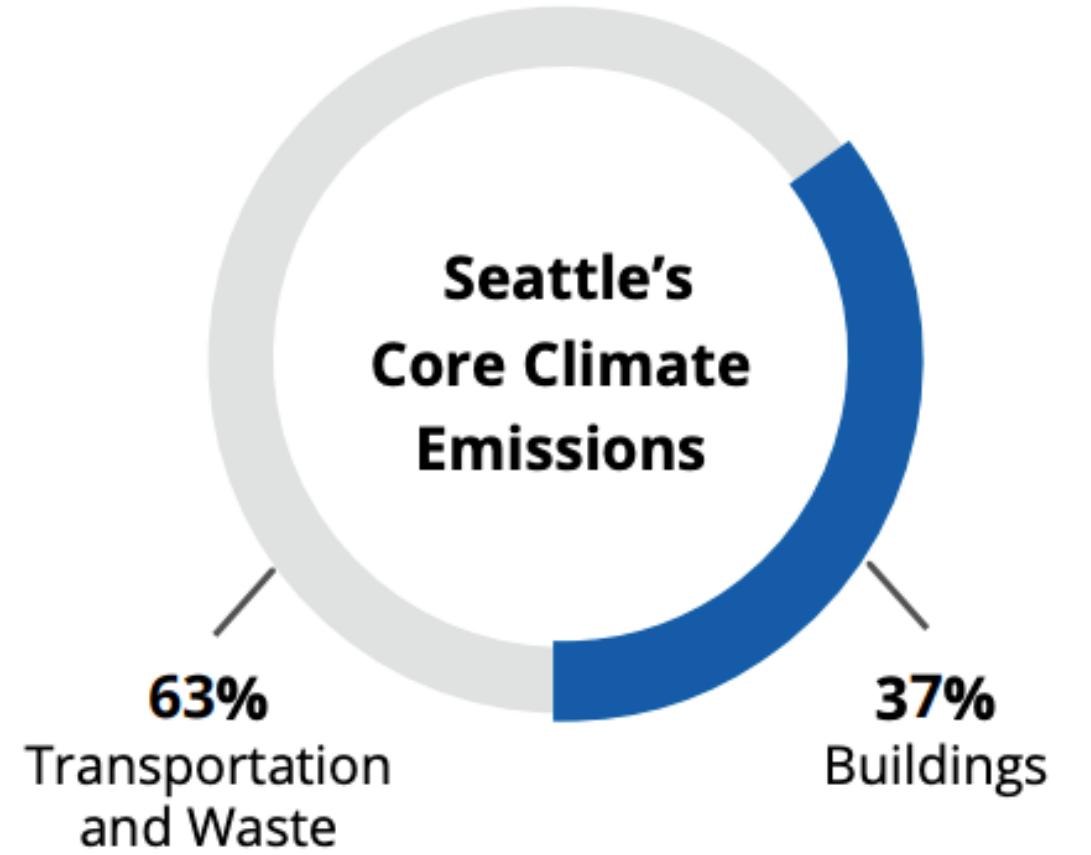
## Sommaire

- Rappel du contexte et des objectifs
  - Neutralité carbone 2050
  - Overview des données
- Analyse exploratoire & Feature engineering
- Prédiction de la consommation énergétique totale
- Prédiction des émissions de gaz à effets de serre
- Conclusion

# Rappel du contexte et des objectifs

## Neutralité carbone - 2050

- Objectif de ville neutre en émissions de carbone en 2050
  - Bâtiments : 1/3 des émissions de Seattle
  - Politique de benchmarking et de suivi des bâtiments non résidentiels et multi-familiaux (Building Energy Benchmarking and Reporting Program)
- ➡ Enjeu : modèles de prédiction des consommations énergétique et émissions de GES des **bâtiments non destinés à l'habitation**



# Rappel du contexte et des objectifs

## Overview de données



- **Seattle Open Data - Données 2016 - Mise à jour le 21 avril 2023**
- **3 376 bâtiments et 46 variables :**
  - *Identification des bâtiments* : ID, Building Type, etc.
  - *Année de construction*
  - *Utilisation* : 3 premiers types d'utilisations et surfaces associées
  - *Données physique* :
    - Superficies : total, bâtiment, parking, etc.
    - Architecture du bâtiment : nombre de bâtiment, nombre d'étage, etc.
  - *Géographie* : Adresse, Latitude, Longitude, Quartier, etc.
  - *Consommation énergétique* : Consommation totale et surfacique, mix énergétique etc.
  - *Emission de GES* : Emission totale et surfacique
  - *Autres* : Outliers, données manquantes, commentaires, etc.

# Analyse exploratoire & Feature engineering

## Sommaire

- Rappel du contexte et des objectifs
- Analyse exploratoire & Feature engineering
  - Qualité des données & Sélection des donnée
  - Targets
  - Features : Données physiques
  - Features : Âge des bâtiments
  - Features : Utilisation des bâtiments
  - Features : Géographie
  - Features : Mix énergétique
  - Features : EnergyStar Score
- Prédiction de la consommation énergétique totale
- Prédiction des émissions de gaz à effets de serre
- Conclusion

# Analyse exploratoire & Feature engineering

## Qualité des données & Sélection des données

### QUALITE DES DONNEES :

- *Doublons* : 0 doublons
- *Erreurs de type* : 0 erreurs de type
- *Outliers* : suppression des outliers identifiés dans la BDD et sélection des bâtiments avec « Compliant » status
- *Valeurs manquantes* : peu, traitement des valeurs manquantes pour les indicateurs sélectionnés à l'aide de la variable 'Default Data', hormis pour 'EnergyStarScore' traité séparément

### SELECTION DES CATEGORIES DE DONNES DES BATIMENT NON RESIDENTIELS :

Age

Utilisation

EnergyStarScore

Consommation  
énergétique

Données physique

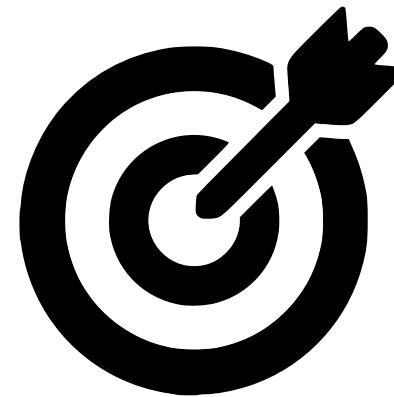
Géographie

Mix énergétique

Emission GES

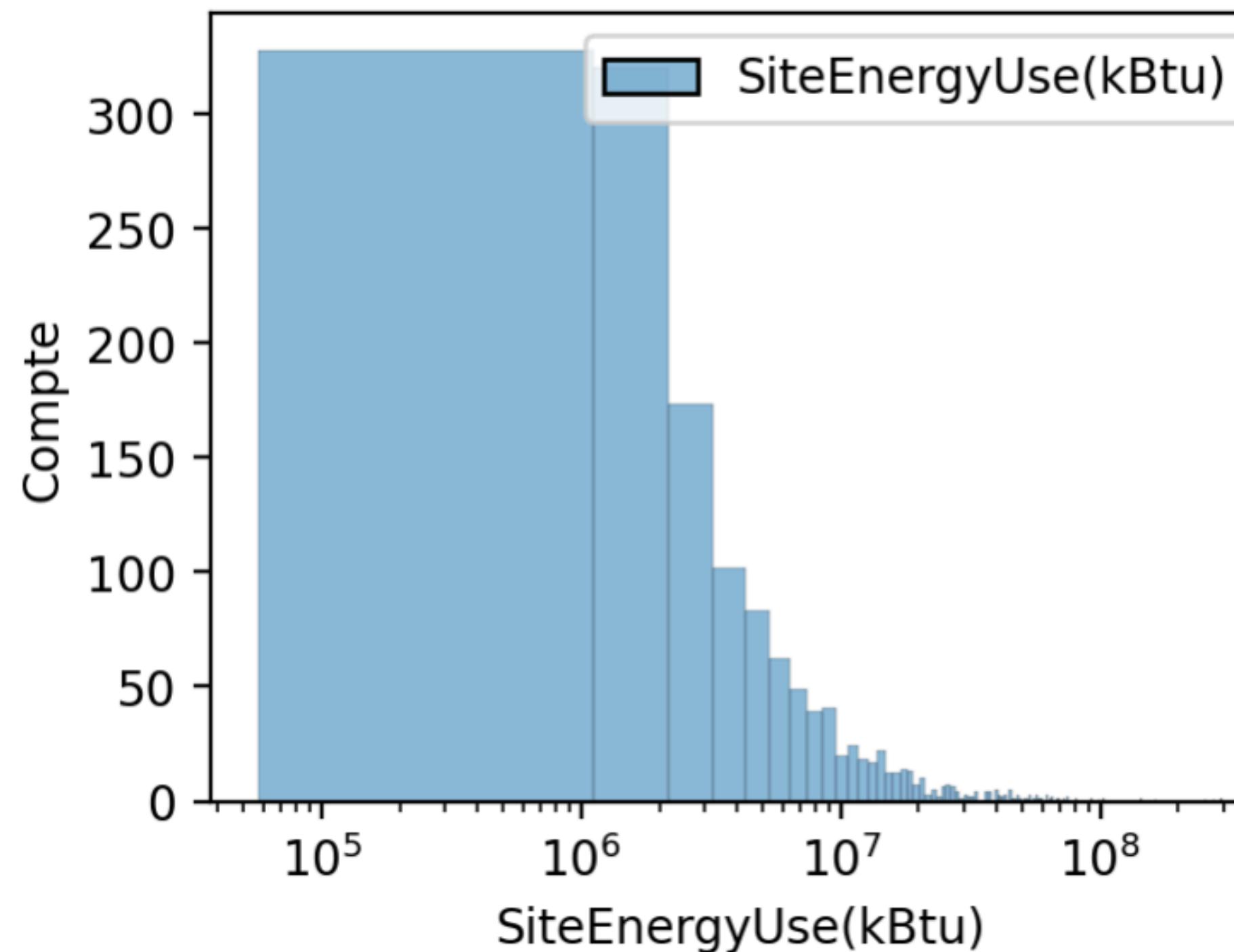
# Analyse exploratoire & Feature engineering

## Targets

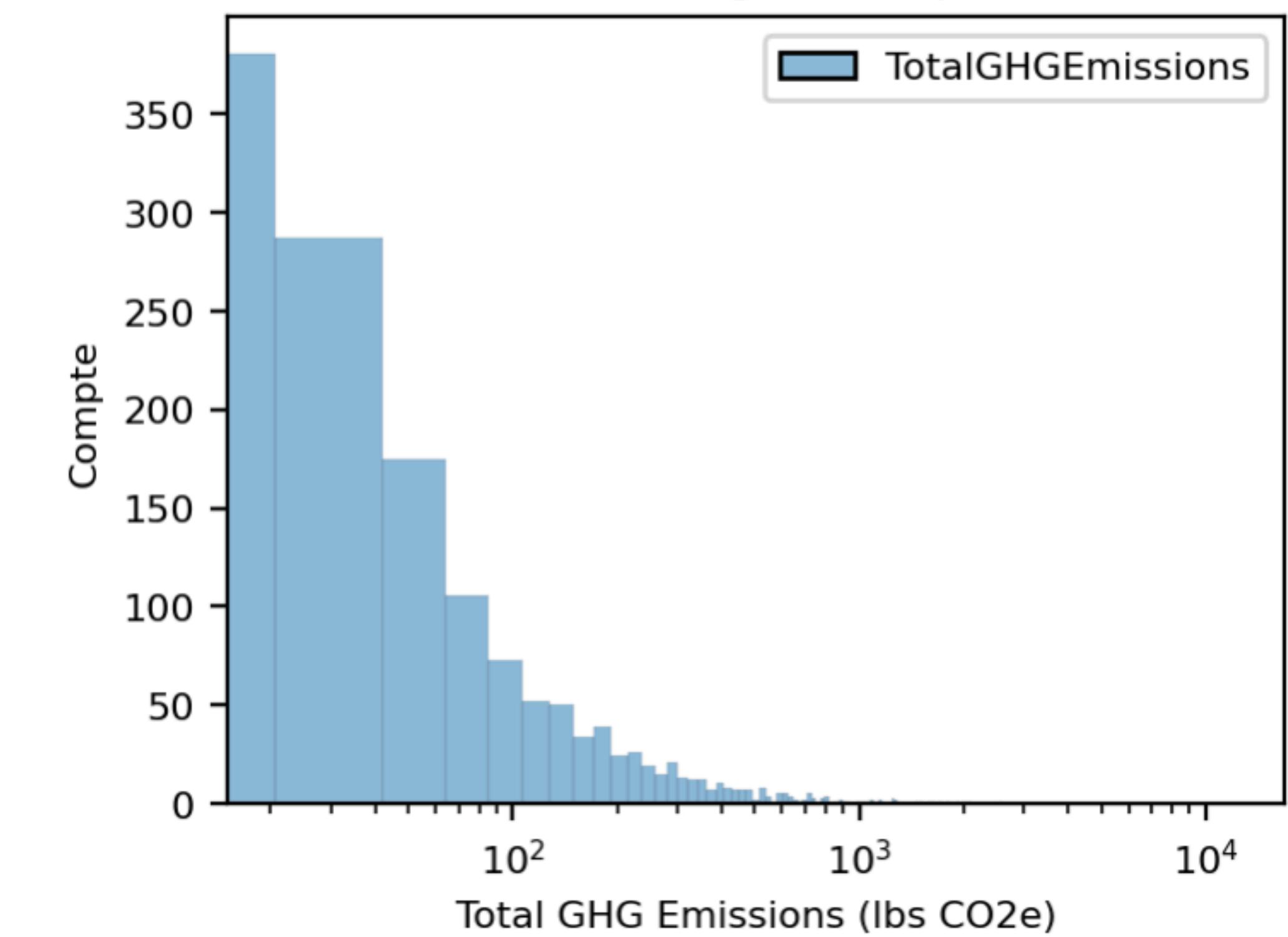


- Consommation énergétique : consommation totale ‘SiteEnergyUse(kBtu)’
- Émissions de GES : émission totale ‘TotalGHGEmissions’

Répartition de la variable SiteEnergyUse(kBtu)  
(échelle logarithmique)



Répartition de la variable émission des GES  
(échelle logarithmique)

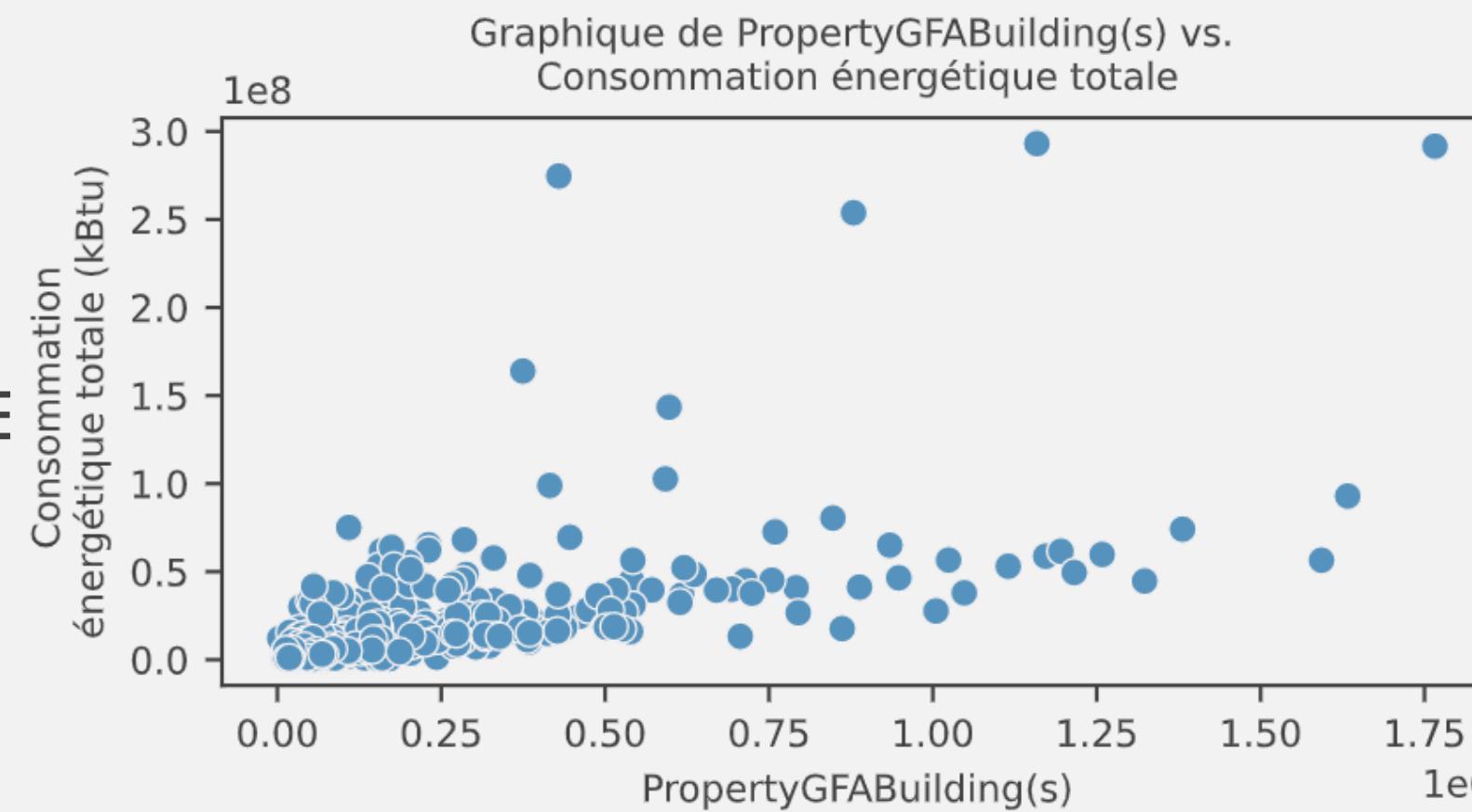


# Analyse exploratoire & Feature engineering

## Features : Données physiques

Catégorie :

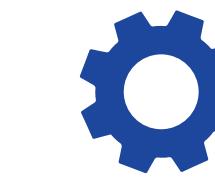
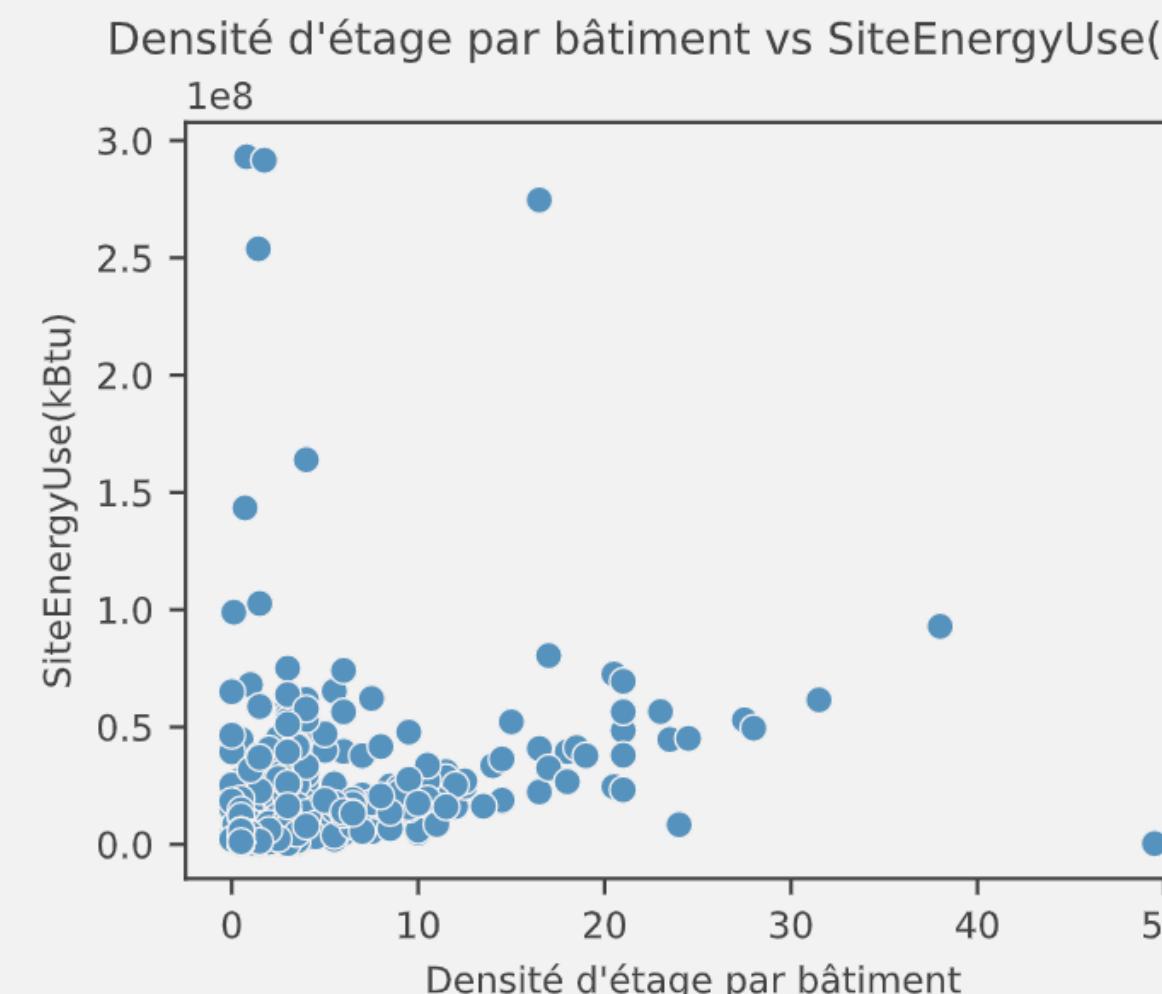
SURFACE



Variables sélectionnées :

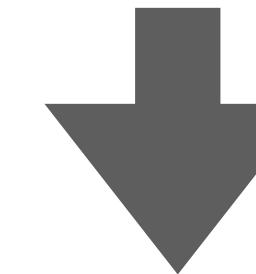
PropertyGFABuilding(s)  
Property GFAParking

ARCHITECTURE



Traitement :

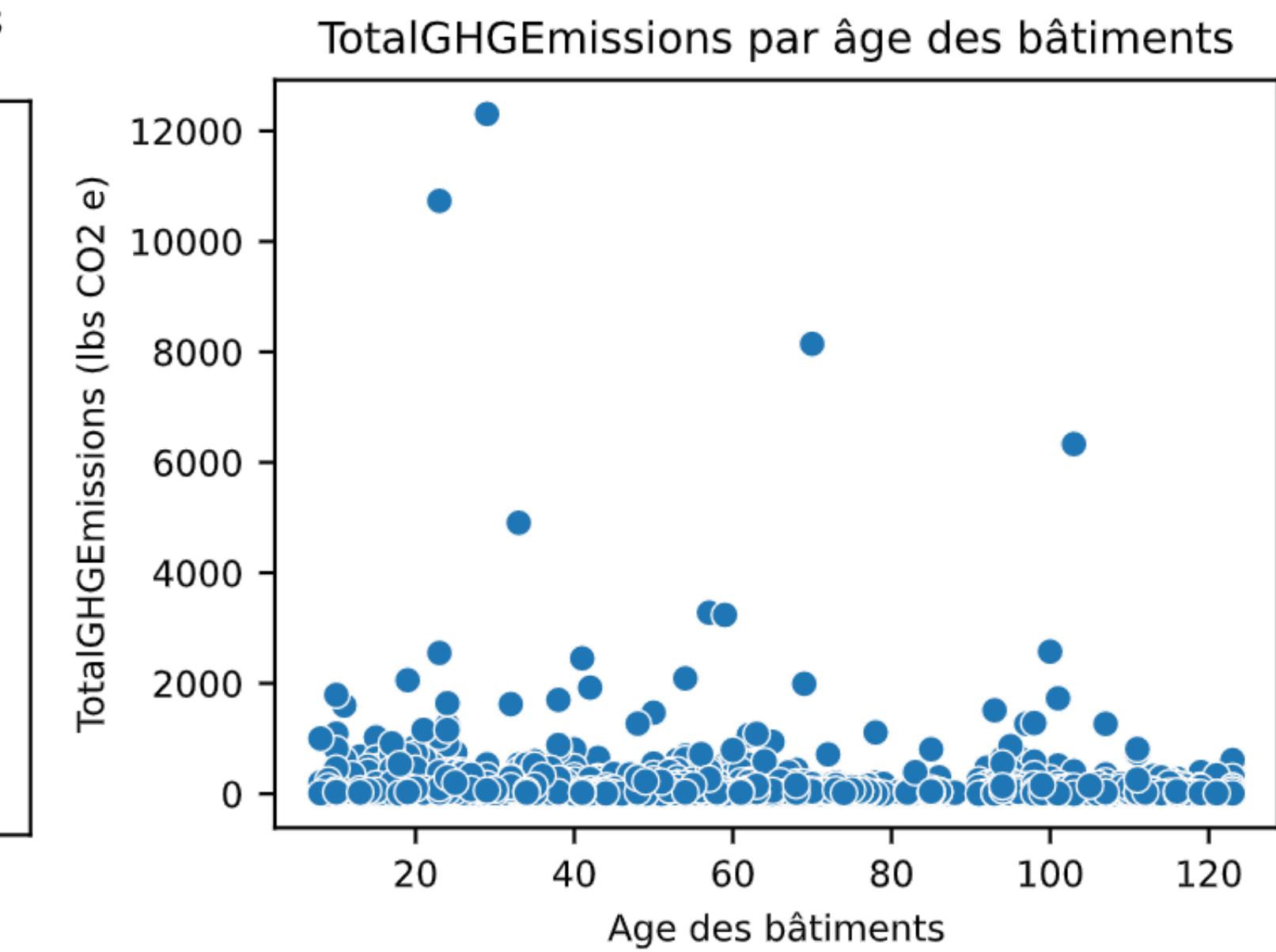
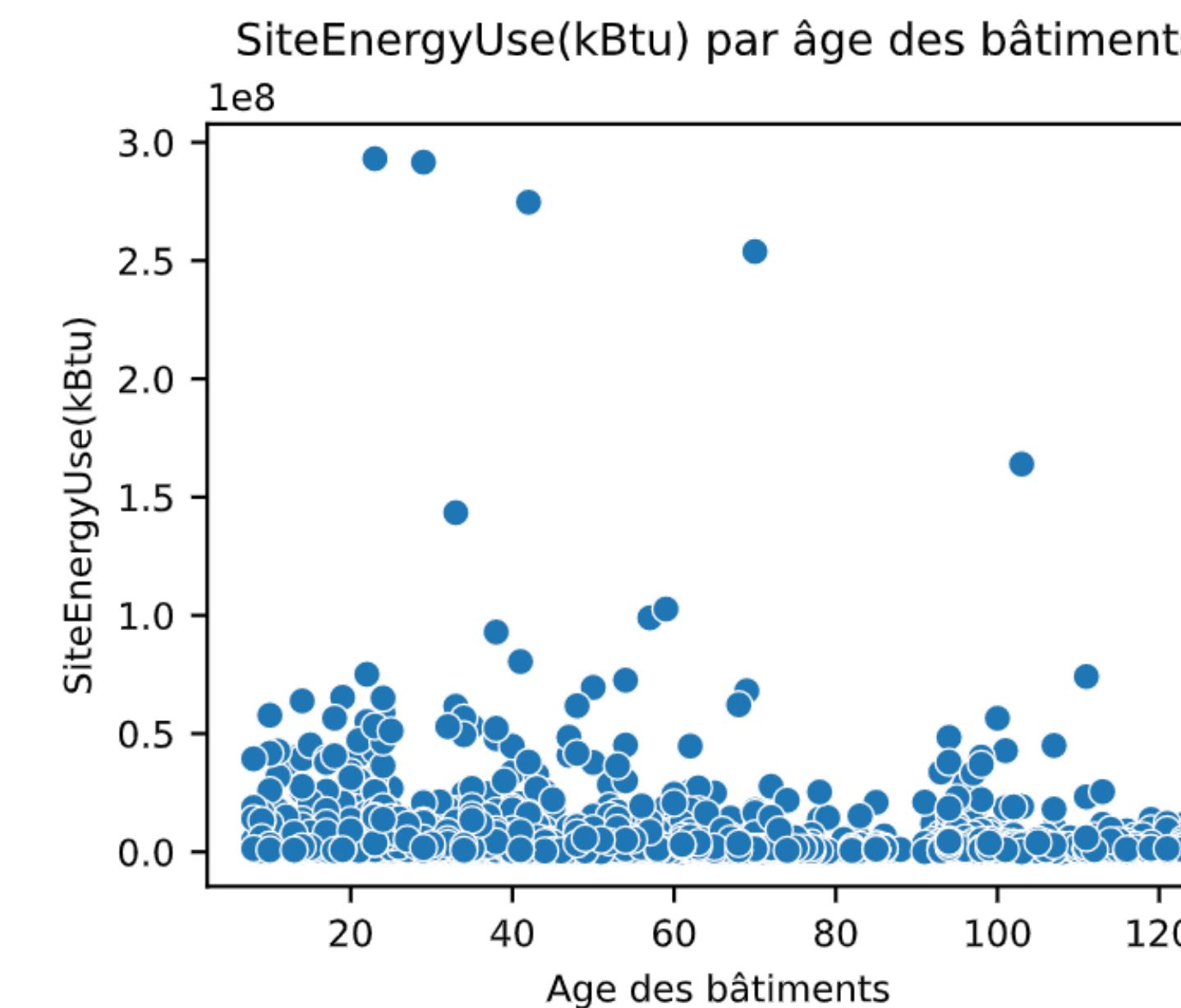
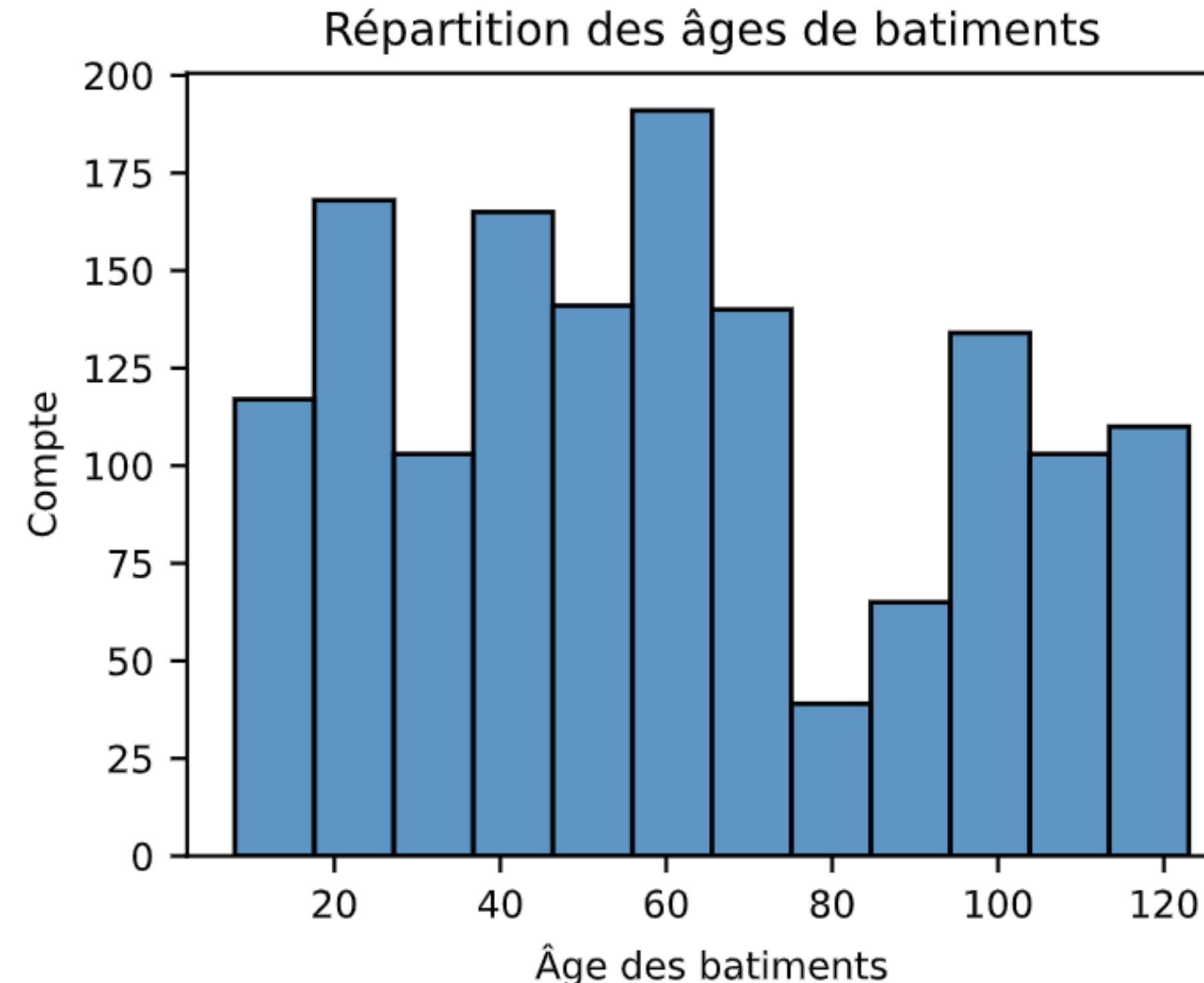
	PropertyGFABuilding(s)	PropertyGFAParking	Densite_etage
count	1.476000e+03	1476.000000	1476.000000
mean	9.855650e+04	14245.642276	2.195381
std	1.667536e+05	44525.035195	3.504821
min	3.636000e+03	0.000000	0.000000
25%	2.790000e+04	0.000000	0.500000
50%	4.585300e+04	0.000000	1.000000
75%	9.356875e+04	0.000000	2.500000
max	1.765970e+06	512608.000000	49.500000



Asymétrie de distribution et passage au log

# Analyse exploratoire & Feature engineering

## Features : Âge des bâtiments

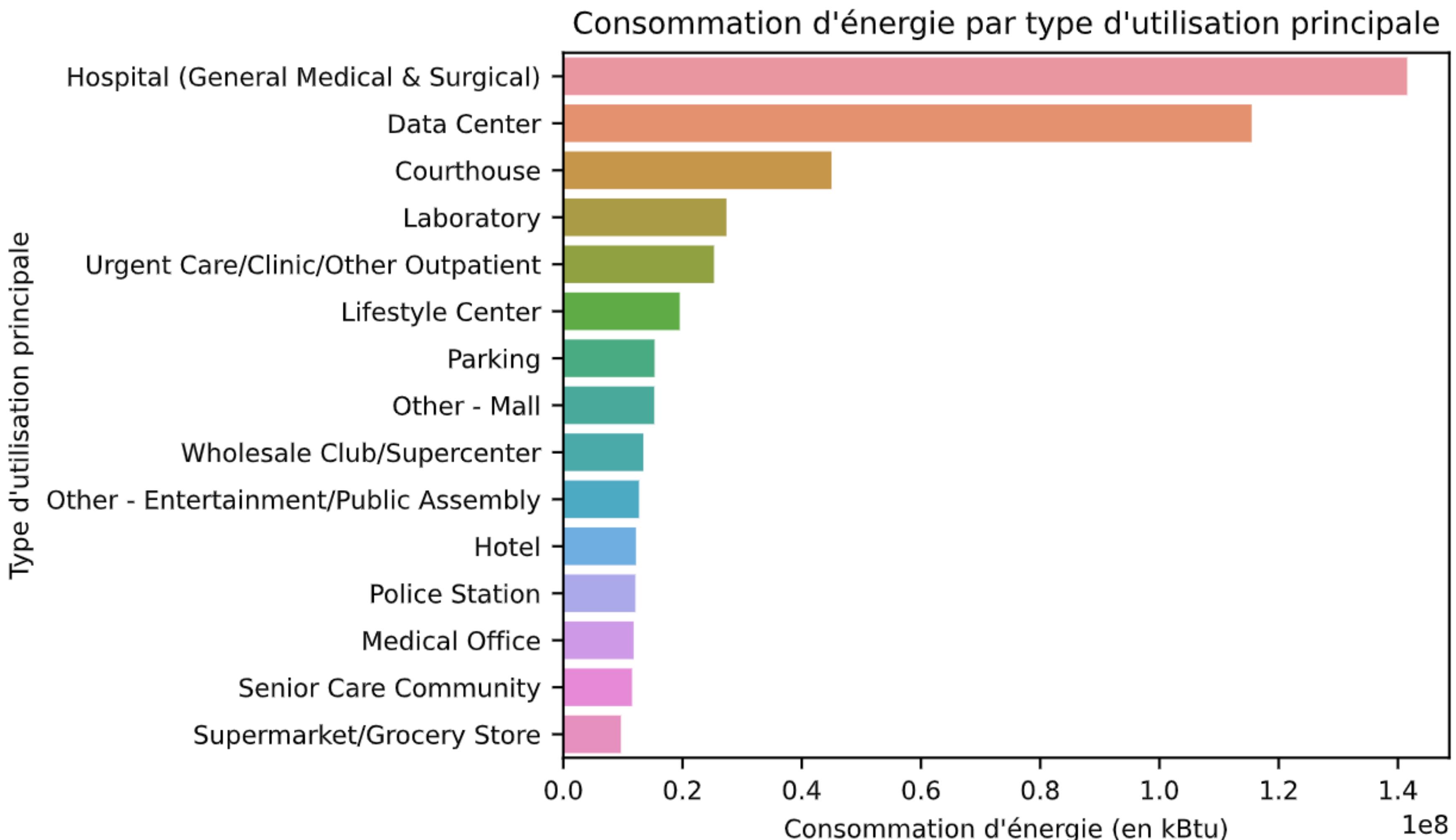


Traitement :

*Calcul à partir de l'année de construction du bâtiment*

# Analyse exploratoire & Feature engineering

## Features : Utilisation des bâtiments

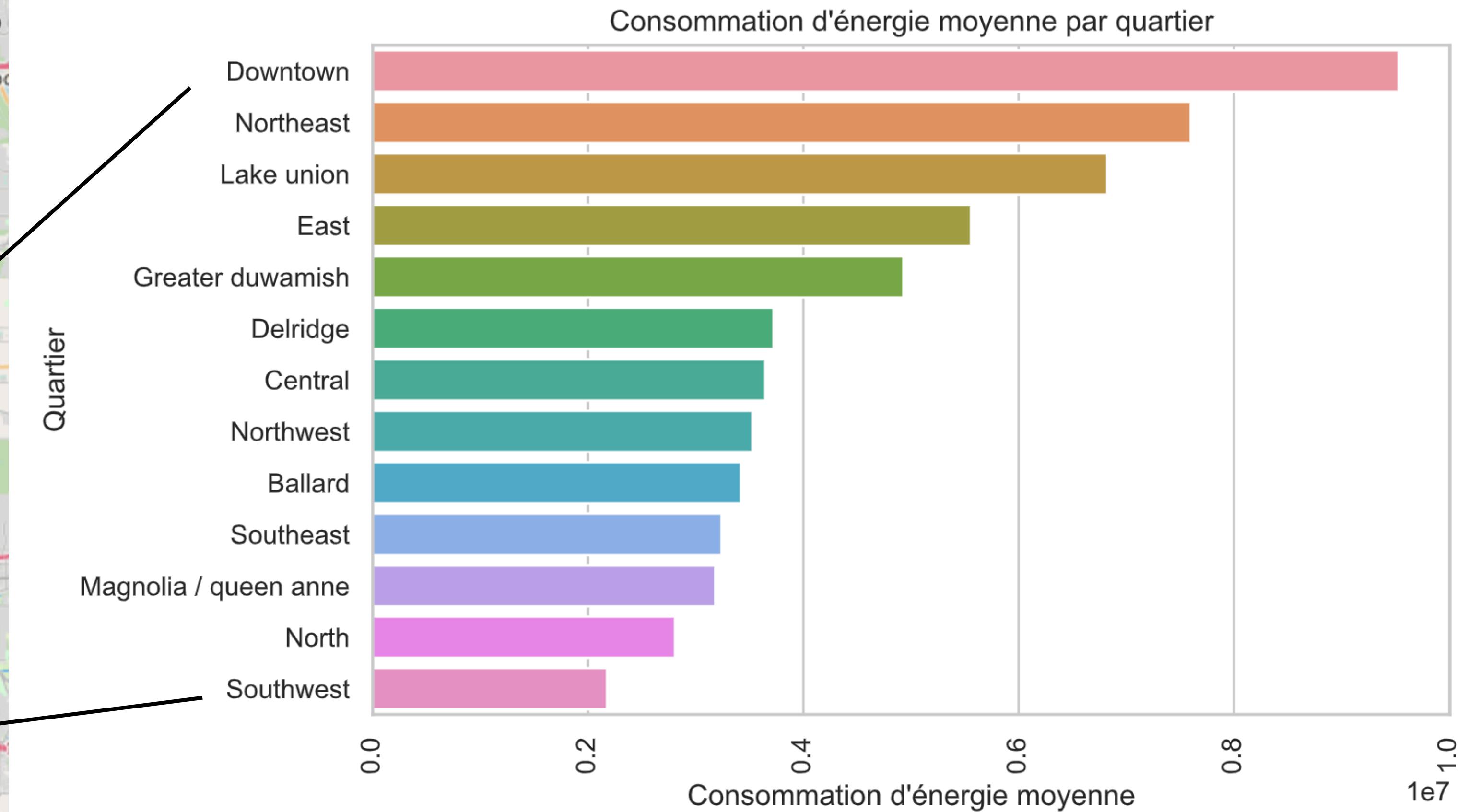
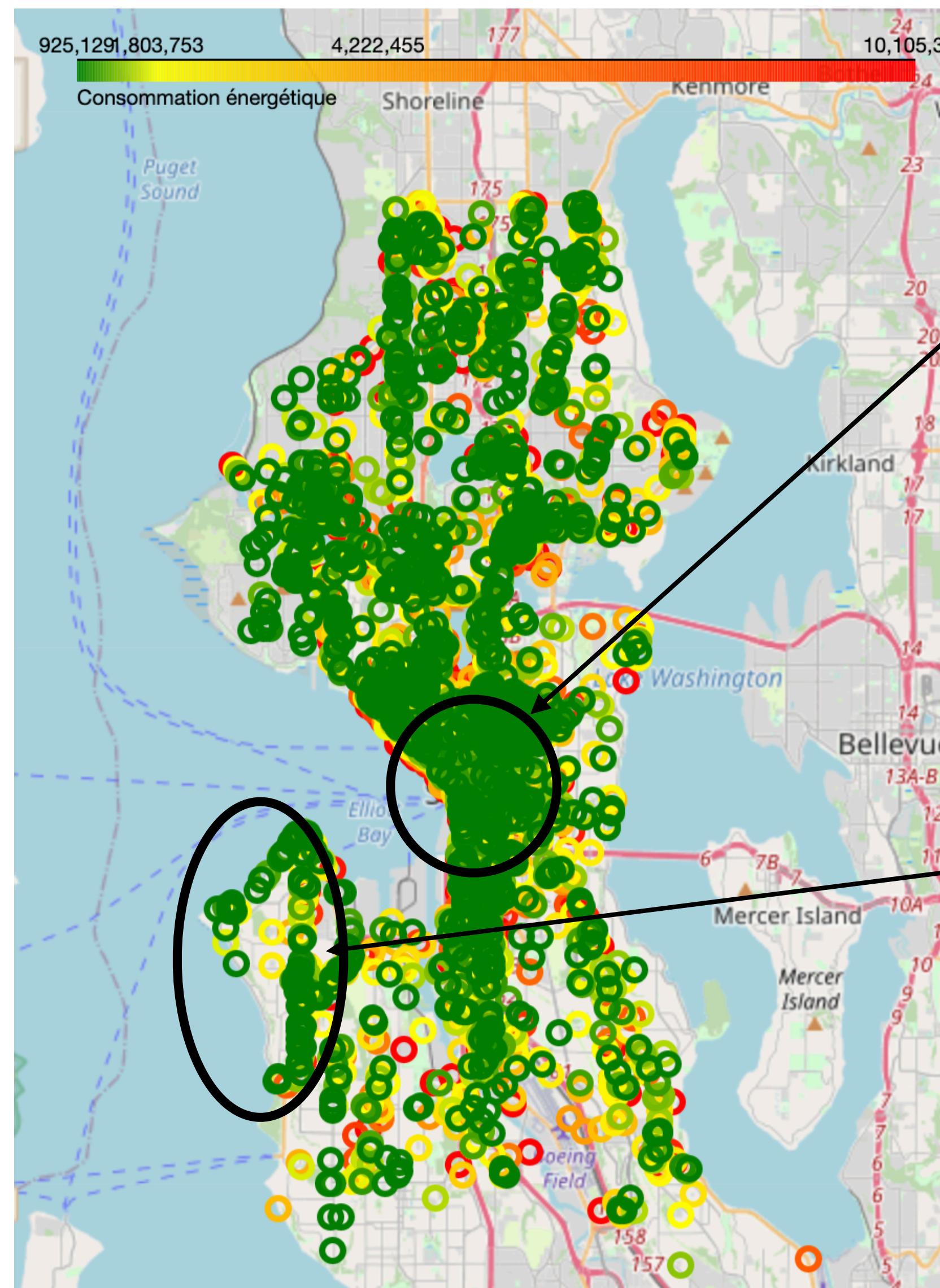


### Traitement :

- *One Hot Encoding*
  - ACP
  - Regroupement des catégories
- *Target Encoding*
- *Encoding Manuel*

# Analyse exploratoire & Feature engineering

## Features : Géographie



### Classes de variables géographiques :

- Quartier du jeu de données
- Nouveaux clusters de quartiers (K-Means)

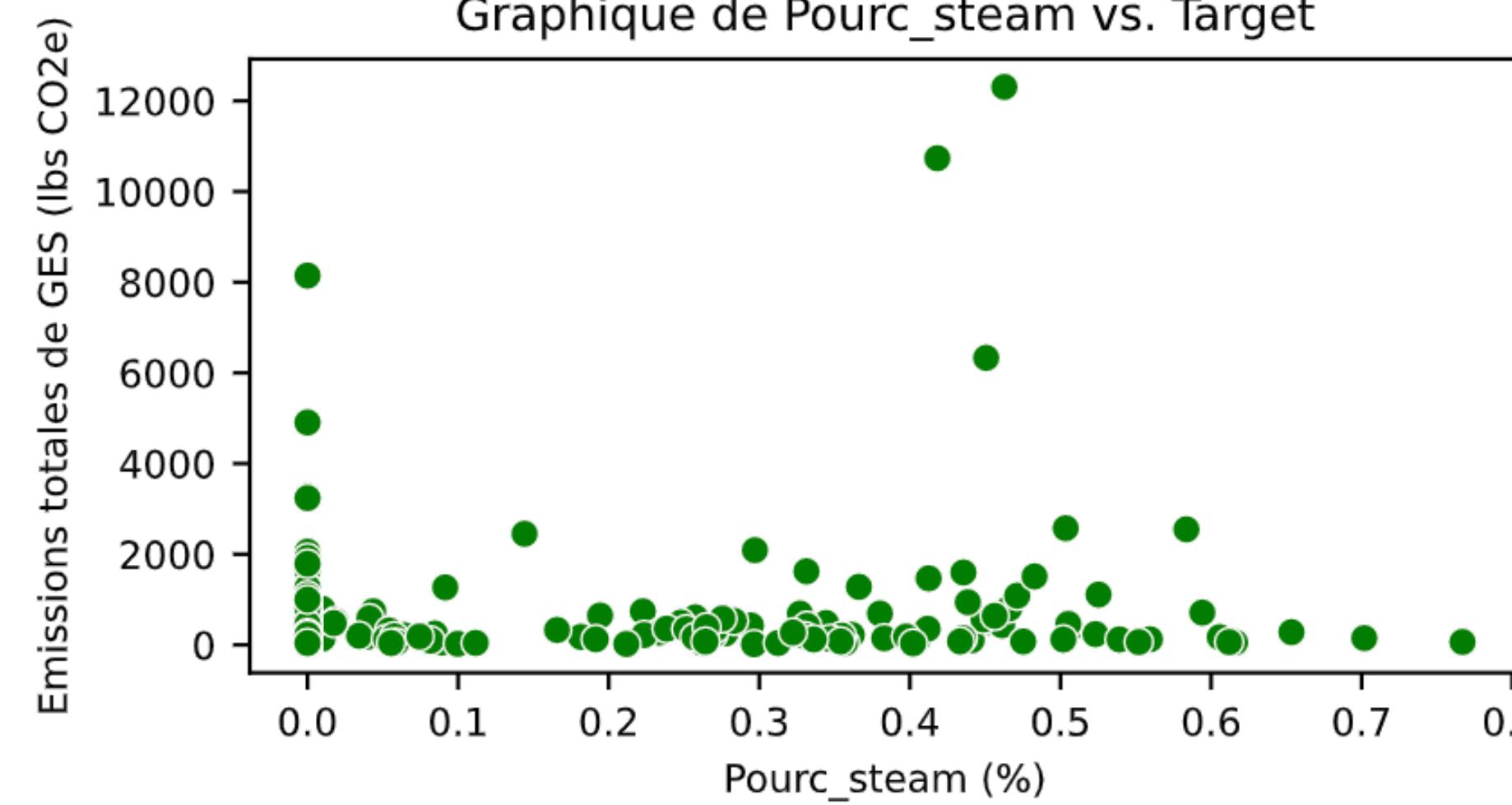
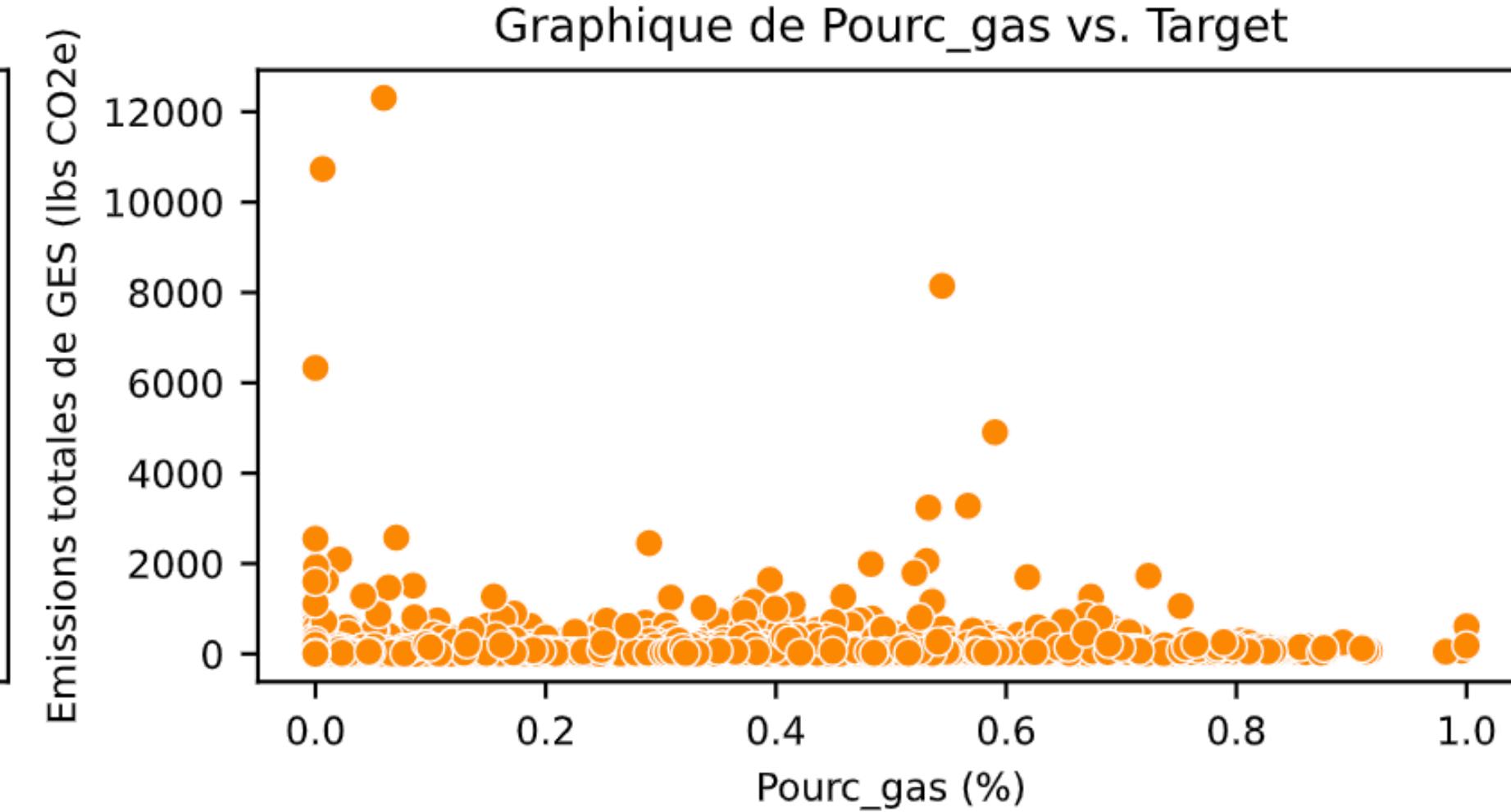
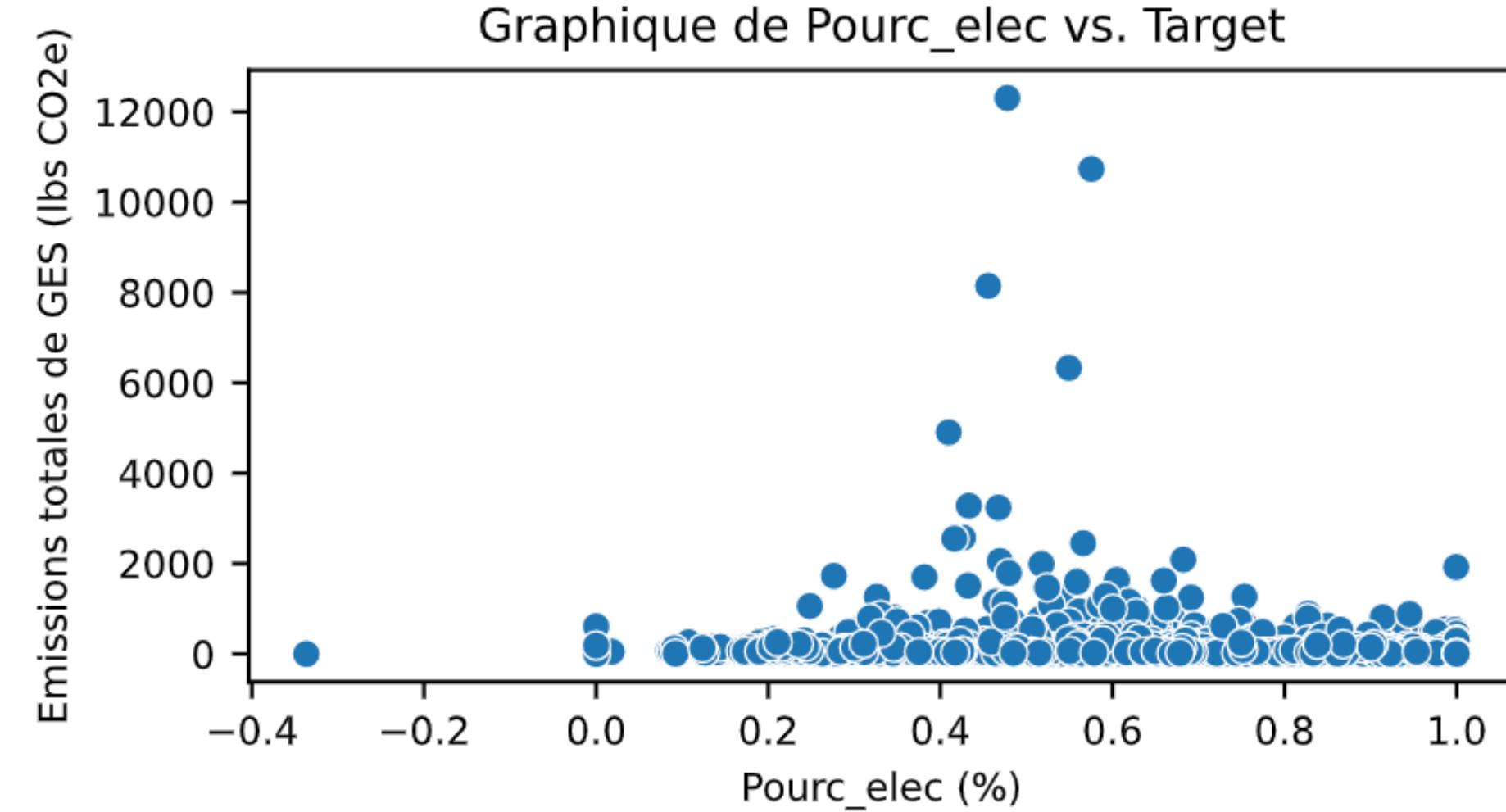
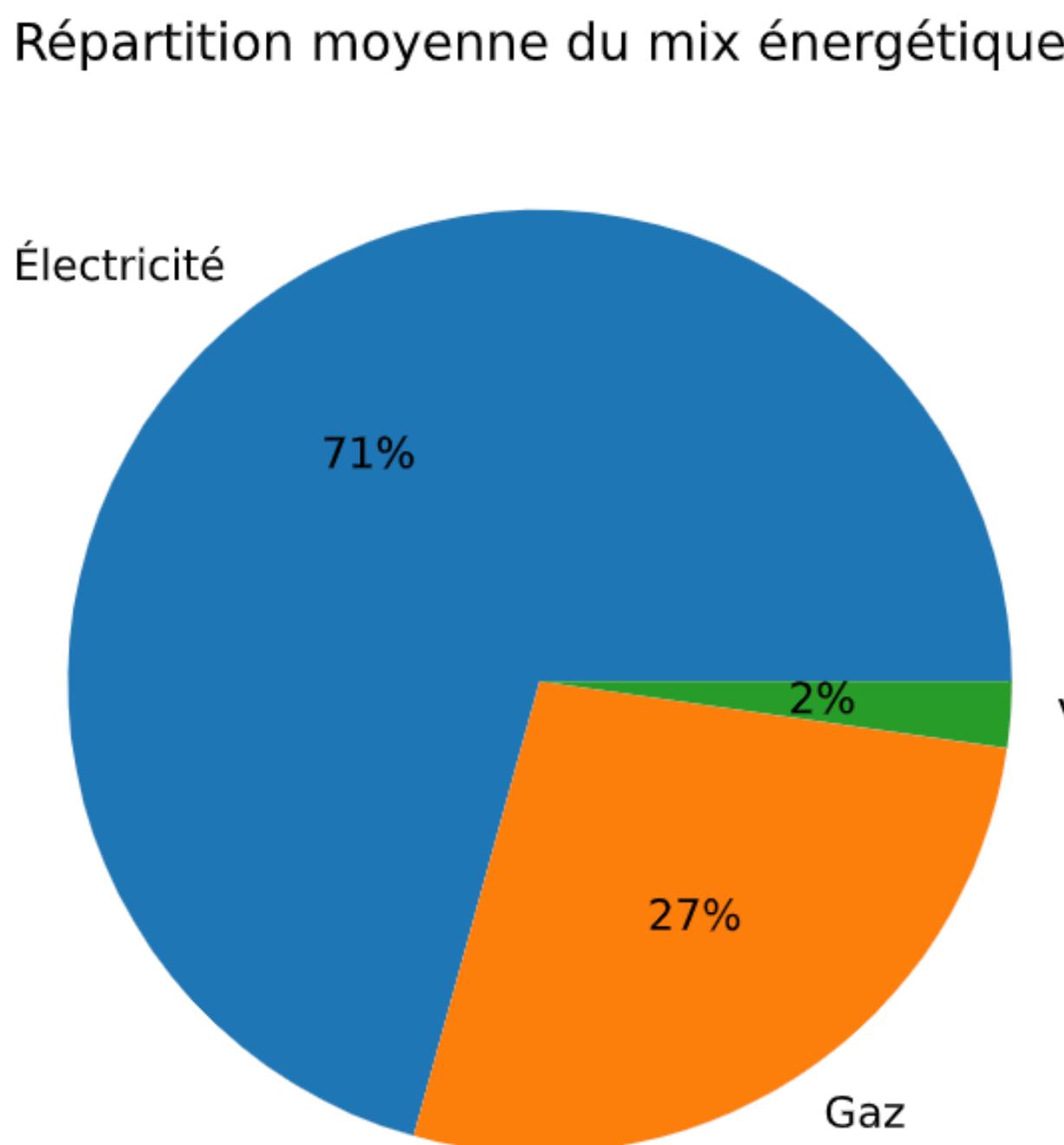


### Traitement :

- One Hot Encoding
- Target Encoding

# Analyse exploratoire & Feature engineering

## Features : Mix énergétique

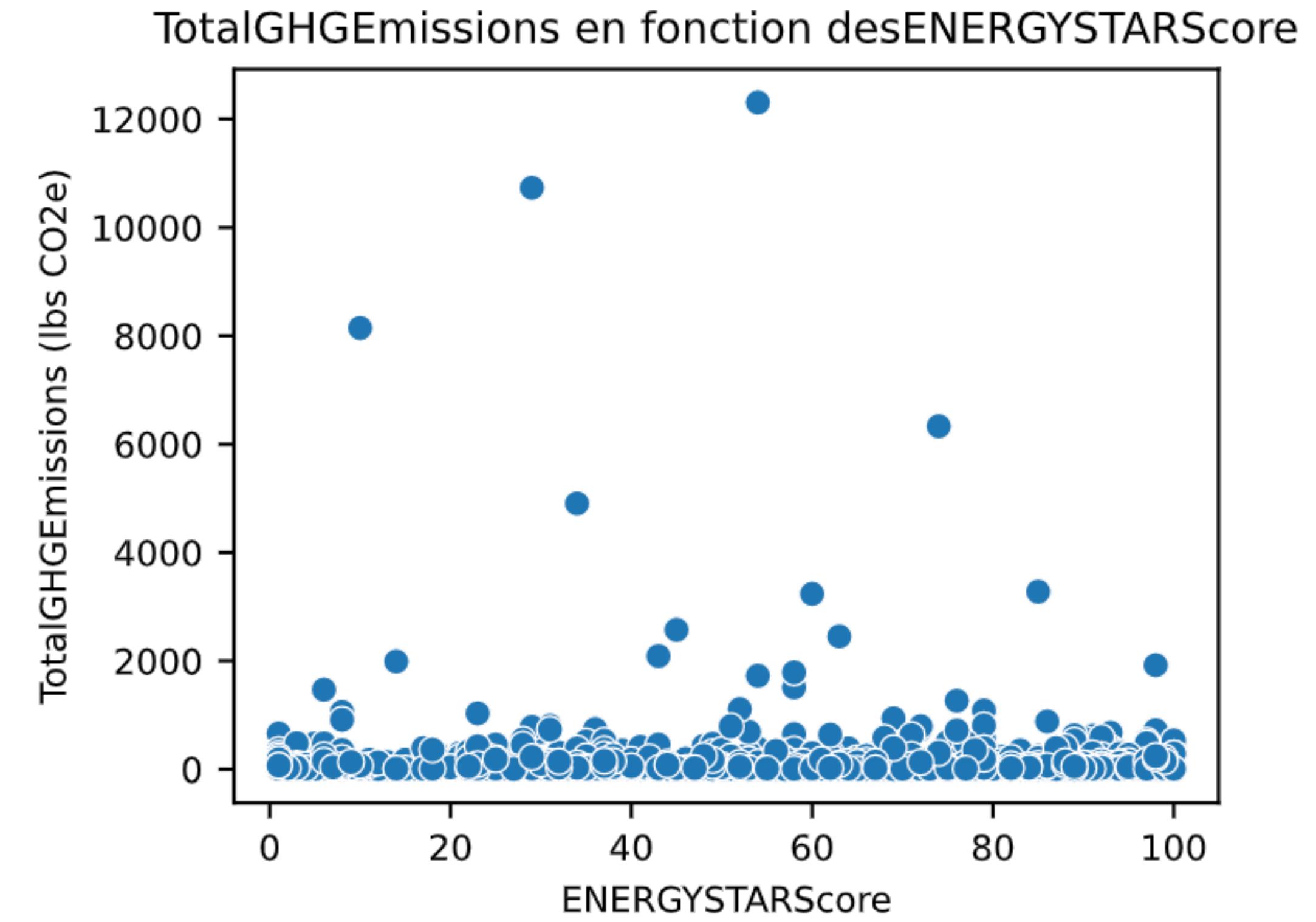
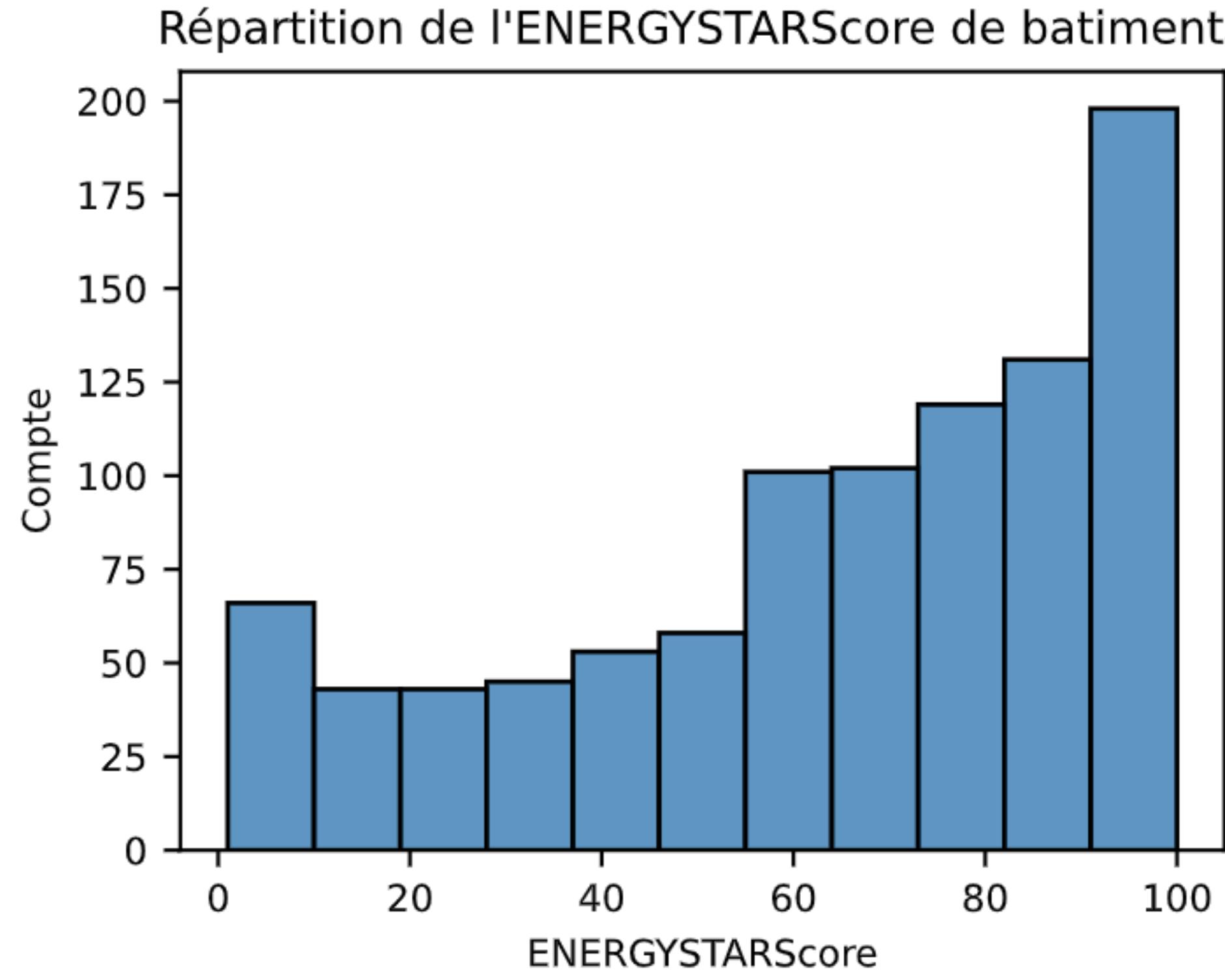


### Traitement :

*Calcul du mix énergétique à partir des données de benchmark pour chaque bâtiment*

# Analyse exploratoire & Feature engineering

## Features : Energy Star Score



Traitemet : 34% de valeurs manquantes :

- *Imputation*
- *Suppression des bâtiments avec valeurs manquantes (cf. Impact de l'EnergyScore sur modélisation des émissions de GES)*

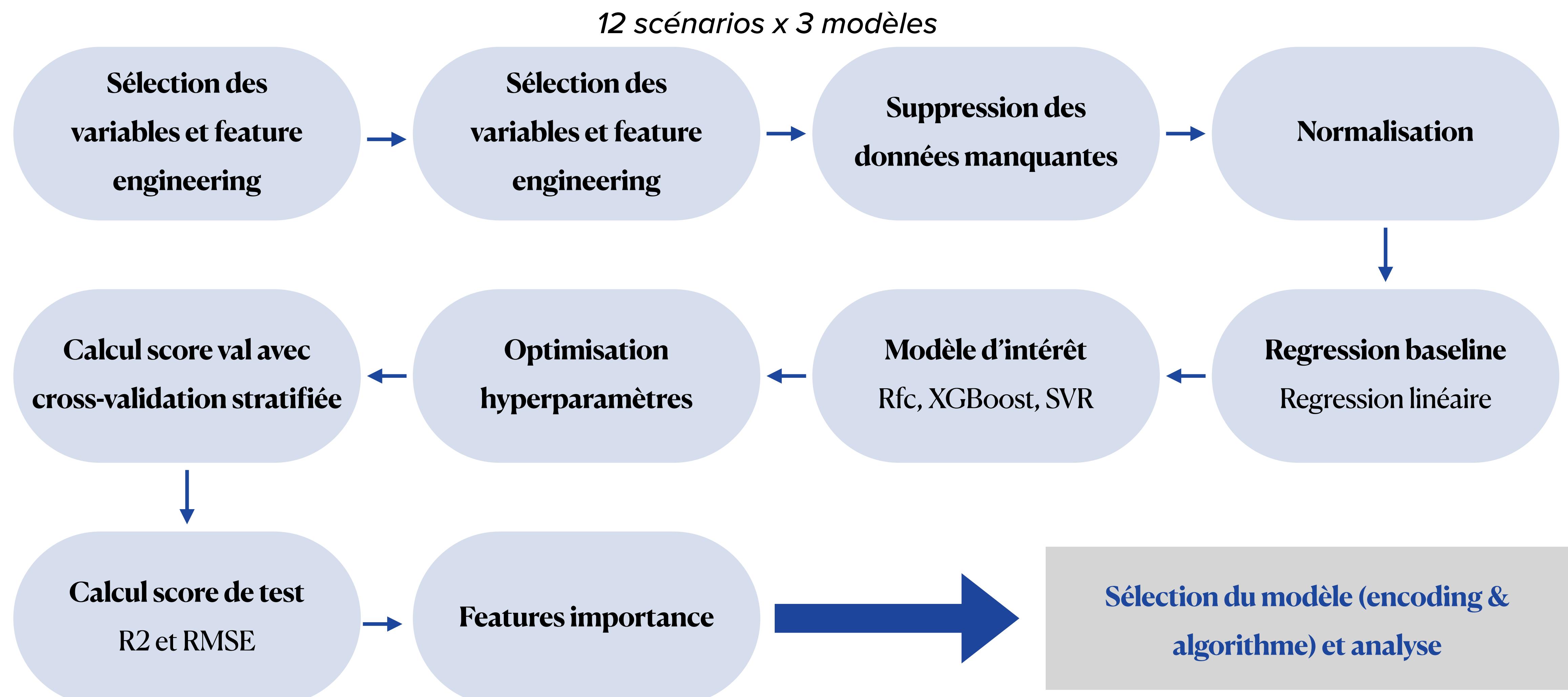
# Prédiction de la consommation énergétique totale

## Sommaire

- Rappel du contexte et des objectifs
- Analyse exploratoire & Feature engineering
- Prédiction de la consommation énergétique totale
  - Etapes de modélisation
  - Résultat de random forest
  - Résultat XGBoost
  - Analyse du modèle sélectionné
- Prédiction des émissions de gaz à effets de serre
- Conclusion

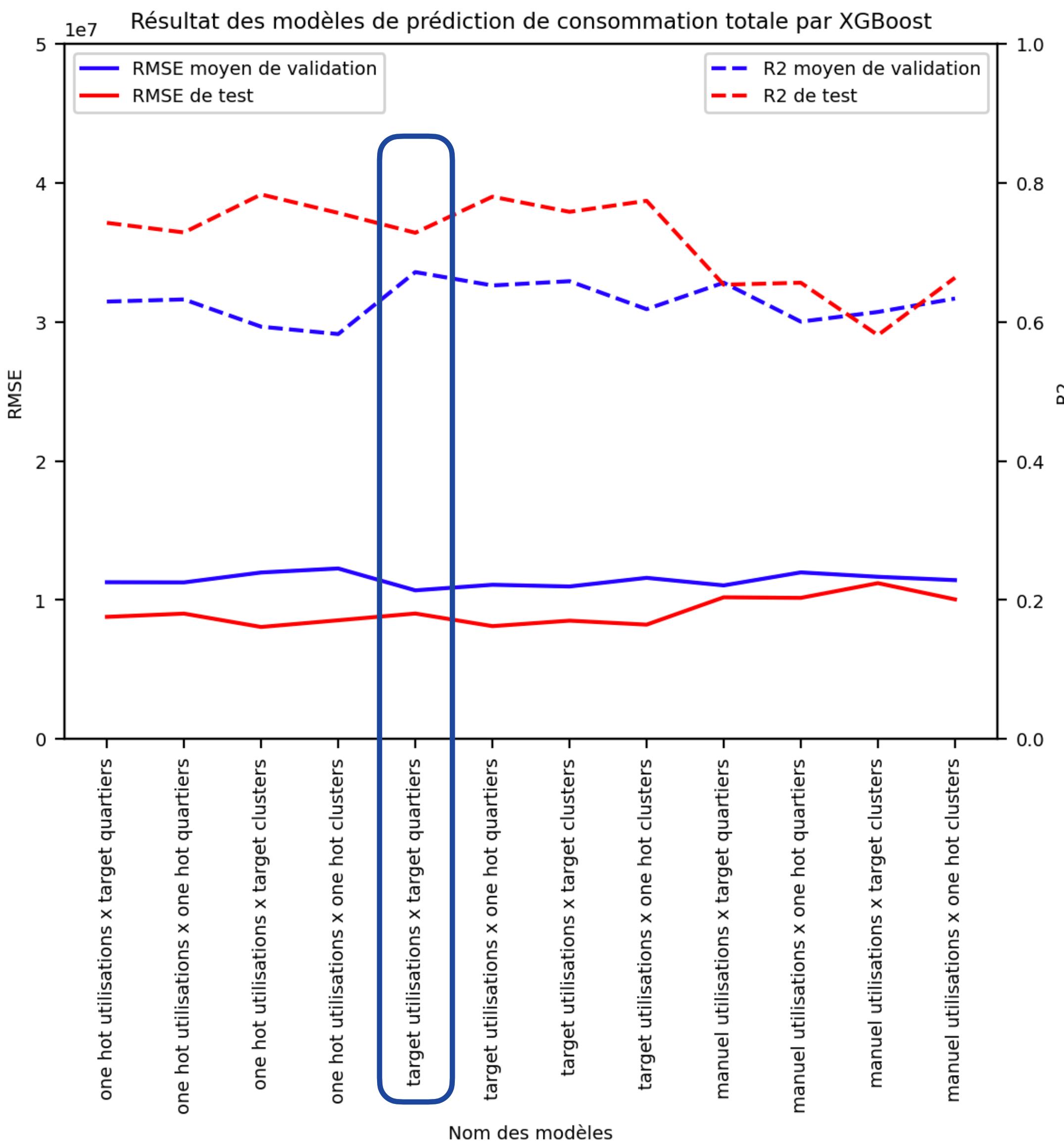
# Prédiction de la consommation énergétique totale

## Etapes de modélisation



# Prédiction de la consommation énergétique totale

## Résultat XGBoost



**Résultats :**

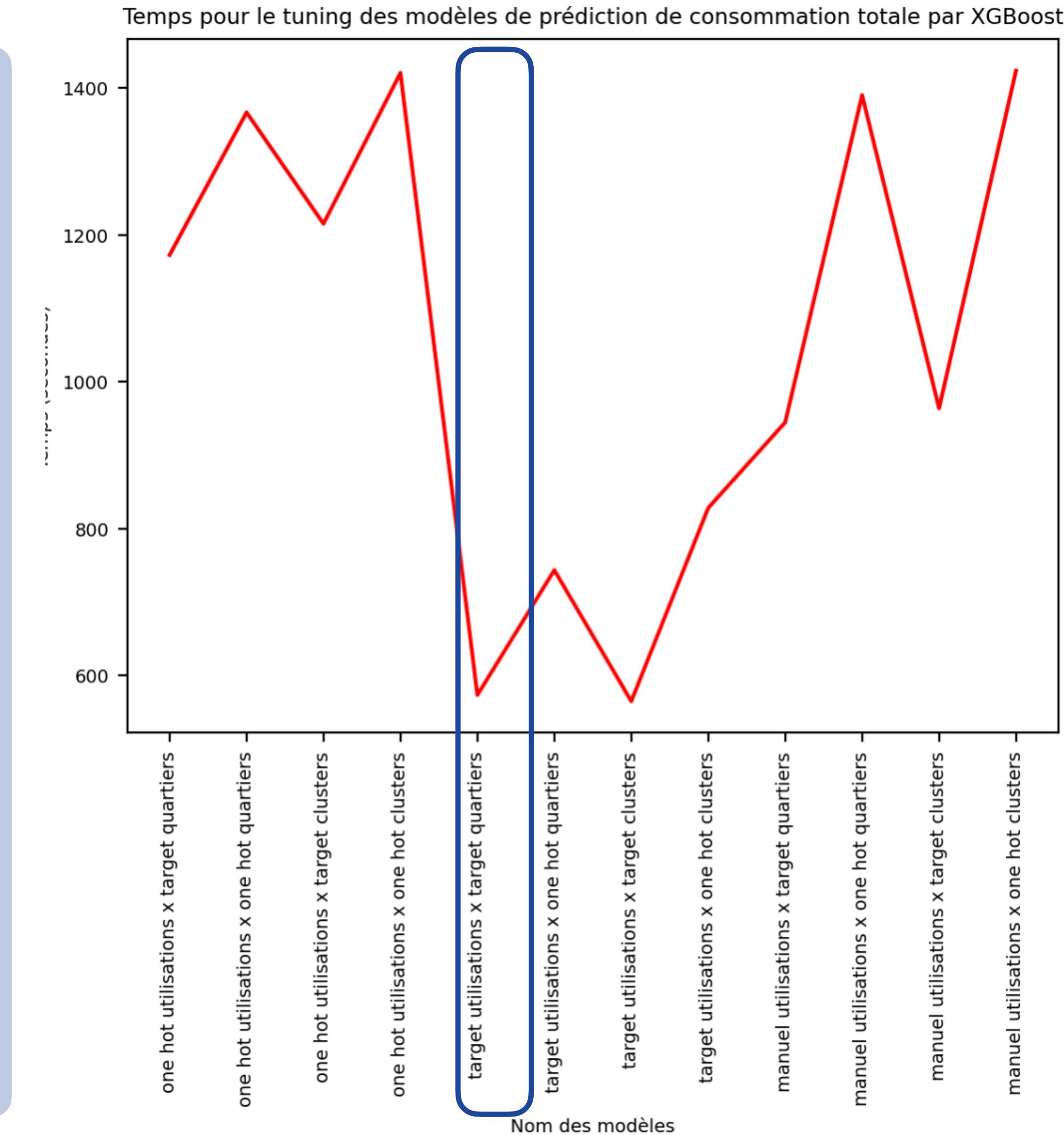
**Baseline - RL :**  
RMSE : 14 815 515 kBtu  
R2 : 0.28

**XGBoost :**

*Meilleurs hyperparamètres :*  
{'colsample\_bytree': 1.0, 'gamma': 0, 'learning\_rate': 0.01, 'max\_depth': 3, 'n\_estimators': 500, 'subsample': 0.8}

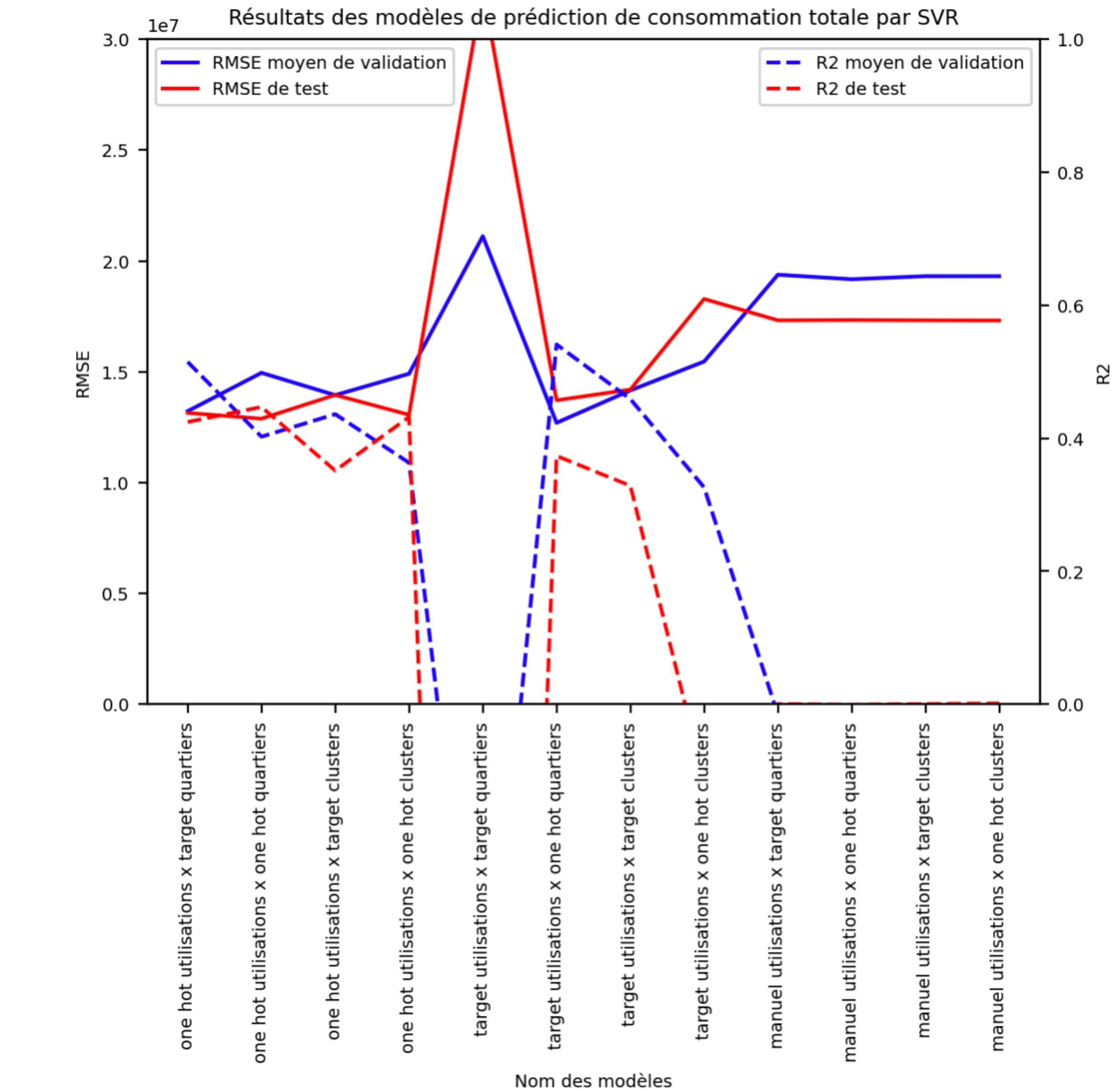
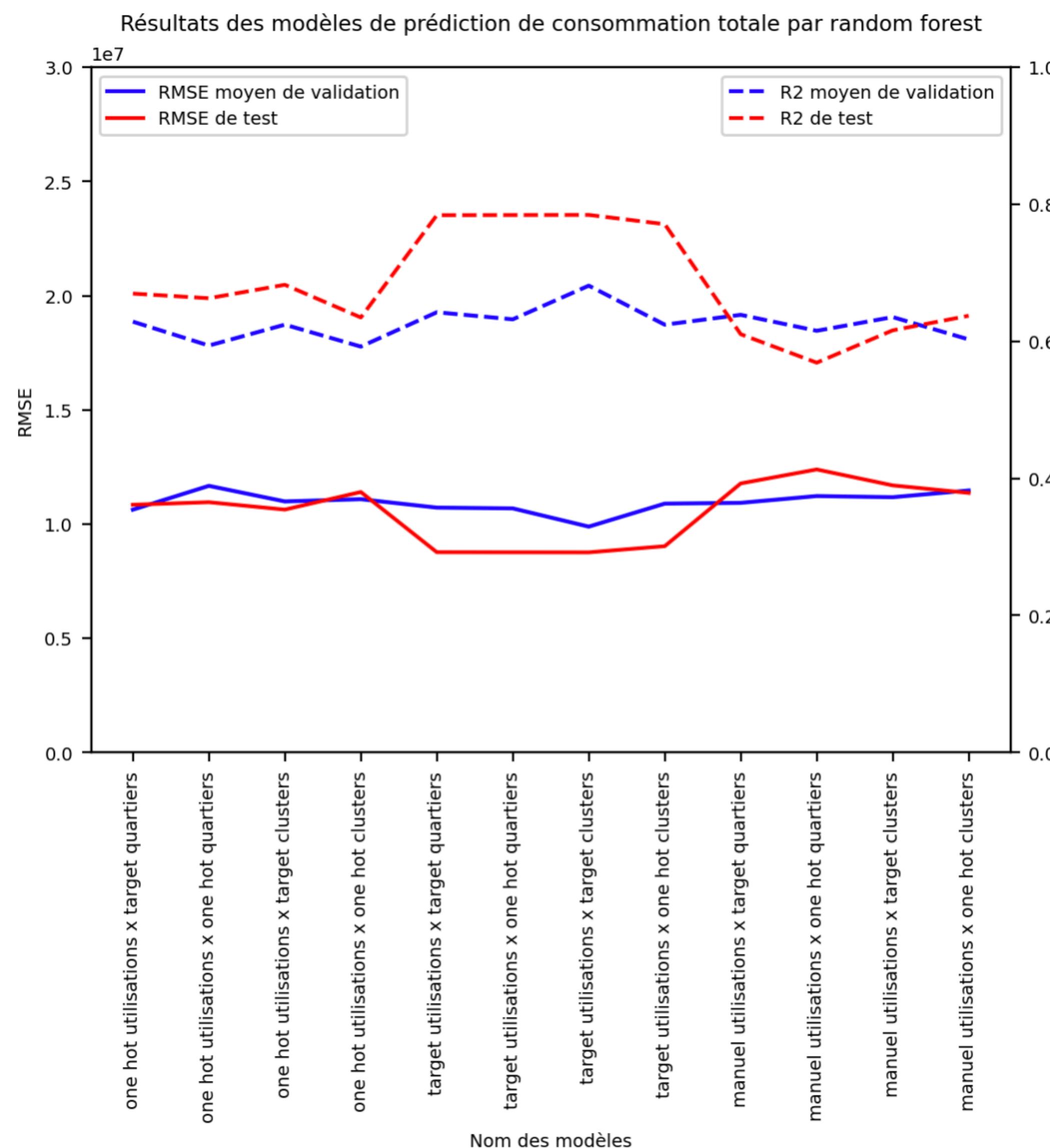
*Score de validation :*  
RMSE : 10 695 758 kBtu  
R2 : 0.67

*Score de test :*  
RMSE : 7 523 168 kBtu  
R2 : 0.81



# Prédiction de la consommation énergétique totale

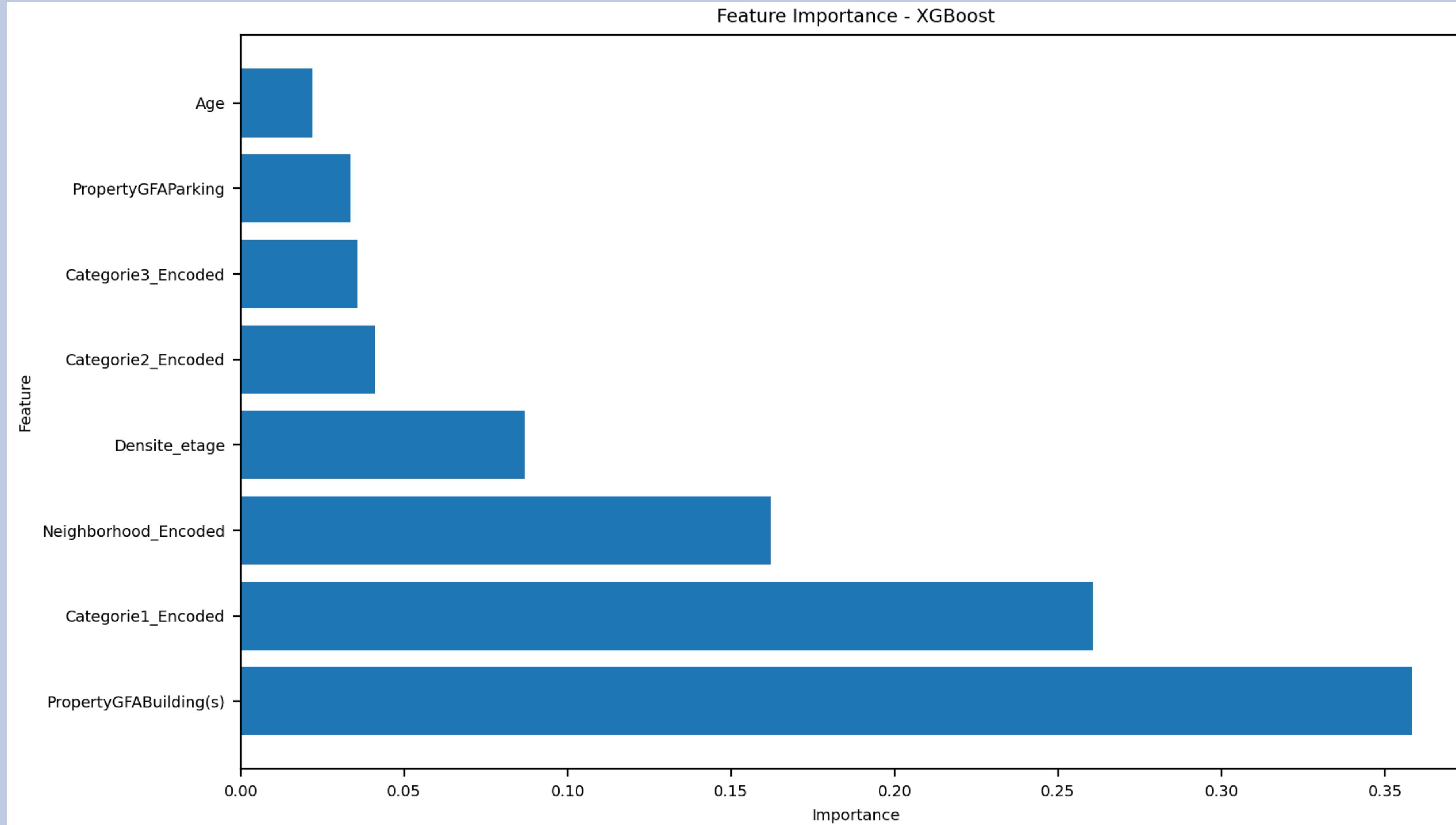
## Résultat random forest & SVR



# Prédiction de la consommation énergétique totale

## Analyse du modèle sélectionné

### XGBoost : Target encoding des utilisations x Target encoding des quartiers



# Prédiction des émissions de gaz à effets de serre

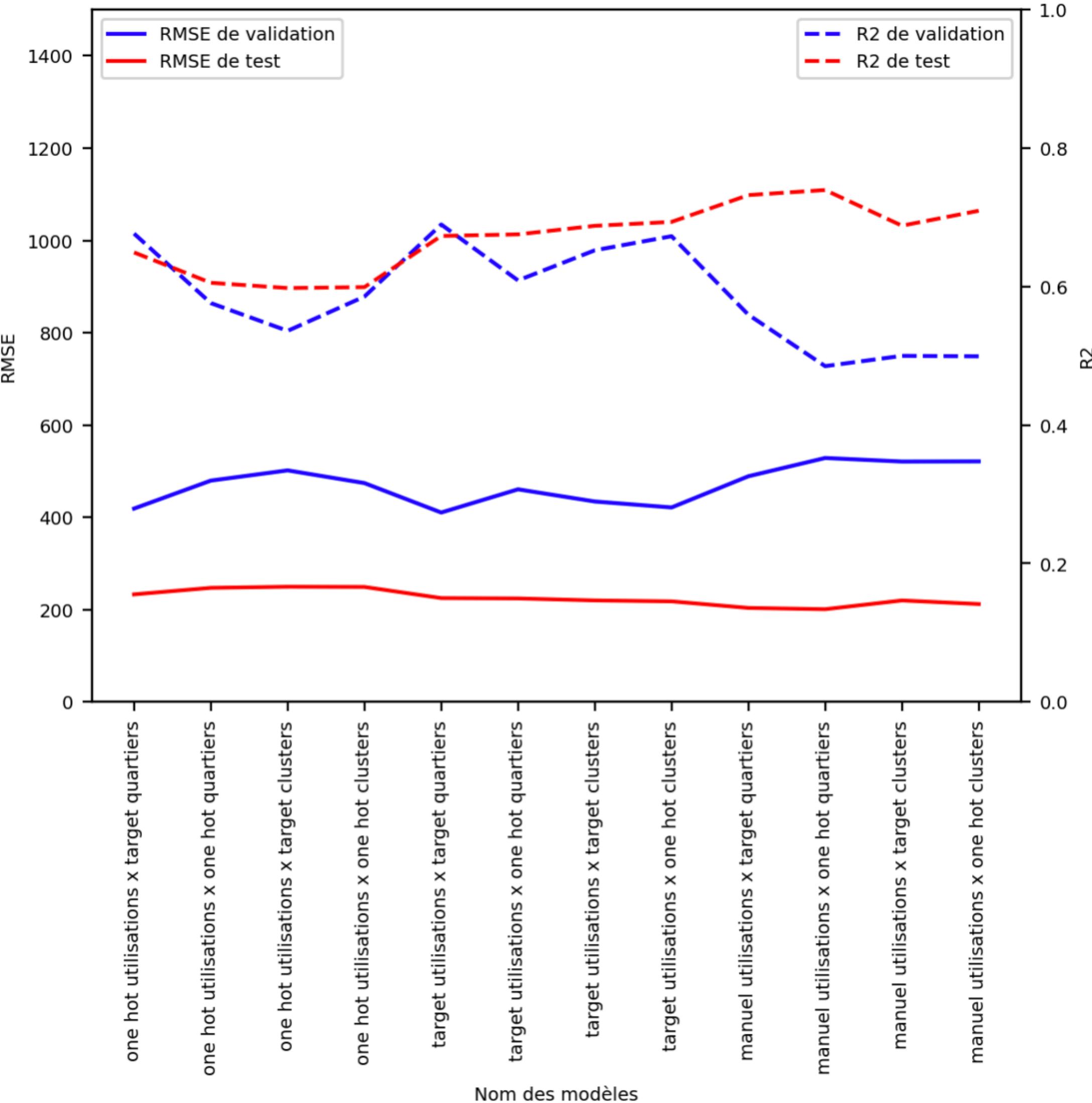
## Sommaire

- Rappel du contexte et des objectifs
- Analyse exploratoire
- Analyse exploratoire & Feature engineering
- **Prédiction des émissions de gaz à effets de serre**
  - Résultats
  - Impact de l'Energy Star Score
  - Analyse du modèle sélectionné
- Conclusion

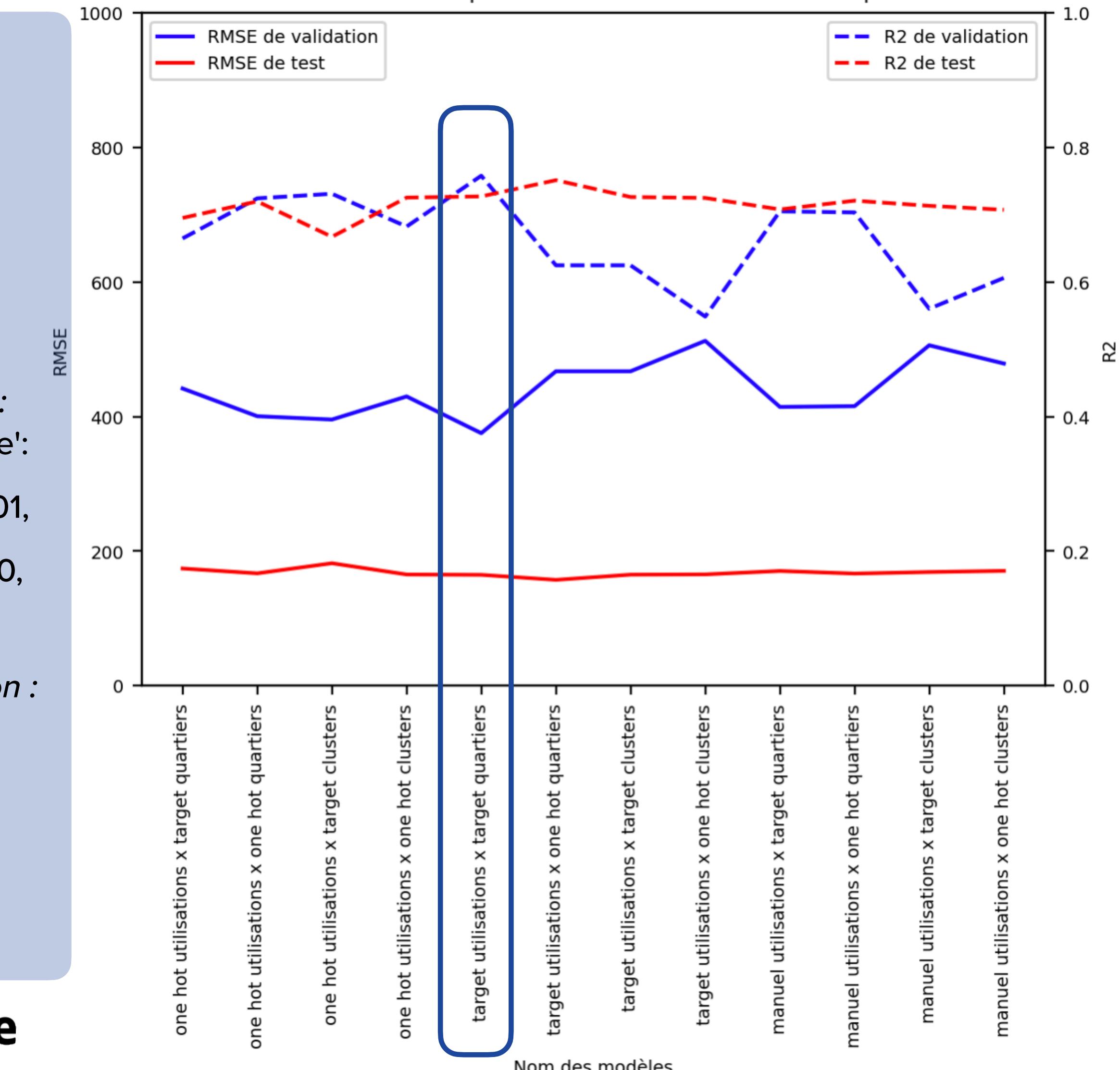
# Prédiction des émissions de gaz à effets de serre

## Résultats XGBoost & Random forest

Résultats des modèles de prédiction d'émissions totale de GES par random forest



Résultats des modèles de prédiction d'émissions totale de GES par XGBoost



### Résultats :

#### Baseline - RL :

RMSE : 408 kBtu

R2 : 0.38

#### XGBoost :

*Meilleurs hyperparamètres :*  
`{'colsample_bytree': 0.8, 'gamma': 5, 'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 200, 'subsample': 0.8}`

#### Score de validation :

RMSE : 387 kBtu

R2 : 0.80

#### Score de test :

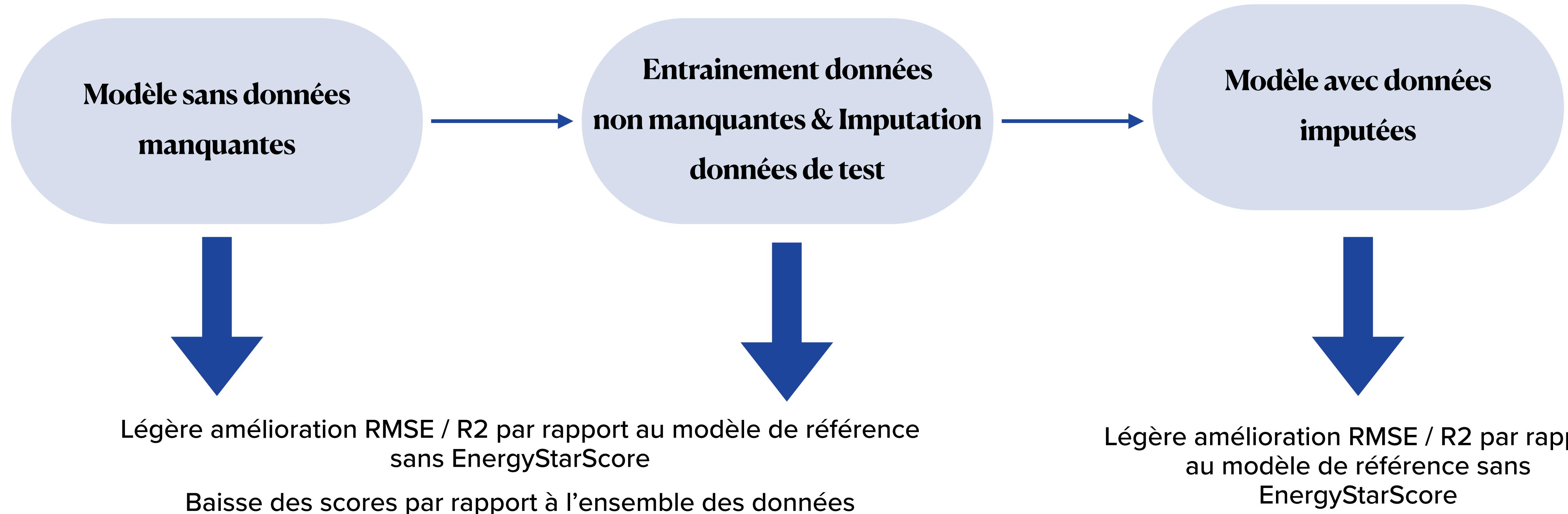
RMSE : 192 kBtu

R2 : 0.77



# Prédiction des émissions de gaz à effets de serre

## Impact de l'Energy Star Score

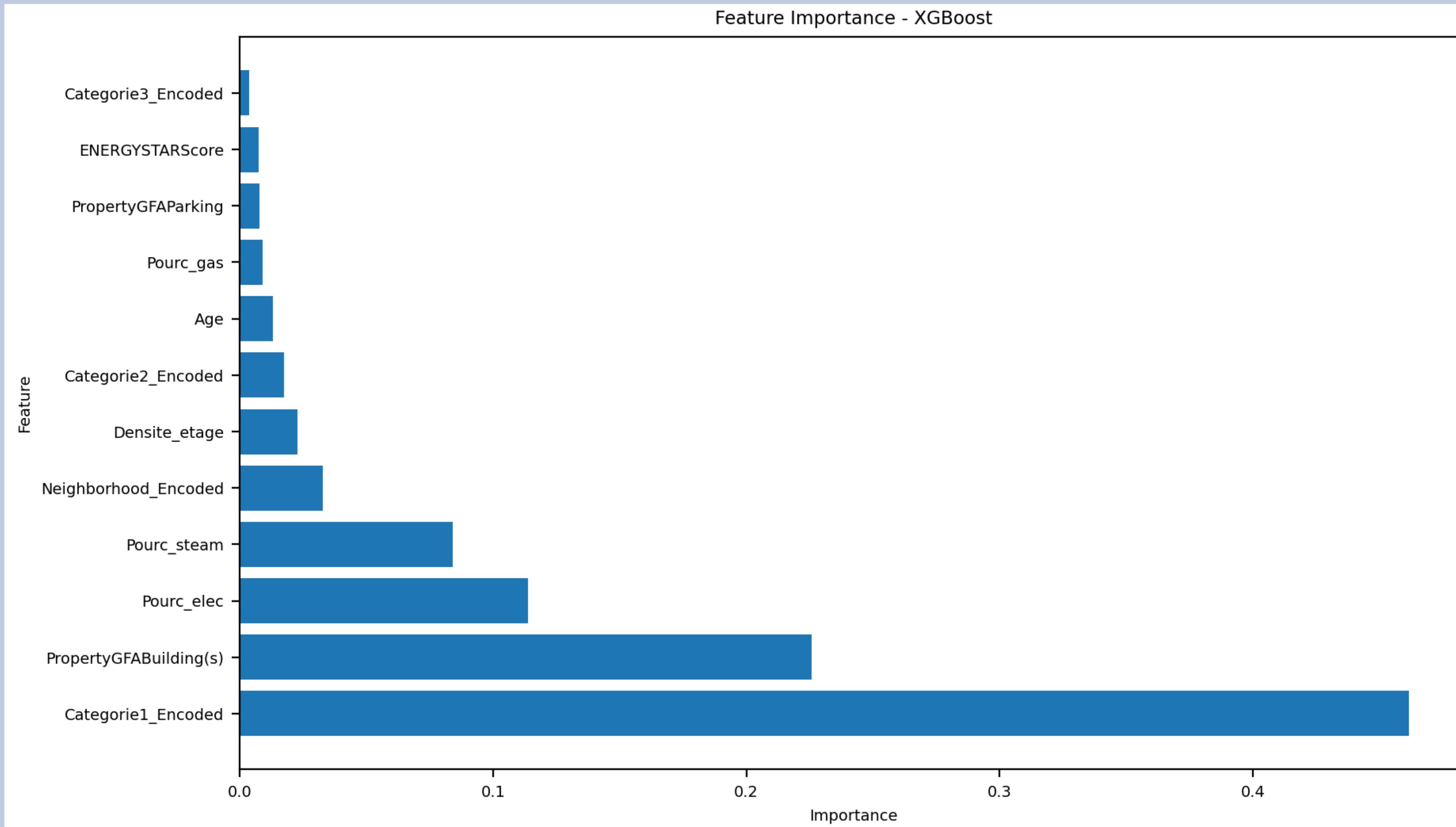


**Intégration de l'EnergyStarScore avec imputation des données manquantes par la moyenne**  
(très légère amélioration, feature avec peu d'importance)

# Prédiction des émissions de gaz à effets de serre

## Analyse du modèle sélectionné

### XGBoost : Target des utilisations x Target encoding des quartiers



# Conclusion

## Sommaire

---

- Rappel du contexte et des objectifs
- Analyse exploratoire & Feature engineering
- Prédiction de la consommation énergétique totale
- Prédiction des émissions de gaz à effets de serre
- Conclusion
  - Résultats & Limites

# Conclusion

## Résultats & Limites

### Résultats :

- Modèle de prédictions consommation énergétique totale et émissions totales GES
- EnergyStarScore : *légère amélioration des prédictions d'émissions totales*

### Limites des modèles :

- RMSE élevé : *mauvaise performance de prédiction individuelle du modèle*
- Variabilité des résultats & différences scores de validation et de tests

### Raisons identifiées :

- Variabilité des données & distribution des variables & bruits hétérogènes
- Phénomène difficile à modéliser

### Actions prises :

- Suppression des valeurs extrêmes, régularisation et sélection des caractéristiques : *non concluant et abandonné*
- Passage au log des variables asymétriques, regroupement des classes d'utilisation, stratification des jeux de données, cross-validation, tuning des HP : *concluant*

### Axes d'amélioration :

- Tester d'autres modèles (Elastic Net, réseau de neurone, etc.)
- Exploration plus approfondie des données et application des techniques appropriées corriger ou supprimer les données aberrantes
- Augmenter la taille des échantillons