PROJET 6

OPENCLASSROOMS: FORMATION

INGÉNIEUR DATA

Contexte:

Data Engineer pour la ville de Seattle.

But du projet:

Prédire les émissions de CO2 et la consommation totale d'énergie de bâtiments non destinés à l'habitation pour lesquels elles n'ont pas encore été mesurées.

PREMIER NETTOYAGE DES DONNÉES

3376 lignes, 46 colonnes

- Sélection bâtiments non résidentiels
- Suppression des lignes ayant des Outlier déclarés
- Suppression lignes n'ayant pas de données
- Remplacement des valeurs manquantes par des 0 pour les colonnes représentant des surfaces
- Suppression d'une colonne n'ayant que des valeurs manquantes

994 lignes, 45 colonnes

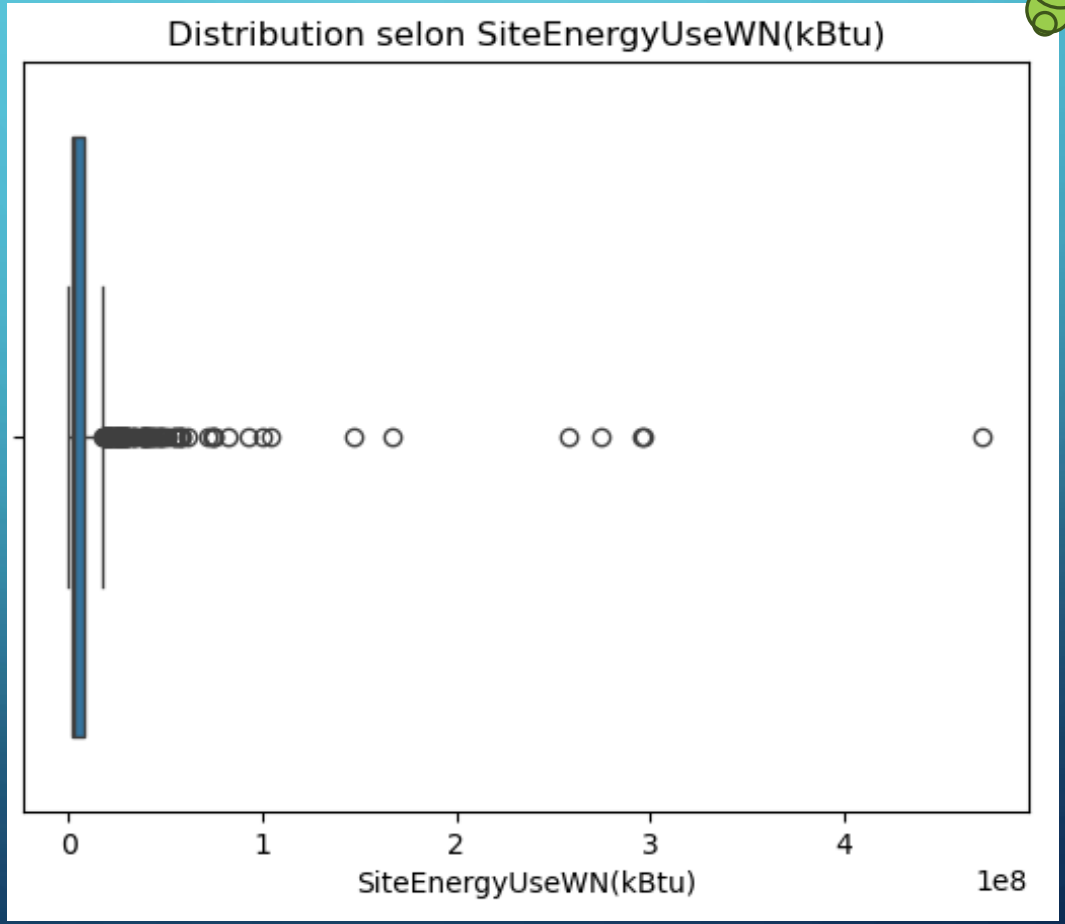
FEATURE ENGINEERING

- Création des colonnes booléennes: SteamUsed, ElectricityUsed, NaturalGasUsed
- Encodage des colonnes catégorielles avec un encodeur ordinal
- Suppression des colonnes inutiles (IDs, data leakage, valeurs identiques, valeurs non structurées comme l' adresse du bâtiment)

994 lignes, 24 colonnes numériques

AFFINAGE DU JEU DE DONNÉES

Trop dispersé! On peut sans doute faire mieux



AFFINAGE DU JEU DE DONNÉES

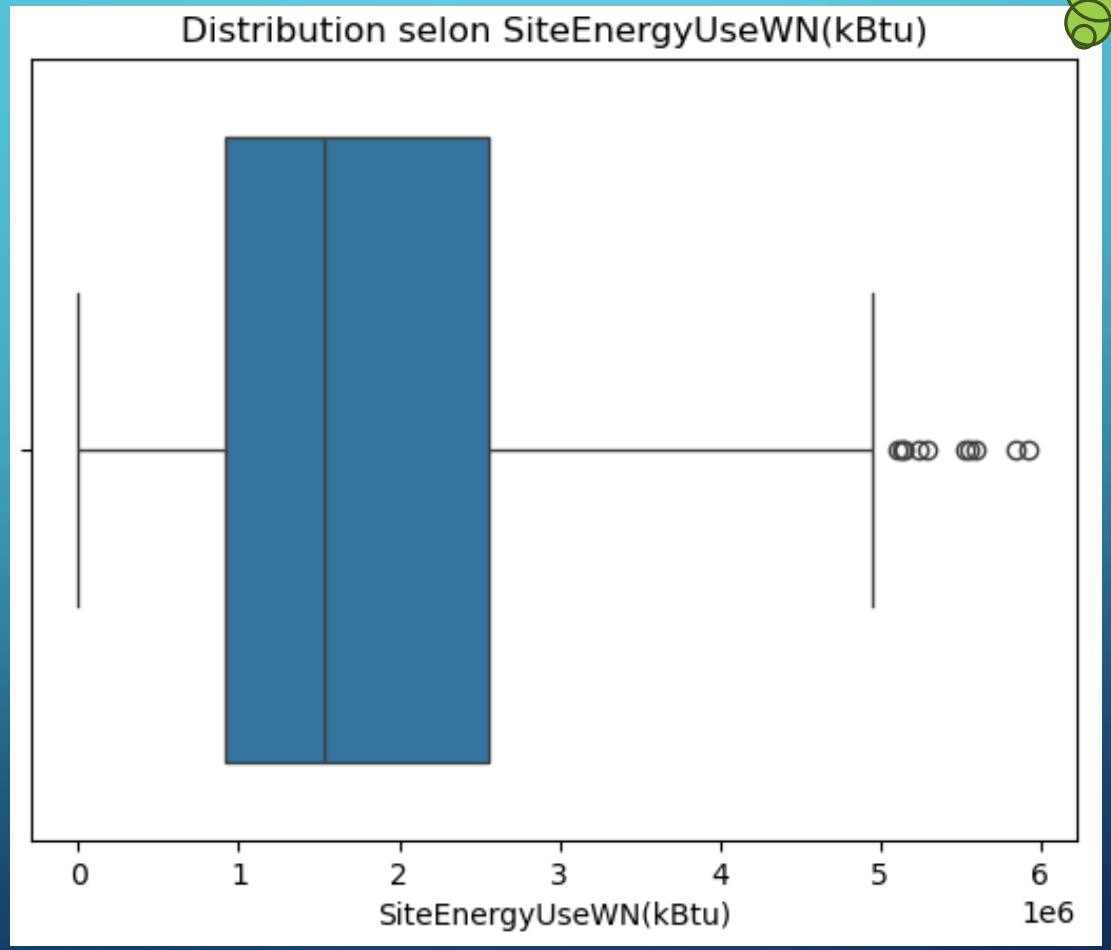
REDUCTION DES OUTLIERS AVEC LA MÉTHODE DE L'IQR

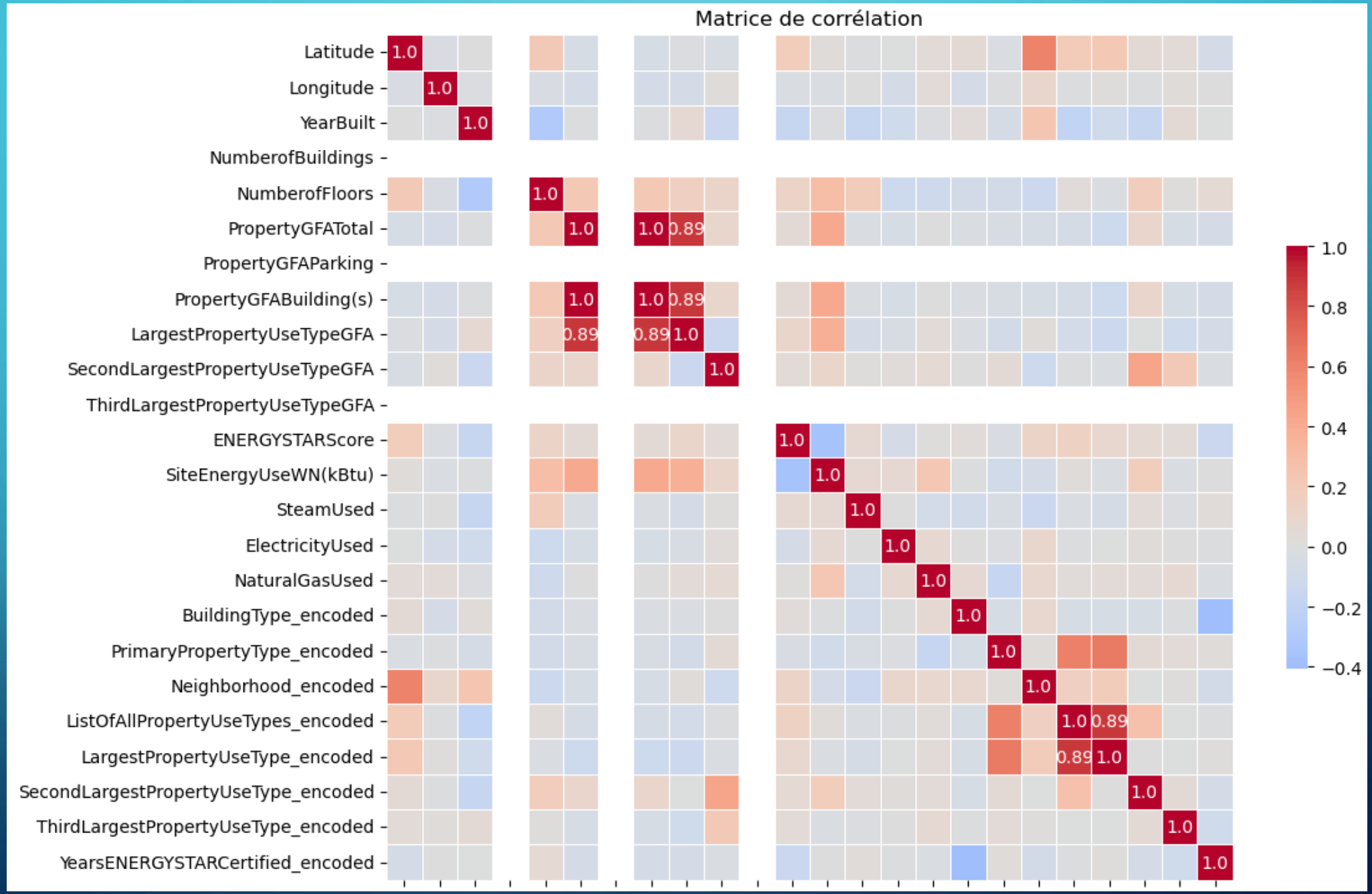
Q1 = 1er quartile (25 %)
Q3 = 3e quartile (75 %)
IQR = Q3 - Q1
Une valeur est considérée comme **outlier** si elle est :
< Q1 - 1.5×IQR
> Q3 + 1.5×IQR

376 lignes, 24 colonnes numériques

AFFINAGE DU JEU DE DONNÉES

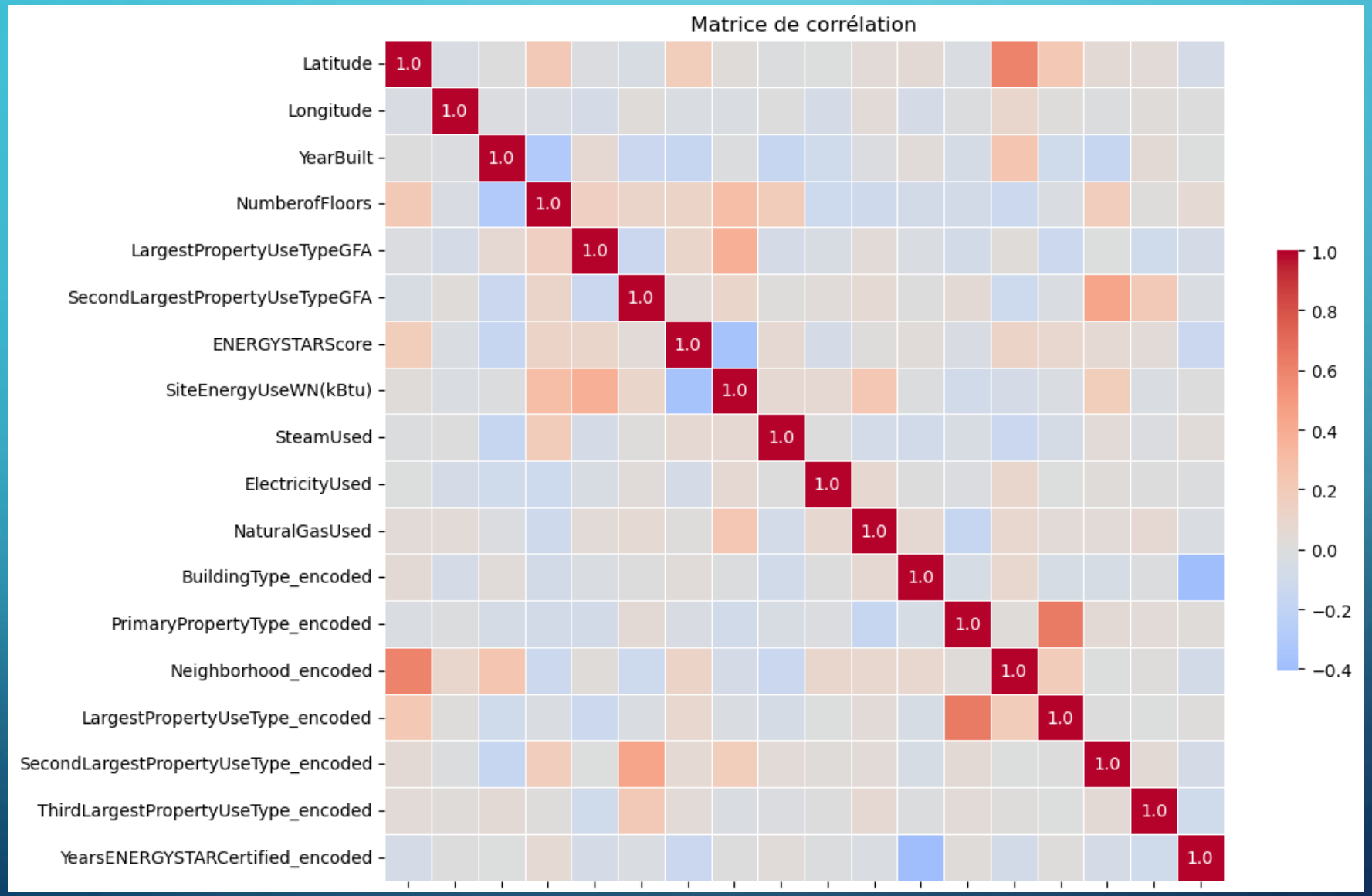
Mieux!




Matrice de corrélation
Latitude
Longitude
YearBuilt
NumberofBuildings
NumberofFloors
PropertyGFATotal
PropertyGFAParking
PropertyGFABuilding(s)
LargestPropertyUseTypeGFA
SecondLargestPropertyUseTypeGFA
ThirdLargestPropertyUseTypeGFA
ENERGYSTARScore
SiteEnergyUseWN(kBtu)
SteamUsed
ElectricityUsed
NaturalGasUsed
BuildingType_encoded
PrimaryPropertyType_encoded
Neighborhood_encoded
ListOfAllPropertyUseTypes_encoded
LargestPropertyUseType_encoded
SecondLargestPropertyUseType_encoded
ThirdLargestPropertyUseType_encoded
YearsENERGYSTARCertified_encoded
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

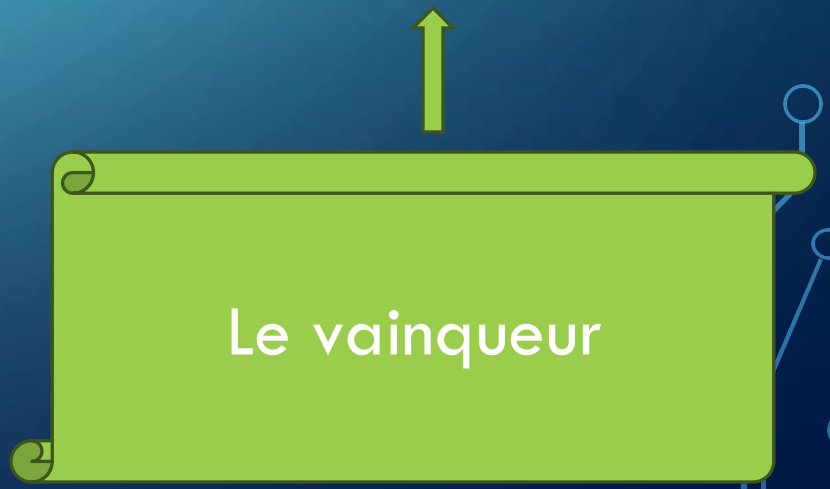
Suppression des colonnes trop corrélées (>= 0,7)
ou non pertinentes (mêmes valeurs ou manquantes)

376 lignes, 18 colonnes numériques

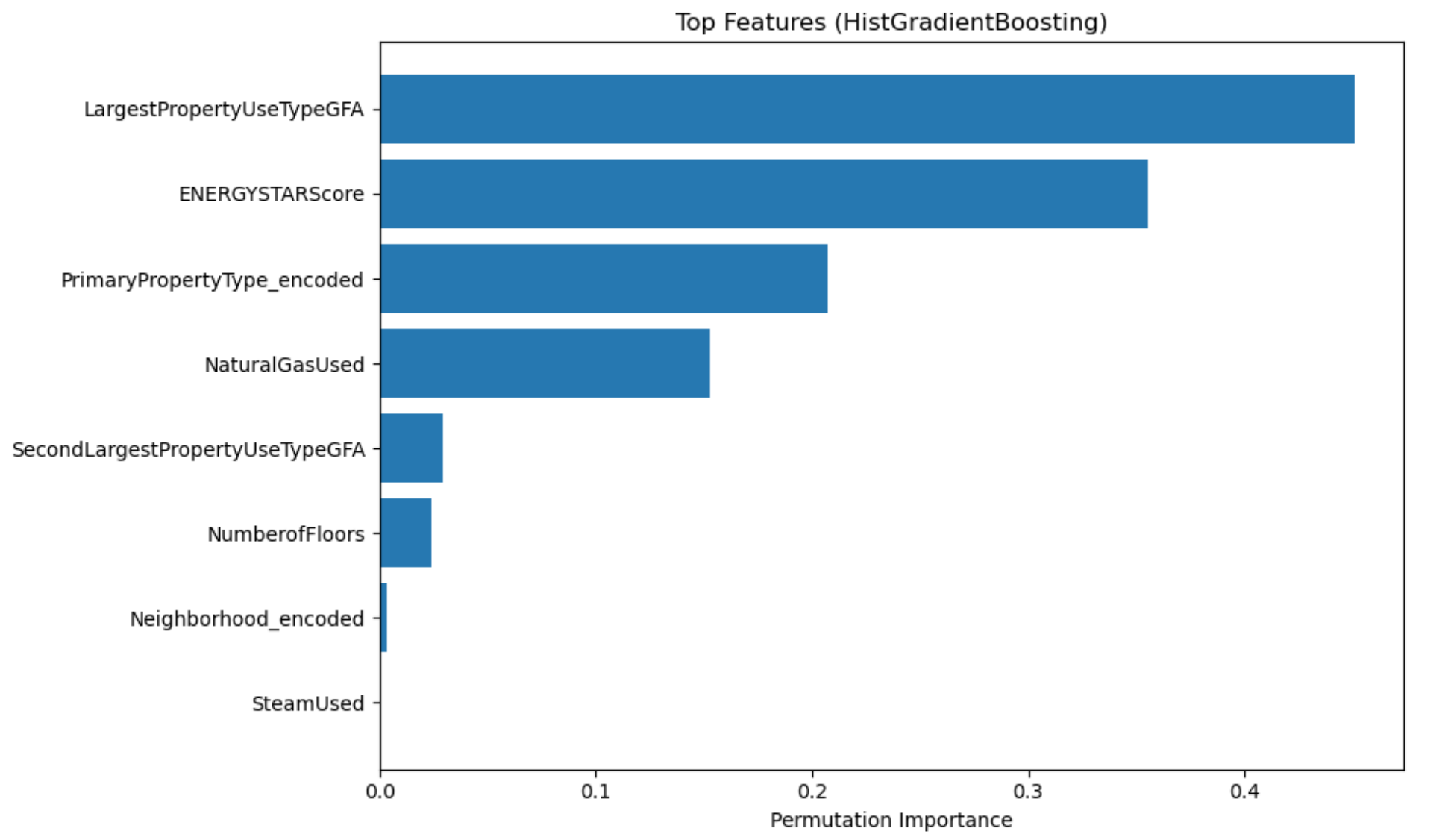

Matrice de corrélation
Latitude
Longitude
YearBuilt
NumberofFloors
LargestPropertyUseTypeGFA
SecondLargestPropertyUseTypeGFA
ENERGYSTARScore
SiteEnergyUseWN(kBtu)
SteamUsed
ElectricityUsed
NaturalGasUsed
BuildingType_encoded
PrimaryPropertyType_encoded
Neighborhood_encoded
LargestPropertyUseType_encoded
SecondLargestPropertyUseType_encoded
ThirdLargestPropertyUseType_encoded
YearsENERGYSTARCertified_encoded
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

CHOIX DU MODÈLE (TARGET = SiteEnergyUseWN(kBtu))

Scoring	Modèle linéaire	Modèle à base d'arbres	Modèle de type SVM	Modèle de type Gradient Boosting
R2	0.39	0.45	-1.62	0.51
MAE	672571.27	621774.10	956350.00	579614.11
RSME	900788.15	844660.04	1254318.00	798006.73



IMPORTANCE DES VARIABLES



AU FINAL

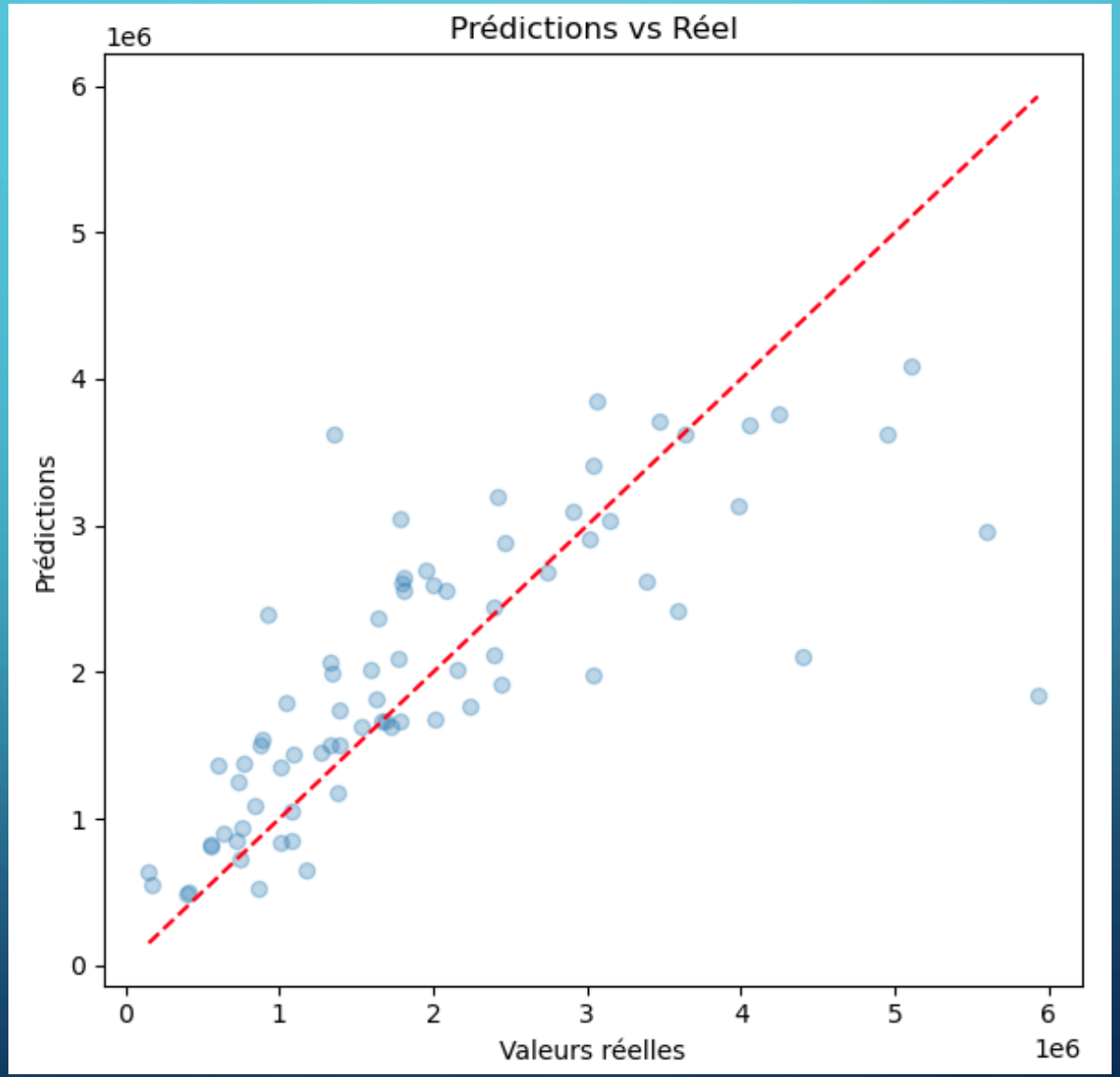
Jeu de 376 lignes, 8 colonnes numériques

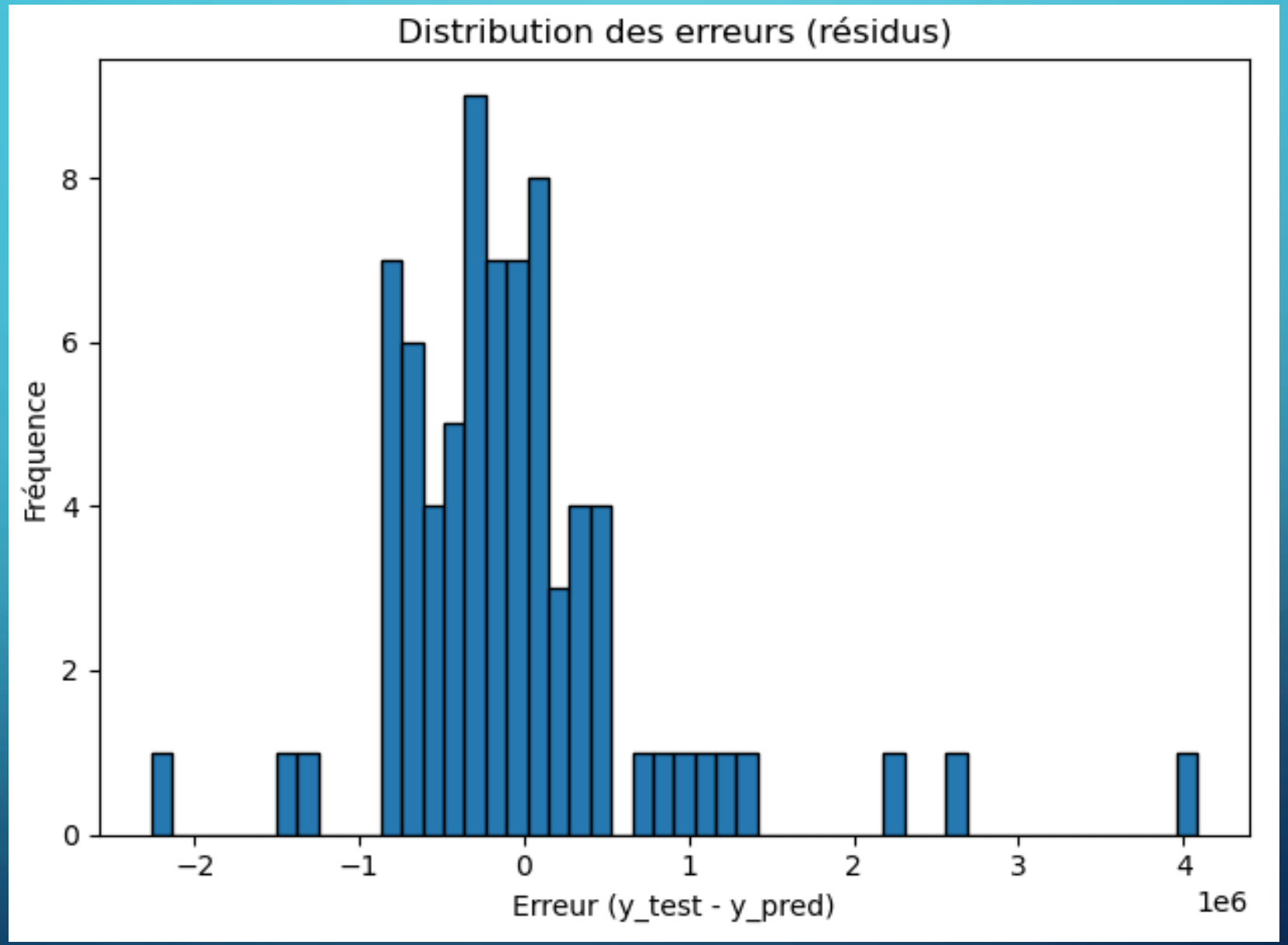
+

Modèle HistGradientBoosting avec scoring de :

- R^2 : 0.56
- MAE : 564029.82 kBtu

PERFORMANCE DU MODÈLE




Distribution des erreurs (résidus)
Fréquence
0
2
4
6
8
−2
−1
0
1
2
3
4
Erreur (y_test - y_pred)
1e6

DÉPLOIEMENT

- Dockerisation d'un service bentoml (API + modèle)
- Image poussée sur Google Cloud Platform(GCP)

DÉPLOIEMENT

Build api => bentoml build

Build image Docker => bentoml containerize --opt platform=linux/amd64 projet6-ml-service:<id_bentoml>

Test image Docker => docker run --rm -p **8080:8080** projet6-ml-servi**ce:**:<id_bentoml>

Tag image Docker => projet6-ml-servi**ce:**:<id_bentoml> gcr.io/nau-projet6/predict

Déploiement sur GCP =>

gcloud auth login
gcloud config set project nau-projet6
gcloud auth configure-docker
docker push gcr.io/nau-projet6/predict
gcloud run deploy predict --image gcr.io/nau-projet6/predict --platform managed --allow-unauthenticated

DÉPLOIEMENT

Données d'entrée de l'API:

- Nombre d'étages
- Surface d'usage principale du/des bâtiment(s)
- Surface d'usage secondaire du/des bâtiment(s)
- Score ENERGYSTAR
- Vapeur utilisée ou non
- Gaz utilisé ou non
- Type principal de propriété
- Quartier

Prédiction sur **https://predict-409714602166.us-central1.run.app/predict**