



Prévision de la consommation d'énergie et des émissions
de CO₂ des bâtiments non résidentiels

SOMMAIRE

- **Présentation** Présentation de la problématique;
Découverte du jeu de données.
- **Traitemet des données** Nettoyage des données;
Analyse exploratoire sur le jeu de données.
- **Feature ingénierie** Création des nouveaux features;
Preprocessing.
- **Modélisation** Mise en place de plusieurs modèles;
Sélection du meilleur modèle;
Evaluation des modèles avec
Energy Star Score.
- **Conclusion**

Présentation



Présentation de la problématique

- Objectif : créer une ville neutre en carbone d'ici 2050.
- En 2016, une collecte manuelle approfondie des données a été réalisée. Cette méthode est coûteuse et prend du temps.
- Il existe encore de nombreux bâtiments non couverts par la recherche.
- Tâche : prédire les émissions de CO₂ et la consommation totale d'énergie.
- Testez différents modèles de prédiction pour répondre au mieux au problème.
- Évaluer les valeurs ENERGY STAR pour la prévision des émissions en les intégrant au processus de modélisation.
- Pour cela, nous avons à notre disposition la base de données [2016_Building_Energy_Benchmarking.csv](#), qui se trouve sur le site officiel de la ville de Seattle :
<https://data.seattle.gov/dataset/2016-Building-Energy-Benchmarking/2bpz-gwpy>

Le fichier de données "[2016_Building_Energy_Benchmarking.csv](#)" est une base de données sur la consommation énergétique des bâtiments de la ville de Seattle, aux États-Unis, pour l'année 2016.

Rows:3376 lignes

Columns:46 colonnes

Information géographique

- Adresse,
- Coordonnées GPS,
- Code postale, etc.

Décrivants caractéristiques

- Le nombre d'étages et des bâtiments,
- Le type d'utilisation,
- La surface brute de plancher etc.

Variables énergétiques

- La consommation électrique,
- La consommation de gaz,
- L'émissions de CO₂,
- Le ENERGYSTAR score etc.

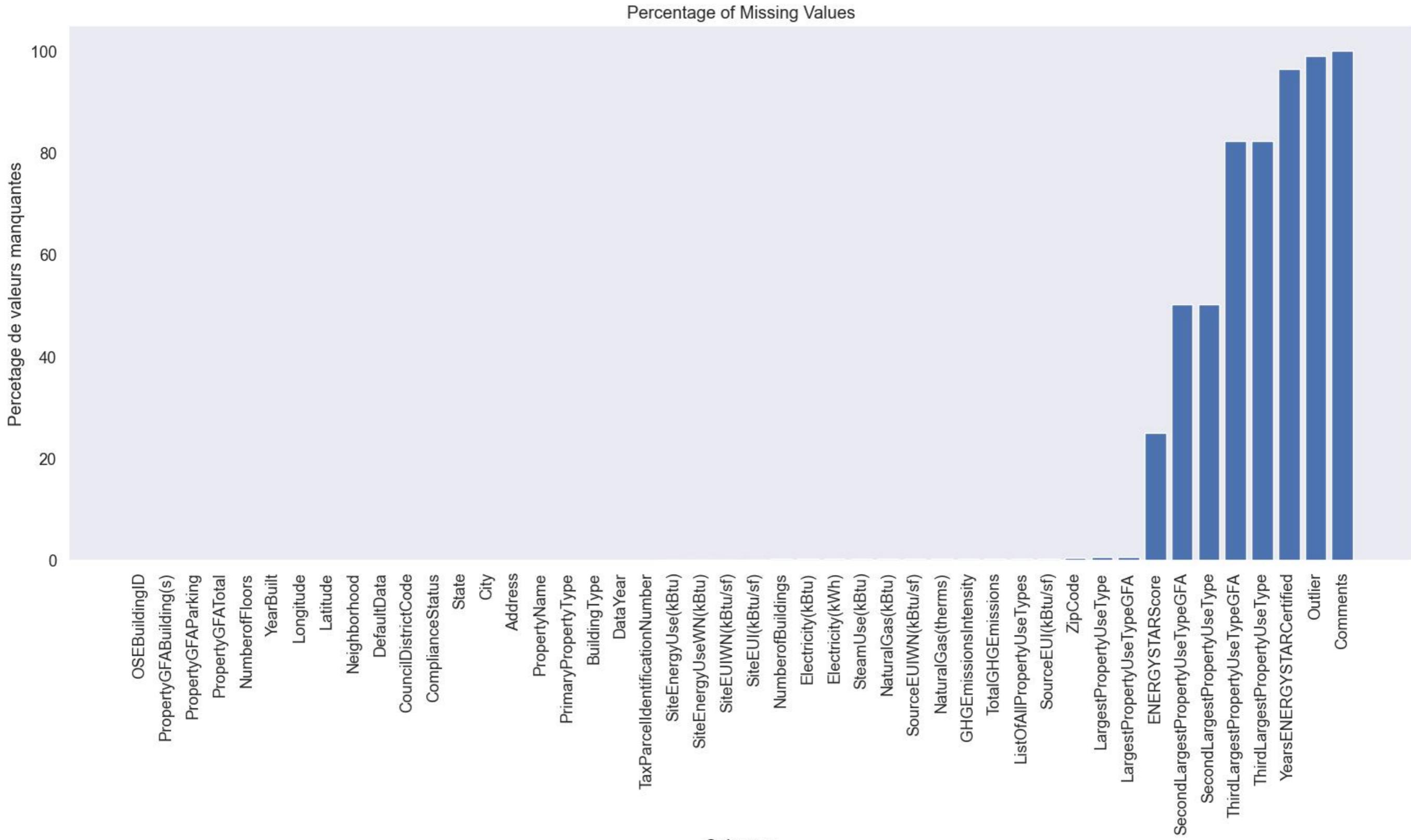


Nettoyage des données



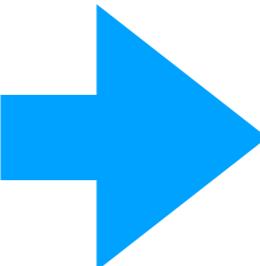
Étape 1 : nettoyage

Nous supprimerons toutes les lignes et colonnes contenant plus de 50 % de données manquantes, ainsi que les variables qui ne sont pas importantes pour une enquête plus approfondie.



● Elimination données des bâtiments résidentiels

multifamily



BuildingType	
NonResidential	1487
Multifamily LR (1-4)	1036
Multifamily MR (5-9)	583
Multifamily HR (10+)	110
SPS-District K-12	93
Nonresidential COS	82
Campus	25
Nonresidential WA	1

● Elimination des outliers

Les valeurs négatives

Les valeurs non conformes

ComplianceStatus

Compliant

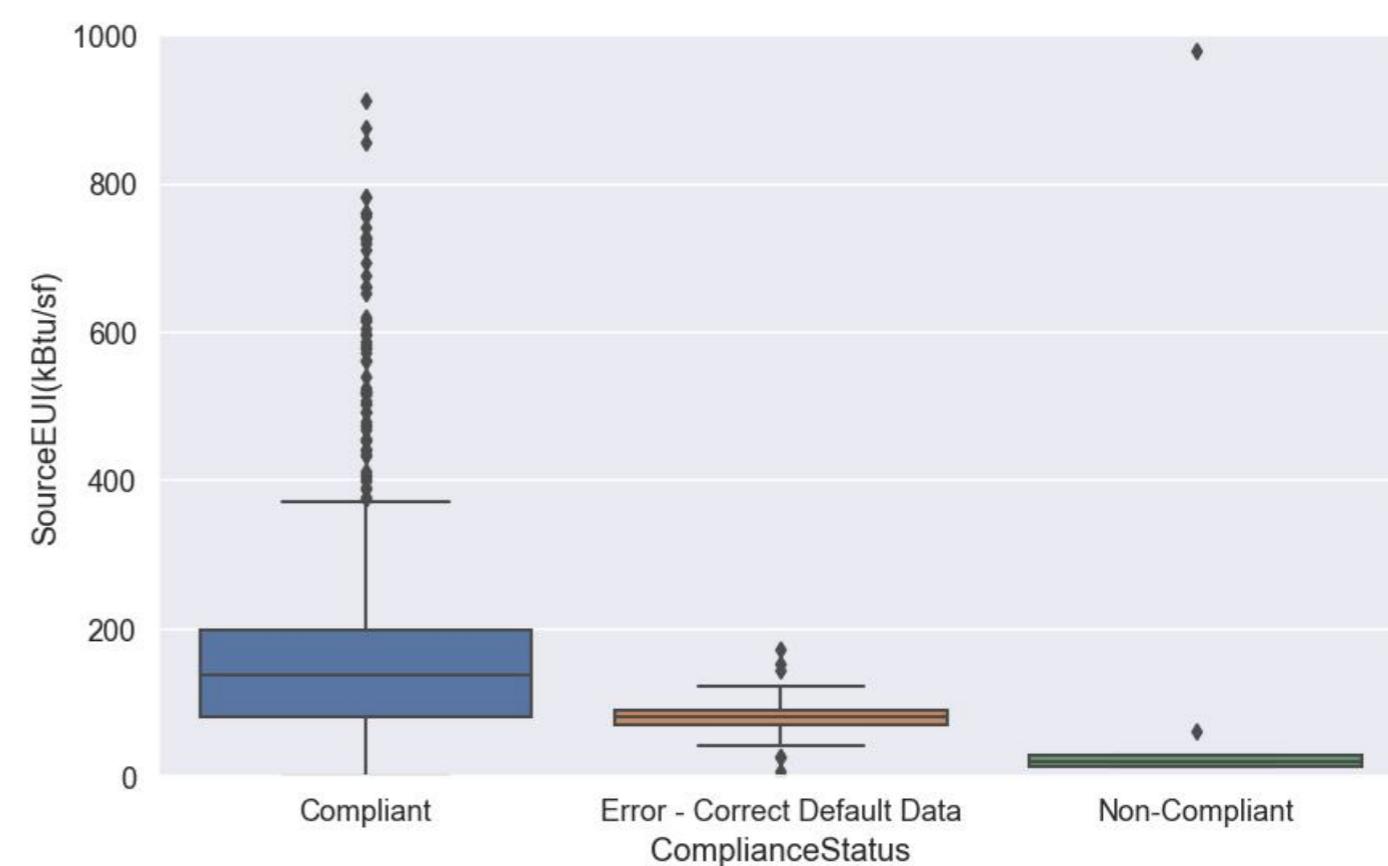
Error - Correct Default Data

Non-Compliant

989

85

9

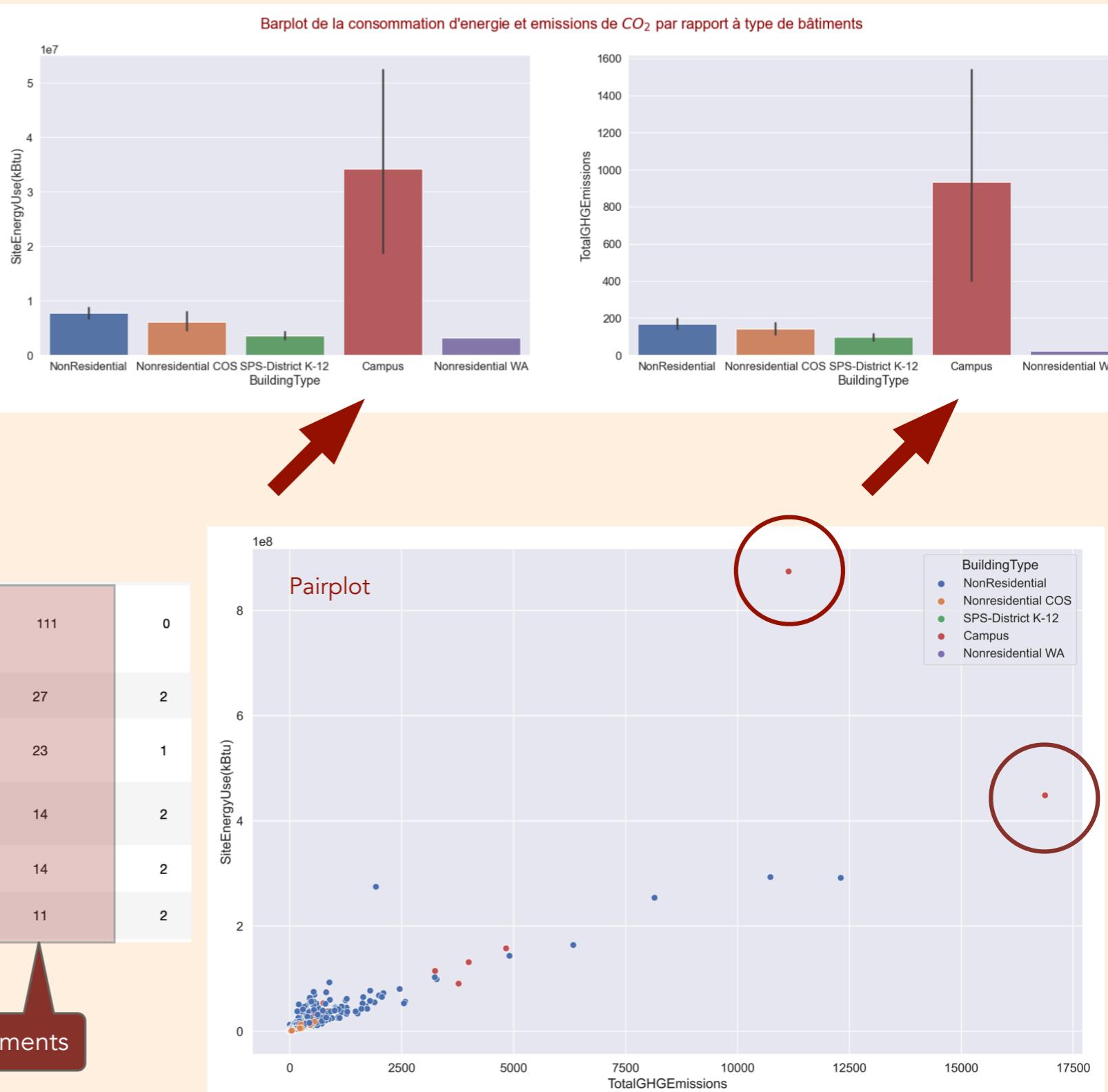


Optimisation

- Elimination des **outliers** (le problème de nombre de bâtiments)
- La méthode ANOVA

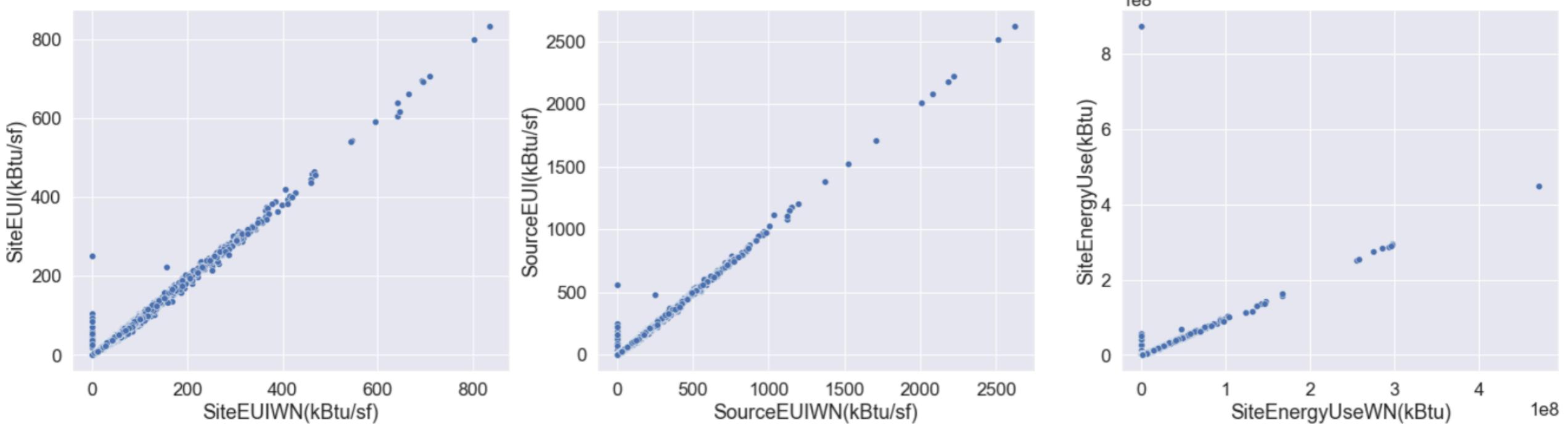
49967	Campus	University	University of Washington - Seattle Campus	4	Northeast	1900	111	0
172	Campus	University	SSCC MAIN CAMPUS	a	Delridge		27	2
23622	Campus	Other	FT C15 Fishermen's Center	g	Magnolia / queen anne		23	1
261	Campus	Large Office	South Park	a	Greater duwamish		14	2
25251	Campus	University	5th Avenue Master Meter	g	Magnolia / queen anne		14	2
211	Campus	University	NSCC MAIN CAMPUS	e	Northwest		11	2

Nombre de bâtiments

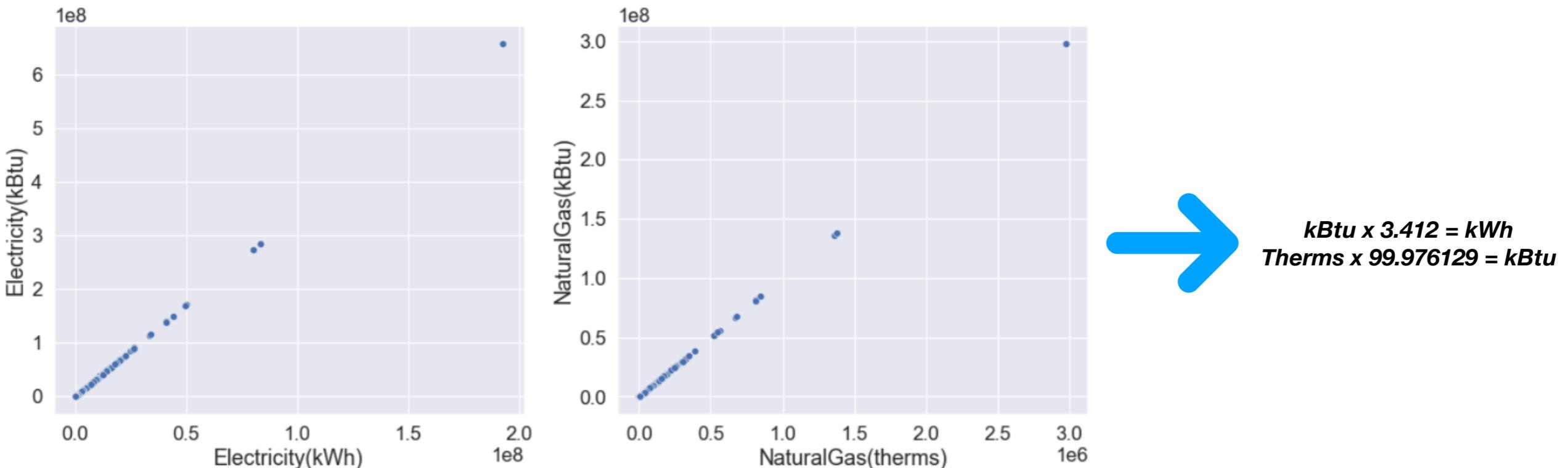


Étape 2: suppression des variables redondantes

Elimination des variables avec le suffixe “WN” (weather normalised)



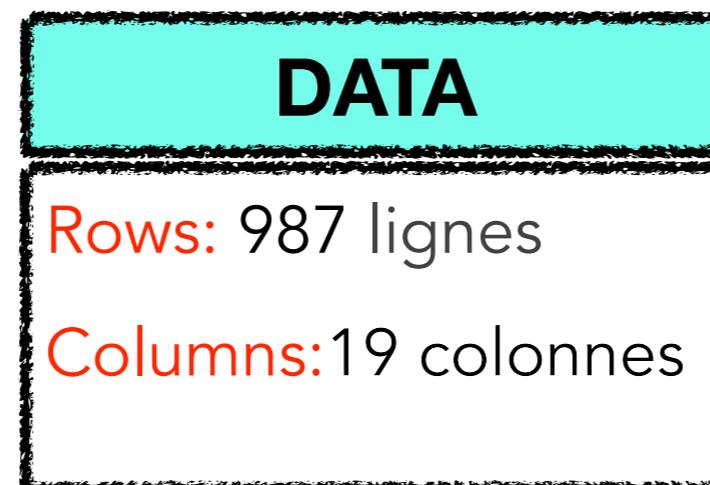
Suppression des variables qui sont fournies en plusieurs unités



Étape 3 : Exclusion de variables des relevés

- Regroupement des variables
- L'étude ne se base pas sur les relevés (électricité, gaz, etc.)

10	ThirdLargestPropertyUseType
17	ThirdLargestPropertyUseTypeGFA
✓18	ENERGYSTARScore
✓19	SiteEnergyUse(kBtu)
✓20	SteamUse(kBtu)
✓21	Electricity(kBtu)
✓22	NaturalGas(kBtu)
23	TotalGHGEmissions
24	ZipCode
25	BuildingAge

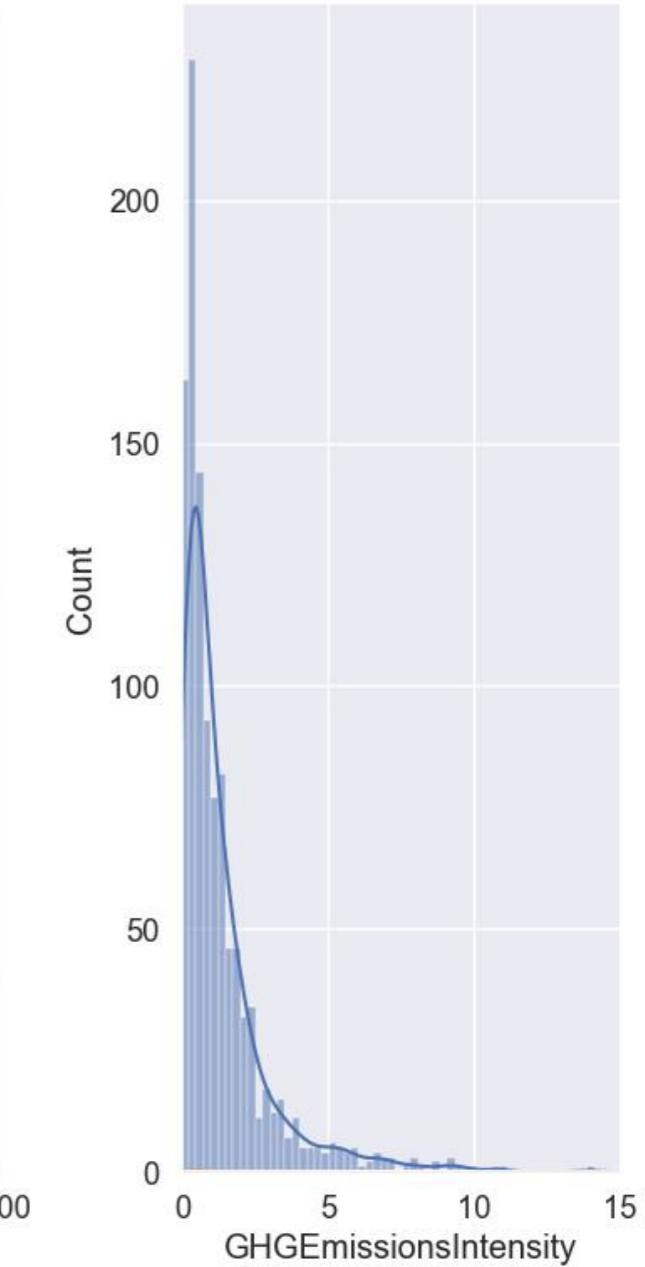
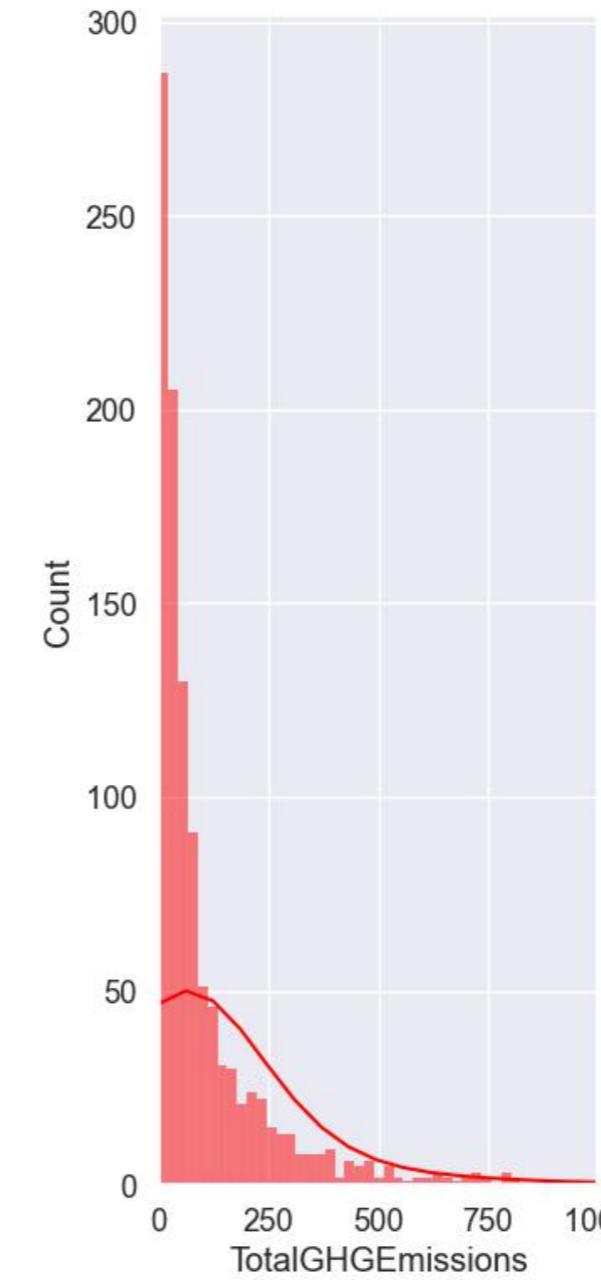
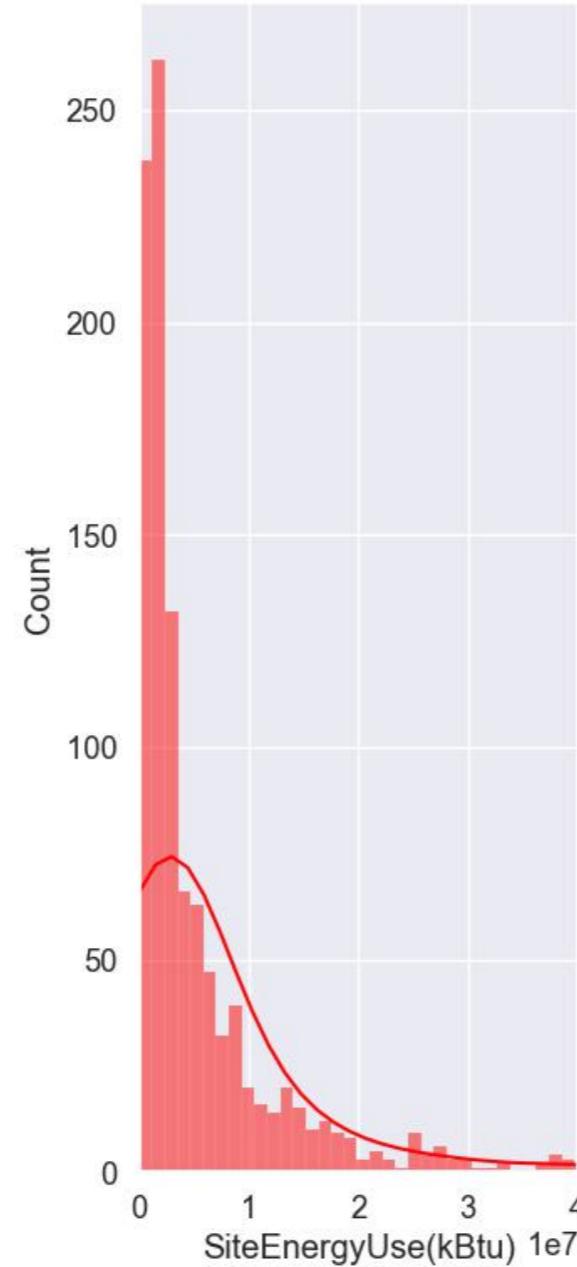
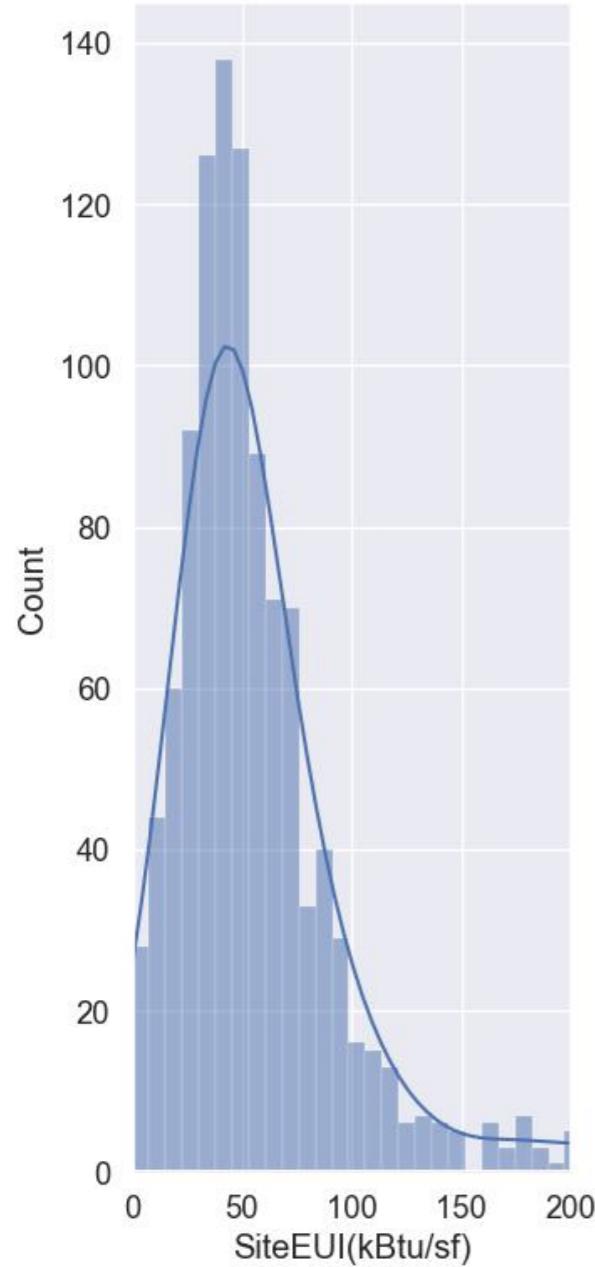


Analyse exploratoire

Analyse exploratoire

Analyse de la distribution des cibles

Les distributions des deux variables ne sont pas Gaussienne



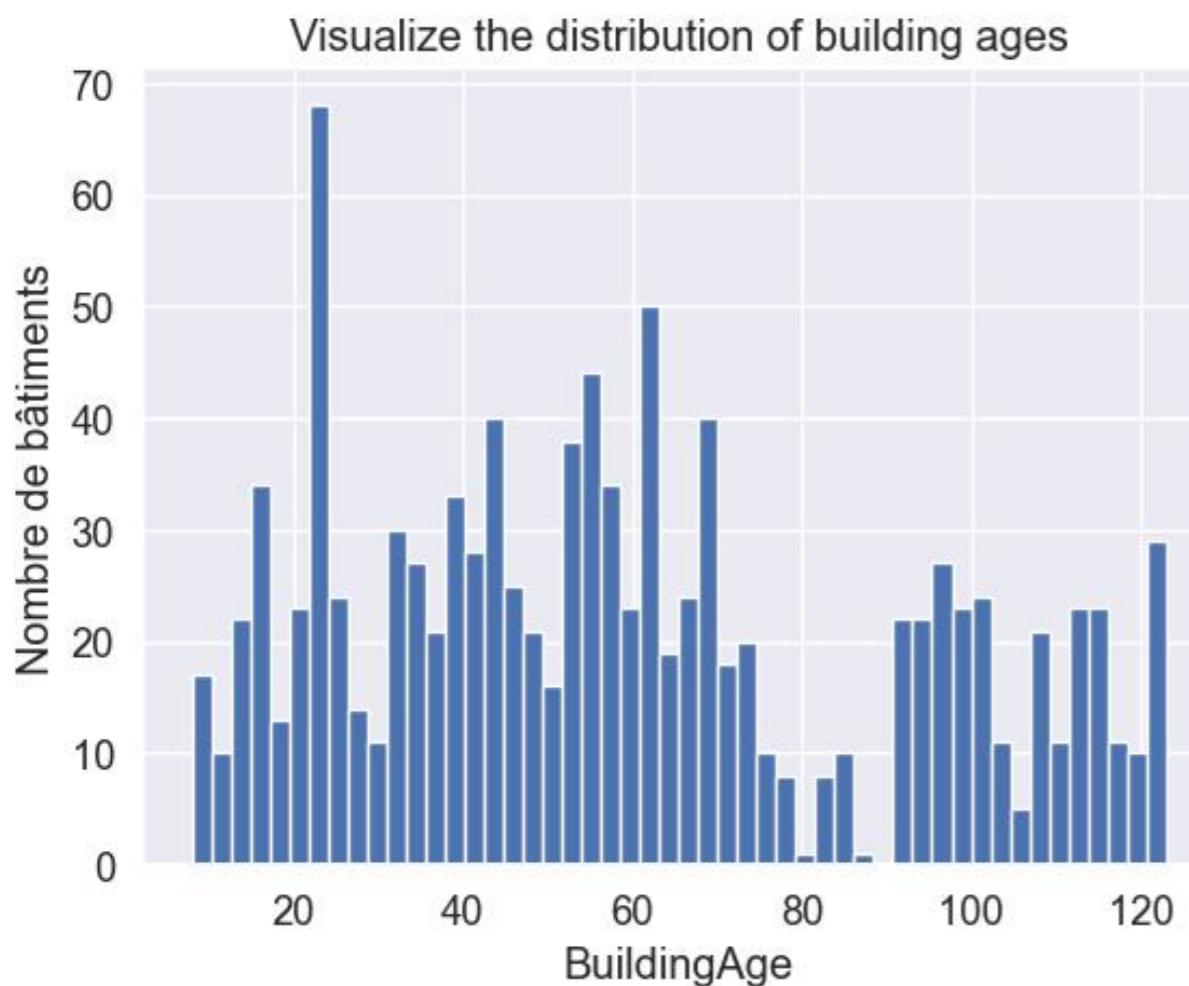
```
TotalGHGEmissions (normaltest):  
Statistics=3041.926, p=0.000  
Sample does not look Gaussian (reject H0)  
=====  
TotalGHGEmissions (Shapiro):  
Statistics=0.229, p=0.000  
Sample does not look Gaussian (reject H0)  
=====  
SiteEnergyUse(kBtu) (normaltest):  
Statistics=2495.587, p=0.000  
Sample does not look Gaussian (reject H0)  
=====  
SiteEnergyUse(kBtu) (Shapiro):  
Statistics=0.344, p=0.000  
Sample does not look Gaussian (reject H0)  
=====
```

Feature engineering

- Transformation des données brutes en la forme la plus appropriée pour les algorithmes de machine learning
- 3 catégories de nouveaux features

1) Nous utilisons la bibliothèque GeoPy pour trouver **la distance** entre le point dans le DataFrame et le centre de Seattle: `df['DistanceToSeattle'] = df.apply(lambda row: geodesic((row['Latitude'], row['Longitude']), seattle_center).miles, axis=1)`

2) Calculons l'âge des bâtiments



#calculer les Score

`df[,GasScore']`
`df['ElectricityScore']`
`df['SteamUse(kBtu)Score']`
`df['ParkingScore']`
`df[,BuildingScore']`

#surface moyenne par bâtiment et par étage

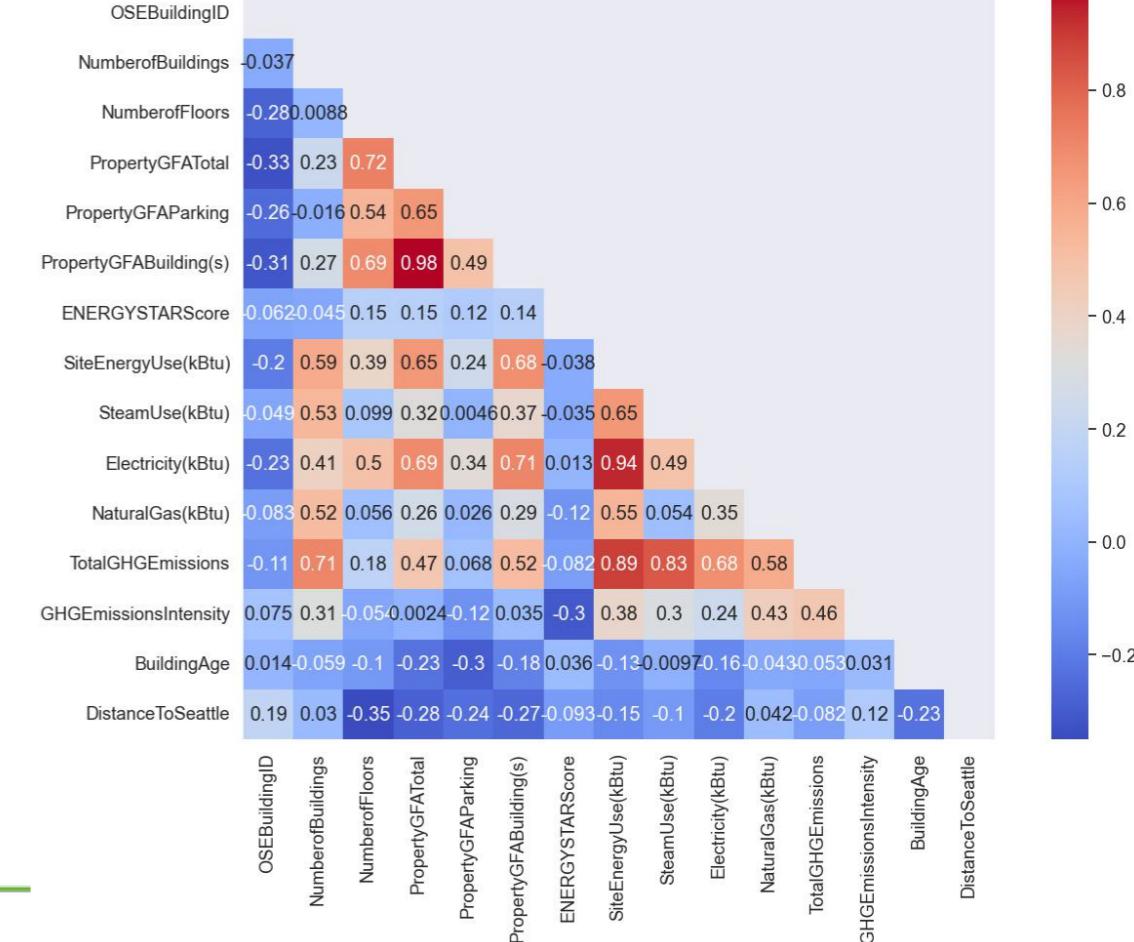
`df['NumberofBuildings'].replace(0,1)`
`df['NumberofFloors'].replace(0,1)`
`df[,SurfacePerBuilding']`
`df['SurfacePerFloor']`

- Log transform the target variables: `y_train_site_log = np.log1p(y_train_site)`

Preprocessing

Réduction de dimension

Elimination des variables fortement corrélées et simplification des features



Variables qualitatives

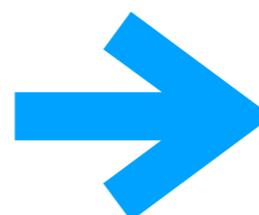
L'encodage **OneHotEncoder**



sur les variables qualitatives

Variables quantitatives

La standardisation **StandardScaler** sur les variables quantitatives.



Échantillonnage

Split du jeu des données en deux parties : données d'entraînement et données de test

Modélisation

Les mesures

de la performance

- R2 — coefficient de détermination

Meilleur score = plus élevé

- RMSE — Racine de l'erreur quadratique moyenne

Meilleur score = plus faible

- Temps moyen de calcul

Les modèles testés

- **LinearRegression**
- **Lasso Regression**
- **Ridge Regression**
- **ElasticNet**
- **Support Vector Regression**
- **k-Nearest Neighbours**
- **Decision Tree Regressor**

- **AdaBoostRegressor**
- **MLPRegressor**
- **Random Forest Regression**
- **Gradient Boosting Regressor**
- **Extra Trees Regressor**

Avec une procédure de grille de recherche pour choisir les hyperparamètres

Les scores pour TotalGHGEmissions

	modele	R2	RMSE	MAE	time
0	Decision Tree Regressor ENERGYSTARScore	0.853000	389.043	102.490	0.009
1	Decision Tree Regressor	0.813000	438.935	126.034	0.009
2	Gradient Boosting Regressor	0.767000	489.567	106.340	0.860
3	MLPRegressor	0.747629	509.443	117.076	0.016
4	Extra Trees Regressor	0.746000	511.075	113.606	0.548
5	RandomForestRegressor	0.566824	667.433	127.445	0.008
6	AdaBoost Regressor	0.565000	668.734	143.153	0.494
7	SVR	0.394000	789.676	138.203	0.374
8	K Neighbors Regressor	0.387000	794.259	161.458	0.001
9	ElasticNet	0.259294	872.767	200.053	0.008
10	Lasso	-0.036027	1032.193	220.372	0.005
11	Ridge	-0.390000	1195.701	20.697	0.007
12	Linear Regression	-0.859636	1382.896	239.792	0.016

Gradient Boosting Regressor 

Les scores obtenus ont été vérifiés par **validation croisée** (5 échantillonnages)



Decision Tree Regressor

Les scores pour SiteEnergyUse

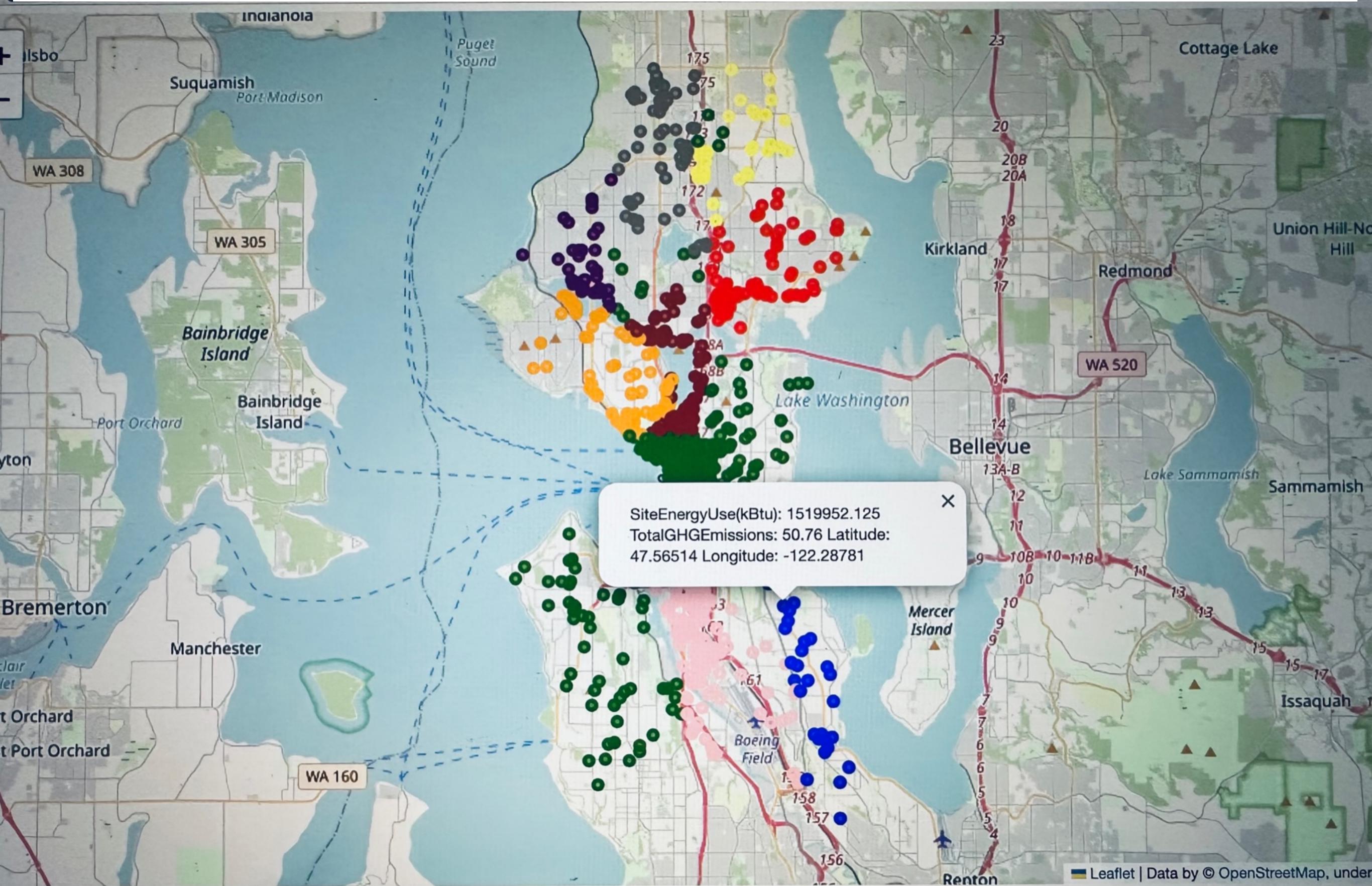
	modele	R2	RMSE	MAE	time
0	Gradient Boosting Regressor ENERGYSTARScore	0.657000	1.884755e+07	4.025102e+06	0.777
1	Gradient Boosting Regressor	0.624000	1.974095e+07	4.662064e+06	0.777
2	Decision Tree Regressor	0.555000	2.148535e+07	5.171799e+06	0.008
3	Extra Trees Regressor	0.552000	2.155473e+07	4.973834e+06	0.508
4	MLPRegressor	0.496095	2.285341e+07	5.684656e+06	0.015
5	RandomForestRegressor	0.430555	2.429421e+07	5.330557e+06	0.015
6	Bagging Regressor	0.424000	2.443508e+07	5.457322e+06	0.756
7	SVR1	0.396000	6.257545e+14	5.838264e+06	0.364
8	AdaBoost Regressor	0.335000	2.624749e+07	6.695223e+06	0.480
9	K Neighbors Regressor	0.323000	2.649612e+07	6.439078e+06	0.001
10	Lasso	-0.059046	3.313098e+07	9.443971e+06	0.002
11	Ridge	-3.186000	6.586693e+07	1.136689e+07	0.006
12	ElasticNet	-3.273915	6.655638e+07	1.161955e+07	0.010
13	Linear Regression	-3.592405	6.899170e+07	1.129509e+07	0.015



Intérêt de l'Energy Star Score

- L'Energy Star Score est un outil de dépistage aidant à évaluer les performances d'émission de CO₂ d'un bâtiment par rapport aux établissements similaires.
- En utilisant la cote ENERGYSTAR, nous avons fait de bonnes prévisions pour la réduction des émissions de C₂O dans les bâtiments non résidentiels de la ville de Seattle.
- Les prédictions de la consommation totale d'énergie avec la feature Energy Star score sont légèrement améliorées.
- Decision Tree Regressor est l'algorithme qui nous donne les meilleurs résultats pour la prédiction de la réduction des émissions de C₂O, et le temps calculé est beaucoup plus petit que les algorithmes Gradient Boosting, RandomForest, MLPRegressor et ExtraTrees.
- Gradient Boosting est l'algorithme qui nous donne les meilleurs résultats pour la prédiction de la consommation totale d'énergie, et le temps calculé est beaucoup plus petit que les algorithmes Decision Tree Regressor , RandomForest, MLPRegressor et ExtraTrees.
- Les modèles que nous avons créé dans ce notebook peuvent être utile si la classification ENERGYSTAR est utilisée à l'avenir.

Une carte interactive de la ville de Seattle



A nighttime photograph of a city skyline, likely Montreal, featuring a prominent illuminated Ferris wheel in the foreground. The city is filled with numerous skyscrapers, their windows glowing with lights. The sky is dark blue.

Merci de votre attention