OPENCLASSROOMS

Projet 4

OpenClassrooms

Kevin

Parcours Data Scientist

Rappel du sujet/problématique

Ville de Seattle

Neutralité carbone en 2050

Emission des batiments non destinés à l’habitation

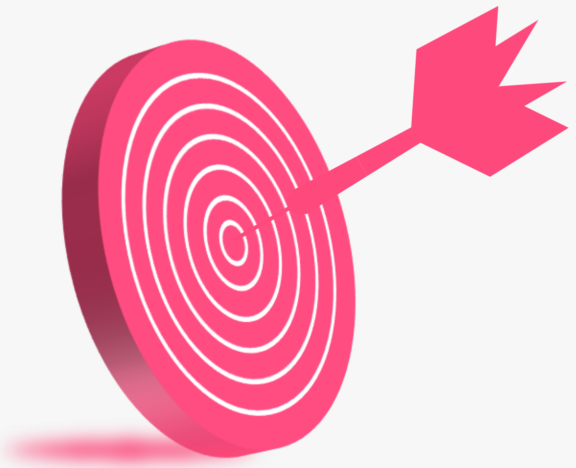
Problématique

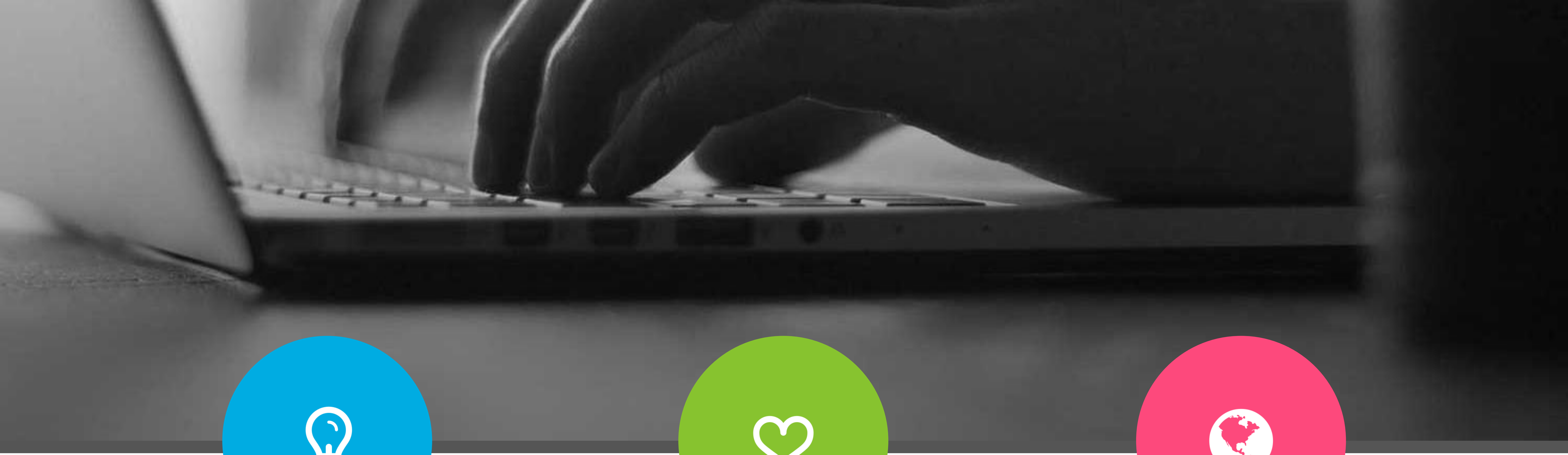
Relevés de consommation en 2015 et 2016

Couteux à obtenir

Objectifs

- Prédire les émissions de CO_2
- Prédire la consommation énergétique
- Réduire les coûts
- Variable « ENERGYSTARSCORE »





Dataset et Réflexion

Pistes de modélisation

Modèle final

1 Dataset et Réflexion

Dataset

2 fichiers

2 périodes différentes :

2015 – 2016

Nombre de batiments différents :

3340 – 3376

Variables différentes :

Address ou City (2016), introuvable en 2015

Ajustements

Présence de dictionnaires :

Extraire les variables

{'latitude': '47.61219025', 'longitude': '-122.33799744', 'human_address': '{"address": "405 OLIVE WAY", "city": "SEATTLE", "state": "WA", "zip": "981
01"}'}

Noms différents :

GHGEmissions(MetricTonCO2e) – TotalGHGEmissions

Types de variables :

Types différents

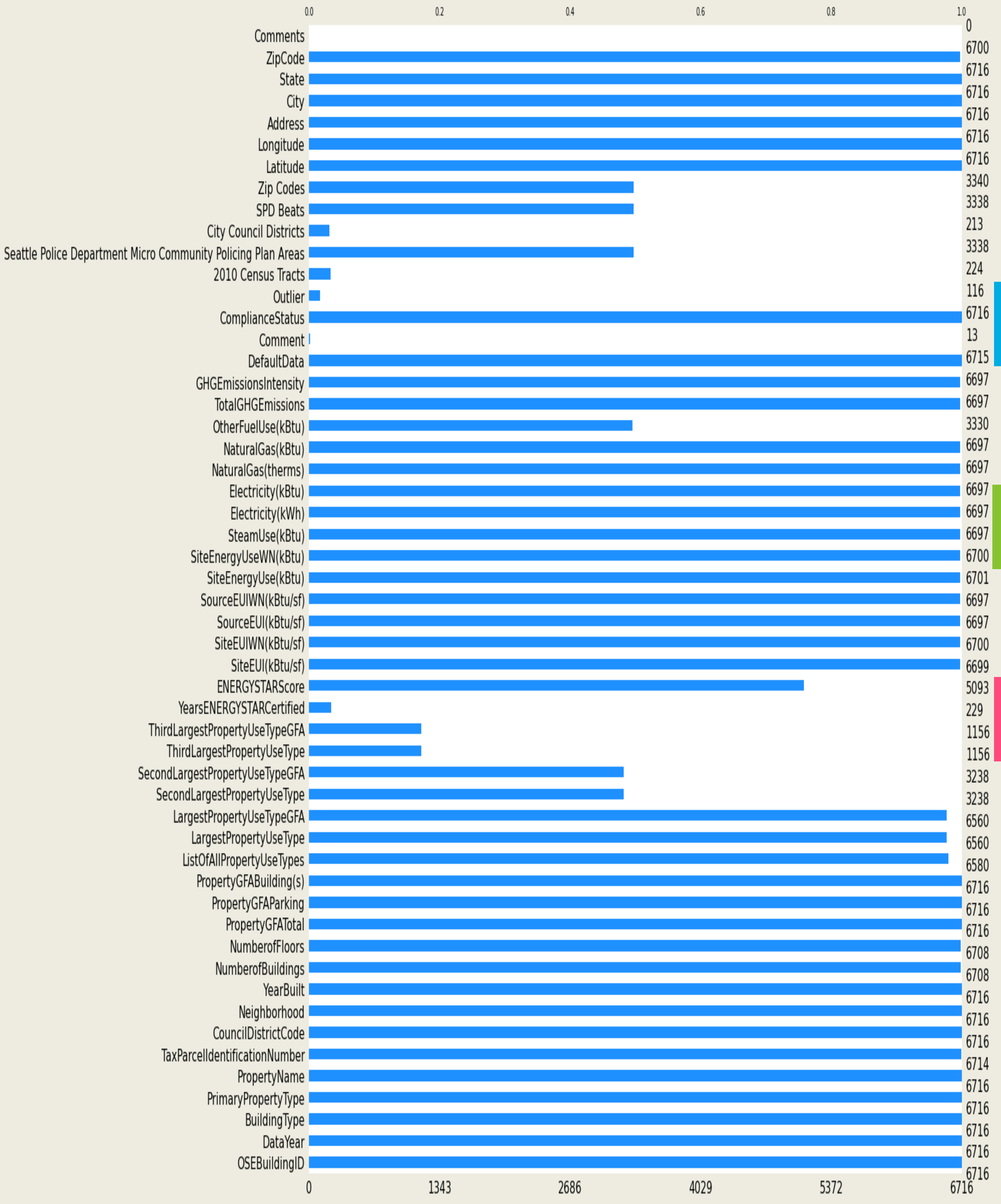
Analyse exploratoire

1 Données manquantes

Suppression des colonnes avec plus de 70% de données manquantes

2 Doublons

Même batiment en 2015 et 2016 : 3300



Analyse exploratoire

1

Données manquantes

Suppression des colonnes avec plus de 70% de données manquantes

2

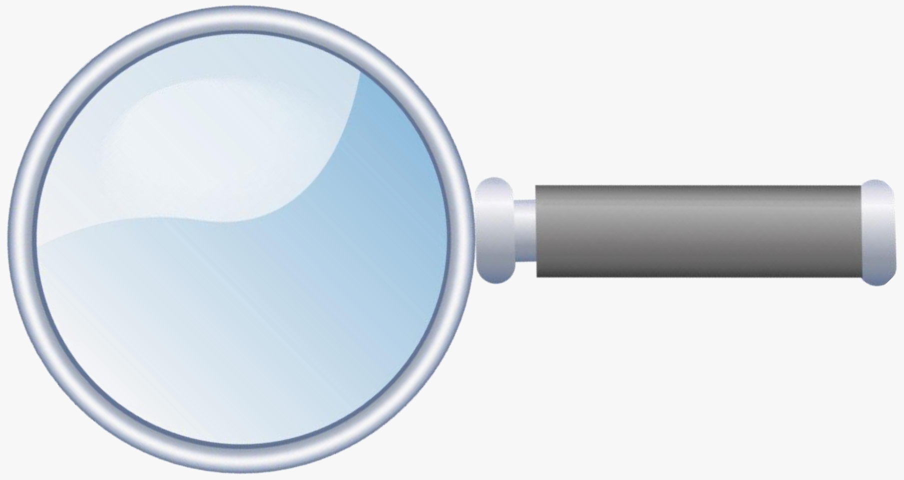
Doublons

Même batiment en 2015 et 2016 : 3300

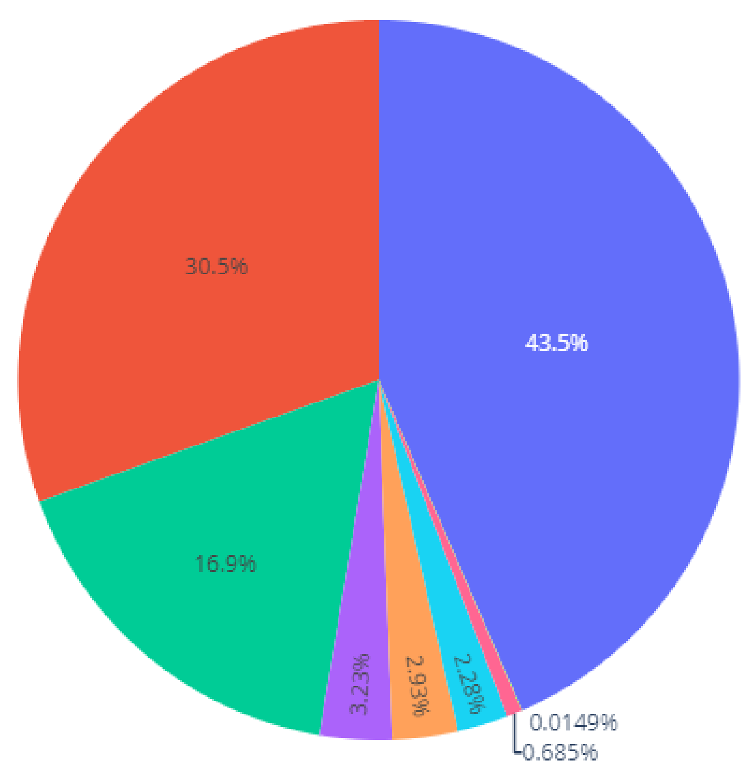
3

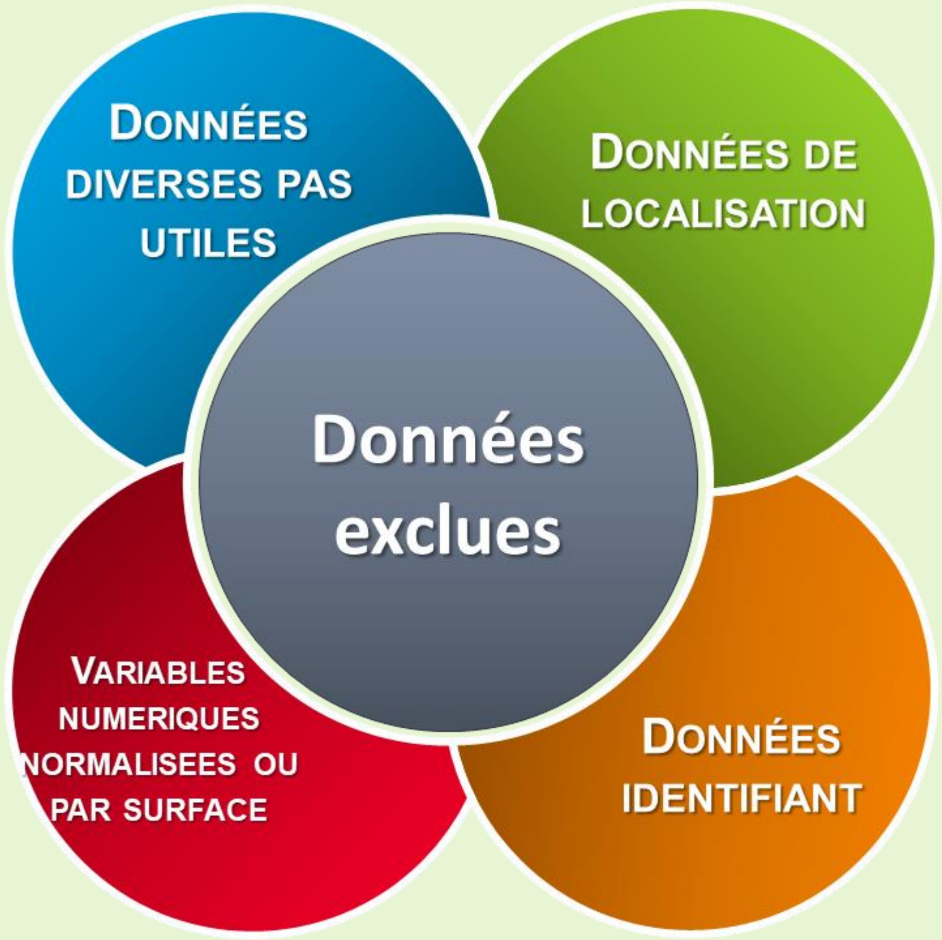
Réflexion sur la problématique

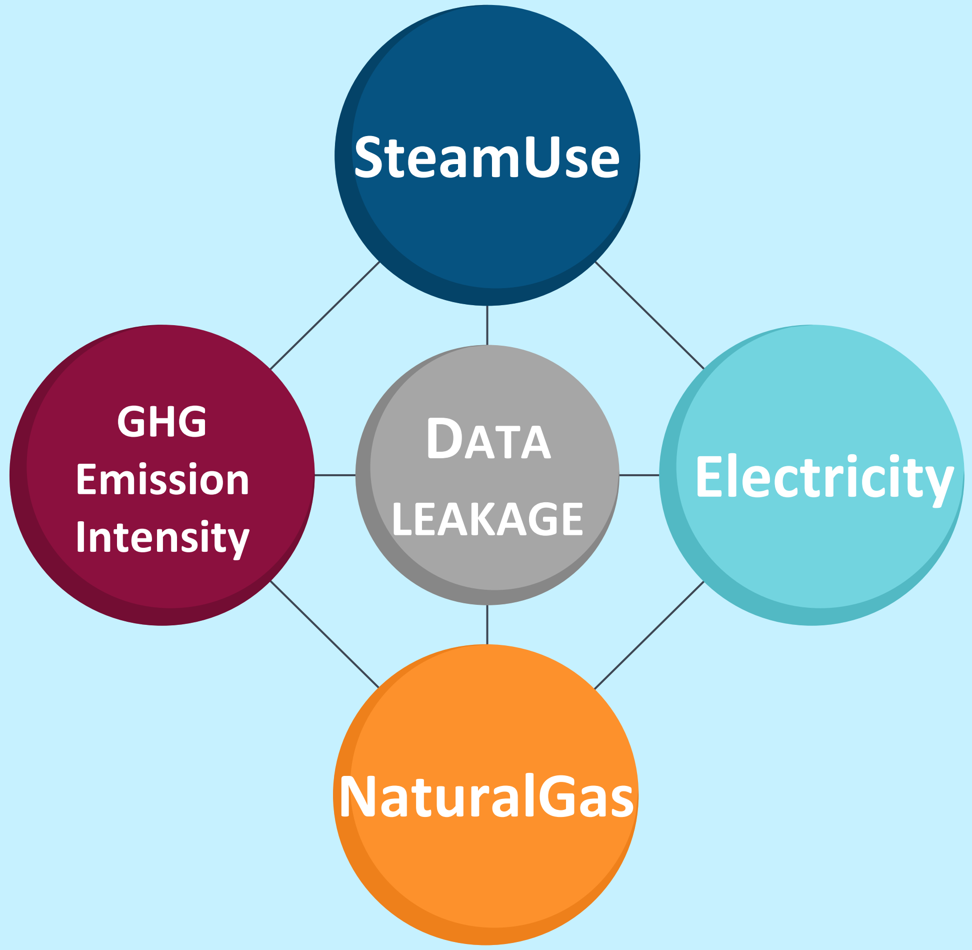
Prédire les émissions



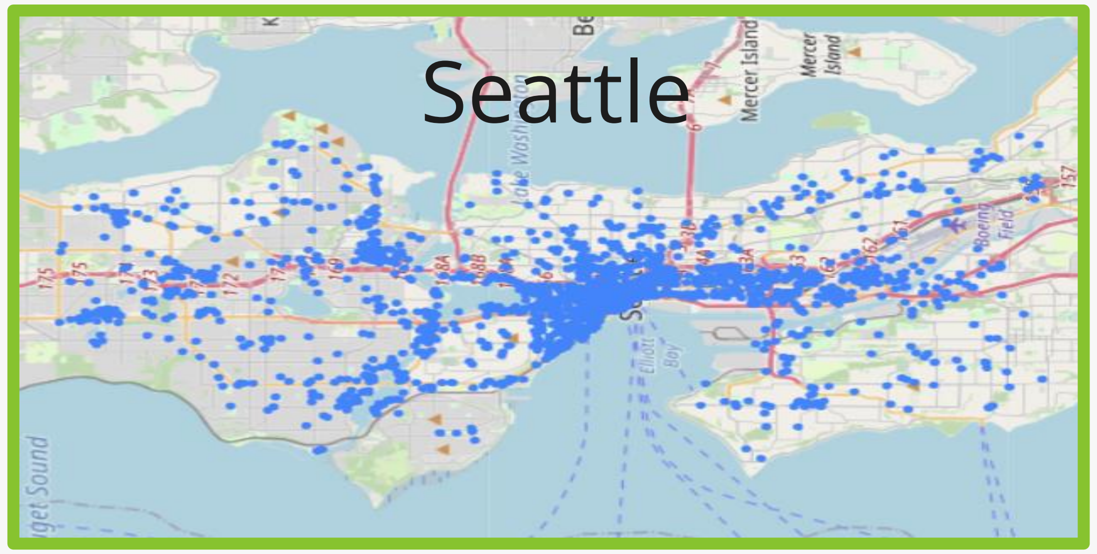
Réflexion sur la problématique

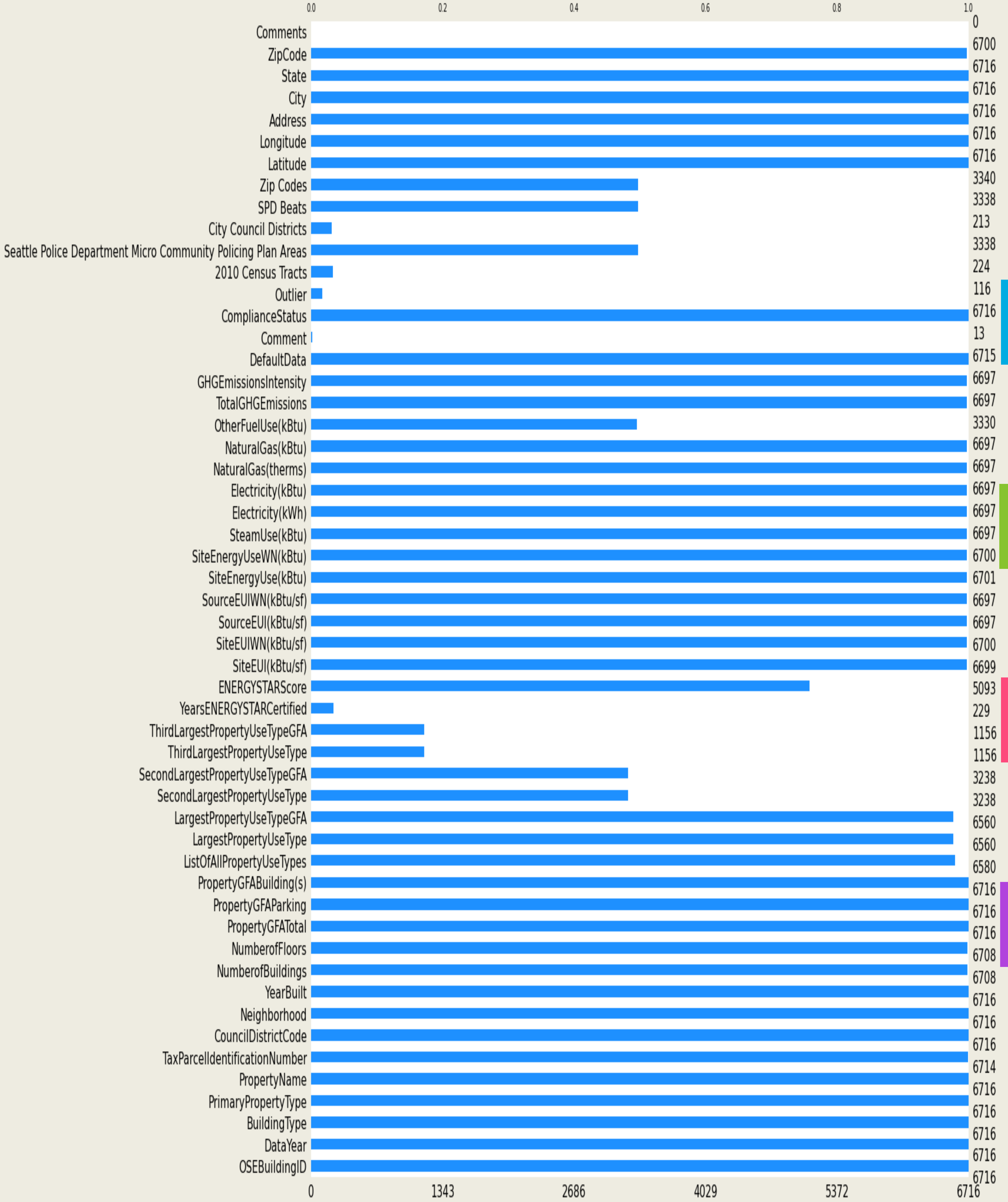






- NonResidential
- Multifamily LR (1-4)
- Multifamily MR (5-9)
- Multifamily HR (10+)
- SPS-District K-12
- Nonresidential COS
- Campus
- Nonresidential WA





Analyse exploratoire

1 Données manquantes

Suppression des colonnes avec plus de 70% de données manquantes

2 Doublons

Même batiment en 2015 et 2016 : 3300

3 Réflexion sur la problématique

Te mel movet equidem vivendum

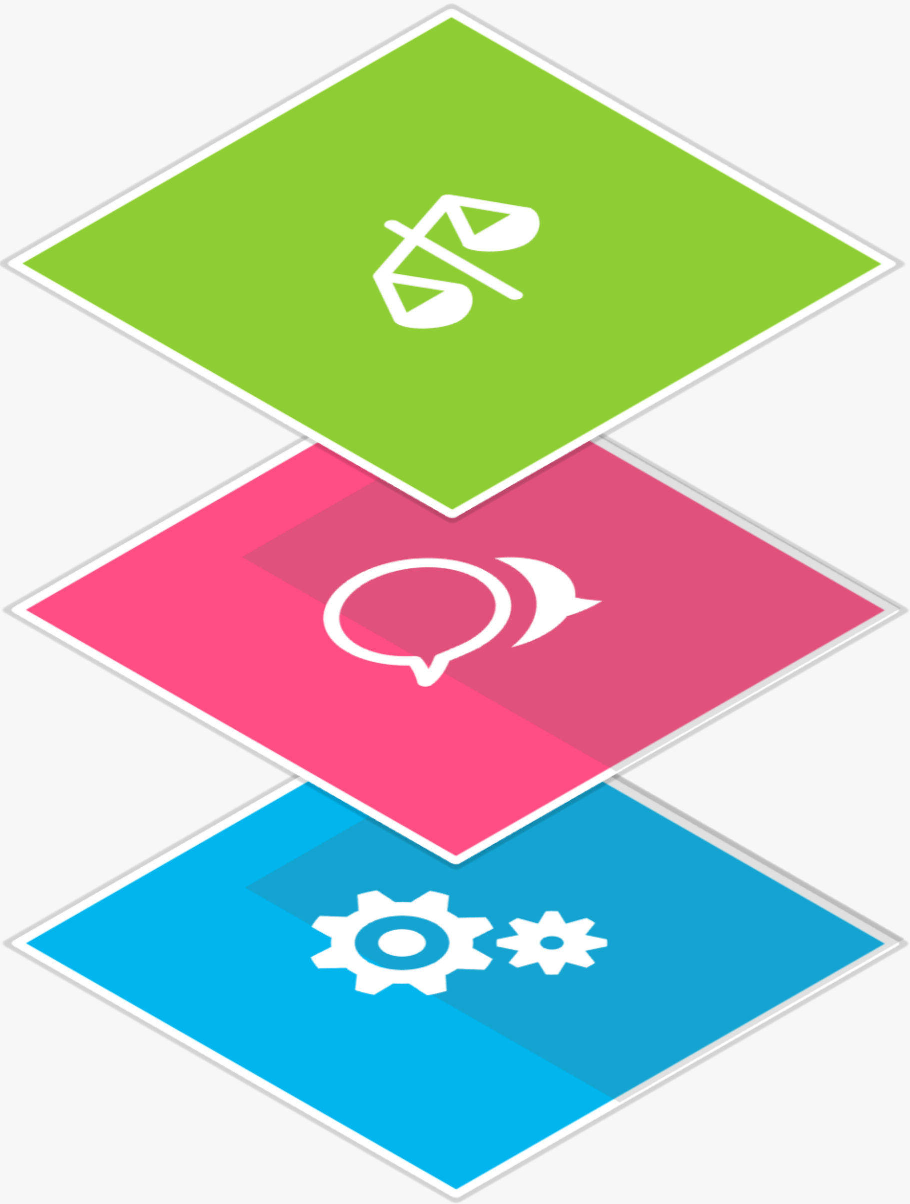
4 Outliers

Etude des outliers

Outliers

Nombre de batiments/etages

Valeurs negatives, nulls ou NAN



PropertyGFATotal et Emission

Valeurs négatives

Identification des cas particuliers

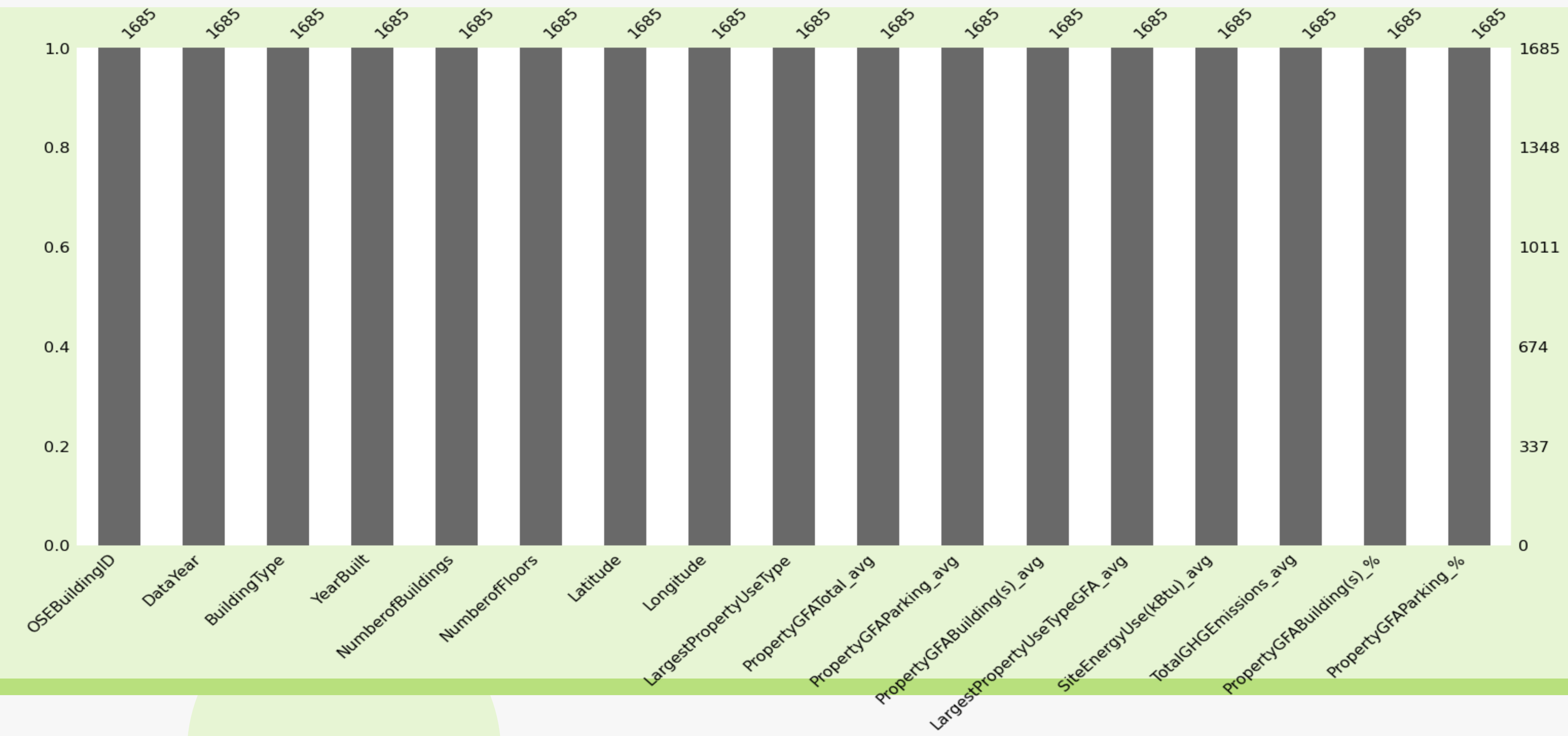
Analyse des écarts...

Dataset final

1. **Suppression de variables redondantes**
   Seconde et troisième type de surface les plus larges
2. **Regroupement par ID**
   Regrouper les variables 2015-2016
3. **Variables écarts 2015/2016**
   Ecart de consommation énergétique, émission carbone et superficie
4. **Simplification des variables de surface**
   Superficie en % du total
5. **Simplification des types de batiment**
   Regroupement dans des thématiques
6. **Age des batiments**

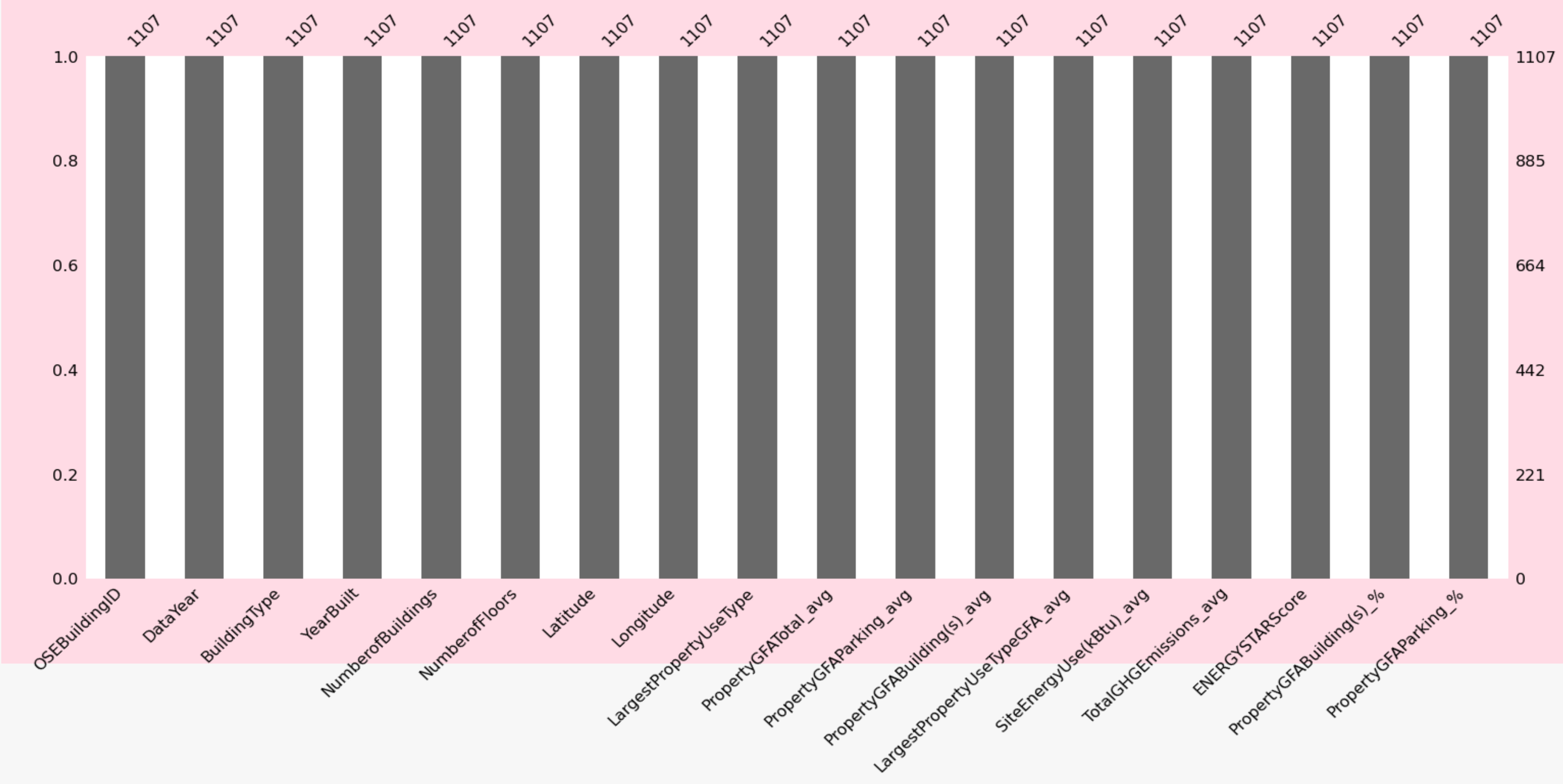
Distinction de deux datasets

One Column

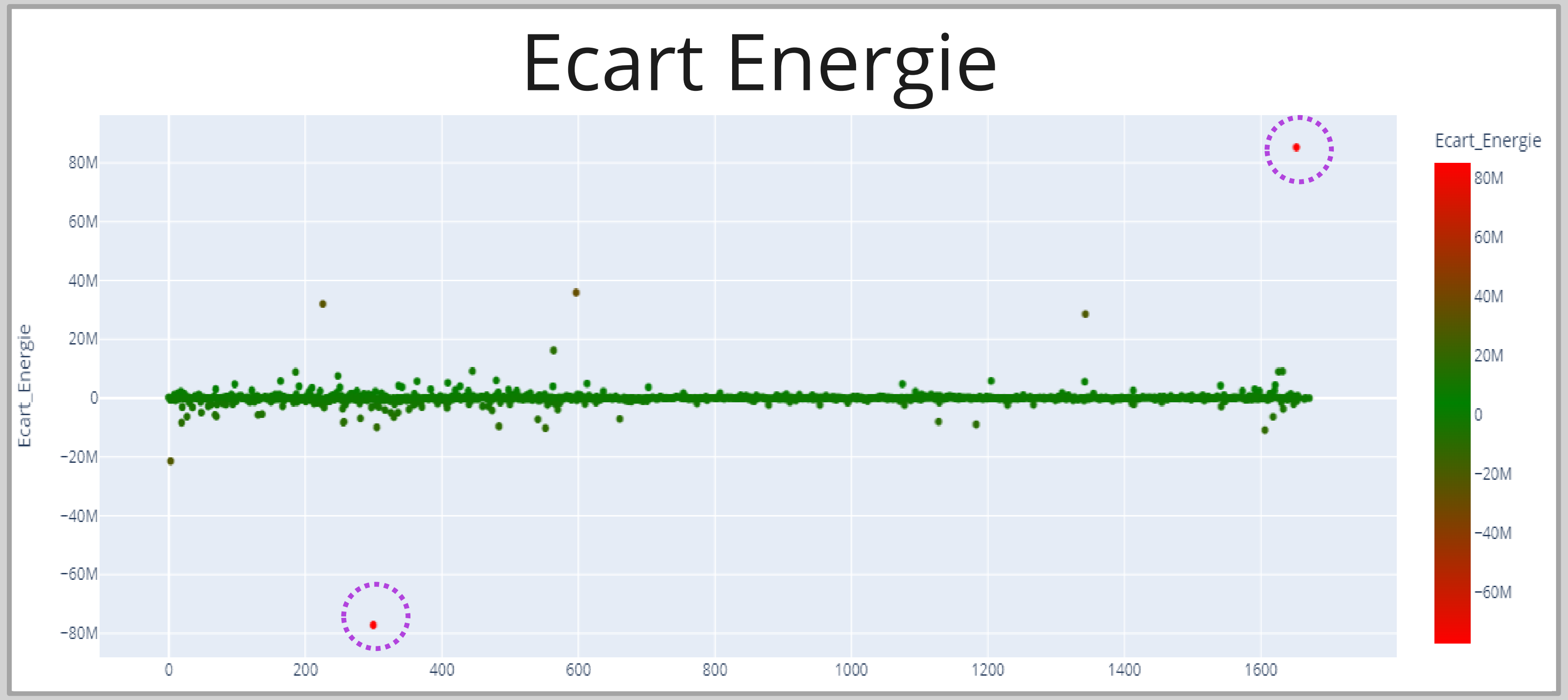


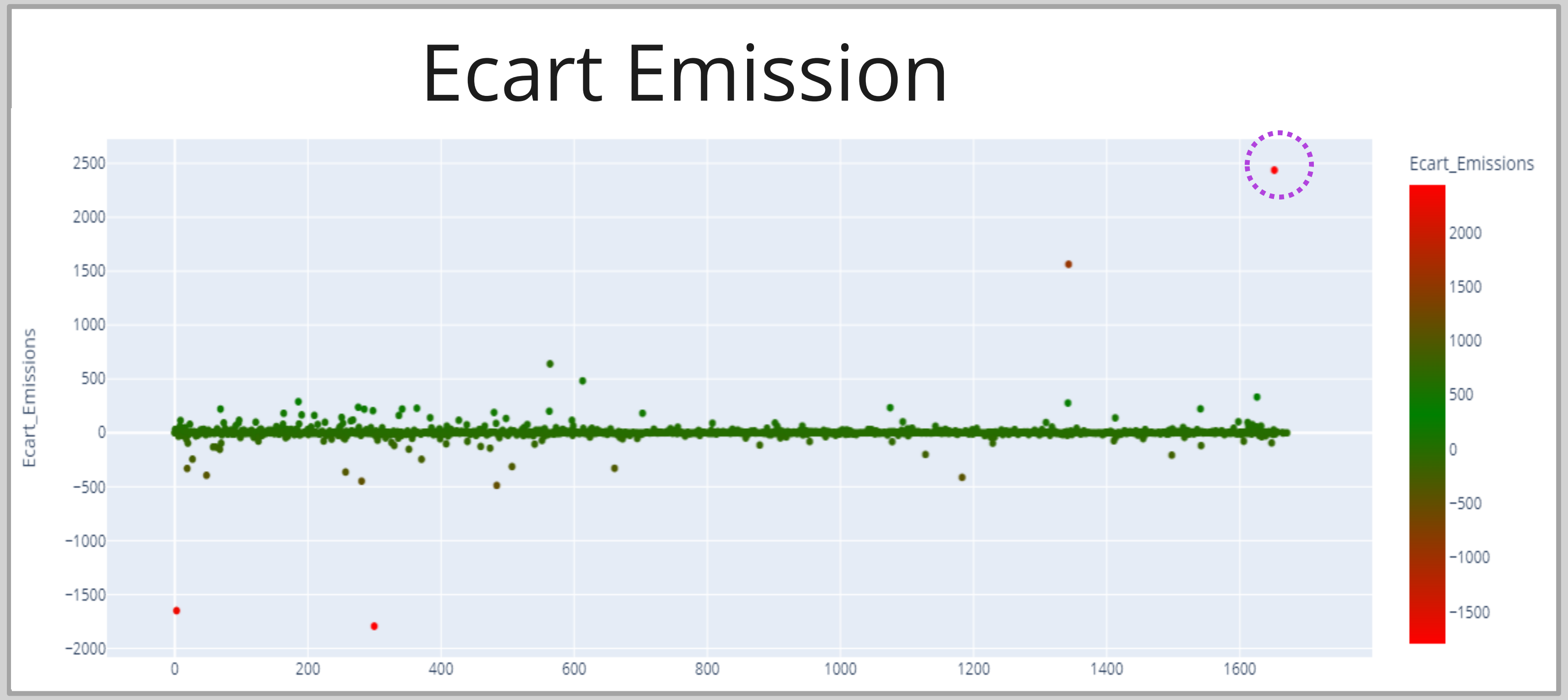
Sans EnergyStarScore

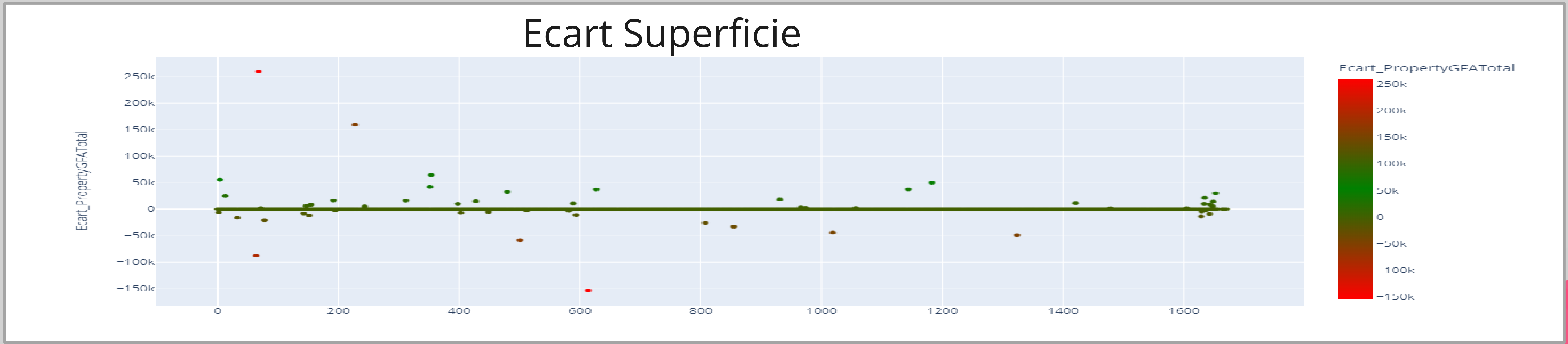
Avec EnergyStarScore



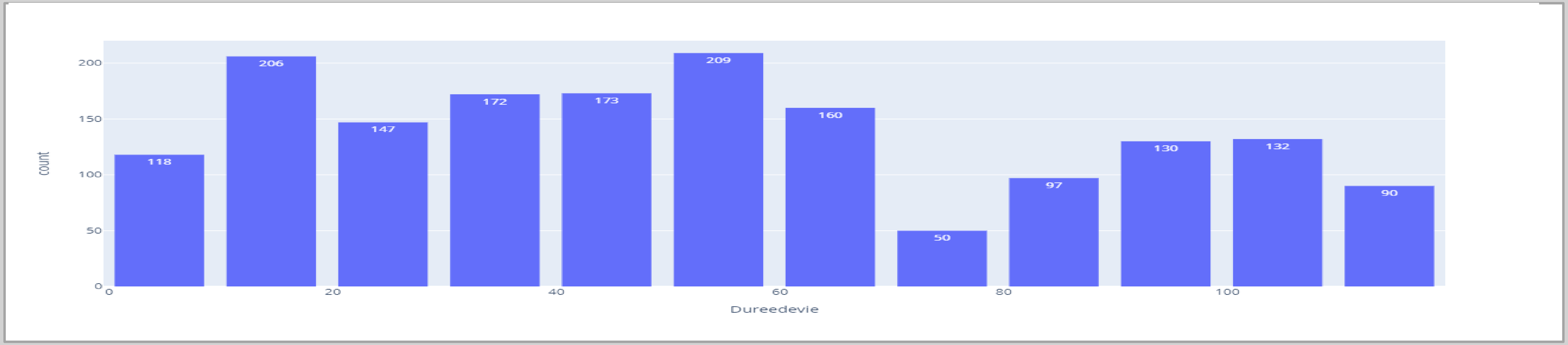
Identification des cas particuliers



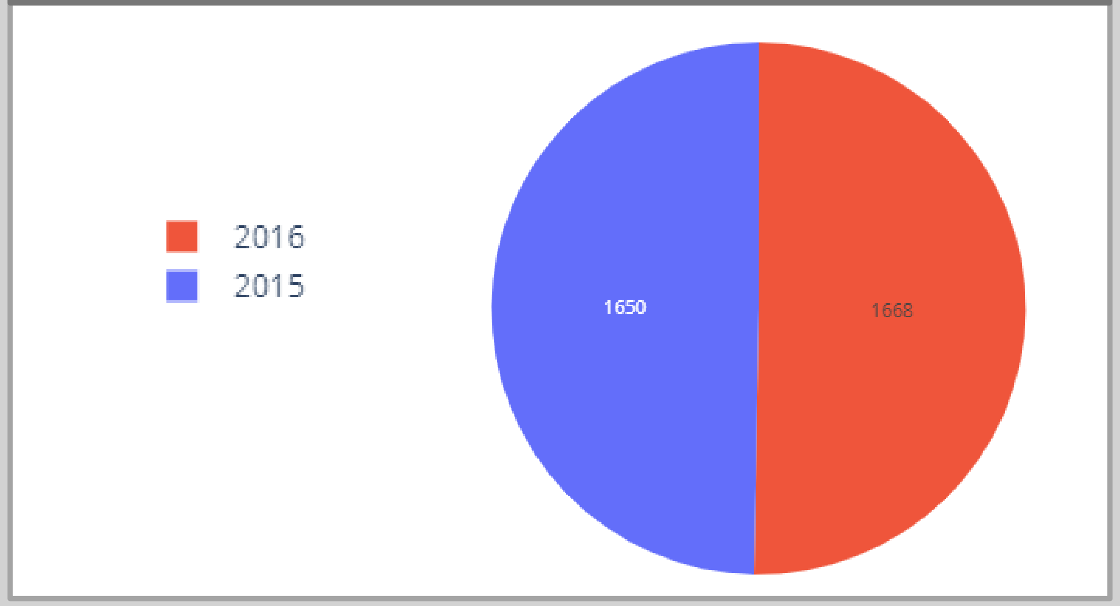




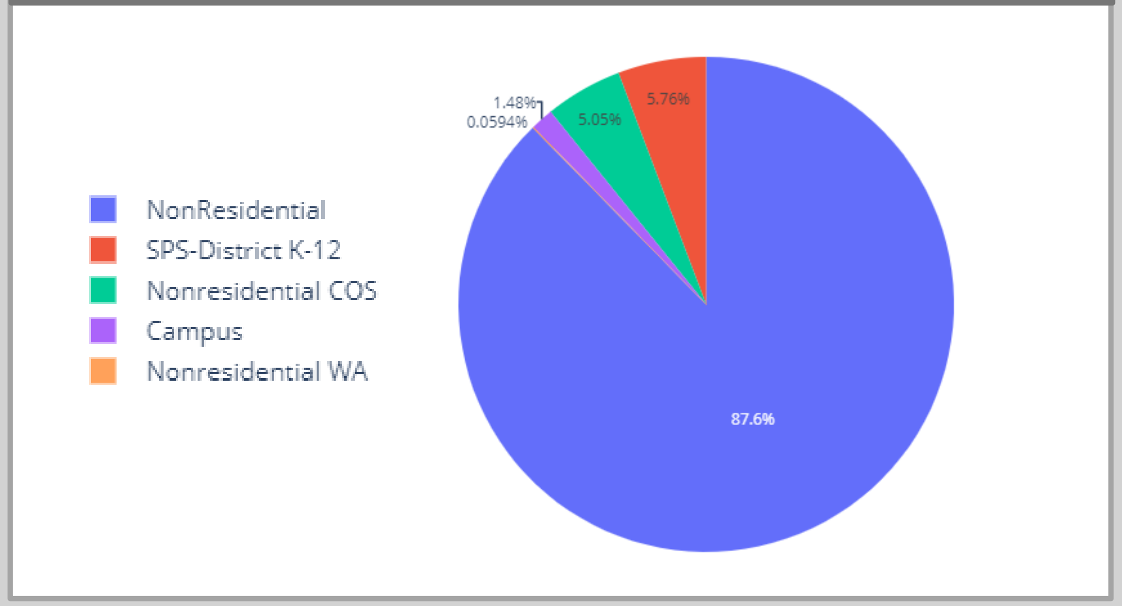
Caractéristiques Batiment



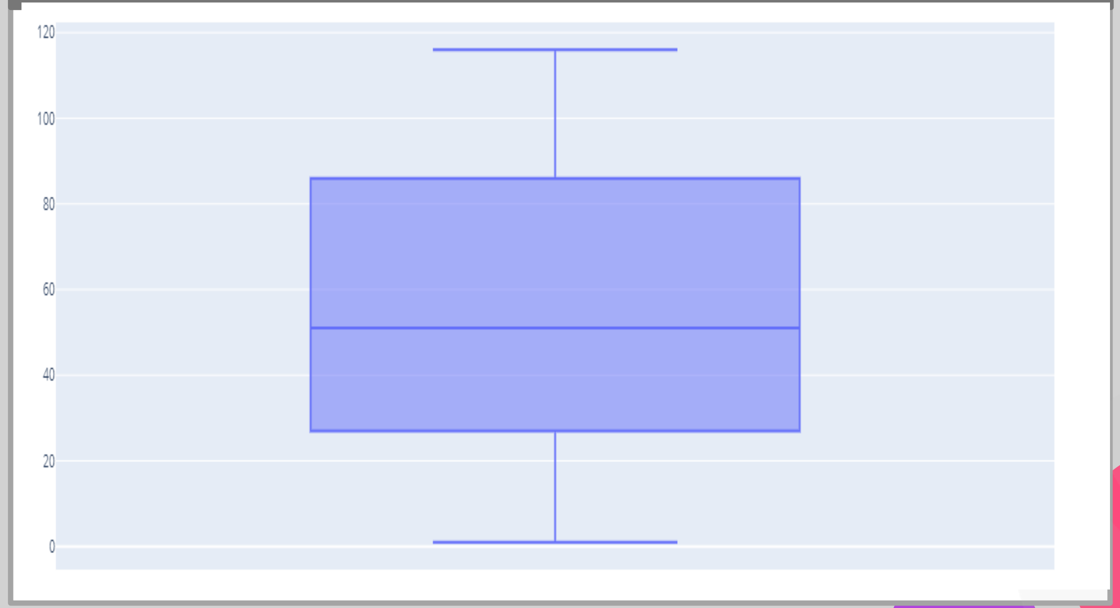
Occurrences par année



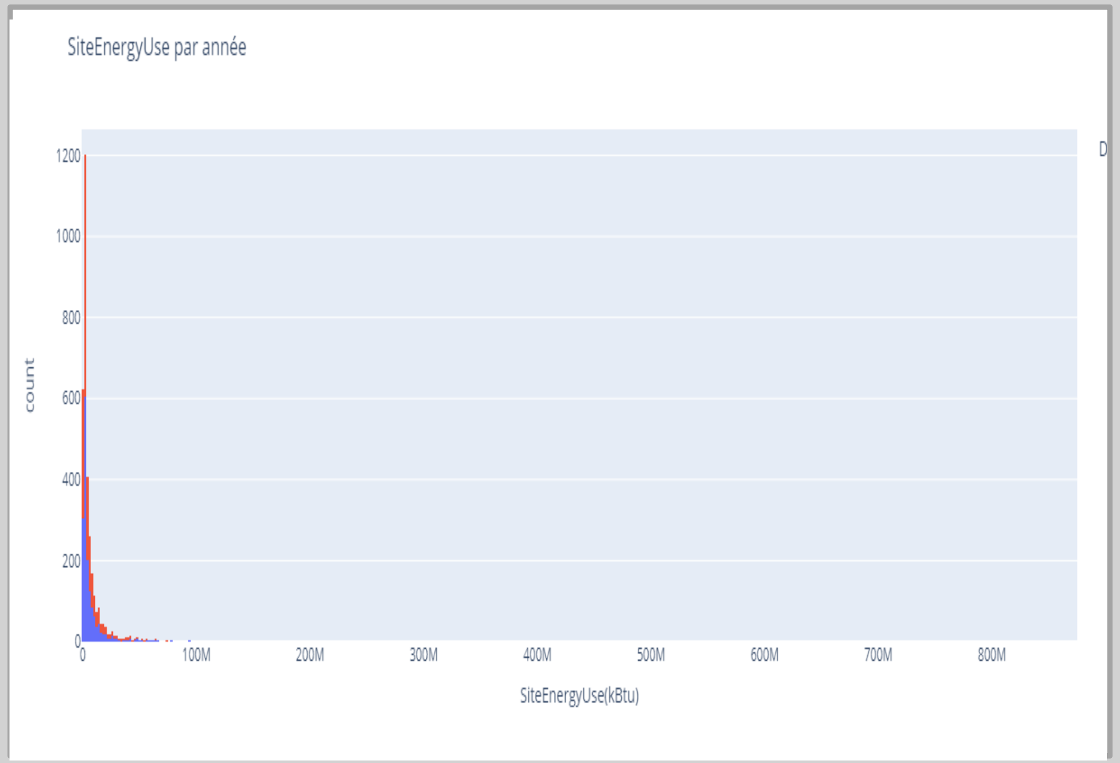
Type de batiment

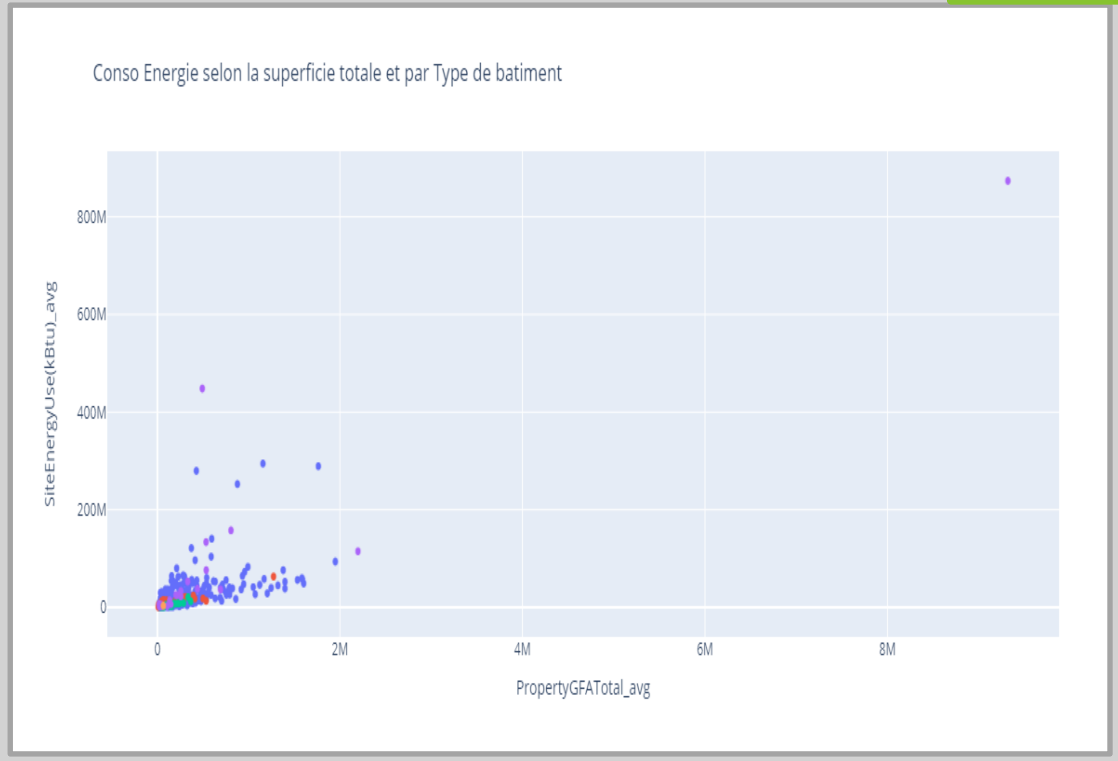


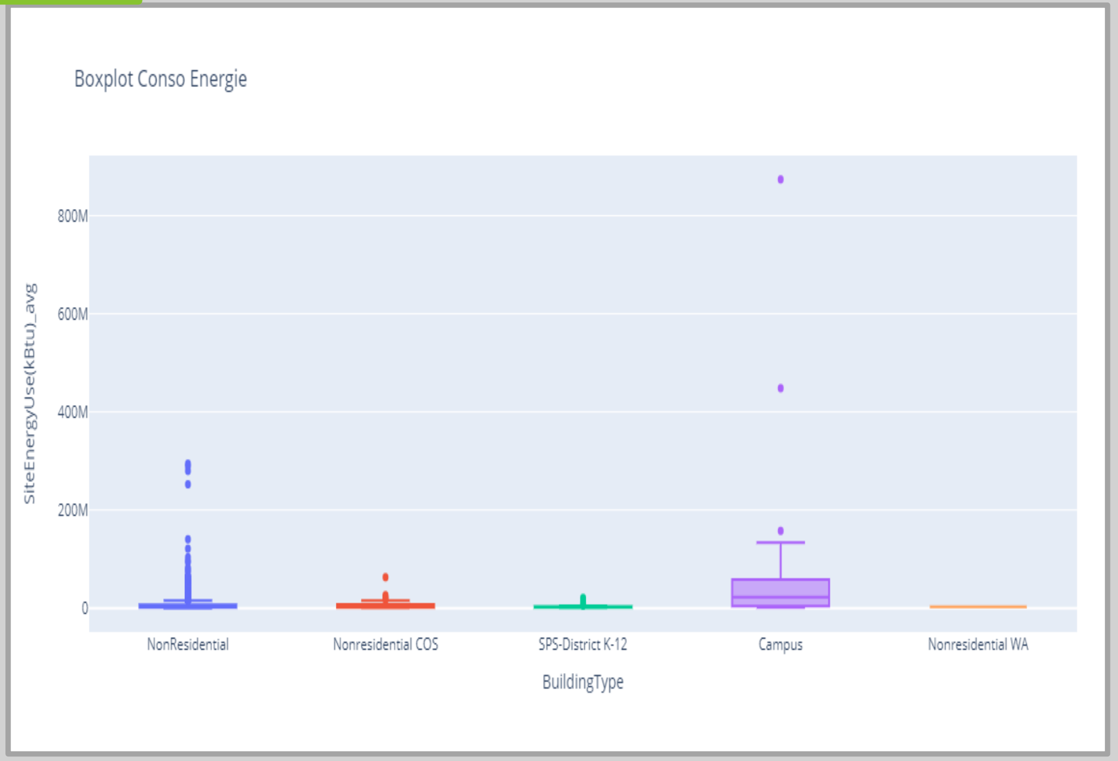
Age des batiments

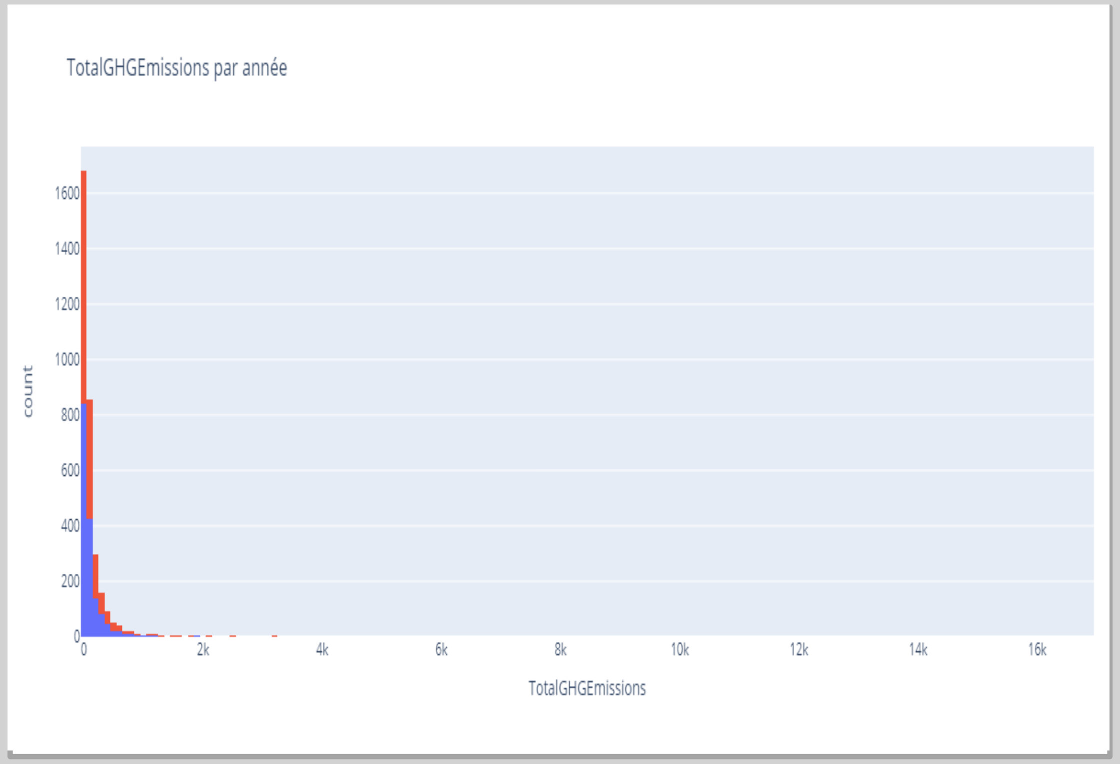


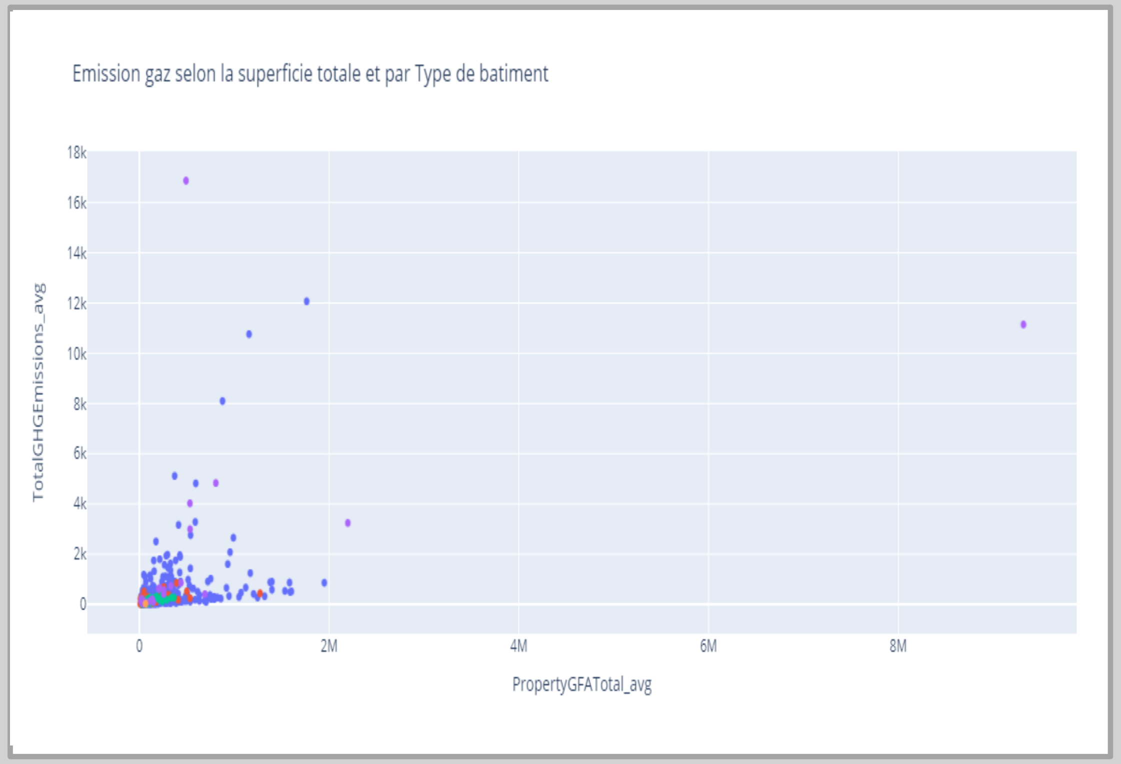
Consommation énergétique et émission de carbone

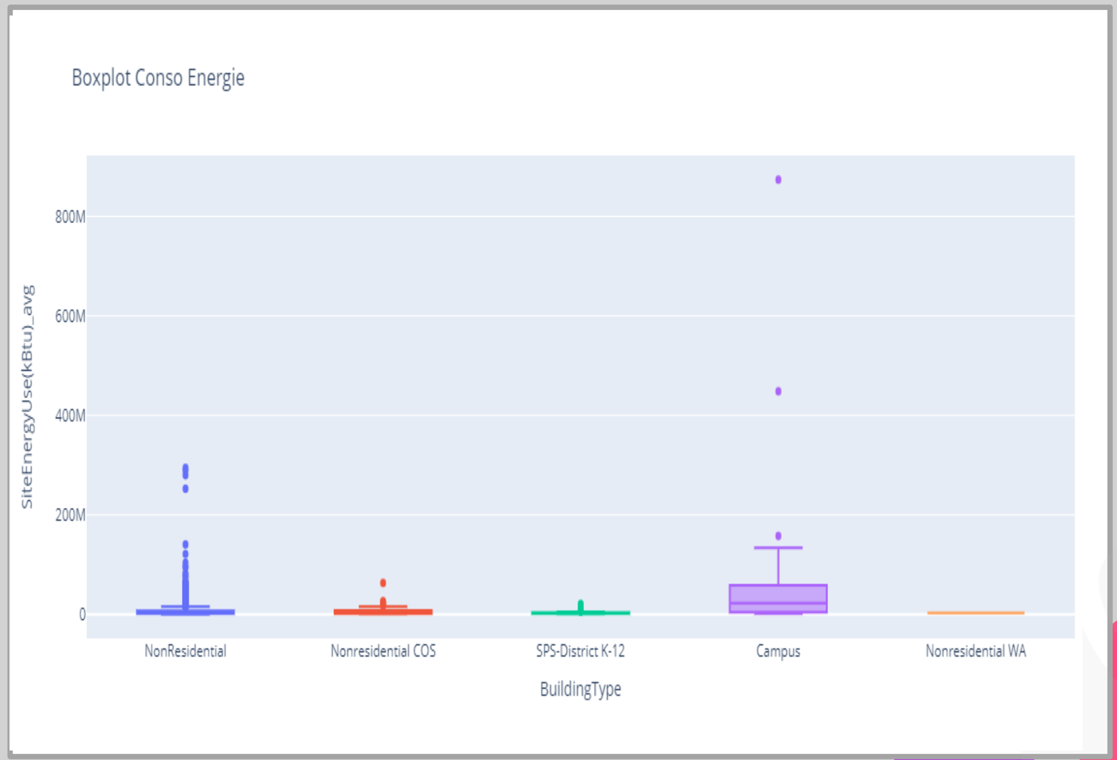




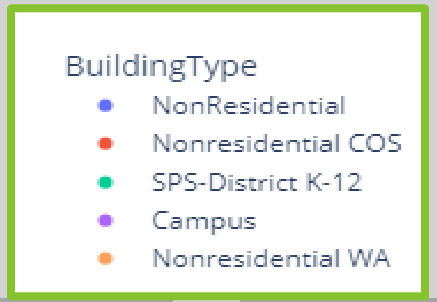


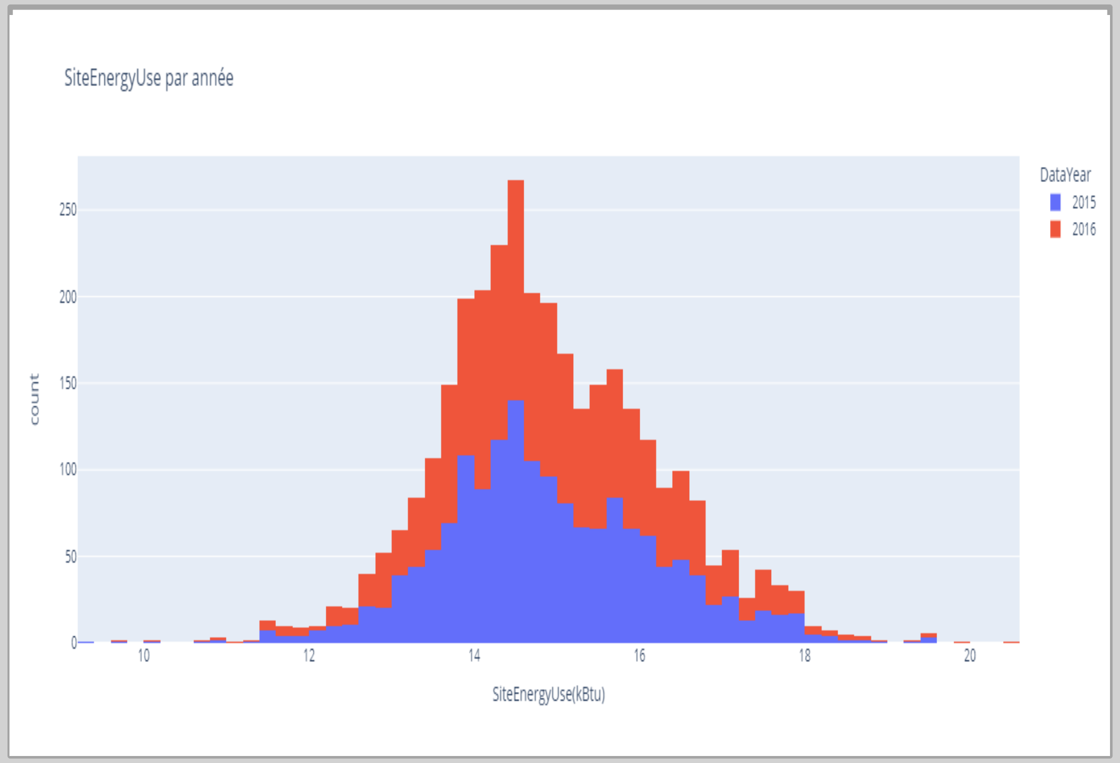


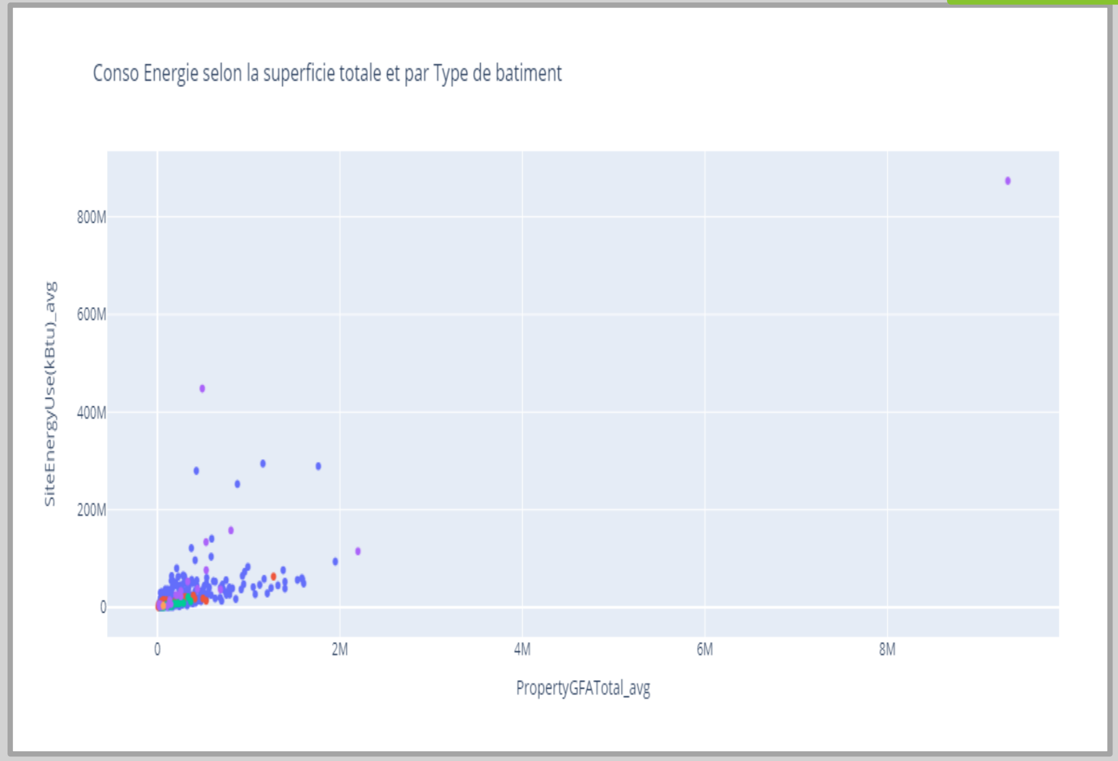


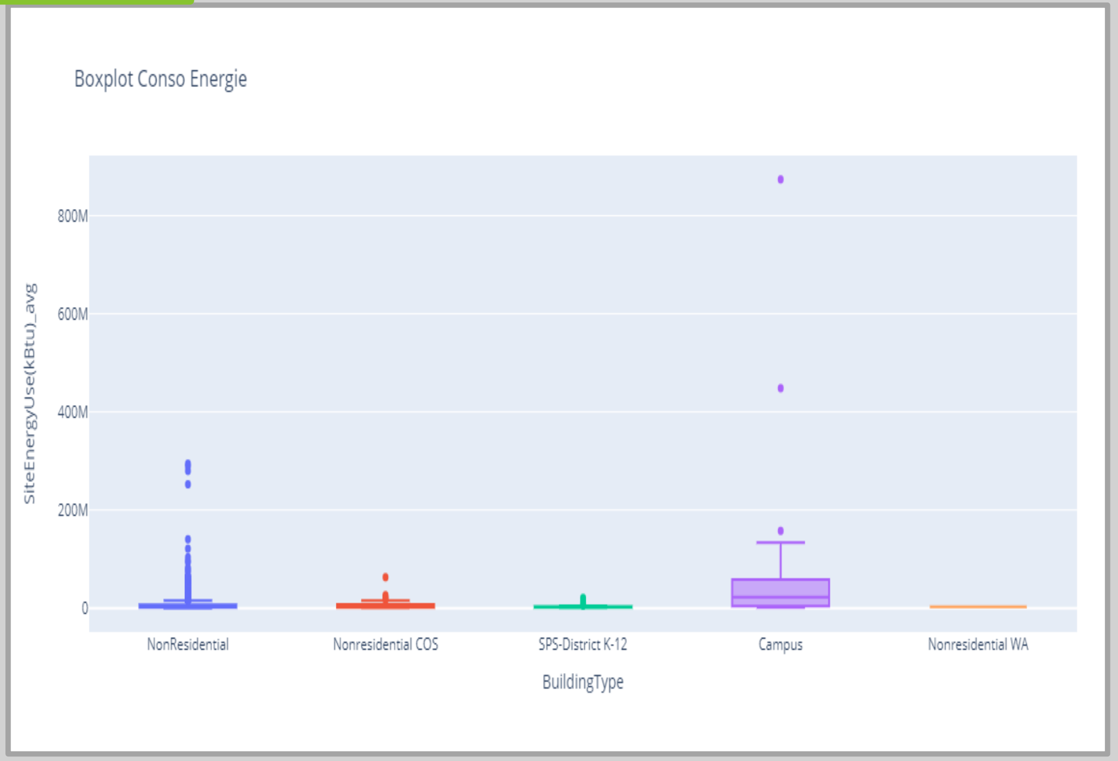


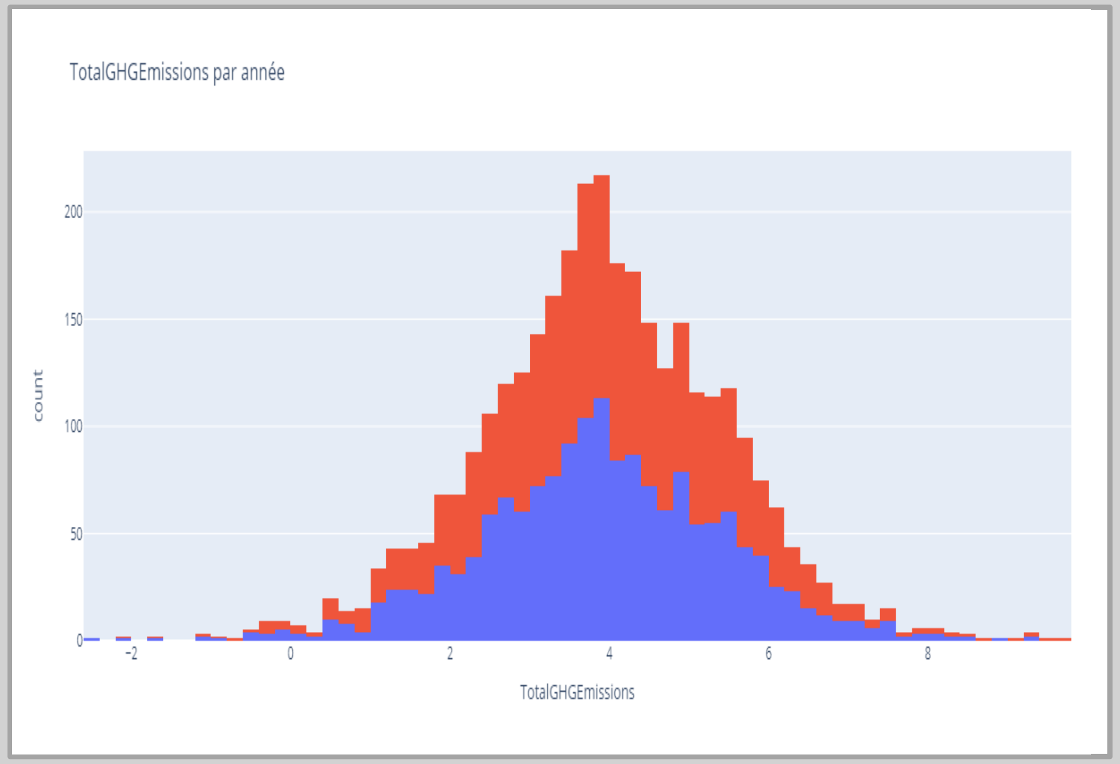
Consommation énergétique et émission de carbone

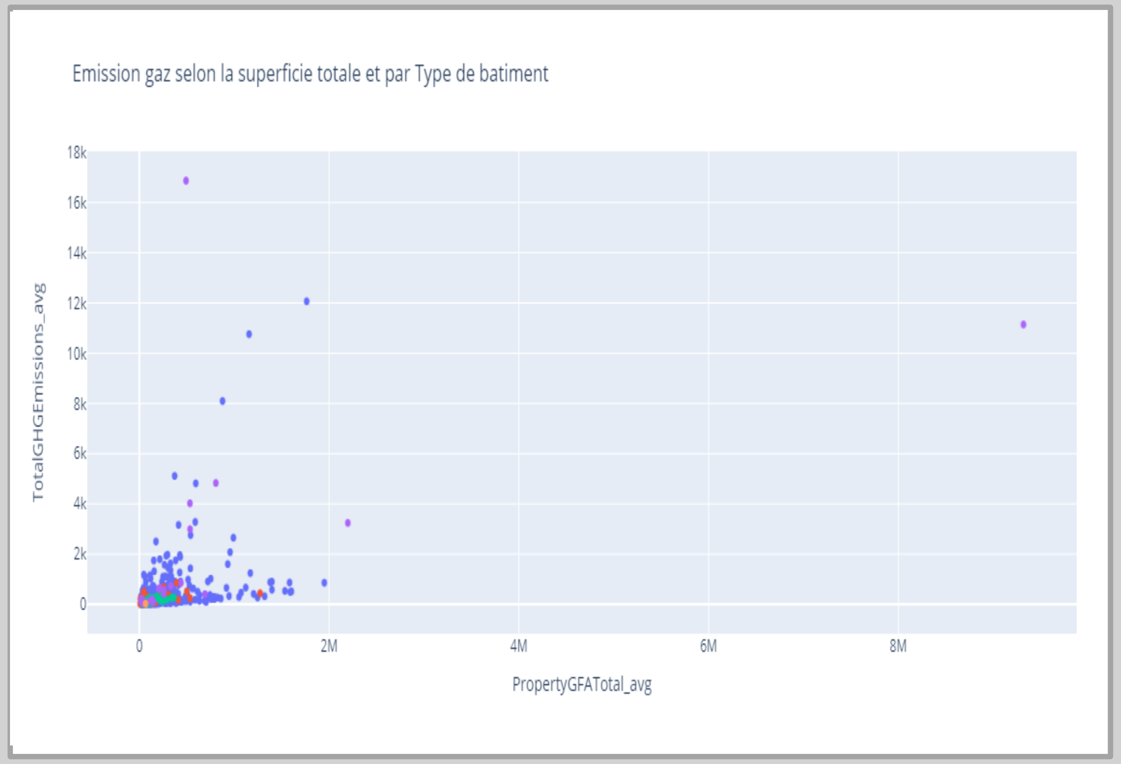


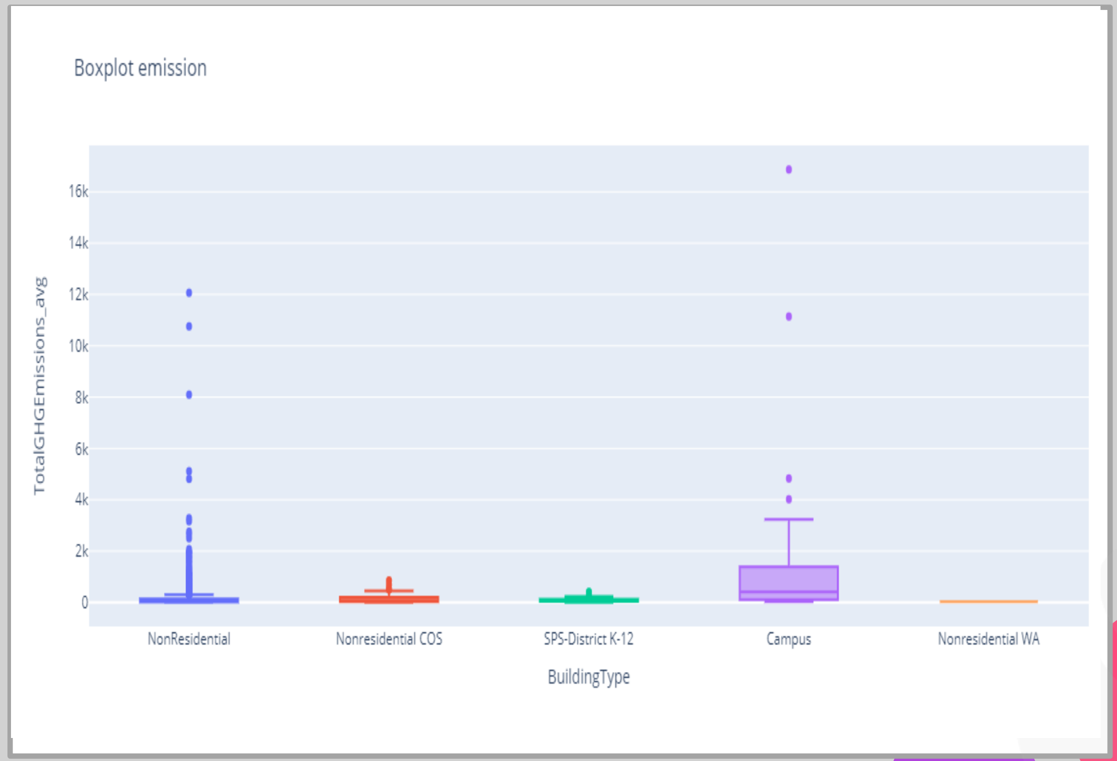


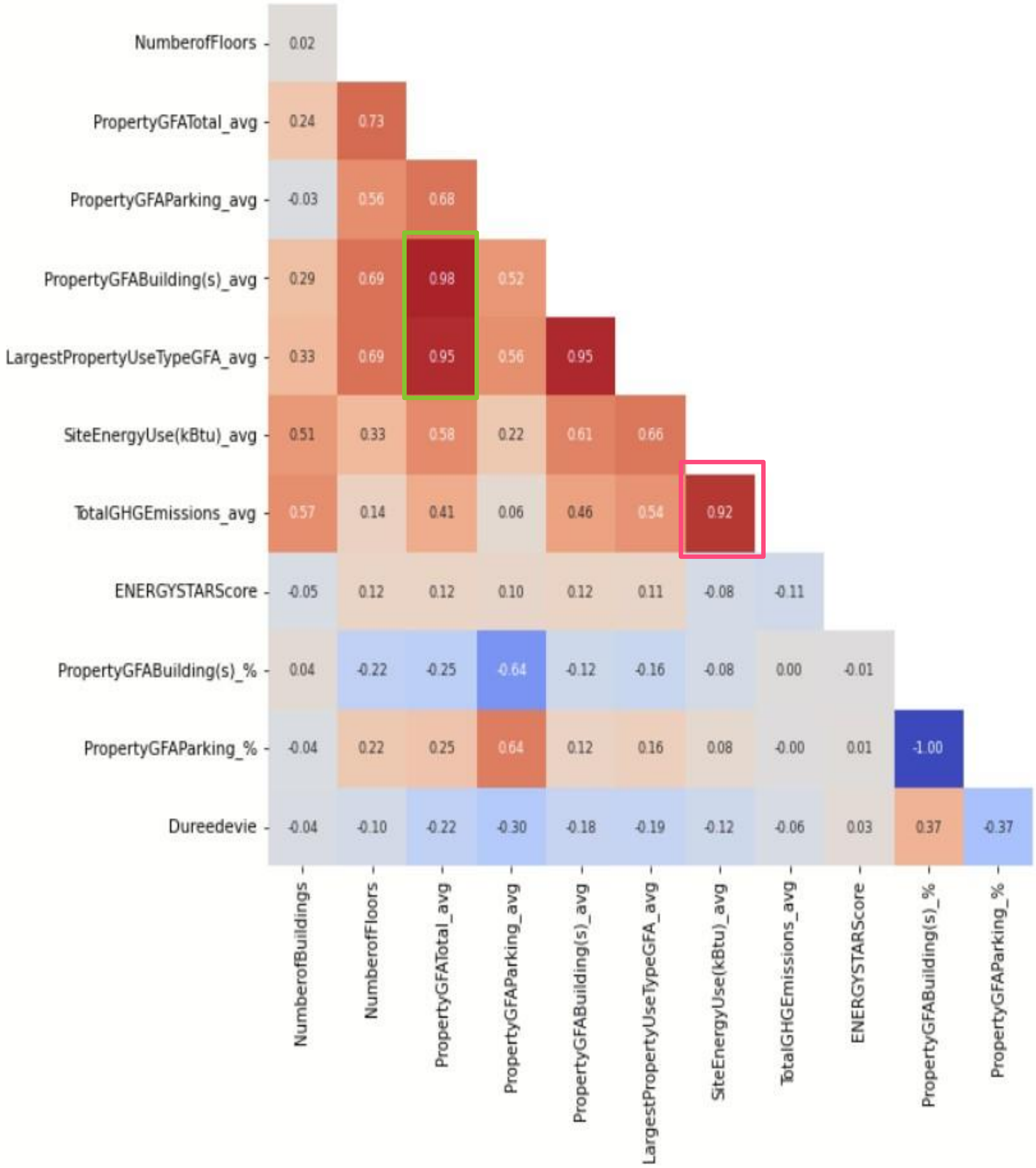










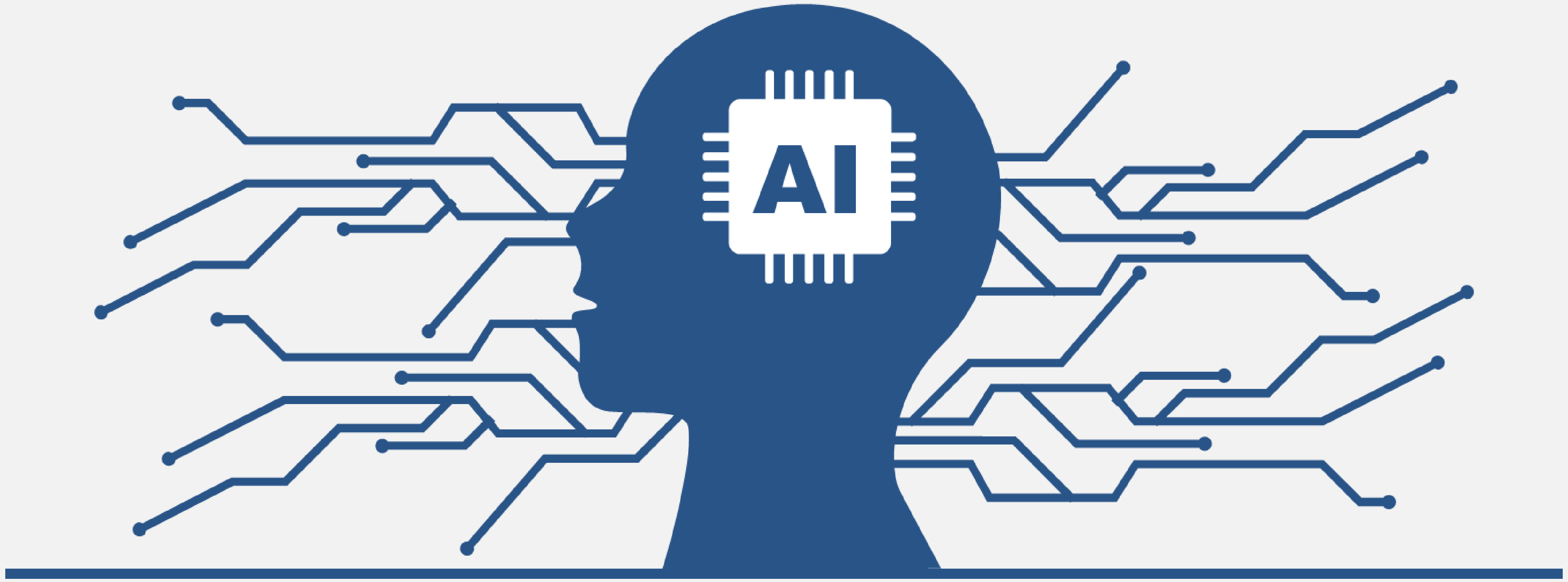


Corrélation

Fortes correlations entre :

- **PropertyGFATotal / PropertyGFABuilding / LargestProperty**

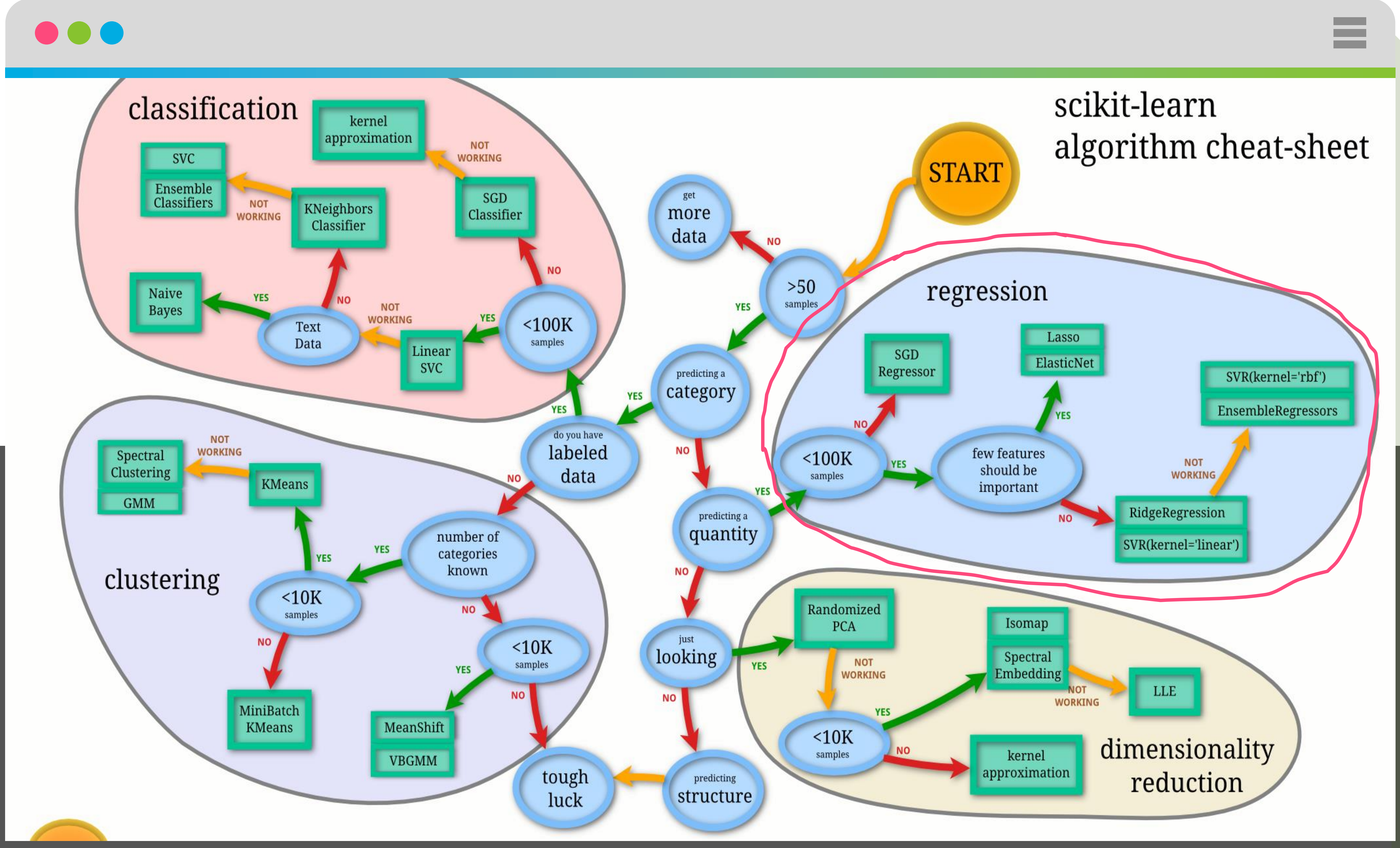
- **TotalGHGEmissions / SiteEnergyUse**



2

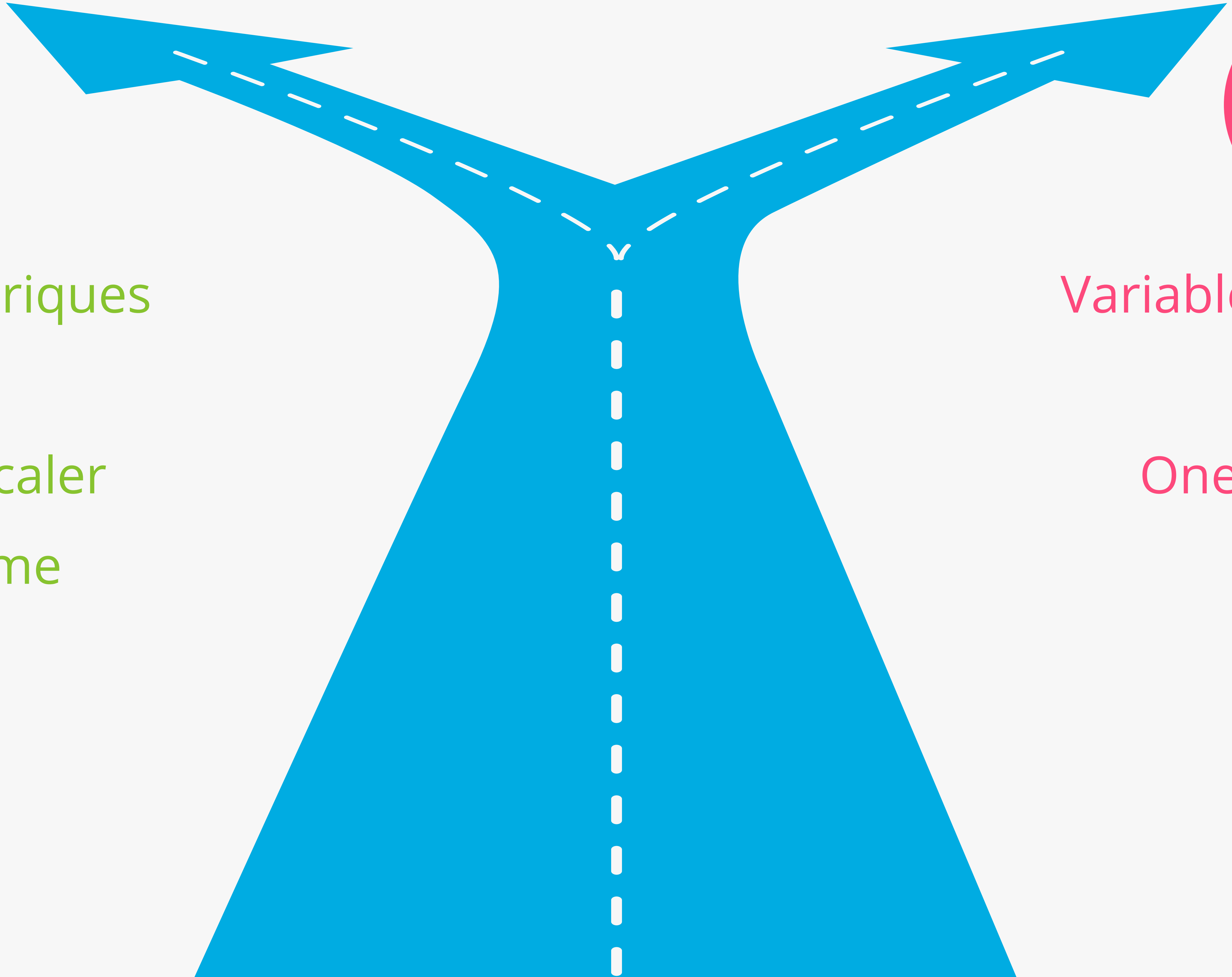
Piste de modélisation

Machine Learning



Régression

Split de la data



01

Variables numériques

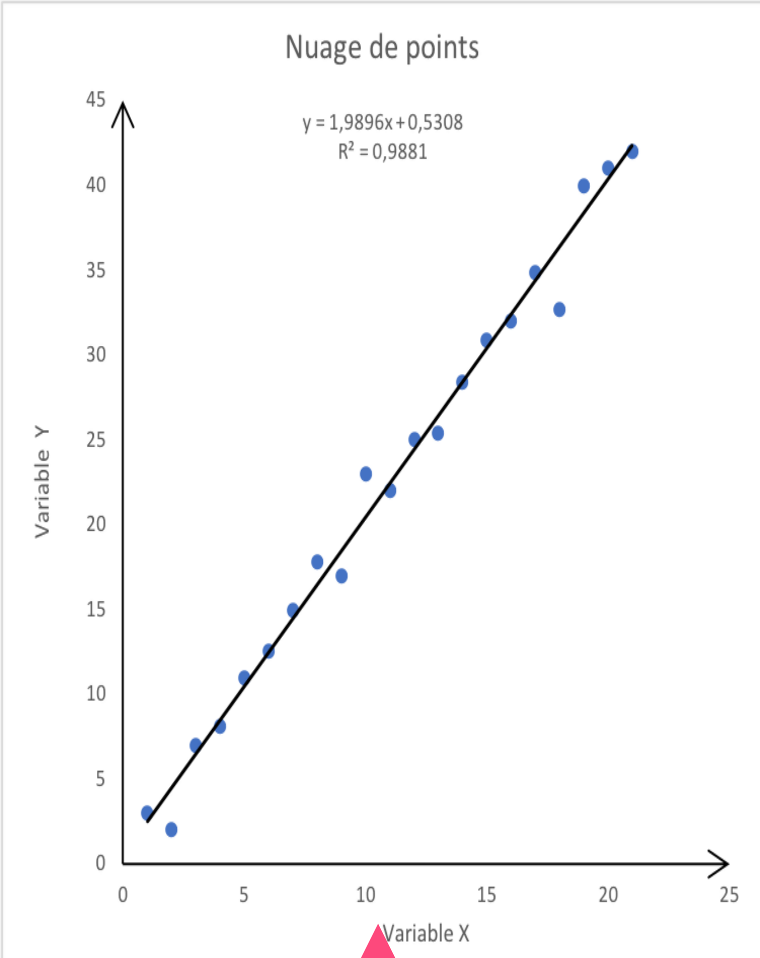
X : StandardScaler

Y : Logarithme

02

Variables catégoriques

OneHotEncoder



Régression linéaire

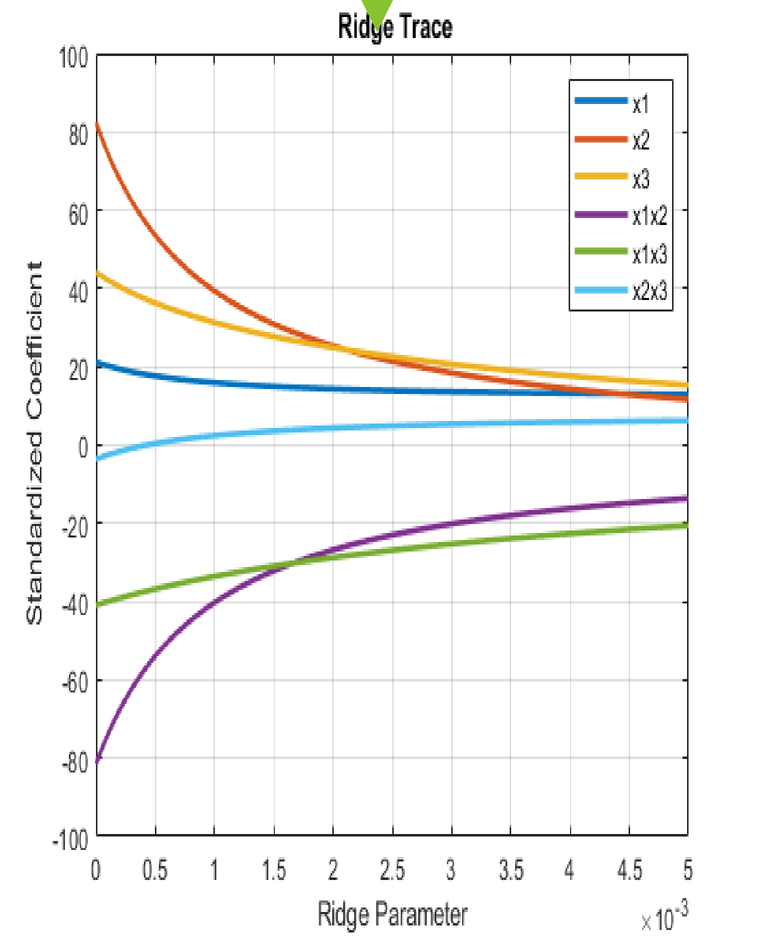
Trouver une fonction linéaire pour trouver y en fonction de x

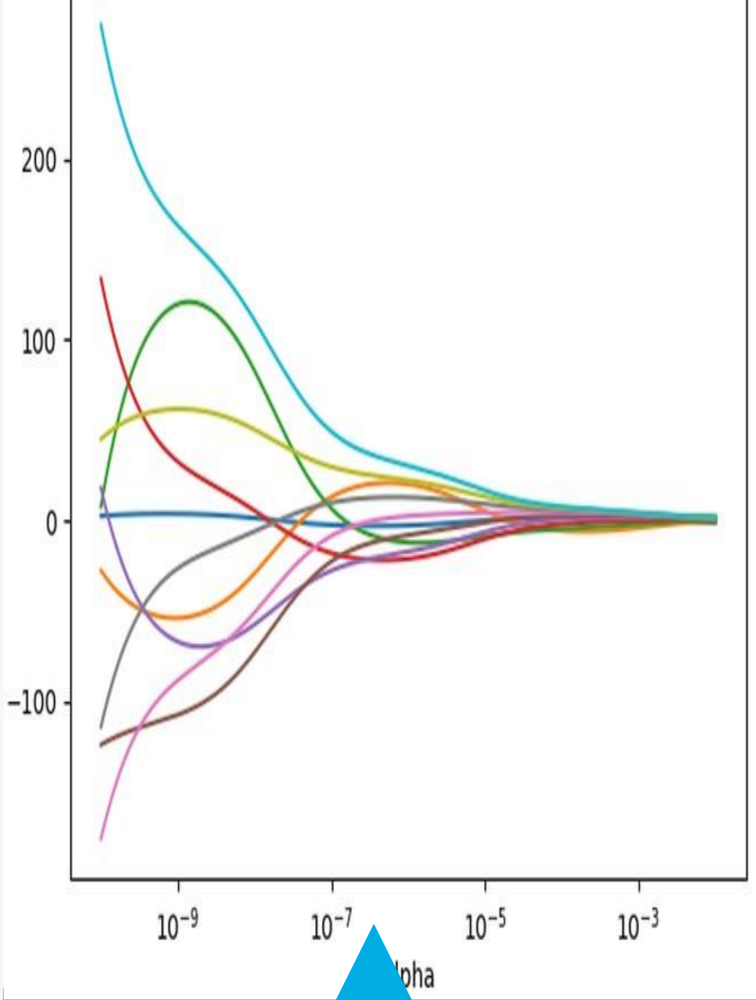
Limites : Instable avec Corrélation

Ridge

Grouper les variables corrélés

Solution unique





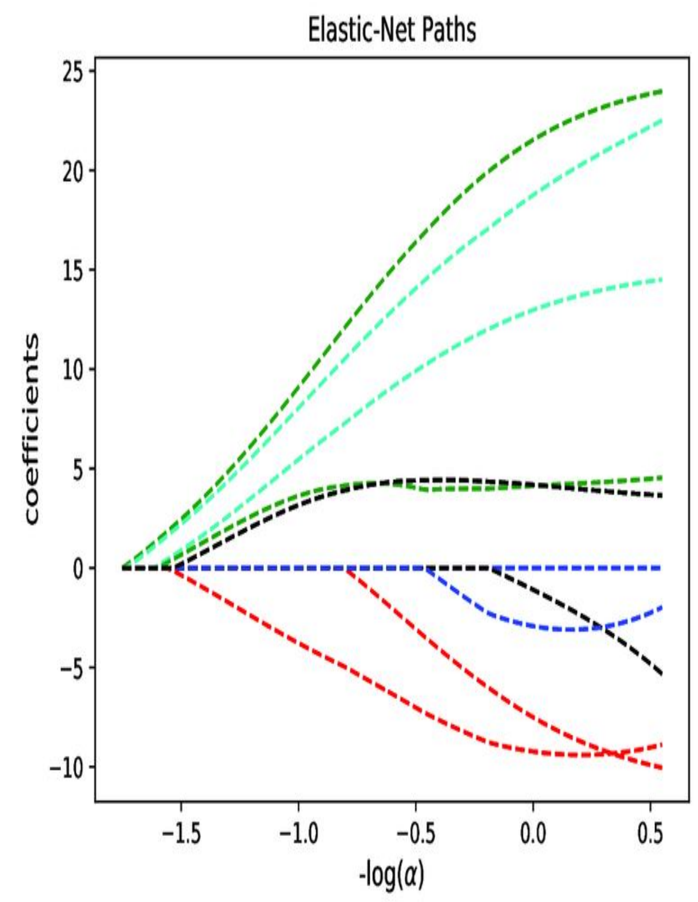
Lasso

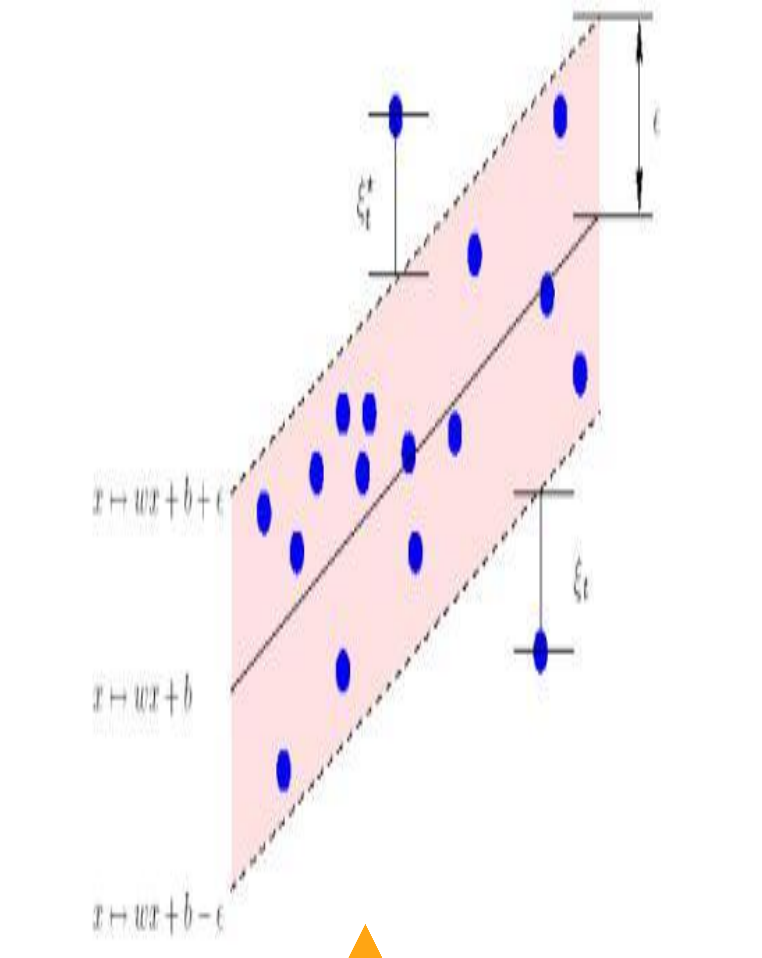
Annule le coefficient des variables inutiles

Limites : Selection aléatoire

Elastic Net

Regroupe Ridge et Lasso





SVR

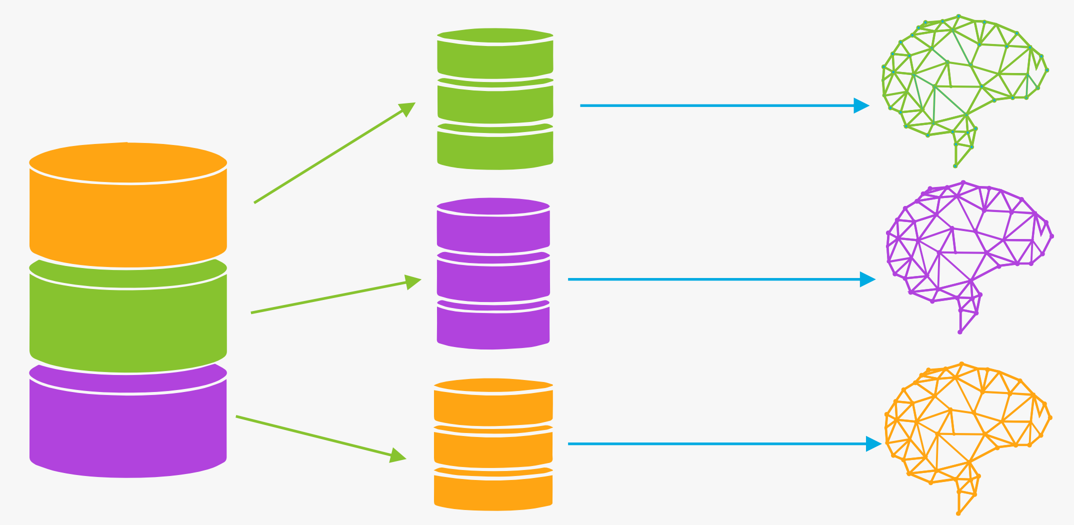
Espace à plusieurs dimensions

Se focalise sur la marge d'erreur, plutôt qu'à la réduire

Régression linéaire et modèles régularisées

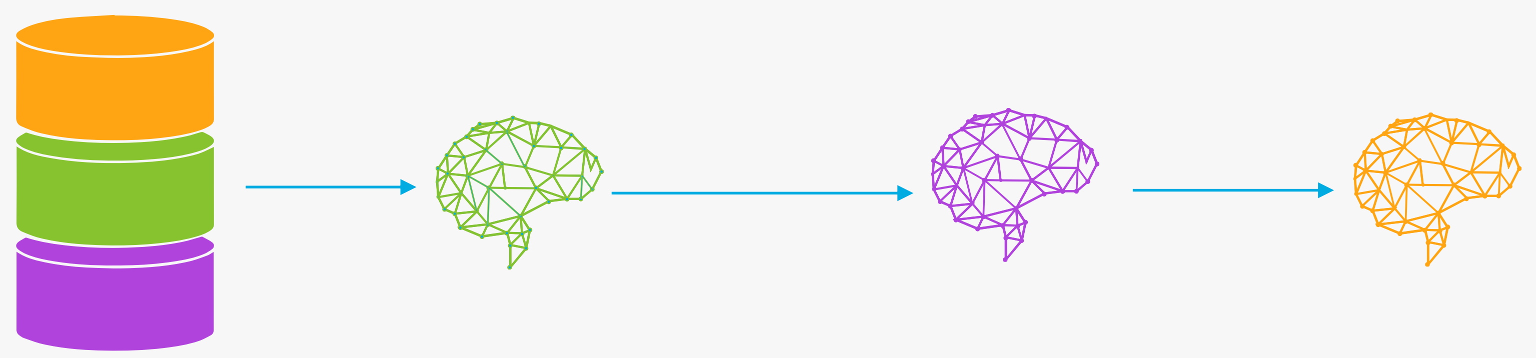
Modèles ensemblistes

BAGGING



- Créer plusieurs copies avec une partie aléatoire du dataset
- Création d'un ensemble de modèles
- Exemple : RandomForest
- Processus parallèle

BOOSTING



- Entrainer des modèles faibles successivement
- Combiner les modèles pour en obtenir un meilleur
- Exemple : XGBoosting
- Processus en série

Process

6 steps

Apprentissage

Score d’entrainement

Cross validation

Plusieurs tests…

MAE

Différence entre les valeurs mesurés et la vraie valeur

MSE

Moyenne de la distance au carré entre les valeurs prédites et vraies valeurs

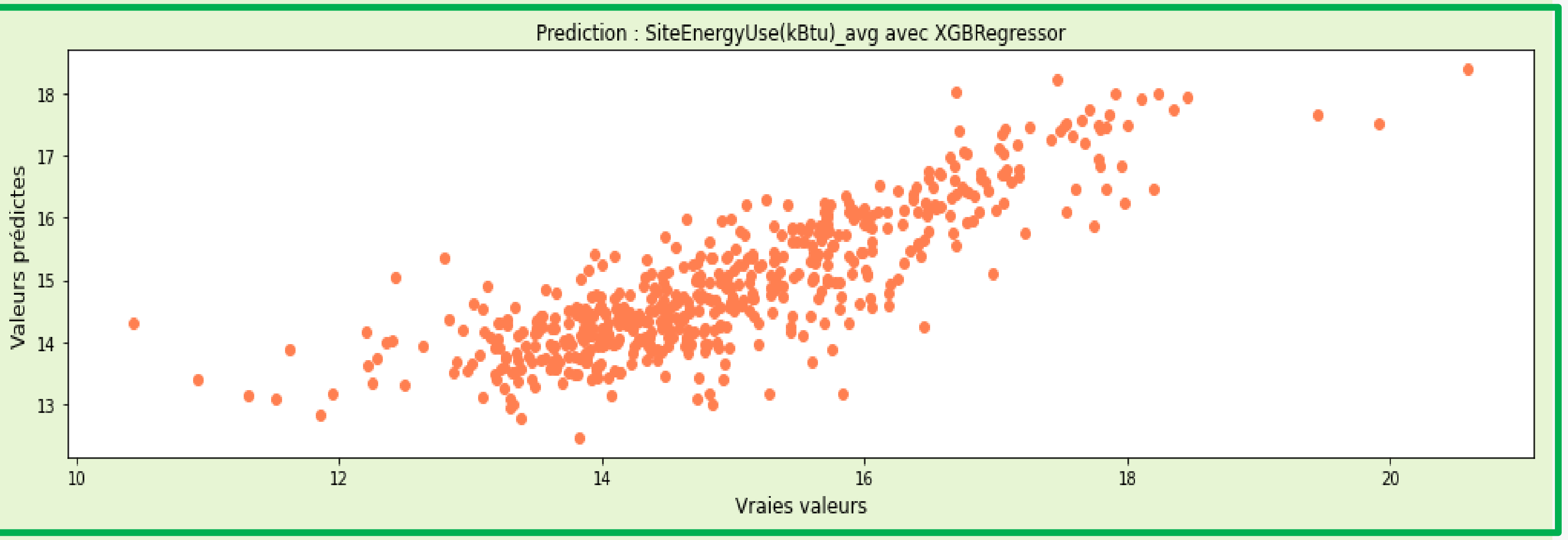
RMSE

Racine carré de la MSE

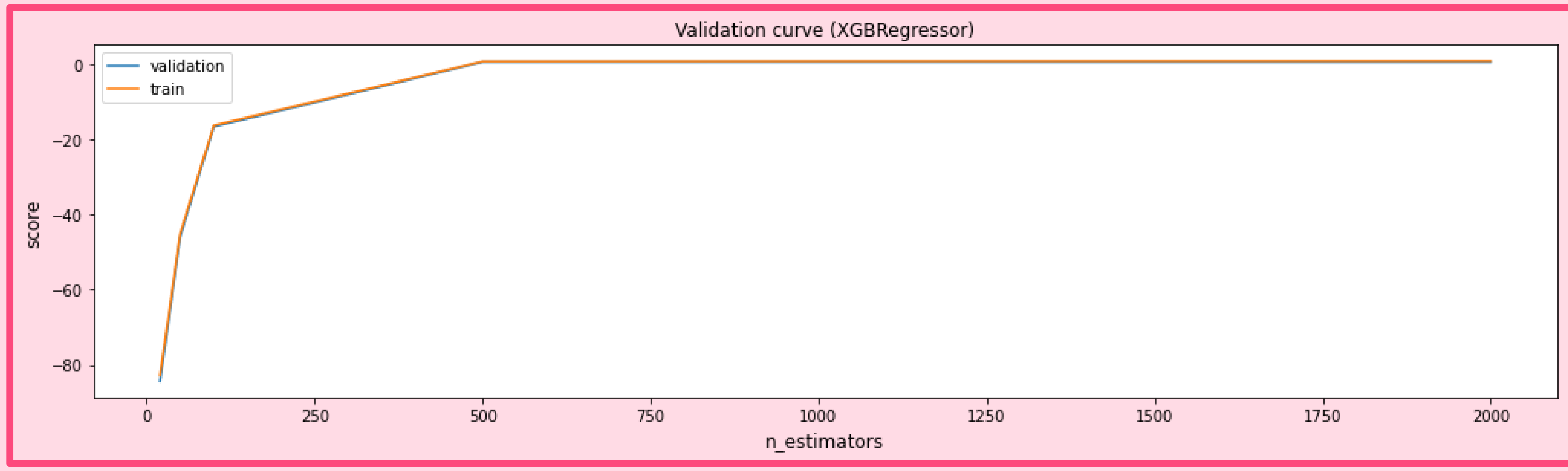
Visuels

Graphiques

Test des algorithmes



2 graphiques



Résultat

```
XGBRegressor
Score entrainement: 0.7052837776110792
Cross: 0.680862695326796
MAE : 0.5433318175982714
MSE : 0.5542381524503868
RMSE : 0.7444717270994157
```

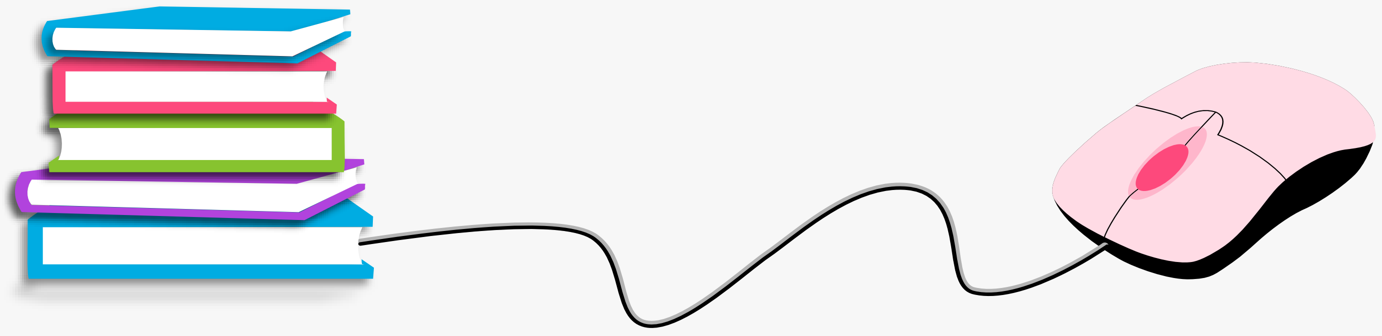
Scoring

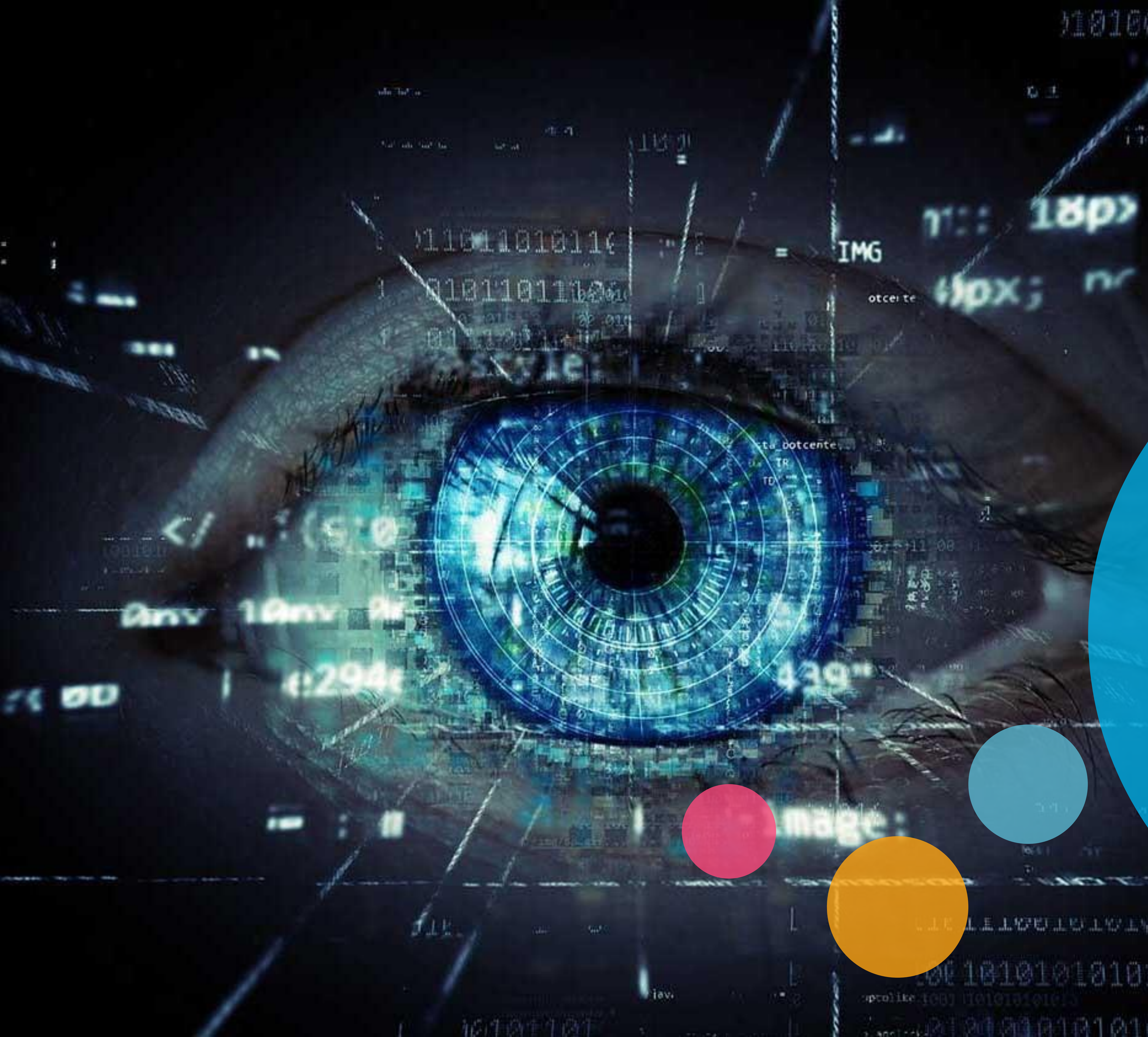
Model	Score training	Cross Validation	MAE	MSE	RMSE
LinearRegression	-4693778974802968576000.0	-5536524415656781676544.0	7954665586.506675	8827038314749600202752.0	93952319368.654221
Lasso	-0.000986	-0.001383	1.078563	1.882437	1.372019
Ridge	-0.887509	0.496017	0.781481	3.549616	1.884043
ElasticNet	0.003046	0.007333	1.077122	1.874854	1.369253
RandomForestRegressor	0.693399	0.658051	0.550178	0.576588	0.759334
XGBRegressor	0.705284	0.680863	0.543332	0.554238	0.744472
SVR	0.65869	0.623746	0.59383	0.641861	0.801162

Model	Score training	Cross Validation	MAE	MSE	RMSE
LinearRegression	-4220373226575075213312.0	-9933369119509288321024.0	8018020645.294805	8968204414529112637440.0	94700604087.456131
Lasso	-0.000231	-0.004947	1.145639	2.12547	1.457899
Ridge	-1.032397	0.381262	0.949412	4.3188	2.078172
ElasticNet	-0.000231	-0.004947	1.145639	2.12547	1.457899
RandomForestRegressor	0.532993	0.494586	0.77685	0.992381	0.996183
XGBRegressor	0.528401	0.495126	0.787816	1.002138	1.001068
SVR	0.485482	0.46662	0.809052	1.093341	1.045629

Energy

GHG





3

Modèle final

Optimisation des paramètres

1 XGBRegressor

Paramètres

N_estimators : 100 – 2000

Learning rate : 0,01 – 0,3

Max_depth : 5 – 10

N_jobs : -1

Energie : 0,681 -> 0,684

GHG : 0,496

2 RandomForest

Paramètres

N_estimators : 100 – 2000

Max_depth : 5 – 10

N_jobs : -1

Energie : 0,656

GHG : 0,492 -> 0,504

3 SVR

Paramètres

Gamma : Auto/Scale

C : 0,1 – 1

Epsilon : 0,01 – 1

Kernel : Linear, Poly, RBF

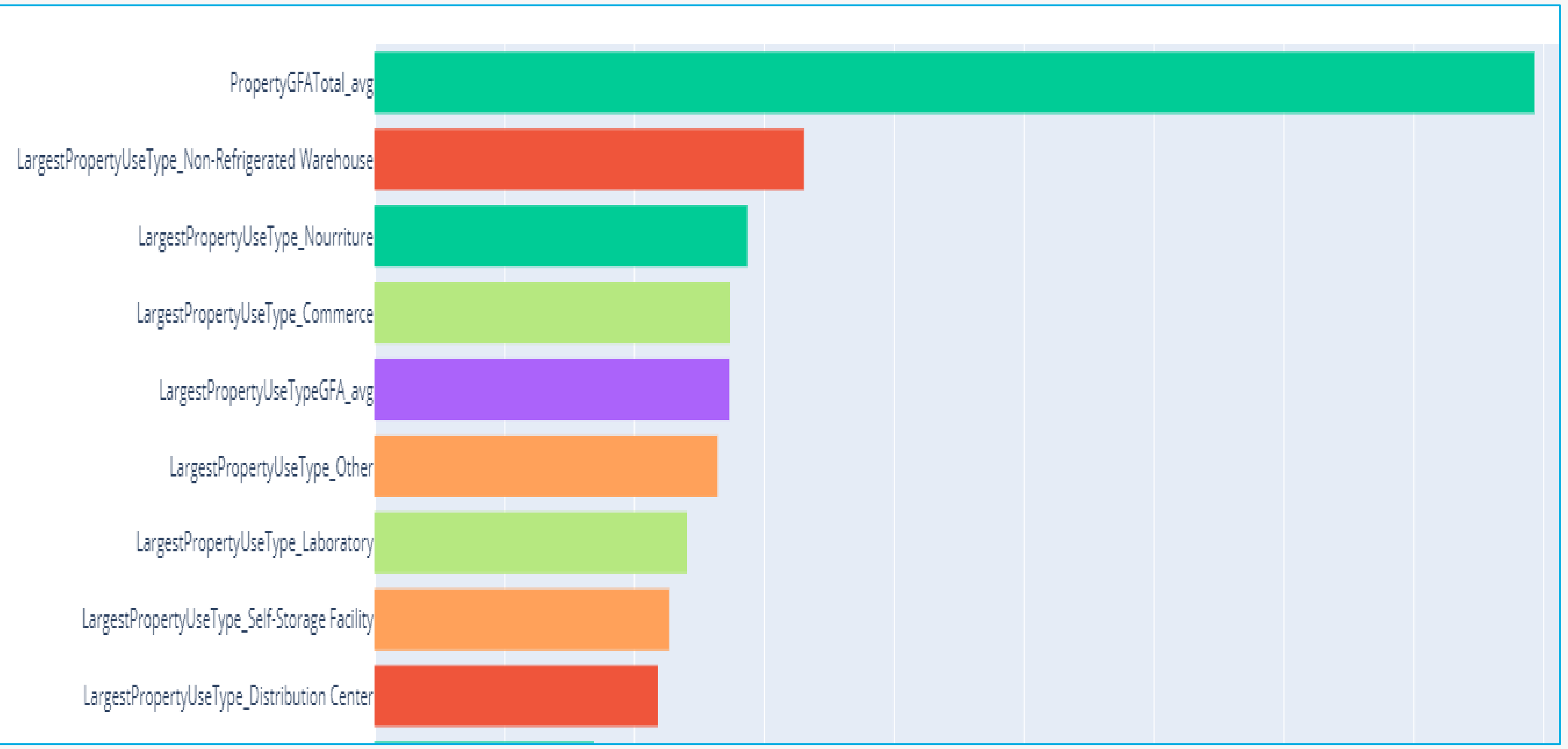
Max_inter : 100-1000

Energie : 0,62

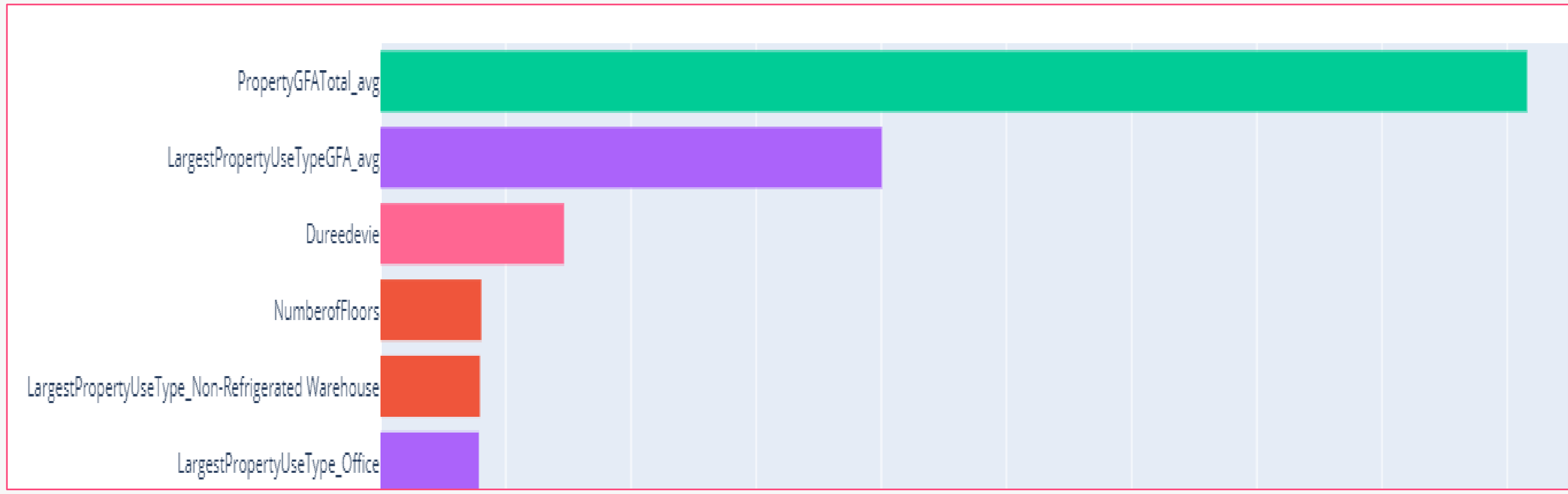
GHG : 0,467

Features **Importance**

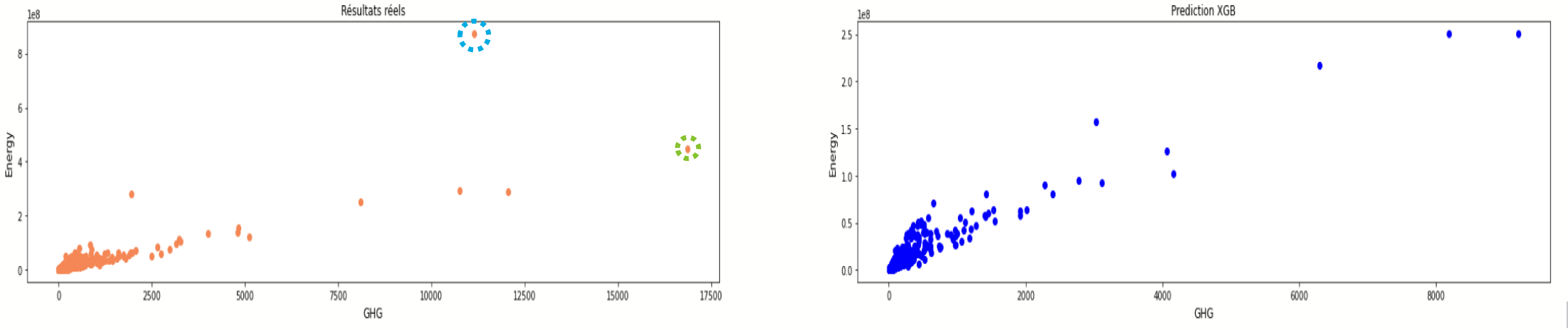
Energy



GHG



Comparaison reel / prédiction



Comparaison reels / prediction

Des differences notables...

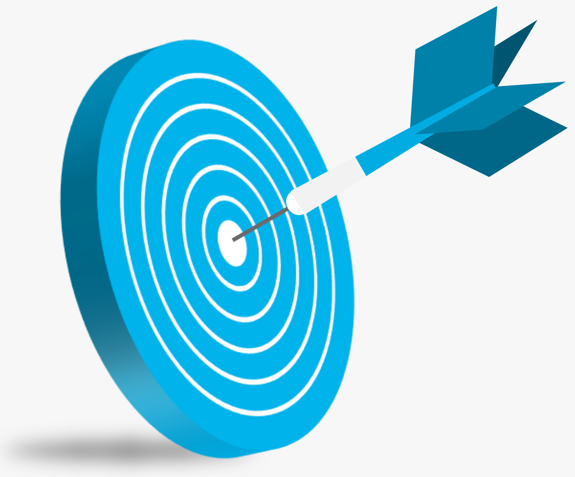
Scoring avec Energystarscore

Model	Score training	Cross Validation	MAE	MSE	RMSE
LinearRegression	0.695837	0.610591	0.546249	0.471367	0.686562
Lasso	-0.012123	-0.030673	1.003819	1.568506	1.2524
Ridge	0.690576	0.610967	0.548206	0.479521	0.692474
ElasticNet	0.235596	0.201093	0.884094	1.184611	1.088398
RandomForestRegressor	0.834737	0.810588	0.347975	0.256111	0.506074
XGBRegressor	0.853892	0.844447	0.335249	0.226426	0.475843
SVR	0.789656	0.750229	0.401018	0.325974	0.570942

Model	Score training	Cross Validation	MAE	MSE	RMSE
LinearRegression	0.528562	0.417705	0.77189	0.926478	0.962537
Lasso	-0.008682	-0.030533	1.107271	1.982278	1.407934
Ridge	0.524627	0.421336	0.77455	0.93421	0.966545
ElasticNet	0.131206	0.099447	1.03333	1.707367	1.306663
RandomForestRegressor	0.64888	0.602874	0.654122	0.690028	0.830679
XGBRegressor	0.632104	0.595027	0.671448	0.722994	0.850291
SVR	0.565771	0.524265	0.706768	0.853354	0.923772

Energy

Cross validation (sans):
RF : 0,66
XGB : 0,68
SVR : 0,62



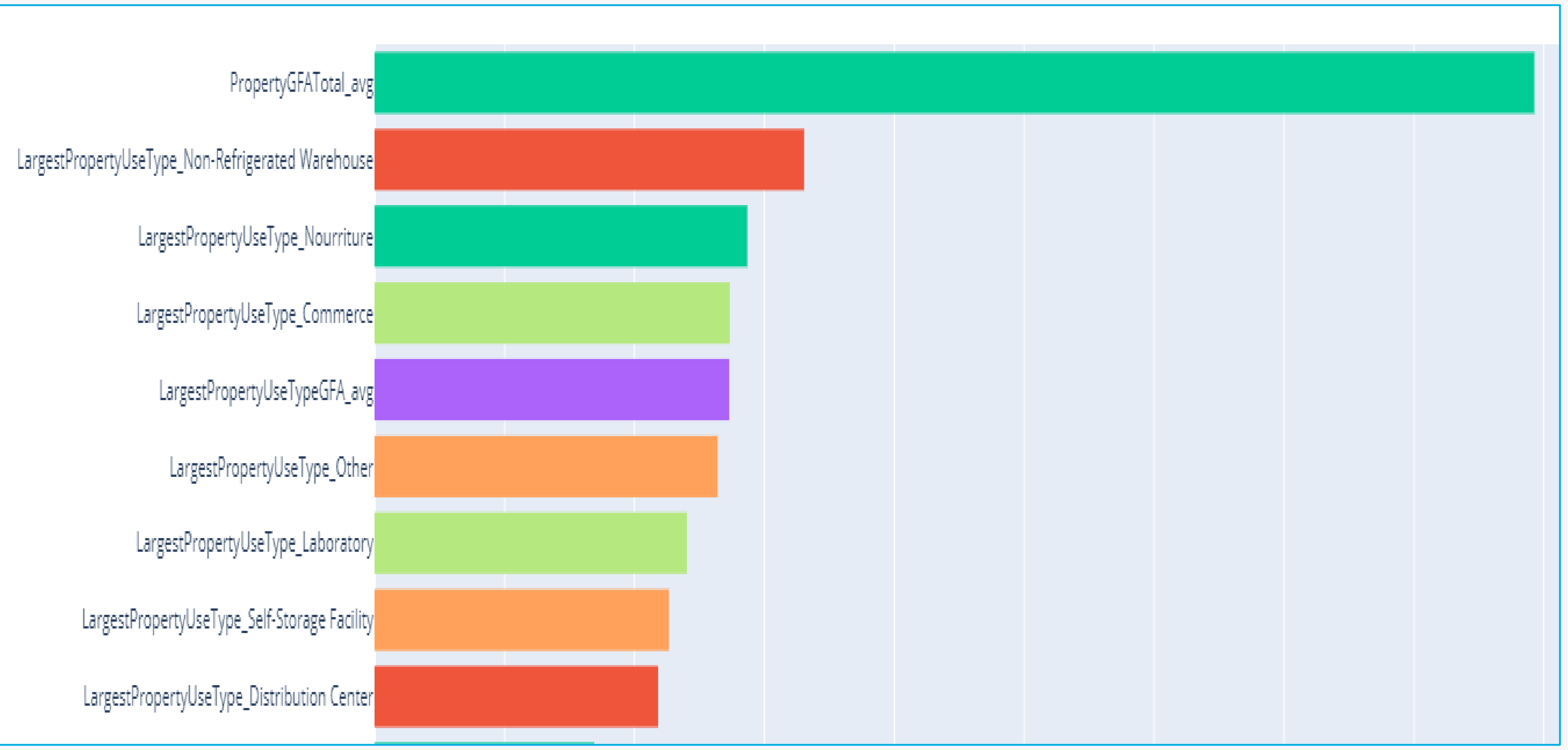
GHG

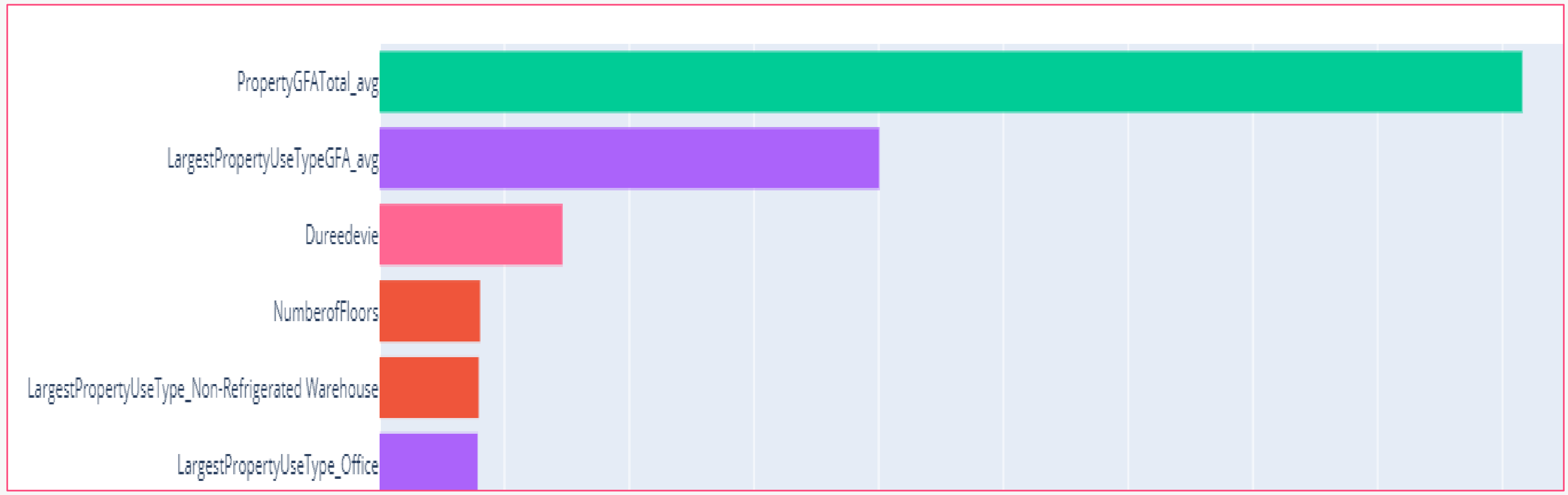
Cross validation (sans) :
RF : 0,491
XGB : 0,495
SVR : 0,46

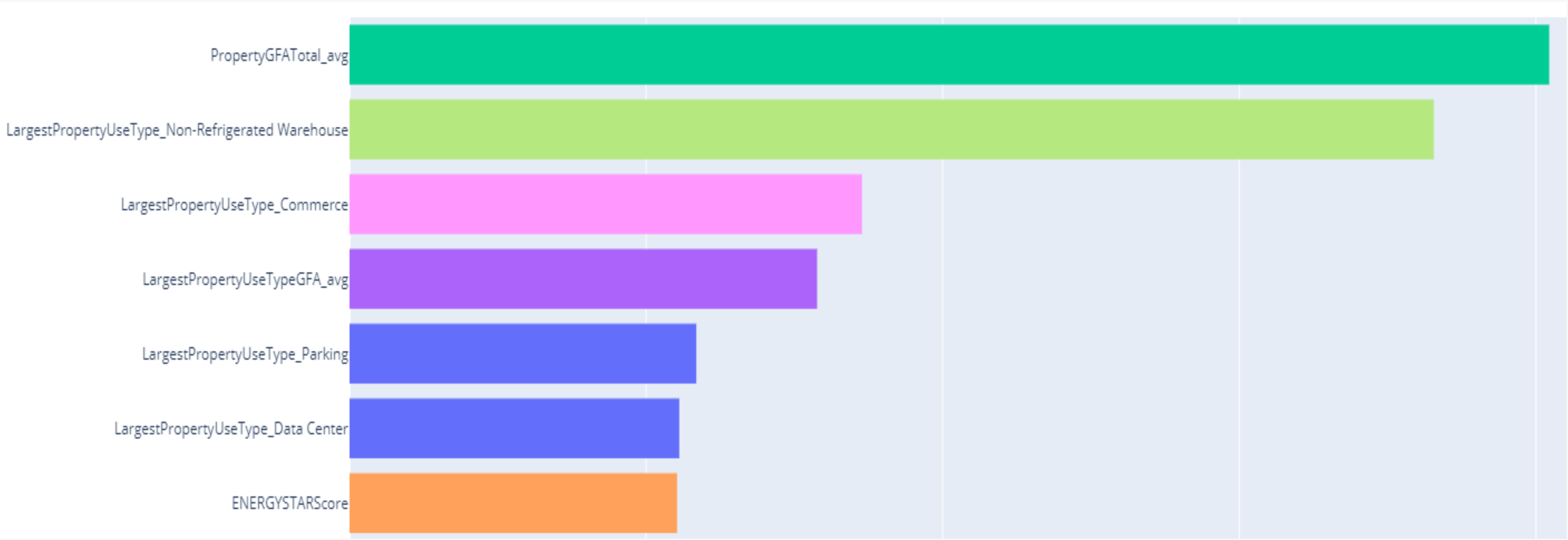
Features **Importance** (energystarscore)

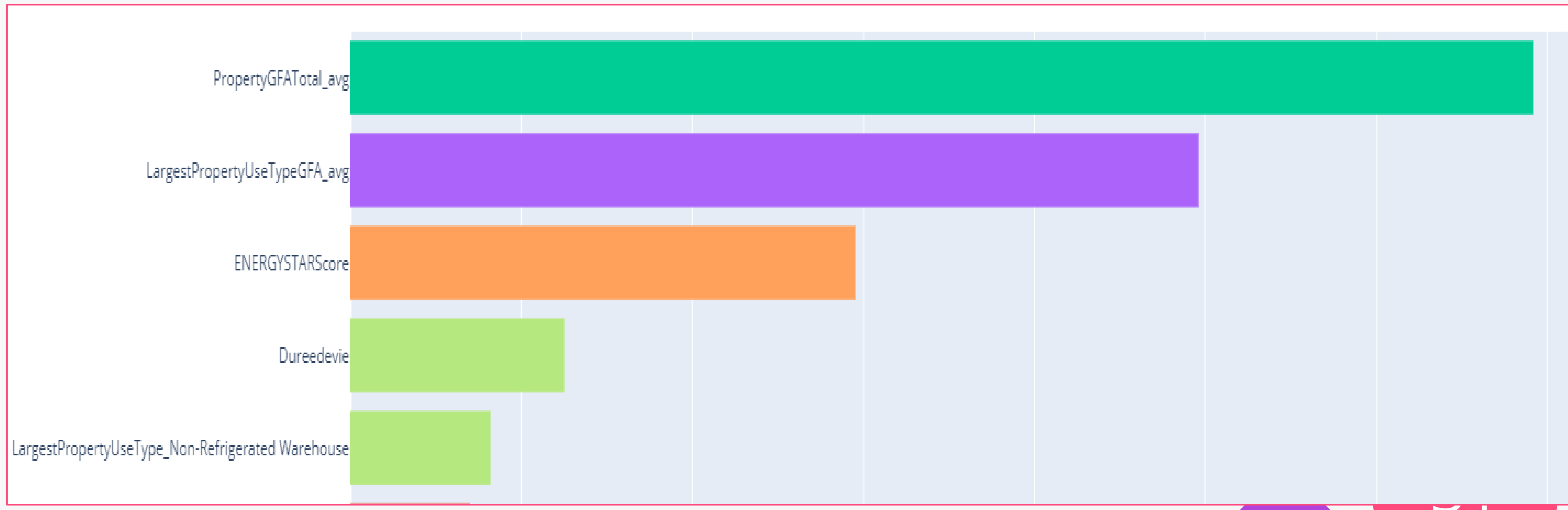
Energy

GHG

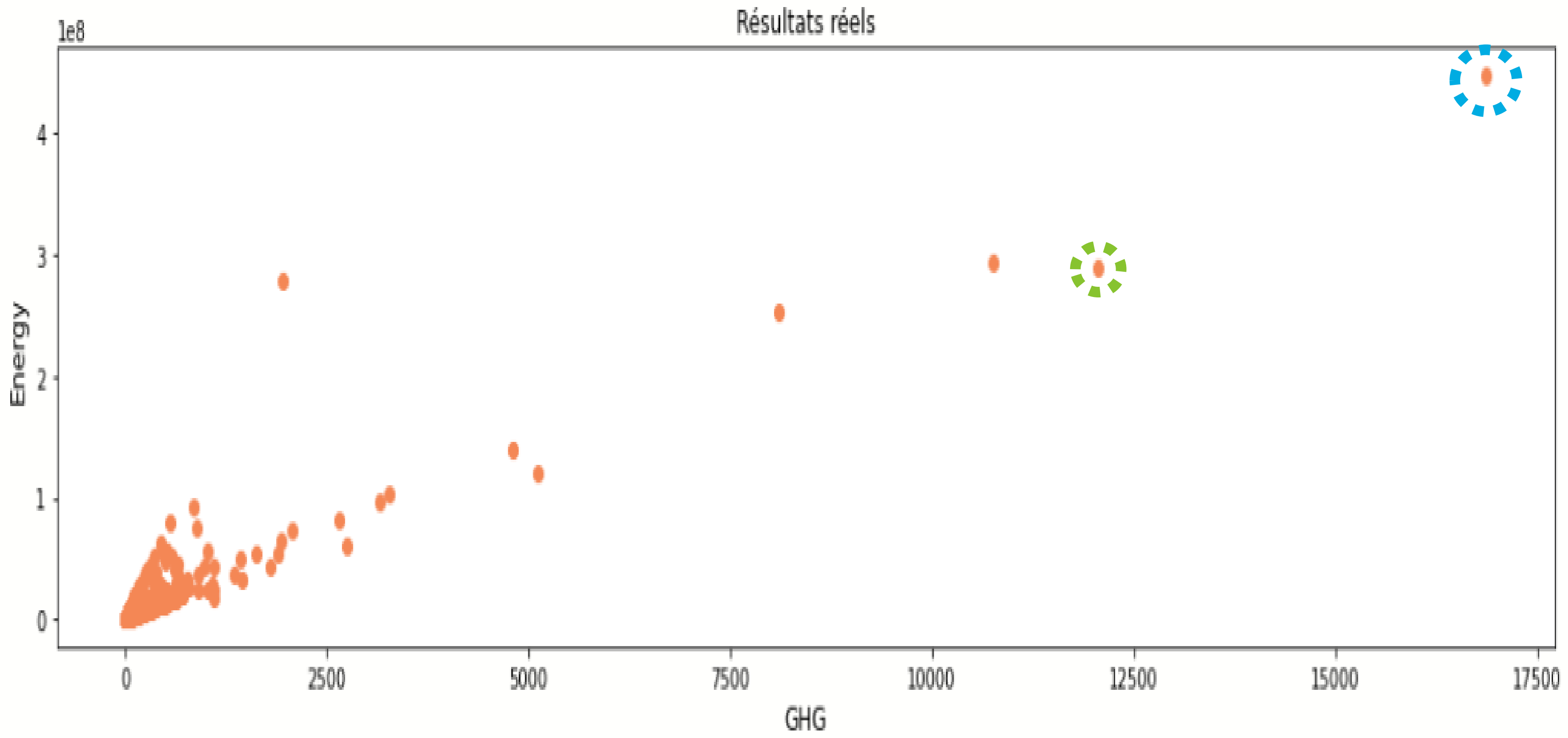


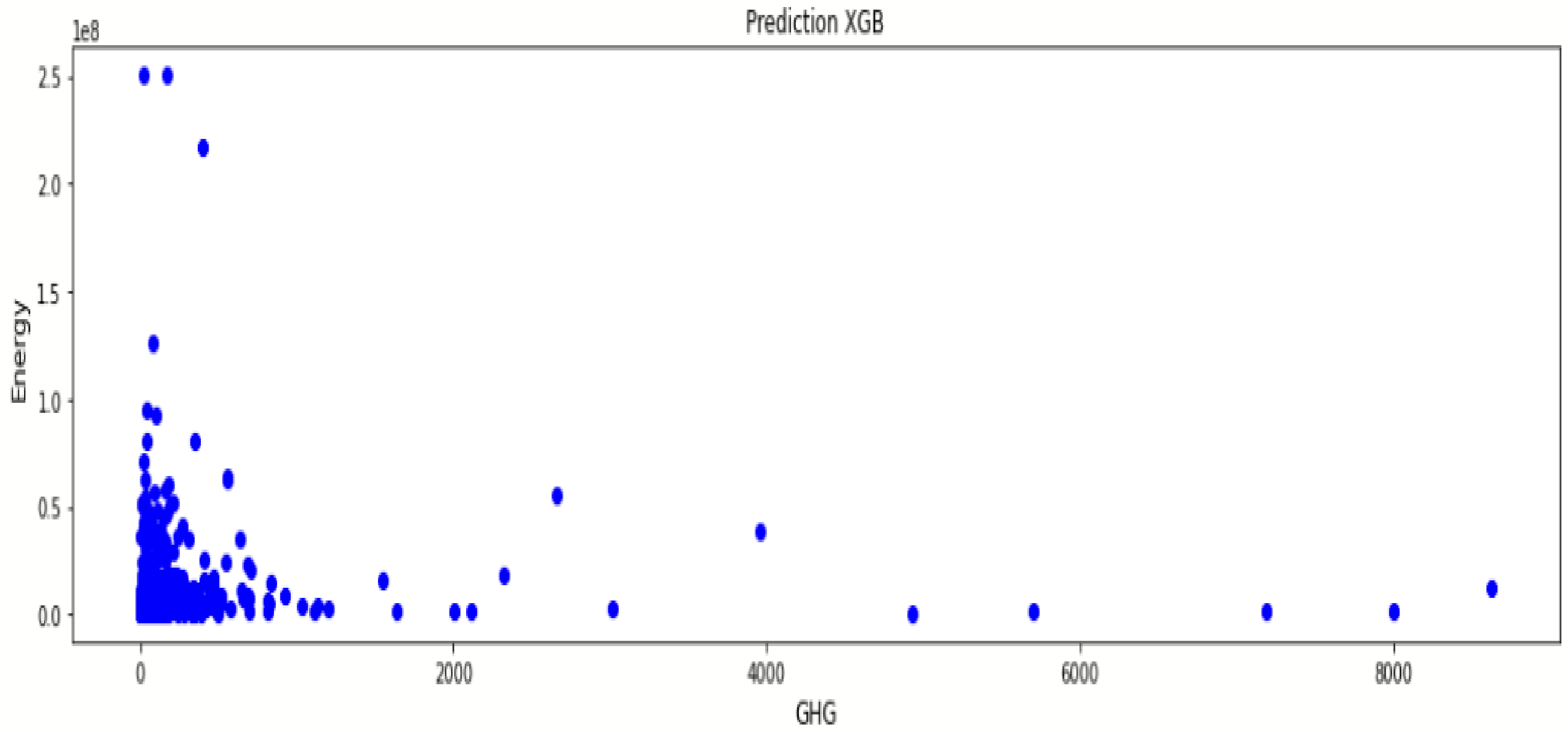






Comparaison reel / prediction (energystarscore)





Comparaison reels / prediction

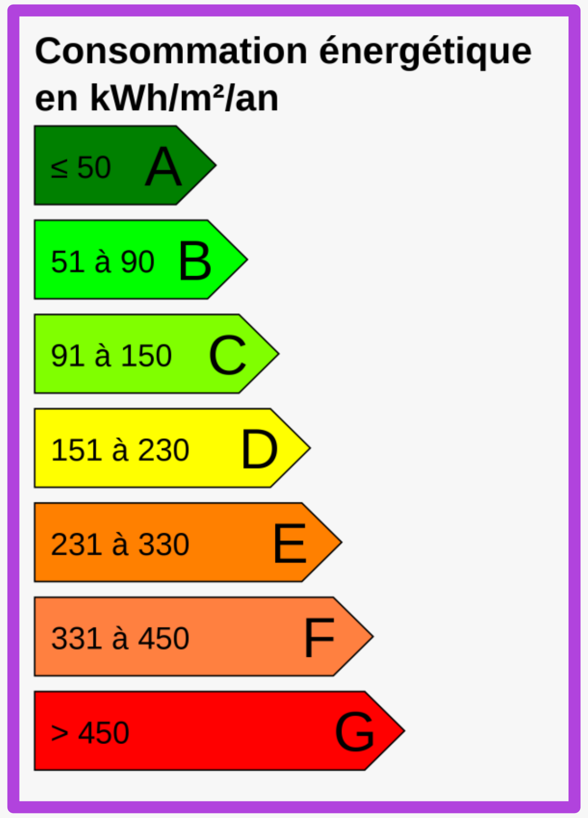
Des differences notables...

Conclusion

Modèle

Perfectible

Valeurs extrêmes



Variable Energystarscore

Très couteuse.

N’apporte pas suffisamment.