ANTICIPEZ LES BESOINS EN CONSOMMATION ÉLECTRIQUE DE BÂTIMENTS

Jeu de données de la ville de Seattle

https://www.kaggle.com/city-of-seattle/sea-building-energy-benchmarking%232015-building-energy-benchmarking.csv

Présentation par Hortense Monnard

Objectif : Une ville neutre en émissions de carbone en 2050

Variables d'intérêt :

- Consommation totale d'énergie ;
- Emissions de CO2.

Problématique :

- Prédiction des variables d'intérêt pour tous les bâtiments non-résidentiels à partir des données déclaratives du permis d'exploitation commerciale, en basant un modèle de prédiction sur les relevés 2015 et 2016 ;
- Evaluer l'intérêt de l'ENERGY STAR Score pour la prédiction d'émissions de CO2.

PLAN

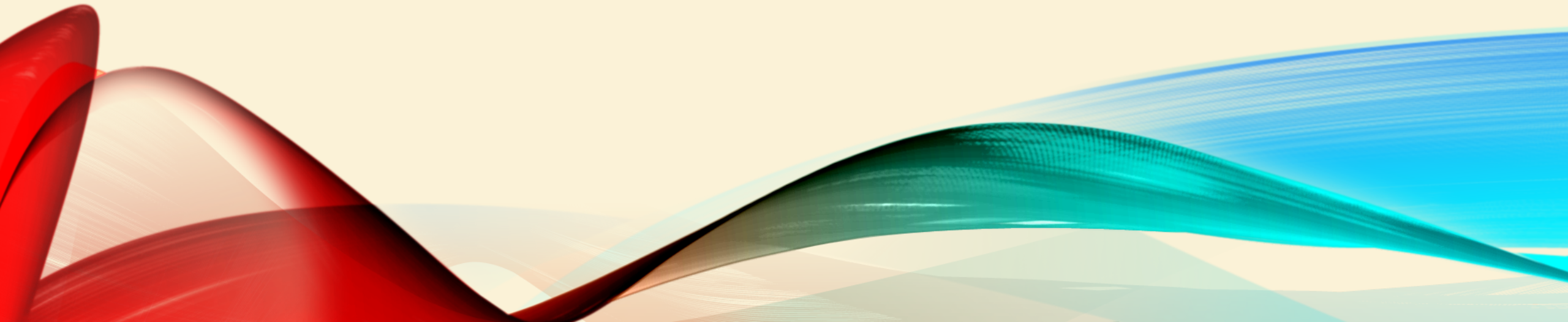
1. Présentation des Jeux de Données

2. Analyse Exploratoire
    2. a. Analyses univariées
    2. b. Analyses bivariées

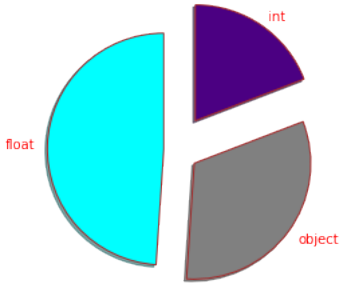
3. Modèles de Prédiction
    3. a. Sélection des features et des modèles
    3. b. Sans prise en compte de l'ENERGYSTARScore
    3. c. Avec prise en compte de l'ENERGYSTARScore

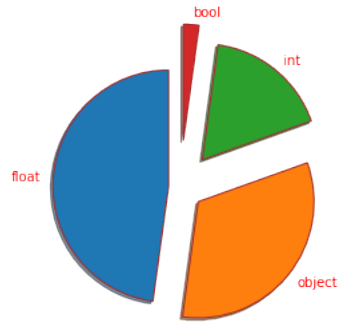
4. Conclusions et Perspectives

1. Présentation des Jeux de Données



Pré-nettoyage



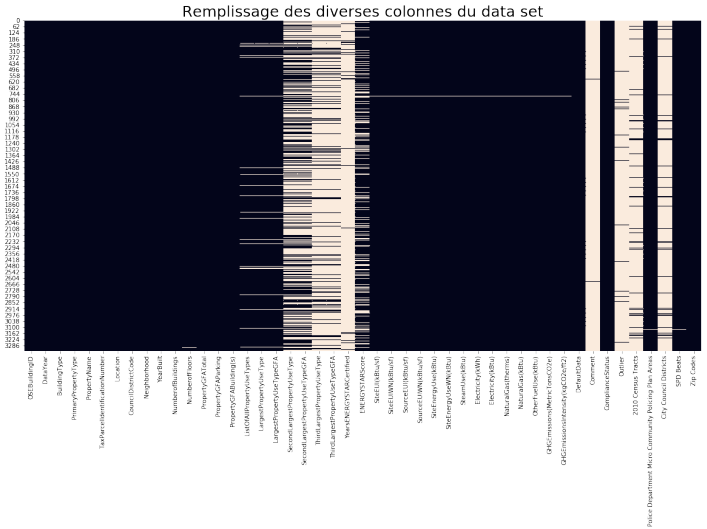


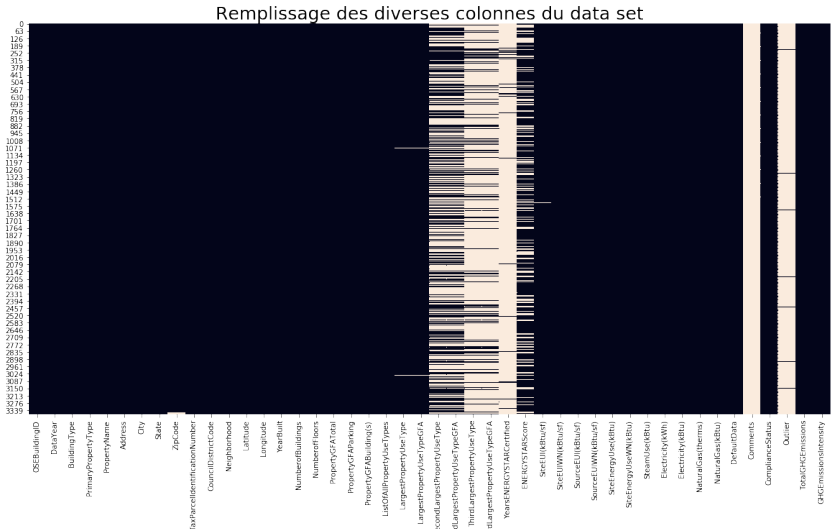
Dimension des DS :

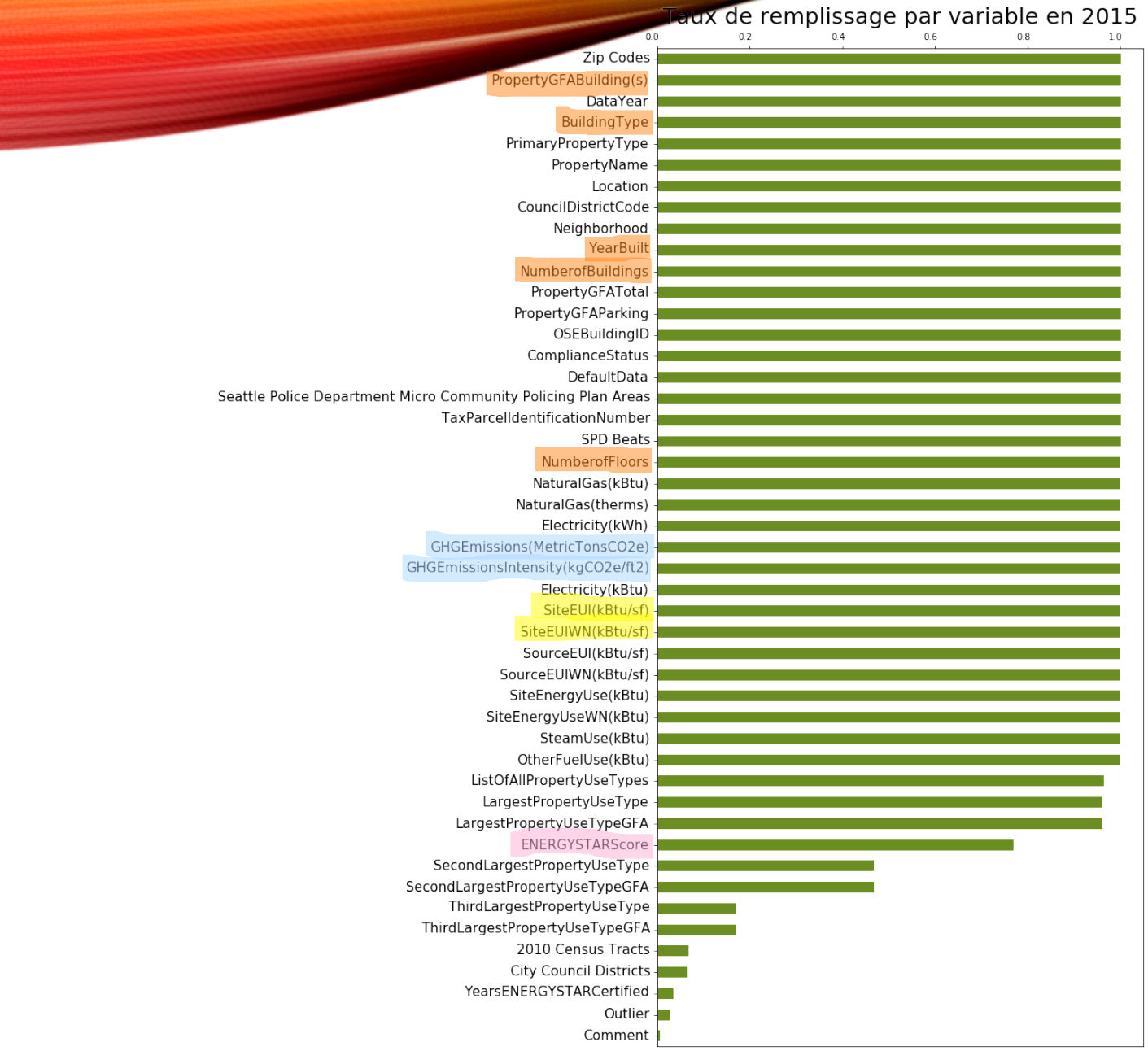
2015 : 3340 lignes, 47 colonnes.

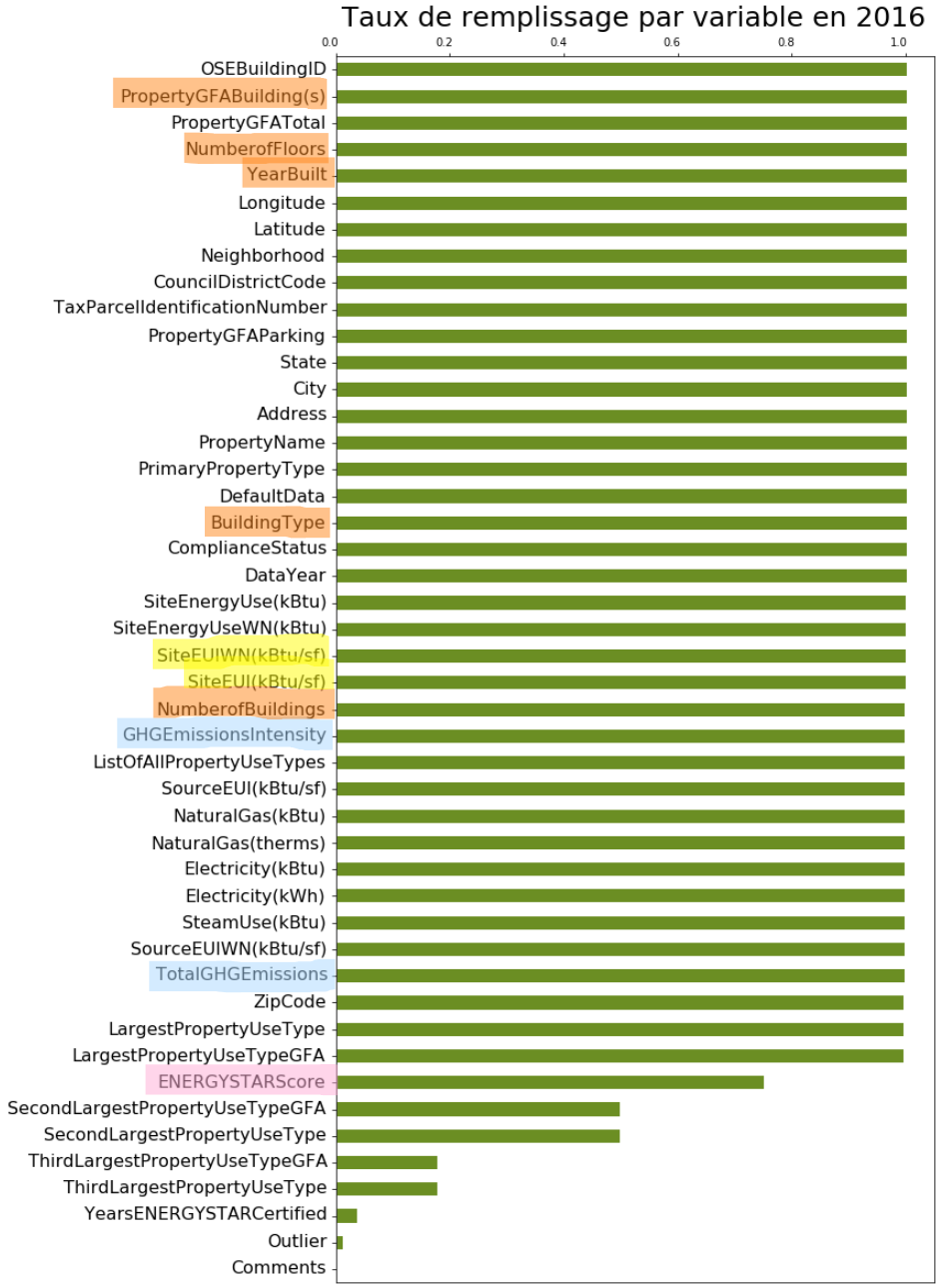
2016 : 3376 lignes, 46 colonnes.

Existe des colonnes avec des noms différents : pour une target notamment
Besoin de vérifier comment se comporte les données par des analyses univariées





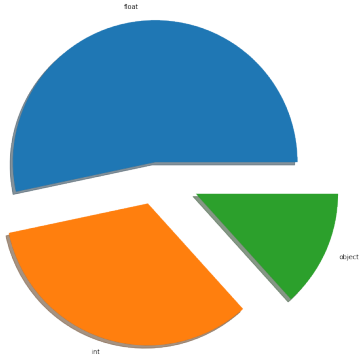

Taux de remplissage par variable en 2015
0.0
0.2
0.4
0.6
0.8
1.0
Zip Codes
PropertyGFABuilding(s)
DataYear
BuildingType
PrimaryPropertyType
PropertyName
Location
CouncilDistrictCode
Neighborhood
YearBuilt
NumberofBuildings
PropertyGFATotal
PropertyGFAParking
OSEBuildingID
ComplianceStatus
DefaultData
Seattle Police Department Micro Community Policing Plan Areas
TaxParcelIdentificationNumber
SPD Beats
NumberofFloors
NaturalGas(kBtu)
NaturalGas(therms)
Electricity(kWh)
GHGEmissions(MetricTonsCO2e)
GHGEmissionsIntensity(kgCO2e/ft2)
Electricity(kBtu)
SiteEUI(kBtu/sf)
SiteEUIWN(kBtu/sf)
SourceEUI(kBtu/sf)
SourceEUIWN(kBtu/sf)
SiteEnergyUse(kBtu)
SiteEnergyUseWN(kBtu)
SteamUse(kBtu)
OtherFuelUse(kBtu)
ListOfAllPropertyUseTypes
LargestPropertyUseType
LargestPropertyUseTypeGFA
ENERGYSTARScore
SecondLargestPropertyUseType
SecondLargestPropertyUseTypeGFA
ThirdLargestPropertyUseType
ThirdLargestPropertyUseTypeGFA
2010 Census Tracts
City Council Districts
YearsENERGYSTARCertified
Outlier
Comment


Taux de remplissage par variable en 2016
0.0
0.2
0.4
0.6
0.8
1.0
OSEBuildingID
PropertyGFABuilding(s)
PropertyGFATotal
NumberofFloors
YearBuilt
Longitude
Latitude
Neighborhood
CouncilDistrictCode
TaxParcelIdentificationNumber
PropertyGFAParking
State
City
Address
PropertyName
PrimaryPropertyType
DefaultData
BuildingType
ComplianceStatus
DataYear
SiteEnergyUse(kBtu)
SiteEnergyUseWN(kBtu)
SiteEUIWN(kBtu/sf)
SiteEUI(kBtu/sf)
NumberofBuildings
GHGEmissionsIntensity
ListOfAllPropertyUseTypes
SourceEUI(kBtu/sf)
NaturalGas(kBtu)
NaturalGas(therms)
Electricity(kBtu)
Electricity(kWh)
SteamUse(kBtu)
SourceEUIWN(kBtu/sf)
TotalGHGEmissions
ZipCode
LargestPropertyUseType
LargestPropertyUseTypeGFA
ENERGYSTARScore
SecondLargestPropertyUseTypeGFA
SecondLargestPropertyUseType
ThirdLargestPropertyUseTypeGFA
ThirdLargestPropertyUseType
YearsENERGYSTARCertified
Outlier
Comments

Nettoyage

- Suppression des lignes avec toutes les données manquantes;
- Vérification de la présence de doublons;
- Filtrage des bâtiments non-résidentiels;
- Conserver les variables qui pourrait expliquer l'évolution de nos 2 variables d'intérêt :
  - Consommation totale d'énergie ;
  - Emissions de CO2.

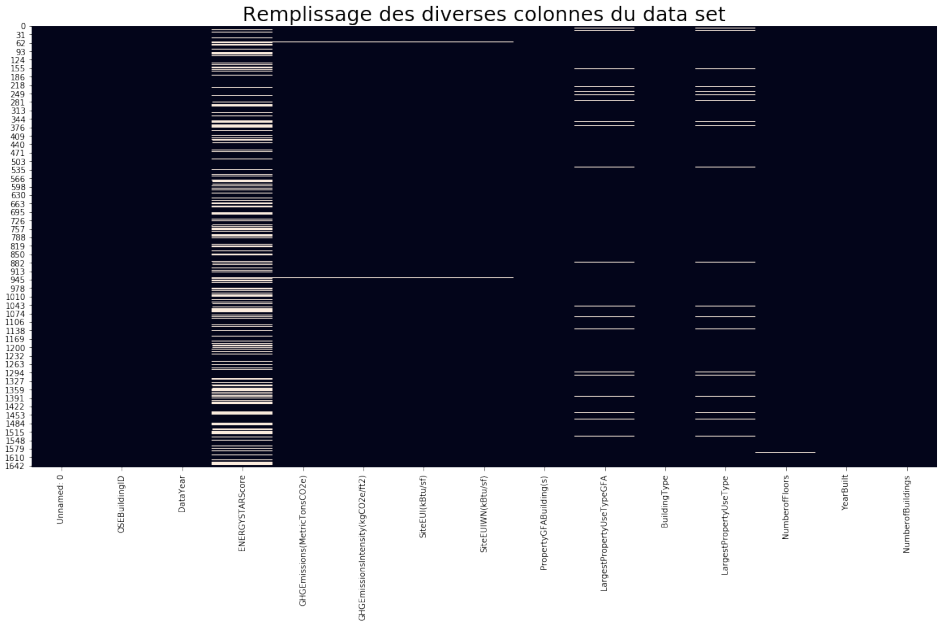
Après-nettoyage



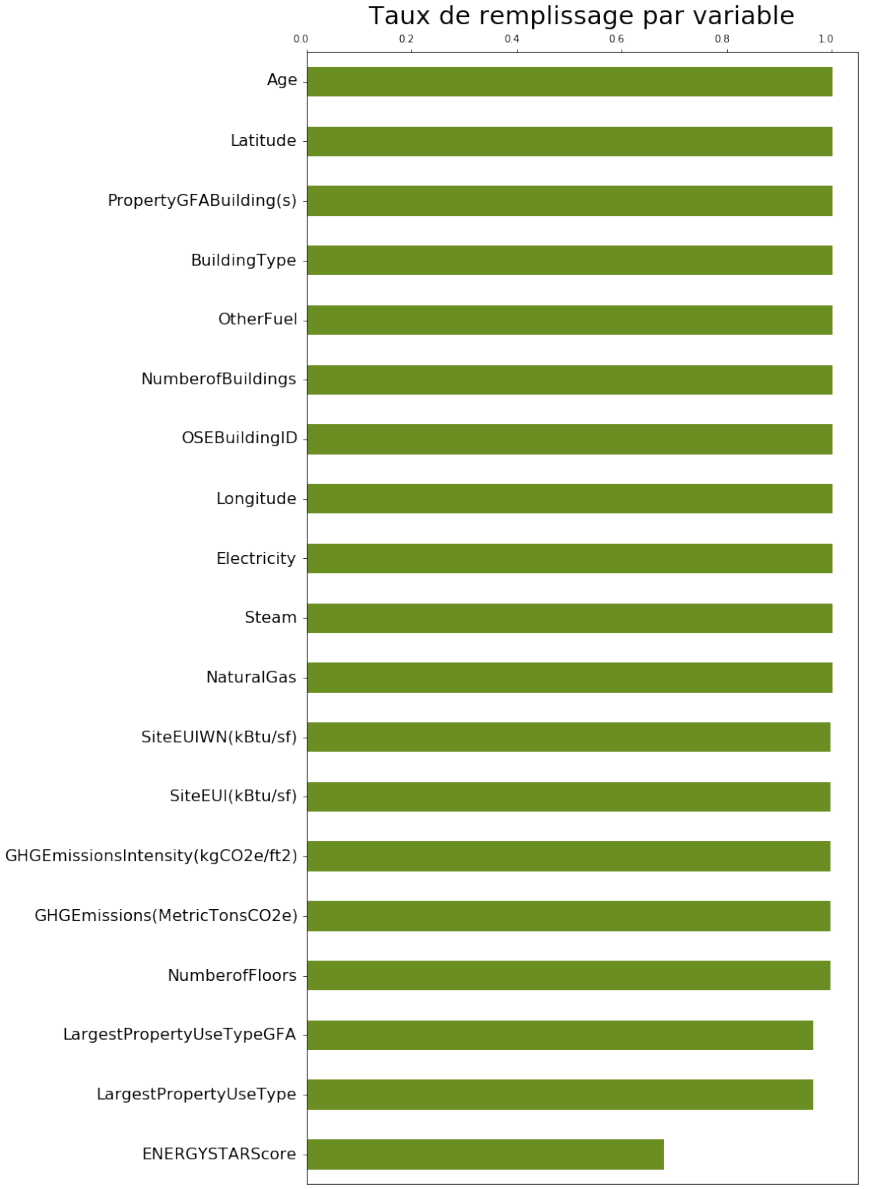
Dimension du DS : 1620 lignes, 19 colonnes

2nd nettoyage : 1547 lignes, 19 colonnes

3ème nettoyage : 1049 lignes, 19 colonnes



Variable	Taux de Remplissage
ENERGYSTARScore	0.678395
LargestPropertyUseType	0.963580
LargestPropertyUseTypeGFA	0.963580
NumberofFloors	0.995062
GHGEmissions(MetricTonsCO2e)	0.996296
GHGEmissionsIntensity(kgCO2e/ft2)	0.996296
SiteEUI(kBtu/sf)	0.996296
SiteEUIWN(kBtu/sf)	0.996296
NaturalGas	1.000000
Steam	1.000000
Electricity	1.000000
Longitude	1.000000
OSEBuildingID	1.000000
NumberofBuildings	1.000000
OtherFuel	1.000000
BuildingType	1.000000
PropertyGFABuilding(s)	1.000000
Latitude	1.000000
Age	1.000000

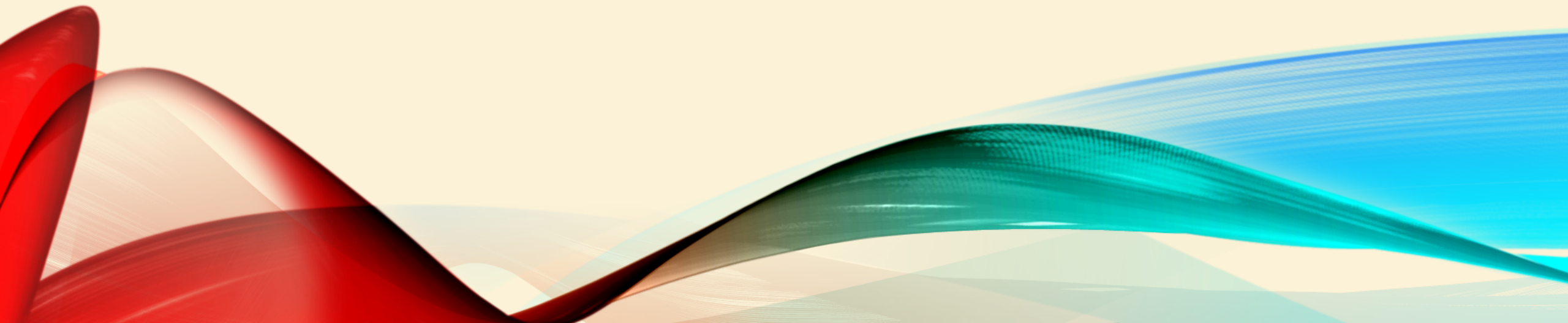


Création de Variables

Utilisation des relevés d'énergie pour créer des variables :

- Présence ou Absence d'une source d'énergie (booléen) :
  - Electricité
  - Gaz naturel
  - Vapeur
  - Autres sources

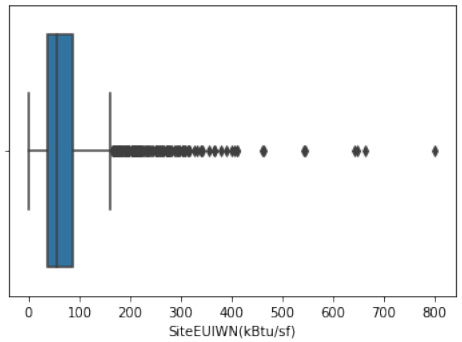
2. Analyse Exploratoire

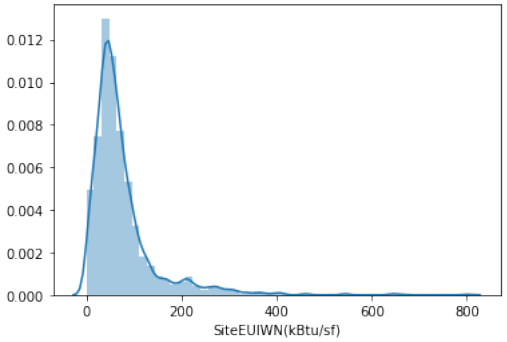


2. a. Analyses univariées

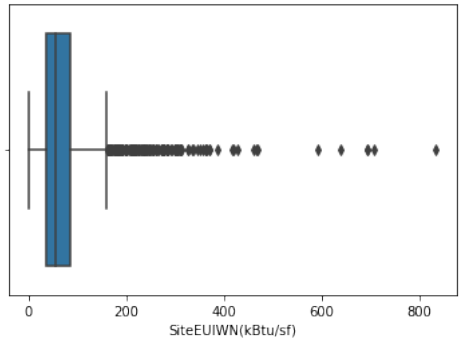
Variables Quantitatives

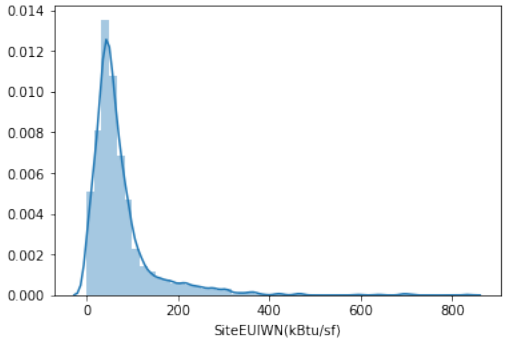
2015



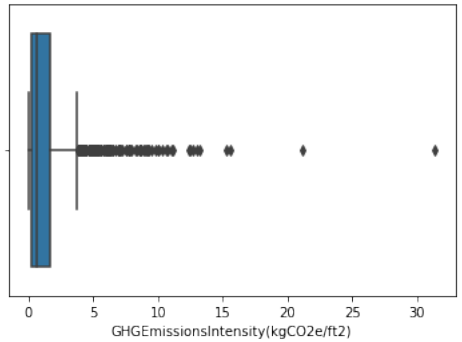


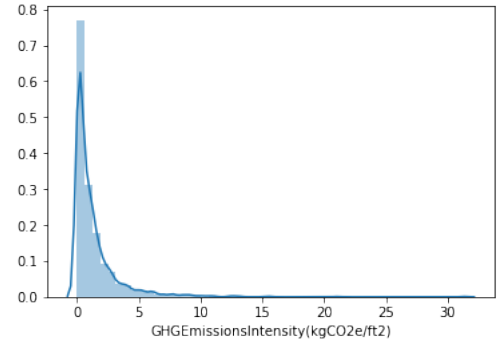
2016

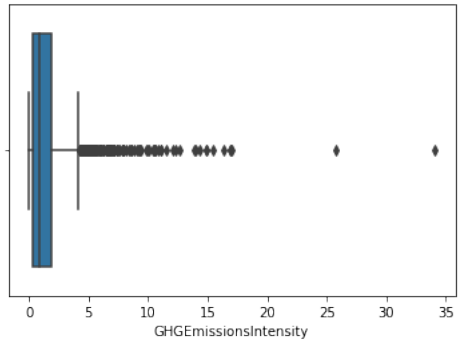


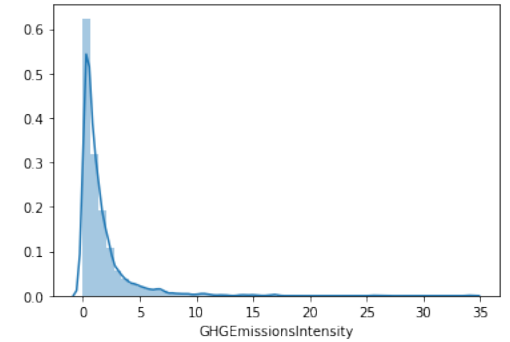


Targets





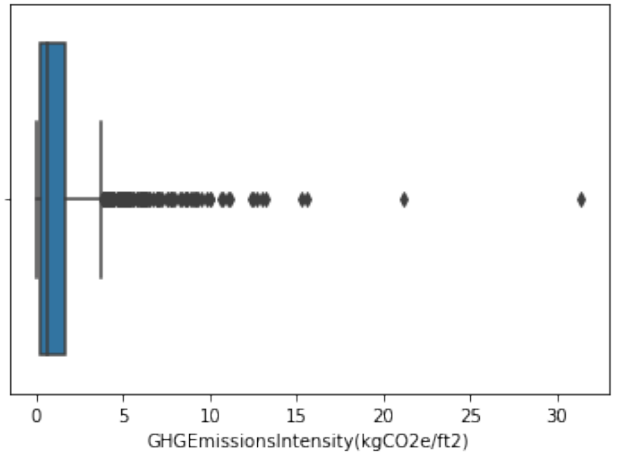


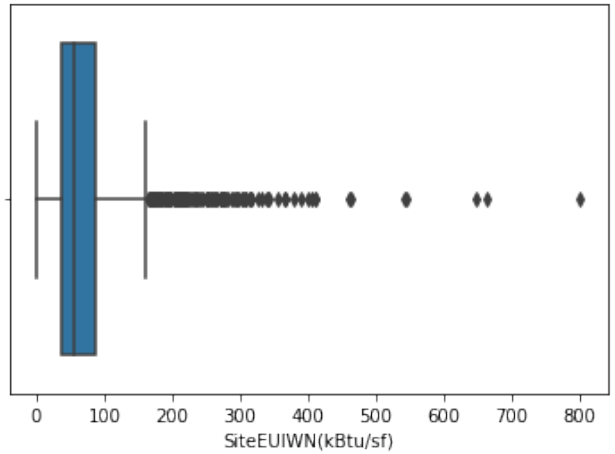


2. a. Analyses univariées

Variables Quantitatives

Targets

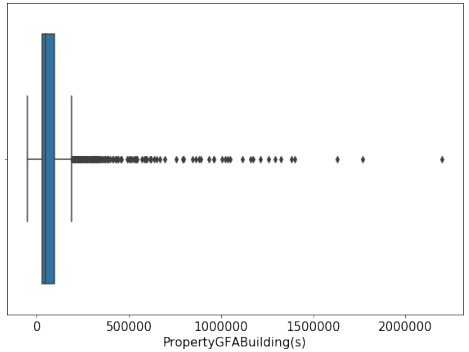


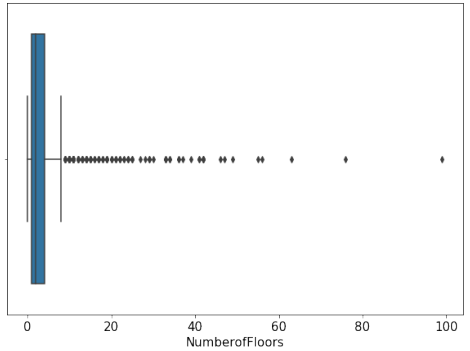


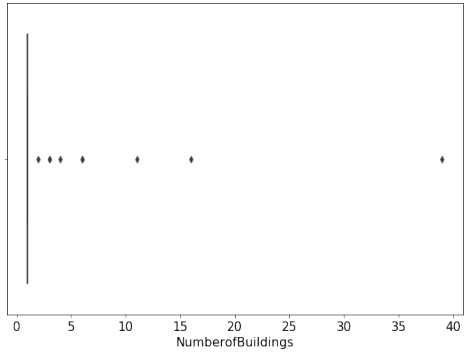
2. a. Analyses univariées

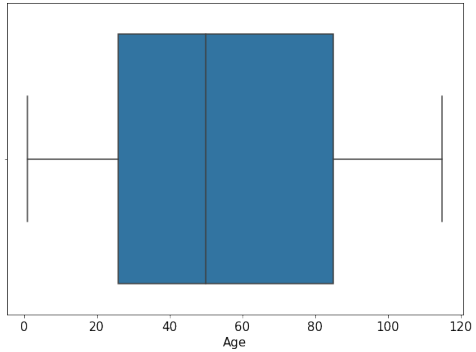
Variables Quantitatives

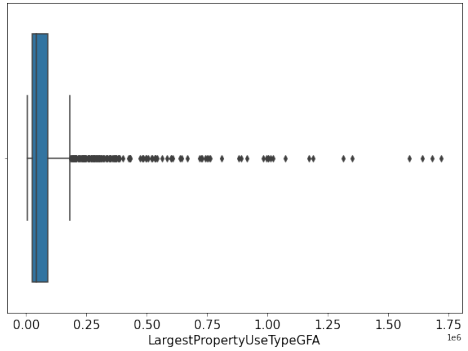
Features

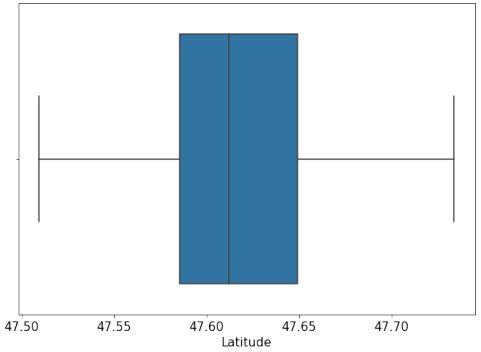


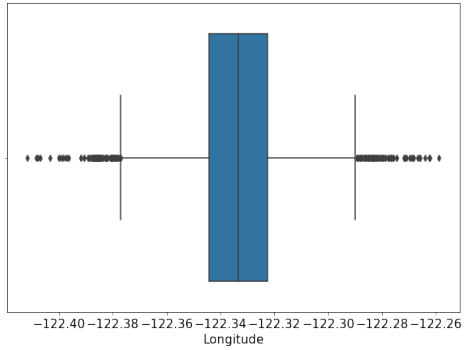












2. a. Analyses univariées

Variables Qualitatives

BuildingType	n	f	F
Campus	24	0.014388	0.014388
NonResidential	1460	0.875300	0.889688
Nonresidential COS	85	0.050959	0.940647
Nonresidential WA	1	0.000600	0.941247
SPS-District K-12	98	0.058753	1.000000

Electricity	n	f	F
0	2	0.001293	0.001293
1	1545	0.998707	1.000000

NaturalGas	n	f	F
0	436	0.281836	0.281836
1	1111	0.718164	1.000000

Steam	n	f	F
0	1442	0.932127	0.932127
1	105	0.067873	1.000000

OtherFuel	n	f	F
0	1533	0.99095	0.99095
1	14	0.00905	1.00000

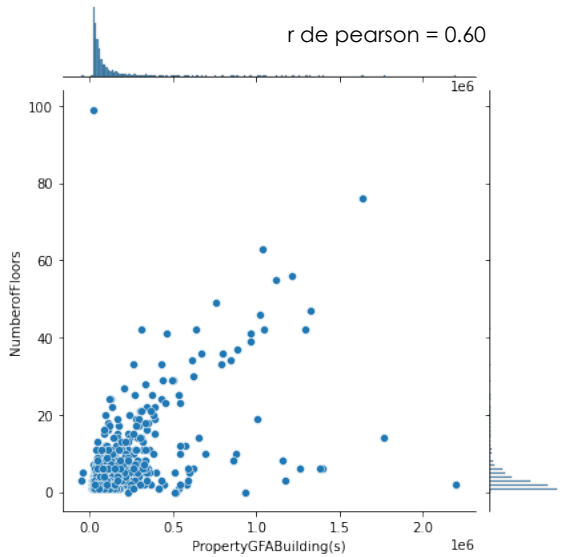
LargestPropertyUseType	n	f	F
Adult Education		20.001199	0.001199
Automobile Dealership		50.002998	0.004197
Bank Branch		40.002398	0.006595
College/University		240.014388	0.020983
Convention Center		10.000600	0.021583
Courthouse		10.000600	0.022182
Data Center		30.001799	0.023981
Distribution Center		540.032374	0.056355
Financial Office		40.002398	0.058753
Fire Station		10.000600	0.059353
Fitness Center/Health Club/Gym		50.002998	0.062350
Food Service		10.000600	0.062950
Hospital (General Medical & Surgical)		100.005995	0.068945
Hotel		760.045564	0.114508
K-12 School		1390.083333	0.197842
Laboratory		130.007794	0.205635
Library		40.002398	0.208034
Lifestyle Center		20.001199	0.209233
Manufacturing/Industrial Plant		80.004796	0.214029
Medical Office		410.024580	0.238609
Movie Theater		10.000600	0.239209
Multifamily Housing		120.007194	0.246403
Museum		50.002998	0.249400
Non-Refrigerated Warehouse		1990.119305	0.368705
Office		4970.297962	0.666667
Other		980.058753	0.725420
Other - Education		40.002398	0.727818
Other - Entertainment/Public Assembly		210.012590	0.740408
Other - Lodging/Residential		50.002998	0.743405
Other - Mall		40.002398	0.745803
Other - Public Services		20.001199	0.747002
Other - Recreation		310.018585	0.765588
Other - Restaurant/Bar		20.001199	0.766787
Other - Services		50.002998	0.769784
Other - Utility		20.001199	0.770983
Other/Specialty Hospital		40.002398	0.773381
Parking		300.017986	0.791367
Performing Arts		30.001799	0.793165
Personal Services (Health/Beauty, Dry Cleaning...)		10.000600	0.793765
Police Station		10.000600	0.794365
Pre-school/Daycare		20.001199	0.795564
Prison/Incarceration		30.001799	0.797362
Refrigerated Warehouse		120.007194	0.804556
Repair Services (Vehicle, Shoe, Locksmith, etc)		60.003597	0.808153
Residence Hall/Dormitory		220.013189	0.821343
Residential Care Facility		10.000600	0.821942
Restaurant		120.007194	0.829137
Retail Store		990.059353	0.888489
Self-Storage Facility		270.016187	0.904676
Senior Care Community		200.011990	0.916667
Social/Meeting Hall		100.005995	0.922662
Strip Mall		60.003597	0.926259
Supermarket/Grocery Store		410.024580	0.950839
Urgent Care/Clinic/Other Outpatient		40.002398	0.953237
Wholesale Club/Supercenter		10.000600	0.953837
Worship Facility		710.042566	0.996403

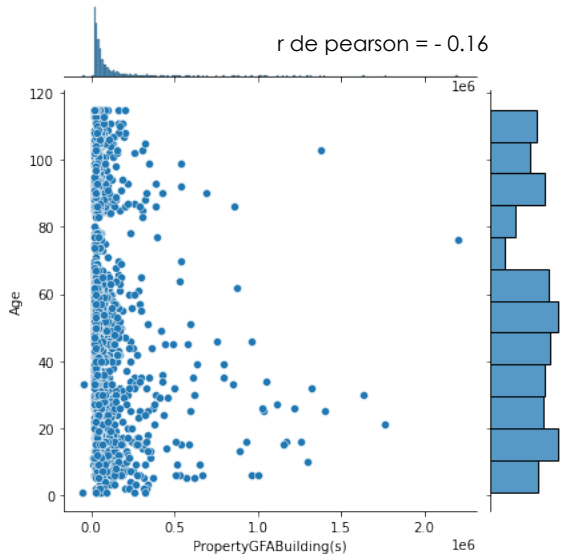
2. b. Analyses bivariées

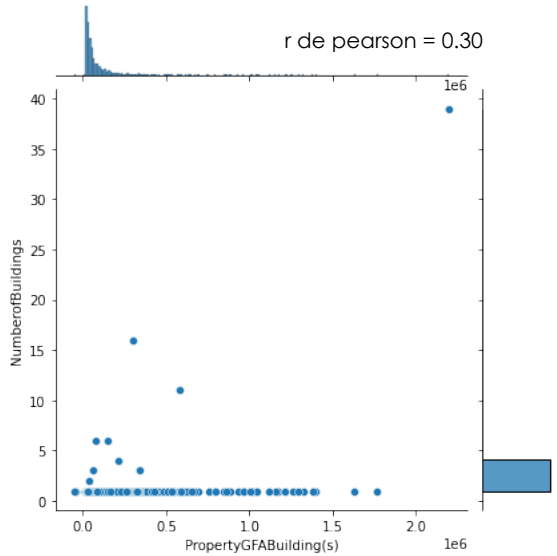
Variables Quantitatives

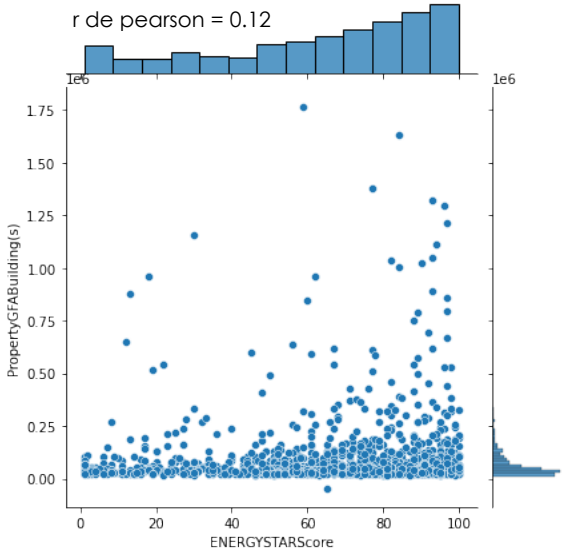
Les features suivantes sont indépendantes les unes des autres :

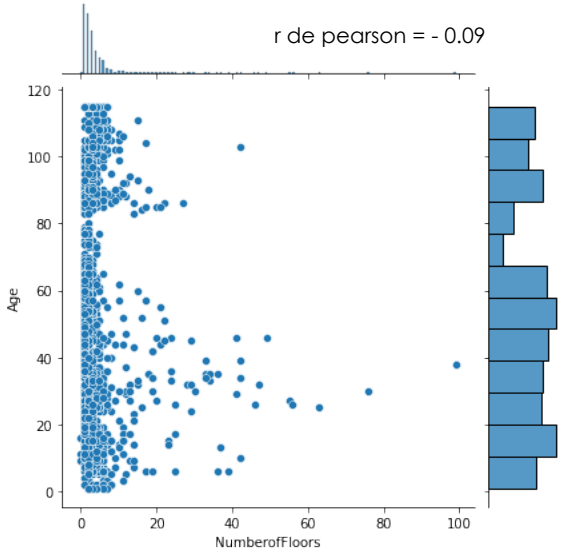
- Surface des bâtiments
- Nombre d'étages
- Nombre de bâtiments
- Age
- ENERGYSTARScore

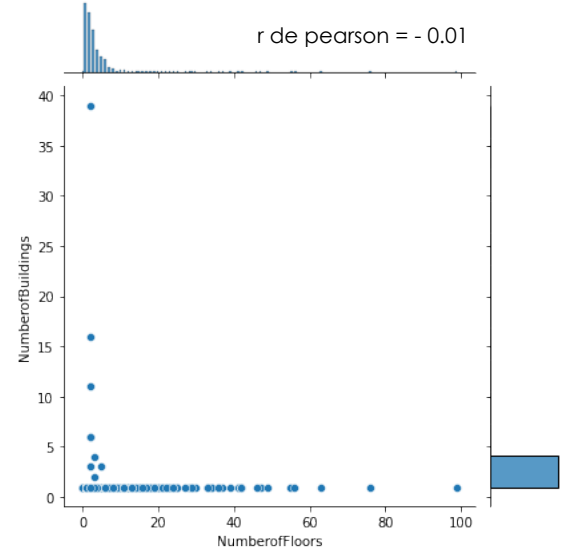


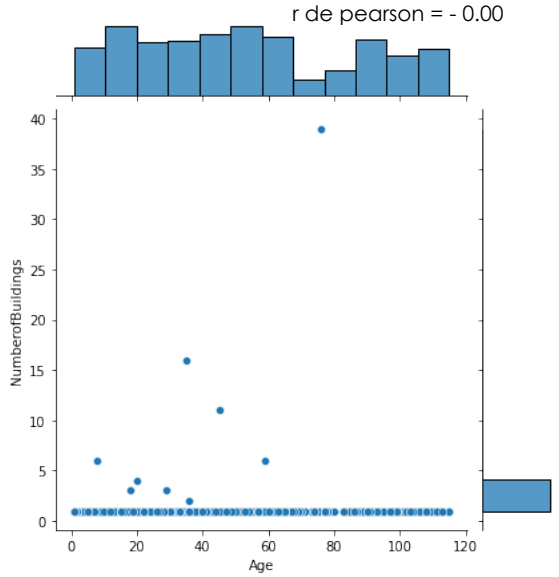






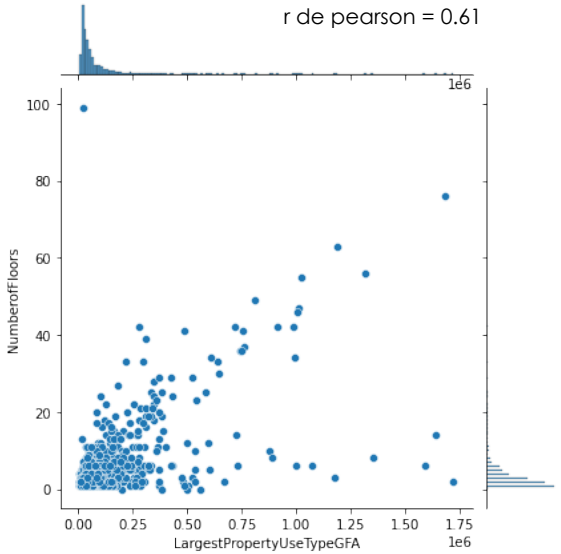


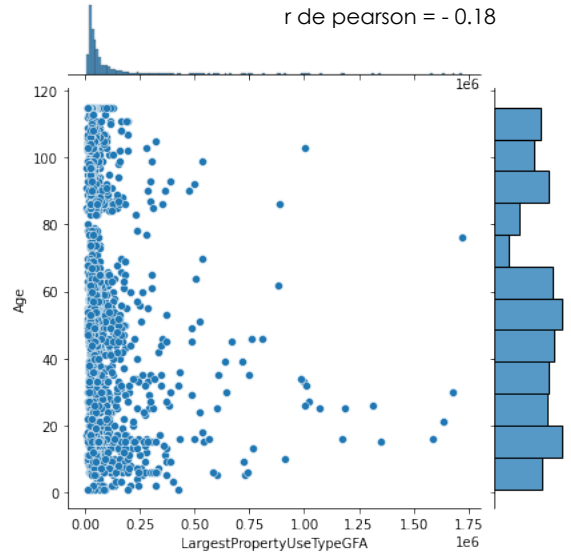


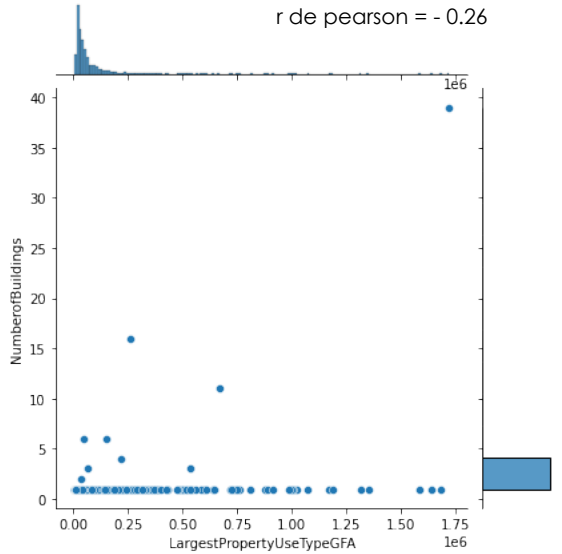


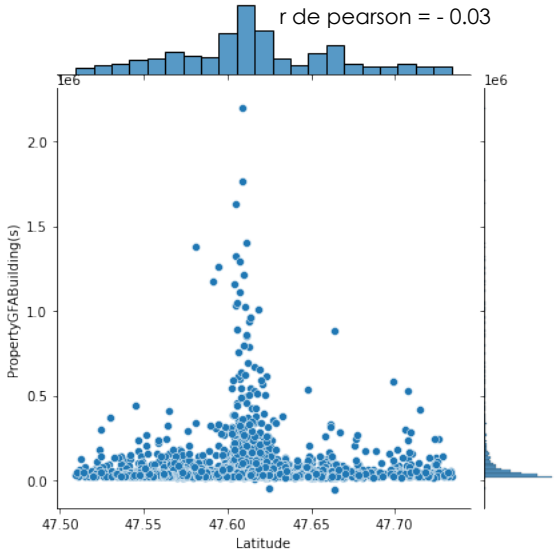
2. b. Analyses bivariées

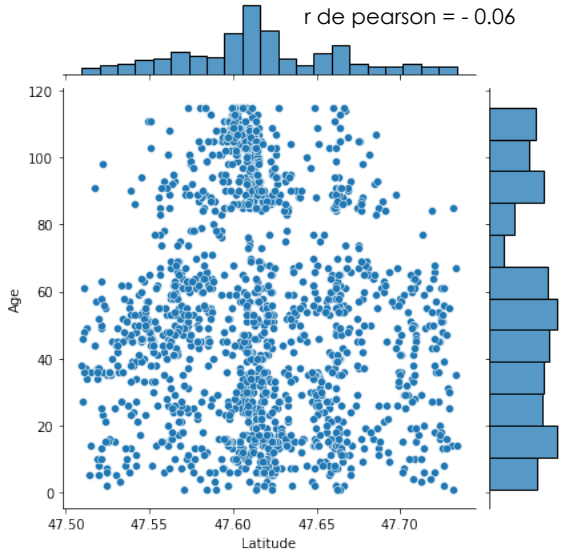
Aucune feature n'est corrélée à absolument toutes les autres features :

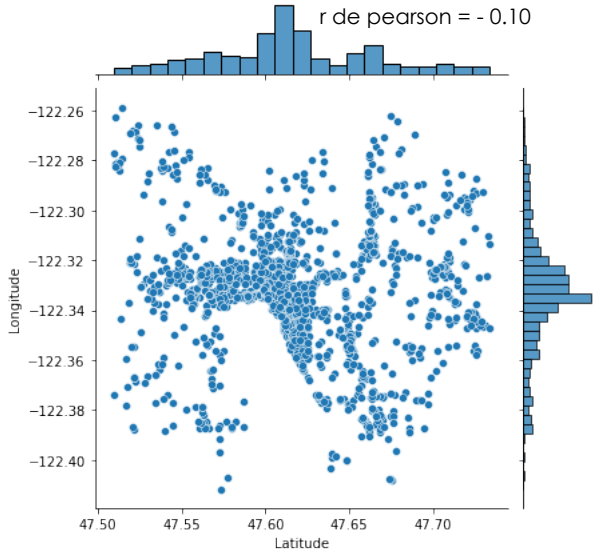


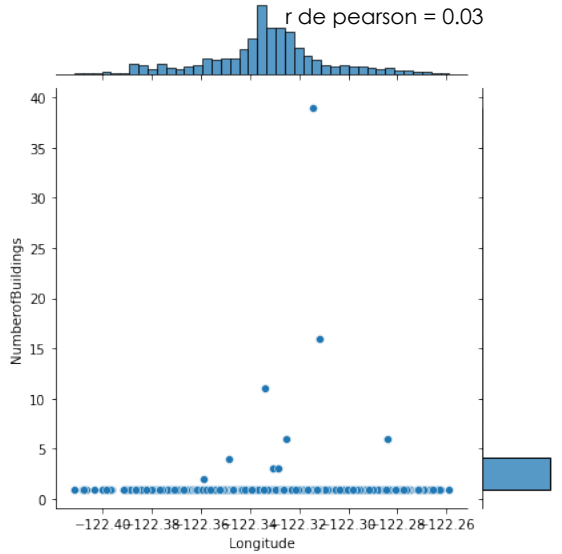






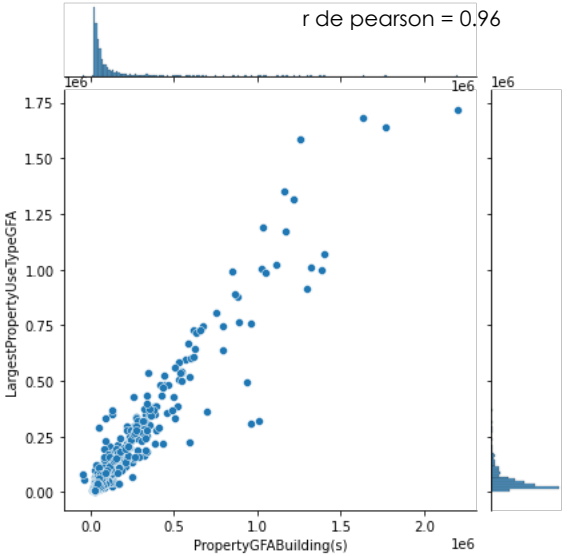




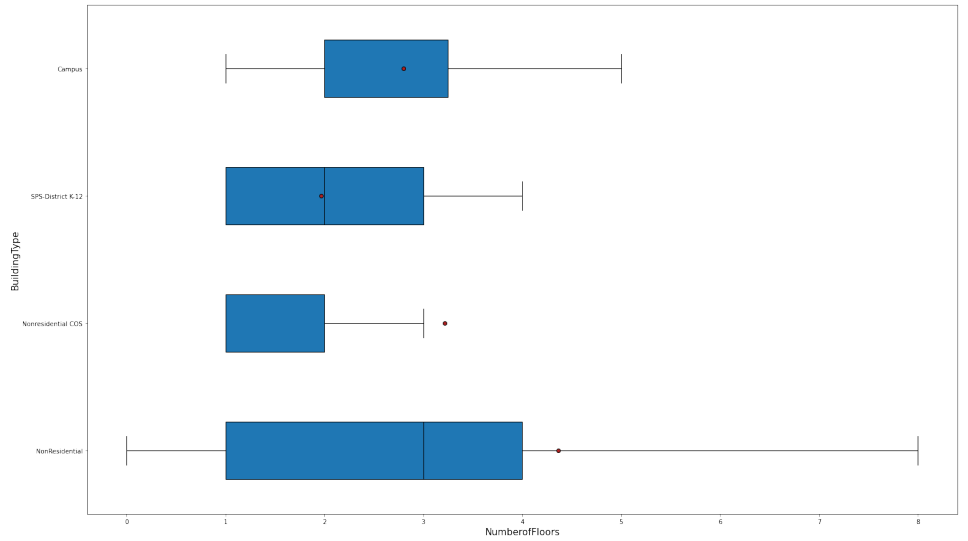


2. b. Analyses bivariées

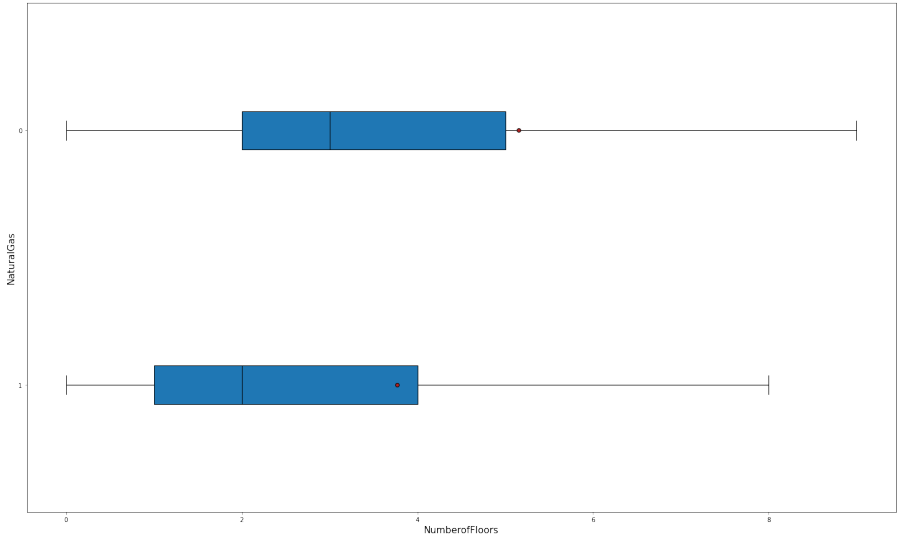
Cependant, certaines features ne sont pas indépendantes entre elles :



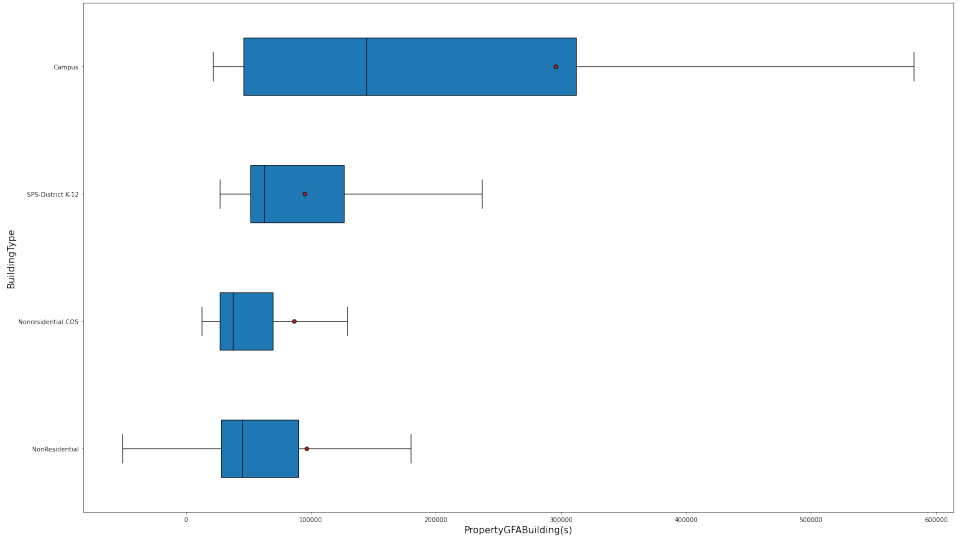
p=0.004



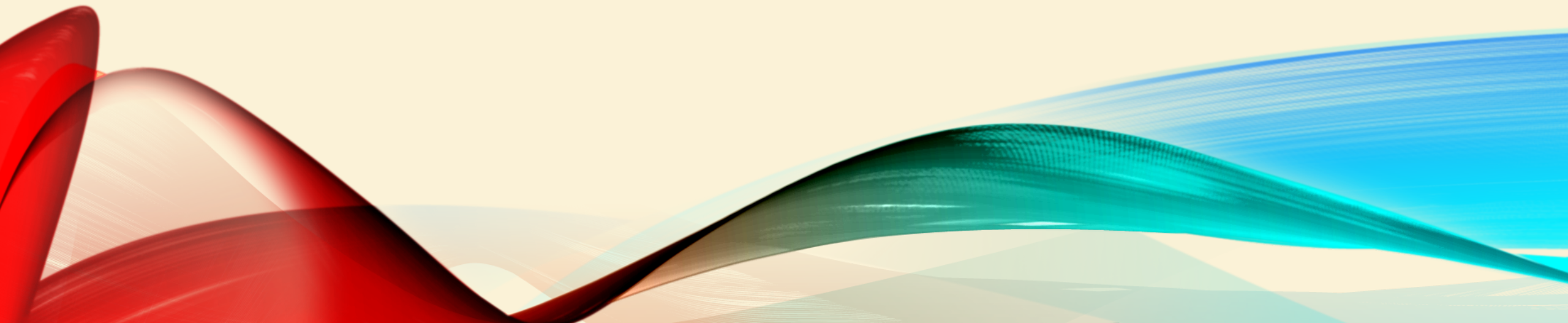
p < 0,001



p < 0,001



3. Modèles de Prévision



3. a. Sélection des features et des modèles

5 est le nombre optimal de features.

- Meilleures features pour modéliser la consommation totale d'énergie :
  - BuildingType,
  - LargestPropertyUseType,
  - NumberofBuildings,
  - Latitude,
  - Longitude.

- Meilleures features pour modéliser les émissions de CO2 :
  - BuildingType,
  - LargestPropertyUseType,
  - NumberofBuildings,
  - Steam,
  - OtherFuel.

3. a. Sélection des features et des modèles

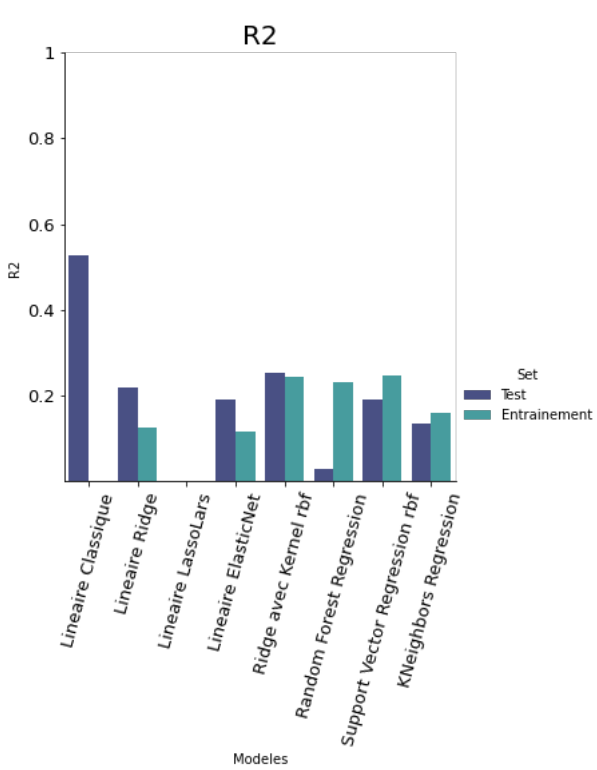
Comparaison de 8 types de modèles :

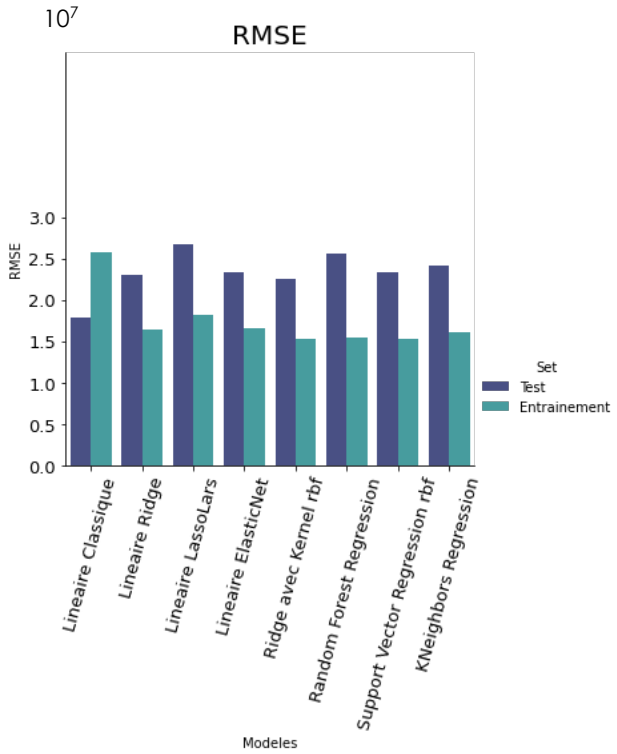
- 4 types de modèles linéaires :
  - Linéaire classique,
  - Ridge,
  - Lasso-Lars,
  - Elastic Net.

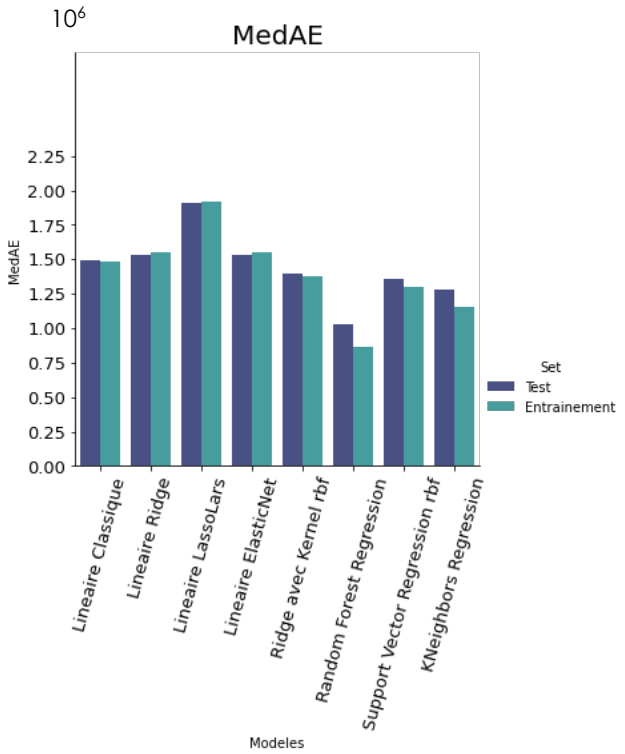
- 4 types de modèles non-linéaires :
  - Ridge avec un kernel rbf,
  - Random forest regression,
  - Support vecteur regression (SVR) avec un kernel rbf,
  - K Nearest Neighbors regression (KNN).

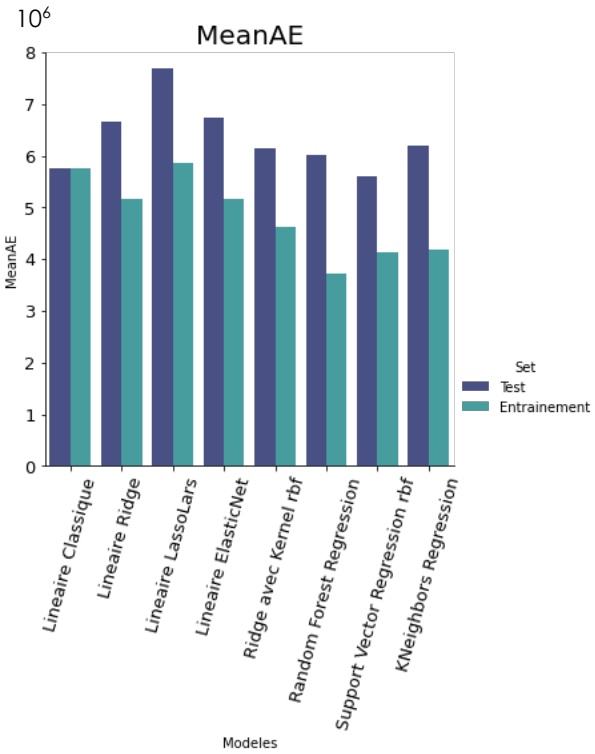
3. b. Sans prise en compte de l'ENERGYSTARScore

Consommation totale d'énergie









Les modèles sont peu performants.
On observe des biais ou du surapprentissage.

3. b. Sans prise en compte de l'ENERGYSTARScore

Consommation totale d'énergie

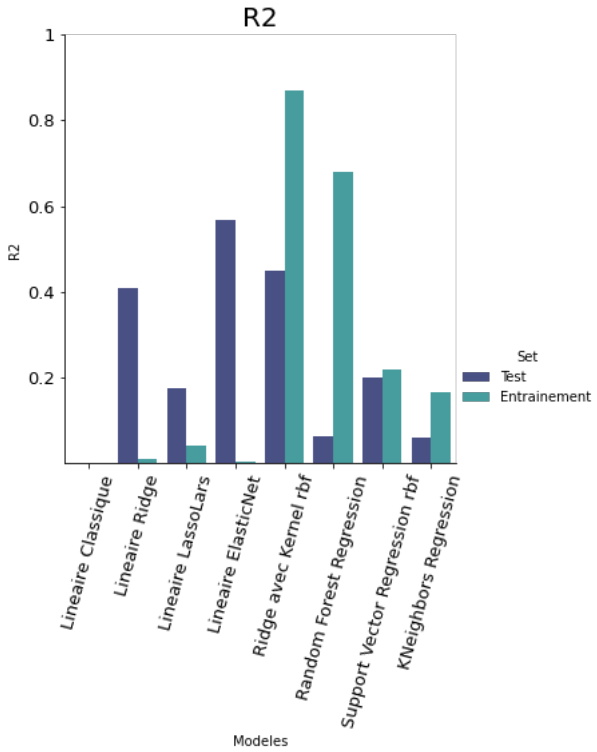
Les modèles Ridge avec noyau rbf et SVR avec noyau rbf sont les plus adaptés.

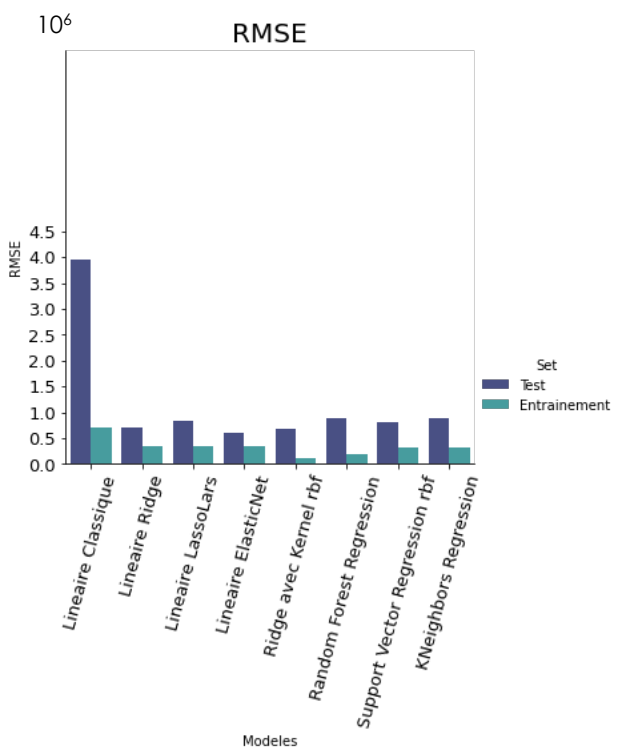
Cependant, leurs performances restent faibles.

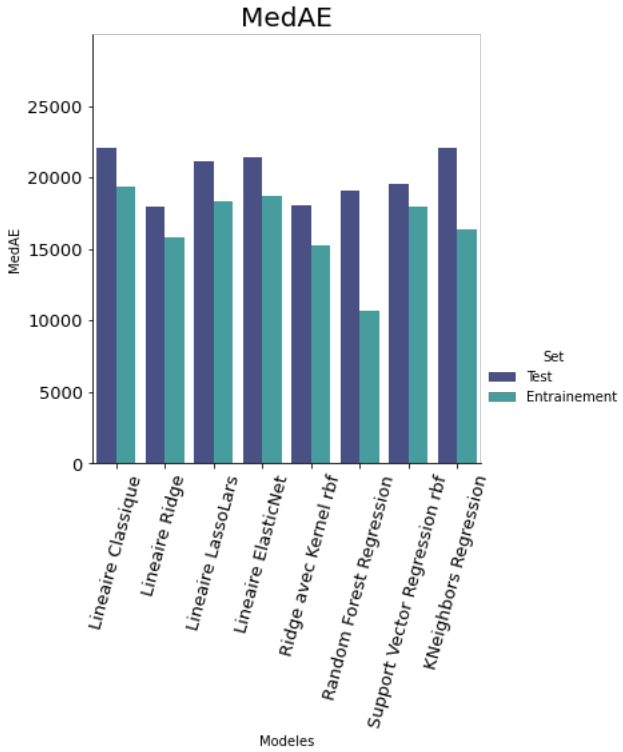
	MedAE	MeanAE	MSE	RMSE	R2	Temps (en s)
Linéaire classique (Test)	1.490609e+06	5.751175e+06	3.223453e+14	1.795398e+07	0.527902	0.000369
Linéaire classique (Entrainement)	1.487885e+06	5.758885e+06	6.652279e+14	2.579201e+07	-1.134902	0.011612
Ridge linéaire (Test)	1.527991e+06	6.660434e+06	5.329401e+14	2.308550e+07	0.219471	0.000387
Ridge linéaire (Entrainement)	1.553576e+06	5.157241e+06	2.725757e+14	1.650987e+07	0.125228	0.094487
Lasso-Lars (Test)	1.907940e+06	7.679855e+06	7.222928e+14	2.687551e+07	-0.057850	0.000463
Lasso-Lars (Entrainement)	1.916620e+06	5.874390e+06	3.317858e+14	1.821499e+07	-0.064793	0.095742
ElasticNet (Test)	1.532678e+06	6.750823e+06	5.518882e+14	2.349230e+07	0.191720	0.000113
ElasticNet (Entrainement)	1.550962e+06	5.164263e+06	2.752631e+14	1.659106e+07	0.116604	0.093564
Ridge avec Kernel rbf (Test)	1.394682e+06	6.146526e+06	5.099358e+14	2.258176e+07	0.253162	0.003920
Ridge avec Kernel rbf (Entrainement)	1.378845e+06	4.635625e+06	2.350995e+14	1.533295e+07	0.245500	0.524764
Random Forest Regression (Test)	1.027543e+06	6.016153e+06	6.632419e+14	2.575348e+07	0.028635	0.105300
Random Forest Regression (Entrainement)	8.618888e+05	3.726778e+06	2.393041e+14	1.546946e+07	0.232006	13.388667
Support Vector Regression (Test)	1.361838e+06	5.594370e+06	5.525877e+14	2.350718e+07	0.190696	0.005184
Support Vector Regression (Entrainement)	1.297760e+06	4.123273e+06	2.341992e+14	1.530357e+07	0.248389	1.172515
K Nearest Neighbors (Test)	1.284725e+06	6.209950e+06	5.894548e+14	2.427869e+07	0.136701	0.004657
K Nearest Neighbors (Entrainement)	1.152409e+06	4.198881e+06	2.618791e+14	1.618268e+07	0.159557	0.184386

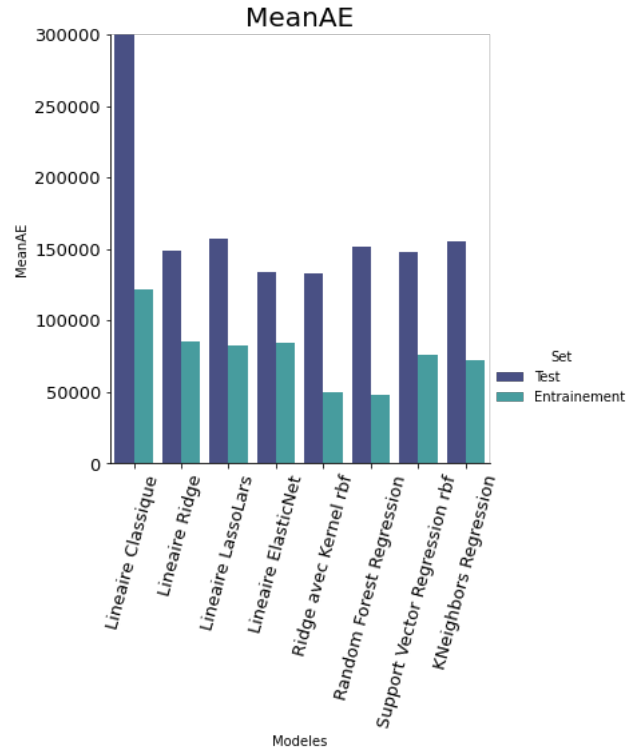
3. b. Sans prise en compte de l'ENERGYSTARScore

Emissions de CO2









Le modèle SVR avec noyau rbf semble le plus adapté mais reste peu performant.
Il y a un surapprentissage ou un biais sur les autres modèles.

3. b. Sans prise en compte de l'ENERGYSTARScore

Emissions de CO2

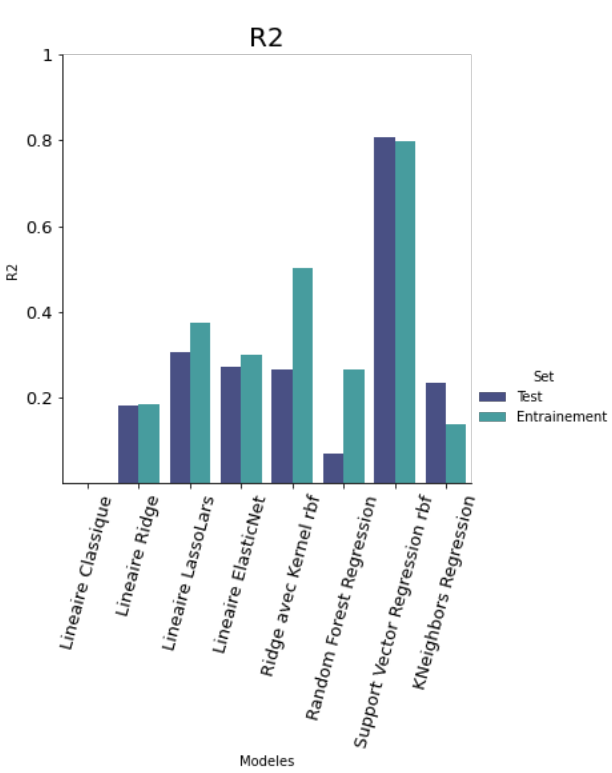
Le modèle SVR avec noyau rbf est le plus adapté.

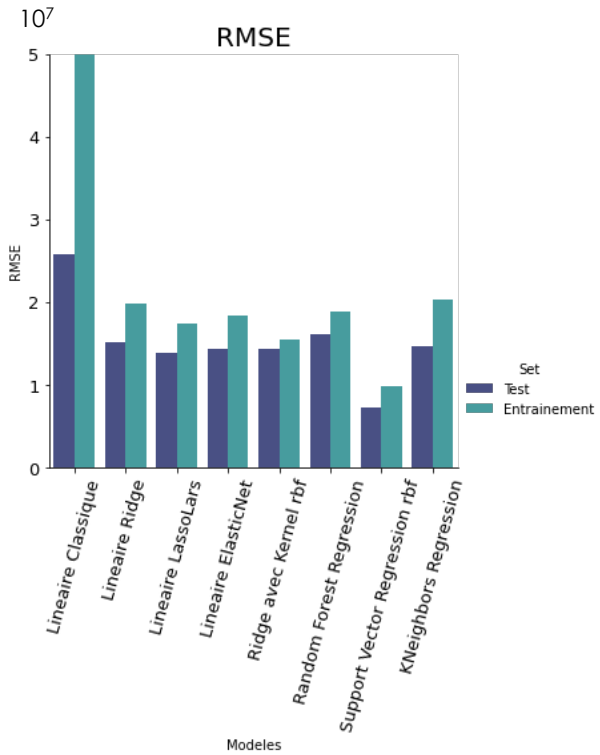
Cependant, ses performances restent faibles.

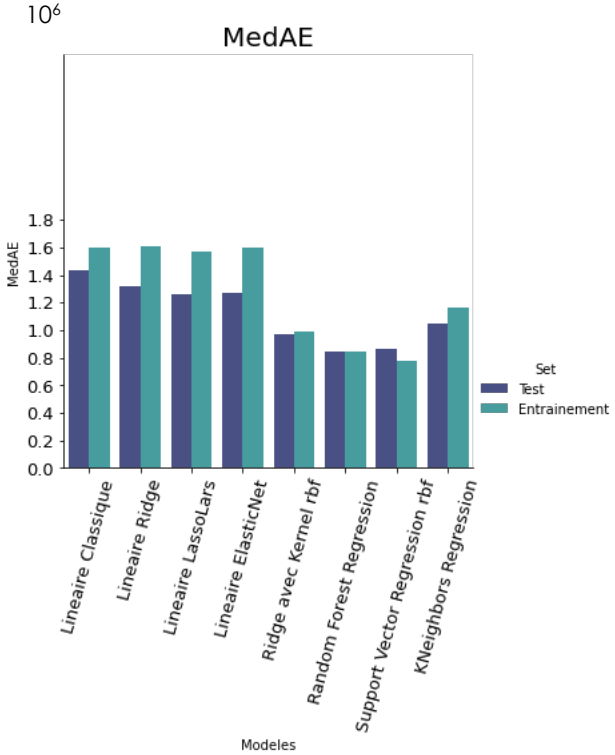
	MedAE	MeanAE	MSE	RMSE	R2	Temps (en s)
Linéaire classique (Test)	22028.091940	347395.993211	1.562723e+13	3.953129e+06	-17.492465	0.000352
Linéaire classique (Entrainement)	19387.197730	121661.741067	4.864866e+11	6.974859e+05	-2.965497	0.011868
Ridge linéaire (Test)	17967.244626	149008.446358	5.003373e+11	7.073453e+05	0.407926	0.000122
Ridge linéaire (Entrainement)	15813.918619	85648.457306	1.214263e+11	3.484627e+05	0.010218	0.100960
Lasso-Lars (Test)	21128.253352	157423.298059	6.956848e+11	8.340772e+05	0.176762	0.000317
Lasso-Lars (Entrainement)	18343.507327	82435.961432	1.177228e+11	3.431076e+05	0.040406	0.111070
ElasticNet (Test)	21402.083419	134015.427849	3.642505e+11	6.035317e+05	0.568964	0.000118
ElasticNet (Entrainement)	18704.981182	84863.612021	1.222874e+11	3.496962e+05	0.003198	0.091439
Ridge avec Kernel rbf (Test)	18091.716946	133184.617886	4.642450e+11	6.813553e+05	0.450636	0.003642
Ridge avec Kernel rbf (Entrainement)	15253.977060	50023.441318	1.619157e+10	1.272461e+05	0.868018	0.519123
Random Forest Regression (Test)	19089.954158	151929.850805	7.920723e+11	8.899844e+05	0.062702	0.106585
Random Forest Regression (Entrainement)	10673.511015	47876.305229	3.944147e+10	1.985988e+05	0.678501	12.870946
Support Vector Regression (Test)	19511.521955	147680.027047	6.761310e+11	8.222718e+05	0.199901	0.005593
Support Vector Regression (Entrainement)	17992.288101	75635.833295	9.591145e+10	3.096957e+05	0.218197	30.552934
K Nearest Neighbors (Test)	22118.815070	154996.672837	7.941747e+11	8.911648e+05	0.060214	0.004866
K Nearest Neighbors (Entrainement)	16377.428146	72355.858799	1.022330e+11	3.197389e+05	0.166669	0.183040

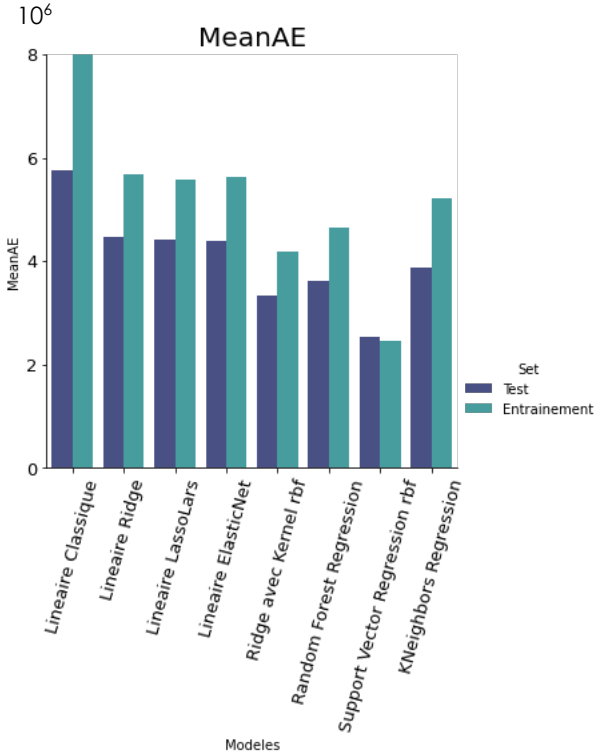
3. c. Avec prise en compte de l'ENERGYSTARScore

Consommation totale d'énergie









Le modèle SVR avec noyau rbf se détache par ses performances.

3. c. Avec prise en compte de l'ENERGYSTARScore

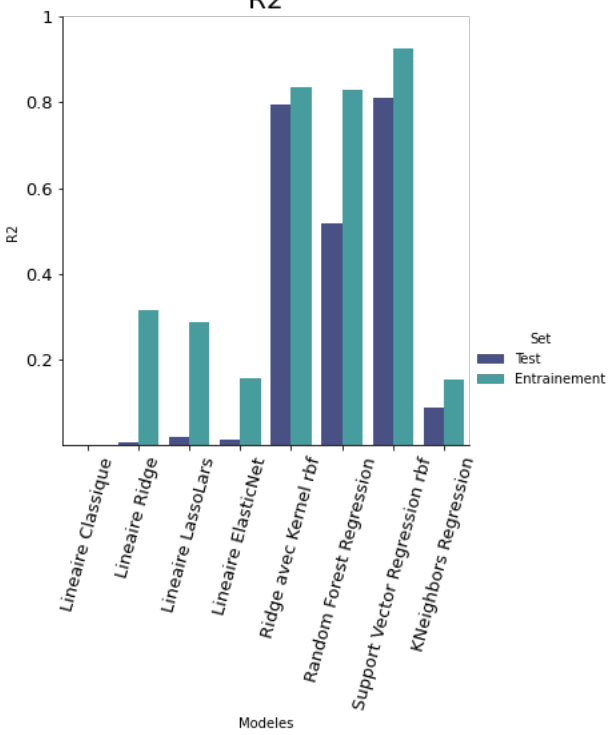
Consommation totale d'énergie

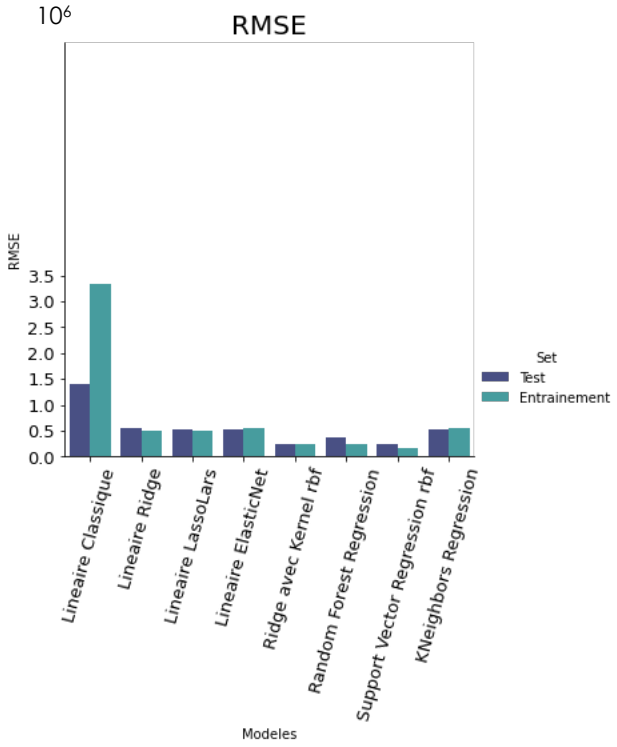
Le modèle SVR avec noyau rbf est le plus adapté.

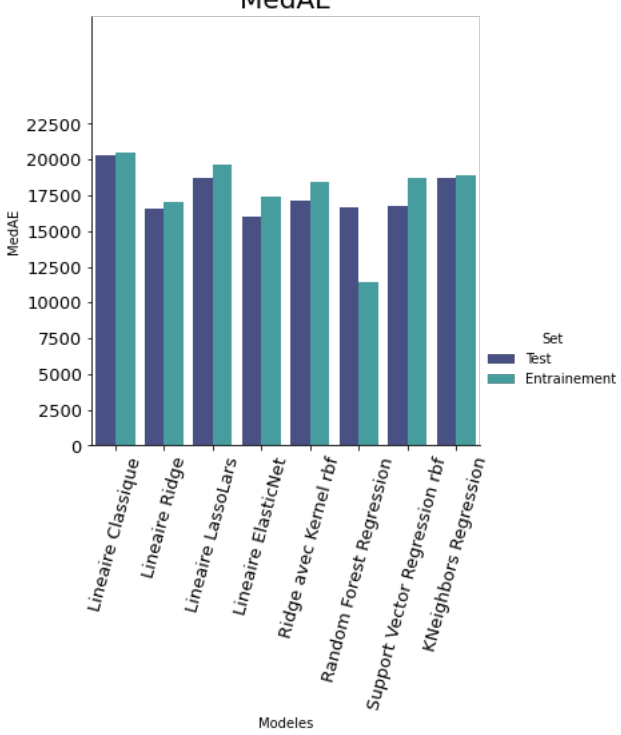
	MedAE	MeanAE	MSE	RMSE	R2	Temps (en s)
Linéaire classique (Test)	1.439133e+06	5.766929e+06	6.698949e+14	2.588233e+07	-1.377019	0.000367
Linéaire classique (Entrainement)	1.595315e+06	9.205579e+06	4.875109e+15	6.982198e+07	-9.077232	0.013422
Ridge linéaire (Test)	1.318553e+06	4.469947e+06	2.308880e+14	1.519500e+07	0.180729	0.000132
Ridge linéaire (Entrainement)	1.605244e+06	5.676568e+06	3.938924e+14	1.984672e+07	0.185793	0.088417
Lasso-Lars (Test)	1.258880e+06	4.406632e+06	1.957015e+14	1.398934e+07	0.305583	0.000294
Lasso-Lars (Entrainement)	1.565641e+06	5.572058e+06	3.026432e+14	1.739664e+07	0.374413	0.155612
ElasticNet (Test)	1.273120e+06	4.387752e+06	2.055535e+14	1.433714e+07	0.270625	0.000114
ElasticNet (Entrainement)	1.601743e+06	5.634841e+06	3.385714e+14	1.840031e+07	0.300146	0.093456
Ridge avec Kernel rbf (Test)	9.728592e+05	3.338382e+06	2.069246e+14	1.438487e+07	0.265760	0.003657
Ridge avec Kernel rbf (Entrainement)	9.860876e+05	4.175670e+06	2.409701e+14	1.552321e+07	0.501896	0.518688
Random Forest Regression (Test)	8.461790e+05	3.622685e+06	2.623241e+14	1.619642e+07	0.069183	0.108516
Random Forest Regression (Entrainement)	8.499980e+05	4.642997e+06	3.555352e+14	1.885564e+07	0.265081	13.176617
Support Vector Regression (Test)	8.639565e+05	2.545404e+06	5.429781e+13	7.368705e+06	0.807333	0.004921
Support Vector Regression (Entrainement)	7.823148e+05	2.471279e+06	9.796927e+13	9.897943e+06	0.797490	2.004681
K Nearest Neighbors (Test)	1.052163e+06	3.888171e+06	2.153105e+14	1.467346e+07	0.236004	0.004981
K Nearest Neighbors (Entrainement)	1.165175e+06	5.227093e+06	4.169647e+14	2.041971e+07	0.138101	0.194166

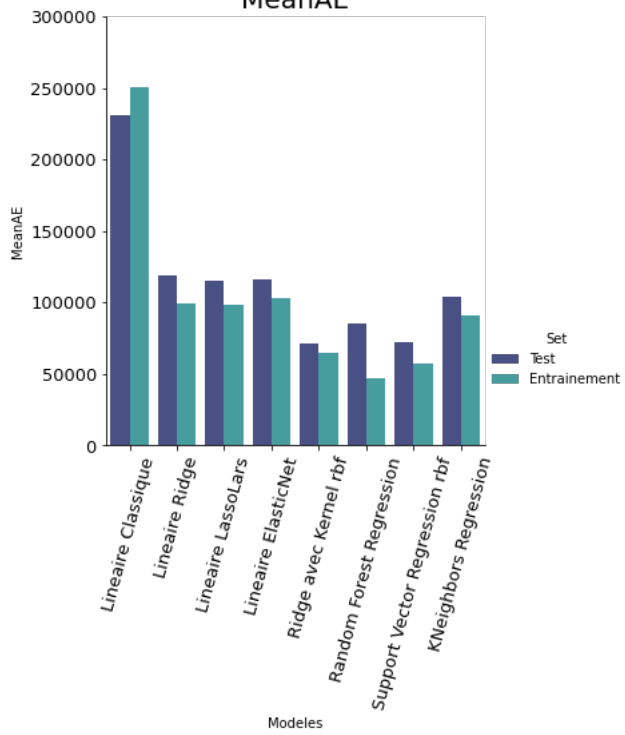
3. c. Avec prise en compte de l'ENERGYSTARScore

Emissions de CO2









Les modèles Ridge avec noyau rbf et SVR avec noyau rbf sont plus adaptés
Les modèles autres modèles montrent un surapprentissage ou des performances trop faibles.

3. c. Avec prise en compte de l'ENERGYSTARScore

Emissions de CO2

Les modèles Ridge avec noyau rbf et SVR avec noyau rbf sont les plus performants.

Le modèle SVR est 60 fois plus long à entraîner
et
a une différence légèrement plus grande entre le r^2 à l'entraînement et au test.

	MedAE	MeanAE	MSE	RMSE	R2	Temps (en s)
Linéaire classique (Test)	20336.845443	231307.198267	1.970755e+12	1.403836e+06	-5.668067	0.000390
Linéaire classique (Entrainement)	20528.143739	250737.212209	1.108735e+13	3.329767e+06	-29.710128	0.012507
Ridge linéaire (Test)	16530.766644	118947.469897	2.928541e+11	5.411600e+05	0.009125	0.000419
Ridge linéaire (Entrainement)	16997.850531	98989.924359	2.471661e+11	4.971580e+05	0.315391	0.085968
Lasso-Lars (Test)	18729.831876	115386.356323	2.897386e+11	5.382738e+05	0.019667	0.000437
Lasso-Lars (Entrainement)	19668.419548	98294.772784	2.570284e+11	5.069797e+05	0.288074	0.109990
ElasticNet (Test)	16041.115431	116328.765367	2.914874e+11	5.398957e+05	0.013750	0.000124
ElasticNet (Entrainement)	17367.959813	103301.135187	3.044992e+11	5.518145e+05	0.156587	0.087854
Ridge avec Kernel rbf (Test)	17087.740243	71692.104708	6.037091e+10	2.457049e+05	0.795734	0.003965
Ridge avec Kernel rbf (Entrainement)	18435.512594	64878.690110	5.989517e+10	2.447349e+05	0.834100	0.559485
Random Forest Regression (Test)	16702.104634	85758.206641	1.429153e+11	3.780414e+05	0.516445	0.106036
Random Forest Regression (Entrainement)	11394.232279	47226.126929	6.121353e+10	2.474137e+05	0.830449	12.842847
Support Vector Regression (Test)	16717.138087	71908.607023	5.616209e+10	2.369854e+05	0.809975	0.005118
Support Vector Regression (Entrainement)	18737.035095	57132.538718	2.647285e+10	1.627048e+05	0.926675	57.235097
K Nearest Neighbors (Test)	18685.706930	103657.316864	2.693054e+11	5.189464e+05	0.088803	0.004861
K Nearest Neighbors (Entrainement)	18886.097059	91062.055387	3.053599e+11	5.525938e+05	0.154203	0.187346

5. Conclusions et Perspectives

Conclusions et Perspectives

Conclusions:

- La prévision d'émissions de CO2 est optimale avec 5 features;
- L'ENERGYSTARScore améliore significativement les prédictions;
- Le modèle Kernel avec un noyau rbf est les modèle optimal pour prévoir la consommation totale d'énergie et les émissions de CO2.

Perspectives :

- Augmenter le nombre de données pour alimenter les modèles, en prenant en compte de plus de bâtiments.

MERCI POUR VOTRE ATTENTION !