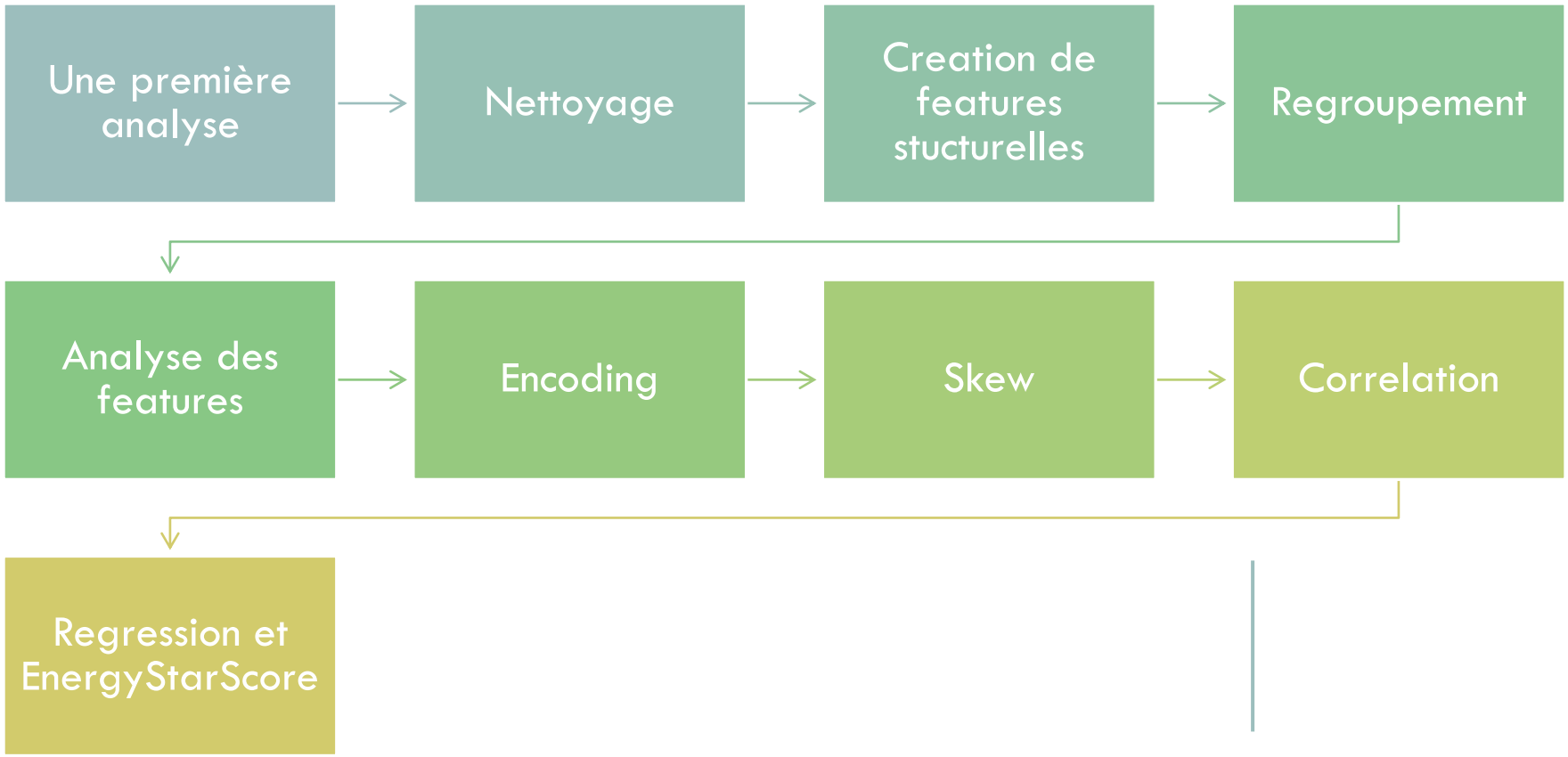
ANTICIPEZ LES BESOINS EN CONSOMMATION DE BÂTIMENTS

Emanuele Partenza

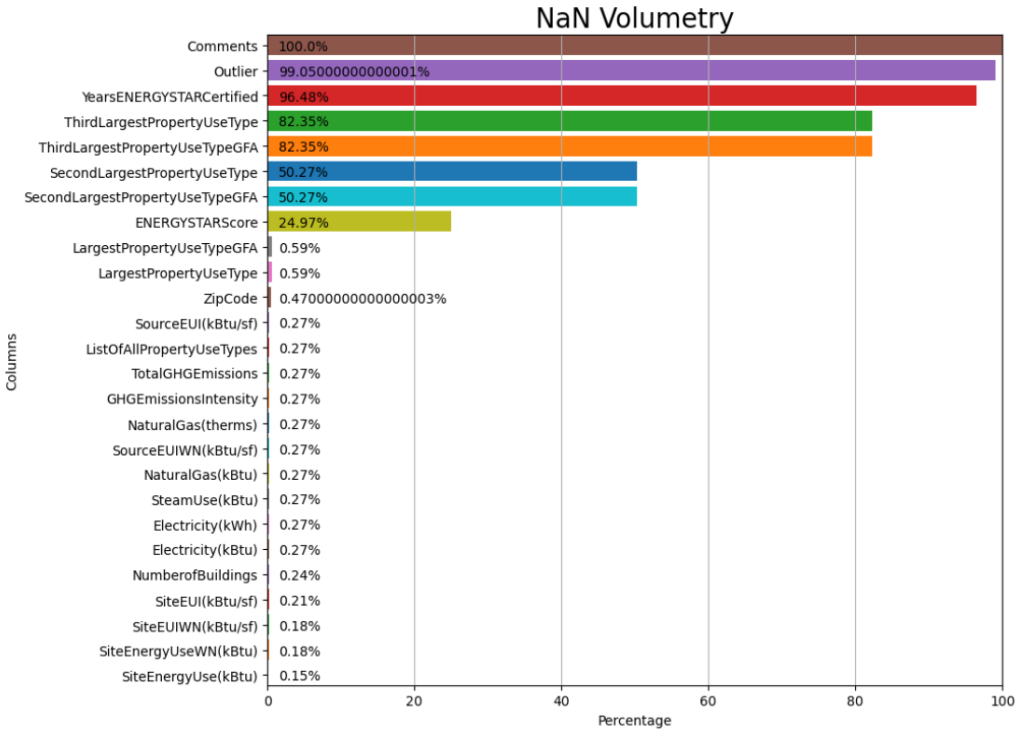
ANTICIPEZ LES BESOINS EN CONSOMMATION DE BÂTIMENTS:



UNE PREMIÈRE ANALYSE

- Shape
- Nan volumetry
- Dtypes
- Describe
- Duplicated
- Nunique

```
LargestPropertyUseType             object
YearsENERGYSTARCertified           object
Neighborhood                       object
TaxParcelIdentificationNumber      object
SecondLargestPropertyUseType       object
State                              object
City                               object
Address                            object
PropertyName                       object
PrimaryPropertyType                object
BuildingType                       object
ComplianceStatus                   object
Outlier                            object
ListOfAllPropertyUseTypes          object
ThirdLargestPropertyUseType        object
```

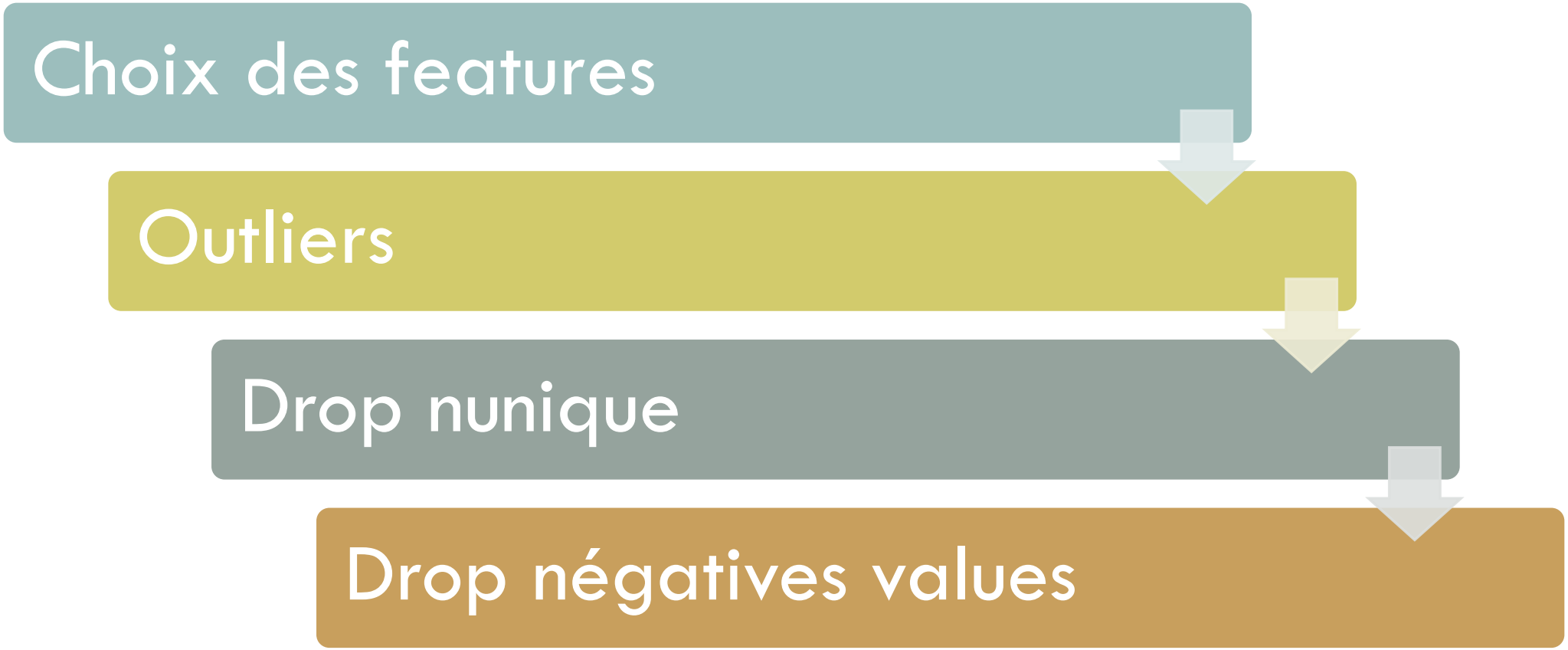



```
DefaultData
OSEBuildingID
DataYear
PropertyGFABuilding(s)
PropertyGFAParking
PropertyGFATotal
NumberofFloors
CouncilDistrictCode
YearBuilt
ENERGYSTARScore
SiteEUI(kBtu/sf)
SiteEUIWN(kBtu/sf)
SourceEUI(kBtu/sf)
SiteEnergyUse(kBtu)
ThirdLargestPropertyUseTypeGFA
SiteEnergyUseWN(kBtu)
SteamUse(kBtu)
Electricity(kWh)
Electricity(kBtu)
NaturalGas(therms)
NaturalGas(kBtu)
Comments
SourceEUIWN(kBtu/sf)
SecondLargestPropertyUseTypeGFA
LargestPropertyUseTypeGFA
ZipCode
Latitude
Longitude
NumberofBuildings
GHGEmissionsIntensity
TotalGHGEmissions
```

LEGEND :

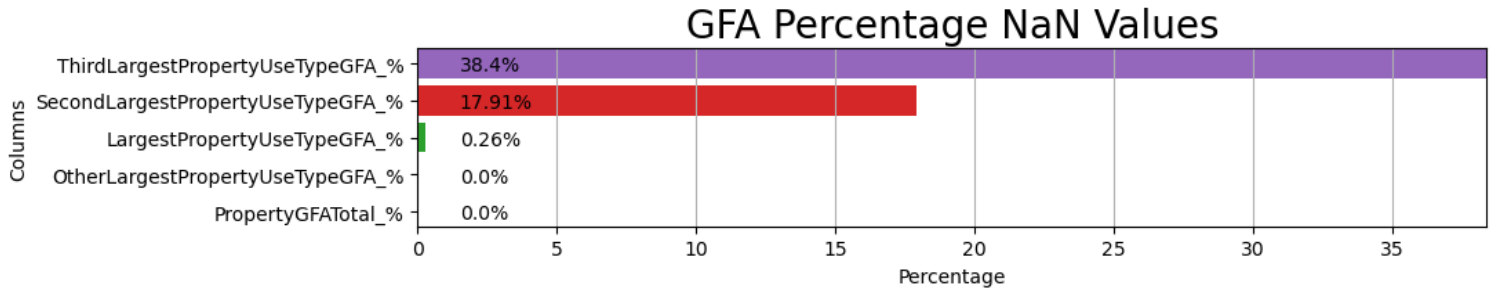
GFA: Groos Floor Area
Btu : British thermal unit
WN : Weather Normalized
EUI : Energy Use Intensity

NETTOYAGE


Choix des features
Outliers
Drop nunique
Drop négatives values

GROSS FLOOR AREA FEATURES

- PropertyGFATotal < LargestPropertyUseTypeGFA in 180 buildings
- ThirdLargestPropertyUseTypeGFA < OtherLargestPropertyUseTypeGFA in 97 buildings
- PropertyGFATotal_% > 105 in 228 buildings
- PropertyGFATotal_% < 95 in 313 buildings

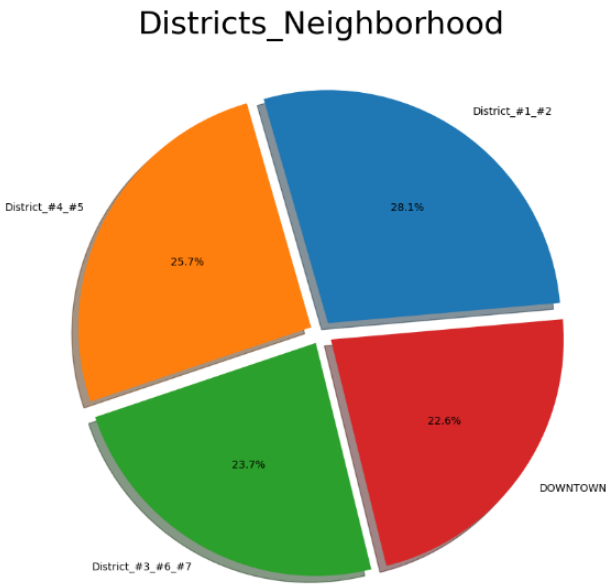


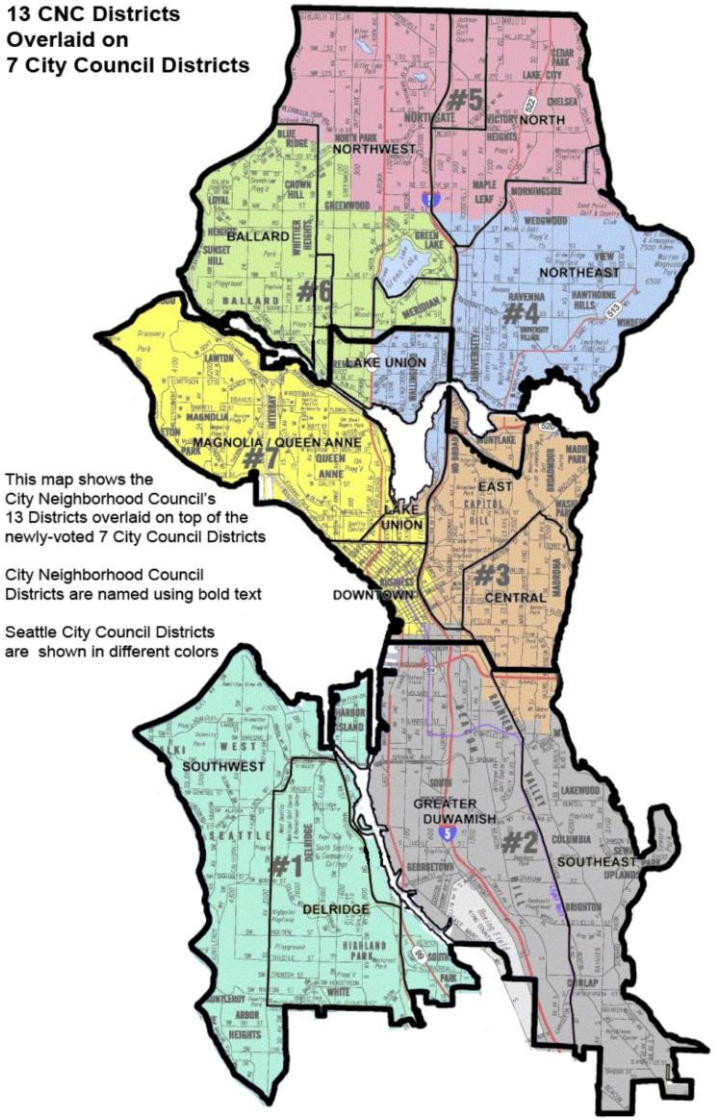
ENERGY TYPE PERCENTAGE

	SiteEnergyUse(kBtu)_%	SteamUse(kBtu)_%	Electricity(kBtu)_%	NaturalGas(kBtu)_%
count	1525.000000	1525.000000	1525.000000	1525.000000
mean	99.999591	2.108971	70.219995	27.670625
std	0.014466	9.134812	26.321659	26.443417
min	99.435800	0.000000	0.000000	0.000000
25%	99.999982	0.000000	49.136821	0.000000
50%	99.999991	0.000000	71.017542	23.833650
75%	100.000000	0.000000	99.999790	49.279076
max	100.000342	76.698738	100.000342	100.000000

DE NEIGHBORHOOD A DISTRICTS_NEIGHBORHOOD

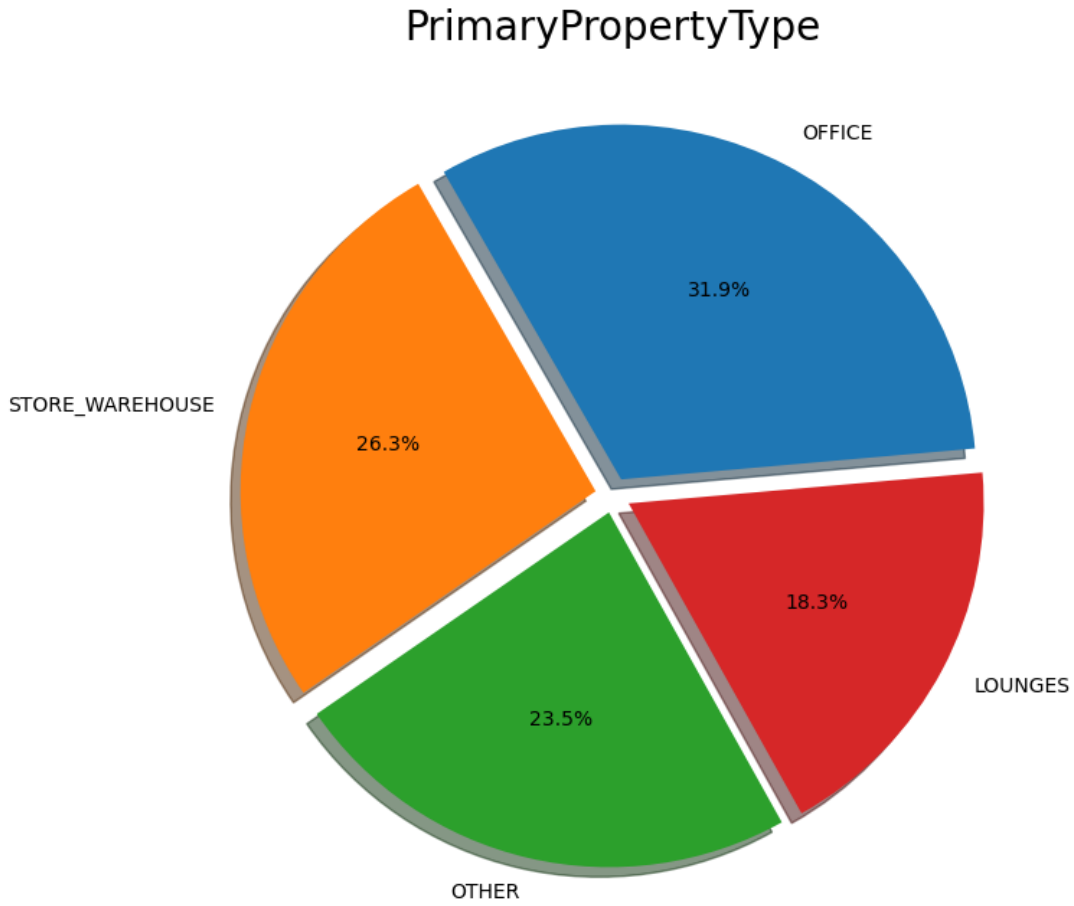
Regroupement des neighborhoods selon leur distribution à l'intérieur des districts

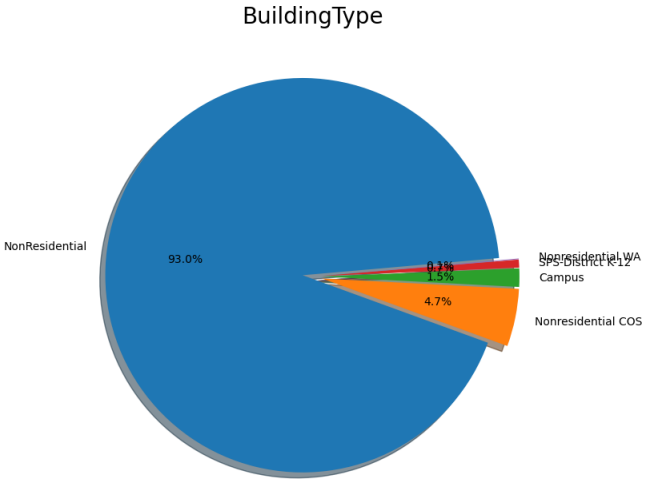




PRIMARY PROPERTY TYPE

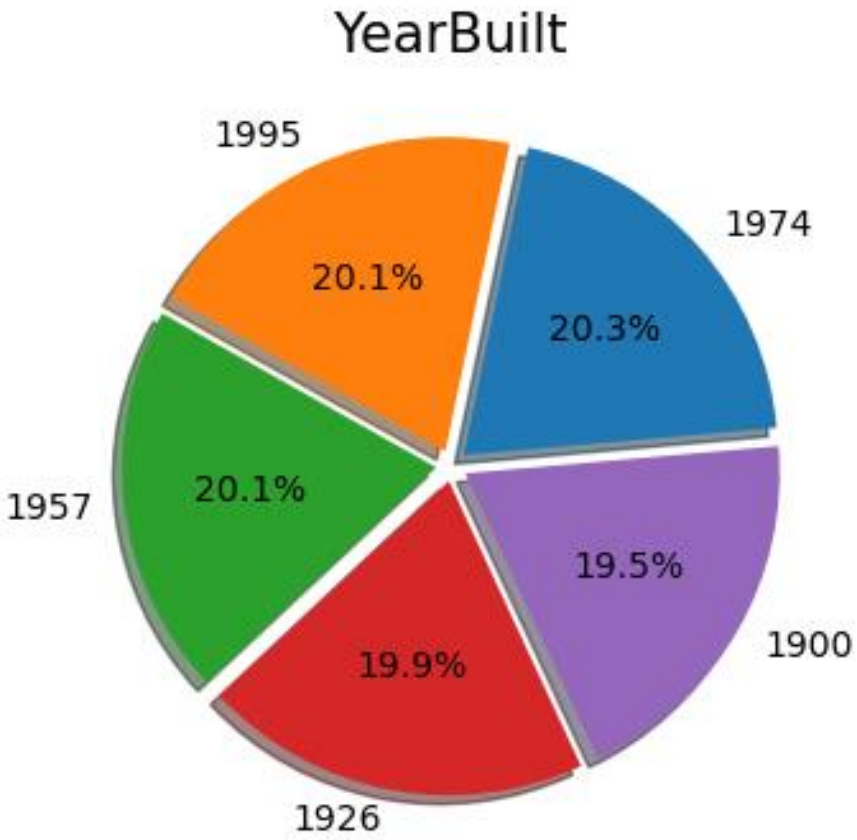
Regroupement des types de bâtiments selon leur fonction ou structure, avec le but d'obtenir des classes plus homogènes.
Suppression de la feature BuildingType



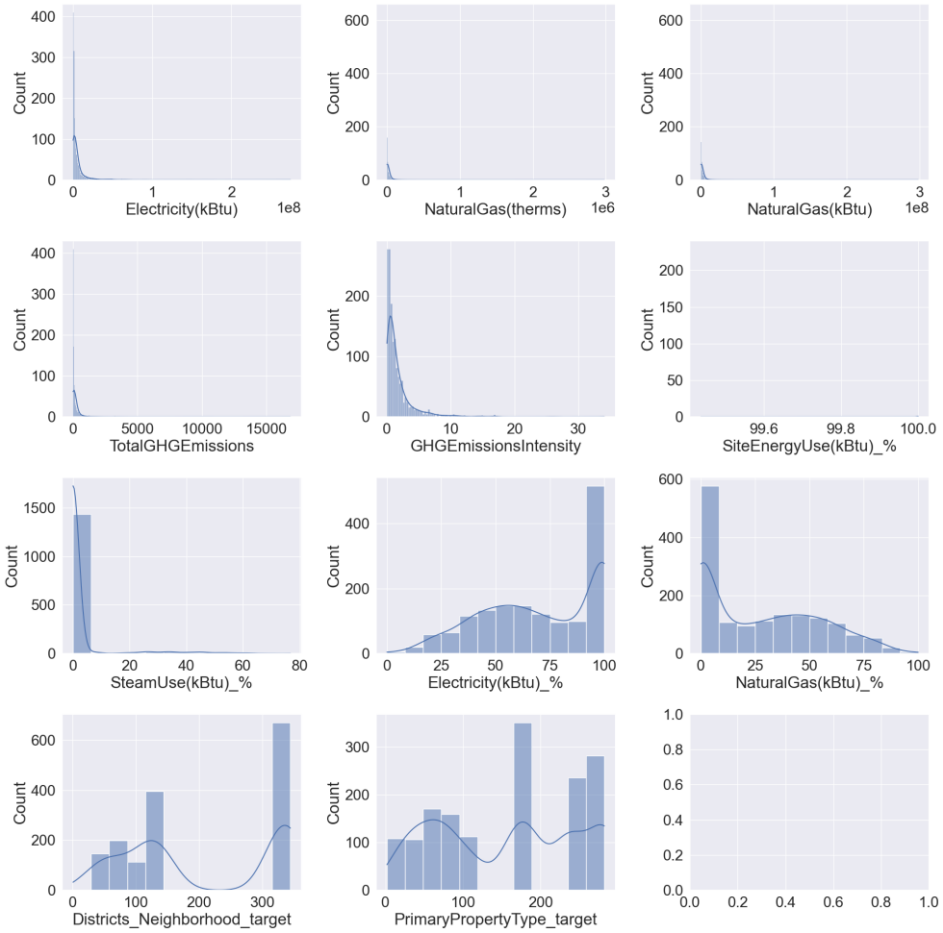


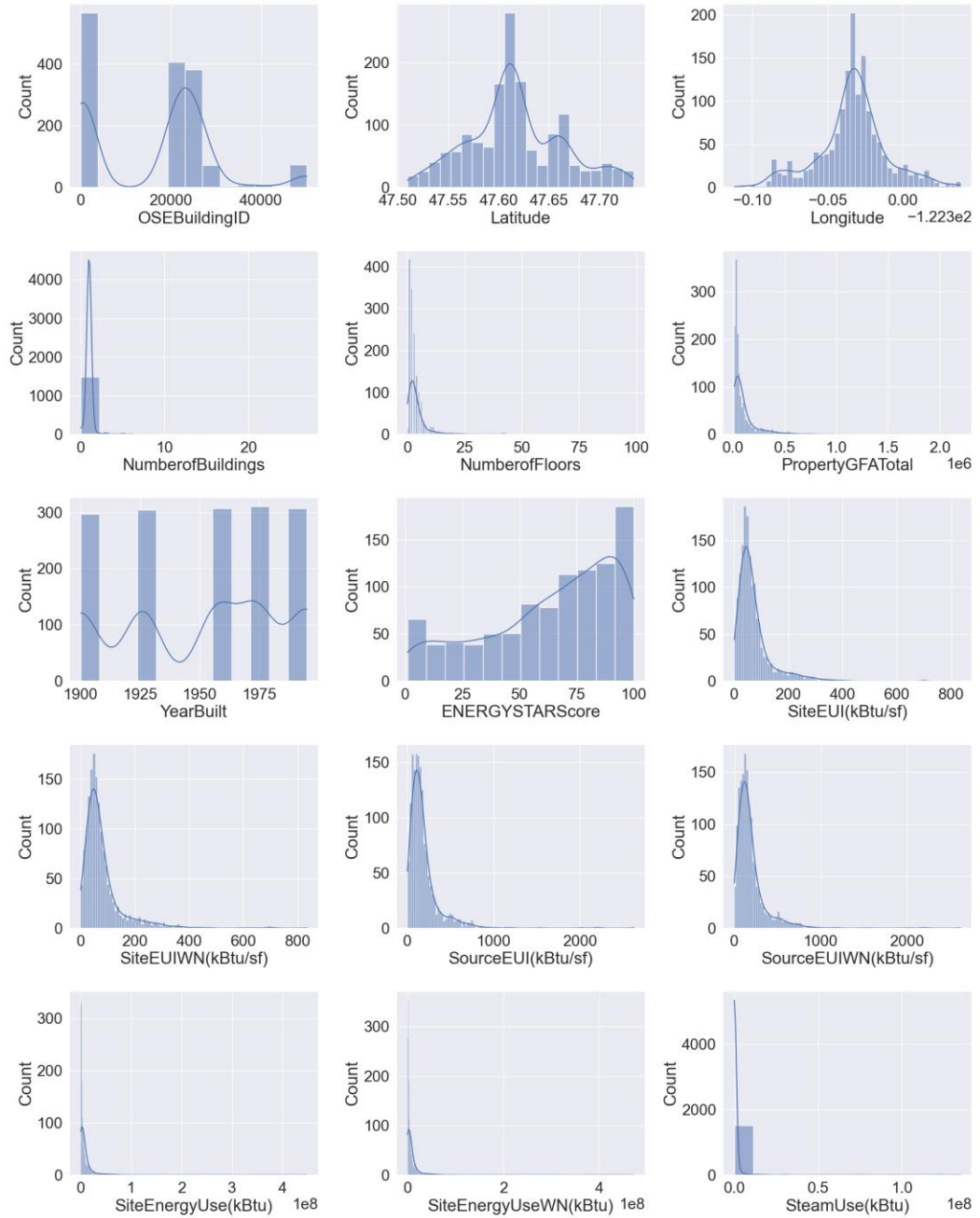
YEAR BUILT

Cinq tranches crées à partir du nombre de bâtiments par année, ceci pour éviter le surapprentissage et créer une distribution homogène.



HISTOGRAMMES





ENCODING

One Hot Encoding

Target Encoding

Transforment les features de type objet en types numériques, de façon à pouvoir les utiliser dans les algorithmes de régression.

NORMALTEST

	statistic	pvalue
OSEBuildingID	36.161842	1.404610e-08
Latitude	17.905665	1.293702e-04
Longitude	31.253766	1.634297e-07
NumberofBuildings	2838.688564	0.000000e+00
NumberofFloors	1757.809762	0.000000e+00
PropertyGFATotal	1531.268667	0.000000e+00
YearBuilt	3467.931217	0.000000e+00
ENERGYSTARScore	NaN	NaN
SiteEUI(kBtu/sf)	1252.958866	8.383769e-273
SiteEUIWN(kBtu/sf)	1220.927470	7.567781e-266
SourceEUI(kBtu/sf)	1595.263141	0.000000e+00
SourceEUIWN(kBtu/sf)	1588.771001	0.000000e+00
SiteEnergyUse(kBtu)	2522.597032	0.000000e+00
SiteEnergyUseWN(kBtu)	2555.719263	0.000000e+00
SteamUse(kBtu)	3443.277375	0.000000e+00
Electricity(kBtu)	2303.157697	0.000000e+00
NaturalGas(therms)	3503.900604	0.000000e+00
NaturalGas(kBtu)	3503.900603	0.000000e+00
TotalGHGEmissions	2945.912519	0.000000e+00
GHGEmissionsIntensity	1484.513311	0.000000e+00
SiteEnergyUse(kBtu)_%	4466.586845	0.000000e+00
SteamUse(kBtu)_%	1452.938012	0.000000e+00
Electricity(kBtu)_%	527.175871	3.351305e-115
NaturalGas(kBtu)_%	379.705415	3.531967e-83
Districts_Neighborhood_target	6688.715723	0.000000e+00
PrimaryPropertyType_target	18951.472593	0.000000e+00

SKEWTEST

	statistic	pvalue
OSEBuildingID	4.602405	4.176394e-06
Latitude	4.063533	4.833555e-05
Longitude	-1.901107	5.728799e-02
NumberofBuildings	46.012501	0.000000e+00
NumberofFloors	35.075591	1.588421e-269
PropertyGFATotal	32.737281	4.606531e-235
YearBuilt	-3.591914	3.282581e-04
ENERGYSTARScore	NaN	NaN
SiteEUI(kBtu/sf)	29.135630	1.270528e-186
SiteEUIWN(kBtu/sf)	28.753718	8.138308e-182
SourceEUI(kBtu/sf)	32.934324	7.093456e-238
SourceEUIWN(kBtu/sf)	32.853335	1.020707e-236
SiteEnergyUse(kBtu)	42.946838	0.000000e+00
SiteEnergyUseWN(kBtu)	43.259049	0.000000e+00
SteamUse(kBtu)	51.529865	0.000000e+00
Electricity(kBtu)	40.624112	0.000000e+00
NaturalGas(therms)	51.944108	0.000000e+00
NaturalGas(kBtu)	51.944108	0.000000e+00
TotalGHGEmissions	47.020115	0.000000e+00
GHGEmissionsIntensity	31.741015	4.224551e-221
SiteEnergyUse(kBtu)_%	-59.935984	0.000000e+00
SteamUse(kBtu)_%	32.298253	7.401358e-229
Electricity(kBtu)_%	-5.551423	2.833530e-08
NaturalGas(kBtu)_%	7.457747	8.801442e-14
Districts_Neighborhood_target	1.169043	2.423863e-01
PrimaryPropertyType_target	-0.426949	6.694168e-01

KURTOSISTEST

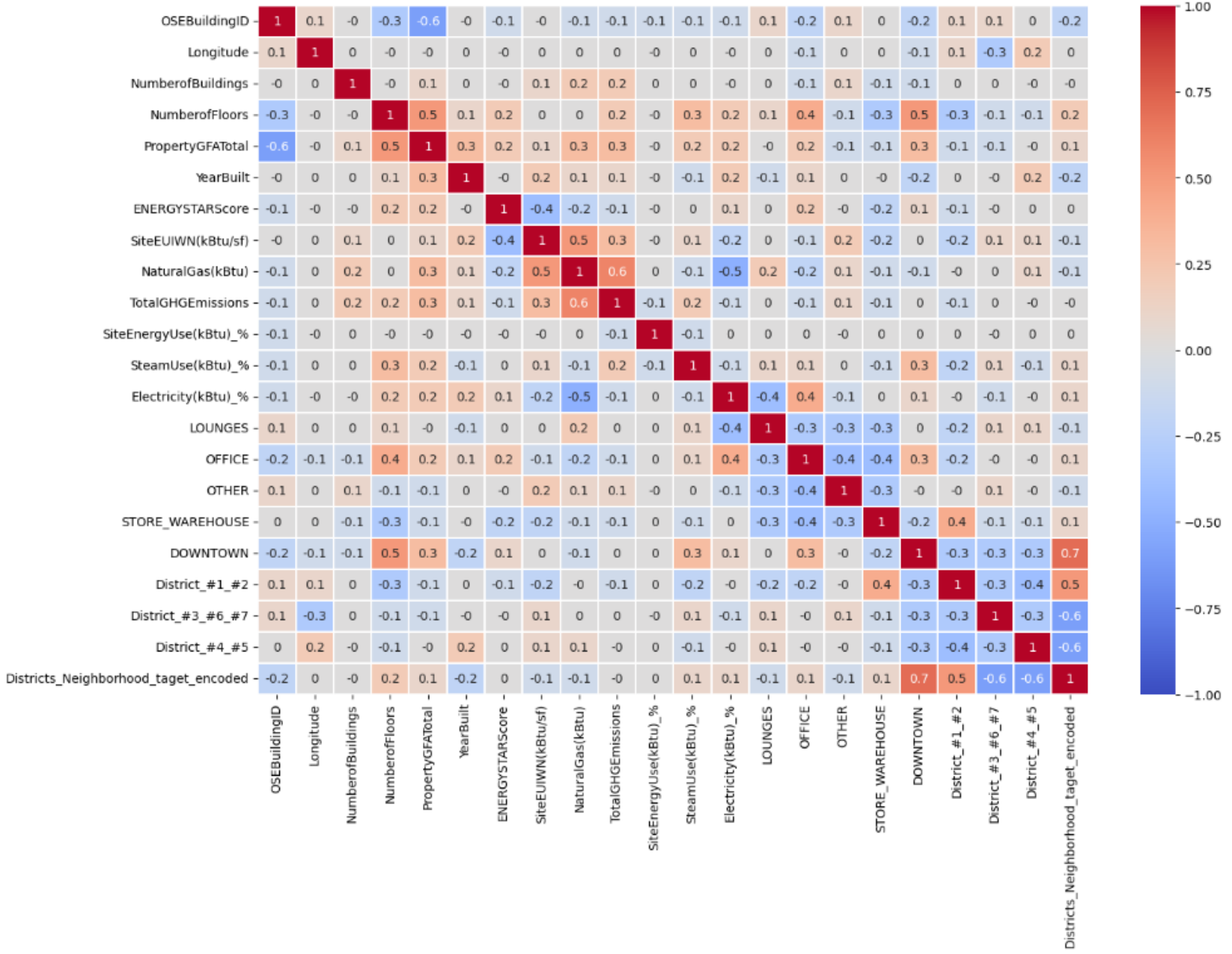
	statistic	pvalue
OSEBuildingID	-3.870363	1.086736e-04
Latitude	-1.180410	2.378373e-01
Longitude	5.257334	1.461589e-07
NumberofBuildings	26.861466	6.196514e-159
NumberofFloors	22.967644	9.819322e-117
PropertyGFATotal	21.436863	6.056084e-102
YearBuilt	-58.779498	0.000000e+00
ENERGYSTARScore	NaN	NaN
SiteEUI(kBtu/sf)	20.101590	7.146639e-90
SiteEUIWN(kBtu/sf)	19.853241	1.033167e-87
SourceEUI(kBtu/sf)	22.596315	4.710860e-113
SourceEUIWN(kBtu/sf)	22.570542	8.440504e-113
SiteEnergyUse(kBtu)	26.041624	1.673877e-149
SiteEnergyUseWN(kBtu)	26.160542	7.477281e-151
SteamUse(kBtu)	28.070455	2.248444e-173
Electricity(kBtu)	25.550718	5.390699e-144
NaturalGas(therms)	28.385036	3.094210e-177
NaturalGas(kBtu)	28.385036	3.094210e-177
TotalGHGEmissions	27.111277	7.250166e-162
GHGEmissionsIntensity	21.840817	9.503961e-106
SiteEnergyUse(kBtu)_%	29.567967	3.858866e-192
SteamUse(kBtu)_%	20.242551	4.132140e-91
Electricity(kBtu)_%	-22.279084	5.895154e-110
NaturalGas(kBtu)_%	-18.002428	1.864615e-72
Districts_Neighborhood_target	81.776213	0.000000e+00
PrimaryPropertyType_target	137.663686	0.000000e+00

ASYMÉTRIES

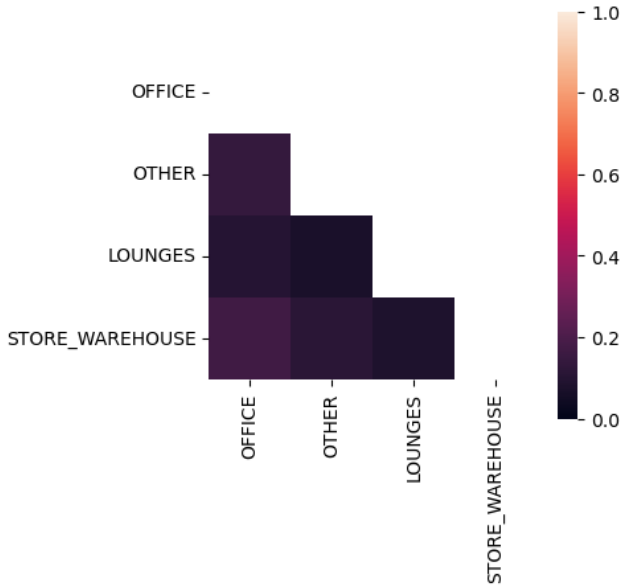
	Columns	Original_skew	Log_skew	SquareRoot_skew	CubeRoot_skew	Boxcox_skew
0	OSEBuildingID	0.293455	-0.827404	-0.331499	-0.452405	-0.448429
1	Latitude	0.257946	0.255293	0.256619	0.256177	0.000000
2	Longitude	-0.119200	NaN	NaN	-0.118645	0.000000
3	NumberofBuildings	13.510621	NaN	4.093332	-0.769598	NaN
4	NumberofFloors	5.867909	NaN	2.609776	1.592276	NaN
5	PropertyGFATotal	4.903511	1.034078	2.505290	1.913537	0.175036
6	YearBuilt	-0.227224	-0.244204	-0.235721	-0.238551	-0.075698
7	ENERGYSTARScore	-0.667878	-2.407296	-1.260024	-1.556984	1.358086
8	SiteEUIWN(kBtu/sf)	3.603128	NaN	1.395261	0.665306	NaN
9	SteamUse(kBtu)	20.547174	NaN	9.379922	5.719591	NaN
10	NaturalGas(kBtu)	21.203906	NaN	4.393886	1.324584	NaN
11	TotalGHGEmissions	14.586217	0.047308	4.734000	2.460896	-0.001522
12	GHGEmissionsIntensity	4.541240	-0.172930	1.634307	0.993154	-0.000564
13	SiteEnergyUse(kBtu)_%	-38.898905	-38.899723	-38.899315	-38.899451	NaN
14	SteamUse(kBtu)_%	4.740538	NaN	4.008736	3.778450	NaN
15	Electricity(kBtu)_%	-0.357223	NaN	-0.822874	-1.228787	NaN
16	NaturalGas(kBtu)_%	0.491263	NaN	-0.132409	-0.398561	NaN
17	LOUNGES	1.641696	NaN	1.641602	1.641602	NaN
18	OFFICE	0.775740	NaN	0.775879	0.775879	NaN
19	OTHER	1.252850	NaN	1.252930	1.252930	NaN
20	STORE_WAREHOUSE	1.077979	NaN	1.078125	1.078125	NaN
21	DOWNTOWN	1.314464	NaN	1.314453	1.314453	NaN
22	District_#1_#2	0.977301	NaN	0.977539	0.977539	NaN
23	District_#3_#6_#7	1.239975	NaN	1.240234	1.240234	NaN
24	District_#4_#5	1.112981	NaN	1.113281	1.113281	NaN
25	PrimaryPropertyType_taget_encoded	-0.449488	-0.904751	-0.665611	-0.744135	-0.355889
26	Districts_Neighborhood_taget_encoded	0.263698	0.098157	0.169468	0.143154	0.072882

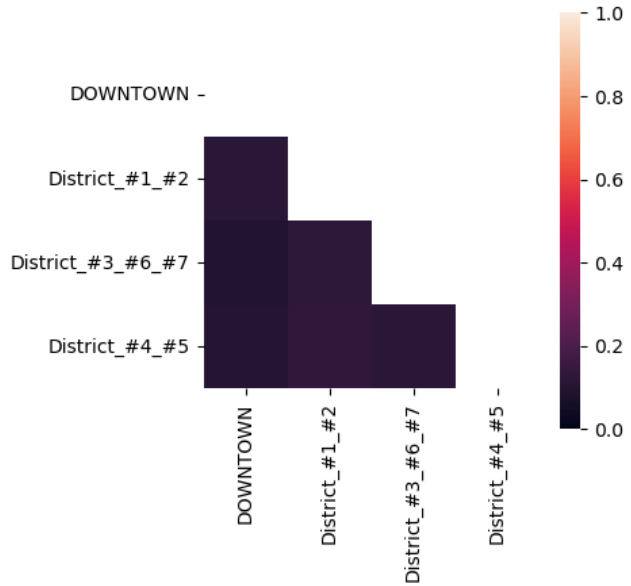
MATRICE DE CORRÉLATION PEARSON

Après avoir supprimer les features avec une corrélation supérieure à 0.7



MATRICE DE CORRELATION CRAMER'S V

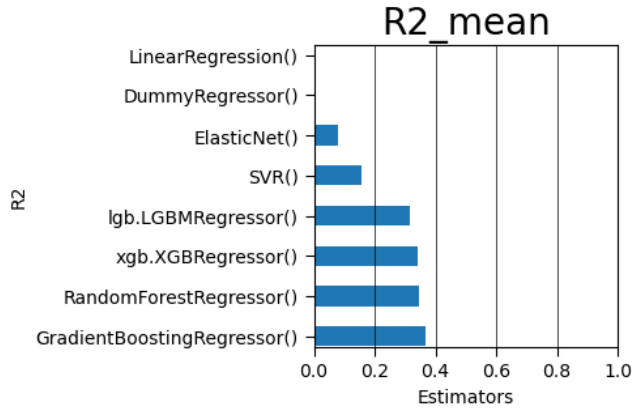




	OFFICE	OTHER	LOUNGES	STORE_WAREHOUSE
OFFICE	1.00	0.14	0.10	0.17
OTHER	0.14	1.00	0.07	0.11
LOUNGES	0.10	0.07	1.00	0.08
STORE_WAREHOUSE	0.17	0.11	0.08	1.00

	DOWNTOWN	District_#1_#2	District_#3_#6_#7	District_#4_#5
DOWNTOWN	1.00	0.11	0.09	0.10
District_#1_#2	0.11	1.00	0.12	0.13
District_#3_#6_#7	0.09	0.12	1.00	0.11
District_#4_#5	0.10	0.13	0.11	1.00

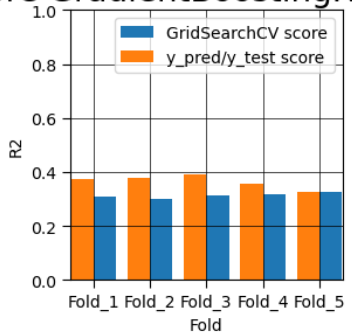
	Estimator	Fold_1_R2	Fold_2_R2	Fold_3_R2	Fold_4_R2	Fold_5_R2	R2_mean	Time_mean	RMSE_mean
5	GradientBoostingRegressor()	0.373450	0.377476	0.391481	0.358335	3.286906e-01	0.365886	3.674357	2.732659
4	RandomForestRegressor()	0.333687	0.373740	0.360351	0.349391	3.147489e-01	0.346384	1.614347	2.773943
6	xgb.XGBRegressor()	0.287628	0.377690	0.362300	0.337107	3.335936e-01	0.339664	0.615677	2.786240
7	lgb.LGBMRegressor()	0.294736	0.323149	0.354246	0.288417	3.160561e-01	0.315321	0.403780	2.838255
3	SVR()	0.184527	0.189626	0.165784	0.127647	1.210337e-01	0.157724	0.043672	3.149473
2	ElasticNet()	0.128253	0.129855	-0.040181	0.087964	7.895251e-02	0.076969	0.001497	3.298140
0	DummyRegressor()	-0.000300	-0.000935	-0.000256	-0.000003	-7.607102e-07	-0.000299	0.000396	3.433353
1	LinearRegression()	0.173266	0.207886	-2.692997	0.095604	8.066102e-02	-0.427116	0.000559	3.892538



SITEEUIWN(KBTU/SF)

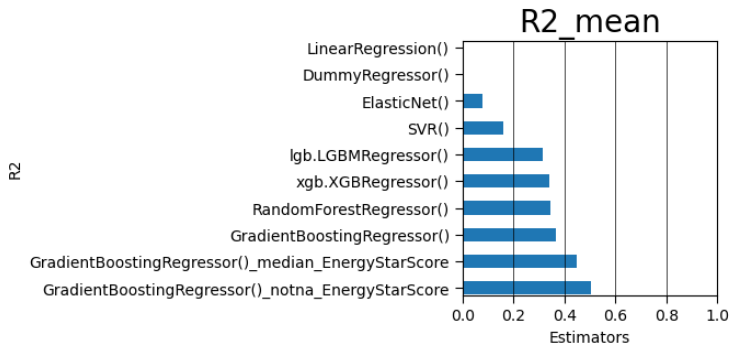
	Fold#	grid_best_score	score	loss	subsample	criterion	learning_rate	n_estimators	max_depth
0	Fold_1	0.309433	0.373450	huber	0.5	friedman_mse	0.01	500	5
1	Fold_2	0.302866	0.377476	huber	0.5	friedman_mse	0.01	500	5
2	Fold_3	0.312706	0.391481	huber	0.5	friedman_mse	0.01	500	5
3	Fold_4	0.319686	0.358335	huber	0.5	friedman_mse	0.01	500	5
4	Fold_5	0.325314	0.328691	huber	0.5	friedman_mse	0.01	500	5

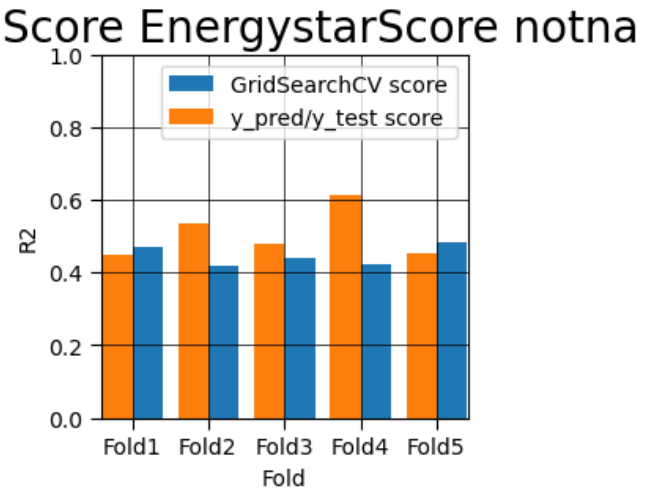
R2 Score GradientBoostingRegressor()



GRADIENT BOOSTING REGRESSOR GRID SEARCH POUR SITEEUIWN(KBTU/SF)

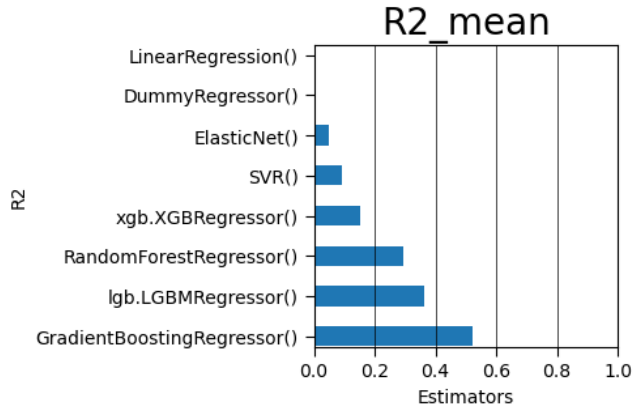
	Fold	shap_rank	shap_value
0	Fold_1	2	0.634627
1	Fold_2	2	0.626482
2	Fold_3	2	0.652988
3	Fold_4	2	0.550023
4	Fold_5	3	0.555273





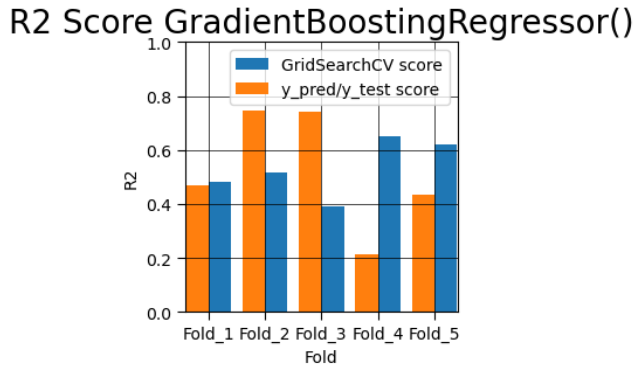
ENERGY STAR SCORE

	Estimator	Fold_1_R2	Fold_2_R2	Fold_3_R2	Fold_4_R2	Fold_5_R2	R2_mean	Time_mean	RMSE_mean
5	GradientBoostingRegressor()	0.468099	0.745385	0.743595	0.216205	0.436017	0.521860	3.671400	465.005383
7	lgb.LGBMRegressor()	0.344679	0.661110	0.330255	0.212863	0.261328	0.362047	0.425645	555.517500
4	RandomForestRegressor()	0.444415	-0.015612	0.556732	0.165978	0.317585	0.293819	1.639300	543.155057
6	xgb.XGBRegressor()	-0.064463	0.305794	-0.069132	0.169949	0.429513	0.154332	0.615753	616.426195
3	SVR()	0.084233	0.185115	0.070290	0.095613	0.022876	0.091625	0.038777	653.978915
2	ElasticNet()	0.222607	-0.421124	0.198053	0.174005	0.077244	0.050157	0.002118	639.934787
0	DummyRegressor()	-0.033886	-0.110610	-0.040804	-0.051509	-0.027596	-0.052881	0.000386	693.385175
1	LinearRegression()	0.282430	-2.045413	0.286818	0.138321	0.104039	-0.246761	0.000552	659.596415



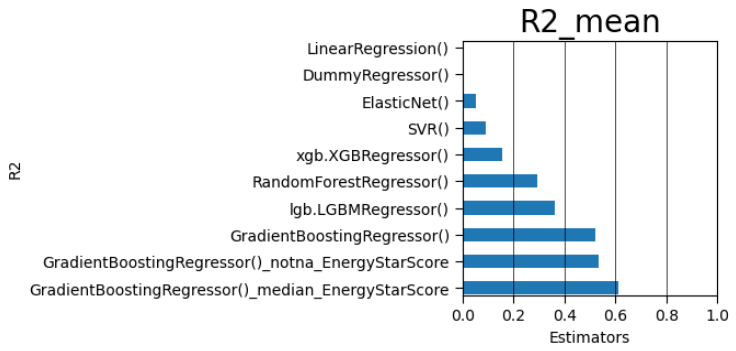
TOTALGHGEMISSIONS

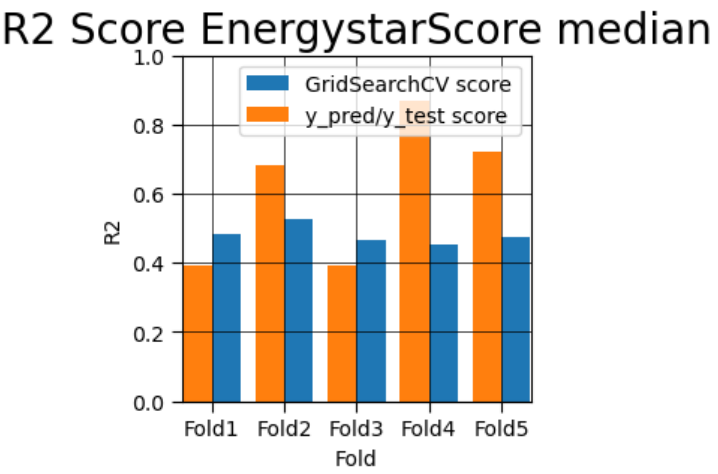
	Fold#	grid_best_score	score	loss	subsample	criterion	learning_rate	n_estimators	max_depth
0	Fold_1	0.481058	0.468099	huber	1.0	friedman_mse	0.01	500	5
1	Fold_2	0.519088	0.745385	huber	1.0	friedman_mse	0.01	500	5
2	Fold_3	0.392016	0.743595	huber	1.0	friedman_mse	0.01	500	5
3	Fold_4	0.652025	0.216205	squared_error	1.0	friedman_mse	0.01	500	3
4	Fold_5	0.622986	0.436017	squared_error	1.0	friedman_mse	0.01	500	5



GRADIENT BOOSTING REGRESSOR GRID SEARCH POUR TOTALGHGEMISSIONS

	Fold	shap_rank	shap_value
0	Fold_1	3	29.788801
1	Fold_2	3	27.710358
2	Fold_3	3	27.647438
3	Fold_4	3	21.751805
4	Fold_5	3	22.156282





ENERGY STAR SCORE

CONCLUSIONS

le modèle d'apprentissage n'est pas très performant

il s'améliore nettement en introduisant l'indicateur *Energy Star Score*

On ne remarque pas de surapprentissage su certaines *features*