

DATA SCIENTIST - PROJET 4

Anticipez les besoins en consommation de bâtiments

Arnaud CHOUX

Formation OPENCLASSROOMS
Financement région IdF



Septembre 2022

Plan

I Présentation du projet

II Nettoyage

III Traitement

IV Conclusions

Plan

I Présentation du projet

1. Objectifs

II Nettoyage

2. Compétences abordées

III Traitement

3. Dataset

IV Conclusions

I.1. Objectifs

1. Prédire les émissions de CO₂ et la consommation totale d'énergie de bâtiments non destinés à l'habitation pour lesquels les relevés sur place, coûteux, n'ont pas encore été réalisés.
2. Au cours de cette étude: Tester différents modèles de prédiction afin de répondre au mieux à la problématique.
3. Finalement, se servir de cette étude pour évaluer l'importance de l'indicateur "ENERGY STAR Score" qui est fastidieux à calculer et dont on aimerait donc se passer s'il influe peu sur nos prédictions.

I.2. Compétences abordées

- collecter des data (sans phase de recherche car la source est donnée)
- comprendre ce que représente en pratique chacun des indicateurs
- nettoyer la data
- fit et predict par régression
- comparer l'importance des indicateurs (shap)

1.3. Dataset

BuildingType	PrimaryPropertyType	PropertyName	Address	City	State	ZipCode	TaxParcelIdentificationNumber	CouncilDistrictCode	Neighborhood	Latitude	Longitude
NonResidential	Hotel	Mayflower park hotel	405 Olive way	Seattle	WA	98101.0	0659000030	7	DOWNTOWN	47.61220	-122.33799
NonResidential	Hotel	Paramount Hotel	724 Pine street	Seattle	WA	98101.0	0659000220	7	DOWNTOWN	47.61317	-122.33393
NonResidential	Hotel	5673-The Westin Seattle	1900 5th Avenue	Seattle	WA	98101.0	0659000475	7	DOWNTOWN	47.61393	-122.33810
NonResidential	Hotel	HOTEL MAX	620 STEWART ST	Seattle	WA	98101.0	0659000640	7	DOWNTOWN	47.61412	-122.33664
NonResidential	Hotel	WARWICK SEATTLE HOTEL (DOR)	401 LENORA ST	Seattle	WA	98121.0	0659000970	7	DOWNTOWN	47.61375	-122.34047

→ 3376 individus, 46indicateurs

Plan

I Présentation du projet

II Nettoyage

III Traitement

IV Conclusions

1. Indicateurs
catégoriels

2. Indicateurs
quantitatifs

3. Indicateurs vides

4. Suppression
d'éléments

5. Imputation

II.1. Indicateurs catégoriels

OSEBuildingID:	1	len([1])
DataYear:	1	len([3376])
BuildingType:	8	len([1460 1018 580 110 98 85 24 1])
PrimaryPropertyType:	22	len([987 564 293 256 187 173 139 133 105 91 77 71 53 45 40...])
PropertyName:	3	len([3 2 1])
Address:	3	len([4 2 1])
City:	1	len([3376])
State:	1	len([3376])
ZipCode:	28	len([294 251 243 230 191 186 169 167 161 152 129 124 101 99 93...])
TaxParcelIdentificationNumber:	6	len([8 5 4 3 2 1])
CouncilDistrictCode:	7	len([1037 596 509 367 338 282 247])
Neighborhood:	19	len([573 453 423 375 280 251 210 166 145 126 107 95 80 42 27...])
ListOfAllPropertyUseTypes:	36	len([866 464 139 135 120 101 61 56 52 51 48 44 43 40 26...])
LargestPropertyUseType:	27	len([1667 498 199 139 102 99 77 71 54 46 41 32...])
SecondLargestPropertyUseType:	21	len([976 215 155 59 40 33 18 17 14 13 12 11 10 8 7...])
ThirdLargestPropertyUseType:	18	len([110 105 71 56 49 29 18 17 14 12 11 9 8 6 5...])
YearsENERGYSTARCertified:	8	len([14 8 7 6 4 3 2 1])
DefaultData:	2	len([3263 113])
Comments:	0	len([])
ComplianceStatus:	4	len([3211 113 37 15])
Outlier:	2	len([23 9])

II.1.1. Mon « Indice de variance »

```
for co in cols_cat:
    vcu = df.loc[:,co].value_counts().unique()
    indice_de_variance = len(vcu)
```

```
OSEBuildingID:      1      len([1])
DataYear:           1      len([3376])
BuildingType:       8      len([1460 1018 580 110 98 85 24 1])
PrimaryPropertyType: 22     len([987 564 293 256 187 173 139 133 105 91 77 71 53 45 40...])
PropertyName:       3      len([3 2 1])
Address:            3      len([4 2 1])
City:               1      len([3376])
State:              1      len([3376])
ZipCode:            28     len([294 251 243 230 191 186 169 167 161 152 129 124 101 99 93...])
TaxParcelIdentificationNumber: 6     len([8 5 4 3 2 1])
CouncilDistrictCode: 7      len([1037 596 509 367 338 282 247])
Neighborhood:       19     len([573 453 423 375 280 251 210 166 145 126 107 95 80 42 27...])
ListOfAllPropertyUseTypes: 36     len([866 464 139 135 120 101 61 56 52 51 48 44 43 40 26...])
LargestPropertyUseType: 27     len([1667 498 199 139 102 99 77 71 54 46 41 32...])
SecondLargestPropertyUseType: 21     len([976 215 155 59 40 33 18 17 14 13 12 11 10 8 7...])
ThirdLargestPropertyUseType: 18     len([110 105 71 56 49 29 18 17 14 12 11 9 8 6 5...])
YearsENERGYSTARCertified: 8      len([14 8 7 6 4 3 2 1])
DefaultData:        2      len([3263 113])
Comments:           0      len([])
ComplianceStatus:   4      len([3211 113 37 15])
Outlier:            2      len([23 9])
```

II.1.2. Indicateurs sans info

```
for co in cols_cat:
    vcu = df.loc[:,co].value_counts().unique()
    indice_de_variance = len(vcu)
```

1 valeur unique
100% de valeurs uniques } → aucune info

OSEBuildingID:	1	len([1])
DataYear:	1	len([3376])
BuildingType:	8	len([1460 1018 580 110 98 85 24 1])
PrimaryPropertyType:	22	len([987 564 293 256 187 173 139 133 105 91 77 71 53 45 40...])
PropertyName:	3	len([3 2 1])
Address:	3	len([4 2 1])
City:	1	len([3376])
State:	1	len([3376])
ZipCode:	28	len([294 251 243 230 191 186 169 167 161 152 129 124 101 99 93...])
TaxParcelIdentificationNumber:	6	len([8 5 4 3 2 1])
CouncilDistrictCode:	7	len([1037 596 509 367 338 282 247])
Neighborhood:	19	len([573 453 423 375 280 251 210 166 145 126 107 95 80 42 27...])
ListOfAllPropertyUseTypes:	36	len([866 464 139 135 120 101 61 56 52 51 48 44 43 40 26...])
LargestPropertyUseType:	27	len([1667 498 199 139 102 99 77 71 54 46 41 32...])
SecondLargestPropertyUseType:	21	len([976 215 155 59 40 33 18 17 14 13 12 11 10 8 7...])
ThirdLargestPropertyUseType:	18	len([110 105 71 56 49 29 18 17 14 12 11 9 8 6 5...])
YearsENERGYSTARCertified:	8	len([14 8 7 6 4 3 2 1])
DefaultData:	2	len([3263 113])
Comments:	0	len([])
ComplianceStatus:	4	len([3211 113 37 15])
Outlier:	2	len([23 9])

→ suppression des indices < 2

II.1.3. Indicateurs presque sans info

```
for co in cols_cat:
    vcu = df.loc[:,co].value_counts().unique()
    indice_de_variance = len(vcu)
```

1 valeur unique
100% de valeurs uniques } → aucune info

OSEBuildingID:	1	len([1])
DataYear:	1	len([3376])
BuildingType:	8	len([1460 1018 580 110 98 85 24 1])
PrimaryPropertyType:	22	len([987 564 293 256 187 173 139 133 105 91 77 71 53 45 40...])
PropertyName:	3	len([3 2 1])
Address:	3	len([4 2 1])
City:	1	len([3376])
State:	1	len([3376])
ZipCode:	28	len([294 251 243 230 191 186 169 167 161 152 129 124 101 99 93...])
TaxParcelIdentificationNumber:	6	len([8 5 4 3 2 1])
CouncilDistrictCode:	7	len([1037 596 509 367 338 282 247])
Neighborhood:	19	len([573 453 423 375 280 251 210 166 145 126 107 95 80 42 27...])
ListOfAllPropertyUseTypes:	36	len([866 464 139 135 120 101 61 56 52 51 48 44 43 40 26...])
LargestPropertyUseType:	27	len([1667 498 199 139 102 99 77 71 54 46 41 32...])
SecondLargestPropertyUseType:	21	len([976 215 155 59 40 33 18 17 14 13 12 11 10 8 7...])
ThirdLargestPropertyUseType:	18	len([110 105 71 56 49 29 18 17 14 12 11 9 8 6 5...])
YearsENERGYSTARCertified:	8	len([14 8 7 6 4 3 2 1])
DefaultData:	2	len([3263 113])
Comments:	0	len([])
ComplianceStatus:	4	len([3211 113 37 15])
Outlier:	2	len([23 9])

→ suppression car tous uniques à 4 ou 5 voisins près.

II.1.4. Indicateurs inintéressants

```
for co in cols_cat:
    vcu = df.loc[:,co].value_counts().unique()
    indice_de_variance = len(vcu)
```

OSEBuildingID:	1	len([1])
DataYear:	1	len([3376])
BuildingType:	8	len([1460 1018 580 110 98 85 24 1])
PrimaryPropertyType:	22	len([987 564 293 256 187 173 139 133 105 91 77 71 53 45 40...])
PropertyName:	3	len([3 2 1])
Address:	3	len([4 2 1])
City:	1	len([3376])
State:	1	len([3376])
ZipCode:	28	len([294 251 243 230 191 186 169 167 161 152 129 124 101 99 93...])
TaxParcelIdentificationNumber:	6	len([8 5 4 3 2 1])
CouncilDistrictCode:	7	len([1037 596 509 367 338 282 247])
Neighborhood:	19	len([573 453 423 375 280 251 210 166 145 126 107 95 80 42 27...])
ListOfAllPropertyUseTypes:	36	len([866 464 139 135 120 101 61 56 52 51 48 44 43 40 26...])
LargestPropertyUseType:	27	len([1667 498 199 139 102 99 77 71 54 46 41 32...])
SecondLargestPropertyUseType:	21	len([976 215 155 59 40 33 18 17 14 13 12 11 10 8 7...])
ThirdLargestPropertyUseType:	18	len([110 105 71 56 49 29 18 17 14 12 11 9 8 6 5...])
YearsENERGYSTARCertified:	8	len([14 8 7 6 4 3 2 1])
DefaultData:	2	len([3263 113])
Comments:	0	len([])
ComplianceStatus:	4	len([3211 113 37 15])
Outlier:	2	len([23 9])

→ suppression d'informations moins utiles, inutiles, redondantes

II.1.5. Indicateurs conservés

```
for co in cols_cat:
    vcu = df.loc[:,co].value_counts().unique()
    indice_de_variance = len(vcu)
```

OSEBuildingID:	1	len([1])
DataYear:	1	len([3376])
BuildingType:	8	len([1460 1018 580 110 98 85 24 1])
PrimaryPropertyType:	22	len([987 564 293 256 187 173 139 133 105 91 77 71 53 45 40...])
PropertyName:	3	len([3 2 1])
Address:	3	len([4 2 1])
City:	1	len([3376])
State:	1	len([3376])
ZipCode:	28	len([294 251 243 230 191 186 169 167 161 152 129 124 101 99 93...])
TaxParcelIdentificationNumber:	6	len([8 5 4 3 2 1])
CouncilDistrictCode:	7	len([1037 596 509 367 338 282 247])
Neighborhood:	19	len([573 453 423 375 280 251 210 166 145 126 107 95 80 42 27...])
ListOfAllPropertyUseTypes:	36	len([866 464 139 135 120 101 61 56 52 51 48 44 43 40 26...])
LargestPropertyUseType:	27	len([1667 498 199 139 102 99 77 71 54 46 41 32...])
SecondLargestPropertyUseType:	21	len([976 215 155 59 40 33 18 17 14 13 12 11 10 8 7...])
ThirdLargestPropertyUseType:	18	len([110 105 71 56 49 29 18 17 14 12 11 9 8 6 5...])
YearsENERGYSTARCertified:	8	len([14 8 7 6 4 3 2 1])
DefaultData:	2	len([3263 113])
Comments:	0	len([])
ComplianceStatus:	4	len([3211 113 37 15])
Outlier:	2	len([23 9])

- BuildingType et PrimaryPropertyType semblent importants pour la régression.
- dans l'idéal: traitement des [...] pour éviter handle_unknown dans le OHE.

II.2. Indicateurs quantitatifs: corrélations

SecondLargestPropertyUseTypeGFA	Longitude	1.00	0.03	-0.03	0.03	0.03	0.03	0.02	0.03	0.01	-0.05	-0.03	0.04	0.03	0.03	-0.00	0.04	0.03	0.02	0.02	0.02	0.03	0.03	0.02	0.03	0.03
	PropertyGFATotal	0.03	1.00	0.40	0.80	0.85	0.18	0.81	0.99	-0.02	0.10	0.07	0.53	0.52	0.04	0.40	0.02	0.97	0.44	0.05	0.08	0.85	0.07	0.69	0.18	0.40
	NumberofFloors	-0.03	0.40	1.00	0.21	0.25	0.07	0.47	0.36	-0.02	0.15	0.02	0.14	0.22	-0.00	0.42	-0.04	0.34	0.08	0.03	0.04	0.25	0.01	-0.03	0.07	0.29
	SiteEnergyUse(kBtu)	0.03	0.80	0.21	1.00	0.96	0.51	0.63	0.81	-0.02	0.03	-0.09	0.86	0.75	0.27	0.17	0.31	0.84	0.60	0.27	0.30	0.96	0.30	0.69	0.51	0.72
	Electricity(kBtu)	0.03	0.85	0.25	0.96	1.00	0.29	0.63	0.86	-0.02	0.04	-0.06	0.69	0.68	0.25	0.22	0.18	0.88	0.55	0.29	0.32	1.00	0.29	0.74	0.29	0.59
	NaturalGas(therms)	0.03	0.18	0.07	0.51	0.29	1.00	0.39	0.18	-0.02	0.02	-0.10	0.73	0.65	0.26	0.06	0.49	0.20	0.03	0.18	0.18	0.29	0.26	0.06	1.00	0.73
	PropertyGFABuilding(s)	0.02	0.81	0.47	0.63	0.63	0.39	1.00	0.79	-0.05	0.20	0.08	0.51	0.66	0.08	0.48	0.11	0.77	0.26	0.10	0.11	0.63	0.10	0.11	0.39	0.63
	PropertyGFABuilding(s)	0.03	0.99	0.36	0.81	0.86	0.18	0.79	1.00	-0.02	0.08	0.06	0.55	0.54	0.03	0.27	0.03	0.98	0.46	0.04	0.07	0.86	0.06	0.73	0.18	0.38
	Latitude	0.01	-0.02	-0.02	-0.02	-0.02	-0.02	-0.05	-0.02	1.00	0.12	0.08	-0.03	-0.11	-0.02	-0.00	-0.04	-0.02	-0.02	-0.00	-0.00	-0.02	-0.01	0.02	-0.02	-0.04
	YearBuilt	-0.05	0.10	0.15	0.03	0.04	0.02	0.20	0.08	0.12	1.00	0.03	0.01	0.09	-0.03	0.18	-0.15	0.07	-0.02	0.04	0.04	0.04	-0.02	-0.02	0.02	0.07
ThirdLargestPropertyUseTypeGFA	ENERGYSTARScore	-0.03	0.07	0.02	-0.09	-0.06	-0.10	0.08	0.06	0.08	0.03	1.00	-0.10	-0.01	-0.35	0.05	-0.27	0.06	-0.04	-0.31	-0.31	-0.06	-0.34	-0.00	-0.10	-0.09
	TotalGHGEmissions	0.04	0.53	0.14	0.86	0.69	0.73	0.51	0.55	-0.03	0.01	-0.10	1.00	0.68	0.27	0.09	0.47	0.58	0.68	0.22	0.23	0.69	0.29	0.41	0.73	0.86
	PropertyGFABuilding(s)	0.03	0.52	0.22	0.75	0.68	0.65	0.66	0.54	-0.11	0.09	-0.01	0.68	1.00	0.13	0.21	0.33	0.46	0.04	0.14	0.14	0.68	0.13	0.00	0.65	0.75
	SiteEUIWN(kBtu/sf)	0.03	0.04	-0.00	0.27	0.25	0.26	0.08	0.03	-0.02	-0.03	-0.35	0.27	0.13	1.00	0.09	0.75	0.03	0.09	0.94	0.93	0.25	0.99	0.01	0.26	0.39
	PropertyGFAParking	-0.00	0.40	0.42	0.17	0.22	0.06	0.48	0.27	-0.00	0.18	0.05	0.09	0.21	0.09	1.00	-0.04	0.30	0.01	0.13	0.13	0.22	0.10	-0.00	0.06	0.24
	GHGEmissionsIntensity	0.04	0.02	-0.04	0.31	0.18	0.49	0.11	0.03	-0.04	-0.15	-0.27	0.47	0.33	0.75	-0.04	1.00	0.05	0.19	0.53	0.52	0.18	0.73	0.03	0.49	0.43
	LargestPropertyUseTypeGFA	0.03	0.97	0.34	0.84	0.88	0.20	0.77	0.98	-0.02	0.07	0.06	0.58	0.46	0.03	0.30	0.05	1.00	0.50	0.03	0.06	0.88	0.06	0.76	0.20	0.39
	SteamUse(kBtu)	0.02	0.44	0.08	0.60	0.55	0.03	0.26	0.46	-0.02	-0.02	-0.04	0.68	0.04	0.09	0.01	0.19	0.50	1.00	0.08	0.09	0.55	0.11	0.40	0.03	0.47
	SourceEUIWN(kBtu/sf)	0.02	0.05	0.03	0.27	0.29	0.18	0.10	0.04	-0.00	0.04	-0.31	0.22	0.14	0.94	0.13	0.53	0.03	0.08	1.00	0.99	0.29	0.94	0.00	0.18	0.39
	SourceEUI(kBtu/sf)	0.02	0.08	0.04	0.30	0.32	0.18	0.11	0.07	-0.00	0.04	-0.31	0.23	0.14	0.93	0.13	0.52	0.06	0.09	0.99	1.00	0.32	0.94	0.03	0.18	0.39
1stPropertyUseTypeGFA	Electricity(kWh)	0.03	0.85	0.25	0.96	1.00	0.29	0.63	0.86	-0.02	0.04	-0.06	0.69	0.68	0.25	0.22	0.18	0.88	0.55	0.29	0.32	1.00	0.29	0.74	0.29	0.59
	SiteEUI(kBtu/sf)	0.03	0.07	0.01	0.30	0.29	0.26	0.10	0.06	-0.01	-0.02	-0.34	0.29	0.13	0.99	0.10	0.73	0.06	0.11	0.94	0.94	0.29	1.00	0.03	0.26	0.40
	NumberofBuildings	0.02	0.69	-0.03	0.69	0.74	0.06	0.11	0.73	0.02	-0.02	-0.00	0.41	0.00	0.01	-0.00	0.03	0.76	0.40	0.00	0.03	0.74	0.03	1.00	0.06	0.09
	NaturalGas(kBtu)	0.03	0.18	0.07	0.51	0.29	1.00	0.39	0.18	-0.02	0.02	-0.10	0.73	0.65	0.26	0.06	0.49	0.20	0.03	0.18	0.18	0.29	0.26	0.06	1.00	0.73
	SiteEnergyUseWN(kBtu)	0.03	0.40	0.29	0.72	0.59	0.73	0.63	0.38	-0.04	0.07	-0.09	0.86	0.75	0.39	0.24	0.43	0.39	0.47	0.39	0.39	0.59	0.40	0.09	0.73	1.00
	Longitude	Longitude	PropertyGFATotal	NumberofFloors	SiteEnergyUse(kBtu)	Electricity(kBtu)	NaturalGas(therms)	1stPropertyUseTypeGFA	PropertyGFABuilding(s)	Latitude	YearBuilt	ENERGYSTARScore	TotalGHGEmissions	1stPropertyUseTypeGFA	SiteEUIWN(kBtu/sf)	PropertyGFAParking	GHGEmissionsIntensity	1stPropertyUseTypeGFA	SteamUse(kBtu)	SourceEUIWN(kBtu/sf)	SourceEUI(kBtu/sf)	Electricity(kWh)	SiteEUI(kBtu/sf)	NumberofBuildings	NaturalGas(kBtu)	SiteEnergyUseWN(kBtu)

II.2.1. Targets et feature particulière

SecondLargestPropertyUseTypeGFA	Longitude	1.00	0.03	-0.03	0.03	0.03	0.03	0.02	0.03	0.01	-0.05	-0.03	0.04	0.03	0.03	-0.00	0.04	0.03	0.02	0.02	0.02	0.03	0.03	0.02	0.03	0.03
	PropertyGFATotal	0.03	1.00	0.40	0.80	0.85	0.18	0.81	0.99	-0.02	0.10	0.07	0.53	0.52	0.04	0.40	0.02	0.97	0.44	0.05	0.08	0.85	0.07	0.69	0.18	0.40
	NumberofFloors	-0.03	0.40	1.00	0.21	0.25	0.07	0.47	0.36	-0.02	0.15	0.02	0.14	0.22	-0.00	0.42	-0.04	0.34	0.08	0.03	0.04	0.25	0.01	-0.03	0.07	0.29
	SiteEnergyUse(kBtu)	0.03	0.80	0.21	1.00	0.96	0.51	0.63	0.81	-0.02	0.03	-0.09	0.86	0.75	0.27	0.17	0.31	0.84	0.60	0.27	0.30	0.96	0.30	0.69	0.51	0.72
	Electricity(kBtu)	0.03	0.85	0.25	0.96	1.00	0.29	0.63	0.86	-0.02	0.04	-0.06	0.69	0.68	0.25	0.22	0.18	0.88	0.55	0.29	0.32	1.00	0.29	0.74	0.29	0.59
	NaturalGas(therms)	0.03	0.18	0.07	0.51	0.29	1.00	0.39	0.18	-0.02	0.02	-0.10	0.73	0.65	0.26	0.06	0.49	0.20	0.03	0.18	0.18	0.29	0.26	0.06	1.00	0.73
	PropertyGFABuilding(s)	0.02	0.81	0.47	0.63	0.63	0.39	1.00	0.79	-0.05	0.20	0.08	0.51	0.66	0.08	0.48	0.11	0.77	0.26	0.10	0.11	0.63	0.10	0.11	0.39	0.63
	PropertyGFABuilding(s)	0.03	0.99	0.36	0.81	0.86	0.18	0.79	1.00	-0.02	0.08	0.06	0.55	0.54	0.03	0.27	0.03	0.98	0.46	0.04	0.07	0.86	0.06	0.73	0.18	0.38
	Latitude	0.01	-0.02	-0.02	-0.02	-0.02	-0.02	-0.05	-0.02	1.00	0.12	0.08	-0.03	-0.11	-0.02	-0.00	-0.04	-0.02	-0.02	-0.00	-0.00	-0.02	-0.01	0.02	-0.02	-0.04
	YearBuilt	-0.05	0.10	0.15	0.03	0.04	0.02	0.20	0.08	0.12	1.00	0.03	0.01	0.09	-0.03	0.18	-0.15	0.07	-0.02	0.04	0.04	0.04	-0.02	-0.02	0.02	0.07
ThirdLargestPropertyUseTypeGFA	ENERGYSTARScore	-0.03	0.07	0.02	-0.09	-0.06	-0.10	0.08	0.06	0.08	0.03	1.00	-0.10	-0.01	0.35	0.05	0.27	0.06	-0.04	-0.31	-0.31	-0.06	-0.34	-0.00	-0.10	-0.09
	TotalGHGEmissions	0.04	0.53	0.14	0.86	0.69	0.73	0.51	0.55	-0.03	0.01	-0.10	1.00	0.68	0.27	0.09	0.47	0.58	0.68	0.22	0.23	0.69	0.29	0.41	0.73	0.86
	PropertyGFAParking	0.03	0.52	0.22	0.75	0.68	0.65	0.66	0.54	-0.11	0.09	-0.01	0.68	1.00	0.13	0.21	0.33	0.46	0.04	0.14	0.14	0.68	0.13	0.00	0.65	0.75
	SiteEUIWN(kBtu/sf)	0.03	0.04	-0.00	0.27	0.25	0.26	0.08	0.03	-0.02	-0.03	0.35	0.27	0.13	1.00	0.09	0.75	0.03	0.09	0.94	0.93	0.25	0.99	0.01	0.26	0.39
	PropertyGFAParking	-0.00	0.40	0.42	0.17	0.22	0.06	0.48	0.27	-0.00	0.18	0.05	0.09	0.21	0.09	1.00	0.04	0.30	0.01	0.13	0.13	0.22	0.10	-0.00	0.06	0.24
LargestPropertyUseTypeGFA	GHGEmissionsIntensity	0.04	0.02	-0.04	0.31	0.18	0.49	0.11	0.03	-0.04	-0.15	-0.27	0.47	0.33	0.75	-0.04	1.00	0.05	0.19	0.53	0.52	0.18	0.73	0.03	0.49	0.43
	PropertyGFAParking	0.03	0.97	0.34	0.84	0.88	0.20	0.77	0.98	-0.02	0.07	0.06	0.58	0.46	0.03	0.30	0.05	1.00	0.50	0.03	0.06	0.88	0.06	0.76	0.20	0.39
	SteamUse(kBtu)	0.02	0.44	0.08	0.60	0.55	0.03	0.26	0.46	-0.02	-0.02	-0.04	0.68	0.04	0.09	0.01	0.19	0.50	1.00	0.08	0.09	0.55	0.11	0.40	0.03	0.47
	SourceEUIWN(kBtu/sf)	0.02	0.05	0.03	0.27	0.29	0.18	0.10	0.04	-0.00	0.04	-0.31	0.22	0.14	0.94	0.13	0.53	0.03	0.08	1.00	0.99	0.29	0.94	0.00	0.18	0.39
	SourceEUI(kBtu/sf)	0.02	0.08	0.04	0.30	0.32	0.18	0.11	0.07	-0.00	0.04	-0.31	0.23	0.14	0.93	0.13	0.52	0.06	0.09	0.99	1.00	0.32	0.94	0.03	0.18	0.39
	Electricity(kWh)	0.03	0.85	0.25	0.96	1.00	0.29	0.63	0.86	-0.02	0.04	-0.06	0.69	0.68	0.25	0.22	0.18	0.88	0.55	0.29	0.32	1.00	0.29	0.74	0.29	0.59
	SiteEUI(kBtu/sf)	0.03	0.07	0.01	0.30	0.29	0.26	0.10	0.06	-0.01	-0.02	-0.34	0.29	0.13	0.99	0.10	0.73	0.06	0.11	0.94	0.94	0.29	1.00	0.03	0.26	0.40
	NumberofBuildings	0.02	0.69	-0.03	0.69	0.74	0.06	0.11	0.73	0.02	-0.02	-0.00	0.41	0.00	0.01	-0.00	0.03	0.76	0.40	0.00	0.03	0.74	0.03	1.00	0.06	0.09
	NaturalGas(kBtu)	0.03	0.18	0.07	0.51	0.29	1.00	0.39	0.18	-0.02	0.02	-0.10	0.73	0.65	0.26	0.06	0.49	0.20	0.03	0.18	0.18	0.29	0.26	0.06	1.00	0.73
	SiteEnergyUseWN(kBtu)	0.03	0.40	0.29	0.72	0.59	0.73	0.63	0.38	-0.04	0.07	-0.09	0.86	0.75	0.39	0.24	0.43	0.39	0.47	0.39	0.39	0.59	0.40	0.09	0.73	1.00
	Longitude	PropertyGFATotal	NumberofFloors	SiteEnergyUse(kBtu)	Electricity(kBtu)	NaturalGas(therms)	1stPropertyUseTypeGFA	PropertyGFABuilding(s)	Latitude	YearBuilt	ENERGYSTARScore	TotalGHGEmissions	1stPropertyUseTypeGFA	SiteEUIWN(kBtu/sf)	PropertyGFAParking	GHGEmissionsIntensity	1stPropertyUseTypeGFA	SteamUse(kBtu)	SourceEUIWN(kBtu/sf)	SourceEUI(kBtu/sf)	Electricity(kWh)	SiteEUI(kBtu/sf)	NumberofBuildings	NaturalGas(kBtu)	SiteEnergyUseWN(kBtu)	

II.2.2. Features inintéressantes

	Longitude	1.00	0.03	-0.03	0.03	0.03	0.03	0.02	0.03	0.01	-0.05	-0.03	0.04	0.03	0.03	-0.00	0.04	0.03	0.02	0.02	0.02	0.03	0.03	0.02	0.03	0.03
	PropertyGFATotal	0.03	1.00	0.40	0.80	0.85	0.18	0.81	0.99	-0.02	0.10	0.07	0.53	0.52	0.04	0.40	0.02	0.97	0.44	0.05	0.08	0.85	0.07	0.69	0.18	0.40
	NumberofFloors	-0.03	0.40	1.00	0.21	0.25	0.07	0.47	0.36	-0.02	0.15	0.02	0.14	0.22	-0.00	0.42	-0.04	0.34	0.08	0.03	0.04	0.25	0.01	-0.03	0.07	0.29
	SiteEnergyUse(kBtu)	0.03	0.80	0.21	1.00	0.96	0.51	0.63	0.81	-0.02	0.03	-0.09	0.86	0.75	0.27	0.17	0.31	0.84	0.60	0.27	0.30	0.96	0.30	0.69	0.51	0.72
	Electricity(kBtu)	0.03	0.85	0.25	0.96	1.00	0.29	0.63	0.86	-0.02	0.04	-0.06	0.69	0.68	0.25	0.22	0.18	0.88	0.55	0.29	0.32	1.00	0.29	0.74	0.29	0.59
	NaturalGas(therms)	0.03	0.18	0.07	0.51	0.29	1.00	0.39	0.18	-0.02	0.02	-0.10	0.73	0.65	0.26	0.06	0.49	0.20	0.03	0.18	0.18	0.29	0.26	0.06	1.00	0.73
SecondLargestPropertyUseTypeGFA		0.02	0.81	0.47	0.63	0.63	0.39	1.00	0.79	-0.05	0.20	0.08	0.51	0.66	0.08	0.48	0.11	0.77	0.26	0.10	0.11	0.63	0.10	0.11	0.39	0.63
	PropertyGFABuilding(s)	0.03	0.99	0.36	0.81	0.86	0.18	0.79	1.00	-0.02	0.08	0.06	0.55	0.54	0.03	0.27	0.03	0.98	0.46	0.04	0.07	0.86	0.06	0.73	0.18	0.38
	Latitude	0.01	-0.02	-0.02	-0.02	-0.02	-0.02	-0.05	-0.02	1.00	0.12	0.08	-0.03	-0.11	-0.02	-0.00	-0.04	-0.02	-0.02	-0.00	-0.00	-0.02	-0.01	0.02	-0.02	-0.04
	YearBuilt	-0.05	0.10	0.15	0.03	0.04	0.02	0.20	0.08	0.12	1.00	0.03	0.01	0.09	-0.03	0.18	-0.15	0.07	-0.02	0.04	0.04	0.04	-0.02	-0.02	0.02	0.07
	ENERGYSTARScore	-0.03	0.07	0.02	-0.09	-0.06	-0.10	0.08	0.06	0.08	0.03	1.00	-0.10	-0.01	-0.35	0.05	-0.27	0.06	-0.04	-0.31	-0.31	-0.06	-0.34	-0.00	-0.10	-0.09
	TotalGHGEmissions	0.04	0.53	0.14	0.86	0.69	0.73	0.51	0.55	-0.03	0.01	-0.10	1.00	0.68	0.27	0.09	0.47	0.58	0.68	0.22	0.23	0.69	0.29	0.41	0.73	0.86
ThirdLargestPropertyUseTypeGFA		0.03	0.52	0.22	0.75	0.68	0.65	0.66	0.54	-0.11	0.09	-0.01	0.68	1.00	0.13	0.21	0.33	0.46	0.04	0.14	0.14	0.68	0.13	0.00	0.65	0.75
	SiteEUIWN(kBtu/sf)	0.03	0.04	-0.00	0.27	0.25	0.26	0.08	0.03	-0.02	-0.03	-0.35	0.27	0.13	1.00	0.09	0.75	0.03	0.09	0.94	0.93	0.25	0.99	0.01	0.26	0.39
	PropertyGFAParking	-0.00	0.40	0.42	0.17	0.22	0.06	0.48	0.27	-0.00	0.18	0.05	0.09	0.21	0.09	1.00	-0.04	0.30	0.01	0.13	0.13	0.22	0.10	0.00	0.06	0.24
	GHGEmissionsIntensity	0.04	0.02	-0.04	0.31	0.18	0.49	0.11	0.03	-0.04	-0.15	-0.27	0.47	0.33	0.75	-0.04	1.00	0.05	0.19	0.53	0.52	0.18	0.73	0.03	0.49	0.43
LargestPropertyUseTypeGFA		0.03	0.97	0.34	0.84	0.88	0.20	0.77	0.98	-0.02	0.07	0.06	0.58	0.46	0.03	0.30	0.05	1.00	0.50	0.03	0.06	0.88	0.06	0.76	0.20	0.39
	SteamUse(kBtu)	0.02	0.44	0.08	0.60	0.55	0.03	0.26	0.46	-0.02	-0.02	-0.04	0.68	0.04	0.09	0.01	0.19	0.50	1.00	0.08	0.09	0.55	0.11	0.40	0.03	0.47
	SourceEUIWN(kBtu/sf)	0.02	0.05	0.03	0.27	0.29	0.18	0.10	0.04	-0.00	0.04	-0.31	0.22	0.14	0.94	0.13	0.53	0.03	0.08	1.00	0.99	0.29	0.94	0.00	0.18	0.39
	SourceEUI(kBtu/sf)	0.02	0.08	0.04	0.30	0.32	0.18	0.11	0.07	-0.00	0.04	-0.31	0.23	0.14	0.93	0.13	0.52	0.06	0.09	0.99	1.00	0.32	0.94	0.03	0.18	0.39
	Electricity(kWh)	0.03	0.85	0.25	0.96	1.00	0.29	0.63	0.86	-0.02	0.04	-0.06	0.69	0.68	0.25	0.22	0.18	0.88	0.55	0.29	0.32	1.00	0.29	0.74	0.29	0.59
	SiteEUI(kBtu/sf)	0.03	0.07	0.01	0.30	0.29	0.26	0.10	0.06	-0.01	-0.02	-0.34	0.29	0.13	0.99	0.10	0.73	0.06	0.11	0.94	0.94	0.29	1.00	0.03	0.26	0.40
	NumberofBuildings	0.02	0.69	-0.03	0.69	0.74	0.06	0.11	0.73	-0.02	-0.02	-0.00	0.41	0.00	0.01	-0.00	0.03	0.76	0.40	0.00	0.03	0.74	0.03	1.00	0.06	0.09
	NaturalGas(kBtu)	0.03	0.18	0.07	0.51	0.29	1.00	0.39	0.18	-0.02	0.02	-0.10	0.73	0.65	0.26	0.06	0.49	0.20	0.03	0.18	0.18	0.29	0.26	0.06	1.00	0.73
	SiteEnergyUseWN(kBtu)	0.03	0.40	0.29	0.72	0.59	0.73	0.63	0.38	-0.04	0.07	-0.09	0.86	0.75	0.39	0.24	0.43	0.39	0.47	0.39	0.39	0.59	0.40	0.09	0.73	1.00
	Longitude																									
	PropertyGFATotal																									
	NumberofFloors																									
	SiteEnergyUse(kBtu)																									
	Electricity(kBtu)																									
	NaturalGas(therms)																									
	istPropertyUseTypeGFA																									
	PropertyGFABuilding(s)																									
	Latitude																									
	YearBuilt																									
	ENERGYSTARScore																									
	TotalGHGEmissions																									
	istPropertyUseTypeGFA																									
	SiteEUIWN(kBtu/sf)																									
	PropertyGFAParking																									
	GHGEmissionsIntensity																									
	istPropertyUseTypeGFA																									
	SteamUse(kBtu)																									
	SourceEUIWN(kBtu/sf)																									
	SourceEUI(kBtu/sf)																									
	Electricity(kWh)																									
	SiteEUI(kBtu/sf)																									
	NumberofBuildings																									
	NaturalGas(kBtu)																									
	SiteEnergyUseWN(kBtu)																									

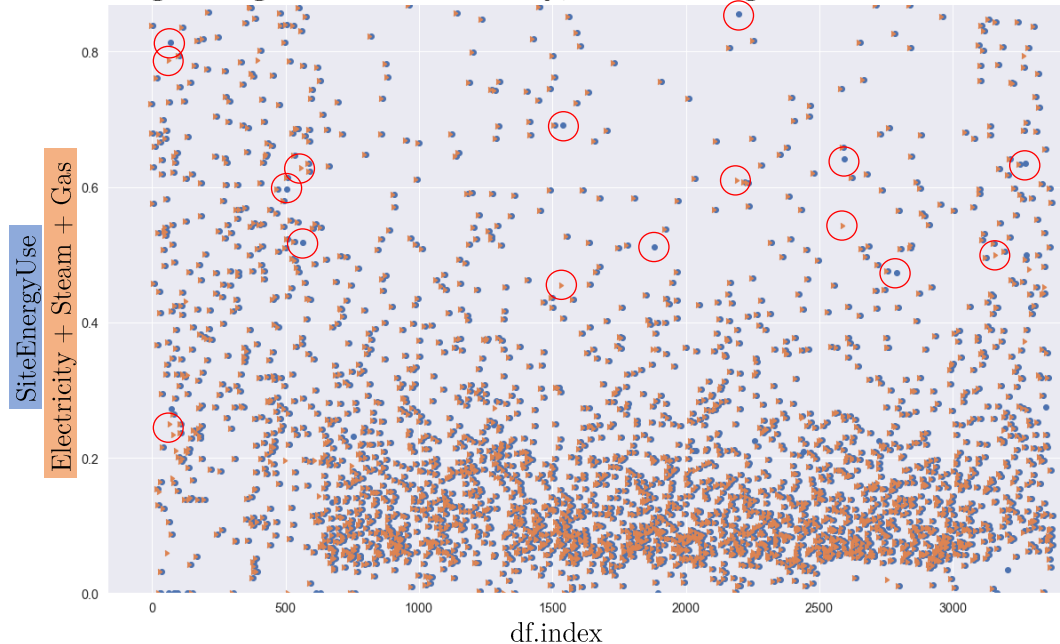
II.2.3. Features redondantes

	Longitude	1.00	0.03	-0.03	0.03	0.03	0.03	0.02	0.03	0.01	-0.05	-0.03	0.04	0.03	0.03	-0.00	0.04	0.03	0.02	0.02	0.02	0.03	0.03	0.02	0.03	0.03
	PropertyGFATotal	0.03	1.00	0.40	0.80	0.85	0.18	0.81	0.99	-0.02	0.10	0.07	0.53	0.52	0.04	0.40	0.02	0.97	0.44	0.05	0.08	0.85	0.07	0.69	0.18	0.40
	NumberofFloors	-0.03	0.40	1.00	0.21	0.25	0.07	0.47	0.36	-0.02	0.15	0.02	0.14	0.22	-0.00	0.42	-0.04	0.34	0.08	0.03	0.04	0.25	0.01	-0.03	0.07	0.29
	SiteEnergyUse(kBtu)	0.03	0.80	0.21	1.00	0.96	0.51	0.63	0.81	-0.02	0.03	-0.09	0.86	0.75	0.27	0.17	0.31	0.84	0.60	0.27	0.30	0.96	0.30	0.69	0.51	0.72
	Electricity(kBtu)	0.03	0.85	0.25	0.96	1.00	0.29	0.63	0.86	-0.02	0.04	-0.06	0.69	0.68	0.25	0.22	0.18	0.88	0.55	0.29	0.32	1.00	0.29	0.74	0.29	0.59
	NaturalGas(therms)	0.03	0.18	0.07	0.51	0.29	1.00	0.39	0.18	-0.02	0.02	-0.10	0.73	0.65	0.26	0.06	0.49	0.20	0.03	0.18	0.18	0.29	0.26	0.06	1.00	0.73
SecondLargestPropertyUseTypeGFA		0.02	0.81	0.47	0.63	0.63	0.39	1.00	0.79	-0.05	0.20	0.08	0.51	0.66	0.08	0.48	0.11	0.77	0.26	0.10	0.11	0.63	0.10	0.11	0.39	0.63
	PropertyGFABuilding(s)	0.03	0.99	0.36	0.81	0.86	0.18	0.79	1.00	-0.02	0.08	0.06	0.55	0.54	0.03	0.27	0.03	0.98	0.46	0.04	0.07	0.86	0.06	0.73	0.18	0.38
	Latitude	0.01	-0.02	-0.02	-0.02	-0.02	-0.02	-0.05	-0.02	1.00	0.12	0.08	-0.03	-0.11	-0.02	-0.00	-0.04	-0.02	-0.02	-0.00	-0.00	-0.02	-0.01	0.02	-0.02	-0.04
	YearBuilt	-0.05	0.10	0.15	0.03	0.04	0.02	0.20	0.08	0.12	1.00	0.03	0.01	0.09	-0.03	0.18	-0.15	0.07	-0.02	0.04	0.04	0.04	-0.02	-0.02	0.02	0.07
	ENERGYSTARScore	-0.03	0.07	0.02	-0.09	-0.06	-0.10	0.08	0.06	0.08	0.03	1.00	-0.10	-0.01	-0.35	0.05	-0.27	0.06	-0.04	-0.31	-0.31	-0.06	-0.34	-0.00	-0.10	-0.09
	TotalGHGEmissions	0.04	0.53	0.14	0.86	0.69	0.73	0.51	0.55	-0.03	0.01	-0.10	1.00	0.68	0.27	0.09	0.47	0.58	0.68	0.22	0.23	0.69	0.29	0.41	0.73	0.86
ThirdLargestPropertyUseTypeGFA		0.03	0.52	0.22	0.75	0.68	0.65	0.66	0.54	-0.11	0.09	-0.01	0.68	1.00	0.13	0.21	0.33	0.46	0.04	0.14	0.14	0.68	0.13	0.00	0.65	0.75
	SiteEUIWN(kBtu/sf)	0.03	0.04	-0.00	0.27	0.25	0.26	0.08	0.03	-0.02	-0.03	-0.35	0.27	0.13	1.00	0.09	0.75	0.03	0.09	0.94	0.93	0.25	0.99	0.01	0.26	0.39
	PropertyGFAParking	-0.00	0.40	0.42	0.17	0.22	0.06	0.48	0.27	-0.00	0.18	0.05	0.09	0.21	0.09	1.00	-0.04	0.30	0.01	0.13	0.13	0.22	0.10	-0.00	0.06	0.24
	GHGEmissionsIntensity	0.04	0.02	-0.04	0.31	0.18	0.49	0.11	0.03	-0.04	-0.15	-0.27	0.47	0.33	0.75	-0.04	1.00	0.05	0.19	0.53	0.52	0.18	0.73	0.03	0.49	0.43
LargestPropertyUseTypeGFA		0.03	0.97	0.34	0.84	0.88	0.20	0.77	0.98	-0.02	0.07	0.06	0.58	0.46	0.03	0.30	0.05	1.00	0.50	0.03	0.06	0.88	0.06	0.76	0.20	0.39
	SteamUse(kBtu)	0.02	0.44	0.08	0.60	0.55	0.03	0.26	0.46	-0.02	-0.02	-0.04	0.68	0.04	0.09	0.01	0.19	0.50	1.00	0.08	0.09	0.55	0.11	0.40	0.03	0.47
	SourceEUIWN(kBtu/sf)	0.02	0.05	0.03	0.27	0.29	0.18	0.10	0.04	-0.00	0.04	-0.31	0.22	0.14	0.94	0.13	0.53	0.03	0.08	1.00	0.99	0.29	0.94	0.00	0.18	0.39
	SourceEUI(kBtu/sf)	0.02	0.08	0.04	0.30	0.32	0.18	0.11	0.07	-0.00	0.04	-0.31	0.23	0.14	0.93	0.13	0.52	0.06	0.09	0.99	1.00	0.32	0.94	0.03	0.18	0.39
	Electricity(kWh)	0.03	0.85	0.25	0.96	1.00	0.29	0.63	0.86	-0.02	0.04	-0.06	0.69	0.68	0.25	0.22	0.18	0.88	0.55	0.29	0.32	1.00	0.29	0.74	0.29	0.59
	SiteEUI(kBtu/sf)	0.03	0.07	0.01	0.30	0.29	0.26	0.10	0.06	-0.01	-0.02	-0.34	0.29	0.13	0.99	0.10	0.73	0.06	0.11	0.94	0.94	0.29	1.00	0.03	0.26	0.40
	NumberofBuildings	0.02	0.69	-0.03	0.69	0.74	0.06	0.11	0.73	0.02	-0.02	-0.00	0.41	0.00	0.01	-0.00	0.03	0.76	0.40	0.00	0.03	0.74	0.03	1.00	0.06	0.09
	NaturalGas(kBtu)	0.03	0.18	0.07	0.51	0.29	1.00	0.39	0.18	-0.02	0.02	-0.10	0.73	0.65	0.26	0.06	0.49	0.20	0.03	0.18	0.18	0.29	0.26	0.06	1.00	0.73
	SiteEnergyUseWN(kBtu)	0.03	0.40	0.29	0.72	0.59	0.73	0.63	0.38	-0.04	0.07	-0.09	0.86	0.75	0.39	0.24	0.43	0.39	0.47	0.39	0.39	0.59	0.40	0.09	0.73	1.00
	Longitude																									
	PropertyGFATotal																									
	NumberofFloors																									
	SiteEnergyUse(kBtu)																									
	Electricity(kBtu)																									
	NaturalGas(therms)																									
	astPropertyUse TypeGFA																									
	PropertyGFABuilding(s)																									
	Latitude																									
	YearBuilt																									
	ENERGYSTARScore																									
	TotalGHGEmissions																									
	astPropertyUse TypeGFA																									
	SiteEUIWN(kBtu/sf)																									
	PropertyGFAParking																									
	GHGEmissionsIntensity																									
	astPropertyUse TypeGFA																									
	SteamUse(kBtu)																									
	SourceEUIWN(kBtu/sf)																									
	SourceEUI(kBtu/sf)																									
	Electricity(kWh)																									
	SiteEUI(kBtu/sf)																									
	NumberofBuildings																									
	NaturalGas(kBtu)																									
	SiteEnergyUseWN(kBtu)																									

II.2.3. Features redondantes

	Longitude	1.00	0.03	-0.03	0.03	0.03	0.03	0.02	0.03	0.01	-0.05	-0.03	0.04	0.03	0.03	-0.00	0.04	0.03	0.02	0.02	0.02	0.03	0.03	0.02	0.03	0.03
	<div>PropertyGFATotal</div>	0.03	1.00	0.40	0.80	0.85	0.18	0.81	0.99	-0.02	0.10	0.07	0.53	0.52	0.04	0.40	0.02	0.97	0.44	0.05	0.08	0.85	0.07	0.69	0.18	0.40
	NumberofFloors	-0.03	0.40	1.00	0.21	0.25	0.07	0.47	0.36	-0.02	0.15	0.02	0.14	0.22	-0.00	0.42	-0.04	0.34	0.08	0.03	0.04	0.25	0.01	-0.03	0.07	0.29
	<div>SiteEnergyUse(kBtu)</div>	0.03	0.80	0.21	1.00	0.96	0.51	0.63	0.81	-0.02	0.03	-0.09	0.86	0.75	0.27	0.17	0.31	0.84	0.60	0.27	0.30	0.96	0.30	0.69	0.51	0.72
	Electricity(kBtu)	0.03	0.85	0.25	0.96	1.00	0.29	0.63	0.86	-0.02	0.04	-0.06	0.69	0.68	0.25	0.22	0.18	0.88	0.55	0.29	0.32	1.00	0.29	0.74	0.29	0.59
	<div>NaturalGas(therms)</div>	0.03	0.18	0.07	0.51	0.29	1.00	0.39	0.18	-0.02	0.02	-0.10	0.73	0.65	0.26	0.06	0.49	0.20	0.03	0.18	0.18	0.29	0.26	0.06	1.00	0.73
SecondLargestPropertyUseTypeGFA	<div>PropertyGFABuilding(s)</div>	0.02	0.81	0.47	0.63	0.63	0.39	1.00	0.79	-0.05	0.20	0.08	0.51	0.66	0.08	0.48	0.11	0.77	0.26	0.10	0.11	0.63	0.10	0.11	0.39	0.63
	<div>PropertyGFABuilding(s)</div>	0.03	0.99	0.36	0.81	0.86	0.18	0.79	1.00	-0.02	0.08	0.06	0.55	0.54	0.03	0.27	0.03	0.98	0.46	0.04	0.07	0.86	0.06	0.73	0.18	0.38
	Latitude	0.01	-0.02	-0.02	-0.02	-0.02	-0.02	-0.05	-0.02	1.00	0.12	0.08	-0.03	-0.11	-0.02	-0.00	-0.04	-0.02	-0.02	-0.00	-0.00	-0.02	-0.01	0.02	-0.02	-0.04
	YearBuilt	-0.05	0.10	0.15	0.03	0.04	0.02	0.20	0.08	0.12	1.00	0.03	0.01	0.09	-0.03	0.18	-0.15	0.07	-0.02	0.04	0.04	0.04	-0.02	-0.02	0.02	0.07
	ENERGYSTARScore	-0.03	0.07	0.02	-0.09	-0.06	-0.10	0.08	0.06	0.08	0.03	1.00	-0.10	-0.01	-0.35	0.05	-0.27	0.06	-0.04	-0.31	-0.31	-0.06	-0.34	-0.00	-0.10	-0.09
	<div>TotalGHGEmissions</div>	0.04	0.53	0.14	0.86	0.69	0.73	0.51	0.55	-0.03	0.01	-0.10	1.00	0.68	0.27	0.09	0.47	0.58	0.68	0.22	0.23	0.69	0.29	0.41	0.73	0.86
ThirdLargestPropertyUseTypeGFA	<div>SiteEUIWN(kBtu/sf)</div>	0.03	0.52	0.22	0.75	0.68	0.65	0.66	0.54	-0.11	0.09	-0.01	0.68	1.00	0.13	0.21	0.33	0.46	0.04	0.14	0.14	0.68	0.13	0.00	0.65	0.75
	<div>SiteEUIWN(kBtu/sf)</div>	0.03	0.04	-0.00	0.27	0.25	0.26	0.08	0.03	-0.02	-0.03	-0.35	0.27	0.13	1.00	0.09	0.75	0.03	0.09	0.94	0.93	0.25	0.99	0.01	0.26	0.39
	<div>PropertyGFAParking</div>	-0.00	0.40	0.42	0.17	0.22	0.06	0.48	0.27	-0.00	0.18	0.05	0.09	0.21	0.09	1.00	-0.04	0.30	0.01	0.13	0.13	0.22	0.10	-0.00	0.06	0.24
	GHGEmissionsIntensity	0.04	0.02	-0.04	0.31	0.18	0.49	0.11	0.03	-0.04	-0.15	-0.27	0.47	0.33	0.75	-0.04	1.00	0.05	0.19	0.53	0.52	0.18	0.73	0.03	0.49	0.43
	<div>LargestPropertyUseTypeGFA</div>	0.03	0.97	0.34	0.84	0.88	0.20	0.77	0.98	-0.02	0.07	0.06	0.58	0.46	0.03	0.30	0.05	1.00	0.50	0.03	0.06	0.88	0.06	0.76	0.20	0.39
	SteamUse(kBtu)	0.02	0.44	0.08	0.60	0.55	0.03	0.26	0.46	-0.02	-0.02	-0.04	0.68	0.04	0.09	0.01	0.19	0.50	1.00	0.08	0.09	0.55	0.11	0.40	0.03	0.47
	<div>SourceEUIWN(kBtu/sf)</div>	0.02	0.05	0.03	0.27	0.29	0.18	0.10	0.04	-0.00	0.04	-0.31	0.22	0.14	0.94	0.13	0.53	0.03	0.08	1.00	0.99	0.29	0.94	0.00	0.18	0.39
	<div>SourceEUI(kBtu/sf)</div>	0.02	0.08	0.04	0.30	0.32	0.18	0.11	0.07	-0.00	0.04	-0.31	0.23	0.14	0.93	0.13	0.52	0.06	0.09	0.99	1.00	0.32	0.94	0.03	0.18	0.39
	<div>Electricity(kWh)</div>	0.03	0.85	0.25	0.96	1.00	0.29	0.63	0.86	-0.02	0.04	-0.06	0.69	0.68	0.25	0.22	0.18	0.88	0.55	0.29	0.32	1.00	0.29	0.74	0.29	0.59
	<div>SiteEUI(kBtu/sf)</div>	0.03	0.07	0.01	0.30	0.29	0.26	0.10	0.06	-0.01	-0.02	-0.34	0.29	0.13	0.99	0.10	0.73	0.06	0.11	0.94	0.94	0.29	1.00	0.03	0.26	0.40
	NumberofBuildings	0.02	0.69	-0.03	0.69	0.74	0.06	0.11	0.73	0.02	-0.02	-0.00	0.41	0.00	0.01	-0.00	0.03	0.76	0.40	0.00	0.03	0.74	0.03	1.00	0.06	0.09
	NaturalGas(kBtu)	0.03	0.18	0.07	0.51	0.29	1.00	0.39	0.18	-0.02	0.02	-0.10	0.73	0.65	0.26	0.06	0.49	0.20	0.03	0.18	0.18	0.29	0.26	0.06	1.00	0.73
	<div>SiteEnergyUseWNI(kBtu)</div>	0.03	0.40	0.29	0.72	0.59	0.73	0.63	0.38	-0.04	0.07	-0.09	0.86	0.75	0.39	0.24	0.43	0.39	0.47	0.39	0.39	0.59	0.40	0.09	0.73	1.00
	Longitude																									
	PropertyGFATotal																									
	NumberofFloors																									
	SiteEnergyUse(kBtu)																									
	Electricity(kBtu)																									
	NaturalGas(therms)																									
	1stPropertyUseTypeGFA																									
	PropertyGFABuilding(s)																									
	Latitude																									
	YearBuilt																									
	ENERGYSTARScore																									
	TotalGHGEmissions																									
	1stPropertyUseTypeGFA																									
	SiteEUIWN(kBtu/sf)																									
	PropertyGFAParking																									
	GHGEmissionsIntensity																									
	1stPropertyUseTypeGFA																									
	SteamUse(kBtu)																									
	SourceEUIWN(kBtu/sf)																									
	SourceEUI(kBtu/sf)																									
	Electricity(kWh)																									
	SiteEUI(kBtu/sf)																									
	NumberofBuildings																									
	NaturalGas(kBtu)																									
	SiteEnergyUseWNI(kBtu)																									

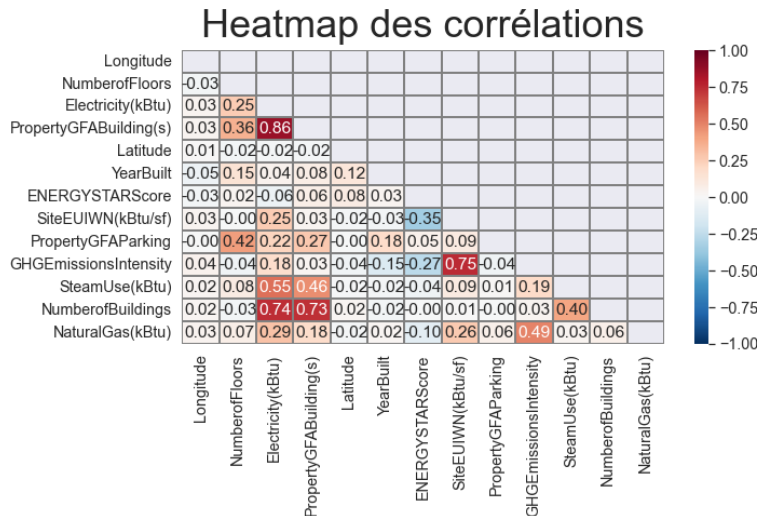
II.2.4. Leakage en gardant electricity, steam et gas ?



→ données identiques à qqes exceptions près

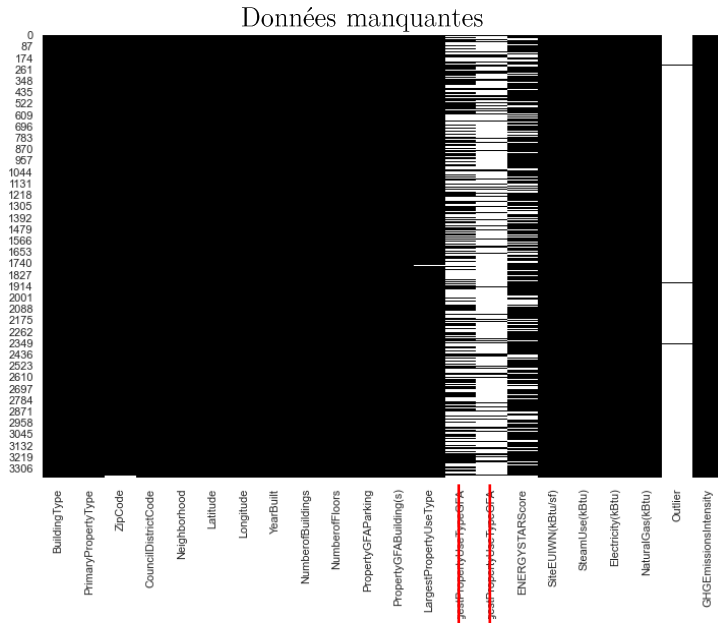
→ $\text{SiteEUIWN} \approx (\text{Electricity} + \text{Steam} + \text{Gas})/\text{GFA}$

II.2.5. Corrélations



→ corrélations de {n_buildings, GFA_buildings et Electricity} et {GHGEI et SiteEUIWN} élevées mais pas alarmantes.

II.3. Indicateurs peu remplis



→ ENERGYSTARScore voulu.

→ Outlier va nous servir.

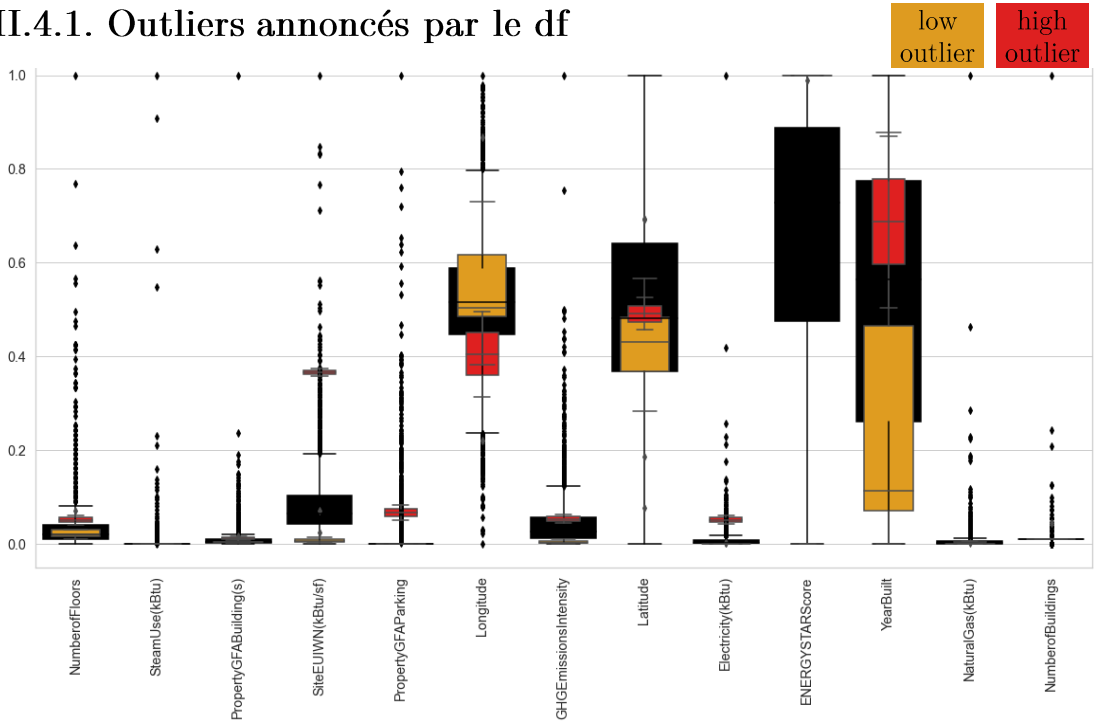
II.4. Suppression d'éléments

- Bâtiments non résidentiels (énoncé) → suppression de la moitié des lignes du df.
- Deux éléments supprimés car presque vides

Parking	PropertyGFABuilding(s)	LargestPropertyUseType	ENERGYSTARScore	SiteEUIWN(kBtu/sf)	SteamUse(kBtu)	Electricity(kBtu)	NaturalGas(kBtu)	Outlier	GHGEmissionsIntensity
0	63150	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0	20760	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

- Outliers

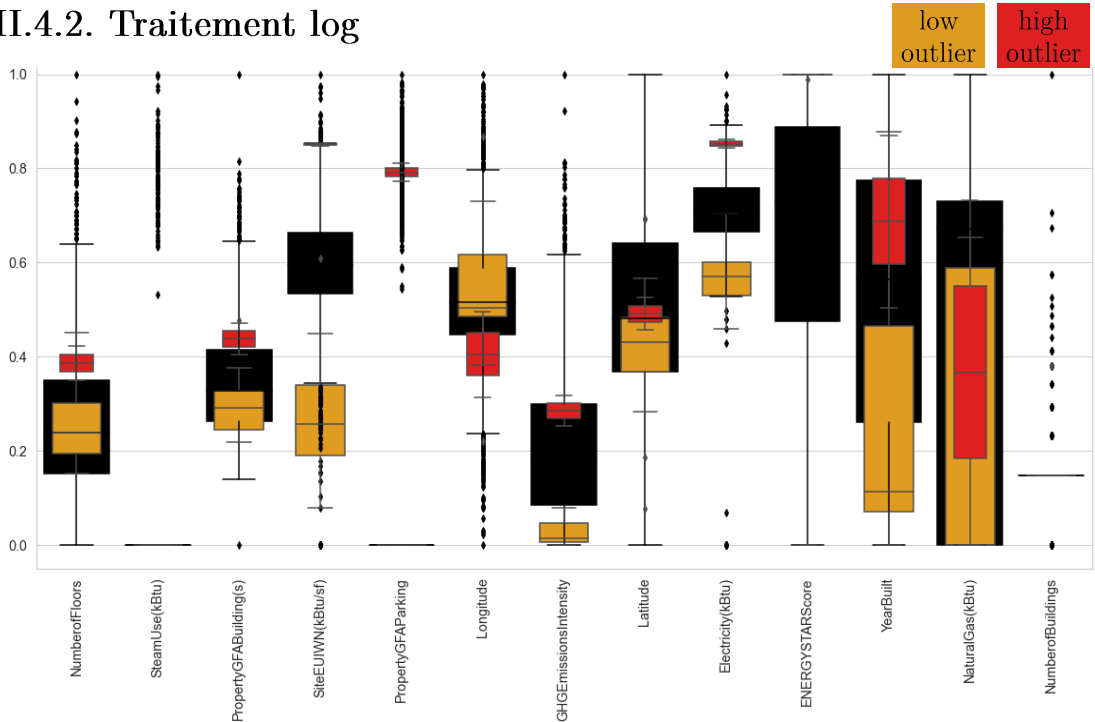
II.4.1. Outliers annoncés par le df



→ les outliers annoncés ne semblent pas être des outliers.

→ illisible, il faut certainement regarder en échelle log.

II.4.2. Traitement log



→ Je n'en ai finalement supprimé aucun.

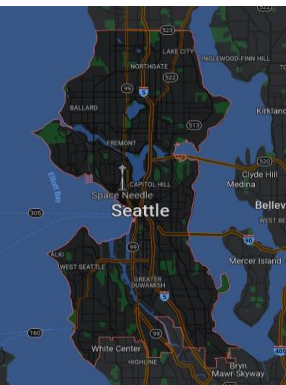
→ Faut-il garder ce traitement log pour la pipeline de régression ?

II.5. Imputation

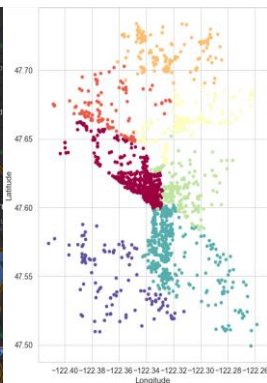
- Imputation de LargestPropertyType par most frequent (mode).
- Imputation de ZipCode par une constante (car OHE ensuite).
- Imputation de SiteEUIWN(kBtu/sf), Electricity(kBtu) et GHGEmissionsIntensity par knn.

Interlude

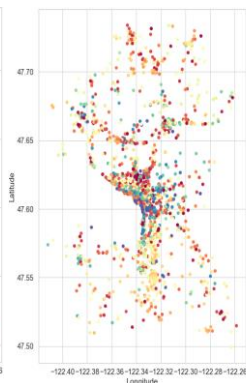
Plan de ville



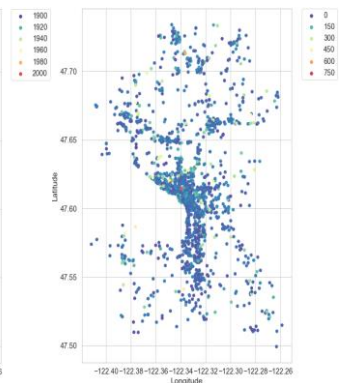
CouncilDistrictCode



YearBuilt



SiteEUIWN(kBtu/sf)



La géographie de la ville semble (encore une fois) peu influencer sur nos targets.

Plan

I Présentation du projet

II Nettoyage

III Traitement

IV Conclusions

1. best_estimator__

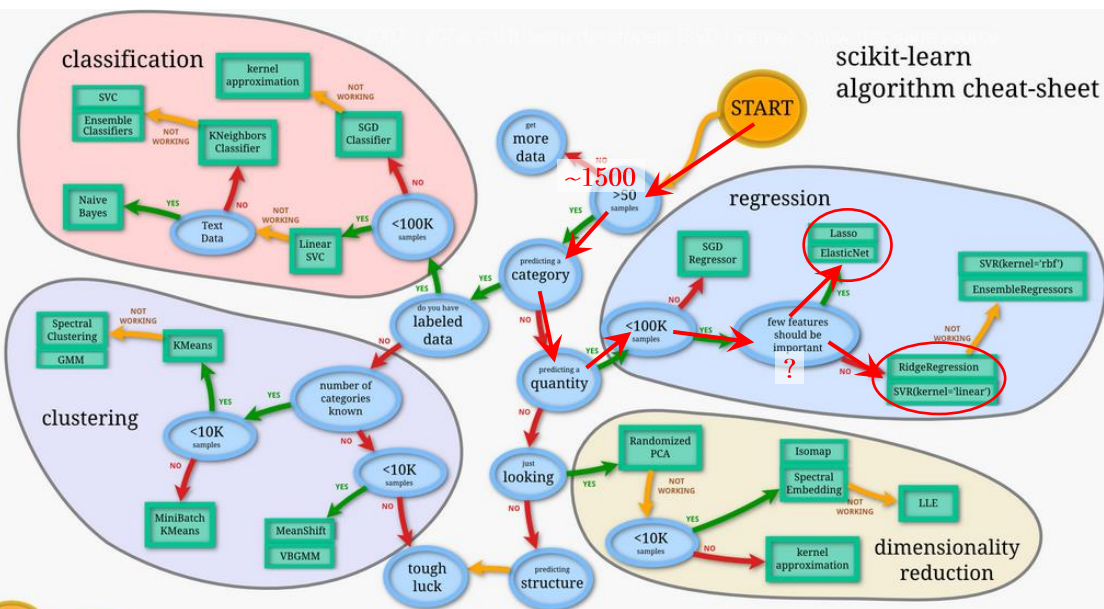
2. Importance de
l'ENERGYSTARScore

3. Plus de nettoyage

4. best_estimator__ bis

5. Empreinte carbone

III.1.1. Identification du problème



Back

scikit
learn

III.1.2. Paramétrage

`train__test__split()`

```
encoders = [OrdinalEncoder(handle_unknown='use_encoded_value', unknown_value=-1), OneHotEncoder(sparse=False, handle_unknown='ignore')]
scalers = [RobustScaler(), StandardScaler()]
```

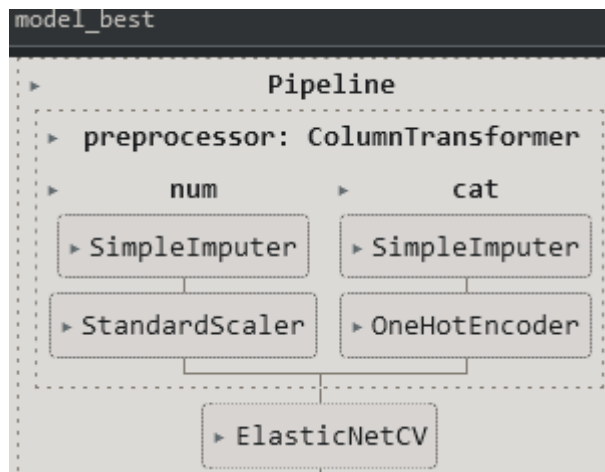
	estimator	param_names	param_nums
0	Ridge()	[ridge_alpha]	[[0.1, 0.3, 1, 3, 10]]
1	Lasso()	[lasso_alpha]	[[0.1, 0.3, 1, 3, 10]]
2	DummyRegressor()	[dummyregressor_strategy]	[[mean], [median]]
3	LinearRegression()	[linearregression_fit_intercept]	[[True]]
4	ElasticNetCV()	[elasticnetcv_l1_ratio, elasticnetcv_n_alphas]	[[0.1, 0.3, 1], [30, 100]]

III.1.3. Scoring

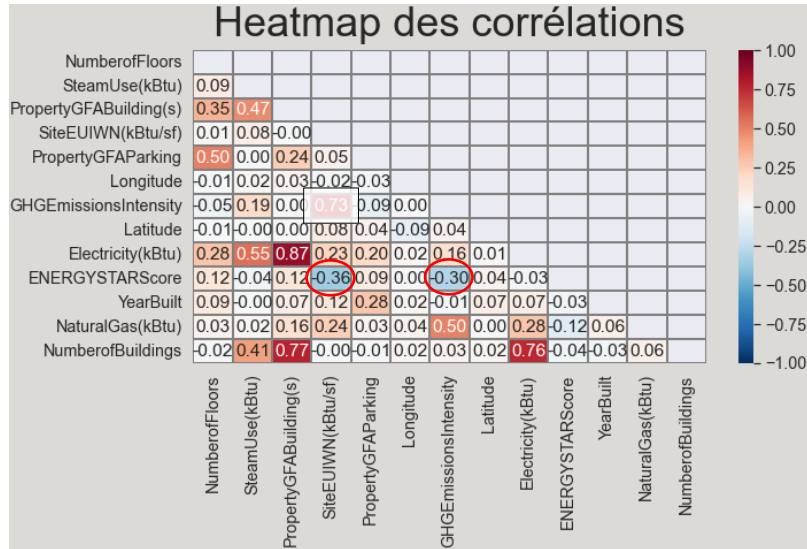
Estimator	best score	r2	mae	msqe	time
Ridge()	1.279617e-01	-1.670998e+00	4.305065e+01	1.303636e+04	0.078462
Lasso()	2.896550e-01	-1.192557e+00	4.271999e+01	1.070123e+04	0.079444
DummyRegressor()	-2.075314e-03	-4.763885e-04	4.547045e+01	4.883032e+03	0.089119
LinearRegression()	1.182682e-01	-1.674003e+00	4.305699e+01	1.305103e+04	0.088264
ElasticNetCV()	-6.414048e-03	-4.763885e-04	4.547045e+01	4.883032e+03	0.201605

```
time_1 = time_passed/(param_loops)
```

III.1.4. best_estimator__

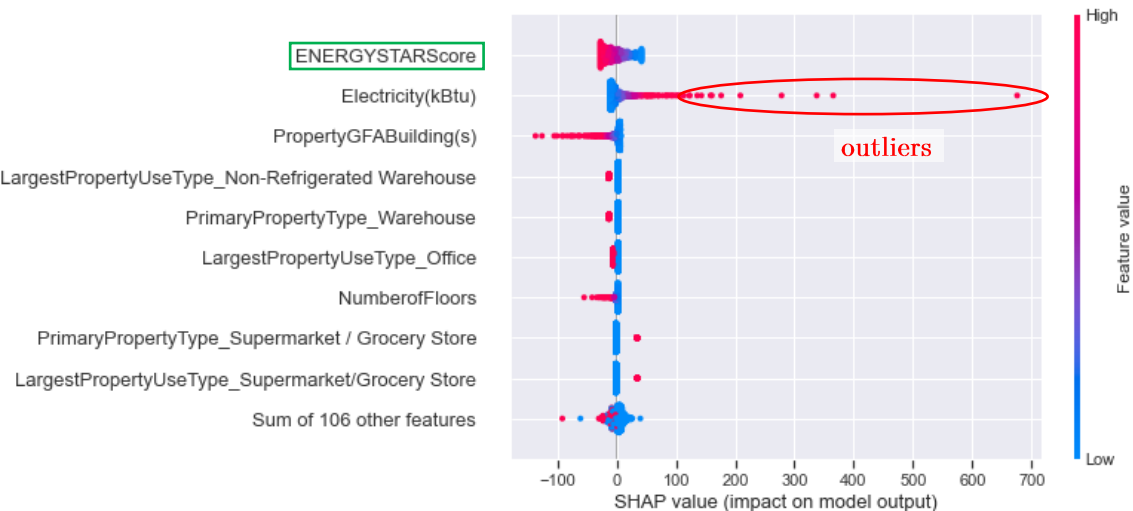


III.2.1. Corrélation



- Feature principale pour **SiteEUIWN(kBtu/sf)** et majeure pour **GHGEmissionsIntensity**.

III.2.2. shapley values



- Feature principale pour SiteEUIWN(kBtu/sf)

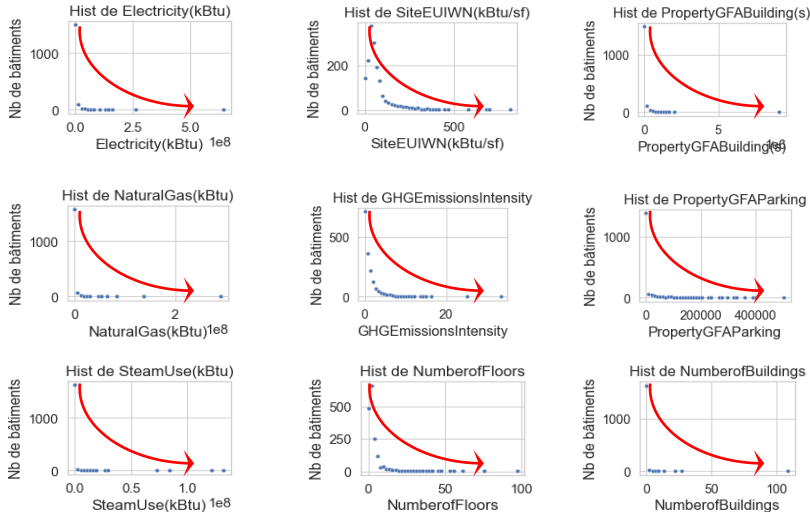
III.2.3. r^2 AVEC/SANS ENERGYSTARScore

```
r2_E-r2_0, mae_E-mae_0, mse_E-mse_0
```

```
(0.05465912749765389, -5.119671419303817, -300.27286418069525)
```

- Scores différents \rightarrow feature non-négligeable
- r^2 supérieur, mae et mse inférieures \rightarrow E* aide la prédiction

III.3.1. Nettoyages nécessaires



- Distribution exp des features
- r^2 très bas ($< .5$)
- affichage des shapley values illisible (outliers)

III.3.2. log

indicateur	skewness		skewness
NumberofFloors	5.106580	→	1.222773
SteamUse(kBtu)	18.518288	→	3.485820
PropertyGFABuilding(s)	4.667818	→	0.963639
SiteEUIWN(kBtu/sf)	4.685968	→	-0.835211
PropertyGFAParking	4.912275	→	1.360725
Longitude	0.030541		0.030541
GHGEmissionsIntensity	6.380721	→	1.213944
Latitude	0.174232		0.174232
Electricity(kBtu)	9.139249	→	-1.332960
ENERGYSTARScore	-0.775344		-0.775344
YearBuilt	-0.392232		-0.392232
NaturalGas(kBtu)	21.417662	→	-0.744709
NumberofBuildings	10.426625	→	2.711552

III.3.3. interquartile

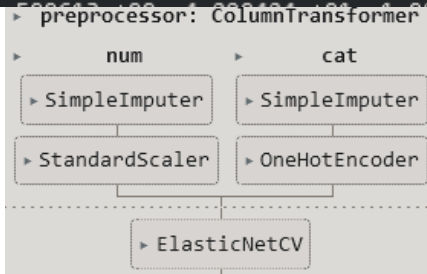
indicateur	# outliers bas	# outliers hauts
YearBuilt	0	0
Latitude	0	0
NaturalGas(kBtu)	0	0
PropertyGFABuilding(s)	1	23
SiteEUIWN(kBtu/sf)	48	33
Longitude	65	62
Electricity(kBtu)	2	2
ENERGYSTARScore	0	0
GHGEmissionsIntensity	0	7

- Suppression seulement des éléments qui sont toujours hors interquartiles même après le log.
- (suppressions selon Longitude pas forcément nécessaires)

III.4.1. best_estimator__

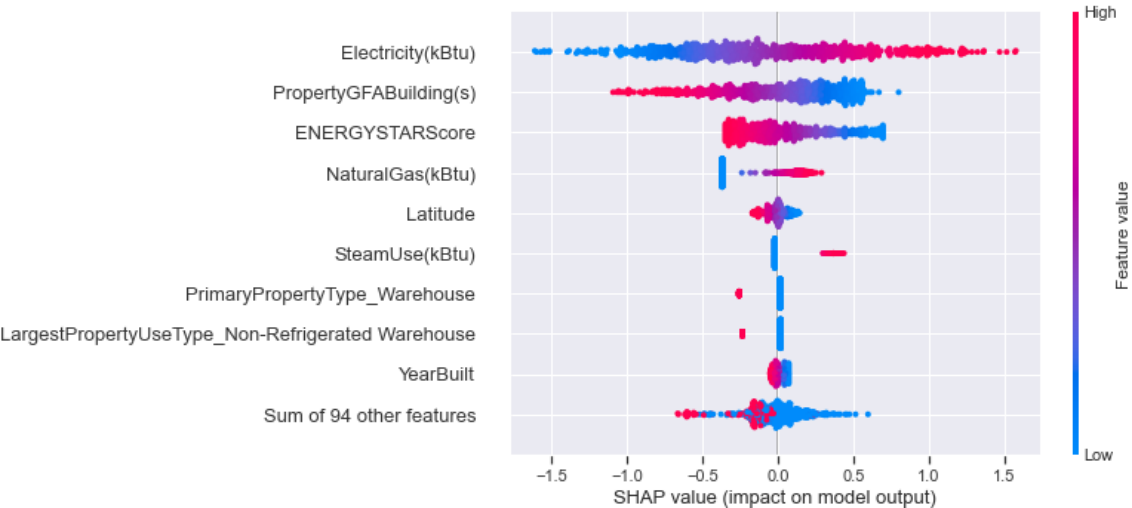
```
print(dfscores2[col_score] - dfscores[col_score])
```

	best score	r2	mae	msqe	time
0	5.456720e-01	2.220879e+00	-4.266704e+01	-1.303612e+04	-0.009930
1	8.160592e-02	1.537467e+00	-4.227454e+01	-1.070087e+04	-0.012977
2	-6.066227e-03	4.737353e-04	-4.491818e+01	-4.882494e+03	-0.022371
3	5.547913e-01	2.221856e+00	-4.267244e+01	-1.305078e+04	-0.018669
4	6.804202e-01	5.530310e-01	-4.508878e+01	-4.882791e+03	-0.009507
5	4.792426e-01	2.145800e+00	-4.280415e+01	-1.266811e+04	-0.006643



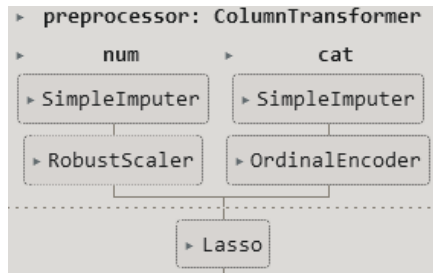
- Meilleurs scores dans les pipelines (sauf le Dummy)
- best_estimator__ similaire
- Shapley values mieux représentées

III.4.2. shapley values

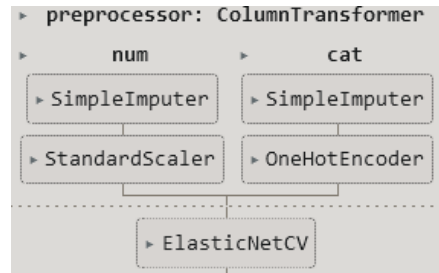


- E^* n'apparaît plus comme la feature principale pour **SiteEUIWN(kBtu/sf)** mais reste majeure.

III.5.1 GHGEmissionsIntensity



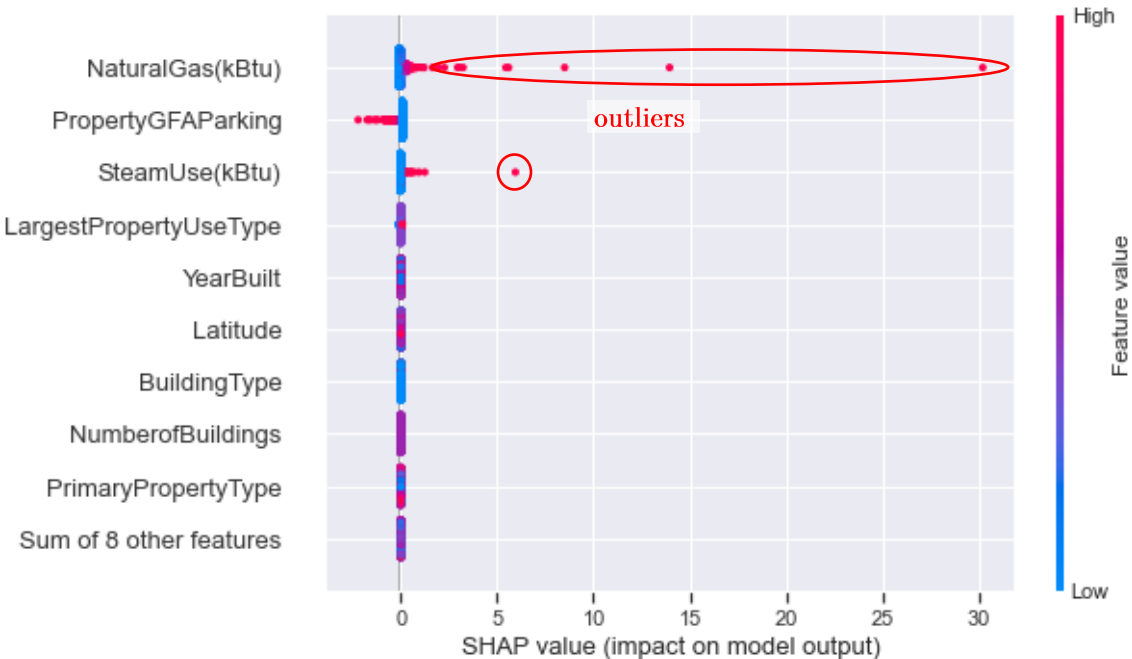
→
nettoyage



```
r2_E-r2_0, mae_E-mae_0, mse_E-mse_0
(0.0, 0.0, 0.0)
```

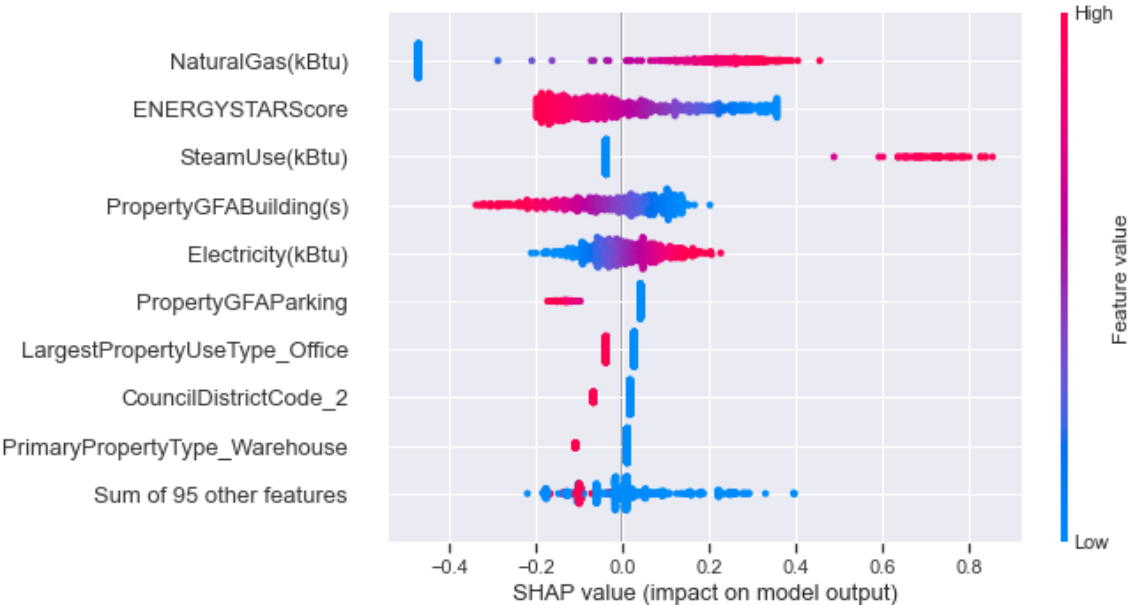
- Pour cette target je trouve une importance nulle de l'ENERGYSTARScore avant nettoyage {log + outliers} malgré ce qu'annonce le tableau de corrélations.
- Le `best_estimator` est modifié après nettoyage.

III.5.2. shapley values



- ENERGYSTARScore négligeable avant nettoyage

III.5.3. shapley values



- ENERGYSTARScore majeur après nettoyage

CCL

- La meilleure pipeline est trouvée.
- Preuves que l'ENERGYSTARScore est indispensable aux prédictions.
- Note: 1^{ère} recherche de meilleure pipeline sur tous éléments mais sans E*.

Questions ?

V. Approfondissement suite à la soutenance

V.1. Recommandations ajoutées

V.1.1. Correction d'un data leakage

SteamUse(y/n)	Electricity(y/n)	NaturalGas(y/n)
True	True	True
False	True	True
True	True	True
False	True	True

V. Approfondissement suite à la soutenance

V.1. Recommandations ajoutées

V.1.2. Ajout d'estimateurs non-linéaires

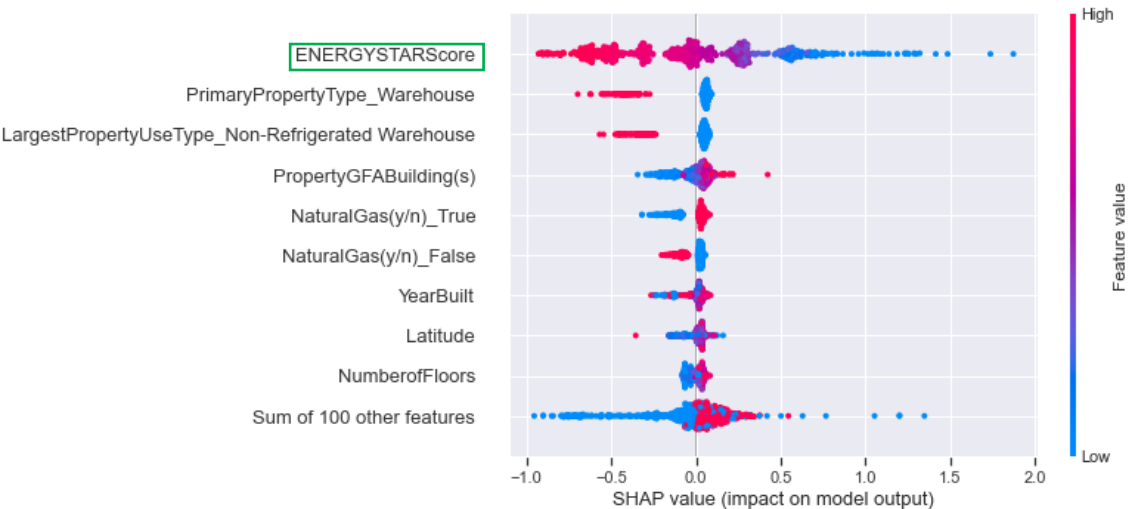
	estimator	param_names	param_nums
0	Ridge()	[ridge_alpha]	[[0.1, 0.3, 1, 3, 10]]
1	Lasso()	[lasso_alpha]	[[0.1, 0.3, 1, 3, 10]]
2	DummyRegressor()	[dummyregressor_strategy]	[[mean], [median]]
3	LinearRegression()	[linearregression_fit_intercept]	[[True]]
4	ElasticNetCV()	[elasticnetcv_l1_ratio, elasticnetcv_n_alphas]	[[0.1, 0.3, 1], [30, 100]]
5	RandomForestRegressor()	[randomforestregressor_max_depth, randomfores...	[[2], [0]]
6	GradientBoostingRegressor()	[gradientboostingregressor_random_state]	[[0]]
7	AdaBoostRegressor()	[adaboostregressor_n_estimators, adaboostregr...	[[30, 100], [0]]

V.2. Résultats

V.2.1. best__estimator__ (SiteEUIWN)

	Encoder	Scaler	Estimator	best score
	OneHotEncoder(handle_unknown='ignore', sparse=...	StandardScaler()	Ridge()	0.59
	OneHotEncoder(handle_unknown='ignore', sparse=...	RobustScaler()	ElasticNetCV()	0.60
	OneHotEncoder(handle_unknown='ignore', sparse=...	RobustScaler()	GradientBoostingRegressor()	0.60
	OrdinalEncoder(handle_unknown='use_encoded_val...	StandardScaler()	GradientBoostingRegressor()	0.60
	OneHotEncoder(handle_unknown='ignore', sparse=...	StandardScaler()	ElasticNetCV()	0.60
	OrdinalEncoder(handle_unknown='use_encoded_val...	RobustScaler()	GradientBoostingRegressor()	0.60
	OneHotEncoder(handle_unknown='ignore', sparse=...	StandardScaler()	GradientBoostingRegressor()	0.61

V.2.2. shapley values (SiteEUIWN)



- Feature principale pour SiteEUIWN(kBtu/sf)

V.2.3. r^2 avec/sans E*score (SiteEUIWN)

r^2 avec E*score

r^2 sans E*score

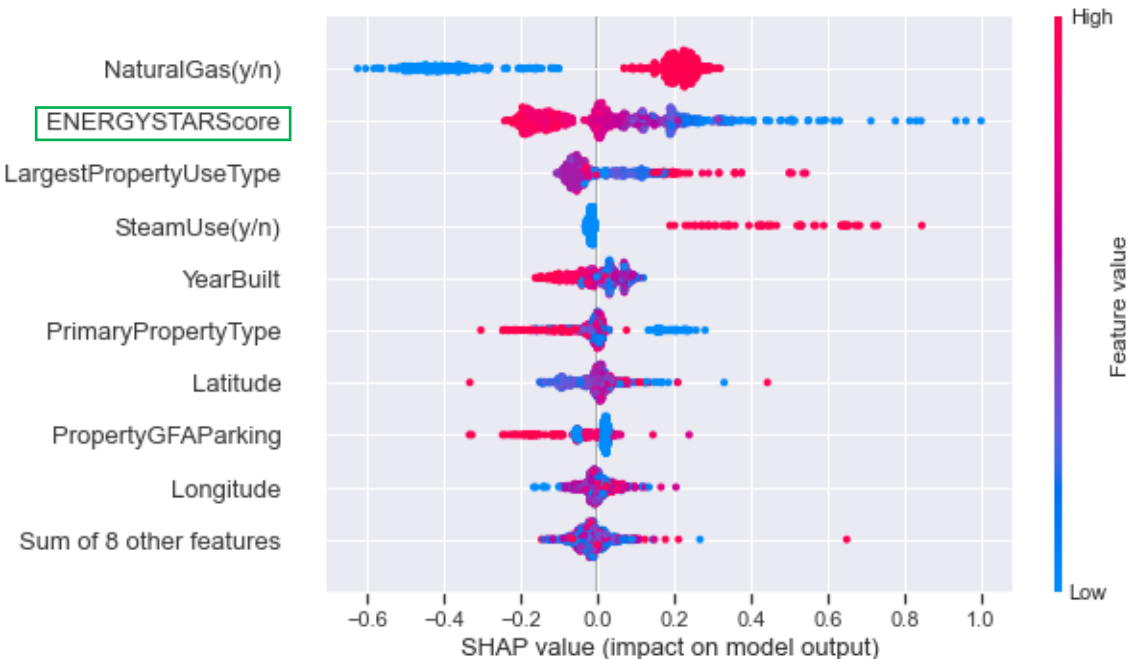
0.6122466721288635, 0.3000463312617414

- E*score indispensable

V.2.4. best_estimator_ (GHGEI)

	Encoder	Scaler	Estimator	best score
OneHotEncoder(handle_unknown='ignore', sparse=...		RobustScaler()	Ridge()	0.50
OneHotEncoder(handle_unknown='ignore', sparse=...		RobustScaler()	ElasticNetCV()	0.51
OneHotEncoder(handle_unknown='ignore', sparse=...		StandardScaler()	ElasticNetCV()	0.51
OneHotEncoder(handle_unknown='ignore', sparse=...		StandardScaler()	GradientBoostingRegressor()	0.52
OrdinalEncoder(handle_unknown='use_encoded_val...		StandardScaler()	GradientBoostingRegressor()	0.52
OneHotEncoder(handle_unknown='ignore', sparse=...		RobustScaler()	GradientBoostingRegressor()	0.52
OrdinalEncoder(handle_unknown='use_encoded_val...		RobustScaler()	GradientBoostingRegressor()	0.52

V.2.5. shapley values (GHGEI)



- Feature importante pour GHGEI

V.2.6. r^2 avec/sans E^* score (GHGEI)

r^2 avec E^* score

r^2 sans E^* score

0.5176801383113236, 0.3851714287519793

- E^* score indispensable

CCL

- Un important data leakage solutionné.
- La meilleure pipeline est trouvée, avec davantage d'estimateurs.
- Preuves que l'ENERGYSTARScore est indispensable aux prédictions.

Questions ?