Anticipez les besoins en consommation de bâtiments

par Ana Bernal

Novembre 2022

OpenClassrooms

Mentor: Samir Tanfous

Programme

1 **Appel** à projets.

2 **Analyse** exploratoire et choix de variables.

3
* **Modélisation** et **prédictions** et **évaluation** des performances de:
  - Consommation totale d'**énergie**
  - Émissions de **gaz a effet de serre**
* Ajout de la variable **EnergyStarScore** et comparaison des performances.

4 **Conclusions**, choix de modèle.

1

Appel à projets

Appel à projets

Ville de

Seattle

objectif 2050

ville neutre
emissions carbone

Appel à projets

Ville de
Seattle

objectif 2050

ville neutre
emissions carbone

bâtiments non
residentiels

relevés
2016

Calcul compliqué

Appel à projets

Ville de

Seattle

objectif 2050

ville neutre
emissions carbone

bâtiments non
residentiels

relevés
2016

ENERGY STAR

Calcul compliqué

MISSION

① Prediction → émissions
gaz effet serre
→ conso. énergie

② Intérêt

2

Analyse Exploratoire

Analyse exploratoire

- Données officielles, ville de Seattle.
- Taille des données brutes

Nombre d'individus	3376
Nombre de variables	46

Analyse exploratoire

- Données officielles, ville de Seattle.
- Taille des données brutes

Nombre d'individus	3376
Nombre de variables	46

dont

Quantitatives	Qualitatives	Boléennes
30	15	1

Analyse exploratoire

- Données officielles, ville de Seattle.
- Taille des données brutes

Nombre d'individus	3376
Nombre de variables	46

Quantitatives	Qualitatives	Booléennes
30	15	1

Variables structurelles: premier filtre	**Variables à prédire**
BuildingType, PrimaryPropertyType, Latitude, Longitude, YearBuilt, NumberofBuildings, NumberofFloors, PropertyGFATotal, PropertyGFAParking, PropertyGFABuilding(s), LargestPropertyUseType, LargestPropertyUseTypeGFA, SecondLargestPropertyUseType, SecondLargestPropertyUseTypeGFA, **ENERGYSTARScore**, Outlier	SiteEnergyUse(kBtu), Total GHGEmissions

Analyse exploratoire

- Données officielles, ville de Seattle.
- Taille des données brutes

Nombre d'individus	3376
Nombre de variables	46

Quantitatives	Qualitatives	Booléennes
30	15	1

Variables structurelles: premier filtre	Variables à prédire
BuildingType, PrimaryPropertyType, Latitude, Longitude, YearBuilt, NumberofBuildings, NumberofFloors, PropertyGFATotal, PropertyGFAParking, PropertyGFABuilding(s), LargestPropertyUseType, LargestPropertyUseTypeGFA, SecondLargestPropertyUseType, SecondLargestPropertyUseTypeGFA, **ENERGYSTARScore**, Outlier	SiteEnergyUse(kBtu), Total GHGEmissions

Analyse exploratoire

- Données officielles, ville de Seattle.
- Taille des données brutes

Nombre d'individus	3376
Nombre de variables	46

Quantitatives	Qualitatives	Booléennes
30	15	1

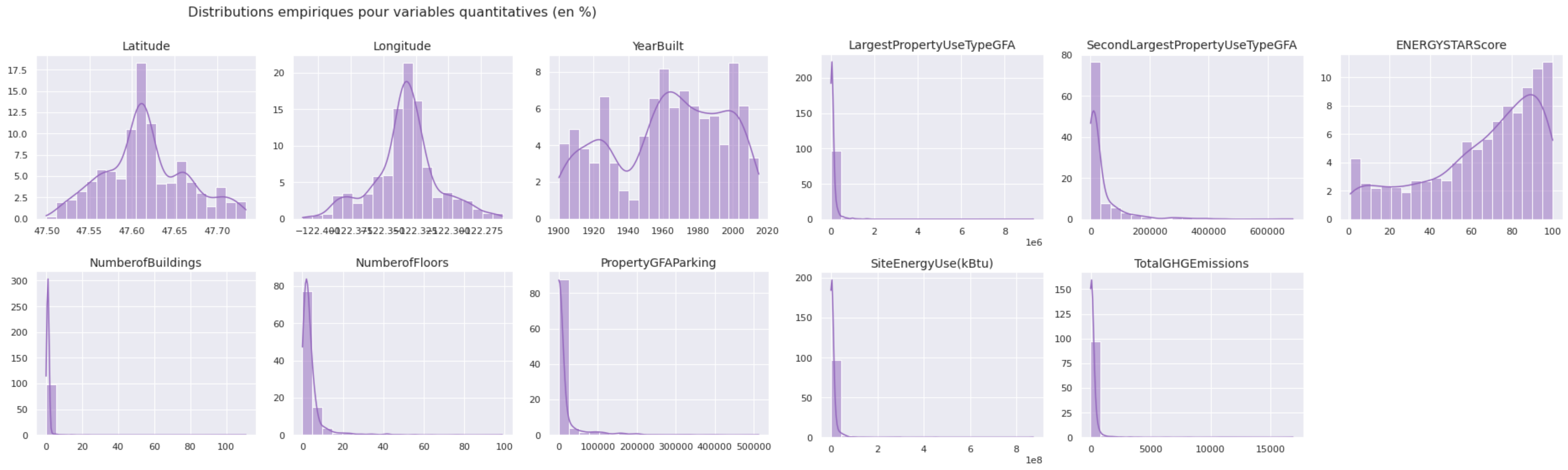
Variables structurelles: premier filtre	**Variables à prédire**
BuildingType, PrimaryPropertyType, Latitude, Longitude, YearBuilt, NumberofBuildings, NumberofFloors, PropertyGFATotal, PropertyGFAParking, PropertyGFABuilding(s), LargestPropertyUseType, LargestPropertyUseTypeGFA, SecondLargestPropertyUseType, SecondLargestPropertyUseTypeGFA, **ENERGYSTARScore**, Outlier	SiteEnergyUse(kBtu), Total GHGEmissions

Analyse exploratoire:

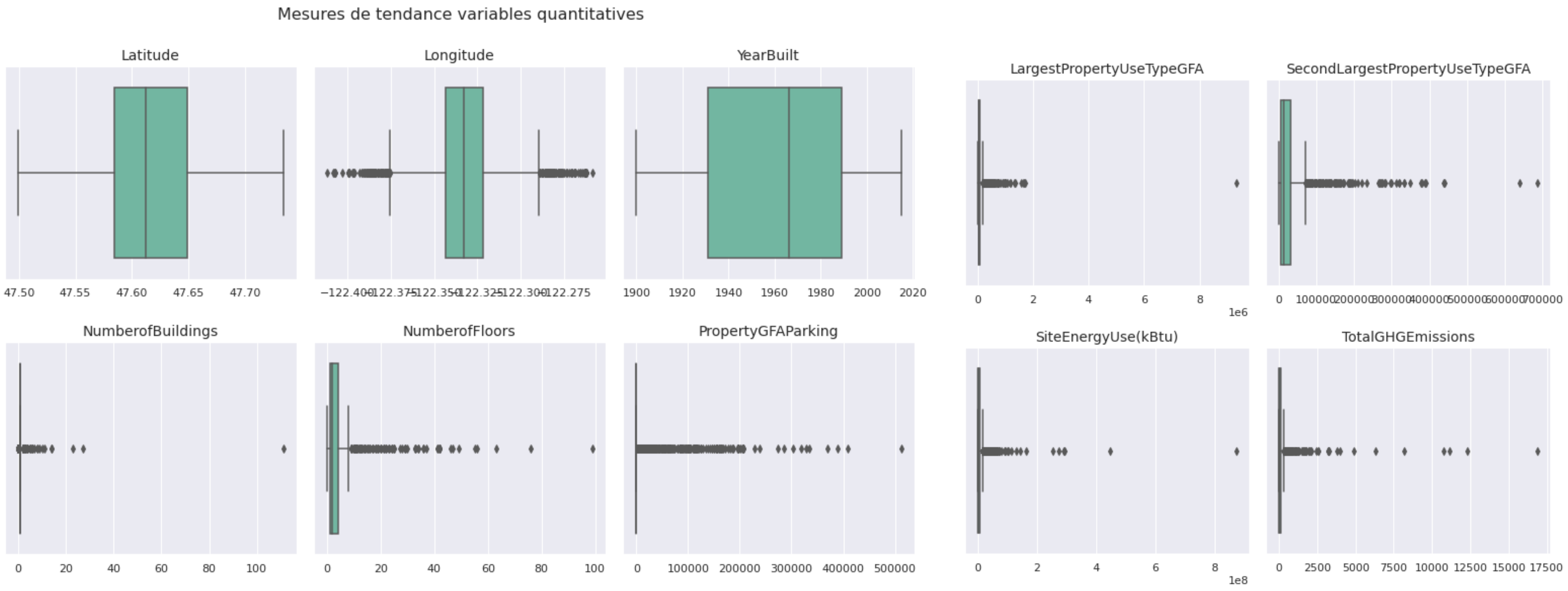
- Premier filtre : bâtiments non destinés à l'habitation

	Nombre d'individus
Avant	3344
Après	1599

Analyse exploratoire: variables quantitatives

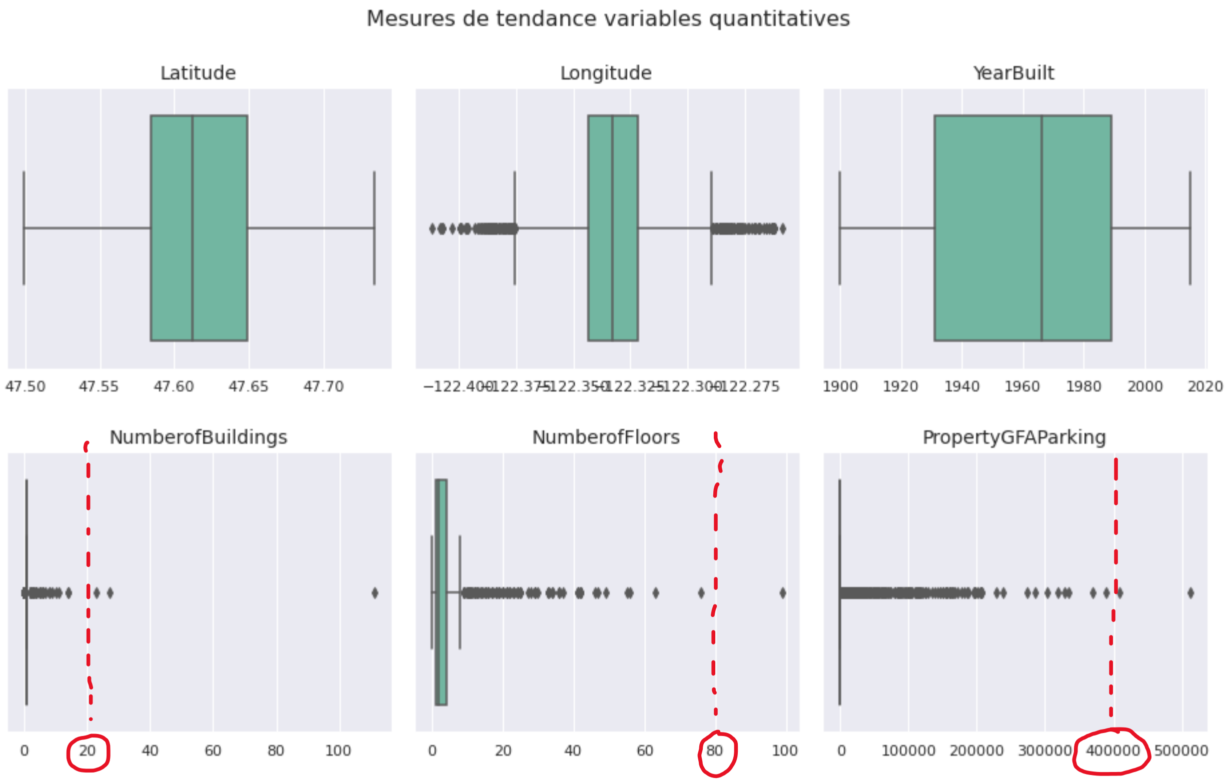


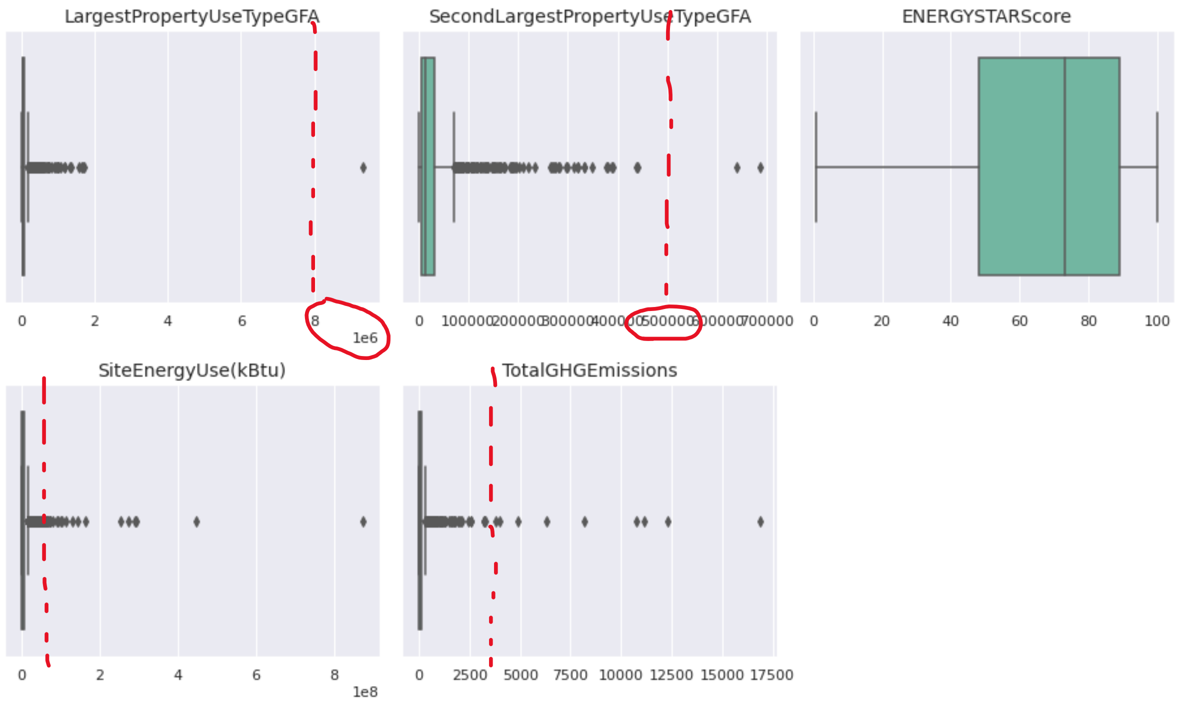
Analyse exploratoire: variables quantitatives 2



Analyse exploratoire: variables quantitatives

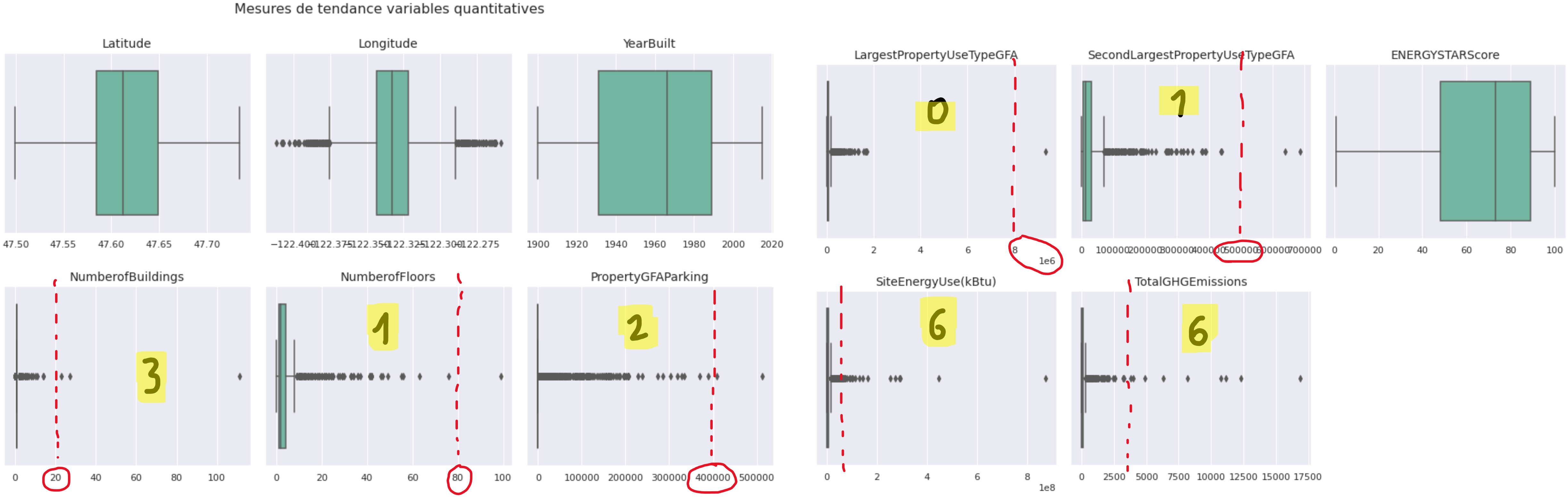
seuil outliers : cas par cas →





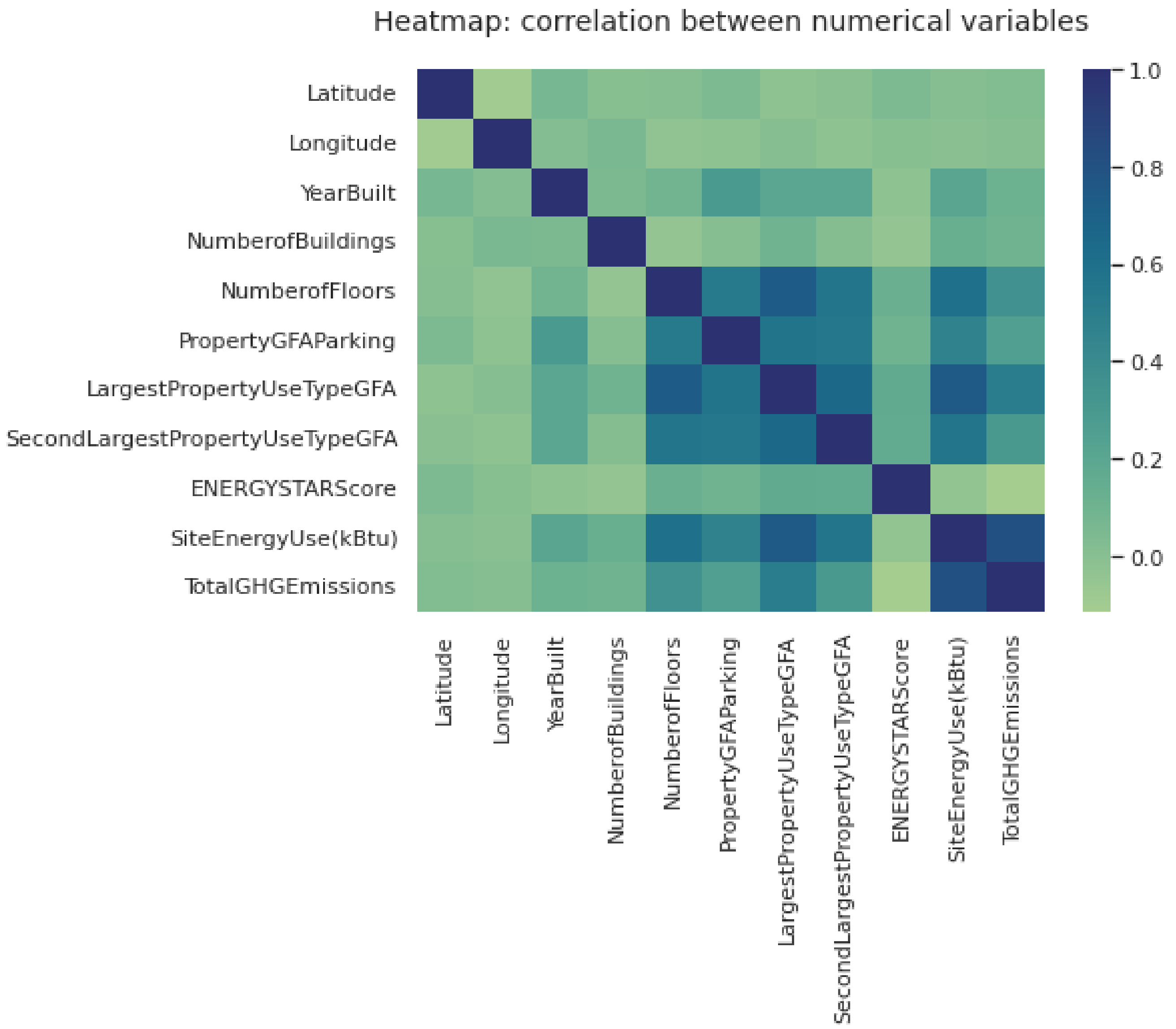
Analyse exploratoire: variables quantitatives

seuil outliers : cas par cas →

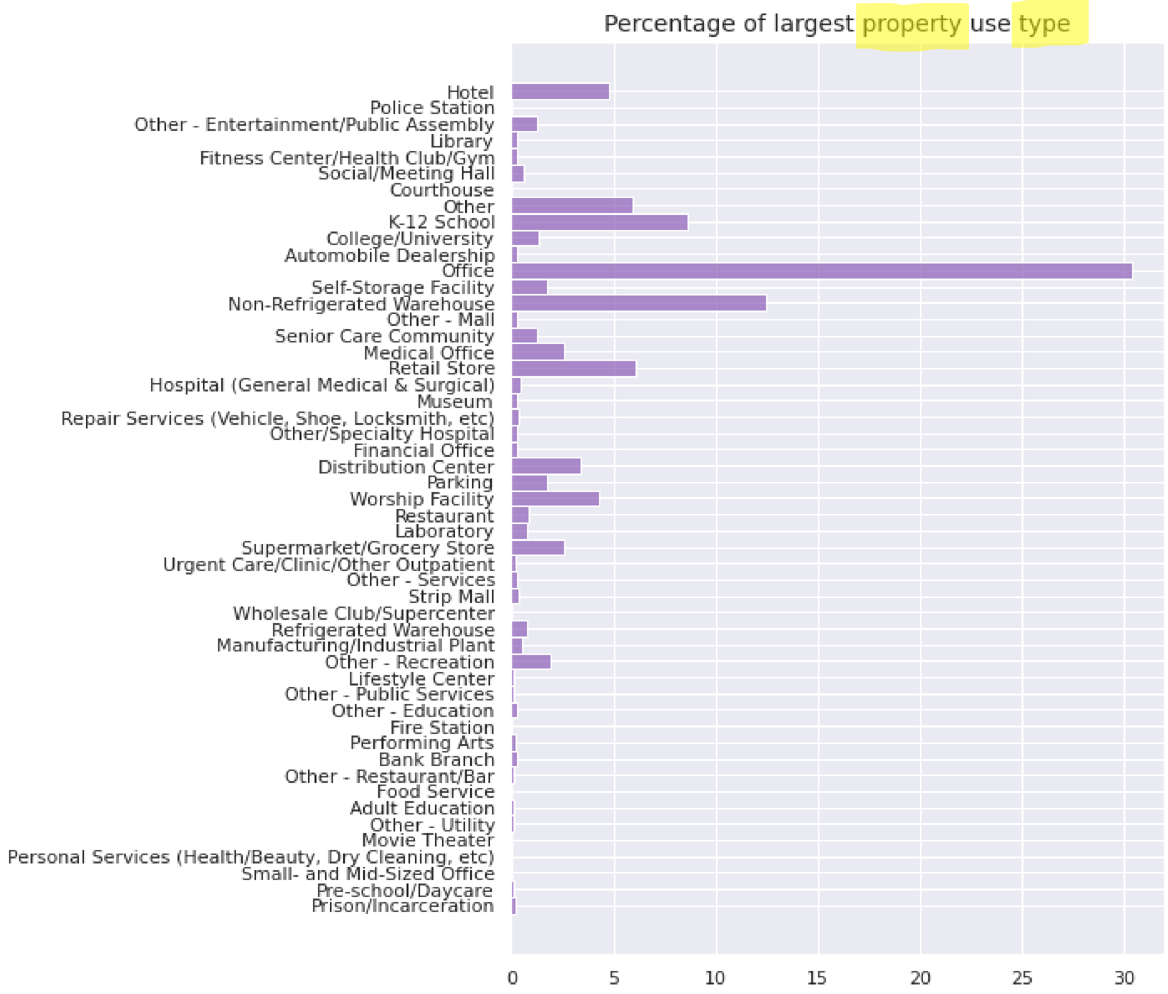


total individus supprimés : 19

Analyse exploratoire: variables quantitatives

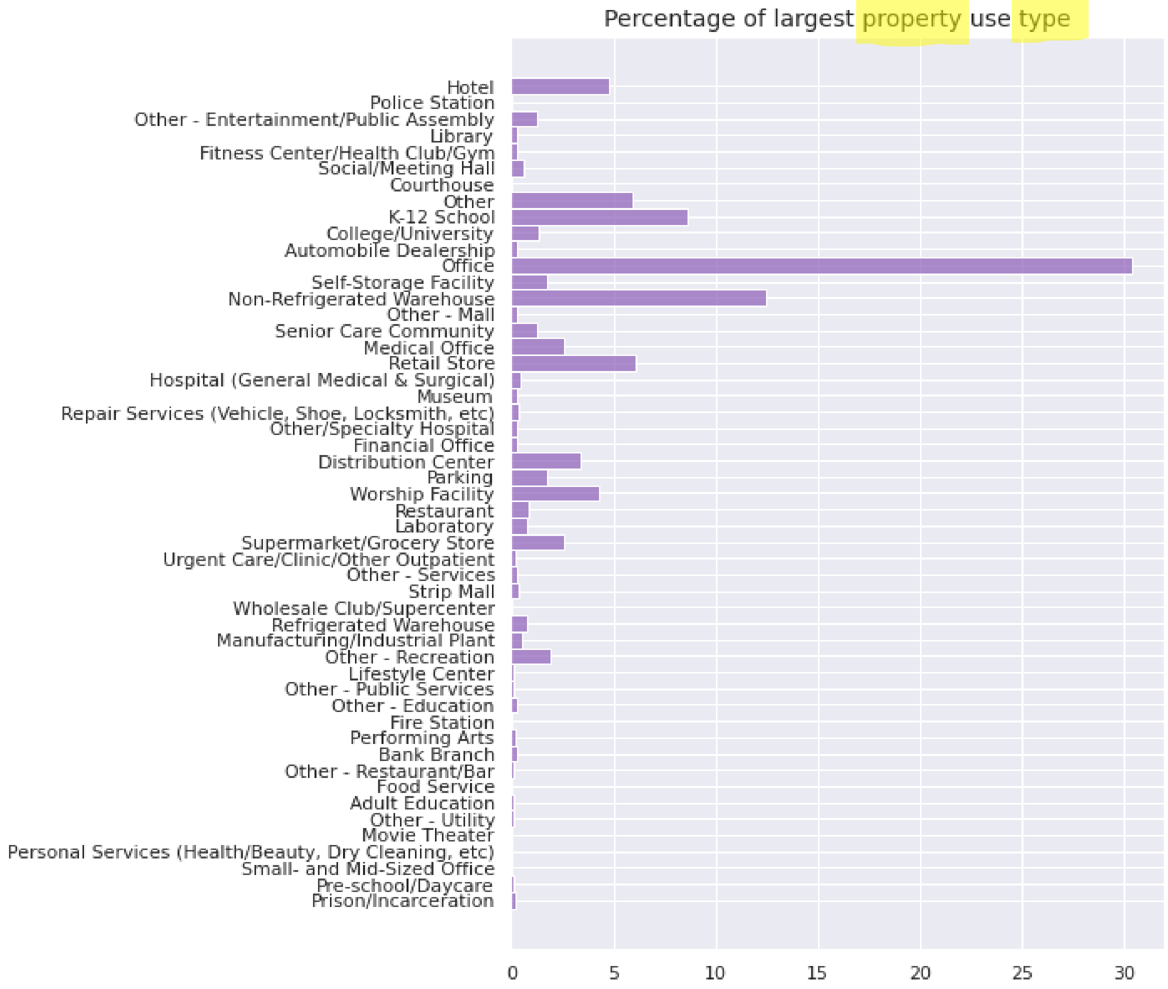


Analyse exploratoire: variables qualitatives



53 catégories

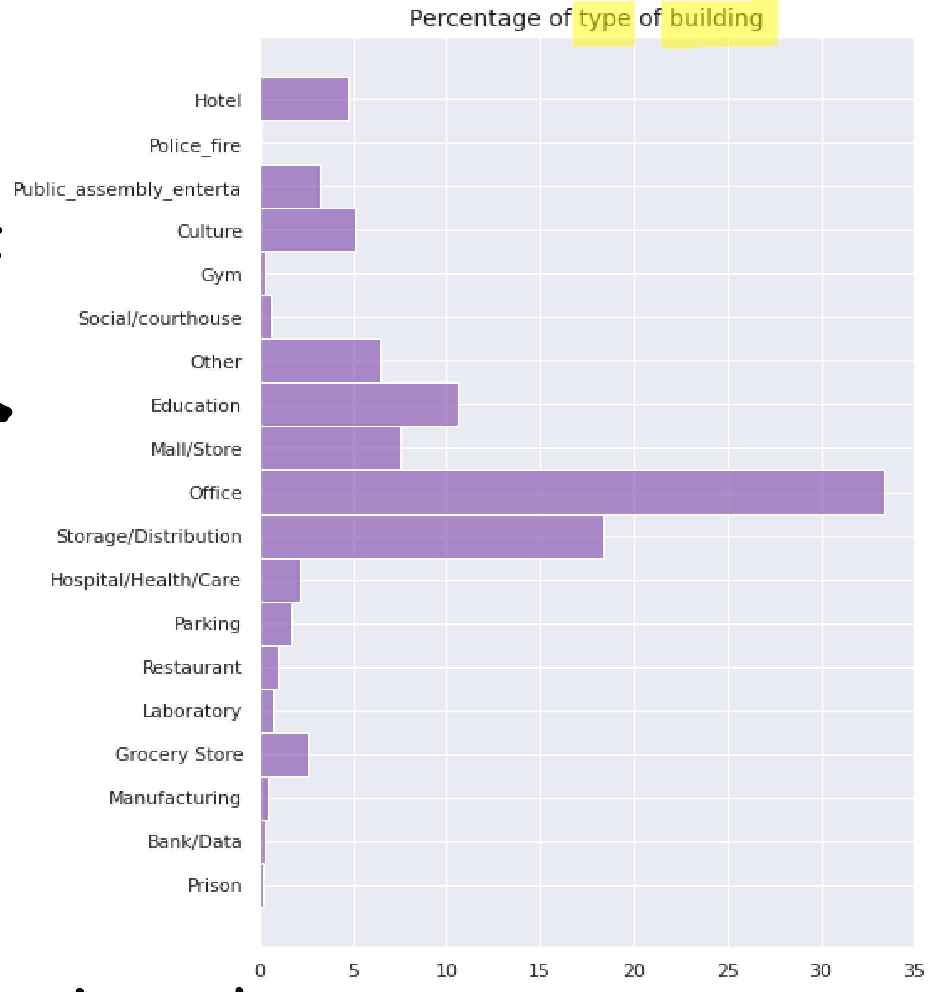
Analyse exploratoire: variables qualitatives



53 catégories

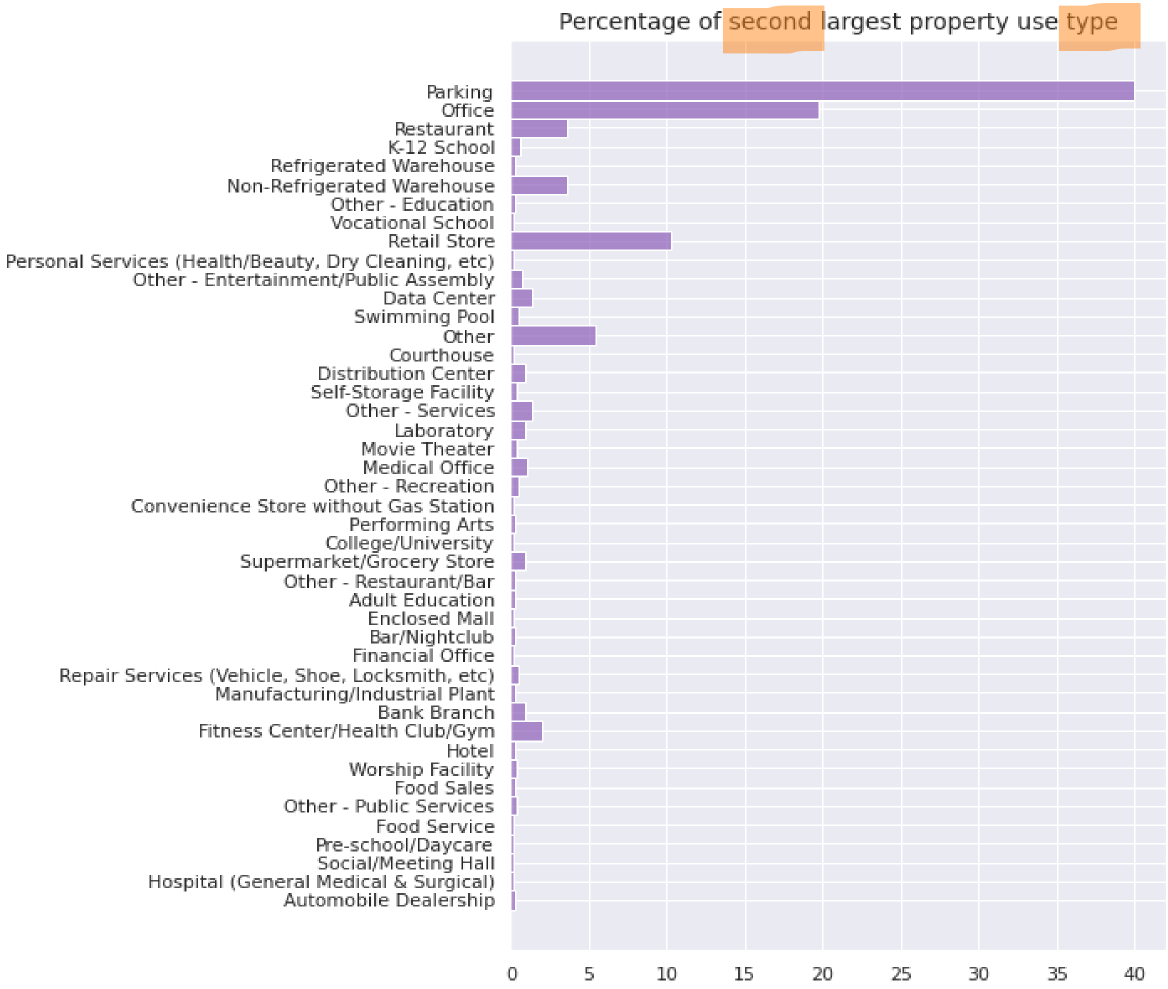
regroupement des catégories

Nouvelle variable



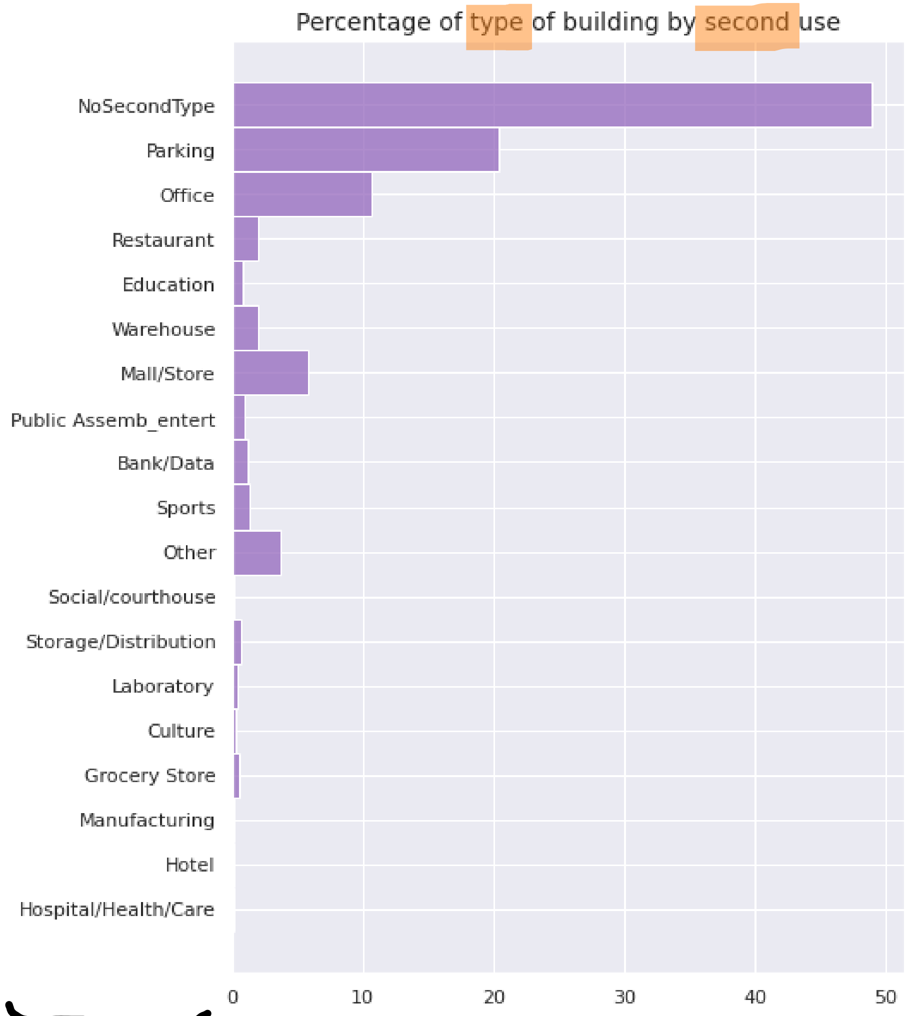
19 catégories

Analyse exploratoire: variables qualitatives



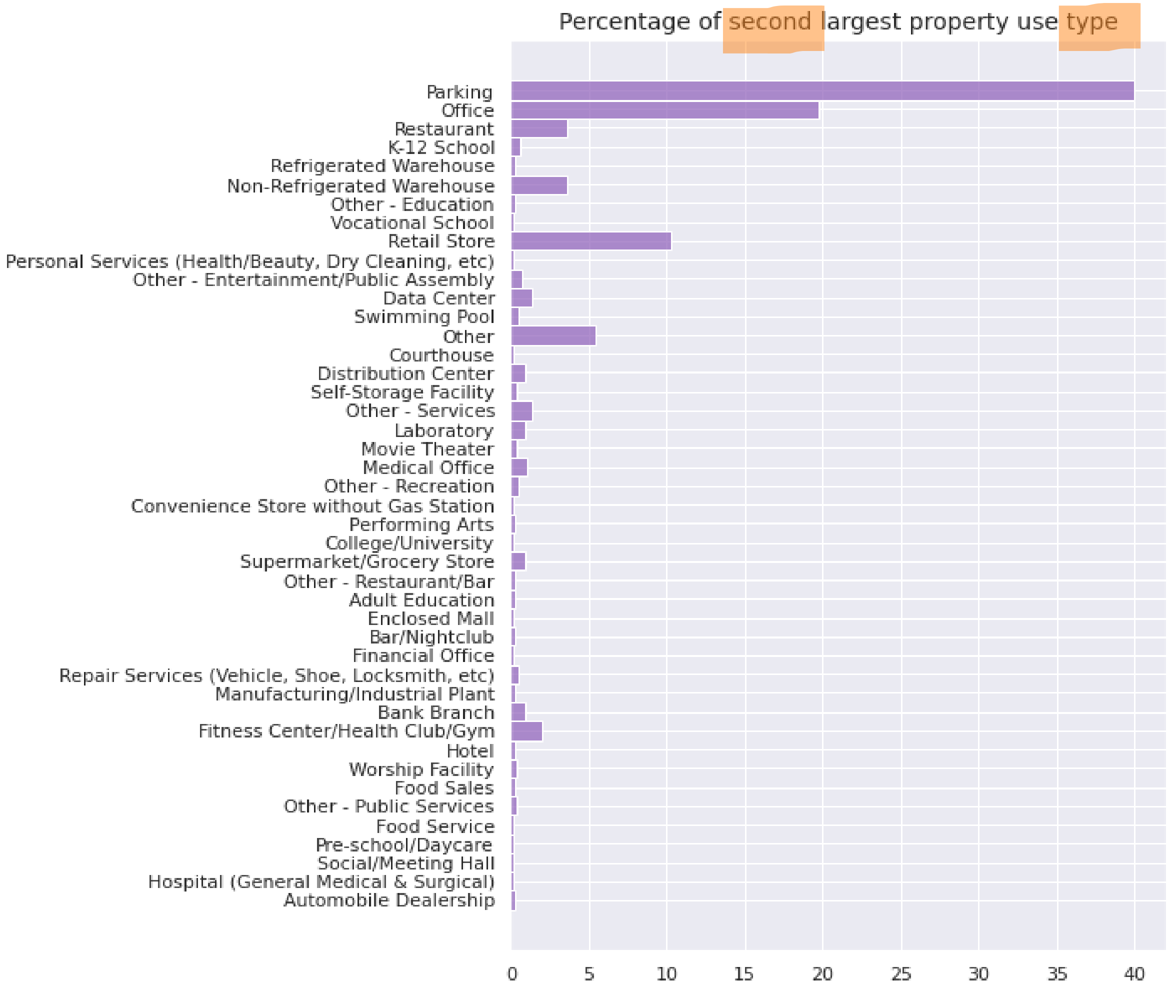
44 catégories

Nouvelle variable



19 catégories

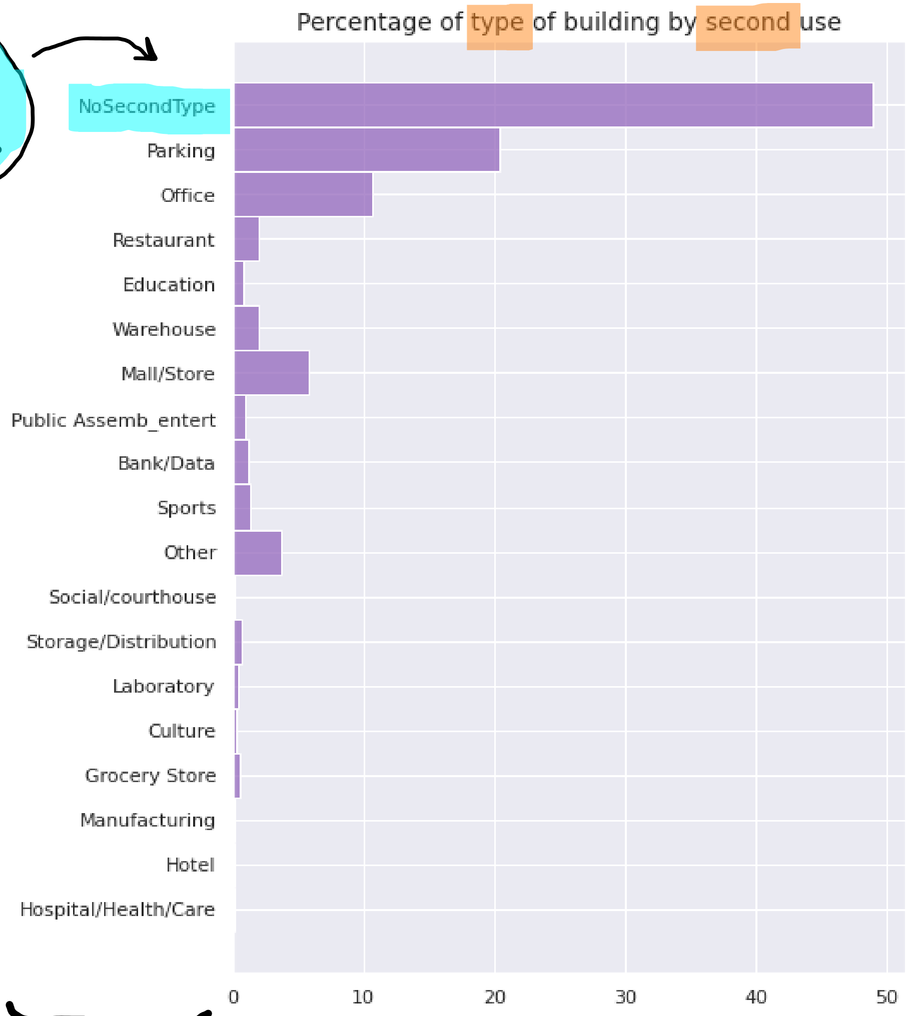
Analyse exploratoire: variables qualitatives



44 catégories

presque 50% sans second type

Nouvelle variable



19 catégories

Taille finale des données

	Avant	Après
Nombre d'individus	3376	1579
Nombre de variables	46	13

3

Prédiction

Prédiction : préparation données

Latitude
Longitude
YearBuilt
NumberofFloors
PropertyGFAParking
BuildType
BuildSecType
LargestPropertyUseTypeGFA
2ndLargestPropertyUseTypeGFA

Prédiction : préparation données

Latitude
Longitude
YearBuilt
NumberofFloors
PropertyGFAParking
BuildType
BuildSecType
LargestPropertyUseTypeGFA
2ndLargestPropertyUseTypeGFA

Consommation Énergie

Émmissions des gaz

Cible 1	Cible 2
SiteEnergyUse(kBtu)	TotalGHGEmissions

Prédiction : préparation données

Latitude
Longitude
YearBuilt
NumberofFloors
PropertyGFAParking
BuildType
BuildSecType
LargestPropertyUseTypeGFA
2ndLargestPropertyUseTypeGFA

Traslation :

1 Age = 2015 - YearBuilt

Cible 1	Cible 2
SiteEnergyUse(kBtu)	TotalGHGEmissions

Prédiction : préparation données

Latitude
Longitude
YearBuilt
NumberofFloors
PropertyGFAParking
BuildType
BuildSecType
LargestPropertyUseTypeGFA
2ndLargestPropertyUseTypeGFA

Cible 1	Cible 2
SiteEnergyUse(kBtu)	TotalGHGEmissions

1 Traslation :

Age = 2015 - YearBuilt

2 Codage avec :

```
X = pd.get_dummies(X)
```

Prédiction : préparation données

Latitude
Longitude
YearBuilt
NumberofFloors
PropertyGFAParking
BuildType
BuildSecType
LargestPropertyUseTypeGFA
2ndLargestPropertyUseTypeGFA

Cible 1	Cible 2
SiteEnergyUse(kBtu)	TotalGHGEmissions

1 Traslation :

Age = 2015 - YearBuilt

2 Codage avec :

```
X = pd.get_dummies(X)
```

3 X, y split X, y train X, y test

Prédiction : préparation données

Latitude
Longitude
YearBuilt
NumberofFloors
PropertyGFAParking
BuildType
BuildSecType
LargestPropertyUseTypeGFA
2ndLargestPropertyUseTypeGFA

Cible 1	Cible 2
SiteEnergyUse(kBtu)	TotalGHGEmissions

① Traslation :

Age = 2015 - YearBuilt

② Codage avec :

```
X = pd.get_dummies(X)
```

③ X,y split

X,y train

X,y test

④ Standardisation avec :

```
sc = StandardScaler()
```

Prédiction : pré-traitement

EnergyStarScore

?

+

Latitude
Longitude
YearBuilt
NumberofFloors
PropertyGFAParking
BuildType
BuildSecType
LargestPropertyUseTypeGFA
2ndLargestPropertyUseTypeGFA

Cible 1	Cible 2
SiteEnergyUse(kBtu)	TotalGHGEmissions

1 Traslation :

Age = 2015 - YearBuilt

2 Codage avec :

```
X = pd.get_dummies(X)
```

3 X, y split

X, y train

x, y test

4 Standardisation avec :

```
sc = StandardScaler()
```

Prédiction : pré-traitement

EnergyStarScore

? ~ 500 individus sans

+

Latitude
Longitude
YearBuilt
NumberofFloors
PropertyGFAParking
BuildType
BuildSecType
LargestPropertyUseTypeGFA
2ndLargestPropertyUseTypeGFA

Cible 1	Cible 2
SiteEnergyUse(kBtu)	TotalGHGEmissions

1 Traslation :

Age = 2015 - YearBuilt

2 Codage avec :

```
X = pd.get_dummies(X)
```

3 X, y split X, y train x, y test

4 Standardisation avec :

```
sc = StandardScaler()
```

Prédiction : sans EnergyStarScore

Énergie

Name	Train_R2_score	Test_R2_score	RMSE	Fit_time	Predict_time
LinearRegression	0.6719	0.6947	6.180426e+06	0.024718	0.013019
Ridge_afterCV	0.6663	0.6895	6.232314e+06	0.013252	0.003780
Lasso_afterCV	0.6571	0.6764	6.362830e+06	0.015944	0.009041
GradientBoosting	0.8448	0.6689	6.436380e+06	0.202172	0.002594
SVR	0.7330	0.6147	9.350737e+06	0.092125	0.018322
RandomForest_afterCV	0.8727	0.6101	6.984421e+06	0.700676	0.018322
XGBoost	0.9978	0.5982	7.090260e+06	0.220546	0.006631
XGBoost_reg	0.8078	0.5905	7.157458e+06	0.146926	0.009786
RandomForest_log_target	0.7347	0.5650	7.377150e+06	0.575726	0.021154
DummyRegressor	0.0000	-0.0001	1.118565e+07	0.000373	0.000211
Lasso_log_target	-0.1388	-0.1173	1.182292e+07	0.010973	0.009646
LinearRegression_log_target	-61.0221	-1052.4623	3.630409e+08	0.031511	0.010318
Ridge_log_target	-60.1682	-1203.0557	3.881229e+08	0.021637	0.004403

~~ENERGY STAR~~

↓

Résultats pas trés bons ni fiables

↑

très sensibles aux outliers

Prédiction : sans EnergyStarScore

Énergie

Name	Train_R2_score	Test_R2_score	RMSE	Fit_time	Predict_time
LinearRegression	0.6719	0.6947	6.180426e+06	0.024718	0.013019
Ridge_afterCV	0.6663	0.6895	6.232314e+06	0.013252	0.003780
Lasso_afterCV	0.6571	0.6764	6.362830e+06	0.015944	0.009041
GradientBoosting	0.8448	0.6689	6.436380e+06	0.202172	0.002594
SVR	0.7330	0.6147	9.350737e+06	0.092125	0.018322
RandomForest_afterCV	0.8727	0.6101	6.984421e+06	0.700676	0.018322
XGBoost	0.9978	0.5982	7.090260e+06	0.220546	0.006631
XGBoost_reg	0.8078	0.5905	7.157458e+06	0.146926	0.009786
RandomForest_log_target	0.7347	0.5650	7.377150e+06	0.575726	0.021154
DummyRegressor	0.0000	-0.0001	1.118565e+07	0.000373	0.000211
Lasso_log_target	-0.1388	-0.1173	1.182292e+07	0.010973	0.009646
LinearRegression_log_target	-61.0221	-1052.4623	3.630409e+08	0.031511	0.010318
Ridge_log_target	-60.1682	-1203.0557	3.881229e+08	0.021637	0.004403

+ bas qu'attendu

Résultats pas trés bons ni fiables

très sensibles aux outliers

Prédiction : sans EnergyStarScore 3

Énergie

Name	Train_R2_score	Test_R2_score	RMSE	Fit_time	Predict_time
LinearRegression	0.6719	0.6947	6.180426e+06	0.024718	0.013019
Ridge_afterCV	0.6663	0.6895	6.232314e+06	0.013252	0.003780
Lasso_afterCV	0.6571	0.6764	6.362830e+06	0.015944	0.009041
GradientBoosting	0.8448	0.6689	6.436380e+06	0.202172	0.002594
SVR	0.7330	0.6147	9.350737e+06	0.092125	0.018322
RandomForest_afterCV	0.8727	0.6101	6.984421e+06	0.700676	0.018322
XGBoost	0.9978	0.5982	7.090260e+06	0.220546	0.006631
XGBoost_reg	0.8078	0.5905	7.157458e+06	0.146926	0.009786
RandomForest_log_target	0.7347	0.5650	7.377150e+06	0.575726	0.021154
DummyRegressor	0.0000	-0.0001	1.118565e+07	0.000373	0.000211
Lasso_log_target	-0.1388	-0.1173	1.182292e+07	0.010973	0.009646
LinearRegression_log_target	-61.0221	-1052.4623	3.630409e+08	0.031511	0.010318
Ridge_log_target	-60.1682	-1203.0557	3.881229e+08	0.021637	0.004403

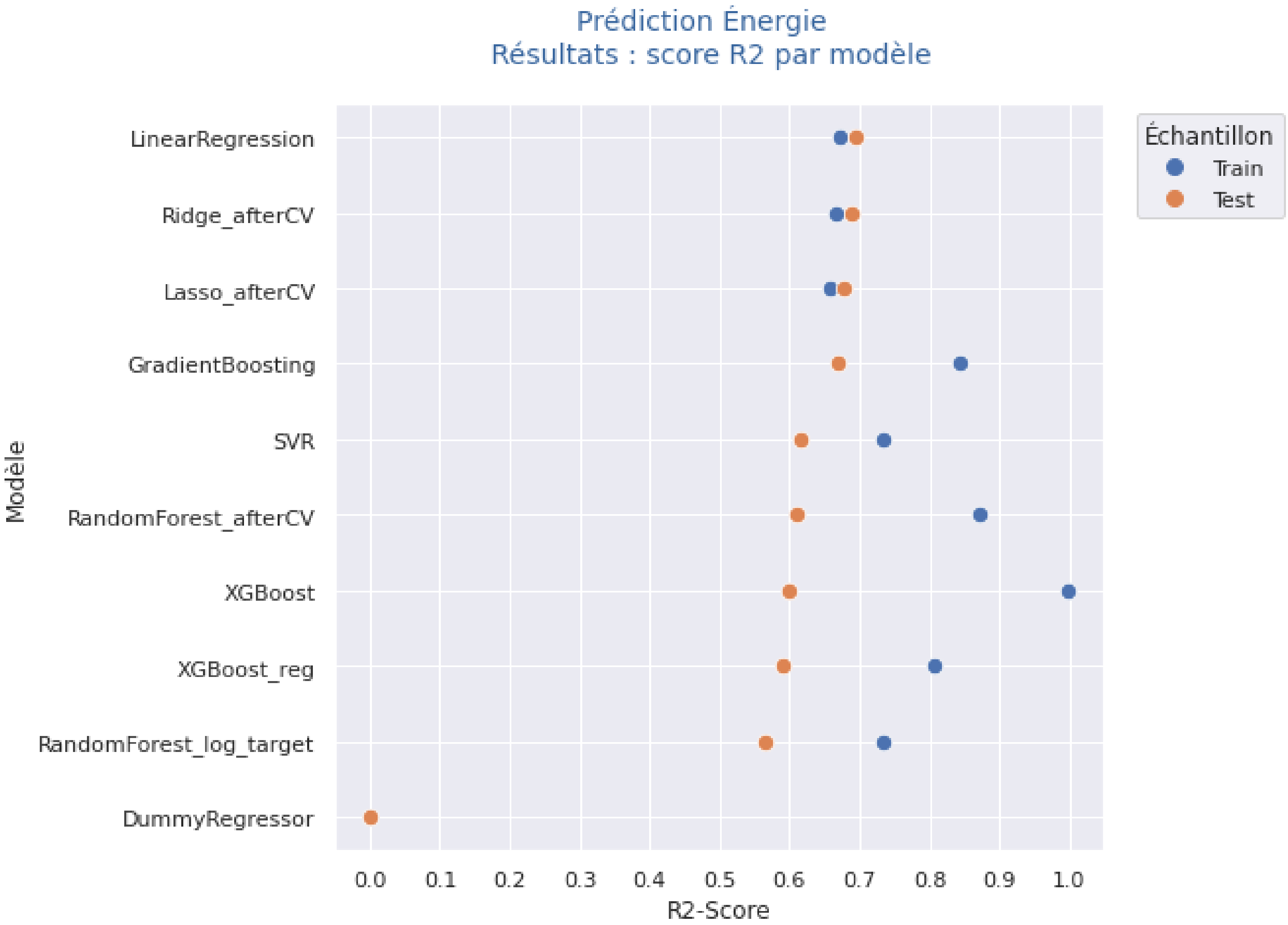
Transformation target : $y \longmapsto \log(1+y)$

Résultats pas trés bons ni fiables

très sensibles aux outliers

Prédiction : sans EnergyStarScore

Énergie



Prédiction : sans EnergyStarScore

Gaz

Name	Train_R2_score	Test_R2_score	RMSE	Fit_time	Predict_time
GradientBoosting	0.9317	0.5826	135.80	0.324040	0.005314
LinearRegression	0.4474	0.5328	143.67	0.013099	0.016180
Lasso_afterCV	0.4131	0.5286	144.31	0.007946	0.009279
Ridge_afterCV	0.4314	0.5195	145.70	0.010202	0.012094
RandomForest_afterCV	0.8725	0.4737	152.48	0.645518	0.024444
RandomForest_log_target	0.7510	0.4430	156.87	0.537519	0.017747
SVR	0.6307	0.4076	154.60	0.119469	0.024444
XGBoost_reg	0.6403	0.3898	164.19	1.241869	0.034905
XGBoost	0.9966	0.3803	165.47	0.180047	0.005178
DummyRegressor	0.0000	-0.0095	211.19	0.000570	0.000257
Lasso_log_target	-0.1072	-0.0892	219.37	0.011554	0.005156
LinearRegression_log_target	-24.5012	-112.0861	2235.24	0.018689	0.007828
Ridge_log_target	-21.8017	-128.1484	2388.72	0.019704	0.003900

ENERGY STAR

↓

Résultats pas trés bons ni fiables

↑

très sensibles aux outliers

Prédiction : sans EnergyStarScore

Gaz

Name	Train_R2_score	Test_R2_score	RMSE	Fit_time	Predict_time
GradientBoosting	0.9317	0.5826	135.80	0.324040	0.005314
LinearRegression	0.4474	0.5328	143.67	0.013099	0.016180
Lasso_afterCV	0.4131	0.5286	144.31	0.007946	0.009279
Ridge_afterCV	0.4314	0.5195	145.70	0.010202	0.012094
RandomForest_afterCV	0.8725	0.4737	152.48	0.645518	0.024444
RandomForest_log_target	0.7510	0.4430	156.87	0.537519	0.017747
SVR	0.6307	0.4076	154.60	0.119469	0.024444
XGBoost_reg	0.6403	0.3898	164.19	1.241869	0.034905
XGBoost	0.9966	0.3803	165.47	0.180047	0.005178
DummyRegressor	0.0000	-0.0095	211.19	0.000570	0.000257
Lasso_log_target	-0.1072	-0.0892	219.37	0.011554	0.005156
LinearRegression_log_target	-24.5012	-112.0861	2235.24	0.018689	0.007828
Ridge_log_target	-21.8017	-128.1484	2388.72	0.019704	0.003900

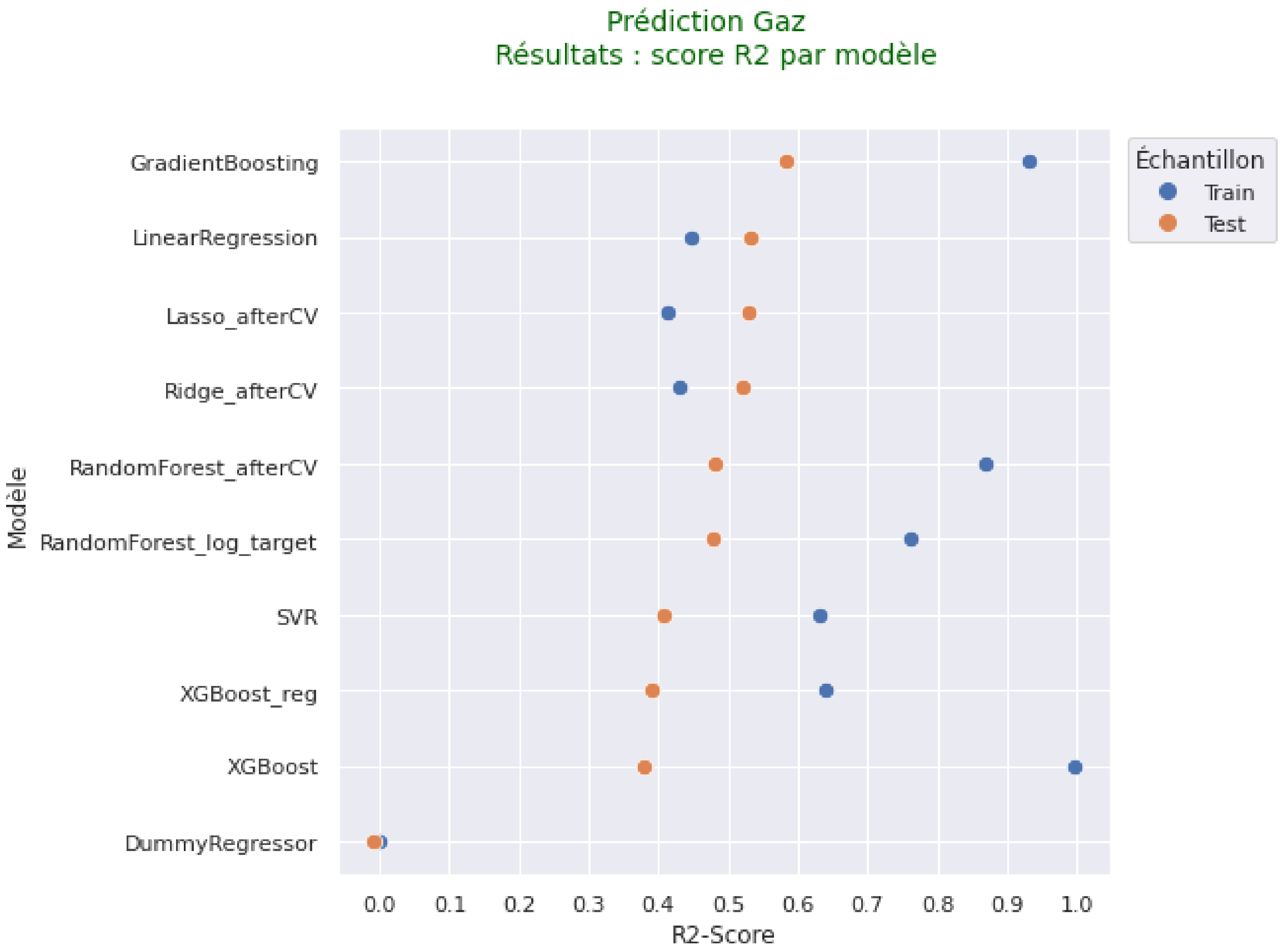
Résultats pas très bons ni fiables

très sensibles aux outliers

Performances moins bonnes en général

Prédiction : sans EnergyStarScore

Gaz



Prédiction :
avec EnergyStarScore

Prédiction : avec EnergyStarScore

Énergie

Name	Train_R2_score	Test_R2_score	RMSE	Fit_time	Predict_time
SVR	0.8136	0.7542	9094546.92	0.060459	0.009005
Ridge_afterCV	0.7930	0.7473	5006566.91	0.008954	0.009605
Lasso_afterCV	0.7704	0.7431	5047915.09	0.009315	0.004007
LinearRegression	0.8011	0.7332	5144866.79	0.006419	0.009903
GradientBoosting	0.8430	0.7193	5276750.41	0.151077	0.002100
XGBoost_reg	0.8464	0.7148	5318697.34	0.106857	0.005229
RandomForest_afterCV	0.8220	0.6538	5860681.32	0.216160	0.009005
DummyRegressor	0.0000	-0.0062	9990923.04	0.000577	0.000342

Prédiction : avec EnergyStarScore

Énergie

Name	Train_R2_score	Test_R2_score	RMSE	Fit_time	Predict_time
SVR	0.8136	0.7542	9094546.92	0.060459	0.009005
Ridge_afterCV	0.7930	0.7473	5006566.91	0.008954	0.009605
Lasso_afterCV	0.7704	0.7431	5047915.09	0.009315	0.004007
LinearRegression	0.8011	0.7332	5144866.79	0.006419	0.009903
GradientBoosting	0.8430	0.7193	5276750.41	0.151077	0.002100
XGBoost_reg	0.8464	0.7148	5318697.34	0.106857	0.005229
RandomForest_afterCV	0.8220	0.6538	5860681.32	0.216160	0.009005
DummyRegressor	0.0000	-0.0062	9990923.04	0.000577	0.000342

Choix d'hyperparamètres :

validation croisée

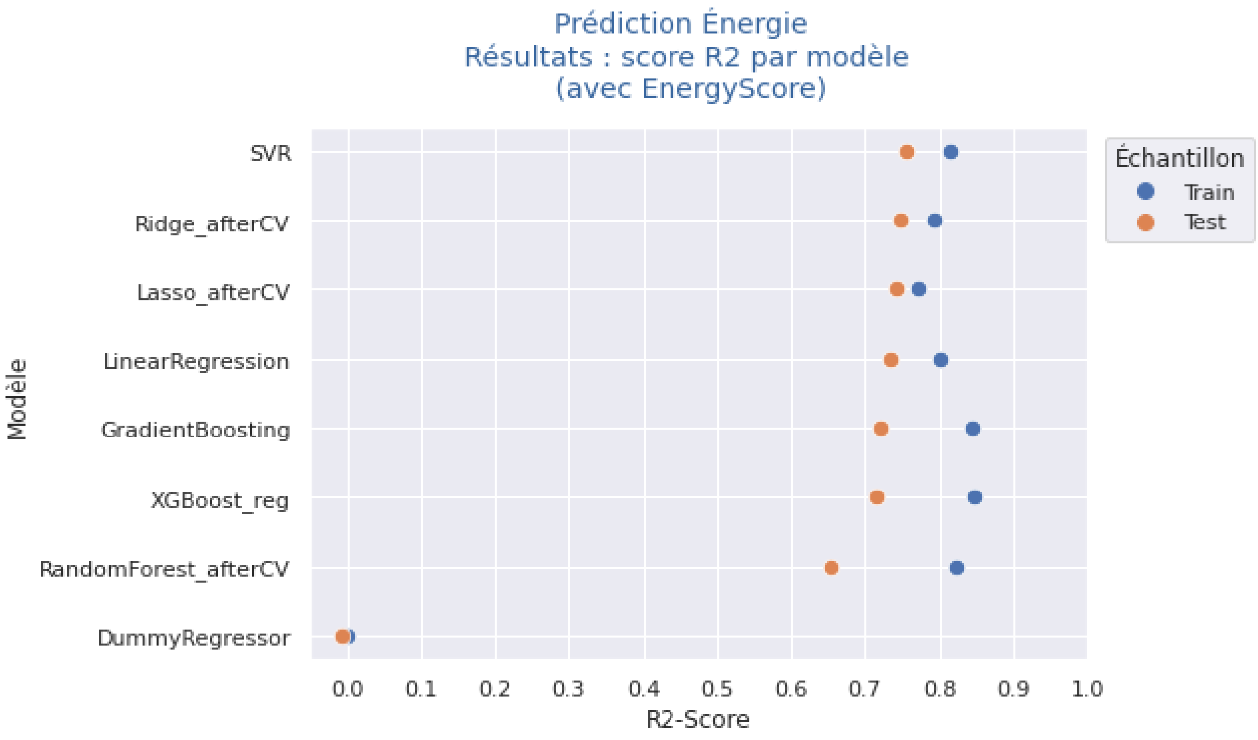
```
grid = GridSearchCV(estimator,
                    params,
                    cv=5,
                    n_jobs=-1,
                    return_train_score=True)
```

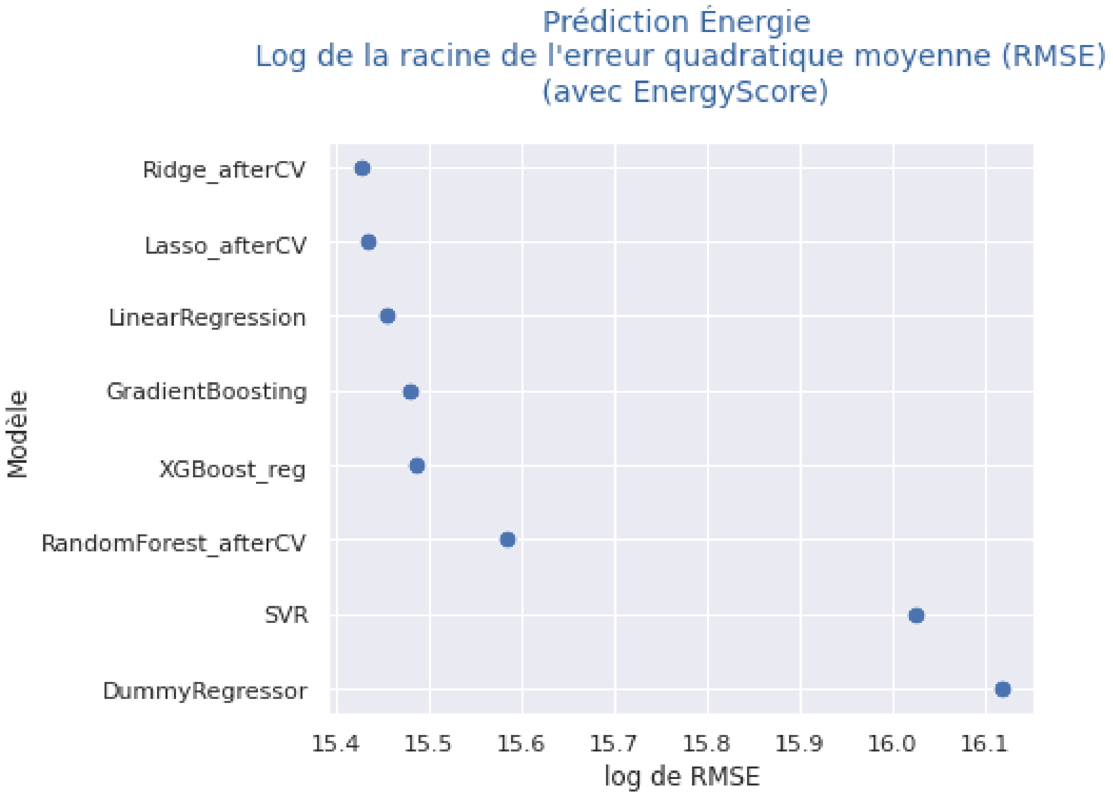
Ex : regularisation alpha, lambda
max depth, max samples, etc.

5, 10

Prédiction : avec EnergyStarScore

Énergie





Prédiction : avec EnergyStarScore

Énergie

Modèle choisi

Energie

Name	Train_R2_score	Test_R2_score	RMSE	Fit_time	Predict_time
SVR	0.8136	0.7542	9094546.92	0.060459	0.009005
Ridge_afterCV	0.7930	0.7473	5006566.91	0.008954	0.009605
Lasso_afterCV	0.7704	0.7431	5047915.09	0.009315	0.004007
LinearRegression	0.8011	0.7332	5144866.79	0.006419	0.009903
GradientBoosting	0.8430	0.7193	5276750.41	0.151077	0.002100
XGBoost_reg	0.8464	0.7148	5318697.34	0.106857	0.005229
RandomForest_afterCV	0.8220	0.6538	5860681.32	0.216160	0.009005
DummyRegressor	0.0000	-0.0062	9990923.04	0.000577	0.000342

Prédiction : avec EnergyStarScore

Gaz

Name	Train_R2_score	Test_R2_score	RMSE	Fit_time	Predict_time
GradientBoosting	0.8163	0.7302	130.79	0.131098	0.001697
RandomForest_afterCV	0.8513	0.6554	147.81	0.226076	0.009129
XGBoost_reg	0.7896	0.5879	161.64	0.060390	0.003654
SVR	0.6917	0.5494	154.22	0.055092	0.009129
LinearRegression	0.5545	0.4999	178.06	0.037742	0.007469
Ridge_afterCV	0.5281	0.4930	179.28	0.005924	0.009302
Lasso_afterCV	0.5027	0.4847	180.74	0.015530	0.003676
DummyRegressor	0.0000	-0.0006	251.86	0.000674	0.000336

Choix d'hyperparamètres :

validation croisée

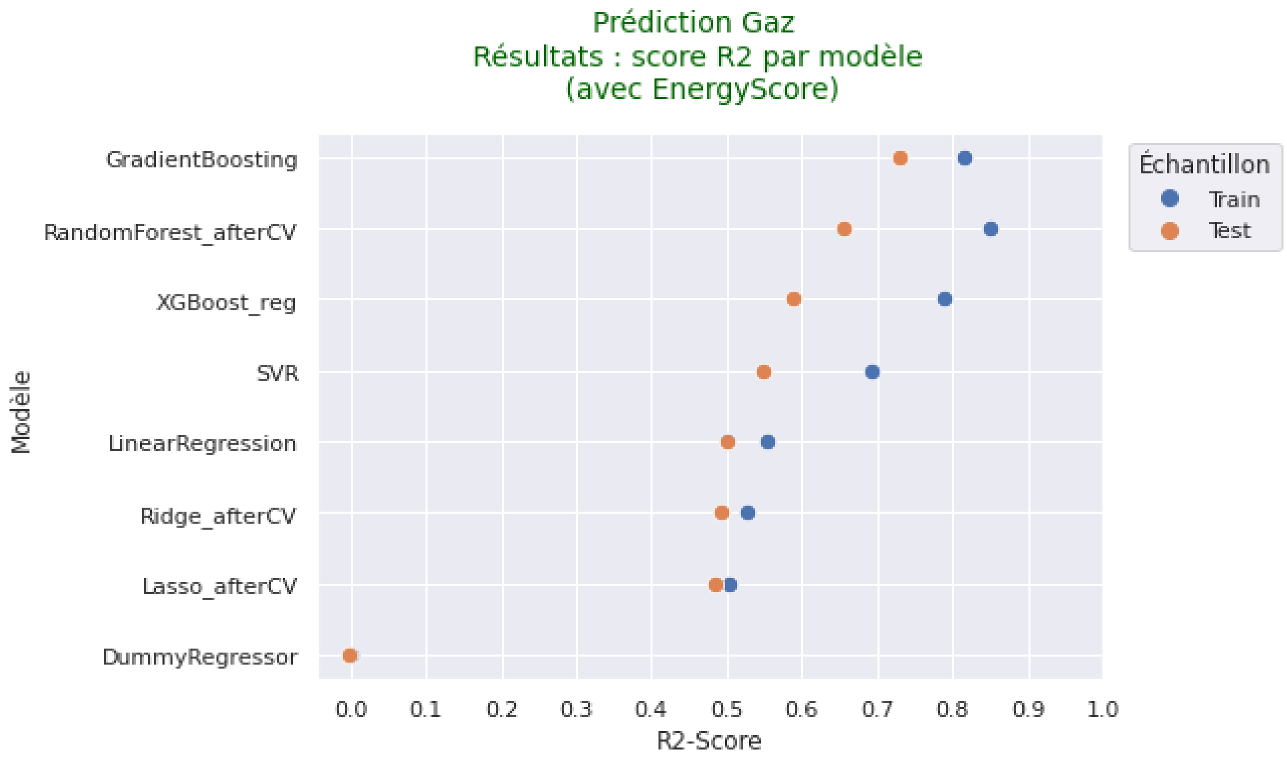
```
grid = GridSearchCV(estimator,
                    params,
                    cv=5,
                    n_jobs=-1,
                    return_train_score=True)
```

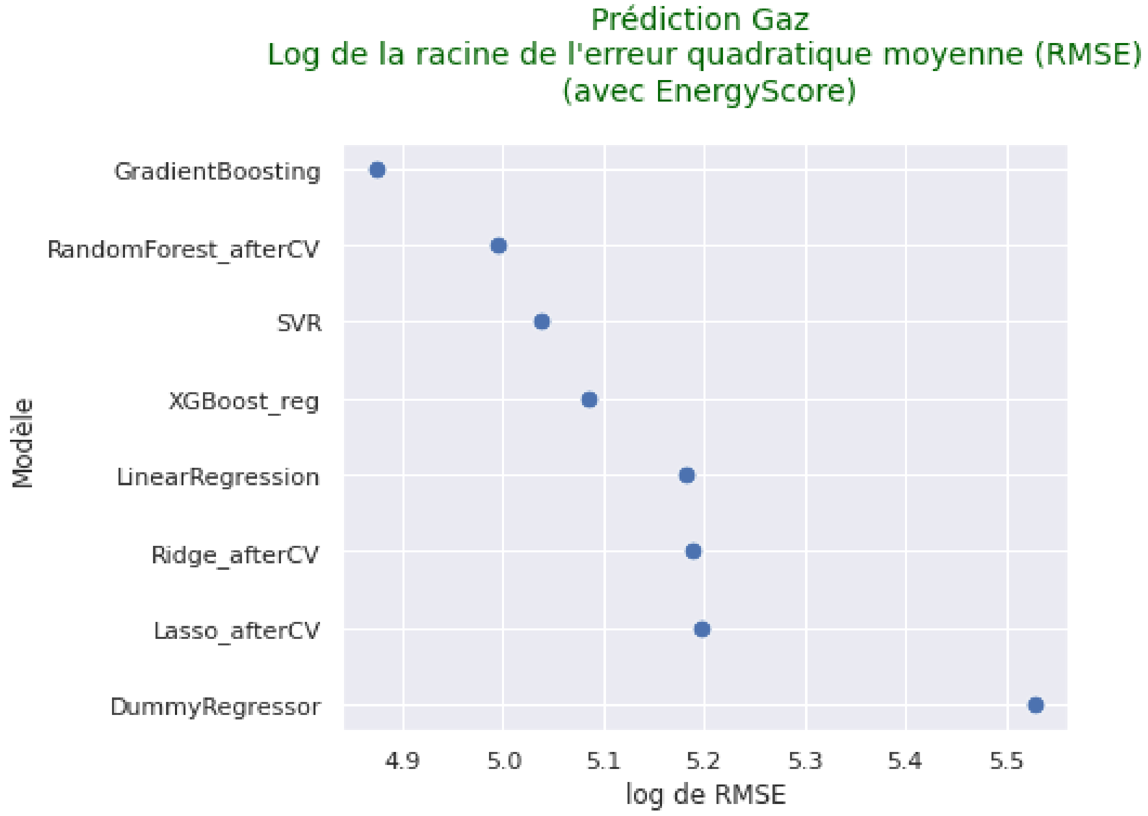
Ex : regularisation alpha, lambda
max depth, max samples, etc.

5, 10

Prédiction : avec EnergyStarScore

Gaz





Prédiction : avec EnergyStarScore

Gaz

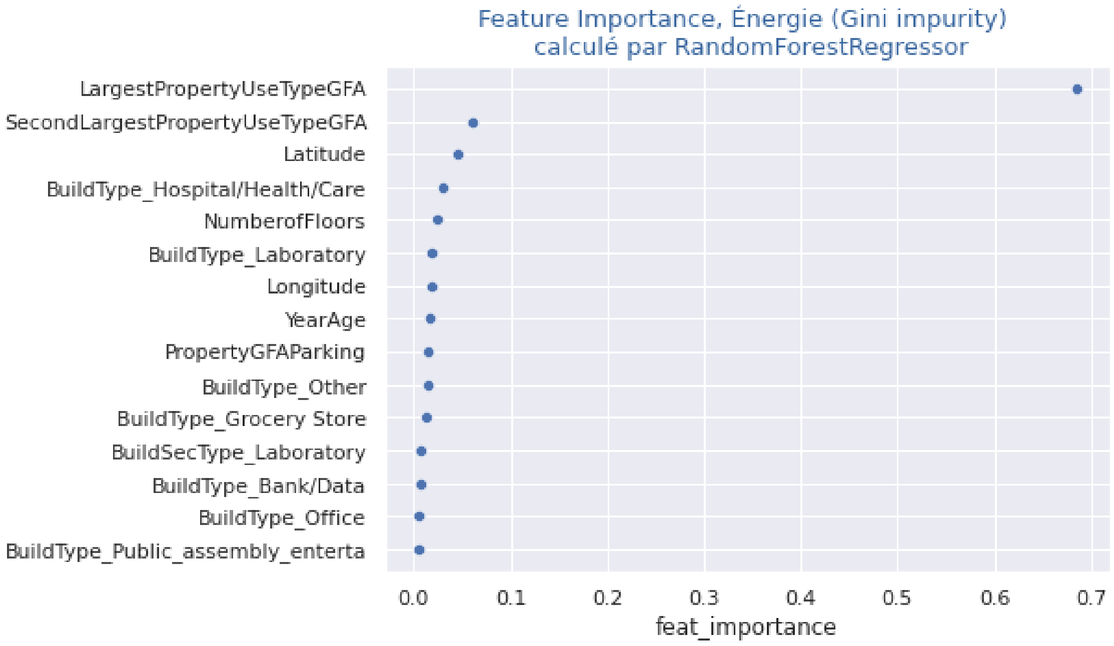
Modèle choisi gaz

Name	Train_R2_score	Test_R2_score	RMSE	Fit_time	Predict_time
GradientBoosting	0.8163	0.7302	130.79	0.131098	0.001697
RandomForest_afterCV	0.8513	0.6554	147.81	0.226076	0.009129
XGBoost_reg	0.7896	0.5879	161.64	0.060390	0.003654
SVR	0.6917	0.5494	154.22	0.055092	0.009129
LinearRegression	0.5545	0.4999	178.06	0.037742	0.007469
Ridge_afterCV	0.5281	0.4930	179.28	0.005924	0.009302
Lasso_afterCV	0.5027	0.4847	180.74	0.015530	0.003676
DummyRegressor	0.0000	-0.0006	251.86	0.000674	0.000336

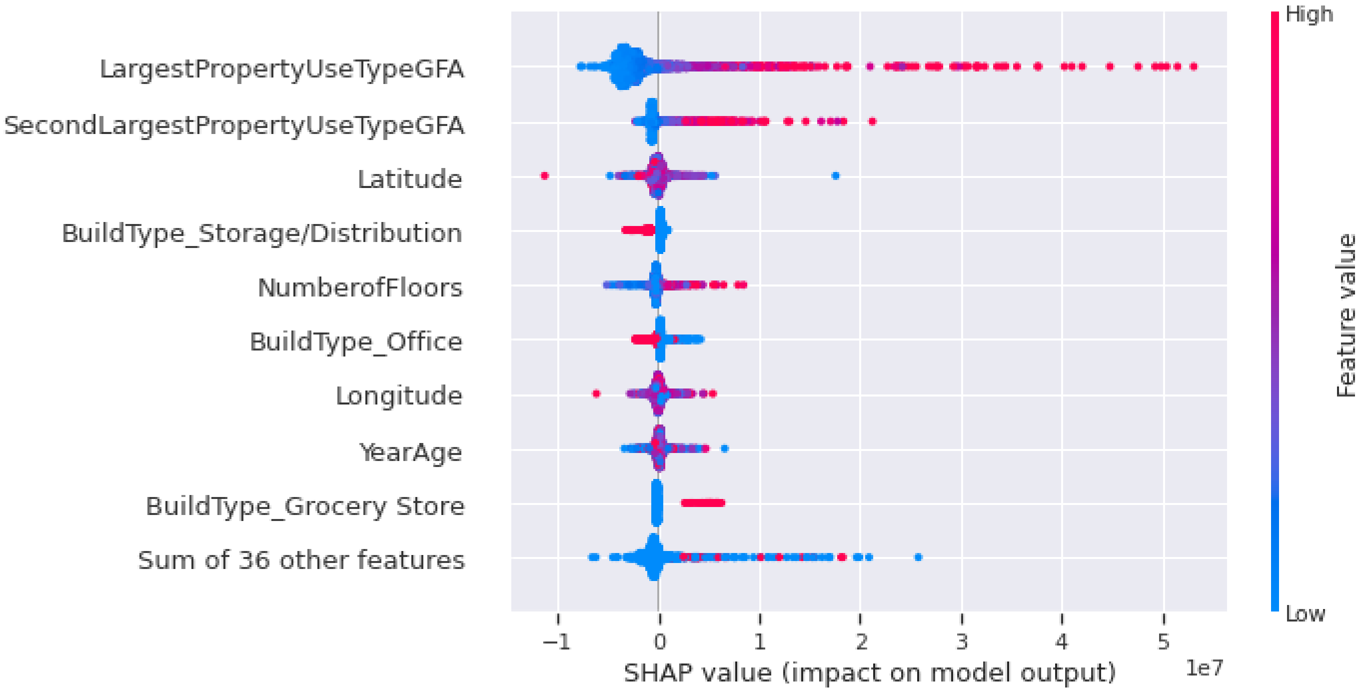
Extras : Feature importance

Pour la consommation d'**énergie**

Avec feature importance de scikit learn



Avec SHAP



4

Conclusions

Conclusions

- Variable **EnergyStarScore** nécessaire.
- Modèles/(nettoyage de données) très **sensibles** !
- **Ajout** de variables -> amélioration importante performances.
- Modèles **linéaires** fonctionnent assez bien pour ce cas.
- Modèle **RandomForest** -> possible amélioration.
- Il n'y a pas eu du data leakage.