PROJET SEATTLE

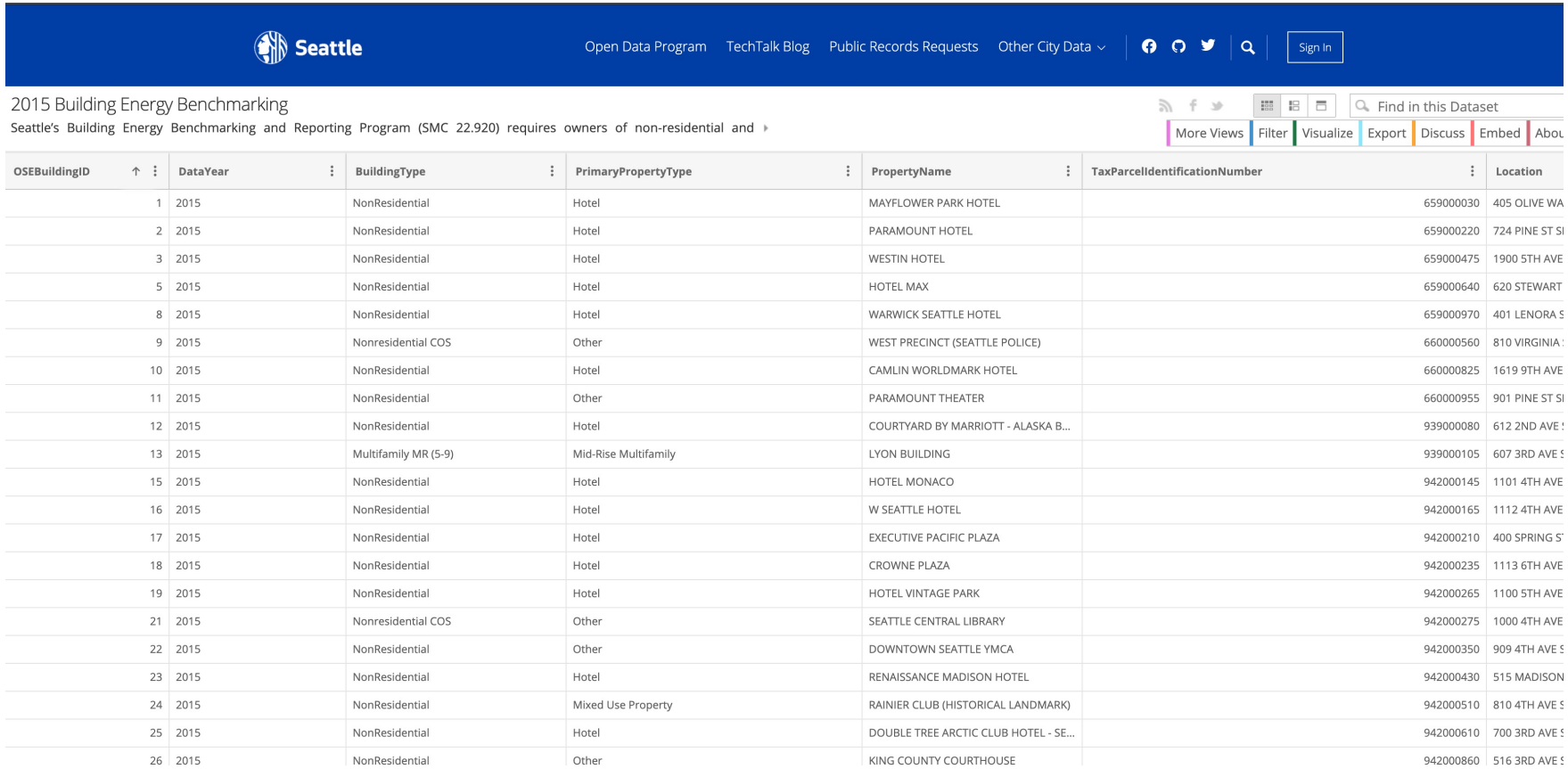
P4 Openclassrooms

PRÉSENTATION DU PROJET

Prédire les émissions de CO2 (+Energystar)

Prédire la consommation totale d'énergie

LA BASE DE DONNÉES



Seattle | Open Data Program | TechTalk Blog | Public Records Requests | Other City Data | Sign In

2015 Building Energy Benchmarking

Seattle's Building Energy Benchmarking and Reporting Program (SMC 22.920) requires owners of non-residential and

Find in this Dataset

More Views | Filter | Visualize | Export | Discuss | Embed | Abou

OSEBuildingID	DataYear	BuildingType	PrimaryPropertyType	PropertyName	TaxParcelIdentificationNumber	Location
1	2015	NonResidential	Hotel	MAYFLOWER PARK HOTEL	659000030	405 OLIVE WA
2	2015	NonResidential	Hotel	PARAMOUNT HOTEL	659000220	724 PINE ST SI
3	2015	NonResidential	Hotel	WESTIN HOTEL	659000475	1900 5TH AVE
5	2015	NonResidential	Hotel	HOTEL MAX	659000640	620 STEWART
8	2015	NonResidential	Hotel	WARWICK SEATTLE HOTEL	659000970	401 LENORA S
9	2015	Nonresidential COS	Other	WEST PRECINCT (SEATTLE POLICE)	660000560	810 VIRGINIA
10	2015	NonResidential	Hotel	CAMLIN WORLDMARK HOTEL	660000825	1619 9TH AVE
11	2015	NonResidential	Other	PARAMOUNT THEATER	660000955	901 PINE ST SI
12	2015	NonResidential	Hotel	COURTYARD BY MARRIOTT - ALASKA B...	939000080	612 2ND AVE
13	2015	Multifamily MR (5-9)	Mid-Rise Multifamily	LYON BUILDING	939000105	607 3RD AVE S
15	2015	NonResidential	Hotel	HOTEL MONACO	942000145	1101 4TH AVE
16	2015	NonResidential	Hotel	W SEATTLE HOTEL	942000165	1112 4TH AVE
17	2015	NonResidential	Hotel	EXECUTIVE PACIFIC PLAZA	942000210	400 SPRING ST
18	2015	NonResidential	Hotel	CROWNE PLAZA	942000235	1113 6TH AVE
19	2015	NonResidential	Hotel	HOTEL VINTAGE PARK	942000265	1100 5TH AVE
21	2015	Nonresidential COS	Other	SEATTLE CENTRAL LIBRARY	942000275	1000 4TH AVE
22	2015	NonResidential	Other	DOWNTOWN SEATTLE YMCA	942000350	909 4TH AVE S
23	2015	NonResidential	Hotel	RENAISSANCE MADISON HOTEL	942000430	515 MADISON
24	2015	NonResidential	Mixed Use Property	RAINIER CLUB (HISTORICAL LANDMARK)	942000510	810 4TH AVE S
25	2015	NonResidential	Hotel	DOUBLE TREE ARCTIC CLUB HOTEL - SE...	942000610	700 3RD AVE S
26	2015	NonResidential	Other	KING COUNTY COURTHOUSE	942000860	516 3RD AVE S

NOS VARIABLES CLÉS

Emission de CO_2 ⇔ TotalGHGEmissions

Production totale d'énergie ⇔ SiteEnergyUseWN(kBtu)

Energystar Score ⇔ ENERGYSTARScore

NOTRE APPROCHE ML

- S’orienter vers un problème de regression
- Utiliser un maximum de données
- Obtenir un modèle performant

SOMMAIRE

ANALYSE EXPLORATOIRE ET NETTOYAGE

NETTOYAGE : GESTION DES VALEURS MANQUANTES

- On a supprimé les colonnes avec plus de 40% de valeurs manquantes
- On a supprimé les lignes ou nos variables clés étaient absentes

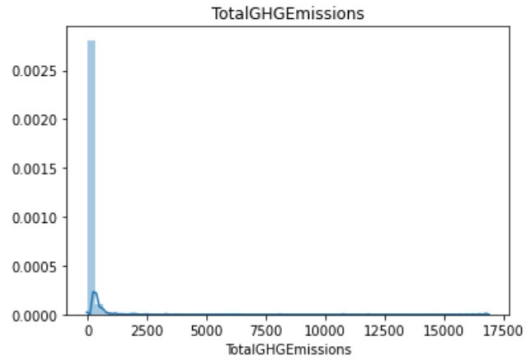
```
memory usage: 2.0+ MB
```

[10]:
```
# Nous supprimons les colonnes avec plus de 40% de valeurs manquantes.
df=df.drop(['ThirdLargestPropertyUseTypeGFA', 'ThirdLargestPropertyUseType',"SecondLargestPropertyUseTypeGFA" ,'Secor
```

[11]:
```
#drop si les variables d'intérêts ne sont pas la
df=df[df['ENERGYSTARScore'].isnull()==0]
df=df[df['SiteEnergyUseWN(kBtu)'].isnull()==0]
df=df[df['TotalGHGEmissions'].isnull()==0]
```

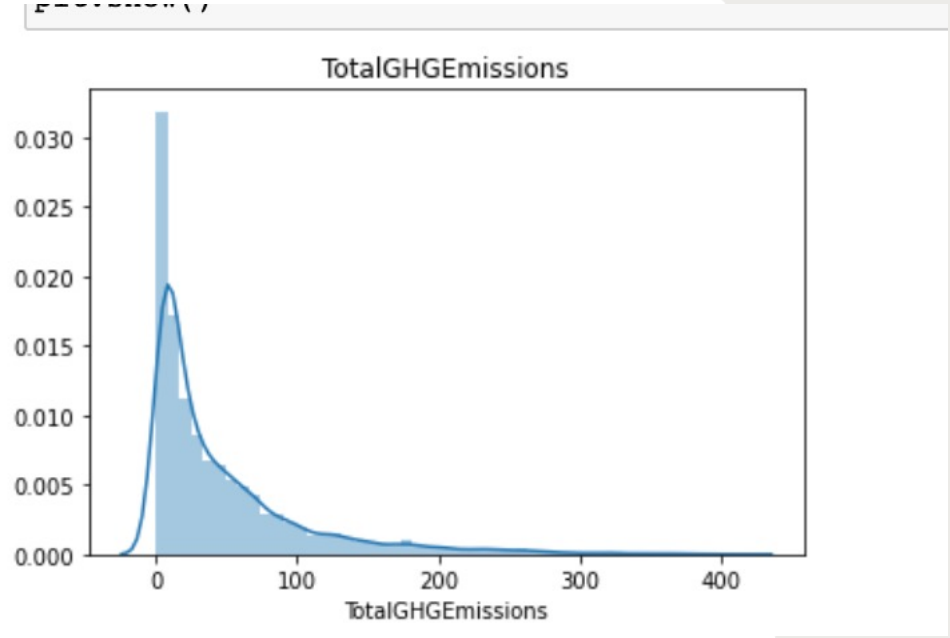
NETTOYAGE : GESTION DES OUTLIERS

Entrée [21]:
```
plt.title("TotalGHGEmissions")
x = pd.Series(df["TotalGHGEmissions"], name="TotalGHGEmissions")
ax = sns.distplot(x)
plt.show()
```



Entrée [22]:
```
x.describe()
```

Out[22]:
```
count     5092.000000
mean       107.449998
std        491.772634
min         -0.800000
25%          8.847500
50%         31.495000
75%         86.967500
max      16870.980000
Name: TotalGHGEmissions, dtype: float64
```



FEATURE ENGINEERING : DICHOTOMISATION

Création variables

Entrée [40]:
```
df['NumberofFloors'].unique()
```

Out[40]:
```
array([12., 11.,  6.,  9., 10.,  2.,  1.,  5.,  4.,  3., 20.,  7., 17.,
       22.,  8., 14., 15., 26., 13., 24., 23., 33., 18., 16., 99., 42.,
       32., 27.,  0.])
```

Entrée [41]:
```
df['NumberofFloors_group'] = np.where(df['NumberofFloors']>3, "Plus de 3 étages", "Moins de 3 étages")
df['NumberofFloors_group'].value_counts()
```

Out[41]:
```
Plus de 3 étages     2464
Moins de 3 étages    2101
Name: NumberofFloors_group, dtype: int64
```

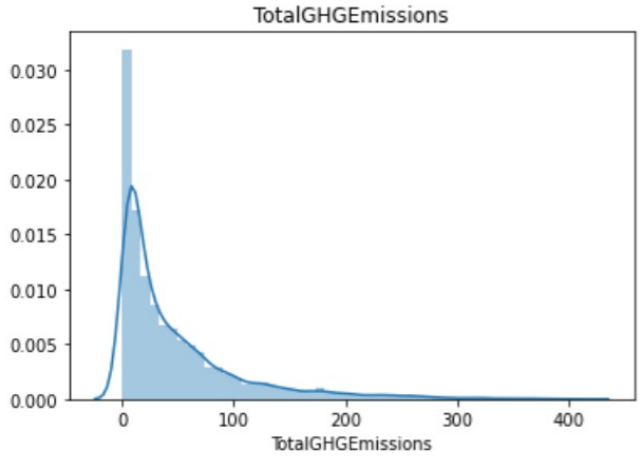
Entrée [42]:
```
df['YearBuilt'] = np.where(df['YearBuilt']>1999, "21ème siècle", "20ème siècle")
df['YearBuilt'].head()
```

Out[42]:
```
0     20ème siècle
1     20ème siècle
6     20ème siècle
9     20ème siècle
12    20ème siècle
Name: YearBuilt, dtype: object
```

FEATURE ENGINEERING : LOGARITHME

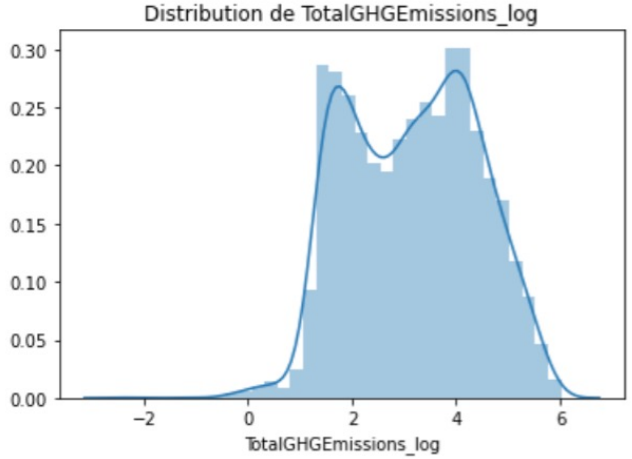
- Pour obtenir de meilleur résultat en modélisation

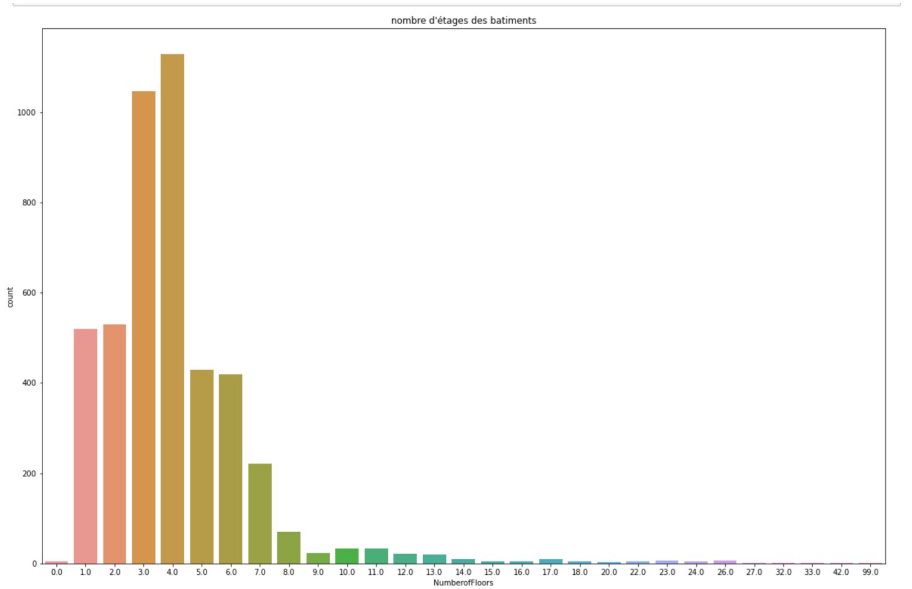
```
]: plt.title("TotalGHGEmissions")
   x = pd.Series(df["TotalGHGEmissions"], name="TotalGHGEmissions")
   ax = sns.distplot(x)
   plt.show()
```

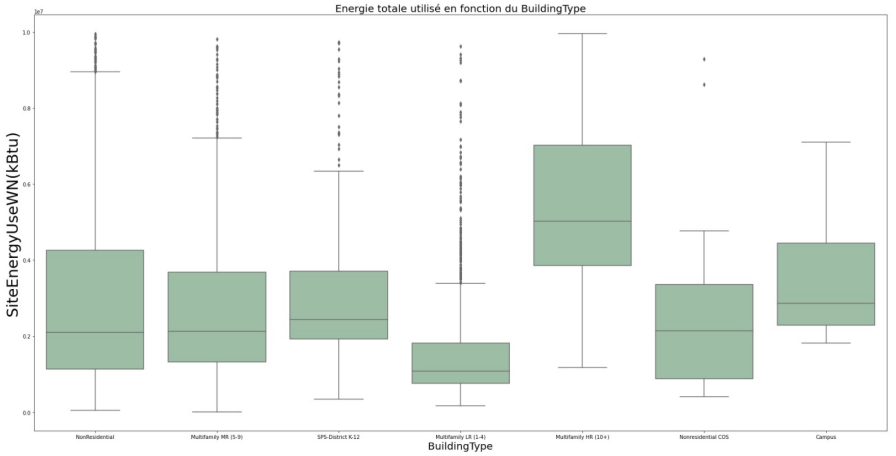



```
Entrée [28]: df["TotalGHGEmissions_log"]=np.log(df["TotalGHGEmissions"])
```

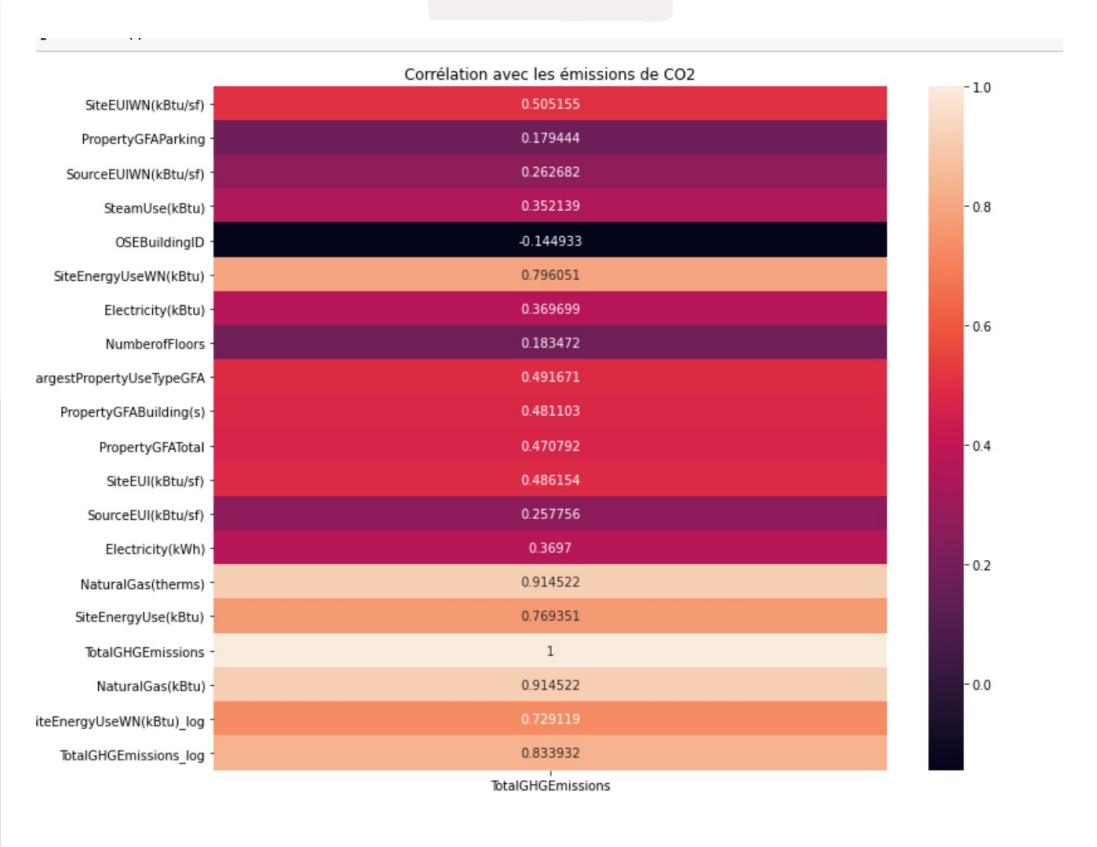
```
Entrée [29]: plt.title("Distribution de TotalGHGEmissions_log")
   x = pd.Series(df["TotalGHGEmissions_log"], name="TotalGHGEmissions_log")
   ax = sns.distplot(x)
   plt.show()
```

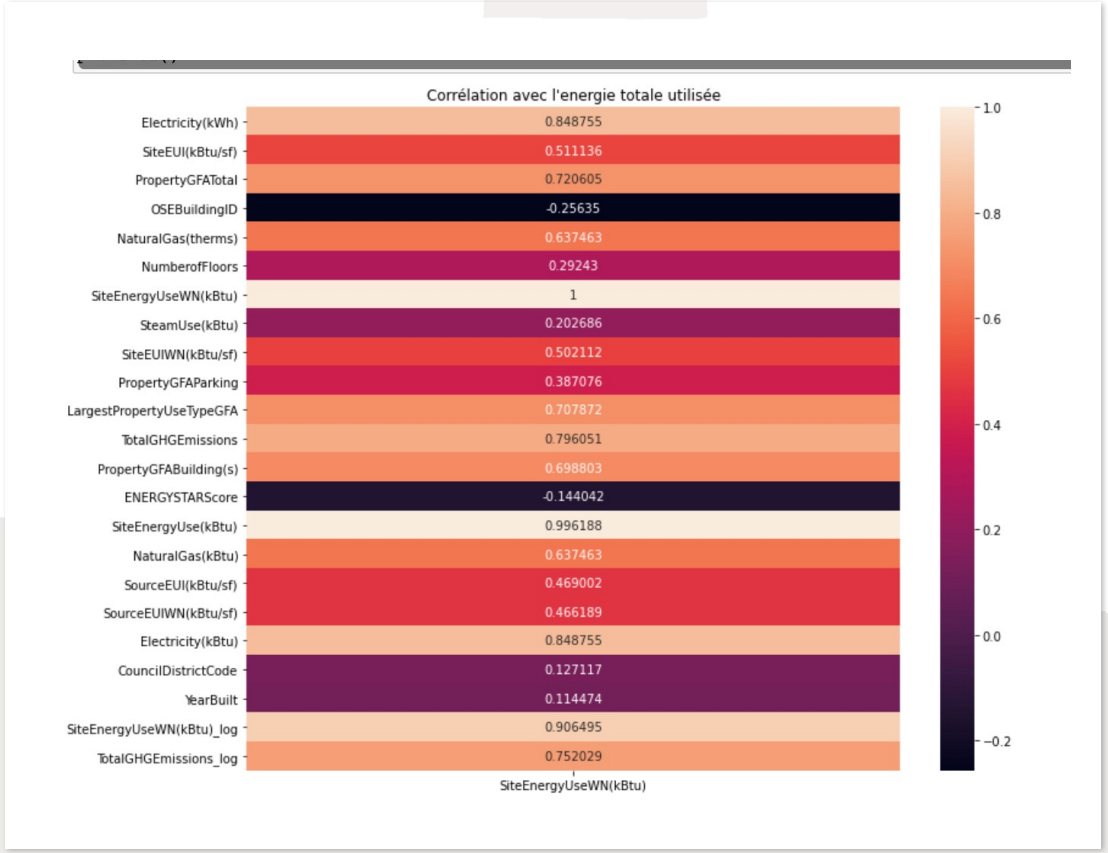






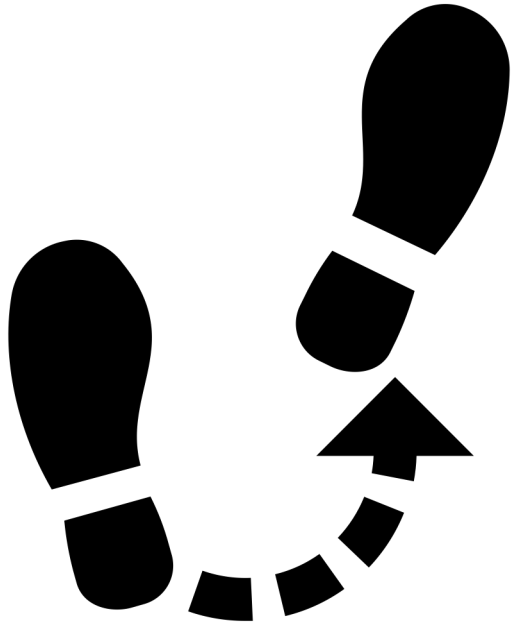
ANALYSE EXPLORATOIRE





ETUDE DES CORRELATIONS

PISTES DE MODELISATION



RÉGRESSION LINEAIRE MULTIPLE

Première approche simpliste

ECHEC

Entrée [19]:
```
MSE_rm = mean_squared_error(y_test, predictions_LM, squared=False)
MSE_rm
```

Out[19]: 49339056797215.13

Entrée [20]:
```
MAE_rm =  mean_absolute_error(y_test, predictions_LM)
MAE_rm
```

Out[20]: 4994412404551.244

Entrée [21]:
```
r2_rm =r2_score(y_test, predictions_LM)
r2_rm
```

Out[21]: -2.3568315084936527e+27

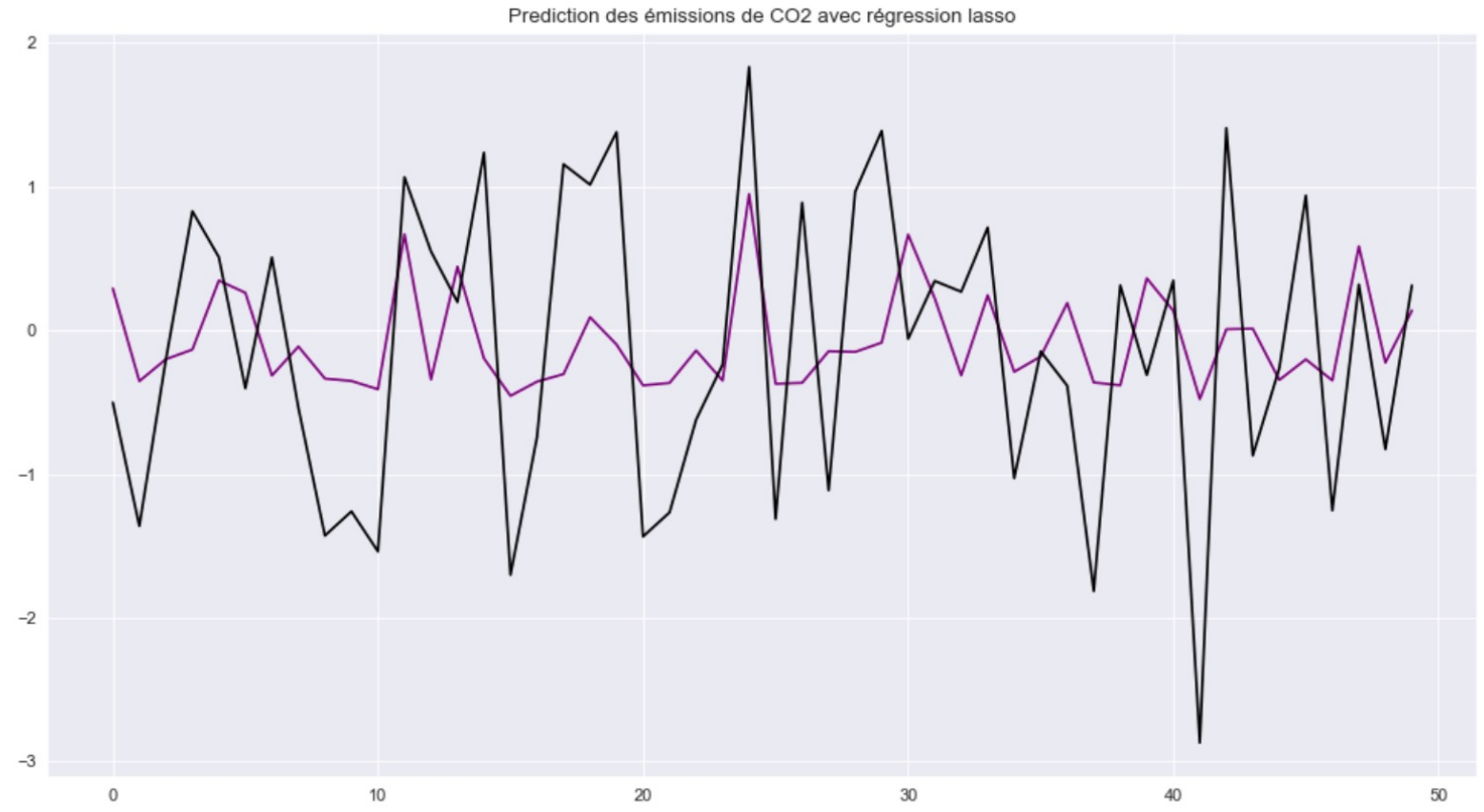
REGRESSION LASSO

L'idée est de restreindre le nombre de variables

]:

	Nom_colonne	coefficient	valeur_absolue
1	SourceEUIWN(kBtu/sf)	0.261183	0.261183
7	PropertyGFATotal	0.138897	0.138897
36	PrimaryPropertyType_K-12 School	0.126219	0.126219
6	LargestPropertyUseTypeGFA	0.120316	0.120316
38	PrimaryPropertyType_Low-Rise Multifamily	-0.058137	0.058137
21	Neighborhood_EAST	0.046756	0.046756
50	PrimaryPropertyType_Senior Care Community	0.029888	0.029888
25	Neighborhood_NORTH	-0.007604	0.007604
3	OSEBuildingID	-0.002253	0.002253
173	ListOfAllPropertyUseTypes_Senior Care Community	0.002108	0.002108
65	LargestPropertyUseType_K-12 School	0.000746	0.000746
0	PropertyGFAParking	-0.000000	0.000000

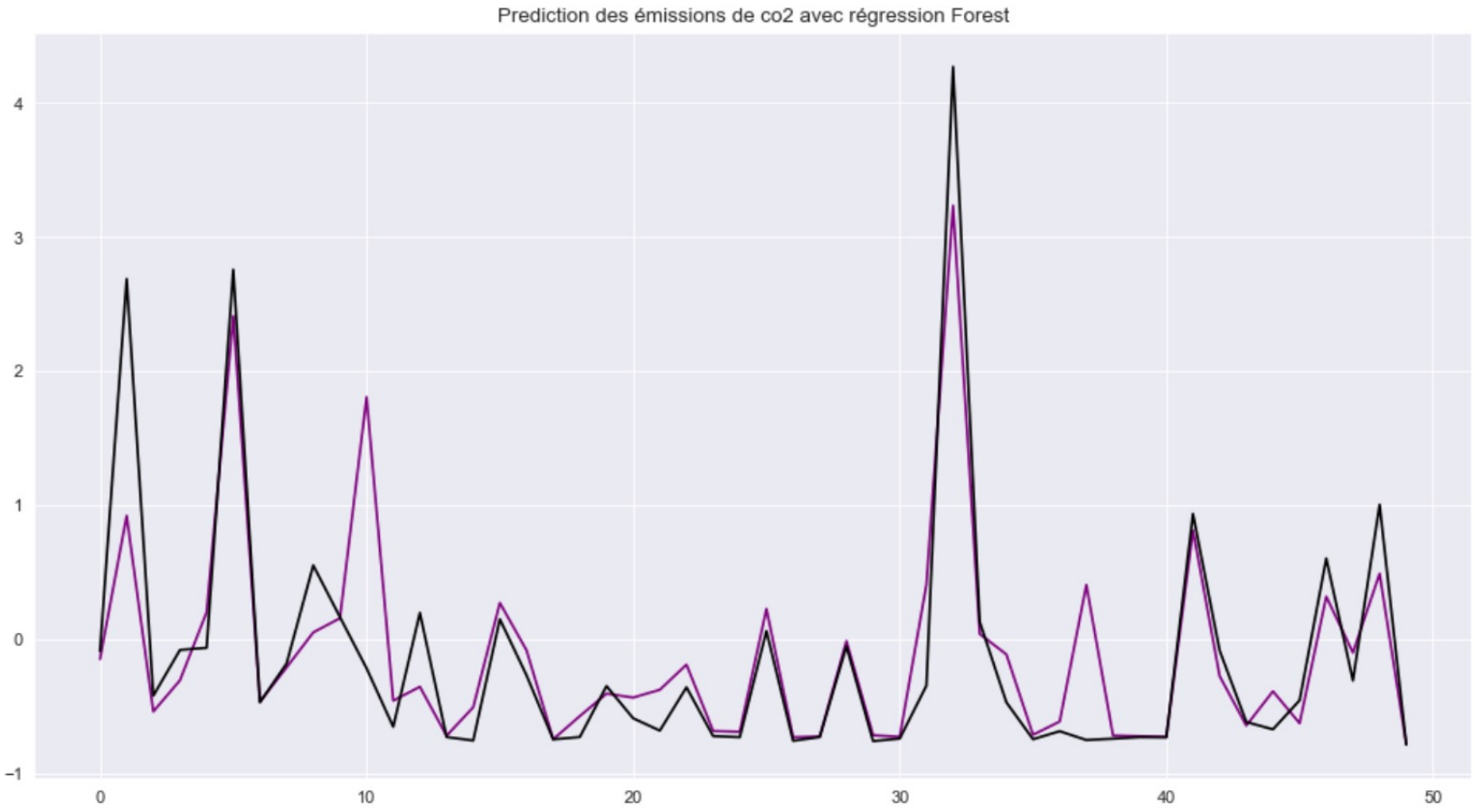
REGRESSION LASSO



En violet les predictions, en noir la réalité.

R2 =0.4

REGRESSION RANDOM FOREST



R2=0.78

TUNAGE

Entrée [43]:
```
start = time.time()
param_grid = {'n_estimators':[120,130,140],'min_samples_split':[2,3],'min_samples_leaf':[1,2],'max_depth': [10, 20,30
score = 'r2'
```

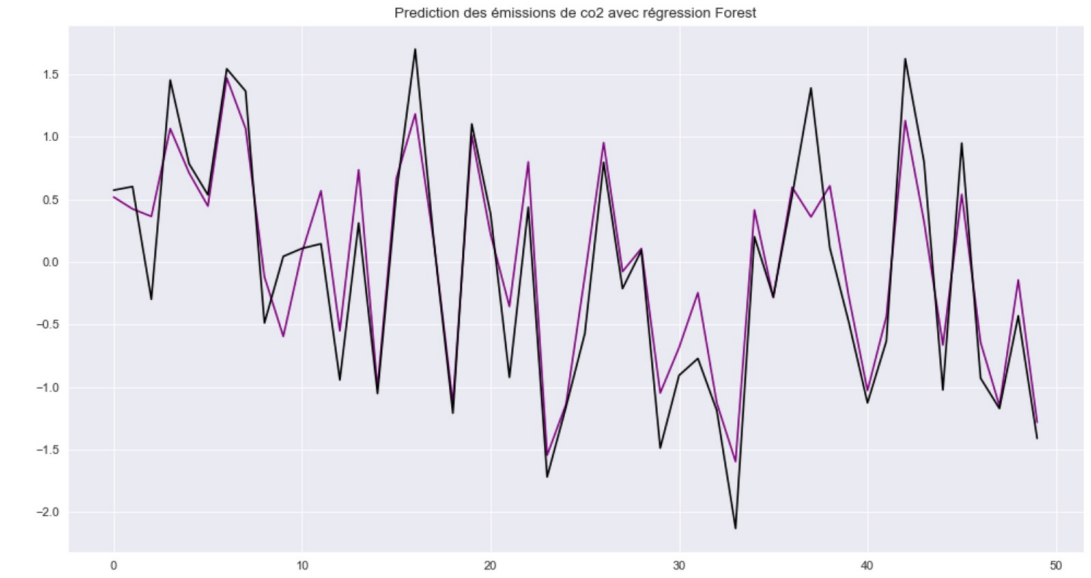
Entrée [44]:
```
clf = model_selection.GridSearchCV(
    RandomForestRegressor(), # un classifieur kNN
    param_grid,     # hyperparamètres à tester
    cv=5,           # nombre de folds de validation croisée
    scoring=score   # score à optimiser
)
clf.fit(X_train,y_train.values.ravel())
```

Out[44]:
```
GridSearchCV(cv=5, estimator=RandomForestRegressor(),
             param_grid={'max_depth': [10, 20, 30], 'min_samples_leaf': [1, 2],
                         'min_samples_split': [2, 3],
                         'n_estimators': [120, 130, 140]},
             scoring='r2')
```

Entrée [45]:
```
clf.best_params_
```

Out[45]:
```
{'max_depth': 30,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'n_estimators': 120}
```

RESULTATS SATISFAISANTS



En violet les predictions, en noir la réalité.

Entrée [48]:
```
MSE_forest = mean_squared_error(y_test,predictions_FOREST, squared=False)
MSE_forest
```

Out[48]: 0.36829249880491133

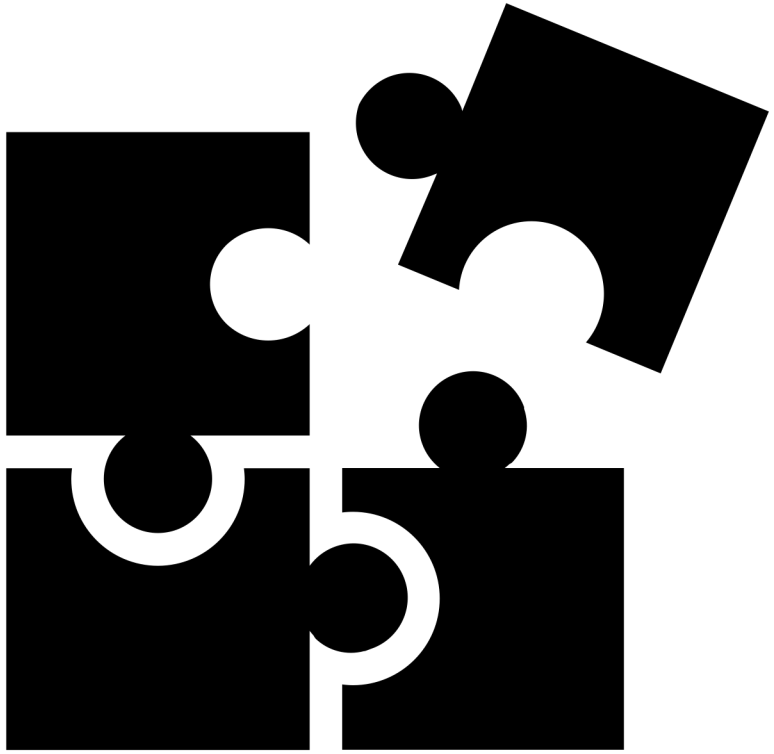
Entrée [49]:
```
MAE_forest =  mean_absolute_error(y_test, predictions_FOREST)
MAE_forest
```

Out[49]: 0.27153246052785823

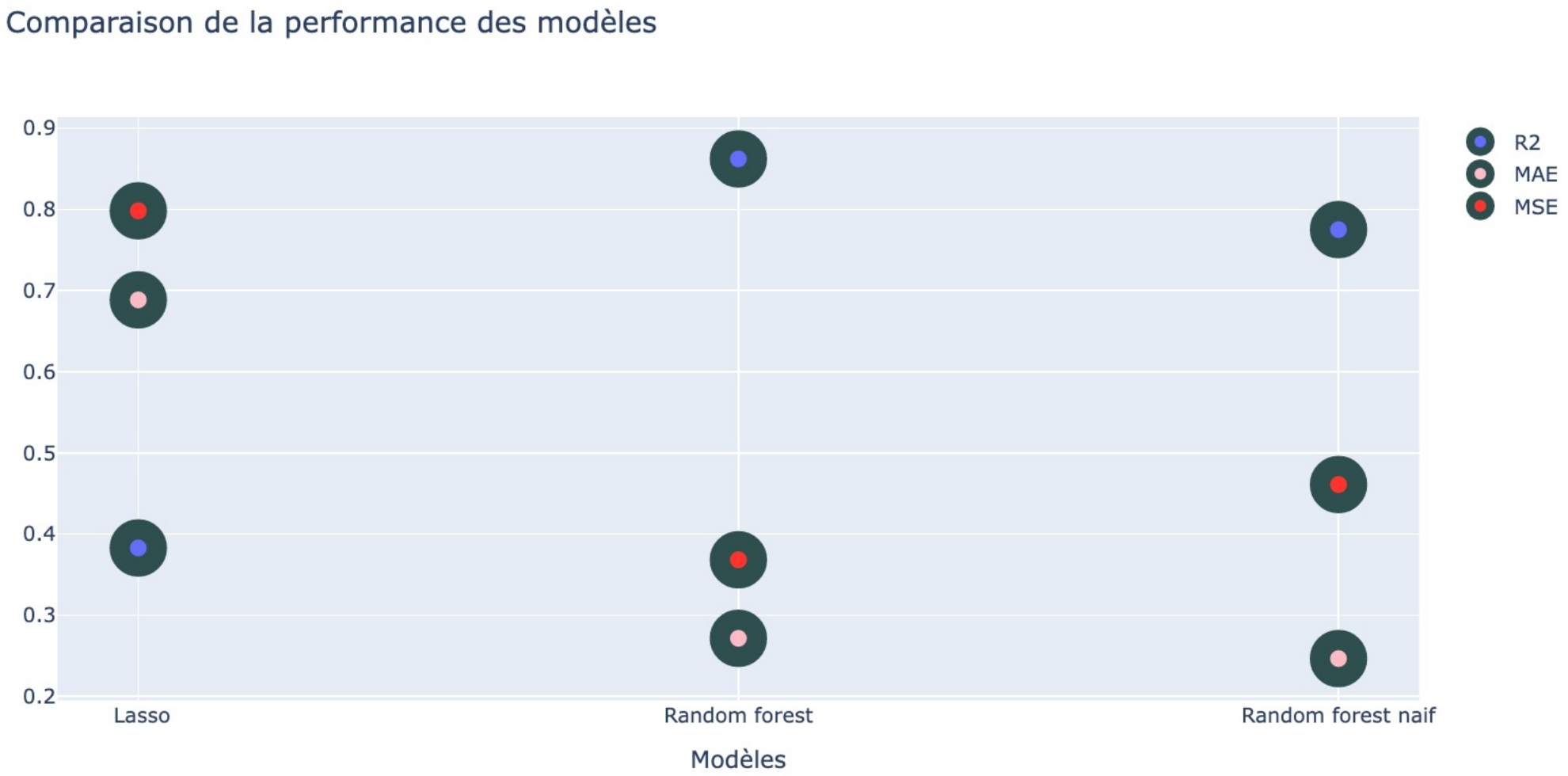
Entrée [50]:
```
r2_forest =r2_score(y_test, predictions_FOREST)
r2_forest
```

Out[50]: 0.8623488619299098

MODELE FINALE



COMPARAISON DES MODELES (CO2) : R2,MAE,MSE



COMPARAISON DES MODELES (PRODUCTION ENERGIE) : R2,MAE,MSE

COMPARAISON DES MODELES (ENERGIE TOTALE) : TEMPS

REGRESSION RANDOM FOREST TUNÉ

Meilleures performances statistiques

Temps raisonnable

PRISE EN COMPTE ENERGYSTARSCORE POUR LE CO2 (R2)

Type	*Energystar Score*	*Sans Energystar Score*
> 3 étages	0.810	0.817
< 3 étages	0.862	0.857

Non significatif

MERCI