Seattle

OPENCLASSROOMS

Anticipez les besoins en consommation de bâtiments

Parcours Data Scientist| projet 4

Rim BAHROUN

Janvier 2023

Anticipez les besoins en consommation de bâtiments

- **Problématique**

Seattle: objectif de ville neutre en émissions de carbone en 2050.

relevés sont coûteux à obtenir

"ENERGY STAR Score" fastidieux à calculer

- **Mission**

prédire les émissions de CO2 et la consommation totale d'énergie de bâtiments non destinés à l'habitation.

évaluer l'intérêt de l'"ENERGY STAR Score" pour la prédiction des émissions.

Anticipez les besoins en consommation de bâtiments

- **Partie 1: Nettoyage et exploration des données**
  - Nettoyage des données
  - Exploration des données
- **Partie 2: Modélisation et présentation des résultats**
  - Prétraitement des données
  - Modélisation de la consommation énergétique
  - Modélisation des émissions de CO2
  - Intérêt de l'ENERGYSTARScore
- **Conclusion**

1. Nettoyage et exploration des données

1. Nettoyage et analyse de la forme des données

Seattle.

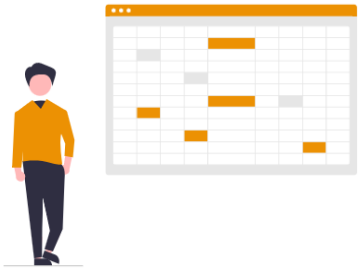
Données à disposition : **1** fichiers **.csv**

3376 lignes, 46 colonnes

Il donne les valeurs des caractéristiques de chaque bâtiment de la ville de Seattle.

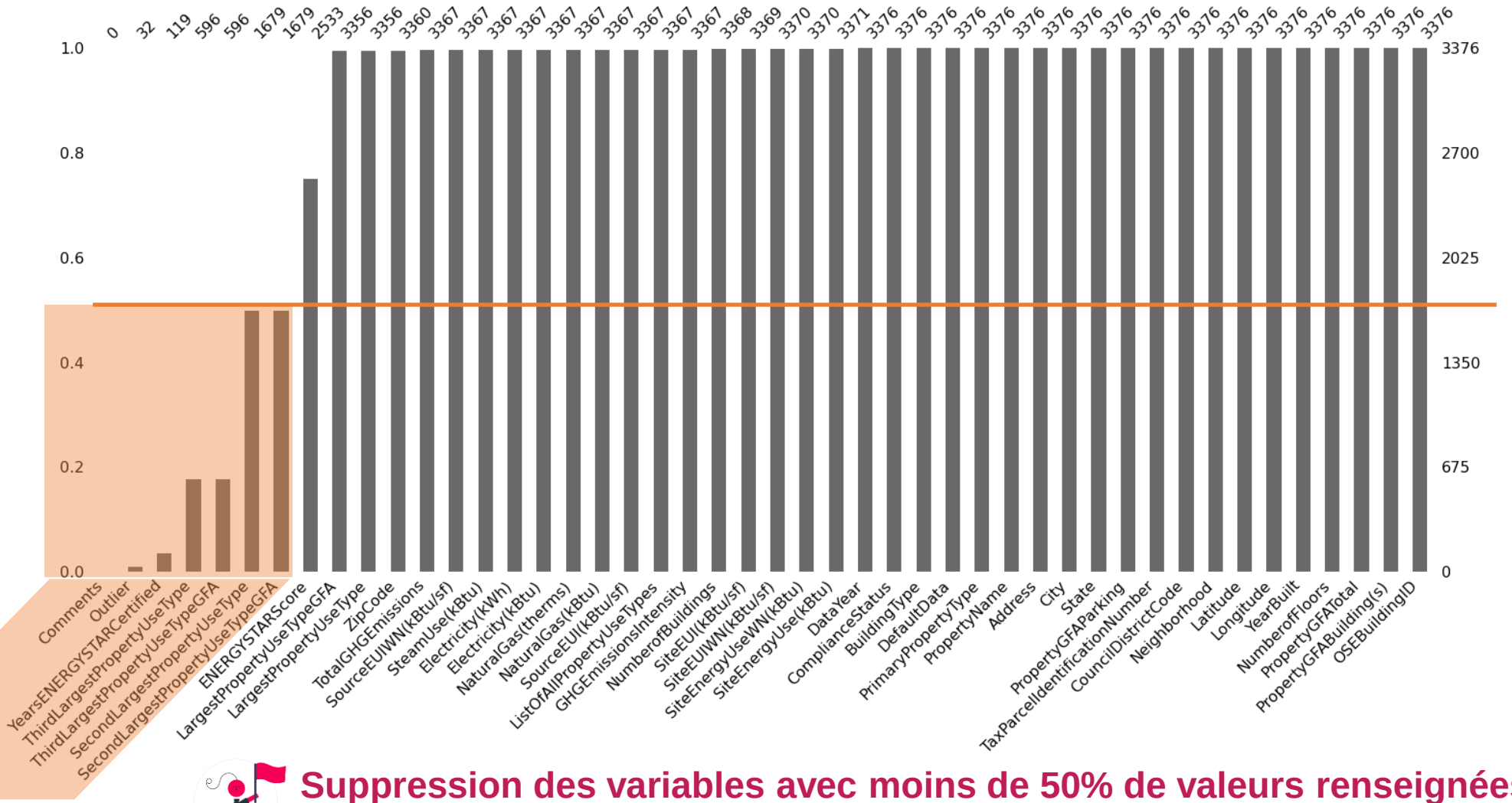
30 variables numériques et 16 catégorielles.

	OSEBuildingID	DataYear	BuildingType	PrimaryPropertyType	PropertyName	Address	City	State	ZipCode
0	1	2016	NonResidential	Hotel	Mayflower park hotel	405 Olive way	Seattle	WA	98101.0
1	2	2016	NonResidential	Hotel	Paramount Hotel	724 Pine street	Seattle	WA	98101.0
2	3	2016	NonResidential	Hotel	5673-The Westin Seattle	1900 5th Avenue	Seattle	WA	98101.0
3	5	2016	NonResidential	Hotel	HOTEL MAX	620 STEWART ST	Seattle	WA	98101.0
4	8	2016	NonResidential	Hotel	WARWICK SEATTLE HOTEL (ID8)	401 LENORA ST	Seattle	WA	98121.0



1. Nettoyage et exploration des données

1. Nettoyage et analyse de la forme des données



Suppression des variables avec moins de 50% de valeurs renseignées.

On passe de **46** variables à **39** variables.

1. Nettoyage et exploration des données

1. Nettoyage et analyse de la forme des données

Identification et localisation d'un bâtiment

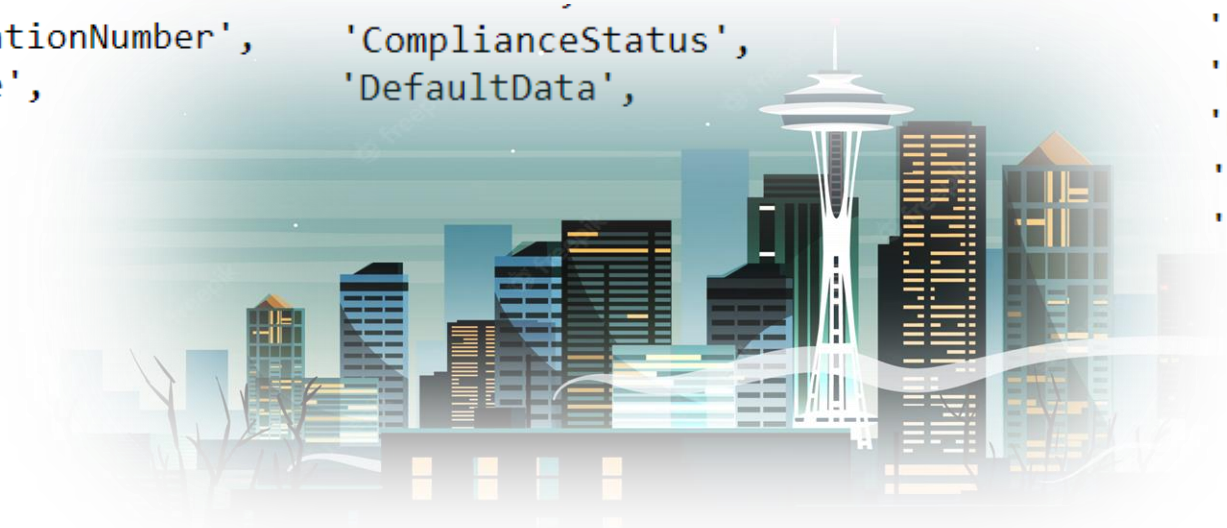
```
'OSEBuildingID',
'DataYear',
'BuildingType',

'PropertyName',
'Address',
'City',
'State',
'ZipCode',
'TaxParcelIdentificationNumber',
'CouncilDistrictCode',
'Neighborhood',
'Latitude',
'Longitude',
'YearBuilt',
'NumberofBuildings',
'NumberofFloors',
```

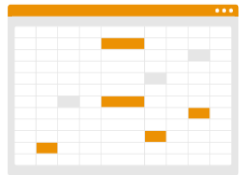
Surface (GFA) et utilisation (type) d'un bâtiment

```
'PropertyGFATotal',
'PropertyGFAParking',
'PropertyGFABuilding(s)',
'LargestPropertyUseTypeGFA',

'PrimaryPropertyType',
'ListOfAllPropertyUseTypes',
'LargestPropertyUseType',

'ComplianceStatus',
'DefaultData',
```

Les relevés de consommation et d'émission d'un bâtiment

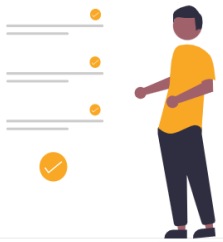
```
'ENERGYSTARScore',
'SiteEUI(kBtu/sf)',
'SiteEUIWN(kBtu/sf)',
'SourceEUI(kBtu/sf)',
'SourceEUIWN(kBtu/sf)',
'SiteEnergyUse(kBtu)',
'SiteEnergyUseWN(kBtu)',
'SteamUse(kBtu)',
'Electricity(kWh)',
'Electricity(kBtu)',
'NaturalGas(therms)',
'NaturalGas(kBtu)',
'TotalGHGEmissions',
'GHGEmissionsIntensity'
```





1. Nettoyage et exploration des données

1. Nettoyage des données: élimination des variables inutiles



Identification et localisation d'un bâtiment

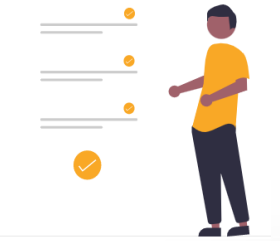
Variables pertinentes

~~'OSEBuildingID',~~	Passé en index
~~'DataYear',~~	'2016' pour tout le jeu de donnée
~~'BuildingType',~~	Sélection des bâtiments Non résidentiel (nettoyage sur les individus)
~~'PropertyName',~~	Localisation
~~'Address',~~	
~~'City',~~	'Seattle' pour tout le jeu de donnée
~~'State',~~	'WA' pour tout le jeu de donnée
~~'ZipCode',~~	
~~'TaxParcelIdentificationNumber',~~	
'CouncilDistrictCode',	**Le quartier où se situe le bâtiment**
~~'Neighborhood',~~	
~~'Latitude',~~	
~~'Longitude',~~	
'YearBuilt',	**Date de construction du bâtiment**
~~'NumberofBuildings',~~	1 pour 97% du jeu de donnée
'NumberofFloors',	**Nombre d'étages dans le bâtiment**

1. Nettoyage et exploration des données

1. Nettoyage des données: élimination des variables inutiles

Surface (GFA) et utilisation (type) d'un bâtiment



Variables pertinentes

```
'PropertyGFATotal',
'PropertyGFAParking',
'PropertyGFABuilding(s)',
'LargestPropertyUseTypeGFA',

'PrimaryPropertyType',
'ListOfAllPropertyUseTypes',
'LargestPropertyUseType',

'ComplianceStatus',
'DefaultData',
```

Surface totale du bâtiment

Très corrélée avec la surface du bâtiment (0,94)

Utilisation du bâtiment

Sélection des bâtiments avec des données conformes (nettoyage sur les individus)

'False' pour tous les bâtiments sélectionnés

1. Nettoyage et exploration des données

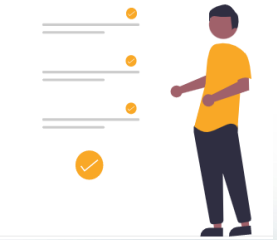
1. Nettoyage des données: élimination des variables inutiles

Les relevés de consommation et d'émission d'un bâtiment

Variables pertinentes

```
'ENERGYSTARScore',
'SiteEUI(kBtu/sf)',
'SiteEUIWN(kBtu/sf)',
'SourceEUI(kBtu/sf)',
'SourceEUIWN(kBtu/sf)',
'SiteEnergyUse(kBtu)',
'SiteEnergyUseWN(kBtu)',
'SteamUse(kBtu)',
'Electricity(kWh)',
'Electricity(kBtu)',
'NaturalGas(therms)',
'NaturalGas(kBtu)',
'TotalGHGEmissions',
'GHGEmissionsIntensity'
```

Variable à étudier son effet sur la prédiction



Variables mesurables à partir de relevés couteux

Ne peuvent pas être utilisé pour ce projet

Variable à prédire

- **Steam_use (0/1)**
- **Electricity_use (0/1)**
- **NaturalGas_use (0/1)**

Variable à prédire

1. Nettoyage et exploration des données

1. Nettoyage des données: élimination des variables inutiles

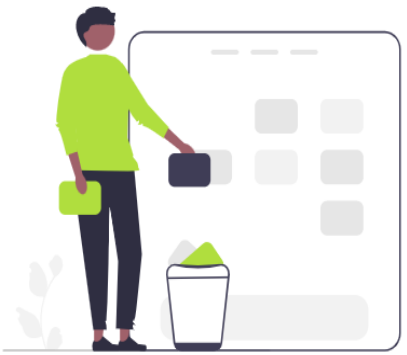
3376 lignes, 46 colonnes → **Nettoyage des données** → 1513 lignes, 11 colonnes

Variables quantitatives
- YearBuilt
- NumberofFloors
- PropertyGFATotal

Variables qualitatives
- CouncilDistrictCode
- PrimaryPropertyType
- Steam_use (0/1)
- Electricity_use (0/1)
- NaturalGas_use (0/1)

Variable quantitative à étudier
- ENERGYSTARScore

Variables quantitatives cibles (targets)
- SiteEnergyUse
- TotalGHGEmissions



Anticipez les besoins en consommation de bâtiments

- **Partie 1: Nettoyage et exploration des données**
  - Nettoyage des données
  - Exploration des données
- **Partie 2: Modélisation et présentation des résultats**
  - Prétraitement des données
  - Modélisation de la consommation énergétique
  - Modélisation des émissions de CO2
  - Intérêt de l'ENERGYSTARScore
- **Conclusion**

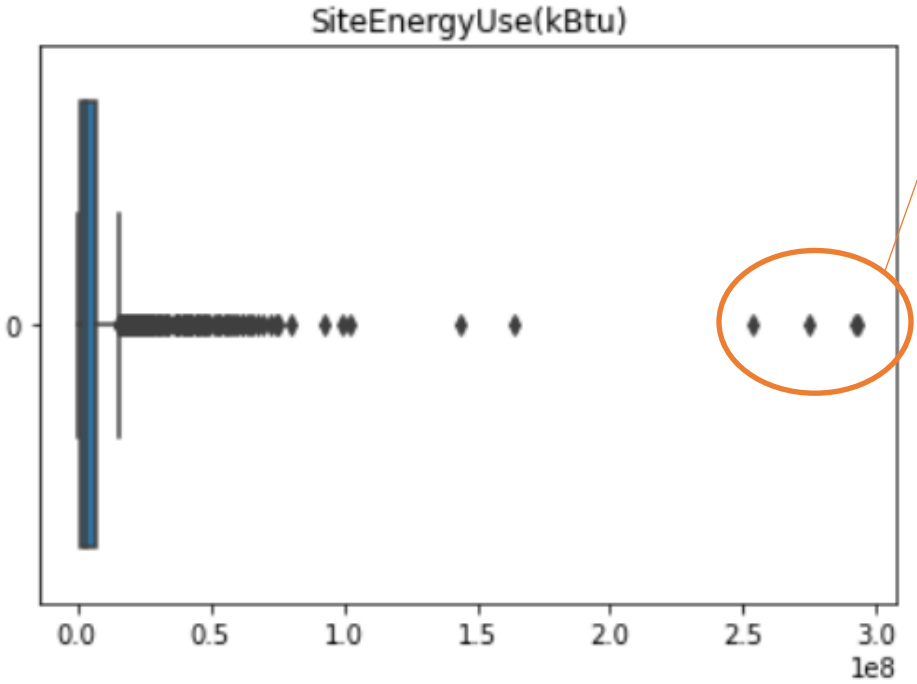
1. Nettoyage et exploration des données

2. Exploration des données

Analyse univariée: variables cibles (targets)

SiteEnergyUse

	SiteEnergyUse(kBtu)
count	1.514000e+03
mean	7.726794e+06
std	1.879214e+07
min	5.713320e+04
25%	1.233141e+06
50%	2.665809e+06
75%	7.064835e+06
max	2.930908e+08



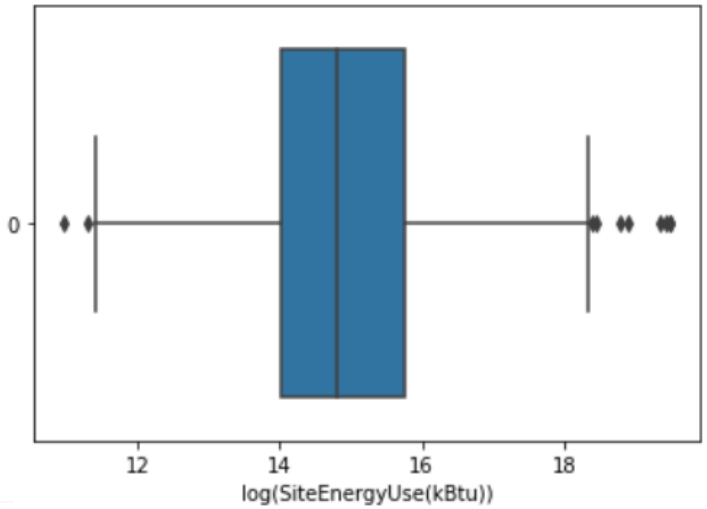
Principalement des hôpitaux

1. Nettoyage et exploration des données

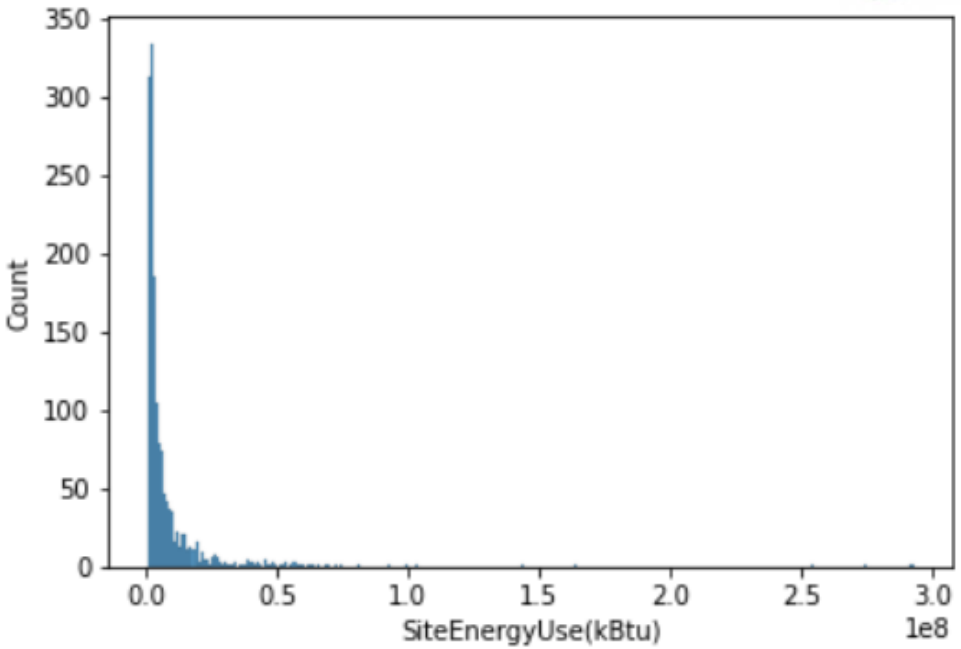
2. Exploration des données

Analyse univariée: variables cibles (targets)

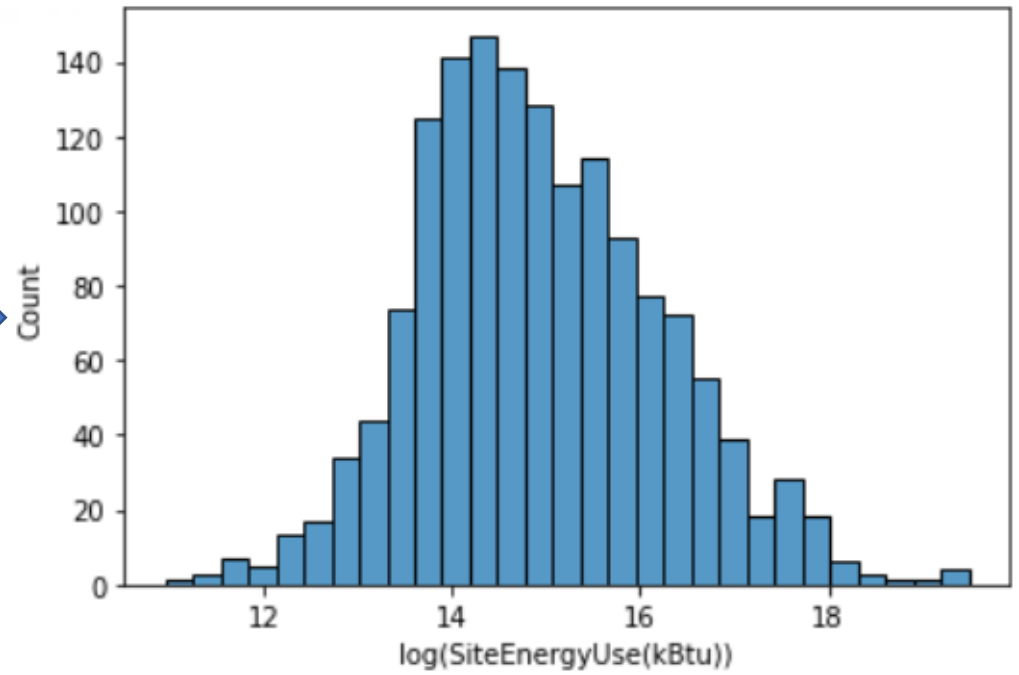
SiteEnergyUse



count	1513.000000
mean	14.928041
std	1.288035
min	10.953158
25%	14.025850
50%	14.797426
75%	15.771512
max	19.495993



?



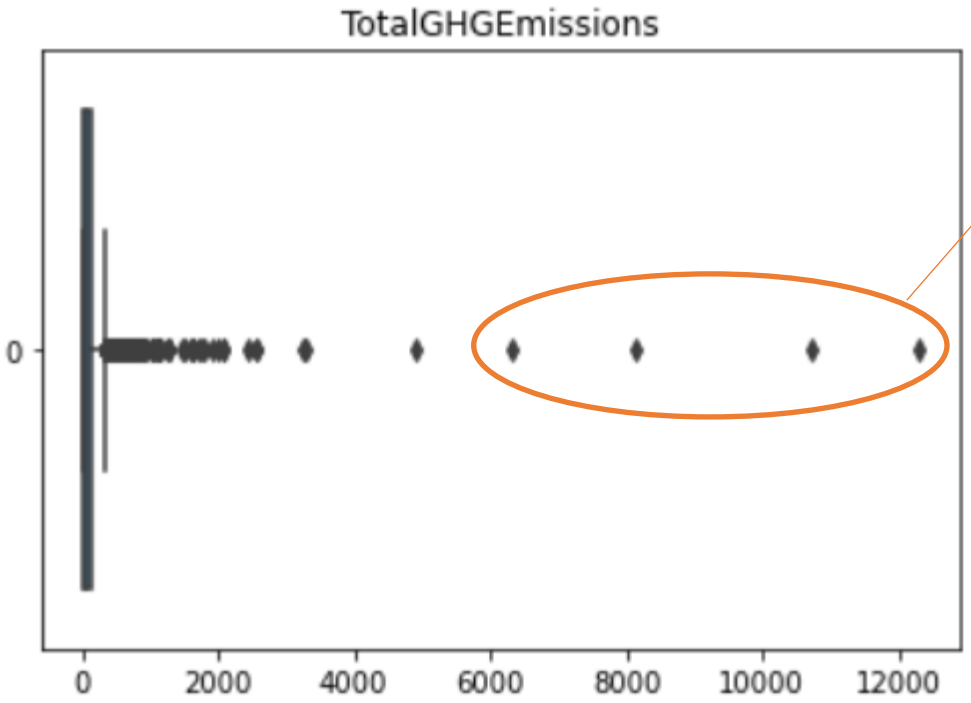
1. Nettoyage et exploration des données

2. Exploration des données

Analyse univariée: variables cibles (targets)

TotalGHGEmissions

	TotalGHGEmissions
count	1514.000000
mean	167.408151
std	575.982871
min	-0.800000
25%	20.092500
50%	49.215000
75%	140.830000
max	12307.160000



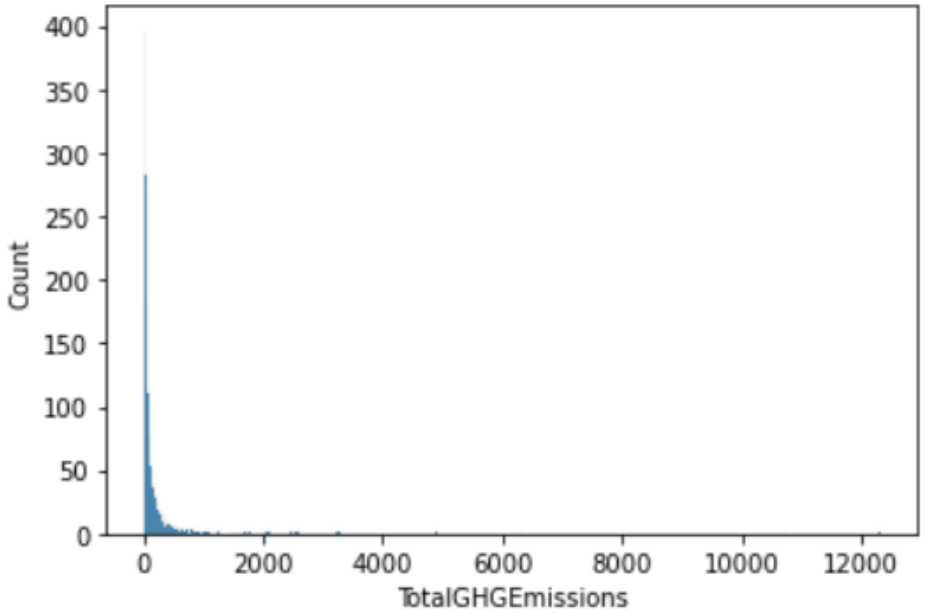
Principalement des hôpitaux

1. Nettoyage et exploration des données

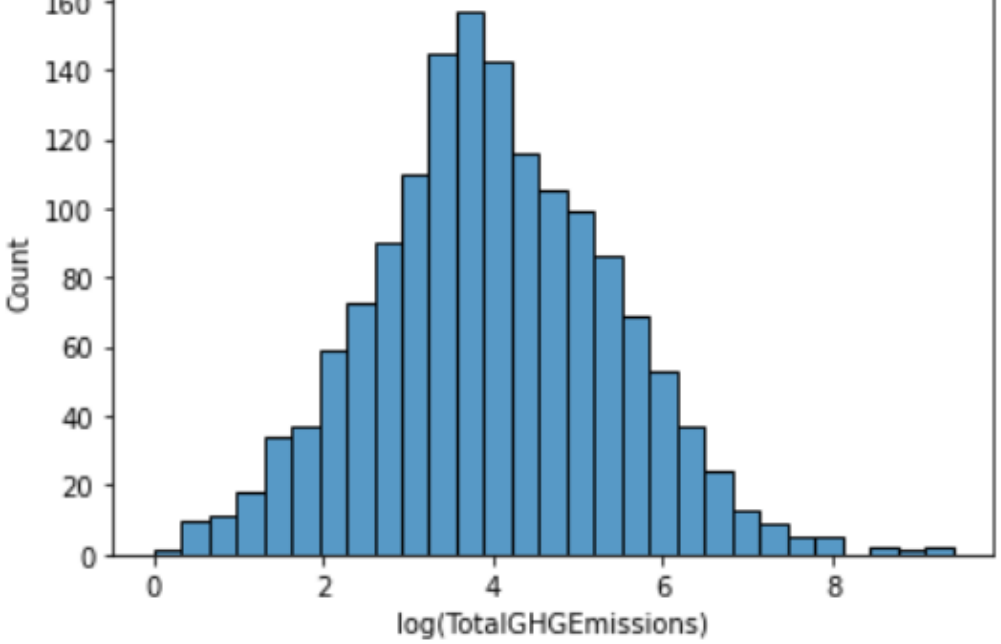
2. Exploration des données

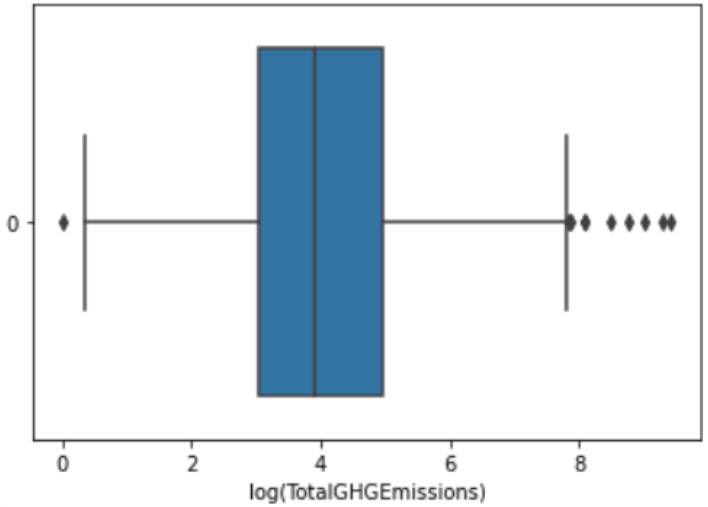
Analyse univariée: variables cibles (targets)

TotalGHGEmissions



?





count	1513.000000
mean	3.993708
std	1.426829
min	0.000000
25%	3.049273
50%	3.917011
75%	4.957235
max	9.418018

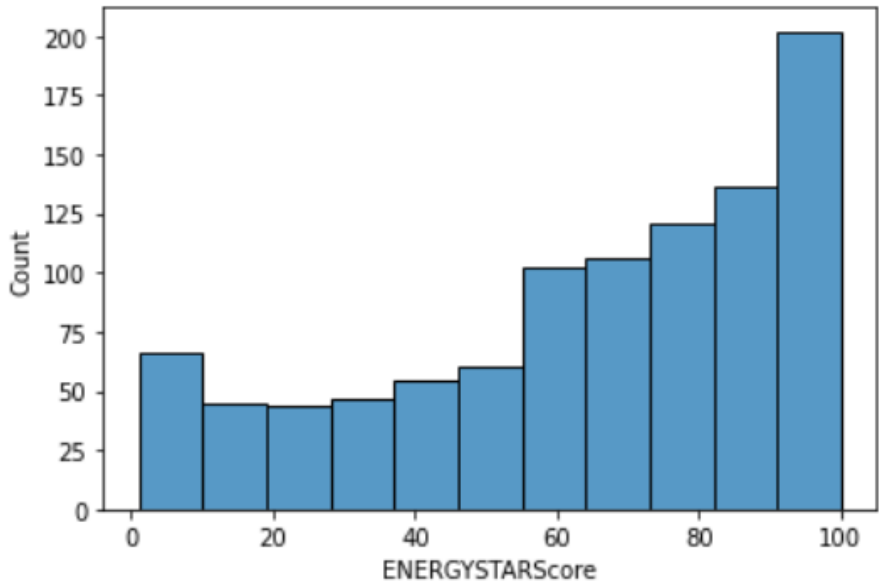
1. Nettoyage et exploration des données

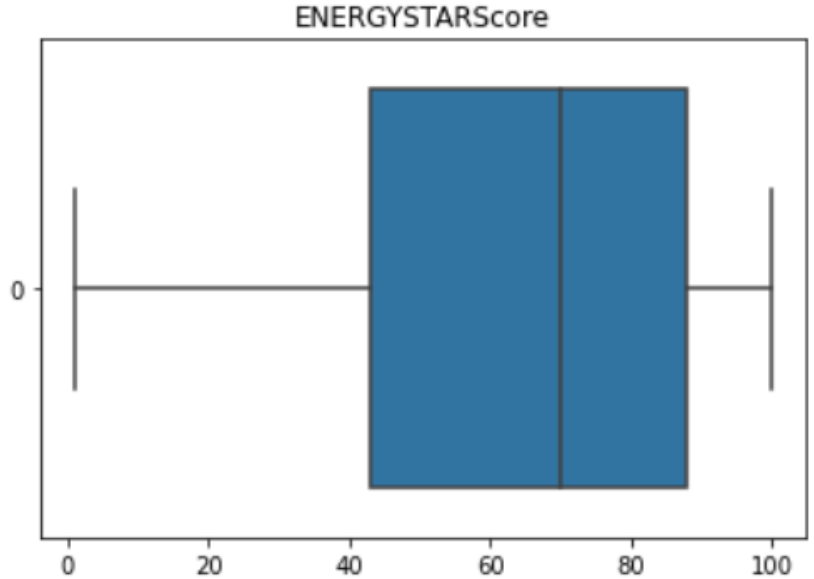
2. Exploration des données

Analyse univariée: ENERGYSTARScore

Il est exprimé sur une échelle de 1 à 100 facile à comprendre, où plus le score est élevé, meilleure est la performance énergétique du bâtiment.

count	983.000000
mean	63.390641
std	28.802615
min	1.000000
25%	43.000000
50%	70.000000
75%	88.000000
max	100.000000

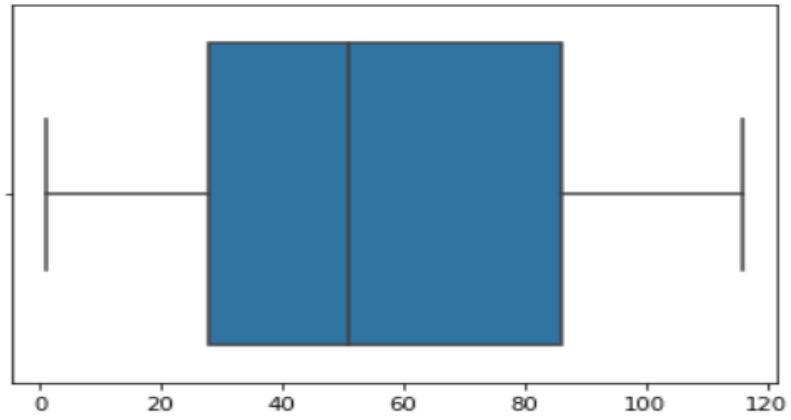


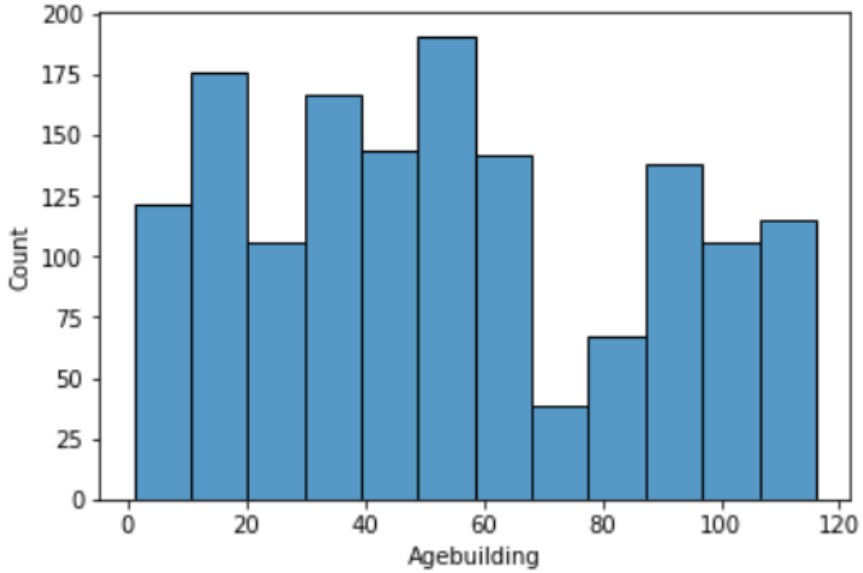


1. Nettoyage et exploration des données

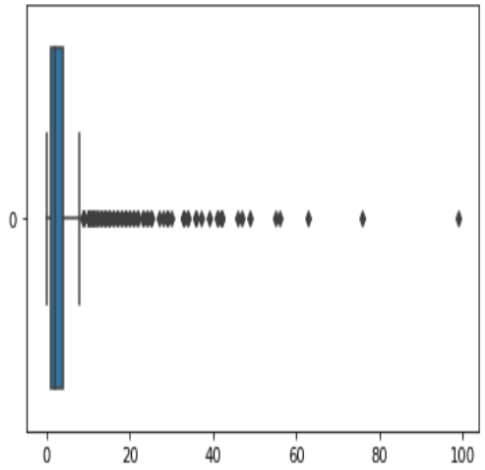
2. Exploration des données: Analyse univariée

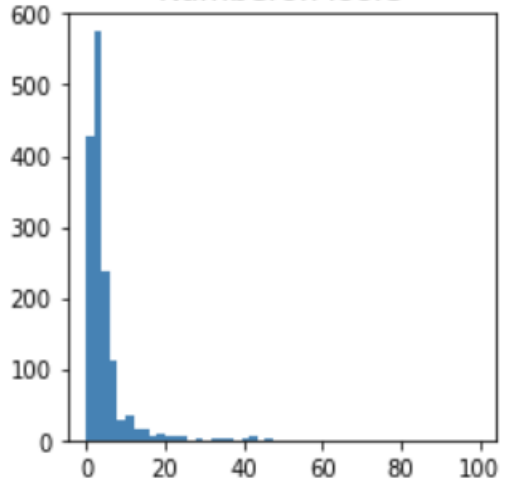
YearBuilt ➡ Agebuilding





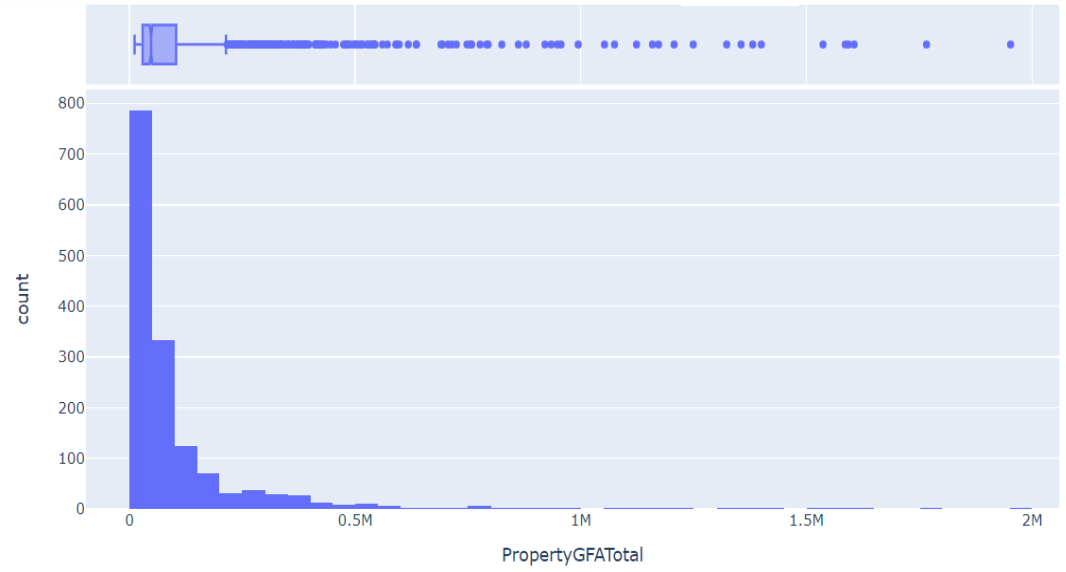
NumberofFloors





PropertyGFATotal

	PropertyGFATotal
count	1.513000e+03
mean	1.122782e+05
std	1.904159e+05
min	1.128500e+04
25%	2.880000e+04
50%	4.768600e+04
75%	1.035660e+05
max	1.952220e+06



1. Nettoyage et exploration des données

2. Exploration des données

Analyse univariée: CouncilDistrictCode (arrondissement)

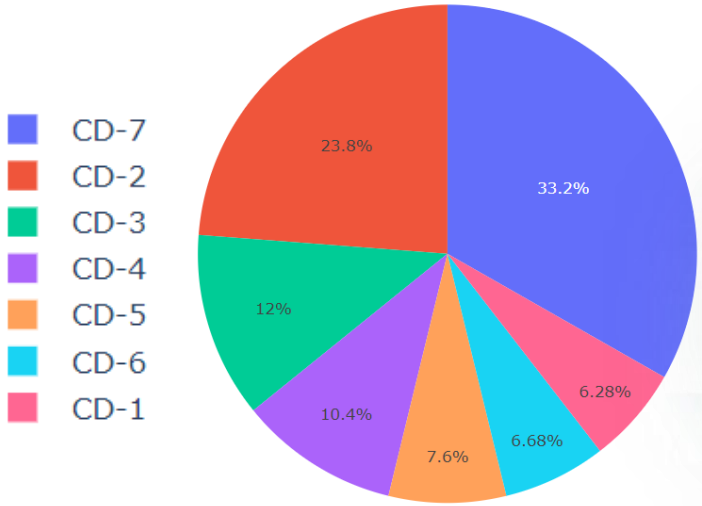
```
data['CouncilDistrictCode'].unique()
```

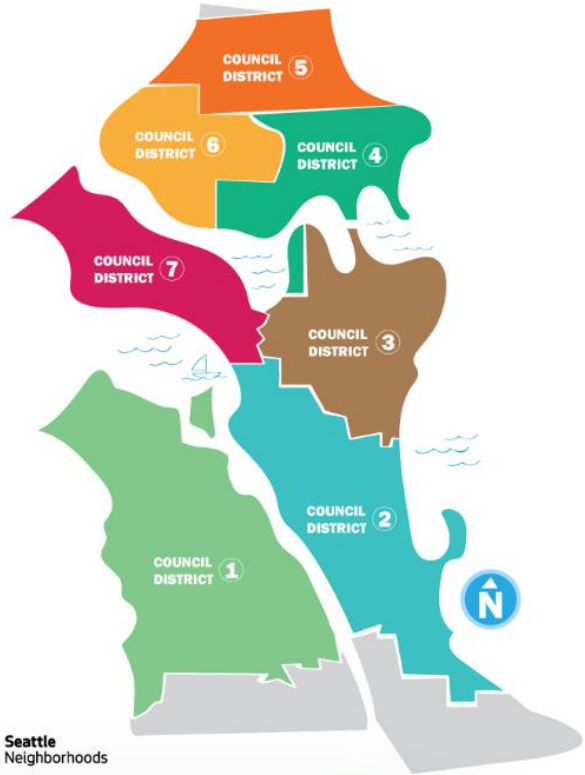
```
array([7, 3, 4, 2, 6, 1, 5], dtype=int64)
```

```
data['CouncilDistrictCode'] = data['CouncilDistrictCode'].apply(lambda x : 'CD-' + str(x))
```

```
data['CouncilDistrictCode'].unique()
```

```
array(['CD-7', 'CD-3', 'CD-4', 'CD-2', 'CD-6', 'CD-1', 'CD-5'],
      dtype=object)
```



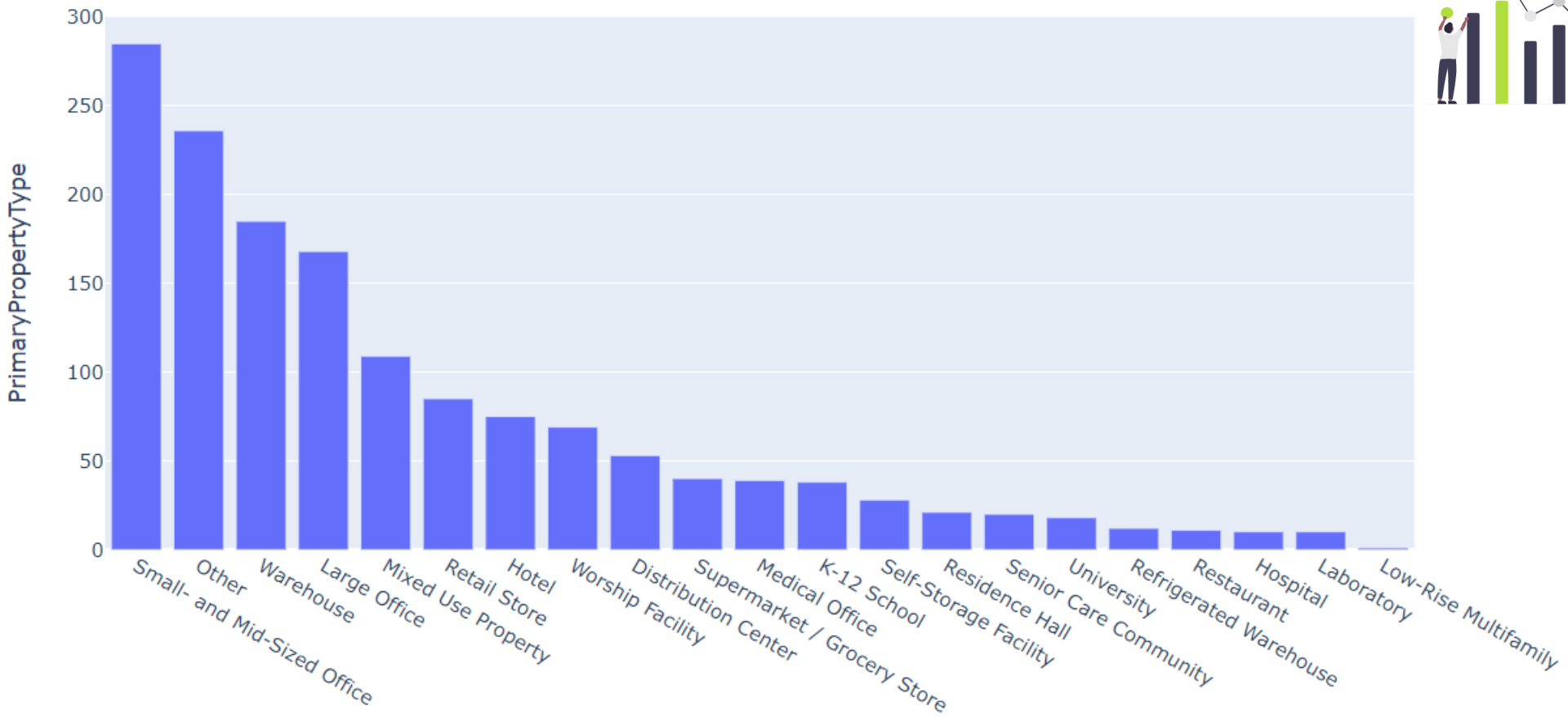


1. Nettoyage et exploration des données

2. Exploration des données

Analyse univariée: PrimaryPropertyType

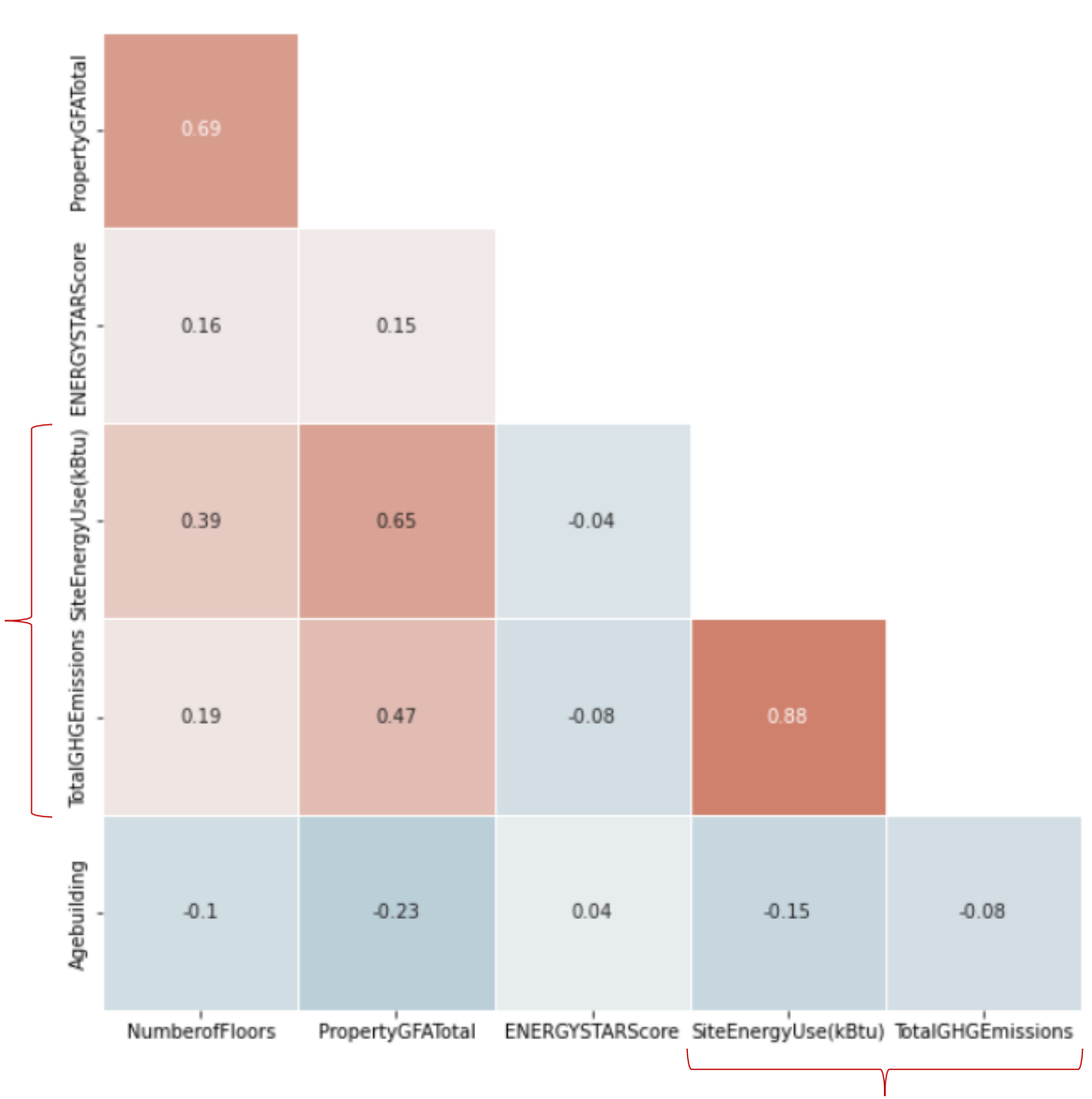
	PrimaryPropertyType	ListOfAllPropertyUseTypes	LargestPropertyUseType
count	1513	1513	1509
unique	21	361	55



1. Nettoyage et exploration des données

2. Exploration des données

Analyse bivariée: matrice de corrélation



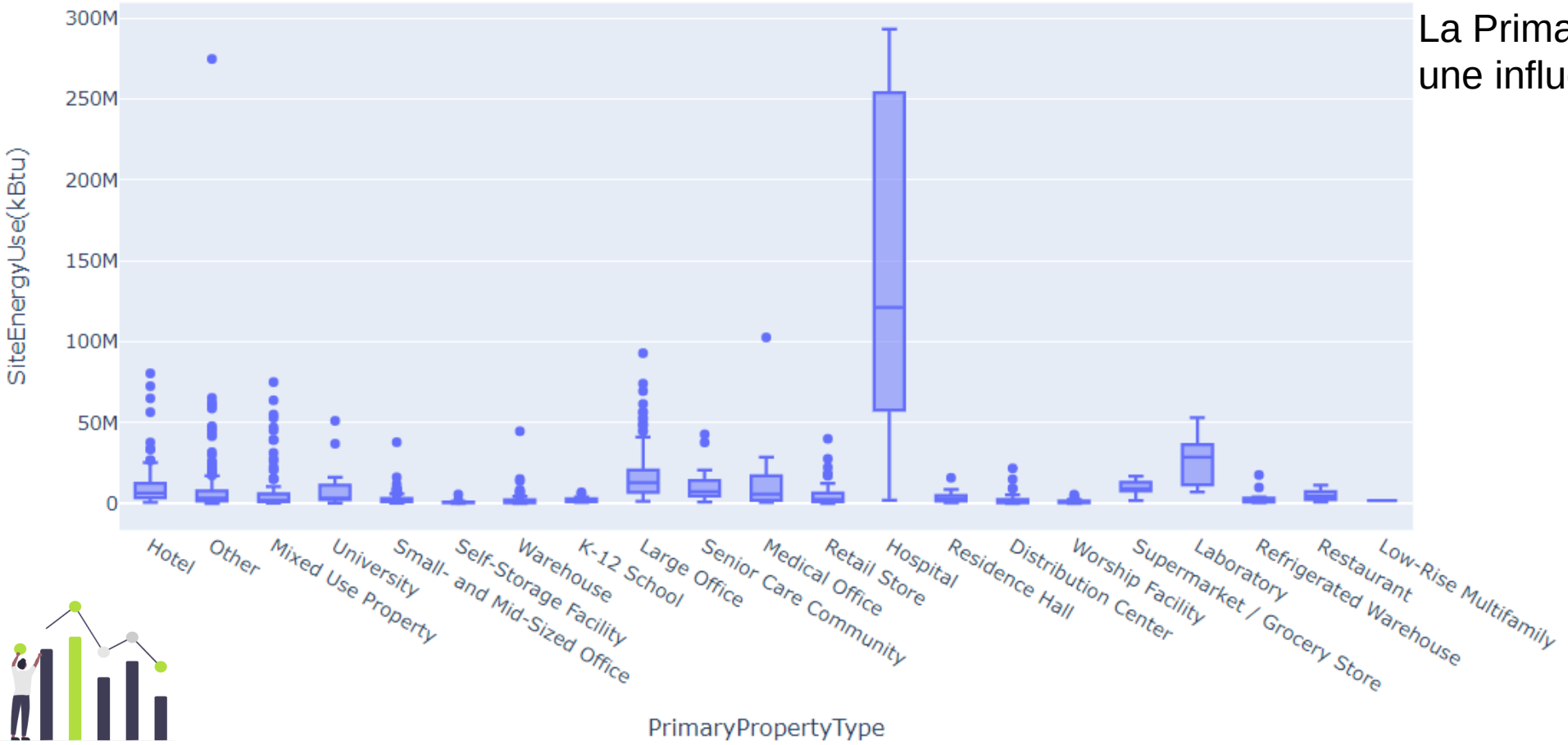
1. Nettoyage et exploration des données

2. Exploration des données

Analyse bivariée : ANOVA

	sum_sq	df	F	PR(>F)
PrimaryPropertyType	2.223177e+17	20.0	53.167757	3.752359e-158
Residual	3.119352e+17	1492.0	NaN	NaN

PrimaryPropertyType



Les résultats du test de Fisher indiquent une p-value inferieur au niveau de test de 5%. Nous rejetons donc l'hypothèse H0 selon laquelle les distributions sont identiques.
La PrimaryPropertyType a donc bien une influence sur la consommation.

1. Nettoyage et exploration des données

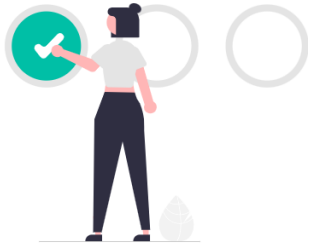
Synthèse

Dans cette première partie, nous avons effectué:

- un repérage des variables cibles (targets)
- un nettoyage sur les variables et sur les individus
- des **analyses univariées** pour des variables **quantitatives** et **qualitatives**
- des **analyses bivariées :**
  - la **corrélation** entre les variables quantitatives pertinentes
  - l'**ANOVA** entre les variables cibles et les variables qualitatives pertinentes

Ces opérations ont permis de préparer le jeu de donnée pour l'étape de modélisation

#	Column	Non-Null Count	Dtype
0	PrimaryPropertyType	1513 non-null	object
1	CouncilDistrictCode	1513 non-null	object
2	NumberofFloors	1513 non-null	int64
3	PropertyGFATotal	1513 non-null	int64
4	ENERGYSTARScore	983 non-null	float64
5	SiteEnergyUse(kBtu)	1513 non-null	float64
6	TotalGHGEmissions	1513 non-null	float64
7	Agebuilding	1513 non-null	int64
8	NaturalGas_use	1513 non-null	int64
9	Electricity_use	1513 non-null	int64
10	Steam_use	1513 non-null	int64



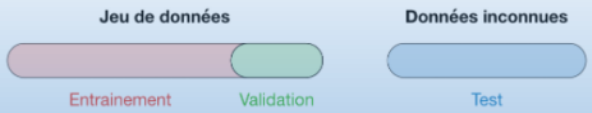
Anticipez les besoins en consommation de bâtiments

- **Partie 1: Nettoyage et exploration des données**
  - Nettoyage des données
  - Exploration des données
- **Partie 2: Modélisation et présentation des résultats**
  - Prétraitement des données
  - Modélisation de la consommation énergétique
  - Modélisation des émissions de CO2
  - Intérêt de l'ENERGYSTARScore
- **Conclusion**

2. Modélisation et présentation des résultats

Jeu de données nettoyé

Prétraitement des données



Encodage des variables catégorielles **OneHotEncoder**

Standardisation des variables numériques **StandardScaler**

y: consommation Totale d'énergie

X: Données de permis de construction

y: émission du co2

Modélisation
- Implémentation des modèles de régression
- Optimisation des hyperparamètres

Evaluations des performances
- Comparaison des modèles
- Choix du modèle final pour la consommation
- Choix du modèle final pour les émissions

Prédiction de la consommation d'énergie

Prédiction de l'émission en CO2

Intérêt de l'ENERGYSTARScore

2. Modélisation et présentation des résultats

1. Prétraitement des données

NaturalGas_use	Electricity_use	Steam_use	PrimaryPropertyType	NumberofFloors	PropertyGFATotal	ENERGYSTARScore	Agebuilding	SiteEnergyUse(kBtu)	TotalGHGEmissions
1	1	1	Hotel	12	88434	60.0	89	7226362.5	249.98
1	1	0	Hotel	11	103566	61.0	20	8387933.0	295.86
1	1	1	Hotel	41	956110	43.0	47	72587024.0	2089.28
1	1	1	Hotel	10	61320	56.0	90	6794584.0	286.43
1	1	0	Hotel	18	175580	75.0	36	14172606.0	505.01

Variables qualitatives

OneHotEncoder()

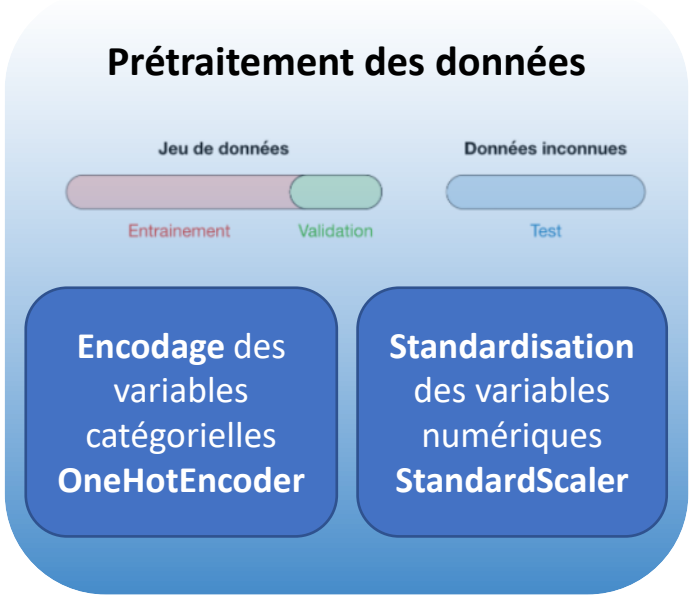
Variables quantitatives

StandardScalar()

Variables cibles (targets)

2. Modélisation et présentation des résultats

1. Prétraitement des données




```
print('X_train.shape = ', X_train.shape)
print('y_train.shape = ', y_train.shape)
print('X_test.shape = ' , X_test.shape)
print('y_test.shape = ', y_test.shape)
```

```
X_train.shape =  (1059, 27)
y_train.shape =  (1059,)
X_test.shape =  (454, 27)
y_test.shape =  (454,)
```

NaturalGas_use	Electricity_use	Steam_use	NumberofFloors	PropertyGFATotal	Agebuilding	_Distribution Center	_Hospital	_Hotel	_K-12 School	_Laboratory	_Large Office	_Low-Rise Multifamily	_Medical Office	_Mixed Use Property	...
1	1	1	-0.484918	-0.022843	-0.960387	0	0	0	0	0	0	0	0	0	
1	1	0	-0.046638	-0.183166	-1.295848	0	0	0	0	0	0	0	0	1	
1	1	1	-0.484918	0.056647	-1.204359	0	0	0	0	0	0	0	0	0	
1	1	1	-0.484918	-0.444007	1.387842	0	0	0	0	0	0	0	0	0	
1	1	0	-0.484918	-0.404455	-1.387338	0	0	0	0	0	0	0	0	0	

Anticipez les besoins en consommation de bâtiments

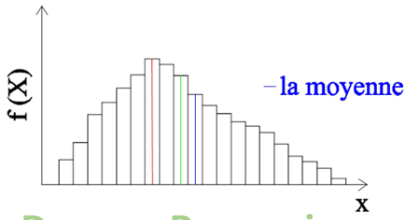
- **Partie 1: Nettoyage et exploration des données**
  - Nettoyage des données
  - Exploration des données
- **Partie 2: Modélisation et présentation des résultats**
  - Prétraitement des données
  - Modélisation de la consommation énergétique
  - Modélisation des émissions de CO2
  - Intérêt de l'ENERGYSTARScore
- **Conclusion**

2. Modélisation et présentation des résultats

2. Modélisation

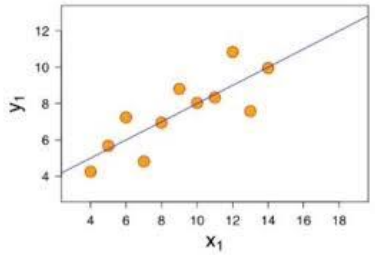
Apprentissage supervisé (Régression)

Modèles triviaux

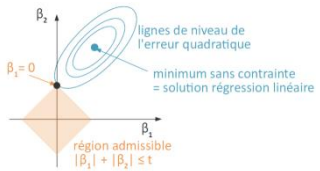


Dummy Regression

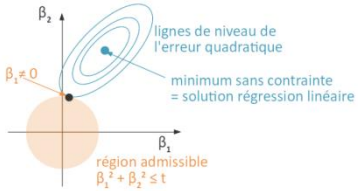
Modèles linéaires



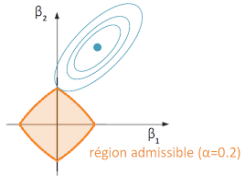
LinearRegression



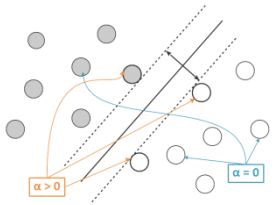
Lasso



Ridge

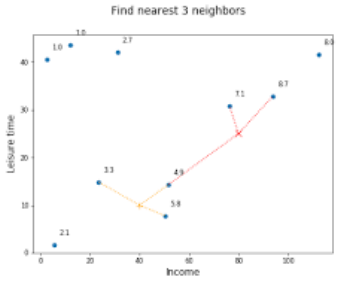


ElasticNet

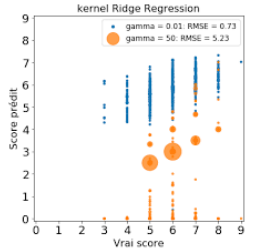


LinearSVR

Modèles non linéaires

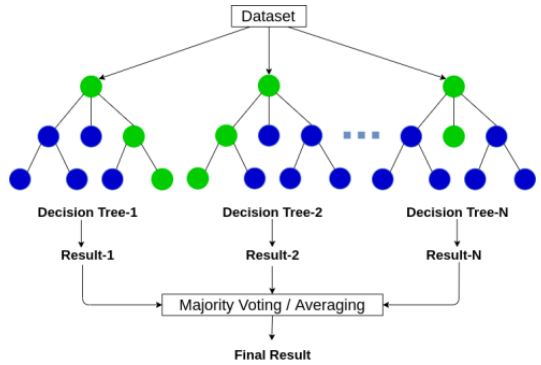


KNeighborsRegressor



KernelRidge

Modèles ensemblistes



Random Forest Regressor

ExtraTreesRegressor

GradientBoostingRegressor

2. Modélisation et présentation des résultats

2. Modélisation

Nous allons calculer **3 métriques** pour évaluer nos modèles :

- y_i : valeure réelle
- $\hat{y}_i$: valeure prédite
- $\bar{y}$: valeure moyenne

- **MAE : Mean Absolute Error**, l'erreur absolue moyenne, est la moyenne des valeurs absolues des erreurs. La MAE est dans la même unité que la variable à prédire. Plus elle est élevée, moins le modèle est performant. La MAE pénalise autant les grandes erreurs que les petites erreurs, contrairement à la RMSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **RMSE : Root Mean Squared Error** est la racine de l'erreur quadratique moyenne. La RMSE est dans la même unité que la variable à prédire. Plus elle est élevée, moins le modèle est performant. La RMSE pénalise plus fortement les grandes erreurs que les petites.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

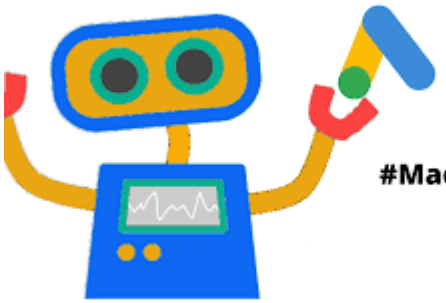
- **R^2 : Coefficient de détermination** est le carré du coefficient de corrélation linéaire. Plus le R^2 est proche de 1 meilleure est notre prédiction. Le coefficient de détermination nous indique à quel point les valeurs prédites sont corrélées aux vraies valeurs.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

2. Modélisation et présentation des résultats

2. Modélisation de la consommation énergétique

SiteEnergyUse(kBtu)



#Machine Learning

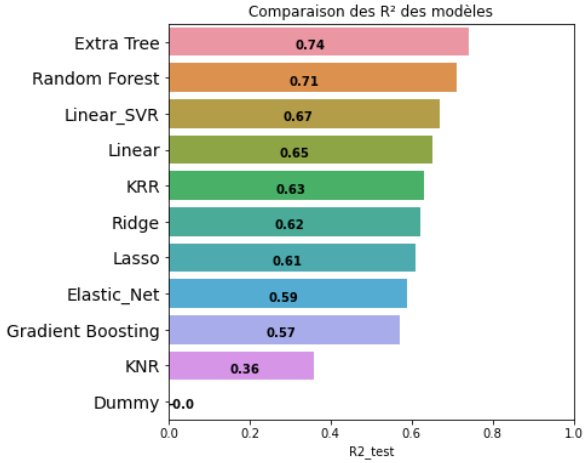
Hyperparamètres

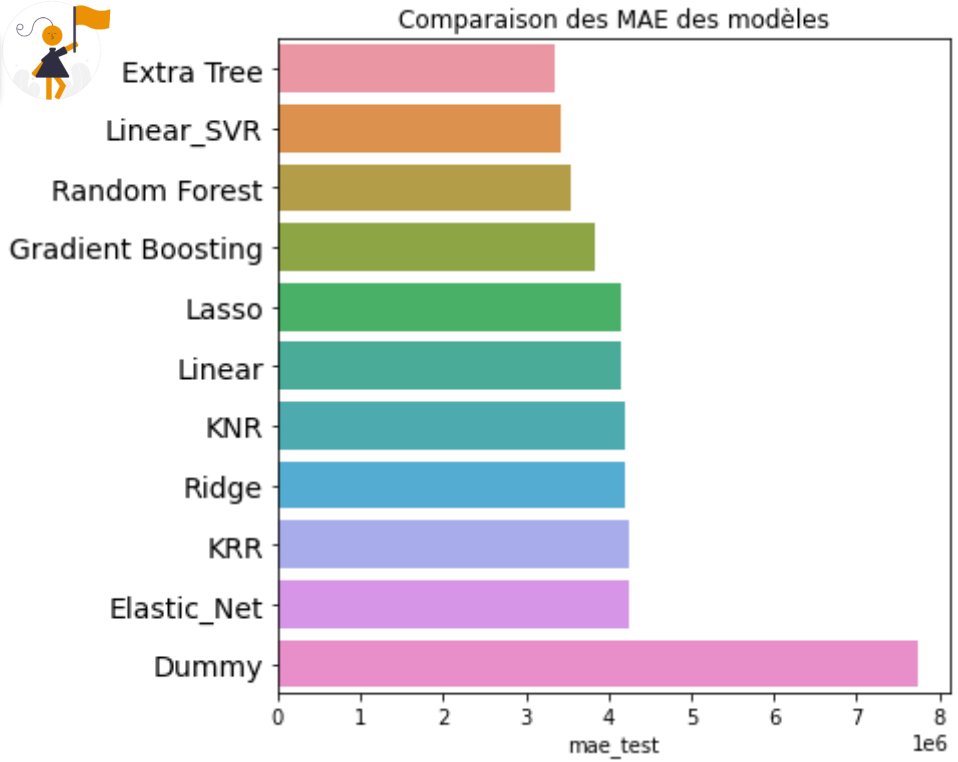
```
Dummy {}
Linear {}
Ridge {'alpha': 1.6681005372000592}
Lasso {'alpha': 215443.46900318822}
Elastic_Net {'alpha': 0.005994842503189421}
Linear_SVR {'C': 59948425.03189421, 'epsilon': 46.41588833612782}
KRR {'gamma': 1e-20}
KNR {'n_neighbors': 8}
Random Forest {'max_depth': 20, 'n_estimators': 100}
Extra Tree {'criterion': 'mae', 'max_depth': 8, 'n_estimators': 100}
Gradient Boosting {'n_estimators': 50}
```

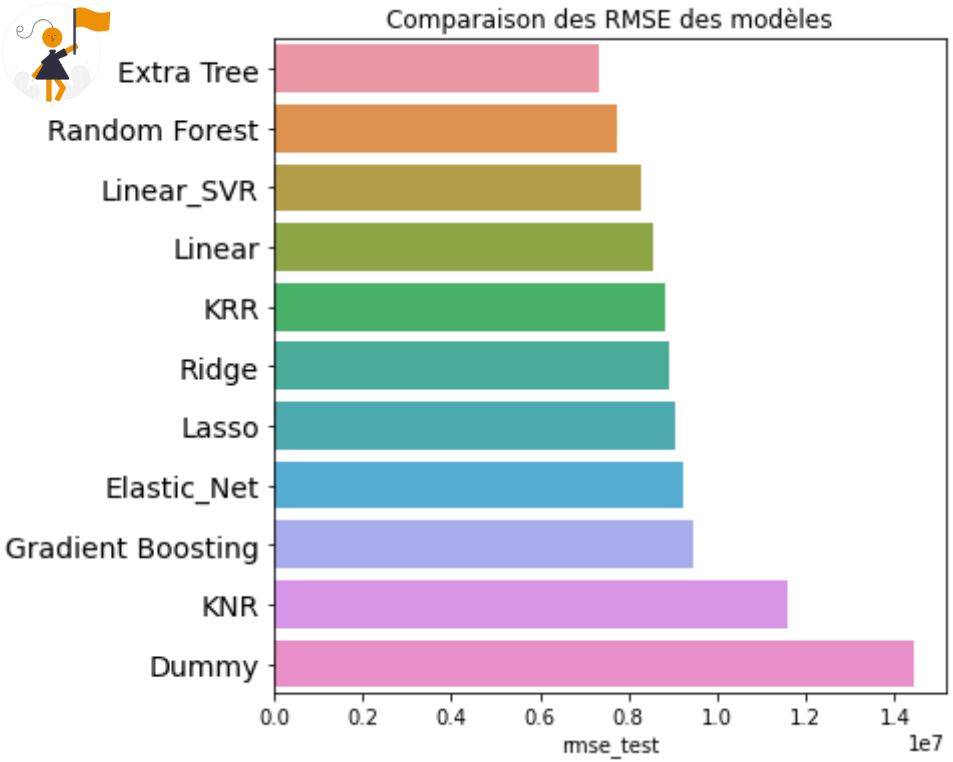
2. Modélisation et présentation des résultats

2. Modélisation de la consommation énergétique

SiteEnergyUse(kBtu)







Modèle retenu pour la prédiction de la consommation énergétique: **Extra Trees**

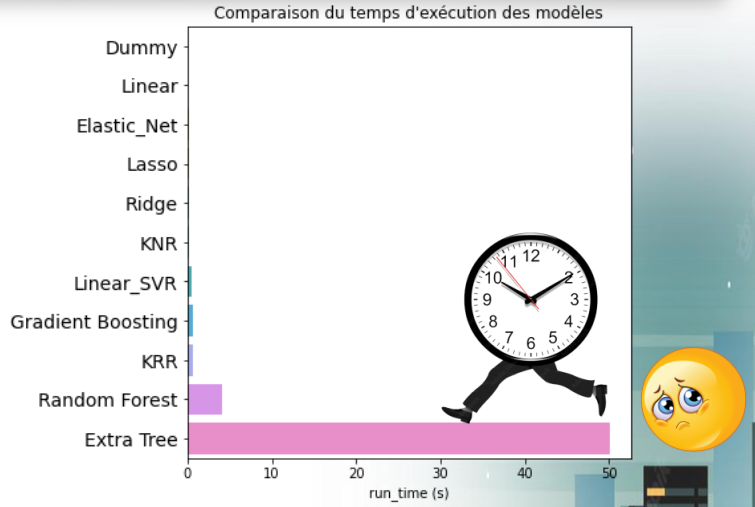
- **Extra Trees hyperparamètres:**`'criterion': 'mae', 'max_depth': 8, 'n_estimators': 100`

2. Modélisation et présentation des résultats

2. Modélisation de la consommation énergétique

Modèle retenu pour la prédiction de la consommation énergétique: **Extra Trees**

SiteEnergyUse(kBtu)



- **sklearn.ensemble.ExtraTreesRegressor**

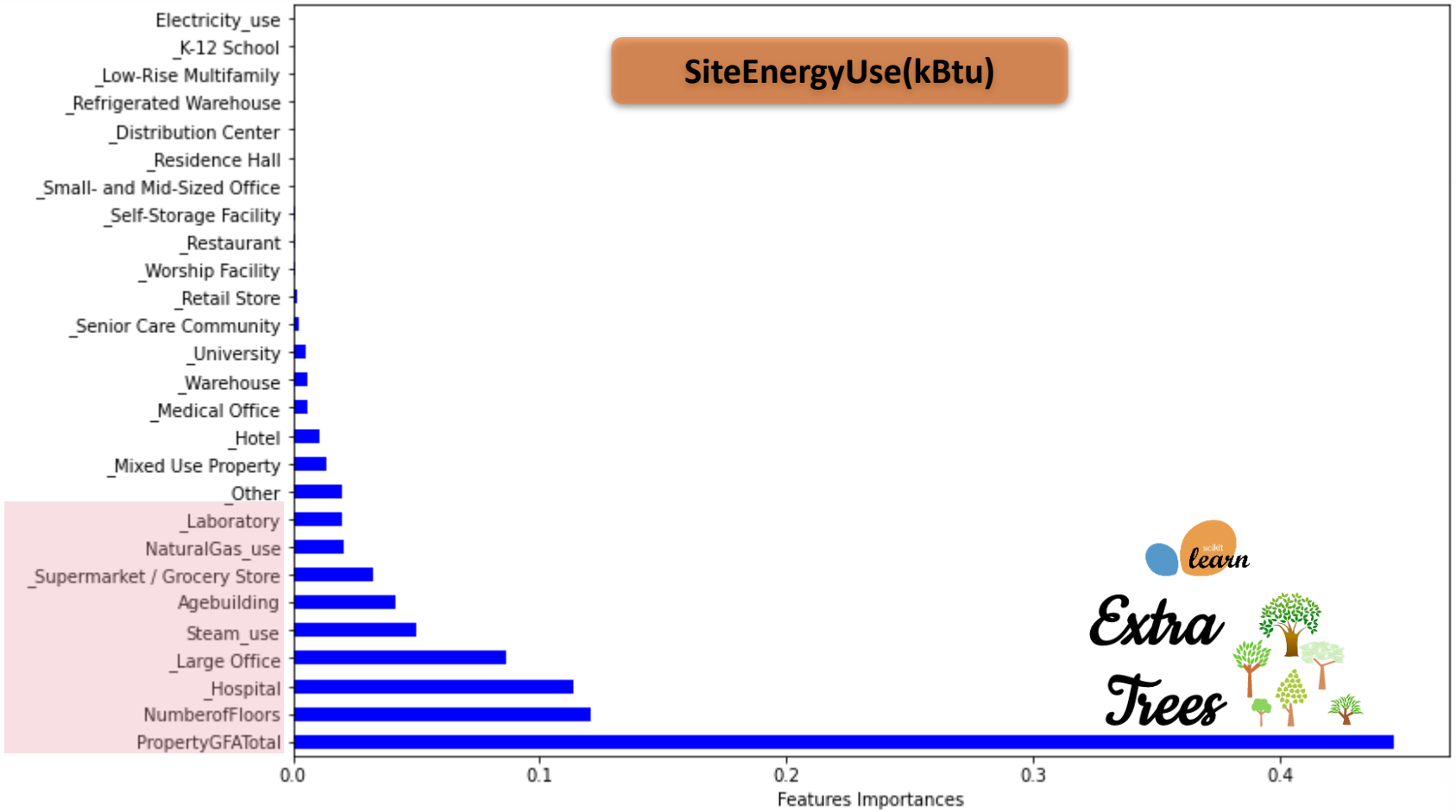
Cette classe implémente un méta-estimateur qui ajuste un certain nombre d'arbres de décision aléatoires sur divers sous-échantillons de l'ensemble de données et utilise la moyenne pour améliorer la précision prédictive et contrôler le sur-ajustement.

- **`Extra Trees hyperparamètres:`**
  - `'n_estimators': 100, le nombre d'arbres dans la forêt.`
  - `'criterion': 'mae', la fonction pour mesurer la qualité d'un fonctionnement.`
  - `'max_depth': 8, la profondeur maximale de l'arbre.`

2. Modélisation et présentation des résultats

2. Modélisation de la consommation énergétique

Modèle retenu pour la prédiction de la consommation énergétique: **Extra Trees**



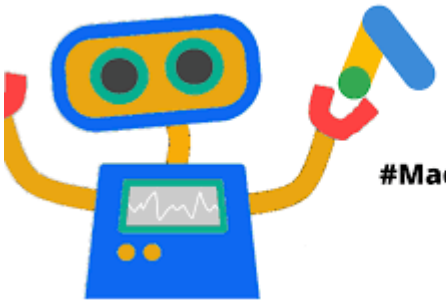
Anticipez les besoins en consommation de bâtiments

- **Partie 1: Nettoyage et exploration des données**
  - Nettoyage des données
  - Exploration des données
- **Partie 2: Modélisation et présentation des résultats**
  - Prétraitement des données
  - Modélisation de la consommation énergétique
  - Modélisation des émissions de CO2
  - Intérêt de l'ENERGYSTARScore
- **Conclusion**

2. Modélisation et présentation des résultats

3. Modélisation des émission de CO2

TotalGHGEmissions



#Machine Learning

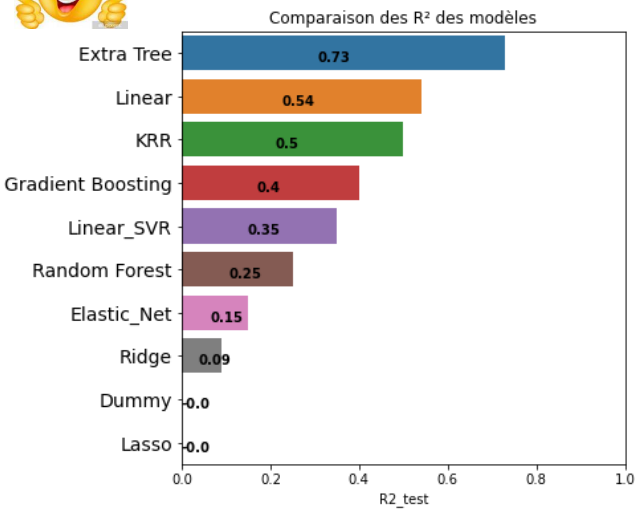
Hyperparamètres

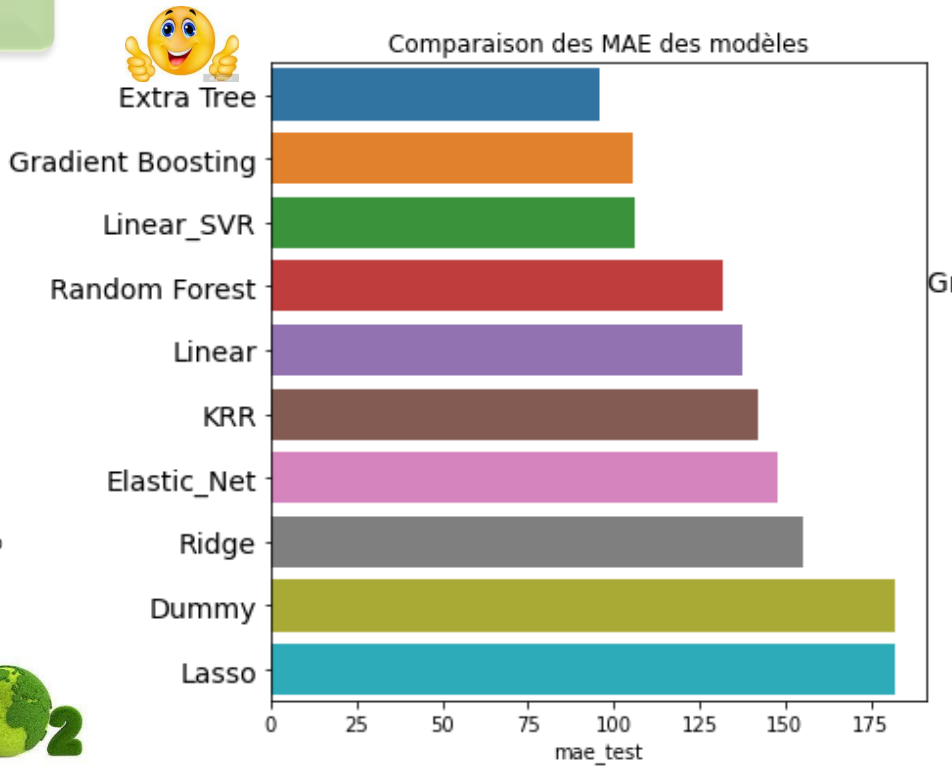
```
Dummy {}
Linear {}
Ridge {'alpha': 100.0}
Lasso {'alpha': 464.15888336127773}
Elastic_Net {'alpha': 1.0}
Linear_SVR {'C': 166.81005372000593, 'epsilon': 46.41588833612782}
KRR {'gamma': 1e-20}
Random Forest {'max_depth': 5, 'n_estimators': 150}
Extra Tree {'criterion': 'mse', 'max_depth': 8, 'n_estimators': 100}
Gradient Boosting {'n_estimators': 50}
```

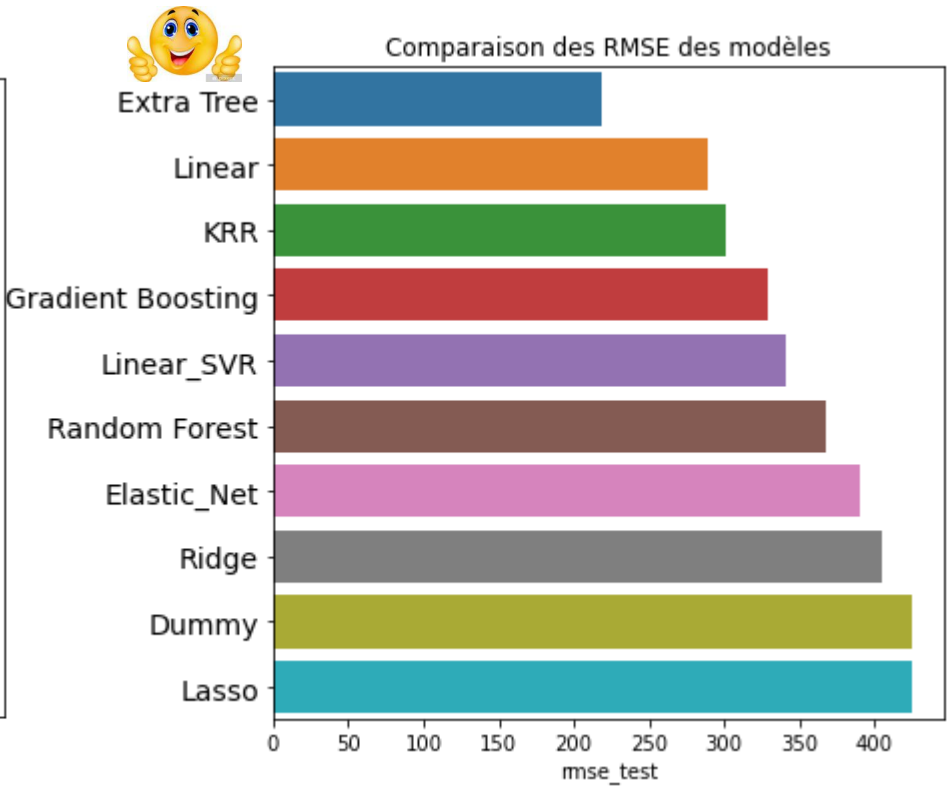
2. Modélisation et présentation des résultats

3. Modélisation des émission de CO2

TotalGHGEmissions







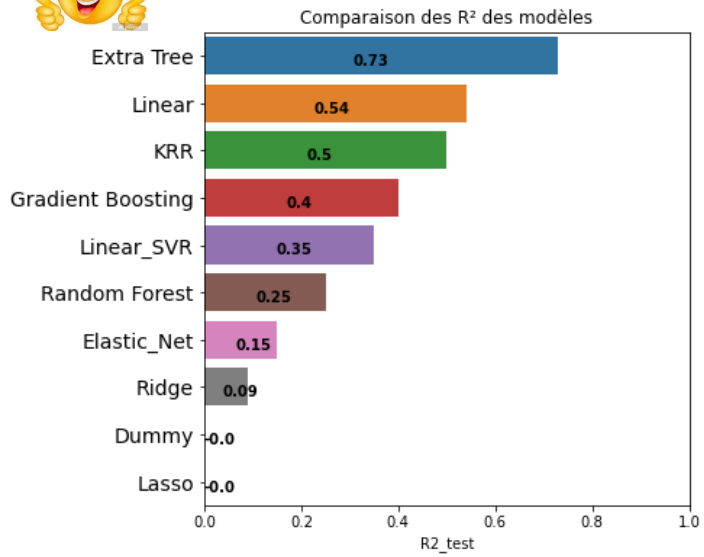
Modèle retenu pour la prédiction des émissions
ExtraTrees

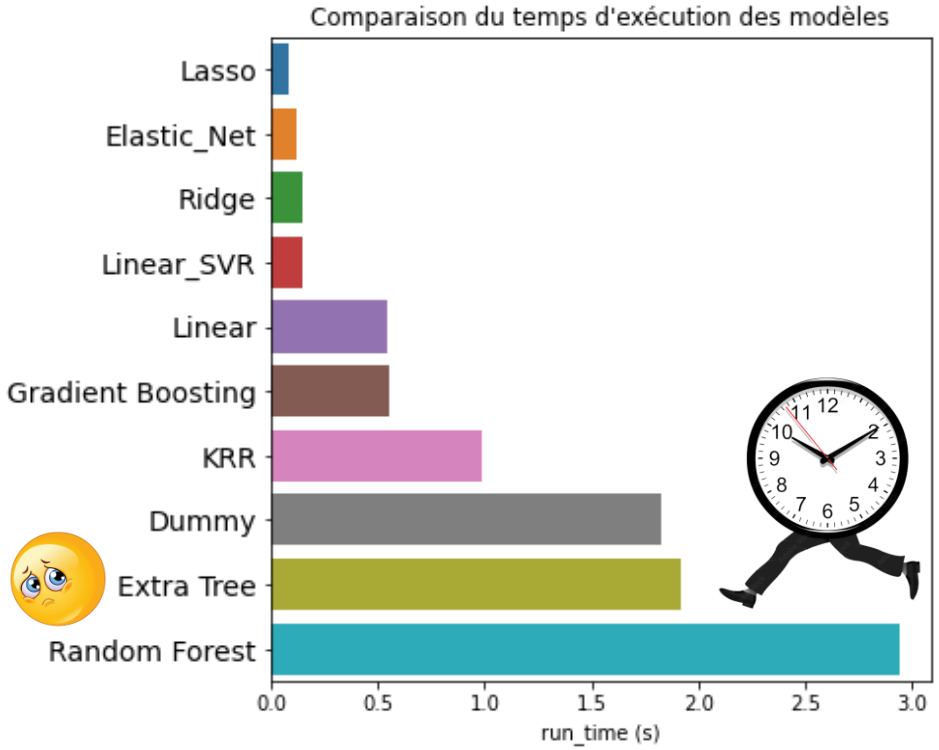
- **Extra Trees hyperparamètres:**`'criterion': 'mse', 'max_depth': 8, 'n_estimators': 100`

2. Modélisation et présentation des résultats

3. Modélisation des émission de CO2

TotalGHGEmissions





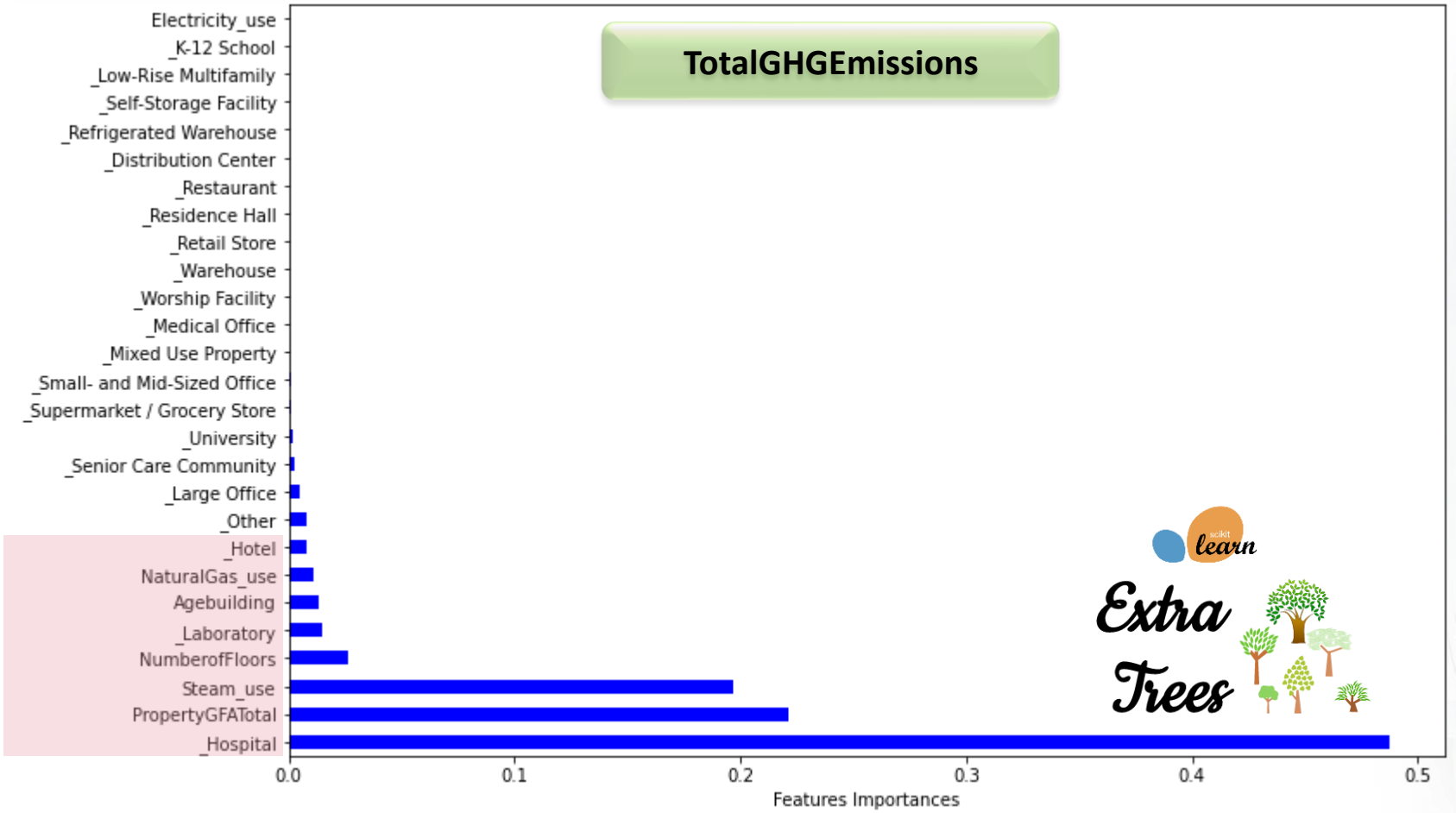
Modèle retenu pour la prédiction des émissions **ExtraTrees**

Extra Trees hyperparamètres:`'criterion': 'mse', 'max_depth': 8, 'n_estimators': 100`

2. Modélisation et présentation des résultats

3. Modélisation des émission de CO2

Modèle retenu pour la prédiction des émissions: **Extra Trees**

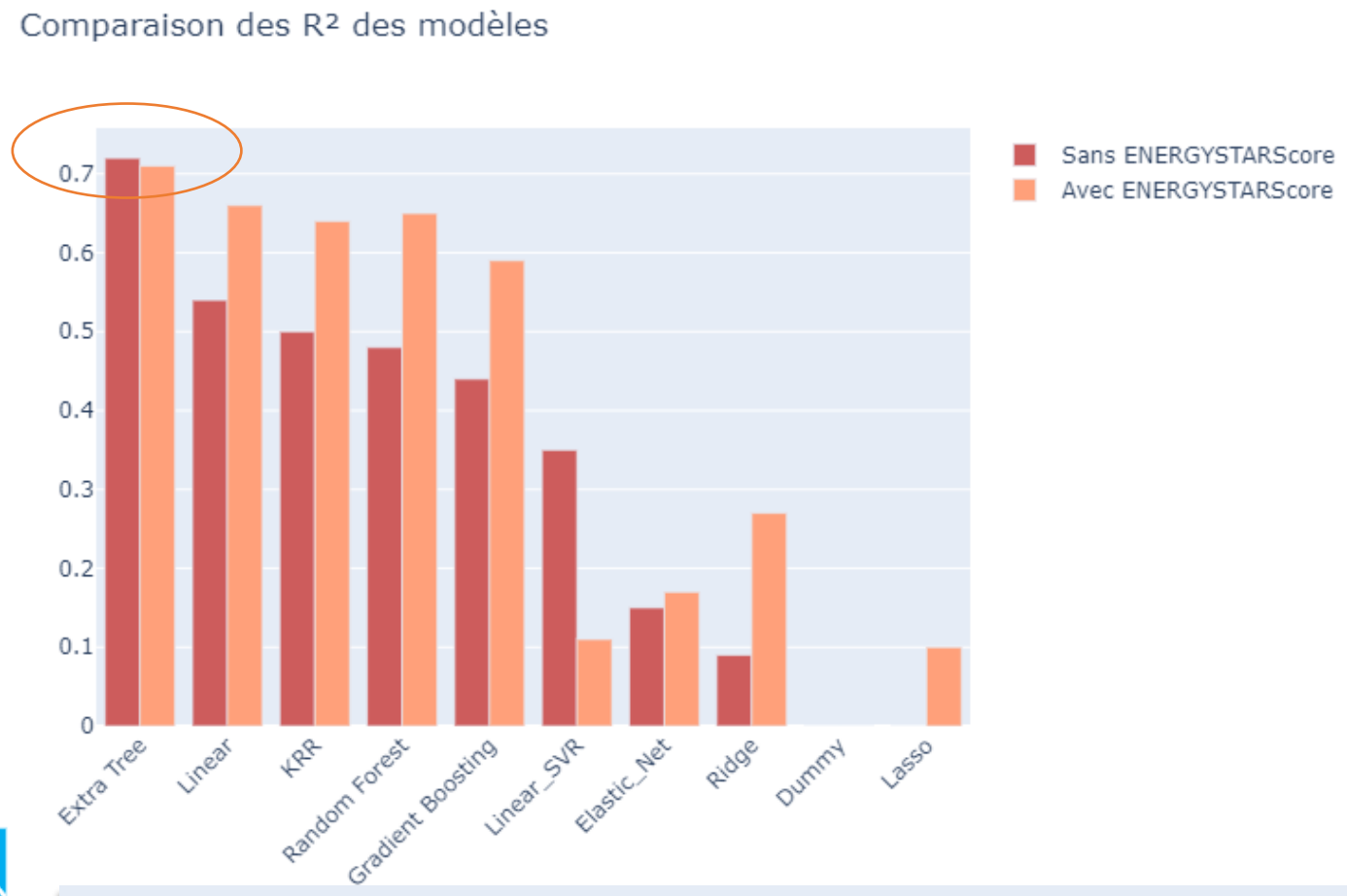


Anticipez les besoins en consommation de bâtiments

- **Partie 1: Nettoyage et exploration des données**
  - Nettoyage des données
  - Exploration des données
- **Partie 2: Modélisation et présentation des résultats**
  - Prétraitement des données
  - Modélisation de la consommation énergétique
  - Modélisation des émissions de CO2
  - Intérêt de l'ENERGYSTARScore
- **Conclusion**

2. Modélisation et présentation des résultats

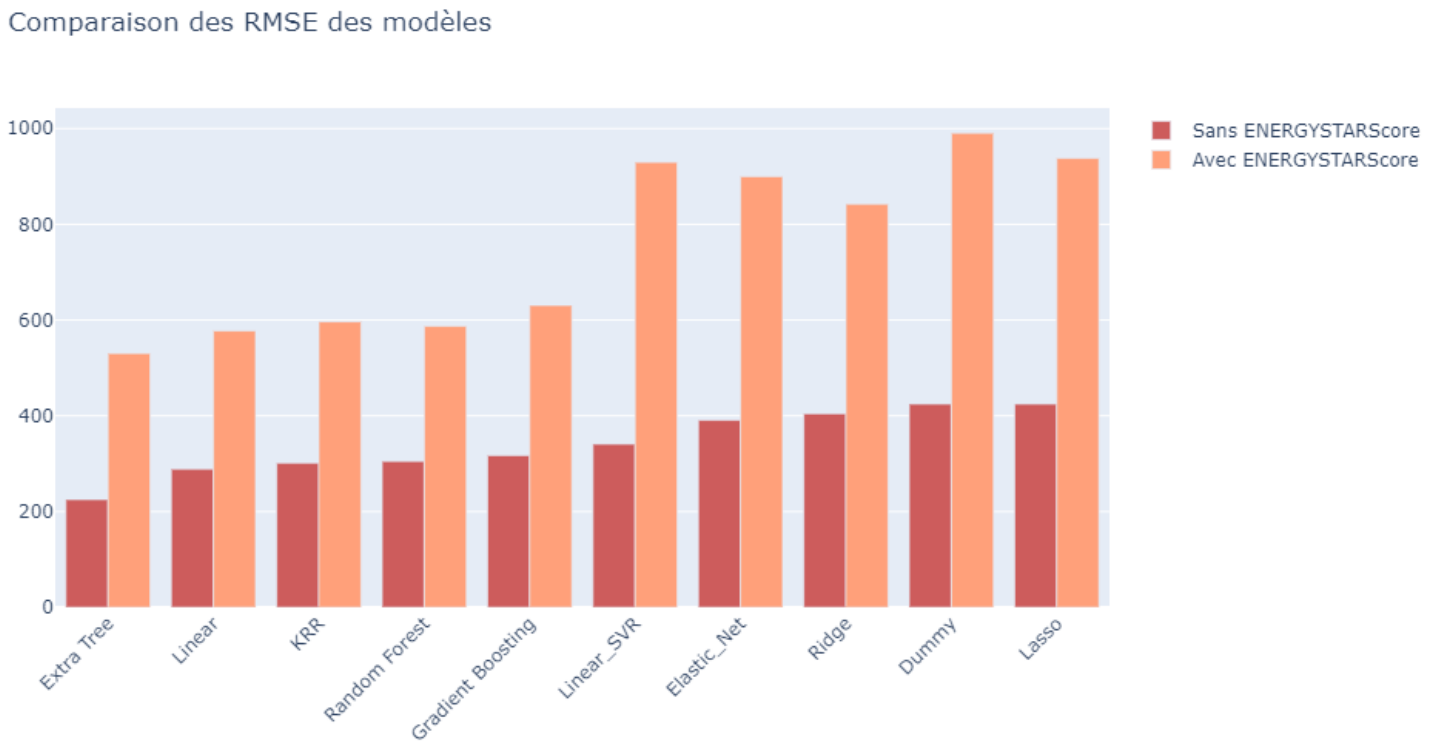
4. Modélisation des émission de CO2: Intérêt de l'ENERGYSTARScore

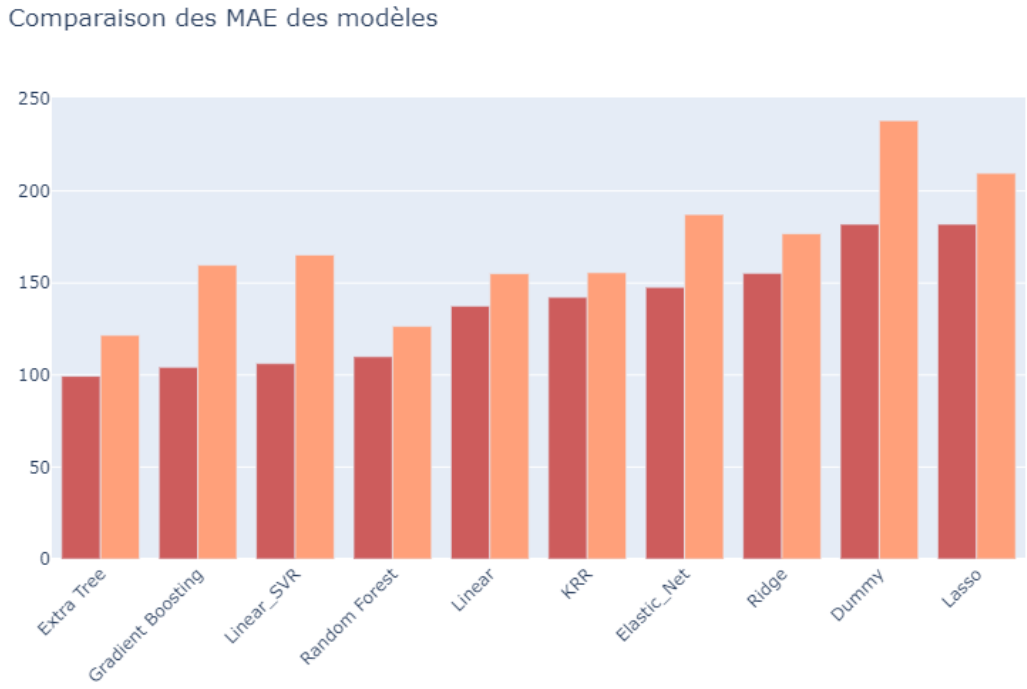


L 'ENERGYSTARScore n'est intéressant pour la prédiction des émissions

2. Modélisation et présentation des résultats

4. Modélisation des émission de CO2: Intérêt de l'ENERGYSTARScore





L 'ENERGYSTARScore n'est intéressant pour la prédiction des émissions

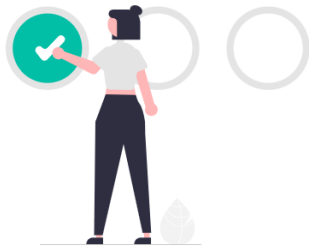
2. Modélisation et présentation des résultats

Synthèse

Dans cette deuxième partie, nous avons effectué:

- un prétraitement des données: **standardisation** et **encodage**
- une séparation des données en **train** set et **test** set
- une implémentation des **modèles de régression** linéaires, non linéaires et ensemblistes
- recherche des hyperparamètres optimaux pour chaque modèle (GridsearchCV)
- calculer pour chaque modèle des métriques suivantes: MAE, RMSE et R^2

Ces opérations ont permis de prédire la consommation énergétique et les émissions et d'étudier l'intérêt de l'ENERGySTARScore



- Modèle final ensembliste: **ExtraTrees**
- **SiteEnergyUse(kBtu): R^2 = 0,74**
- **TotalGHGEmissions: R^2 = 0,73**
- **L'ENERGYSTARScore n'est pas intéressant pour la prédiction des émissions**

Anticipez les besoins en consommation de bâtiments

- **Partie 1: Nettoyage et exploration des données**
  - Nettoyage des données
  - Exploration des données
- **Partie 2: Modélisation et présentation des résultats**
  - Prétraitement des données
  - Modélisation de la consommation énergétique
  - Modélisation des émissions de CO2
  - Intérêt de l'ENERGYSTARScore
- **Conclusion**

Anticipez les besoins en consommation de bâtiments

Conclusion

- Confirmer les compétences acquises en nettoyage et exploration des données
- Transformer les variables pertinentes d'un modèle d'apprentissage
- Mettre en place des **modèles d'apprentissage supervisé** adapté au problème métier
- Evaluer les **performances** de ces modèles
- Adapter les **hyperparamètres** d'un algorithme d'apprentissage supervisé afin de l'améliorer

Seattle

Anticipez les besoins en consommation de bâtiments

Merci de votre attention

Seattle