Projet 3 : Soutenance

Anticipez les besoins en consommation électrique de bâtiments

Gaëtan PELLETIER

Sommaire

- Problématiques, interprétation et pistes de recherche envisagées
- Nettoyage des données, feature engineering et exploration/analyse
- Modélisations effectuées
- Choix du modèle final
- Synthèse

Projet 3 : Soutenance

Problématiques,
interprétation
et pistes de recherche envisagées

Problématiques

En se basant sur les données des bâtiments de Seattle, les questions sont :

- Quelle sera la consommation totale d'énergie ?
- Quelle sera la quantité de CO2 émis ?
- ENERGYSTARScore est-il pertinent ?

Interprétation

- **Quelle sera la consommation totale d'énergie ?**
  - Construire un modèle sans les features liées à l'énergie
  - Cible à prédire : SiteEnergyUse
- **Quelle sera la quantité de CO2 émis ?**
  - Construire un modèle sans les features liées au CO2
  - Cible à prédire : GHGEmissions
- **ENERGY STAR Score est-il pertinent ?**
  - Ajout de cette feature dans les modèles précédents

Pistes de recherche envisagées

- **Nettoyage des données**
- **Analyse des features :**
  - Corrélations avec la cible ?
  - Indépendance des features entre elles ?
- **Transformation des données**
- **Comparer différents modèles**
- **Optimiser le meilleur modèle**
- **Observer les effets d'ENERGYSTAR Score**

Nettoyage des données, feature engineering et exploration/analyse

Nettoyage des données

- **Mémoire Ram :**
  - Les données utilisent 1,2MB de mémoire RAM
- **On supprime les bâtiments « familiaux »**
- **Complétion des features (infos bâtiments + parking) :**
  - surfaces principales, secondaires et tertiaires
- **Séparation de latitude et longitude**
- **Discrétisation :**
  - Yearbuilt, NumberofBuildings, NumberofFloors, CouncilDistrictCode
- **Suppression des « NaN » et « 0 » dans cibles**

Exploration / Analyse

Corrélations avec SiteEnergyUse :

- Features quantitatives

SiteEnergyUse(kBtu) corr with:	corr	p-value
GHGEmissions(MetricTonsCO2e)	0.887696	0.000000e+00
LargestPropertyUseTypeGFA	0.710549	1.812494e-252
PropertyGFABuilding(s)	0.691718	5.787033e-234
PropertyGFATotal	0.676523	4.831835e-220
SecondLargestPropertyUseTypeGFA	0.548749	1.095418e-129
PropertyGFAParking	0.313273	1.078475e-38
ThirdLargestPropertyUseTypeGFA	0.214450	1.592062e-18
OSEBuildingID	-0.205099	4.802256e-17
CouncilDistrictCode	0.106125	1.647699e-05

- Features qualitatives

Feature	eta_squared
PropertyName	1.00
TaxParcelIdentificationNumber	0.94
ListOfAllPropertyUseTypes	0.51
LargestPropertyUseType	0.41
PrimaryPropertyType	0.34

Exploration / Analyse

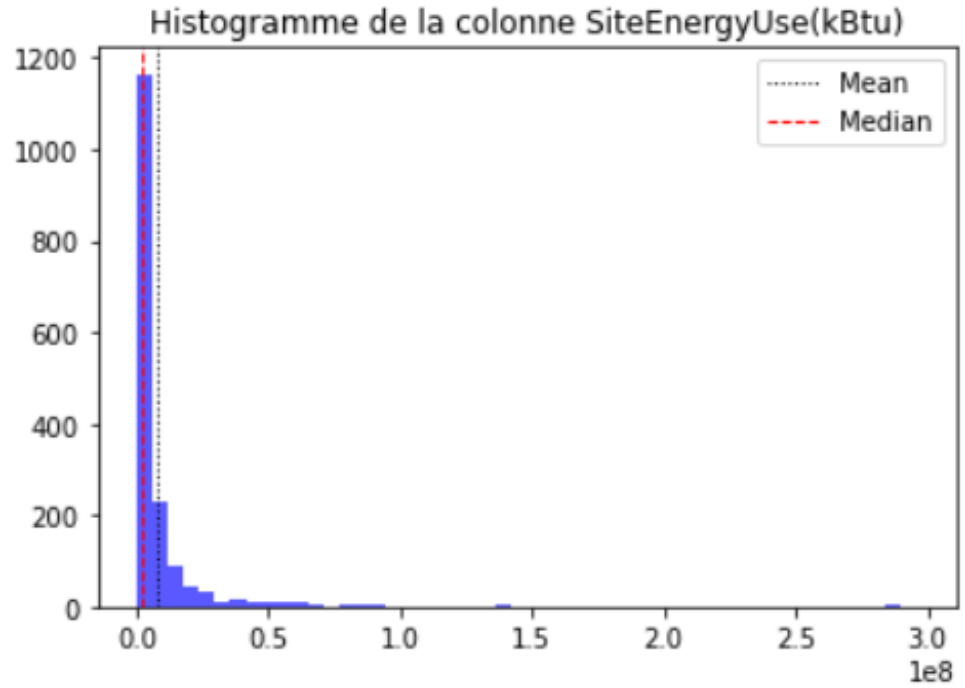
- **Corrélation particulière entre features :**

LargestPropertyUseTypeGFA corr with:	corr	p-value
PropertyGFATotal	0.960100	0.000000e+00
PropertyGFABuilding(s)	0.958684	0.000000e+00
SecondLargestPropertyUseTypeGFA	0.686926	1.760311e-229
PropertyGFAParking	0.531017	4.713097e-120
ThirdLargestPropertyUseTypeGFA	0.306453	5.084811e-37

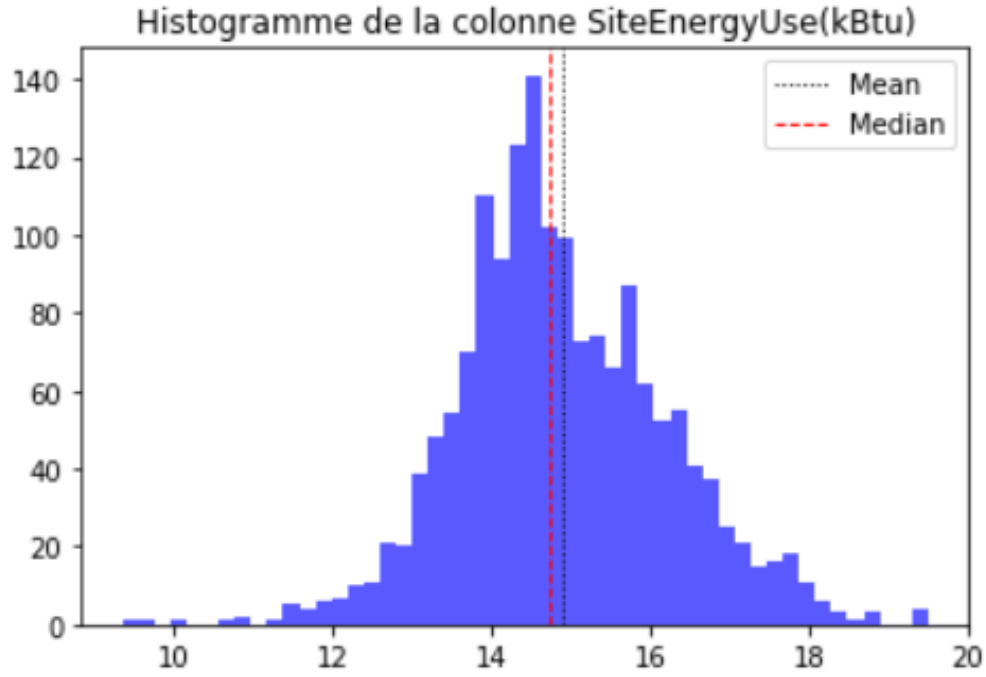
- **3 features sont très fortement corrélées :**
  - LargestPropertyUsetypeGFA
  - PropertyGFATotal
  - PropertyGFABuilding(s)
- **On garde seulement *LargestPropertyUsetypeGFA***

Exploration / Analyse

Analyse univariée de SiteEnergyUse :



→ log :




```
mean:     7703524.236          skewness: 8.980
median:   2474457.000          kurtosis: 112.630
var:      357280469700748.062
ect:      18901864.186
```

```
mean:     14.889     skewness: 0.200
median:   14.722     kurtosis: 0.530
var:      1.751
ect:      1.323
```

Feature engineering

- **Utilisation d'un logarithme pour obtenir une distribution normale :**

  → transformation $x = \log(x + 1)$

- **Pour les données quantitatives :**

  → utilisation de StandardScaler

- **Pour les données qualitatives :**

  → utilisation de OneHotEncoder

Projet 3 : Résumé choix des features

Pour SiteEnergyUse :

- ListOfAllPropertyUseTypes
- LargestPropertyUseType
- NumberofFloors
- LargestPropertyUseTypeGFA
- SecondLargestPropertyUseTypeGFA
- PropertyGFAParking
- ThirdLargestPropertyUseTypeGFA

Projet 3 : Résumé choix des features

Pour GHGEmissions:

- ListOfAllPropertyUseTypes
- LargestPropertyUseType
- LargestPropertyUseTypeGFA
- SecondLargestPropertyUseTypeGFA
- PropertyGFAParking

Projet 3 : Résumé nettoyage

- Dataset pour SiteEnergyUse :

```
data2015 shape:  (1581, 8)

Nombre de lignes   supprimees : 1759
Nombre de colonnes supprimees : 39

Nombre de lignes   supprimees : 52.7 %
Nombre de colonnes supprimees : 83.0 %

memory usage: 191.2+ KB
```

```
data2016 shape:  (1581, 8)

Nombre de lignes   supprimees : 1795
Nombre de colonnes supprimees : 38

Nombre de lignes   supprimees : 53.2 %
Nombre de colonnes supprimees : 82.6 %

memory usage: 111.2+ KB
```

- Dataset pour GHGEmissions :

```
data2015 shape:  (1581, 6)

Nombre de lignes   supprimees : 1759
Nombre de colonnes supprimees : 41

Nombre de lignes   supprimees : 52.7 %
Nombre de colonnes supprimees : 87.2 %

memory usage: 166.5+ KB
```

```
data2016 shape:  (1581, 6)

Nombre de lignes   supprimees : 1795
Nombre de colonnes supprimees : 40

Nombre de lignes   supprimees : 53.2 %
Nombre de colonnes supprimees : 87.0 %

memory usage: 166.5+ KB
```

Modélisations effectuées

Modélisations effectuées

Recherche du meilleur modèle :

- **Création d'une baseline (dummy regressor)**
- **Modèles sélectionnés :**
  - Linéaires (Lasso, Ridge)
  - Non linéaire (SVM à noyau gaussien)
  - Ensemblistes (forêt aléatoire, gradient boosting)
- **Entraînement des modèles :**
  - Données de 2015 : séparation en jeu d'entraînement et jeu de test
  - Transformation des données, via un pipeline (utilisation de la méthode ColumnTransformer)
  - Recherche sur grille + validation croisée
    - 5 plis
    - scoring = 'neg_root_mean_squared_error')
  - Métrique utilisée sur le jeu de test : RMSE
- **Autre scoring utilisé :**
  - Temps d'entraînement,
  - Temps de prédiction

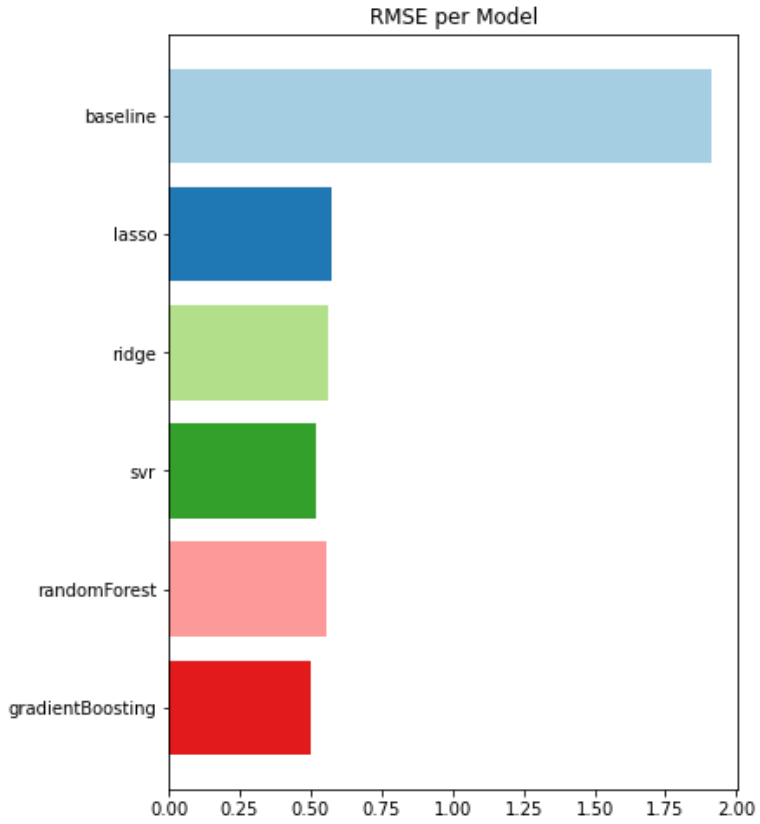
Modélisations effectuées

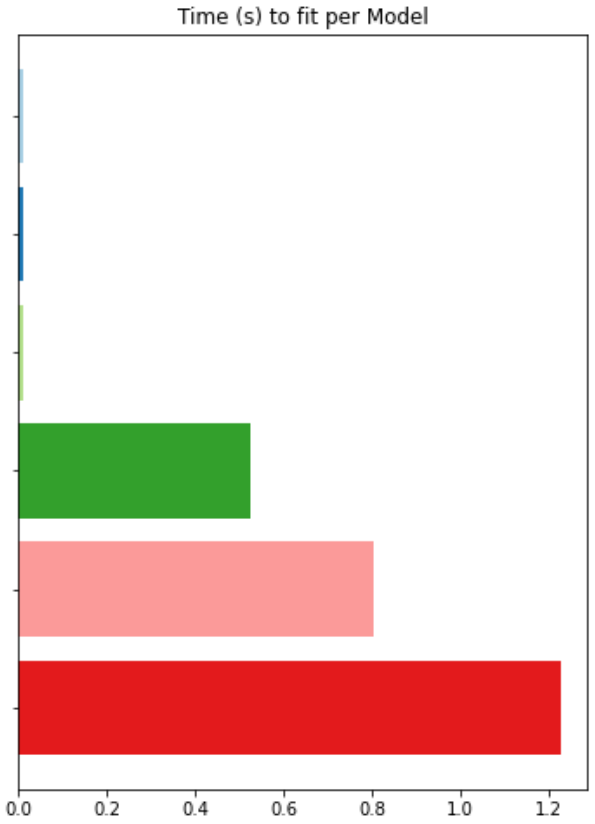
Hyperparamètres :

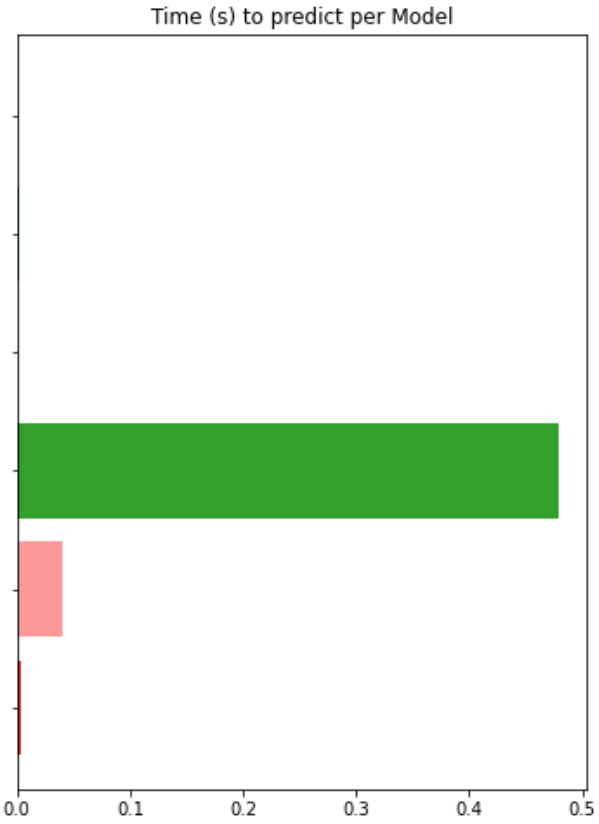
- Lasso :
  - alpha
  - max_iter
- Ridge :
  - alpha
- SVM à noyau gaussien :
  - C
- Forêt aléatoire :
  - n_estimators
  - max_depth
- Gradient Boosting :
  - n_estimarors
  - max_depth
  - learning_rate

Modélisations effectuées

Modélisations pour SiteEnergyUse :

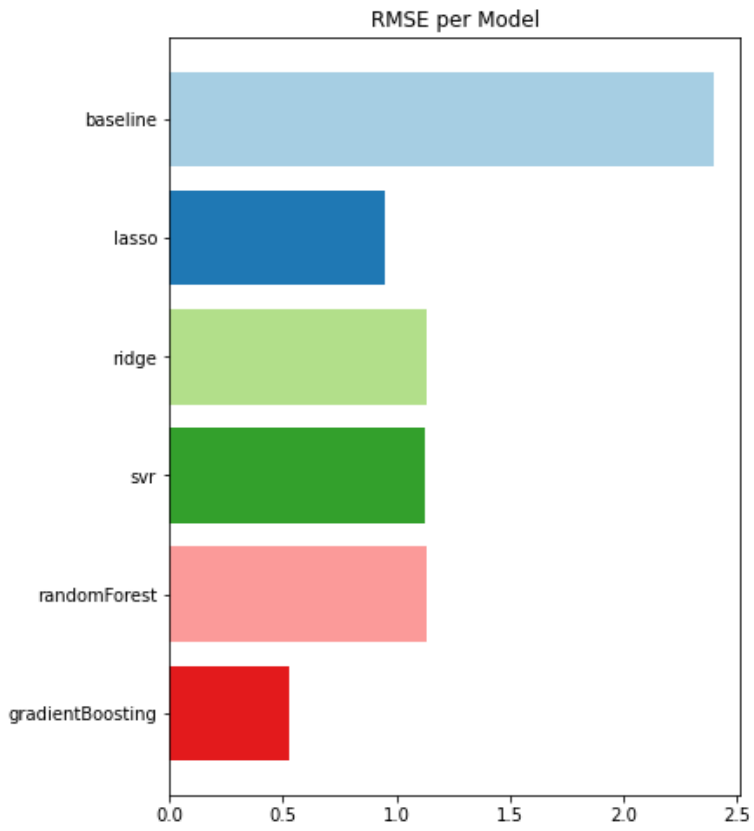


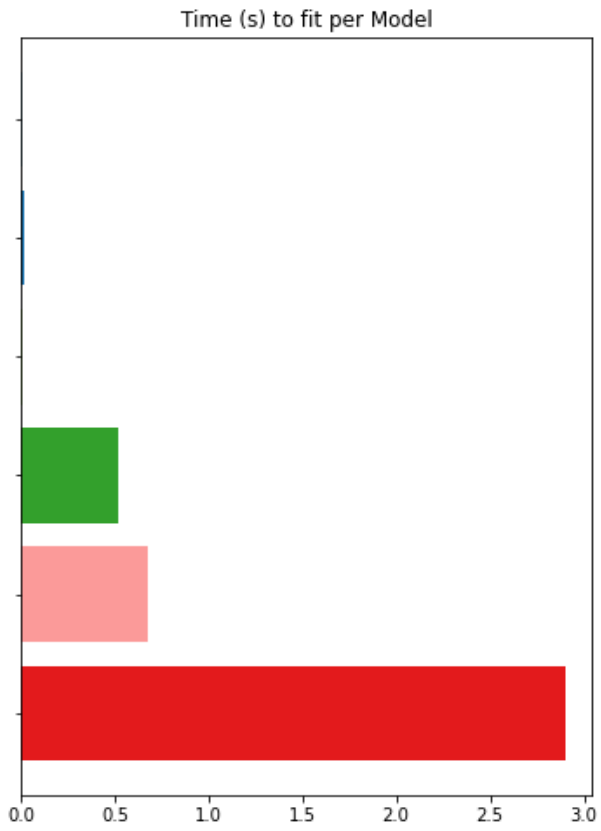


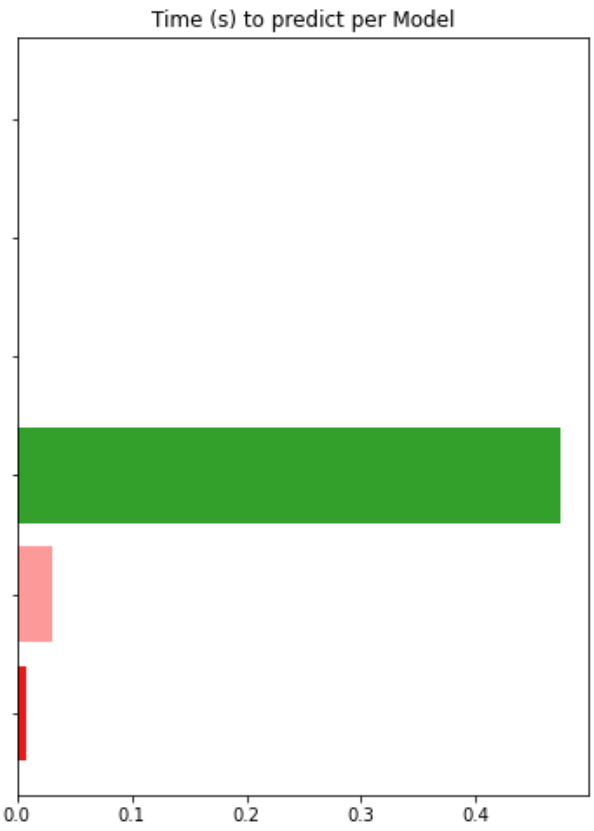


Modélisations effectuées

Modélisations pour GHGEmissions :







Choix du modèle final

Choix du modèle final : Gradient Boosting

Entraînement du gradient boosting :

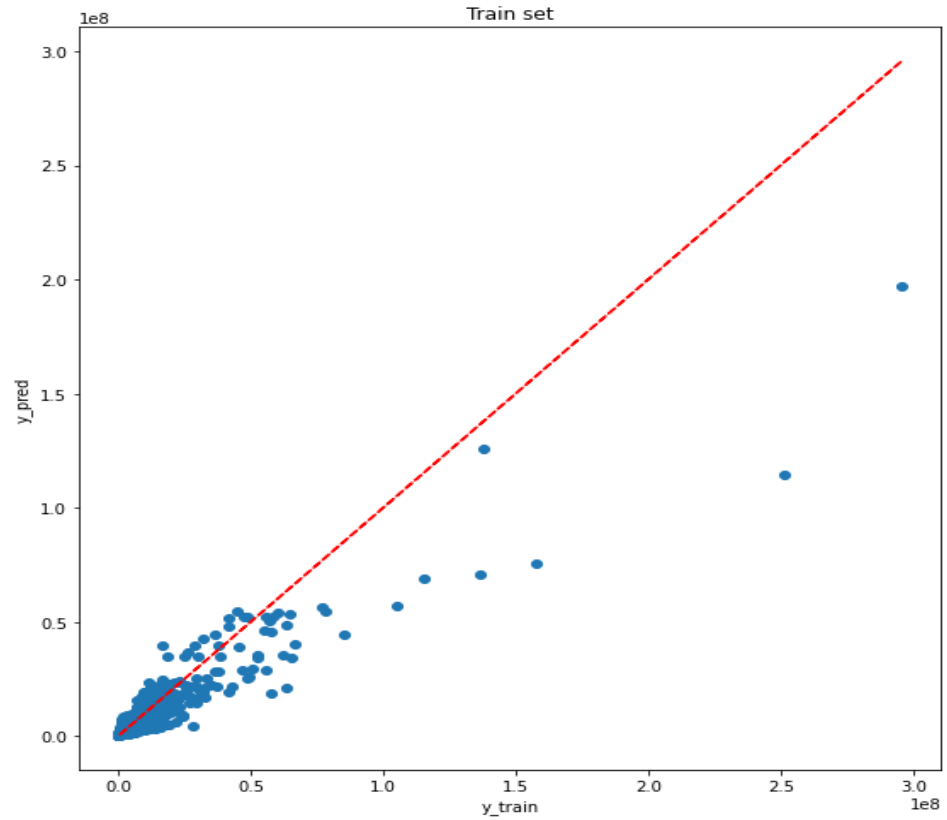
- Données de 2015 :
  - séparation en jeu d'entraînement et jeu de test
- Transformation des données :
  - StandardScaler
  - OneHotEncoder
- Recherche sur grille + validation croisée (5 plis)
- Métrique utilisée : R^2

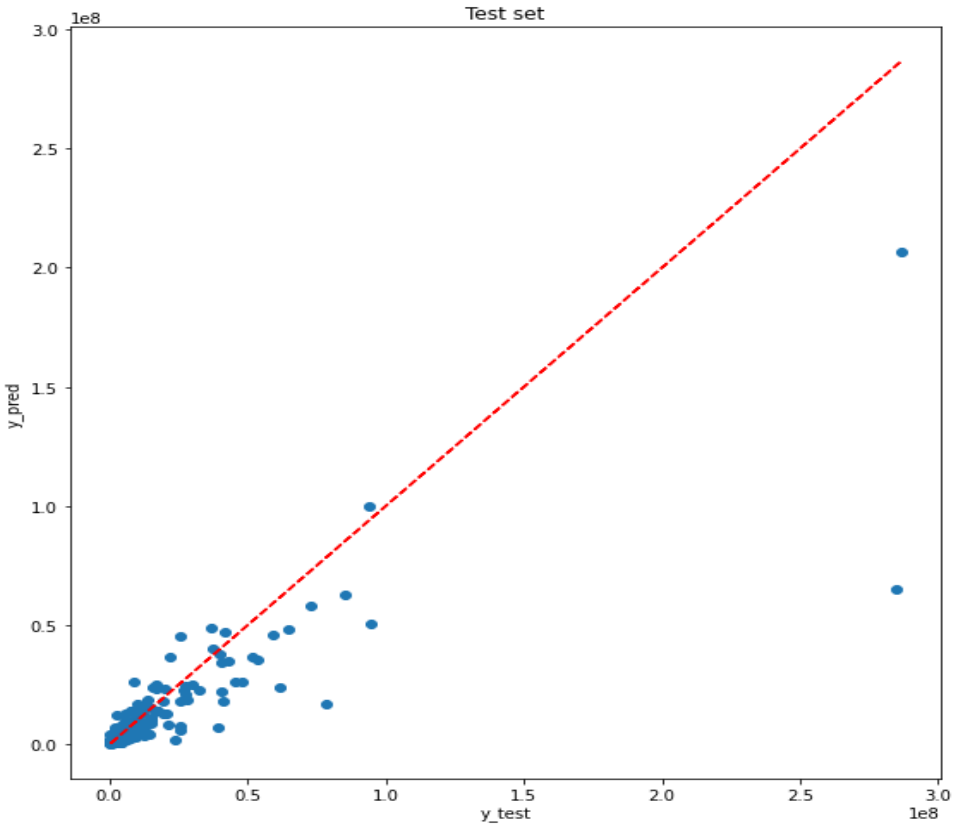
Choix du modèle final : Gradient Boosting

Modélisation SiteEnergyUse sans ENERGYSTARScore :

```
Train set score = 0.82
Test set score  = 0.74
```

```
#data2016
Test set score  = 0.80
```





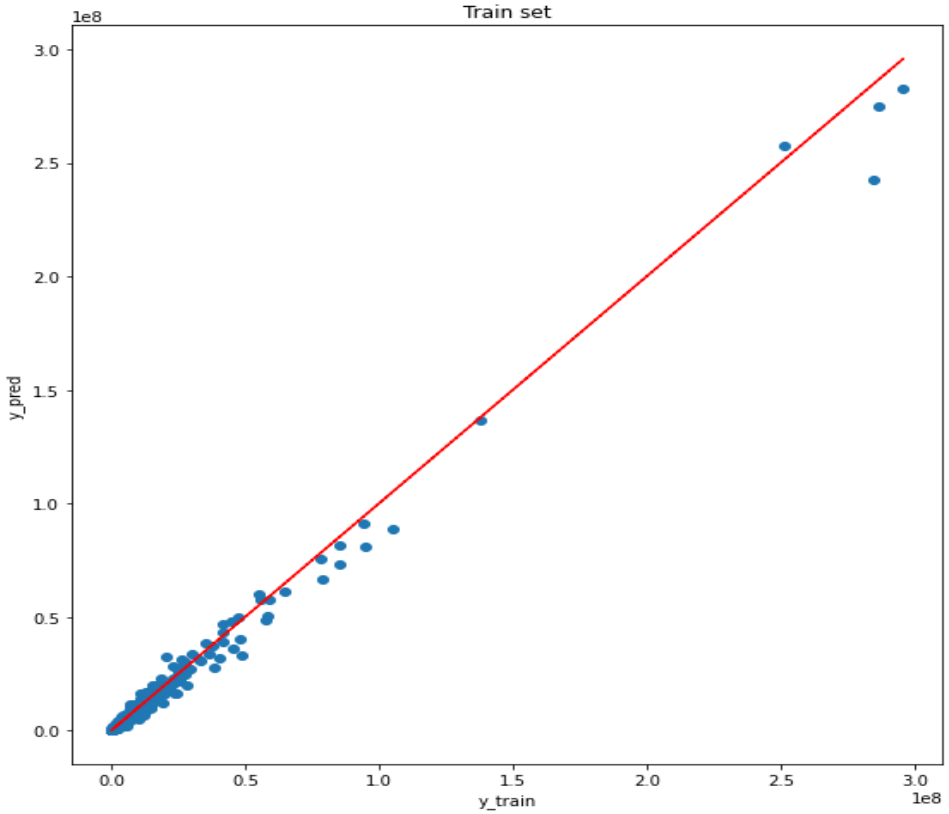
Choix du modèle final : Gradient Boosting

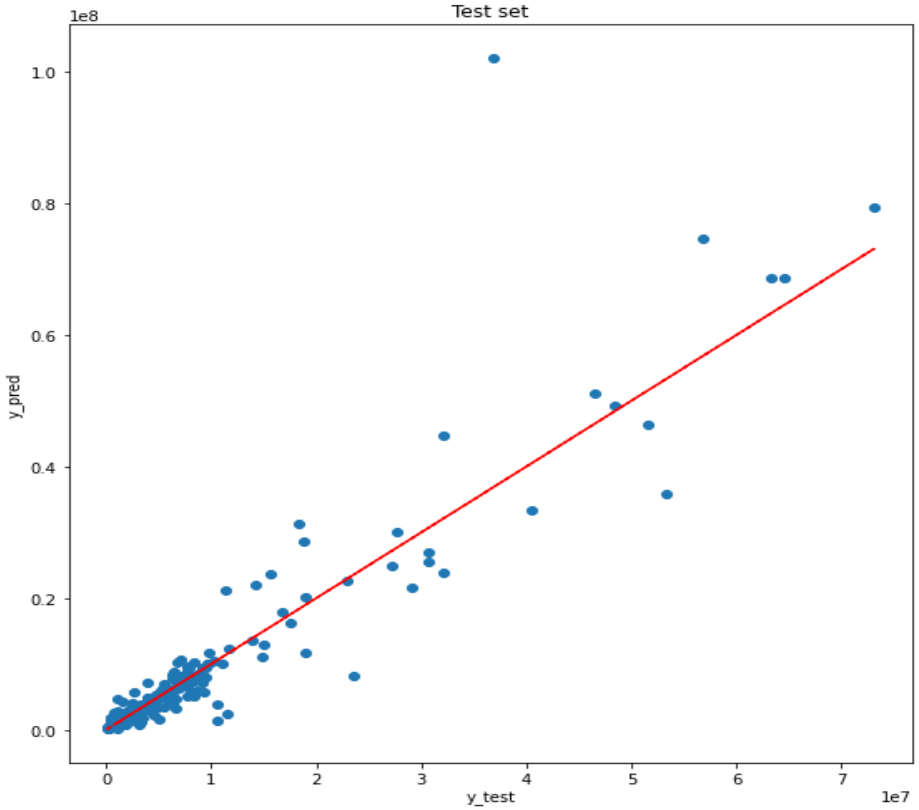
Modélisation SiteEnergyUse avec ENERGYSTARScore :

```
Train set score = 0.96
Test set score  = 0.87
```

(+ 13 points)

```
#data2016
Test set score  = 0.94
```



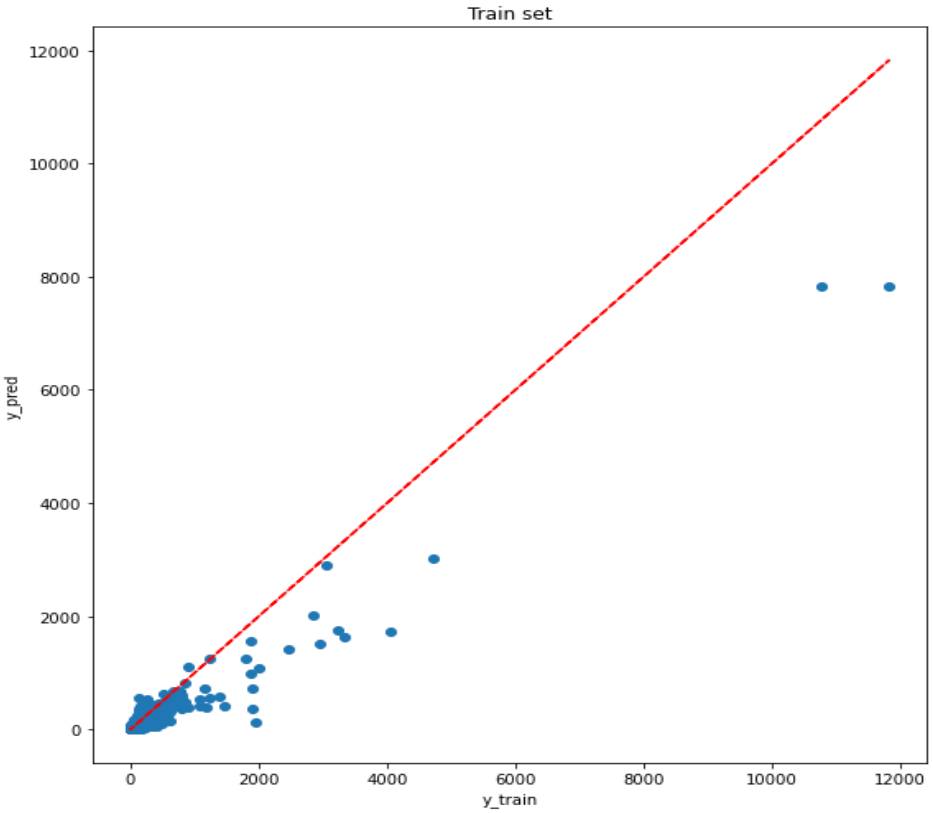


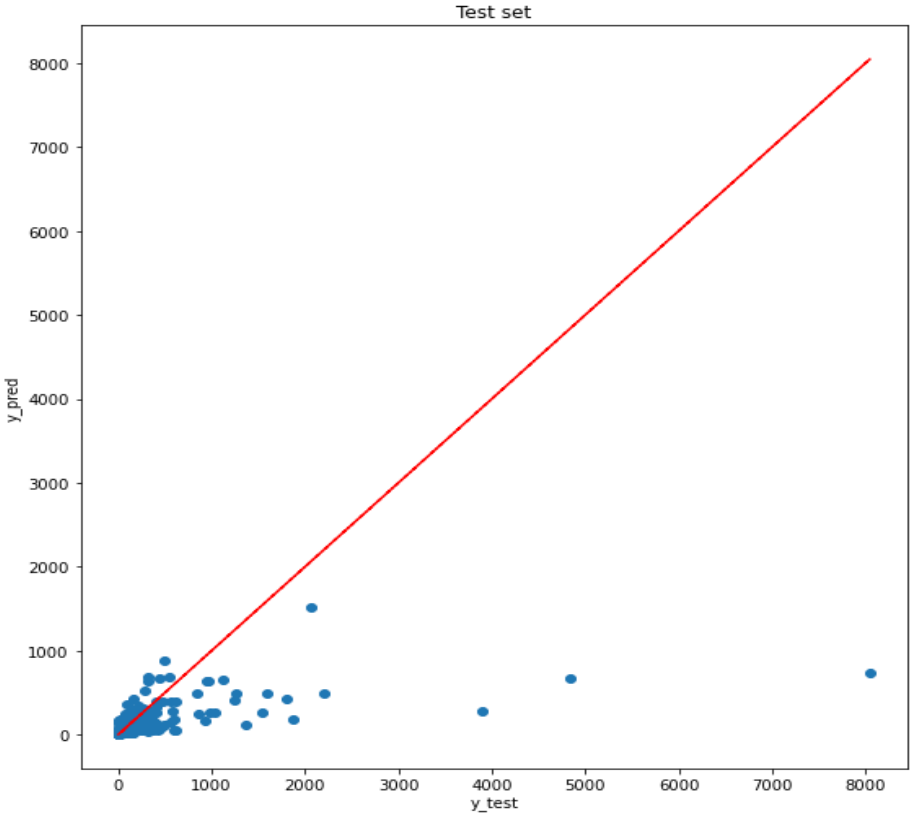
Choix du modèle final : Gradient Boosting

Modélisation GHGEmissions sans ENERGYSTARScore :

```
Train set score = 0.71
Test set score  = 0.53
```

```
#data2016
Test set score  = 0.66
```





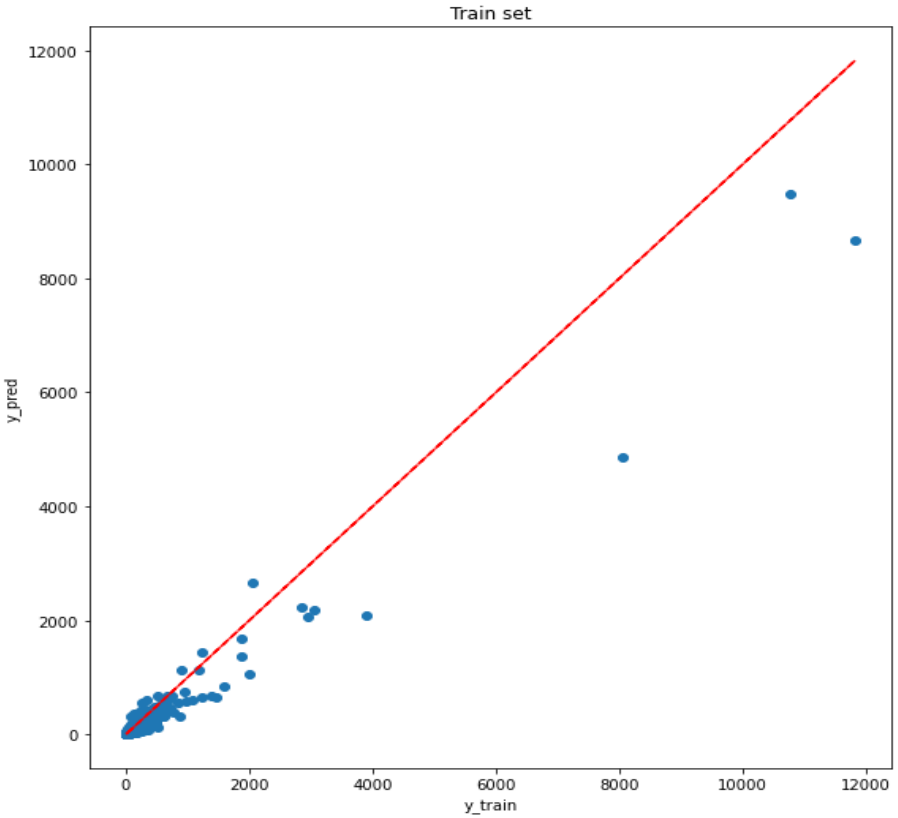
Choix du modèle final : Gradient Boosting

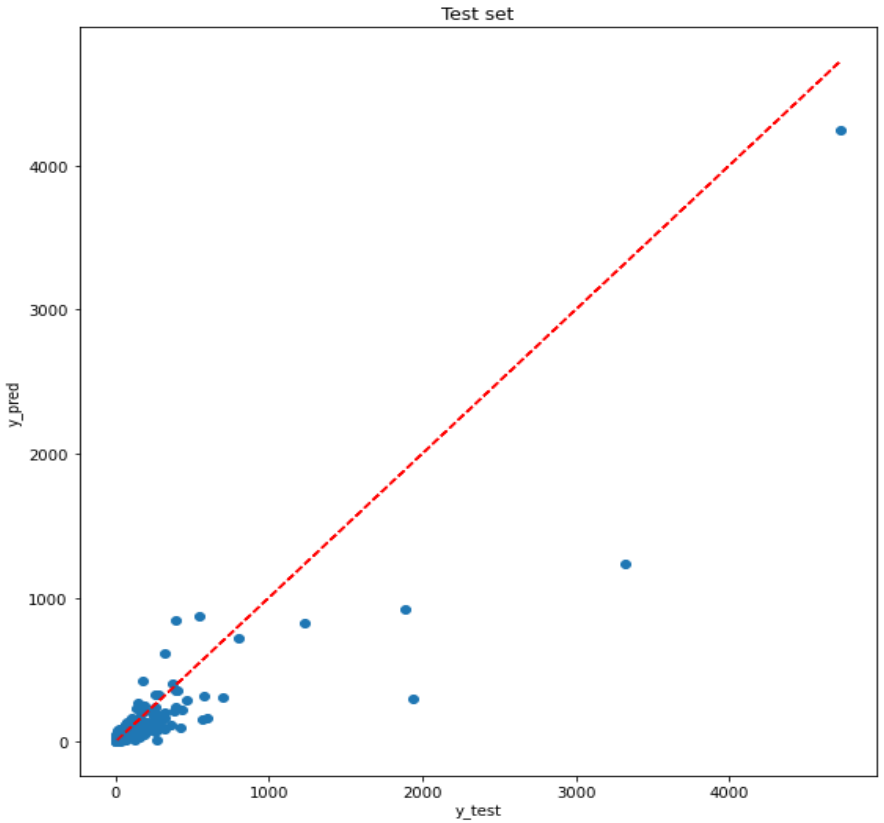
Modélisation GHGEmissions avec ENERGYSTARScore :

```
Train set score = 0.81
Test set score  = 0.62
```

(+ 9 points)

```
#data2016
Test set score  = 0.76
```





Synthèse

Synthèse

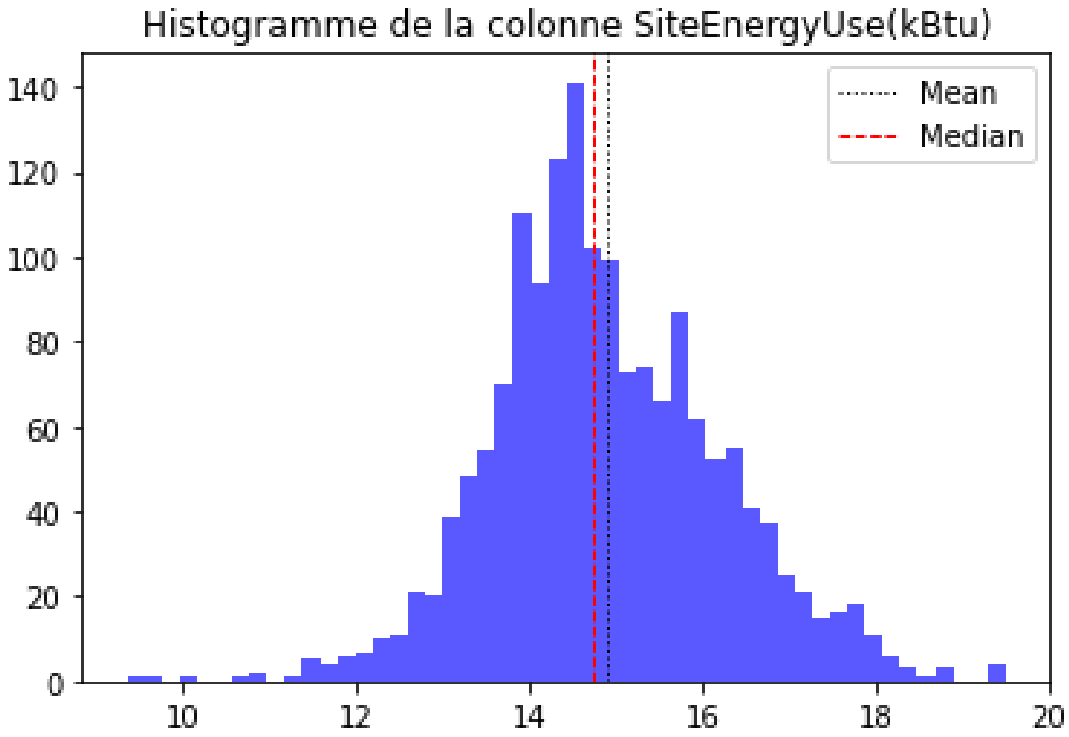
- **Nettoyage/Analyse puis Transformations des features :**
  - log(x + 1)
  - standardisation
  - OneHotEncoder
- **Choix du modèle pour chaque cible**
  - Recherche sur grille avec validation croisée
  - Gradient Boosting présente le meilleur score RMSE
- **Optimisation du gradient boosting**
  - Recherche sur grille avec validation croisée
  - Prédiction SiteEnergyUse → R^2 = 74 % sur le jeu de test
  - Prédiction GHGEmissions → R^2 = 53% sur le jeu de test
- **ENERGYSTARScore augmente la précision des modèles**
  - Prédiction SiteEnergyUse → R^2 = 87 % sur le jeu de test
  - Prédiction GHGEmissions → R^2 = 62% sur le jeu de test

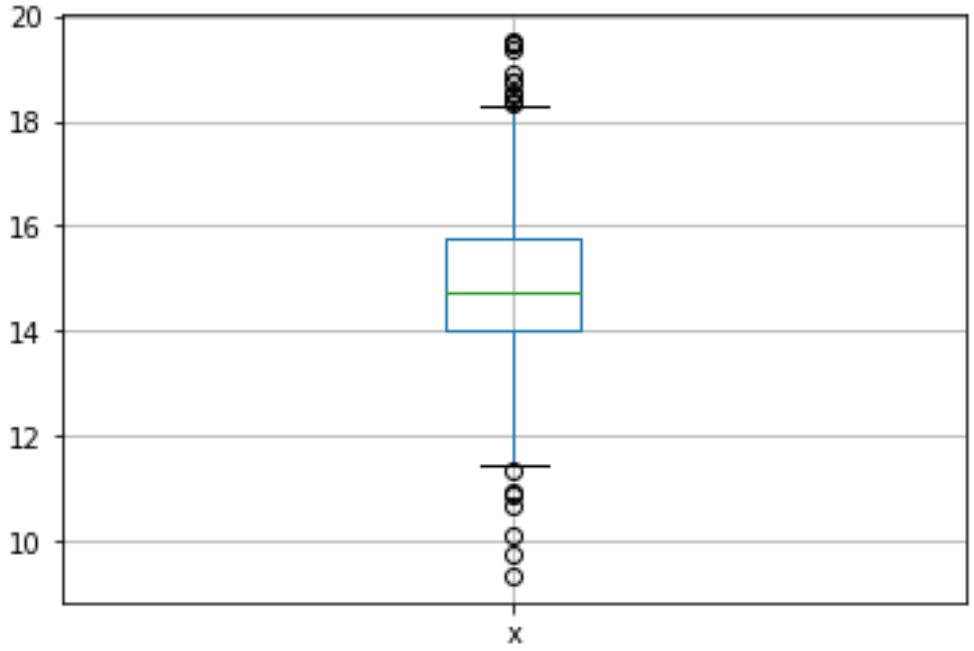
Merci de votre attention

Annexes

Annexes

Analyse univariée de SiteEnergyUse (log)

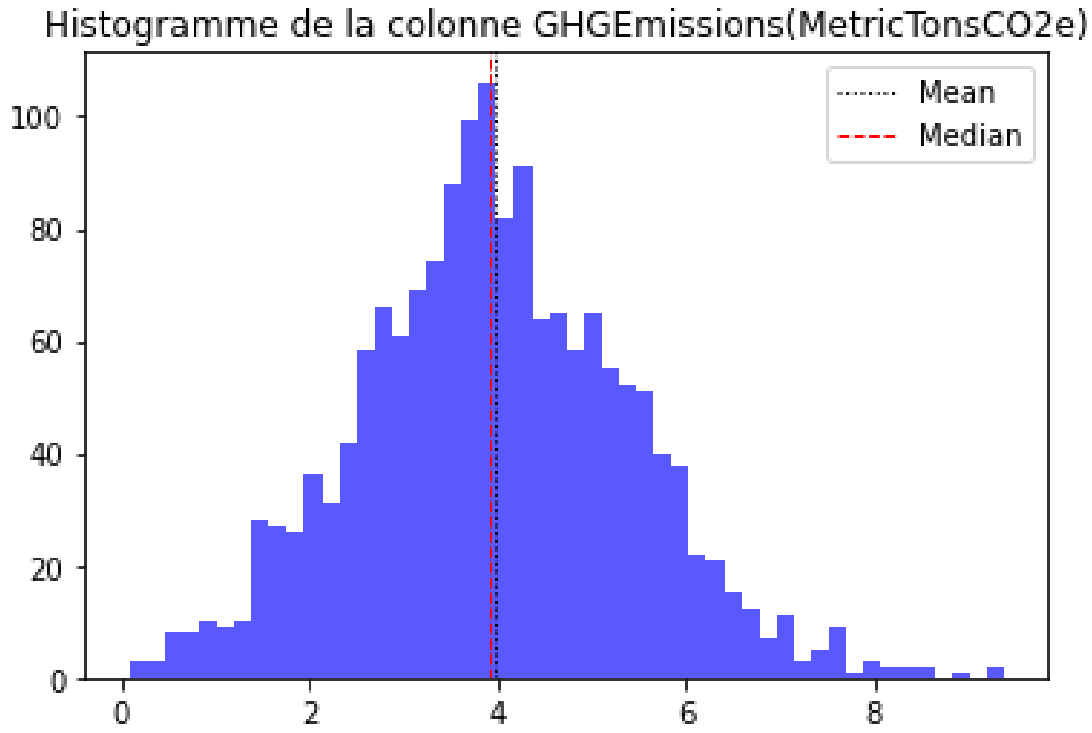


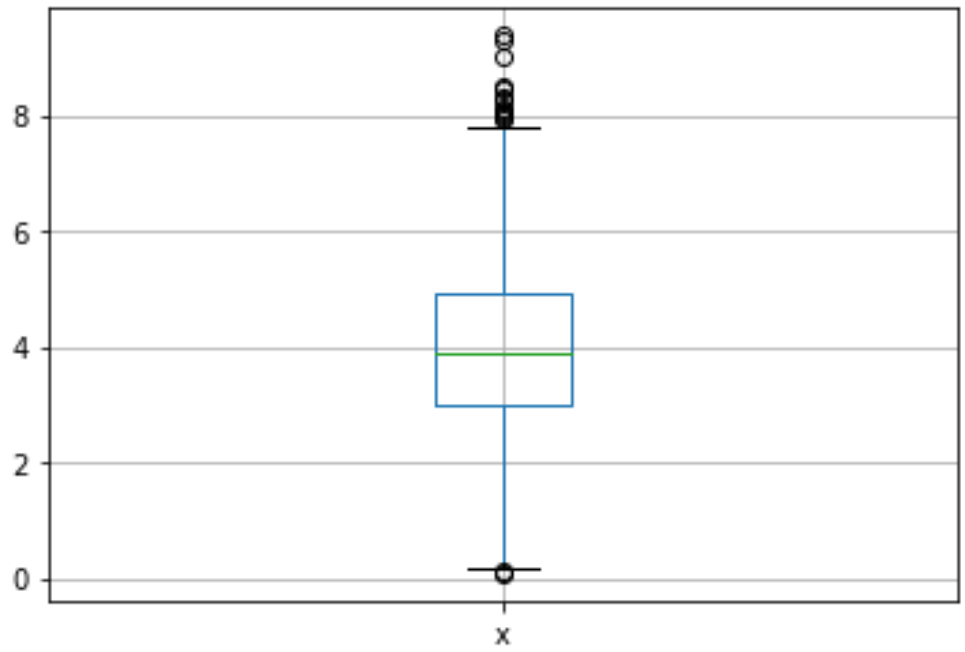


mean:	14.889	var:	1.751	skewness:	0.200
median:	14.722	ect:	1.323	kurtosis:	0.530

Annexes

Analyse univariée GHGEmissions (log)

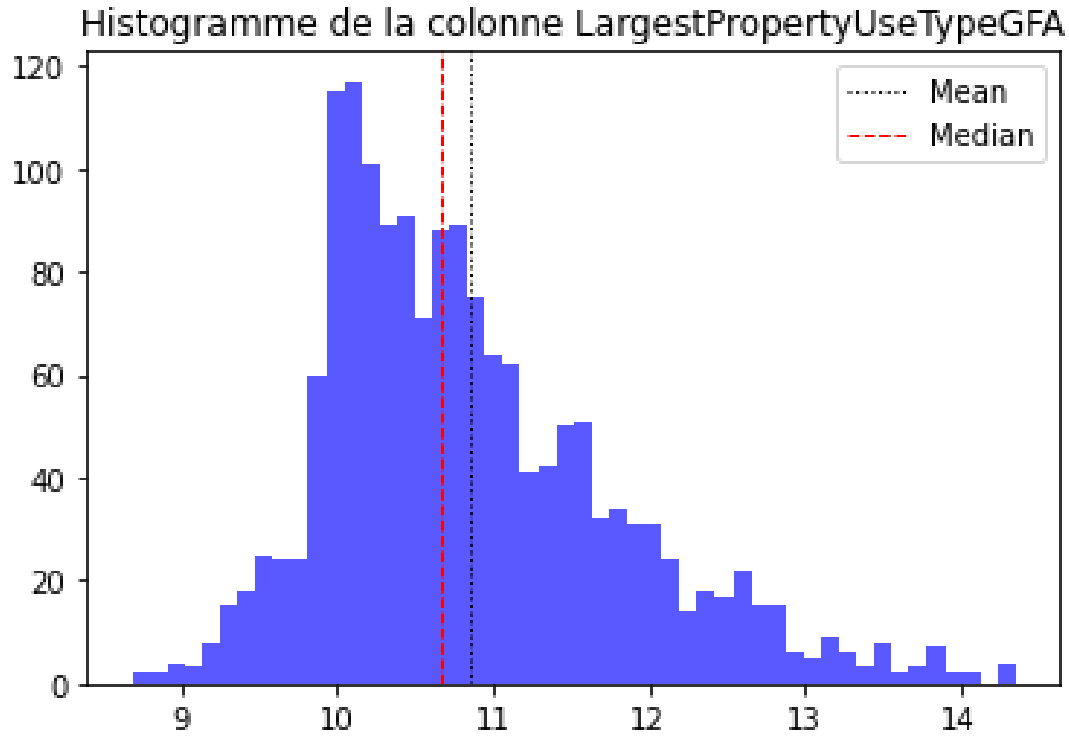




mean:	3.968	var:	2.086	skewness:	0.180
median:	3.904	ect:	1.444	kurtosis:	0.180

Annexes

Analyse univariée LargestPropertyUseTypeGFA (log)



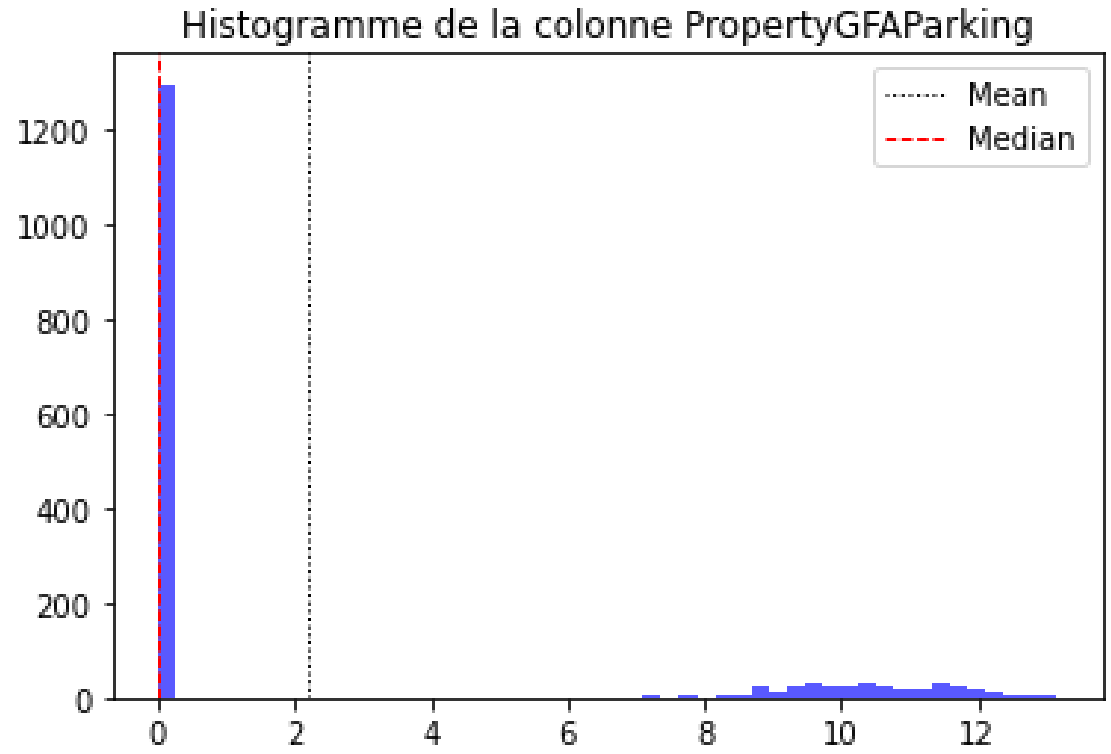

```
mean:    10.853
median:  10.674
var:     0.900
ect:     0.948

Skewness de la colonne [LargestPropertyUseTypeGFA]:
La distribution est etalee a droite.
skewness: 0.890

Kurtosis de la colonne [LargestPropertyUseTypeGFA]:
Les observations sont plus concentrées :
kurtosis: 0.710
```

Annexes

Analyse univariée de PropertyGFAParking (log)



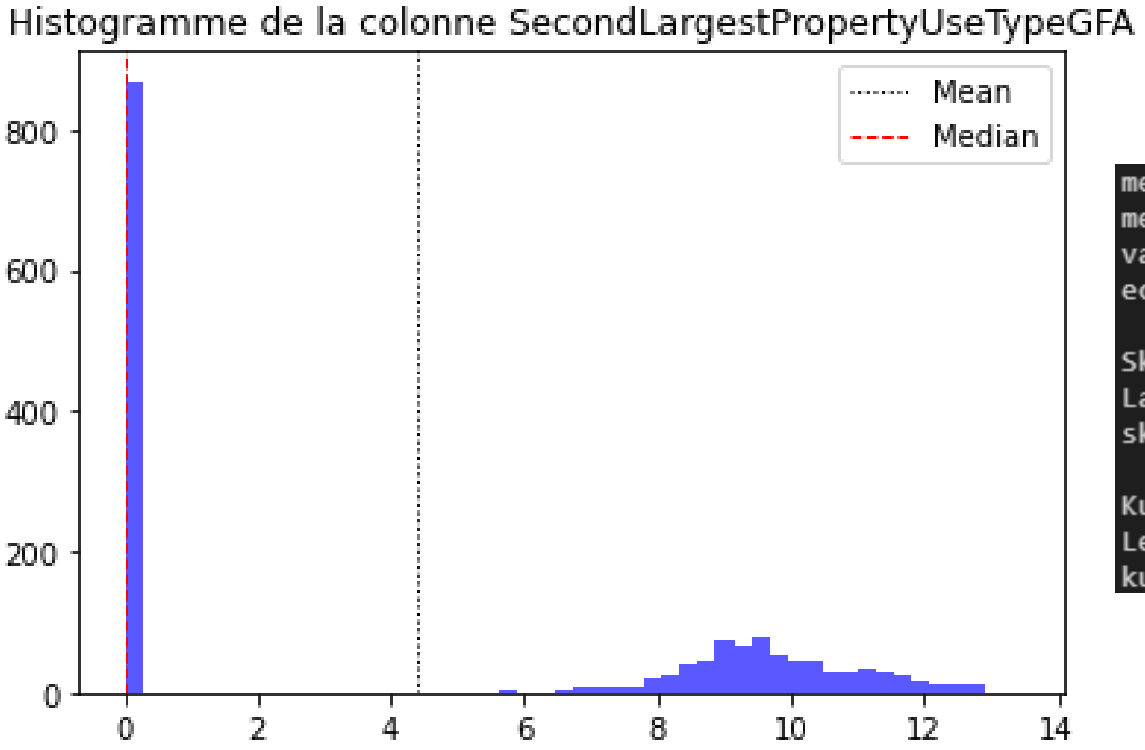

```
mean:    2.198
median:  0.000
var:     18.482
ect:     4.299

Skewness de la colonne [PropertyGFAParking]:
La distribution est etalee a droite.
skewness: 1.480

Kurtosis de la colonne [PropertyGFAParking]:
Les observations sont plus concentrées :
kurtosis: 0.280
```

Annexes

Analyse univariée SecondLargestPropertyUseTypeGFA (log)



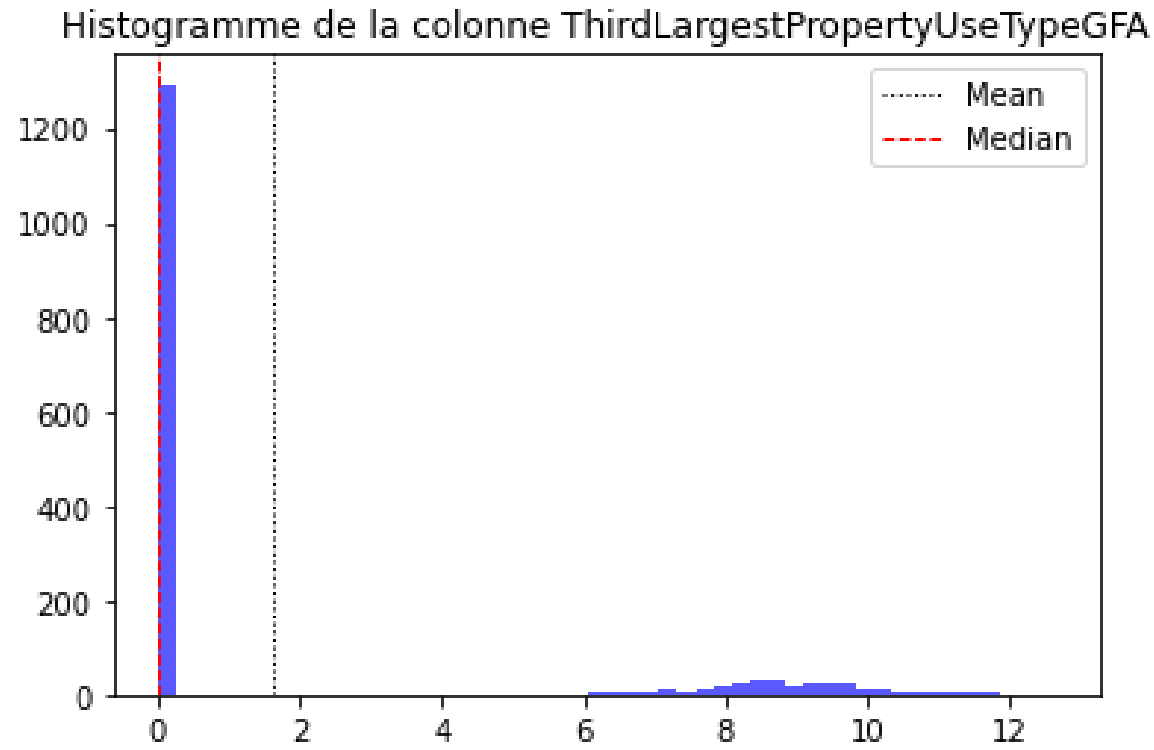

```
mean:    4.395
median:  0.000
var:     24.238
ect:     4.923

Skewness de la colonne [SecondLargestPropertyUseTypeGFA]:
La distribution est etalee a droite.
skewness: 0.290

Kurtosis de la colonne [SecondLargestPropertyUseTypeGFA]:
Les observations sont moins concentrées :
kurtosis: -1.790
```

Annexes

Analyse univariée ThirdLargestPropertyUseTypeGFA (log)



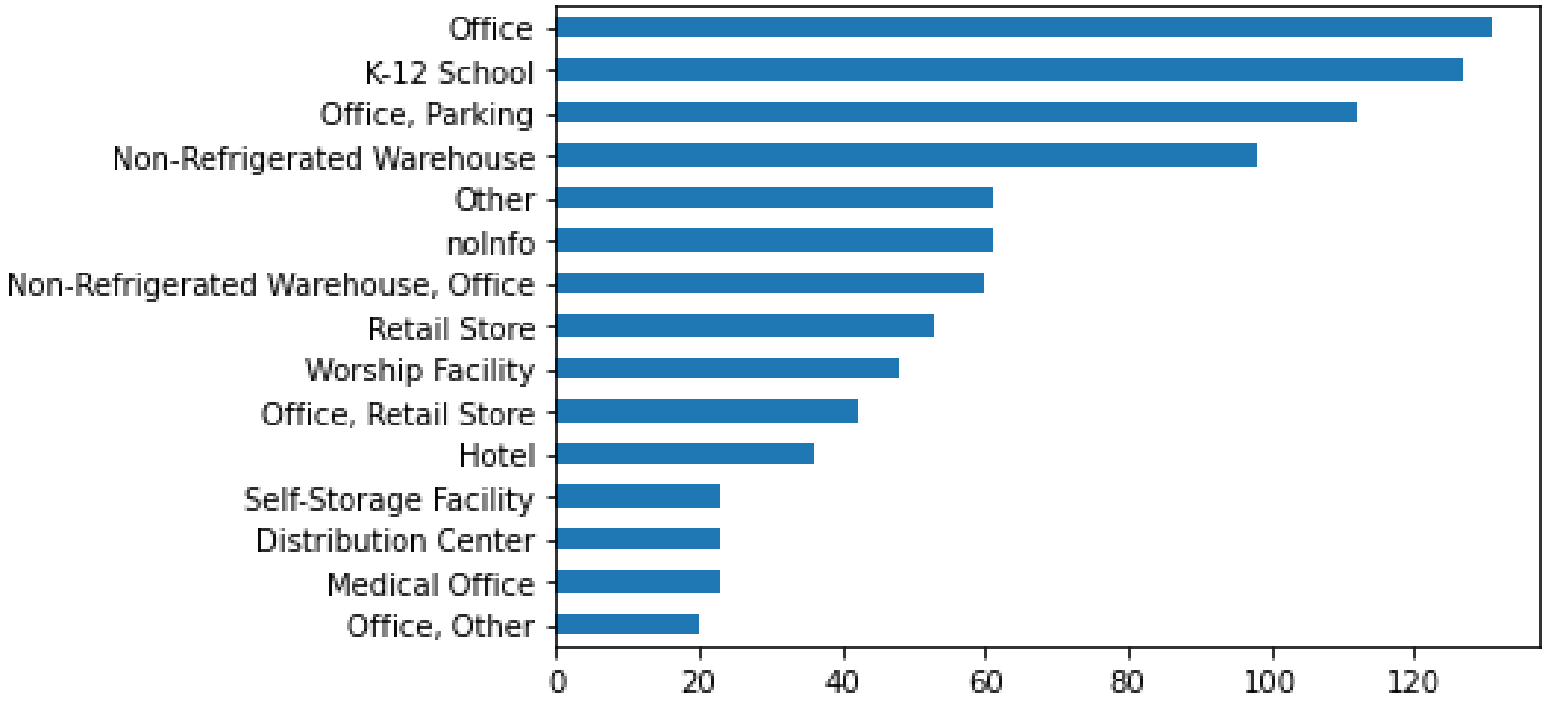

```
mean:    1.604
median:  0.000
var:     11.772
ect:     3.431

Skewness de la colonne [ThirdLargestPropertyUseTypeGFA]:
La distribution est etalee a droite.
skewness: 1.730

Kurtosis de la colonne [ThirdLargestPropertyUseTypeGFA]:
Les observations sont plus concentrées :
kurtosis: 1.130
```

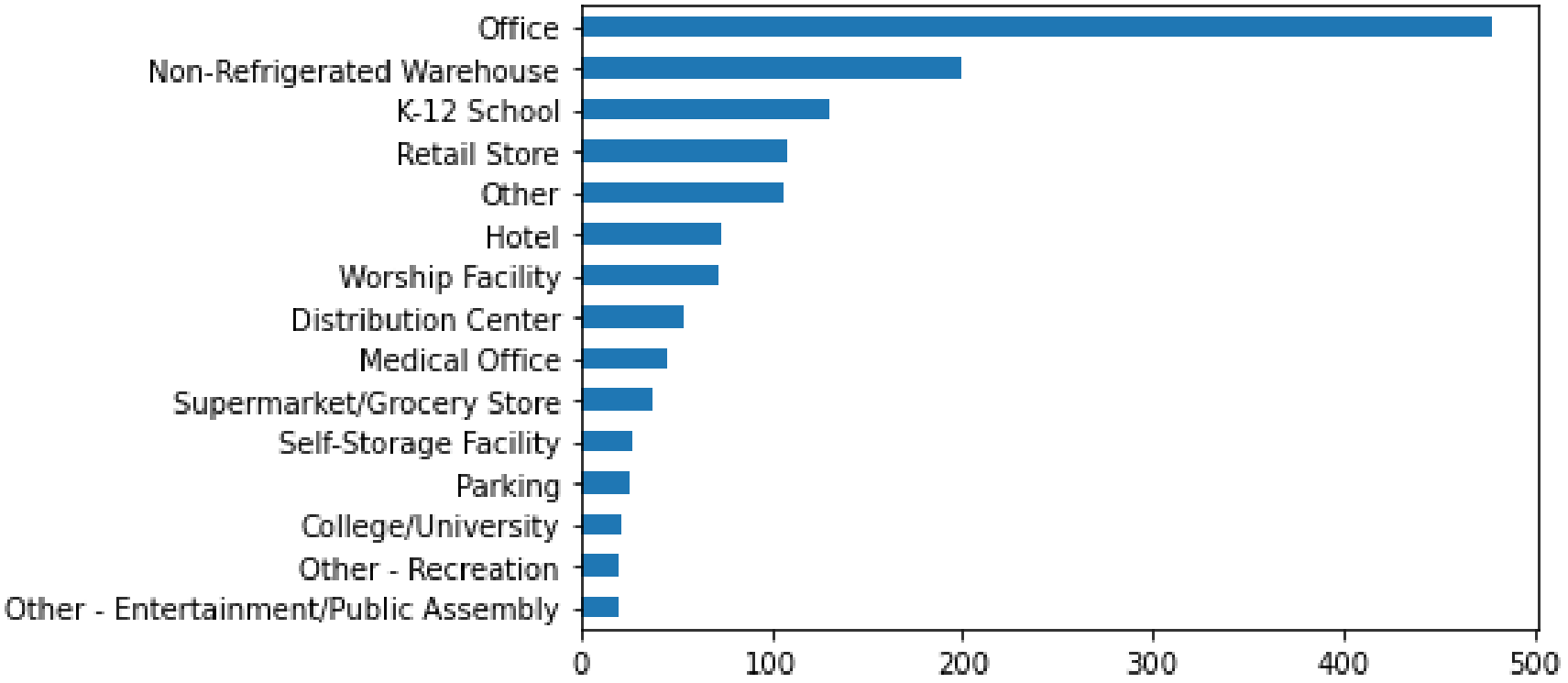
Annexes

Analyse univariée de ListOfAllPropertyUseTypes (top 15)



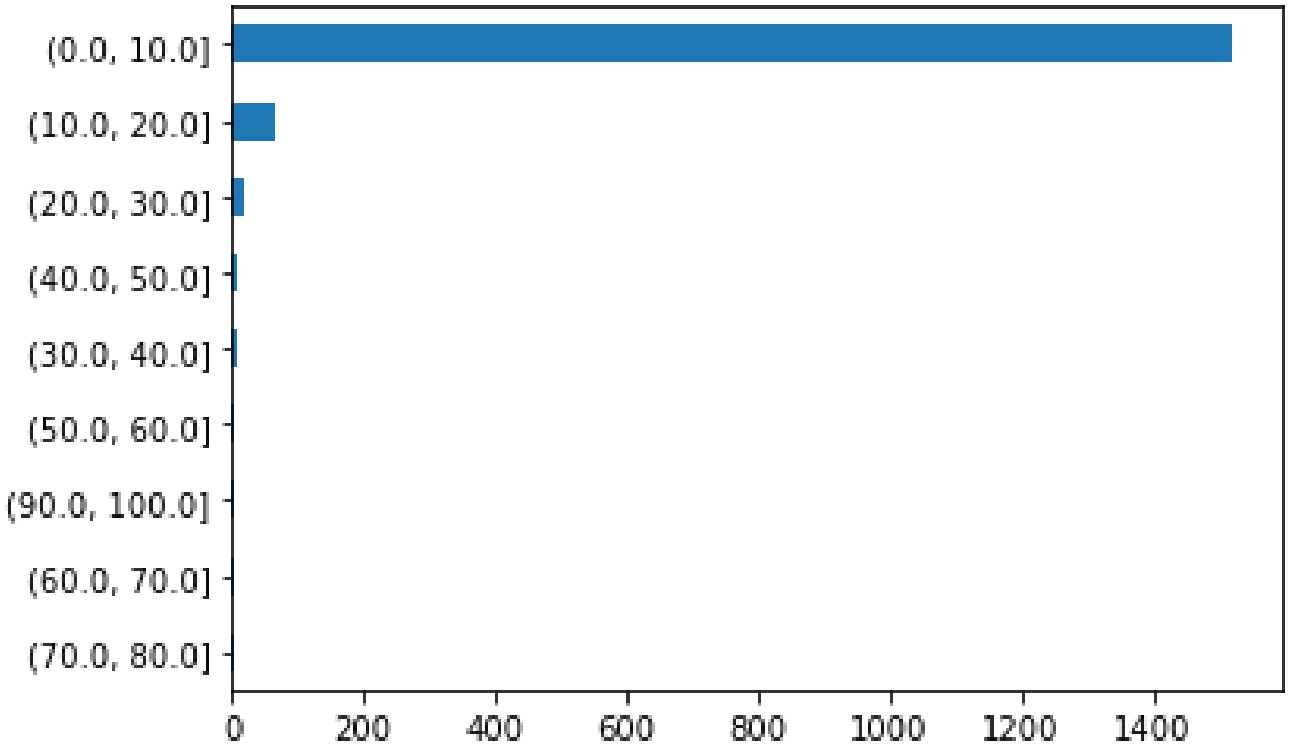
Annexes

Analyse univariée de LargestPropertyUseTypes (top 15)



Annexes

Analyse univariée de NumberofFloors



Annexes

Calcul R^2 :

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ – *predicted value of* y

$\bar{y}$ – *mean value of* y

Annexes

Calcul MSE et RMSE :

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

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

Where,

$\hat{y}$ – *predicted value of* y

$\bar{y}$ – *mean value of* y

Annexes

Calcul MAE :

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

Where,

$\hat{y}$ – *predicted value of* y

$\bar{y}$ – *mean value of* y

Annexes

Calcul R^2 ajusté :

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where n is number of observations in sample and p is number of independent variables in model