Anticipez les besoins en consommation de bâtiments

Noura Romari – Parcours Ingénieur Machine Learning depuis le 04/03/22
Soutenance Projet n°3 du 24/06/22

Mentor : Hamza Tajmouati
Evaluateur : Amosse Edouard

Sommaire

Introduction

Introduction

- **Contexte**

  - A Seattle, les bâtiments sont responsables de **33%** des émissions de gaz à effet de serre.
  - La ville a mis en place un plan d'action pour le climat depuis 2013 pour réduire ces émissions:
    - **Engagement fort** : zéro émission de gaz à effet de serre d'ici 2050
    - Mise en place d'un **programme d'analyse comparative** des performances énergétiques des bâtiments.
      → Tous les bâtiments de plus de 20 000 m^2
      → Remise d'un rapport annuel de faisant état de **l'énergie consommée** et des **gaz à effet de serre émis.**
      → Relevés indispensables pour la mise en place du plan d'action, mais très **couteux**.

- **Objectifs**

  - Mettre en place **des modèles prédictifs** de l'énergie consommée et des gaz à effet de serre émis pour les bâtiments non résidentiels, en se basant sur les informations factuelles disponibles.
  - Se passer des relevés couteux.
  - Evaluer l'intérêt de **l'Energy Star Score** pour la mise en place de ces modèles.

Introduction

- **Les données**

➢ Utiliser les données déjà relevées (**2015** et **2016**) et disponibles sur le site de la ville .

```
print(data_2015.shape)
print(data_2016.shape)
```

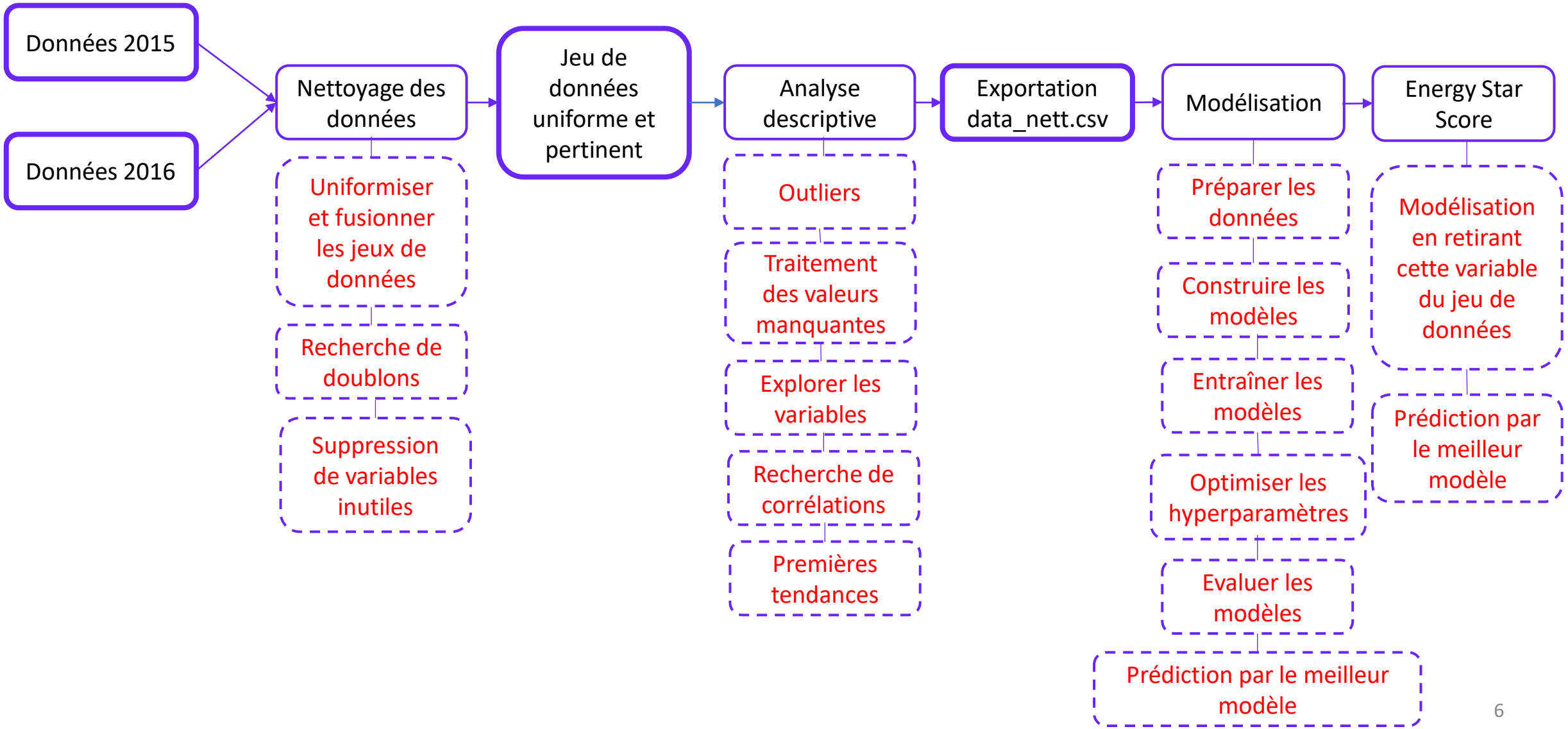
```
(3340, 47)
(3376, 46)
```

- Targets à prédire
- L'Energy Star Score
- Données factuelles
- Données de relevés

```
Data columns (total 47 columns):
 #   Column
---  ------
 0   OSEBuildingID
 1   DataYear
 2   BuildingType
 3   PrimaryPropertyType
 4   PropertyName
 5   TaxParcelIdentificationNumber
 6   Location
 7   CouncilDistrictCode
 8   Neighborhood
 9   YearBuilt
 10  NumberofBuildings
 11  NumberofFloors
 12  PropertyGFATotal
 13  PropertyGFAParking
 14  PropertyGFABuilding(s)
 15  ListOfAllPropertyUseTypes
 16  LargestPropertyUseType
 17  LargestPropertyUseTypeGFA
 18  SecondLargestPropertyUseType
 19  SecondLargestPropertyUseTypeGFA
 20  ThirdLargestPropertyUseType
 21  ThirdLargestPropertyUseTypeGFA
 22  YearsENERGYSTARCertified
 23  ENERGYSTARScore
 24  SiteEUI(kBtu/sf)
 25  SiteEUIWN(kBtu/sf)
 26  SourceEUI(kBtu/sf)
 27  SourceEUIWN(kBtu/sf)
 28  SiteEnergyUse(kBtu)
 29  SiteEnergyUseWN(kBtu)
 30  SteamUse(kBtu)
 31  Electricity(kWh)
 32  Electricity(kBtu)
 33  NaturalGas(therms)
 34  NaturalGas(kBtu)
 35  OtherFuelUse(kBtu)
 36  GHGEmissions(MetricTonsCO2e)
 37  GHGEmissionsIntensity(kgCO2e/ft2)
 38  DefaultData
 39  Comment
 40  ComplianceStatus
 41  Outlier
 42  2010 Census Tracts
 43  Seattle Police Department Micro Community Policing Plan Areas
 44  City Council Districts
 45  SPD Beats
 46  Zip Codes
```

Introduction

- **La stratégie globale**



Nettoyage des données

Nettoyage des données

- **Uniformiser les jeux de données**

```
for var in list(data_2016.columns) :
    if var not in list(data_2015.columns) :
        print(var)
```

```
Address
City
State
ZipCode
Latitude
Longitude
Comments
TotalGHGEmissions
GHGEmissionsIntensity
```

```
for var in list(data_2015.columns) :
    if var not in list(data_2016.columns) :
        print(var)
```

```
Location
OtherFuelUse(kBtu)
GHGEmissions(MetricTonsCO2e)
GHGEmissionsIntensity(kgCO2e/ft2)
Comment
2010 Census Tracts
Seattle Police Department Micro Community Policing Plan Areas
City Council Districts
SPD Beats
Zip Codes
```

```
data_2015['Location'][0]
```

```
: '{\'latitude\': \'47.61219025\', \'longitude\': \'-122.33799744\', \'human_address\': \'{"address": "405 OLIVE WAY", "city":
  "SEATTLE", "state": "WA", "zip": "98101"}\'}'
```

```
print(data_2015.shape)
data_2015 = data_2015.drop(['2010 Census Tracts', 'SPD Beats', 'City Council Districts',
                            'Seattle Police Department Micro Community Policing Plan Areas'], axis=1)
print(data_2015.shape)
```

Nettoyage des données

- **Recherche de doublons**

```
data_2015['OSEBuildingID'].describe()
```

```
count     3340
unique    3340
top          1
freq         1
Name: OSEBuildingID, dtype: object
```

```
data_2016['OSEBuildingID'].describe()
```

```
count     3376
unique    3376
top          1
freq         1
Name: OSEBuildingID, dtype: object
```

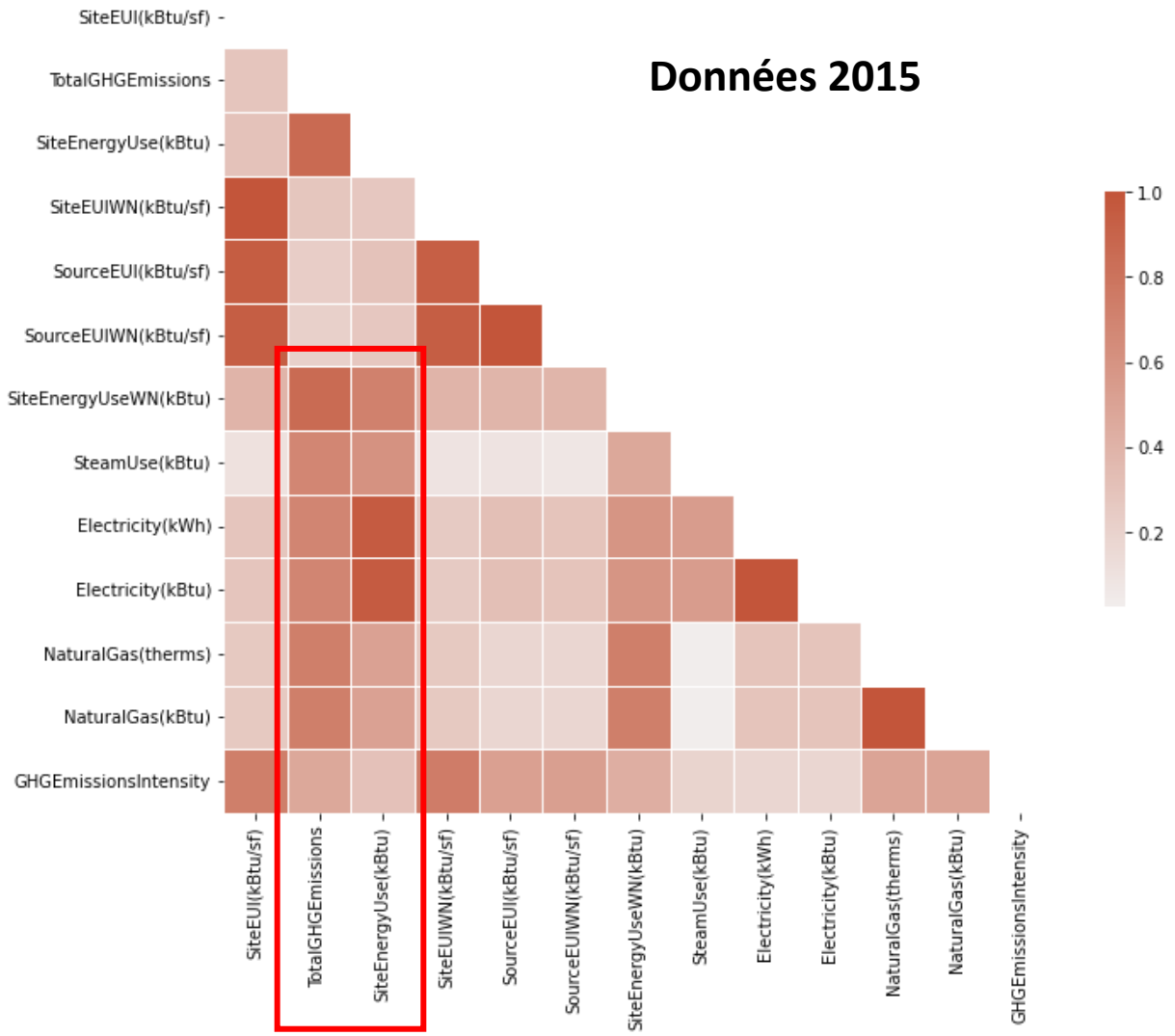
- **Suppression de variables (inutiles et données relevées)**

```
print(data_2015.shape)
data_2015 = data_2015.drop(['PropertyName', 'YearsENERGYSTARCertified', 'SiteEUI(kBtu/sf)',
                            'SiteEUIWN(kBtu/sf)','SourceEUI(kBtu/sf)', 'SourceEUIWN(kBtu/sf)',
                            'SiteEnergyUseWN(kBtu)', 'SteamUse(kBtu)', 'Electricity(kWh)',
                            'Electricity(kBtu)', 'NaturalGas(therms)', 'NaturalGas(kBtu)',
                            'OtherFuelUse(kBtu)', 'GHGEmissionsIntensity(kgCO2e/ft2)', 'DefaultData',
                           'Comment', 'ComplianceStatus'], axis=1)
print(data_2015.shape)
```

```
(3340, 48)
(3340, 31)
```

Nettoyage des données

- **Suppression des données des relevées**



- **Fusion des jeux de données**

```
print(data_2015.shape)
print(data_2016.shape)
print(data.shape)
```

```
(3340, 27)
(3376, 27)
(6716, 27)
```

Analyse descriptive

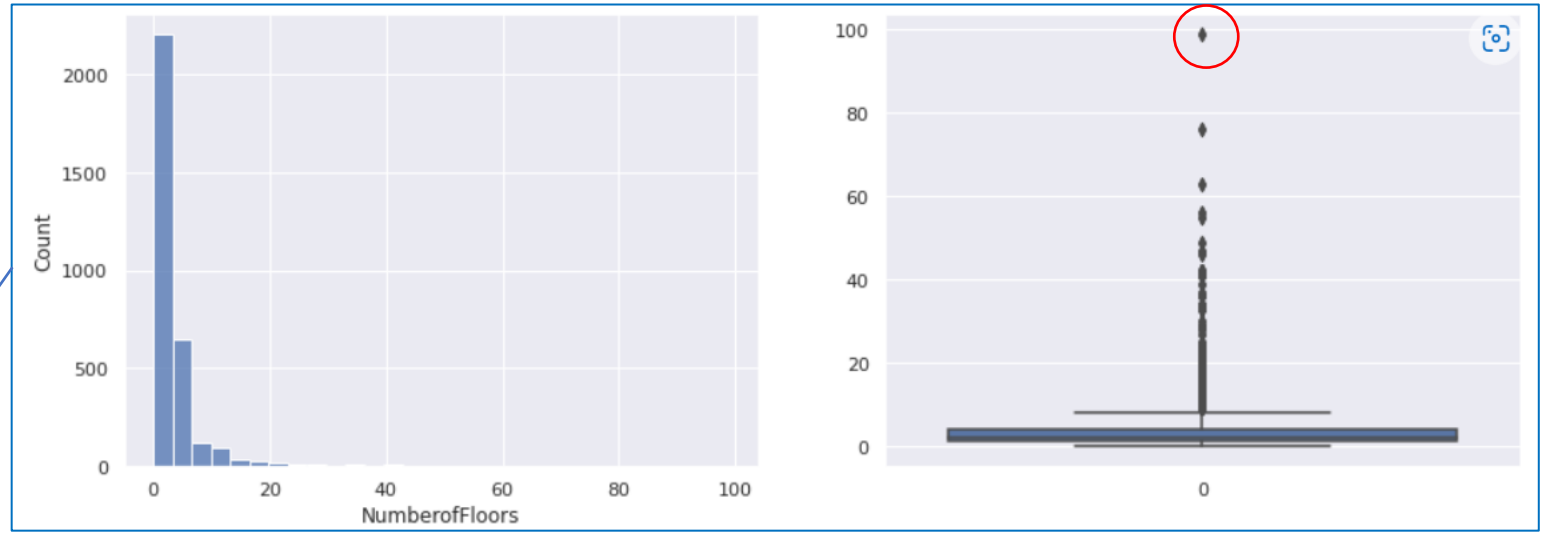
Analyse descriptive

- **Traitement des outliers**

```
count(data['Outlier'])

Counter({nan: 6600,
         'High Outlier': 46,
         'Low Outlier': 38,
         'High outlier': 9,
         'Low outlier': 23})
```

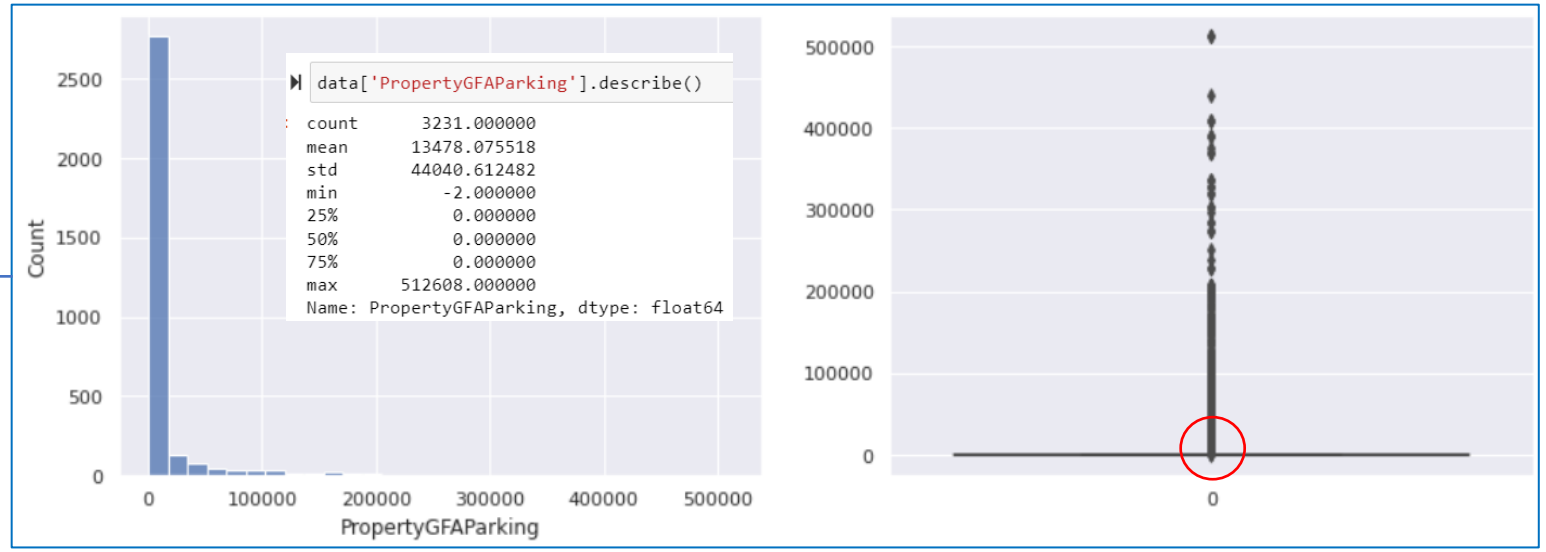
Supprimés



➢ Le plus grand gratte-ciel de Seattle est le Columbia center avec 76 étages.

Valeurs manquantes

KNN imputer



➢ Des variables à valeurs négatives aberrantes

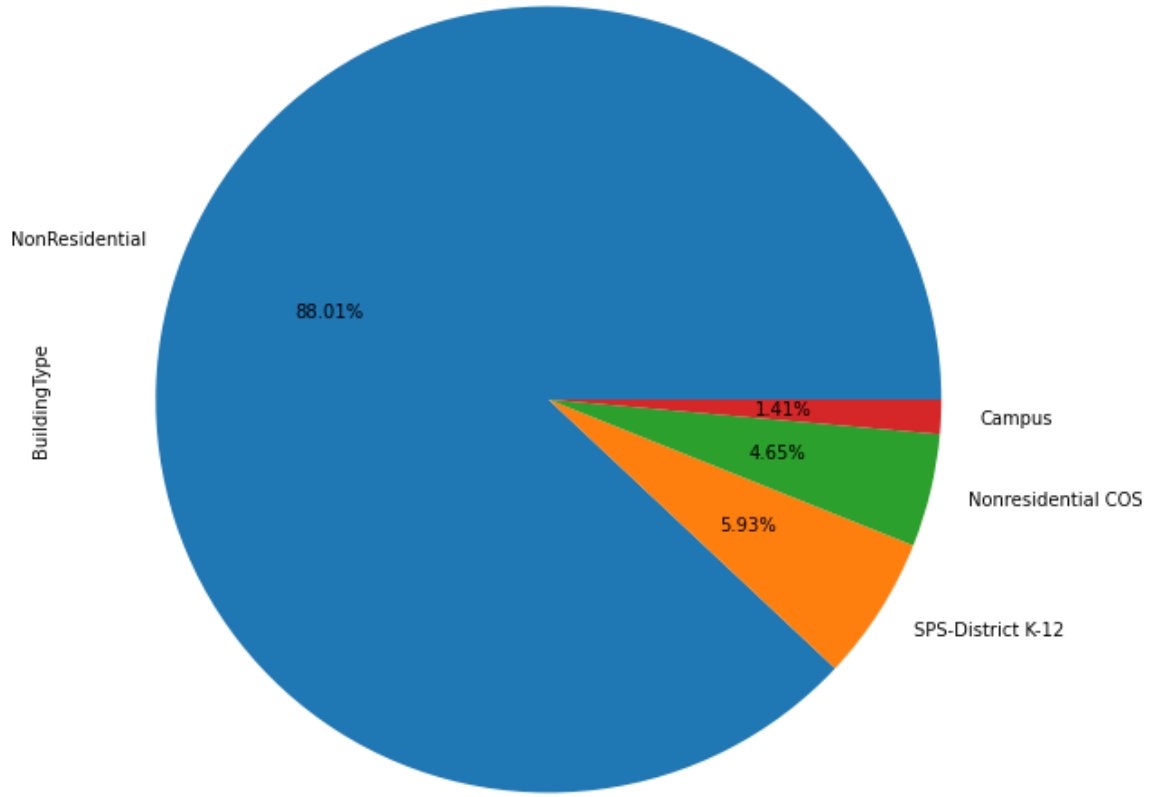
Analyse descriptive

- **Exploration des variables catégorielles**

  - Le type de bâtiment

```
data['BuildingType'].value_counts()
```

```
NonResidential          2877
Multifamily LR (1-4)    2002
Multifamily MR (5-9)    1115
Multifamily HR (10+)     213
SPS-District K-12        194
Nonresidential COS       152
Campus                    46
Nonresidential WA          1
Name: BuildingType, dtype: int64
```



Analyse descriptive

- **Exploration des variables catégorielles**

  - ➢ Les types d'activités (Principale, secondaire, tertiaire)

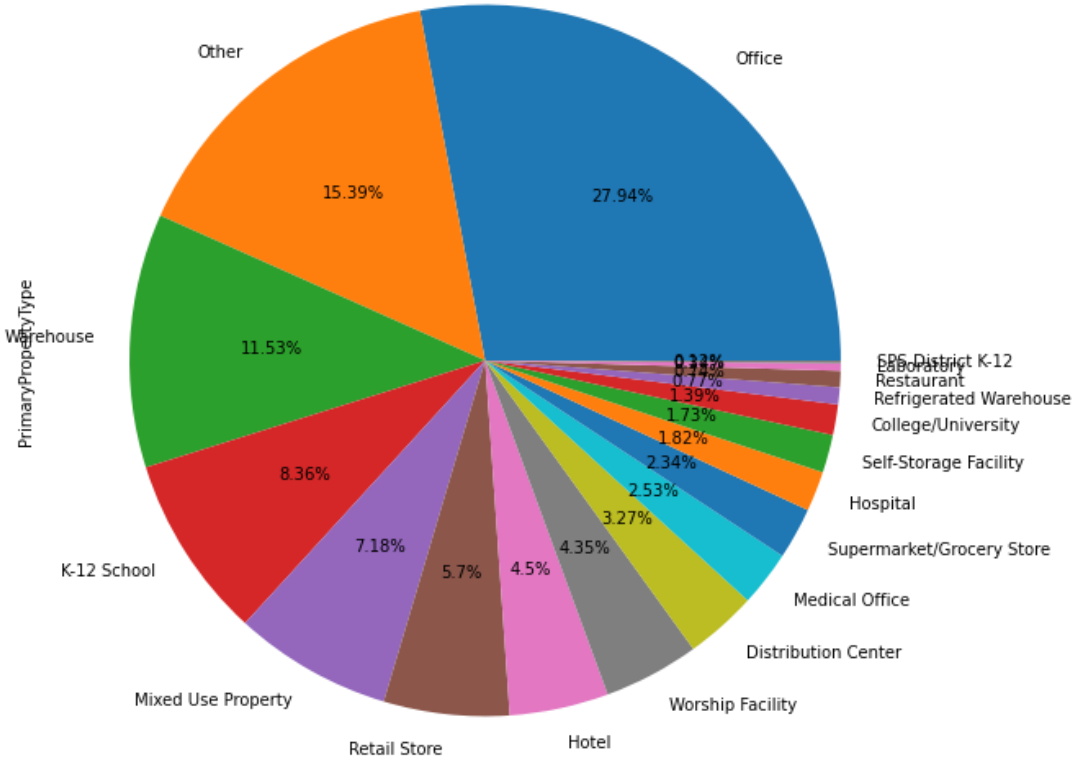
```
data['PrimaryPropertyType'].value_counts()
```

```
Small- and Mid-Sized Office    572
Other                          499
Large Office                   331
K-12 School                    271
Mixed Use Property             220
Non-Refrigerated Warehouse     187
Warehouse                      187
Retail Store                   185
Hotel                          146
Worship Facility               141
Medical Office                  82
Distribution Center             55
Distribution Center\n           51
Supermarket / Grocery Store     40
Residential                     40
Senior Care Community           39
Supermarket/Grocery Store       36
Self-Storage Facility           29
Self-Storage Facility\n         27
Refrigerated Warehouse          25
University                      24
College/University              21
Hospital                        20
Restaurant                      13
Laboratory                      11
Restaurant\n                    11
SPS-District K-12                4
Office                           3
Name: PrimaryPropertyType, dtype: int64
```

- Suppression des bâtiments résidentiels sauf si ils ont une activité secondaire et/ou tertiaire mixte.

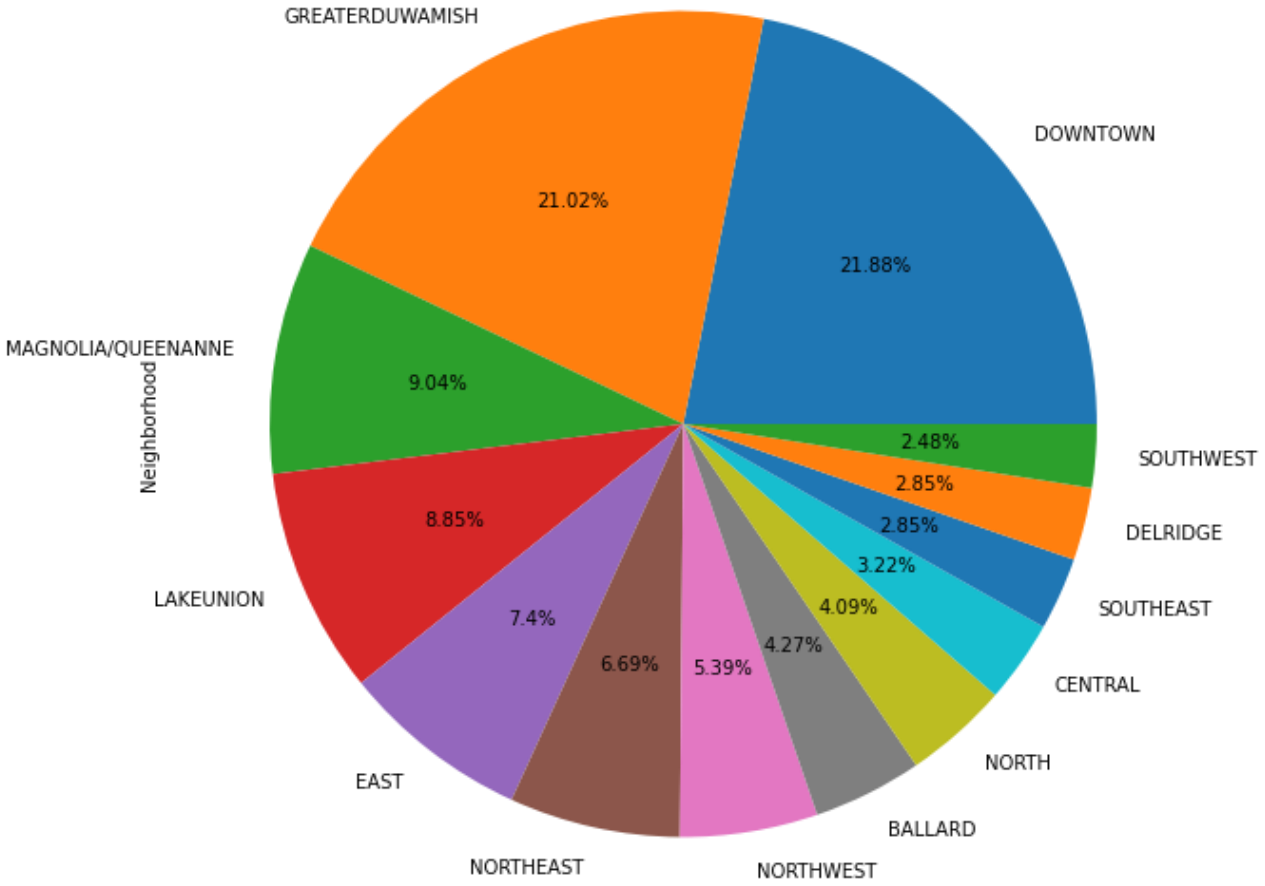
- Fusion de certaines classes afin de réduire le nombre de catégories

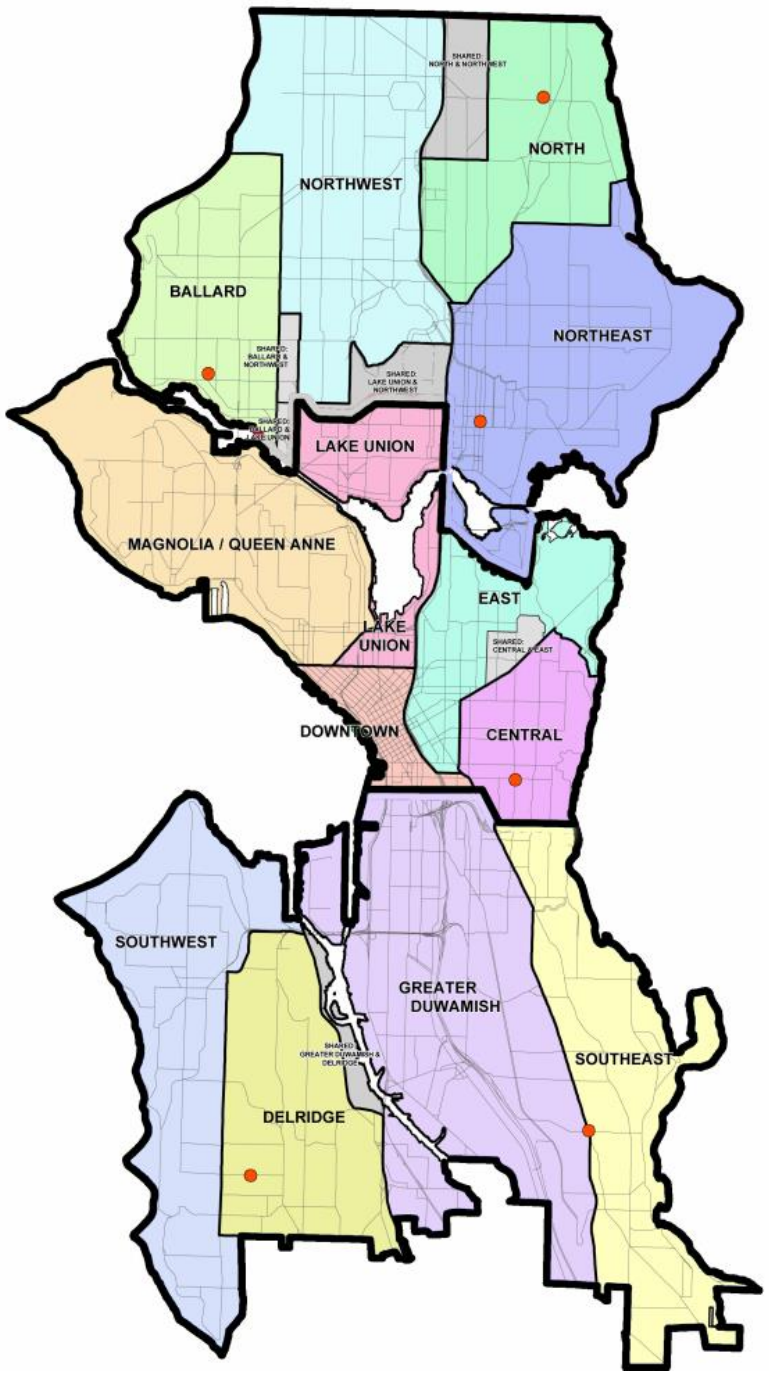
- Uniformisation des doublons



Analyse descriptive

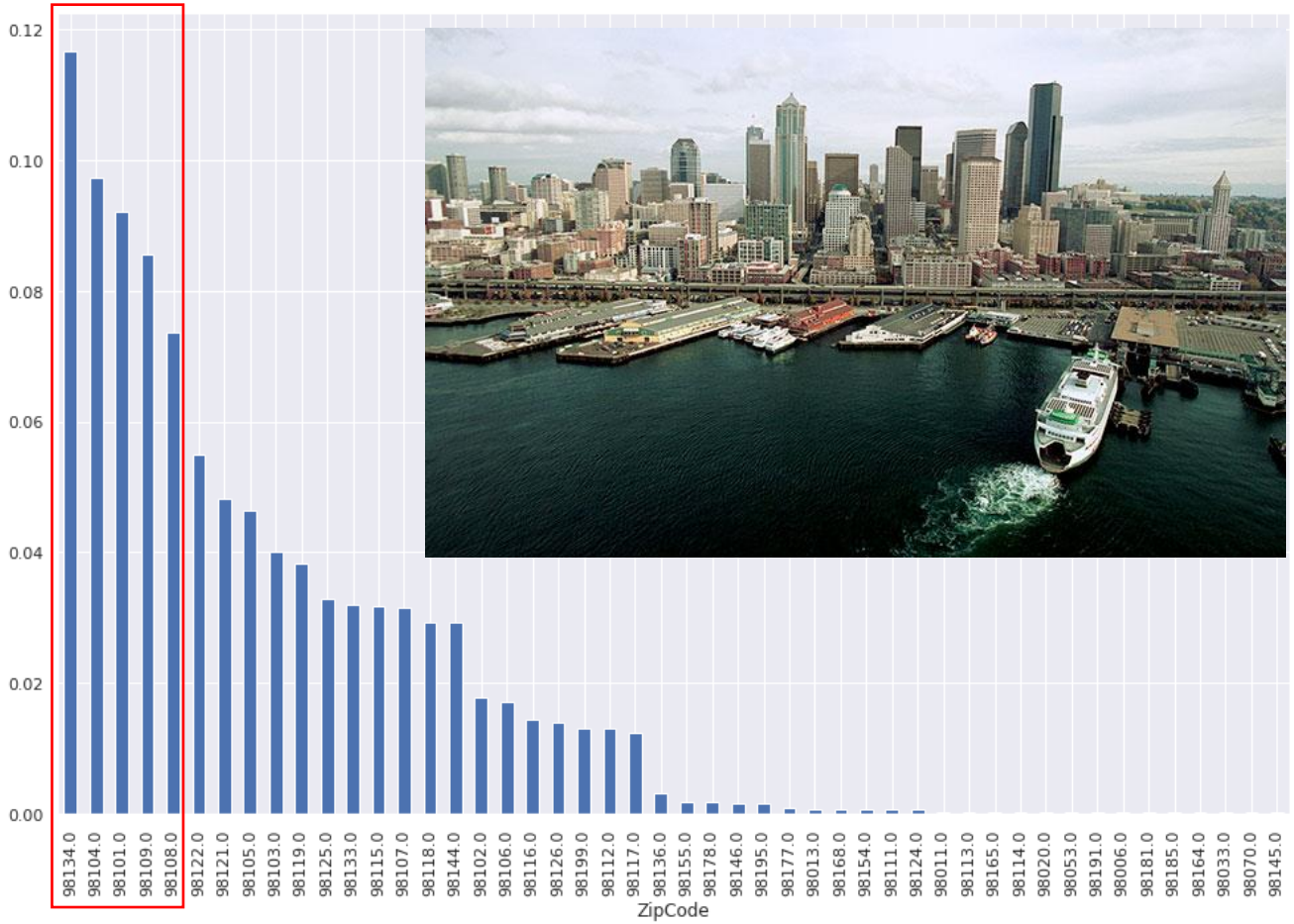
- **Exploration des variables catégorielles**
  - La répartition par quartier

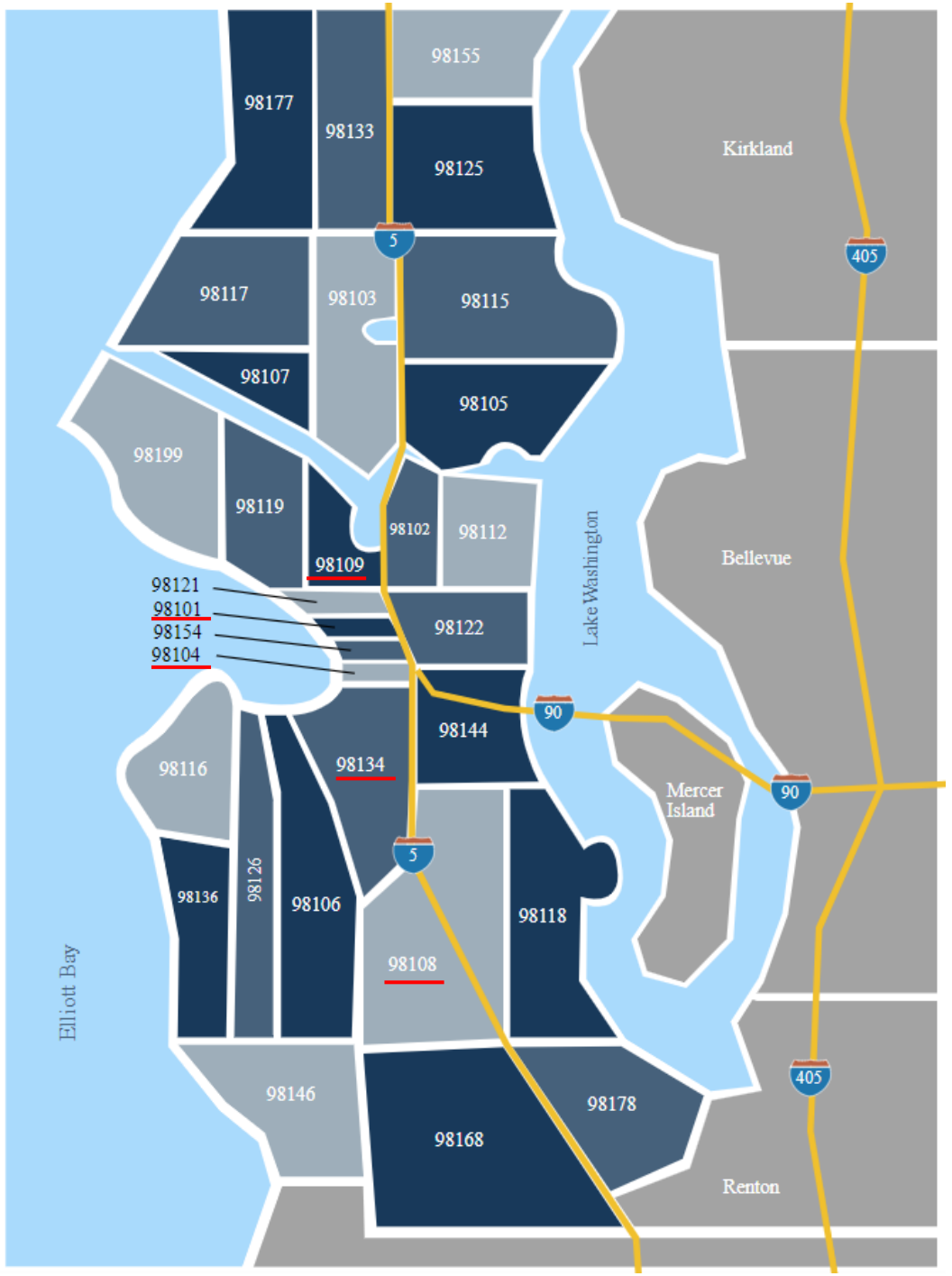




Analyse descriptive

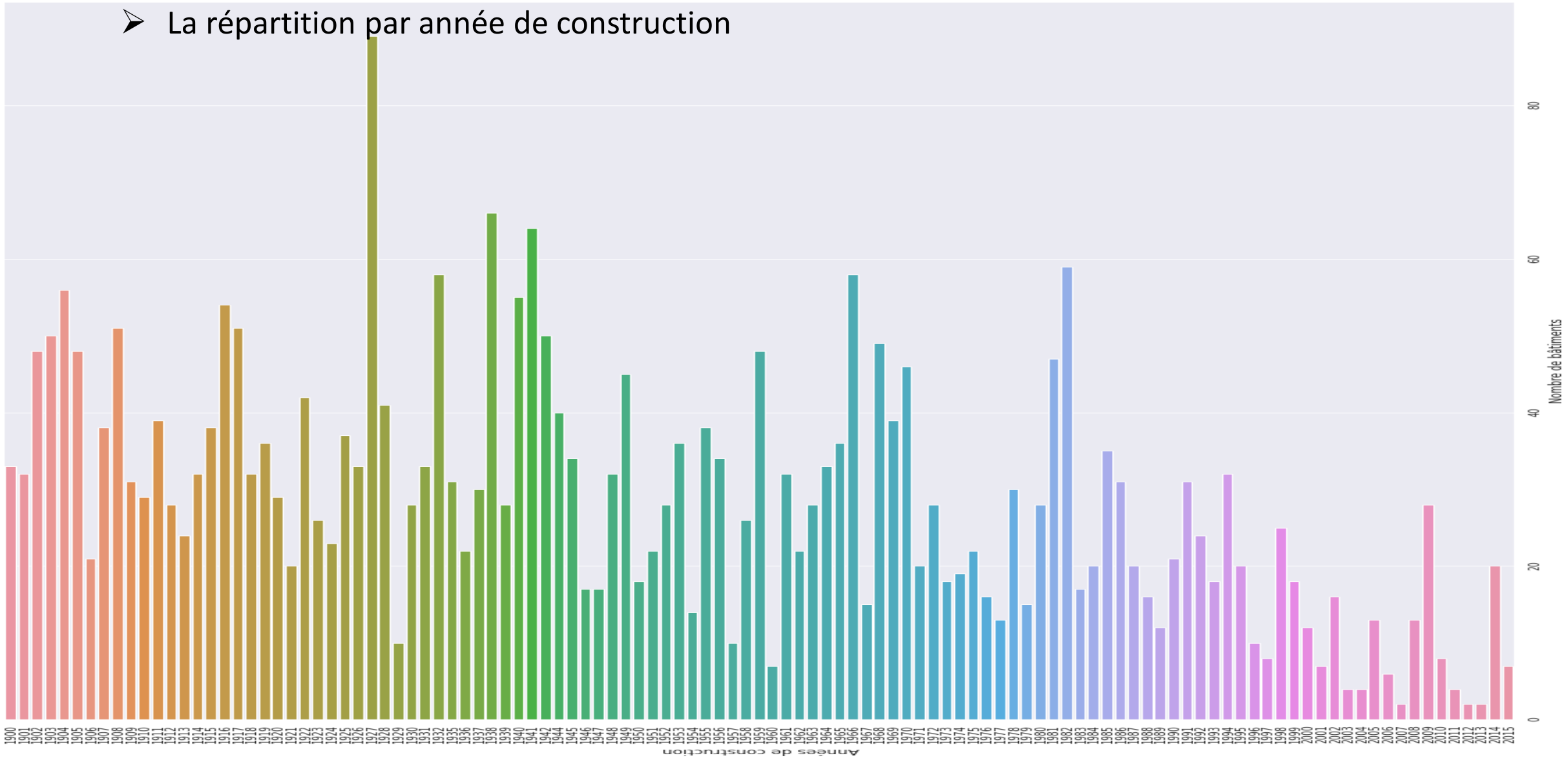
- **Exploration des variables catégorielles**
  - ➢ La répartition par quartier





Analyse descriptive

- **Exploration des variables catégorielles**
  - La répartition par année de construction

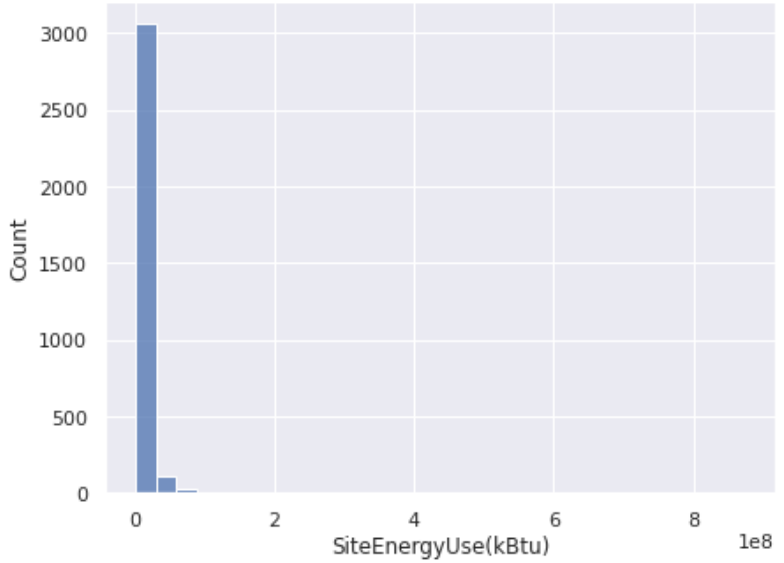


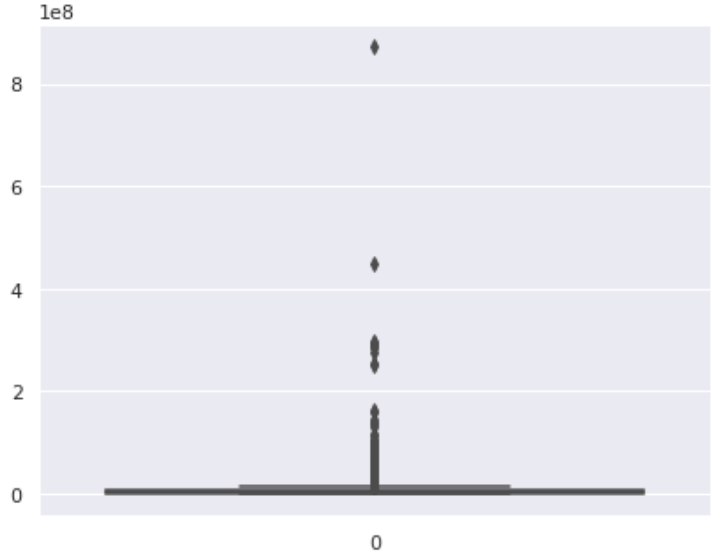
Analyse descriptive

- **Exploration des variables quantitatives**

```
data['SiteEnergyUse(kBtu)'].describe()
```

```
count    3.224000e+03
mean     8.139577e+06
std      2.553620e+07
min      0.000000e+00
25%      1.242800e+06
50%      2.554947e+06
75%      6.962459e+06
max      8.739237e+08
Name: SiteEnergyUse(kBtu), dtype: float64
```

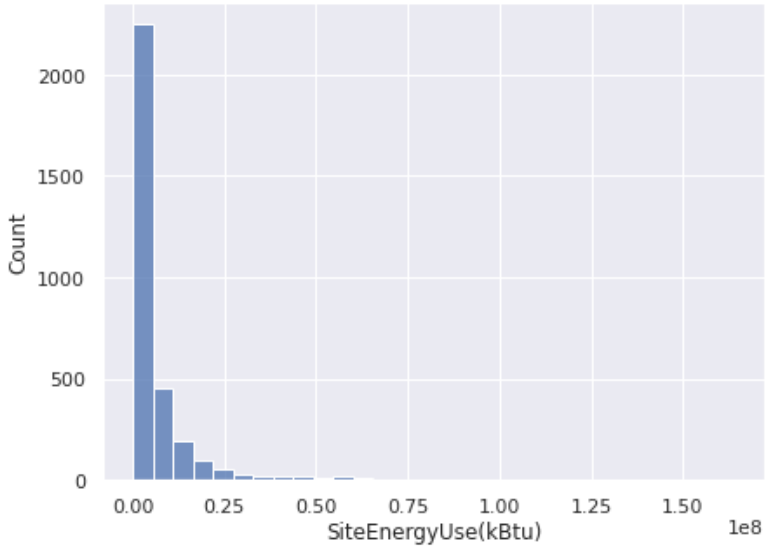


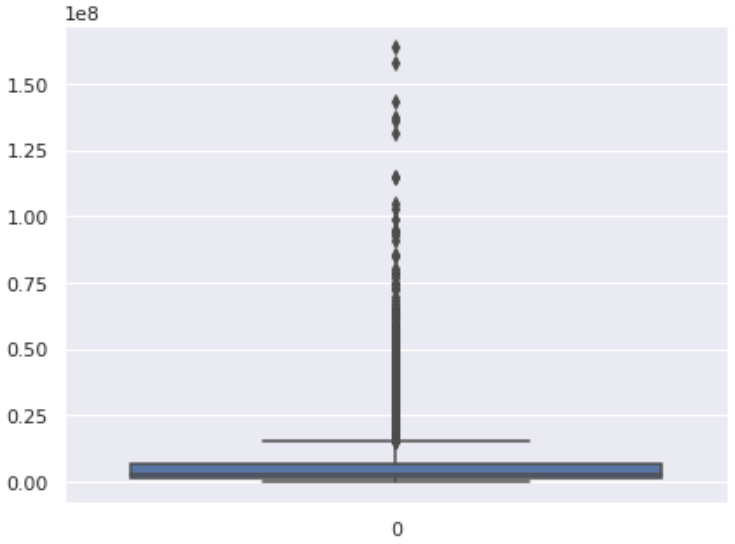


➢ Suppression des valeurs extrêmes susceptibles de perturber les modèles

```
data['SiteEnergyUse(kBtu)'].describe()
```

```
count    3.231000e+03
mean     7.201475e+06
std      1.362199e+07
min      0.000000e+00
25%      1.243047e+06
50%      2.553764e+06
75%      6.926601e+06
max      1.639460e+08
Name: SiteEnergyUse(kBtu), dtype: float64
```



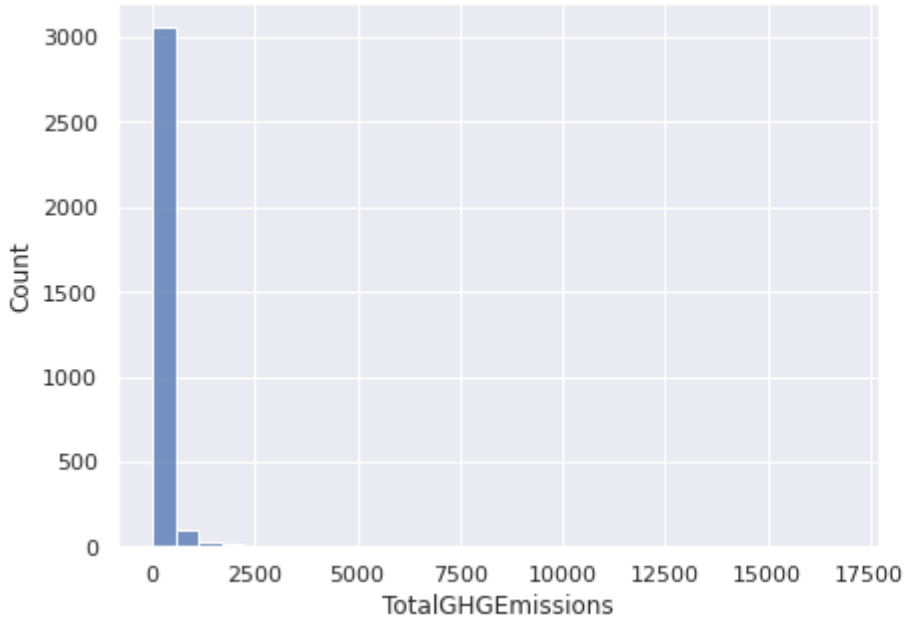


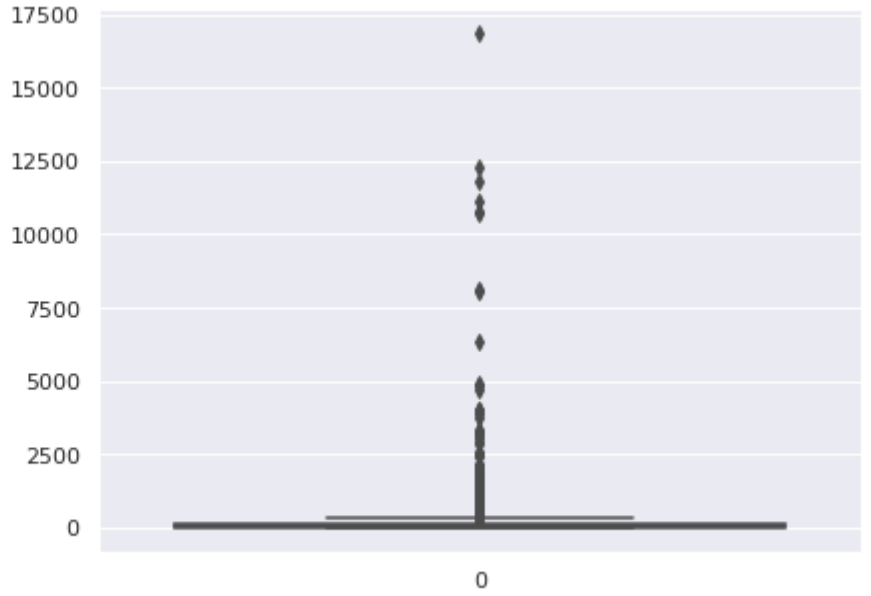
Analyse descriptive

- **Exploration des variables quantitatives**

```
data['TotalGHGEmissions'].describe()
```

```
count     3224.000000
mean       179.128886
std        674.026074
min         -0.800000
25%         20.040000
50%         49.365000
75%        139.350000
max      16870.980000
Name: TotalGHGEmissions, dtype: float64
```





Analyse descriptive

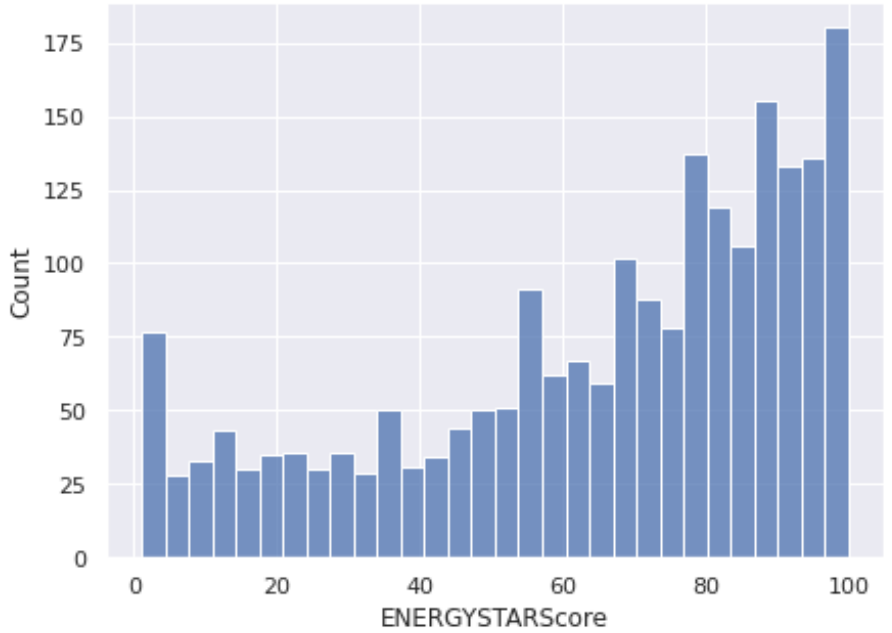
- **Exploration des variables quantitatives**

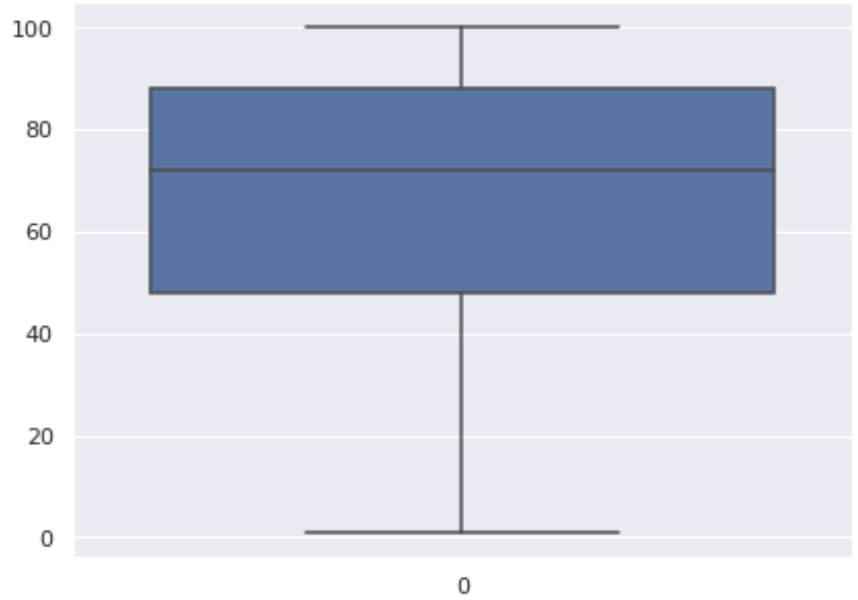
➢ **Définition :**

- Reflète la performance énergétique d'un bâtiment.
- Il tient compte de la morphologie du bâtiment, des activités qui y sont exercés, et du comportement de ses occupants. Il est fastidieux à calculer
- Il se situe entre 1 et 100 (100 = performance énergétique optimale).

```
data['ENERGYSTARScore'].describe()
```

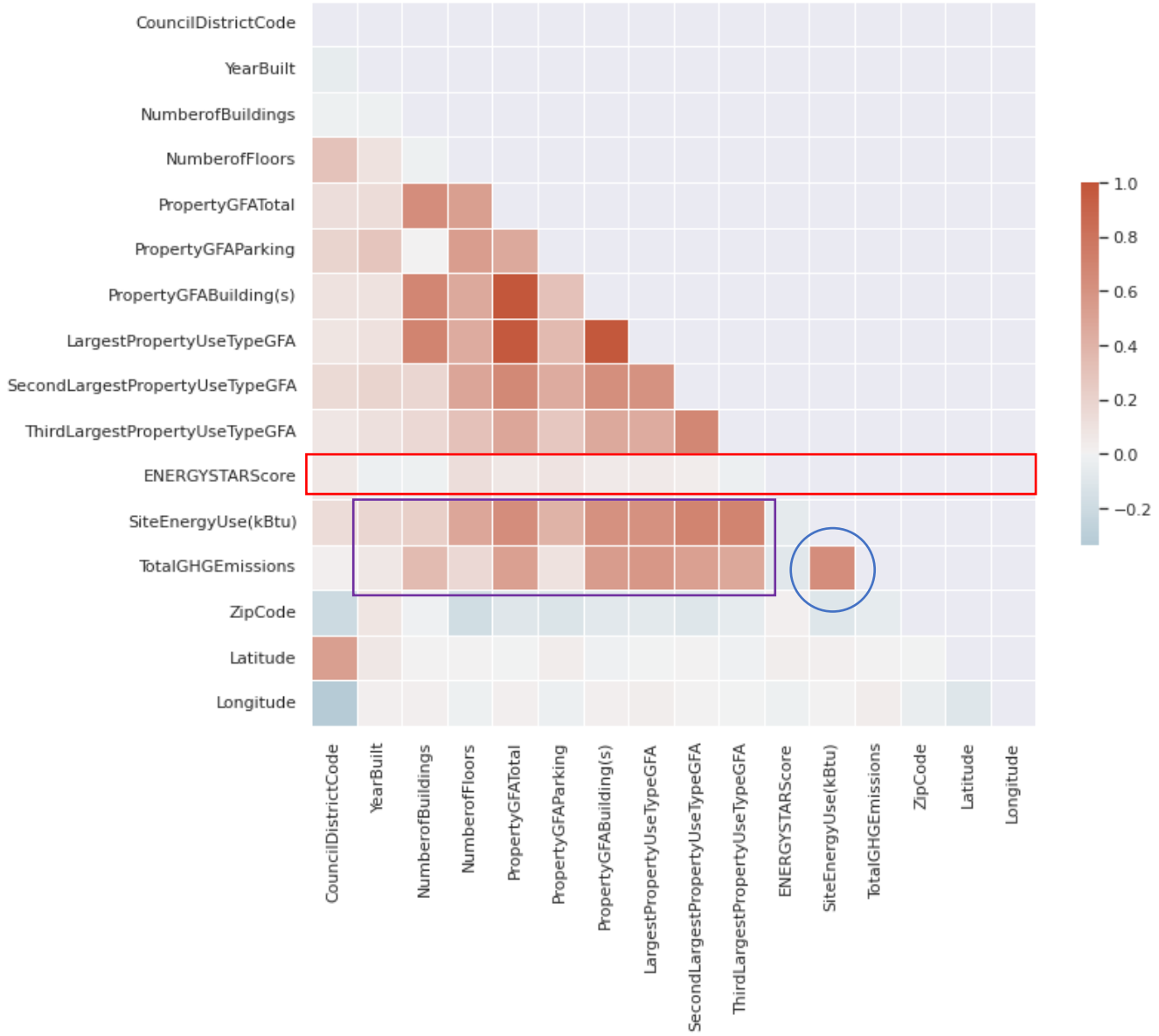
```
count    2150.000000
mean       64.821860
std        28.252524
min         1.000000
25%        48.000000
50%        72.000000
75%        88.000000
max       100.000000
Name: ENERGYSTARScore, dtype: float64
```





Analyse descriptive

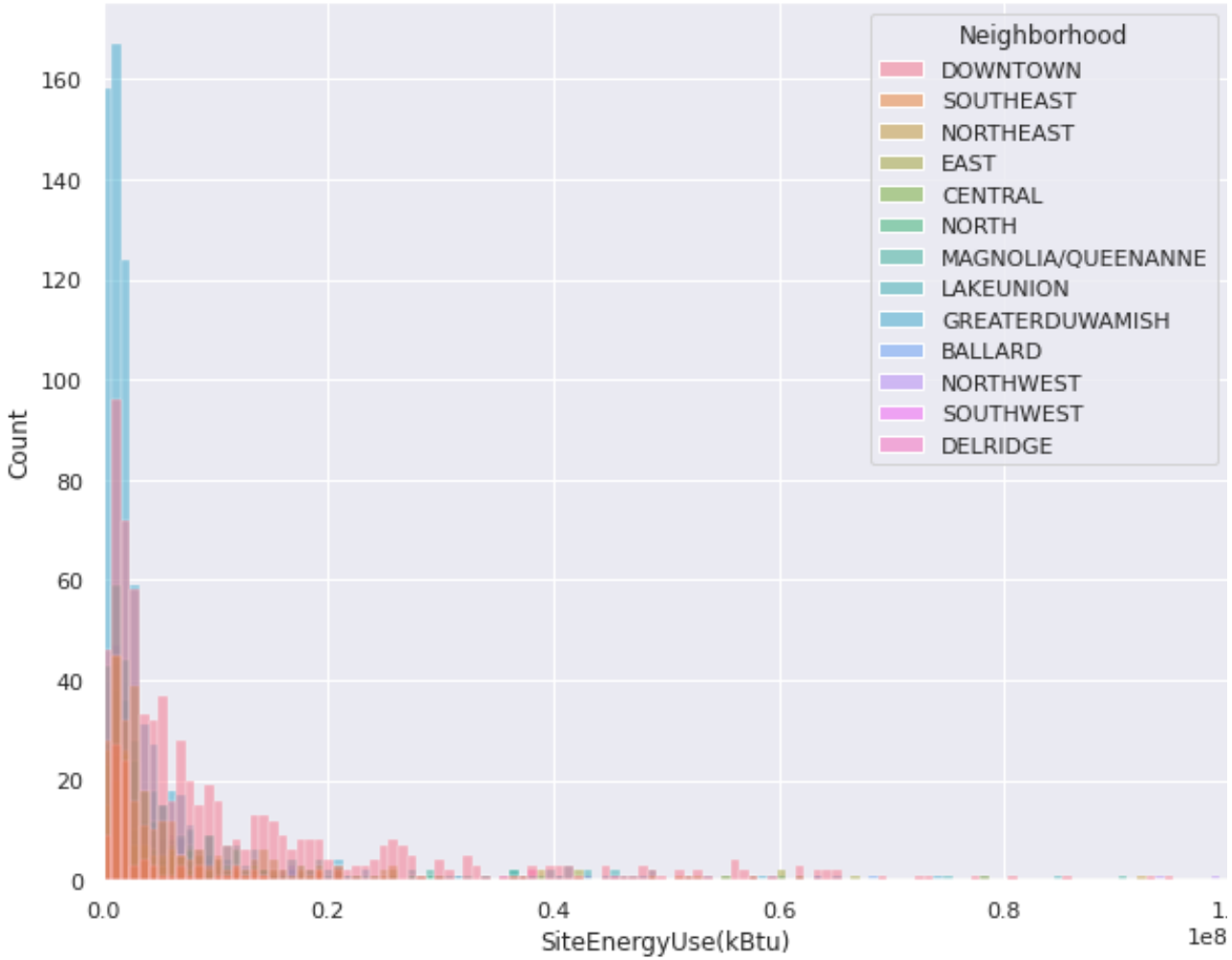
- **Recherche de corrélations**

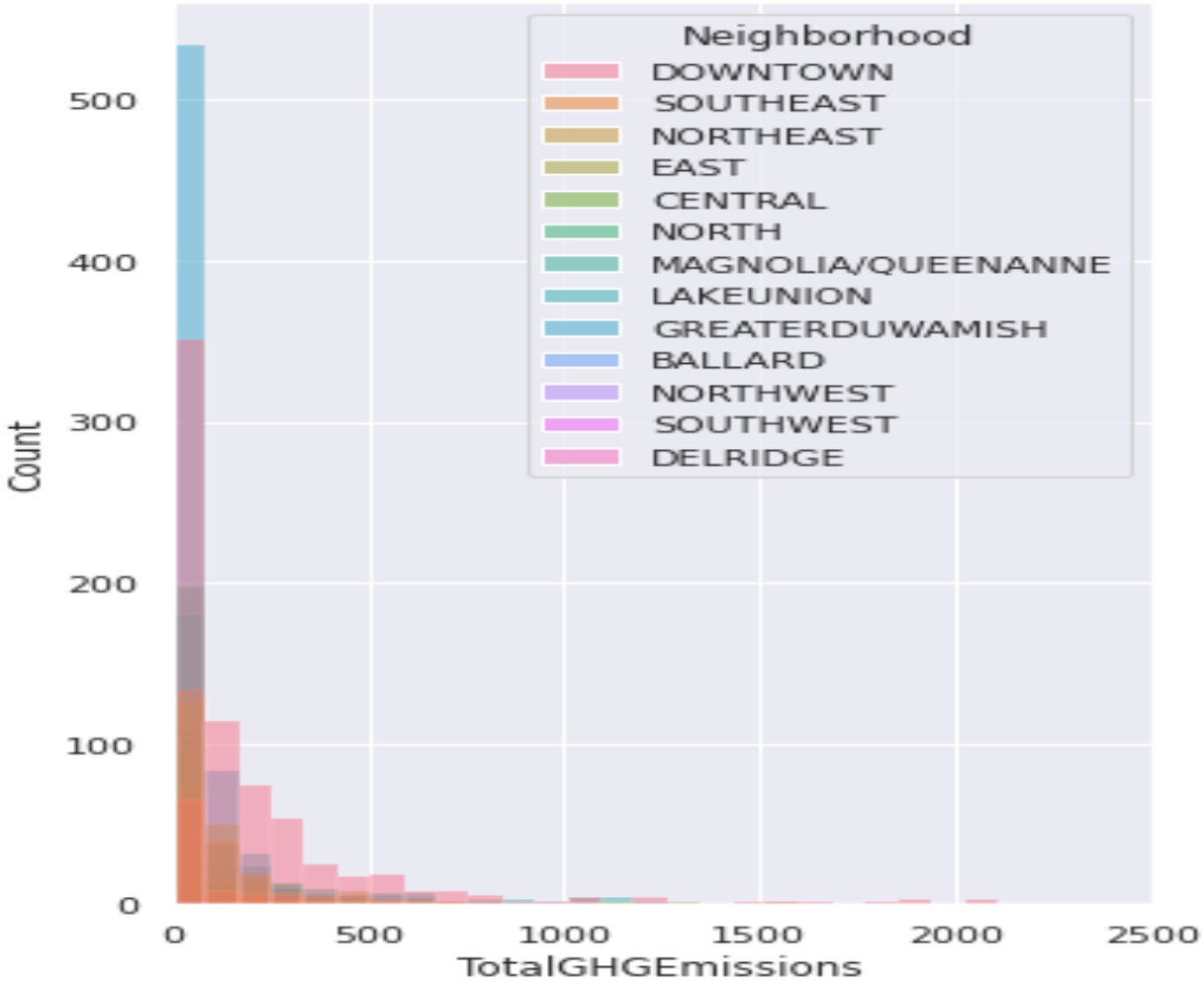


- Les 2 targets à prédire 'SiteEnergyUse(kBtu)' et 'TotalGHGEmissions' sont fortement corrélées entre elles.
- Elles sont aussi corrélées avec la plus part des variables de typologie des bâtiments (surface nombre de bâtiments)
- L'Energy Star Score n'est corrélé à aucune autre variable.

Analyse descriptive

- **Mise en évidence de tendances**

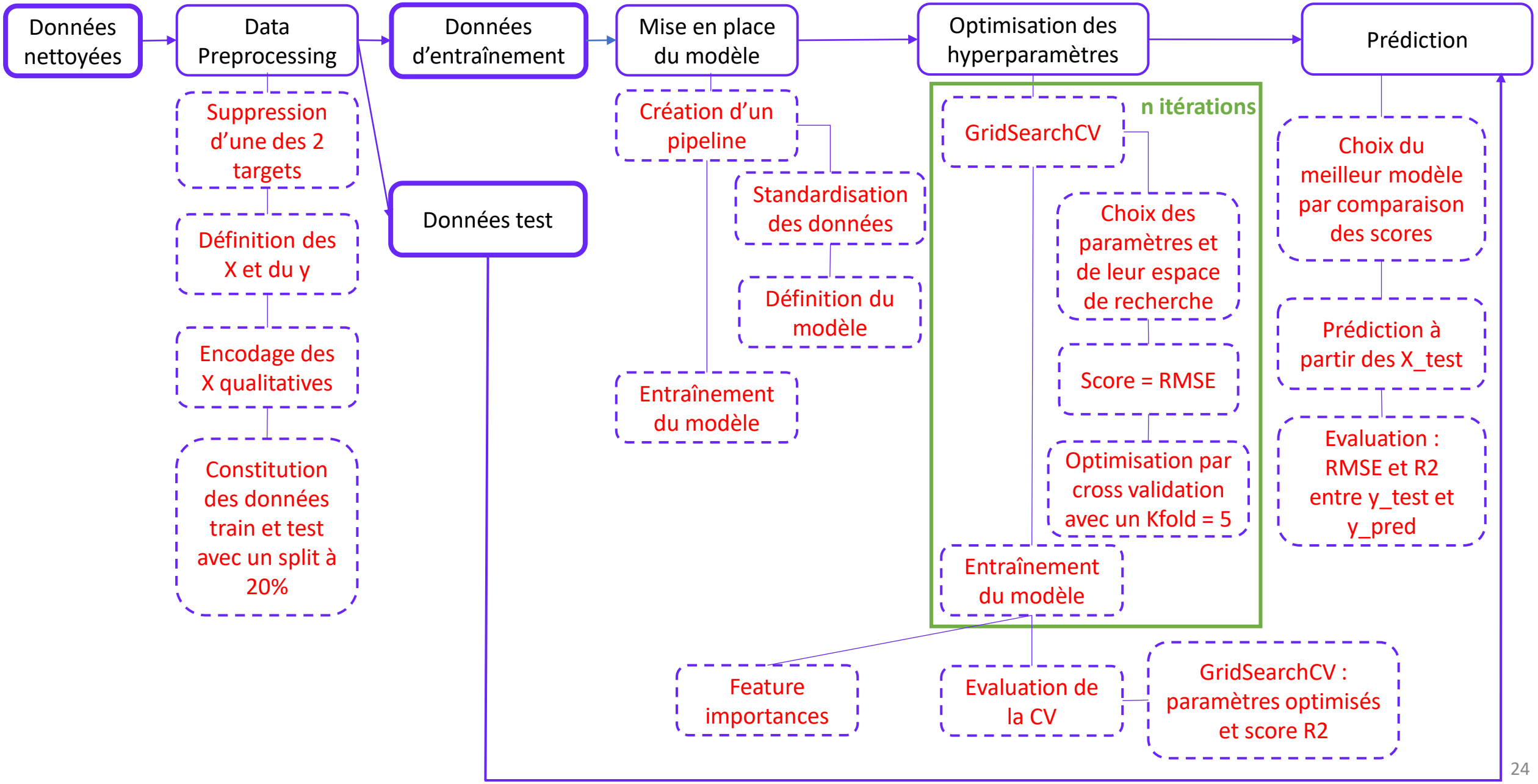




Modélisation

Modélisation

- **La stratégie de modélisation**



Modélisation

- **Choix des modèles et des hyperparamètres**

Les targets à prédire sont des variables **quantitatives**. Il faut donc répondre à des problèmes de **régression supervisées.**

➢ **La Régression Dummy :**

- Régression linéaire basique qui renvoie la moyenne qui sert de Baseline pour la prédiction.
- Pas d'optimisation de paramètres

➢ **La Régression Ridge :**

- Modèle par régression linéaire qui intègre un terme de régulation = pénalité qui réduit le risque d'apprentissage.
- Adapté aux jeux de données avec un nombre important de variables ou lorsque les variables sont corrélées.
- Hyperparamètre :
  - **Alpha** : la constante de régulation. Alpha > 0.

Modélisation

- **Choix des modèles et des hyperparamètres**

➢ **Le Random Forest:**

- Méthode ensembliste par **forêt aléatoire.**
- Combinaison parallèle d'arbres de décision obtenus par **boostrap.**
- La construction des arbres est basée sur un sous ensemble de features sélectionnés de manière **aléatoire** → les arbres sont moins corrélés, il y a moins d'overfitting.
- Hyperparamètres :
  - **n_estimators** : le nombre d'arbres dans la forêt.
  - **max_depth** : la profondeur maximale d'un arbre.
  - **max_features** : le nombre de features choisi alétoirement à chaque embranchement (entre 0 et 1)
  - **min_samples_split** : Le nombre minimum d'échantillons requis à chaque embranchement.

Modélisation

- **Choix des modèles et des hyperparamètres**

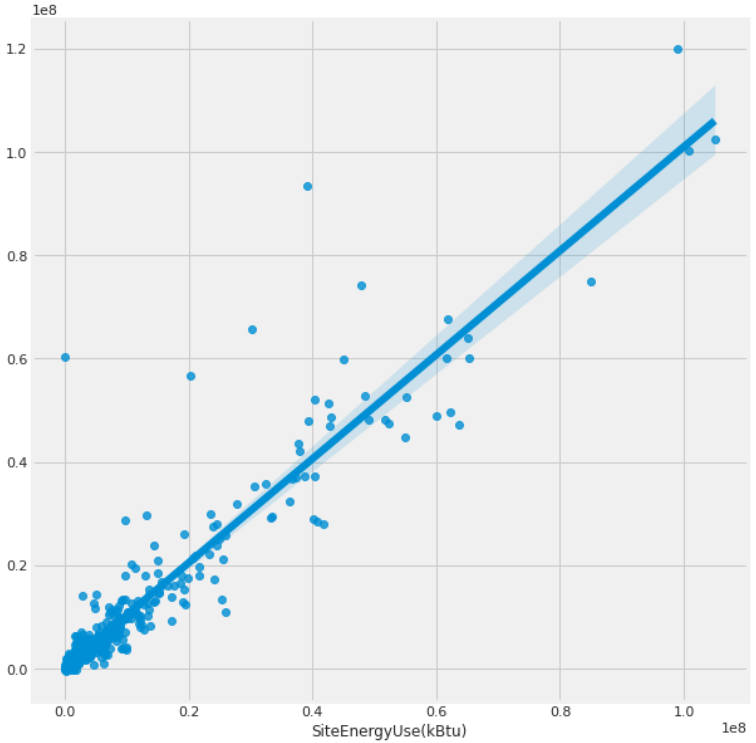
➢ **Le Gradient Boosting:**

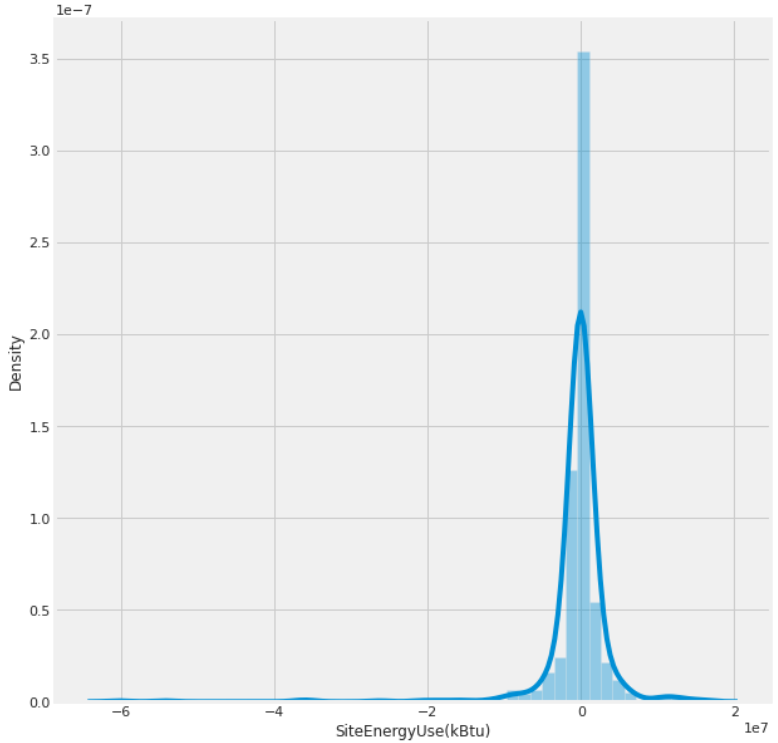
- Méthode ensembliste par **combinaison d'arbres de décision** obtenus par boostrap**.**
- La combinaison est **séquentielle** : les modèles sont construits à la suite, en fonction du modèle précédent.
- La succession des modèles se comporte selon l'algorithme de descente de gradient, avec une fonction de perte à minimiser.
- Hyperparamètres :
  - **n_estimators** : le nombre de boosting.
  - **max_depth** : la profondeur maximale d'un arbre.
  - **learning_rate** : Le taux d'apprentissage qui réduit la contribution de chaque arbre.
  - **min_samples_split** : Le nombre minimum d'échantillons requis à chaque embranchement.

Modélisation

- **Résultats pour la consommation d' énergie.**

	Ridge_regression	Random Forest	Gradient Boosting
Grid search error score	-9302215	-5105747	-4683543
R2 score	0.482	0.857	0.882



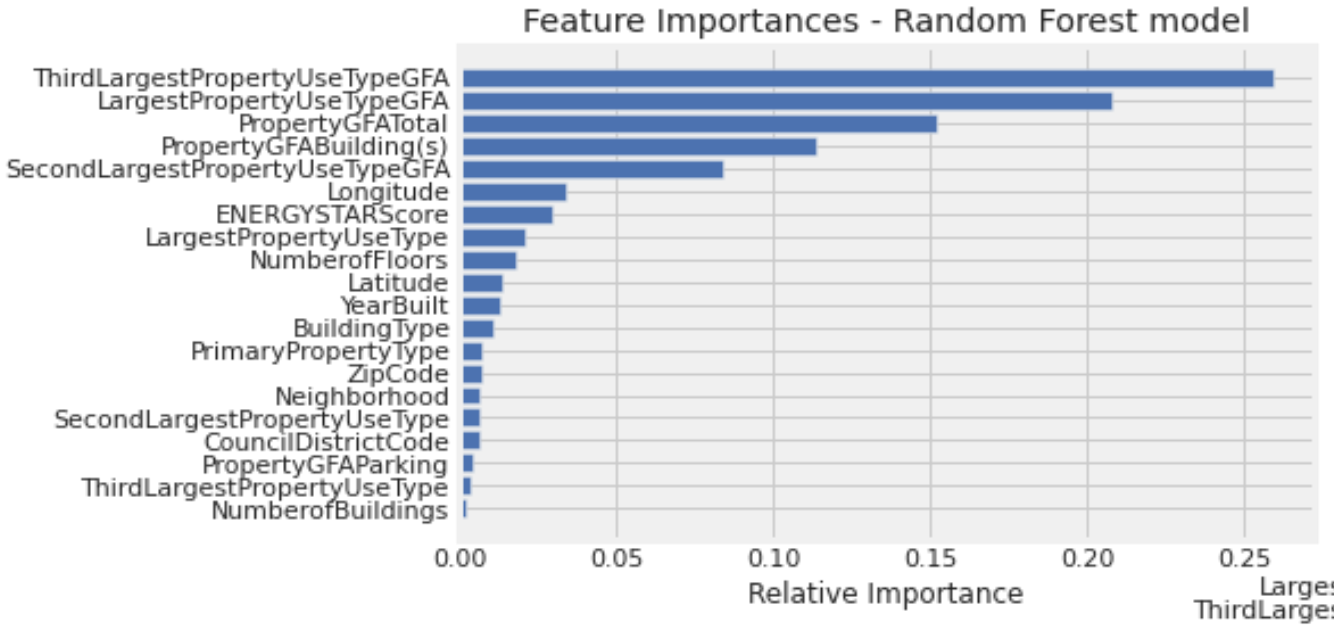



```
RMSE_grid :  4905690.073060225
R2_Score_grid :  0.8598976017578565

RMSE_dummy :  13106457.56421194
R2_Score_dummy :  -3.6856534348927994e-05
```

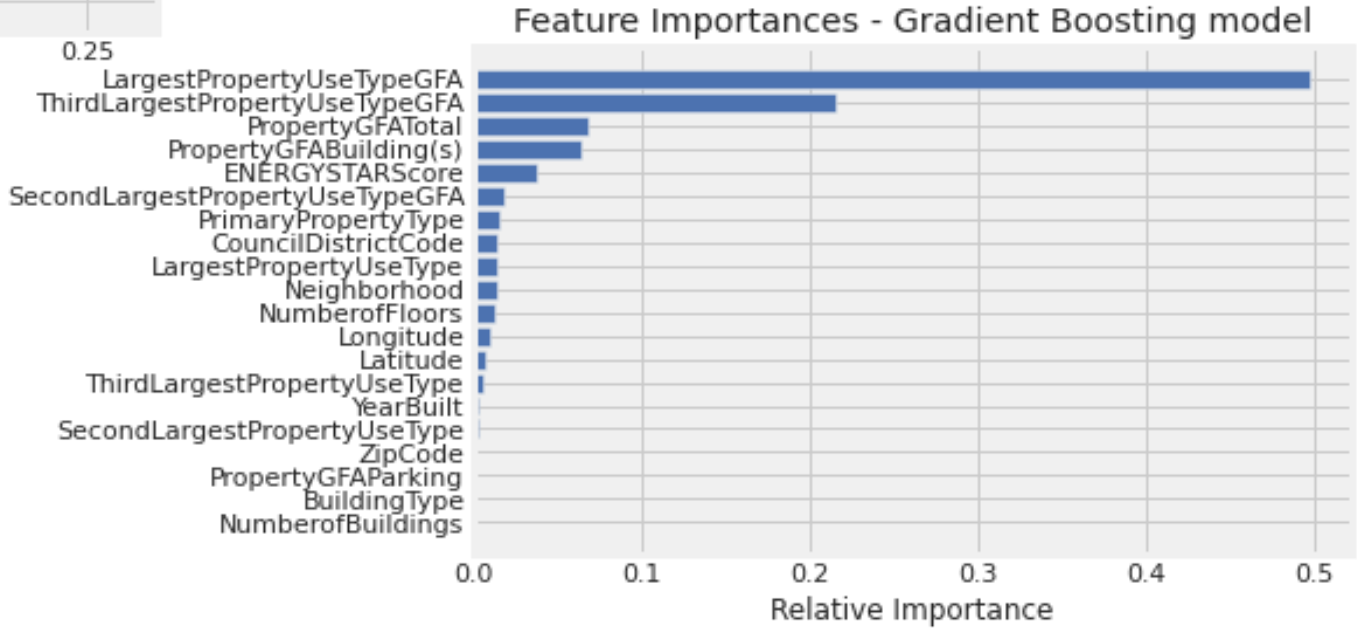
Modélisation

- **Résultats pour la consommation d’ énergie.**



- L’Energy Star Score est 7ème pour le Random Forest et 5ème pour Le Gradient Boosting

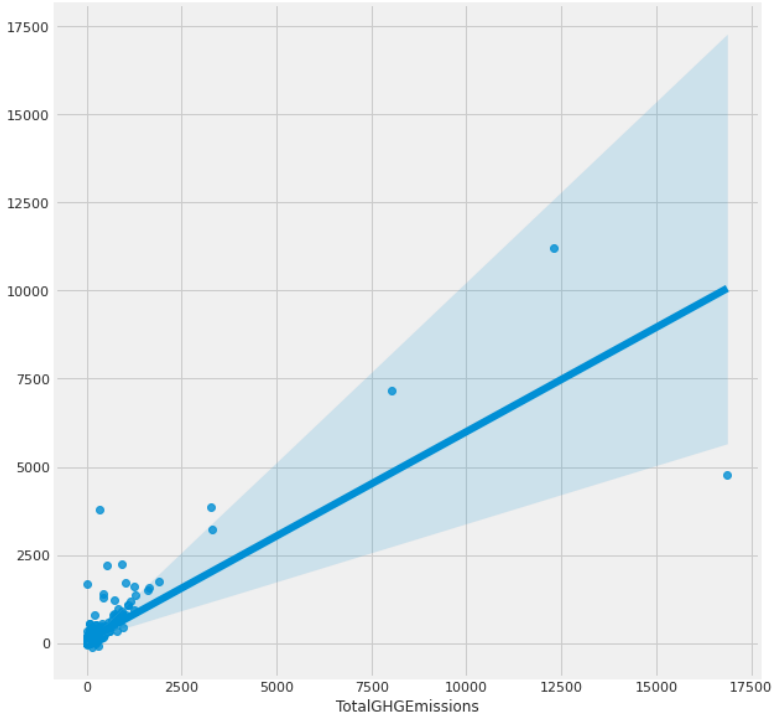
- Les variables de surface, notamment en fonction de l’activité, impact le plus les modèles en arbre.

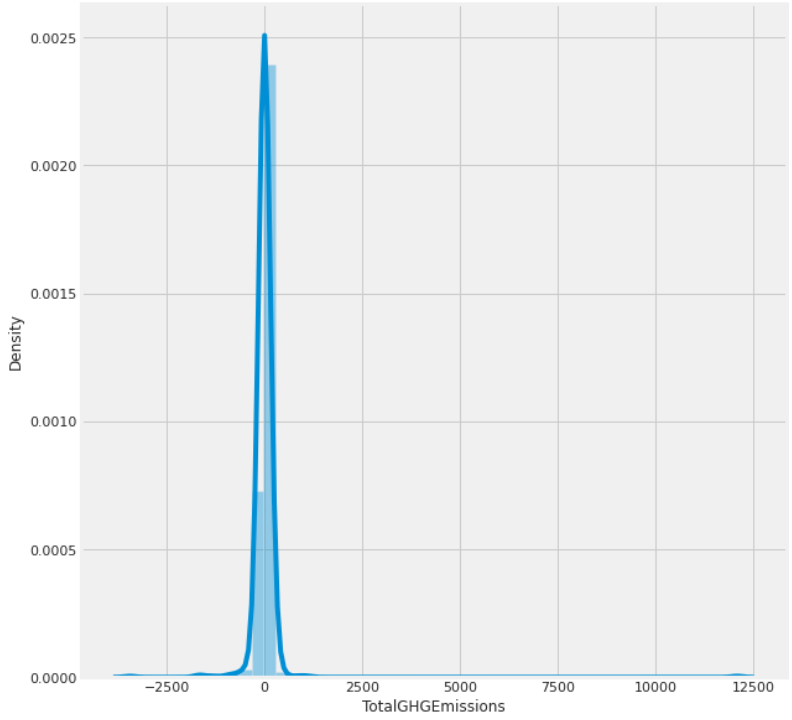


Modélisation

- **Résultats pour l'émission de gaz à effet de serre.**

	Ridge_regression	Random Forest	Gradient Boosting
Grid search error score	-416	-303	-235
R2 score	0.404	0.694	0.787



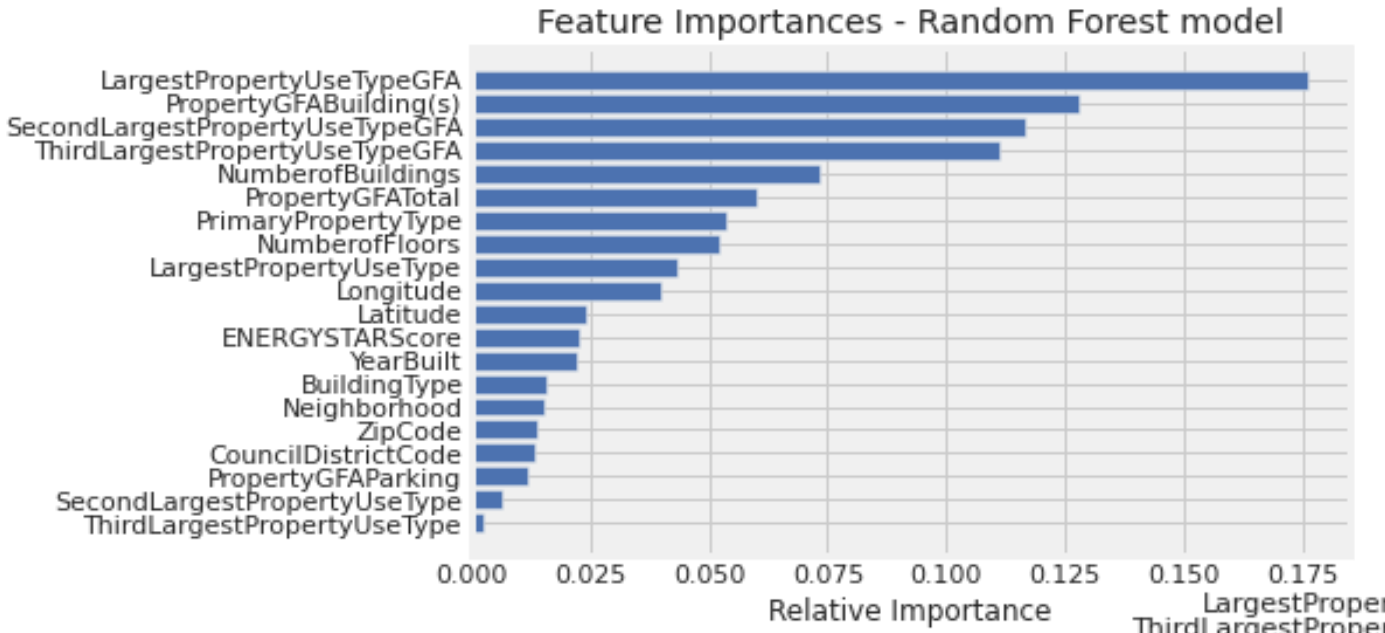



```
RMSE_grid :  519.1551770816351
R2 Score grid :  0.6771139322913269

RMSE_dummy :  13106457.56421194
R2_Score_dummy :  -3.6856534348927994e-05
```

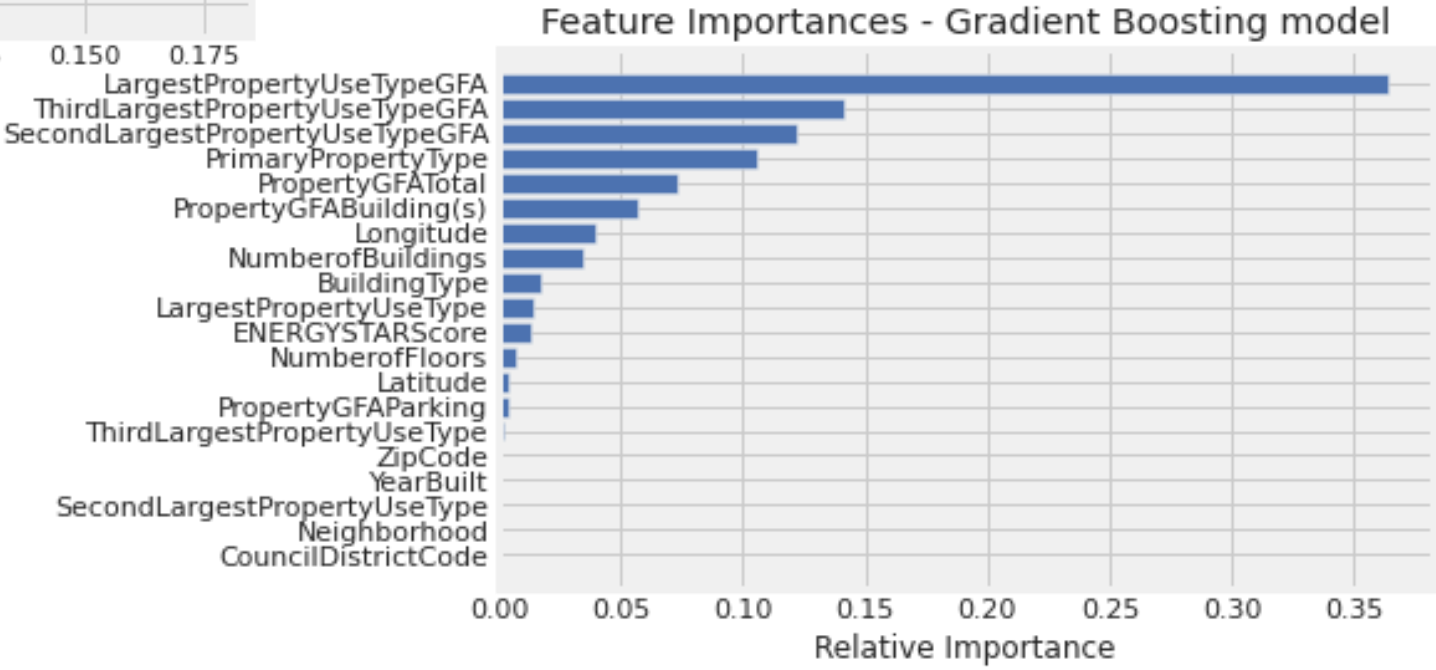
Modélisation

- **Résultats pour l'émission de gaz à effet de serre.**



- L'Energy Star Score a moins d'importance pour la target gaz émis.

- Les variables de surface, notamment en fonction de l'activité, impact le plus les modèles en arbre.



Intérêt de l'ENERGY STAR SCORE

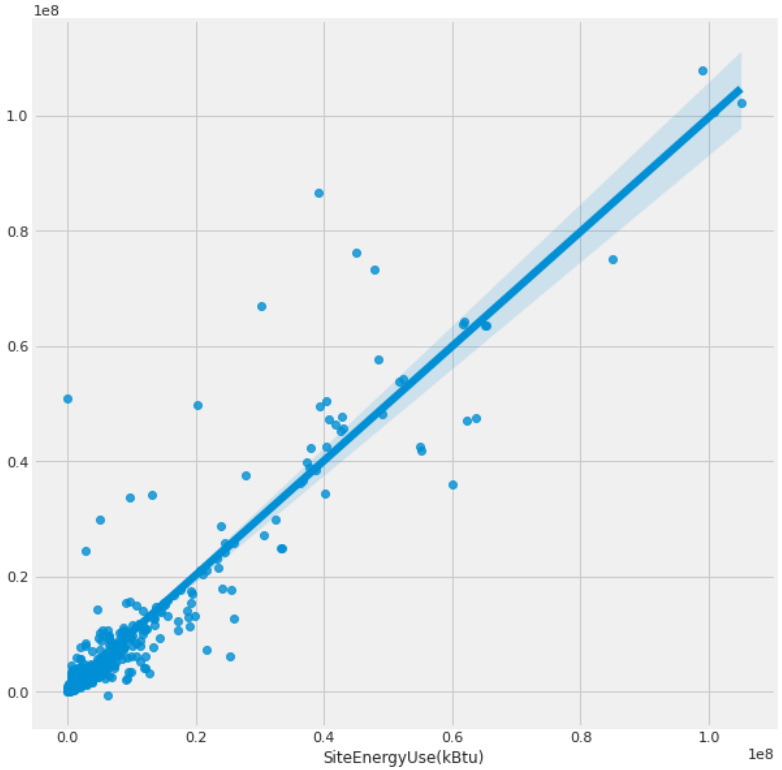
➢ **Stratégie :**

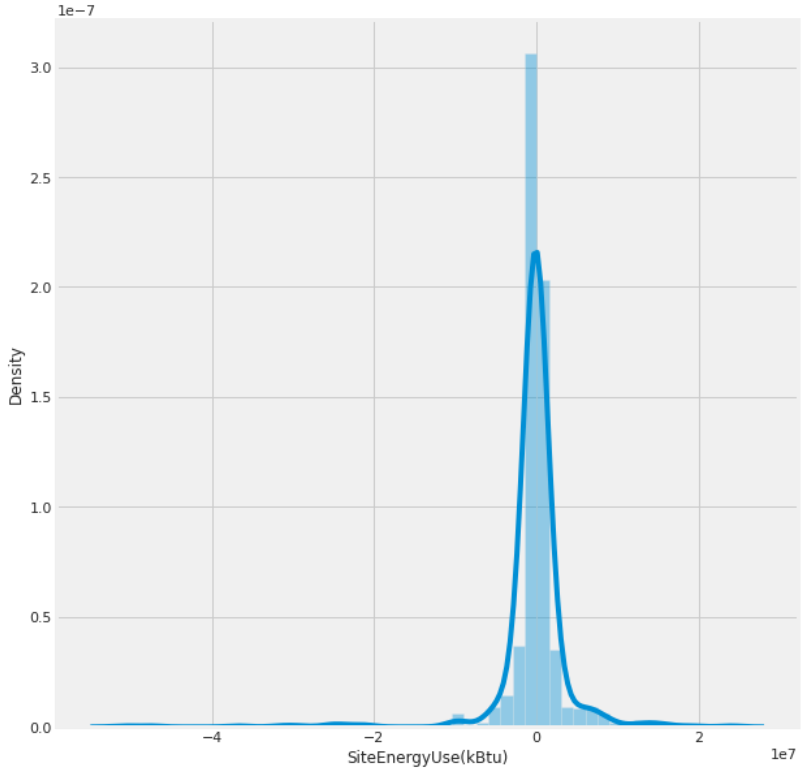
- Relancer les modélisations selon la même démarche que pour le jeux de données complet, mais sur un jeux de données sans l'Energy Star Score.

Intérêt de l'Energy Star Score

- **Résultats pour la consommation d' énergie.**

	Ridge_regression	Random Forest	Gradient Boosting
Grid search error score	-9393239	-5300234	-4895498
R2 score	0.475	0.846	0.87



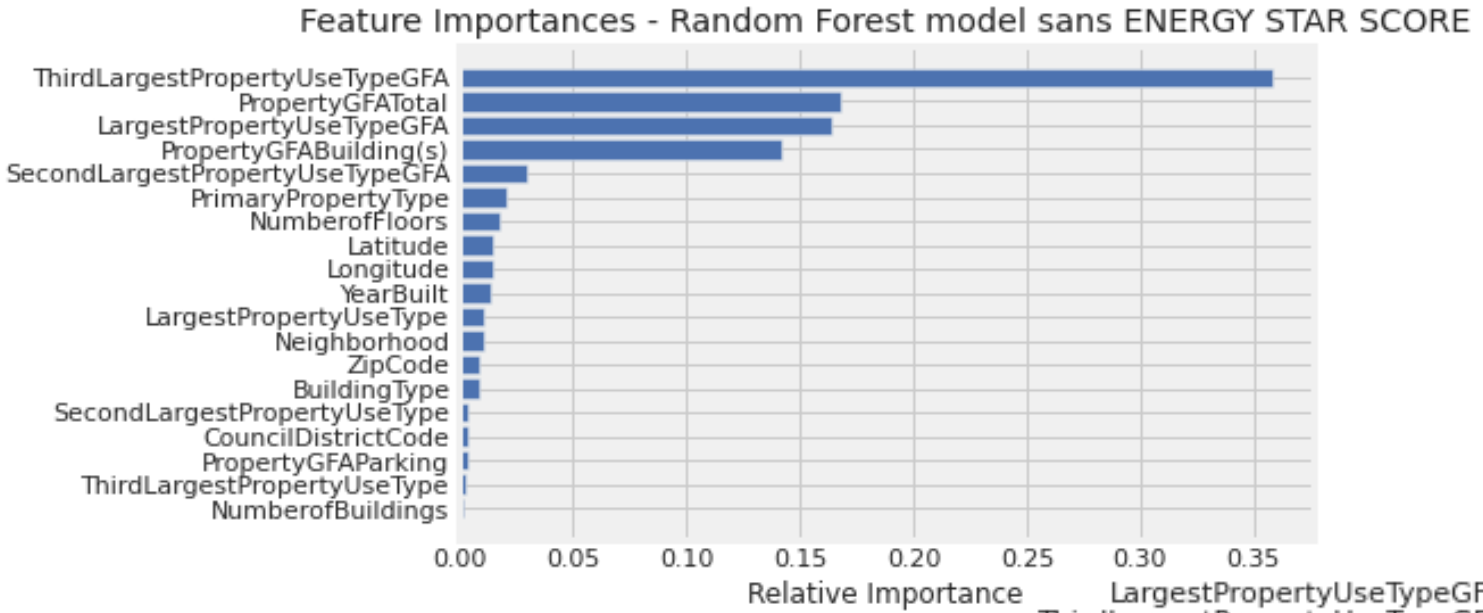


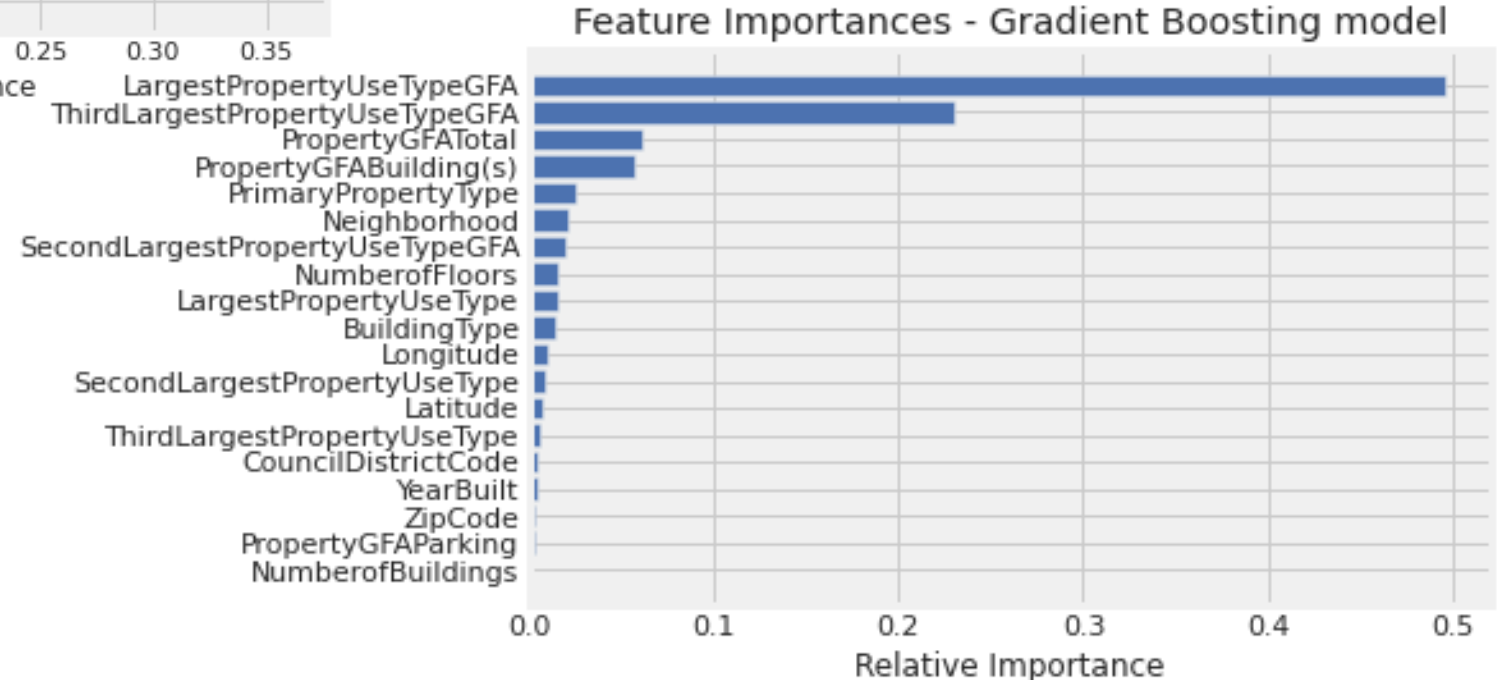

```
RMSE_grid :  4948571.200043767
R2_Score_grid :  0.8574375988609136

RMSE_dummy :  13106457.56421194
R2_Score_dummy :  -3.6856534348927994e-05
```

Intérêt de l'Energy Star Score

- **Résultats pour la consommation d' énergie.**



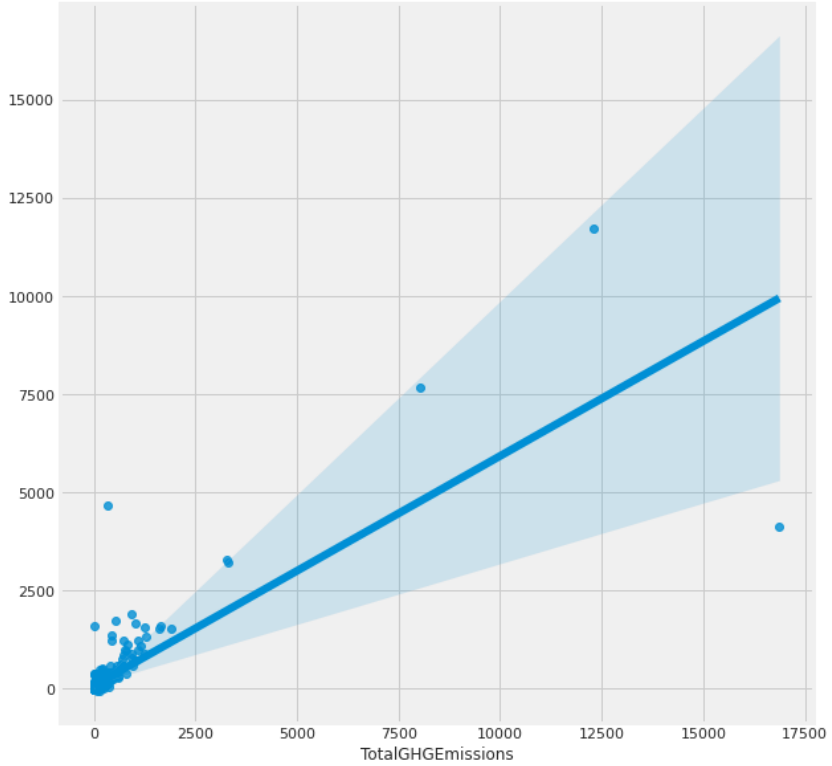


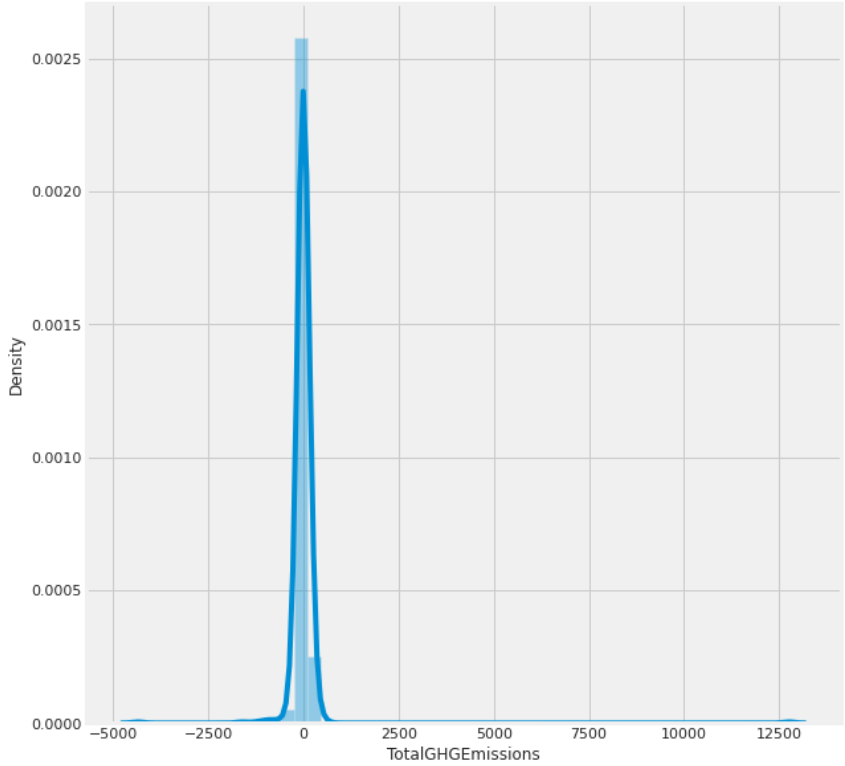
- Les variables de surface, restent les plus importantes de façon plus nette.

Intérêt de l'Energy Star Score

- **Résultats pour l'émission de gaz à effet de serre.**

	Ridge_regression	Random Forest	Gradient Boosting
Grid search error score	-200813	-296	-247
R2 score	0.369	0.707	0.759



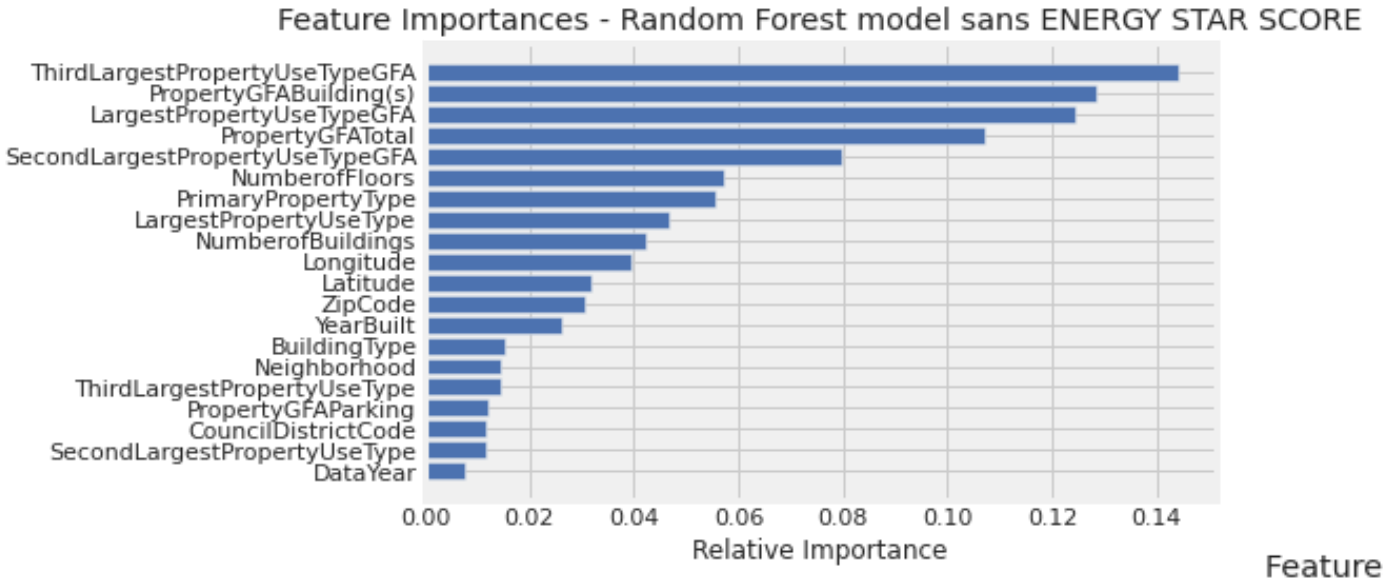


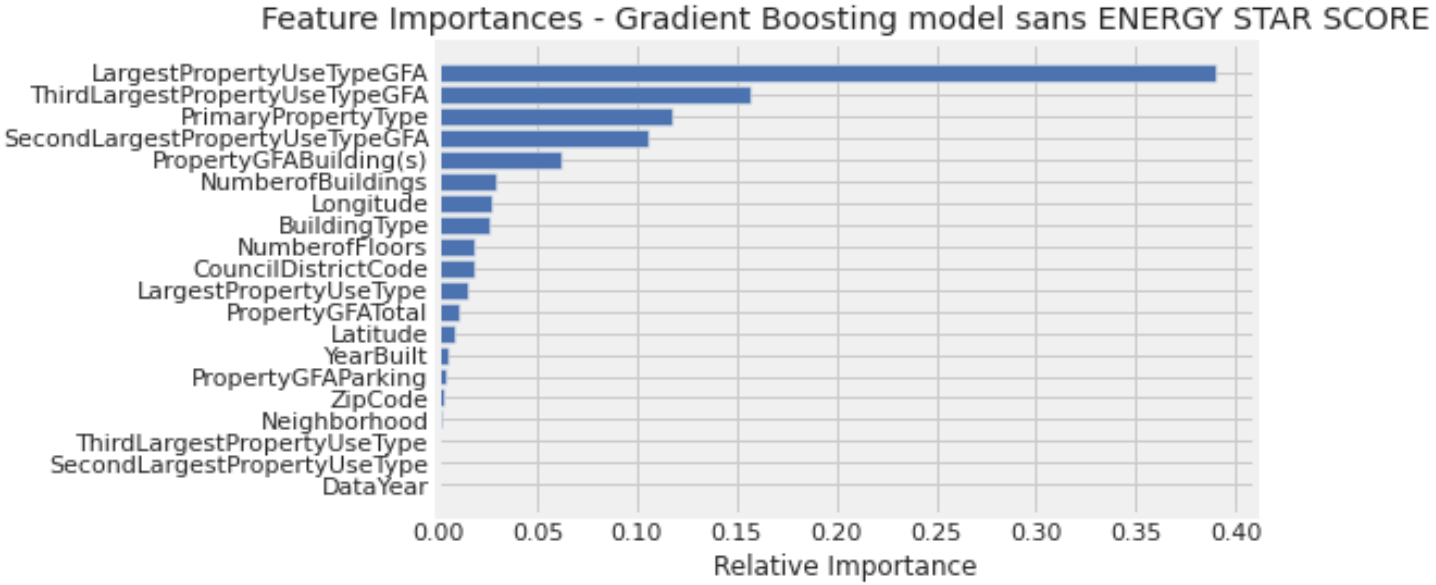

```
RMSE_grid :  547.0149351100725
R2_Score_grid :  0.6415296040657681

RMSE_dummy :  913.7825121742885
R2_Score_dummy :  -0.0003238239700298351
```

Intérêt de l'Energy Star Score

- **Résultats pour l'émission de gaz à effet de serre.**





- Les variables de surface, notamment en fonction de l'activité, impact le plus les modèles en arbre.

Conclusion

Conclusions

➢ La modélisation :

- Il est possible de proposer un modèle prédictif fiable, sur la base des données de 2015 et 2016, afin d'anticiper les consommations en énergie et les émissions de gaz à effet de serre pour les bâtiments non résidentiels de la ville de Seattle.

- Les modèles en arbre sont significativement plus performants que le modèle linéaire par Ridge régression. C'est la modélisation par Gradient Boosting qui a les performances les plus optimales en cross validation (erreur minimale et R2 maximale), pour les deux targets à prédire.

- En terme de temps de traitement, j'ai constaté que le Gradient Boosting nécessite le double de temps que le modèle Random Forest. Au vu des performances très proches de celles obtenues avec le gradient Boosting, Le Random Forest est un bon compromis entre délais et robustesse.

- L'analyse des Feature importances a montré que les surfaces par type d'activité étaient les variables qui impactaient le plus les modèles.

Conclusions

➢ L'Energy Star Score :

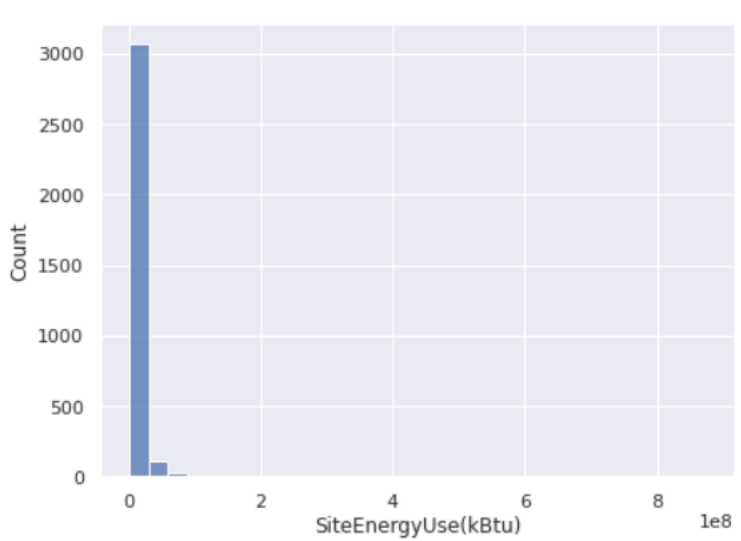
- L'Energy Star Score a peu d'impact sur la performance des modèles, pour les deux targets à prédire.
- Il ne présente pas d'intérêt pour la prédiction des émissions de gaz à effet de serre, ainsi que pour l'énergie consommée.
- Il peut néanmoins continuer à sensibiliser les propriétaires sur l'impact environnemental de leurs bâtiments.

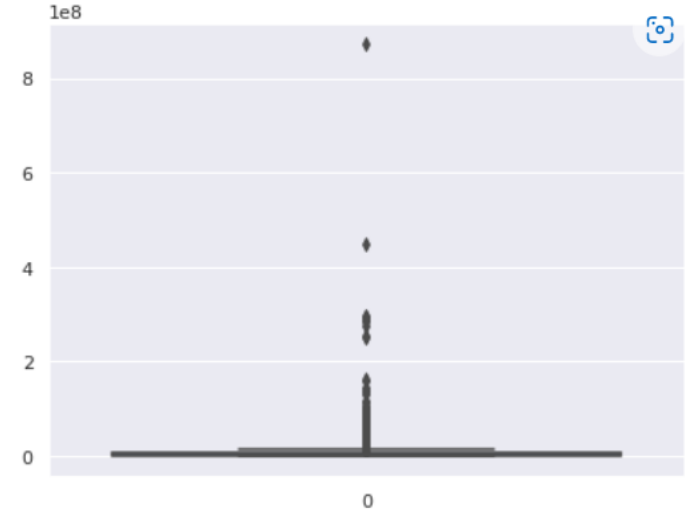
➢ Voies d'amélioration :

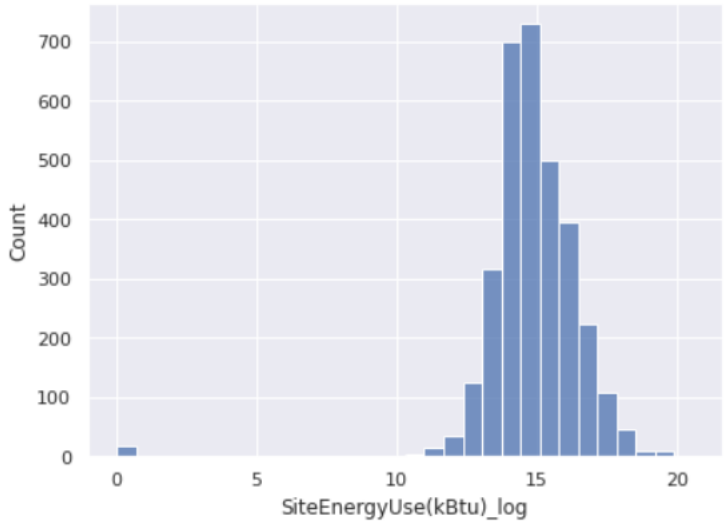
- Il serait intéressant de voir dans quel sens les superficies par type d'activité impact la consommation d'énergie et l'émission de gaz à effet de serre.
- La mise en place des modèles pour l'énergie consommée a nécessité la suppression des valeurs extrêmes. Une transformation mathématique des données (passage en log) aurait permis de les lisser et d'éviter la suppression de données.

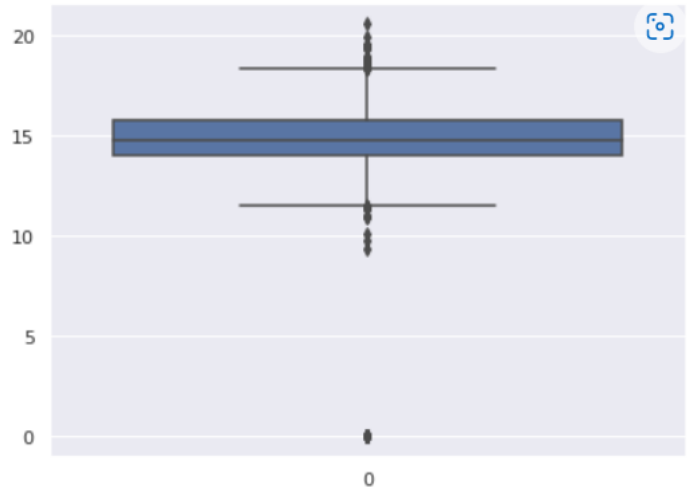
Conclusions

- **L’effet du passage en log pour la target energy :**









Conclusions

- Modélisation sur données complètes non transformées :

	Ridge_regression tot	Random Forest tot
Grid search error score	-1.211245e+07	-12203983.01
R2 score	6.220000e-01	0.74

- Modélisation sans données extrêmes non transformées :

	Ridge_regression	Random Forest
Grid search error score	-9302215	-5105747
R2 score	0.482	0.857

- Modélisation sur données complètes transformées (log) :

	Ridge_regression log	Random Forest log
Grid search error score	-1.405	-0.699
R2 score	0.303	0.820

Merci pour votre attention