# OpenClassrooms

**Soutenance:** Projet 3

## Consommation Energetique à Seattle

---

Edward Levavasseur

Updated: 2021/03/06

# Problématique de la ville de Seattle

Des relevés minutieux ont été effectués par vos agents en 2015 et en 2016. Cependant, ces relevés sont coûteux à obtenir, et à partir de ceux déjà réalisés, **vous voulez tenter de prédire les émissions de CO2 et la consommation totale d'énergie** de bâtiments pour lesquels elles n'ont pas encore été mesurées.

Vous cherchez également à **évaluer l'intérêt de l'"ENERGY STAR Score" pour la prédiction d'émissions**, qui est fastidieux à calculer avec l'approche utilisée actuellement par votre équipe.

# Objectifs

- Nettoyer les données

# Objectifs

- Nettoyer les données
- Identifier les variables principales

# Objectifs

- Nettoyer les données
- Identifier les variables principales
- Utiliser différentes méthodes pour prédire la consommation énergétique
  - regréssion linéaire, random forest, XG boost...

# Objectifs

- Nettoyer les données
- Identifier les variables principales
- Utiliser différentes méthodes pour prédire la consommation énergétique
  - regréssion linéaire, random forest, XG boost...
- Evaluer le bénéfice du EnergyStar Score pour évaluer la consommation réelle d'énergie

# Objectifs

- Nettoyer les données
- Identifier les variables principales
- Utiliser différentes méthodes pour prédire la consommation énergétique
  - regréssion linéaire, random forest, XG boost...
- Evaluer le bénéfice du EnergyStar Score pour évaluer la consommation réelle d'énergie

# Plan

1. Nettoyage des données
2. Relation entre les differentes variables
3. Prediction de la consommation d'energie
4. Evaluation du EnergyStar Score
5. Conclusion

# Nettoyage des données

# Nettoyage des données (1/3)

- Importation des données de 2015 et de 2016
- Concatenation des 2 bases de données en une

# Nettoyage des données (1/3)

- Importation des données de 2015 et de 2016
- Concatenation des 2 bases de données en une

- Suppression des observations avec espaces sur
  - **Data["PrimaryPropertyType"]**

# Nettoyage des données (1/3)

- Importation des données de 2015 et de 2016
- Concatenation des 2 bases de données en une

- Suppression des observations avec espaces sur
  - **Data["PrimaryPropertyType"]**

- Supression des observations avec des "inf" et "NaN" sur
  - **Data[["YearBuilt", "NumberofBuildings", "NumberofFloors", "Electricity(kWh)", "PropertyGFATotal", "SiteEnergyUseWN(kBtu)", "SiteEnergyUse(kBtu)"]]**

# Nettoyage des données (2/3)

- Suppression des "Outliers", en utilisant une régression linéaire:

```
import numpy as np
from sklearn.linear_model import LinearRegression

Data['Ones'] = 1

X = Data[['YearBuilt', 'NumberofBuildings', 'NumberofFloors', 'PropertyGFATotal', 'Ones'] + Property_Type + Property_Type_GPA].iloc[0:10000000]
y = Data['Electricity(kWh)'].iloc[0:10000000]

reg = LinearRegression().fit(X, y)

Predicted_y = reg.predict(X)

Data['Predicted_y'] = reg.predict(X)
Data['Difference'] = abs(Data['Predicted_y'] - Data['Electricity(kWh)'])
```

- Suppression des observations dont la consommation d'électricité prédite diverge très fortement de la valeur réelle :
  - > 5 fois le seuil des top 10% d'erreur

# Nettoyage des données (3/3)

- Même approche pour la consommation de gaz.

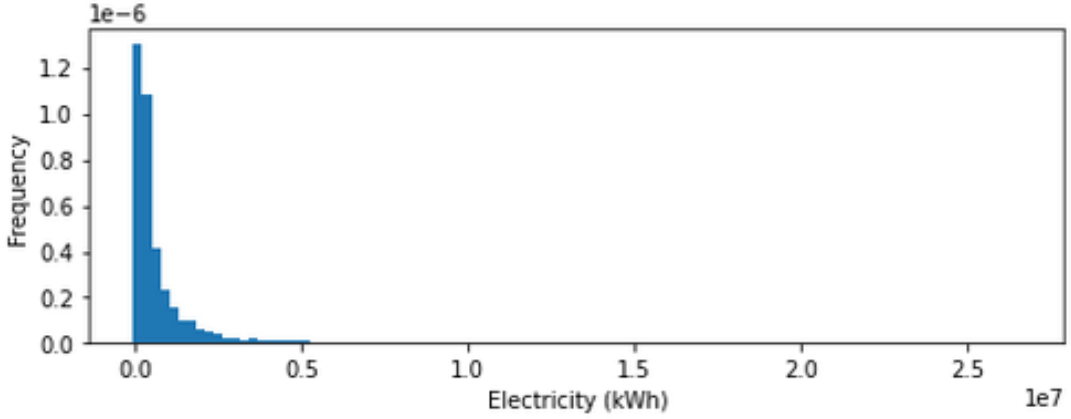
```
Original Number of Obersavtions :  6716
Final Number of Obersavtions :  6575
Percentage Dropped :  2.099 %
```

Après avoir "cleané" les données, 2,099 % des observations ont été supprimées.

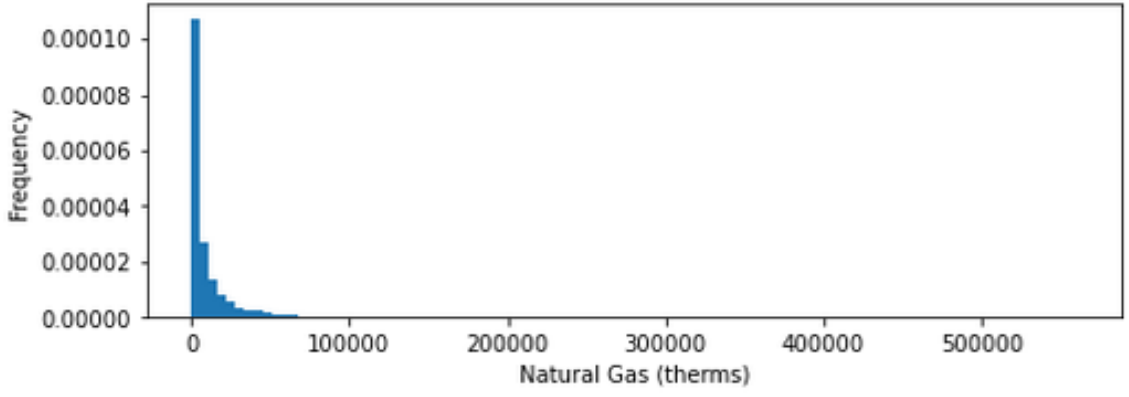
# Relation entre les differentes variables

# Histogrammes des Outputs (1/2)

- Electricité

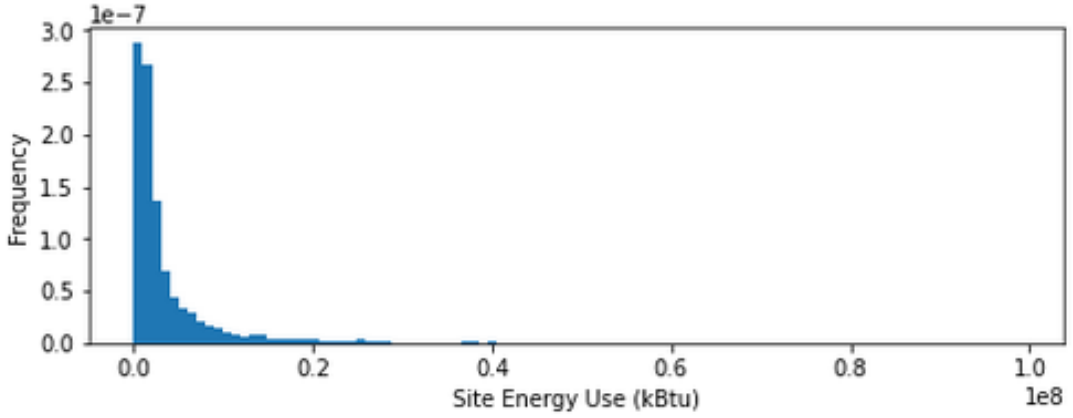



- Gaz

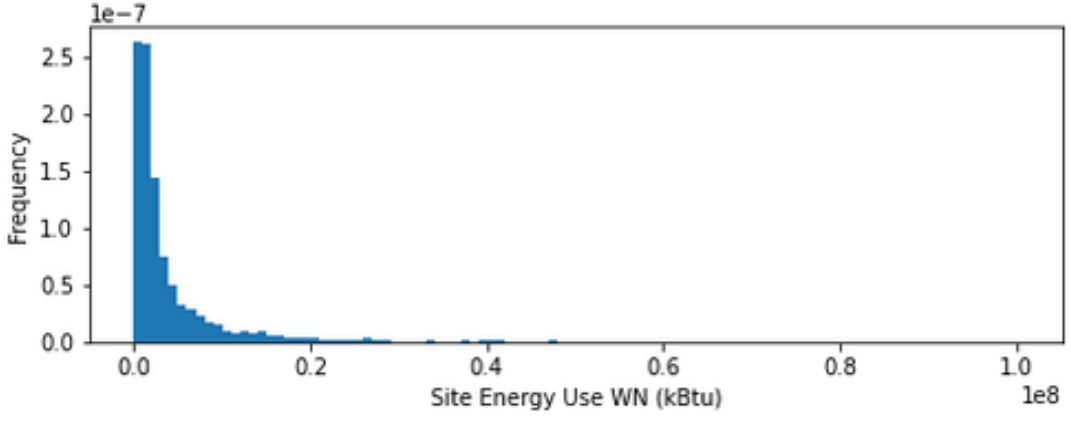


# Histogrammes des Outputs (2/2)

- Energy

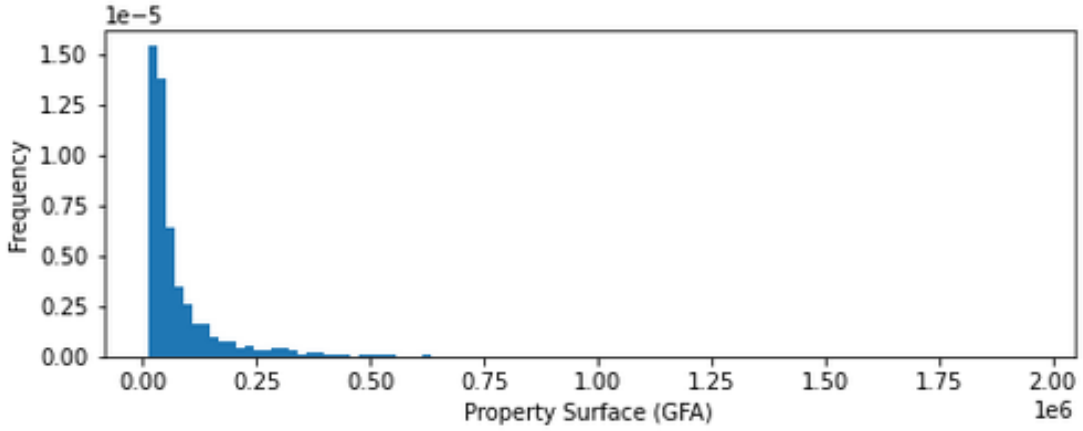



- Energy WN

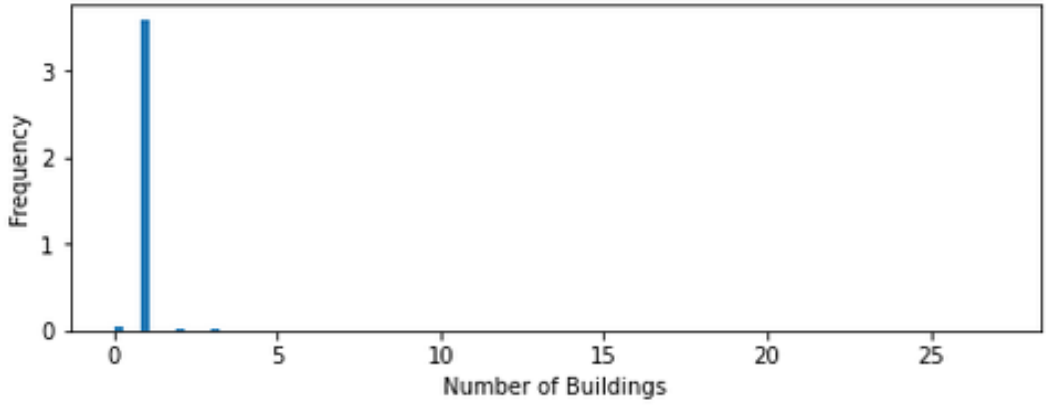


# Histogrammes des Charactéristiques (1/2)

- Surface

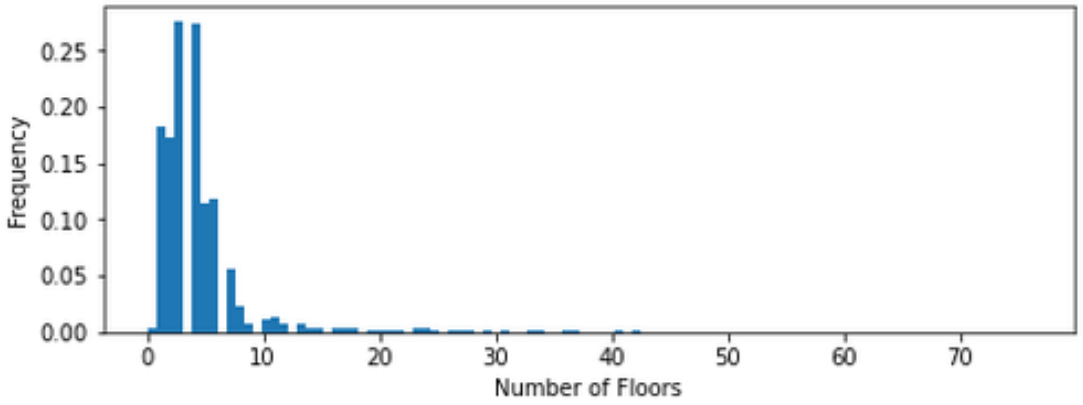



- Nombre de Battiments

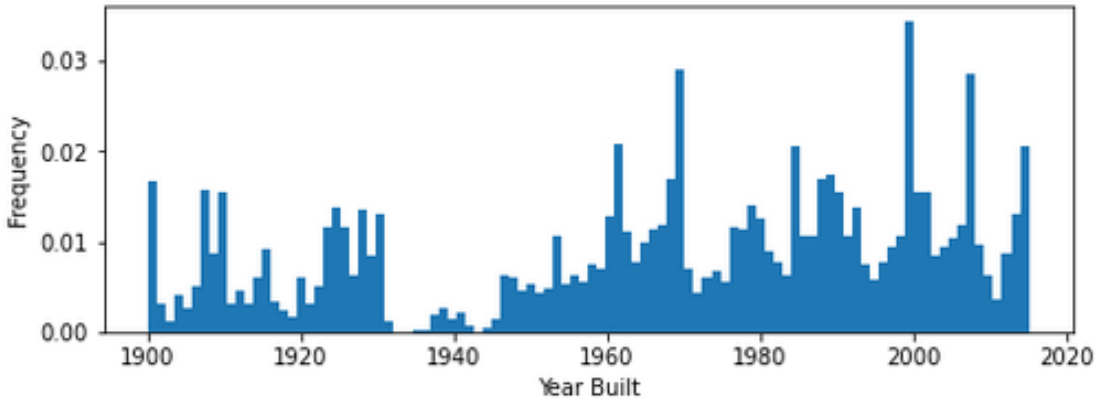


# Histogrammes des Charactéristiques (2/2)

- Nombre d'étages

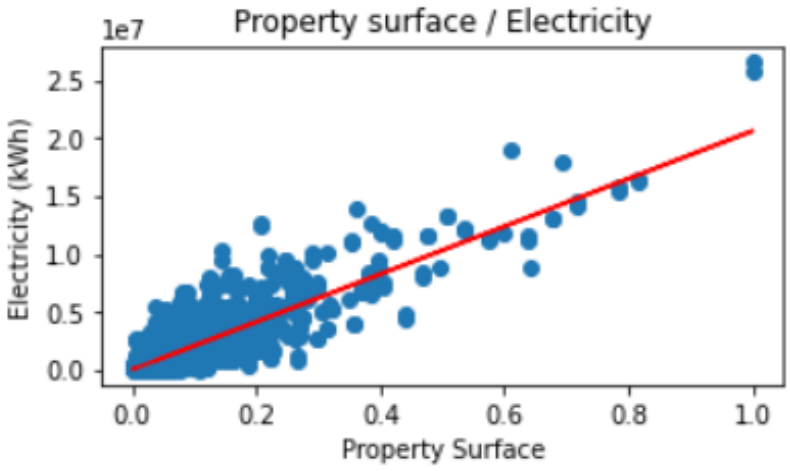


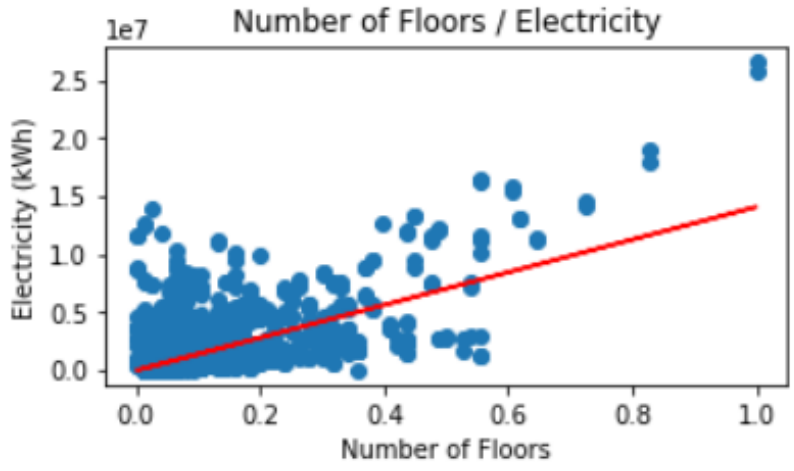

- Année de Construction

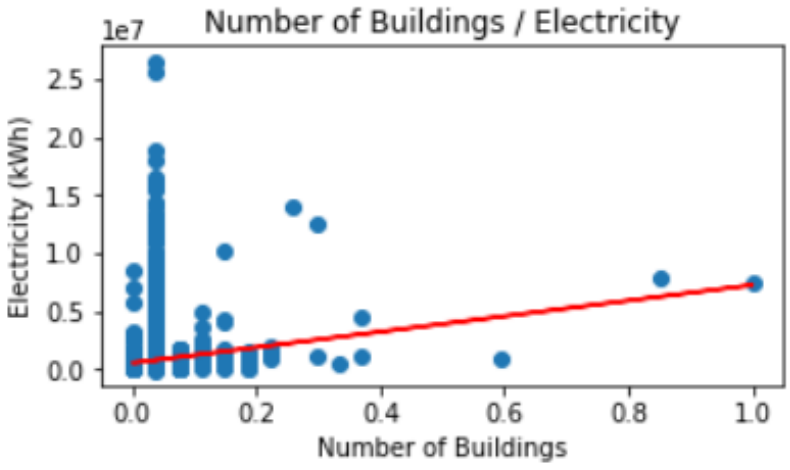


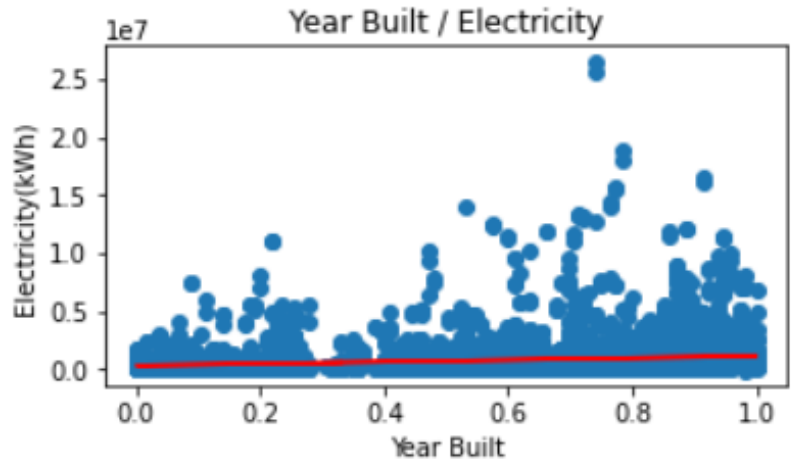
# ScatterPlots pour l'Electricité (1/2)

- Electricité / Charactéristiques



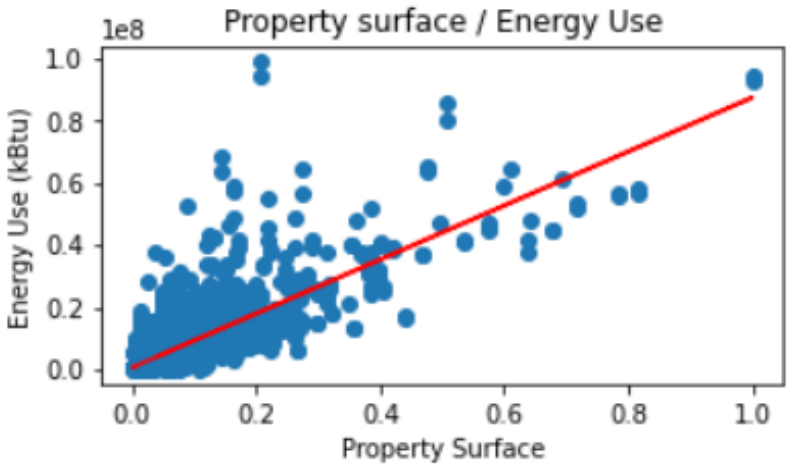



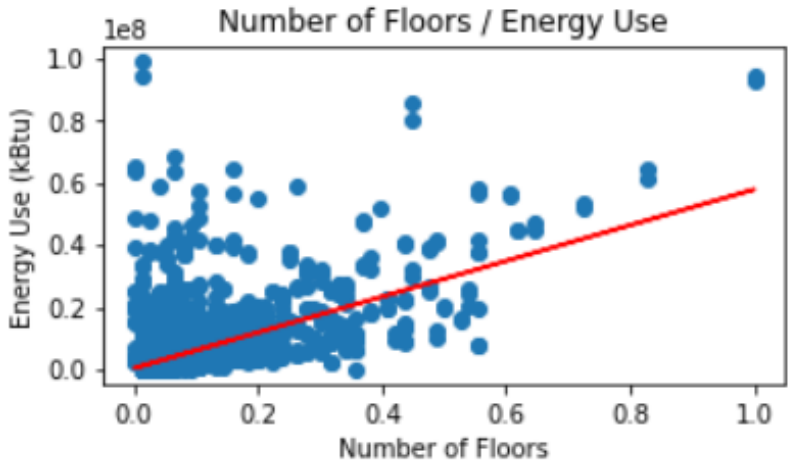



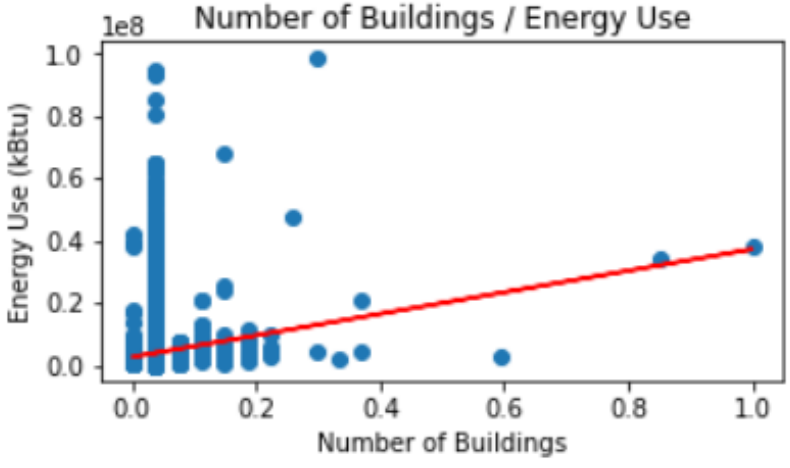



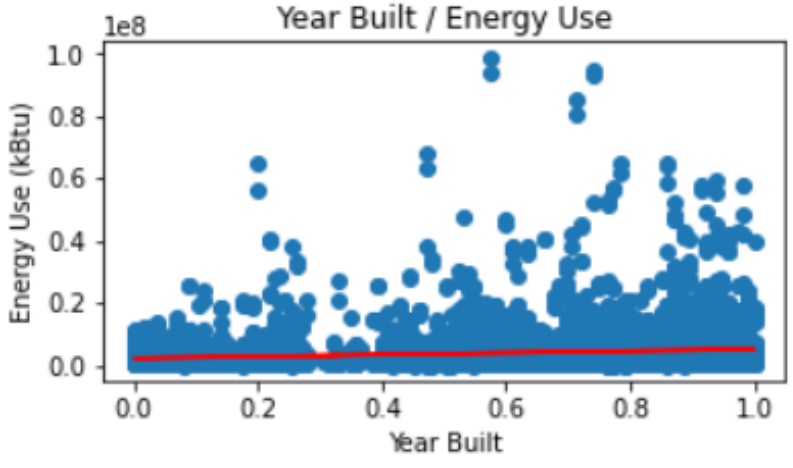
# ScatterPlots pour l'Energie (2/2)

- Energie / Charactéristiques

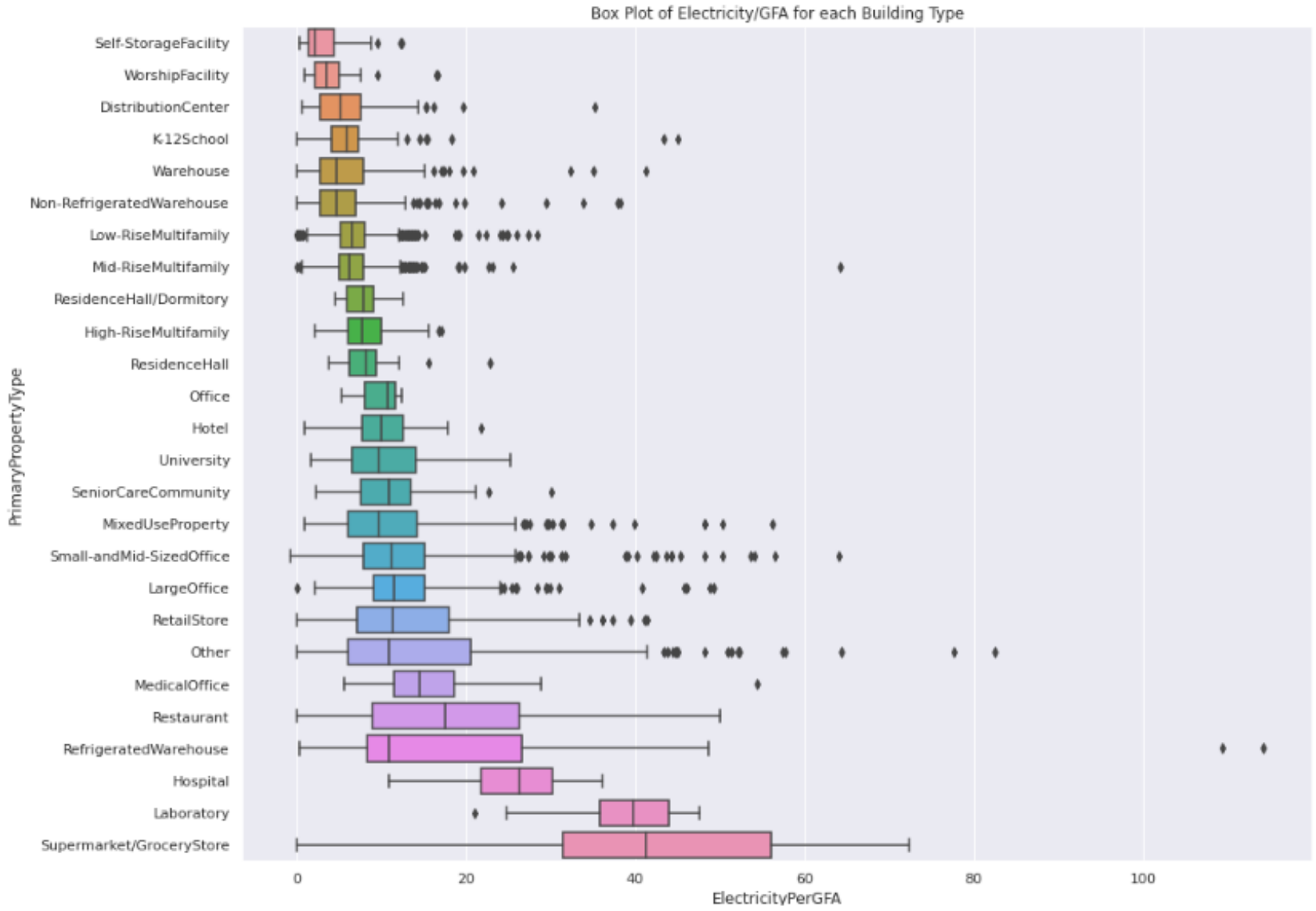


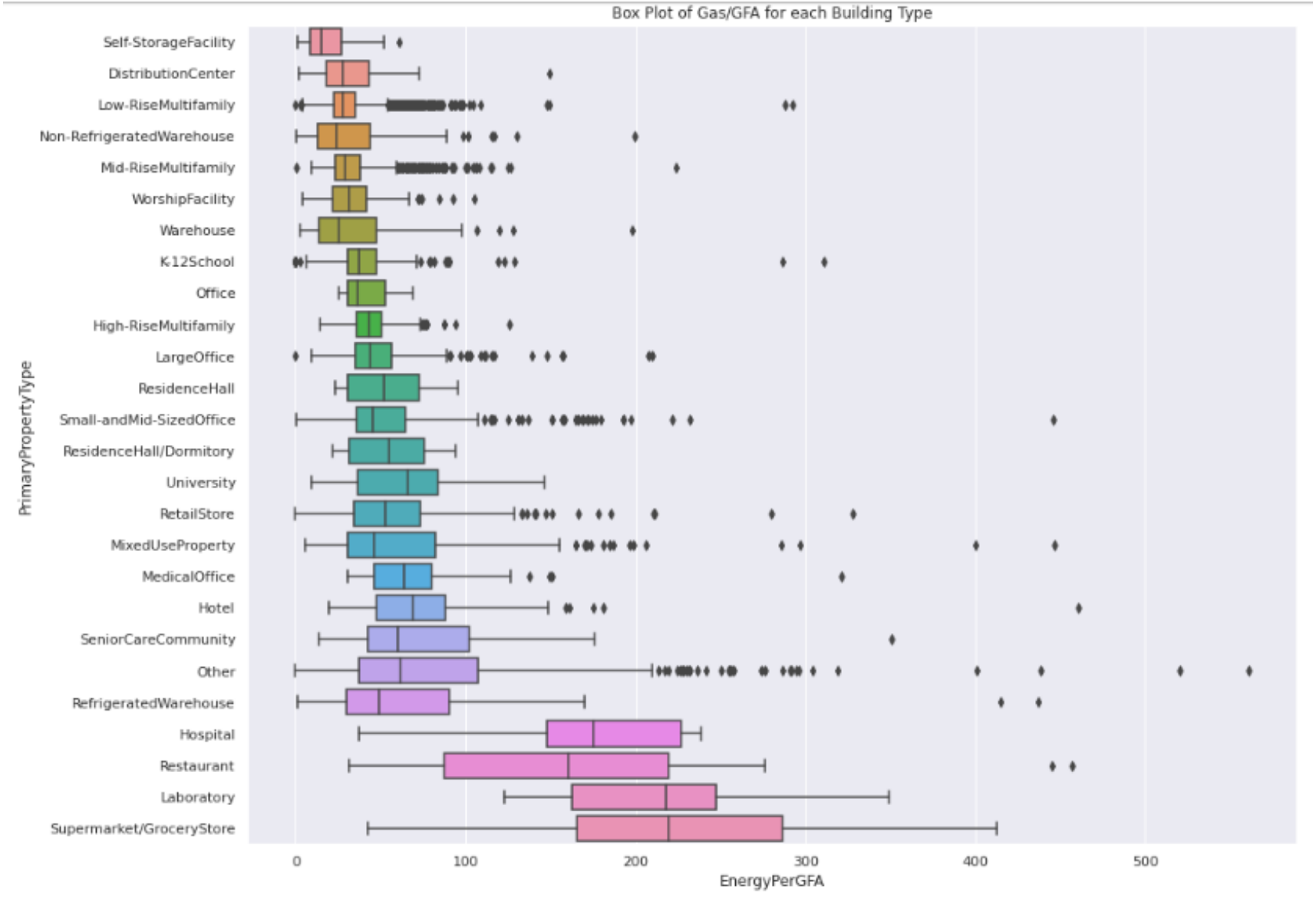



# BoxPlots pour l'Electricité Par Surface, et Acticité (1/2)



# BoxPlots pour l'Energie Par Surface, et Acticité (2/2)



# Prediction de la consommation d'energie

# Selection des variables

- Variables Prédites:
  - 'SiteEnergyUse(kBtu)'
  - 'SiteEnergyUseWN(kBtu)'

# Selection des variables

- Variables Prédites:
  - 'SiteEnergyUse(kBtu)'
  - 'SiteEnergyUseWN(kBtu)'

- Prédicteurs:
  - 'YearBuilt'
  - 'NumberofBuildings'
  - 'NumberofFloors'
  - 'PropertyGFATotal'
  - 'PrimaryPropertyType'

# Selection des variables

- Variables Prédites:
  - 'SiteEnergyUse(kBtu)'
  - 'SiteEnergyUseWN(kBtu)'

- Prédicteurs:
  - 'YearBuilt'
  - 'NumberofBuildings'
  - 'NumberofFloors'
  - 'PropertyGFATotal'
  - 'PrimaryPropertyType'
    - Variable catégorielle
    - Création de variables Dummy pour chaque catégorie

# Regression Linéaire Simple

- Séparation des données en Training set / Test Set
- Calibrage sur données Training
- Application sur données Test

# Regression Linéaire Simple

- Séparation des données en Training set / Test Set
- Calibrage sur données Training
- Application sur données Test

Linear Regressions:

| | R2 : | Adj_R2 : | MSE : | MAE : | Time : |
|---|---|---|---|---|---|
| Site Energy Training : | 0.874 | 0.871 | 6349188783594 | 1202060 | |
| Site Energy Test : | 0.846 | 0.842 | 6826114074273 | 1243040 | 0:00:00 |
| Site Energy WN Training : | 0.851 | 0.848 | 7637384738935 | 1303953 | |
| Site Energy WN Test : | 0.767 | 0.76 | 10341917225045 | 1372828 | 0:00:00 |

# Regression KNN

## Résultats

KNN Regressions:

| | R2 : | Adj_R2 : | MSE : | MAE : | Time : | Best_Parameters : |
|---|---|---|---|---|---|---|
| Site Energy Training : | 0.879 | 0.877 | 4988598186153 | 949235 | | n_neighbors |
| Site Energy Test : | 0.751 | 0.745 | 13390075616126 | 1459303 | 0:00:10 | 3 |
| Site Energy WN Training : | 0.847 | 0.845 | 6425990927808 | 1053602 | | n_neighbors |
| Site Energy WN Test : | 0.701 | 0.694 | 16364801787257 | 1600953 | 0:00:10 | 3 |

- Les résultats sur Training set sont meilleurs qu'avec une regression linéaire
- Les résultats sur Test set sont moins qu'avec une regression linéaire

# Regression KNN

## Résultats

KNN Regressions:

| | R2 : | Adj_R2 : | MSE : | MAE : | Time : | Best_Parameters : |
|---|---|---|---|---|---|---|
| Site Energy Training : | 0.879 | 0.877 | 4988598186153 | 949235 | | n_neighbors |
| Site Energy Test : | 0.751 | 0.745 | 13390075616126 | 1459303 | 0:00:10 | 3 |
| Site Energy WN Training : | 0.847 | 0.845 | 6425990927808 | 1053602 | | n_neighbors |
| Site Energy WN Test : | 0.701 | 0.694 | 16364801787257 | 1600953 | 0:00:10 | 3 |

- Les résultats sur Training set sont meilleurs qu'avec une regression linéaire
- Les résultats sur Test set sont moins qu'avec une regression linéaire
  - $\Longrightarrow$ Overfitting

# Random Forest

## Résultats

Random Forests:

| | R2 : | Adj_R2 : | MSE : | MAE : | Time : |
|---|---|---|---|---|---|
| Site Energy Training : | 0.993 | 0.992 | 305109543659 | 260440 | |
| Site Energy Test : | 0.901 | 0.898 | 5338028930968 | 758795 | 0:00:23 |
| Site Energy WN Training : | 0.985 | 0.985 | 630241648627 | 315132 | |
| Site Energy WN Test : | 0.844 | 0.841 | 8533227522055 | 891306 | 0:00:24 |

Les résultats sont nettement meilleurs qu'avec une régression Linéaire

# XG Boost

## Résultats

```
XG Boost :

                          Best_Parameters :

                          ['colsample_bytree', 'learning_rate', 'max_depth', 'n_estimators']
Site Energy :             [0.7, 0.3, 10, 100]
Site EnergyWN :           [0.7, 0.1, 6, 100]


                                  R2 :          Adj_R2 :         MSE :           MAE :          Time :

Site Energy Training :            0.992           0.992       336352840005       327117
Site Energy Test :                0.828           0.824       9245957687979      1095179        0:08:12

Site Energy WN Training :         0.923           0.922       3248206192913      1002228
Site Energy WN Test :             0.685           0.678      17252060399108      1478353        0:00:24
```

Les résultats sont moins bons qu'avec une Random Forest, et le calibrage des hyper-paramêtres est très long.

# Résumé des résultats

Energy:

| | R2 : | Adj_R2 : | MSE : | MAE : | Time : |
|---|---|---|---|---|---|
| Linear Regression Training : | 0.828 | 0.825 | 7106407037678 | 1374362 | |
| Linear regression Test : | 0.798 | 0.793 | 10860701658017 | 1481664 | 0:00:00 |
| KNN Training : | 0.879 | 0.877 | 4988598186153 | 949235 | |
| KNN Test : | 0.751 | 0.745 | 13390075616126 | 1459303 | 0:00:10 |
| Forest Random Training : | 0.993 | 0.992 | 305109543659 | 260440 | |
| Site Energy Test : | 0.901 | 0.898 | 5338028930968 | 758795 | 0:00:23 |
| XG Boost Training : | 0.992 | 0.992 | 336352840005 | 327117 | |
| XG Boost Test : | 0.828 | 0.824 | 9245957687979 | 1095179 | 0:08:12 |

Energy WN:

| | R2 : | Adj_R2 : | MSE : | MAE : | Time : |
|---|---|---|---|---|---|
| Linear Regression Training : | 0.778 | 0.775 | 9315952695173 | 1511172 | |
| Linear Regression Test : | 0.764 | 0.759 | 12927794952783 | 1624568 | 0:00:00 |
| KNN Training : | 0.847 | 0.845 | 6425990927808 | 1053602 | |
| KNN Test : | 0.701 | 0.694 | 16364801787257 | 1600953 | 0:00:10 |
| Random Forest Training : | 0.985 | 0.985 | 630241648627 | 315132 | |
| Random Forest Test : | 0.844 | 0.841 | 8533227522055 | 891306 | 0:00:24 |
| XG Boost Training : | 0.923 | 0.922 | 3248206192913 | 1002228 | |
| XG Boost Test : | 0.685 | 0.678 | 17252060399108 | 1478353 | 0:00:24 |

# Mises à jour

- Je n'ai pas utilisé les outliers dans les données Trainning
  - Suppression des Outliers
  - Division Training / Test set
  - Utilisation du Training Set pour le calibrage des modèles

- J'ai remis les outliers dans le Test set
  - Utilisation des données contenant les Outliers
  - Division Training / Test set
  - Utilisation du Test Set pour évaluer la prédictivité des modèles

# NOUVEAU Résumé des résultats (1/2)

Energy:

| | R2 : | Adj_R2 : | MSE : | MAE : | Time : |
|---|---|---|---|---|---|
| Linear Regression Training : | 0.882 | 0.88 | 8344451785965 | 1428852 | |
| Linear regression Test : | 0.606 | 0.598 | 62531297942341 | 2120524 | 0:00:00 |
| KNN Training : | 0.888 | 0.886 | 7922830897272 | 975517 | |
| KNN Test : | 0.507 | 0.496 | 78326708820622 | 1948605 | 0:00:19 |
| Forest Random Training : | 0.985 | 0.985 | 1060286947585 | 276125 | |
| Site Energy Test : | 0.843 | 0.839 | 24977898085900 | 872741 | 0:00:37 |
| XG Boost Training : | 0.995 | 0.995 | 342005361962 | 362004 | |
| XG Boost Test : | 0.596 | 0.587 | 64267155512356 | 1451984 | 0:41:48 |

Energy WN:

| | R2 : | Adj_R2 : | MSE : | MAE : | Time : |
|---|---|---|---|---|---|
| Linear Regression Training : | 0.85 | 0.848 | 10711487075682 | 1564062 | |
| Linear Regression Test : | 0.595 | 0.586 | 67124137496466 | 2273833 | 0:00:00 |
| KNN Training : | 0.873 | 0.872 | 9039260314990 | 1049636 | |
| KNN Test : | 0.487 | 0.475 | 85056151119334 | 2090616 | 0:00:19 |
| Random Forest Training : | 0.977 | 0.977 | 1631438683884 | 328852 | |
| Random Forest Test : | 0.801 | 0.796 | 33008623954404 | 983507 | 0:00:37 |
| XG Boost Training : | 0.993 | 0.992 | 534092004929 | 361101 | |
| XG Boost Test : | 0.588 | 0.579 | 68227550433224 | 1485396 | 0:00:37 |

# NOUVEAU Résumé des résultats (2/2)

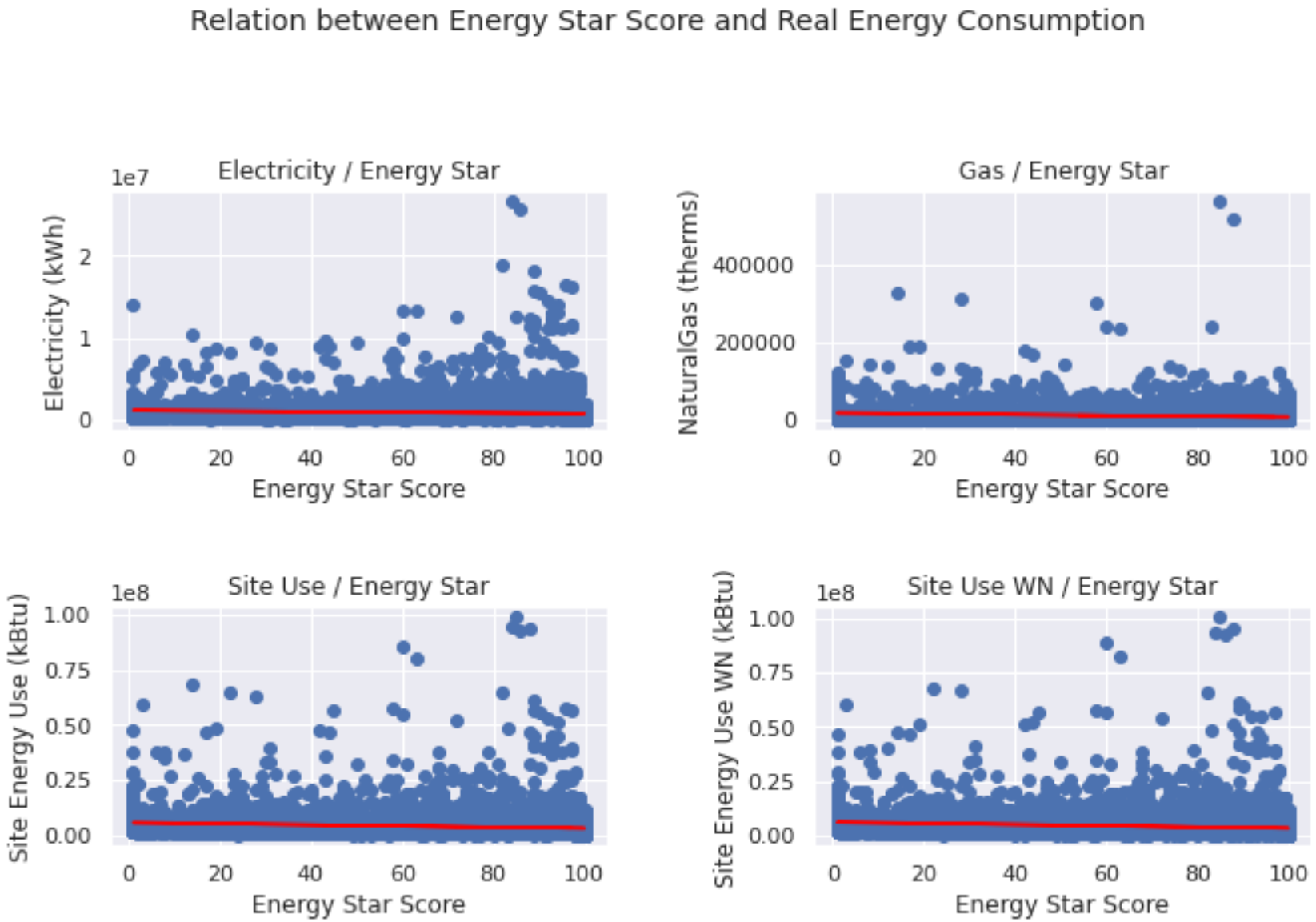
CO2:

| | R2 : | Adj_R2 : | MSE : | MAE : | Time : |
|---|---|---|---|---|---|
| Linear Regression Training : | 0.79 | 0.787 | 11270 | 50 | |
| Linear Regression Test : | 0.618 | 0.61 | 39892 | 67 | 0:00:00 |
| KNN Training : | 0.822 | 0.819 | 9552 | 31 | |
| KNN Test : | 0.528 | 0.518 | 49270 | 56 | 0:00:54 |
| Random Forest Training : | 0.961 | 0.961 | 2079 | 14 | |
| Random Forest Test : | 0.738 | 0.732 | 27398 | 39 | 0:00:37 |
| XG Boost Training : | 0.944 | 0.944 | 2976 | 32 | |
| XG Boost Test : | 0.679 | 0.672 | 33563 | 55 | 0:00:37 |

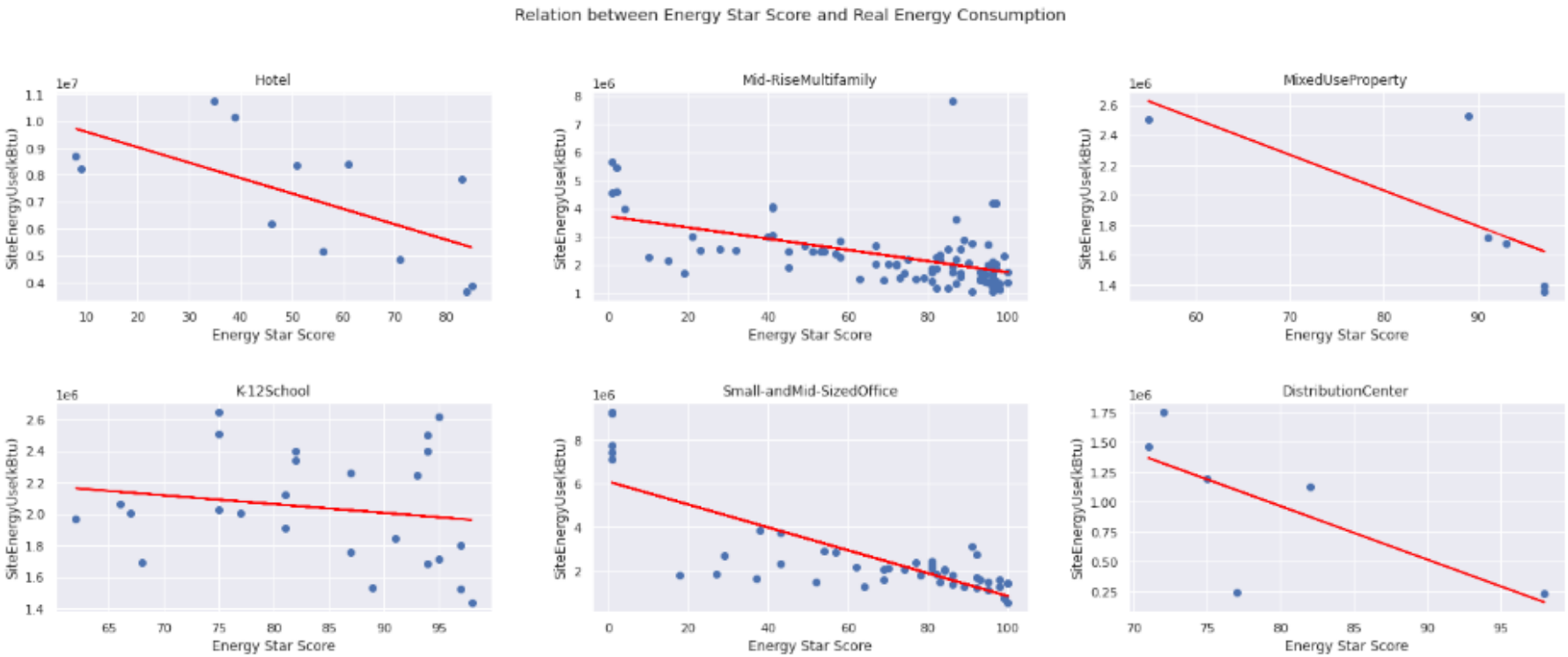
- Les résultats sur Training Set sont à peu près similaires aux résultats précédemment obtenus
  - car le training set ne contient pas les outliers
- Les résultats sur Test Set sont significativement descendus
  - car le test set contient les outliers

# Evaluation du EnergyStar Score

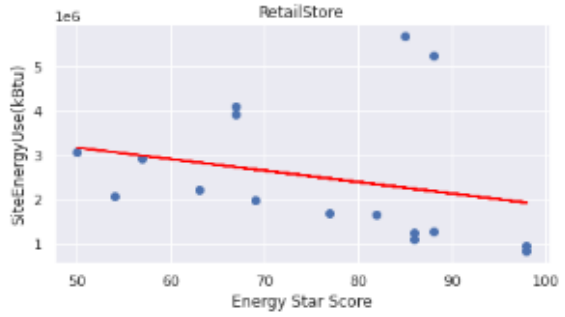
# Energy Star Score et consommation réelle d'énergie

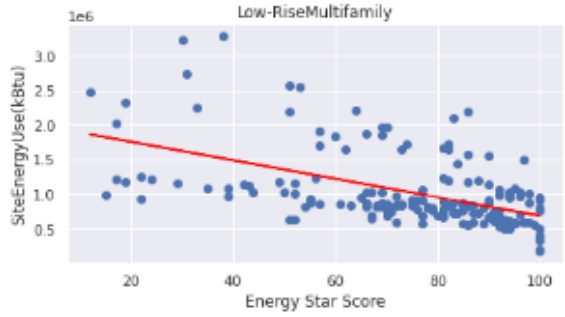


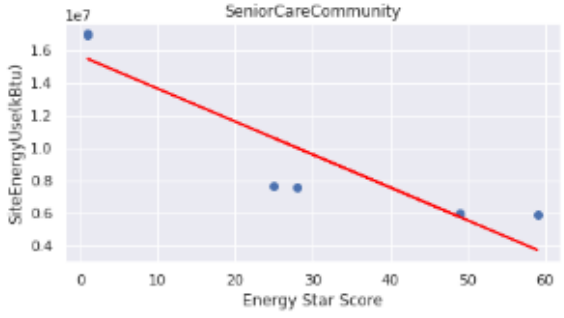
# Energy Star Score et consommation réelle d'énergie (1/2)

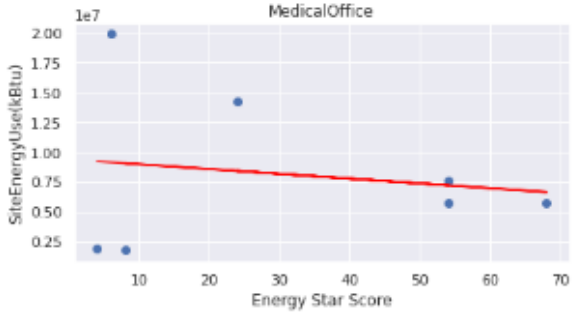


# Energy Star Score et consommation réelle d'énergie (2/2)

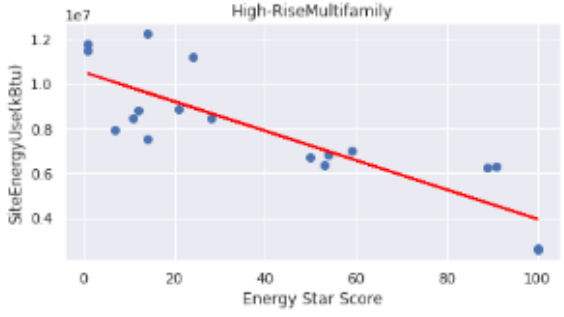



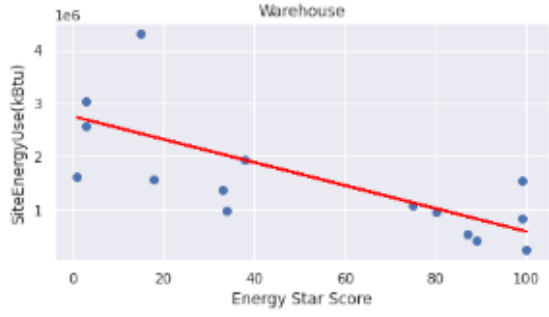







# Inclure Energy Star Score pour Prédire CO2

Energy:

| | R2 : | Adj_R2 : | MSE : | MAE : | Time : |
|---|---|---|---|---|---|
| Linear Regression: | | | | | |
| Training : | 0.874 | 0.871 | 6349188783594 | 1202060 | |
| Training with Energy Star Score : | 0.889 | 0.887 | 5563153807345 | 1191838 | |
| Test : | 0.846 | 0.842 | 6826114074273 | 1243040 | 0:00:00 |
| Test with Energy Star Score : | 0.869 | 0.865 | 5833684856705 | 1252979 | 0:00:00 |
| Random Forest : | | | | | |
| Training : | 0.993 | 0.992 | 305109543659 | 260440 | |
| Training with Energy Star Score : | 0.992 | 0.992 | 392474532589 | 239609 | |
| Test : | 0.901 | 0.898 | 5338028930968 | 758795 | 0:00:23 |
| Test with Energy Star Score : | 0.925 | 0.923 | 3324540215284 | 671789 | 0:00:16 |

# Conclusion

# Conclusion

- Nettoyage des données
  - Suppression de 2.099 % des observations
- Prédiction de la consommation énergétique
  - Régression Linéaire
  - Régression KNN
  - Random Forest
    - $R^2 = 0.92$ pour 'SiteEnergy'
  - XG Boost
- Inclure le Energy Star Score?
  - Data Leakage
  - Apporte seulement 0.03 supplémentaire sur le $R^2$