

# **Formation de Data Science - Openclassrooms**

Formation Ouverte et à Distance – FOAD par Pôle Emploi  
Solutions 100% à distance

## **Projet 4 : Anticipez les besoins en consommation électrique de bâtiments**

Étudiant : Maria Daniela Barrios  
Mentor : Dan Slama

Février, 2022

# Contexte du problème

- Le projet consiste à créer des stratégies basées sur des données pour la ville de Seattle, aux États-Unis. Pour atteindre l'objectif d'une ville neutre en carbone d'ici 2050, l'équipe porte une attention particulière aux émissions des bâtiments non résidentiels
- Mission : prédire les **émissions de CO<sub>2</sub>** et la **consommation totale d'énergie** de bâtiments pour lesquels elles n'ont pas encore été mesurées.
- Les données de consommation peuvent être téléchargées à cette adresse : <https://www.kaggle.com/city-of-seattle/sea-building-energy-benchmarking#2015-building-energy-benchmarking.csv>
- Les prédictions seront basées sur les données déclaratives du permis d'exploitation commerciale (taille et utilisation des bâtiments, mention de travaux récents, date de construction, etc.)

# Stratégie pour réaliser la mission

- Phase pré-exploratoire : Analyse générale et découverte des fichiers
  - Décrire les informations contenues dans l'ensemble de données: nombre de lignes et de colonnes
  - Sélection et explication d'indicateurs (variables) pertinents
  
- Analyse exploratoire et nettoyage des données
  - Exploration des valeurs manquantes et nettoyage des données
  - Analyse de la corrélation des variables
  - Détection des valeurs aberrantes et normalisation des données
  
- Analyse des prédictions
  - Test de deux algorithmes de prédiction différents: Ordinary Least Squares Linear Regression et Random Forest Regressor
  - Prédiction de la consommation et des émissions de CO<sub>2</sub>

# Phase pré-exploratoire : Analyse générale et découverte des fichiers

Deux fichiers contenant :

**2015-building-energy-benchmarking.csv :**

- **3340** lignes et **47** colonnes

**2016-building-energy-benchmarking.csv :**

- **3376** lignes et **47** colonnes

Les fichiers de données des années 2015 et 2016 sont différents. Nous pouvons vérifier quelles colonnes sont différentes :

## Colonnes communes

```
(Index(['OSEBuildingID', 'DataYear', 'BuildingType', 'PrimaryPropertyType',  
       'PropertyName', 'TaxParcelIdentificationNumber', 'CouncilDistrictCode',  
       'Neighborhood', 'YearBuilt', 'NumberofBuildings', 'NumberofFloors',  
       'PropertyGFATotal', 'PropertyGFAParking', 'PropertyGFABuilding(s)',  
       'ListOfAllPropertyUseTypes', 'LargestPropertyUseType',  
       'LargestPropertyUseTypeGFA', 'SecondLargestPropertyUseType',  
       'SecondLargestPropertyUseTypeGFA', 'ThirdLargestPropertyUseType',  
       'ThirdLargestPropertyUseTypeGFA', 'YearsENERGYSTARCertified',  
       '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)',  
       'DefaultData', 'ComplianceStatus', 'Outlier'],  
      dtype='object'))
```

## Colonnes présentes dans le fichier 2015 et non dans le fichier 2016 :

```
Index(['2010 Census Tracts', 'City Council Districts', 'Comment',  
       'GHGEmissions(MetricTonsCO2e)', 'GHGEmissionsIntensity(kgCO2e/ft2)',  
       'Location', 'OtherFuelUse(kBtu)', 'SPD Beats',  
       'Seattle Police Department Micro Community Policing Plan Areas',  
       'Zip Codes'],  
      dtype='object'))
```

## Colonnes présentes dans le fichier 2016 et non dans le fichier 2015 :

```
Index(['Address', 'City', 'Comments', 'GHGEmissionsIntensity', 'Latitude',  
       'Longitude', 'State', 'TotalGHGEmissions', 'ZipCode'],  
      dtype='object'))
```

# Phase pré-exploratoire : Analyse générale et découverte des fichiers

Comme la mission du projet concerne les émissions de CO<sub>2</sub>, nous devons accorder une attention particulière aux variables :

- **GHGEmissionsIntensity** : total des émissions Greenhouse Gas divisé par la surface brute de la propriété (mesurée en kilogrammes d'équivalent de dioxyde de carbone par pied carré)
- **TotalGHGEmissions** : quantité totale d'émissions de Greenhouse Gases, y compris le dioxyde de carbone, le méthane et l'oxyde nitreux libérés dans l'atmosphère en raison de la consommation d'énergie de la propriété (mesurée en tonnes métriques d'équivalent de dioxyde de carbone)
- **ENERGYSTARScore** : Une note de 1 à 100 qui évalue la performance énergétique globale d'un bien immobilier
- **SiteEUI(kBtu/sf)** : Energy Use Intensity du site (EUI) divisée par sa surface brute
- **SourceEUI(kBtu/sf)** : Energy Use Intensity à la source (EUI) divisée par la surface
- **SiteEnergyUse(kBtu)** : La quantité annuelle d'énergie consommée par la propriété, toutes sources d'énergie comprises
- **Electricity(kBtu)** : La quantité annuelle d'électricité consommée par le bien sur place
- **NaturalGas(kBtu)** : La quantité annuelle de gaz naturel fourni par le distributeur et consommé par la propriété.

<https://data.seattle.gov/dataset/2016-Building-Energy-Benchmarking/2bpz-gwpy>

# Phase pré-exploratoire : Analyse générale et découverte des fichiers

D'autres variables importantes :

- **OSEBuildingID**
- **DataYear**
- **BuildingType**
- **PrimaryPropertyType**
- **Latitude**
- **Longitude**
- **Address**
- **Neighborhood**
- **YearBuilt**
- **NumberOfBuildings**
- **NumberOfFloors**
- **PropertyGFATotal**
- **PropertyGFABuilding(s)**

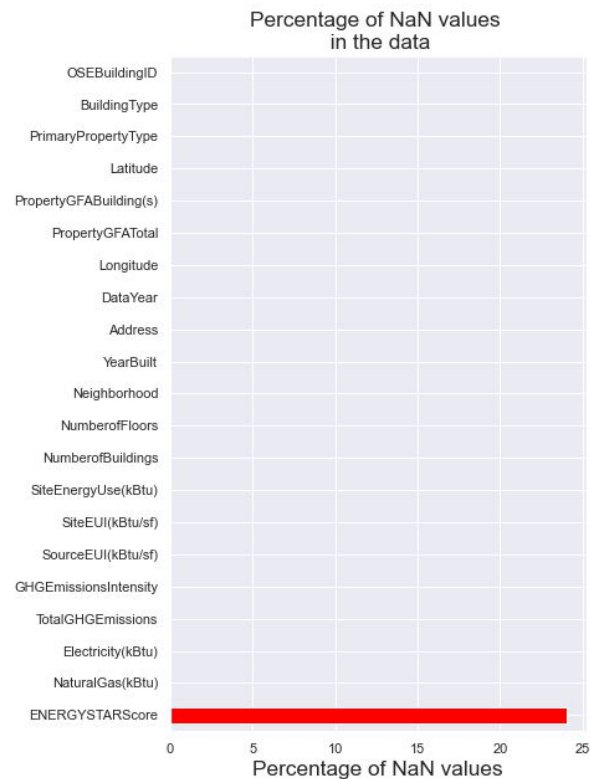
Identifiant unique attribué à chaque propriété : utile pour trouver des données doublées

Le nombre de valeurs dupliquées dans les variables "OSEBuildingID" du fichiers pour 2015 et pour 2016 était de 0

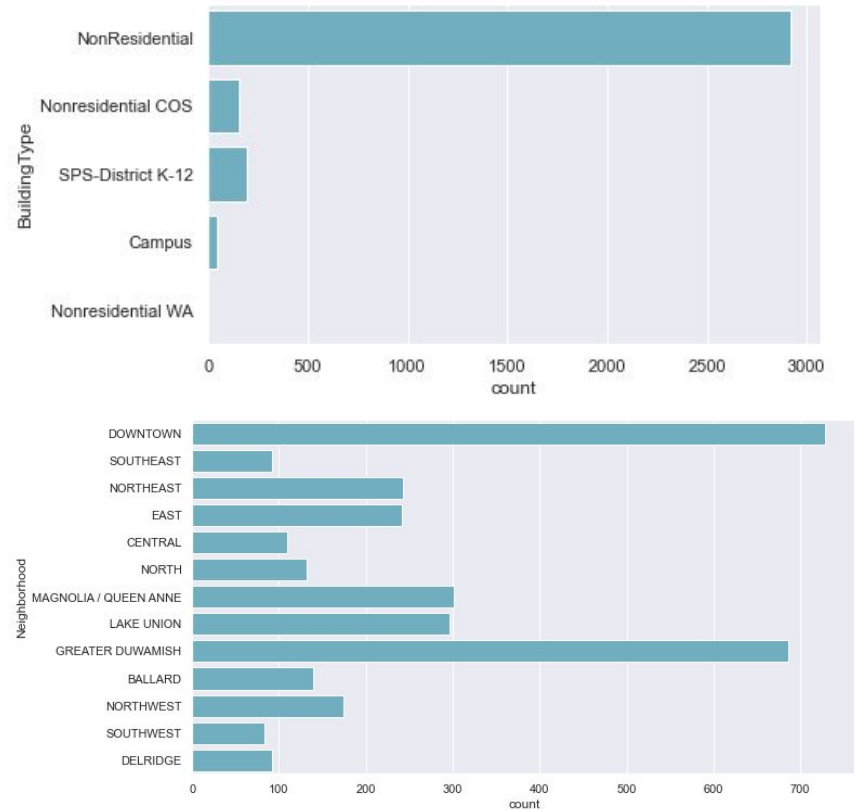
Fichier de données unique avec les données des années 2015 et 2016 avec les variables pertinentes contient **6716** lignes et 19 colonnes

# Analyse exploratoire et nettoyage des données

## Valeurs manquantes parmi les variables

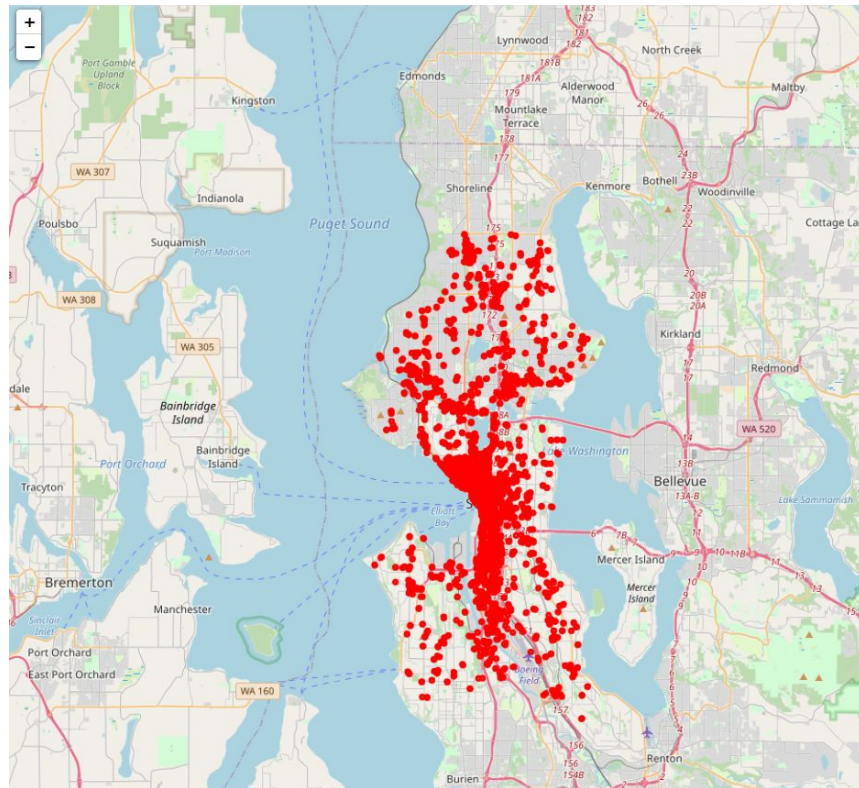
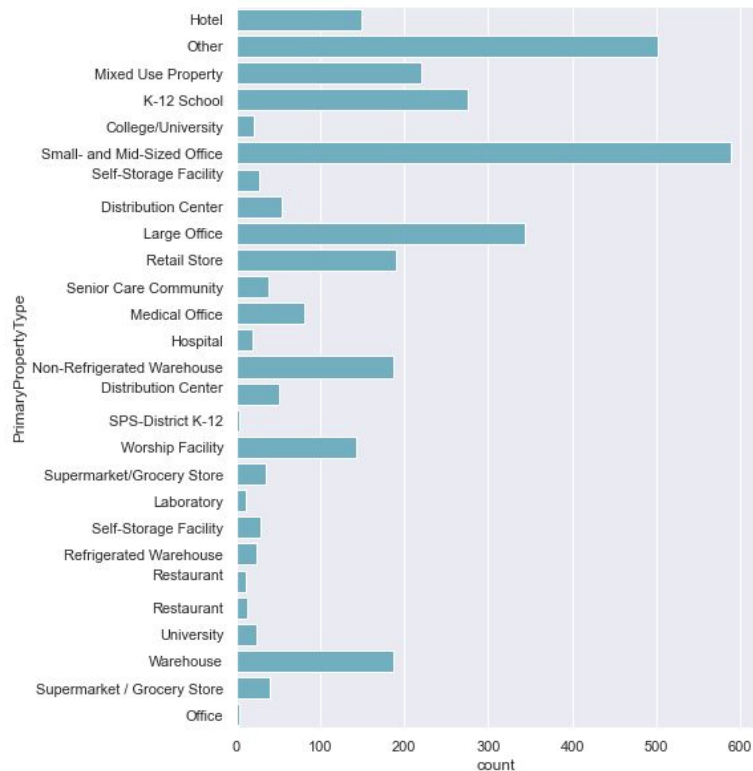


## Bâtiments non résidentiels



# Analyse exploratoire et nettoyage des données

## Bâtiments non résidentiels et son emplacement sur la carte de Seattle





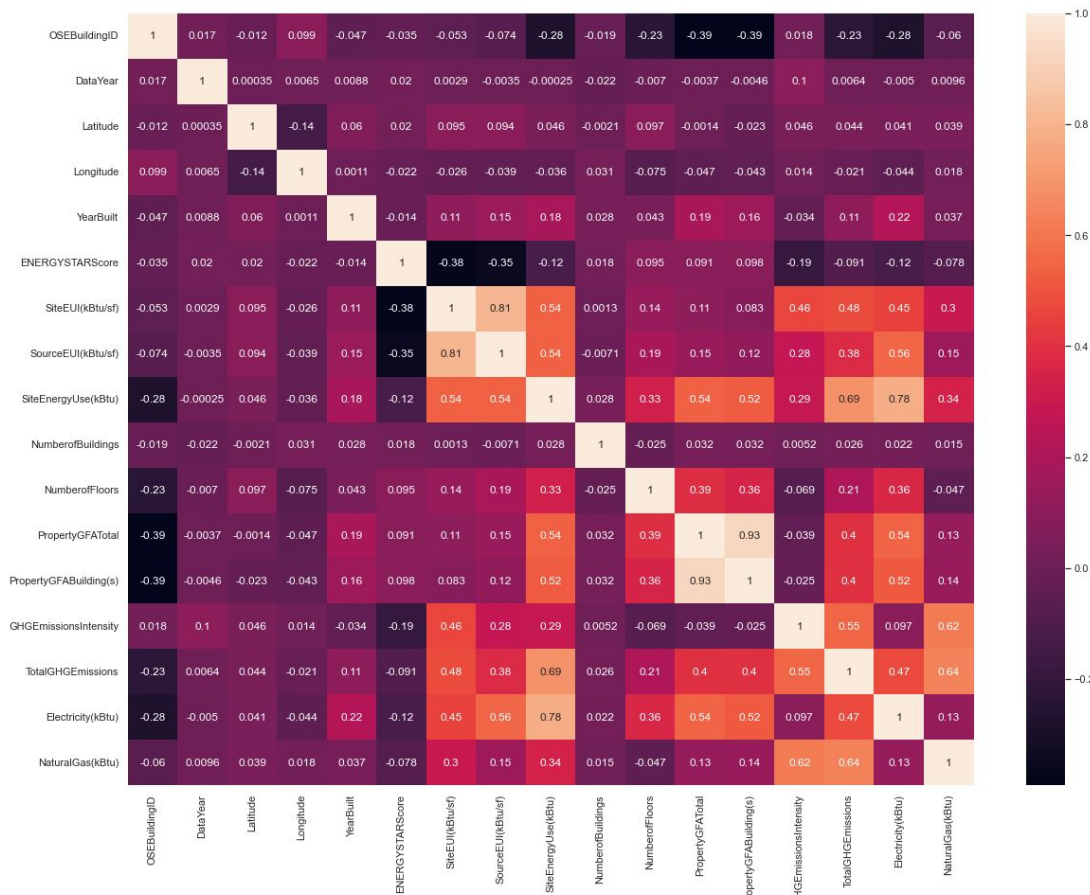
# Analyse exploratoire et nettoyage des données

**Méthode de Kendall** : le coefficient de corrélation mesure la relation monotone entre deux variables. Il n'est pas nécessaire que les variables soient normalement distribuées

PropertyGFATotal	PropertyGFABuilding(s)	0.927988
SourceEUI(kBtu/sf)	SiteEUI(kBtu/sf)	0.809493
Electricity(kBtu)	SiteEnergyUse(kBtu)	0.782865

Afin d'éviter le sur-apprentissage, nous devons **éliminer** les variables à forte corrélation :

**'SourceEUI(kBtu/sf)' et  
'PropertyGFABuilding(s)',  
'Electricity(kBtu)'**



# Analyse exploratoire et nettoyage des données

- Avant de traiter les valeurs aberrantes, nous avons remplacé les valeurs manquantes par la valeur des médianes
- Nous avons également transformé les variables catégorielles en variables numériques

Int64Index: 3264 entries, 0 to 6715

Data columns (total 9 columns):

#	Column	Non-Null	Count	Dtype
0	PrimaryPropertyType	3264	non-null	int8
1	Neighborhood	3264	non-null	int8
2	YearBuilt	3264	non-null	int64
3	SiteEnergyUse(kBtu)	3264	non-null	float64
4	NumberOfBuildings	3264	non-null	float64
5	NumberOfFloors	3264	non-null	float64
6	PropertyGFATotal	3264	non-null	int64
7	TotalGHGEmissions	3264	non-null	float64
8	NaturalGas(kBtu)	3264	non-null	float64

dtypes: float64(5), int64(2), int8(2)

memory usage: 339.4 KB

## Pourquoi faire feature engineering ?

Nous avons aussi transformé les variables qui sont fortement asymétriques en faisant une transformation logarithmique

## Pourquoi normaliser ou transformer ?

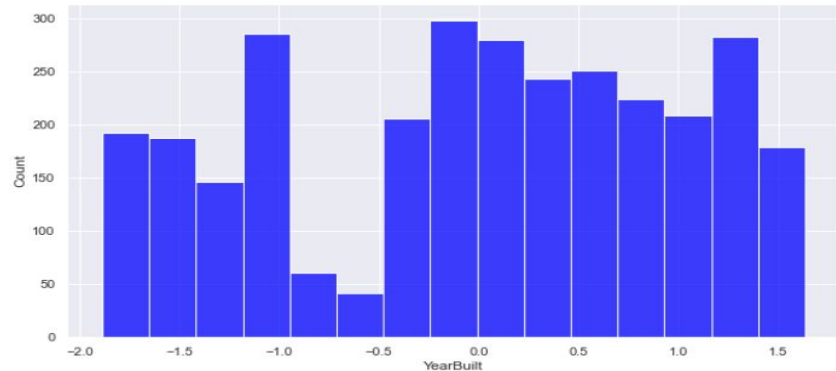
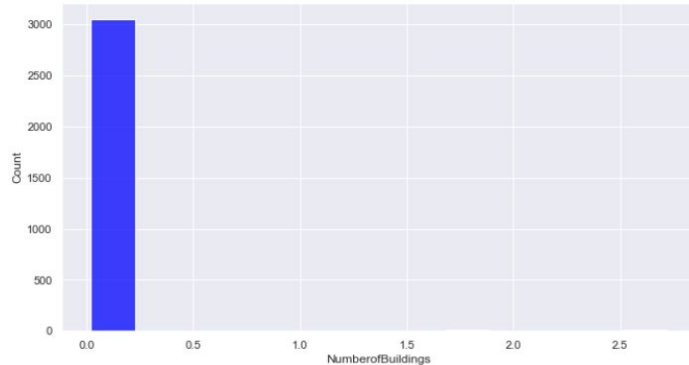
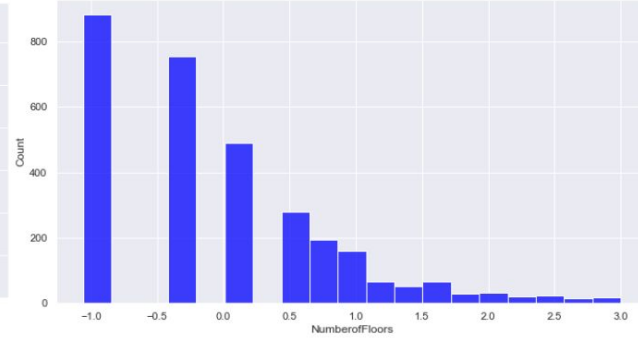
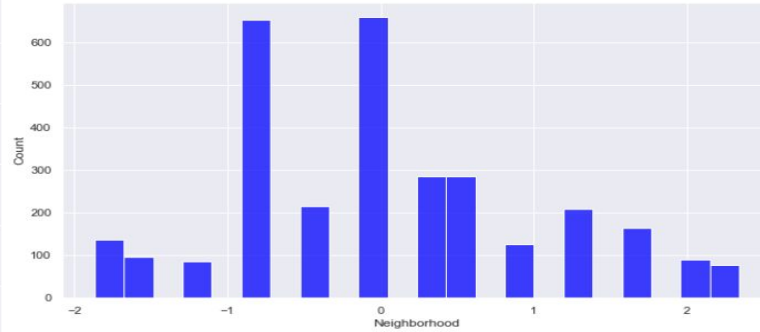
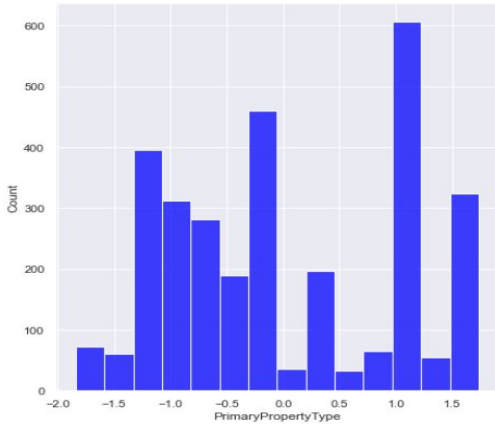
De nombreux algorithmes d'apprentissage automatique tentent de trouver des tendances dans les données en comparant les caractéristiques des points de données. Cependant, un problème se pose lorsque les caractéristiques sont à des échelles radicalement différentes

## Z-scores pour détecter les valeurs aberrantes

Nous avons utilisé la standardisation des z-scores en considérant une tolérance de  $\pm 3$  des z-scores

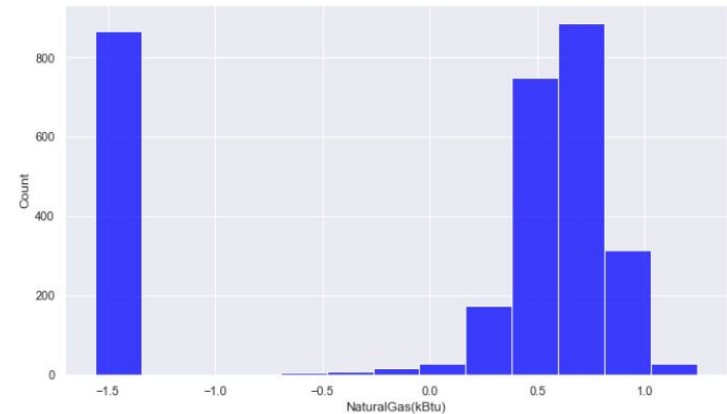
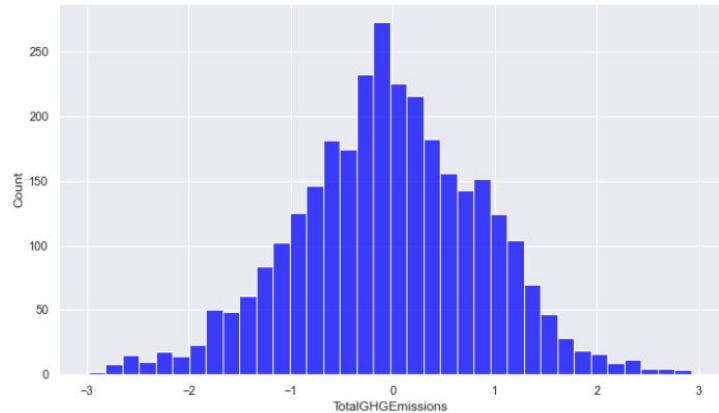
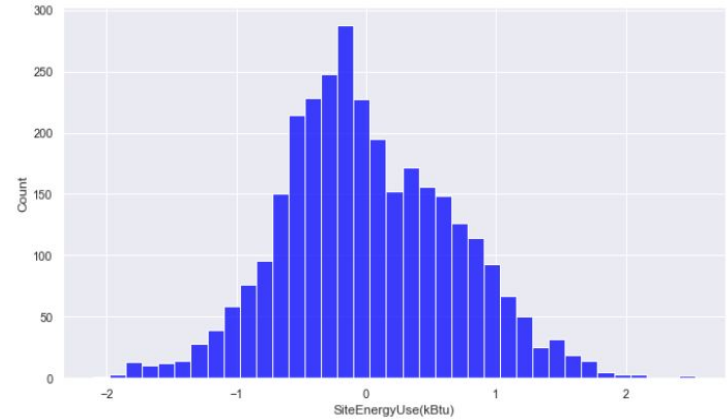
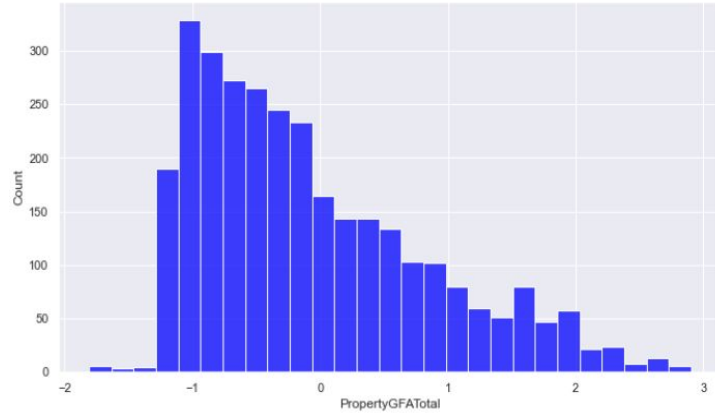
# Analyse exploratoire et nettoyage des données

## Données transformées en utilisant log-transform

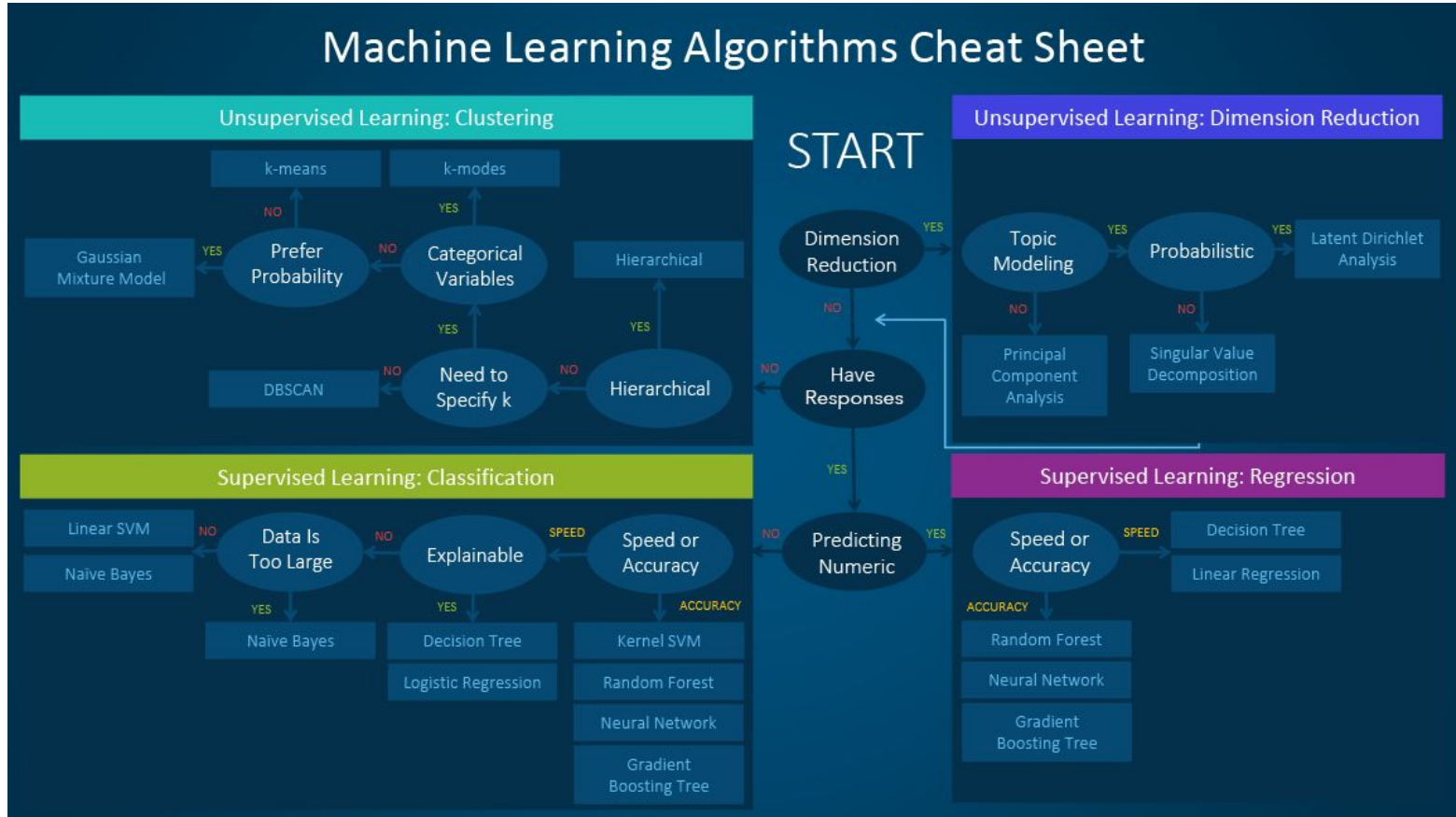


# Analyse exploratoire et nettoyage des données

Données transformées en utilisant log-transform



# Analyse exploratoire et nettoyage des données



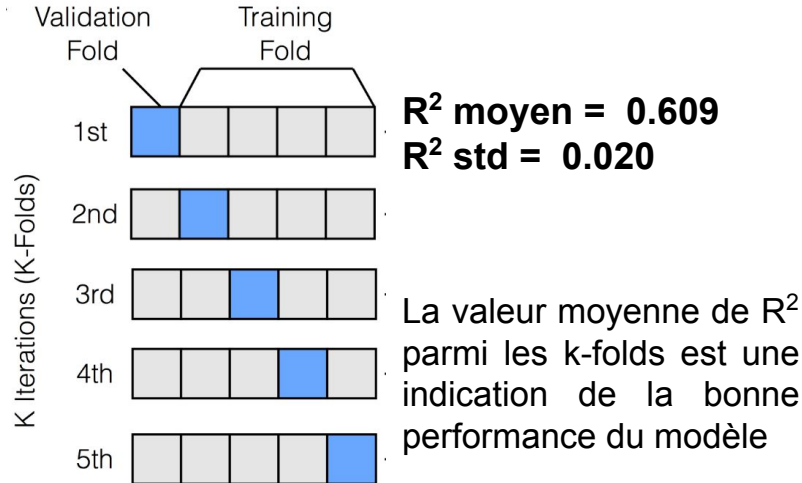
<https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/>

# Analyse des prédictions

- Fractionnement des données : nous utilisons 80 % pour la training et 20 % pour le test
- Nous avons éliminé des données d'entrée les deux variables cibles : 'SiteEnergyUse(kBtu)' et 'TotalGHGEmissions' - Nous avons d'abord choisi une variable pour tester les algorithmes: 'TotalGHGEmissions'

## Analyse par régression linéaire (moindres carrés ordinaires - ordinary least squares)

Nous effectuons une validation croisée K-Fold pour évaluer la performance du modèle de **régression linéaire (moindres carrés ordinaires)**



Après d'entraîner le modèle de régression linéaire, on teste le modèle dans des données non vues et les résultats sont les suivants

**Données de test  $R^2 = 0.598$**

**MSE = 0.337**

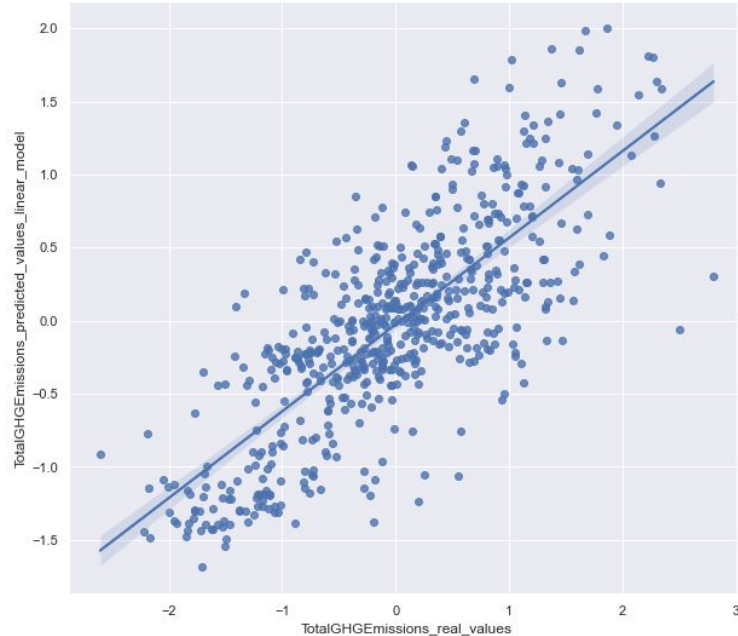
**MAE = 0.440**

La valeur  $R^2$  dans le sous-ensemble de données de test suggère que notre modèle régressif s'est bien adapté aux données non vues

# Analyse des prédictions

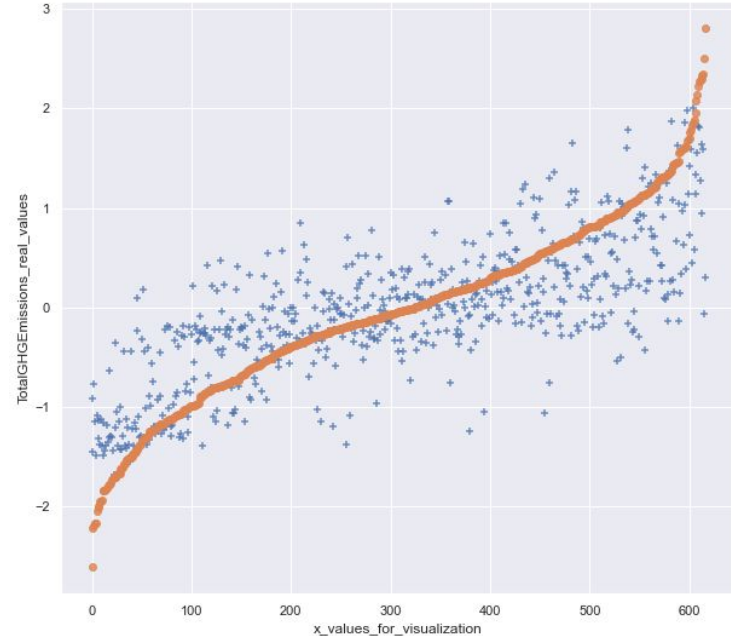
## Analyse par régression linéaire (moindres carrés ordinaires - ordinary least squares)

Nous constatons la tendance à une corrélation linéaire entre les valeurs prédites et les valeurs réelles



Points bleus : données prédites

Points orange : données réelles

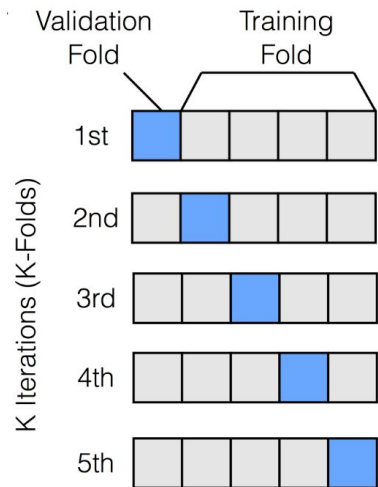


# Analyse des prédictions

## Analyse utilisant une régression par forêt aléatoire (random forest regressor)



Nous avons effectué aussi une validation croisée K-Fold pour évaluer la performance du modèle de **régression par forêt aléatoire**



**$R^2$  moyen = 0.893**

**$R^2$  std = 0.023**

La valeur moyenne de  $R^2$  parmi les k-folds est encore mieux que  $R^2$  que du modèle de régression linéaire (moindres carrés ordinaires)

Après avoir optimisé le modèle avec une validation croisée et après l'avoir entraîné, on teste le modèle dans des données non vues et les résultats sont les suivants

**Données de test  $R^2 = 0.912$**

**MSE = 0.074**

**MAE = 0.158**

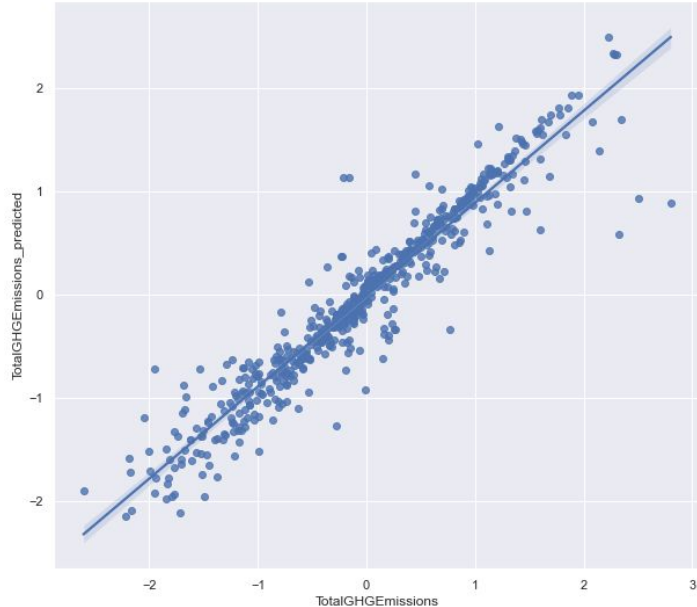
La valeur  $R^2$  dans le sous-ensemble de données de test suggère que le modèle de régression par forêt aléatoire s'est bien adapté aux données non vues est encore mieux que le modèle de régression linéaire (moindres carrés ordinaires)



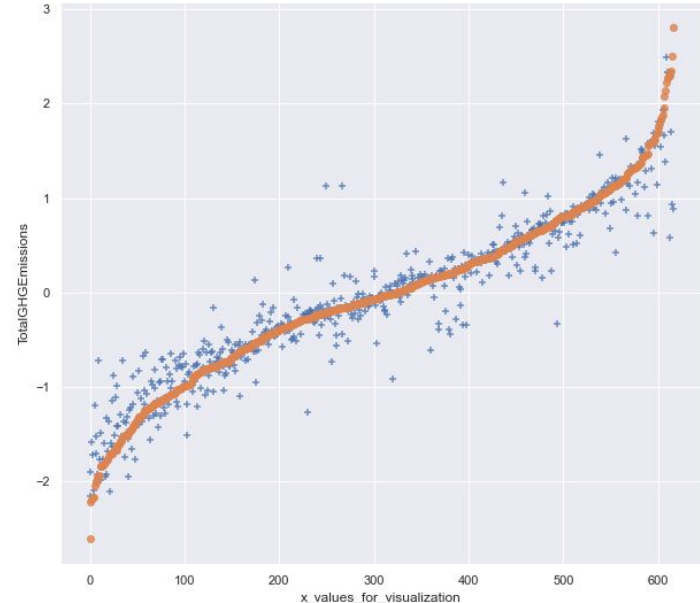
# Analyse des prédictions

## Analyse utilisant une régression par forêt aléatoire (random forest regressor)

Nous observons que la performance du régresseur de la forêt aléatoire est supérieure à celle de la régression linéaire des moindres carrés ordinaires



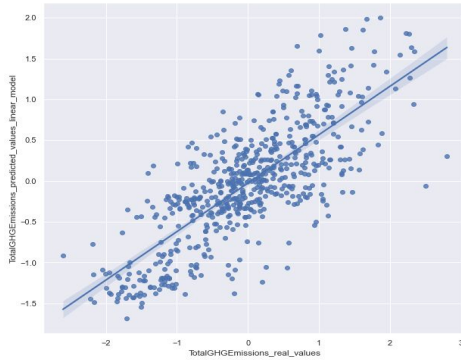
Points bleus : données prédites  
Points orange : données réelles



# Analyse des prédictions

Comparaison entre les résultats des prédictions utilisant la régression linéaire des moindres carrés et le régresseur de la forêt aléatoire

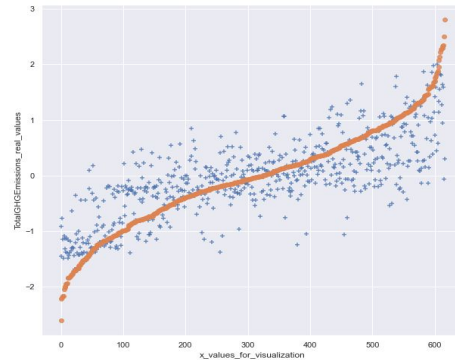
## Linear regression- ordinary least squares



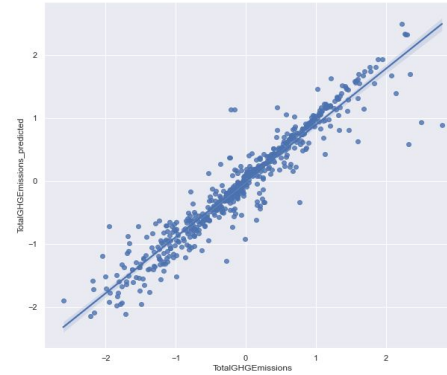
Données de test  $R^2 = 0.598$

**MSE = 0.337**

**MAE = 0.440**



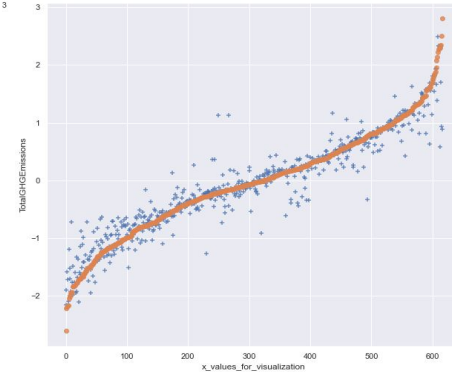
## Random forest regressor



Données de test  $R^2 = 0.912$

**MSE = 0.074**

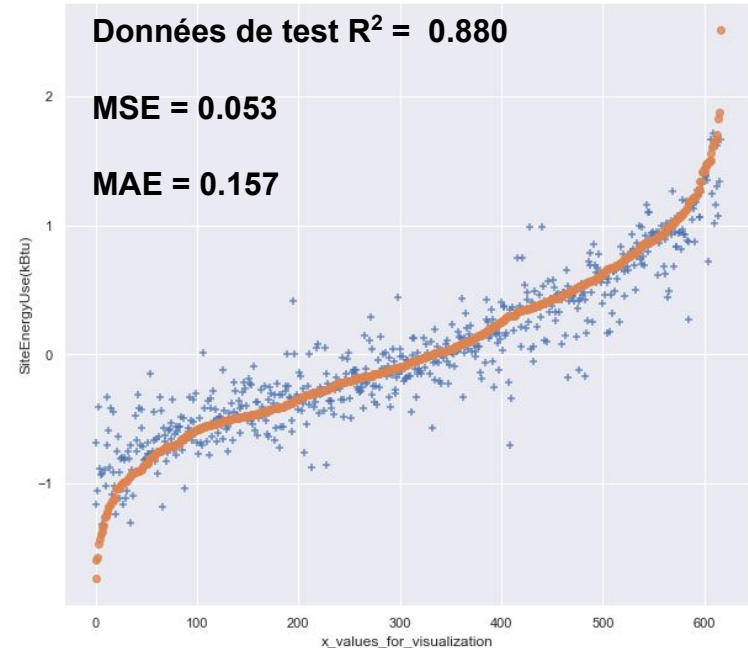
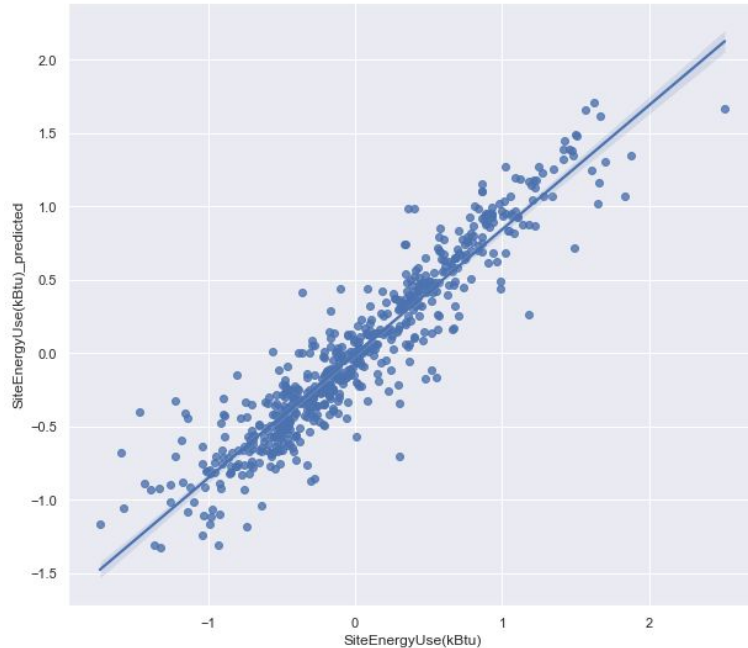
**MAE = 0.158**



# Analyse des prédictions

## Analyse utilisant une régression par forêt aléatoire (random forest regressor)

En utilisant le régresseur de la forêt aléatoire, nous allons montrer la prédiction de la consommation d'énergie en choisissant la variable '**SiteEnergyUse(kBtu)**'



# Conclusions

- Après une sélection appropriée des variables d'intérêts pour réaliser cette mission, nous avons évalué deux modèles de prédiction, basés sur le besoin de prédictions numériques :
  - ◆ Régression linéaire (moindres carrés ordinaires - ordinary least squares)
  - ◆ Régression par forêt aléatoire (random forest regressor)
- Nous avons obtenu une meilleure performance du modèle random forest regressor en termes de fidélité des prédictions
  - ◆ Il est important de noter qu'en choisissant ce modèle, on compromet les ressources informatiques en augmentant le temps de calcul
  - ◆ Bien que la régression par forêt aléatoire a eu une performance élevée, nous avons effectué une validation croisée pour optimiser le modèle et nous avons obtenu des résultats proches
- Nous pouvons utiliser un régresseur de forêt aléatoire pour prédire les valeurs d'ENERGY STAR Score pour la performance énergétique des bâtiments
- On recommande également d'utiliser une approche en utilisant des réseaux neuronaux, qui offrent également une bonne précision de prédiction