

# Anticipez les besoins en consommation électrique de bâtiments

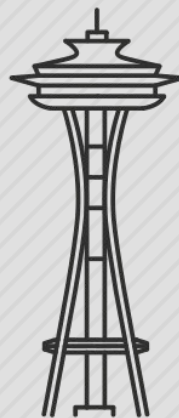
...

Projet 4 Parcours Data Scientist  
Python - Support Jupyter Notebook

Nalron - Novembre 2020  
OpenClassrooms . Centrale Supélec

# Mission abordée en 4 principales étapes

- Réflexion sur la problématique et présentation des données
- Traitement, exploration et feature engineering
- Phase de modélisation avec testing des hyperparamètres
- Évaluation des performances et choix du modèle final





# Problématique de la ville de Seattle



**Objectif : Ville neutre en émissions de carbone en 2050.**  
Traitement des émissions des bâtiments non destinés à l'habitation.  
Interprétation et pistes de recherche envisagées.

# Présentation de la problématique

## Relevés effectués par des agents sur les années 2015 et 2016 :

- Coûts élevés des relevés de consommation annuels
- Opérations fastidieuses
- Caractères limitatifs

## Solutions envisagées :

- Prédire la consommation totale d'énergie de bâtiments
- Prédire les émissions de CO2 avec ou sans ENERGY STAR Score
- Modèles de prédiction performants et réutilisables



# Présentation des données






Relevés de consommation annuels. Données déclaratives du permis d'exploitation commerciale (taille et usage des bâtiments, mention de travaux récents, date de construction..)

# Données de consommation

Source :

<https://www.kaggle.com/city-of-seattle/sea-building-energy-benchmarking#2015-building-energy-benchmarking.csv>

Open Data Seattle : <https://data.seattle.gov/>

			
2015	3256	24	99.9%
2016	3336	27	96.0%
	Bâtiments	Millions M2	Tx de conformité

## Données de consommation 2015

- Observations : 3340
- Variables : 47
- % NaN : 16,88
- Aucune observation dupliquée

## Données de consommation 2016

- Observations : 3376
- Variables : 46
- % NaN : 12,84
- Aucune observation dupliquée



# Traitement des données

...

Nettoyage, duplications, outliers, valeurs manquantes, etc...  
Agrégation des données pour établir un nouvel échantillon de travail  
adapté à la problématique métier.

# Concaténation des données 2015-2016

Création d'un échantillon de référence nécessitant des opérations de cleaning, traitement de valeurs manquantes, etc...

---

*Un échantillon de 46 variables est obtenu,  
mais sont-elles toutes pertinentes?*

## Traitement des colonnes :

- Identification des écarts
- Renommage des colonnes

## Données obtenues :

- Observations : 6716
- Variables : 46
- % NaN : 13
- Aucun relevé de conso. dupliqué



# % des données manquantes par variable

Première sélection des variables selon un seuil fixé à **50%** des données inconnues.

*La variable 'ENERGYSTARScore' est conservée pour être optimisée par imputation méthode KNN.*

	Total	%
Comment	6703	99.81
Outlier	6600	98.27
YearsENERGYSTARCertified	6487	96.59
ThirdLargestPropertyUseType	5560	82.79
ThirdLargestPropertyUseTypeGFA	5560	82.79
SecondLargestPropertyUseTypeGFA	3478	51.79
SecondLargestPropertyUseType	3478	51.79
ENERGYSTARScore	1623	24.17
LargestPropertyUseType	156	2.32
LargestPropertyUseTypeGFA	156	2.32
ListOfAllPropertyUseTypes	136	2.03
Electricity(kBtu)	19	0.28
Electricity(kWh)	19	0.28
NaturalGas(therms)	19	0.28
NaturalGas(kBtu)	19	0.28
SourceEUIWN(kBtu/sf)	19	0.28
SourceEUI(kBtu/sf)	19	0.28
TotalGHGEmissions	19	0.28
GHGEmissionsIntensity	19	0.28
SteamUse(kBtu)	19	0.28
SiteEUI(kBtu/sf)	17	0.25
SiteEUIWN(kBtu/sf)	16	0.24
SiteEnergyUseWN(kBtu)	16	0.24

ZipCode	16	0.24
SiteEnergyUse(kBtu)	15	0.22
NumberOfFloors	8	0.12
NumberOfBuildings	8	0.12
TaxParcelIdentificationNumber	2	0.03
DefaultData	1	0.01
DataYear	0	0.00
ComplianceStatus	0	0.00
State	0	0.00
City	0	0.00
Address	0	0.00
Longitude	0	0.00
Latitude	0	0.00
BuildingType	0	0.00
PropertyName	0	0.00
PrimaryPropertyType	0	0.00
PropertyGFABuilding(s)	0	0.00
CouncilDistrictCode	0	0.00
Neighborhood	0	0.00
YearBuilt	0	0.00
PropertyGFATotal	0	0.00
PropertyGFAParking	0	0.00
OSEBuildingID	0	0.00

# Outliers?

La variable '**Outlier**' est utilisée pour exclure les valeurs non souhaitables **1.73%**.

Suppression de quelques valeurs négatives considérées comme aberrantes.

---

*Variable 'Outlier' utile pour différencier les observations plus influentes.*

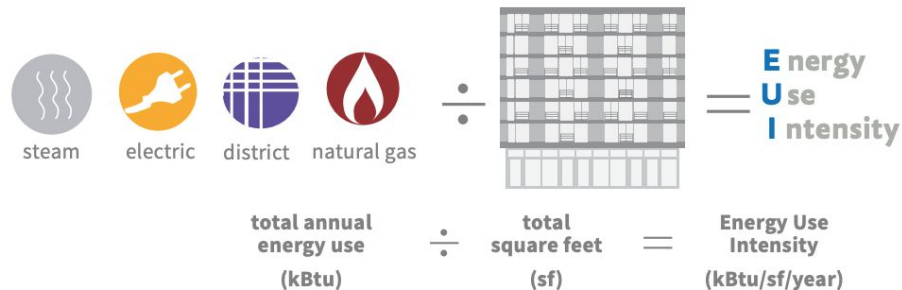
**Plus de 3000 buildings à Seattle**  
une large diversité d'occupants  
impliquent des valeurs atypiques à  
prendre en compte.

---

# Choix des variables selon une logique métier

La sélection des caractéristiques techniques est faite pour conserver uniquement des variables explicatives métiers.

*Lutte contre la fuite de données, coûts des relevés, calculs complexes...*



source : <http://www.seattle.gov/>

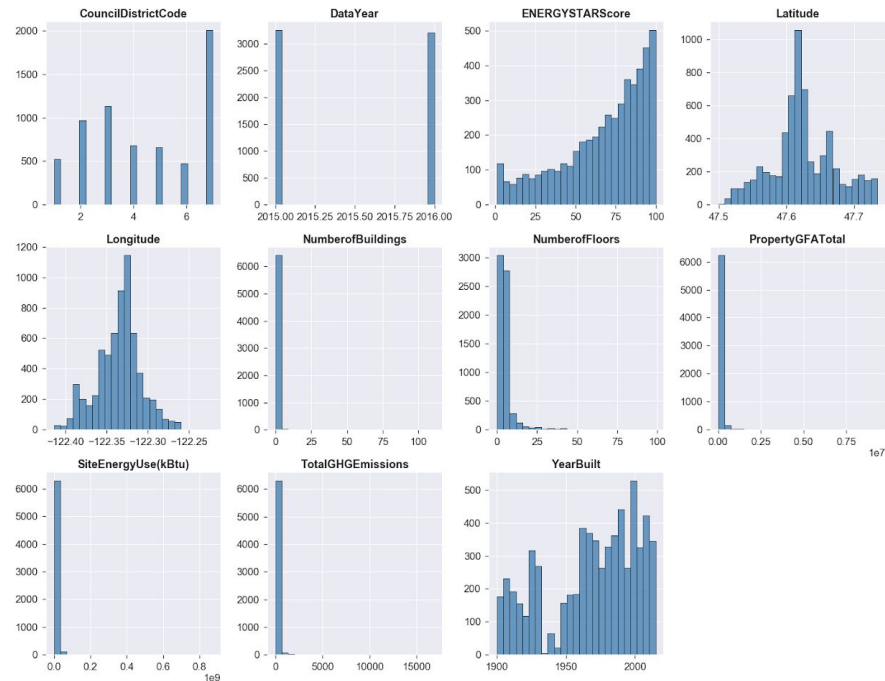
**Variable cible : SiteEnergyUse(kBtu)**  
**Autre target : TotalGHGEmissions**

	count	mean	std	min	25%	50%	75%	max
ENERGYSTARScore	4887.0	6.771619e+01	2.674852e+01	1.000000	52.000000	7.500000e+01	9.000000e+01	1.000000e+02
TotalGHGEmissions	6450.0	1.154833e+02	4.860752e+02	0.080000	9.522500	3.309000e+01	9.030250e+01	1.687098e+04
SiteEnergyUse(kBtu)	6450.0	5.242291e+06	1.843106e+07	11441.000000	930482.500000	1.787052e+06	4.140252e+06	8.739237e+08
NumberofFloors	6442.0	4.746042e+00	5.545013e+00	0.000000	2.000000	4.000000e+00	5.000000e+00	9.900000e+01
NumberofBuildings	6450.0	1.073643e+00	1.624171e+00	0.000000	1.000000	1.000000e+00	1.000000e+00	1.110000e+02
Longitude	6450.0	-1.223350e+02	2.686159e-02	-122.414250	-122.350455	-1.223326e+02	-1.223199e+02	-1.222205e+02
Latitude	6450.0	4.762439e+01	4.760876e-02	47.499331	47.600411	4.761890e+01	4.765717e+01	4.773387e+01
DataYear	6450.0	2.015497e+03	5.000282e-01	2015.000000	2015.000000	2.015000e+03	2.016000e+03	2.016000e+03
CouncilDistrictCode	6450.0	4.460310e+00	2.119925e+00	1.000000	3.000000	4.000000e+00	7.000000e+00	7.000000e+00
YearBuilt	6450.0	1.968360e+03	3.296413e+01	1900.000000	1948.000000	1.975000e+03	1.997000e+03	2.015000e+03
PropertyGFATotal	6450.0	9.355935e+04	1.901342e+05	11285.000000	28371.000000	4.399800e+04	9.000000e+04	9.320156e+06

# Distribution des variables quantitatives

Type de distribution spécifique aux caractéristiques techniques expliquées.

*La diversité des bâtiments implique nécessairement quelques valeurs atypiques.*



## Nécessité de transformation

Problèmes d'échelle et de distribution

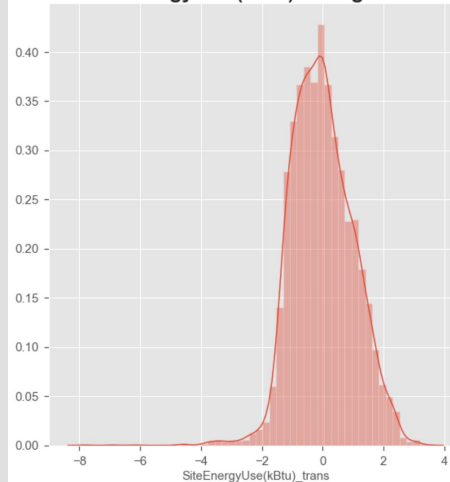
# Distribution des variables cibles

Transformation par méthode **Box-Cox**.

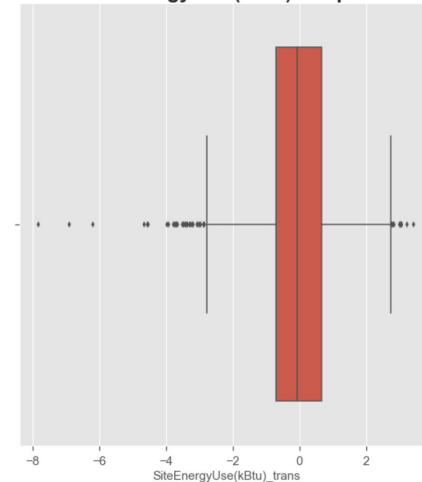
Valeurs négatives aberrantes exclues dans le traitement des outliers.

*Des valeurs extrêmes sont toujours palpables mais acceptables dans le contexte de la diversité des types de bâtiments.*

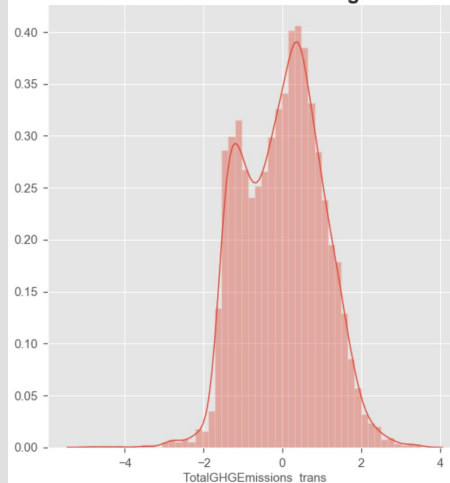
Site Energy Use(kBtu) Histogramme



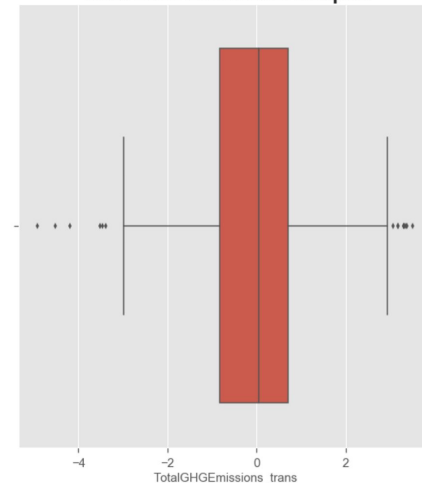
Site Energy Use(kBtu) Boxplot



Total GHG Emissions Histogramme



Total GHG Emissions Boxplot



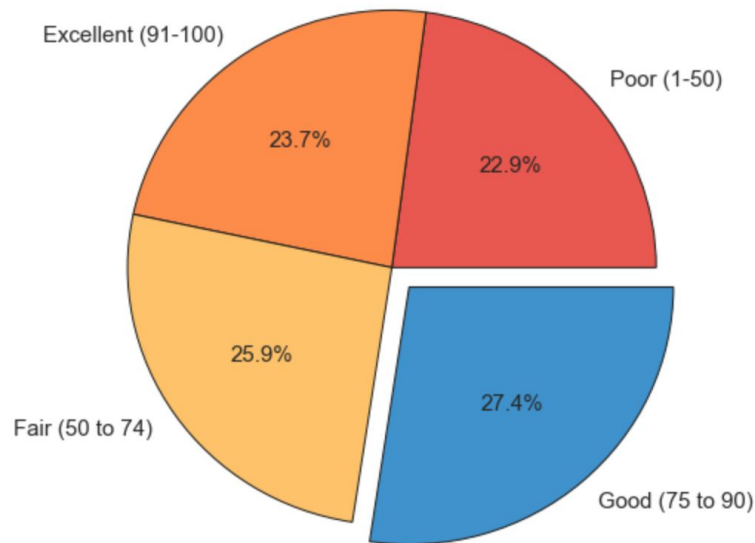
# Répartition de la performance énergétique

L'ENERGY STAR score indique le niveau de performance énergétique d'un bâtiment.

---

*Indicateur représentatif de l'activité et du comportement des occupants (en autre...).*

ENERGY STAR Performance Category



Fonctionnement des bâtiments de Seattle relativement homogène

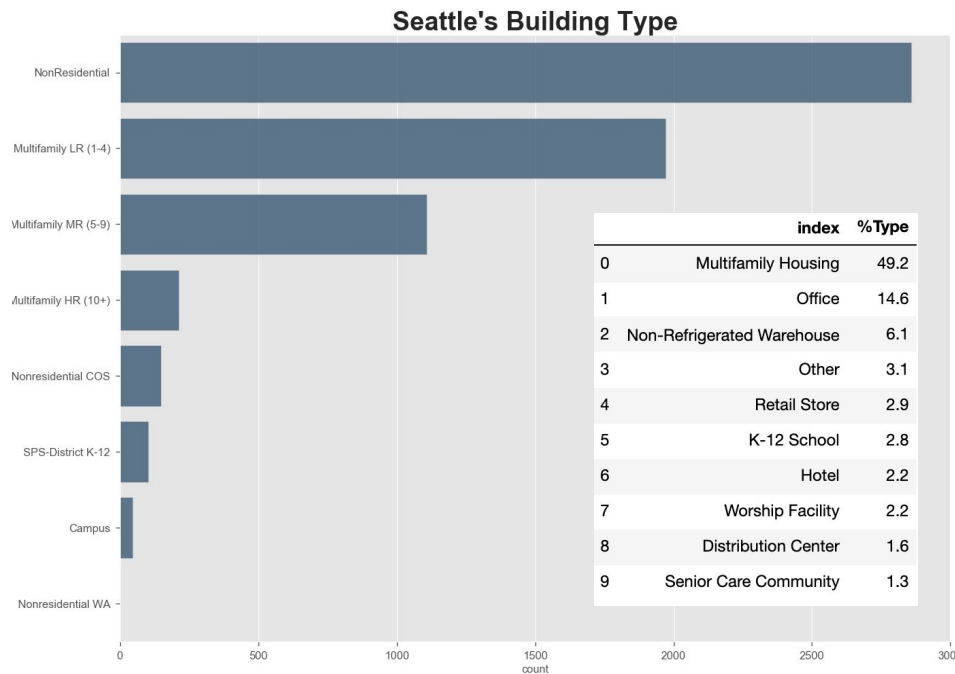
---

# Répartition du type de bâtiments

≈ 65% occupés par le secteur du Multifamilial et du Bureau.

Attention aux autres gros consommateurs identifiés dans l'analyse : hôtels et retailers

*Il est nécessaire d'identifier ces usages non destinés à l'habitation pour atteindre l'objectif fixé à 2050.*



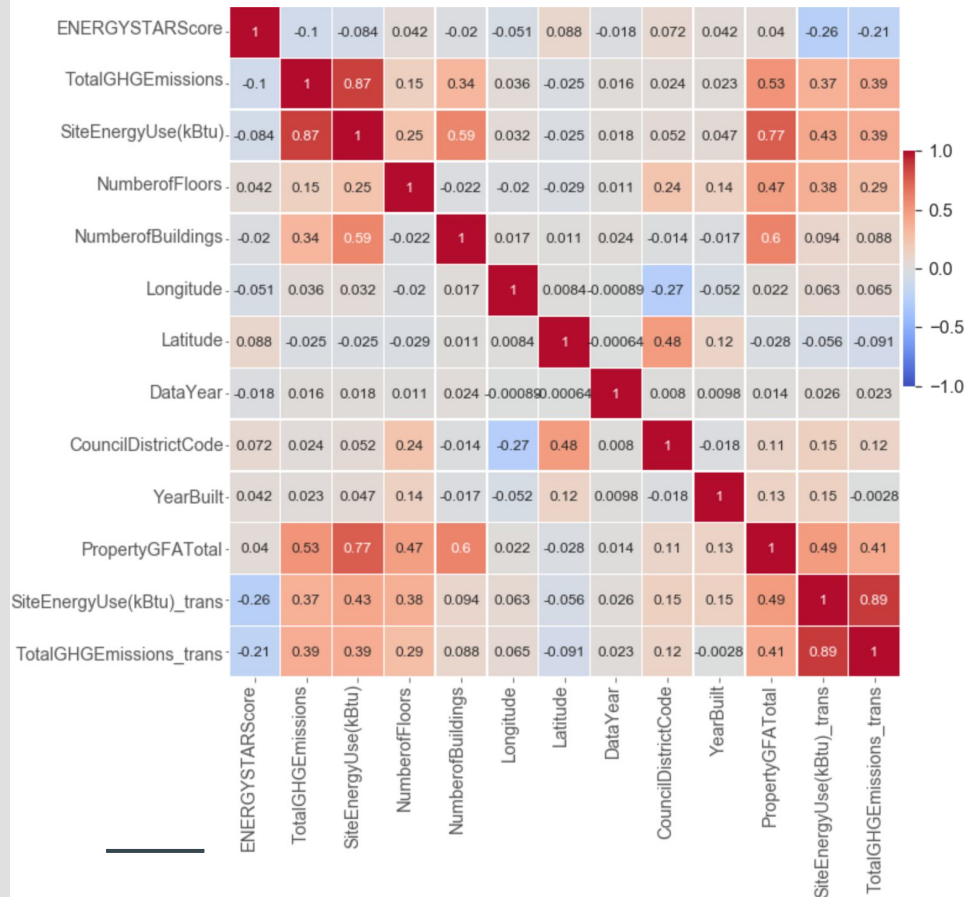
Moins représentés ici, d'autres types de bâtiments énergivores existent !

# Analyse des corrélations

Consommation corrélée avec la superficie totale : 'PropertyGFATotal'.

Émissions corrélées avec la conso. totale : 'SiteEnergyUse(kBtu)'.

*Aucune corrélation linéaire avec la métrique de l'"ENERGY STAR Score".*





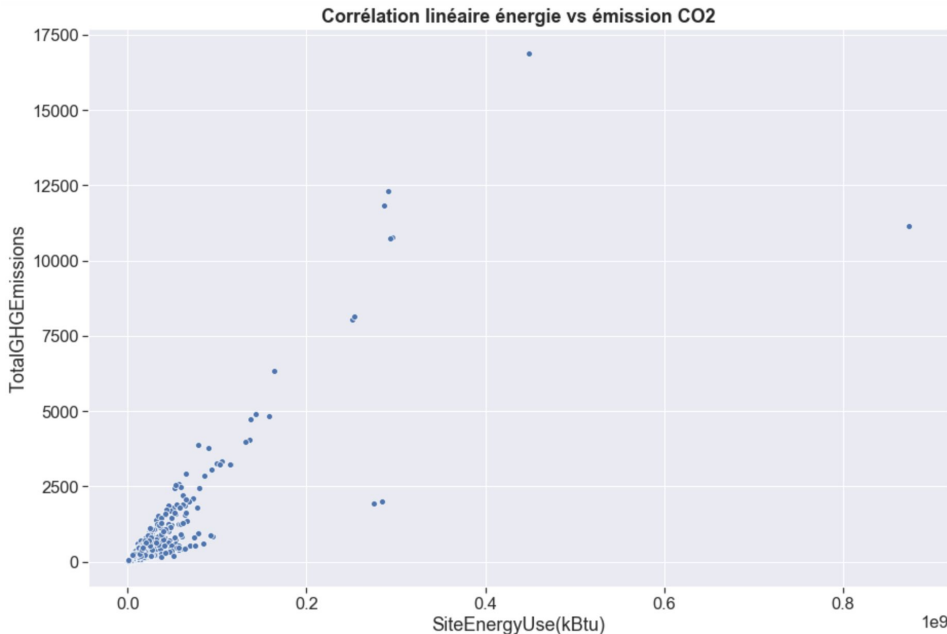
# Visualisation Régression linéaire

L'échantillon comporte quelques valeurs extrêmes...

**Choix de conserver ces valeurs**, décision prise en réponse au contexte métier.

---

*Les algorithmes robustes permettront de modéliser sans exclure ces valeurs.*



*Valeurs atypiques visibles .*  
**Visuellement on retrouve ces formes de distribution à longues queues**



# Modélisation



Le problème posé doit prédire des valeurs.  
Plusieurs indicateurs explicatifs doivent-êtré pris en compte.

Des modèles d'apprentissage supervisé de régression multivariée seront testés et optimisés.

# Optimisation des modalités de l'échantillon

Encodage avec pré-traitement?  
Seuil proposé : **occurrence > 100**

---

*L'avantage est de pouvoir prendre en compte toutes observations en limitant les nouvelles colonnes et en limitant les modalités à un seuil fixé.*

Comptage des modalités...  
Remplacement des  
plus faibles modalités par "**Others**"

Limiter les colonnes  
créées

Éliminer les modalités  
moins significatives

Multifamily Housing	3172
Office	940
Others	844
Non-Refrigerated Warehouse	390
Other	200
Retail Store	187
K-12 School	180
Hotel	142
Worship Facility	139
Distribution Center	106
Name: LargestPropertyUseType, dtype	
Low-Rise Multifamily	1869
Mid-Rise Multifamily	1056
Small- and Mid-Sized Office	556
Others	553
Other	500
Large Office	323
Mixed Use Property	253
High-Rise Multifamily	195
Warehouse	185
Non-Refrigerated Warehouse	181
K-12 School	180
Retail Store	170
Hotel	140
Worship Facility	139
Name: PrimaryPropertyType, dtype: i	

# Feature Scaling

Nos variables quantitatives ne sont pas toutes à la même échelle.

---

*Sans ce pré-traitement on expose la modélisation à des lenteurs et des pertes de performance.*

## Identification des différences :

- Histogrammes de l'analyse des variables démontrent des échelles propres aux métriques

## Technique StandardScaler() :

- Données centrées autour de 0
  - Écart-Type de 1
-

# Transformation des features catégorielles

Des données sur le type de bâtiments, l'usage affecté à prendre en compte.

**4 variables catégorielles** : LargestPropertyUseType, PrimaryPropertyType, BuildingType, Neighborhood

---

*Créer et ajouter des colonnes binaires qui réfèrent ou non la donnée par un 0 ou 1.*

## Variables qualitatives :

- Comptage des modalités
- Restriction selon un seuil
- Catégorisation / remplacement des plus faibles en "Others"

## Encodage One-Hot :

- 17 variables initiales
- 47 variables obtenues
- Taille d'échantillon limitée par l'étape précédente > "Others"

# Découpage de l'échantillon

Données d'entraînement vs. données de test

Utilisation de l'ensemble des données  
d'entraînement et de test par  
**Validation croisée.**

---

*Le but est de conserver une certaine cohérence et  
surtout une représentativité...*

**Fonction `train_test_split()` :**

- Ici **20%** attribués au test
- Est-ce suffisant ?

**Validation croisée :**

- k folds fixé à **5**
- éviter le biais potentiel
- pas d'évaluation unique

# Choix des modèles

Plusieurs algorithmes sont utilisés afin de pouvoir comparer leurs performances.

Critères de performances traités :  
**Time -  $R^2$  - RMSE**

---

*L'erreur des modèles les plus performants sera visualisée graphiquement.*

## Modèles simples :

- LinearRegression
- Ridge
- Lasso
- ElasticNet
- KNeighborsRegressor
- SVR

## Méthodes ensemblistes :

- RandomForestRegressor
  - GradientBoostingRegressor
-

# Hyperparamètres

Après l'ajustement des caractéristiques  
métier, optimisation des principaux  
hyperparamètres : régresseur (*alpha*), arbres  
de décision (*n\_estimators*)...

---

*Approche par dichotomie :  
valeurs espacées > écart à réduire*

```
#HyperParamètres
lr_params = {'fit_intercept': [True, False], 'normalize': [True, False], 'copy_X': [True]}

ridge_params = {'alpha': [1, 0.1, 0.01, 0.001], 'max_iter': [1000], 'random_state': [42], 'tol': [0.001]}

lasso_params = {'alpha': [1, 0.1, 0.01, 0.001], 'max_iter': [1000], 'random_state': [42], 'tol': [0.0001]}

elastic_params = {'alpha': [1, 0.1, 0.01, 0.001], 'max_iter': [1000], 'random_state': [42], 'tol': [0.0001]}

knn_params = {'n_neighbors': list(range(1, 30))}

svr_params = {'gamma': ['scale'], 'epsilon': [0.001, 0.01, 0.1, 1],
              'C': [0.001, 0.01, 0.1, 1, 10], 'tol': [0.001]}

rfr_params = {'n_estimators': [100, 500, 1000], 'max_features': ['auto'], 'n_jobs': [-1],
              'random_state': [42], 'max_depth': [None]}

gradboost_params = {'n_estimators': [100, 500, 1000], 'random_state': [42], 'max_depth': [None]}

mygrids = [lr_params, ridge_params, lasso_params, elastic_params, knn_params, svr_params,
           rfr_params, gradboost_params]
```

2 méthodes sont appliquées :  
**GridSearch** et **RandomizesSearch**  
pour couvrir l'espace des valeurs  
pertinentes des hyperparamètres



# Performances de Modélisation conso. énergie

8 Algorithmes testés et comparés.  
Sans et avec validation croisée.

*La tendance générale est meilleure avec  
**GridSearch.***

## Cible: consommation d'énergie

	lr	ridge	lasso	elastic	knn	svr	rfr	gradboost
Standard Time	0.0278831	2.07178	2.36884	2.86369	3.40243	16.3233	83.1562	184.2
GridSearch Time	1.95915	2.2857	2.63854	3.13648	10.9655	49.3754	125.553	319.882
RandomSearch Time	2.06316	2.36206	2.85771	3.36965	14.3121	79.7276	183.147	605.46
Standard R <sup>2</sup>	0.498417	0.498571	-0.000311991	0.000388134	0.488313	0.65608	0.81523	0.693749
GridSearch R <sup>2</sup>	0.498417	0.498417	0.496414	0.498074	0.488816	0.668196	0.821004	0.762261
RandomSearch R <sup>2</sup>	0.498417	0.498417	0.496414	0.498074	0.488816	0.666121	0.821004	0.762261
Standard RMSE	0.707922	0.707814	0.999728	0.999379	0.715018	0.586197	0.429665	0.553163
GridSearch RMSE	0.707922	0.707922	0.709335	0.708165	0.714665	0.575778	0.422899	0.487377
RandomSearch RMSE	0.707922	0.707922	0.709335	0.708165	0.714665	0.577575	0.422899	0.487377



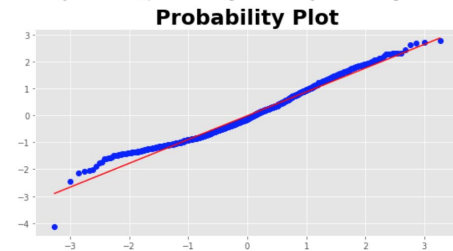
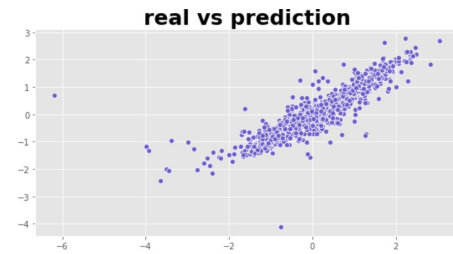
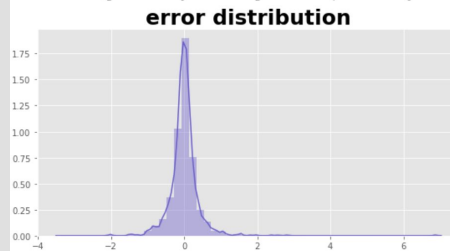
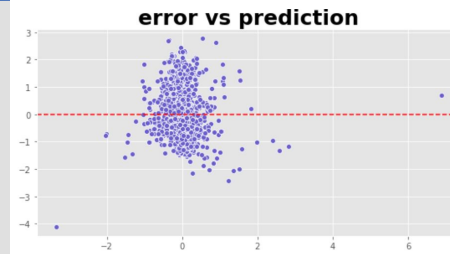
	lr	ridge	lasso	elastic	knn	svr	rfr	gradboost
GridSearch Time	1.95915	2.2857	2.63854	3.13648	10.9655	49.3754	125.553	319.882
GridSearch R <sup>2</sup>	0.498417	0.498417	0.496414	0.498074	0.488816	0.668196	0.821004	0.762261
GridSearch RMSE	0.707922	0.707922	0.709335	0.708165	0.714665	0.575778	0.422899	0.487377

Une **méthode ensembliste** permet de  
gagner en précision avec un temps de calcul  
acceptable pour le **Random Forest.**

# Visualisation de l'erreur et feature importance

Distribution gaussienne des résidus.  
Quelques valeurs outliers...

*Représentation satisfaisante du modèle Random Forest en GridSearch.*



PropertyGFATotal	67.261458
YearBuilt	6.000162
Longitude	5.733221
Latitude	5.239361
x0_Multifamily Housing	3.281412
x0_Non-Refrigerated Warehouse	2.178511
NumberOfFloors	2.023985
x0_Distribution Center	0.849948
CouncilDistrictCode	0.717666

**Cohérence des principales caractéristiques pouvant impacter la consommation d'énergie**

# Performances de Modélisation émissions CO2

8 Algorithmes testés et comparés.  
Sans et avec validation croisée.

*La tendance générale est meilleure avec  
GridSearch.*

## Cible: émissions de CO2

	lr	ridge	lasso	elastic	knn	svr	rfr	gradboost
Standard Time	0.00579596	2.09872	2.4141	2.96731	3.52457	16.5881	73.0243	171.013
GridSearch Time	1.96739	2.30933	2.69027	3.24176	11.6801	46.6119	116.062	273.239
RandomSearch Time	2.0918	2.40588	2.96128	3.48797	14.5816	69.6739	169.779	559.525
Standard R²	0.436805	0.436975	-0.000781605	-0.000781605	0.394081	0.565222	0.816049	0.623747
GridSearch R²	0.436805	0.436975	0.434267	0.436136	0.406709	0.565222	0.819213	0.758478
RandomSearch R²	0.436805	0.436975	0.434267	0.436136	0.406709	0.559876	0.819213	0.758478
Standard RMSE	0.751999	0.751886	1.00244	1.00244	0.780001	0.660727	0.429773	0.614651
GridSearch RMSE	0.751999	0.751886	0.753692	0.752446	0.771831	0.660727	0.426061	0.492454
RandomSearch RMSE	0.751999	0.751886	0.753692	0.752446	0.771831	0.664776	0.426061	0.492454



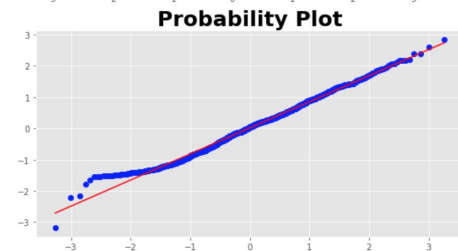
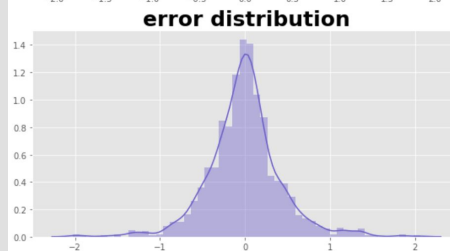
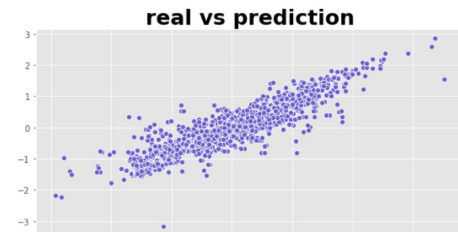
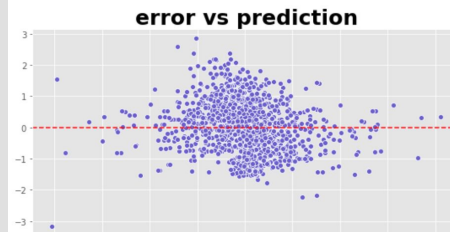
	lr	ridge	lasso	elastic	knn	svr	rfr	gradboost
GridSearch Time	1.96739	2.30933	2.69027	3.24176	11.6801	46.6119	116.062	273.239
GridSearch R²	0.436805	0.436975	0.434267	0.436136	0.406709	0.565222	0.819213	0.758478
GridSearch RMSE	0.751999	0.751886	0.753692	0.752446	0.771831	0.660727	0.426061	0.492454

Une **méthode ensembliste** permet de  
gagner en précision avec un temps de calcul  
acceptable pour le **Random Forest**

# Visualisation de l'erreur et feature importance

Distribution gaussienne des résidus.  
Quelques valeurs outliers...

*Représentation satisfaisante du modèle Random Forest en GridSearch.*



PropertyGFATotal	47.084799
YearBuilt	14.406709
Longitude	8.002508
Latitude	7.948130
x1_Low-Rise Multifamily	6.334390
NumberofFloors	3.082281
x0_Non-Refrigerated Warehouse	1.398868
CouncilDistrictCode	1.041052
x0_Multifamily Housing	0.994943
x0_Office	0.823064
x2_NonResidential	0.688978

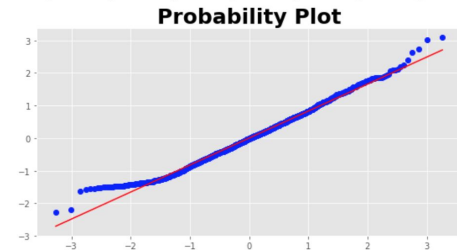
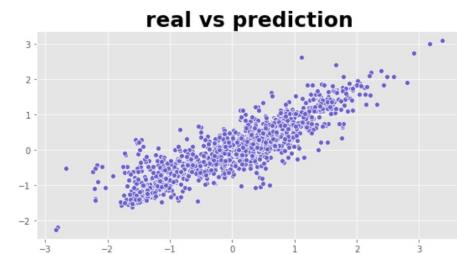
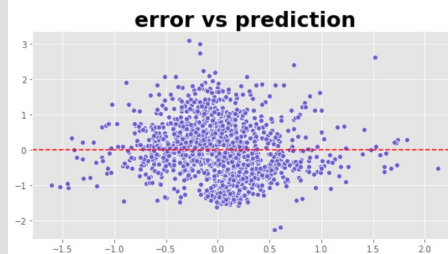
**Cohérence des principales caractéristiques  
pouvant impacter les émissions CO2**

# Modélisation des émissions CO2 avec l'énergyStarScore

Si les calculs de l'indicateur sont complexes, son intérêt peut-être remis en question.

*D'un point de vue modélisation statistique l'énergyStarScore n'a pas d'intérêt significatif.*

	lr	ridge	lasso	elastic	knn	svr	rfr	gradboost
GridSearch Time	1.82868	2.16754	2.54046	2.95865	12.8008	47.0794	110.438	180.138
GridSearch R <sup>2</sup>	0.45975	0.45975	0.455794	0.457659	0.460981	0.614614	0.779918	0.553939
GridSearch RMSE	0.734302	0.734302	0.736985	0.735722	0.733465	0.620191	0.468673	0.667229



**Baisse des performances, aucune amélioration significative pouvant justifier l'intérêt de l'énergyStarScore**



# Conclusion

...

Problématique de régression identifiée et traitée  
8 algorithmes traités et optimisés en validation croisée

Meilleures performances en Méthode ensembliste

Choix d'un modèle en Random Forest

Aucun intérêt de l'énergyStarScore pour la prédiction d'émissions



# Réflexion sur les axes d'amélioration



Enrichissement des données par des caractéristiques complémentaires.  
exemples : type d'isolants thermiques des bâtiments, type d'éclairage, ou  
encore des indications météorologiques.