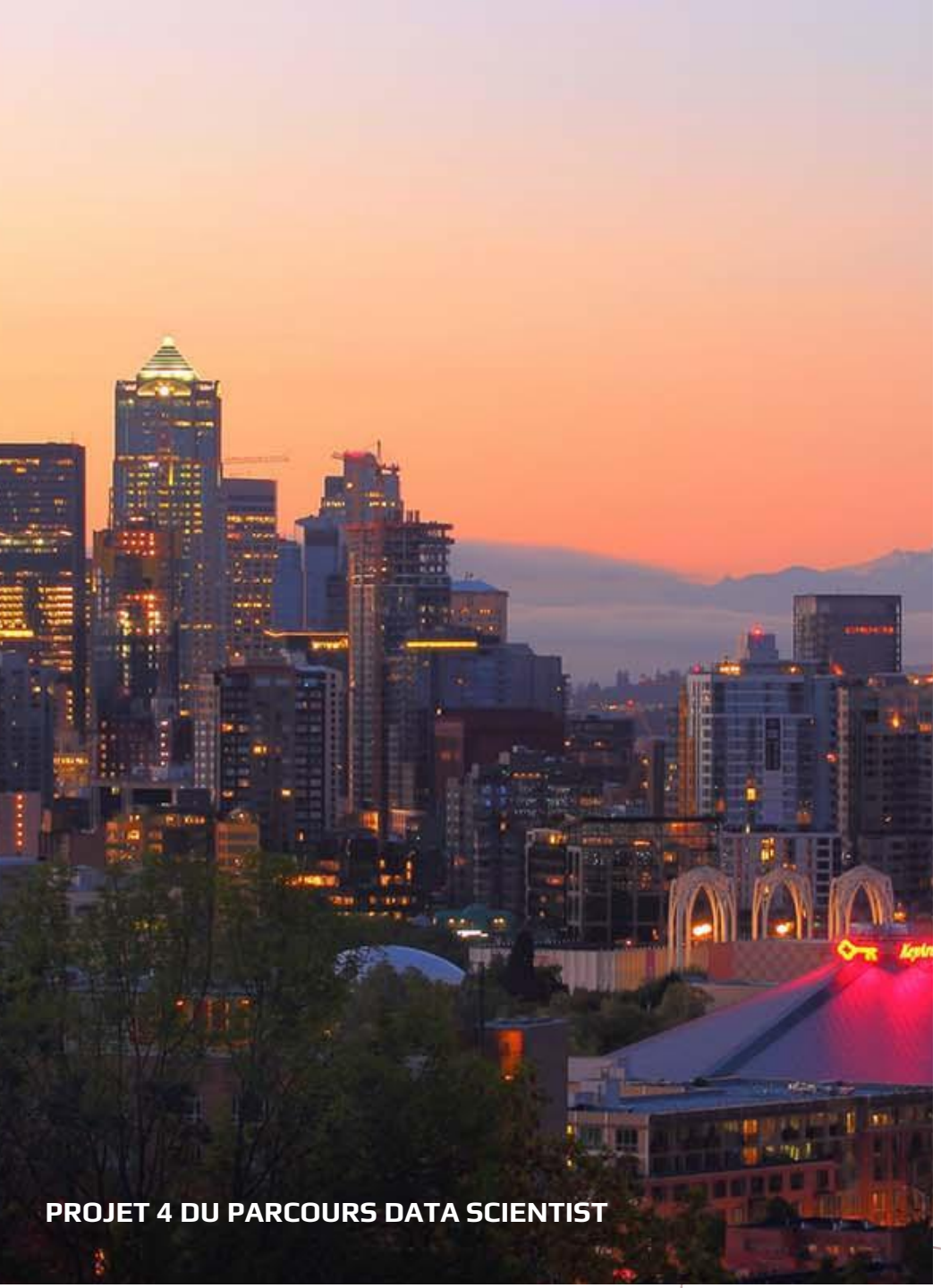
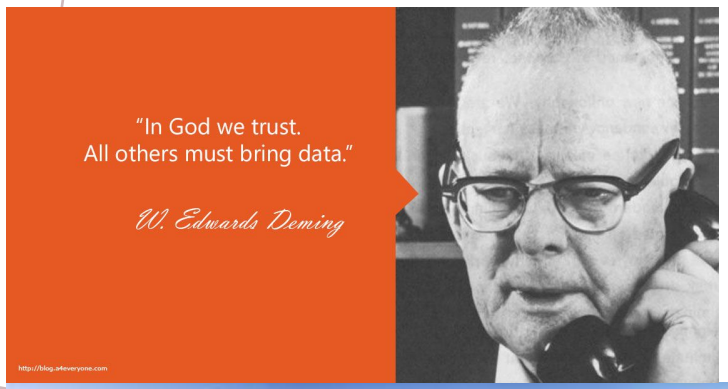


PROJET 4 DU PARCOURS DATA SCIENTIST

ANTICIPEZ LES BESOINS EN CONSOMMATION ÉLECTRIQUE DE BÂTIMENTS

GAUTHIER RAULT
PARCOURS DATA SCIENTIST
CHEZ OPENCLASSROOMS

EVALUATEUR
MONSIEUR CHRISTIAN NOUMSI



Agenda

Ouverture de la problématique - 5 min

Exposition du cleaning - 5 min

Modélisation des données - 10 min

Modèle finale et améliorations - 5 min

Questions-Réponses - 5-10 min

Ouverture de la problématique

Objectifs

1) Tenter de prédire les émissions de **CO2** et la consommation totale d'**énergie** des bâtiments **non résidentiels** de la ville de **Seattle**

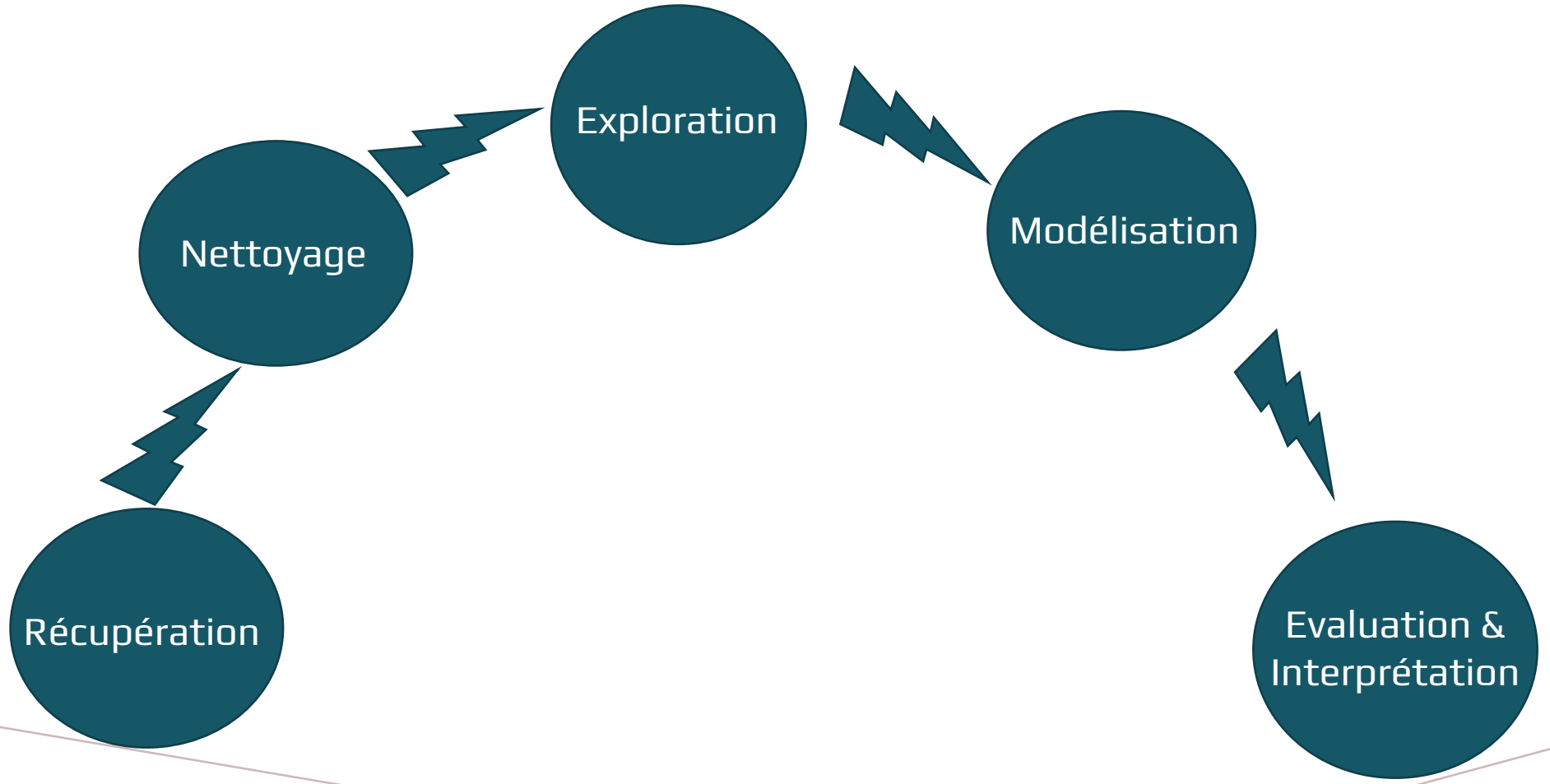
2) Evaluer l'intérêt de l'**ENERGY STAR Score**

Moyens mis à disposition

1) Les données déclaratives du permis d'exploitation commerciale

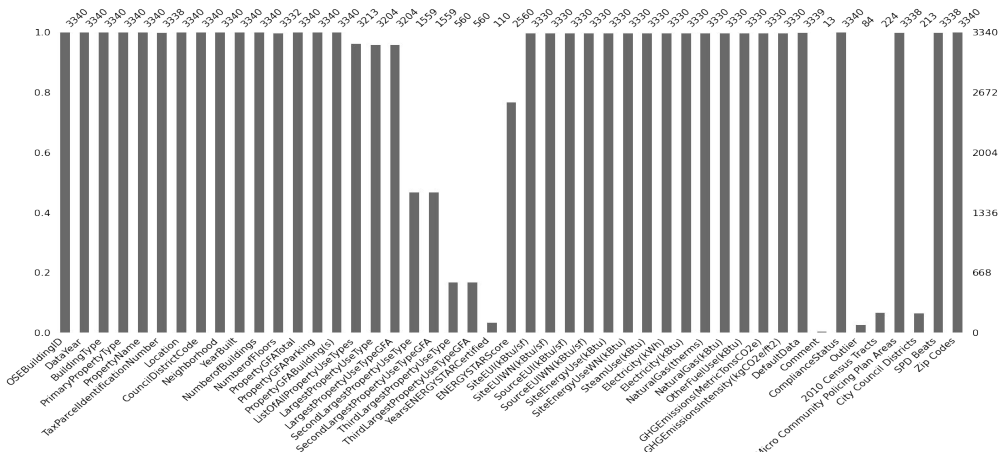
2) Les relevés des agents en 2015 et 2016

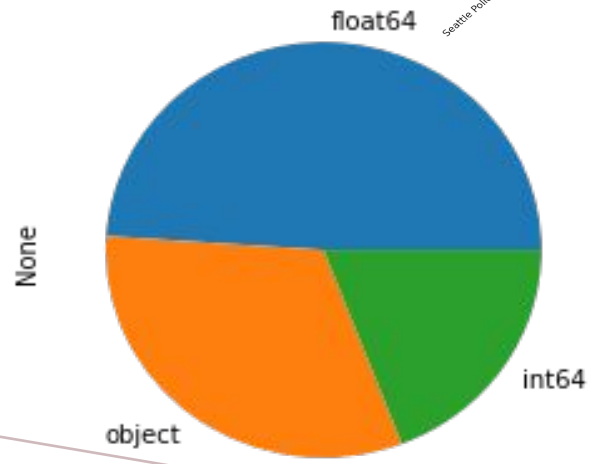
Démarche du projet



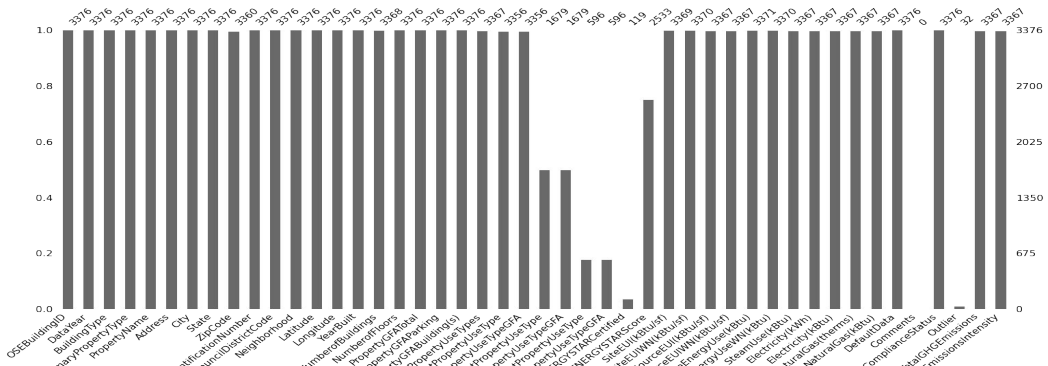
Aperçu des données 2015 et 2016

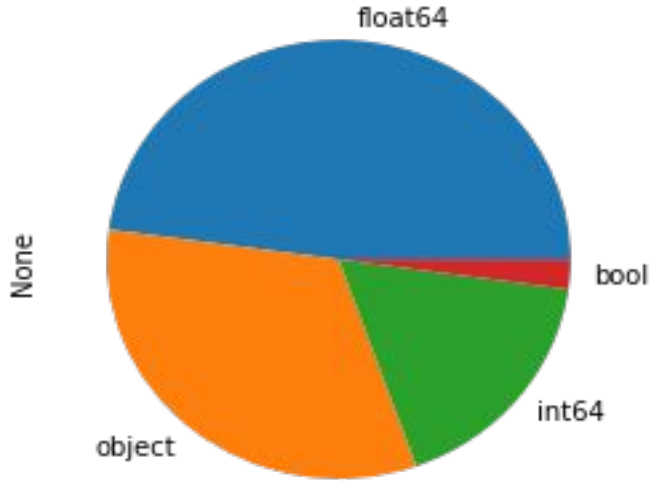
Pour 2015



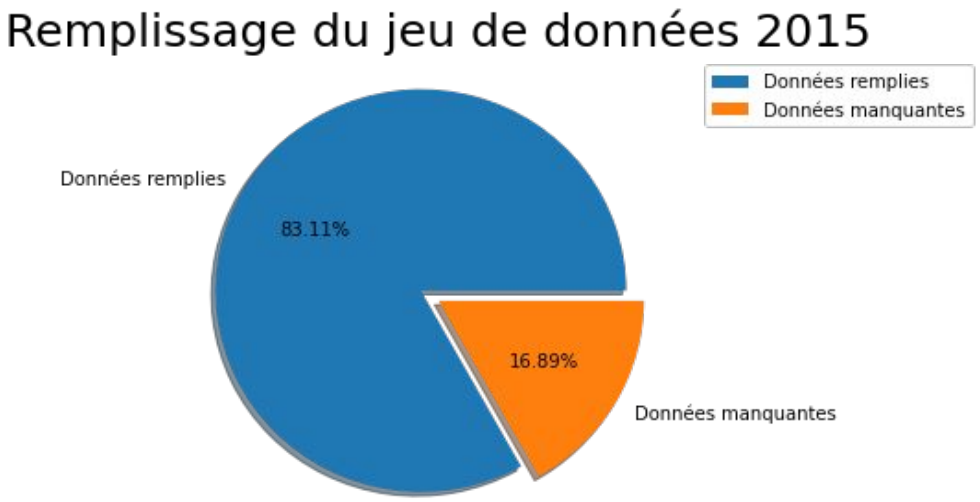


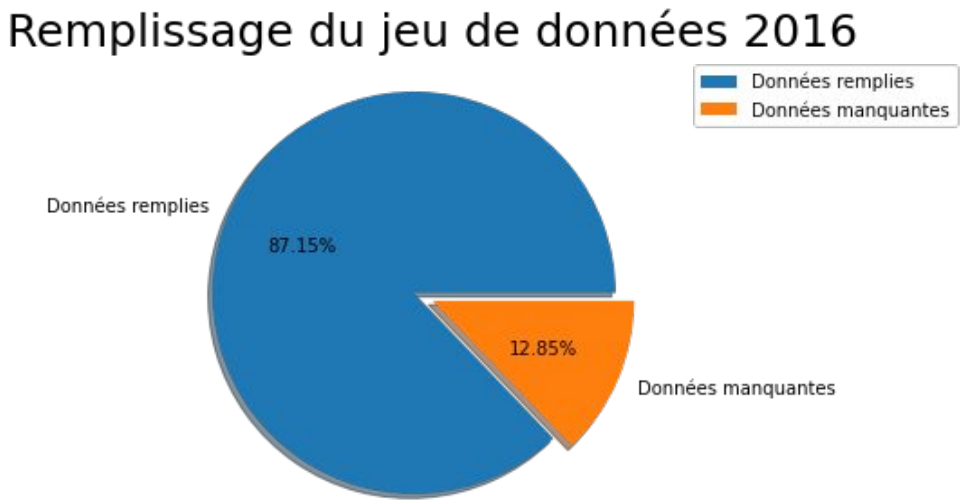
Pour 2016





Aperçu des données 2015 et 2016





Exposition du cleaning

Objectifs

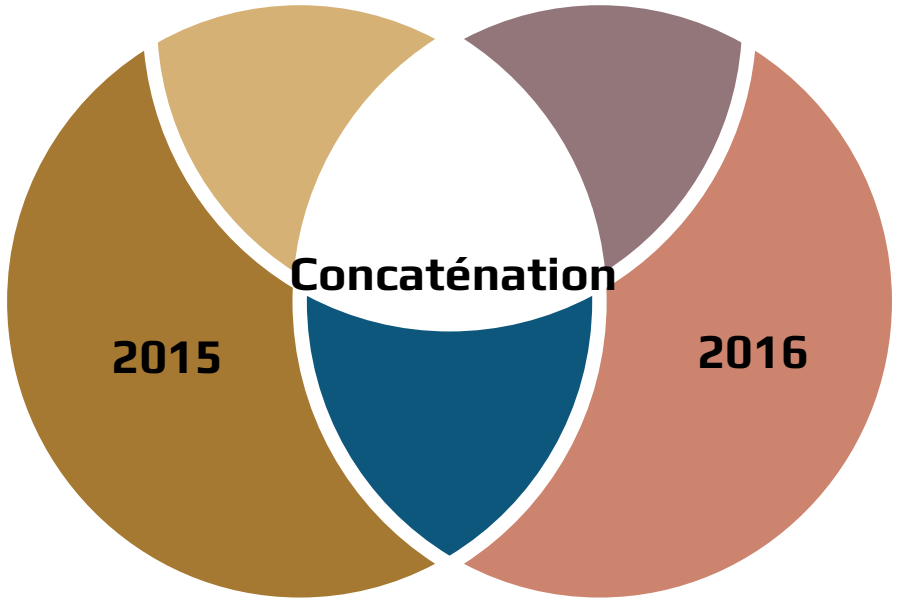
Aligner le nom des variables du dataset 2015 et 2016

Corriger les erreurs d'entrée

Concaténation des données 2015 et 2016 dans un Dataframes

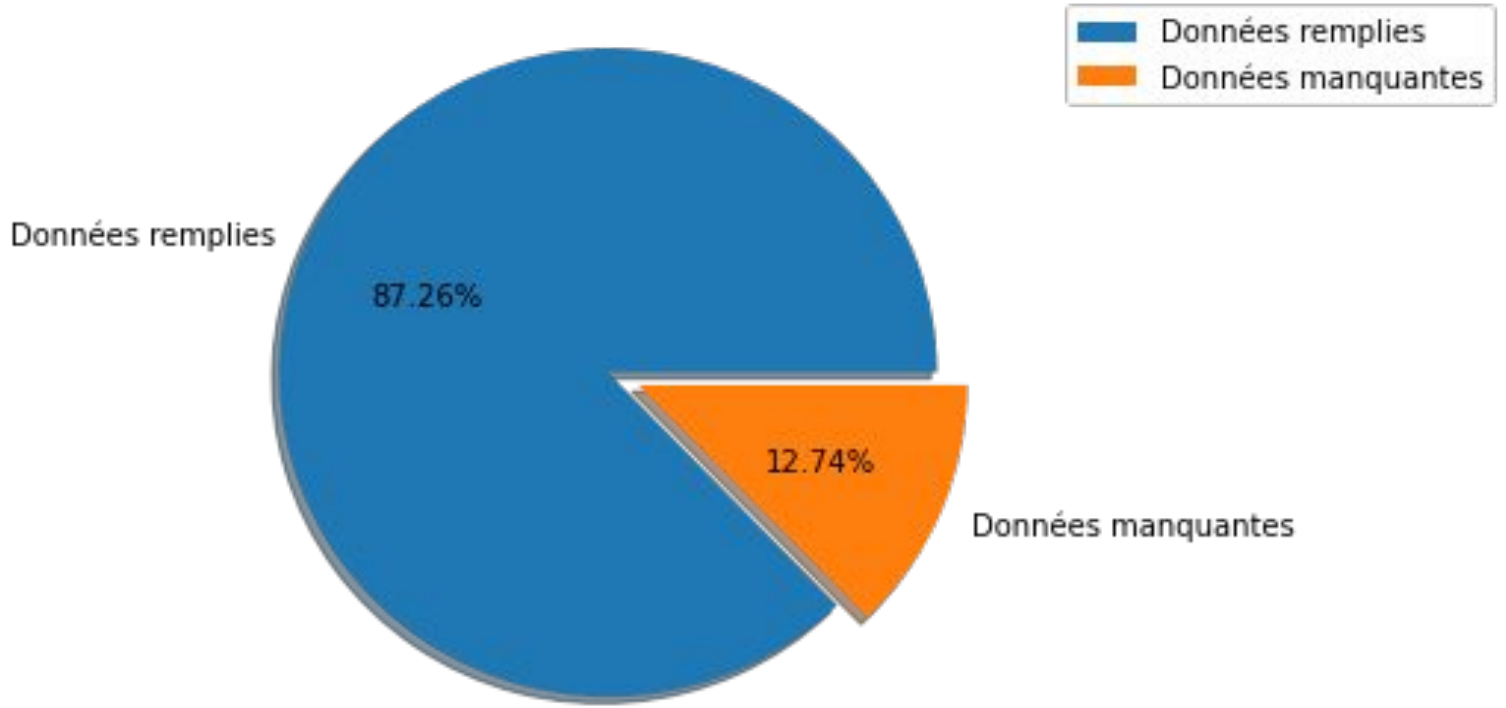
Amputation des colonnes sans correspondances

Renommage des colonnes



Correspondance entre colonnes

Remplissage du jeu de données concaténé



Traitement des erreurs

Suppression des données résidentielles

Suppressions des variables insignifiantes (constante, peu remplie) et redondantes

Vérification de l'unicité des valeurs et de leurs cohérences

Traitement des outliers (aberrations, IQR)
Complétion des NaN (0 / No information)
Suppression des valeurs 0

ENERGYSTARScore

Aucune stratégie retenue pour agir ou compléter cette feature

Filtre sur les valeurs de ENERGYSTARScore remplie

Identification des 2 targets/labels/étiquettes

TotalGHGEmissions

Quantité totale des émissions de gaz à effet de serre, suite à la consommation d'énergie sur le site

Mesurée en tonne d'équivalent dioxyde de carbone

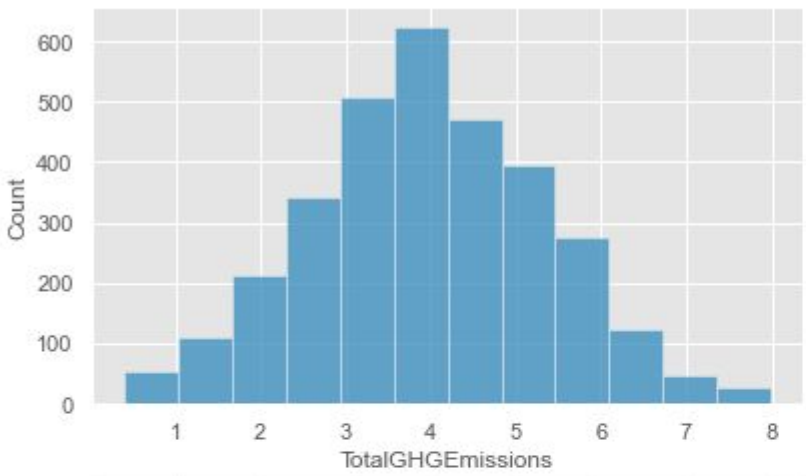
SiteEnergyUse

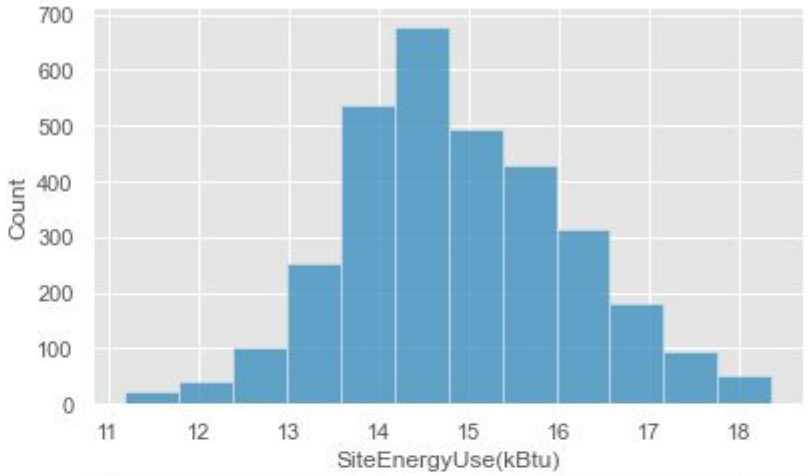
Quantité annuelle d'énergie consommée, toutes sources d'énerg confondues

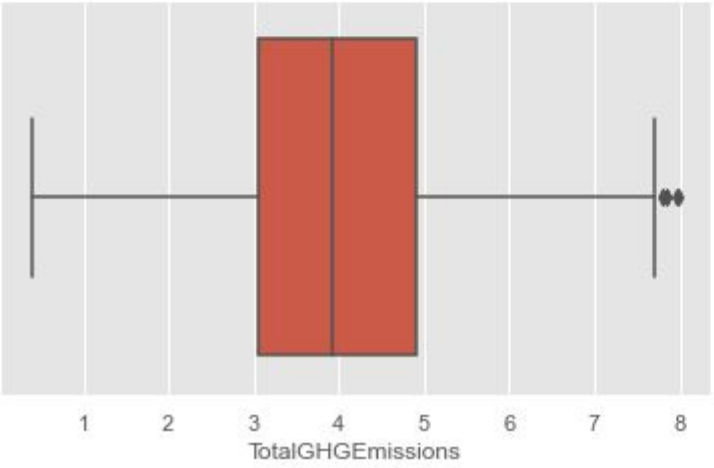
Mesurée en kBtu*

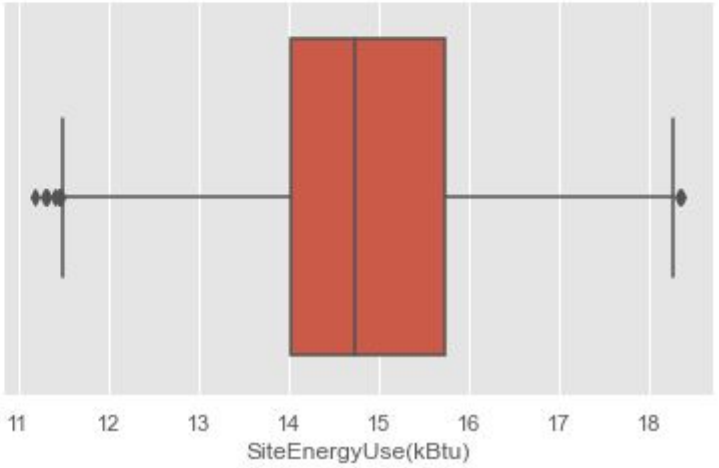
**Signifie kilo British thermal unit*
1 Btu = 0.00029307017975752 55 kwh
1 Btu = 1055.56 joules

Passage au log des targets (pas retenu)





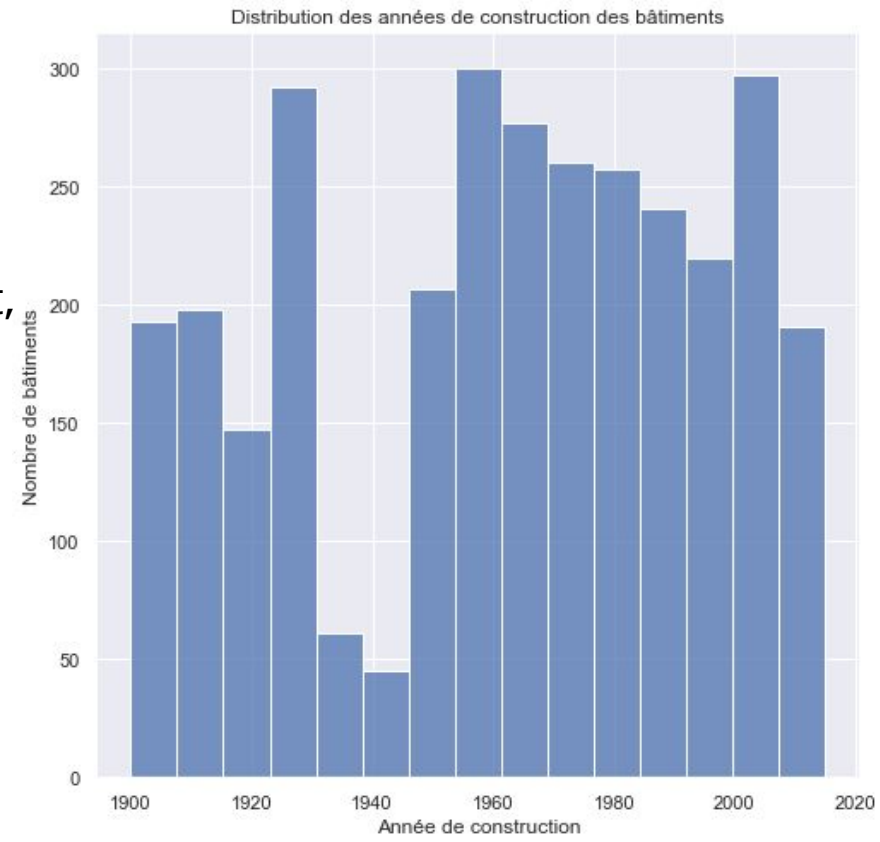




Création de variables

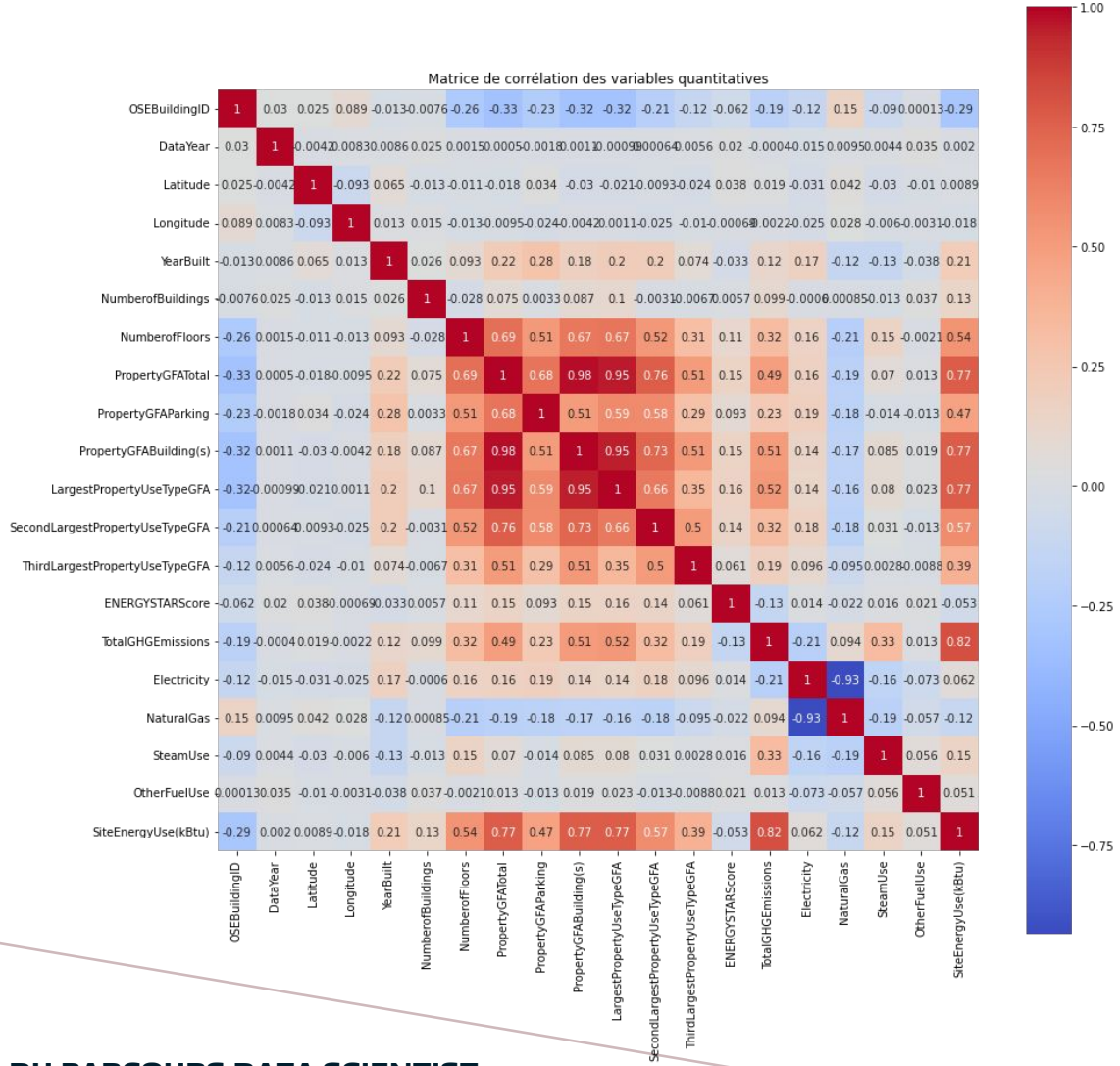
AgeBuilding → DataYear » - « YearBuilt »

Electricity
NaturalGas
SteamUse
OtherFuelUse
→ Indique, pour chaque bâtiment, la proportion d'utilisation de chaque type d'énergie



Exemple de calcul → Electricity = Electricity(kBtu) / SiteEnergyUse(kBtu)

Matrice de corrélation



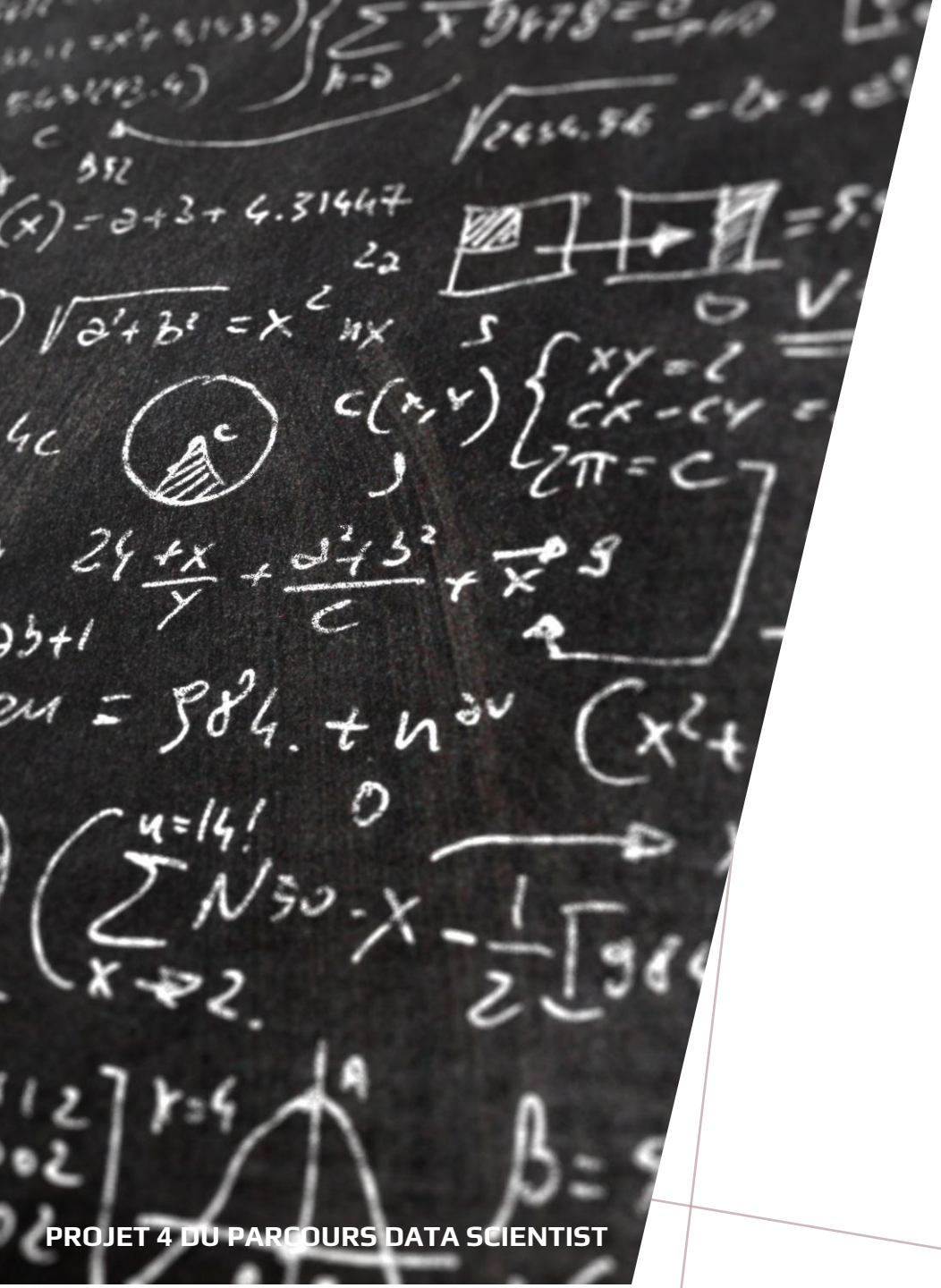
Variables très corrélées entre elles

<u>Suppression :</u>

PropertyGFATotal

PropertyGFABuilding

LargestPropertyUseTypeGFA



Transformation des variables

Encodage des Variables Catégorielles avec OneHotEncoder

Standardisation des Variables numériques avec StandardScaler

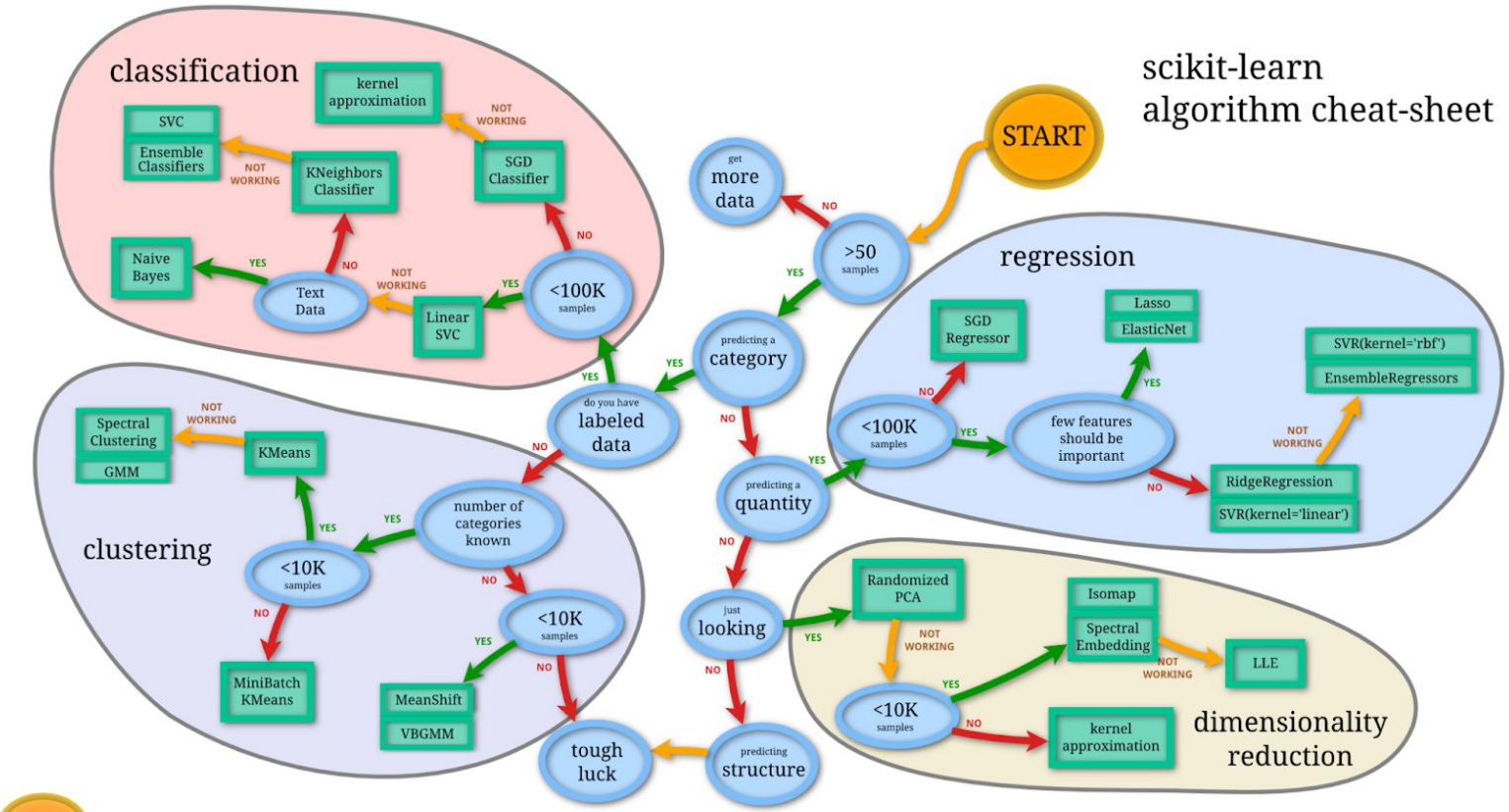
Modélisation des données

Objectifs

Déterminer quel est le meilleur algorithme

Évaluer la pertinence de l'ENERGYSTARScore

Sélection des algorithmes



DummyRegressor

Linéaire

Régression Ridge (Régularisation de TYKHONOV)

Régression Elastic Net

Lasso (poursuite de base)

Non-linéaire via la méthode à noyau

KernelRidge

SVR (Support Vector Regression)

Métriques utilisés

Pénalise les grandes erreurs
Résultat dans la même unité que la target

RMSE
(Root Mean Squared Error)

Pénalise les grandes erreurs
Plus difficile à interpréter

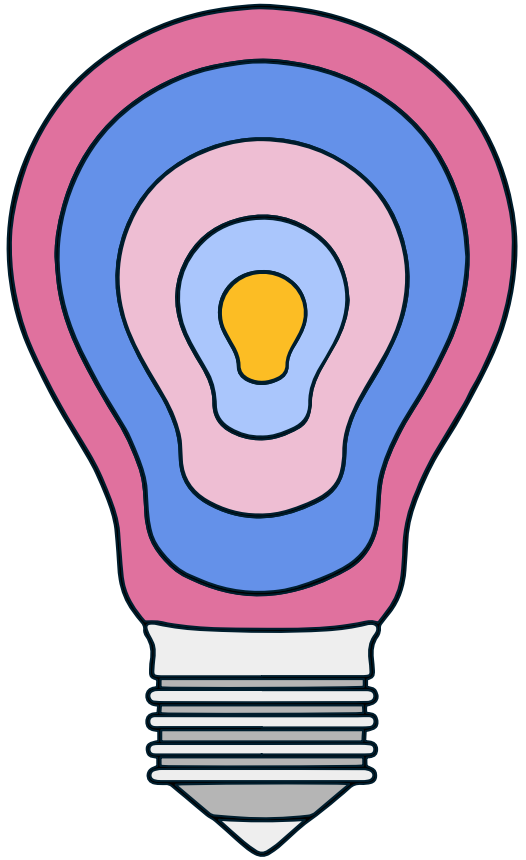
MSE
(Mean Squared Error)

“A quelle distance étions nous en moyenne dans nos prédictions?”

MAE
(Mean Absolute Error)

Représente la proportion de variance expliquée par le modèle

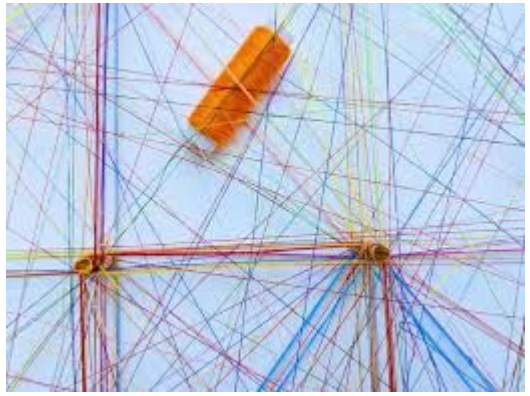
R^2



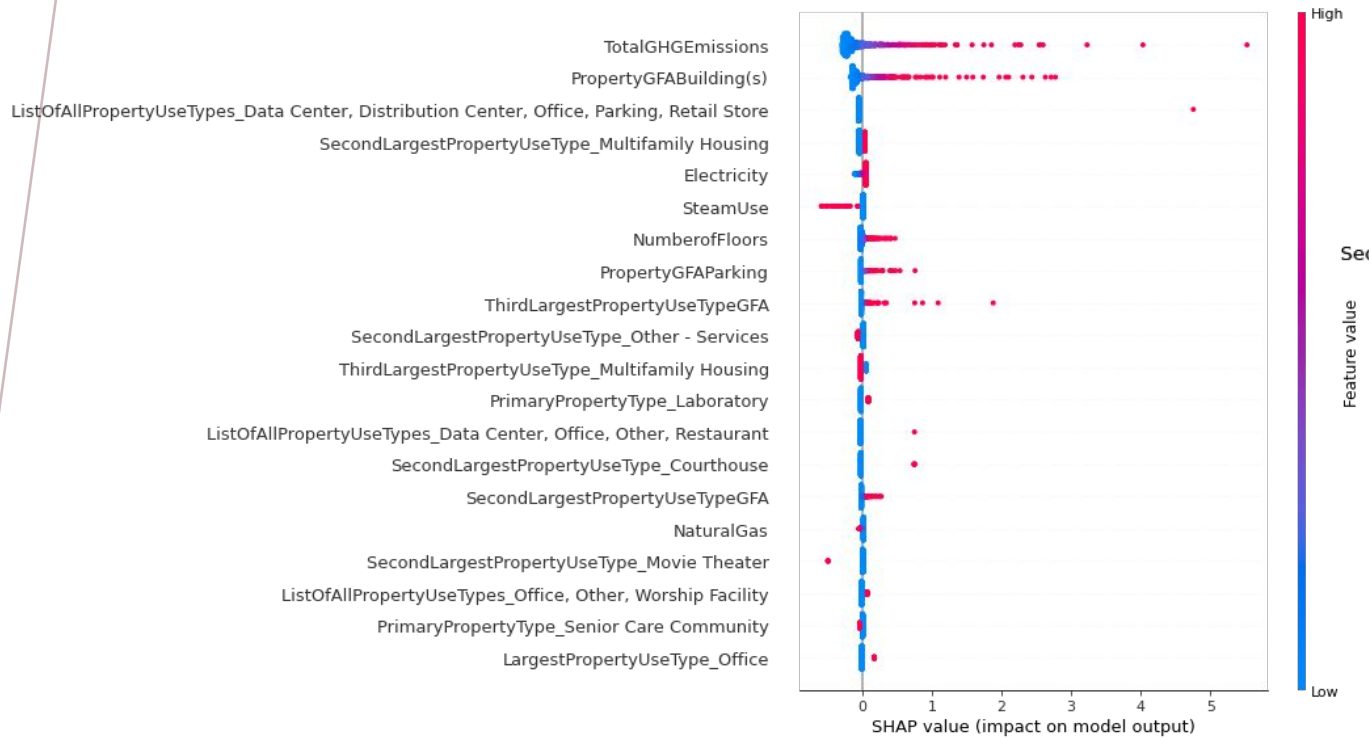
Feature selection via RFECV

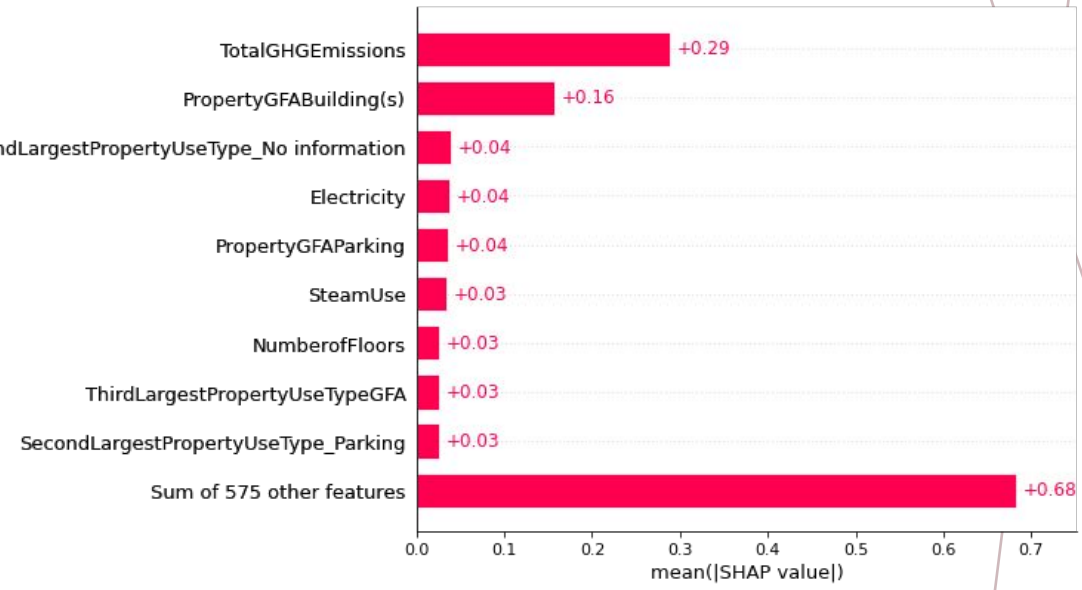
(Recursive feature elimination with cross-validation)

L’utilisation de RFECV n’a pas permis d’augmentation significative des résultats

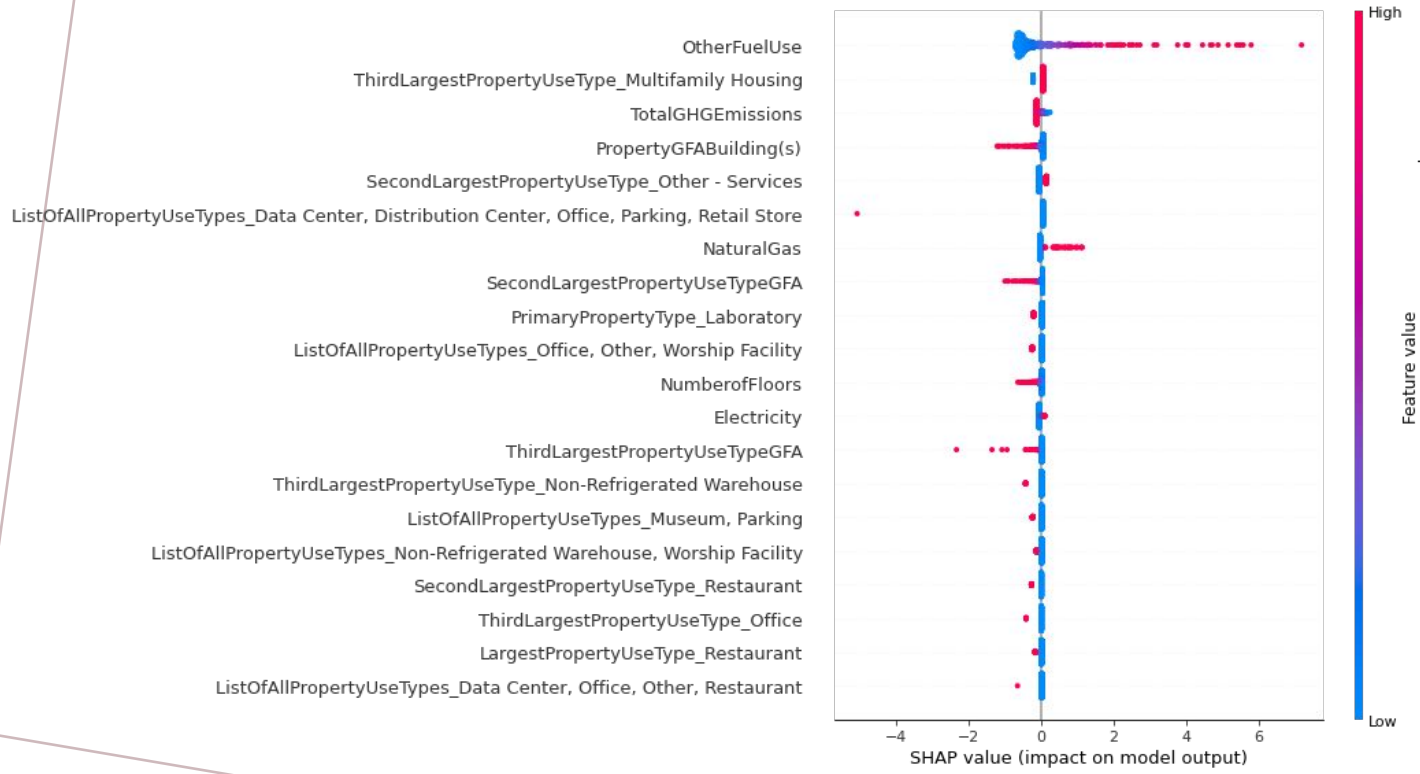


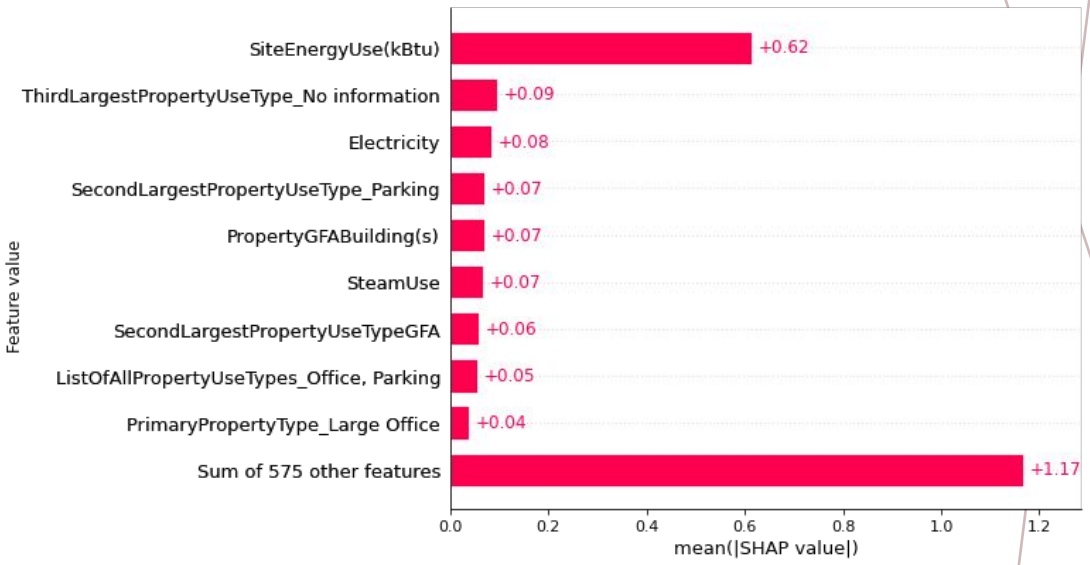
Features importance via SHAP pour SiteEnergyUse(kBtu)





Features importance via SHAP pour TotalGHGEmissions

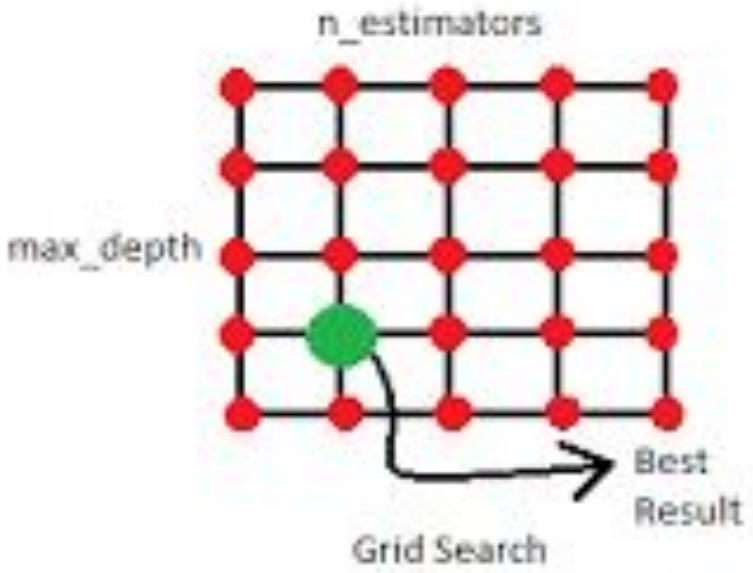




Optimisation des Hyperparamètres (fine tuning)

GridSearchCV

- ✔ Teste l'ensemble des hyperparamètres fournis
- ✔ Meilleure efficacité
- ✔ Peut nécessiter beaucoup de temps



Choix du Meilleur Algorithme

	TotalGHGEmissions			SiteEnergyUse		
Algorithme	**R^2**	**RMSE**	**Temps (fit+pred)**	**R^2**	**RMSE**	**Temps (fit+pred)**
DummyRegressor	0	0,87	0,0003s	0	0,92	0,0003s
Lasso	0	0,87	0,01s	0	0,92	1,2s
Ridge	0,85	0,34	0.03s	0,93	0,25	0,04s
ElasticNet	0,33	0,71	0.01s	0,42	0,7	0,01s
KernelRidge	0,85	0,34	0.23s	0,92	0,25	1,42s
SVR	0,58	0,56	0.84s	0,68	0,52	0,23s

La régression Ridge

Modèle finale et améliorations

	TotalGHGEmissions				SiteEnergyUse			
Algorithme Ridge	**R^2**	**RMSE**	**MSE**	**MAE**	**R^2**	**RMSE**	**MSE**	**MAE**
Ridge	0,85	0,34	0,11	0,17	0,93	0,25	0,06	0,13
RFECV	0,87	0,32	0,1	0,17	0,93	0,24	0,06	0,13
GridSearchCV alpha et solver	0,87	0,31	0,1	0,16	0,93	0,24	0,06	0,13
Avec l'EnergyStarScore	**R^2**	**RMSE**	**MSE**	**MAE**	**R^2**	**RMSE**	**MSE**	**MAE**
Ridge	0,85	0,35	0,12	0,19	0,93	0,26	0,07	0,13
RFECV	0,86	0,34	0,12	0,18	0,93	0,25	0,06	0,13
GridSearchCV alpha et solver	0,87	0,3	0,09	0,17	0,93	0,24	0,06	0,13

Choix de la régression Ridge

Fine tuning et utilisation de RFECV

Evaluation de l'EnergyStarScore

	Métrique : R^2					
	TotalGHGEmissions			SiteEnergyUse		
Algorithme	**Sans ESS**	**Avec ESS**	**Gain (%)**	**Sans ESS**	**Avec ESS**	**Gain(%)**
DummyRegressor	0	0	0%	0	0	0%
Lasso	0	0	0%	0	0	0%
Ridge	0,85	0,85	0%	0,93	0,93	0%
ElasticNet	0,33	0,33	0%	0,42	0,42	0%
KernelRidge	0,85	0,85	0%	0,92	0,92	0%
SVR	0,58	0,58	0%	0,68	0,68	0%

Utilisation de l'ENERGYSTARScore
est non pertinent / non rentable

Conclusion

Modèle choisi : La régression Ridge

- Meilleurs résultats pour toutes les métriques utilisées

TotalGHGEmissions	
R^2 :	**0,85**
RMSE :	**0,34**
Temps :	**0,03s**

SiteEnergyUse	
R^2 :	**0,93**
RMSE :	**0,25**
Temps :	**0,04s**

- Plus rapide à entraîner

Utilisation de l'ENERGY STAR Score

➢ Aucune amélioration des modèles testés

AXES D'AMÉLIORATIONS DE L'ÉTUDE

1. Augmenter la taille du dataset pour améliorer les résultats des algorithmes et permettrait d'en tester d'autres

   1.1. se baser sur d'autres années 2017, 2018 etc.

   1.2. en augmentant la taille du territoire

   1.3. en prenant en compte les données résidentielles

2. Faire des tests :

   2.1. sortir les données catégorielles qui au travers du OneHotEncoder multiplie les features

   2.2. en conservant outliers (valeur atypique)

QUESTIONS ET RÉPONSES (5-10 MIN)

MERCI BEAUCOUP POUR VOTRE ATTENTION