

# PROJET 4

ANTICIPER LES BESOINS EN CONSOMMATION DES BATIMENTS



# PLAN

1. Sujet
2. Problématique
3. Nettoyage
4. Analyse Exploratoire
5. Prédictions
6. Modèle choisi
7. Importance des variables
8. Conclusion

# THÉMATIQUE

Je travaille pour la ville de Seattle, en tant que data scientist.

Pour atteindre son objectif de ville neutre en émissions de carbone en 2050, mon équipe s'intéresse de près à la consommation et aux émissions des bâtiments non destinés à l'habitation.

Des relevés minutieux ont été effectués par les agents de la ville en 2016.





# PROBLÉMATIQUE

Cependant, ces relevés sont coûteux à obtenir, et à partir de ceux déjà réalisés, je veux tenter de prédire les émissions de CO<sub>2</sub> et la consommation totale d'énergie de bâtiments **non destinés à l'habitation** pour lesquels elles n'ont pas encore été mesurées.

Je cherche 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 mon équipe.



# PRÉSENTATION DES DONNÉES

Un seul et unique jeu de données :

- 3376 lignes
- 46 colonnes



Après nettoyage :

- 1548 lignes
- 21 colonnes



Création de plusieurs  
de dataframes pour la  
suite du projet

	data.isnull().sum()
OSEBuildingID	0
DataYear	0
BuildingType	0
PrimaryPropertyType	0
PropertyName	0
Address	0
City	0
State	0
ZipCode	16
TaxParcelIdentificationNumber	0
CouncilDistrictCode	0
Neighborhood	0
Latitude	0
Longitude	0
YearBuilt	0
NumberofBuildings	8
NumberofFloors	0
PropertyGFATotal	0
PropertyGFAParking	0
PropertyGFABuilding(s)	0
ListOfTypePropertyUseTypes	9
LargestPropertyUseType	20
LargestPropertyUseTypeGFA	20
SecondLargestPropertyUseType	1697
SecondLargestPropertyUseTypeGFA	1697
ThirdLargestPropertyUseType	2780
ThirdLargestPropertyUseTypeGFA	2780
YearsENERGYSTARCertified	3257
ENERGYSTARScore	843
SiteEUI(kBtu/sf)	7
SiteEUIWN(kBtu/sf)	6
SourceEUI(kBtu/sf)	9
SourceEUIWN(kBtu/sf)	9
SiteEnergyUse(kBtu)	5
SiteEnergyUseWN(kBtu)	6
SteamUse(kBtu)	9
Electricity(kWh)	9
Electricity(kBtu)	9
NaturalGas(therms)	9
NaturalGas(kBtu)	9
DefaultData	0
Comments	3376
ComplianceStatus	0
Outlier	3344
TotalGHGEmissions	9
GHGEmissionsIntensity	9
<i>dtype: int64</i>	

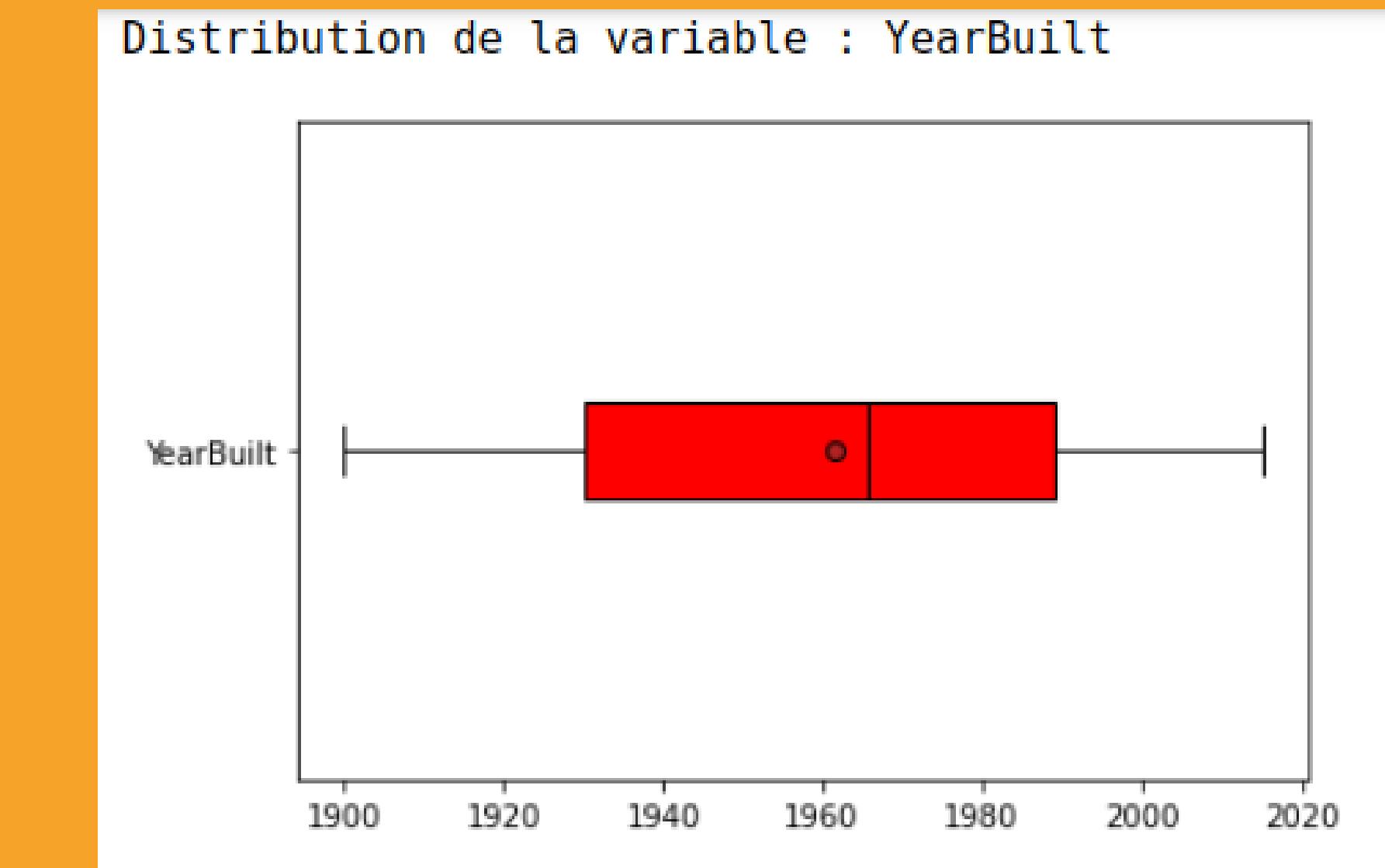
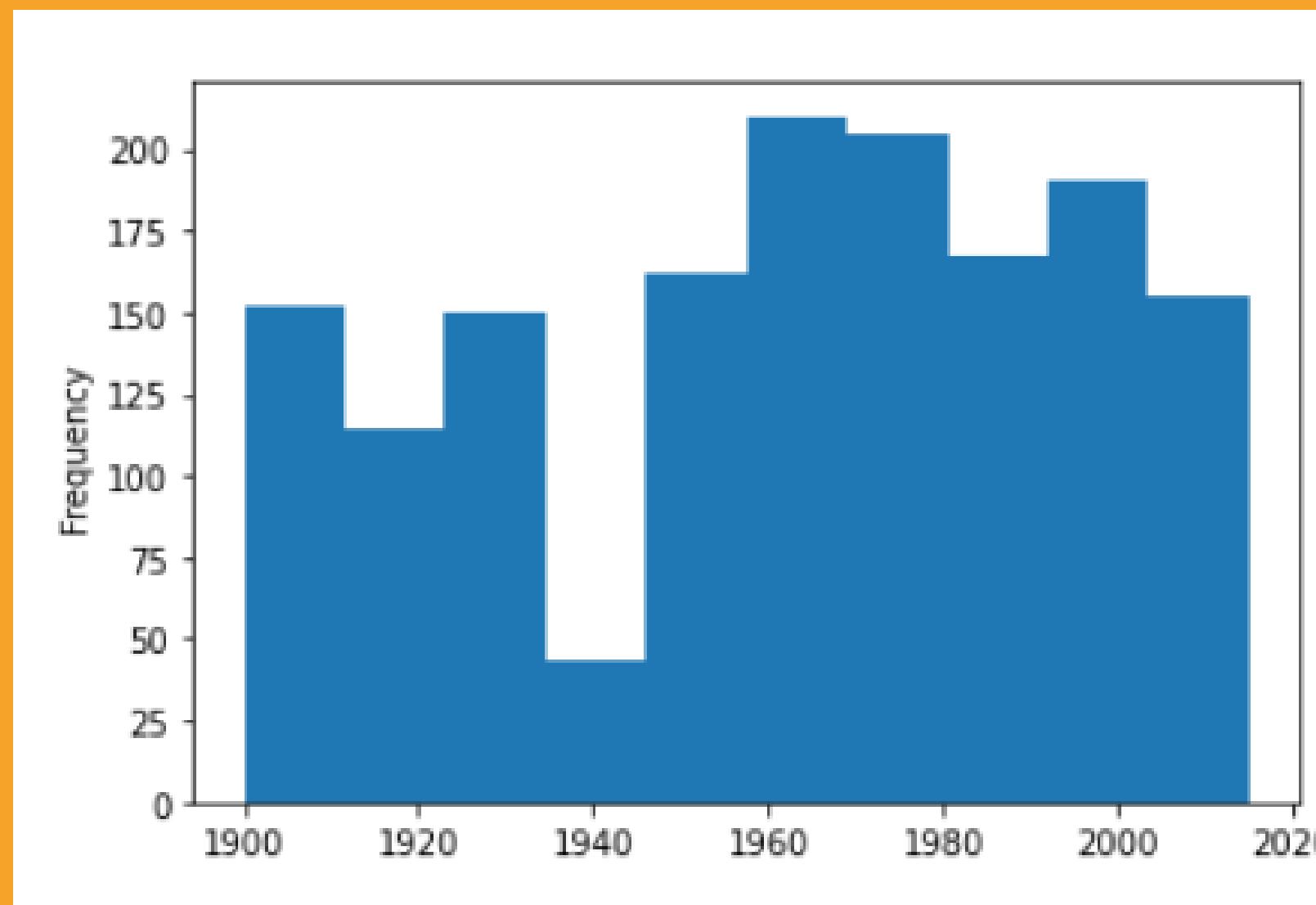
# NETTOYAGE DES DONNÉES

Voici les différents nettoyages effectués :

- Suppression de colonnes inutiles pour la prédiction(ville, adresse, building id...)
- Suppression des habitations, on s'intéresse ici aux bâtiments non destinés à l'habitation
- Suppression des doublons
- Suppression de lignes (ex: variable y non remplie,
- Transformer les variables de consommation d'énergie en boolean (ex: SteamUse(kbtu) devient UsingSteam)
- Imputation des NaN (NumberOfUseTypes = 1, imputer LargestPropertyUseTypeGFA avec PropertyGFABuilding)
- Valeurs aberrantes
- Outliers

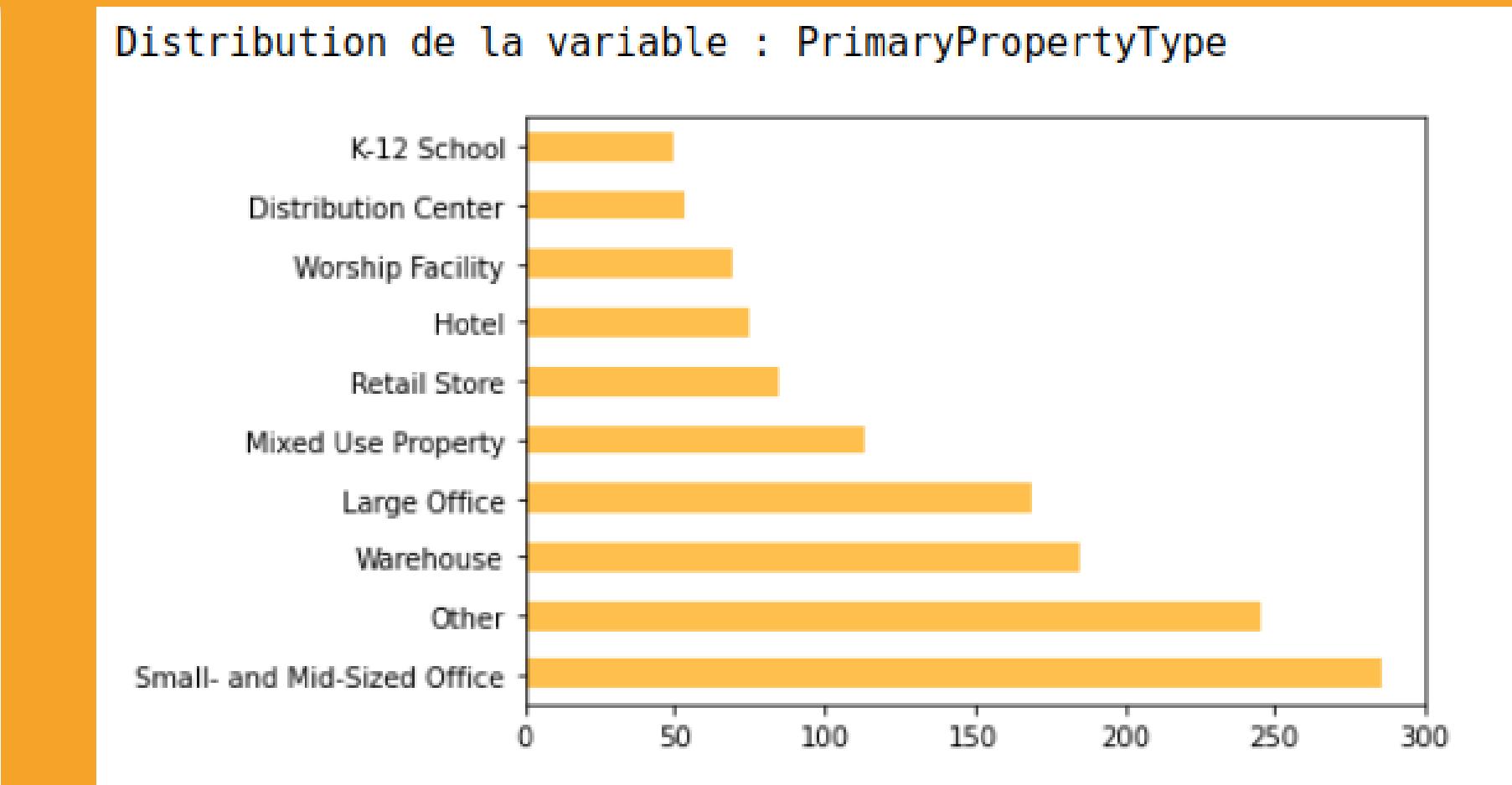
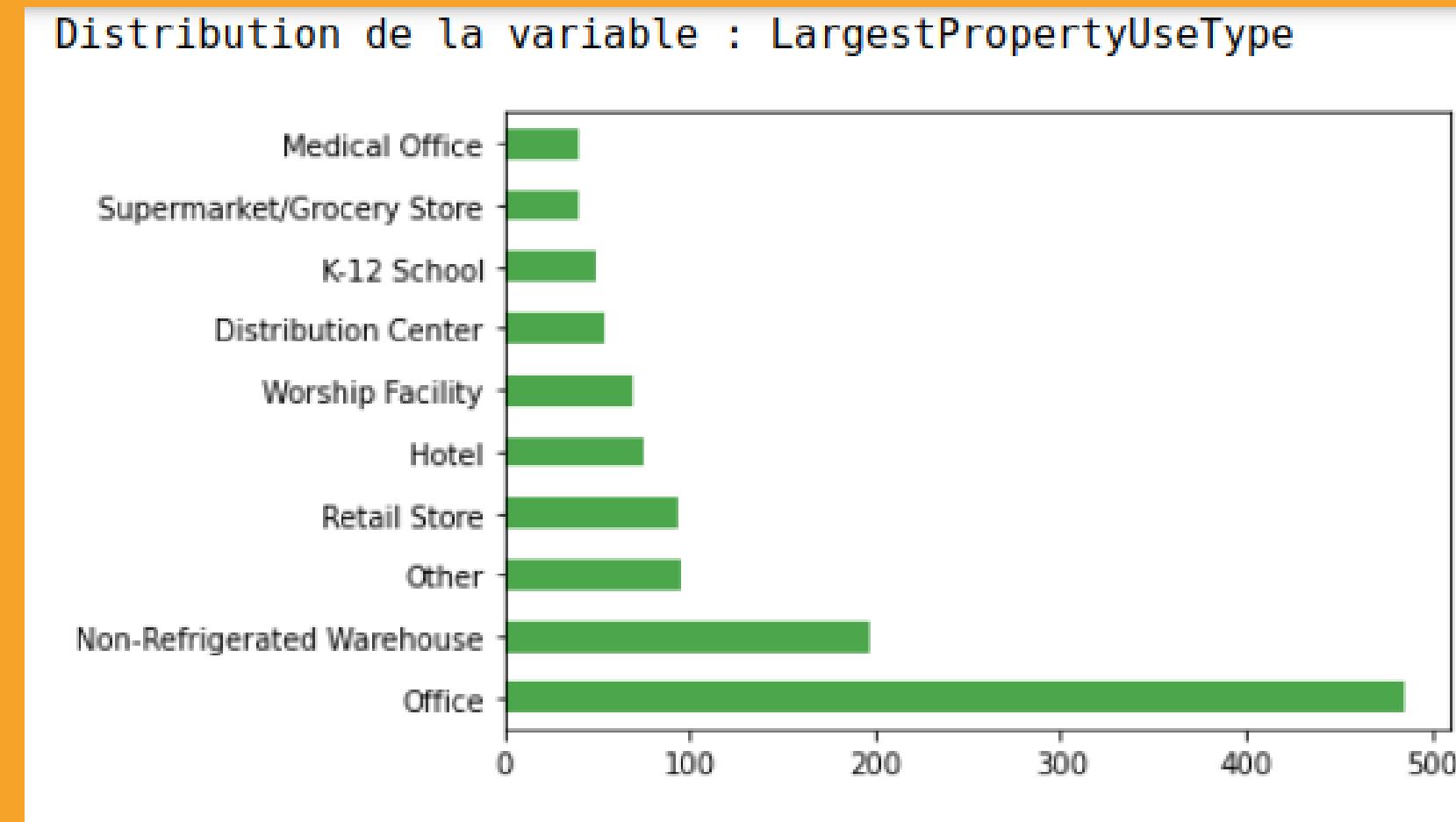
# ANALYSE EXPLORATOIRE

Année de construction



# ANALYSE EXPLORATOIRE

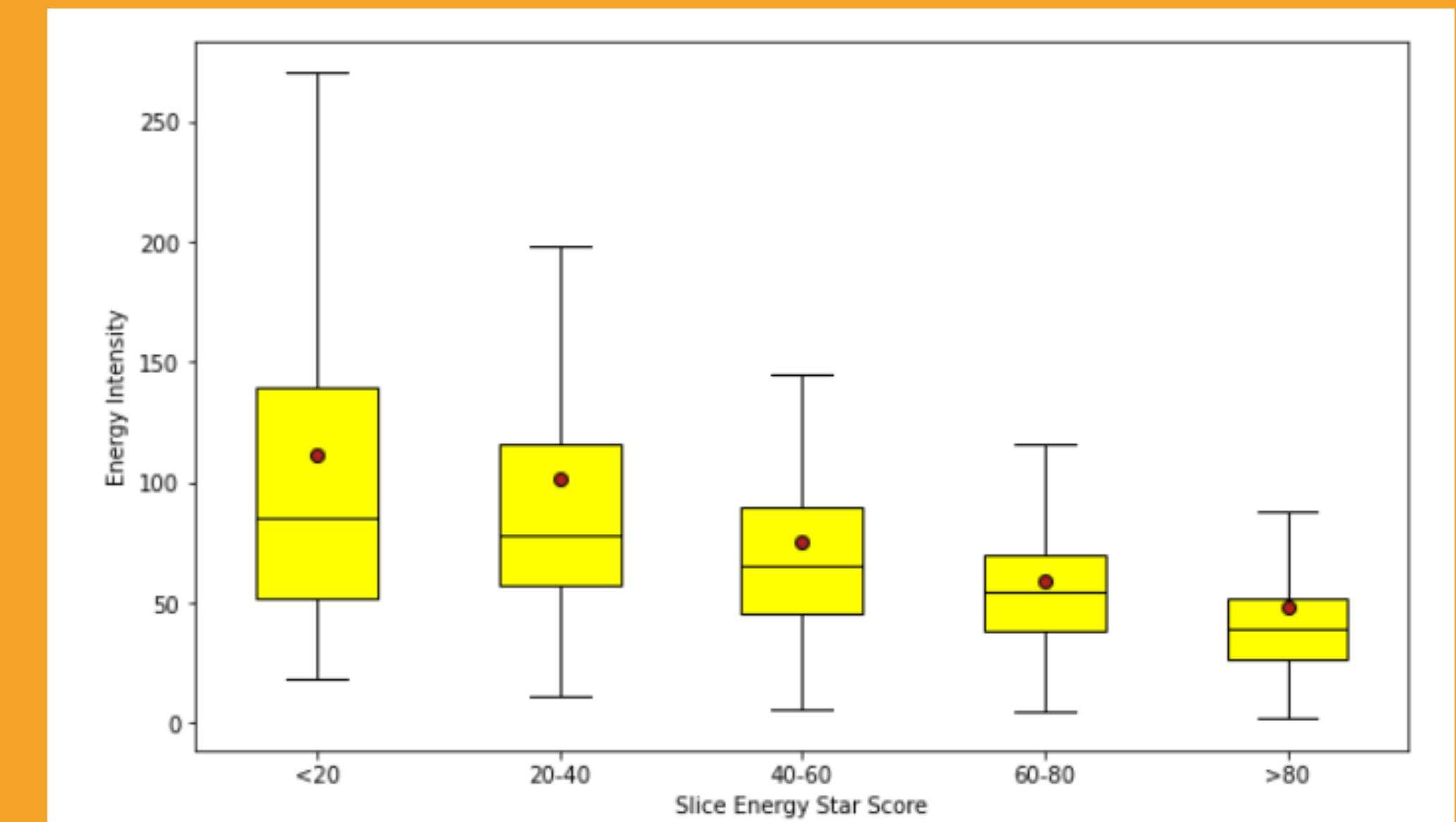
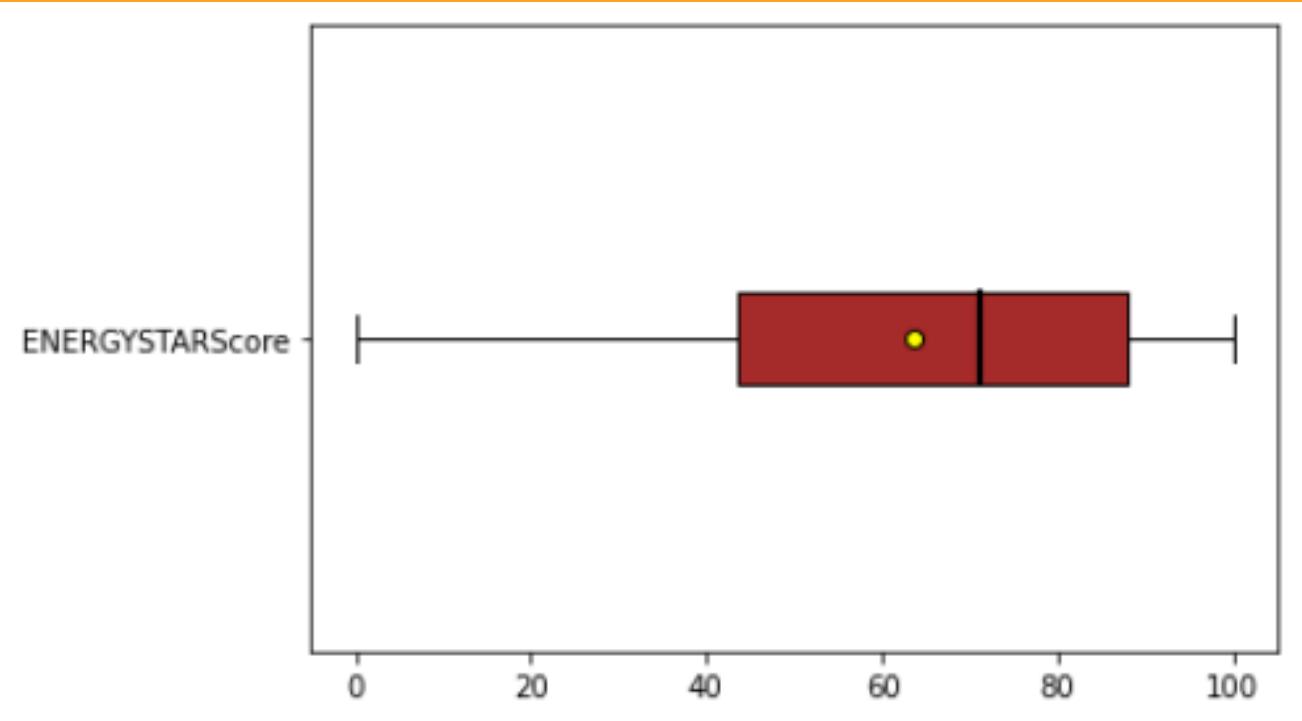
## Utilisation principale des bâtiments



# ANALYSE EXPLORATOIRE

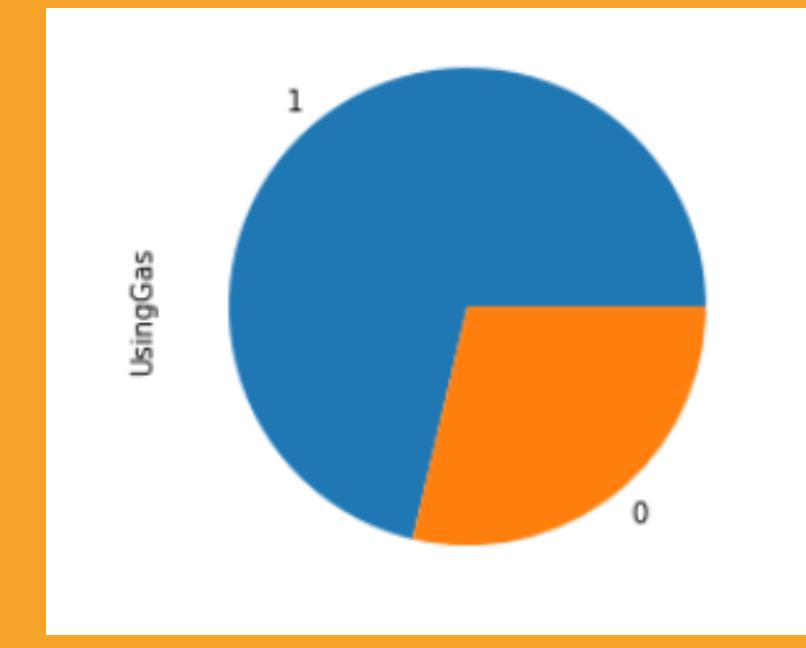
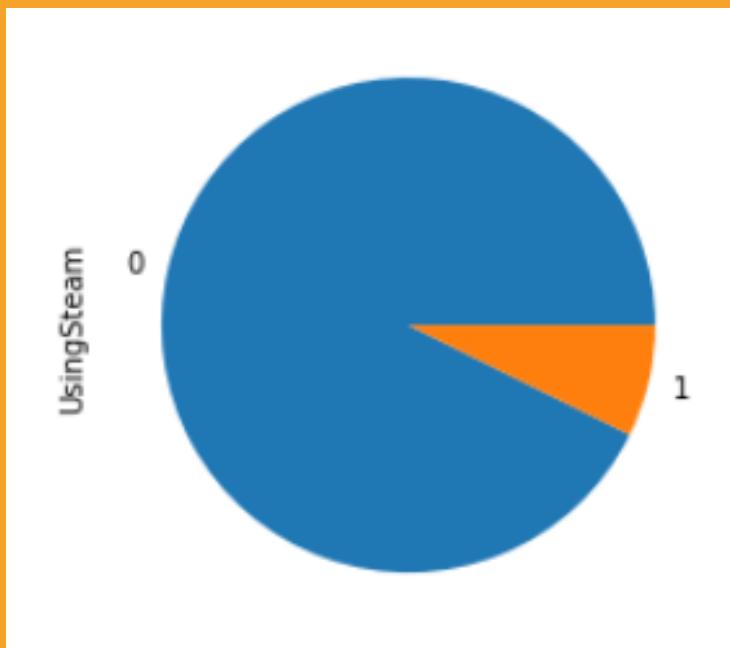
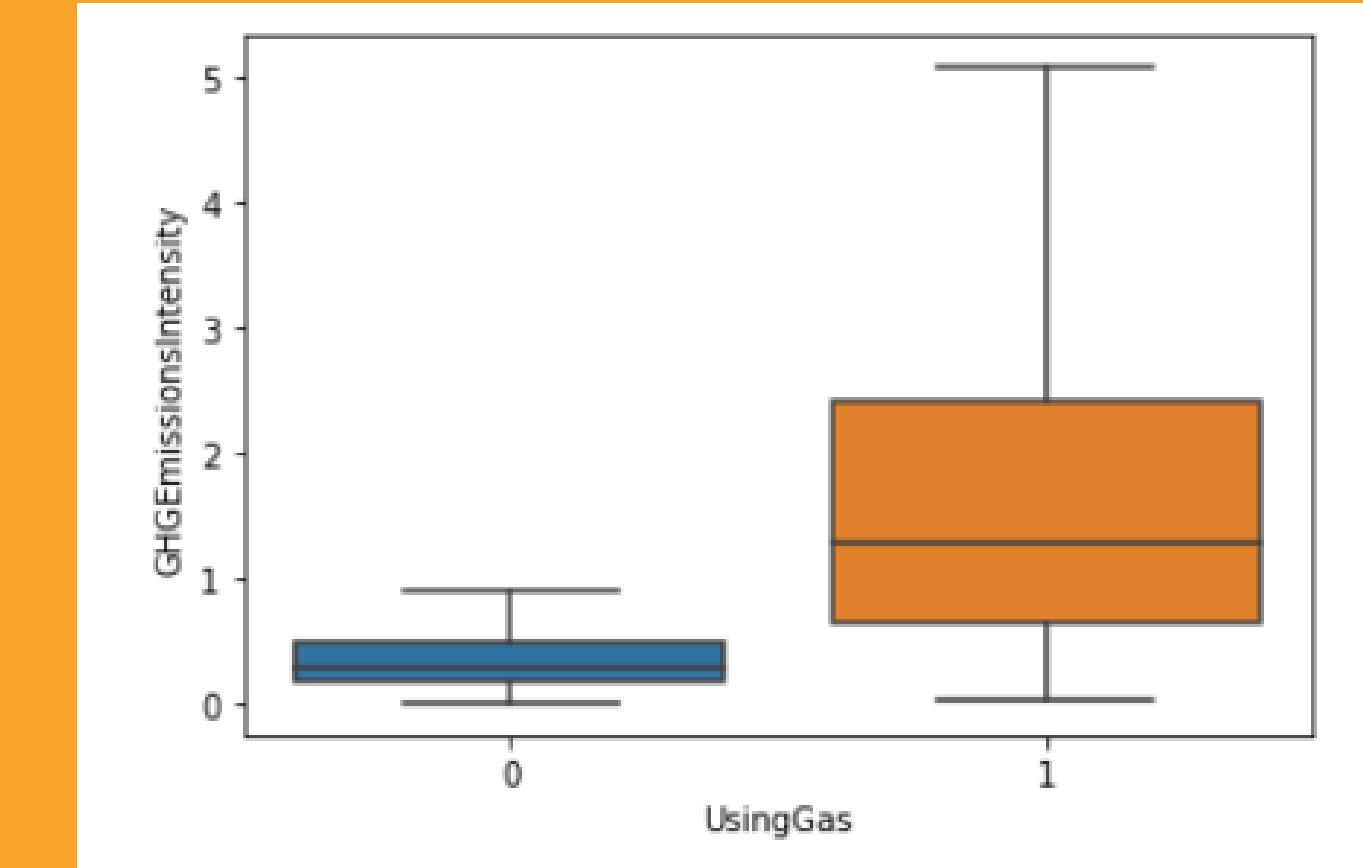
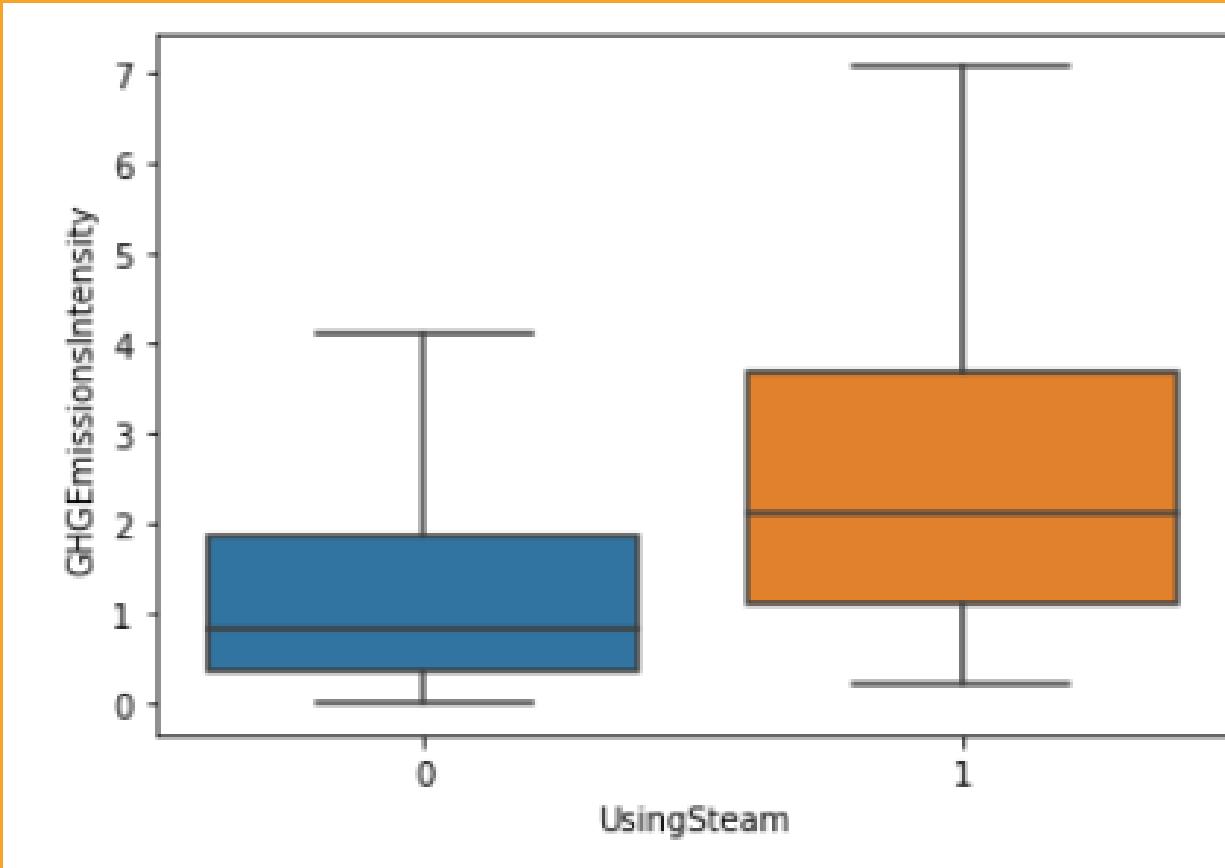


Energy Star Score



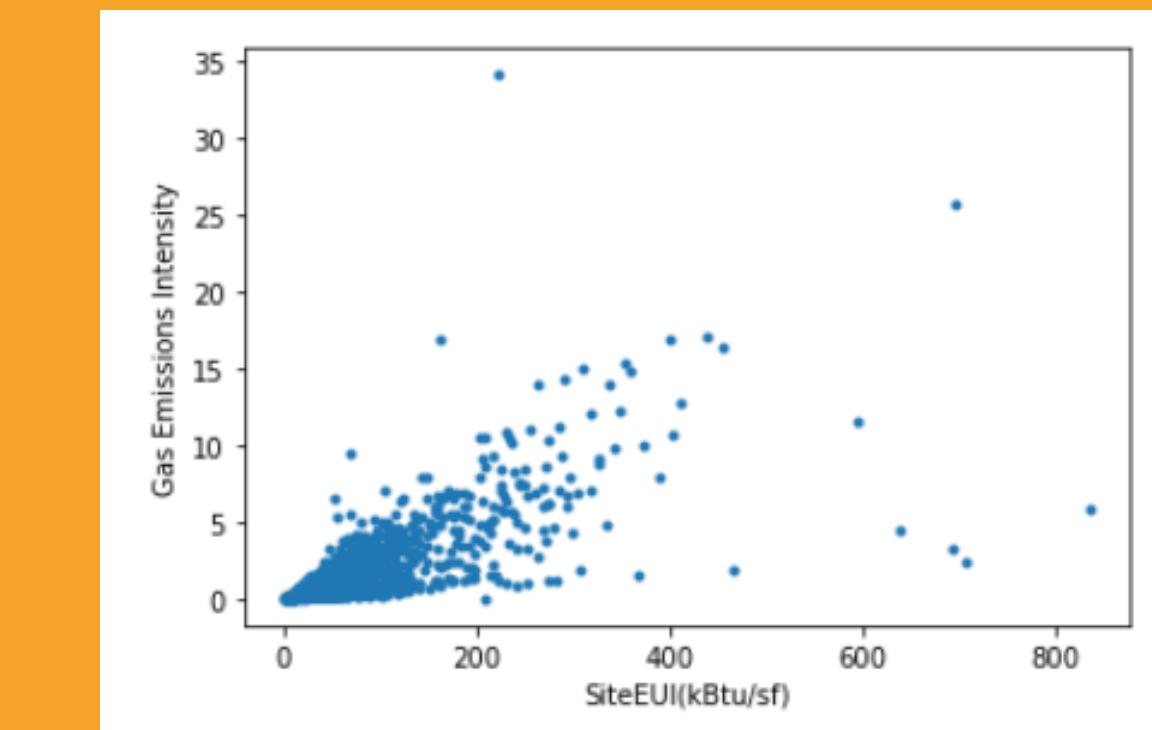
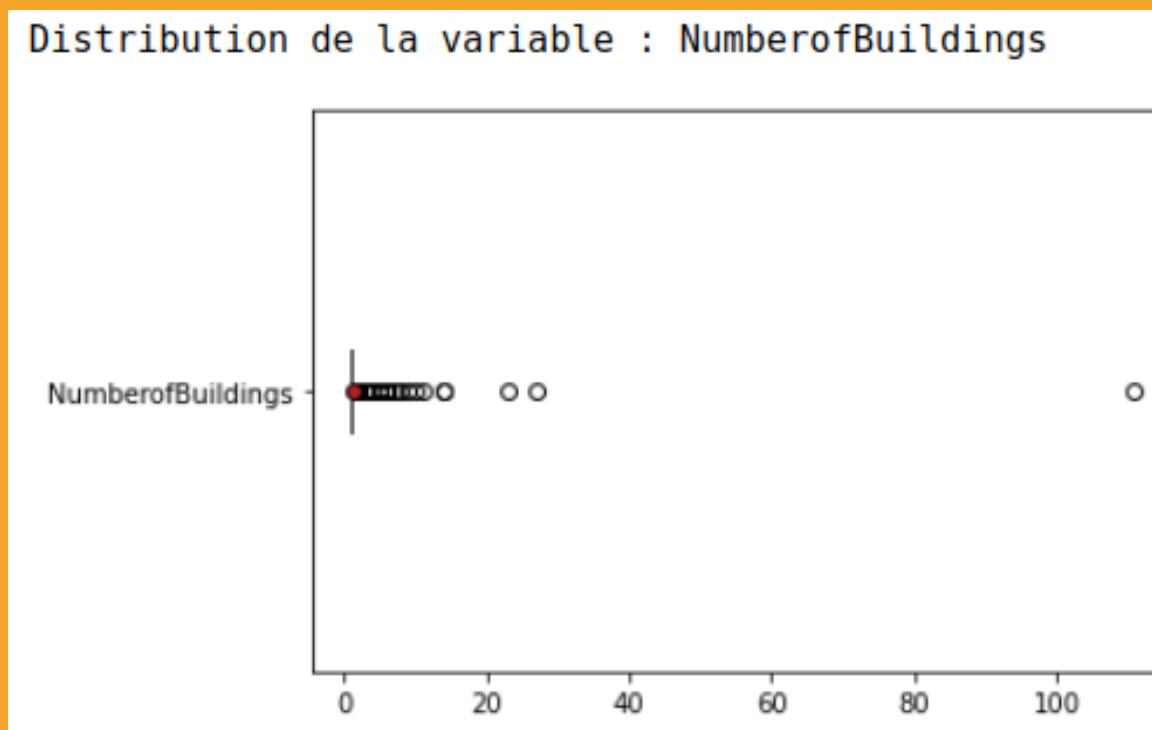
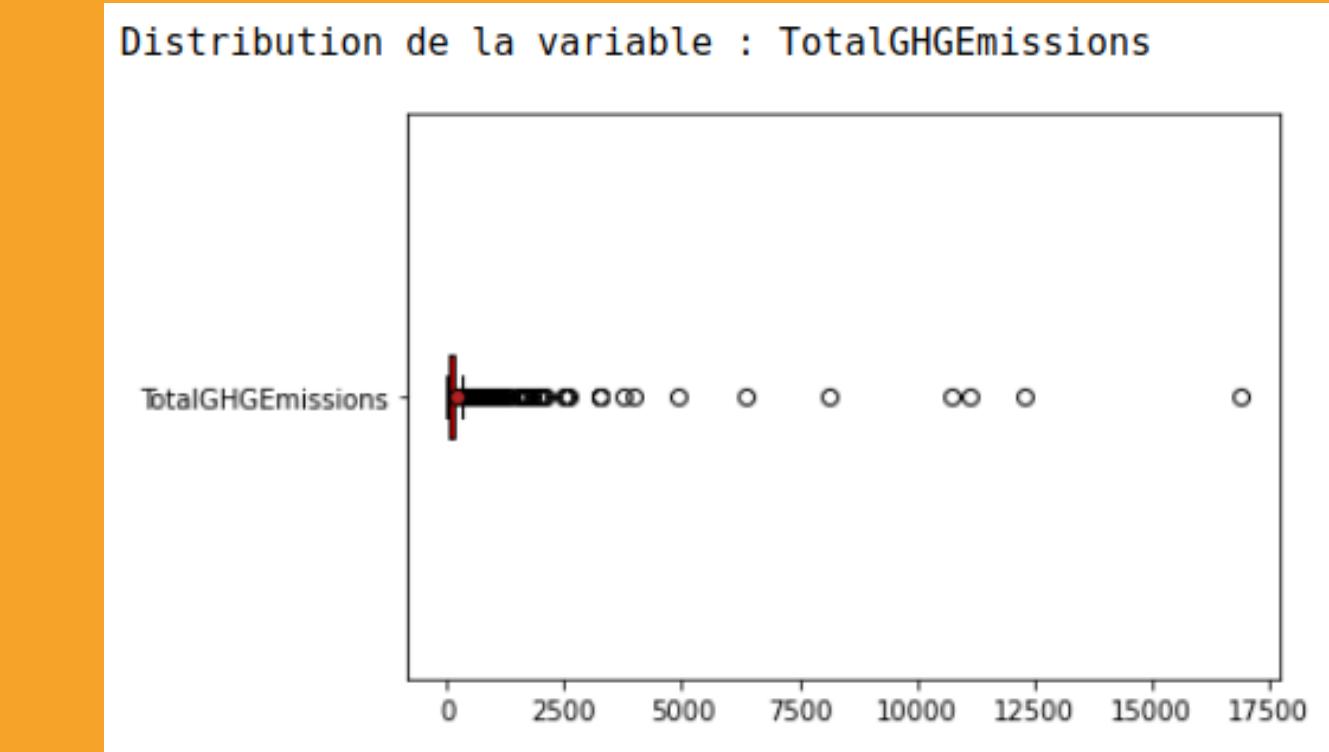
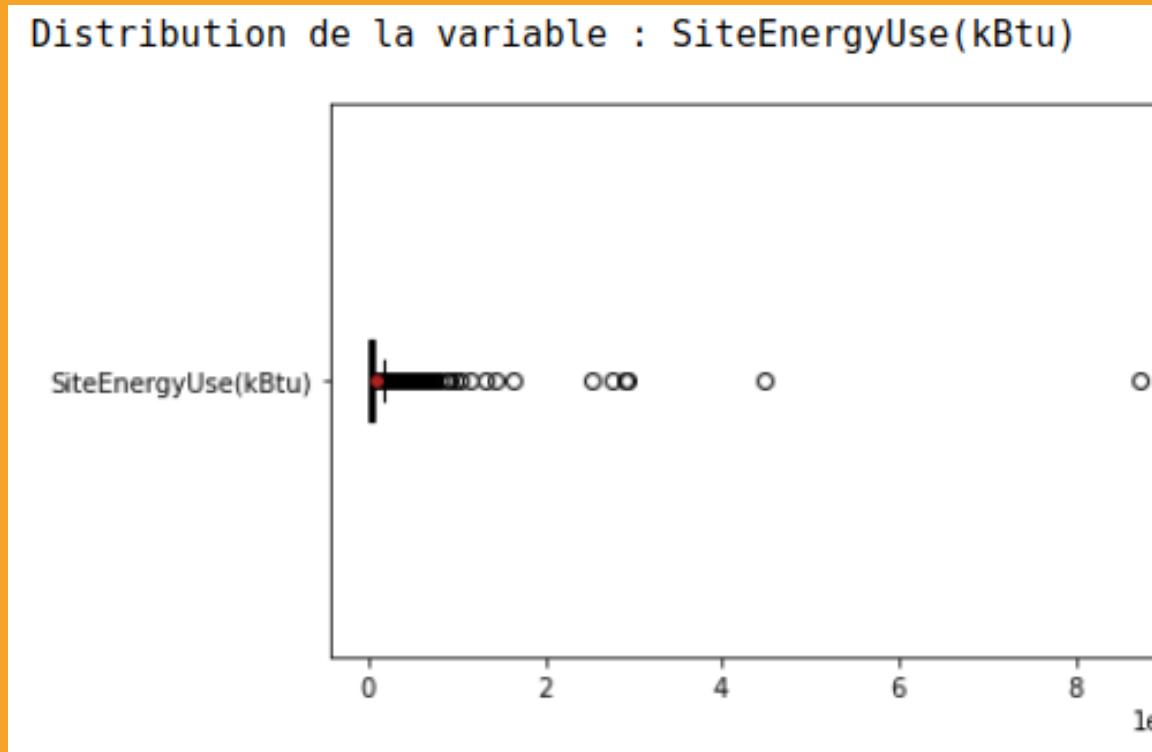
# ANALYSE EXPLORATOIRE

Energies utilisées



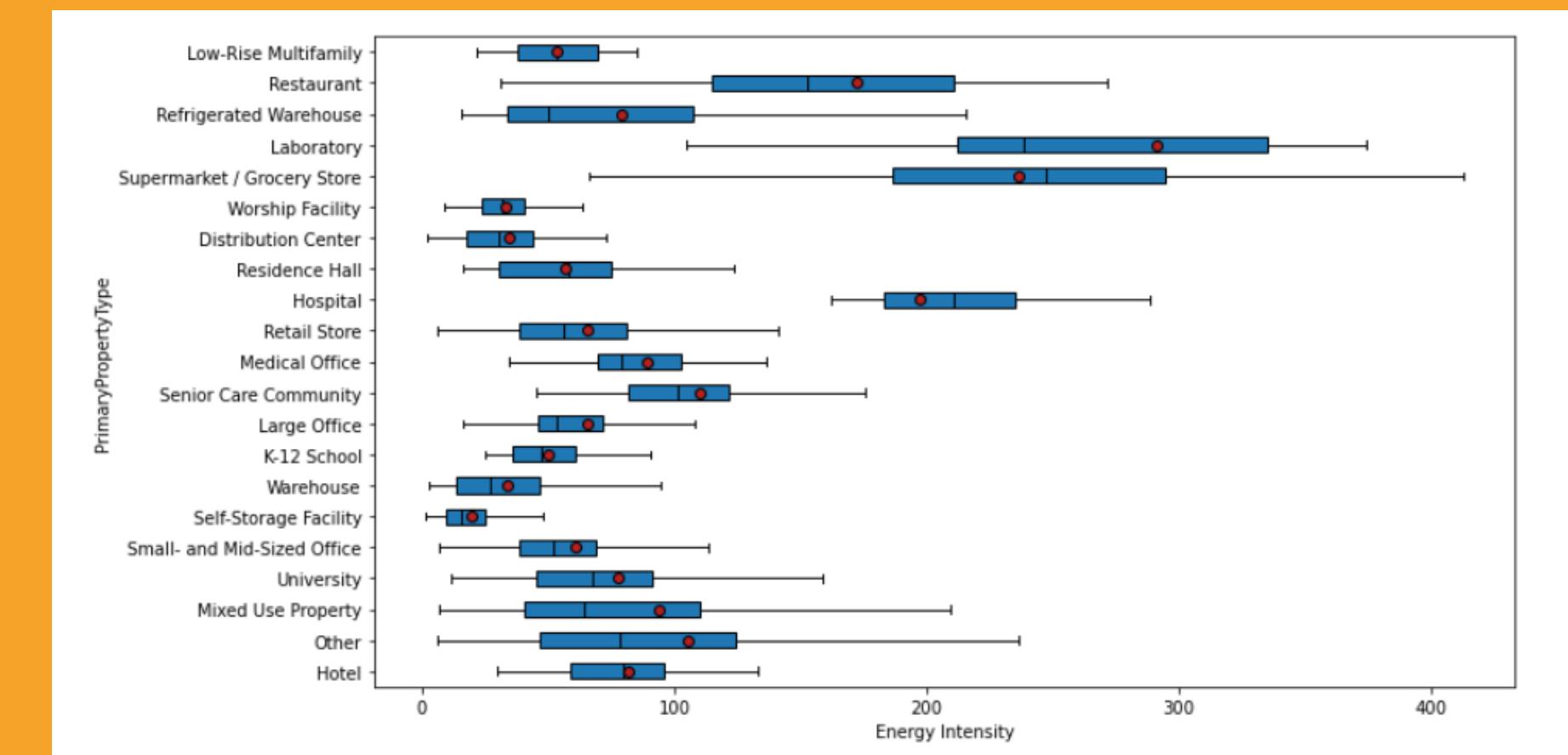
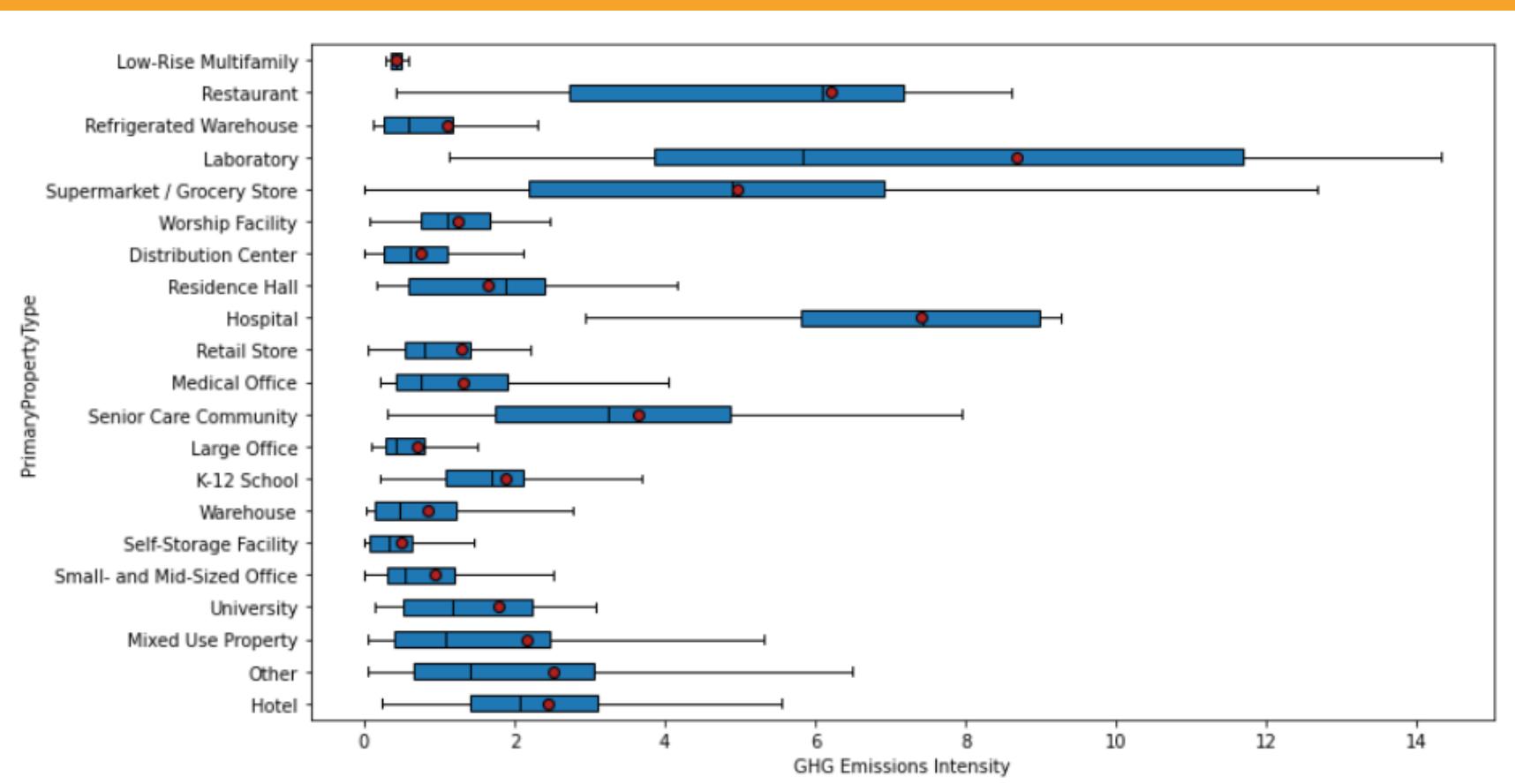
# ANALYSE EXPLORATOIRE

## Variables cibles



# ANALYSE EXPLORATOIRE

## Types de bâtiments énergivores



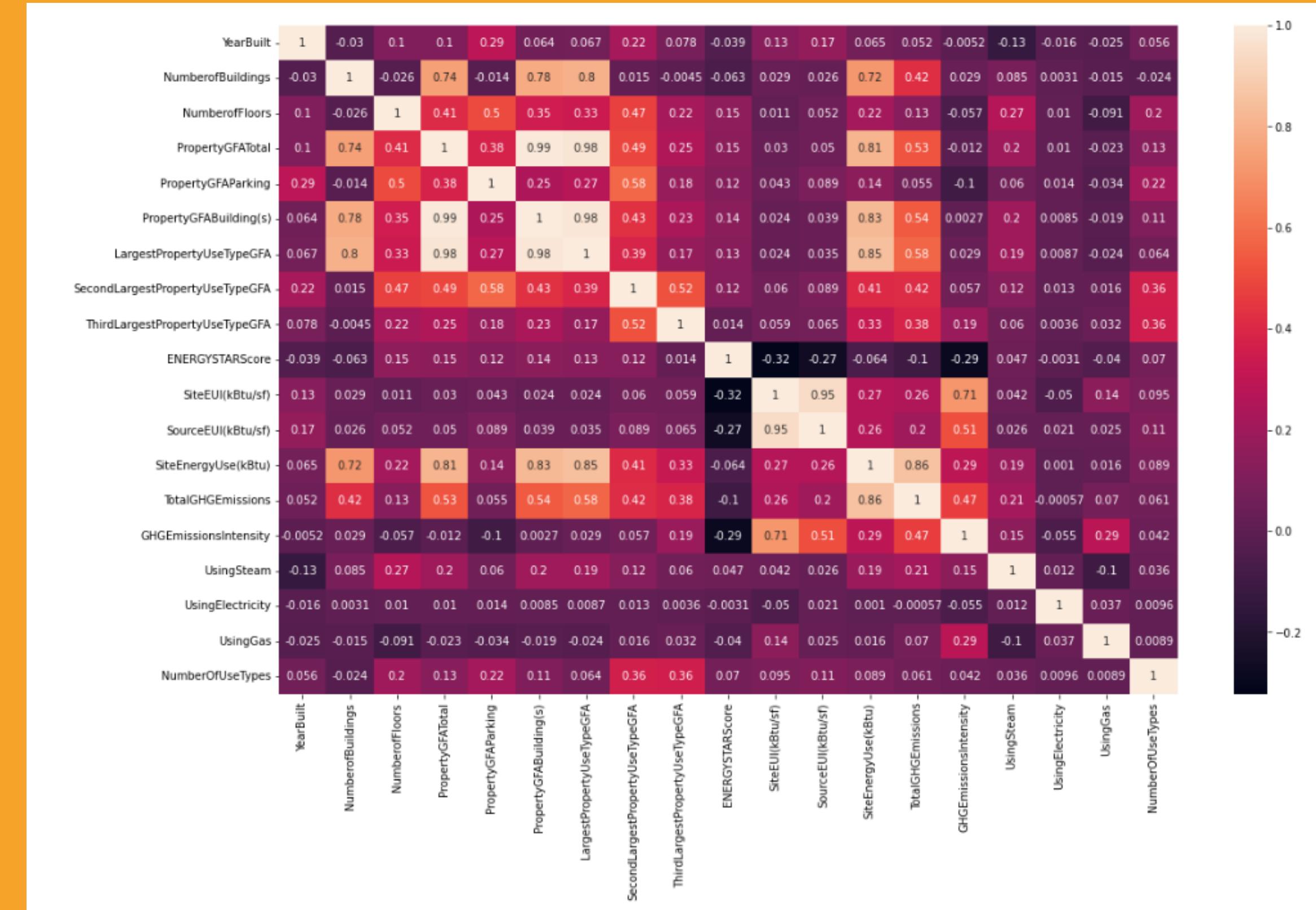
# ANALYSE EXPLORATOIRE

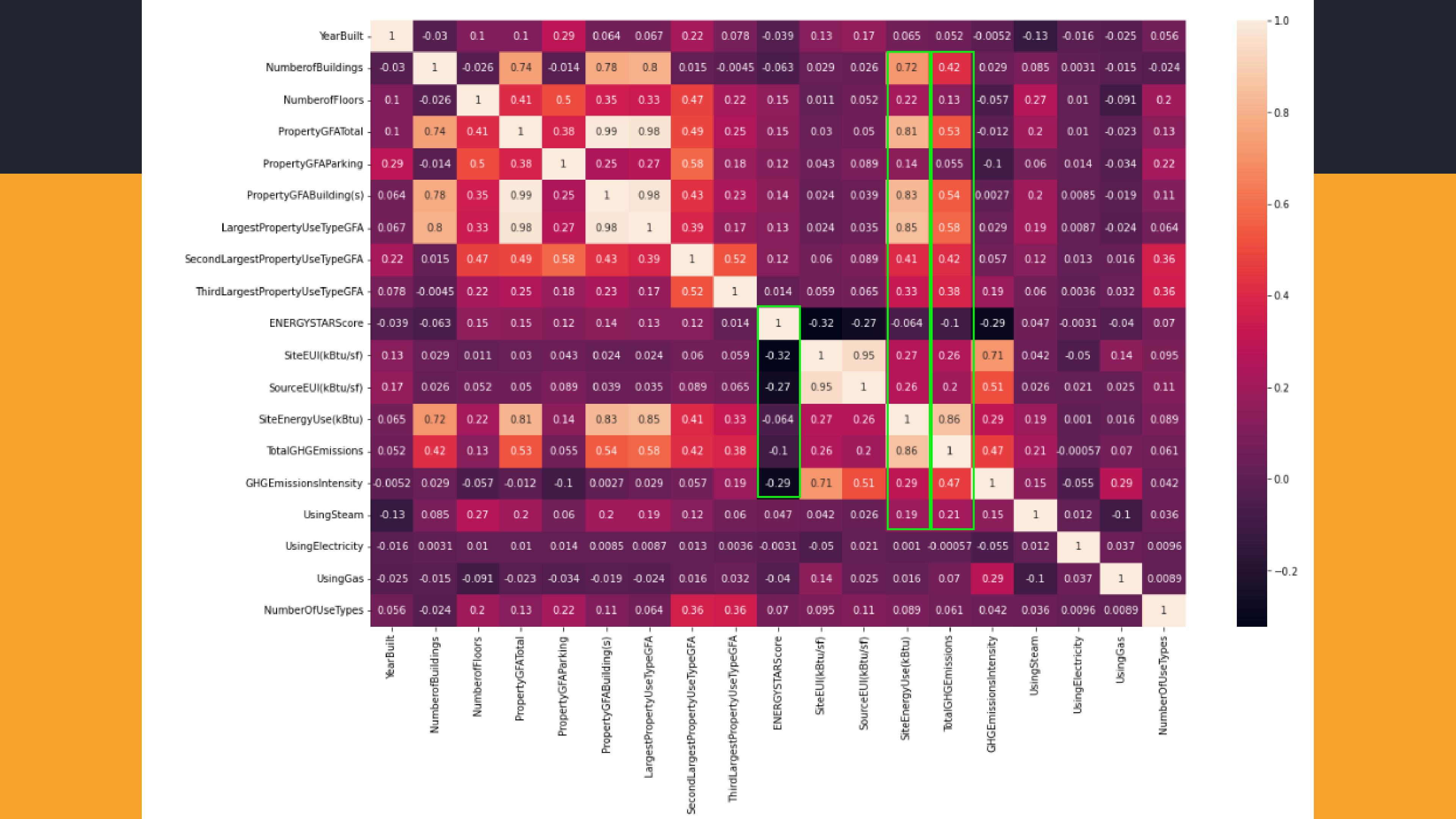
## Corrélations entre variables

Quelles sont les variables ayant un impact sur les variables cibles ?

- Les variables concernant la superficie(nombre de buildings, d'étages, surface au sol)
- Energy star score (anti correlées)
- Type energy use

-> Feature importance





# MODÈLES UTILISÉS POUR LES PRÉDICTIONS

Uniquement pour la consommation d'énergie

- Baseline : random et moyenne
- Régression Lasso
- Régression Ridge
- SVR
- KRR
- Réseau de neurones - MLP Regressor
- Gradient Boosting Regressor - XGBoost
  
- Standard Scaler
- Robust Scaler
- MinMax Scaler
  
- Utilisation de GridSearch pour trouver les hyperparamètres optimaux
- Scoring r2
- RMSE
- Max erreur



# MODÈLES CHOISI

## PRÉDICTION CONSOMMATION D'ÉNERGIE

- Gradient Boosting Regressor - XGBoost
- Dataframe choisi : full\_data\_with\_energy\_score
- Passage au log10 sur ma variable cible (meilleures performances)
- Hyperparamètres sélectionnés par cross validation :
  - learning rate : 0.03
  - loss : 'huber'
  - n estimators : 150
  - max depth : 6
  - random state : 33530
- Performances du modèle **avec** energy star score :
  - r2 train set : 0.97
  - **r2 validation set : 0.86**
  - r2 test set : 0.89
- Performances du modèle **sans** energy star score :
  - r2 train set : 0.95
  - **r2 validation set : 0.78**
  - r2 test set : 0.81



Il semble que la présence/absence de la variable energy star score ait un léger impact sur l'entraînement et les prédictions du modèle.

Cette hypothèse est à vérifier lors de la dernière étape, les feature importances.

# MODÈLES CHOISI

## PRÉDICTION ÉMISSION DE CO2

- Gradient Boosting Regressor - XGBoost
- Dataframe choisi : full\_data\_with\_energy\_score
- Passage au log10 sur ma variable cible (meilleures performances)
- Hyperparamètres sélectionnés par cross validation :
  - learning rate : 0.1
  - loss : 'huber'
  - n estimators : 100
  - max depth : 3
  - min\_samples\_leaf : 5
  - random state : 1
- Performances du modèle **avec** energy star score :
  - r2 train set : 0.91
  - **r2 validation set : 0.79**
  - r2 test set : 0.79
- Performances du modèle **sans** energy star score :
  - r2 train set : 0.86
  - **r2 validation set : 0.75**
  - r2 test set : 0.82



Ici, il semble que la variable energy star score ait un impact modéré sur l'entraînement et les prédictions du modèle. Les r2 scores sont légèrement différents. Cette hypothèse est à vérifier lors de la dernière étape, les feature importances.

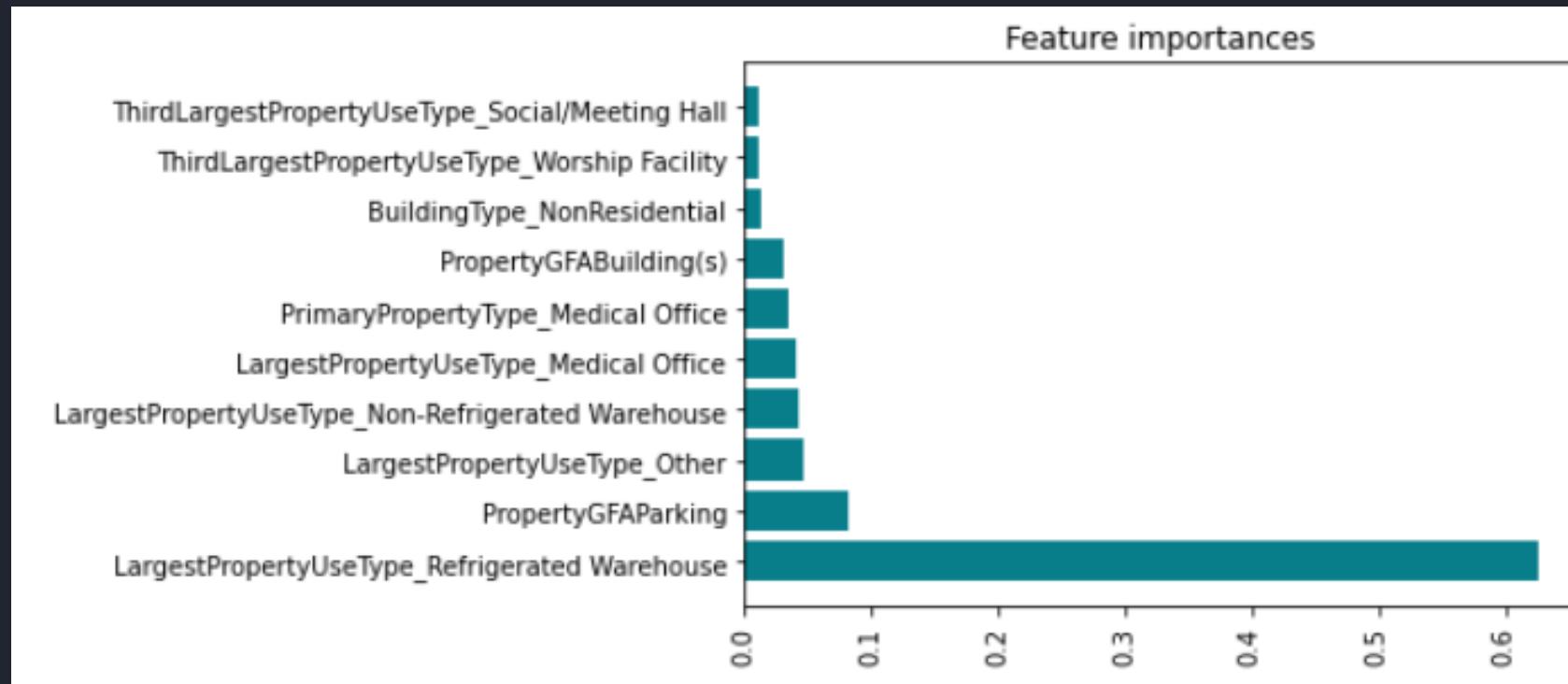


# IMPORTANCE DES VARIABLES

- L'objet GridSearchCV nous donne accès à `best_estimator` qui est le modèle utilisé avec les hyperparamètres choisis
- Notre modèle possède un attribut `feature_importances`
- Cet attribut nous retourne un tableau avec la valeur de l'influence de chaque variables sur notre modèle
- Création d'un dataframe contenant les variables et leur importance lors des prédictions du modèle

# PRÉDICTION CONSOMMATION D'ÉNERGIE

Avec variable energy star score



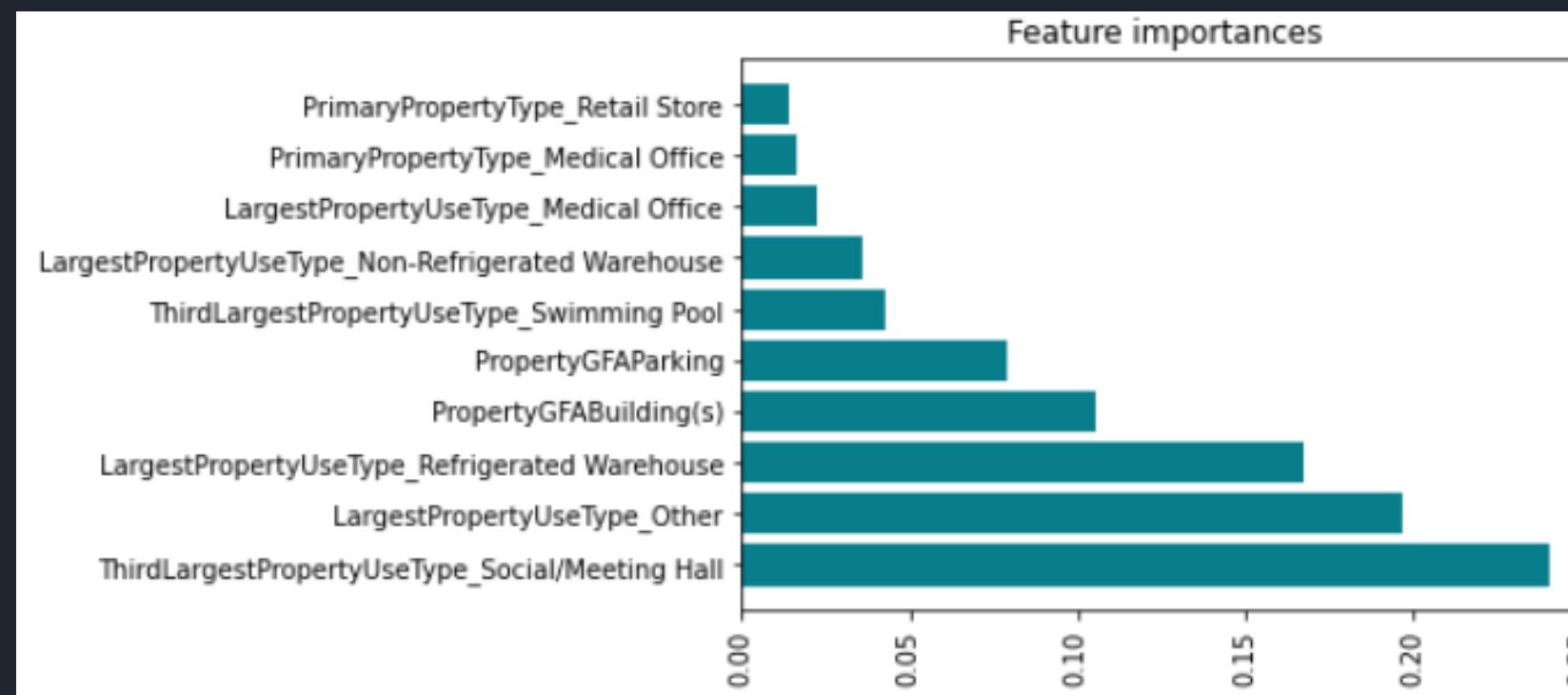
Avec variable energy star score

index		Variable	Importance
0	47	LargestPropertyUseType_Refrigerated Warehouse	0.626670
1	4	PropertyGFAParking	0.082887
2	45	LargestPropertyUseType_Other	0.046509
3	43	LargestPropertyUseType_Non-Refrigerated Warehouse	0.043191
4	42	LargestPropertyUseType_Medical Office	0.040475
5	23	PrimaryPropertyType_Medical Office	0.035585
6	5	PropertyGFABuilding(s)	0.031336
7	15	BuildingType_NonResidential	0.013952
8	116	ThirdLargestPropertyUseType_Worship Facility	0.011737
9	114	ThirdLargestPropertyUseType_Social/Meeting Hall	0.011429

Sans variable energy star score

index		Variable	Importance
0	46	LargestPropertyUseType_Refrigerated Warehouse	0.641682
1	44	LargestPropertyUseType_Other	0.067259
2	22	PrimaryPropertyType_Medical Office	0.042862
3	42	LargestPropertyUseType_Non-Refrigerated Warehouse	0.040587
4	115	ThirdLargestPropertyUseType_Worship Facility	0.036800
5	41	LargestPropertyUseType_Medical Office	0.035886
6	4	PropertyGFAParking	0.032360
7	14	BuildingType_NonResidential	0.016454
8	113	ThirdLargestPropertyUseType_Social/Meeting Hall	0.013638
9	7	SecondLargestPropertyUseTypeGFA	0.006773

Avec variable energy star score



# PRÉDICTION EMISSION DE CO<sub>2</sub>

Avec variable energy star score

index		Variable	Importance
0	114	ThirdLargestPropertyUseType_Social/Meeting Hall	0.240575
1	45	LargestPropertyUseType_Other	0.196632
2	47	LargestPropertyUseType_Refrigerated Warehouse	0.167010
3	5	PropertyGFABuilding(s)	0.105059
4	4	PropertyGFAParking	0.079051
5	115	ThirdLargestPropertyUseType_Swimming Pool	0.042735
6	43	LargestPropertyUseType_Non-Refrigerated Warehouse	0.036136
7	42	LargestPropertyUseType_Medical Office	0.022039
8	23	PrimaryPropertyType_Medical Office	0.016004
9	28	PrimaryPropertyType_Retail Store	0.013864

Sans variable energy star score

index		Variable	Importance
0	113	ThirdLargestPropertyUseType_Social/Meeting Hall	0.233654
1	44	LargestPropertyUseType_Other	0.233553
2	46	LargestPropertyUseType_Refrigerated Warehouse	0.152216
3	4	PropertyGFAParking	0.125460
4	114	ThirdLargestPropertyUseType_Swimming Pool	0.041033
5	42	LargestPropertyUseType_Non-Refrigerated Warehouse	0.024174
6	41	LargestPropertyUseType_Medical Office	0.020963
7	14	BuildingType_NonResidential	0.019854
8	27	PrimaryPropertyType_Retail Store	0.019793
9	32	PrimaryPropertyType_Worship Facility	0.019617



# ENERGY STAR SCORE



Consommation d'énergie

Evaluer l'intérêt de l'ENERGY STAR Score :

La variable energy star score se trouve en 10ème / 80ème position en ce qui concerne l'influence des variables sur mon modèle.

Le label n'est pas à négliger ou dénigrer car il est représentatif de la consommation du bâtiment bien que son influence est infime dans le cadre de ce projet de machine learning.

index	Variable	Importance
10	9 ENERGYSTARScore	0.007324

Emission de CO2

index	Variable	Importance
80	9 ENERGYSTARScore	0.0



**MERCI DE  
VOTRE  
ATTENTION**