P4 PROJET ENERGIE SEATTLE VERTE 2050

CATHERINE BRICE 15/11/2021

OPENCLASSROOMS

FORMATION DATA SCIENTIST

SOMMAIRE

INTRODUCTION

SEATTLE = ville neutre en émissions de carbone en 2050

PREDIRE :

Consommation énergétique de bâtiments non résidentiels 'SiteEnergyUse(kBtu)'

Emissions co2 'TotalGHGEmissions'

Modélisation ML supervisée à partir de 2 relevés 2015-2016 (gas, électricité, vapeur)

NETTOYAGE DES DONNÉES

df_15.shape (3340, 42)

```
]: df_15.sample()
```

	OSEBuildingID	DataYear	BuildingType	PrimaryPropertyType	PropertyName	TaxParcelIdentificationNumber	Locati
1031	20522	2015	NonResidential	Non-Refrigerated Warehouse	MB WAREHOUSE LLC	1498302235	{"human_addres '{"address":"20 21ST AVE S

df_16.shape (3376, 46)

```
]: df_16.sample()
```

	OSEBuildingID	DataYear	BuildingType	PrimaryPropertyType	PropertyName	Address	City	State	ZipCode	TaxParcelIdentificationNumber	Coun
2026	24179	2016	Multifamily LR (1-4)	Low-Rise Multifamily	Northbrook Place	10215 Lake City Way NE	Seattle	WA	98125.0	5101405948	

NETTOYAGE DES DONNÉES

2 structures différentes entre relevé 2015 et relevé 2016

sur 2015 et pas 2016

['Location', 'OtherFuelUse(kBtu)', 'GHGEmissions(MetricTonsCO2e)', 'GHGEmissionsIntensity(kgCO2e/ft2)', 'Comment']

sur 2016 et pas 2015

['Address', 'City', 'State', 'ZipCode', 'Latitude', 'Longitude', 'Comments', 'TotalGHGEmissions', 'GHGEmissionsIntensity']

2015 : 3340 lignes, 42 features

+

2016 : 3376 lignes, 46 features

=

3318 lignes NonResidential , 46 features après nettoyage des données

Les types de batiments

['NonResidential' 'Nonresidential COS' 'Multifamily MR (5-9)' ... 'Campus' 'Multifamily HR (10+)' 'Nonresidential WA']

Batiments non résidentiels

['NonResidential' 'Nonresidential COS' 'SPS-District K-12' 'Campus' 'Nonresidential WA']

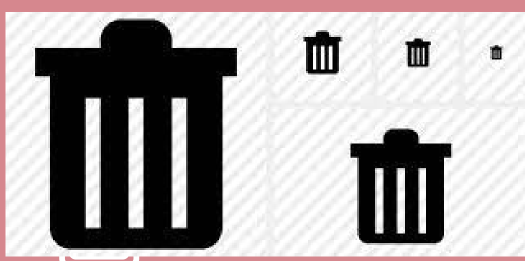
NETTOYAGE DES DONNÉES

```
---  ------                           --------------  -----
 0   OSEBuildingID                    3318 non-null   int64
 1   DataYear                         3318 non-null   int64
 2   BuildingType                     3318 non-null   object
 3   PrimaryPropertyType              3318 non-null   object
 4   PropertyName                     3318 non-null   object
 5   Address                          3318 non-null   object
 6   City                             3318 non-null   object
 7   State                            3318 non-null   object
 8   ZipCode                          3302 non-null   object
 9   TaxParcelIdentificationNumber    3317 non-null   object
 10  CouncilDistrictCode              3318 non-null   int64
 11  Neighborhood                     3318 non-null   object
 12  Latitude                         3318 non-null   object
 13  Longitude                        3318 non-null   object
 14  YearBuilt                        3318 non-null   int64
 15  NumberofBuildings                3316 non-null   float64
 16  NumberofFloors                   3310 non-null   float64
 17  PropertyGFATotal                 3318 non-null   int64
 18  PropertyGFAParking               3318 non-null   int64
 19  PropertyGFABuilding(s)           3318 non-null   int64
 20  ListOfAllPropertyUseTypes        3255 non-null   object
 21  LargestPropertyUseType           3247 non-null   object
 22  LargestPropertyUseTypeGFA        3247 non-null   float64
 23  SecondLargestPropertyUseType     1667 non-null   object
 24  SecondLargestPropertyUseTypeGFA  1667 non-null   float64
 25  ThirdLargestPropertyUseType      684 non-null    object
 26  ThirdLargestPropertyUseTypeGFA   684 non-null    float64
 27  YearsENERGYSTARCertified         188 non-null    object
 28  ENERGYSTARScore                  2211 non-null   float64
 29  SiteEUI(kBtu/sf)                 3308 non-null   float64
 30  SiteEUIWN(kBtu/sf)               3308 non-null   float64
 31  SourceEUI(kBtu/sf)               3309 non-null   float64
 32  SourceEUIWN(kBtu/sf)             3309 non-null   float64
 33  SiteEnergyUse(kBtu)              3309 non-null   float64
 34  SiteEnergyUseWN(kBtu)            3308 non-null   float64
 35  SteamUse(kBtu)                   3309 non-null
 36  Electricity(kWh)                 3309 non-null
 37  Electricity(kBtu)                3309 non-null
 38  NaturalGas(therms)               3309 non-null
 39  NaturalGas(kBtu)                 3309 non-null
 40  DefaultData                      3317 non-null
 41  Comments                         12 non-null
 42  ComplianceStatus                 3318 non-null
 43  Outlier                          48 non-null
 44  TotalGHGEmissions                3309 non-null
 45  GHGEmissionsIntensity            3309 non-null
dtypes: float64(19), int64(7), object(20)
```

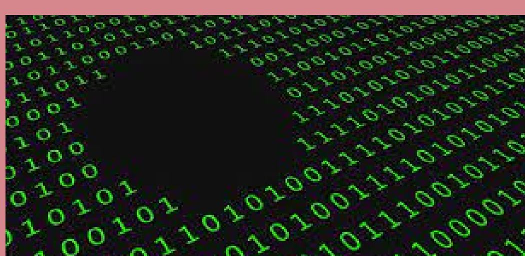
NETTOYAGE DES DONNÉES

Décompactage de la variable Location
---> Latitude", "Longitude"
"Address","City",
"State", "ZipCode"

Harmonisation (renommage) des noms de variables ('TotalGHGEmissions' et 'GHGEmissions(MetricTonsCO2e')

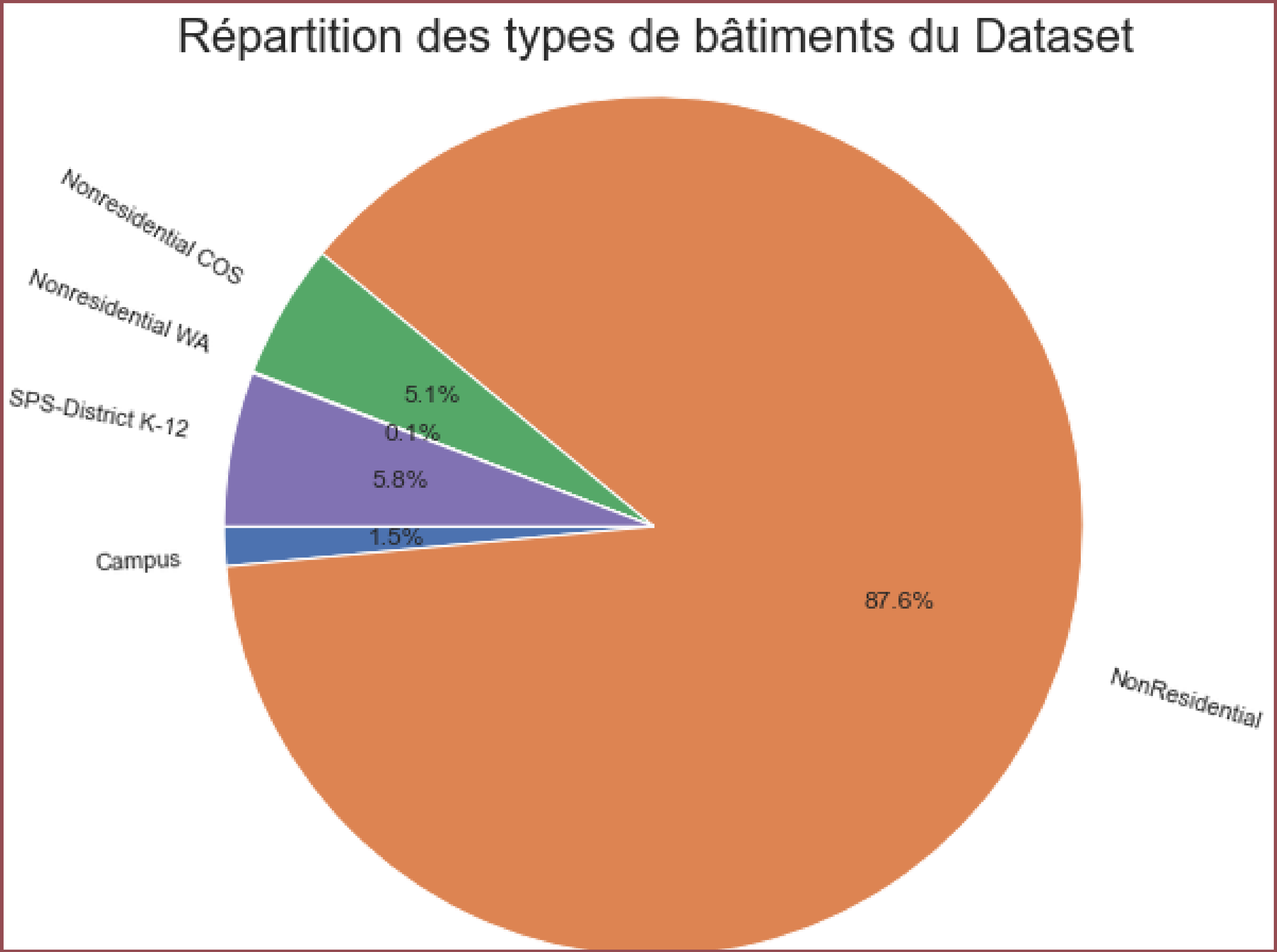


Suppression variables inutiles (Comments, Outliers,YearENERGYSTARScorecertified), redondantes (consommations énergie...)

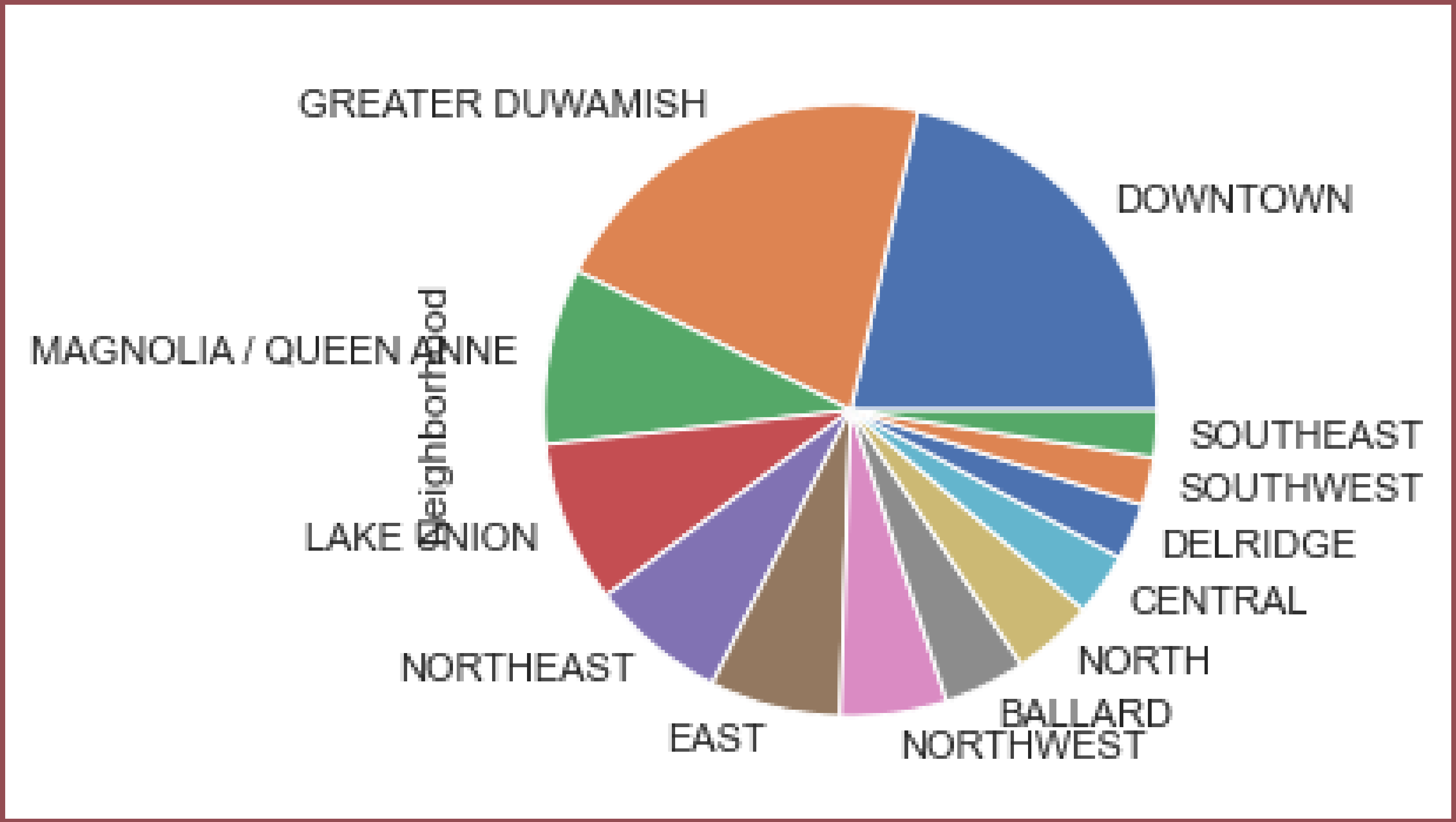


Suppression des valeurs manquantes ou pas remplies à 50%
ListOfAllPropertyUseTypes, Largest..., Third...

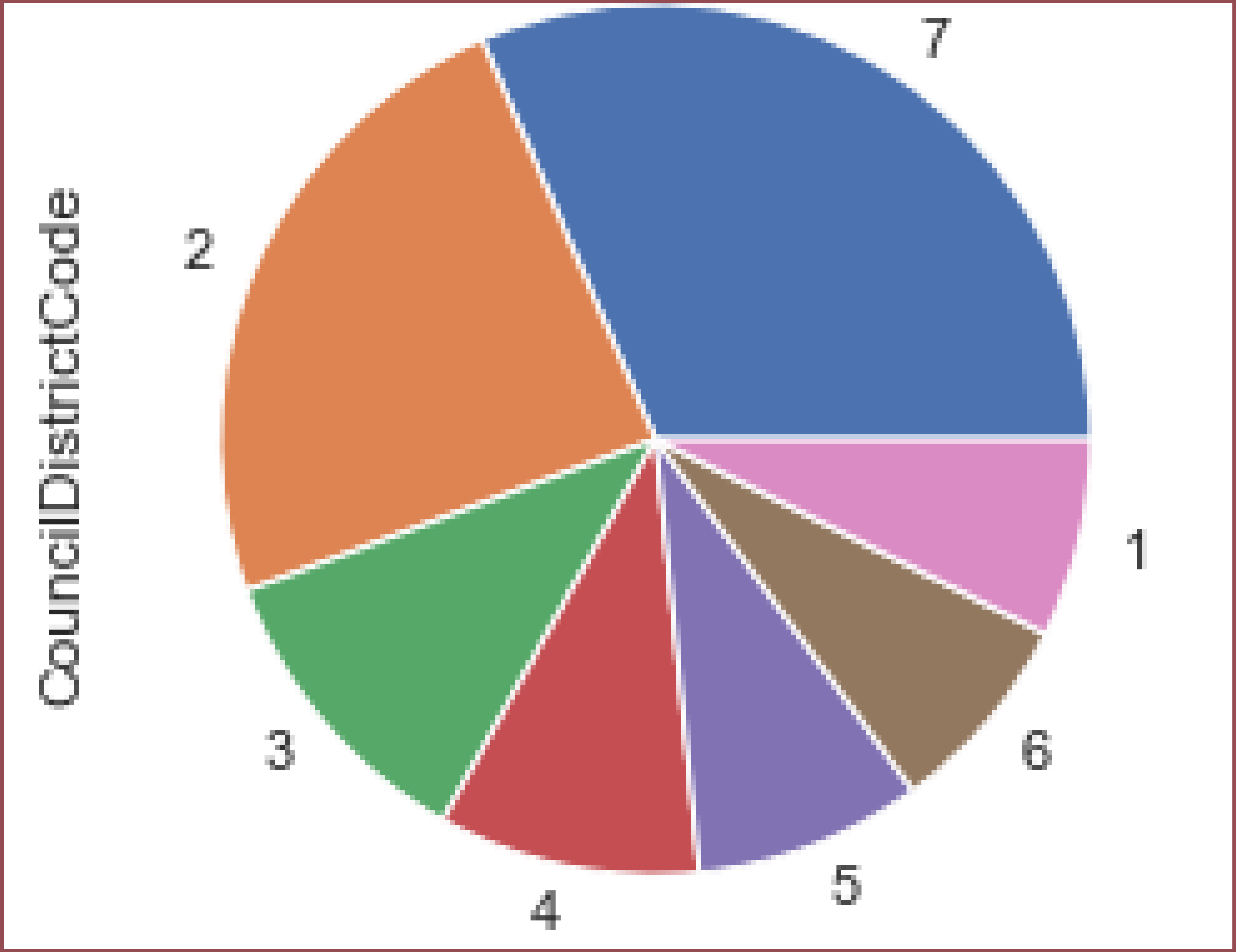
ANALYSE UNIVARIEE



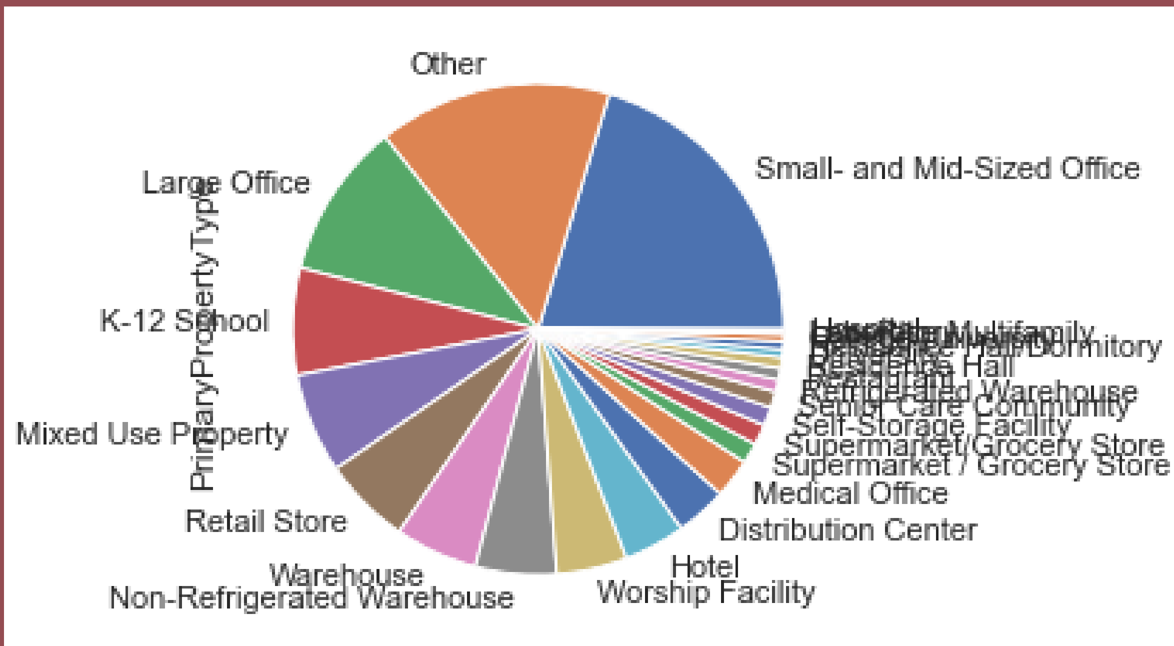
ANALYSE UNIVARIEE - 13 NEIGHBORHOODS



ANALYSE UNIVARIEE - CouncilDistrictCode



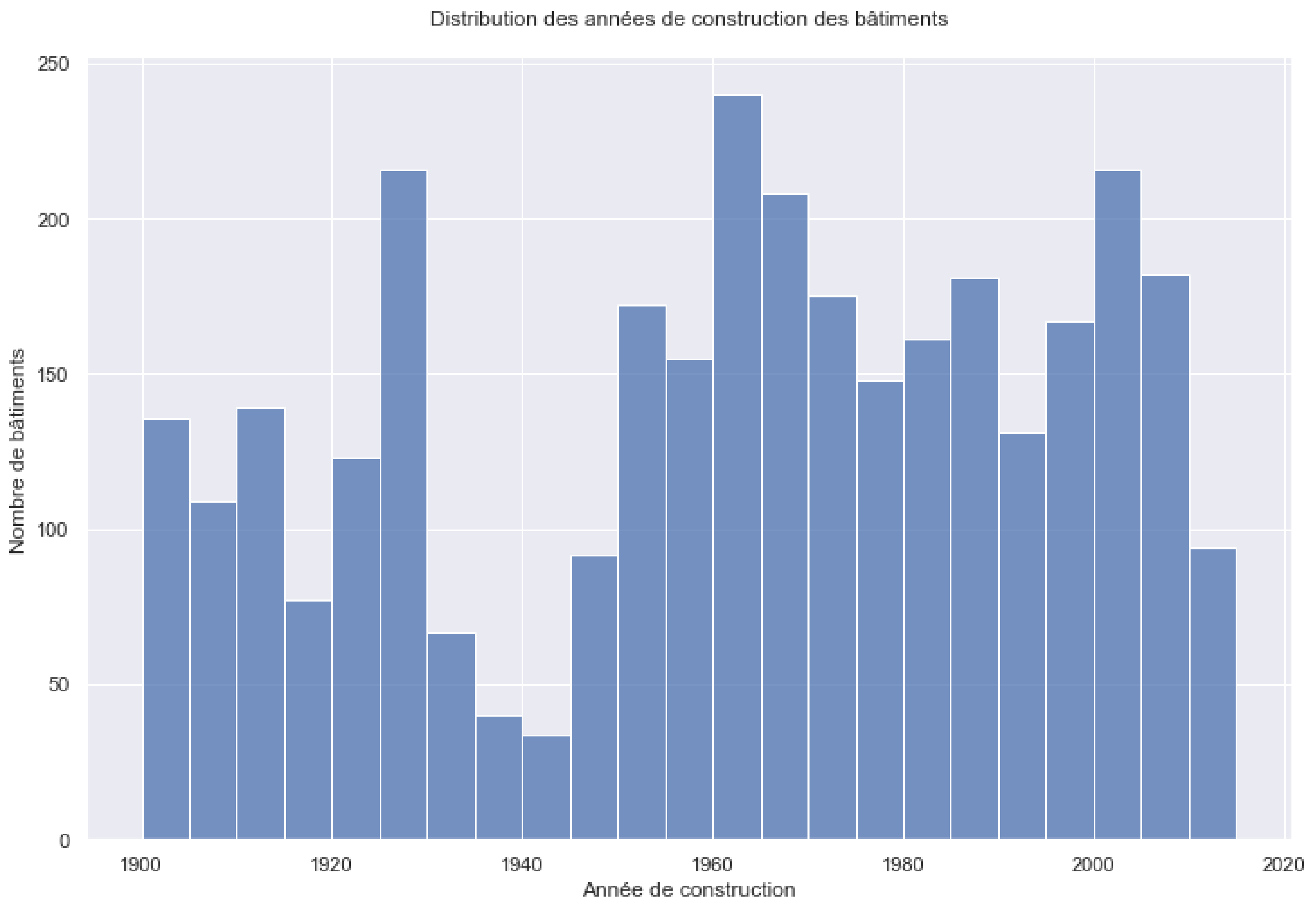
ANALYSE UNIVARIEE - primaryPropertyType



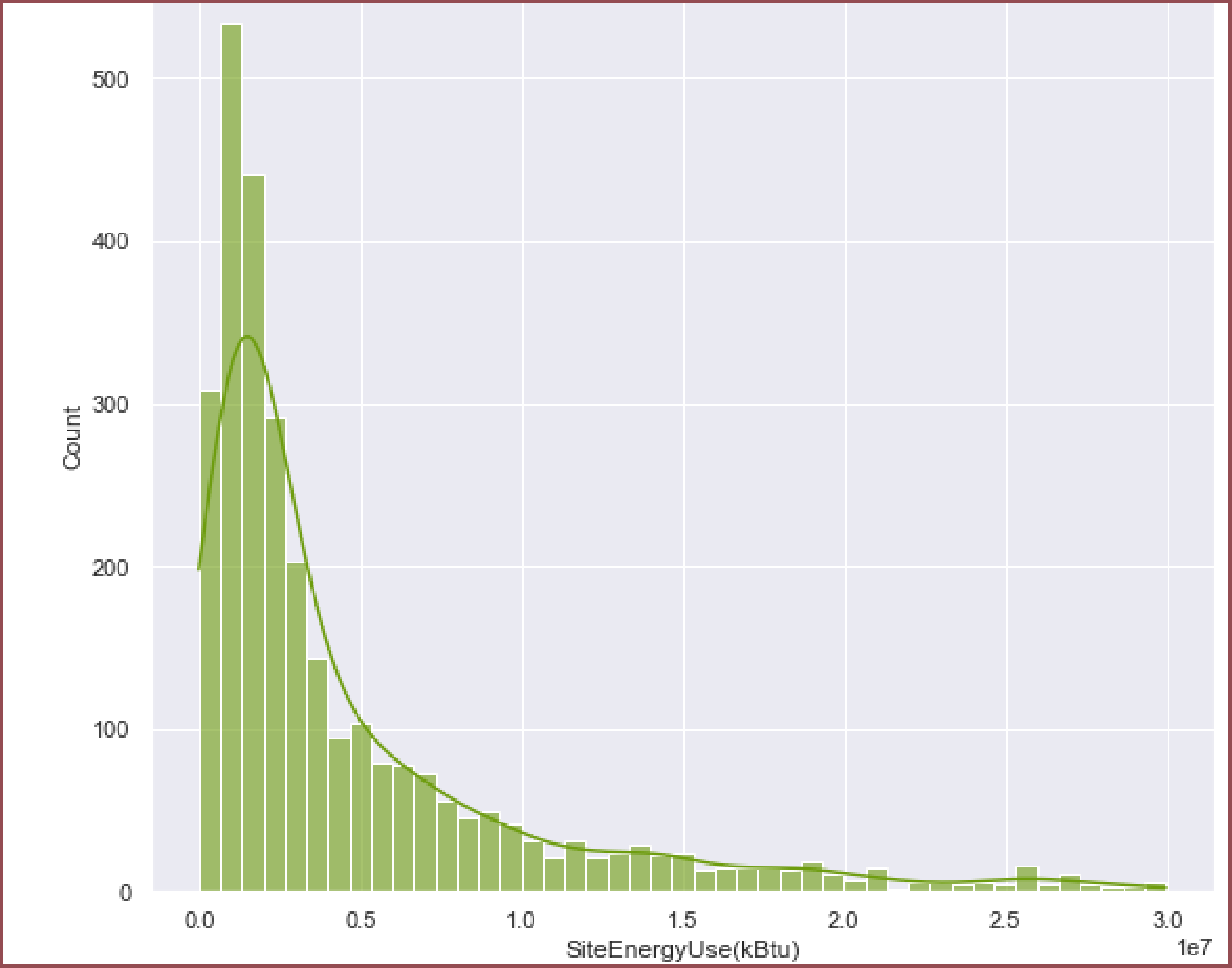
Small- and Mid-Sized Office	571
Other	499
Large Office	331
K-12 School	267
Mixed Use Property	219
Warehouse	187
Non-Refrigerated Warehouse	187
Retail Store	185
Hotel	146
Worship Facility	141
Distribution Center	106
Medical Office	82
Self-Storage Facility	56
Supermarket / Grocery Store	40
Senior Care Community	39
Supermarket/Grocery Store	36
Refrigerated Warehouse	25
University	24
Restaurant	23
College/University	21
Residence Hall	21
Hospital	20
Residence Hall/Dormitory	15
Laboratory	11
Low-Rise Multifamily	4
SPS-District K-12	4
Office	3

Name: PrimaryPropertyType, dtype: int64

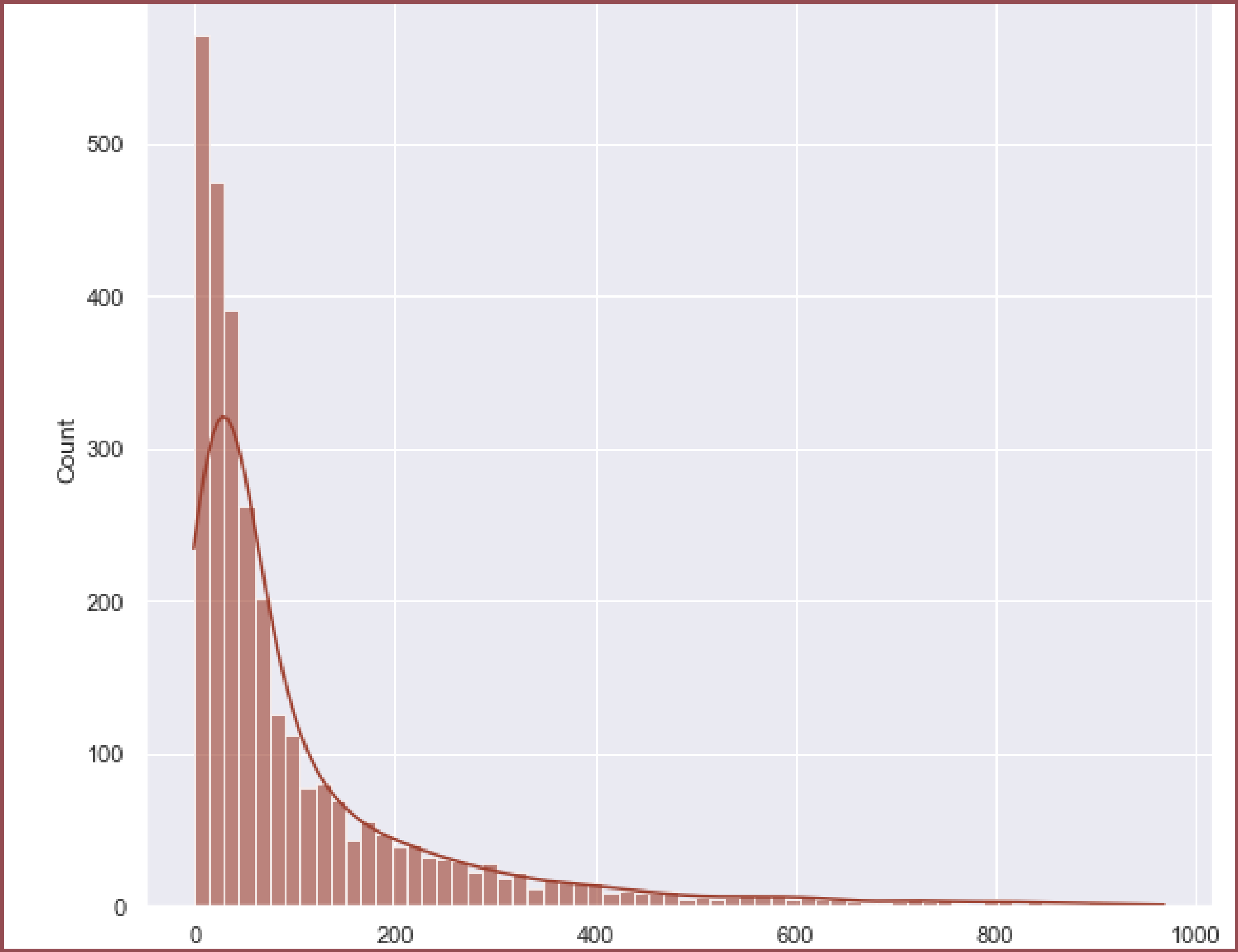
ANALYSE UNIVARIEE - ANNEES DE CONSTRUCTION



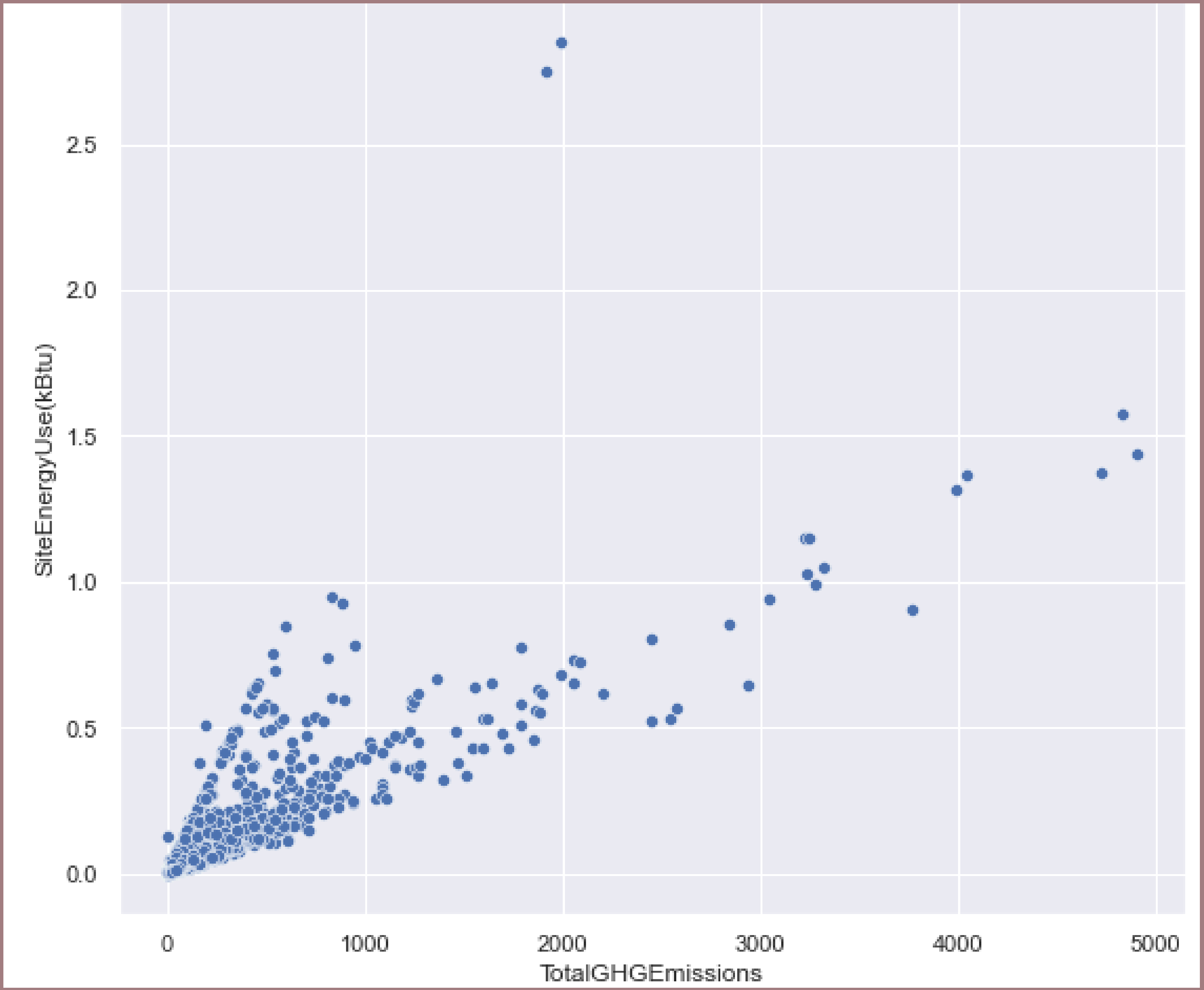
ANALYSE UNIVARIEE - SiteEnergyUse(kBtu)



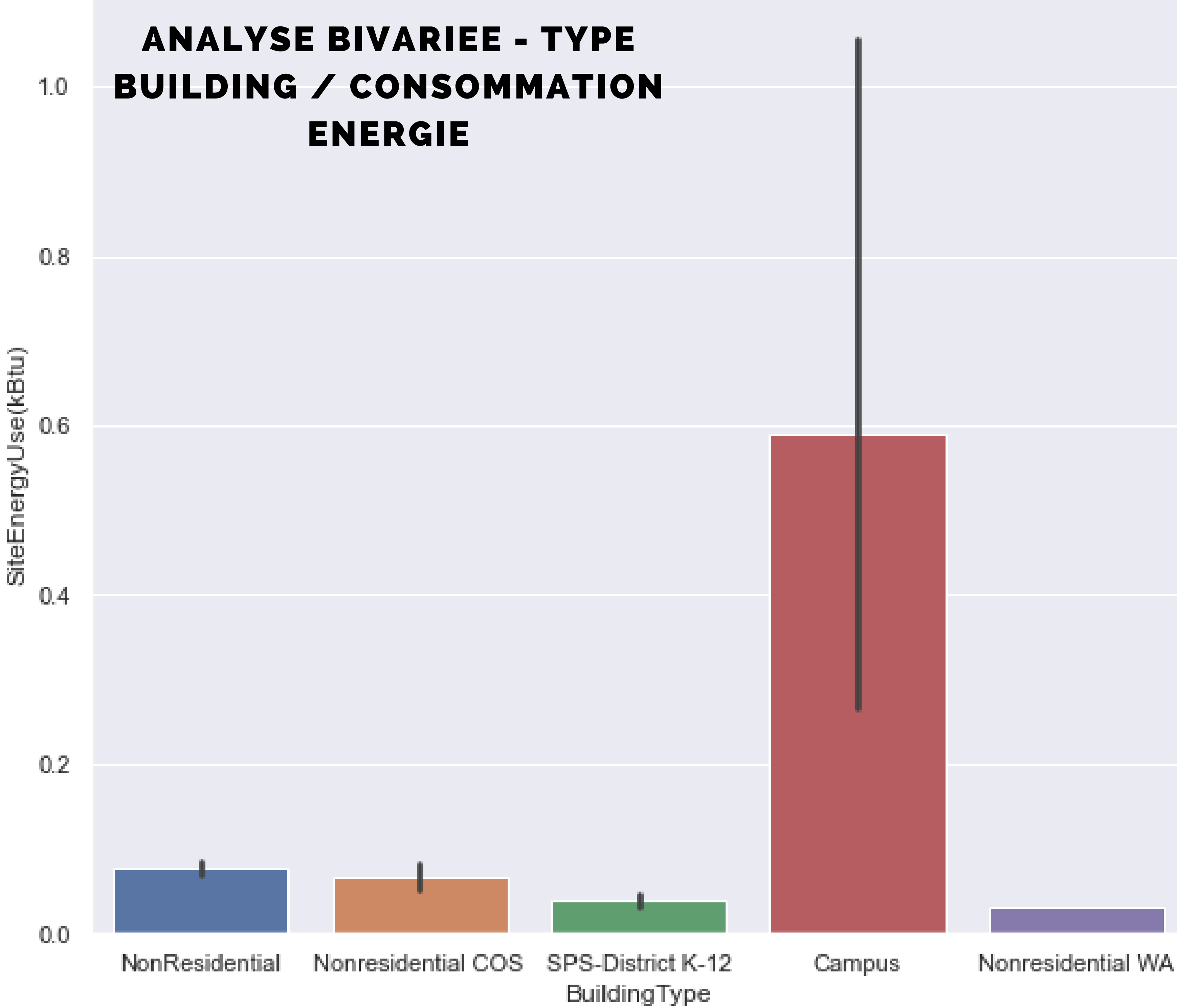
ANALYSE UNIVARIEE - EMISSIONS DE CO2



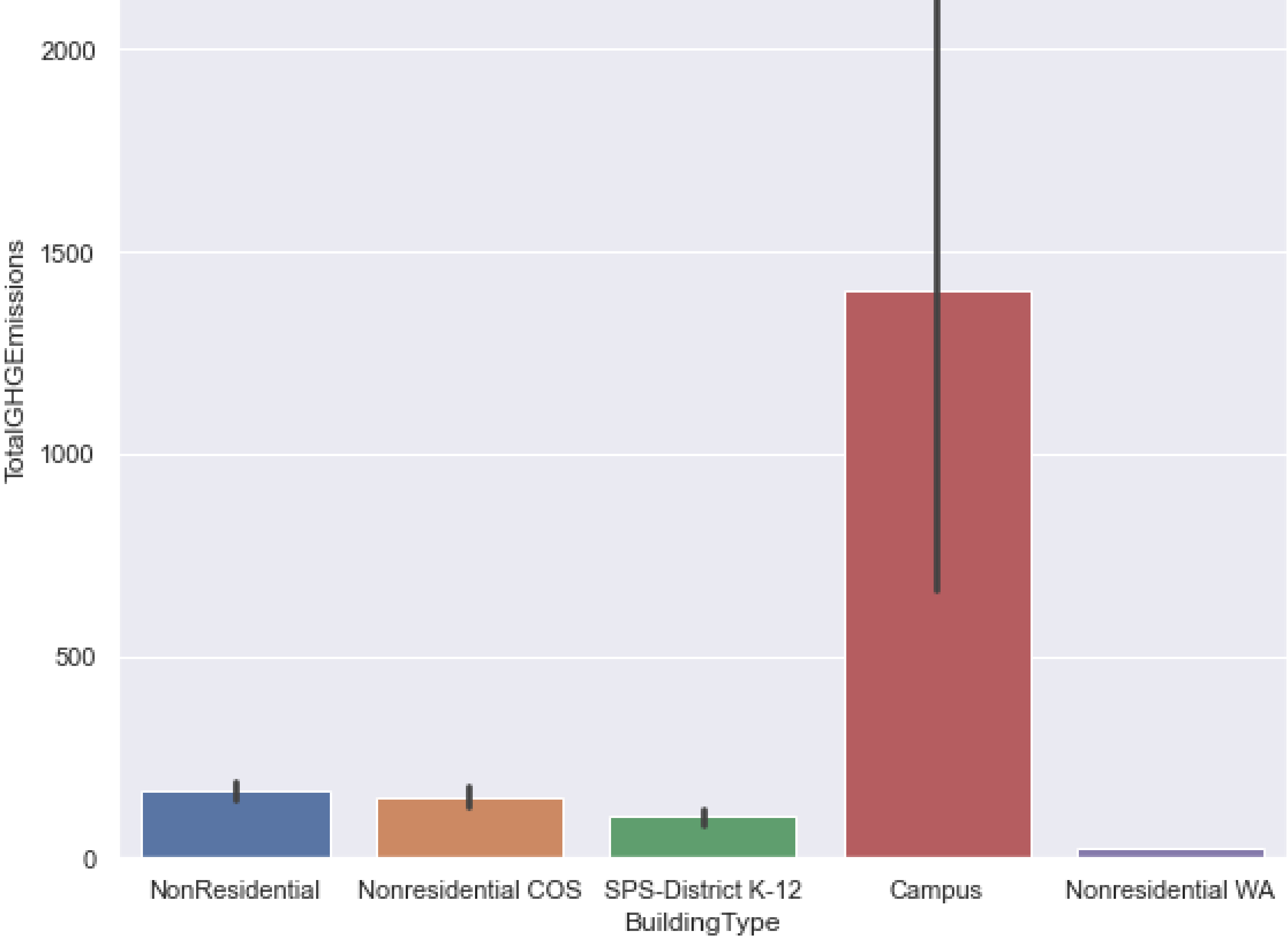
ANALYSE BIVARIEE - REPARTITION CONSOMMATION ENERGIE VS EMISSION CO2



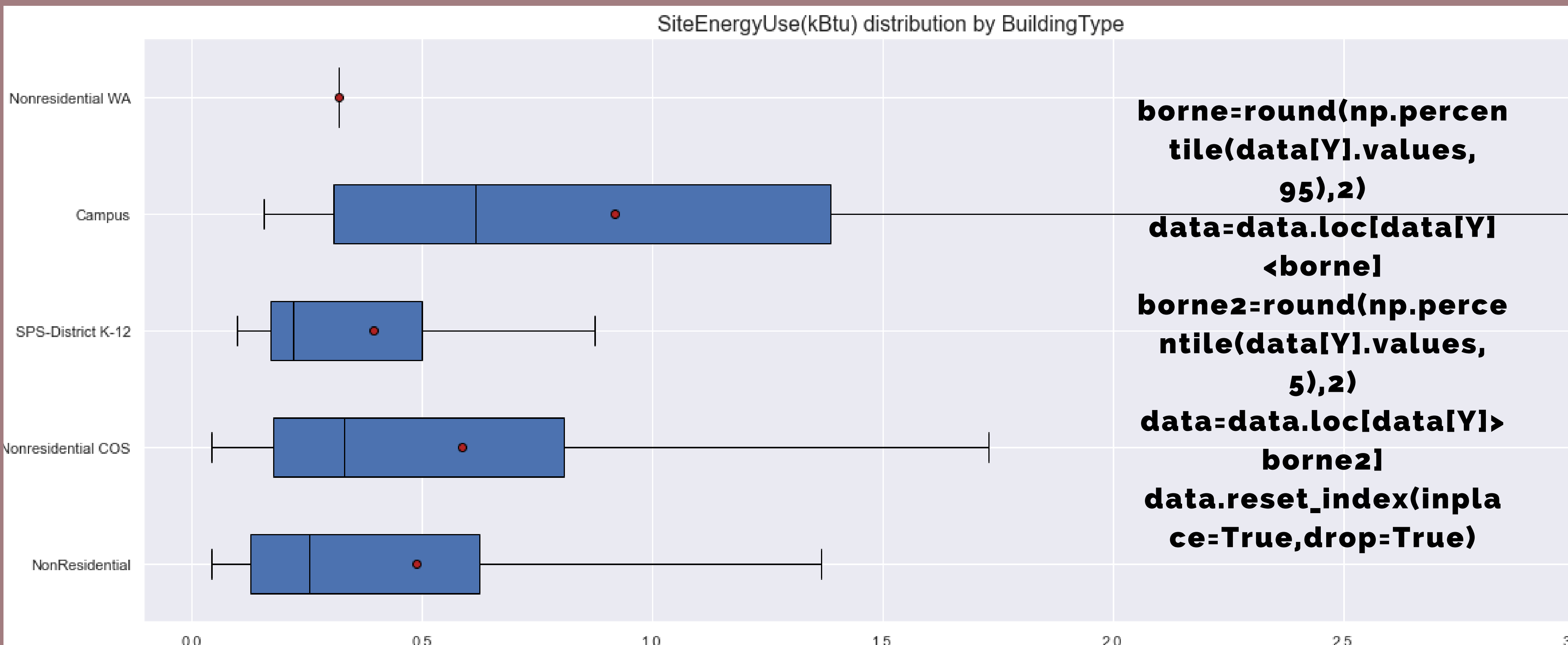
ANALYSE BIVARIEE - TYPE BUILDING / CONSOMMATION ENERGIE



ANALYSE BIVARIEE - TYPE BUILDING /EMISSION CO_2

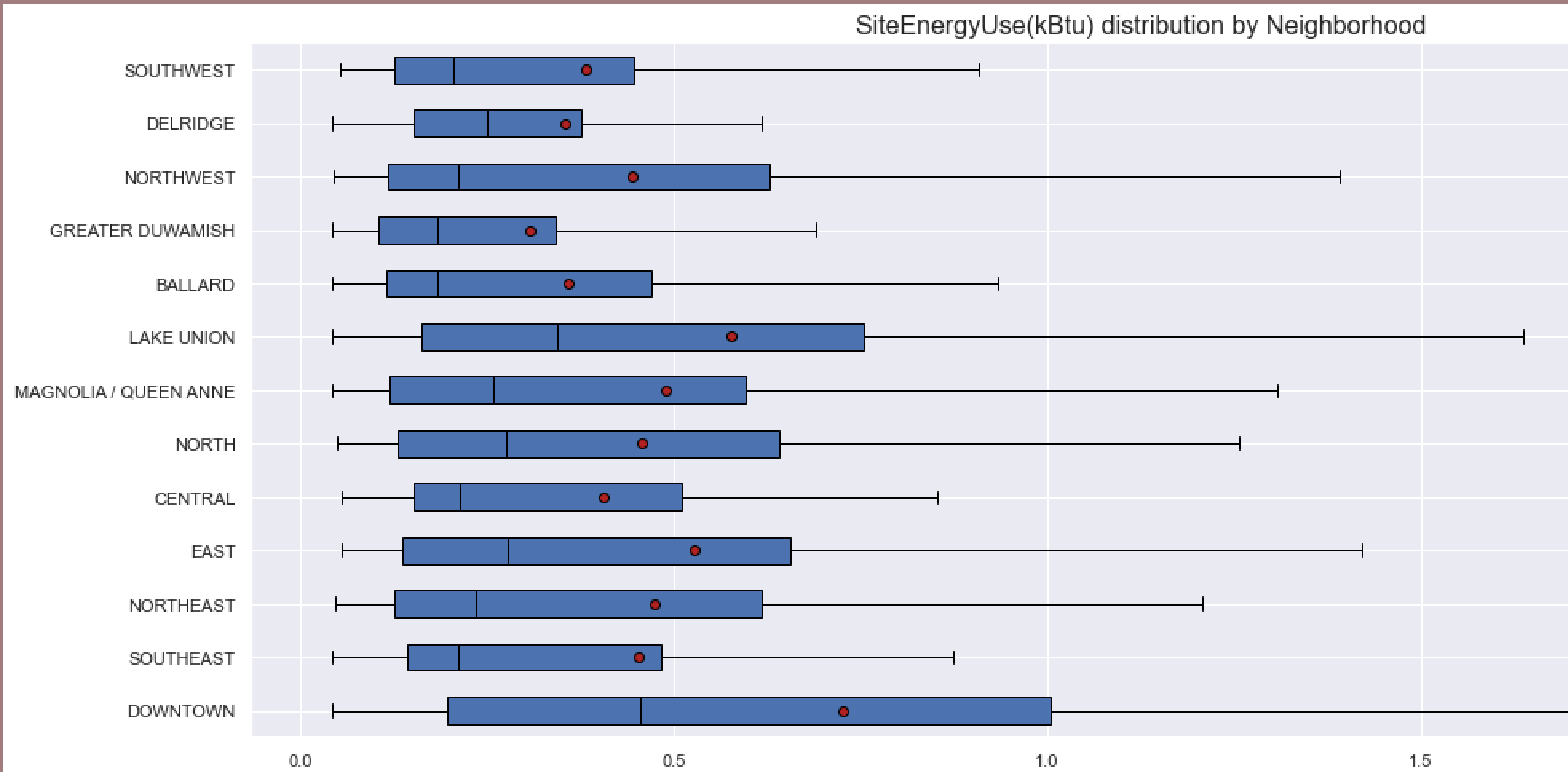


BOXPLOT - TYPE BUILDING /CONSOMMATION ENERGIE

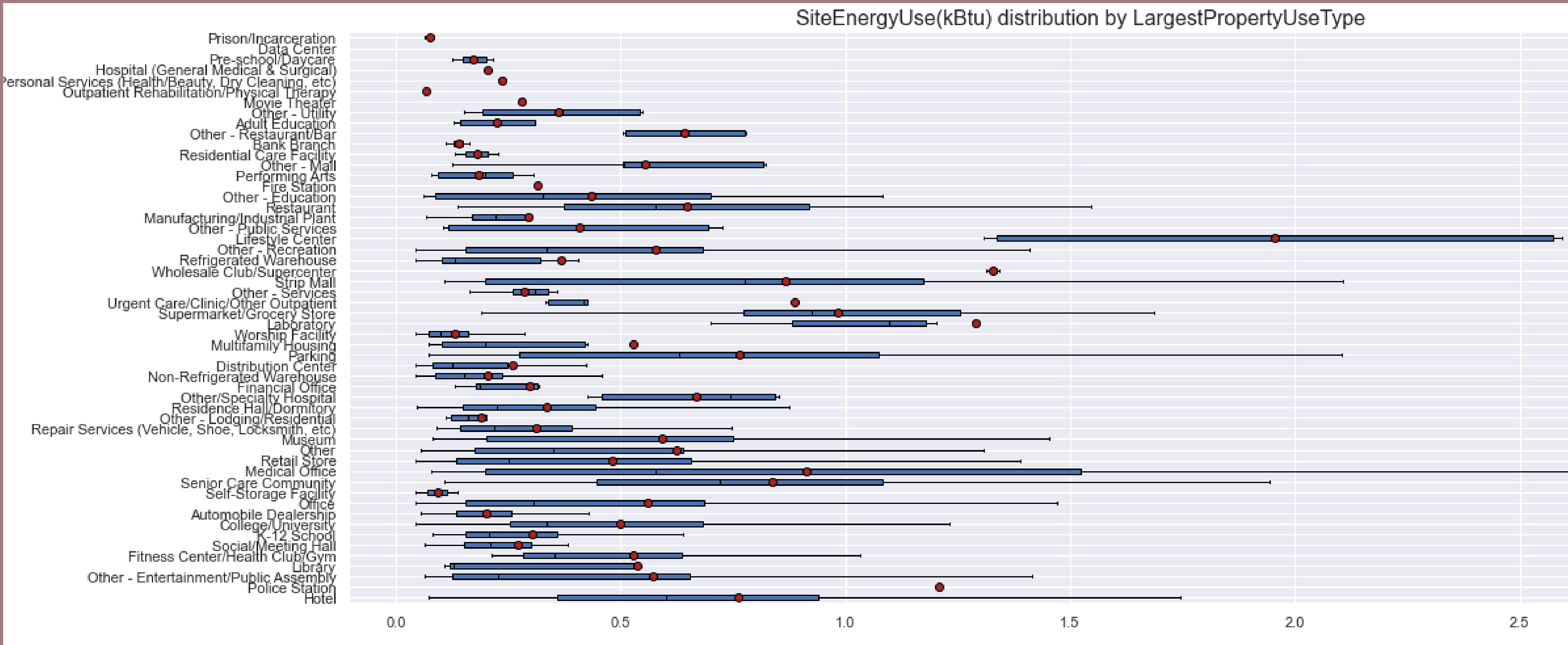



```
borne=round(np.percentile(data[Y].values,95),2)
data=data.loc[data[Y]<borne]
borne2=round(np.percentile(data[Y].values,5),2)
data=data.loc[data[Y]>borne2]
data.reset_index(inplace=True,drop=True)
```

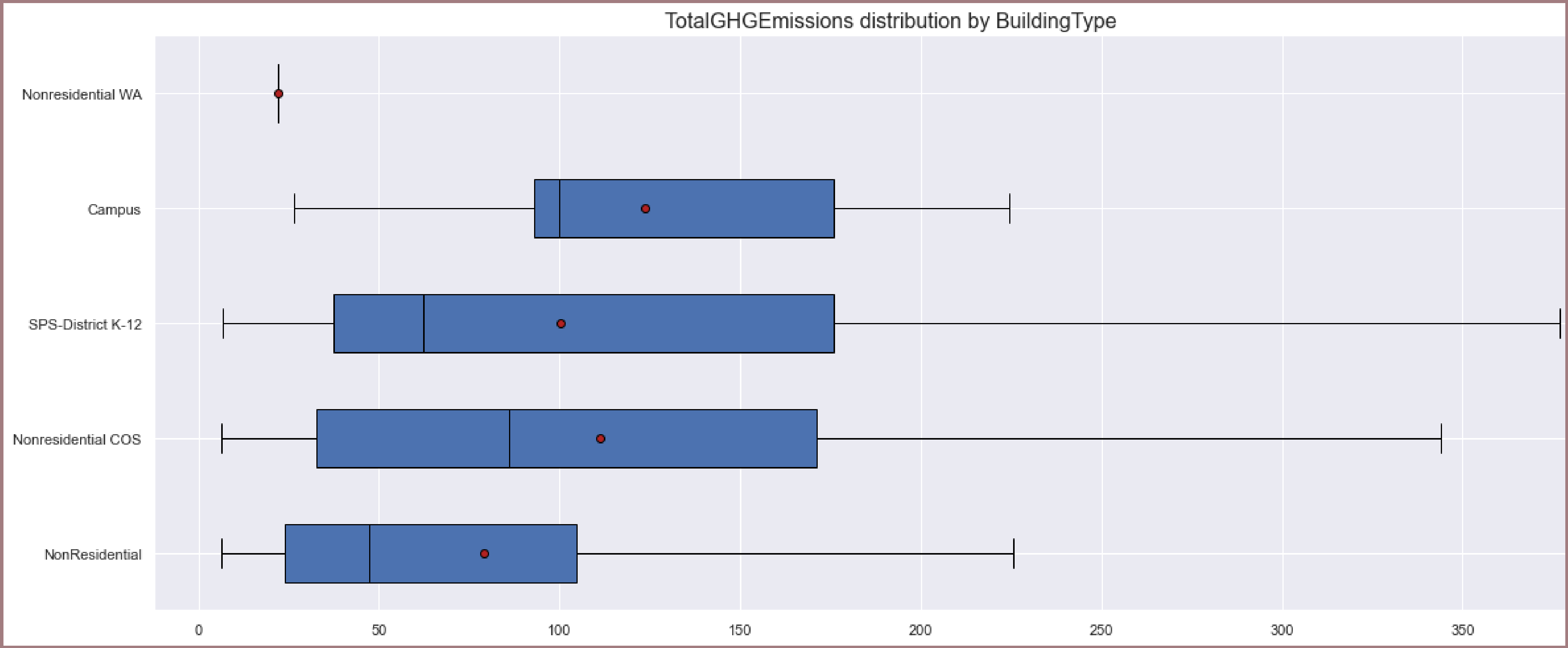
BOXPLOT - neighborhood/CONSOMMATION ENERGIE



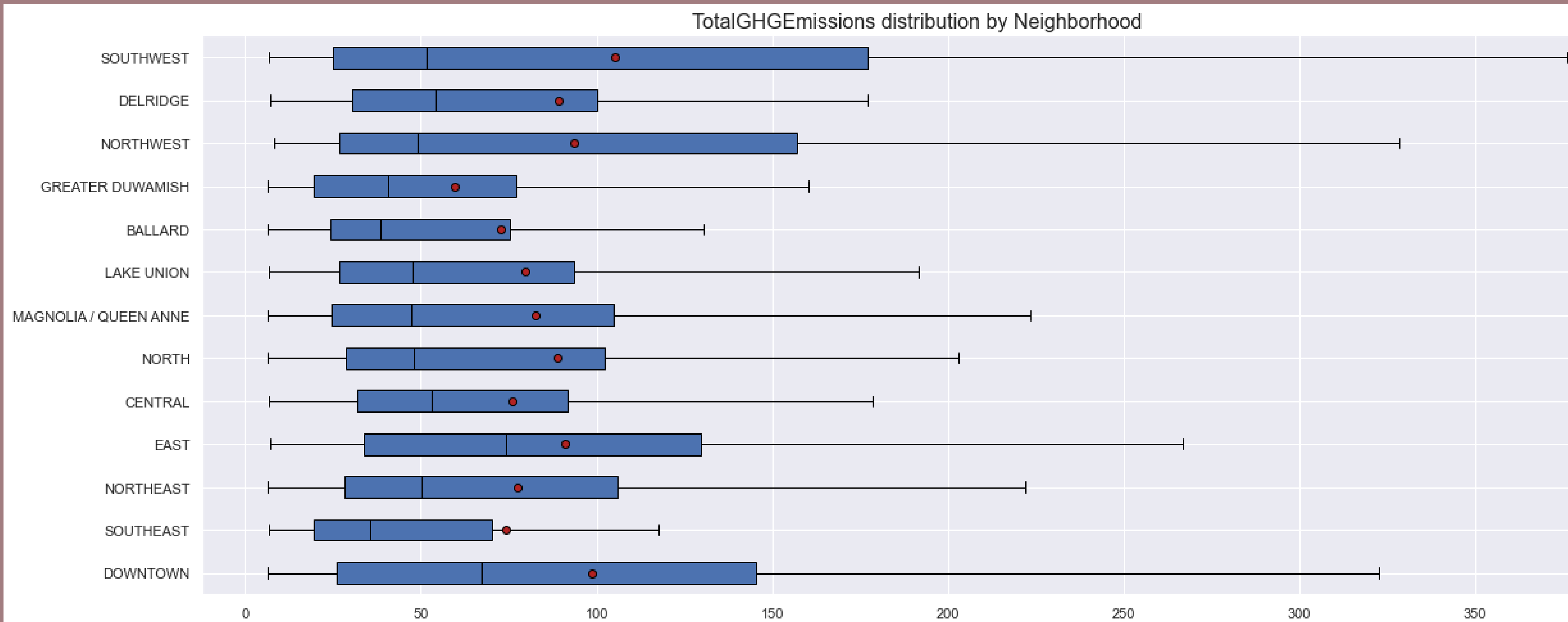
BOXPLOT - largestPropertyUseType/CONSOMMATION ENERGIE



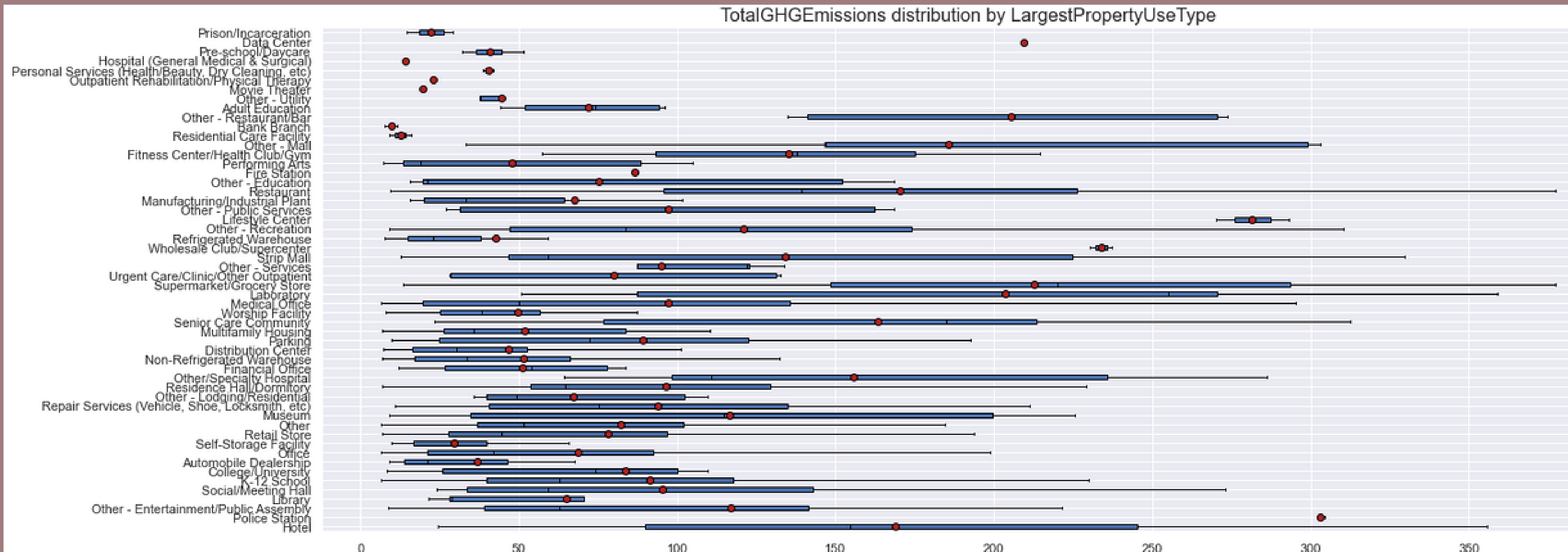
BOXPLOT - TYPE BUILDING /EMISSIONS CO2



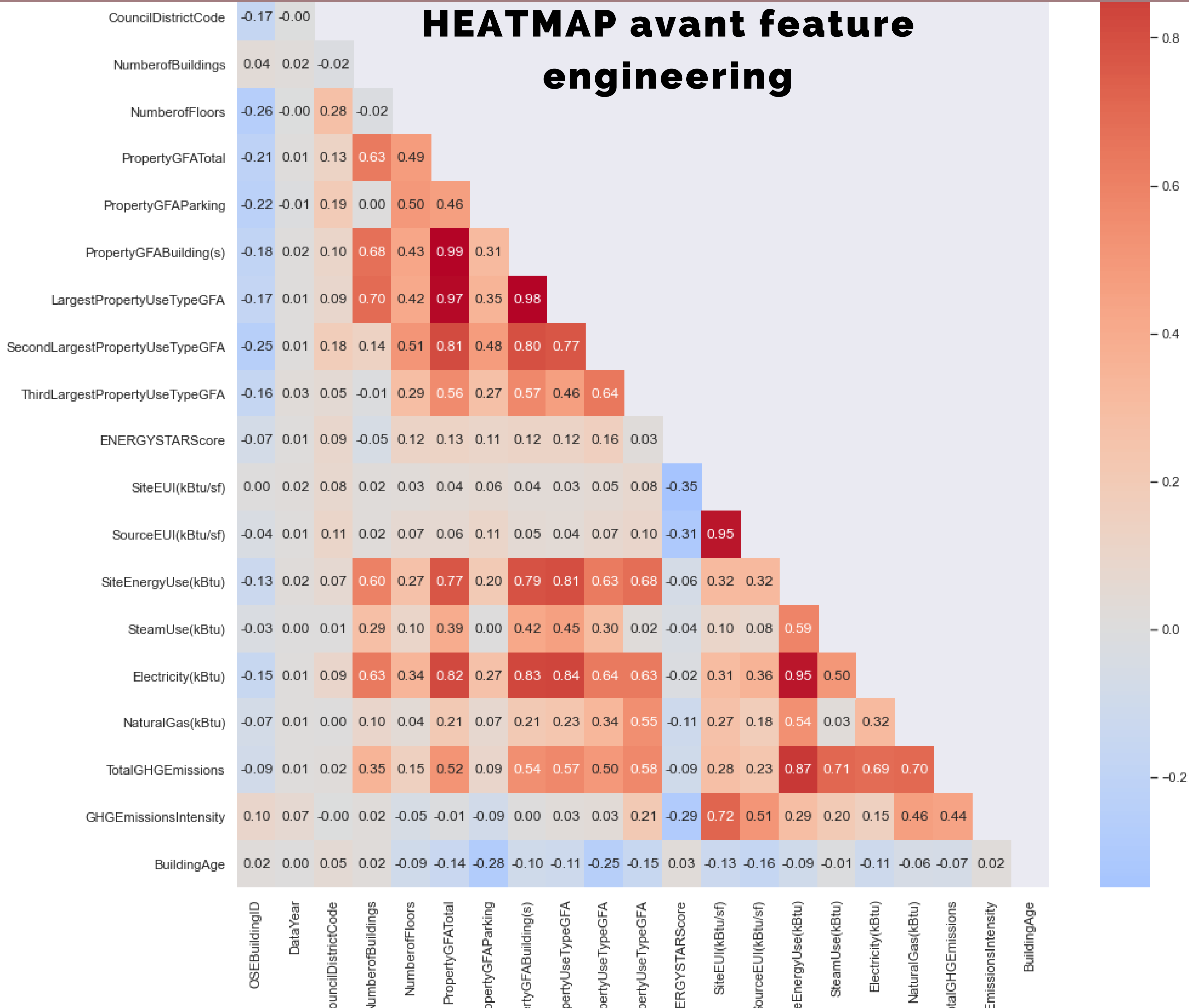
BOXPLOT - neighborhood/emissions co2



BOXPLOT - largestPropertyUseType/CONSOMMATION ENERGIE



HEATMAP avant feature engineering



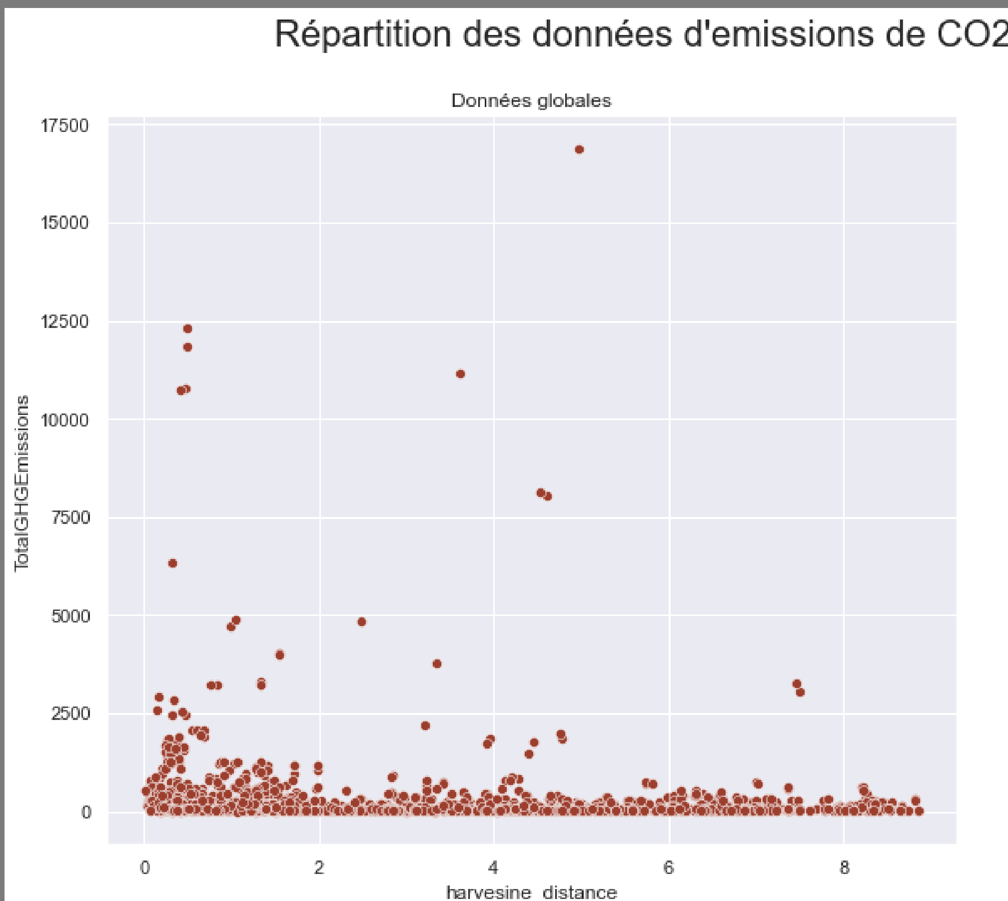
FEATURE ENGINEERING

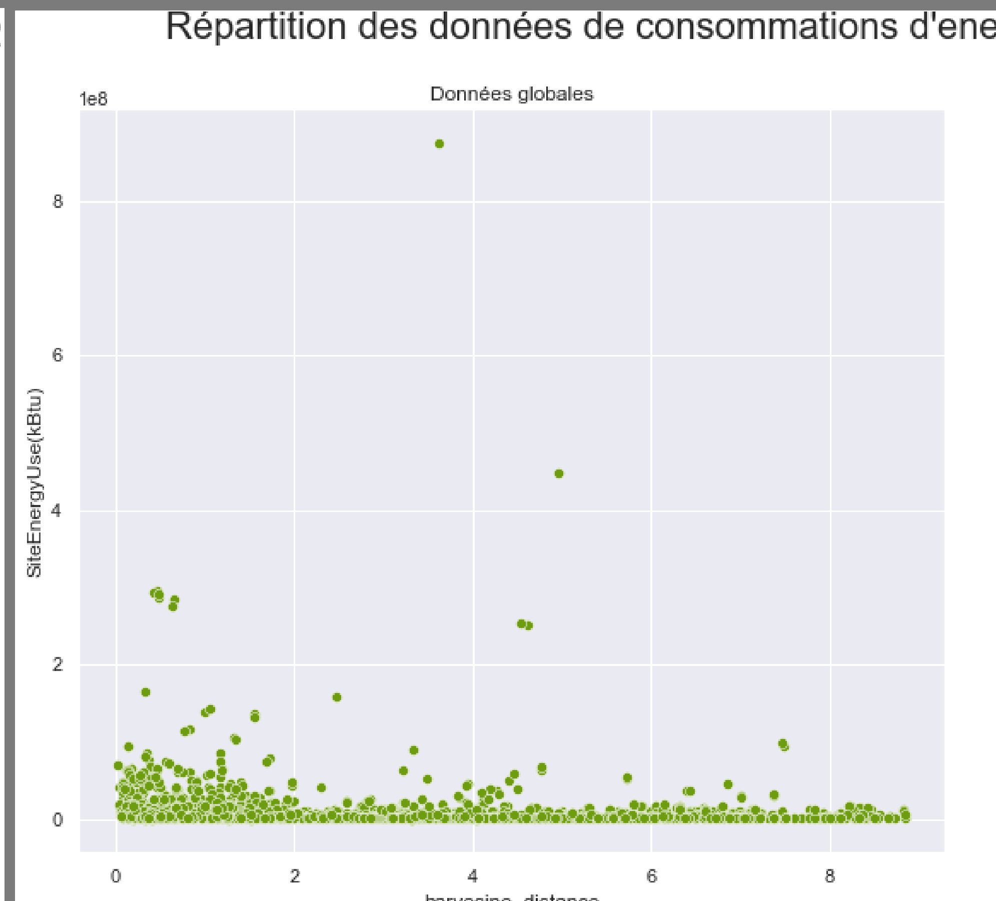
Conversion des différentes surfaces (Buildings et Parking) en pourcentage de la surface totale (GFABuildingRate, GFAParkingRate)
Calcul de la surface moyenne par bâtiment et par étage (GFAPerBuilding, GFAPerFloor)

	TotalUseTypeNumber	PropertyGFATotal	PropertyGFAParking	PropertyGFABuilding(s)	LargestPropertyUseTypeGFA
0	1	88434	0	88434	88434.0
1	3	103566	15064	88502	83880.0
2	3	961990	0	961990	757243.0
4	3	119890	12460	107430	123445.0
5	1	97288	37198	60090	88830.0
6	1	83008	0	83008	81352.0
7	1	102761	0	102761	102761.0
8	1	163984	0	163984	163984.0
10	1	153163	19279	133884	NaN
11	1	333176	61161	272015	336640.0

FEATURE ENGINEERING

Coordonnées géographiques (harvesine_distance) en fonction de la Latitude et Longitude et corrélation avec les targets





FEATURE ENGINEERING

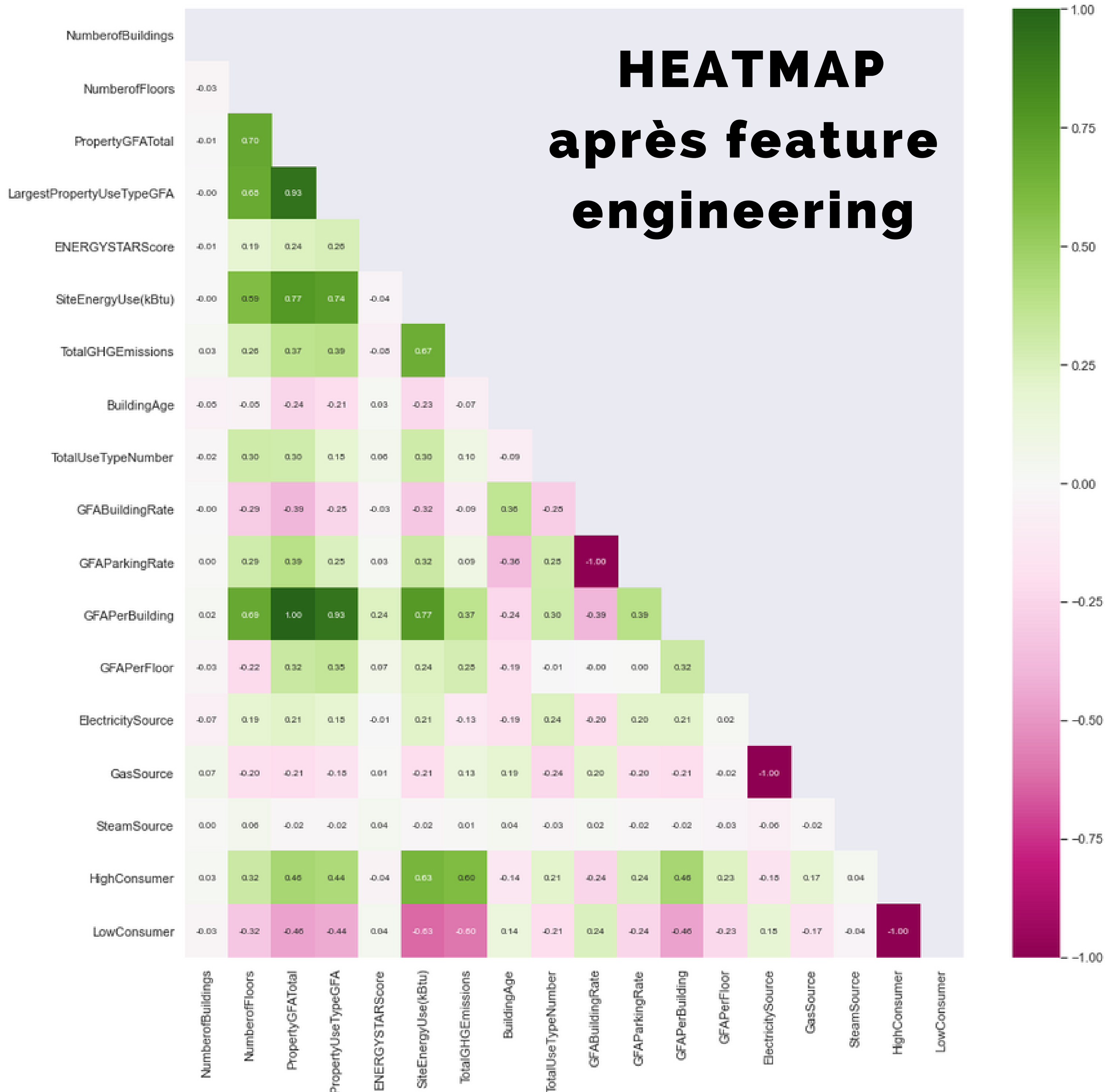
supprimer 'LargestPropertyUseType' redondant avec 'PrimaryPropertyType'

'BuildingType' est de type NonResidential - on le retire pour la modélisation

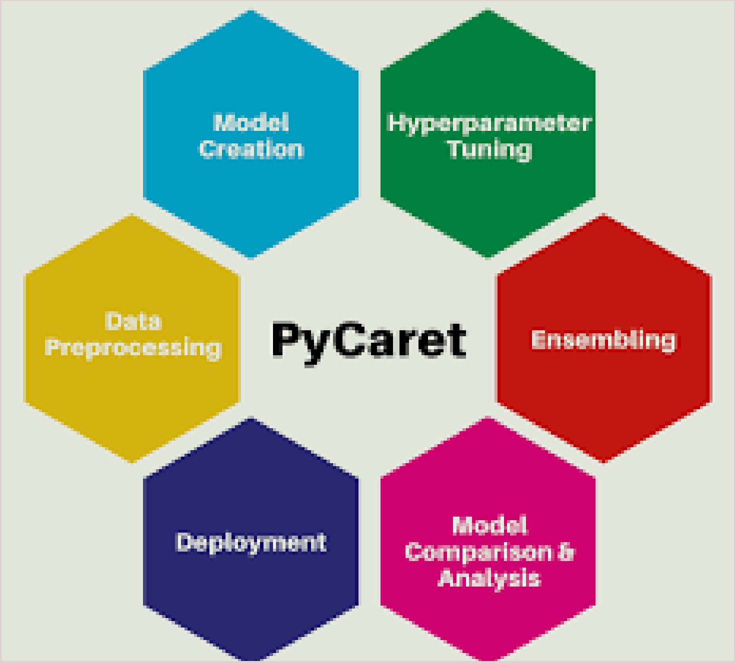
encoder les types object
['PrimaryPropertyType', 'CouncilDistrictCode', 'Neighborhood'

	PrimaryPropertyType__Other	PrimaryPropertyType__Mixed Use Property	PrimaryPropertyType__K-12 School	PrimaryPropertyType__College/University	PrimaryPropertyTy and Mid-!
0	1	0	0	0	
1	1	0	0	0	
2	0	1	0	0	
3	1	0	0	0	
4	0	1	0	0	

HEATMAP après feature engineering



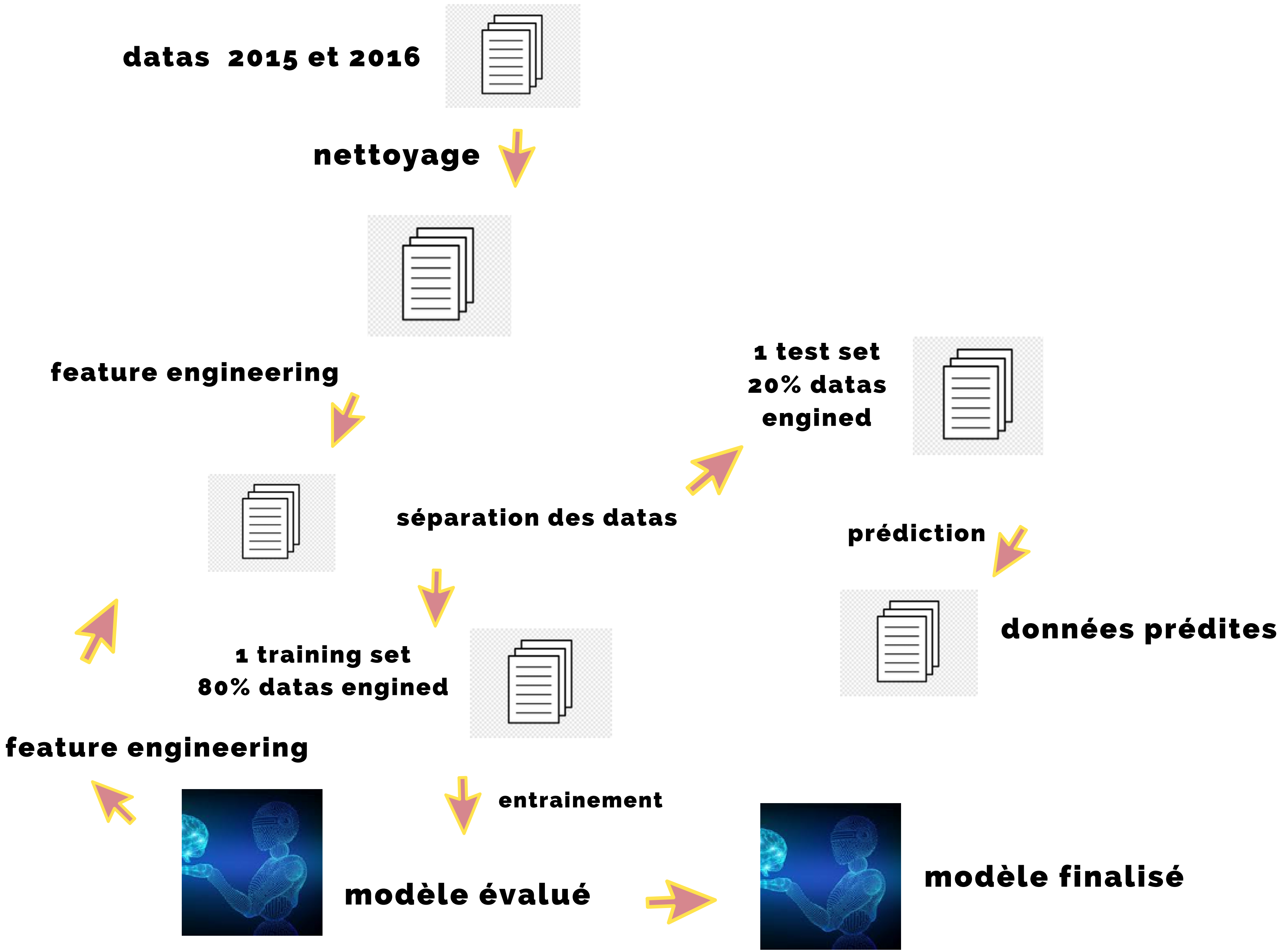
MODELISATION MACHINE LEARNING



MODELISATION MACHINE LEARNING
LE DATAFRAME

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1696 entries, 0 to 2500
Data columns (total 19 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   PrimaryPropertyType        1696 non-null   object
 1   CouncilDistrictCode        1696 non-null   object
 2   Neighborhood               1696 non-null   object
 3   NumberofBuildings          1696 non-null   float64
 4   NumberofFloors             1696 non-null   float64
 5   PropertyGFATotal           1696 non-null   int64
 6   LargestPropertyUseTypeGFA  1696 non-null   float64
 7   ENERGYSTARScore            1696 non-null   float64
 8   SiteEnergyUse(kBtu)        1696 non-null   float64
 9   TotalGHGEmissions          1696 non-null   float64
 10  BuildingAge                1696 non-null   int64
 11  TotalUseTypeNumber         1696 non-null   int64
 12  GFABuildingRate            1696 non-null   float64
 13  GFAParkingRate             1696 non-null   float64
 14  GFAPerBuilding             1696 non-null   float64
 15  GFAPerFloor                1696 non-null   float64
 16  ElectricitySource          1696 non-null   int64
 17  GasSource                  1696 non-null   int64
 18  SteamSource                1696 non-null   int64
dtypes: float64(10), int64(6), object(3)
memory usage: 265.0+ KB
```

MODELISATION MACHINE LEARNING PROCESS



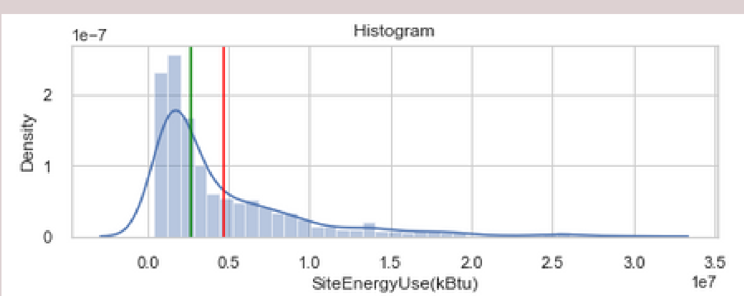
Prediction target 'SiteEnergyUse(kBtu)'

Echantillon Site_Energy_Use

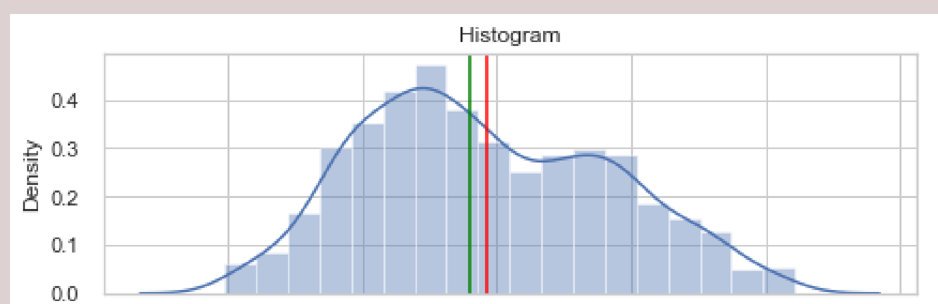
	Model
et	Extra Trees Regressor

	Sans passage au log		avec passage au log	
	R2	RMSLE	R2	RMSLE
Cross validation 10 folds	0.8881	0.3020	0.8883	0.0190
modele tuné (avec optimisation des hyperparamètres)	0.8433	0.4099	0.7738	0.0273
prédictions unseen data	0.7794	0.4650	0.7564	0.0296

Skewness of the SiteEnergyUse(kBtu) is 2.1100300655712854



Skewness of the SiteEnergyUse(kBtu) is 0.2817765168279764



hyperparamètres

sans passage au Log et avec passage au Log

bootstrap	False
ccp_alpha	0.0
criterion	mse
max_depth	10
max_features	1.0
max_leaf_nodes	None
max_samples	None
min_impurity_decrease	0.4
min_impurity_split	None
min_samples_leaf	3
min_samples_split	10
min_weight_fraction_leaf	0.0
n_estimators	70
n_jobs	-1
oob_score	False
random_state	3213
verbose	0
warm_start	False

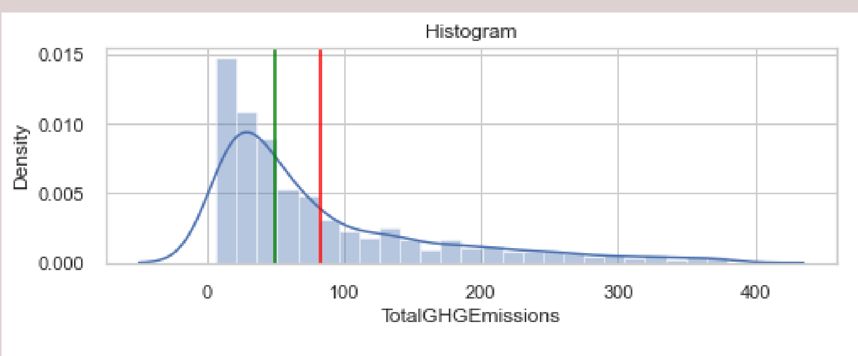
	Parameters
bootstrap	False
ccp_alpha	0.0
criterion	mse
max_depth	7
max_features	1.0
max_leaf_nodes	None
max_samples	None
min_impurity_decrease	0.001
min_impurity_split	None
min_samples_leaf	5
min_samples_split	5
min_weight_fraction_leaf	0.0
n_estimators	80
n_jobs	-1
oob_score	False
random_state	6929
verbose	0
warm_start	False

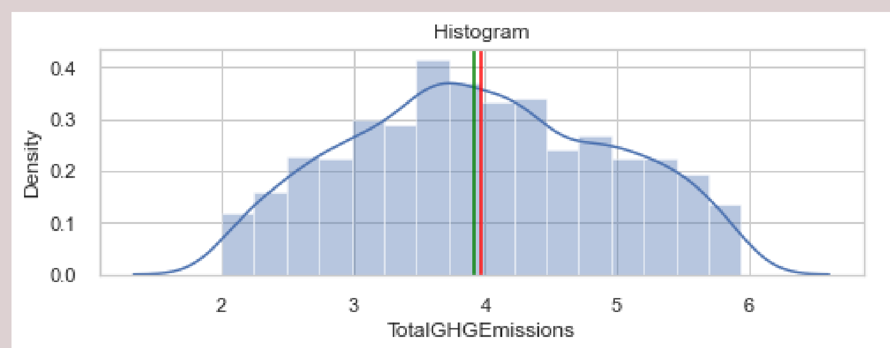
Prediction target TotalGHGEmissions (avec ENERGYSTARScore)

Echantillon

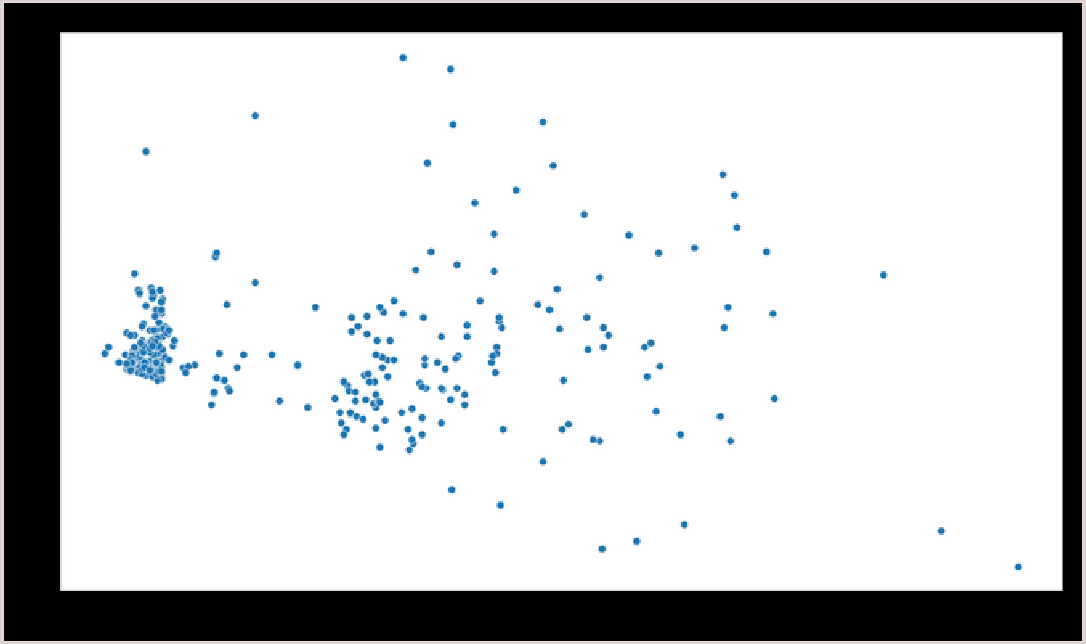
	Model
et	Extra Trees Regressor

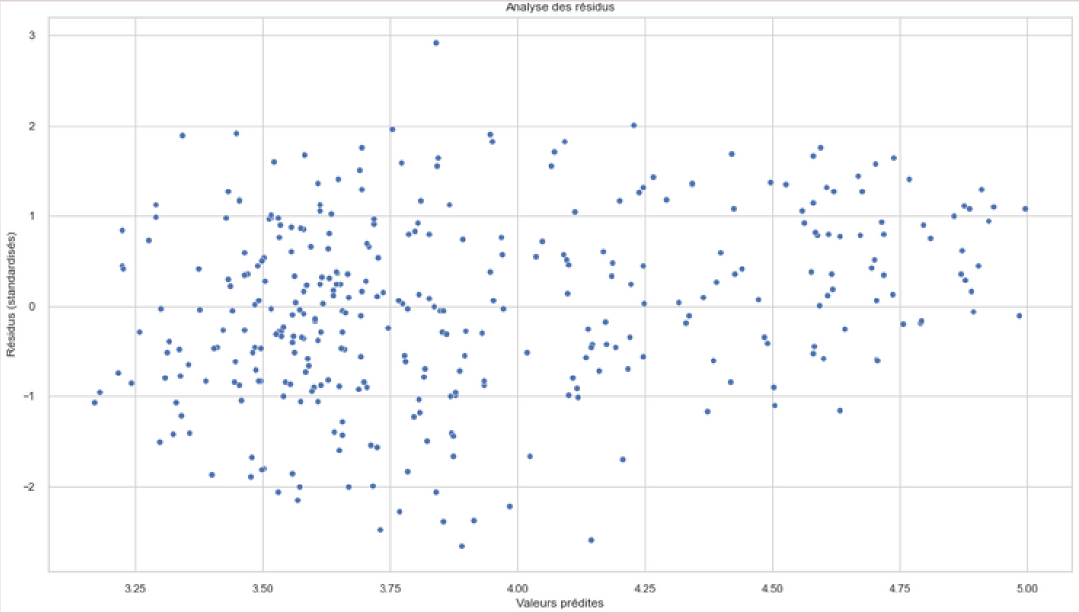
	Sans passage au log		avec passage au log	
	R2	RMSLE	R2	RMSLE
Cross validation 10 folds	0.7023	0.5107	0.7415	0.1067
modele tuné (avec optimisation des hyperparamètres)	0.6123	0.6437	0.4076	0.1609
prédictions unseen data	0.6154	0.6654	0.4319	0.1515



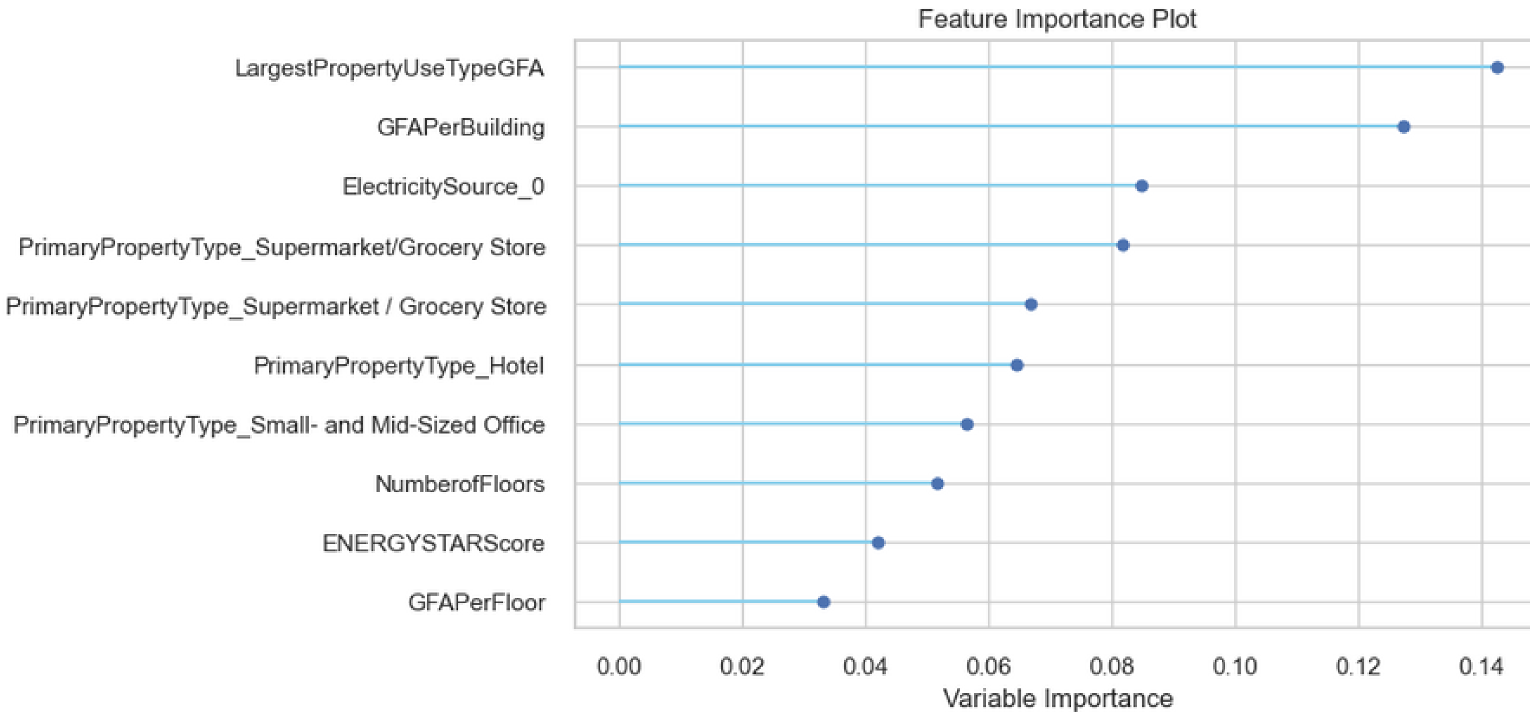


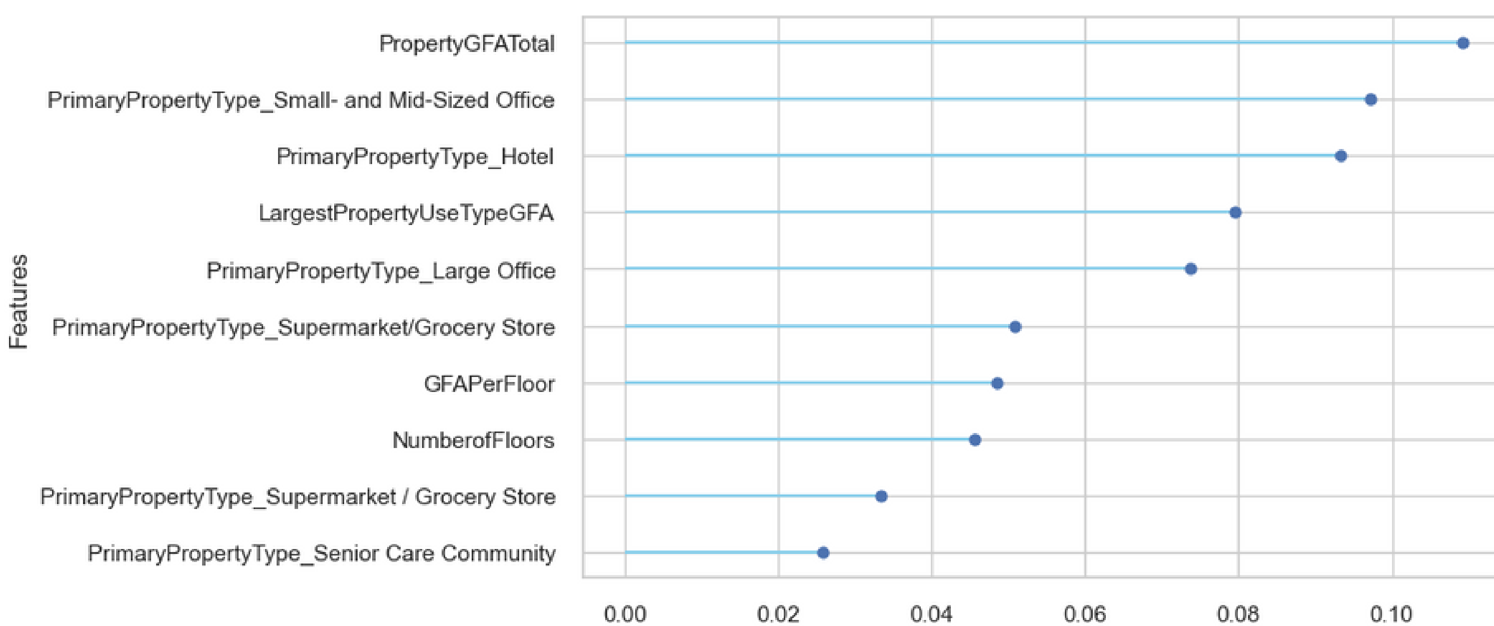
Analyse des résidus
sans passage au Log et avec passage au Log





Feature importance sans passage au Log et avec passage au Log





hyperparamètres sans passage au Log et avec passage au Log

	Parameters
bootstrap	True
ccp_alpha	0.0
criterion	mse
max_depth	11
max_features	1.0
max_leaf_nodes	None
max_samples	None
min_impurity_decrease	0.0005
min_impurity_split	None
min_samples_leaf	2
min_samples_split	9
min_weight_fraction_leaf	0.0
n_estimators	290
n_jobs	-1
oob_score	False
random_state	7001
verbose	0
warm_start	False

	Parameters
bootstrap	False
ccp_alpha	0.0
criterion	mae
max_depth	11
max_features	log2
max_leaf_nodes	None
max_samples	None
min_impurity_decrease	0
min_impurity_split	None
min_samples_leaf	3
min_samples_split	2
min_weight_fraction_leaf	0.0
n_estimators	200
n_jobs	-1
oob_score	False
random_state	2167
verbose	0
warm_start	False

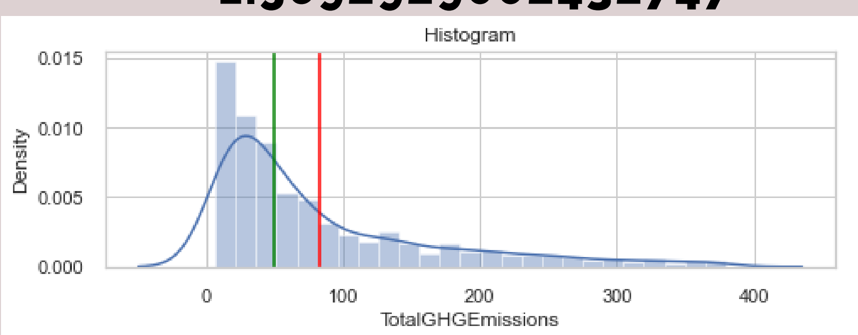
Prediction target TotalGHGEmissions (sans ENERGYSTARScore)

Echantillon

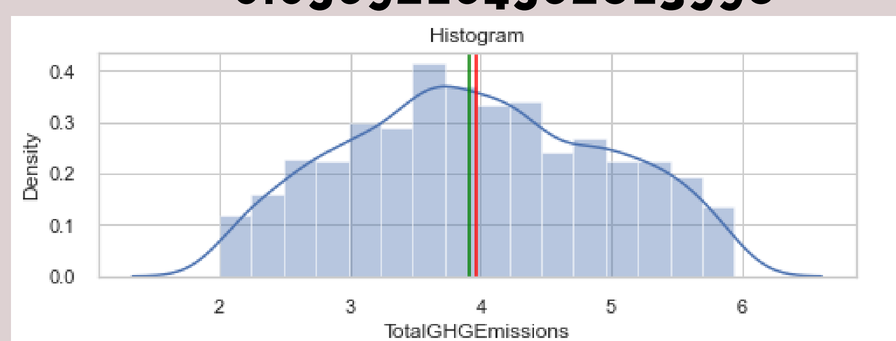
	Model
et	Extra Trees Regressor

	Sans passage au log		avec passage au log	
	R2	RMSLE	R2	RMSLE
Cross validation 10 folds	0.6508	0.5198	0.6320	0.1242
modele tuné (avec optimisation des hyperparamètres)	0.5303	0.7036	0.3962	0.1602
prédictions unseen data	0.6317	0.6598	0.4105	0.1663

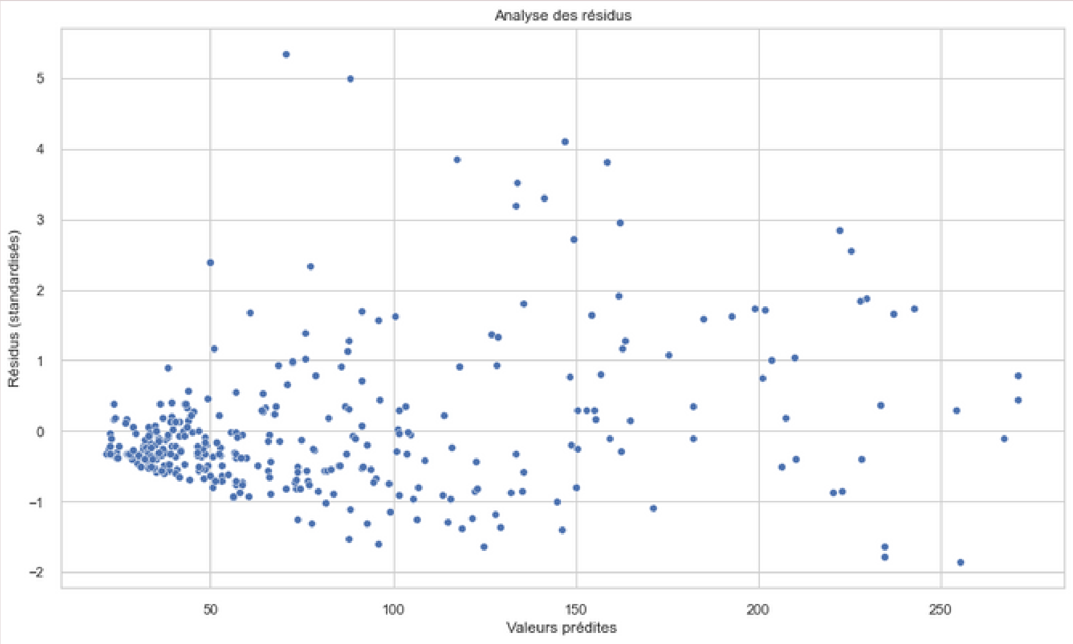
Skewness of TotalGHGEmissions is 1.5692929002432747

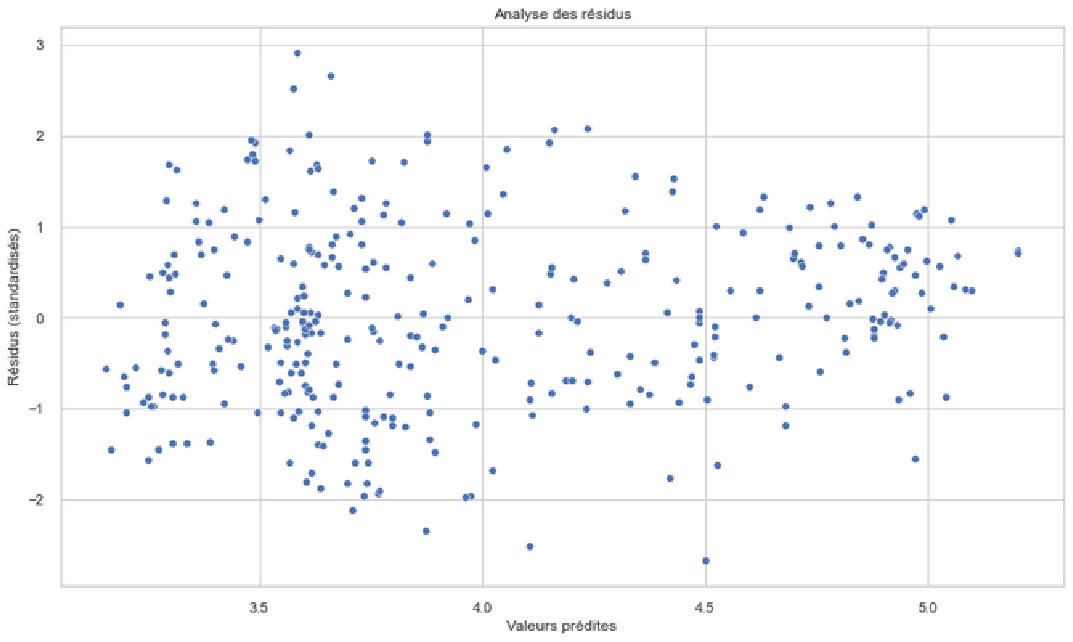


Skewness of TotalGHGEmissions is 0.05692164562813998

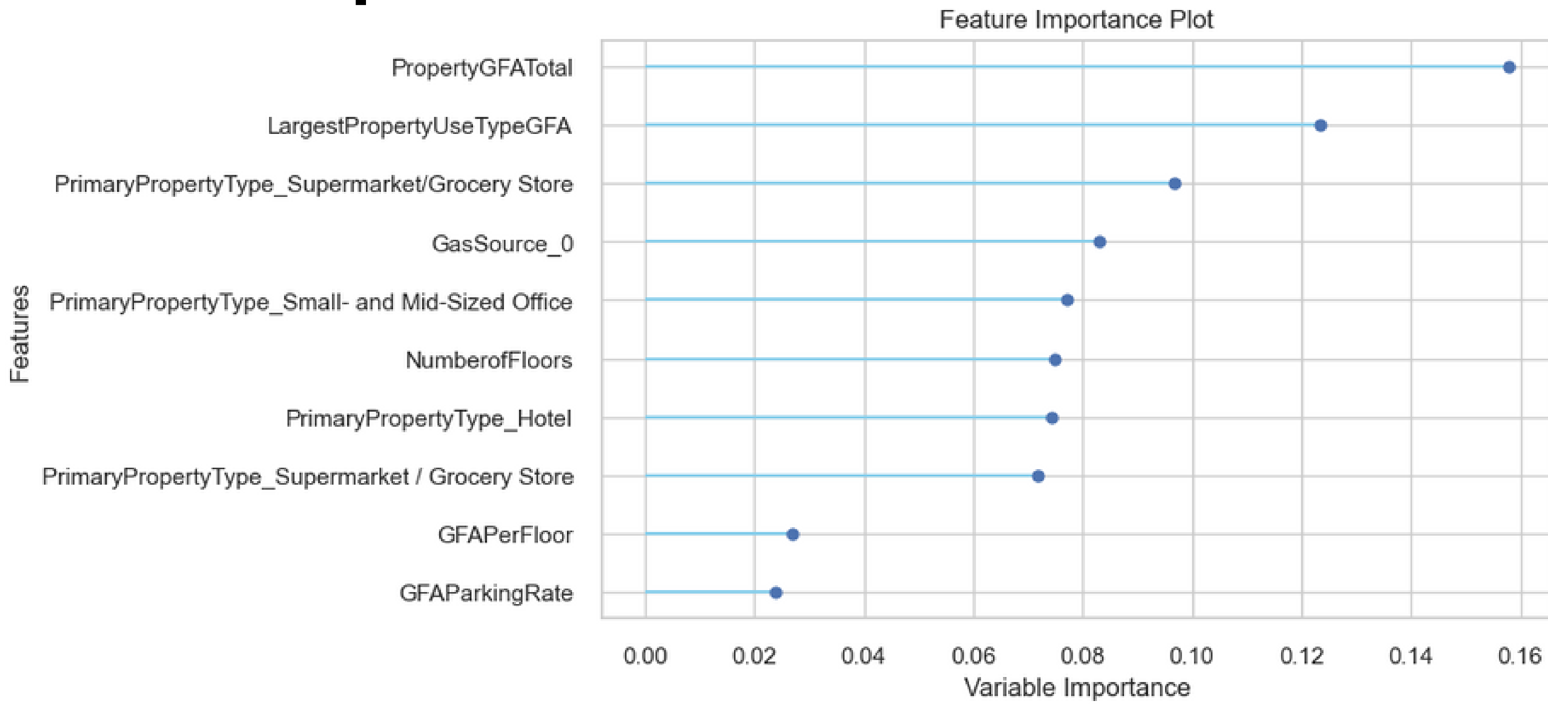


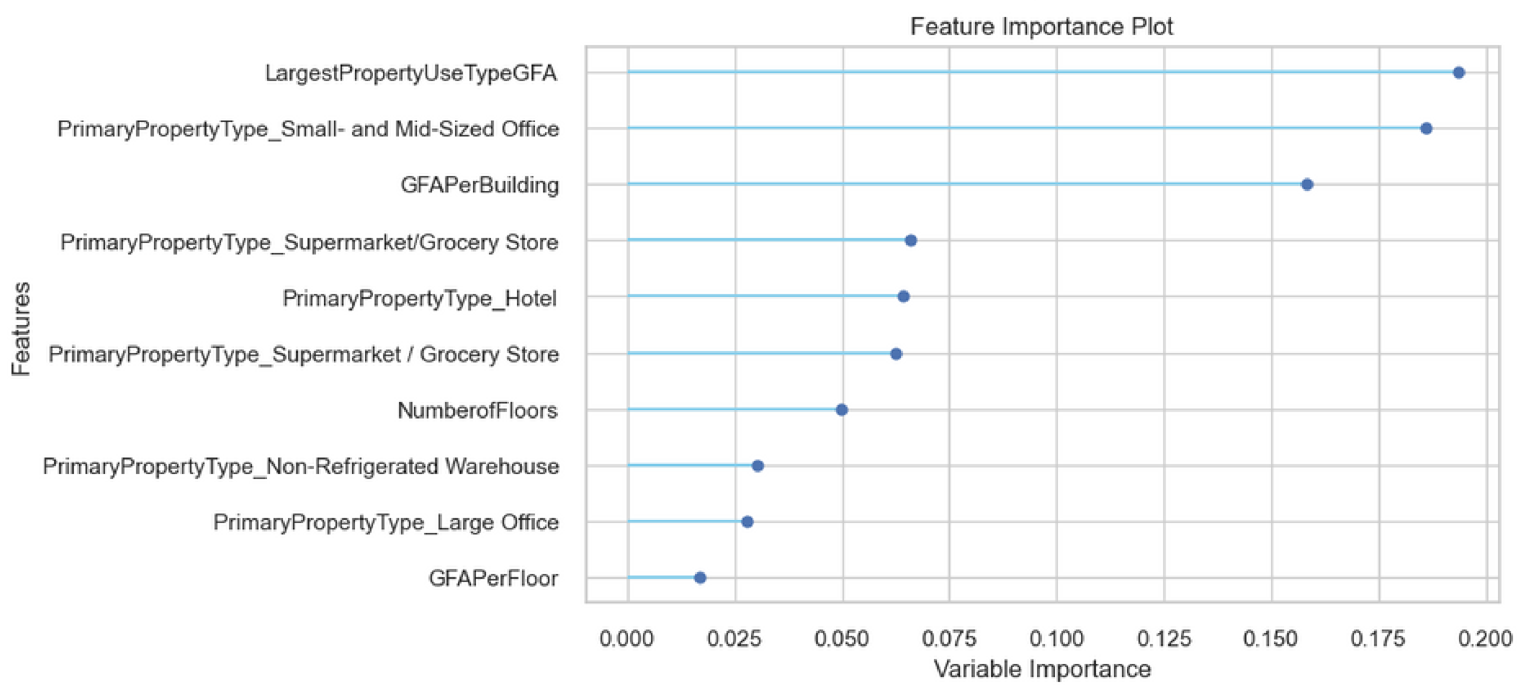
Analyse des résidus sans passage au Log et avec passage au Log





Feature importance sans passage au Log et avec passage au Log





hyperparamètres

sans passage au Log et avec passage au Log

	Parameters
bootstrap	False
ccp_alpha	0.0
criterion	mse
max_depth	8
max_features	1.0
max_leaf_nodes	None
max_samples	None
min_impurity_decrease	0.01
min_impurity_split	None
min_samples_leaf	3
min_samples_split	10
min_weight_fraction_leaf	0.0
n_estimators	280
n_jobs	-1
oob_score	False
random_state	3483
verbose	0
warm_start	False

	Parameters
bootstrap	True
ccp_alpha	0.0
criterion	mae
max_depth	5
max_features	1.0
max_leaf_nodes	None
max_samples	None
min_impurity_decrease	0.002
min_impurity_split	None
min_samples_leaf	2
min_samples_split	9
min_weight_fraction_leaf	0.0
n_estimators	250
n_jobs	-1
oob_score	False
random_state	7632
verbose	0
warm_start	False

CONCLUSION

Prédiction consommation énergie : modèle en surapprentissage qui n'a rien appris. Le passage au log diminue le RMSLE

Prédiction émissions de CO2 : ENERGYstarscore n'apporte pas une grande plus value. Les modèles avec et sans cette variable sont sensiblement les mêmes en terme de performances et erreurs.

Par rapport à ce jeu de données , les passages au log diminuent le RMSLE mais n'améliorent pas beaucoup les prédictions.
On pourrait revoir les features et les recombiner pour trouver des axes d'amélioration.