Projet 4 : Anticipez les besoins en consommation de bâtiments

Présenté par :
Bourama FANE
Etudiant Data Scientist

Dirigé par :
Babou M'BAYE
Mentor chez OpenClassrooms

14 Avril 2023

Sommaire

Plan de la présentation

Problématique

Vous travaillez pour la **ville de Seattle**. Pour atteindre son objectif de **ville neutre en émissions de carbone en 2050**, votre équipe s'intéresse de près à la consommation et aux émissions des bâtiments non destinés à l'habitation.
Vous cherchez 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.

Seattle

Sources de données & mission

Sources de données

Des relevés minutieux ont été effectués par les agents de la ville en 2016. Cependant, ces relevés sont coûteux à obtenir, et à partir de ceux déjà réalisés, vous voulez tenter de prédire les **émissions de CO2** et la **consommation totale d'énergie** de bâtiments non destinés à l'habitation pour lesquels elles n'ont pas encore été mesurées.

Mission

- Réaliser une courte analyse exploratoire ;
- Tester différents modèles de prédiction afin de répondre au mieux à la problématique ;
- Evaluer l'interêt de la variable « **ENERGY STAR Score** ».

Plan de la présentation

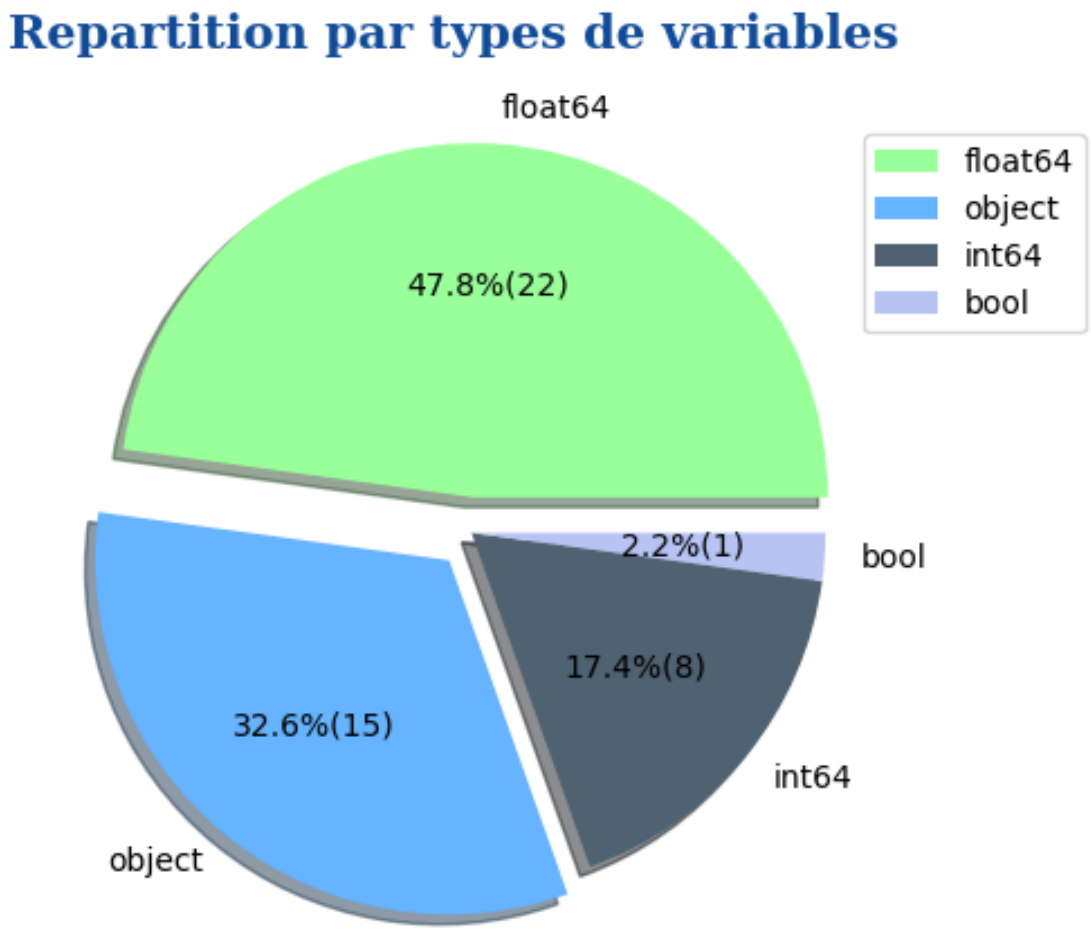
Description de la base

Les données sont organisées en 5 sections listées ci dessous :

- 1. Identification du bâtiments et localisation ;
- 2. Informations sur le type de bâtiments et style de construction et usages ;
- 3. Consommation energie ;
- 4. Emission de gaz a effet de serre ;
- 5. Autres variables diverses.

Description de la base

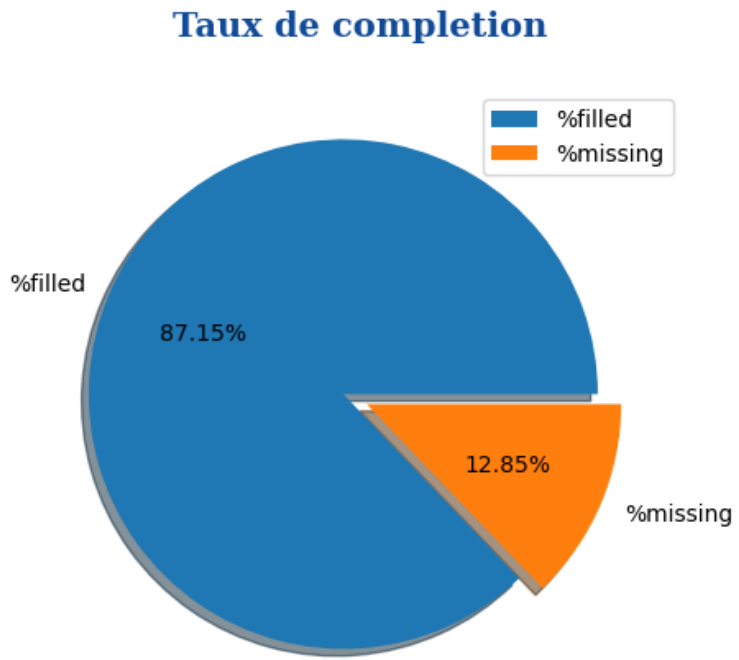
	Variable	nombre
0	lignes	3376
1	colonnes	46



- ☞ La base de données contient **3 376 lignes**, contre **46 variables** ;
- ☞ Nous distinguons quatre (4) types de variables (**object**,**float**, **int**, **bool**).

Description de la base

- Taux de missings de 12.85% .
- Certaines variables sont pratiquement vides.



Filtre sur les lignes

- Nous avons gardé uniquement les bâtiments qui ne sont pas à usage d'habitation (non residentiels).

```
NotHabitation=['NonResidential', 'Nonresidential COS','SPS-District K-12',
               'Campus', 'Nonresidential WA']
dfBuild=filtreModalite(dfBuild, 'BuildingType', NotHabitation)
dfBuild['BuildingType'].unique()
```

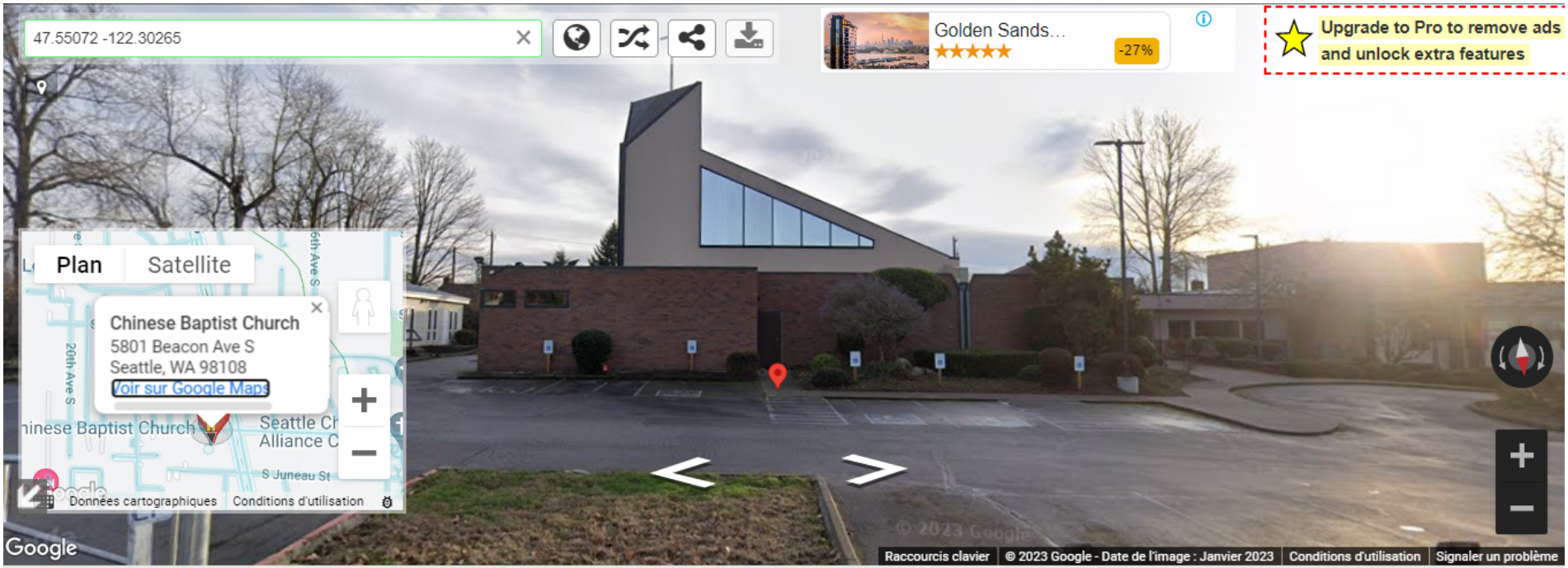
- Nous avons supprimé les lignes avec **TotalGHGEmissions<=0**
- Nous avons supprimé les lignes avec **SiteEnergyUse(kBtu)<=0**
- Nous avons restreint les données aux bâtiments pour lesquels la variable "**ENERGYSTARScore**" est renseignée ;

Filtre sur les colonnes

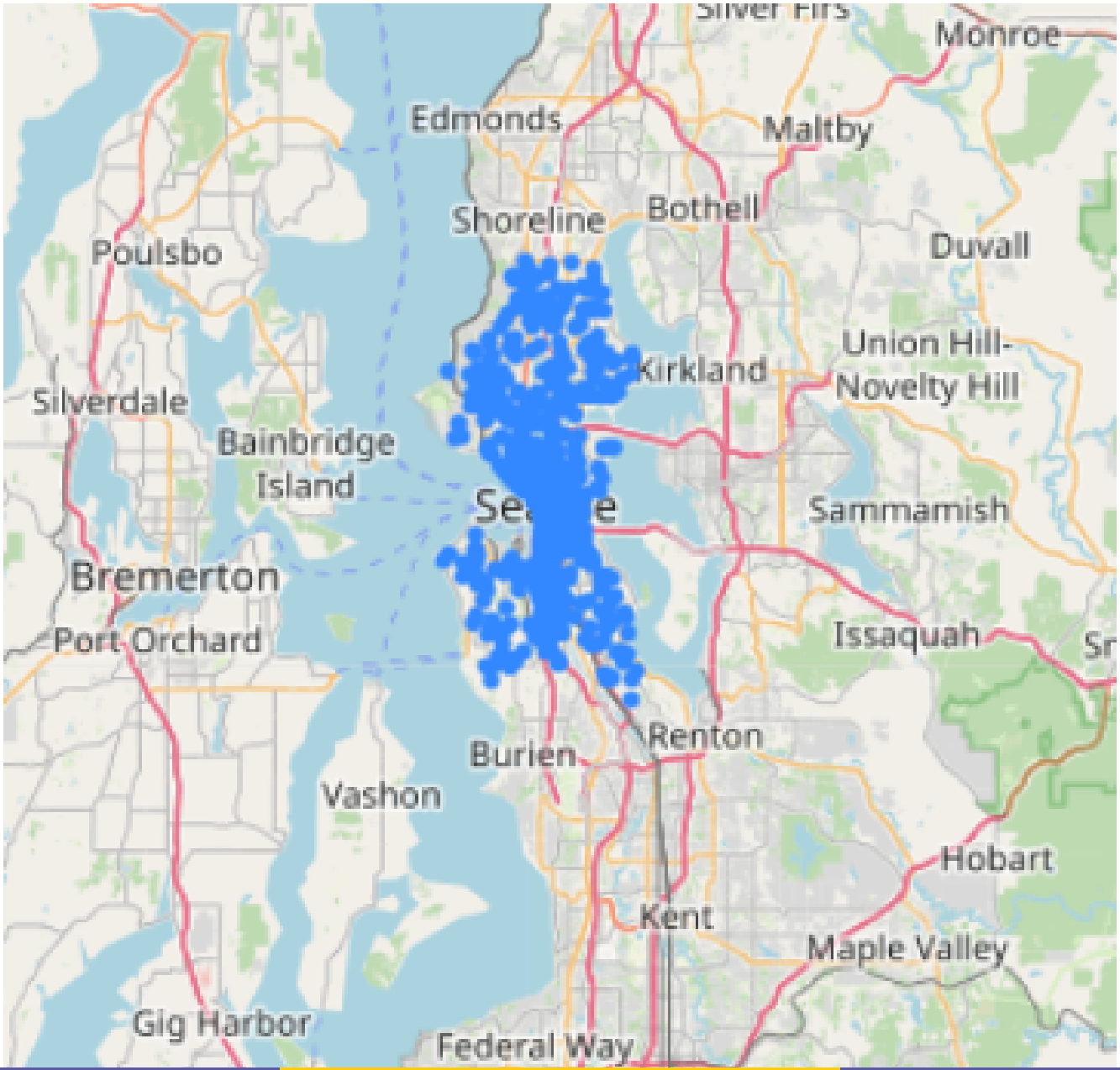
- ☞ Suppression des colonnes avec des « **WN** », des corrections liées au climat (*weather normalised*) : **['SiteEUIWN(kBtu/sf)', 'SourceEUIWN(kBtu/sf)', 'SiteEnergyUseWN(kBtu)']**
- ☞ Suppression des colonnes redondantes : **['NaturalGas(therms)','Electricity(kWh)']**
- ☞ Suppression des colonnes constantes ou difficiles à exploiter : **['State','Comments','ZipCode','City','DataYear','OSEBuildingID', 'TaxParcelIdentificationNumber']**
- ☞ Suppression des colonnes entièrement vides ;
- ☞ Suppression des colonnes ayant un taux de missing supérieur à 50% (choix arbitraire).

Consistance des données

	BuildingType	PrimaryPropertyType	PropertyName	Neighborhood	Latitude	Longitude	NumberofBuildings	NumberofFloors
1359	NonResidential	Worship Facility	Seattle Chinese Baptist Church	GREATER DUWAMISH	47.55072	-122.30265	1.0	99

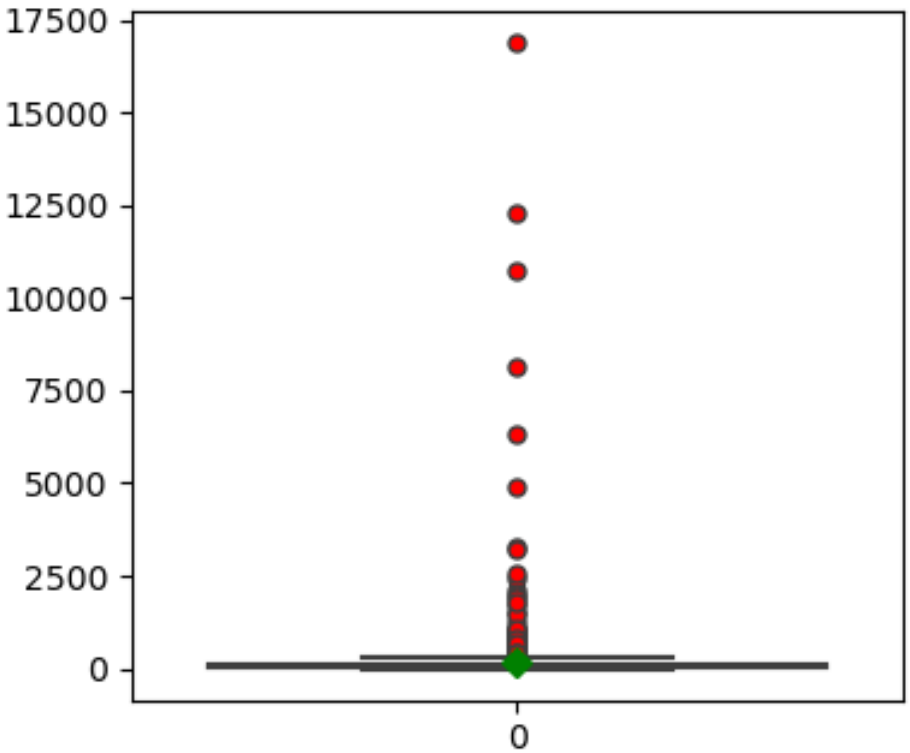


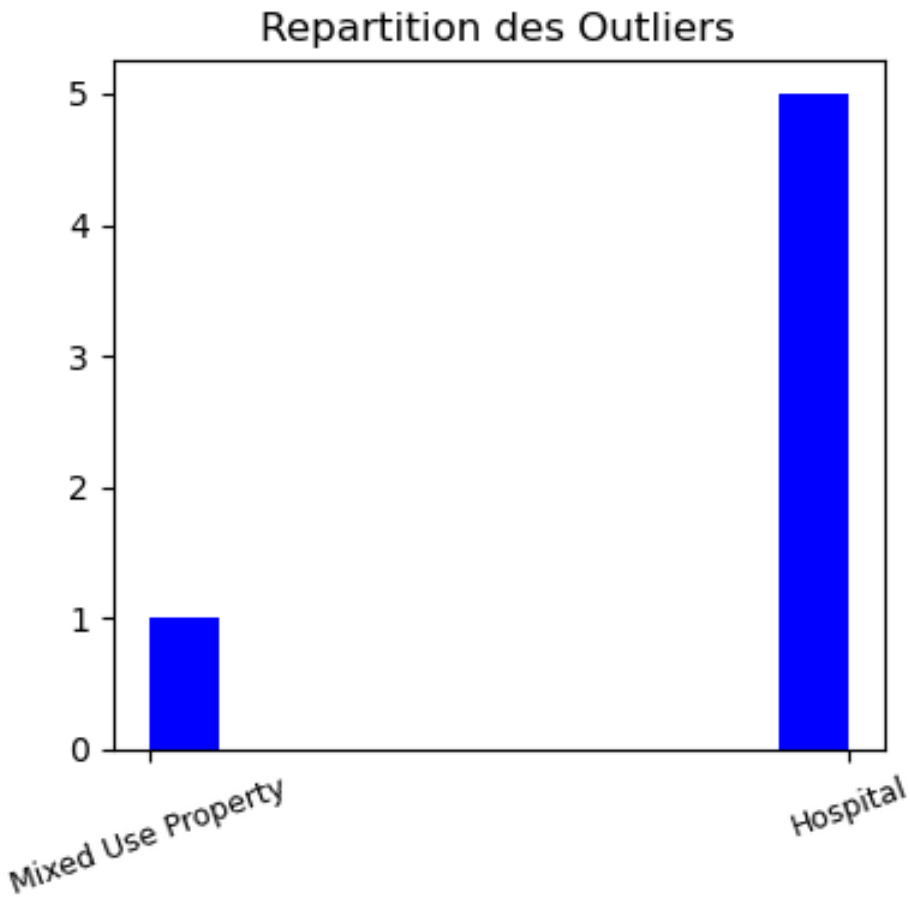
Consistance des données



Valeurs aberrantes

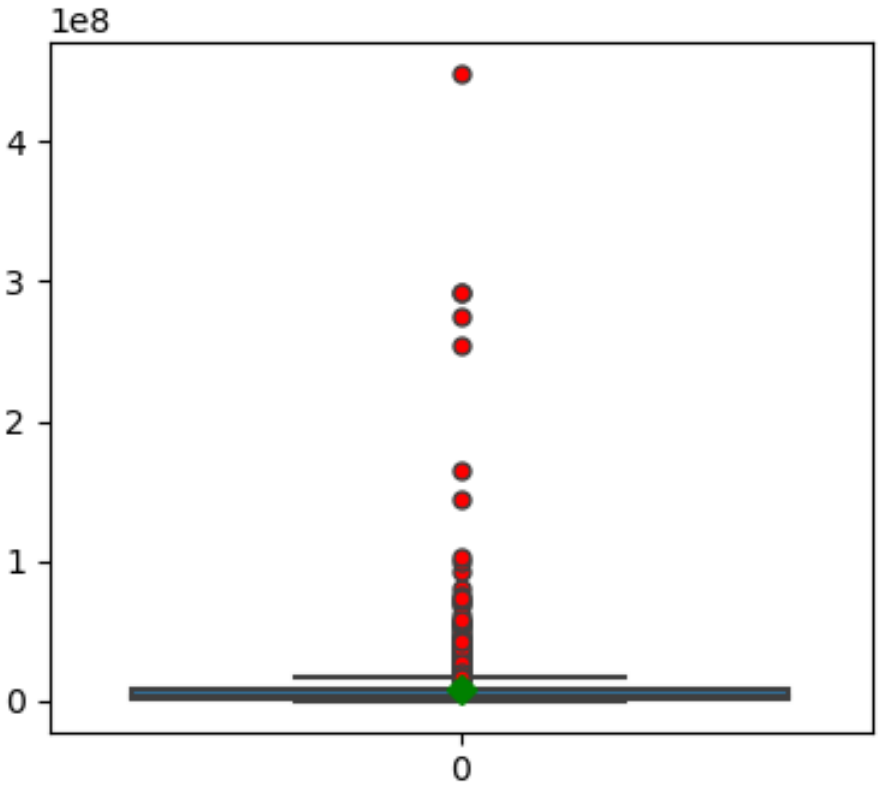
	PrimaryPropertyType	PropertyName	TotalGHGEmissions
35	Mixed Use Property	PLANT 2 SITE	16870.98
618	Hospital	SWEDISH FIRST HILL	12307.16
170	Hospital	HARBORVIEW MEDICAL CENTER	10734.57
124	Hospital	SEATTLE CHILDREN'S HOSPITAL MAIN CAMPUS	8145.52
3264	Hospital	VIRGINIA MASON MEDICAL CENTER - 2149	6330.91
167	Hospital	SWEDISH CHERRY HILL	4906.33

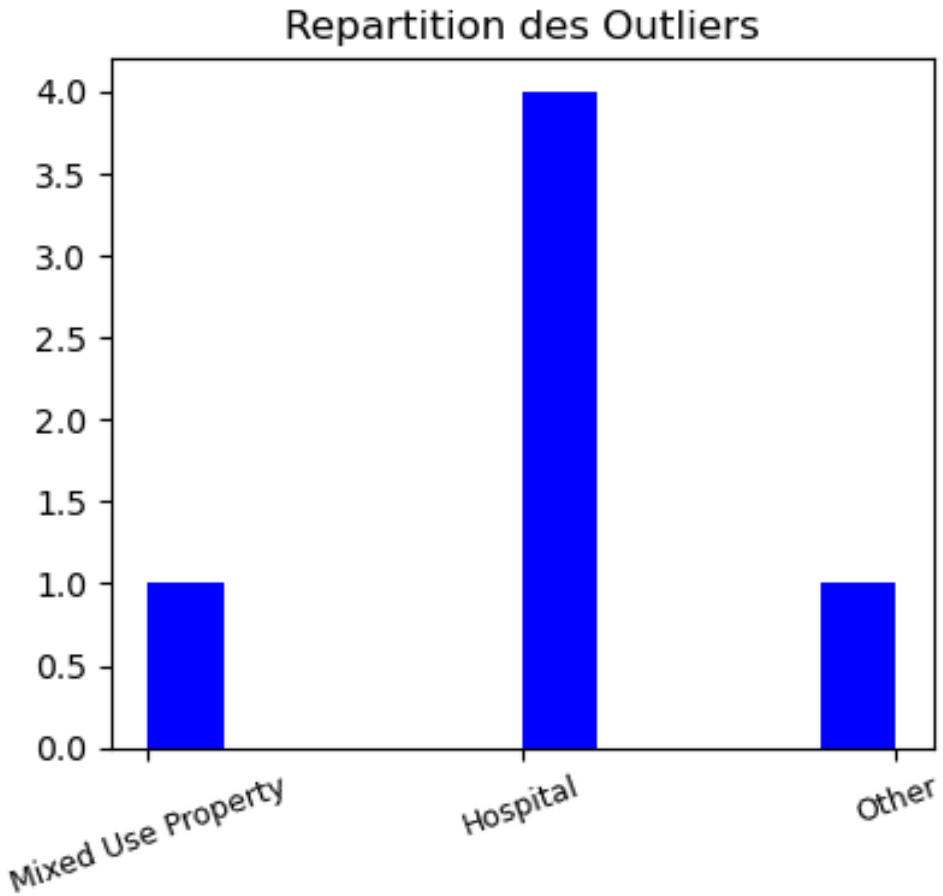




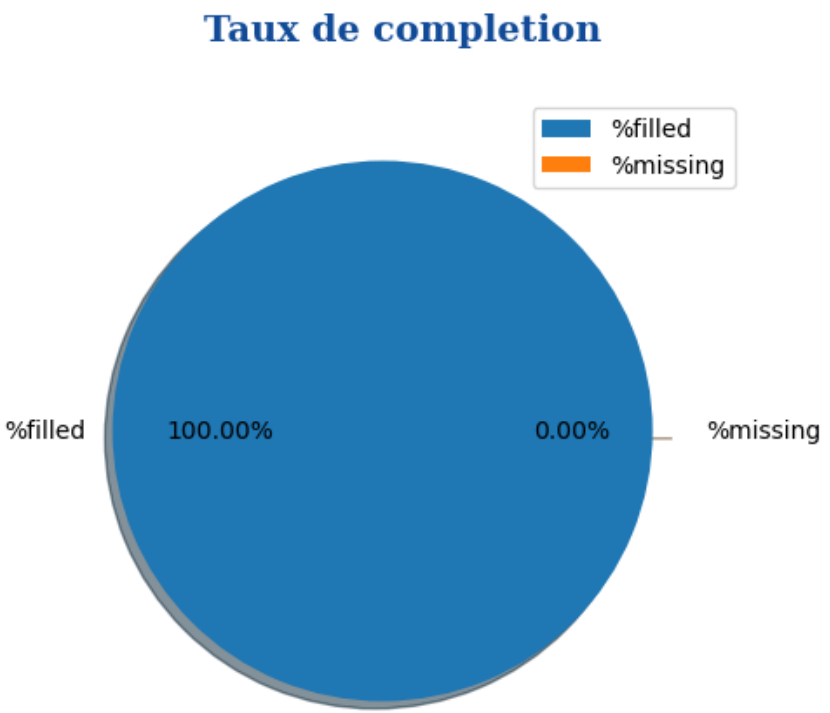
Valeurs aberrantes

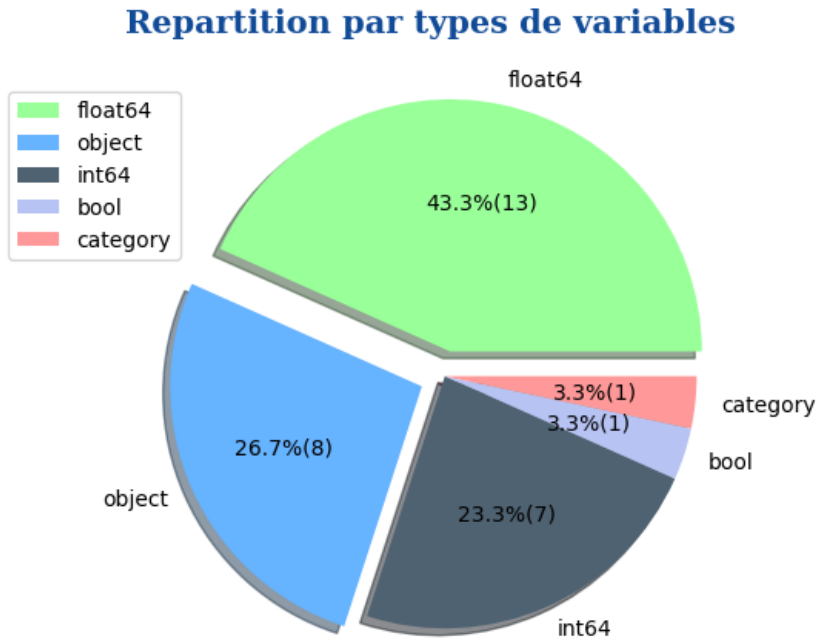
	PrimaryPropertyType	PropertyName	SiteEnergyUse(kBtu)
35	Mixed Use Property	PLANT 2 SITE	448385312.0
170	Hospital	HARBORVIEW MEDICAL CENTER	293090784.0
618	Hospital	SWEDISH FIRST HILL	291614432.0
558	Other	WESTINBUILDING	274682208.0
124	Hospital	SEATTLE CHILDREN'S HOSPITAL MAIN CAMPUS	253832464.0
3264	Hospital	VIRGINIA MASON MEDICAL CENTER - 2149	163945984.0





Bilan nettoyage

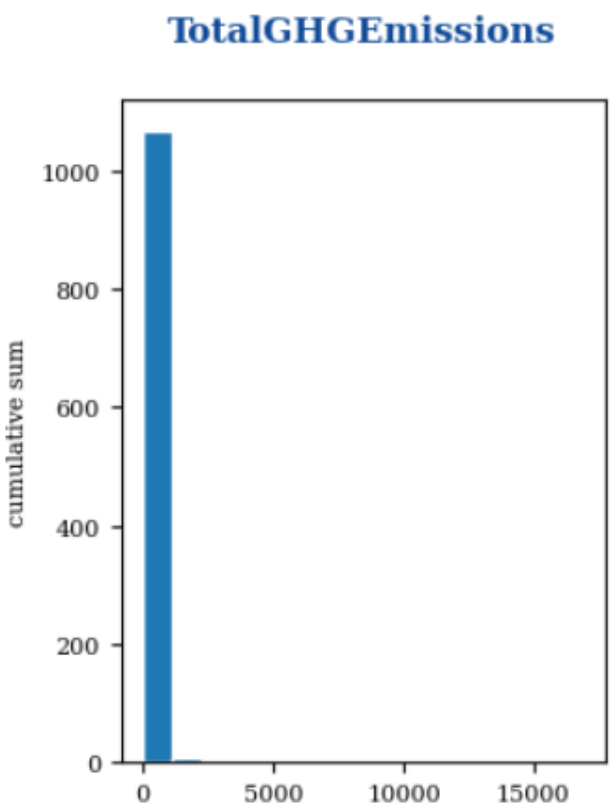


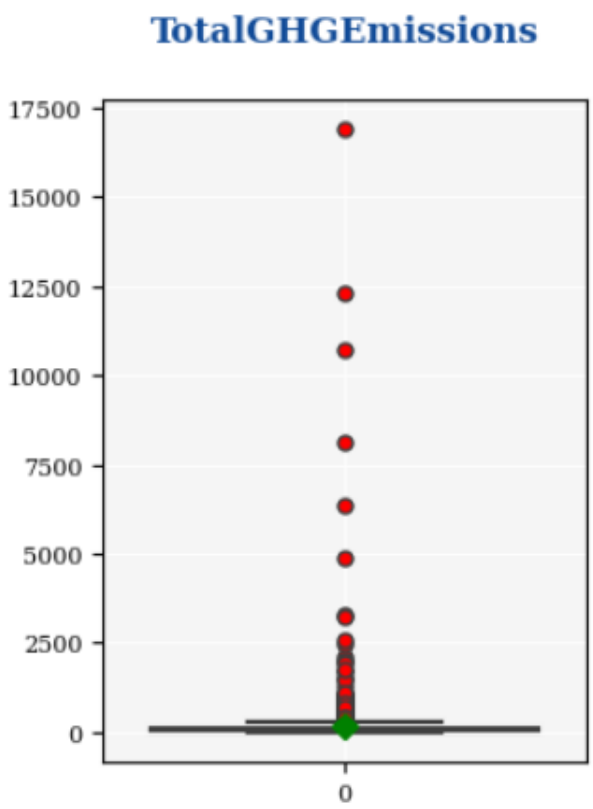


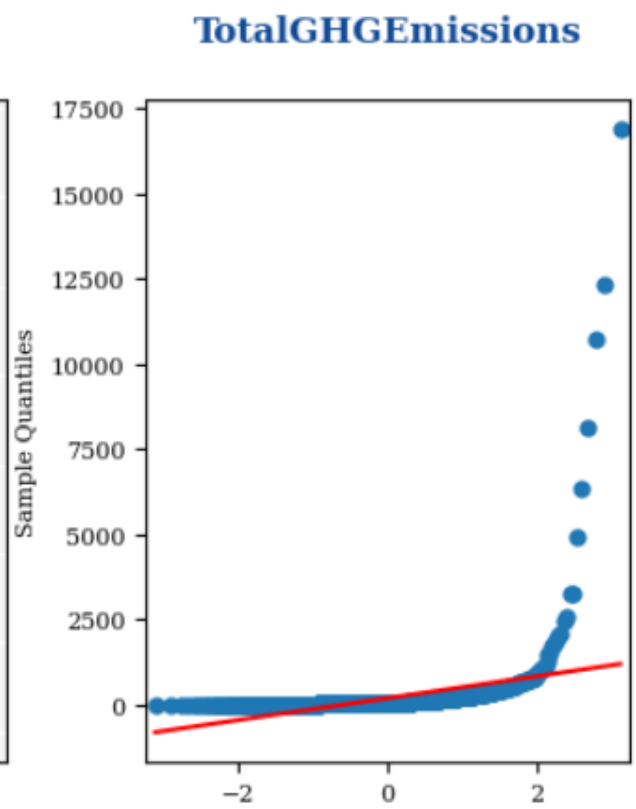
	Variable	nombre
0	lignes	1085
1	colonnes	30

Plan de la présentation

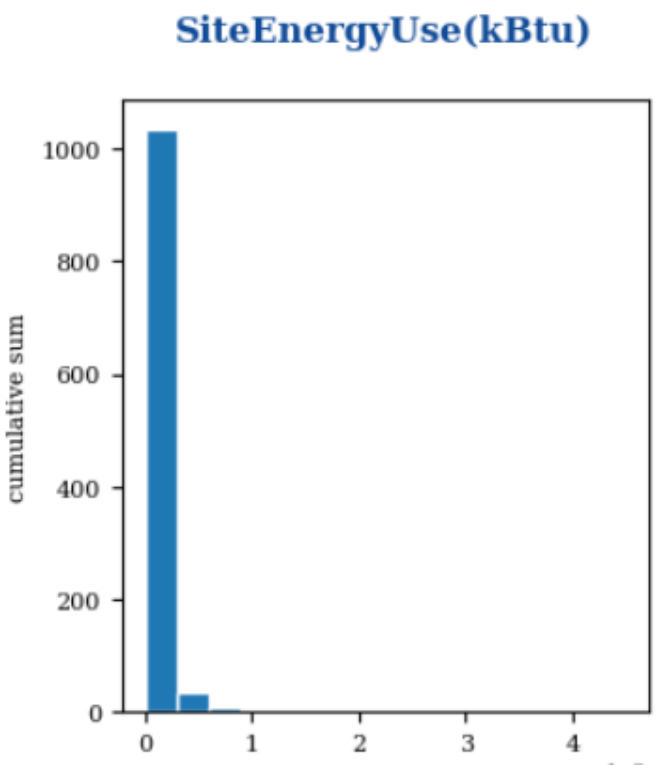
Distribution Emission CO2

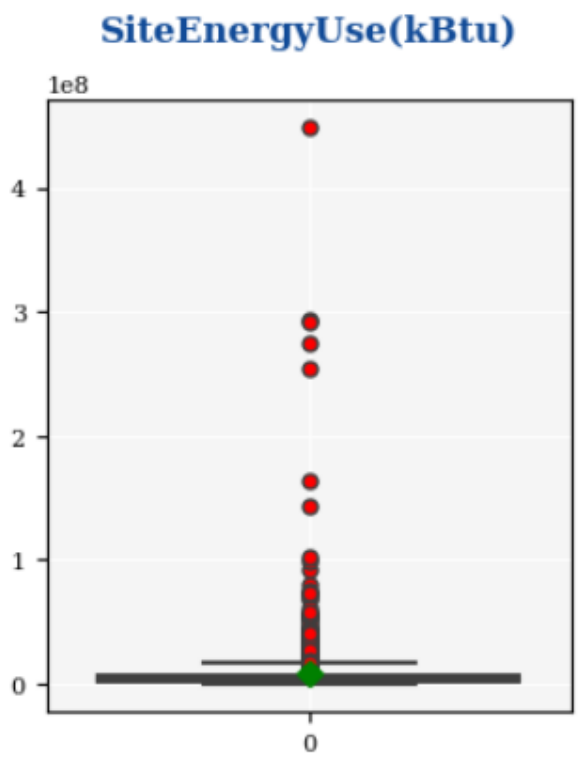


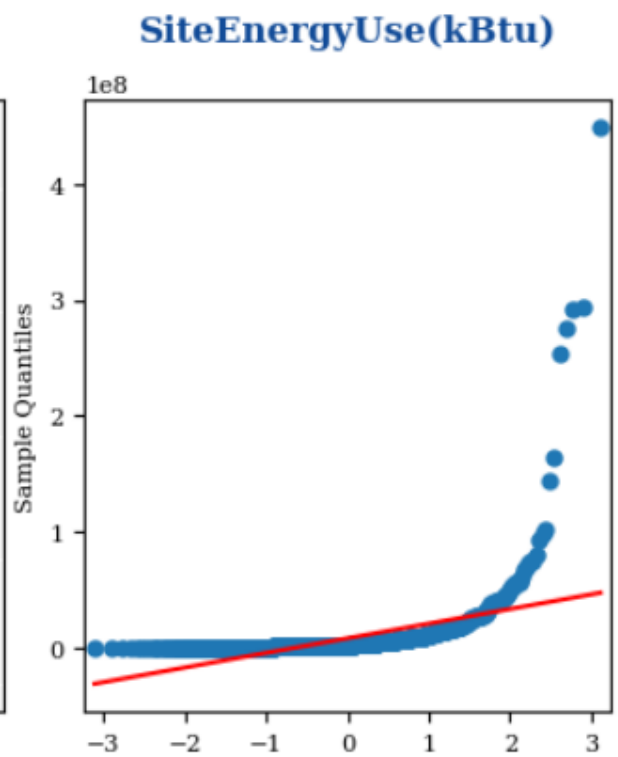




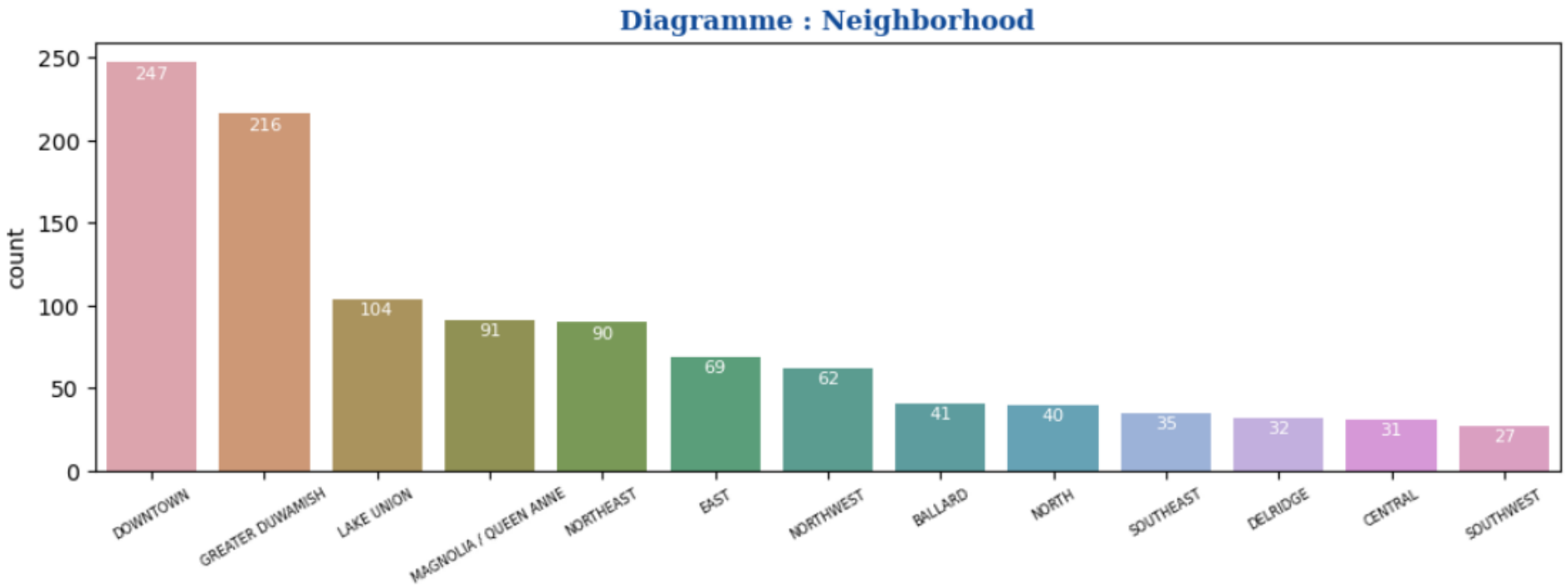
Distribution Consommation d'Energie



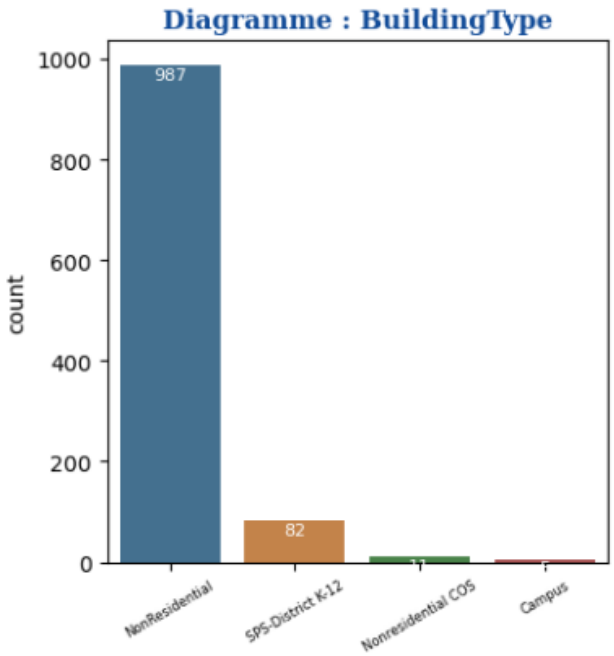


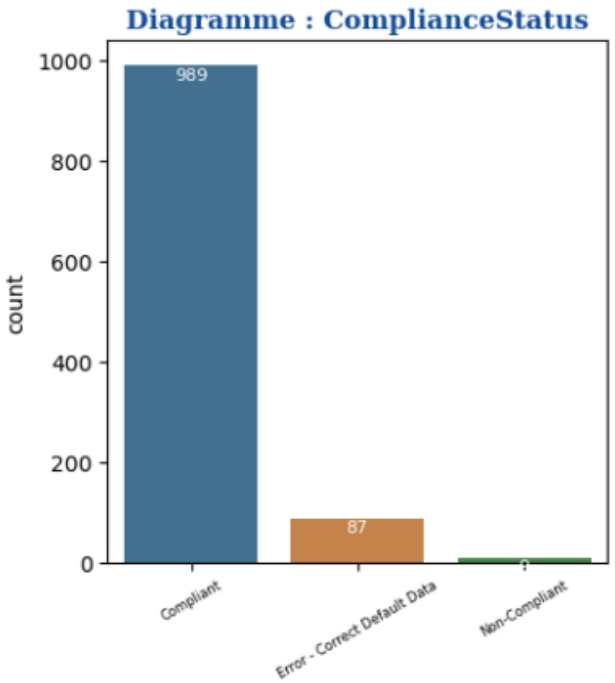


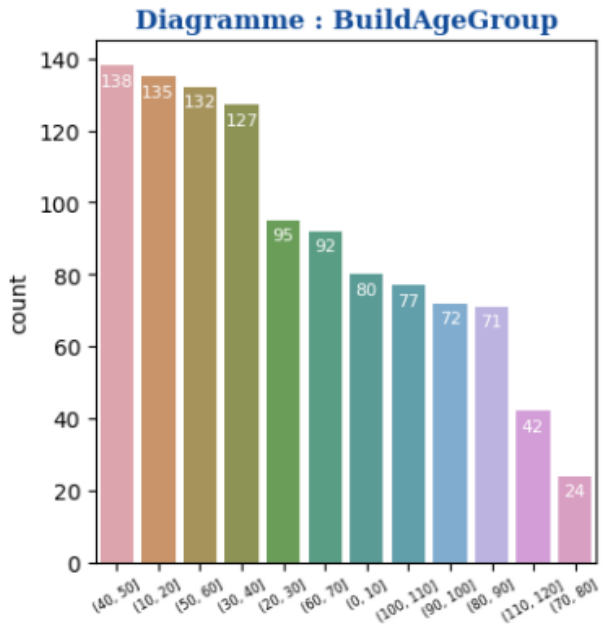
Distribution selon le quartier



Distribution autres variables





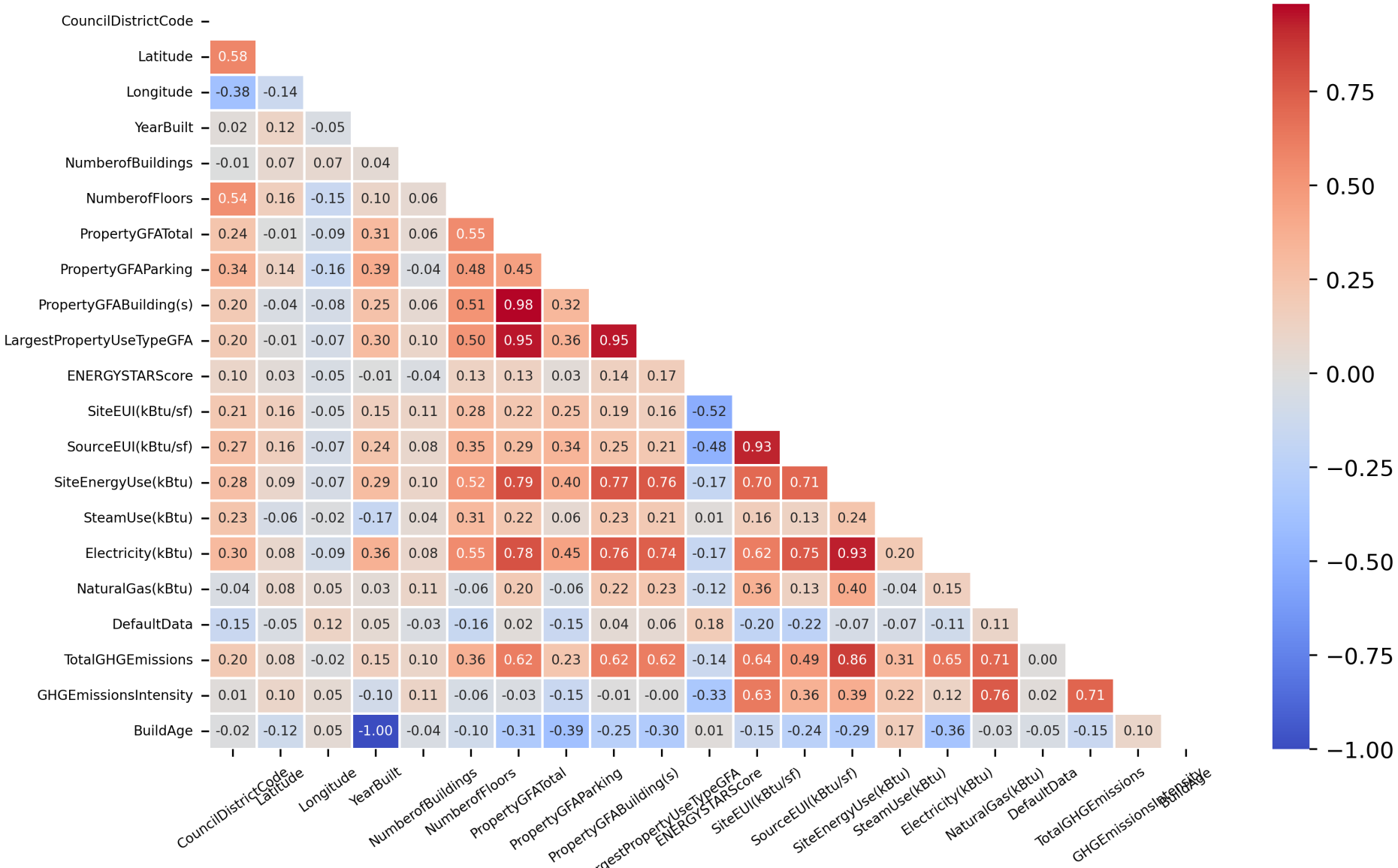


Test de normalité

```
for var in numeric_features:
    test_AndersonDarling(dfBuild, var,seuil=1)
```

```
Distribution normale CouncilDistrictCode : False
Distribution normale Latitude : False
Distribution normale Longitude : False
Distribution normale YearBuilt : False
Distribution normale NumberofBuildings : False
Distribution normale NumberofFloors : False
Distribution normale PropertyGFATotal : False
Distribution normale PropertyGFAParking : False
Distribution normale PropertyGFABuilding(s) : False
Distribution normale LargestPropertyUseTypeGFA : False
Distribution normale ENERGYSTARScore : False
Distribution normale SiteEUI(kBtu/sf) : False
Distribution normale SourceEUI(kBtu/sf) : False
Distribution normale SiteEnergyUse(kBtu) : False
Distribution normale SteamUse(kBtu) : False
Distribution normale Electricity(kBtu) : False
Distribution normale NaturalGas(kBtu) : False
Distribution normale TotalGHGEmissions : False
Distribution normale GHGEmissionsIntensity : False
Distribution normale BuildAge : False
```

Coefficients de corrélation de Spearman



Matrice de corrélation Spearman

- ☞ L'analyse de la matrice de corrélation montre que la variable d'interêt **SiteEnergyUse(kBtu)** est correlée avec les variables *PropertyGFATotal*, *PropertyGFABuilding(s)*, *LargestPropertyUseTypeGFA*, *SiteEUI(kBtu/sf)*, *SourceEUI(kBtu/sf)*, *Electricity(kBtu)*.
- ☞ Quant à la variable d'interêt **TotalGHEmissions**, elle est correlée avec *PropertyGFATotal*, *PropertyGFABuilding(s)*, *LargestPropertyUseTypeGFA*, *SiteEUI(kBtu/sf)*, *SourceEUI(kBtu/sf)*, *NaturalGas(kBtu)*, *Electricity(kBtu)*.

Bilan de l'analyse des correlations

	level_0	level_1	corr_coeff
12	PropertyGFABuilding(s)	PropertyGFATotal	0.983128
10	PropertyGFABuilding(s)	LargestPropertyUseTypeGFA	0.950134
8	PropertyGFATotal	LargestPropertyUseTypeGFA	0.948520
6	PropertyGFATotal	Electricity(kBtu)	0.784482
4	NaturalGas(kBtu)	GHGEmissionsIntensity	0.757432
2	Electricity(kBtu)	PropertyGFABuilding(s)	0.755685
0	Electricity(kBtu)	LargestPropertyUseTypeGFA	0.738973

- corrélations linéaires fortes entre variables.
- ces corrélations peuvent entrainer des problèmes de colinéarité.
- identifier les pairs de variables avec des corrélations supérieures à **0.7**.
- **PropertyGFATotal**, comme explicative quantitative.
- feature engeneering autres variables.

Plan de la présentation

Data preprocessing

- Transformation logarithmique des targets ;
- Normalisation de toutes les données numériques ;
- Années de construction regroupées par décénnies ;
- Données nominales : One Hot Encoding par la méthode **get_dummies** ;
- Données ordinales : **Ordinal Encoding**.

Démarche générale de modélisation

- ☞ Estimation d'un modèle baseline (regression linéaire) ;
- ☞ Estimations de plusieurs autres modèles de predictions ;
- ☞ Recupération des métriques pour chaque modèle estimé (**MAE, MSE, RSME, R2**) ;
- ☞ Comparaison de tous les modèles (métriques) ;
- ☞ Selection du meilleur modèle ;
- ☞ Optimisation des hyperparamètres avec GridSearchCV ;
- ☞ Analyse des résidus.

Modèle Emission C02

	ind	Model	MAE	MSE	R2	RMSE	tempsExecution
6	Emission CO2	randomforest	0.724903	0.881716	0.608622	0.938997	0.681609
7	Emission CO2	xgboost	0.752265	0.920984	0.591192	0.959679	153.470964
5	Emission CO2	knn	0.805327	1.090943	0.515750	1.044482	0.058824
0	Emission CO2	linearregression	0.796248	1.132933	0.497111	1.064393	0.351702
2	Emission CO2	ridge	0.796510	1.133025	0.497070	1.064437	0.027116
4	Emission CO2	linearsvr	0.810837	1.206069	0.464647	1.098212	0.041850
1	Emission CO2	lasso	0.890417	1.352117	0.399819	1.162806	0.024079
3	Emission CO2	elasticnet	1.004934	1.653978	0.265829	1.286071	0.022553
8	Emission CO2	SVR Poly	1.004841	2.392755	-0.062102	1.546853	0.360511

- **RandomForest** présente les meilleures métriques.
- Ce modèle sera retenu comme **meilleur modèle**.

C02 : Impact ENERGY STAR Score

	ind	Model	MAE	MSE	R2	RMSE	tempsExecution
10	Emission CO2	best model with ESS	0.653192	0.660018	0.707030	0.812415	261.605388
9	Emission CO2	best model without ESS	0.692212	0.801988	0.644012	0.895538	415.212450

La variable « **ENERGY STAR Score** » améliore significativement une amélioration de la précision du modèle ; donc, cette variables **reste utile** pour le calcul de l'emission de CO2.

Modèle Consommation d'Energie

	ind	Model	MAE	MSE	R2	RMSE	tempsExecution
6	Conso Energy	randomforest	0.465511	0.395355	0.759866	0.628773	0.418966
7	Conso Energy	xgboost	0.480020	0.435776	0.735315	0.660133	0.145931
5	Conso Energy	knn	0.550043	0.616403	0.625604	0.785114	0.043892
2	Conso Energy	ridge	0.663717	0.772549	0.530764	0.878947	0.025141
0	Conso Energy	linearregression	0.664386	0.774791	0.529402	0.880222	0.052583
4	Conso Energy	linearsvr	0.659926	0.938625	0.429891	0.968827	0.029315
3	Conso Energy	elasticnet	0.805482	1.044239	0.365742	1.021880	0.017284
1	Conso Energy	lasso	0.850251	1.155981	0.297872	1.075166	0.021580
8	Conso Energy	SVR Poly	0.724993	1.477619	0.102513	1.215573	0.243167

- **RandomForest** presente les meilleures métriques.
- Ce modèle sera retenu comme **meilleur modèle**.

Energie : Impact ENERGY STAR Score

	ind	Model	MAE	MSE	R2	RMSE	tempsExecution
10	Conso Energy	best model with ESS	0.454144	0.381084	0.768535	0.61732	5.994788
9	Conso Energy	best model without ESS	0.458921	0.387743	0.764490	0.62269	5.878887

L'apport de la variable « **ENERGY STAR Score** » reste négligeable sur l'amélioration de la précision du modèle ; Donc, **elle n'est pas utile** dans le calcul de la consommation d'Energie des bâtiments.

MERCI POUR VOTRE AIMABLE ATTENTION