Anticipez les besoins en consommation électrique de bâtiments

Seattle

Xavier Barbier - Décembre 2020

CONTEXTE

Seattle

- **Capitale de l'état de Washington**
- **+ grande ville de l'État**

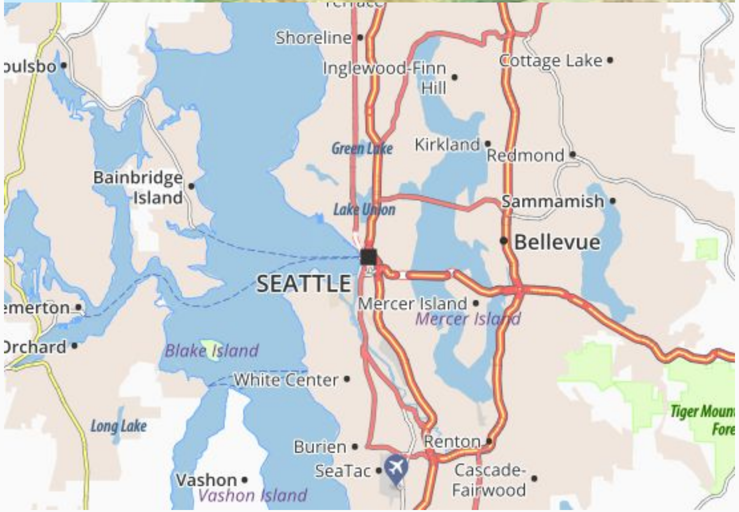
amazon

NORDSTROM

STARBUCKS COFFEE

COSTCO WHOLESALE

CONTEXTE



Historical population		
Census	Pop.	%±
1860	188	—
1870	1,107	488.8%
1880	3,533	219.2%
1890	42,837	1,112.5%
1900	80,671	88.3%
1910	237,194	194.0%
1920	315,312	32.9%
1930	365,583	15.9%
1940	368,302	0.7%
1950	467,591	27.0%
1960	557,087	19.1%
1970	530,831	−4.7%
1980	493,846	−7.0%
1990	516,259	4.5%
2000	563,374	9.1%
2010	608,660	8.0%
2019 (est.)	753,675 [3]	23.8%

U.S. Decennial Census[125]

Objectif de ville neutre en émissions de carbone en 2050

PROJET

Données bâtiments non destinés à l'habitation

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

Modélisation

Prévision

Consommation électrique (ktbu)

Émissions carbone (ghg)

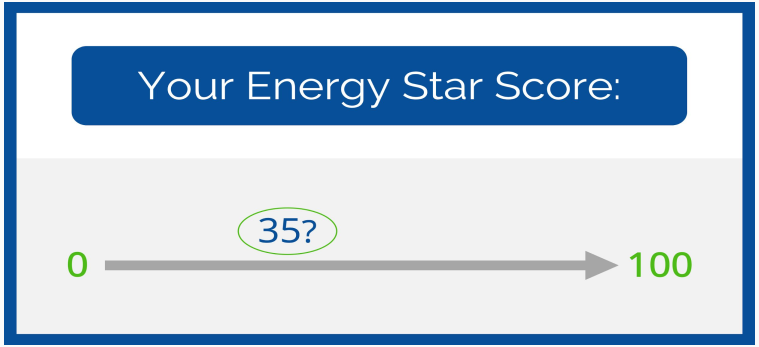
Python / Notebook Jupyter / Colab

QUESTIONS

1. **Quelle est la meilleure modélisation pour une prédiction de la consommation électrique et des émissions carbones de ces bâtiments non destinés à l’habitation ?**

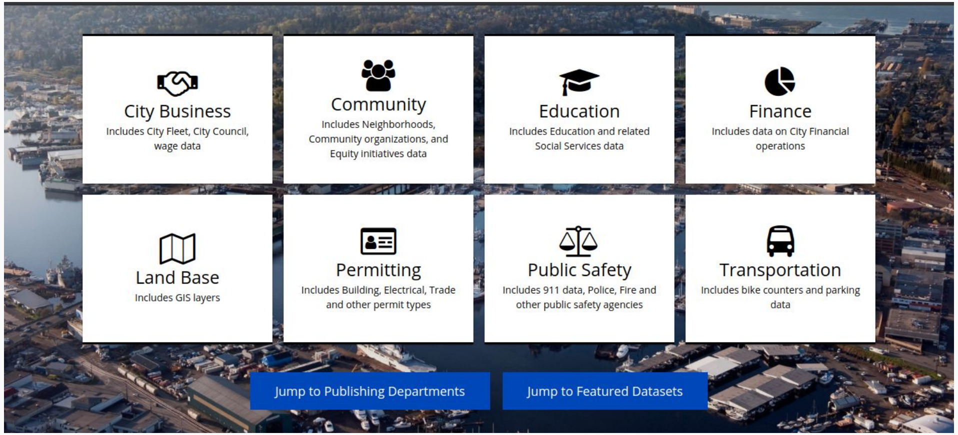
2. **Quel est l’intérêt de l’Energy Star Score ?**

ENERGY STAR



DONNÉES

DONNÉES



Données :

- 2015 (1000 obs / 47 variables)
- 2016 (1000 obs / 45 variables)

Qualitatives :

- étiquettes (type d’activitées)

Quantitatives :

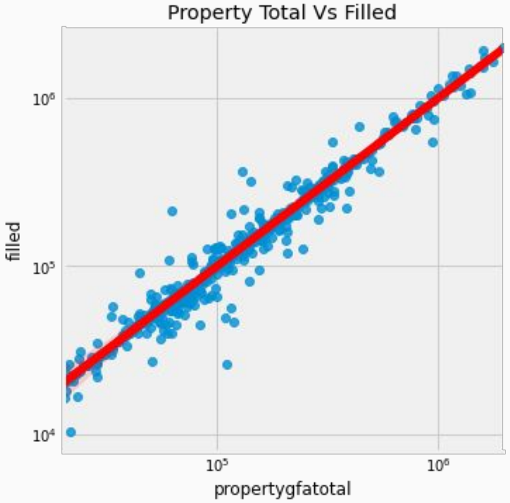
- surfaces
- consommation électrique
- émission carbone

NETTOYAGE

- ❑ **Conservation bâtiments non résidentiels uniquement.**
- ❑ **Conservation variables d’intérêt (type activité, surface, cibles).**
- ❑ **Uniformisation des variables : localisation, id, ghg, use type.**
- ❑ **Merge + suppression doublon (id), conservation ghg supérieur.**
- ❑ **Imputation valeurs aberrantes (ex : 0 floors/building) à partir de sources externes ou par médiane au m² pour ce type de bâtiments (K-12 schools)**
- ❑ **Imputation valeurs manquantes (Nan) pour type d’activités et surface à partir de profils similaires et/ou sources externes, ou logique (parking).**
- ❑ **Imputation surface = 0 quand pas d’activité.**
- ❑ **Nettoyage texte variables qualitatives types d’activités**

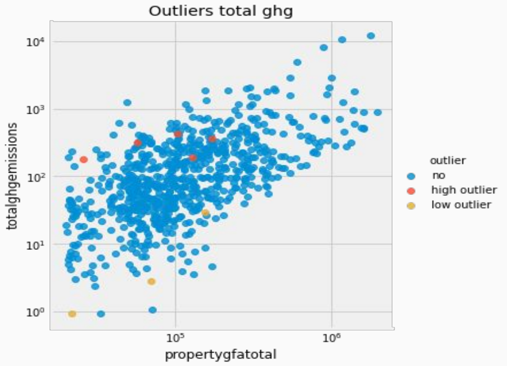
NETTOYAGE

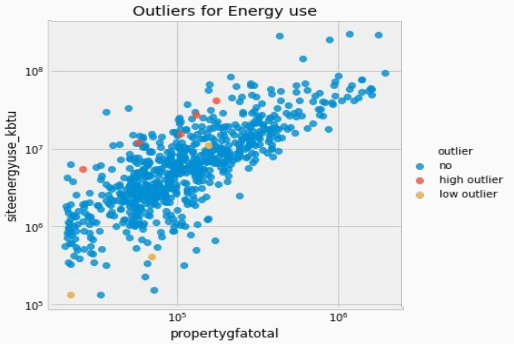
- ❏ **Total = parking + bâtiments: ok**
- ❏ **Largest + second + third > total : 86% des observations**



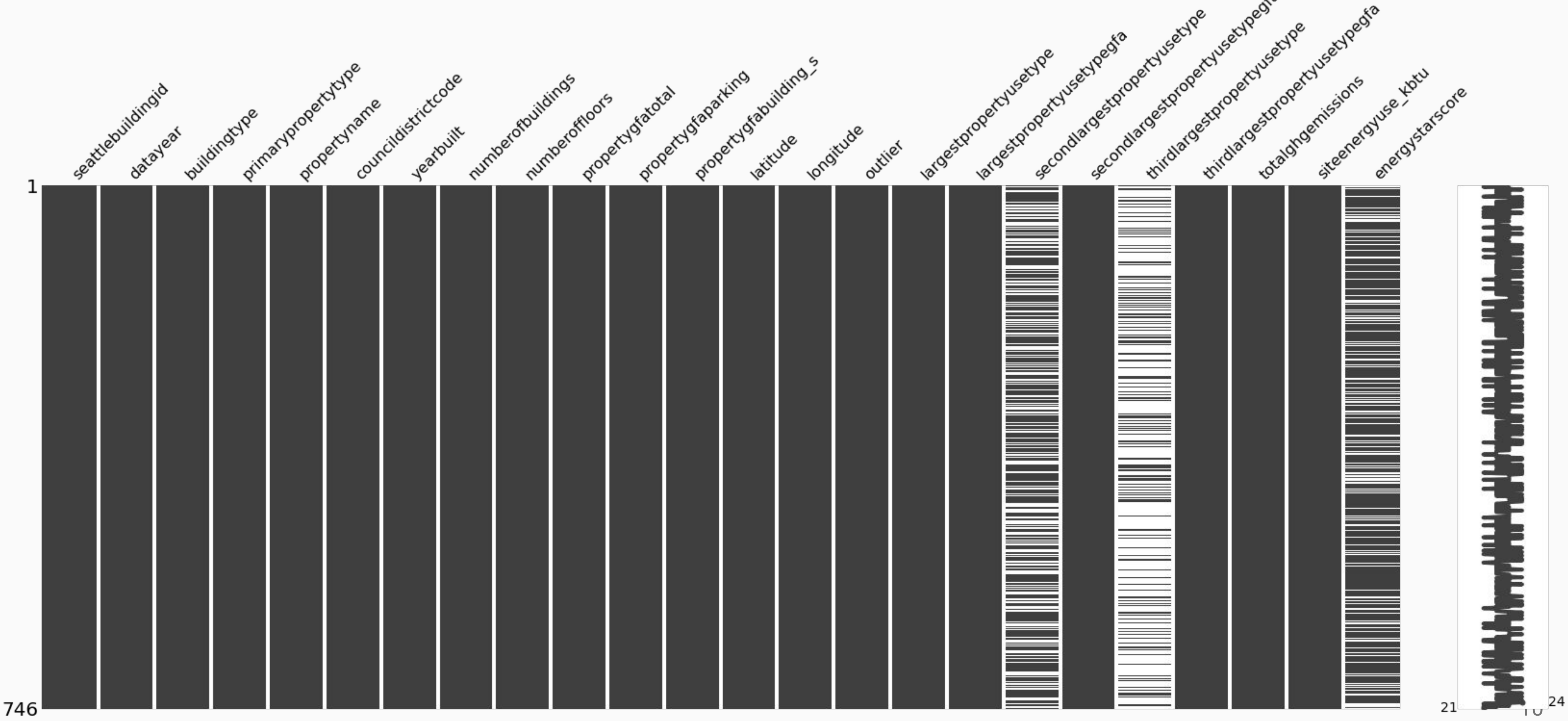
	filled	propertygfatotal
filled	1.000000	0.985713
propertygfatotal	0.985713	1.000000

- ❏ **Outliers variable**





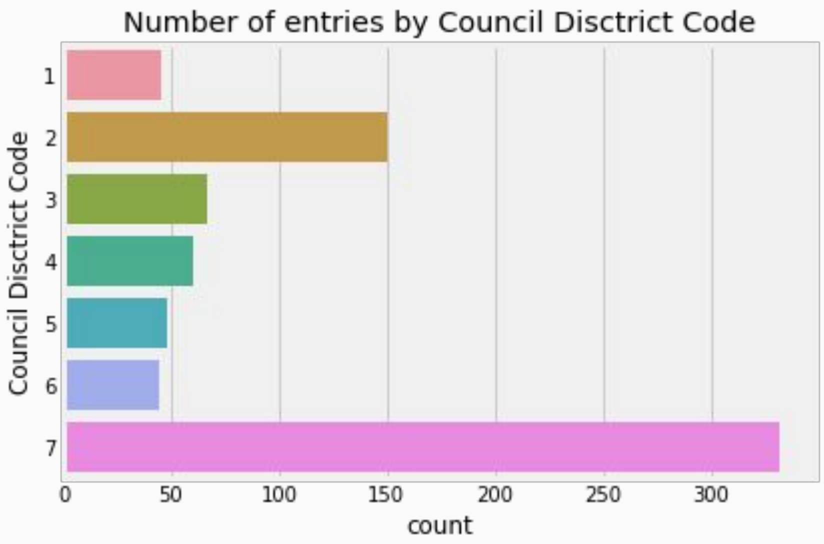
NETTOYAGE

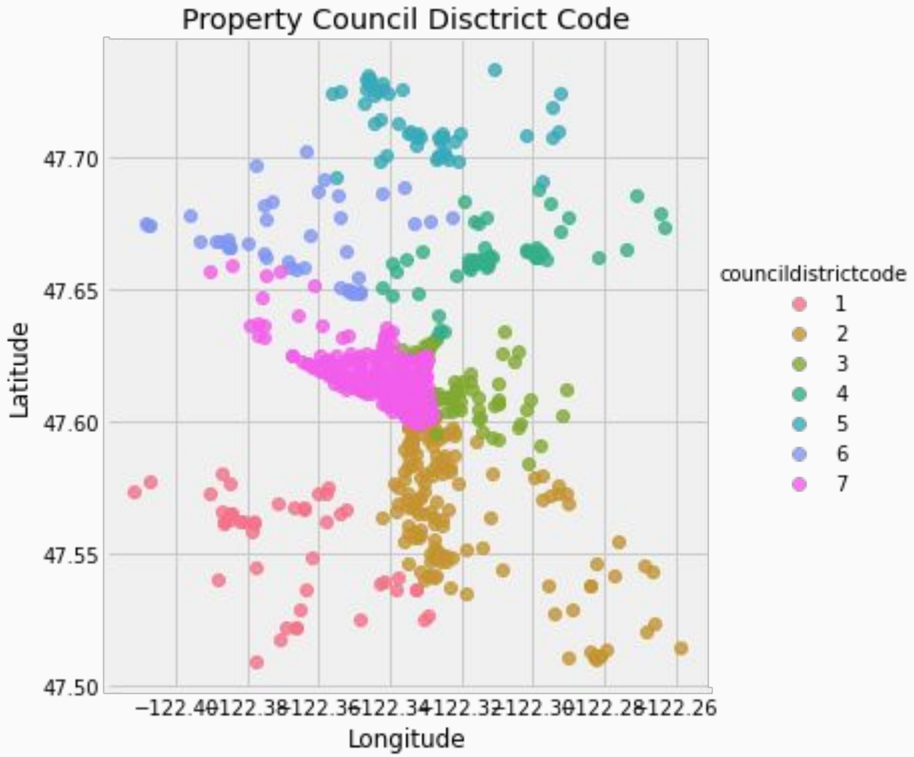

seattlebuildingid
datayear
buildingtype
primarypropertytype
propertyname
councildistrictcode
yearbuilt
numberofbuildings
numberoffloors
propertygfatotal
propertygfaparking
propertygfabuilding_s
latitude
longitude
outlier
largestpropertyusetype
largestpropertyusetypegfa
secondlargestpropertyusetype
secondlargestpropertyusetypegfa
thirdlargestpropertyusetype
thirdlargestpropertyusetypegfa
totalghgemissions
siteenergyuse_kbtu
energystarscore
1
746
21
24

ANALYSE EXPLORATOIRE

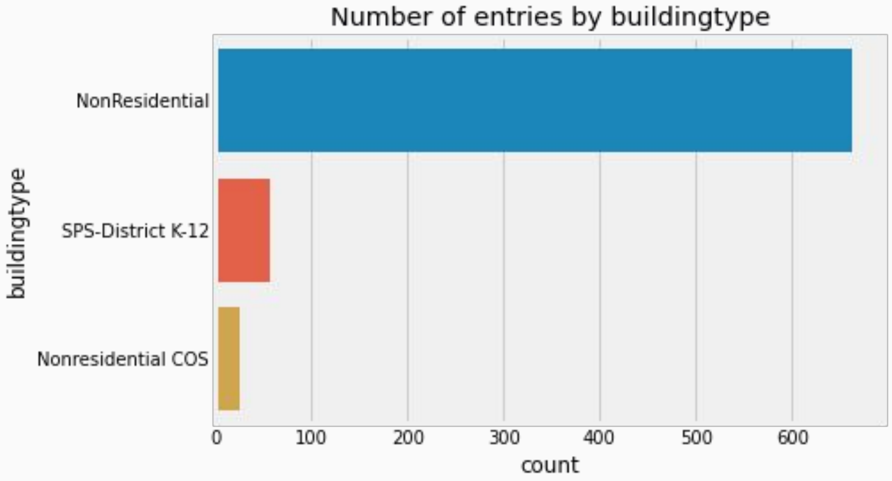
Observations

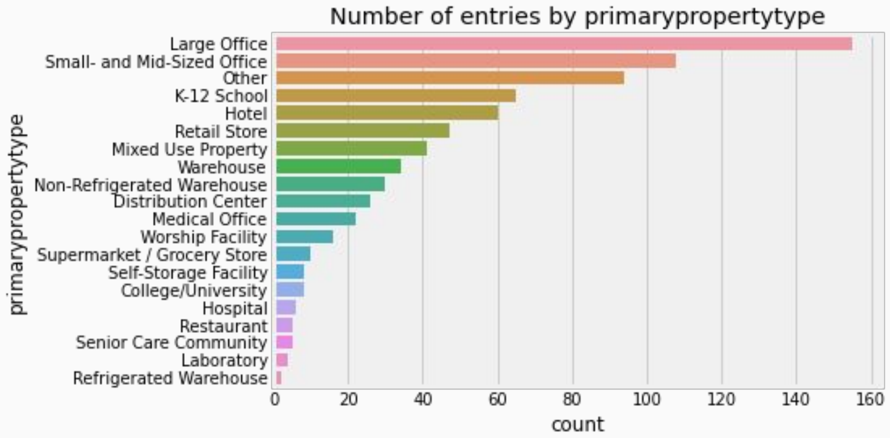
746 obs





Observations

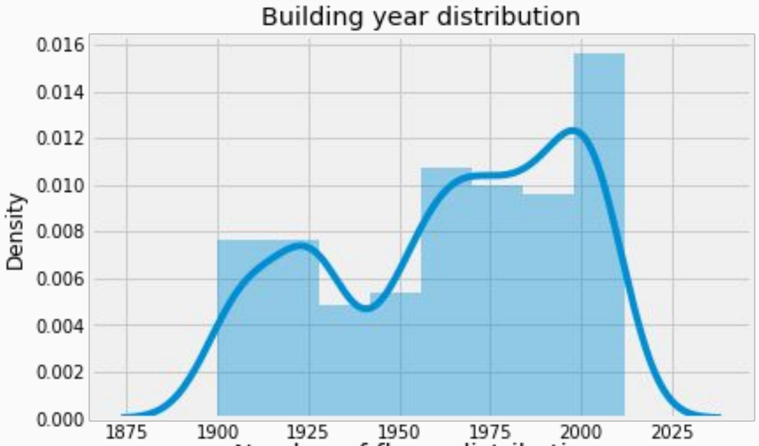


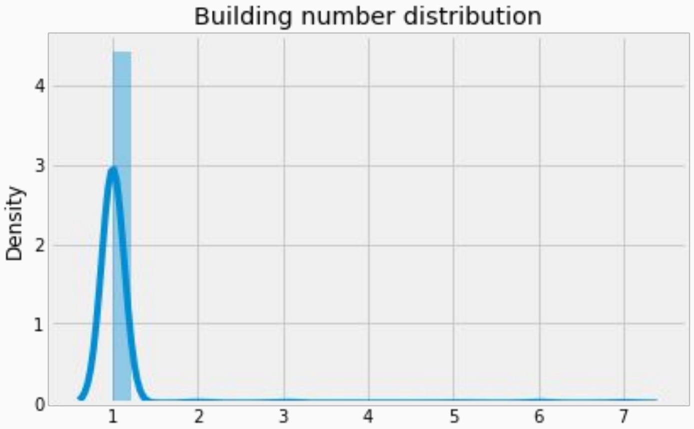


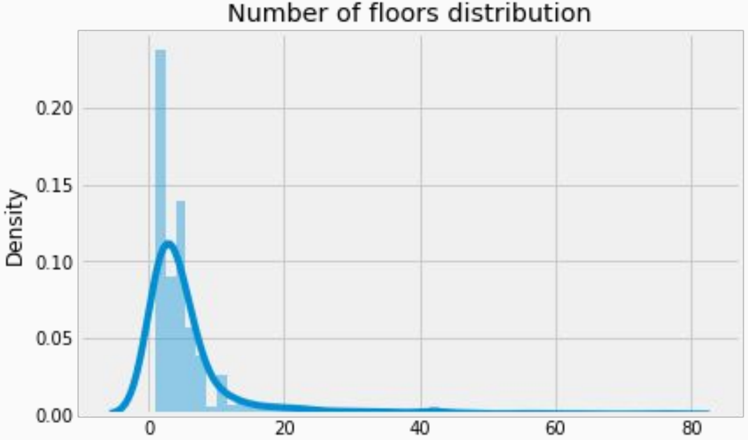
TENDANCES CENTRALES

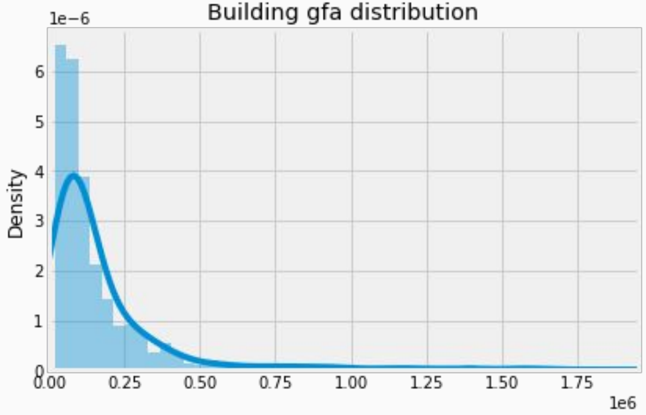
	yearbuilt	numberofbuildings	numberoffloors	propertygfatotal
count	746.000000	746.000000	746.000000	7.460000e+02
mean	1963.733244	1.057641	6.080429	1.788734e+05
std	33.207784	0.497674	8.398879	2.419935e+05
min	1900.000000	1.000000	1.000000	2.002800e+04
25%	1930.000000	1.000000	2.000000	5.848300e+04
50%	1969.000000	1.000000	4.000000	9.842250e+04
75%	1994.000000	1.000000	6.000000	1.907350e+05
max	2012.000000	7.000000	76.000000	1.952220e+06

Distribution

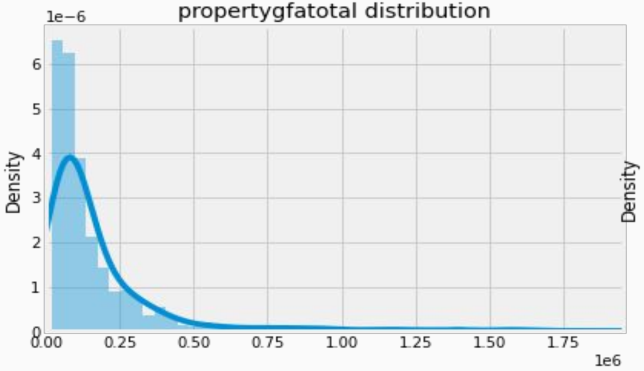


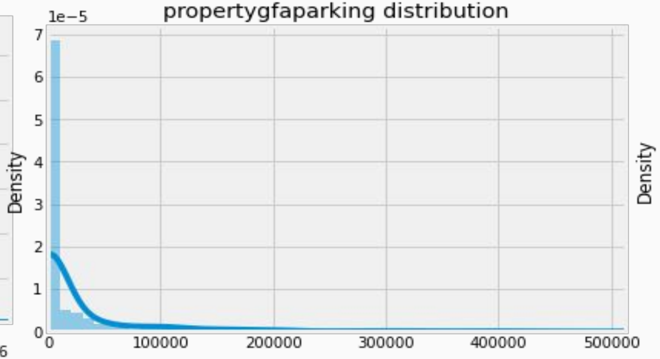


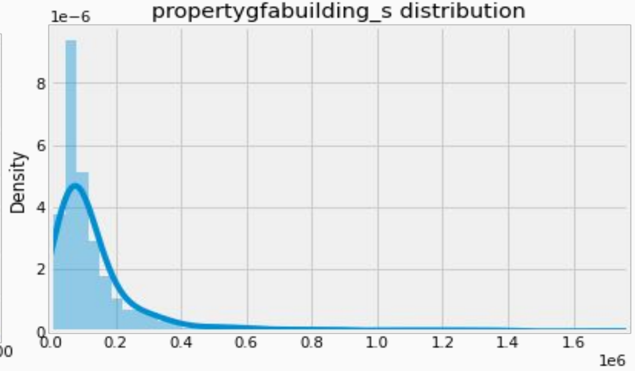


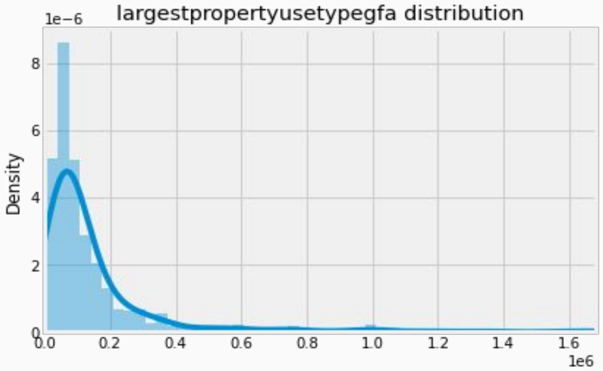


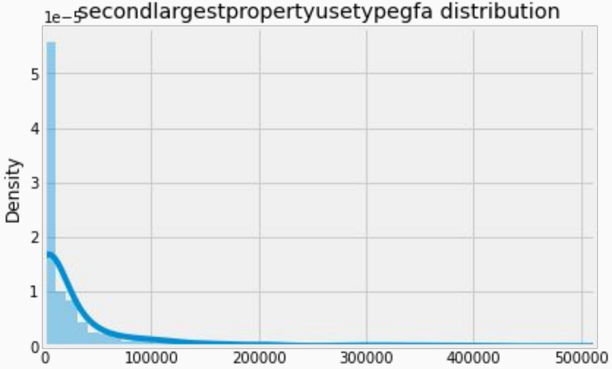
Distribution

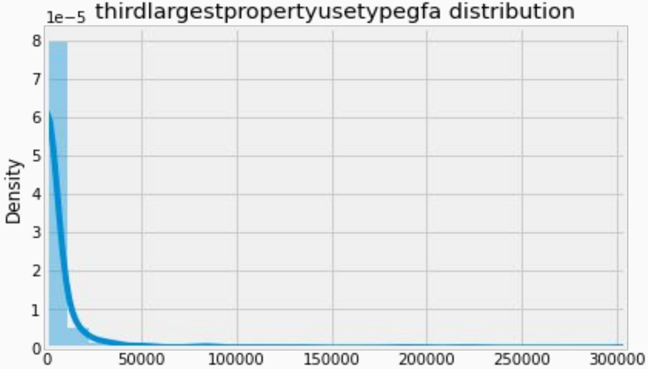




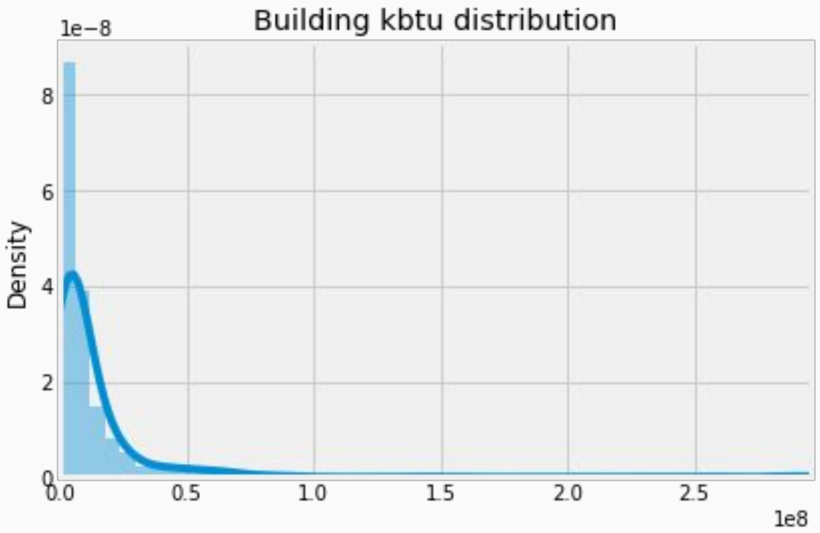


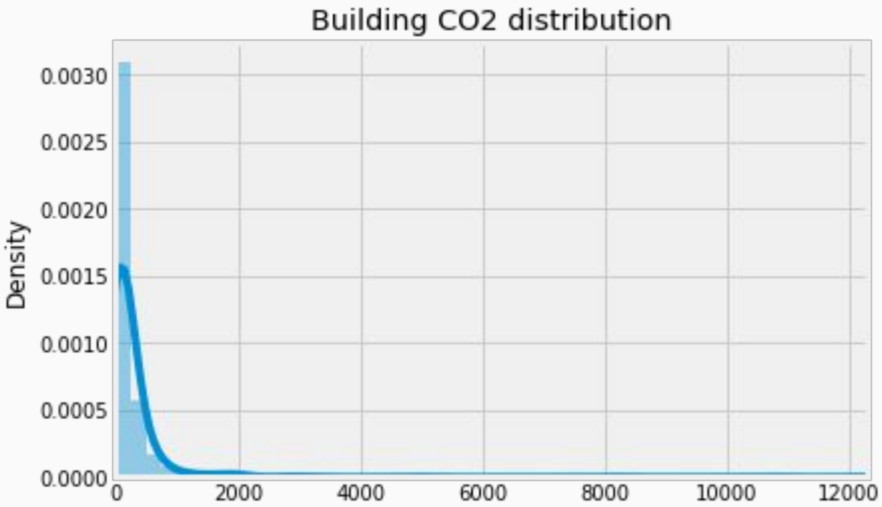






Distribution



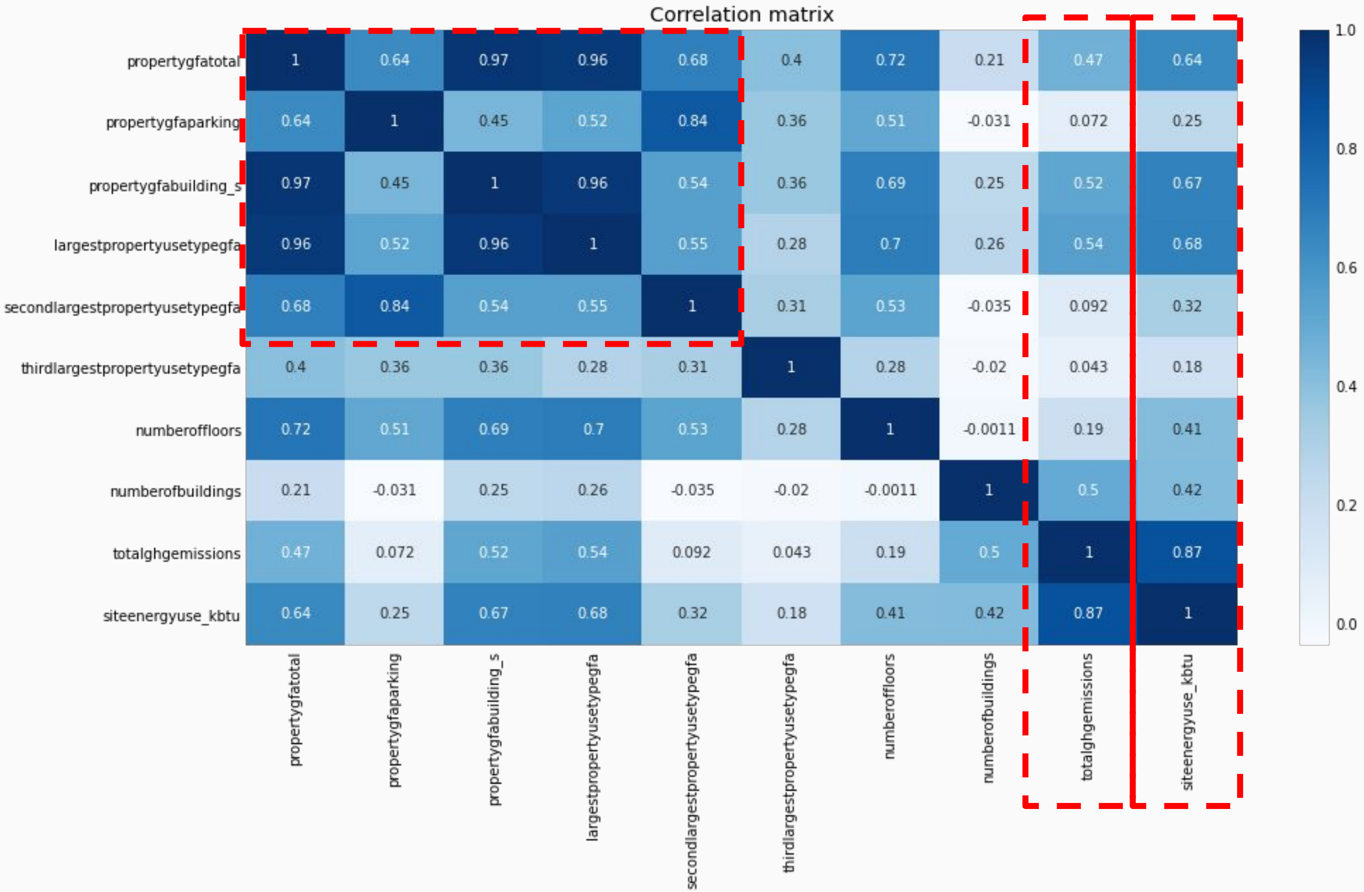


-> 4 hôpitaux dans top 5

-> 5 hôpitaux dans top 10

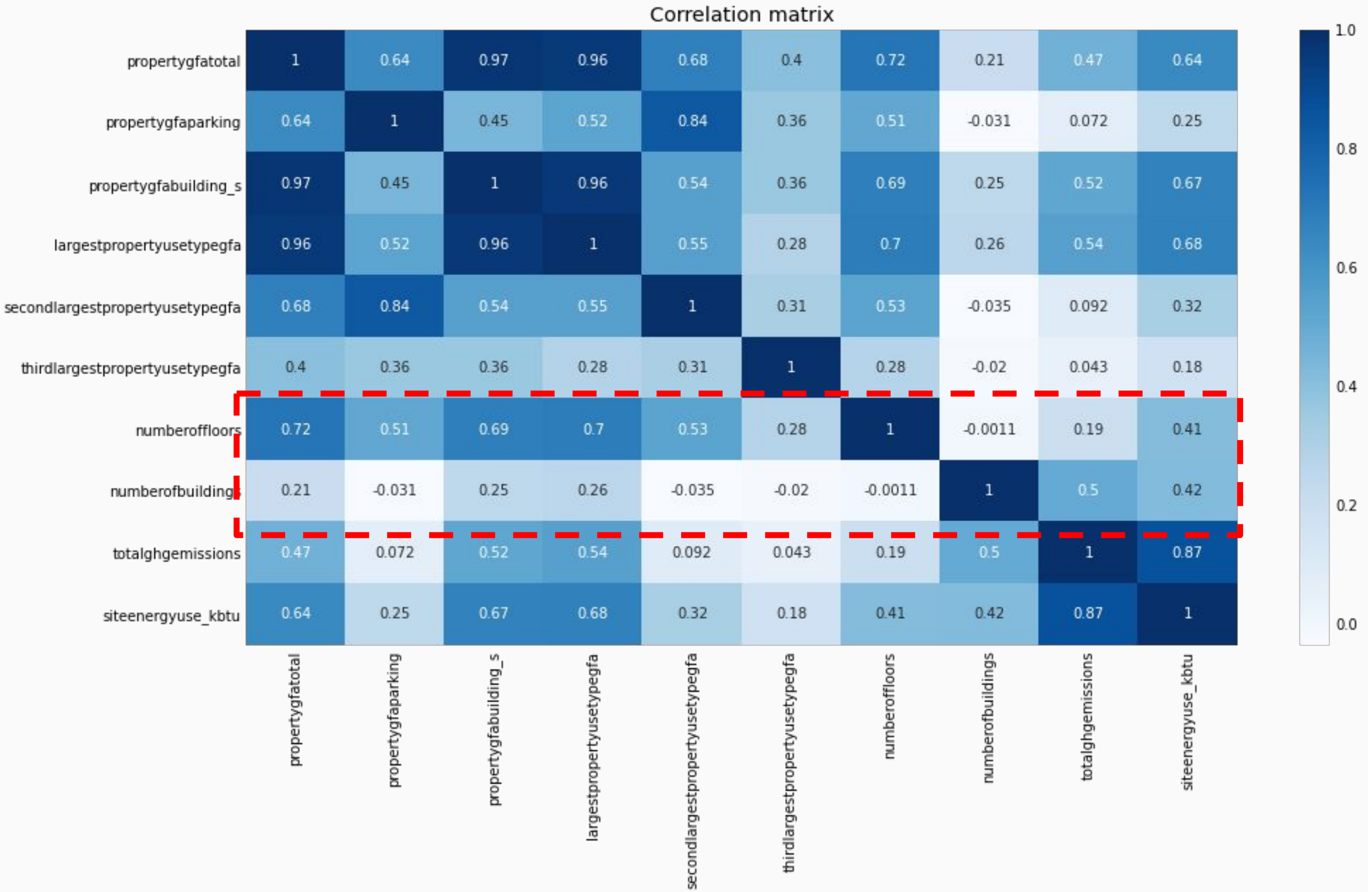
Log10

CORRELATIONS



largestpropertyusetypegfa

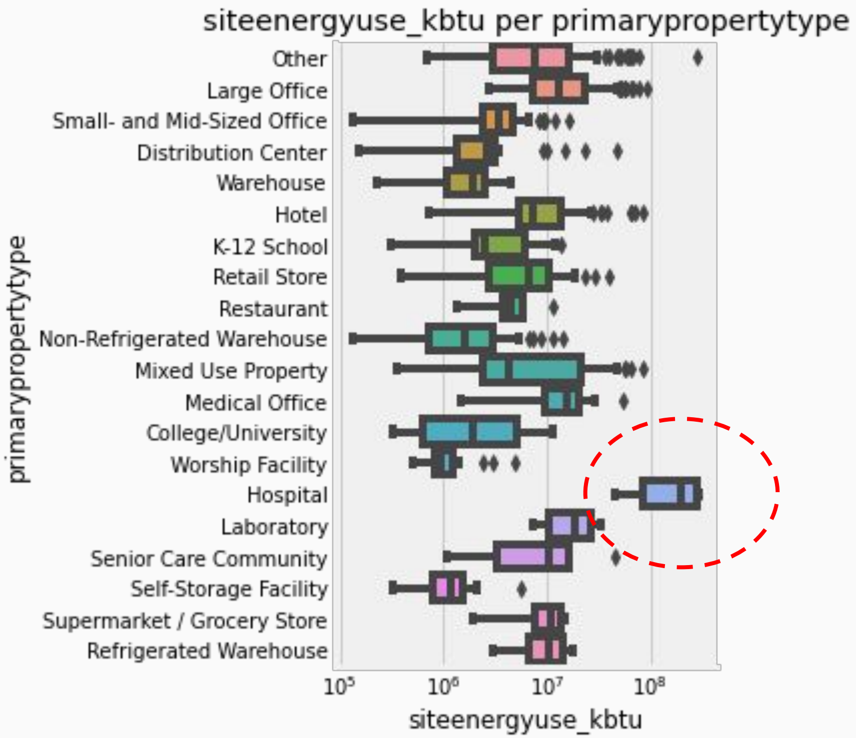
CORRELATIONS

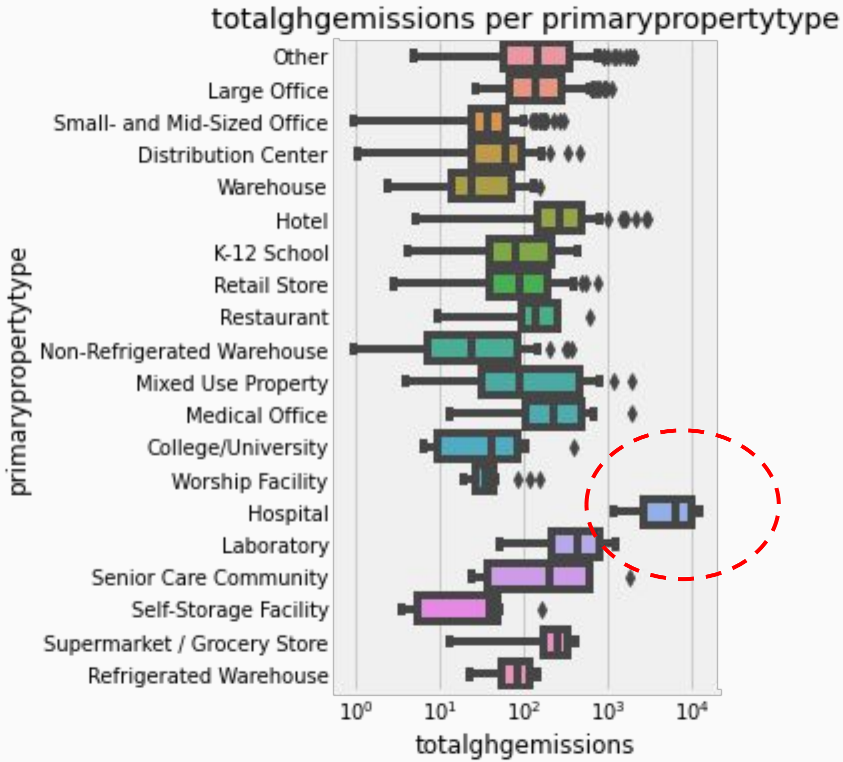


numberoffloors

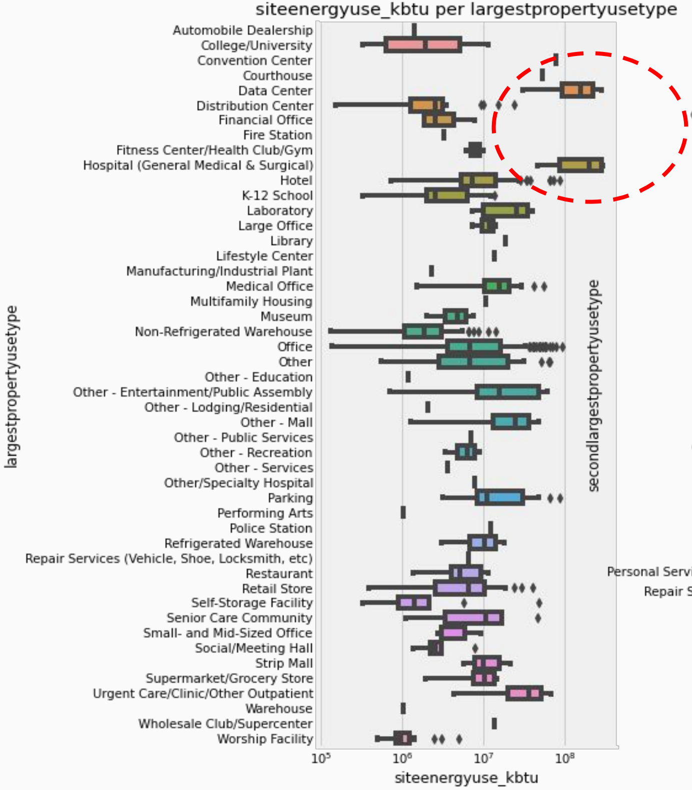
numberofbuildings

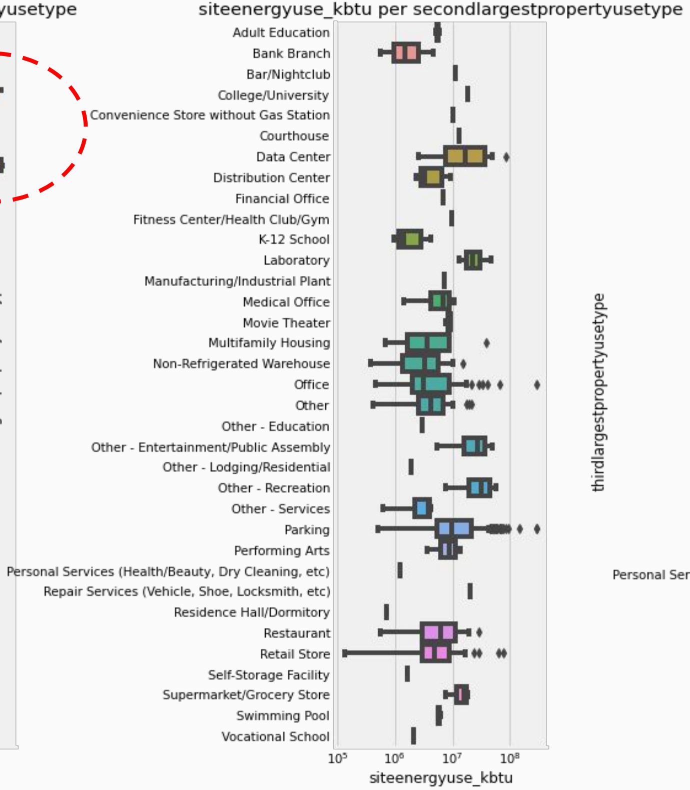
CATÉGORIES : kbtu & ghg per primary property types

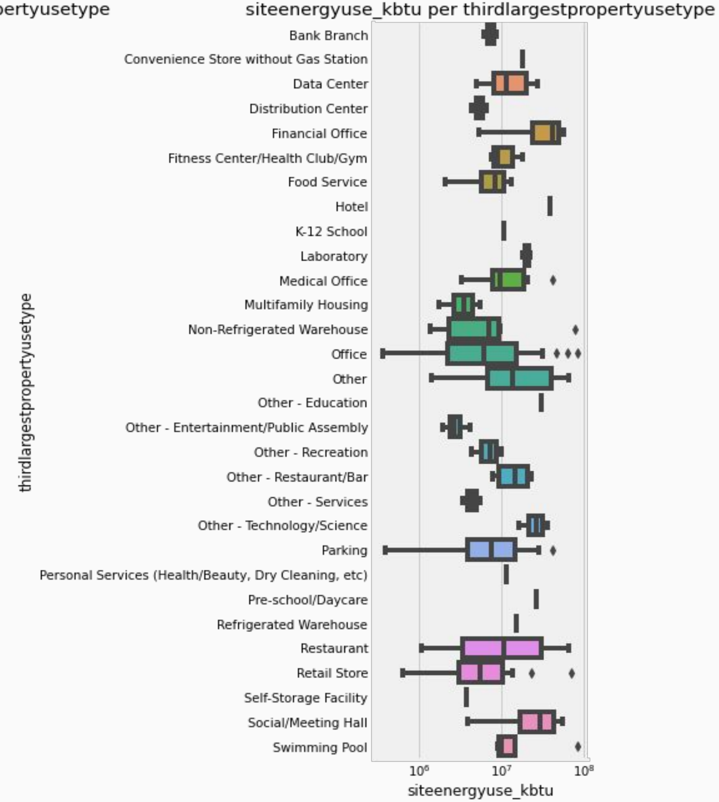




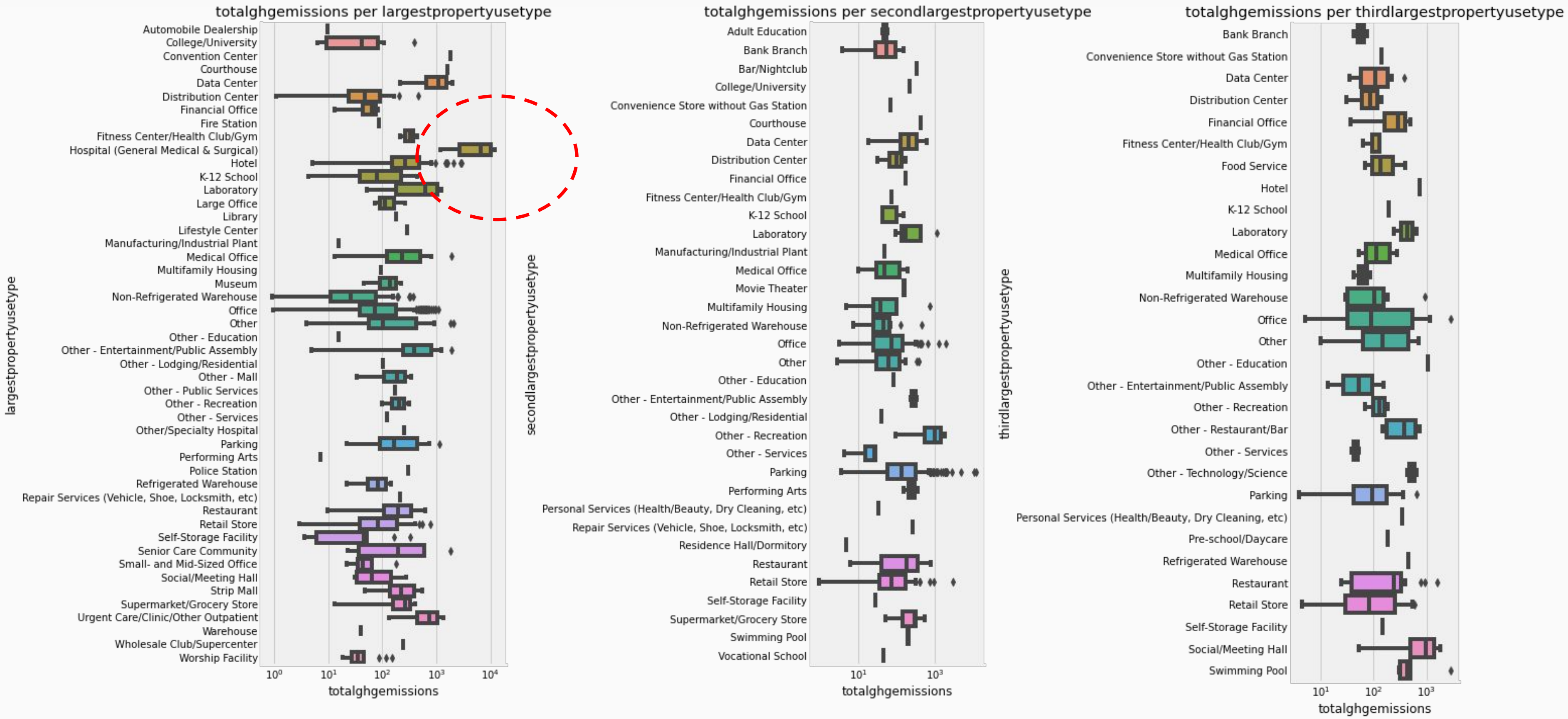
CATÉGORIES : kbtu



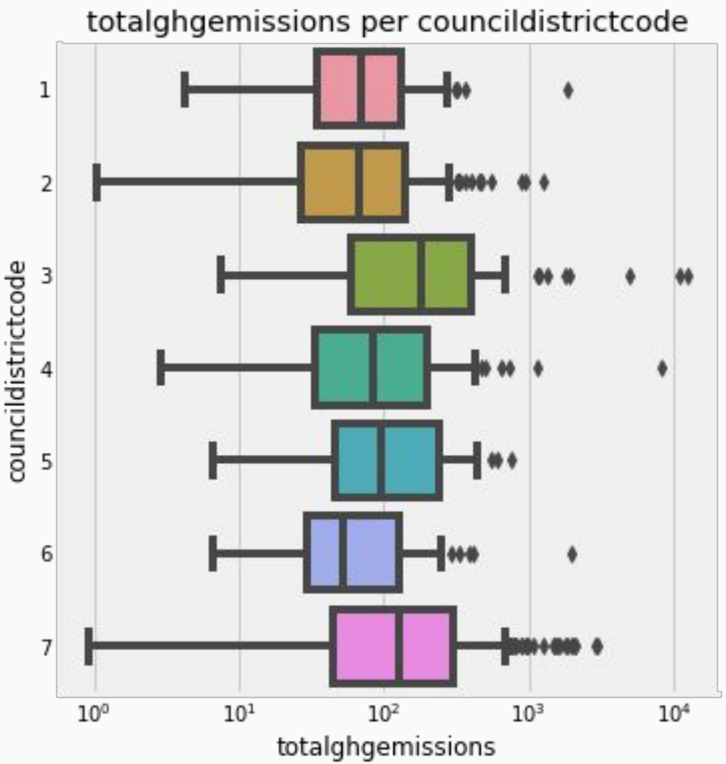


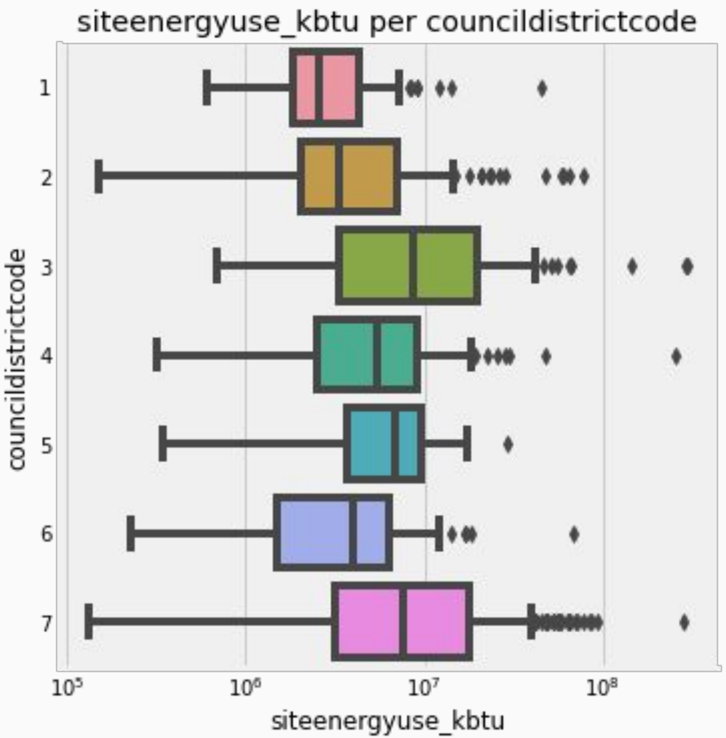


CATÉGORIES : ghg

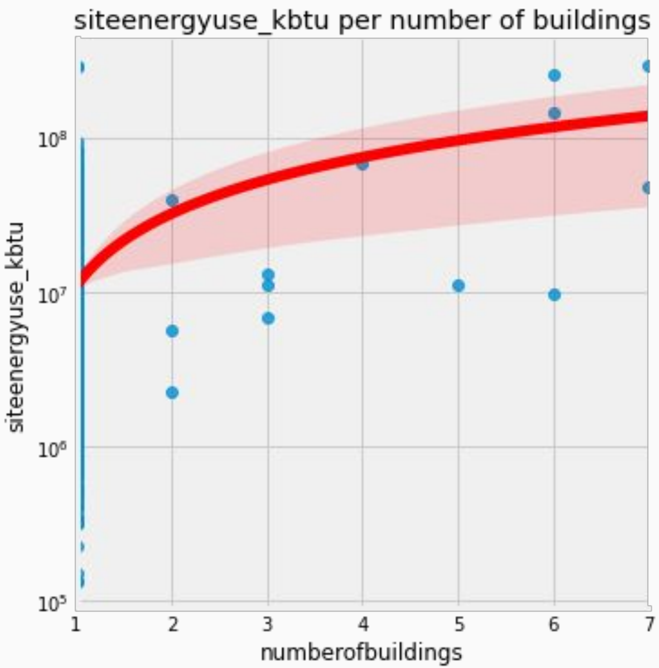


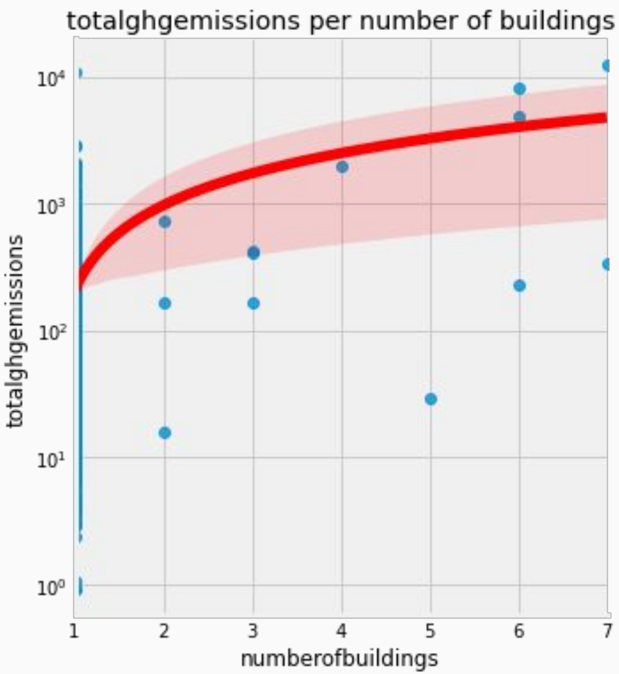
CATÉGORIES : district code





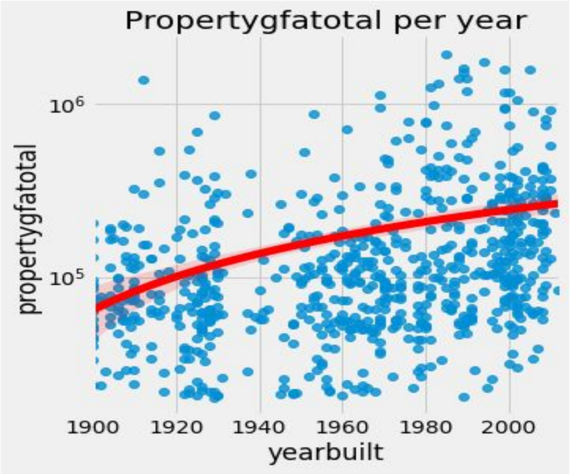
CATÉGORIES : number of building ?

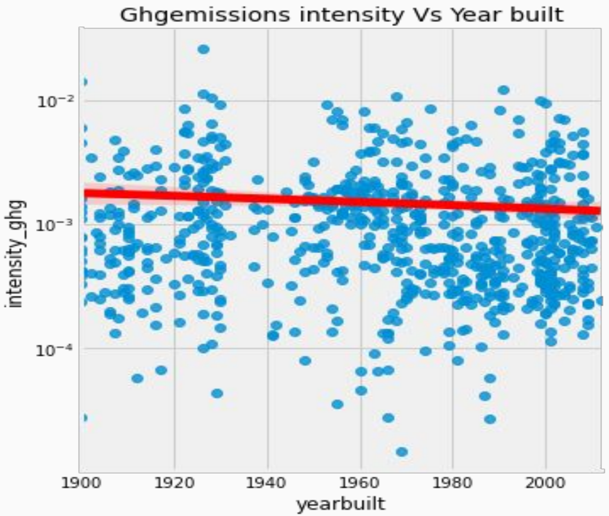


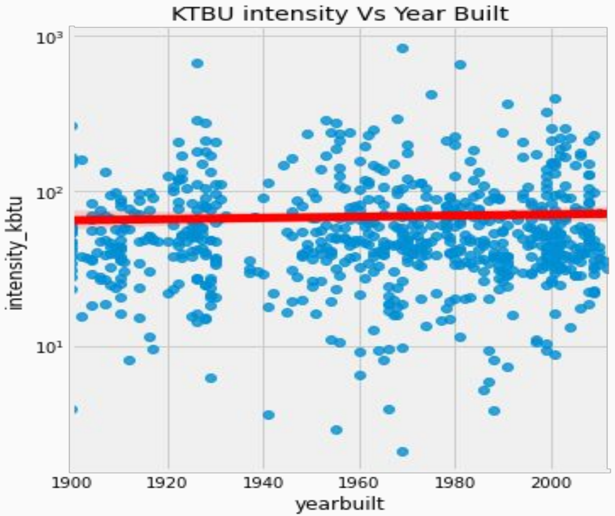


13 bâtiments de +1 étages.
30% hospital.

CATÉGORIES : year built ?







Effets techniques d'isolations

FEATURES ENGINEERING

VARIABLES CATÉGORIELLES

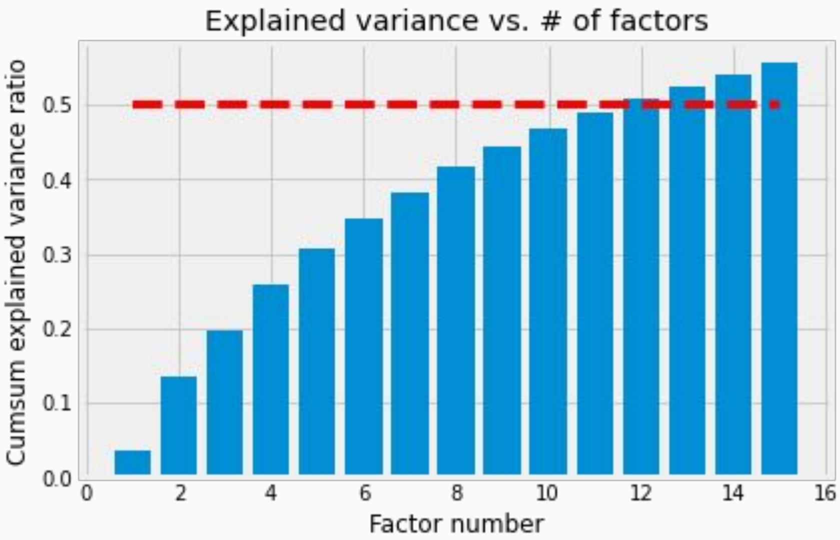
Truncated SVD (Singular Value Decomposition)

- 'buildingtype',
- 'primarypropertytype',
- 'councildistrictcode',
- 'largestpropertyusetype',
- 'secondlargestpropertyusetype',
- 'thirdlargestpropertyusetype',
- "numberofbuildings",
- "yearbuilt"

-> pandas get_dummies (256 variables)

	buildingtype_NonResidential	buildingtype_Nonresidential COS	buildingtype_SPS-District K-12	primarypropertytype_College/University	primarypropertytype_Distribution Center
0	1	0	0	0	0
1	1	0	0	0	0
2	1	0	0	0	0
3	1	0	0	0	0
4	1	0	0	0	0

5 rows × 256 columns



VARIABLES QUANTITATIVES

largestpropertyusetypegfa | **numberoffloors** | **SVD 1-12**

- log10
- Polynomial 2 à 3, + interactions

Consommation électrique (ktbu) | **Émissions carbone (ghg)**

- log10

MODÉLISATION

Méthode

Train, test split (0.33), stratification = primarytype

Pipeline (preprocessing, scaling, gridsearchCV)

Sans polynomial | **Polynomial (2-3)**

Standardisation

GridsearchCV

Modèles | **Paramètres** | **CV 10 fold**

CV

Sans log10 | **log10**

scikit learn

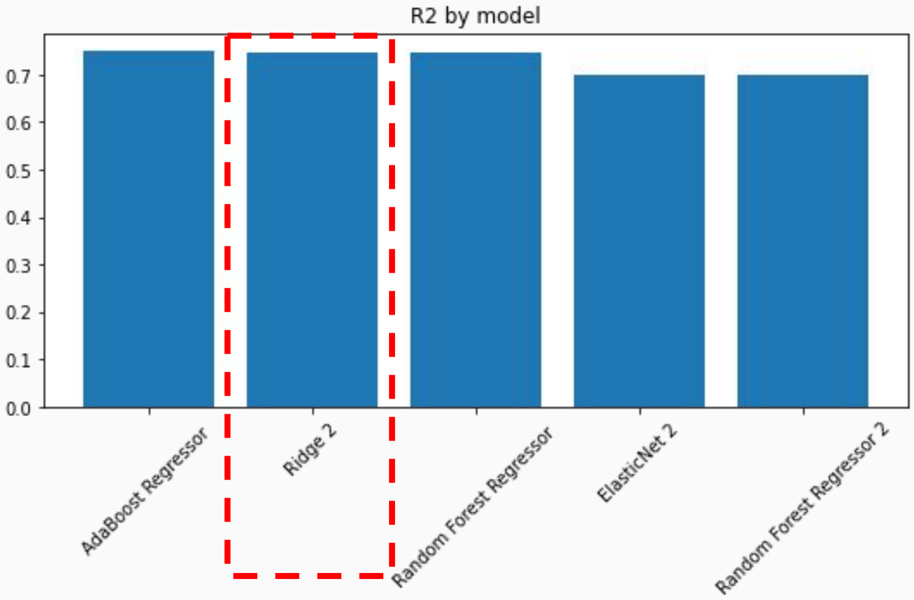
- Linear
- Ridge
- Lasso
- MLP
- Random forest
- Adaboost

43 Modèles

Scores : KTBU

	Model	Degree	Target	RMSE	R2
0	AdaBoost Regressor		siteenergyuse_kbtu	12051718.624488	0.750546
1	Ridge	2	siteenergyuse_kbtu	12119797.784640	0.747719
2	Random Forest Regressor		siteenergyuse_kbtu	12133527.582050	0.747148
3	ElasticNet	2	siteenergyuse_kbtu	13215476.597134	0.700043
4	Random Forest Regressor	2	log10_siteenergyuse_kbtu	0.305668	0.698259

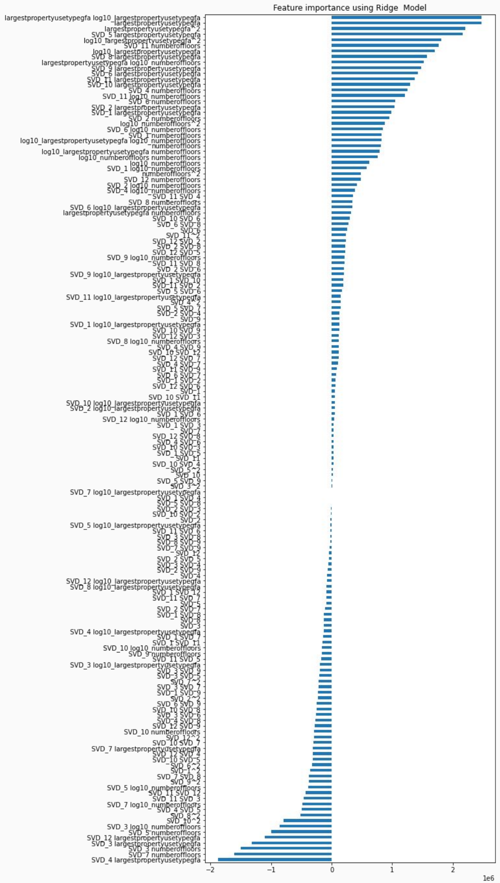
```
R2 X_test: 0.7396431649936854
R2 X_train: 0.8488071097828279
```



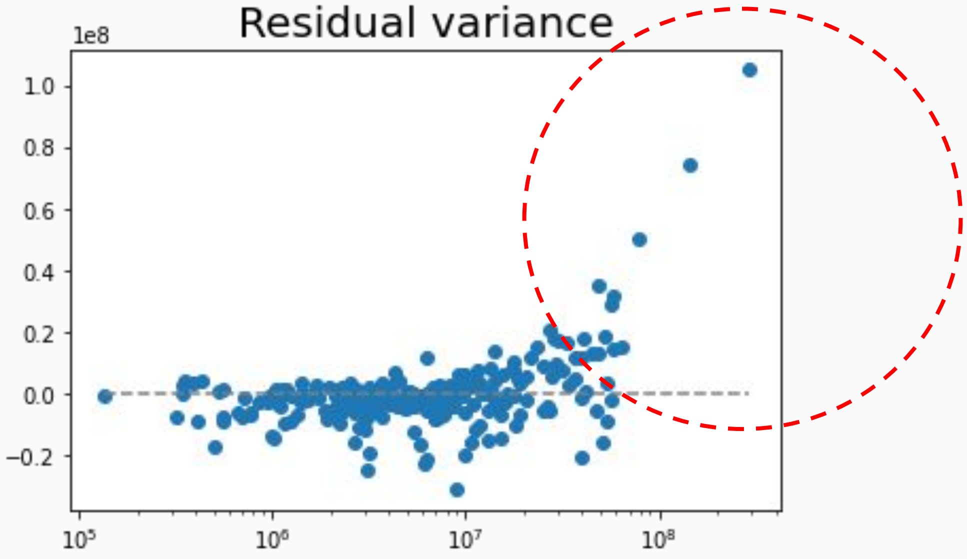
Ridge polynomial 2

```
R2 X_test: 0.7477194211682014
R2 X_train: 0.644218035382027
```

Scores : KTBU



Ridge a sélectionné 152 variables et rejeté 1

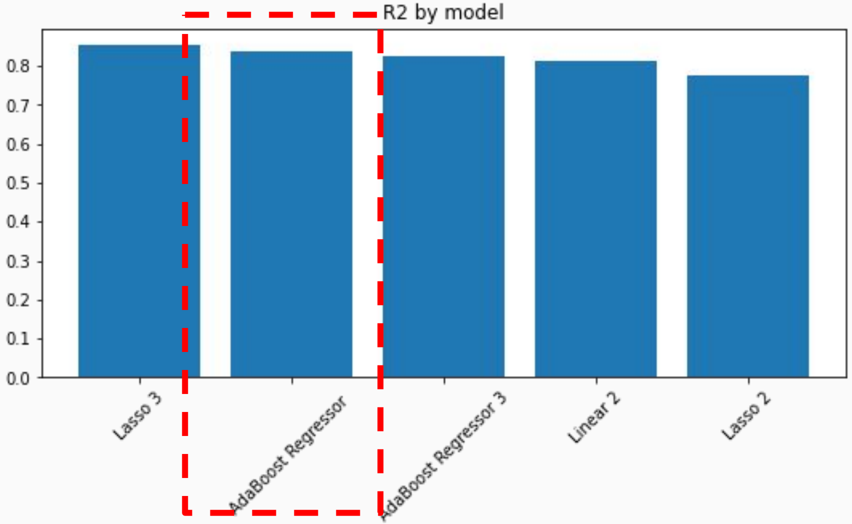


Hôpitaux ? 0.8% des bâtiments / 13% KBTU

Scores : GHG

	Model	Degree	RMSE	R2
0	Lasso	3	336.745104	0.852384
1	AdaBoost Regressor		351.043319	0.839582
2	AdaBoost Regressor	3	367.773840	0.823927
3	Linear	2	379.832834	0.812191
4	Lasso	2	415.345030	0.775431

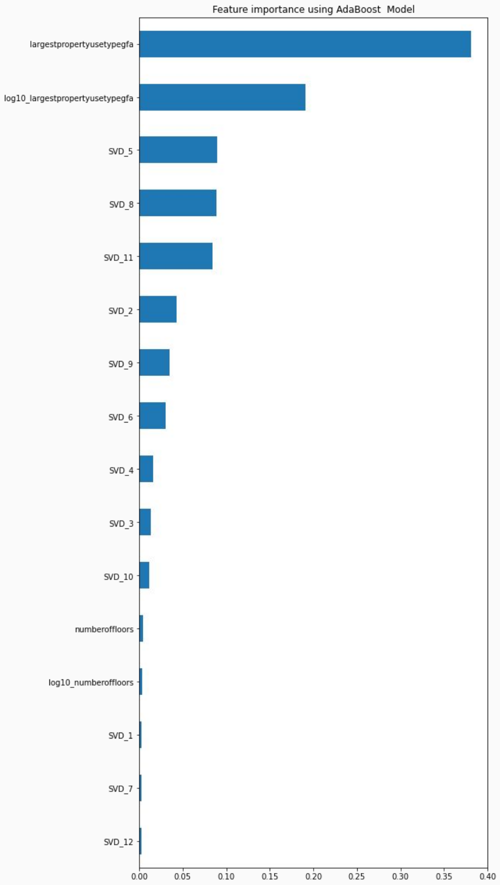
```
R2 X_test: 0.85
R2 X_train: 0.95
```



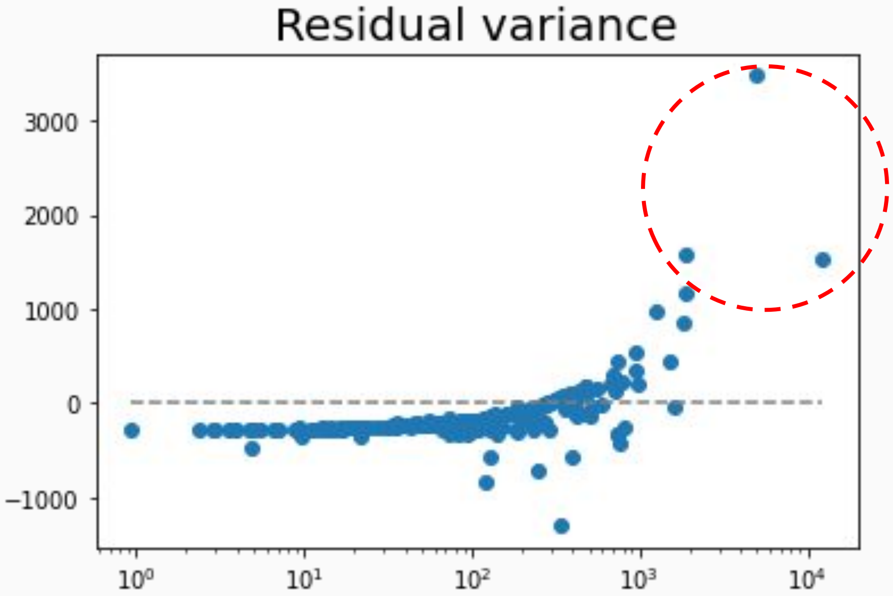
AdaBoost regressor

```
R2 X_test: 0.82
R2 X_train: 0.85
```

Scores : GHG



AdaBoost a sélectionné 16 variables et rejeté 0



Hôpital ?

ENERGY STAR score

ENERGY STAR score

191 valeurs manquantes -> 555 obs.

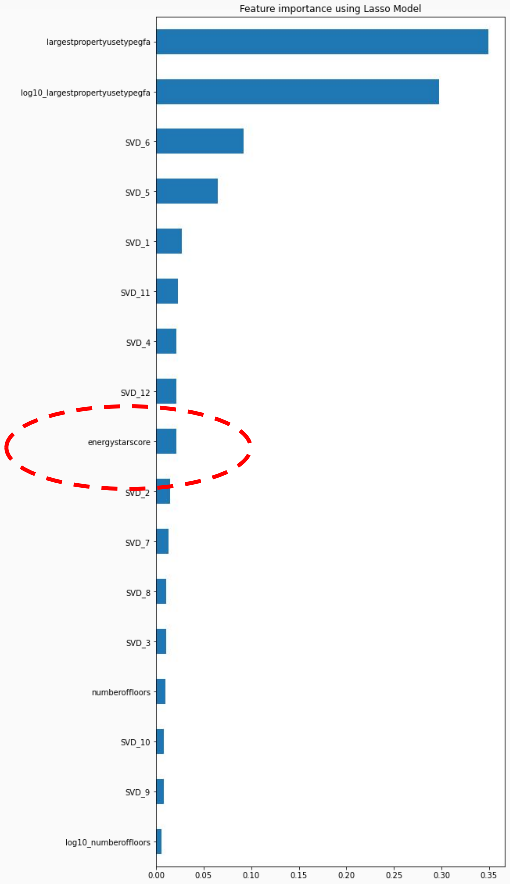
AdaBoost

	SANS Energy Star Score	Energy Star Score
R2	0.82	0.81
RMSE	413	435

ENERGY STAR score:

- 5% des variables retenues
- 2 % du poids

Intérêt relativement limité de l'Energy Star Score.
Évaluation de son ROI.



CONCLUSION

Seattle

Données bâtiments non destinés à l’habitation

Consommation électrique (ktbu)

Ridge polynomial 2

R2 = 0.74

Émissions carbone (ghg)

AdaBoost regressor

R2 = 0.84

ENERGY STAR

AdaBoost regressor

MERCI !

@xavbarbier

https://www.linkedin.com/in/barbierxavier/

https://github.com/xavierbarbier/

contact@xavierbarbier.com