

Anticiper les besoins en  
consommation énergétique  
des bâtiments

# Problématique

## Consommation d'énergie et émissions de CO2



Pour atteindre son objectif de **ville neutre en émissions de carbone en 2050**, la ville de Seattle s'intéresse de près à la **consommation** et aux **émissions** des bâtiments non destinés à l'habitation et nous sollicite afin de :

- *prédir la consommation totale d'énergie et les émissions de CO2* de bâtiments non destinés à l'habitation pour lesquelles elles n'ont pas été mesurées (un relevé de référence est effectué la première année)
- juger de *l'intérêt du calcul de l' ENERGY STAR Score* pour la prédiction des émissions

N.B : Un ENERGY STAR Score > 75 permet d'obtenir un label environnemental, l'Energy Star Certification, attribuée par l'U.S Environmental Protection Agency via ENERGY STAR  
(Source : [ENERGY STAR Certification for Buildings | ENERGY STAR](#))

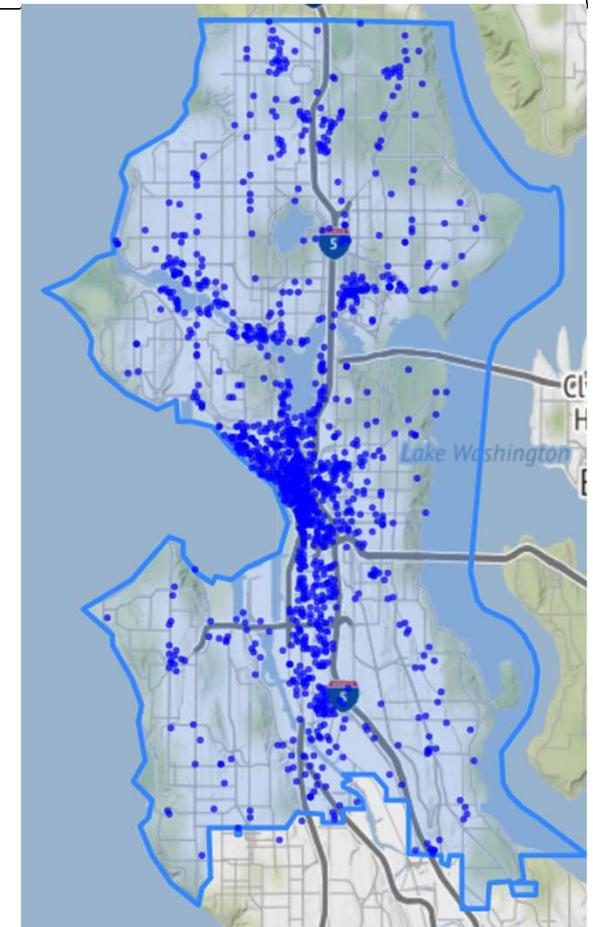
# Les données à disposition

Relevées sur l'année 2016

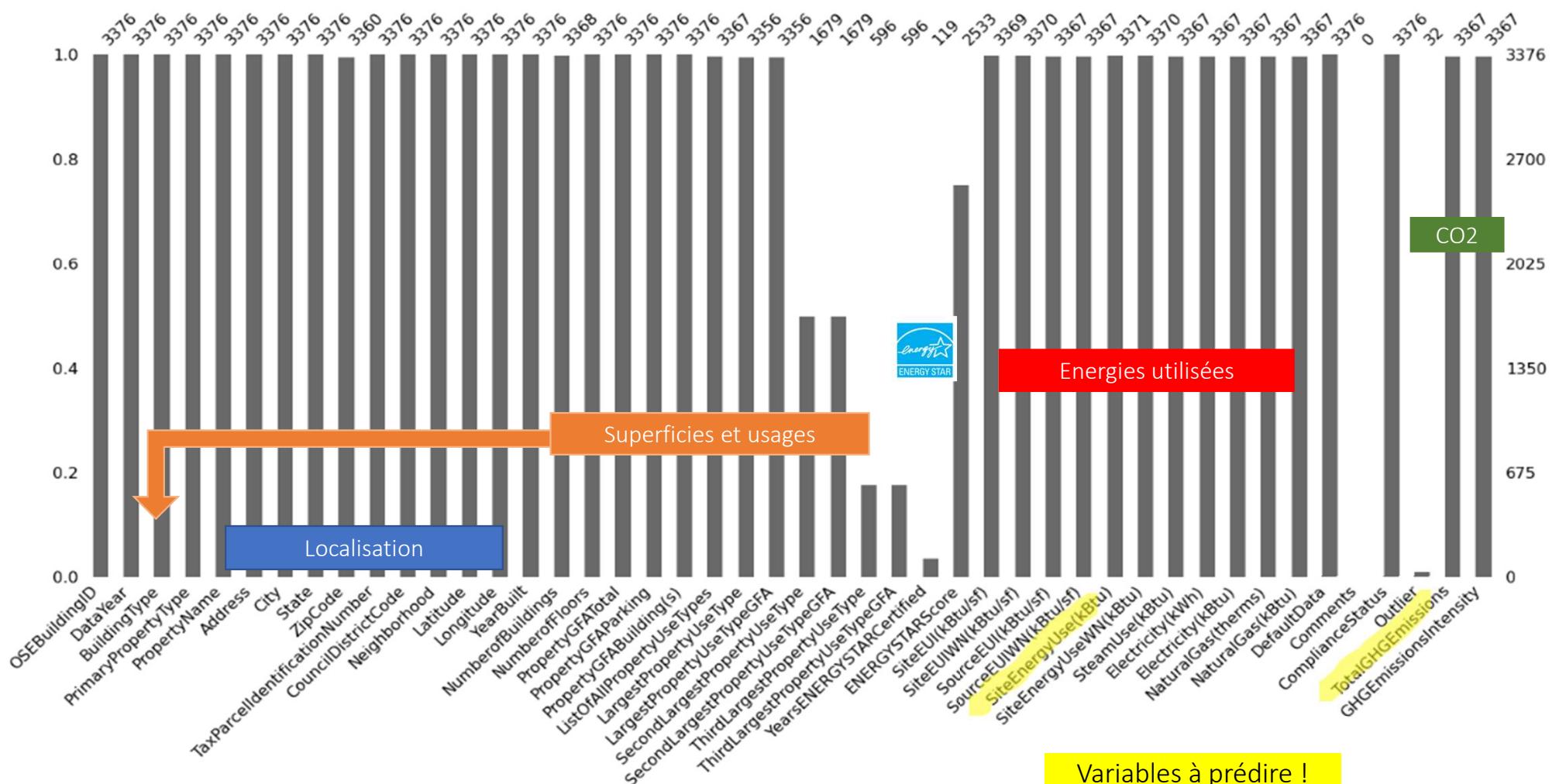


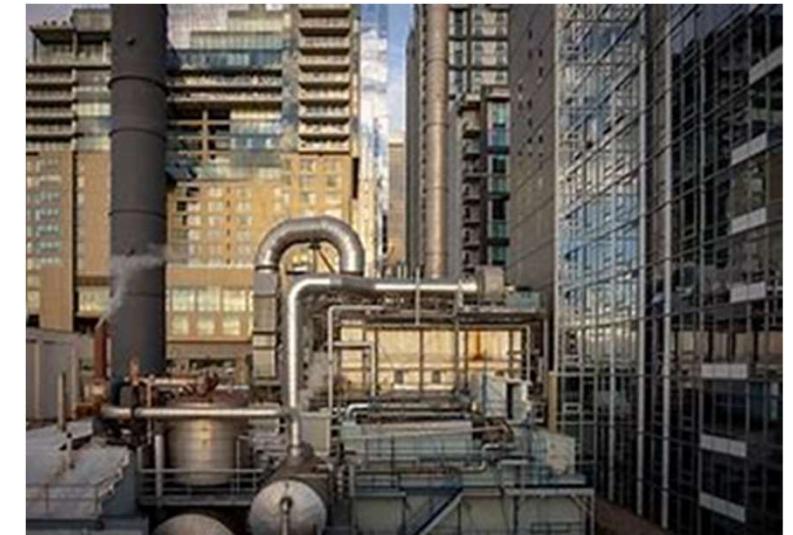
[Source : 2016 Building Energy Benchmarking | City of Seattle Open Data portal](#)

- Localisation (adresse, coordonnées GPS)
- Ancienneté
- Type et usages des bâtiments
- Structure des bâtiments (nombre de niveaux, superficies totales et par usages)
- Consommation d'énergie (différentes sources)
- Energy Star Score lorsqu'il a été attribué
- Emissions de GES (eq. CO<sub>2</sub>)



# Composition du jeu de données





Data Cleaning et Feature  
Engineering

# Premier filtrage des données

## Nettoyage et sélection



### Conformité

Compliant	1546
Error - Correct Default Data	88
Non-Compliant	17
Missing Data	14
Name: ComplianceStatus, dtype: int64	

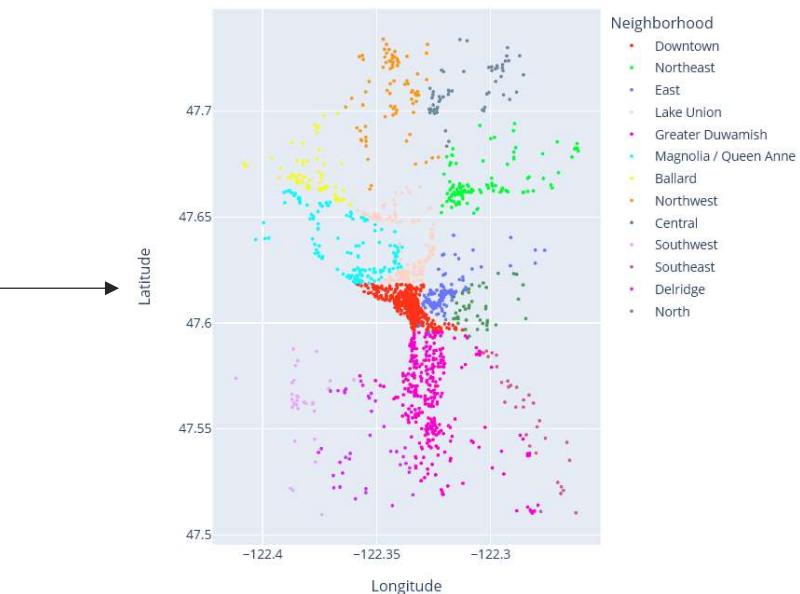
### Sélection des BuildingType (non destinés à l'habitation)

NonResidential	1460
Multifamily LR (1-4)	1018
Multifamily MR (5-9)	580
Multifamily HR (10+)	110
SPS-District K-12	98
Nonresidential COS	85
Campus	24
Nonresidential WA	1
Name: BuildingType, dtype: int64	

La seule information de localisation conservée pour l'apprentissage sera celle du quartier « Neighborhood), dont on corrige les valeurs mal orthographiées ou fausses.

DOWNTOWN	350
GREATER DUWAMISH	328
MAGNOLIA / QUEEN ANNE	144
LAKE UNION	142
NORTHEAST	117
EAST	116
<b>NORTHWEST</b>	74
<b>BALLARD</b>	58
<b>NORTH</b>	51
<b>CENTRAL</b>	42
DELRIDGE	36
SOUTHWEST	31
SOUTHEAST	31
North	8
<b>Ballard</b>	5
<b>Delridge</b>	4
<b>Northwest</b>	4
<b>Central</b>	4
<b>DELRIDGE NEIGHBORHOODS</b>	1
Name: Neighborhood, dtype: int64	

Les différents quartiers (Neighborhood)



# Sélection des features numériques

Suppression d'outliers (faux ou non)...ou correction 😊



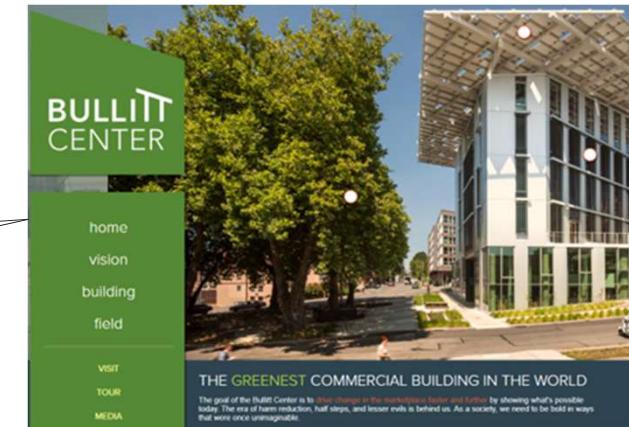
	count	mean	std	min	25%	50%	75%	max
<b>OSEBuildingID</b>	1546.0	1.649282e+04	1.383604e+04	1.00000	6.022500e+02	2.118050e+04	2.461900e+04	5.022600e+04
<b>CouncilDistrictCode</b>	1546.0	4.413972e+00	2.191933e+00	1.00000	2.000000e+00	4.000000e+00	7.000000e+00	7.000000e+00
<b>Latitude</b>	1546.0	4.761647e+01	4.697675e-02	47.50959	4.758783e+01	4.761271e+01	4.764908e+01	4.773387e+01
<b>Longitude</b>	1546.0	-1.223335e+02	2.327804e-02	-122.41182	-1.223431e+02	-1.223332e+02	-1.223227e+02	-1.222618e+02
<b>YearBuilt</b>	1546.0	1.961632e+03	3.288993e+01	1900.00000	1.930000e+03	1.965500e+03	1.989000e+03	2.015000e+03
<b>NumberofBuildings</b>	1546.0	1.179172e+00	3.041181e+00	0.00000	1.000000e+00	1.000000e+00	1.000000e+00	1.100000e+02
<b>NumberofFloors</b>	1546.0	4.277490e+00	6.784999e+00	0.00000	1.000000e+00	2.000000e+00	4.000000e+00	9.900000e+01
<b>PropertyGFATotal</b>	1546.0	1.213994e+05	3.064857e+05	11285.00000	2.884175e+04	4.815900e+04	1.078405e+05	9.320156e+06
<b>PropertyGFAParking</b>	1546.0	1.383690e+04	4.374503e+04	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	5.126080e+05
<b>PropertyGFABuilding(s)</b>	1546.0	1.075625e+05	2.928095e+05	3636.00000	2.793425e+04	4.608400e+04	9.585275e+04	9.320156e+06
<b>LargestPropertyUseTypeGFA</b>	1542.0	1.000015e+05	2.856797e+05	5656.00000	2.503975e+04	4.204300e+04	9.209825e+04	9.320156e+06
<b>SecondLargestPropertyUseTypeGFA</b>	841.0	3.637328e+04	6.666481e+04	0.00000	5.561000e+03	1.210200e+04	3.184500e+04	6.399310e+05
<b>ThirdLargestPropertyUseTypeGFA</b>	347.0	1.503154e+04	3.708548e+04	0.00000	2.613500e+03	6.000000e+03	1.299000e+04	4.597480e+05
<b>ENERGYSTARScore</b>	996.0	6.366767e+01	2.882233e+01	1.00000	4.400000e+01	7.100000e+01	8.800000e+01	1.000000e+02
<b>SiteEUI(kBtu/sf)</b>	1546.0	7.511197e+01	7.524890e+01	1.40000	3.490000e+01	5.375000e+01	8.517500e+01	8.344000e+02
<b>SiteEUIWN(kBtu/sf)</b>	1545.0	7.755618e+01	7.627154e+01	0.00000	3.700000e+01	5.630000e+01	8.790000e+01	8.344000e+02
<b>SourceEUI(kBtu/sf)</b>	1546.0	1.832398e+02	1.880506e+02	0.00000	1.250000e+01	1.386500e+02	2.133500e+02	2.620000e+03
<b>SourceEUIWN(kBtu/sf)</b>	1546.0	1.853849e+02	1.880808e+02	-2.10000	8.432500e+01	1.418000e+02	2.156750e+02	2.620000e+03
<b>SiteEnergyUse(kBtu)</b>	1546.0	8.867785e+06	3.132518e+07	57133.19922	1.248602e+06	2.732167e+06	7.301812e+06	8.739237e+08
<b>SiteEnergyUseWN(kBtu)</b>	1545.0	8.449111e+06	2.279062e+07	0.00000	1.321763e+06	2.824097e+06	7.483350e+06	4.716139e+08
<b>SteamUse(kBtu)</b>	1546.0	5.518961e+05	5.722168e+06	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.349435e+08
<b>Electricity(kWh)</b>	1546.0	1.801059e+06	6.308399e+06	-33826.80078	2.139610e+05	5.109501e+05	1.543338e+06	1.925775e+08
<b>Electricity(kBtu)</b>	1546.0	6.145212e+06	2.152426e+07	-115417.00000	7.300348e+05	1.743362e+06	5.265868e+06	6.570744e+08
<b>NaturalGas(therms)</b>	1546.0	2.040770e+04	9.716705e+04	0.00000	0.000000e+00	4.906580e+03	1.531061e+04	2.979090e+06
<b>NaturalGas(kBtu)</b>	1546.0	2.040770e+06	9.716705e+06	0.00000	0.000000e+00	4.906580e+05	1.531061e+06	2.979090e+08
<b>TotalGHGEmissions</b>	1546.0	1.938258e+02	7.795860e+02	-0.80000	2.064500e+01	4.994000e+01	1.474025e+02	1.687098e+04
<b>GHGEmissionsIntensity</b>	1546.0	1.668959e+00	2.408717e+00	-0.02000	3.600000e-01	8.850000e-01	1.960000e+00	3.409000e+01

?

Erreur



Outlier

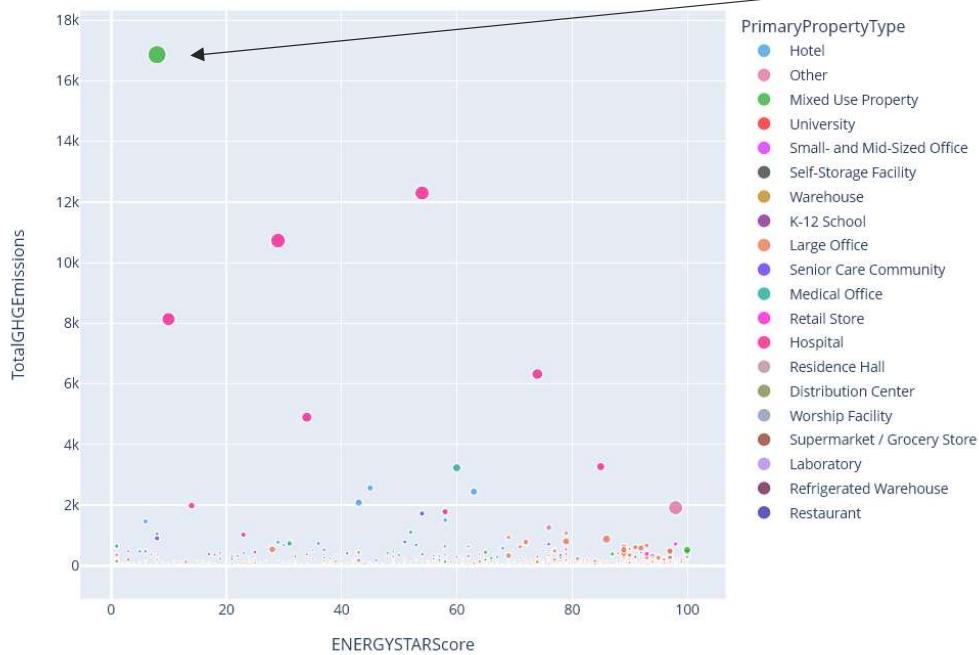


# ENERGY STAR Score

## consommation, émissions et usage principal



Taille du cercle = SiteEnergyUse(kBtu)



Outlier

In US, Boeing tearing down Plant 2, factory where Seattle became a high-tech town

Seattle.com Press

You can now listen to Fox News articles  
SEATTLE, Washington – The dilapidated factory that helped make Seattle a high-tech town is being demolished after 75 years, a casualty of time, technology and tails that grew too tall.



“Your building is *not* compared to the other buildings in Portfolio Manager to determine your ENERGY STAR score.

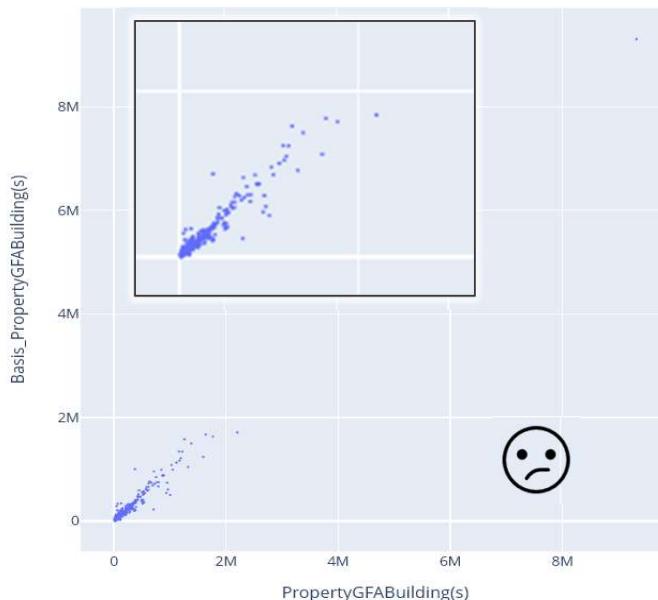
Instead, your building is compared to **other buildings nationwide that have the same primary use.**” (documentation)

# Vérification de la cohérence

Et « mise à jour » des données



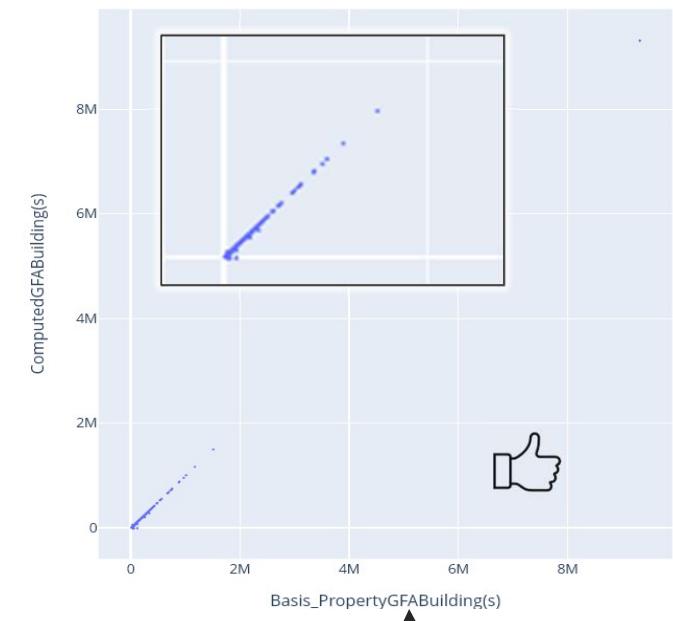
D'après les définitions des données, la variable `PropertyGFABuilding(s)` est renseignée à la création de la propriété, mais ne change pas si les superficies par usages sont mises à jour. On vérifie donc sa pertinence au regard des autres données.



Somme des  
PropertyUseTypeGFA  
(hors parking)

C'est cette valeur qui sera conservée pour l'apprentissage

GFA prise en compte pour le calcul du SiteEUI (kBtu/sf)

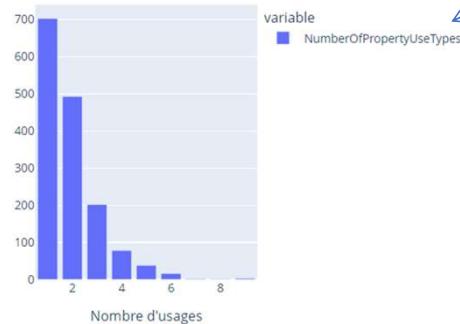


# Création de nouvelles colonnes qui permettront d'exploiter plus d'informations

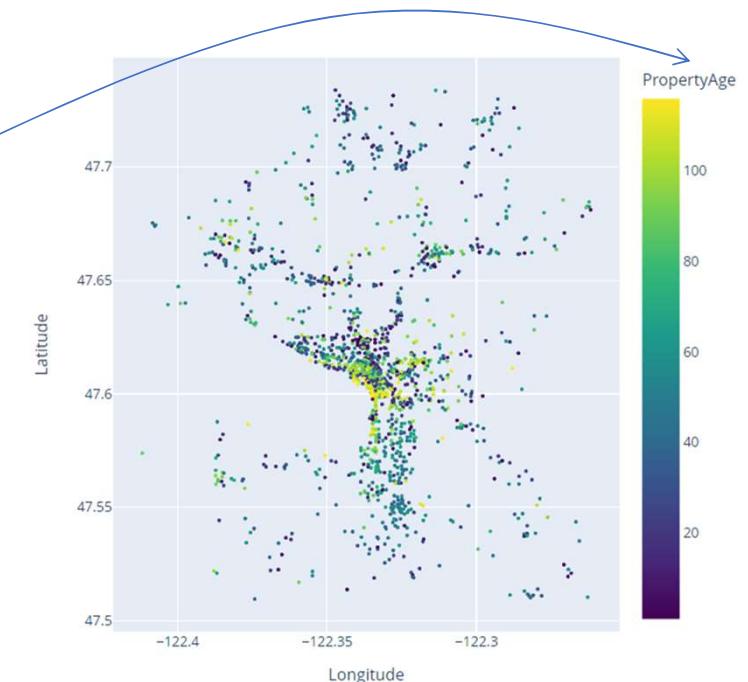


Nombre d'usages différents, à partir des listes d'usages

```
0           Hotel, Parking, Restaurant  
1           Hotel  
2           Hotel  
3           Hotel  
4           Hotel, Parking, Swimming Pool  
1541          ...  
1542          Other - Recreation  
1543 Fitness Center/Health Club/Gym, Other - Recrea...  
1544 Fitness Center/Health Club/Gym, Food Service, ...  
1545 Fitness Center/Health Club/Gym, Food Service, ...  
Name: ListOfAllPropertyUseTypes, Length: 1539, dtype: object
```



Ancienneté des bâtiments (à partir de la date de construction)



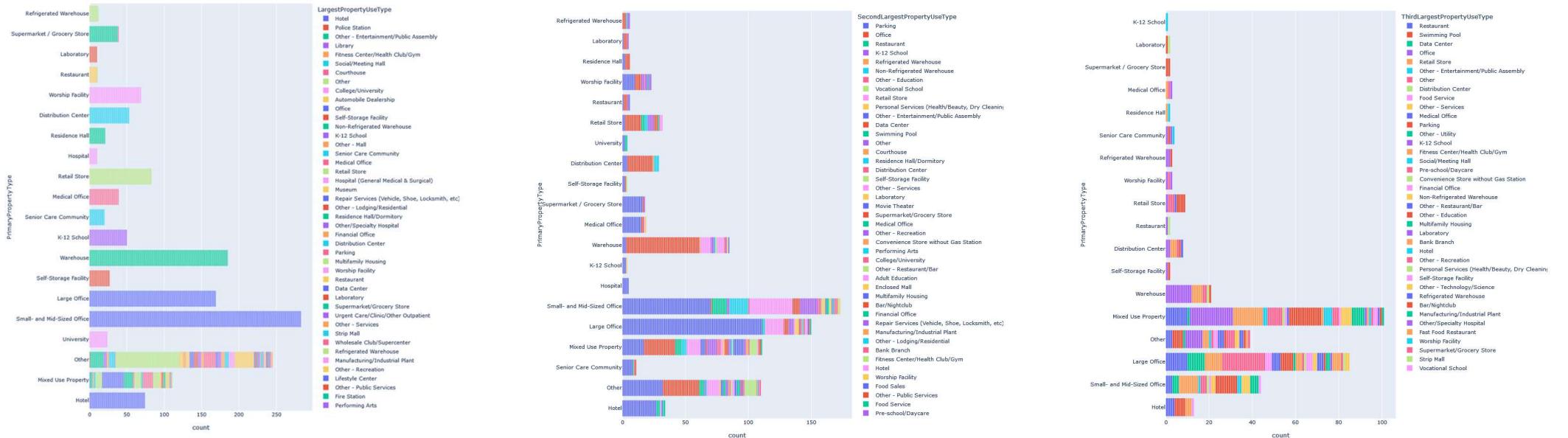
Calcul des ratios d'utilisation des différentes sources d'énergie (permettant d'exploiter un relevé de référence sur une durée inconnue)

	count	mean	std	min	25%	50%	75%	max
<b>ElectricityRatio</b>	1539.0	0.701338	0.262487	0.0	0.491303	0.709146	0.999988	1.000003
<b>NaturalGasRatio</b>	1539.0	0.274718	0.264094	0.0	0.000000	0.233243	0.490002	1.000000
<b>SteamUseRatio</b>	1539.0	0.022642	0.094307	0.0	0.000000	0.000000	0.000000	0.766987

# Prise en compte des usages



LargestPropertyUseType, SecondaryPropertyUseType, ThirdLargestPropertyUseType

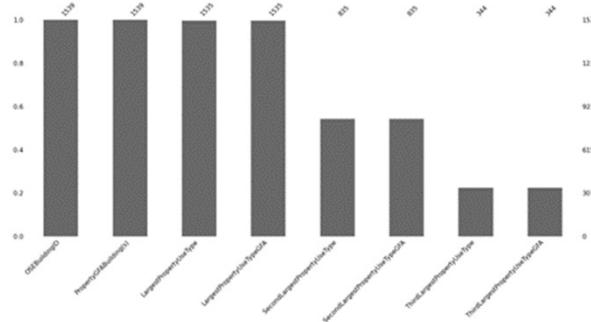


On va affecter à chaque usage un poids qui sera sa superficie dans la propriété documentée (GFA)

- Les valeurs NaN sont mises à zéro.
  - Conservation de l'usage « parking » (en lieu et place de GFAParking)
  - Regroupement de certains usages (par type) pour éviter des ensembles avec trop peu d'occurrences.

# Transformation des features

## Superficies et Usages



	PrimaryPropertyType	PropertyUseType	PropertyUseTypeGFA
OSEBuildingID			
1	Hotel	Hotel	88434.0
2	Hotel	Hotel	83880.0
3	Hotel	Hotel	756493.0
5	Hotel	Hotel	61320.0
8	Hotel	Hotel	123445.0
...	...	...	...
50208	Other	Office	187.0
50219	Mixed Use Property	Office	3779.0
50224	Other	Swimming Pool	0.0
50225	Mixed Use Property	Pre-school/Daycare	484.0
50226	Mixed Use Property	Pre-school/Daycare	1108.0

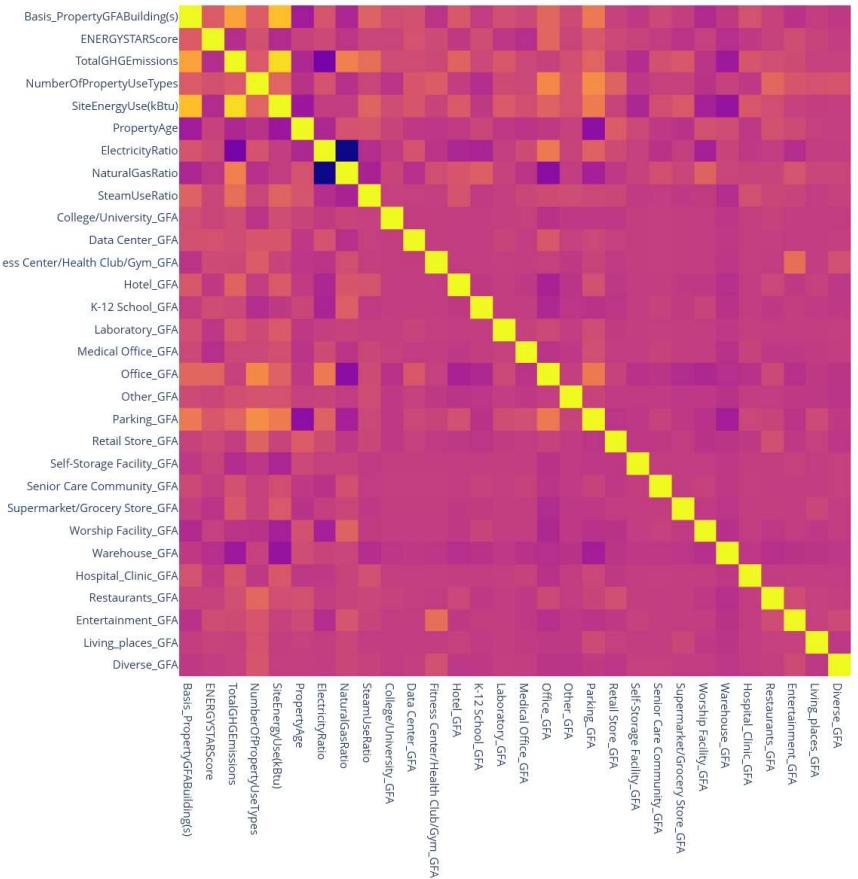
2717 rows × 3 columns



	count	mean	std	min	25%	50%	75%	max
College/University_GFA	25.0	584192.840000	1.855446e+06	21600.0	52611.00	84300.0	203030.00	9320156.0
Data Center_GFA	28.0	21116.107143	4.501199e+04	182.0	575.50	4244.0	12178.50	218997.0
Fitness Center/Health Club/Gym_GFA	25.0	16212.280000	1.975757e+04	1670.0	7501.00	8508.0	14081.00	90000.0
Hotel_GFA	79.0	136710.645570	1.613985e+05	8000.0	52029.00	83880.0	134445.00	994212.0
K-12 School_GFA	57.0	57973.280702	6.608695e+04	3231.0	22860.00	40809.0	55500.00	367884.0
Laboratory_GFA	22.0	76956.909091	5.932480e+04	5000.0	29956.25	61940.0	101692.50	181930.0
Medical Office_GFA	58.0	81941.515490	9.649475e+04	1025.0	19928.50	39545.5	128408.25	520187.0
Office_GFA	709.0	90522.113681	1.721195e+05	187.0	14256.00	32250.0	87925.00	1680937.0
Other_GFA	176.0	51552.937500	8.125734e+04	300.0	9450.00	25321.5	49373.25	535947.0
Parking_GFA	320.0	73659.433740	8.739436e+04	508.0	15000.00	35862.5	98247.50	441551.0
Retail Store_GFA	235.0	40795.661277	7.491753e+04	294.0	8107.00	16260.0	37837.50	561684.0
Self-Storage Facility_GFA	32.0	41162.281250	3.025957e+04	1096.0	22366.50	29894.5	50798.75	130293.0
Senior Care Community_GFA	20.0	118010.000000	1.599853e+05	22583.0	43226.25	57332.5	105372.00	726000.0
Supermarket/Grocery Store_GFA	50.0	42067.460000	2.726152e+04	2000.0	25956.25	39531.0	49678.75	168735.0
Worship Facility_GFA	73.0	31349.657534	1.626329e+04	4500.0	22000.00	26210.0	35844.00	103000.0
Warehouse_GFA	304.0	51191.585526	7.090867e+04	2408.0	21317.00	32373.0	53912.50	892000.0
Hospital_Clinic_GFA	20.0	373538.800000	4.796521e+05	34074.0	50966.00	134232.0	513043.75	1639334.0
Restaurants_GFA	91.0	12143.912088	1.344777e+04	436.0	3690.00	7044.0	15552.50	80000.0
Entertainment_GFA	87.0	70785.896552	2.175662e+05	1000.0	10740.50	20411.0	35623.50	1585960.0
Living_places_GFA	63.0	52097.079365	7.104382e+04	1248.0	11984.00	24338.0	45891.50	340236.0
Diverse_GFA	102.0	39839.529412	5.438318e+04	400.0	12201.50	24627.0	42634.00	364913.0

# Jeu de données final

Et heatmap de corrélation (coefficient de Spearman)



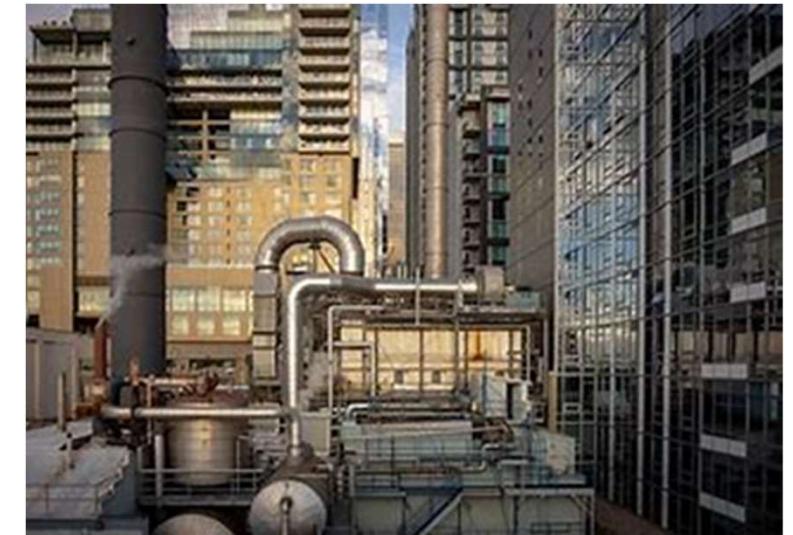
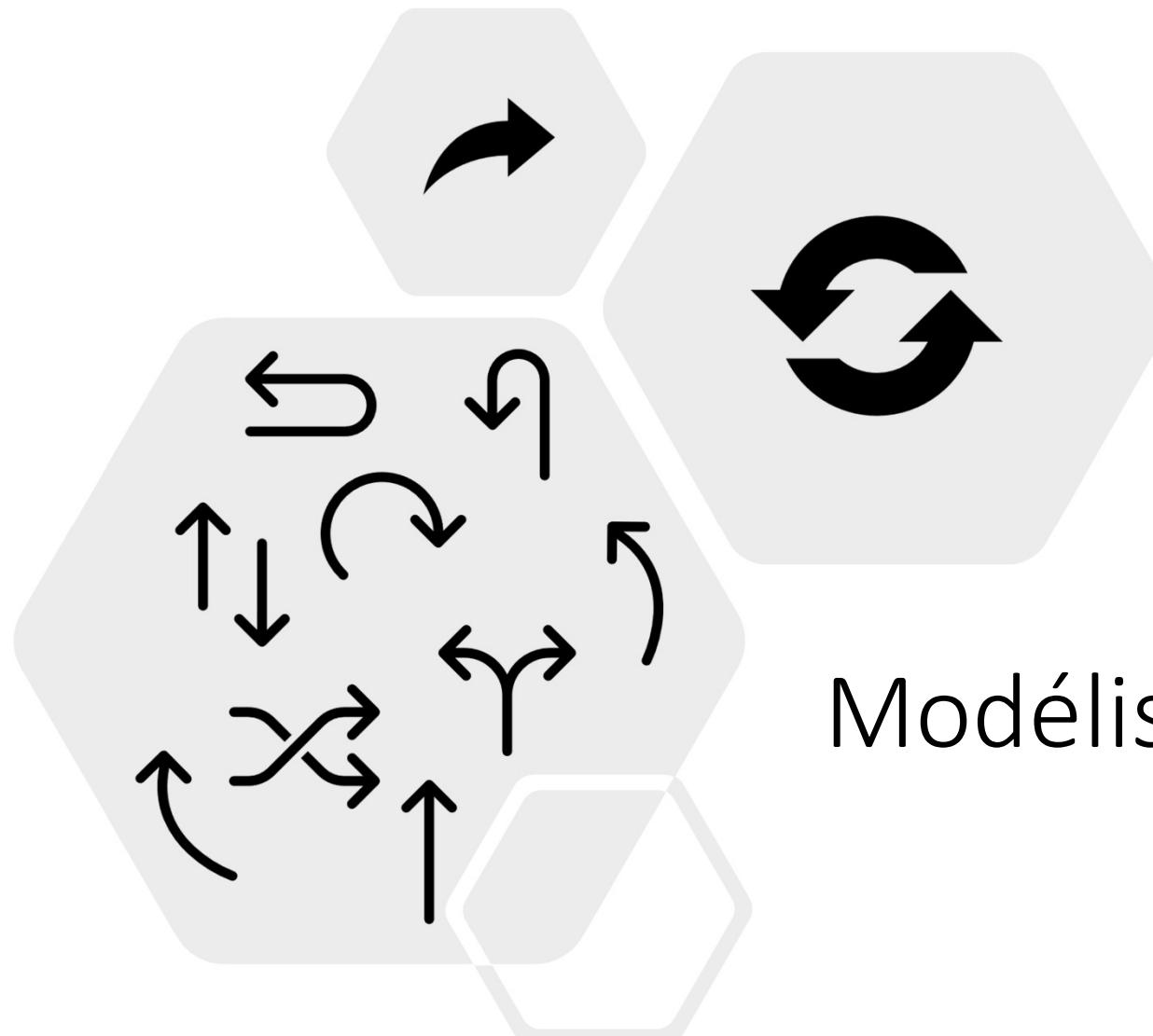
Les variables ne sont pas linéairement dépendantes, excepté qu'on observe le lien fort entre le ratio d'usage de l'Electricité et celui du Gaz.

Toutefois, j'ai préféré conserver malgré tout les deux valeurs.

On peut aussi observer quelques niveaux de corrélations évocatrices (ElectricityRatio et Emissions, par exemple)

#	Column	Non-Null Count	Dtype
0	PropertyName	1538 non-null	object
1	Neighborhood	1538 non-null	object
2	Basis_PropertyGFABuilding(s)	1538 non-null	float64
3	ENERGYSTARScore	990 non-null	float64
4	TotalGHGEmissions	1538 non-null	float64
5	NumberOfPropertyUseTypes	1538 non-null	int64
6	SiteEnergyUse(kBtu)	1538 non-null	float64
7	PropertyAge	1538 non-null	int64
8	ElectricityRatio	1538 non-null	float64
9	NaturalGasRatio	1538 non-null	float64
10	SteamUseRatio	1538 non-null	float64
11	College/University_GFA	1538 non-null	float64
12	Data Center_GFA	1538 non-null	float64
13	Fitness Center/Health Club/Gym_GFA	1538 non-null	float64
14	Hotel_GFA	1538 non-null	float64
15	K-12 School_GFA	1538 non-null	float64
16	Laboratory_GFA	1538 non-null	float64
17	Medical Office_GFA	1538 non-null	float64
18	Office_GFA	1538 non-null	float64
19	Other_GFA	1538 non-null	float64
20	Parking_GFA	1538 non-null	float64
21	Retail Store_GFA	1538 non-null	float64
22	Self-Storage Facility_GFA	1538 non-null	float64
23	Senior Care Community_GFA	1538 non-null	float64
24	Supermarket/Grocery Store_GFA	1538 non-null	float64
25	Worship Facility_GFA	1538 non-null	float64
26	Warehouse_GFA	1538 non-null	float64
27	Hospital_Clinic_GFA	1538 non-null	float64
28	Restaurants_GFA	1538 non-null	float64
29	Entertainment_GFA	1538 non-null	float64
30	Living_places_GFA	1538 non-null	float64
31	Diverse_GFA	1538 non-null	float64

dtypes: float64(28), int64(2), object(2)  
memory usage: 396.5+ KB



Modélisation et résultats  
des prédictions

# Approche de modélisation

Pour la prédiction de SiteEnergyUse(kBtu) et TotalGHGEmissions



- ✓ Critère d'évaluation : R-squared (**R2**)
- ✓ Split des données 80% - 20%
- ✓ Ensemble d'apprentissage et de tests fixés à l'aide d'un random\_state pour stabiliser l'analyse
- ✓ 5 folds pour la cross-validation (valeur adaptée à la limitation des ressources de calcul)
- ✓ Recherche par grille pour l'optimisation des hyper-paramètres
  - ✓ Grid Search Cross-Validation pour les modèles « rapides »
  - ✓ Bayesian Search Cross-Validation pour les régresseurs les plus coûteux en temps d'apprentissage
- ✓ Test de plusieurs régresseurs pour les comparer

```
Dummy()  
KNeighborsRegressor()  
LinearRegression()  
Ridge()  
Lasso()  
SVR(kernel='linear')  
DecisionTreeRegressor()  
GradientBoostingRegressor()  
xgboost.XGBRegressor()  
RandomForestRegressor()
```

# Préprocessing des données



Encodage des variables catégorielles : ici « Neighborhood » uniquement.

- Utilisation de "get\_dummies" (OneHotEncoder peut être plus adapté dans un contexte dynamique)

	Neighborhood	Basis_PropertyGFABuilding(s)	NumberOfPropertyUseTypes	ENERGYScores	Total
0	Downtown	88449.972702	1	60.0	
1	Downtown	88480.303060	3	61.0	
2	Downtown	756114.833333	1	43.0	
3	Downtown	61322.958573	1	56.0	
4	Downtown	123454.752764	3	75.0	

5 rows × 31 columns



	Neighborhood_Downtown	Neighborhood_East	Neighborhood_Greater Duwamish	Neighborhood_Lake Union	Neighborhood_Magnolia / Queen Anne
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

Feature Scaling : Les valeurs des différentes variables s'étendant sur des intervalles différents, il est important d'effectuer une mise à l'échelle.

- Utilisation du StandardScaler

# Temps d'apprentissage

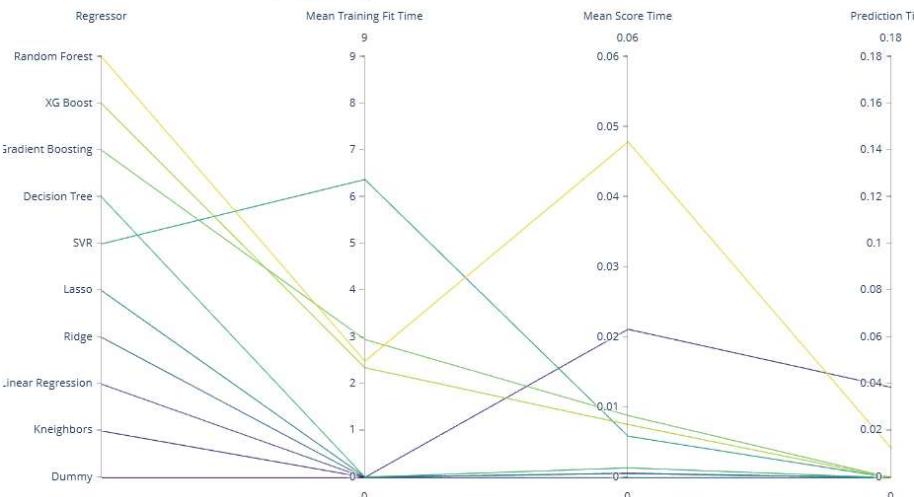
Comparaison des modèles pour les deux régressions



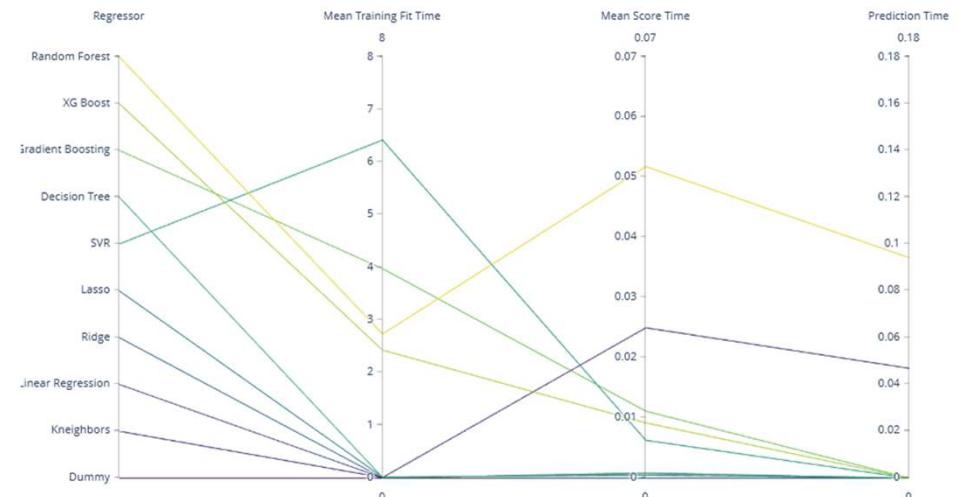
La régression SVR linéaire est la plus coûteuse en temps pour l'apprentissage, même au-delà des modèles basés sur des arbres de décisions (eux-mêmes coûteux).

La prise en compte de ce critère dépend des ressources dont on dispose, et en particulier du nombre de fois où on doit relancer la procédure d'apprentissage au cours de la vie de l'algorithme... (si l'ensemble de données évolue).

Durée des différentes étapes, par régresseur (Consommation)



Durée des différentes étapes, par régresseur (Emissions)

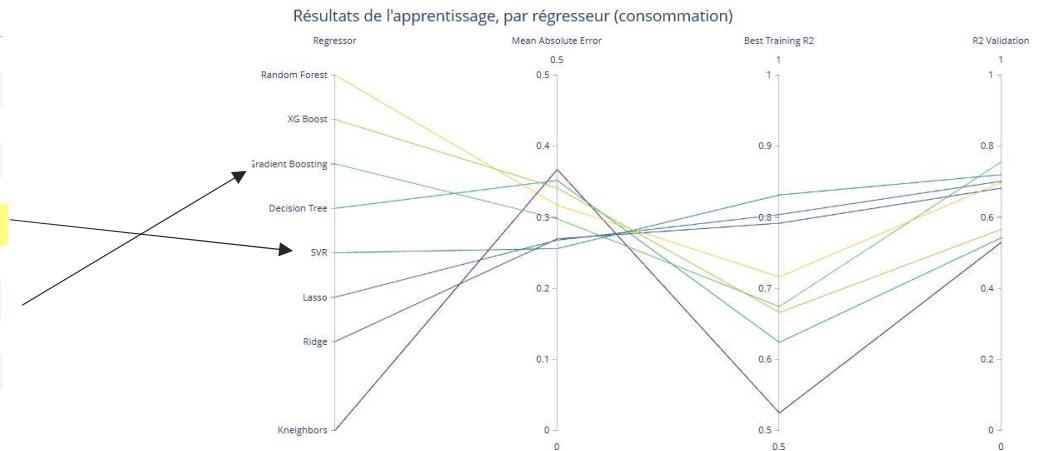


# Prédiction de la consommation d'énergie

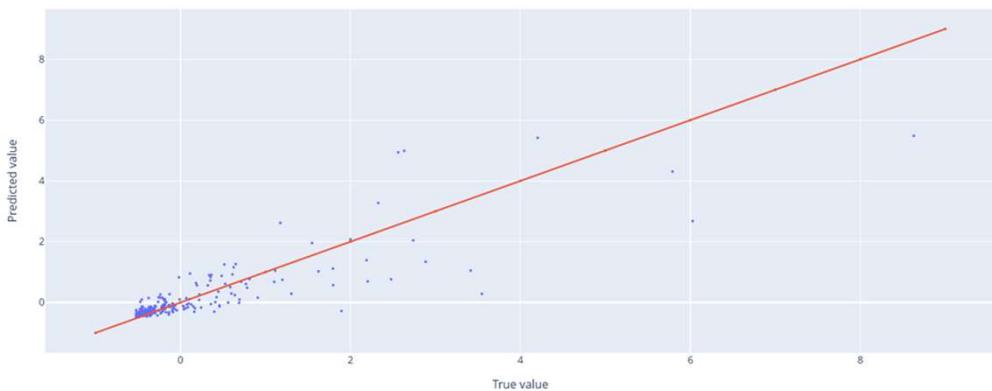
## Comparaison des modèles



Regressor	Hyper params	MSE	MAE	Test Validation R2	Training Best R2
Dummy	None	2.366421e+01	4.115270e+00	-2.266421e+01	-22.664208
Kneighbors	{'n_neighbors': 4}	4.703109e-01	3.672433e-01	5.296891e-01	0.524462
Linear Regression	{'fit_intercept': False}	1.302566e+21	2.157069e+10	-1.302566e+21	0.777185
Ridge	{'alpha': 10.0}	3.181892e-01	2.695954e-01	6.818108e-01	0.791572
Lasso	{'alpha': 0.01}	2.985907e-01	2.677126e-01	7.014093e-01	0.803704
<b>SVR</b>	{'C': 10, 'epsilon': 0.1}	2.803182e-01	2.558743e-01	<b>7.196818e-01</b>	<b>0.831346</b>
Decision Tree	{'ccp_alpha': 4.726931480329788e-06, 'min_impu...	4.579259e-01	3.517051e-01	5.420741e-01	0.623818
Gradient Boosting	{'learning_rate': 0.24543638357341102, 'max_de...	2.437814e-01	2.977956e-01	<b>7.562186e-01</b>	<b>0.674247</b>
XG Boost	{'learning_rate': 0.1701666821364574, 'max_dep...	4.340395e-01	3.409520e-01	5.659605e-01	0.665928
Random Forest	{'max_features': 'auto', 'min_impurity_decreas...	3.016903e-01	3.168865e-01	6.983097e-01	0.716139



Consommation : SVR - R2 = 0.7197



Mon choix s'est porté ici sur **SVR** (Support Vector Regression), malgré la durée de l'apprentissage, dont le meilleur score d'apprentissage est bien supérieur à celui du Gradient Boosting (dont le score de validation est supérieur).

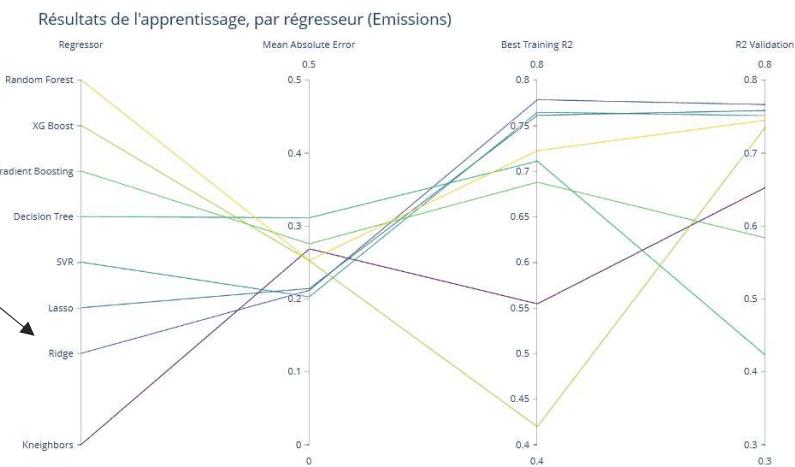
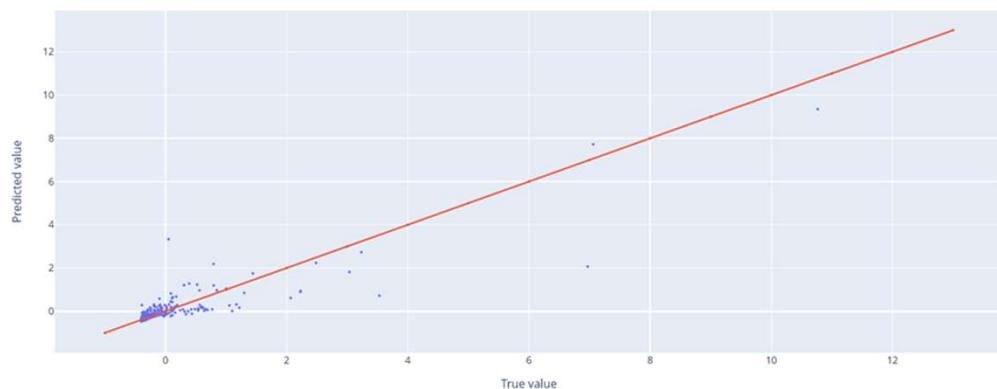
# Prédiction des émissions de CO2

## Comparaison des modèles



Regressor	Hyper params	MSE	MAE	Test Validation R2	Training Best R2
Dummy	None	3.706419e+01	5.095837e+00	-3.606419e+01	-36.064194
Kneighbors	{'n_neighbors': 4}	3.475589e-01	2.682403e-01	6.524411e-01	0.554247
Linear Regression	{'fit_intercept': False}	6.475457e+21	4.809494e+10	-6.475457e+21	0.679179
Ridge	{'alpha': 100.0}	2.337730e-01	2.113342e-01	7.662270e-01	0.778355
Lasso	{'alpha': 0.01}	2.419180e-01	2.138970e-01	7.580820e-01	0.761187
SVR	{'C': 1, 'epsilon': 0.1}	2.489699e-01	2.024703e-01	7.510301e-01	0.764574
Decision Tree	{'ccp_alpha': 0.002236420282054271, 'min_impurity_decreas...}	5.776313e-01	3.109167e-01	4.223687e-01	0.710846
Gradient Boosting	{'learning_rate': 0.1990044301307691, 'max_depth': 6, 'min_samples_leaf': 1, 'n_estimators': 100, 'random_state': 42}	4.167189e-01	2.751406e-01	5.832811e-01	0.687805
XG Boost	{'learning_rate': 0.4999999999999999, 'max_depth': 6, 'min_samples_leaf': 1, 'n_estimators': 100, 'random_state': 42}	2.646997e-01	2.520589e-01	7.353003e-01	0.419395
Random Forest	{'max_features': 'auto', 'min_impurity_decreas...}	2.552463e-01	2.523337e-01	7.447537e-01	0.722324

Emissions | Ridge - R2 = 0.7662



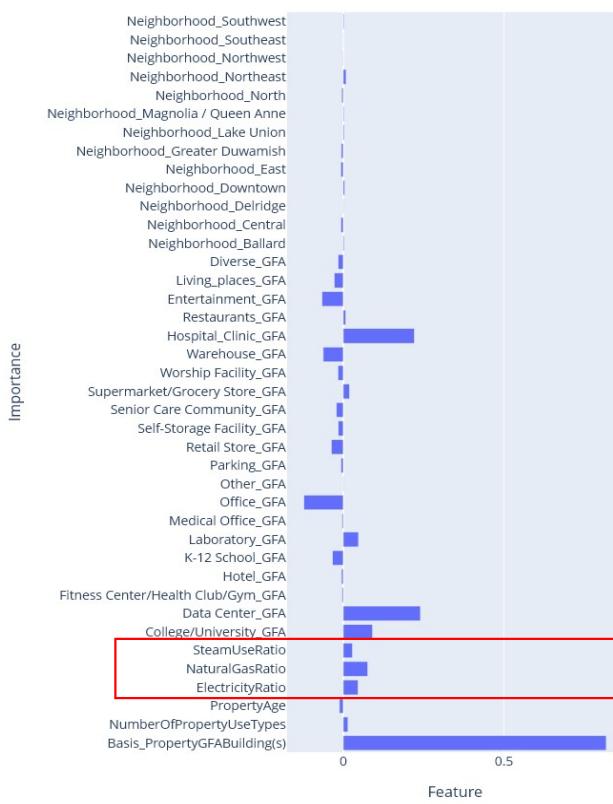
Mon choix s'est porté ici sur la **régression Ridge**, dont les scores sont les meilleurs.

# Importance des features dans les modèles

## Consommation totale d'énergie et émissions de CO2



Visualisation des coefficients de chaque feature (Consommation)



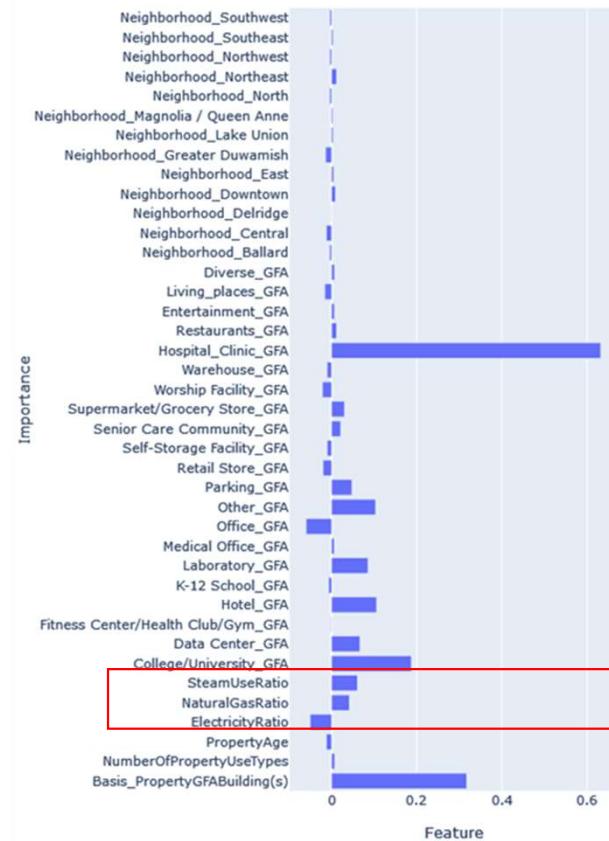
La localisation a peu voire pas d'influence sur la consommation d'énergie, ni sur les émissions de CO2.

La surface totale de la propriété est la variable majeure pour la consommation mais secondaire pour les émissions.

En effet, l'usage fait des bâtiments (en particulier les hôpitaux, les data centers, universités...) joue un rôle important dans les deux cas.

Sans surprise, la nature des sources d'énergie joue un rôle sur les émissions de CO2 plus que sur la consommation totale.

Visualisation de l'importance de chaque feature (Emissions)

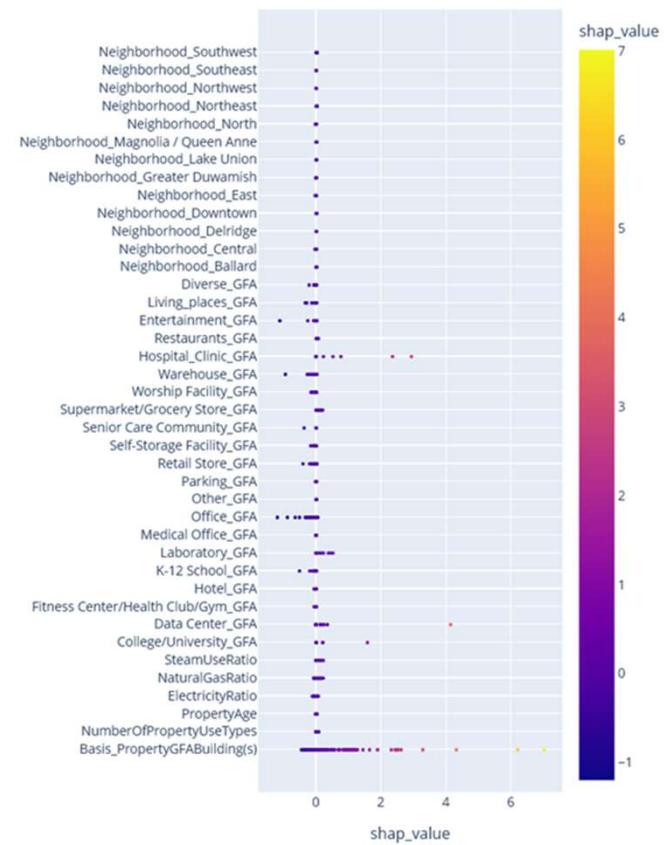


# Importance des features

## Observation fine avec le SHAP explainer



SHAP explainer : influence pour chaque point (Consommation)

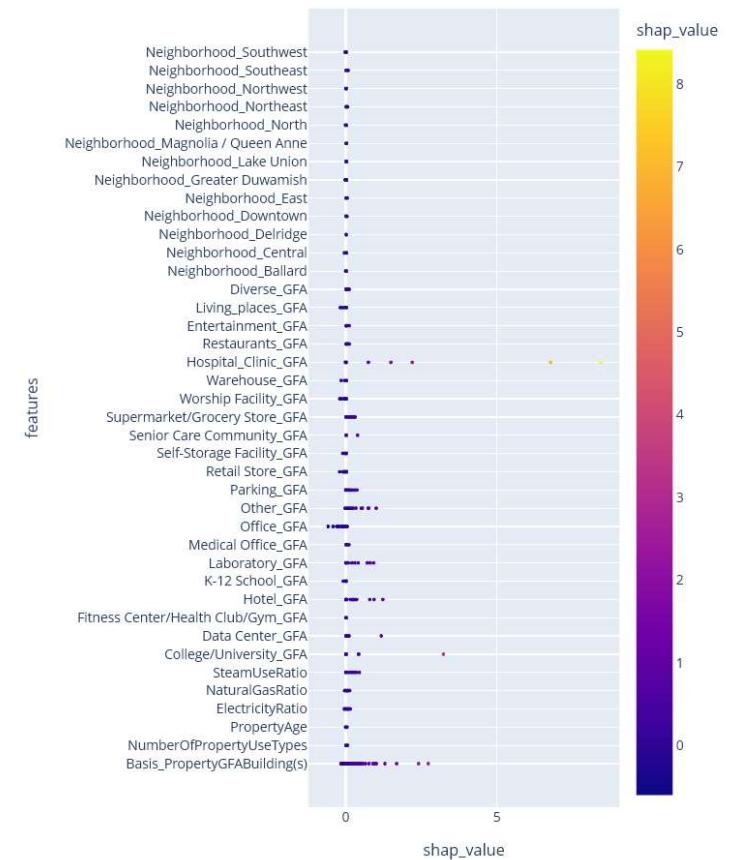


En complément de l'observation des coefficients, les valeurs SHAP calculées pour chaque individu (un sous-ensemble) permettent de visualiser l'influence de chaque variable explicatives pour les deux variables à prédire, par individu.

Les variations sont cohérentes avec les résultats précédemment obtenus.

Elles permettent de mettre en exergue des points qui peut-être gagneraient à être considérés comme des cas particuliers et qui expliquent le score malgré tout insatisfaisant de la régression linéaire.

SHAP explainer : influence pour chaque point (Emissions)



# Quel impact de l'ENERGY STAR SCORE...?

Permet-il de mieux prédire les émissions de CO2 ?



## Méthode

- Sous-ensemble des données initiales (où l'ENERGystarscore est renseigné)
- Split des données(Train/Test) et préprocessing
- Apprentissage avec le modèle optimal de la régression précédente (Ridge)
- Prédiction sur l'ensemble Test et calcul du score.

On effectue deux nouveau apprentissages avec calcul des scores pour éviter tout biais sur ce jeu de données restreints.

Le premier ne prend pas en compte l'ENERGY STAR Score.

Résultats pour le dataset où l'ENERGystarscore est renseigné mais sans le prendre en compte

Mean Squared Error: 0.1160  
Mean Absolute Error: 0.1888  
R2: 0.8840

Résultats avec prise en compte de l'ENERGystarscore  
Mean Squared Error: 0.1147  
Mean Absolute Error: 0.1868  
R2: 0.8853

Les résultats sont très proches !

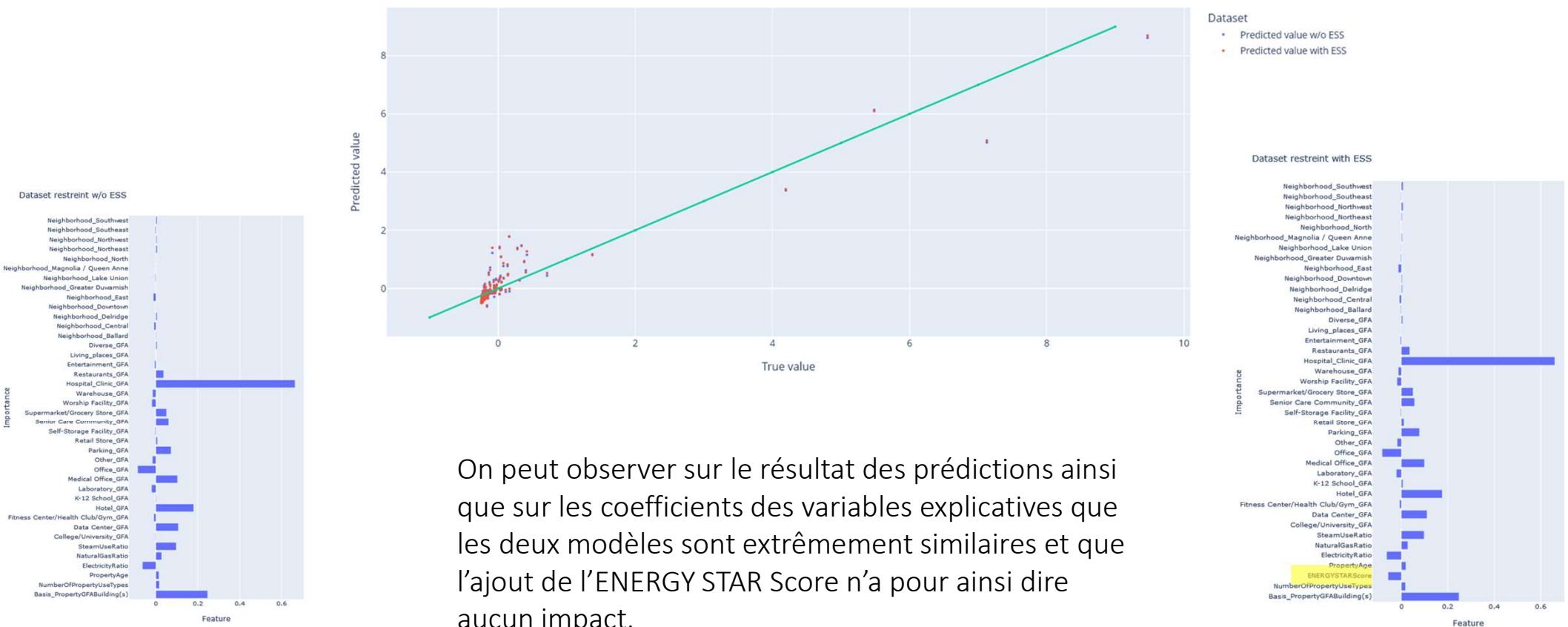
#	Column	Non-Null Count	Dtype
0	Basis_PropertyGFABuilding(s)	990 non-null	float64
1	NumberOfPropertyUseTypes	990 non-null	int64
2	ENERGystarscore	990 non-null	float64
3	TotalGHGEmissions	990 non-null	float64
4	SiteEnergyUse(kBtu)	990 non-null	float64
5	PropertyAge	990 non-null	int64
6	ElectricityRatio	990 non-null	float64
7	NaturalGasRatio	990 non-null	float64
8	SteamUseRatio	990 non-null	float64
9	College/University_GFA	990 non-null	float64
10	Data_Center_GFA	990 non-null	float64
11	Fitness_Center/Health_Club/Gym_GFA	990 non-null	float64
12	Hotel_GFA	990 non-null	float64
13	K-12_School_GFA	990 non-null	float64
14	Laboratory_GFA	990 non-null	float64
15	Medical_Office_GFA	990 non-null	float64
16	Office_GFA	990 non-null	float64
17	Other_GFA	990 non-null	float64
18	Parking_GFA	990 non-null	float64
19	Retail_Store_GFA	990 non-null	float64
20	Self-Storage_Facility_GFA	990 non-null	float64
21	Senior_Care_Community_GFA	990 non-null	float64
22	Supermarket/Grocery_Store_GFA	990 non-null	float64
23	Worship_Facility_GFA	990 non-null	float64
24	Warehouse_GFA	990 non-null	float64
25	Hospital_Clinic_GFA	990 non-null	float64
26	Restaurants_GFA	990 non-null	float64
27	Entertainment_GFA	990 non-null	float64
28	Living_places_GFA	990 non-null	float64
29	Diverse_GFA	990 non-null	float64
30	Neighborhood_Ballard	990 non-null	uint8
31	Neighborhood_Central	990 non-null	uint8
32	Neighborhood_Delridge	990 non-null	uint8
33	Neighborhood_Downtown	990 non-null	uint8
34	Neighborhood_East	990 non-null	uint8
35	Neighborhood_Greater_Duwamish	990 non-null	uint8
36	Neighborhood_Lake_Union	990 non-null	uint8
37	Neighborhood_Magnolia / Queen_Anne	990 non-null	uint8
38	Neighborhood_North	990 non-null	uint8
39	Neighborhood_Northeast	990 non-null	uint8
40	Neighborhood_Northwest	990 non-null	uint8
41	Neighborhood_Southeast	990 non-null	uint8
42	Neighborhood_Southwest	990 non-null	uint8

# Aucun impact de l'ENERGY STAR SCORE

La prédiction des émissions de CO2 n'est pas meilleure.



Emissions with ESS ( $R^2 = 0.884$ ) ou sans ESS - mais existant - ( $R^2 = 0.8853$ )



# CONCLUSION



Les scores obtenus pour les prédictions de la consommation totale d'énergie que des émissions de CO<sub>2</sub> ne sont pas optimaux.

Ce manque de précision peut être préjudiciable dans un contexte aussi crucial.

Par ailleurs, l'ENERGY STAR Score n'a apparemment pas de valeur prédictive des émissions de CO<sub>2</sub>.

Toutefois, il n'est pas impossible que la suppression d'outliers (à déterminer), un feature engineering encore plus poussé (se basant davantage sur les résultats des régressions) et une recherche par grille plus étendue (nécessitant plus de ressources) puisse donner de meilleurs résultats.

MERCI POUR  
VOTRE  
ATTENTION

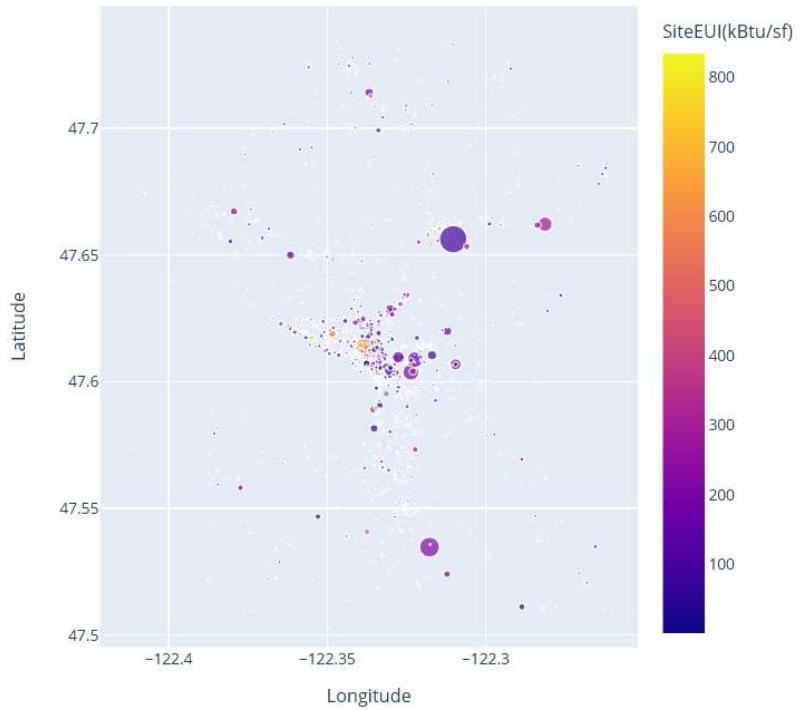


# BACK-UP SLIDES

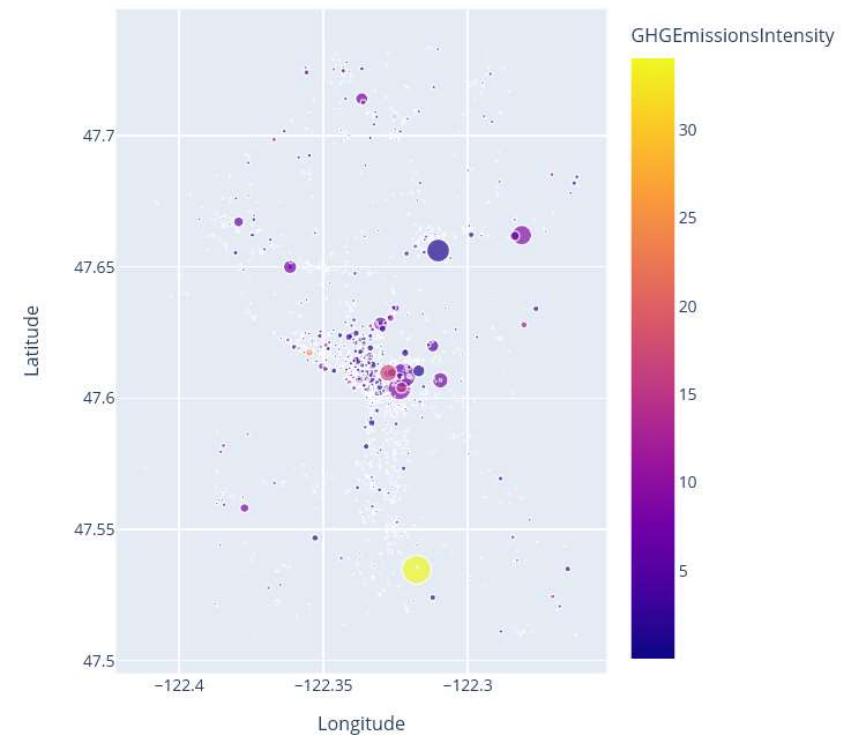
# Consommation énergétique et émissions de CO2

## Visualisation géographique

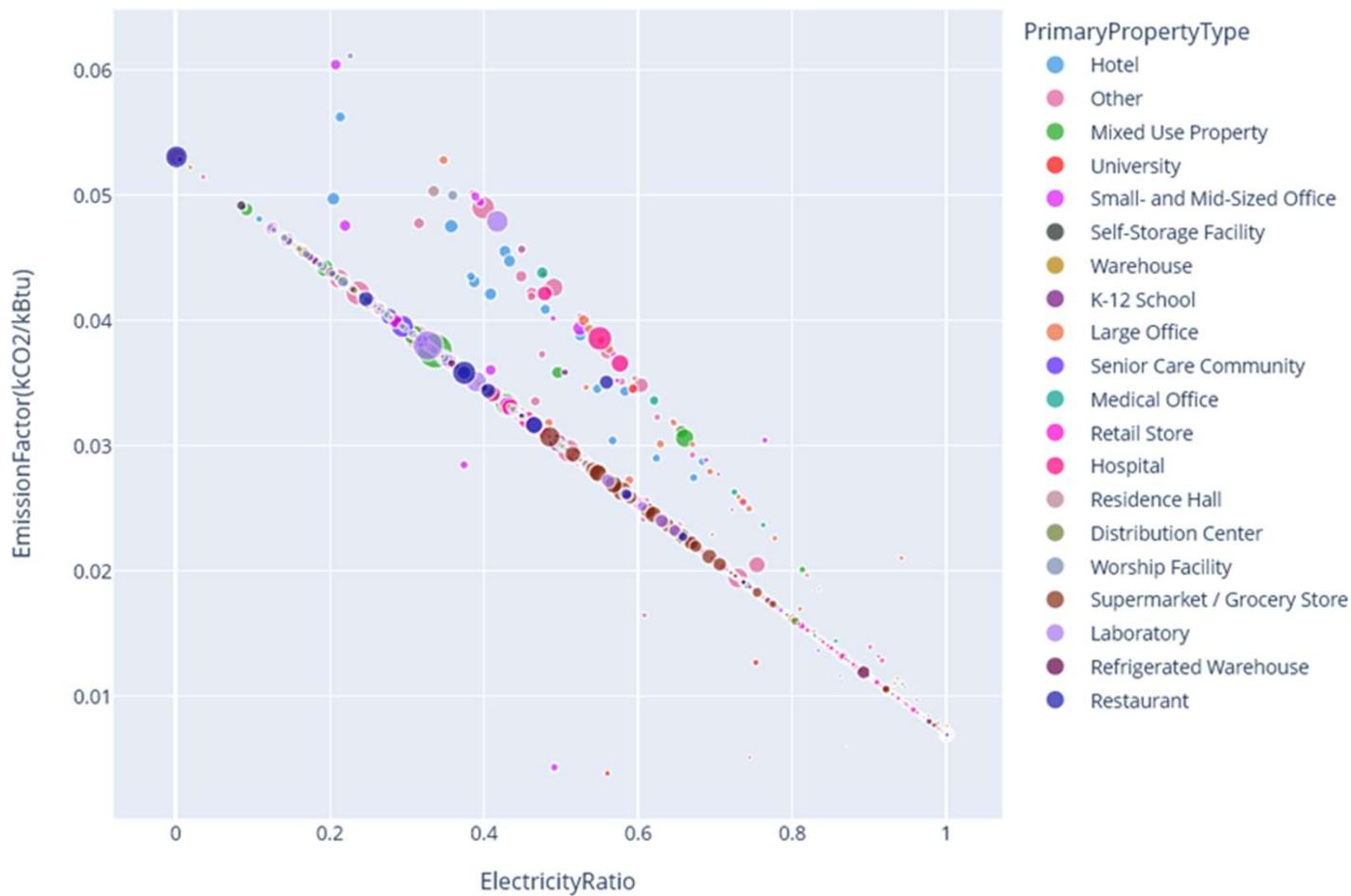
Consommation d'énergie (taille du cercle = consommation totale)



Emissions de CO2 (taille du cercle = émission totale)



Taille du cercle = GHGEmissionsIntensity



Taille du cercle = EmissionFactor(kCO<sub>2</sub>/kBtu)

