PRÉDICTION DES ÉMISSIONS DE CO2 DES BÂTIMENTS

Manu, David

CO2 DES BÂTIMENTS

Le projet

Prospection de modèles

Preprocessing

Application

Analyses

Conclusions

Recommandations
Retour d'expérience

Le projet

Distribution des tâches

Installation Azure (David)

DevOps (David)

Tests (David)

Chargement base (David)

Flow Dataiku (Manu)

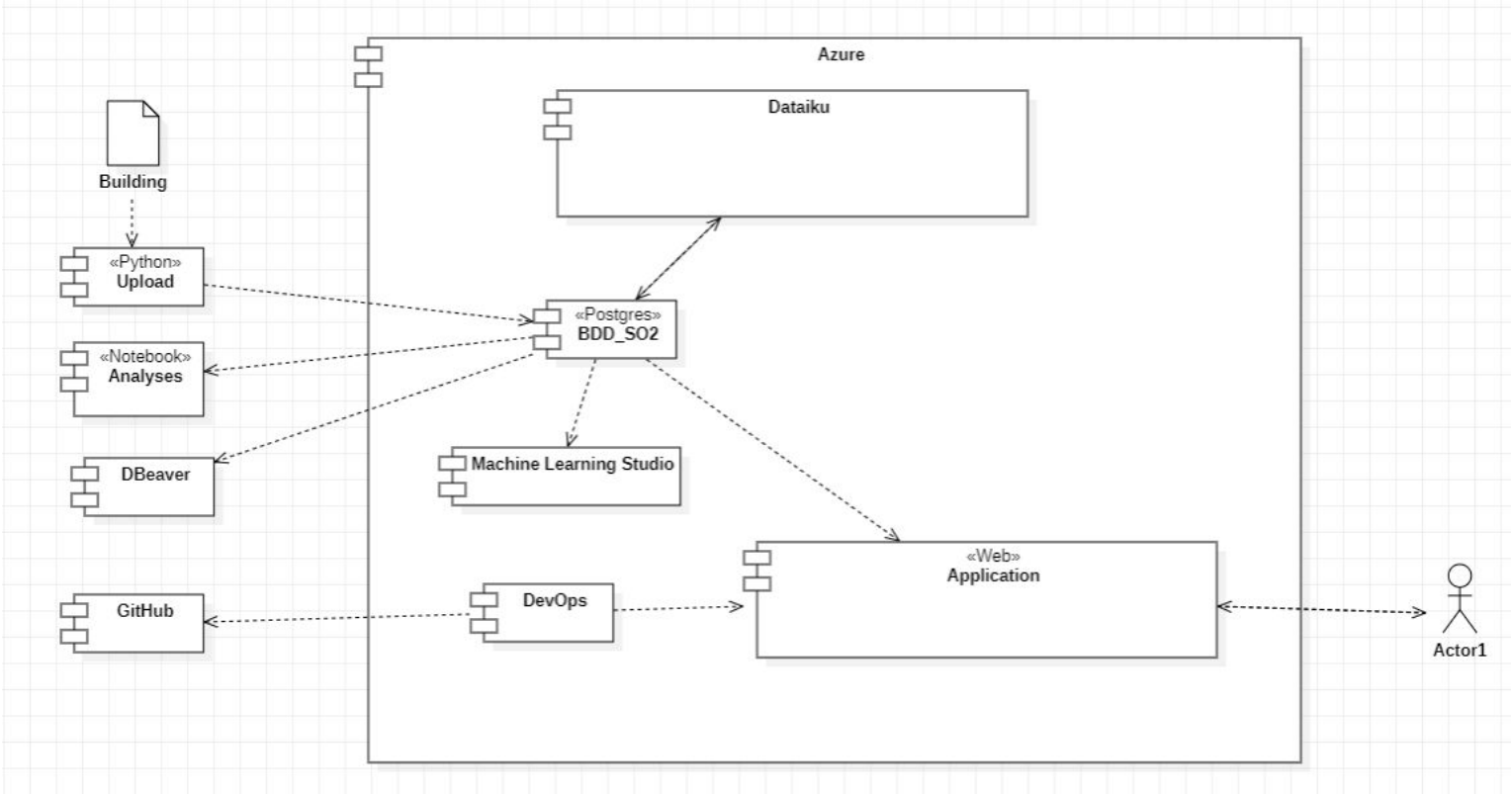
Imputations (Manu)

Analyses (Manu)

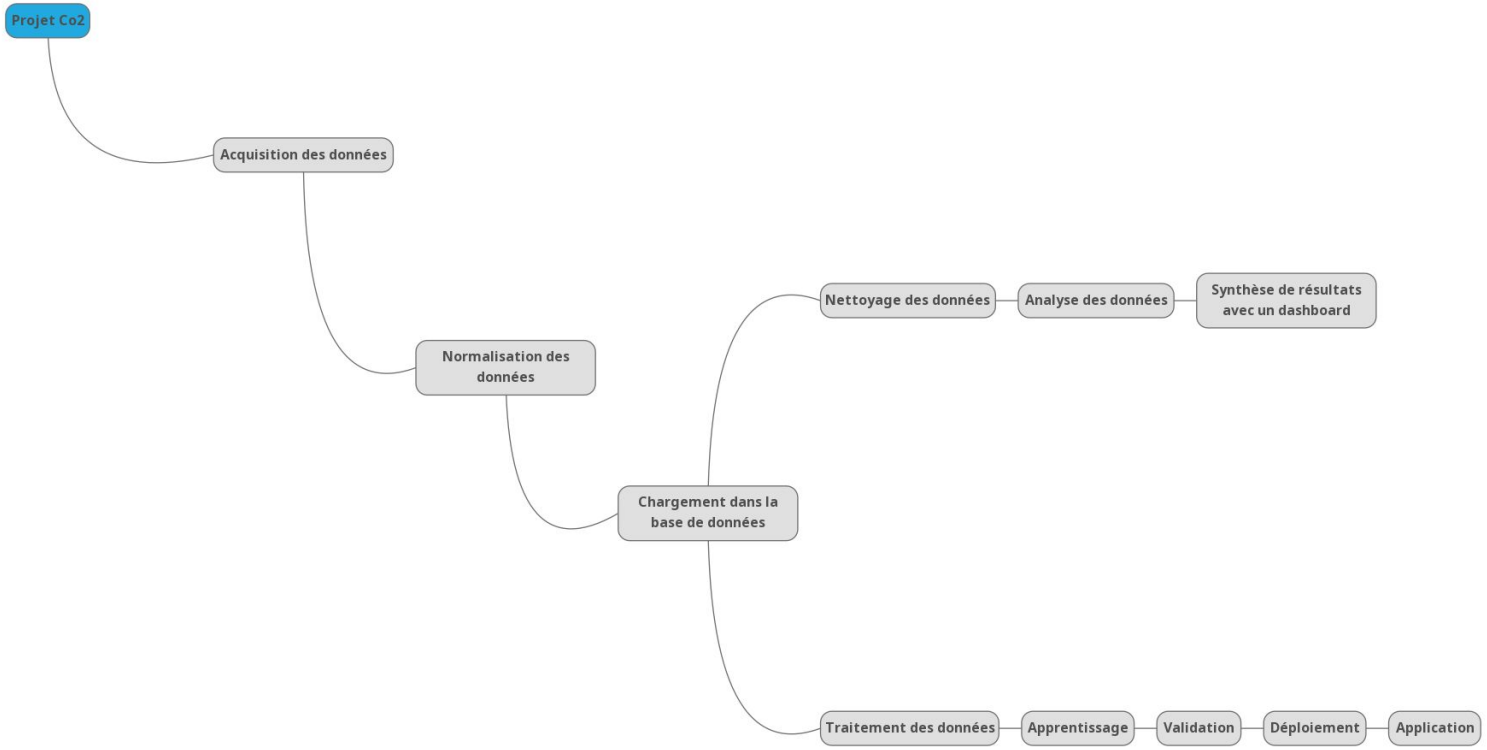
Conception de models (Manu, David)

Présentation (Manu, David)

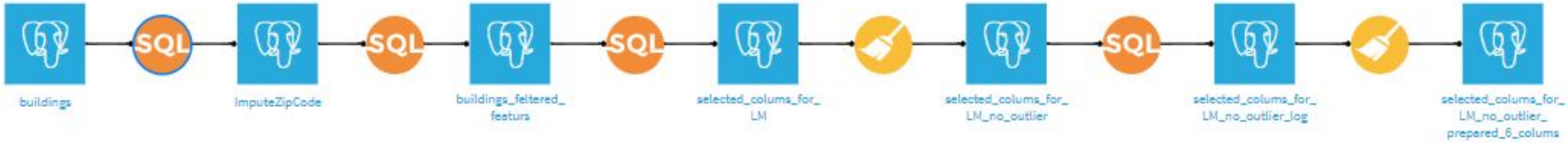
Installation, DevOps, Tests, Chargement Base.



Schema fonctionnel



Flow principal Dataiku



Preprocessing

Nettoyage de base de données

Suppression de colonnes

SUMMARY

Valid	3,376	100.0 %
Hapax	0	0.0 %
Invalid	0	0.0 %
Empty	0	0.0 %
0 HAPAXES		0.0 %
0 INVALIDS		0.0 %

Top 1 out of 1 values in sample	Count	%	Cum. %
2016	3376	100.0	100.0

SUMMARY

Valid	3,376	100.0 %
Hapax	0	0.0 %
Invalid	0	0.0 %
Empty	0	0.0 %
0 HAPAXES		0.0 %
0 INVALIDS		0.0 %

Top 1 out of 1 values in sample	Count	%	Cum. %
Seattle	3376	100.0	100.0

Nettoyage de base de données

Suppression de colonnes

SUMMARY

Valid	3,376	100.0 %
Hapax	3,349	99.2 %
Invalid	0	0.0 %
Empty	0	0.0 %

3349 HAPAXES 99.2 %

- #4706 Bitterlake
- #8944 West Seattle
- (71367A) SEATTLE Macy's
- (71371A) NORTHGATE Macy's

0 INVALIDS 0.0 %

Top 50 out of 3362 values in sample	Count	%	Cum. %
Northgate Plaza	3	0.1	0.1
Airport Way	2	0.1	0.1
Bayview Building	2	0.1	0.2
Canal Building	2	0.1	0.3
Central Park	2	0.1	0.3
Crestview Apartments	2	0.1	0.4
Fairview	2	0.1	0.4

Idem pour adresse

Nettoyage de base de données

Suppression de colonnes

Comments	Text	100.00%
YearsENERGYSTARCertified	Integer	96.48%
ThirdLargestPropertyUseType	Text	82.35%
ThirdLargestPropertyUseType...	Decimal	82.35%
SecondLargestPropertyUseType	Text	50.27%
SecondLargestPropertyUseTyp...	Decimal	50.27%

Nettoyage de base de données

Suppression outliers

Outlier	Text	99.05%

SUMMARY

Valid	3,376	100.0 %
Hapax	0	0.0 %
Invalid	0	0.0 %
Empty	3,344	99.1 %
0 HAPAXES		0.0 %
0 INVALIDS		0.0 %

Top 3 out of 3 values in sample	Count	%	Cum. %
No value	3344	99.1	99.1
Low outlier	23	0.7	99.7
High outlier	9	0.3	100.0

Nettoyage de base de données

Suppression des features quantitatives liées à la target

SiteEUIWN_kBtu_sf_	Decimal	99.82%
SourceEUI_kBtu_sf_	Decimal	99.73%
SourceEUIWN_kBtu_sf_	Decimal	99.73%
SiteEnergyUse_kBtu_	Decimal	99.85%
SiteEnergyUseWN_kBtu_	Decimal	99.82%
SteamUse_kBtu_	Decimal	99.73%
Electricity_kBtu_	Decimal	99.73%
NaturalGas_kBtu_	Decimal	99.73%

Nettoyage de base de données

Imputation NumberofBuildings

```
b."Neighborhood",
CASE WHEN "NumberofBuildings" IS NULL THEN 1 else "NumberofBuildings" END AS "NumberofBuildings",
b."NumberofFloors"
```

SUMMARY

Valid	3,376	100.0 %
Hapax	6	0.2 %
Invalid	0	0.0 %
Empty	8	0.2 %

SUMMARY

Valid	3,367	100.0 %
Hapax	6	0.2 %
Invalid	0	0.0 %
Empty	0	0.0 %

Nb_building = f(Nb_Floors)

	NumberofFloors	round
1	4	1
2	3	1
3	2	1

Nettoyage de base de données

Imputation LargestPropertyUseType

```
b. PropertyGFABuilding_s_ ,
CASE WHEN "LargestPropertyUseTypeGFA" IS NULL THEN "PropertyGFABuilding_s_" else "LargestPropertyUseTypeGFA" end as "LargestPropertyUseTypeGFA",
b."SecondLargestPropertyUseTypeGFA"
```

SUMMARY

Valid	3,376	100.0 %
Hapax	2,977	88.2 %
Invalid	0	0.0 %
Empty	20	0.6 %

Nettoyage de base de données

Imputation ZIPCode

Requêtte SQL.

```
CASE WHEN "ZipCode" IS NULL THEN (
    select
        zipcode
    from
        public."CO2_imputezipcode"
    where
        notzipLong = "Longitude"
        and notzipLat = "Latitude"
) else "ZipCode" END AS "ZipCode",
```

zipcode	notziplong	notziplat
double	double	double
Decimal	Decimal	Decimal
98125.0	-122.32232	47.70541
98144.0	-122.29787	47.59905
98117.0	-122.37717	47.6933
98125.0	-122.29735	47.72126
98107.0	-122.39228	47.67295
98117.0	-122.37624	47.67734
98119.0	-122.37525	47.63572
98112.0	-122.31574	47.63228
98122.0	-122.30225	47.60775
98118.0	-122.27813	47.5644
98126.0	-122.37441	47.54067
98108.0	-122.31154	47.56722
98104.0	-122.32283	47.59625
98109.0	-122.35784	47.63644
98108.0	-122.32431	47.52832
98108.0	-122.29536	47.53939

Nettoyage de base de données

Imputation ENERGYSTARScore

SUMMARY

Valid	3,376	100.0 %
Hapax	0	0.0 %
Invalid	0	0.0 %
Empty	843	25.0 %

	123 energystarscore_imputation	123 count	123 deceny
1	77	138	1 900
2	76	126	1 910
3	78	220	1 920
4	68	53	1 930
5	75	60	1 940
6	75	161	1 950
7	70	356	1 960
8	67	251	1 970
9	74	350	1 980
10	76	265	1 990
11	76	361	2 000
12	93	192	2 010

Transformation de features

Utilise un type d'énergie

SteamUse_kBtu_	Decimal
Electricity_kBtu_	Decimal
NaturalGas_kBtu_	Decimal

Have_Stream_Energy	Boolean
Have_Electricity_Energy	Boolean
Have_NaturalGas_Energy	Boolean

```
b."SteamUse_kBtu_" > 0.0 as "Have_Stream_Energy",
b."Electricity_kBtu_" > 0.0 as "Have_Electricity_Energy",
b."NaturalGas_kBtu_" > 0.0 as "Have_NaturalGas_Energy",
```

SUMMARY

Valid	3,314	100.0 %
Hapax	0	0.0 %
Invalid	0	0.0 %
Empty	0	0.0 %

Top 2 out of 2 values in sample	Count	%	Cum. %
true	2090	63.1	63.1
false	1224	36.9	100.0

Transformation de features

Log de features

```
log ("LargestPropertyUseTypeGFA") as "LargestPropertyUseTypeGFA_log",
log ("TotalGHGEmissions") as "TotalGHGEmissions_log",
log ("SiteEnergyUse_kBtu_") as "SiteEnergyUse_kBtu_log"
```

Sélection de features

YearBuilt	Integer	100.00%
BuildingType	Text	100.00%
Neighborhood	Text	100.00%
Have_Stream_Energy	Boolean	100.00%
Have_Electricity_Energy	Boolean	100.00%
Have_NaturalGas_Energy	Boolean	100.00%
PrimaryPropertyType	Text	100.00%
NumberofBuildings	Integer	100.00%
LargestPropertyUseTypeGFA	Decimal	100.00%
TotalGHGEmissions	Decimal	100.00%
SiteEnergyUse_kBtu_	Decimal	100.00%
LargestPropertyUseTypeGF...	Decimal	100.00%
TotalGHGEmissions_log	Decimal	100.00%
SiteEnergyUse_kBtu_log	Decimal	100.00%

On conserve les targets, TotalGHGEmission et SiteEnergyUse.

On cherche à prédire les deux targets en fonction du projet.

Un projet intègre :

- Des types d'énergie.
- Des usages (résidentiels).
- Une zone géographique (quartier).
  Une surface.
- Un nombre de bâtiments.

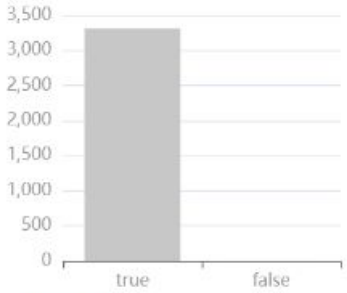
La date de construction de bâtiments est un élément temporel.

Analyses

Analyses

Have_Electricity...

Histogram



Summary stats

N values	3314
N distinct	2
Mode	true
N empty	0

Frequency table

true	100%	3313
false	0%	1
N distinct		2

Have_NaturalGa...

Histogram

Summary stats

N values	3314
N distinct	2
Mode	true
N empty	0

Frequency table

true	63%	2090
false	37%	1224
N distinct		2

Have_Stream_En...

Histogram

Summary stats

N values	3314
N distinct	2
Mode	false
N empty	0

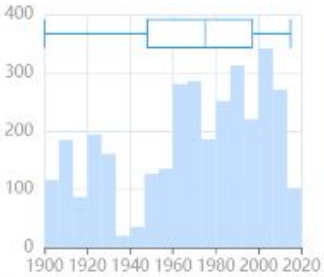
Frequency table

false	96%	3185
true	4%	129
N distinct		2

Analyses

YearBuilt

Histogram

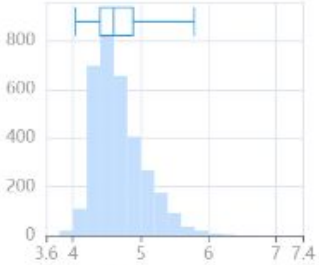


Summary stats

N values	3314
N distinct	113
Mean	1968.6976463
Median	1975
Min	1900
Max	2015

LargestPropertyUse...

Histogram

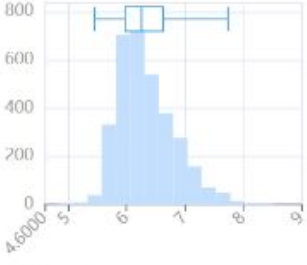


Summary stats

N values	3314
N distinct	3085
N finite	3314
Mean	4.6776795338
Median	4.6009075784
Std Dev	0.370895892
Min	3.7525094008
Max	6.9694231816

SiteEnergyUse_kBt...

Histogram

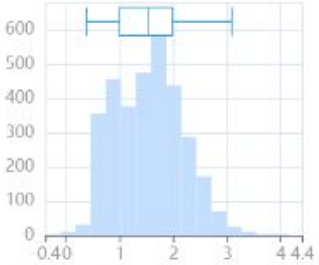


Summary stats

N values	3314
N distinct	3314
N finite	3314
Mean	6.3401236674
Median	6.2604589349
Std Dev	0.493740552
Min	4.7568885434
Max	8.9414735231

TotalGHGEmissions...

Histogram



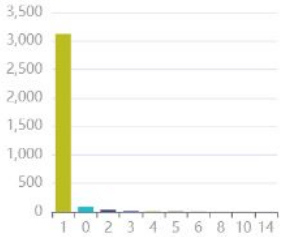
Summary stats

N values	3314
N distinct	2782
N finite	3314
Mean	1.5286686172
Median	1.5350407393
Std Dev	0.6453773932
Min	-0.397940009
Max	4.2271403106

Analyses

NumberofBuildings

Histogram



Summary stats

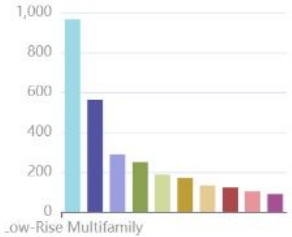
N values	3314
N distinct	17
Mode	1
N empty	0

Frequency table

1	94%	3123
0	3%	92
2	1%	36
3	1%	22
4	0%	12
5	0%	9
6	0%	5
8	0%	3
10	0%	2

PrimaryPropertyTy...

Histogram



Summary stats

N values	3314
N distinct	24
Mode	Low-Rise Multifamily
N empty	0

Frequency table

Low-Rise Multifamily	29%	966
Mid-Rise Multifamily	17%	561
Small- and Mid-Sized Office	9%	288
Other	8%	250
Warehouse	6%	187
Large Office	5%	170
Mixed Use Property	4%	132
K-12 School	4%	123

Neighborhood

Histogram

Summary stats

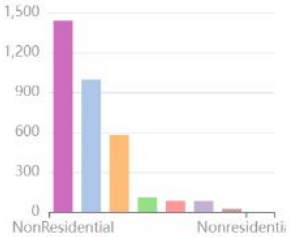
N values	3314
N distinct	19
Mode	DOWNTOWN
N empty	0

Frequency table

DOWNTOWN	17%	562
EAST	14%	448
MAGNOLIA / QUEEN ANNE	13%	415
GREATER DUWAMISH	11%	371
NORTHEAST	8%	274
LAKE UNION	8%	249
NORTHWEST	6%	208
SOUTHWEST	5%	157

BuildingType

Histogram



Summary stats

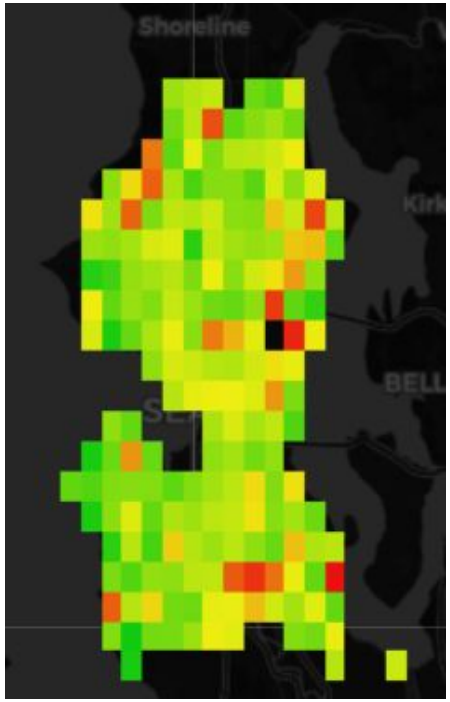
N values	3314
N distinct	8
Mode	NonResidential
N empty	0

Frequency table

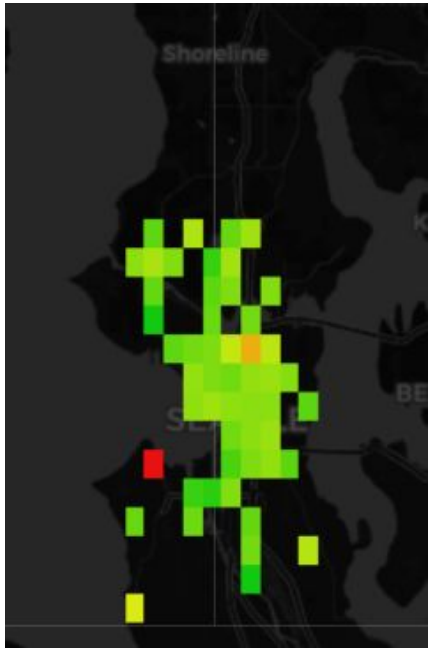
NonResidential	43%	1439
Multifamily LR (1-4)	30%	996
Multifamily MR (5-9)	17%	578
Multifamily HR (10+)	3%	109
Nonresidential COS	3%	84
SPS-District K-12	3%	83
Campus	1%	24
Nonresidential WA	0%	1
N distinct		8

Analyses

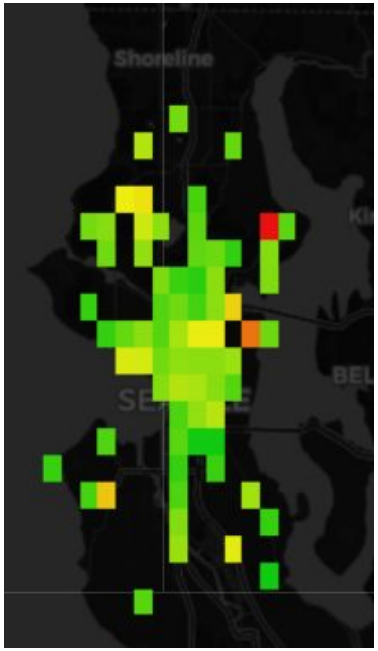
Émissions globales apportées à la surface par tranche d'années de construction



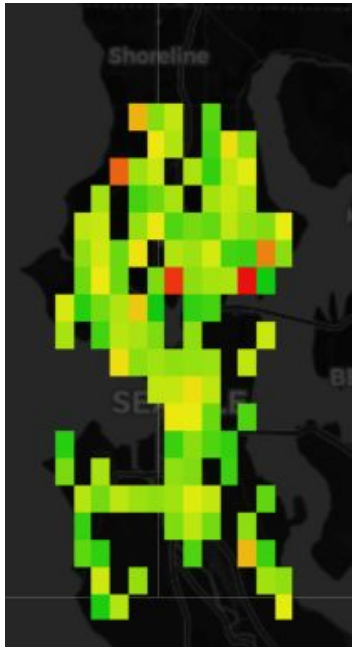
1900-1920



1921-1940

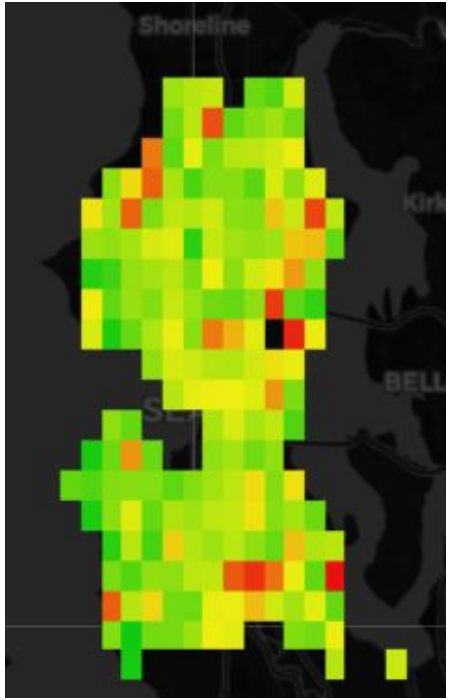


1941-1960



Analyses

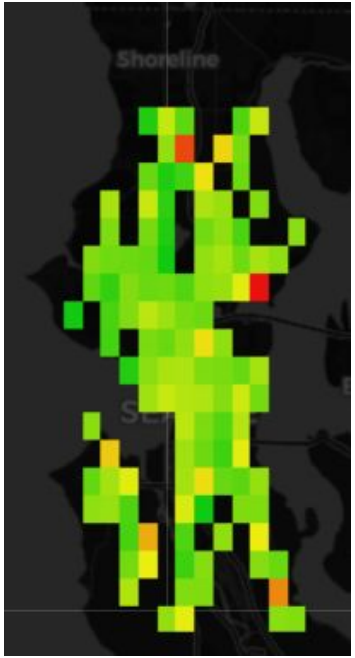
Émissions globales en fonction de la surface par tranche d'années de construction



1961-1980

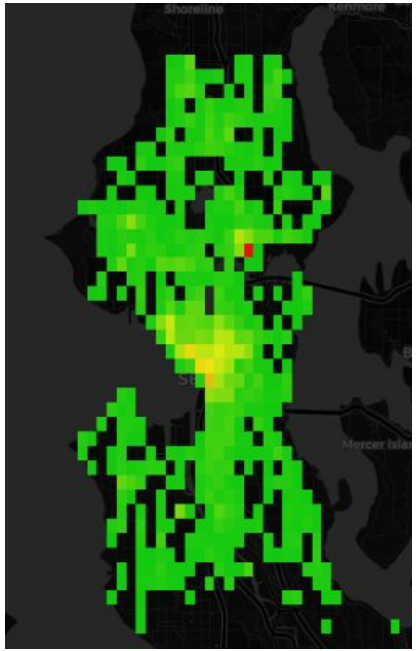
1981-2000

2001- 2015

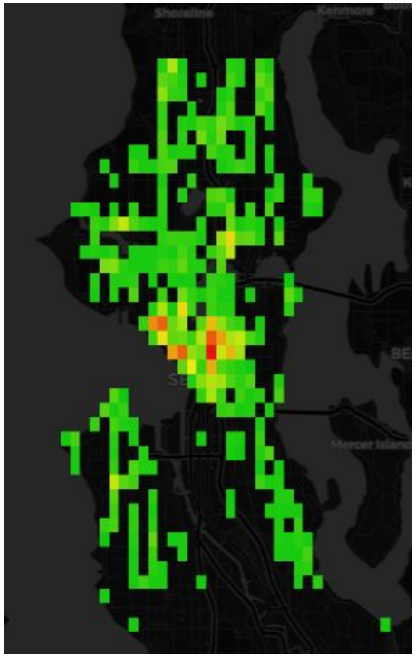


Analyses

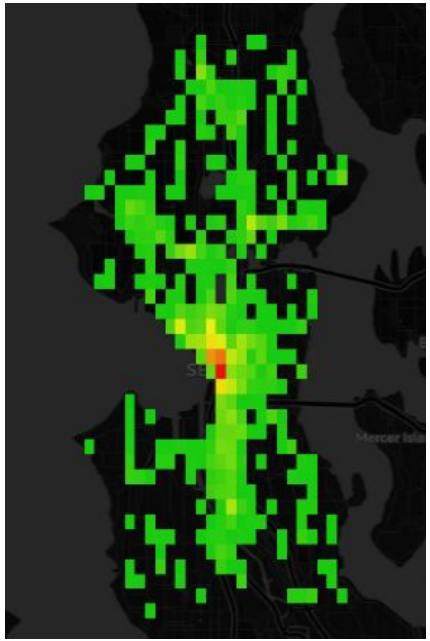
Répartition des bâtiments



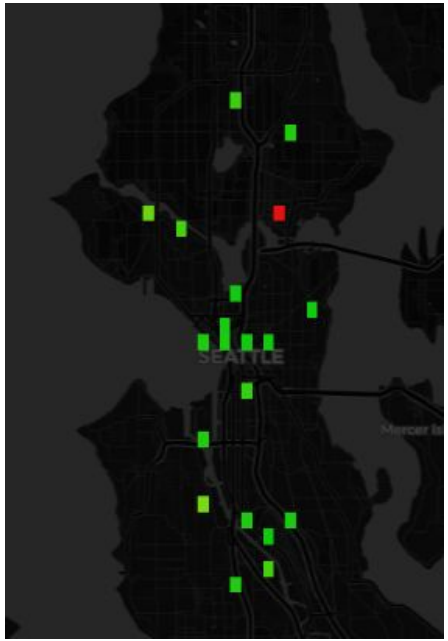
Tout



Multifamily



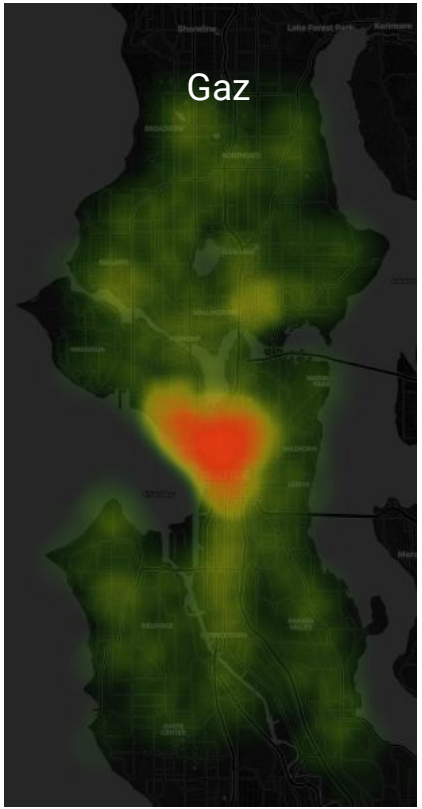
Non Residentiel

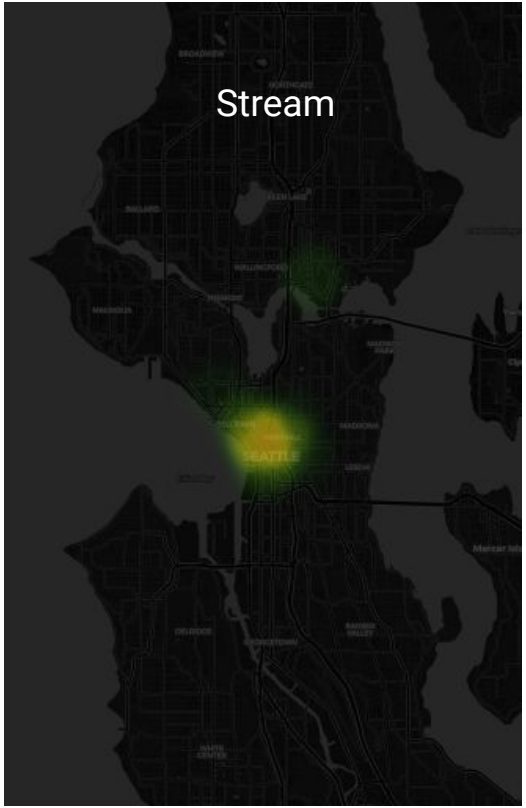


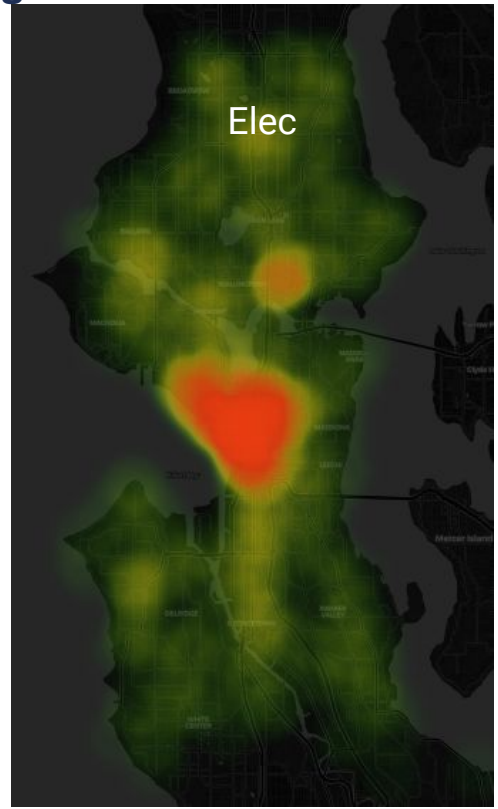
Campus

Analyses

Densité d’utilisation des énergies



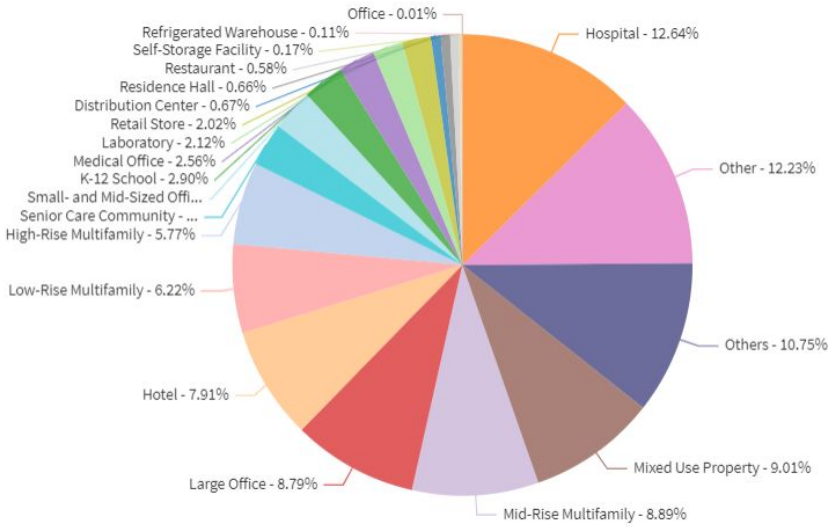




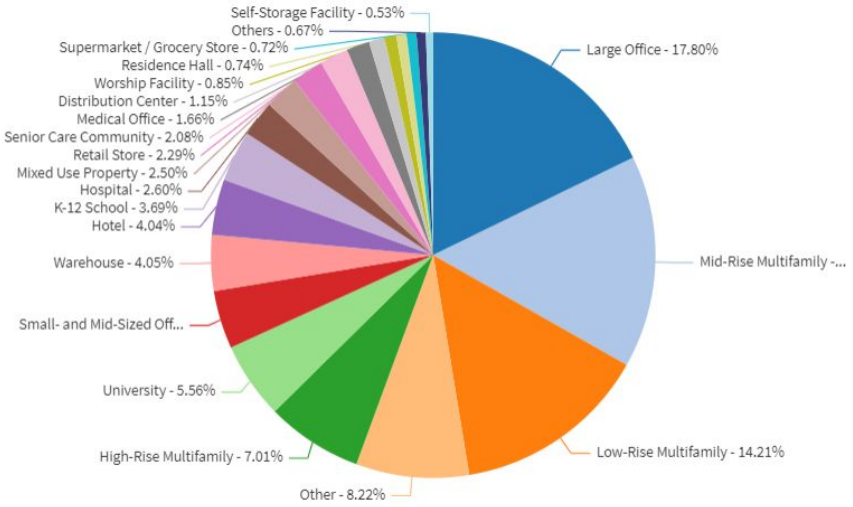
Analyses

Répartition énergétique et surface au sol par usages

Sum of TotalGHGEmissions by PrimaryPropertyType



Sum of LargestPropertyUseTypeGFA by PrimaryPropertyType

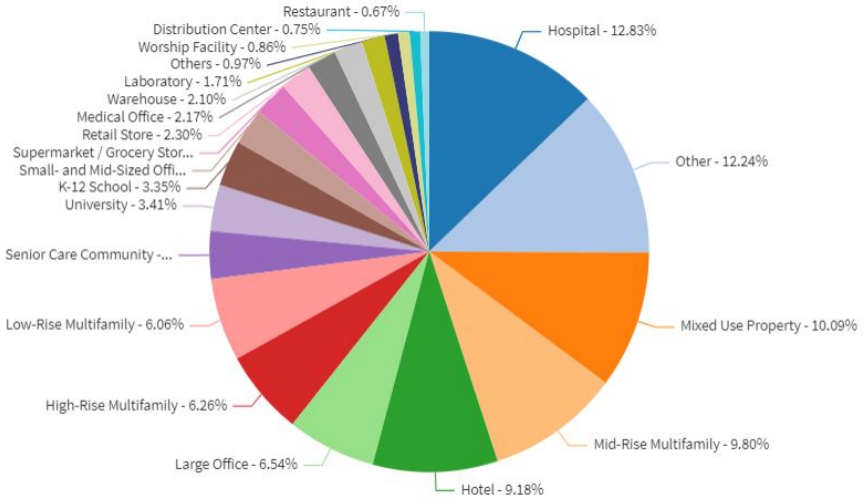


Analyses

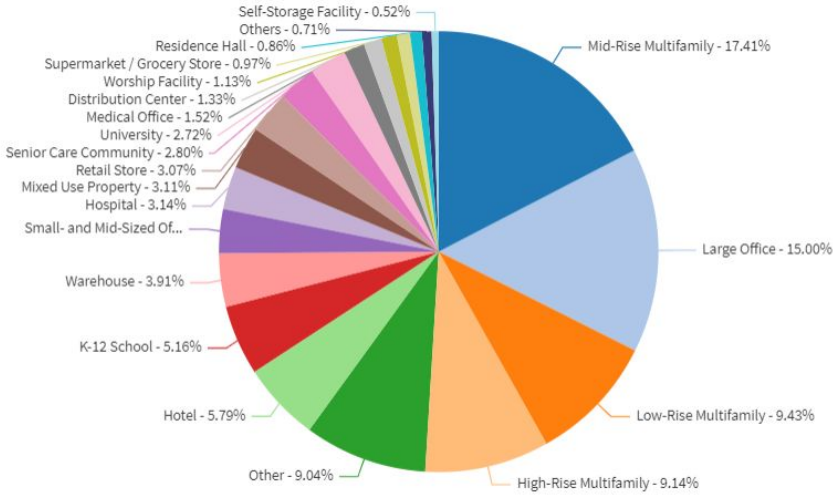
Répartition énergétique et surface au sol par usages en fonction de l'utilisation de gaz

- [x] true (2090)
- [ ] false (1224)

Sum of TotalGHGEmissions by PrimaryPropertyType



Sum of LargestPropertyUseTypeGFA by PrimaryPropertyType

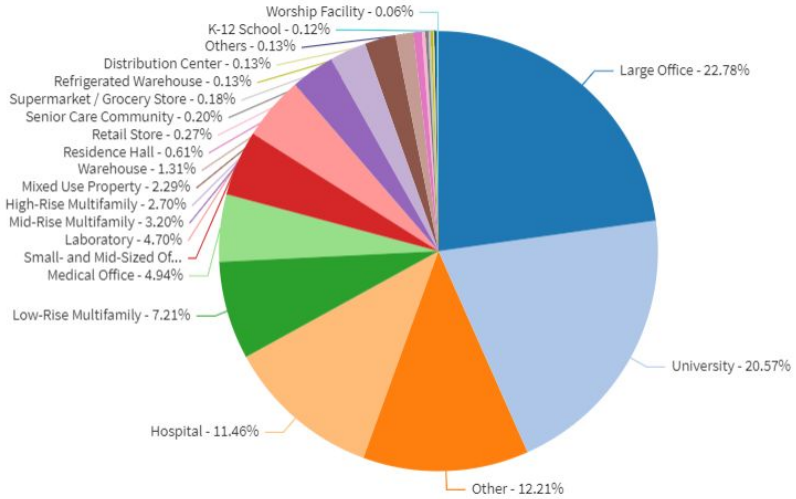


Analyses

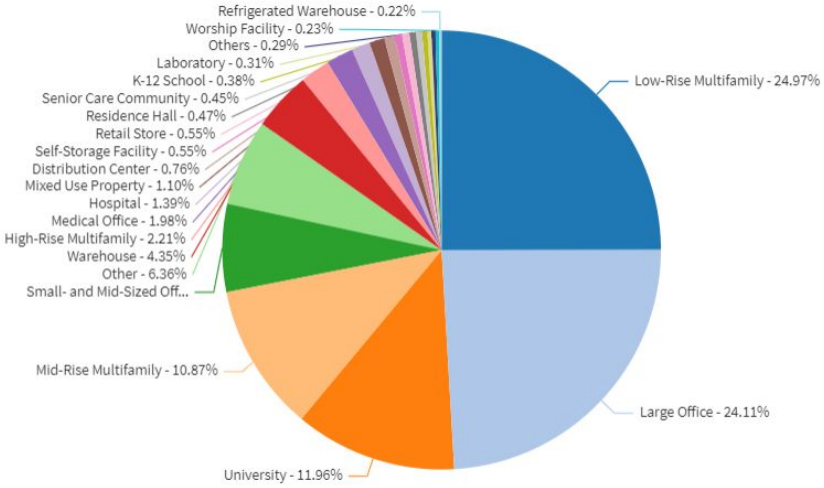
Répartition énergétique et surface au sol par usages en fonction de l'utilisation de gaz

- [ ] true (2090)
- [x] false (1224)

Sum of TotalGHGEmissions by PrimaryPropertyType

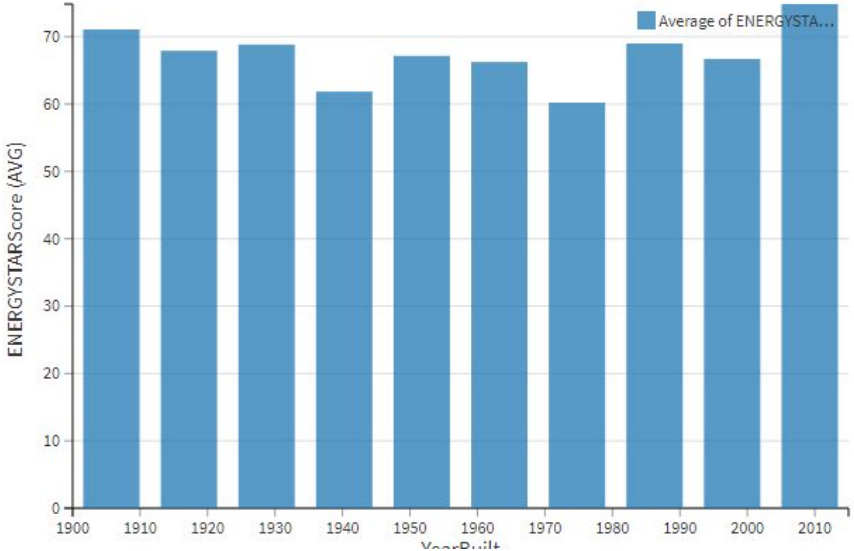


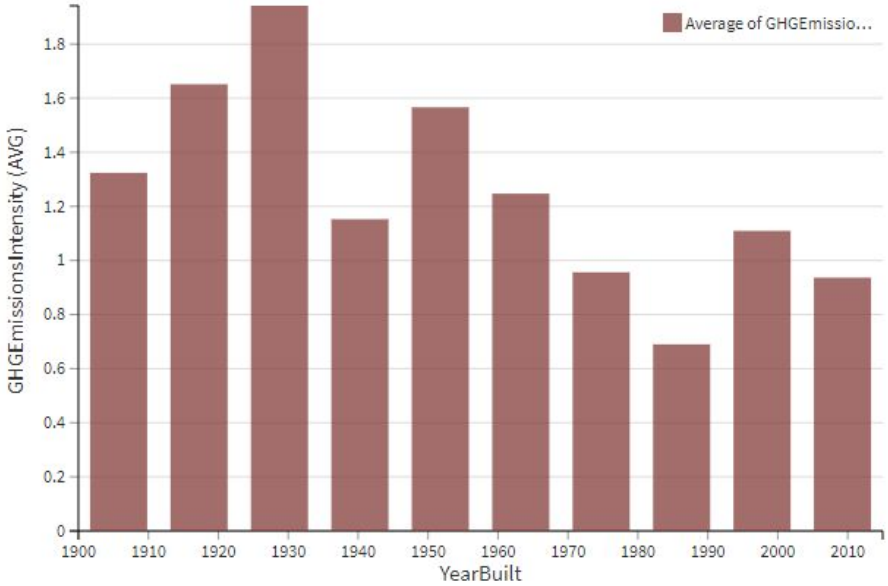
Sum of LargestPropertyUseTypeGFA by PrimaryPropertyType



Analyses

Émissions globales à comparer avec les évaluation de EnergyStarScore





Analyses

Émissions globales à comparer avec les évaluation de EnergyStarScore

Analyses

Le score d’évaluation est inadapté à informer sur le degrés d’émissions le fait d’avoir ou non du gaz semble plus fiable à informer sur le degré d’émission

Analyses

TotalGHGEmissions_log by # SiteEnergyUse_kBtu_log split by BuildingType

All

	LargestPropertyUs...	SiteEnergyUse_kBt...	TotalGHGEmissions...
LargestPro...	1.000	0.736	0.571
SiteEnergy...	0.736	1.000	0.878
TotalGHGE...	0.571	0.878	1.000

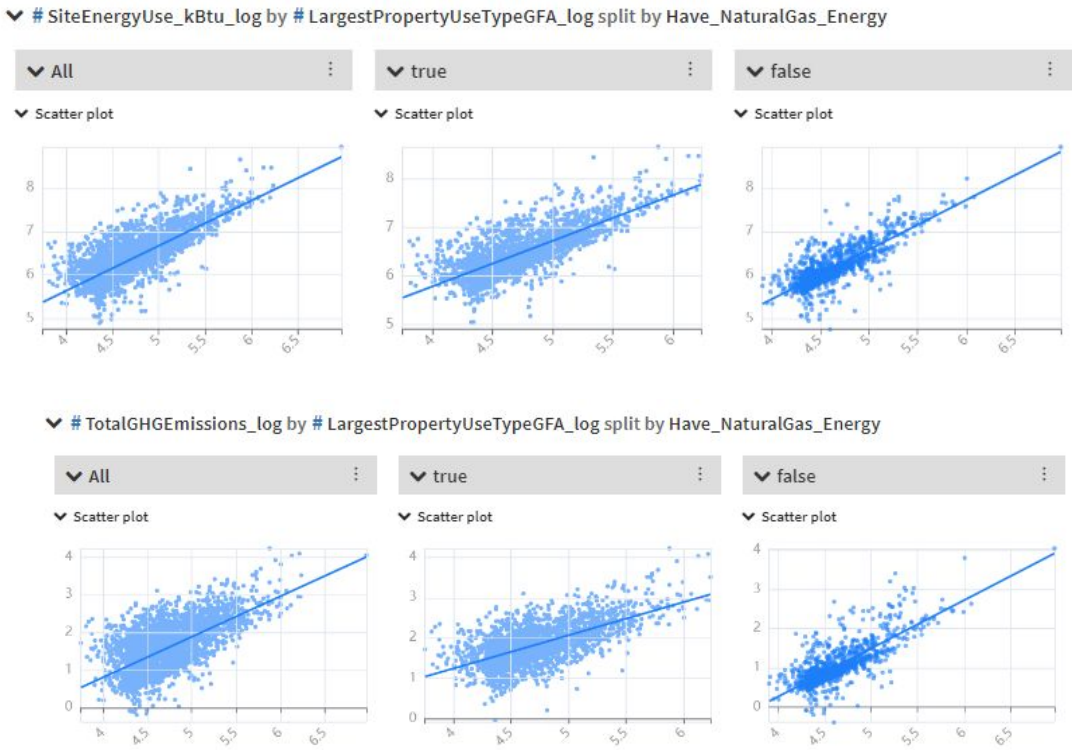
BuildingType: Multifamily LR (1-4)

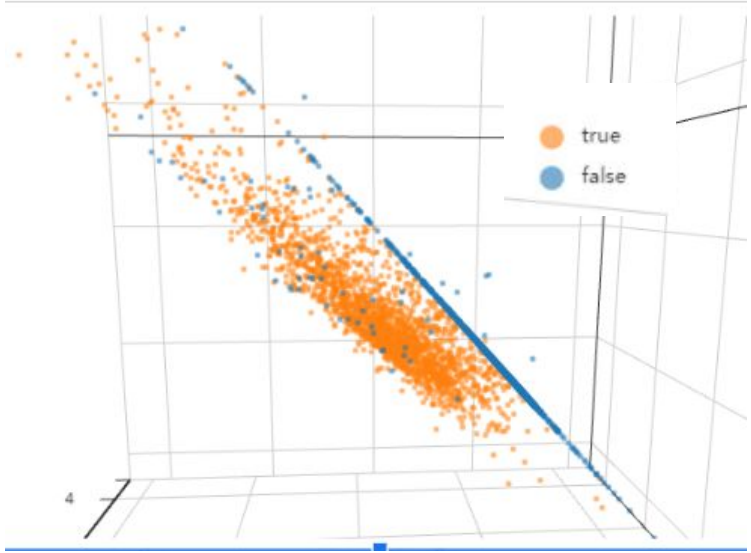
	LargestPropertyUs...	SiteEnergyUse_kBt...	TotalGHGEmissions...
LargestPro...	1.000	0.690	0.428
SiteEnergy...	0.690	1.000	0.847
TotalGHGE...	0.428	0.847	1.000

BuildingType: Multifamily HR (10+)

	LargestPropertyUs...	SiteEnergyUse_kBt...	TotalGHGEmissions...
LargestPro...	1.000	0.907	0.704
SiteEnergy...	0.907	1.000	0.854
TotalGHGE...	0.704	0.854	1.000

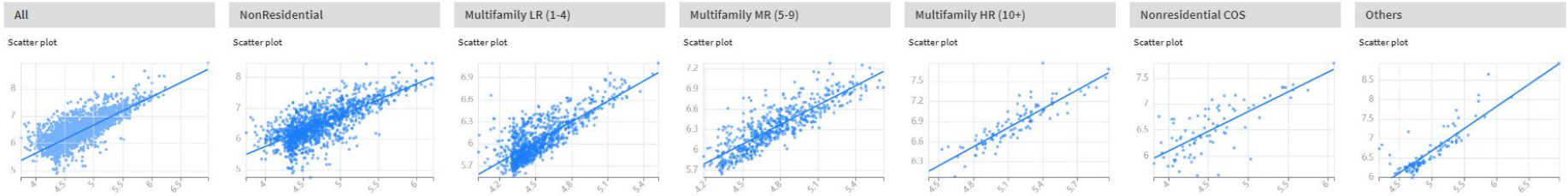
Analyses


SiteEnergyUse_kBtu_log by # LargestPropertyUseTypeGFA_log split by Have_NaturalGas_Energy
All
true
false
Scatter plot
Scatter plot
Scatter plot
TotalGHGEmissions_log by # LargestPropertyUseTypeGFA_log split by Have_NaturalGas_Energy
All
true
false
Scatter plot
Scatter plot
Scatter plot

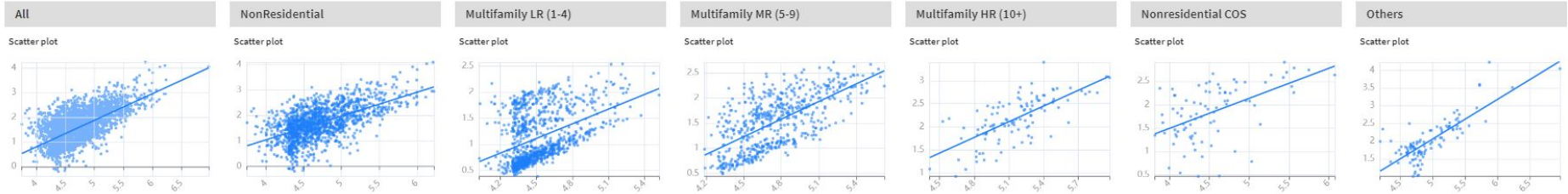

true
false

Analyses

SiteEnergyUse_kBtu_log by # LargestPropertyUseTypeGFA_log split by BuildingType



TotalGHGEmissions_log by # LargestPropertyUseTypeGFA_log split by BuildingType



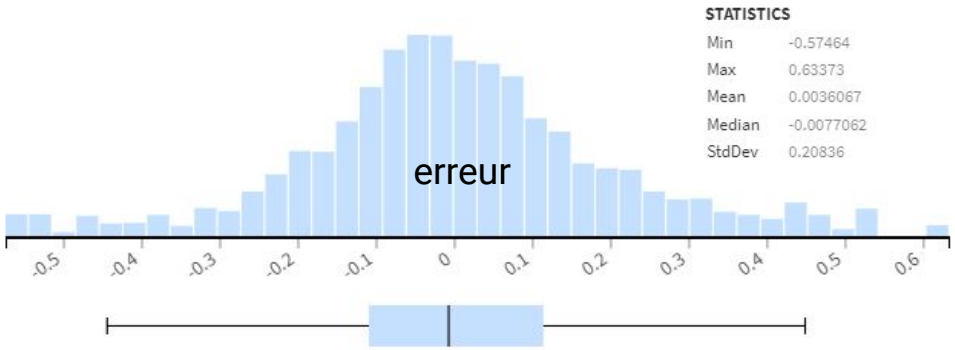
Prospection de modèles

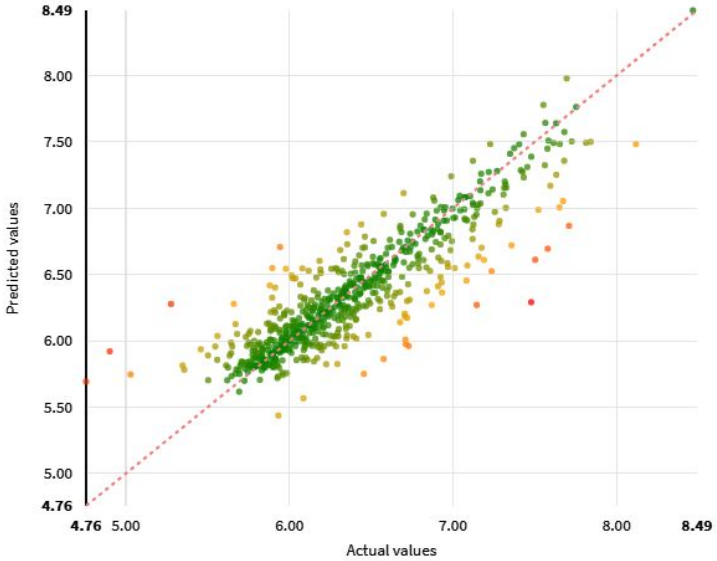
Prospection de modèles

Avec Dataiku (SiteEnergyUse)

R2 Score

Random forest (s19)	0.748 (± 0.027)	
Ridge (L2) regression (s19)	0.780 (± 0.024)	
SVM (s19)	0.785 (± 0.024)	
Single Layer Perceptron (s19)	0.740 (± 0.168)	



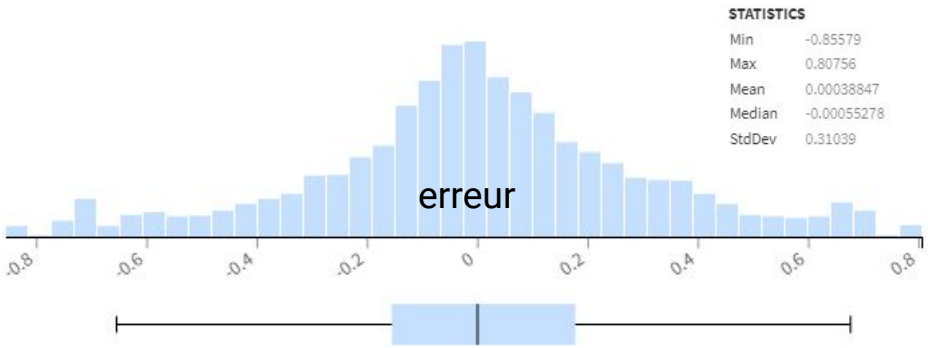


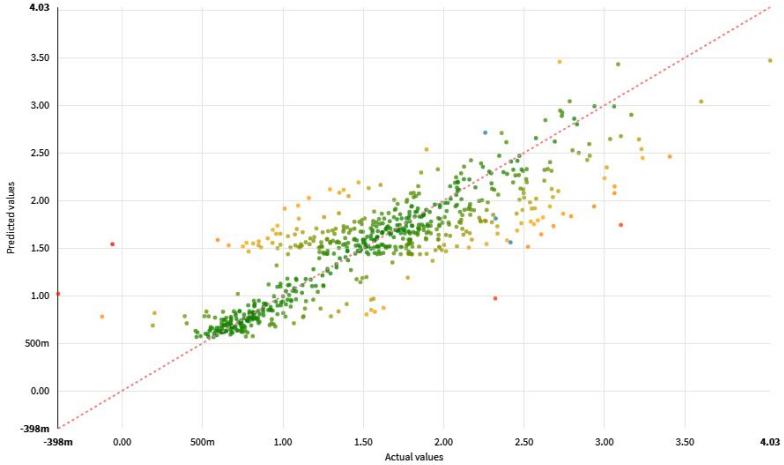
Prospection de modèles

Avec Dataiku (TotalGHGEmission)

R2 Score

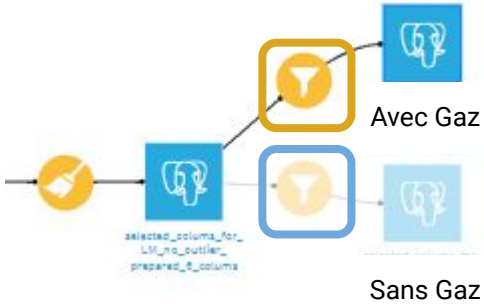
Random forest (s22)	0.724 (± 0.025)	
Ridge (L2) regression (s22)	0.721 (± 0.017)	
SVM (s22)	0.735 (± 0.025)	
Single Layer Perceptron (s22)	0.736 (± 0.022)	

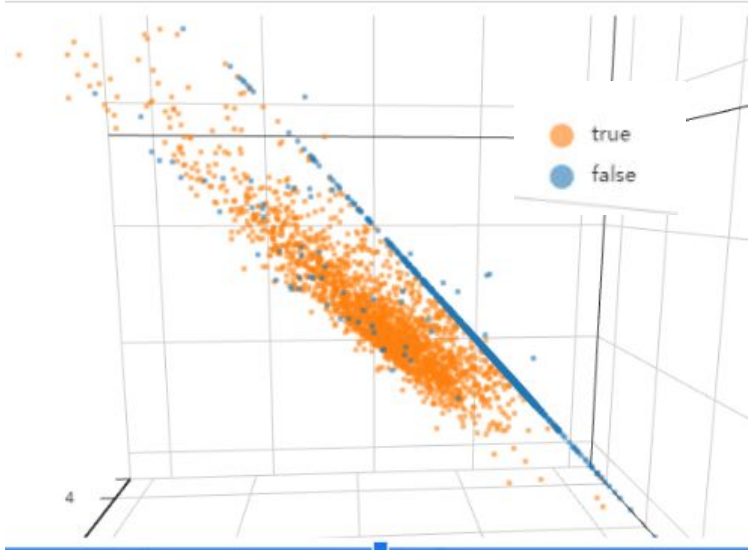




Prospection de modèles

Avec Dataiku (TotalGHGEmission si avec GazEnergy ou sans GazEnergy)



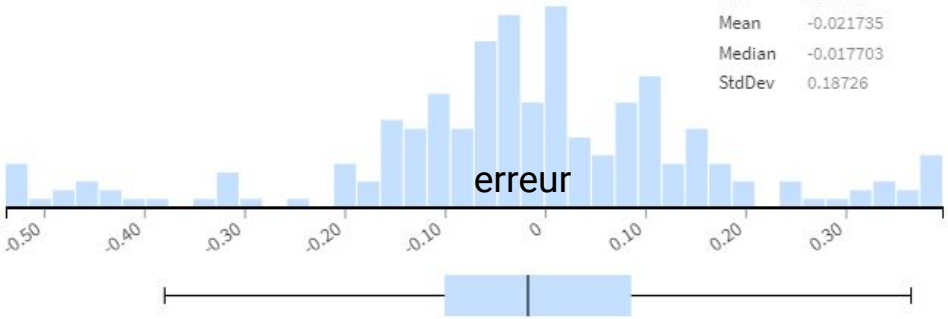


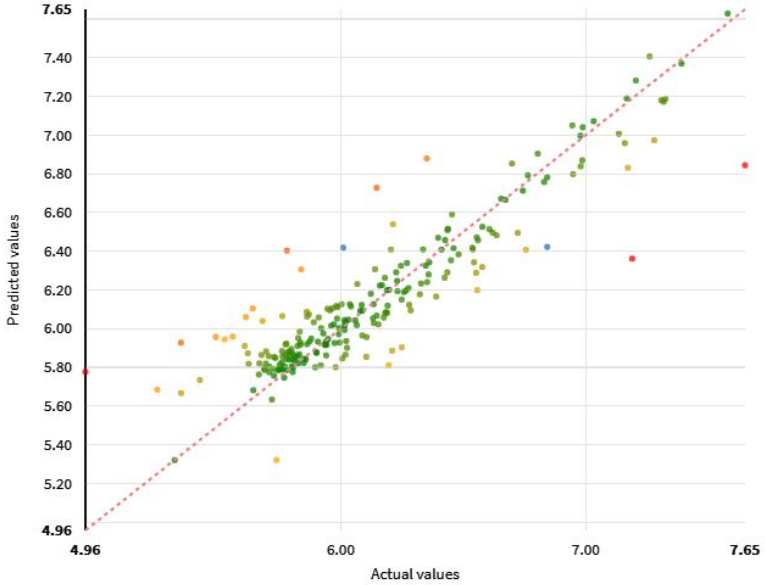
Prospection de modèles

Avec Dataiku (TotalGHGEmission sens GazEnergy)

R2 Score

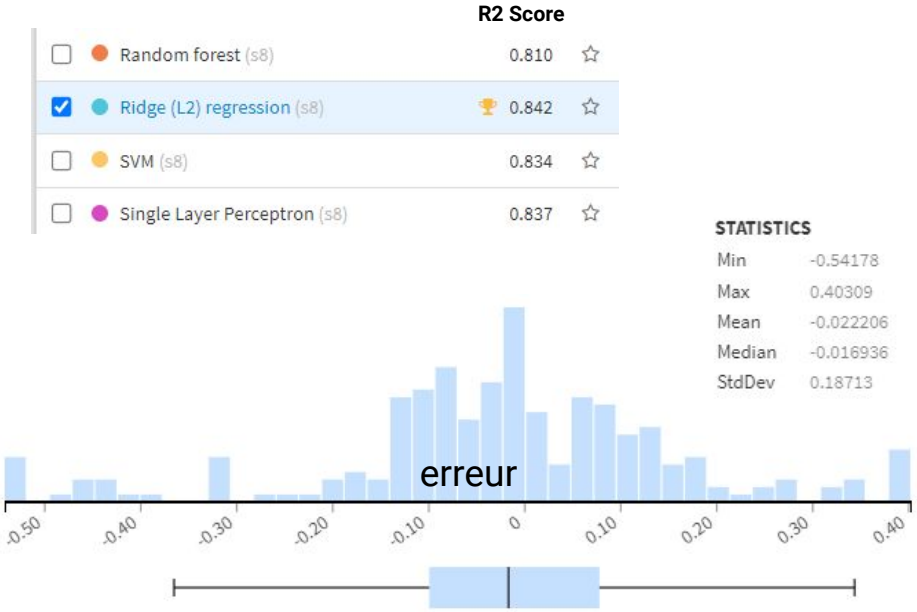
Random forest (s6)	0.809	
Ridge (L2) regression (s6)	0.841	
SVM (s6)	0.831	
Single Layer Perceptron (s6)	0.837	

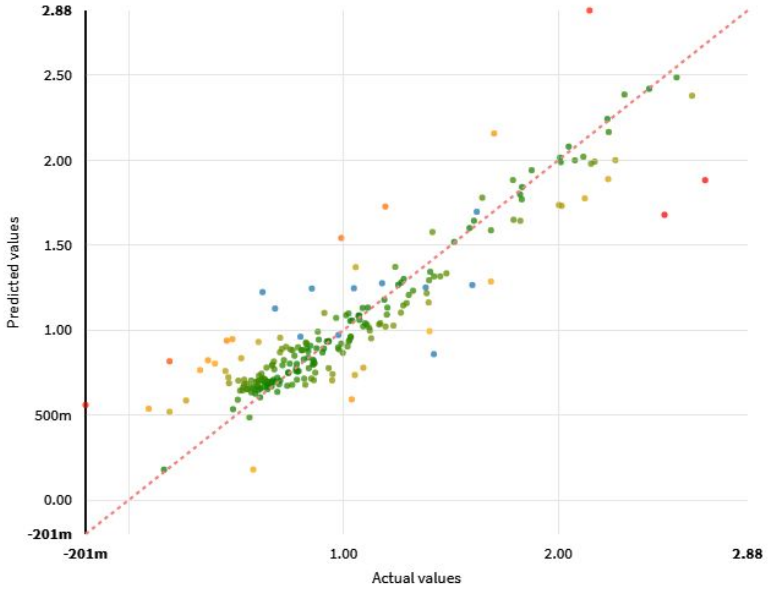




Prospection de modèles

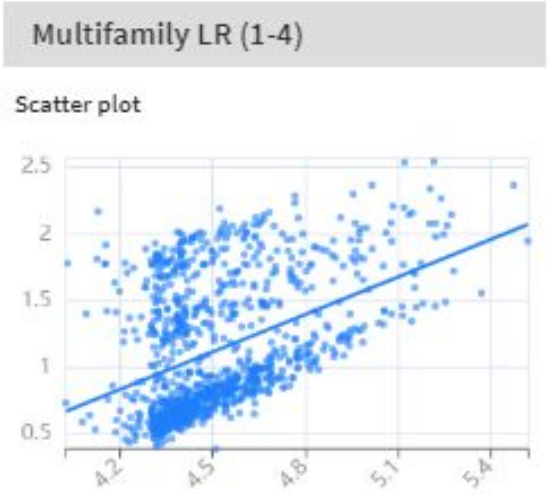
Avec Dataiku (TotalGHGEmission avec GazEnergy)

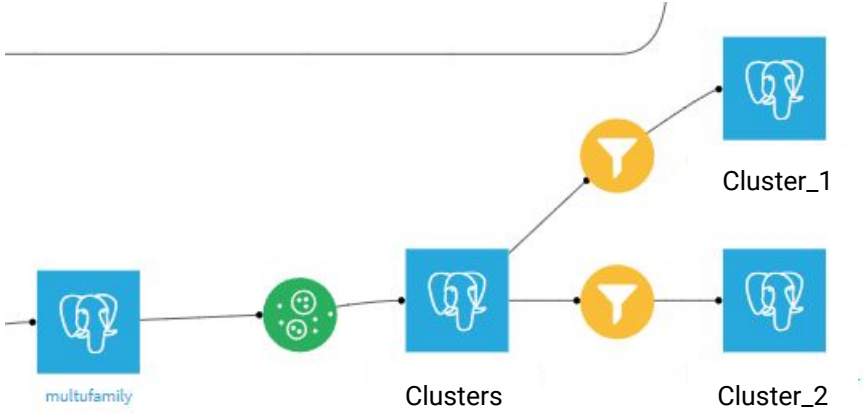




Clustering

Multifamily (1-4) issue de BuildingType



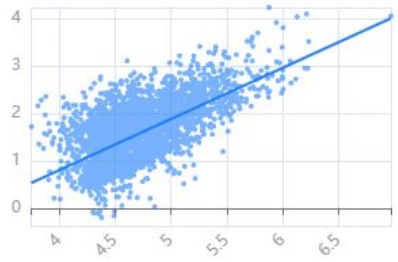


Clustering

Multifamily (1-4) issue de BuildingType

All

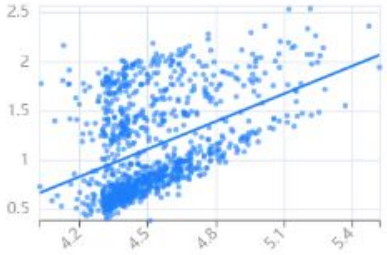
Scatter plot



Single L... 0.736 (± 0.022)

Multifamily LR (1-4)

Scatter plot



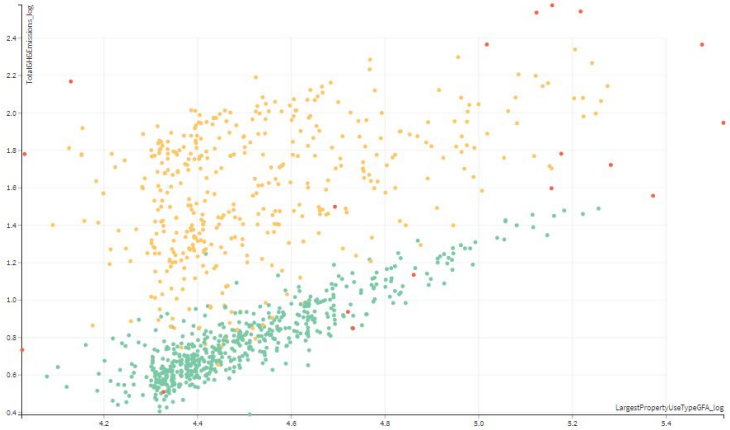
SMV (s1) 0,772

DBScan (s9)

DBScan

Formé en 0 seconde sur 996 enregistrements

grappe_0		19 (1,91 %)
cluster_1		411 (41,27%)
cluster_2		566 (56,83%)

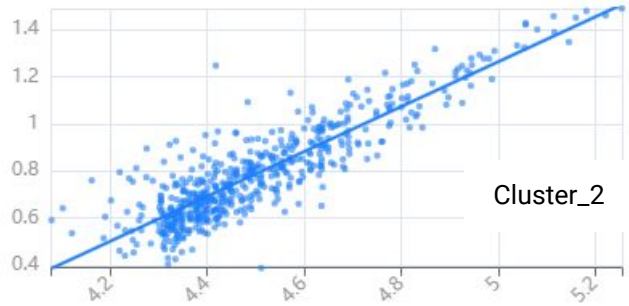


Clustering

Régressions sur les clusters

#TotalGHGEmissions_log par #LargestPro...

Nuage de points



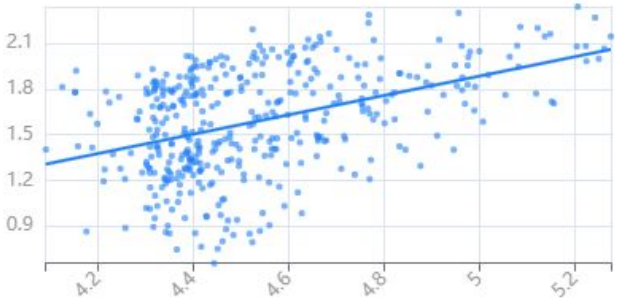
Cluster_2

Single Layer Per... 0.783 (± 0.104)

- 100% of the cluster has false for **Have_NaturalGas_Energy** (against 57.13 % globally)
- **TotalGHGEmissions_log** is in average 30.39% smaller : mean of 0.79 against 1.13 globally
- **LargestPropertyUseTypeGFA_log** is in average 0.43% smaller : mean of 4.50 against 4.52 globally

#TotalGHGEmissions_log par #LargestPro...

Nuage de points



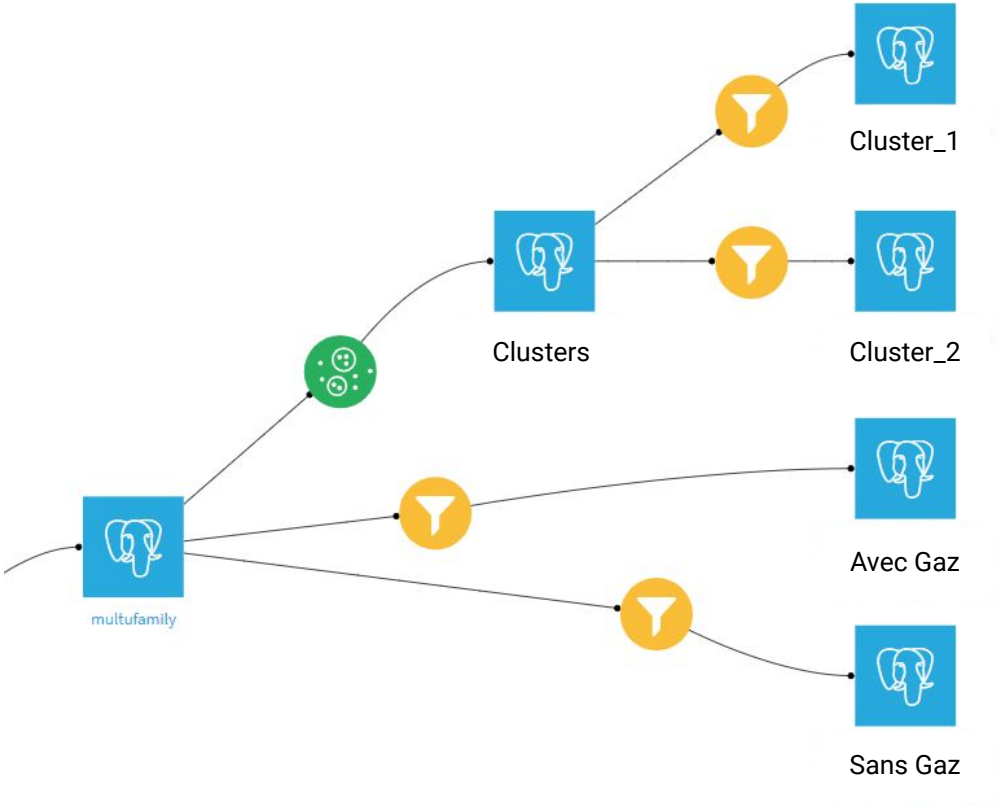
Cluster_1

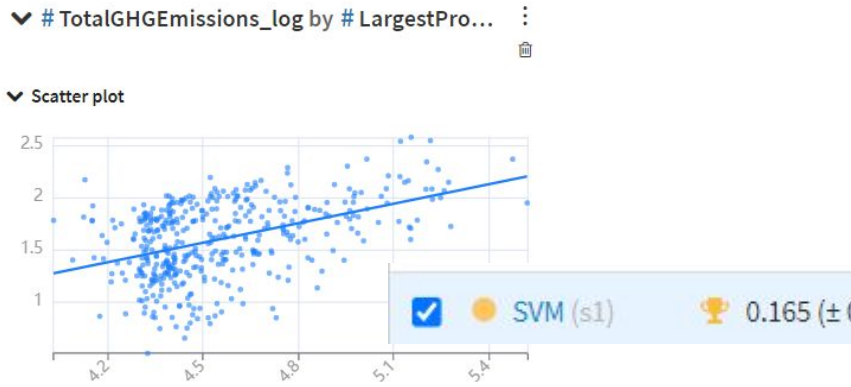
SVM (s2) 0.166 (± 0.089)

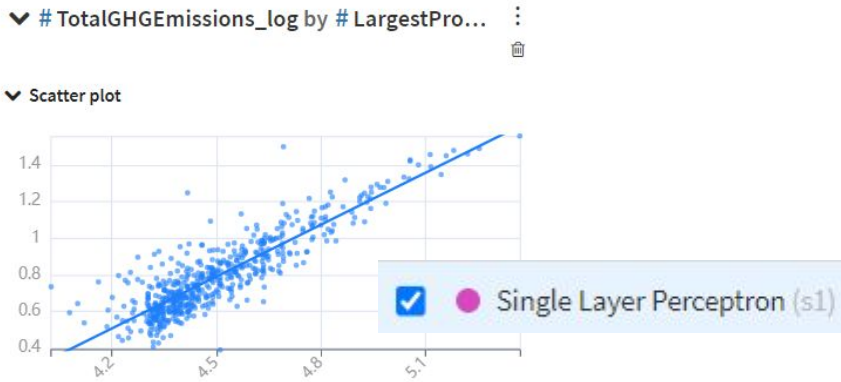
- 100% of the cluster has true for **Have_NaturalGas_Energy** (against 42.87 % globally)
- **TotalGHGEmissions_log** is in average 39.98% greater : mean of 1.59 against 1.13 globally
- **LargestPropertyUseTypeGFA_log** is in average 0.45% greater : mean of 4.54 against 4.52 globally

Prospection de modèles

Multifamily (1-4) issue de BuildingType, split avec ou sans gaz







Applications

Recommandations et retour d’expérience

Recommendations

- Limiter la superficie des nouveaux bâtiments
- Encourager l’utilisation des énergies alternatives au gaz
- Surveiller les bâtiments énergivores (campus & hôpitaux)
- Revoir le mode de notation EnergyStarScore qui est peu représentatif des émissions de CO2.

Retour d’expérience sur azure

- Facile à prendre en main
- Problèmes de rôles qui empéchent la gestion de certaines ressources
- Intégration à azure pose des problèmes de contrôle des données
- Azure ml : efficace mais le modèle ne pouvait pas être déployé
- Découverte d’outils intéressants : mlflow et interpretml

Des questions ?