ANTICIPEZ LES BESOINS EN CONSOMMATION ÉLECTRIQUE DE BÂTIMENTS

OPENCLASSROOMS - INGÉNIEUR MACHINE LEARNING

THIBAUD GROSJEAN - NOVEMBRE 2021

PRÉSENTATION DU PROJET

- Mission pour la ville de Seattle et son *open data*
  - Relevés des consommations d'énergie et d'émissions de gaz à effet de serre
  - Bâtiments non destinés à l'habitation
- Simplification du processus de relevé
  - Prédiction des consommations & émissions
  - Evaluation de l'impact de l'Energy Star Score sur les prédictions d'émissions

MÉTHODOLOGIE

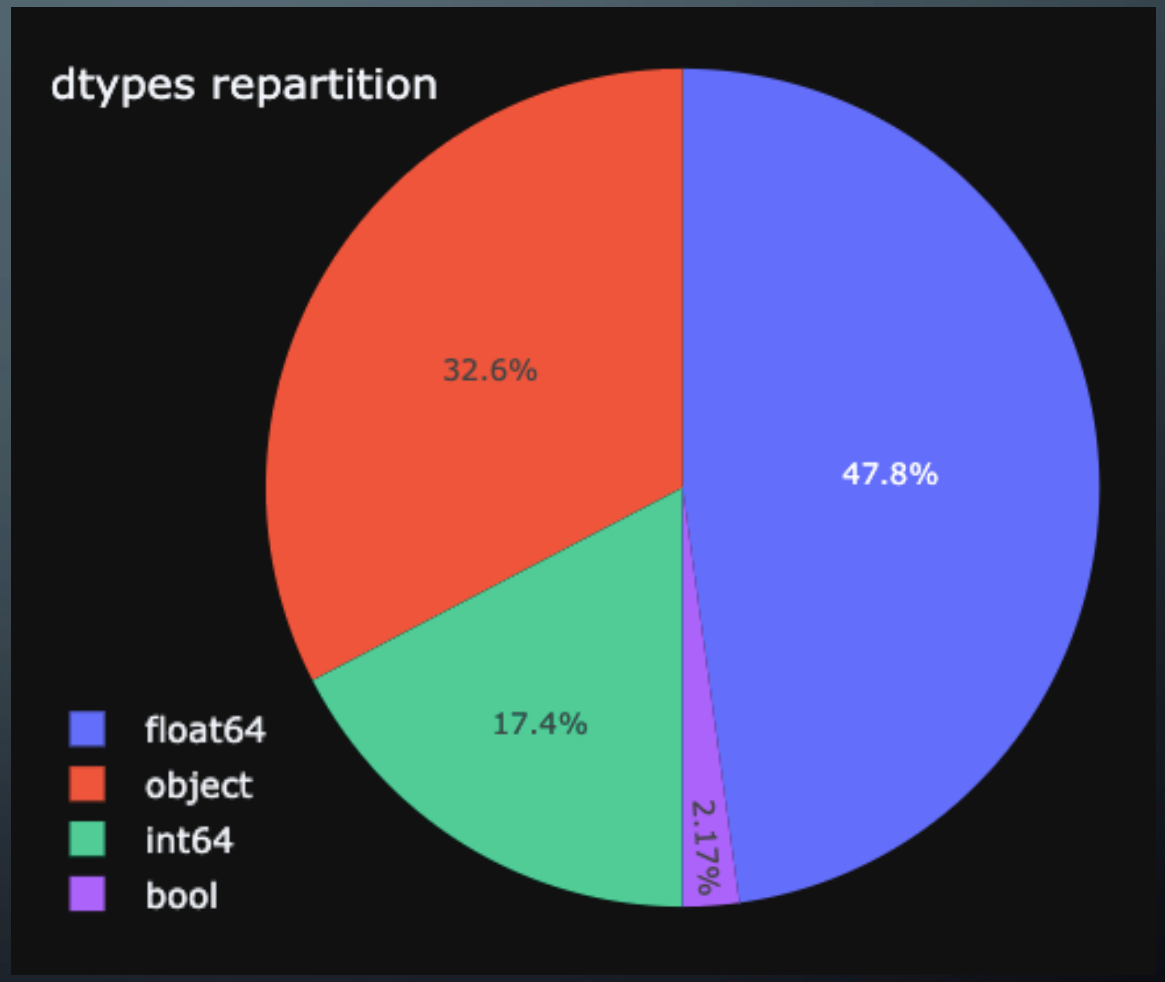
- Description du jeu de données
- Analyse exploratoire
- Feature Engineering
- Machine Learning

DESCRIPTION DU JEU DE DONNÉES

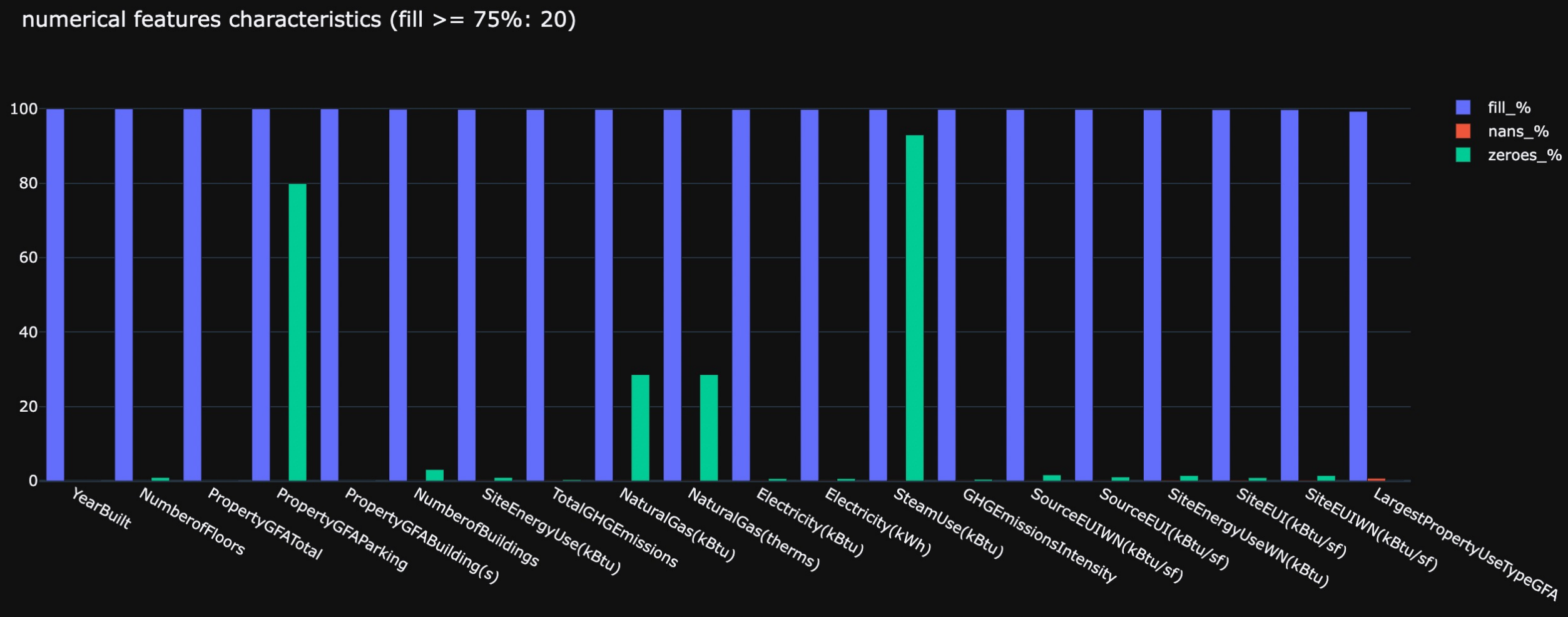
DESCRIPTION DU JEU DE DONNÉES

CARACTÉRISTIQUES DU JEU DE DONNÉES

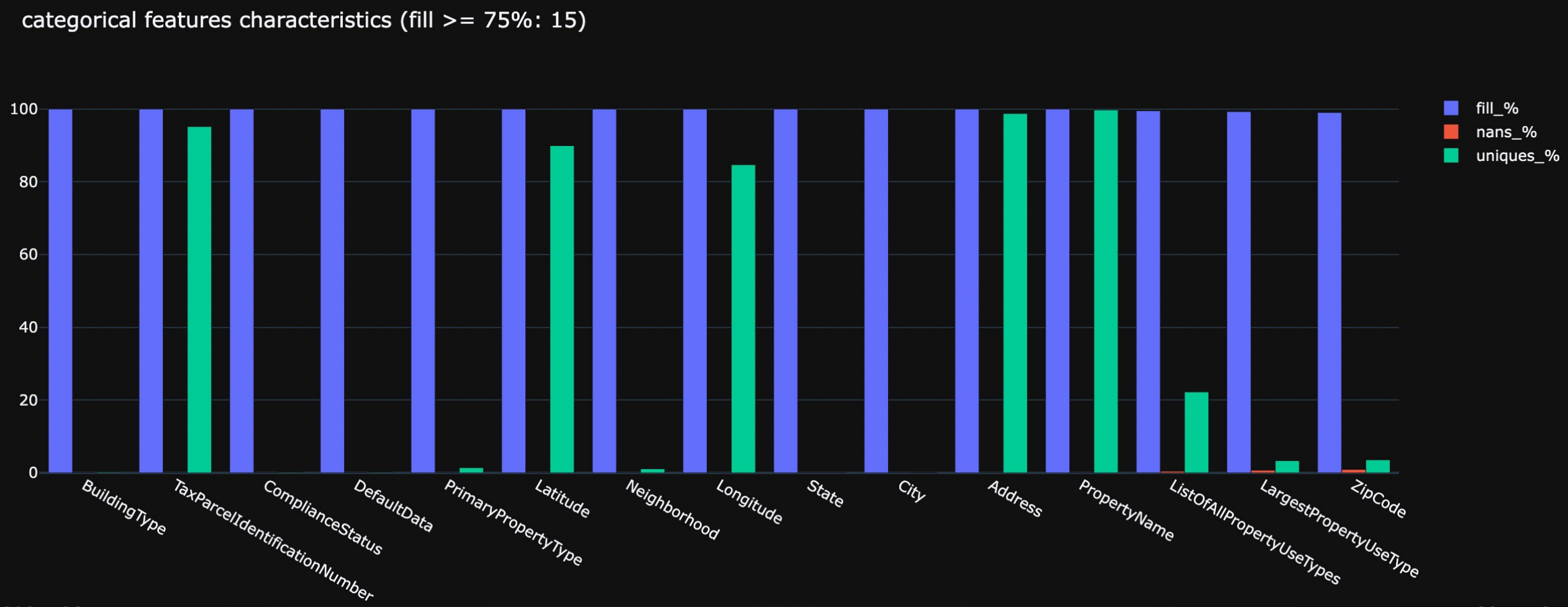
- 1698 individus
- 46 variables
  - Dont 30 variables numériques
    - Dont 22 variables continues
    - Et 8 variables discrètes
  - Et 16 variables qualitatives
- Valeurs manquantes : 13,09 %
- Valeurs uniques : 40.64 %



VARIABLES NUMÉRIQUES



VARIABLES QUALITATIVES

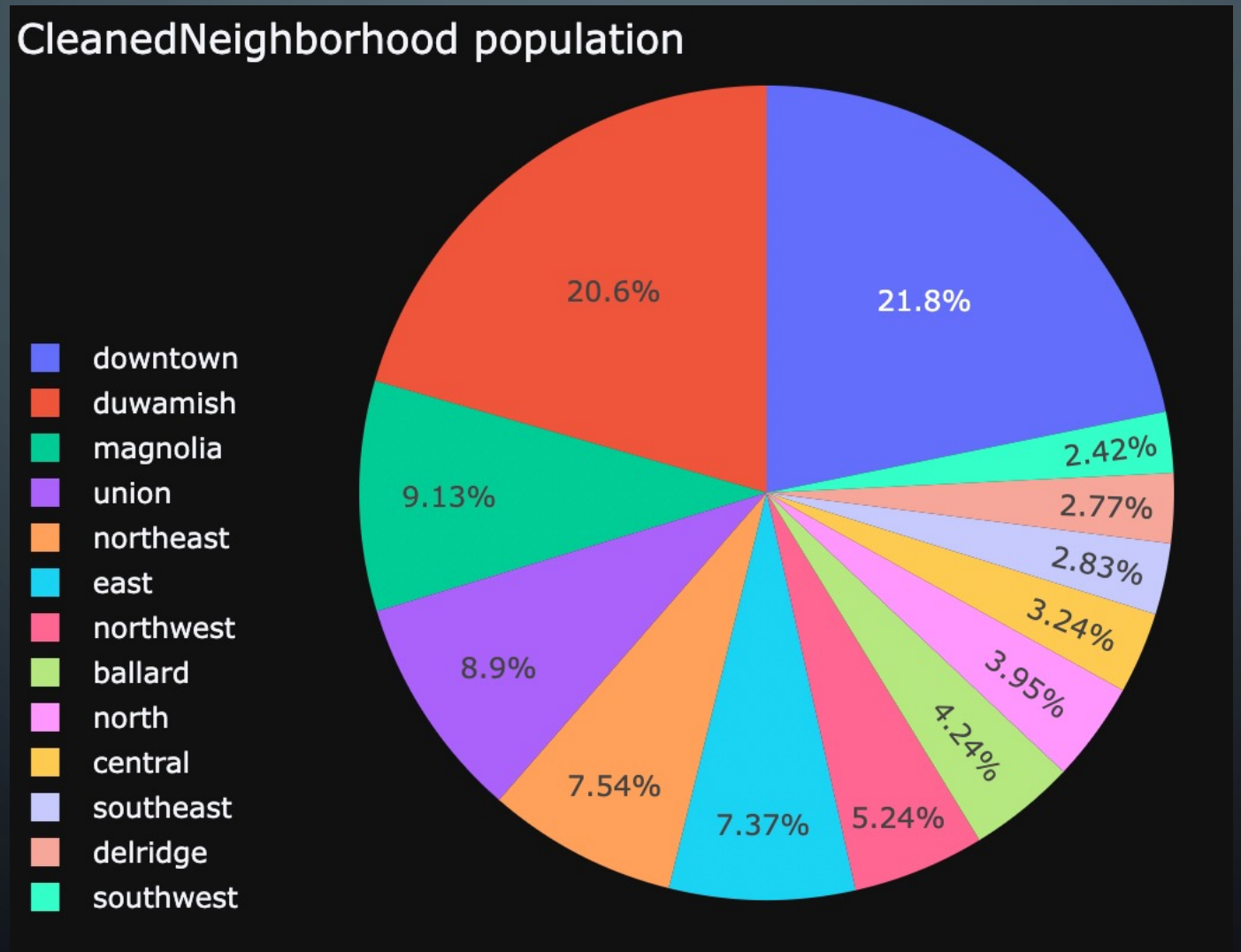


DONNÉES PRÉSENTES

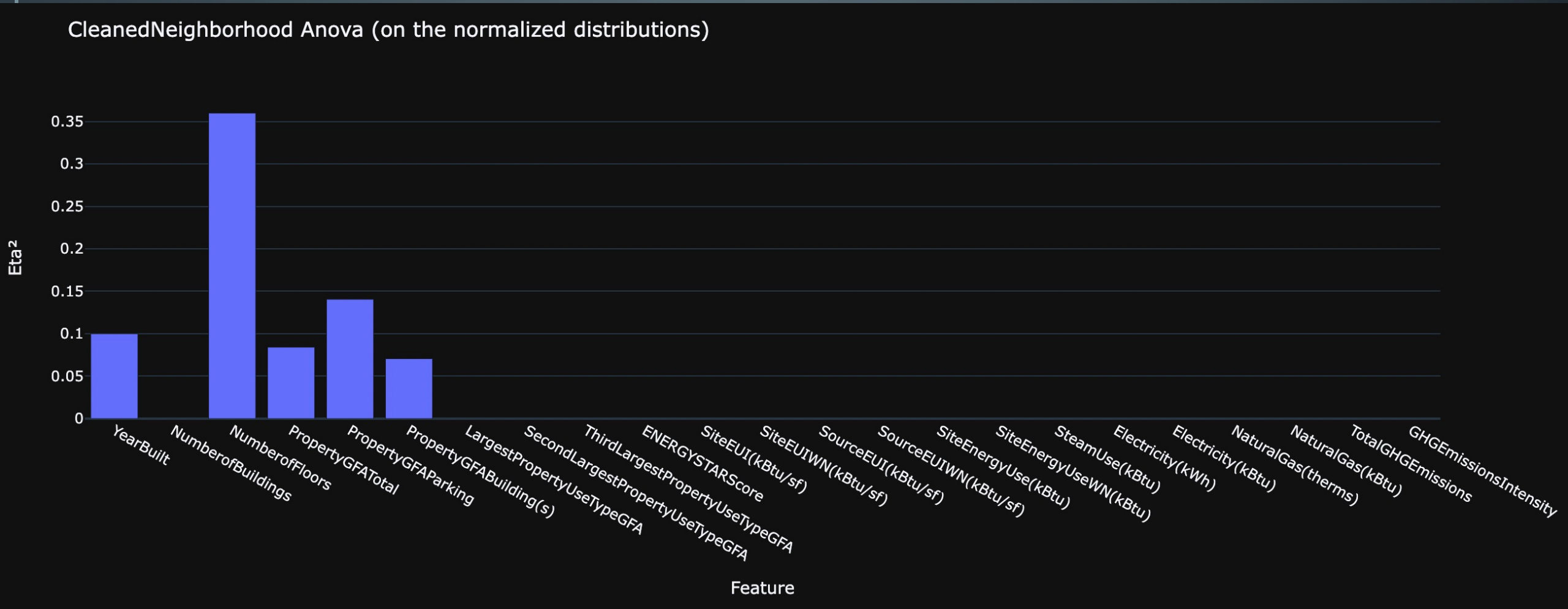
- Méta-données
- Données administratives
- Données de localisation
- Données structurelles
- Données d'usage
- Données cibles

ANALYSE EXPLORATOIRE

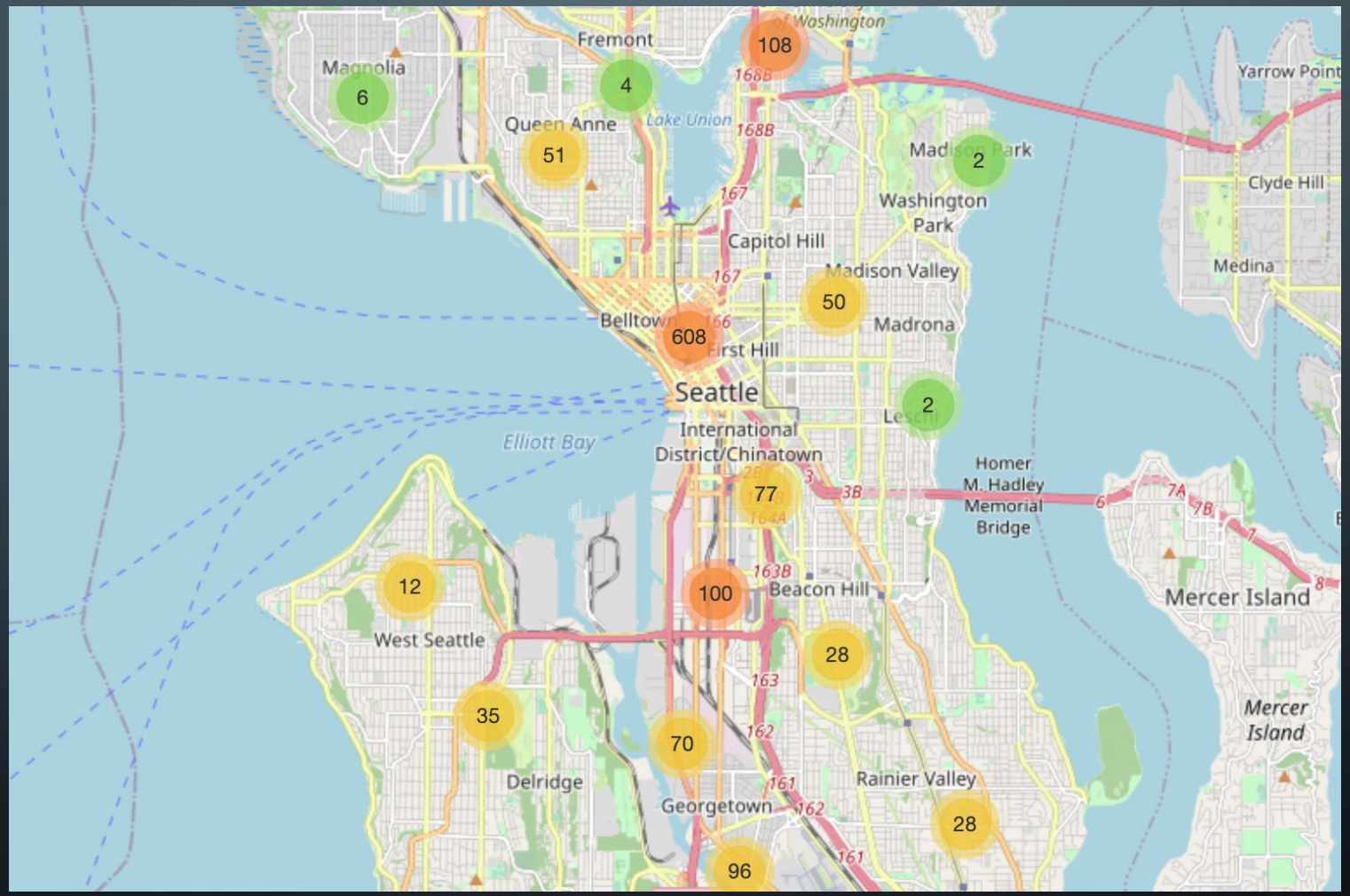
DONNÉES DE LOCALISATION



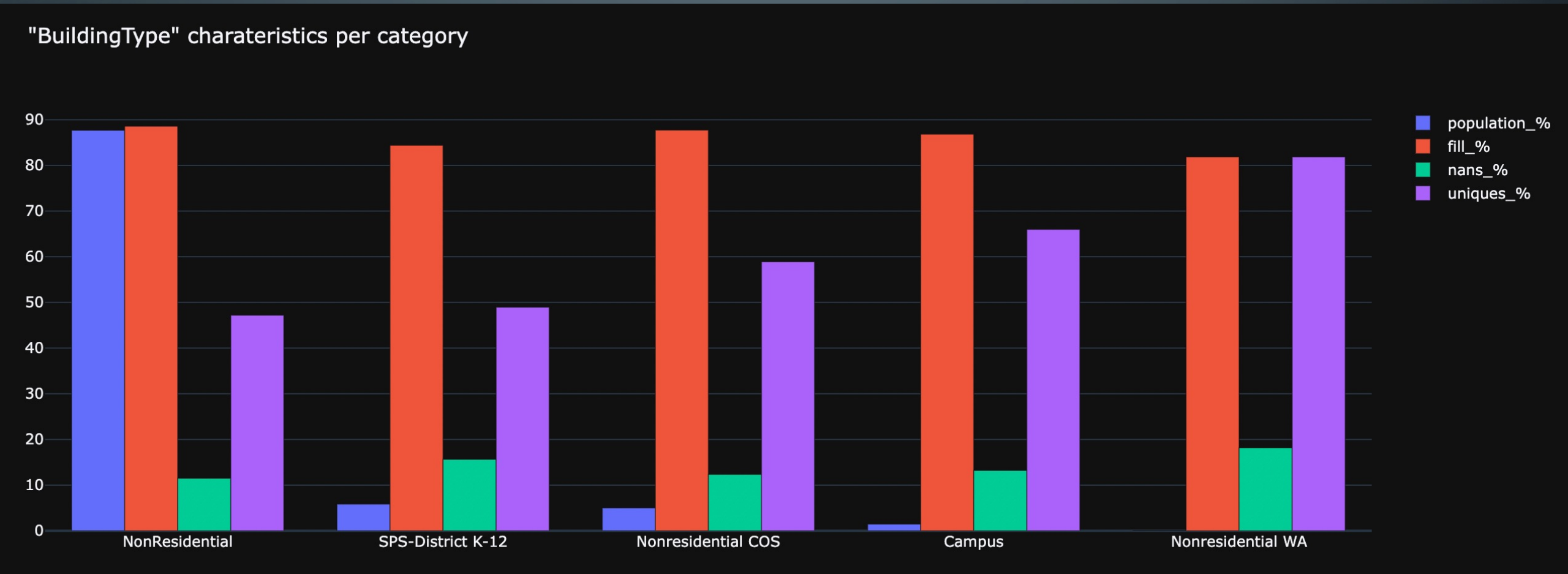
DONNÉES DE LOCALISATION



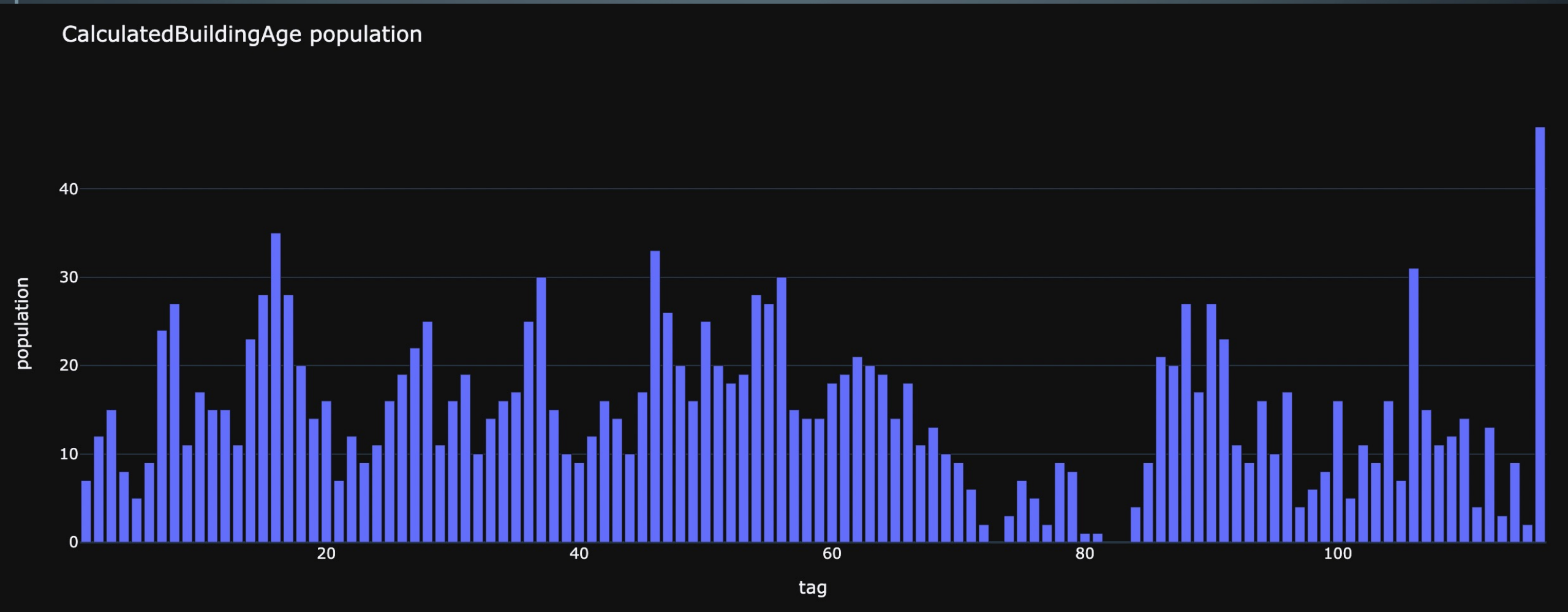
DONNÉES DE LOCALISATION



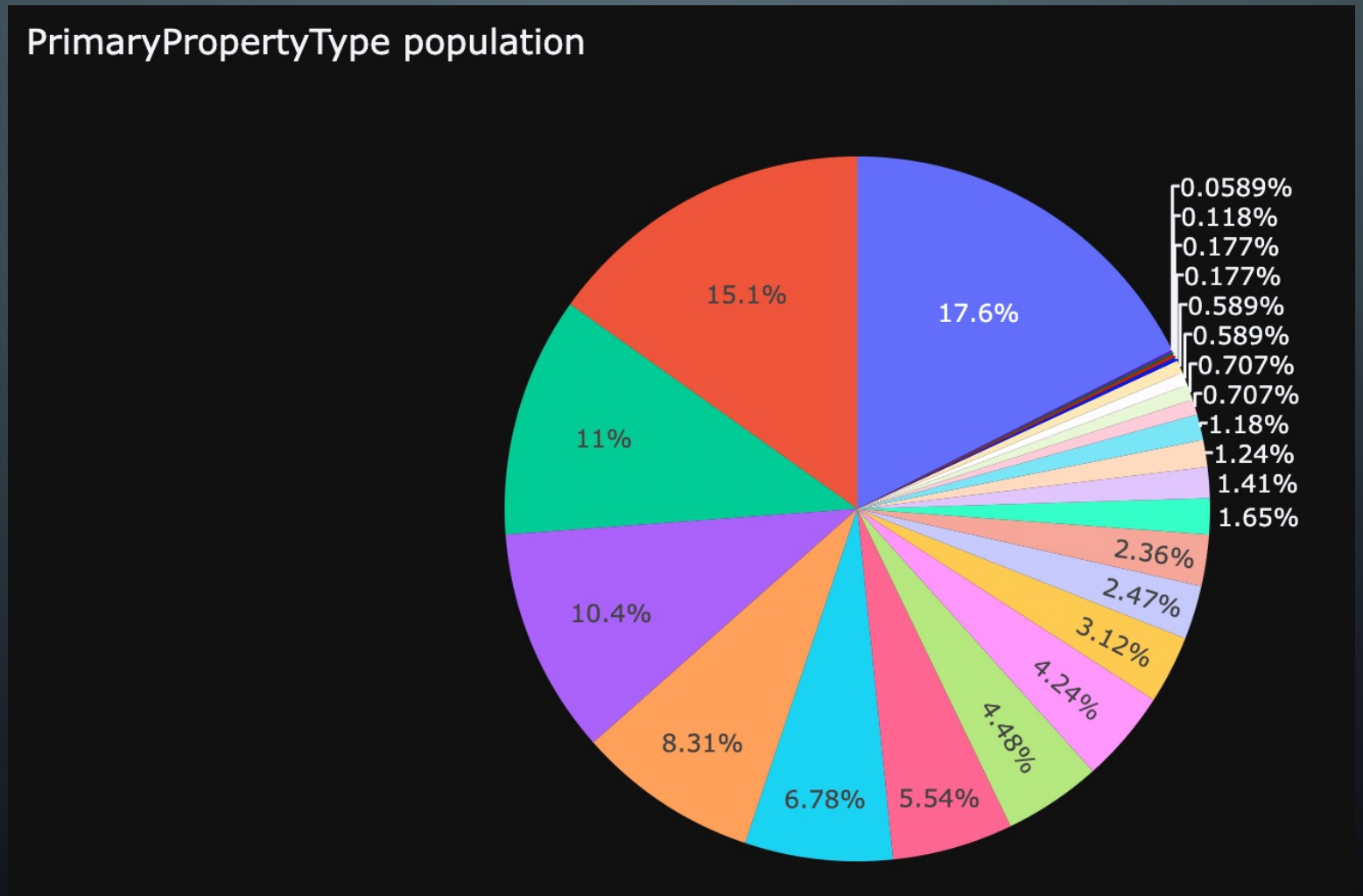
DONNÉES STRUCTURELLES



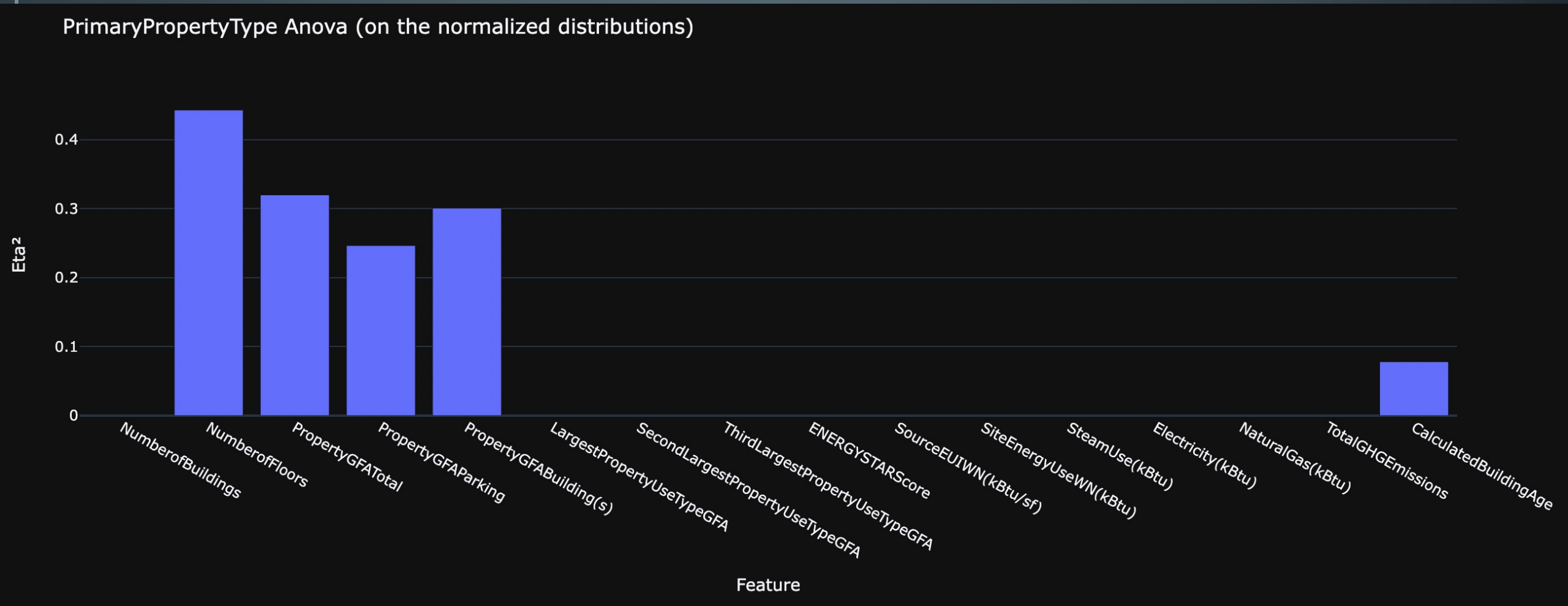
DONNÉES STRUCTURELLES



DONNÉES D'USAGE

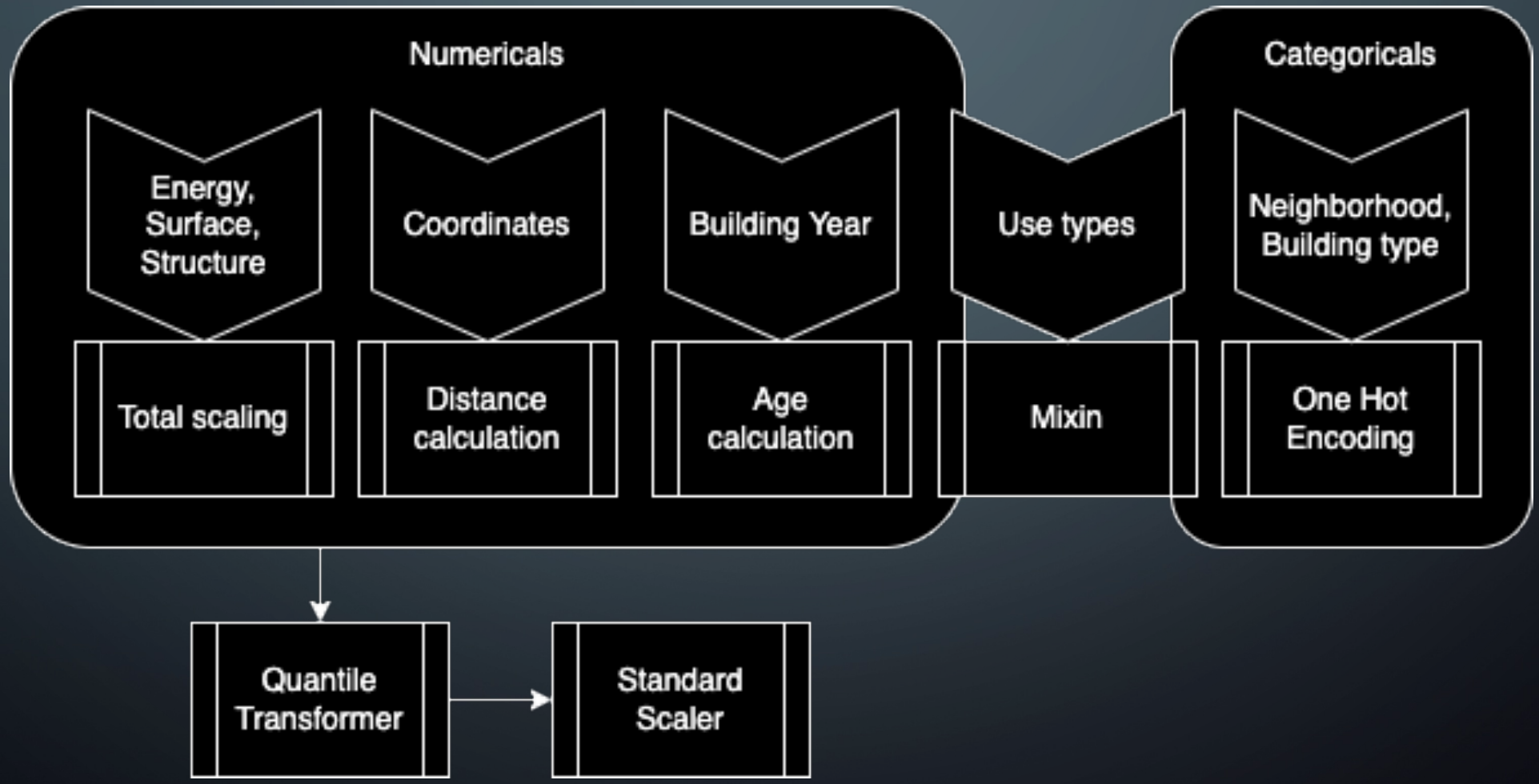


DONNÉES D'USAGE

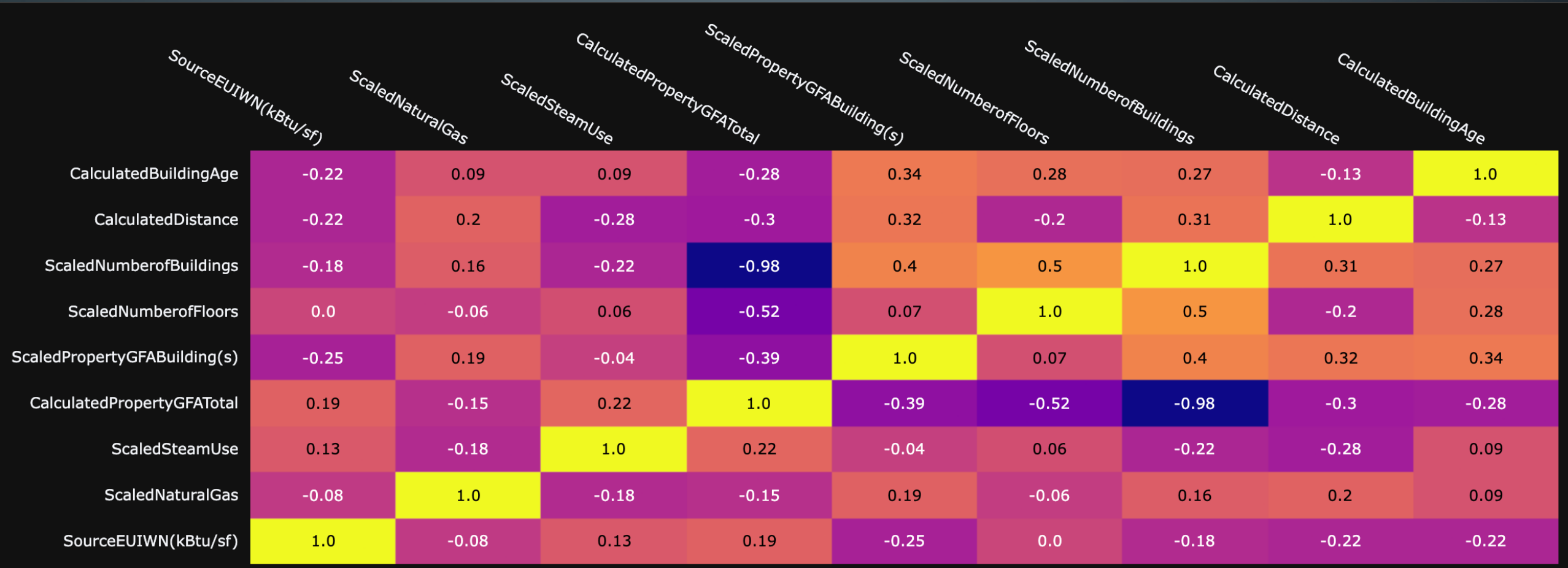


FEATURE ENGINEERING

TRANSFORMERS

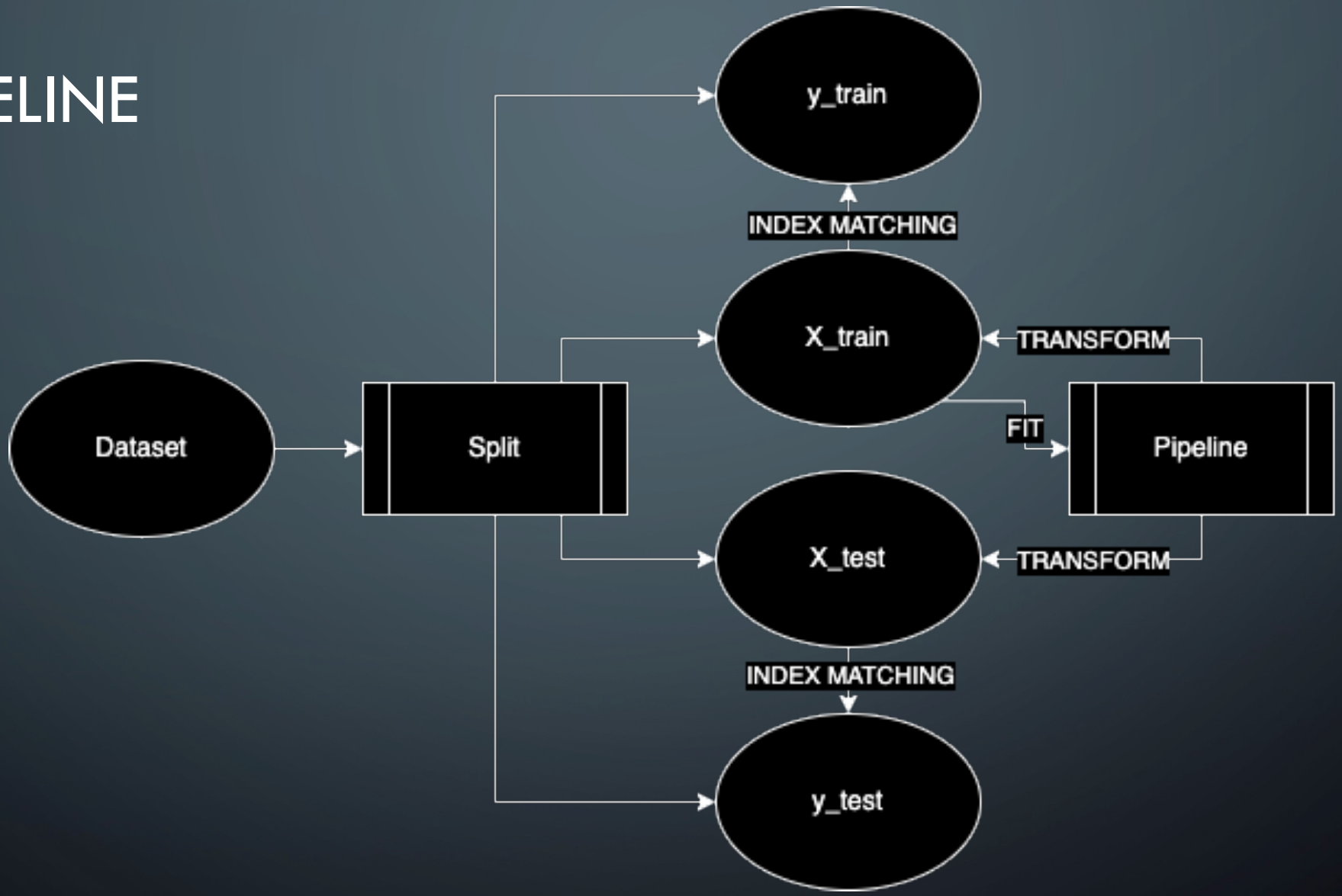


CORRELATIONS



	SourceEUIWN(kBtu/sf)	ScaledNaturalGas	ScaledSteamUse	CalculatedPropertyGFATotal	ScaledPropertyGFABuilding(s)	ScaledNumberofFloors	ScaledNumberofBuildings	CalculatedDistance	CalculatedBuildingAge
CalculatedBuildingAge	-0.22	0.09	0.09	-0.28	0.34	0.28	0.27	-0.13	1.0
CalculatedDistance	-0.22	0.2	-0.28	-0.3	0.32	-0.2	0.31	1.0	-0.13
ScaledNumberofBuildings	-0.18	0.16	-0.22	-0.98	0.4	0.5	1.0	0.31	0.27
ScaledNumberofFloors	0.0	-0.06	0.06	-0.52	0.07	1.0	0.5	-0.2	0.28
ScaledPropertyGFABuilding(s)	-0.25	0.19	-0.04	-0.39	1.0	0.07	0.4	0.32	0.34
CalculatedPropertyGFATotal	0.19	-0.15	0.22	1.0	-0.39	-0.52	-0.98	-0.3	-0.28
ScaledSteamUse	0.13	-0.18	1.0	0.22	-0.04	0.06	-0.22	-0.28	0.09
ScaledNaturalGas	-0.08	1.0	-0.18	-0.15	0.19	-0.06	0.16	0.2	0.09
SourceEUIWN(kBtu/sf)	1.0	-0.08	0.13	0.19	-0.25	0.0	-0.18	-0.22	-0.22

PIPELINE


y_train
INDEX MATCHING
X_train
TRANSFORM
Dataset
Split
FIT
Pipeline
X_test
TRANSFORM
INDEX MATCHING
y_test

```
[('FeaturesFilter',
  FunctionTransformer(func=<function FeaturesFilter at 0x7f7fb83124d0>)),
 ('SamplesFilter',
  FunctionTransformer(func=<function SamplesFilters at 0x7f7fb8308b00>)),
 ('CampusFilter',
  FunctionTransformer(func=<function CampusFilter at 0x7f7fb8312560>)),
 ('EnergyStarFilter',
  FunctionTransformer(func=<function EnergyStarFilter at 0x7f7fb8312830>)),
 ('EnergyTransformer',
  FunctionTransformer(func=<function EnergyTransformer at 0x7f7fb8403200>)),
 ('StructureTransformer',
  FunctionTransformer(func=<function StructureTransformer at 0x7f7fb7fe0b00>)),
 ('DistanceCalculator',
  FunctionTransformer(func=<function DistanceCalculator at 0x7f7fb7fe0e60>)),
 ('AgeCalculator',
  FunctionTransformer(func=<function AgeCalculator at 0x7f7fb8308ef0>)),
 ('NeighborhoodCleaner',
  FunctionTransformer(func=<function NeighborhoodCleaner at 0x7f7fb6fa0050>)),
 ('CatEncoder', CatEncoder(features=['CleanedNeighborhood', 'BuildingType'])),
 ('TypeEncoder',
  TypeEncoder(features=['LargestPropertyUseType', 'SecondLargestPropertyUseType',
                        'ThirdLargestPropertyUseType'],
              frequencies=['LargestPropertyUseTypeGFA',
                           'SecondLargestPropertyUseTypeGFA',
                           'ThirdLargestPropertyUseTypeGFA'])),
 ('QuantTransform', QuantTransform()),
 ('StdScaler', StdScaler())]
```

PIPELINE

MACHINE LEARNING

CONSOMATION D'ÉNERGIE

MÉTHODOLOGIE

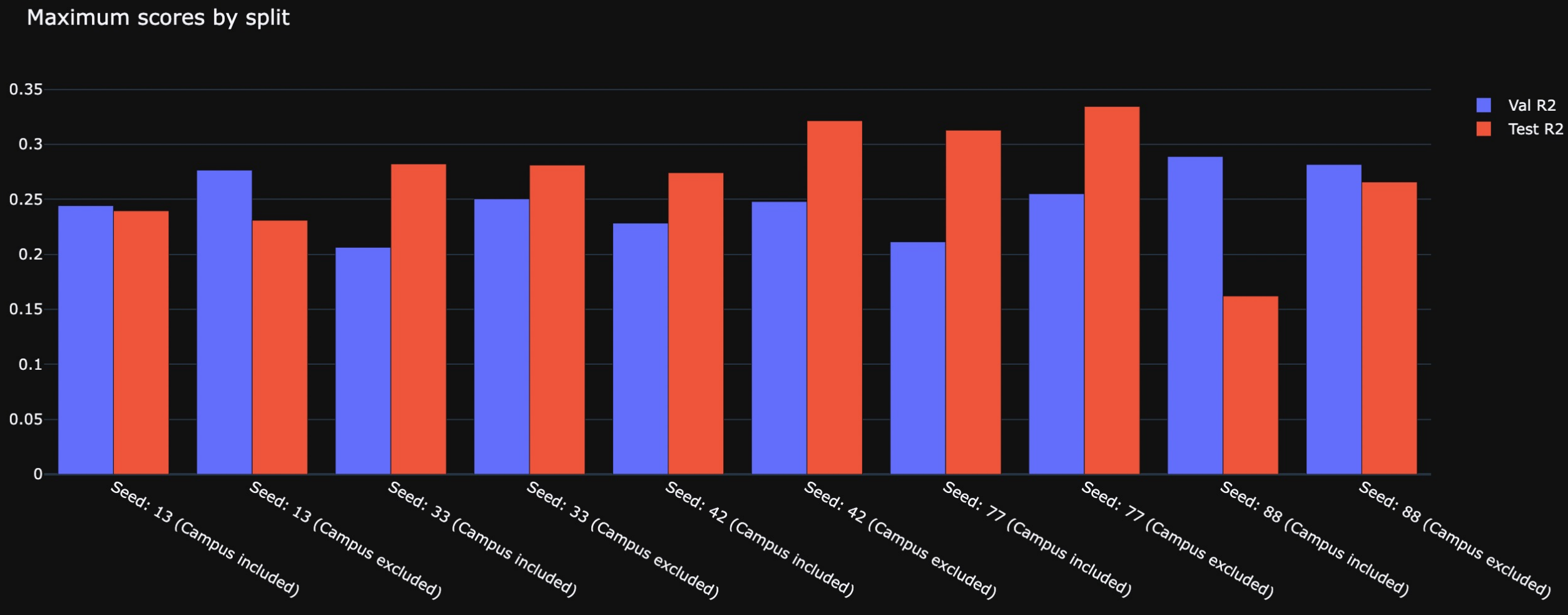
MÉTHODOLOGIE

```
metric_1 = 'neg_mean_absolute_error'
metric_2 = 'r2'
```

```
model = XGBRegressor(objective ='reg:squarederror')
grid, model_best = grid_search(model, X_train, y_train, metrics=metrics,
  n_estimators=[100, 200, 300],
  max_depth=[5, 20, 50],
  min_samples_leaf=[1, 2, 5, 10],
  colsample_bytree=[0.25, 0.5, 0.75, 1]
  )
```

ALGORITHMES

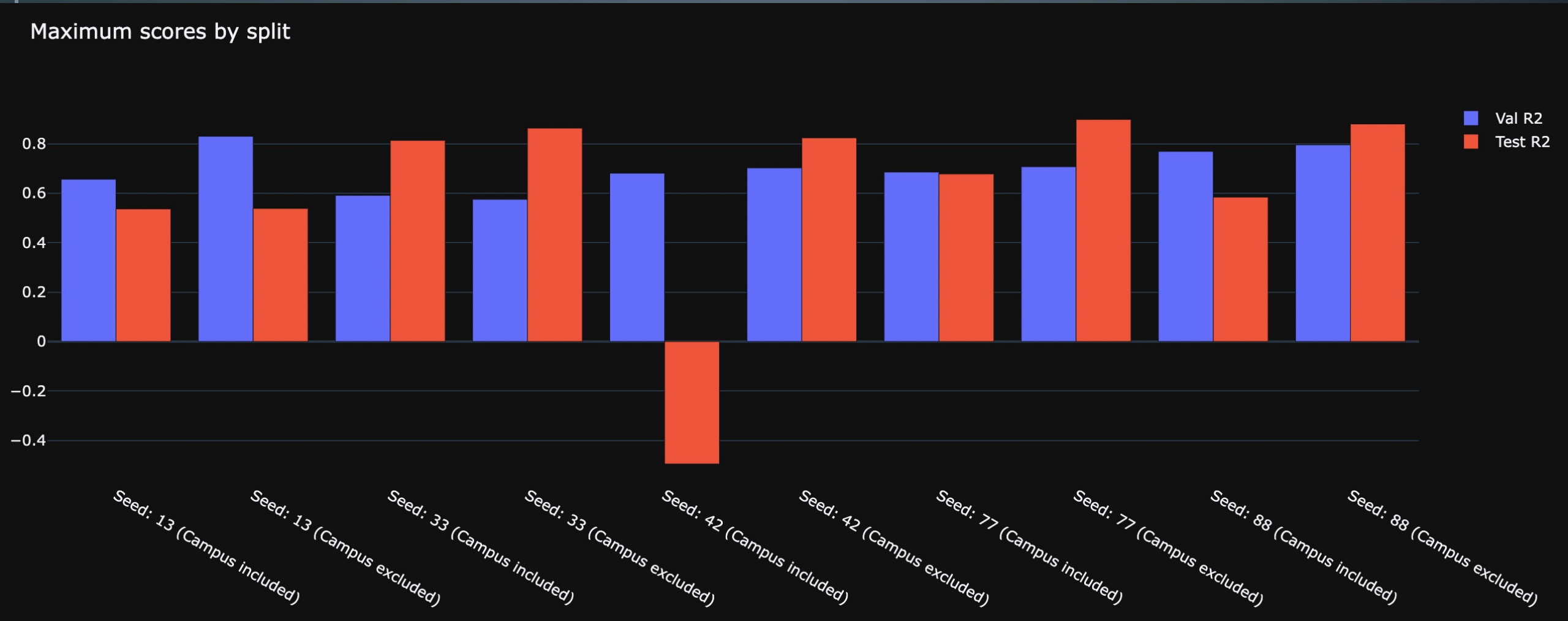
CAMPUS & SPLIT



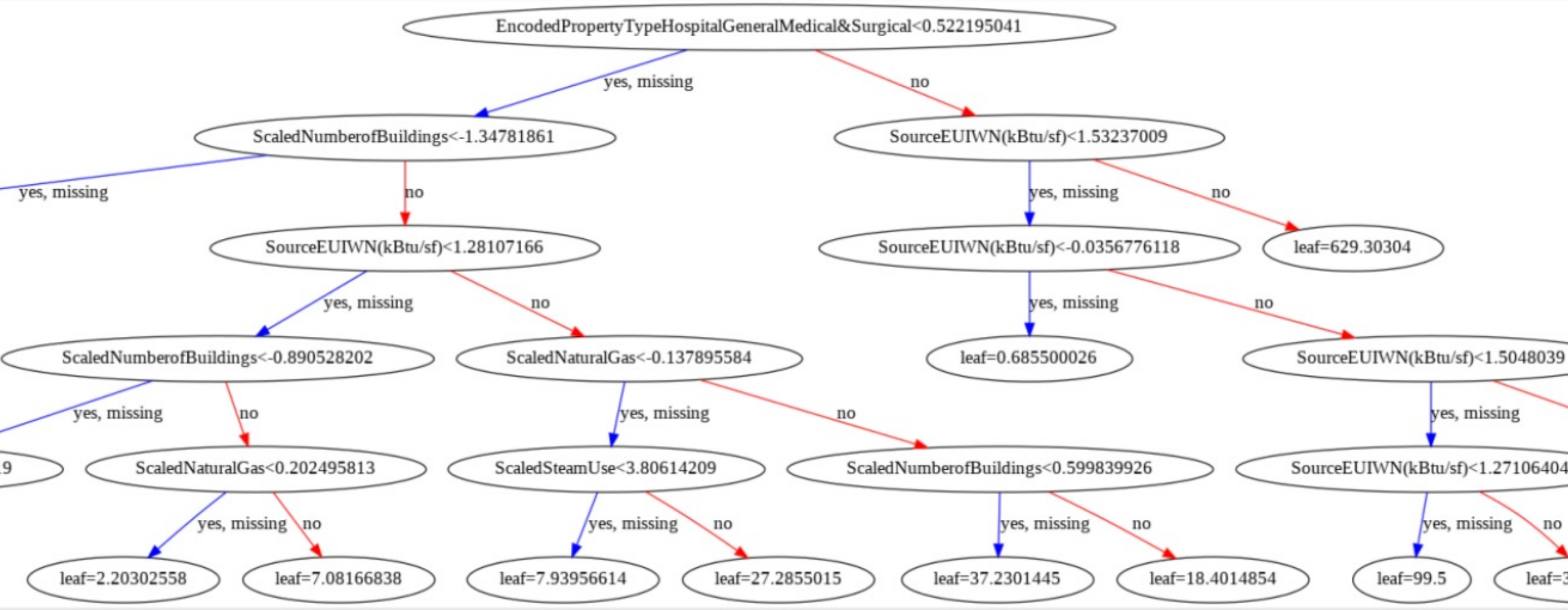
RÉSULTATS

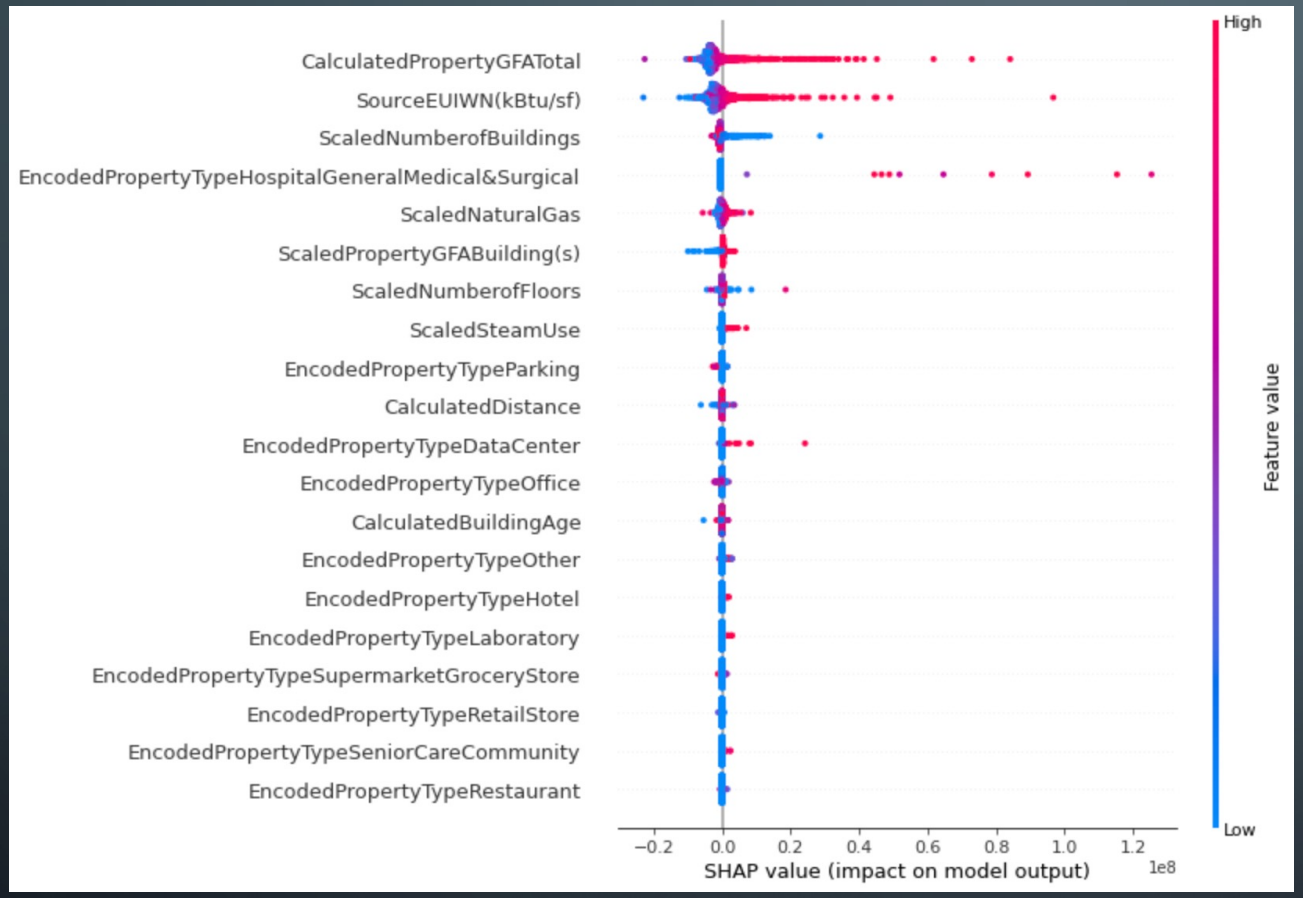
Algorithm	Mean fit time	Std fit time	Mean score time	Std score time	Params	Mean test neg mean absolute error	Std test neg mean absolute error	Rank test neg mean absolute error	Mean test r2	Std test r2	Rank test r2
Mean Dummy Regressor	0.001802	0.000213	0.001231	0.000088	{}	-9.096172e+06	9.135289e+05	1	-5.053937e-03	4.565300e-03	1
Median Dummy Regressor	0.002222	0.000517	0.001281	0.000341	{}	-7.016261e+06	1.234512e+06	1	-7.255148e-02	2.269244e-02	1
Linear Regression	0.025132	0.006169	0.009015	0.003559	{'fit_intercept': False}	-7.267874e+16	7.512659e+16	2	-1.890239e+21	1.597899e+21	1
Linear Regression (subset)	0.007464	0.003206	0.003426	0.000149	{'fit_intercept': True}	-8.539746e+06	5.351346e+05	1	2.283483e-01	5.485025e-02	1
Elastic Net	0.068964	0.015485	0.006432	0.002094	{'alpha': 0.004750810162102793}	-5.683225e+06	5.938999e+05	41	4.258168e-01	7.844371e-02	1
Support Vector Machine	0.007108	0.000119	0.004306	0.000044	{'C': 0.05623413251903491, 'epsilon': 1, 'loss...	-5.829721e+06	6.608362e+05	49	4.011776e-01	7.980976e-02	1
Random Forest	0.862856	0.018728	0.016827	0.003258	{'bootstrap': True, 'max_depth': 20, 'min_samp...	-2.026343e+06	8.099661e+05	1	7.837507e-01	1.526679e-01	1
XGBoost Regressor	0.893621	0.003445	0.007244	0.000311	{'colsample_bytree': 0.75, 'max_depth': 5, 'mi...	-1.871089e+06	6.710531e+05	1	7.956711e-01	1.178603e-01	1

CAMPUS & SPLIT



EXPLICABILITÉ





EXPLICABILITÉ

MACHINE LEARNING

EMISSIONS DE GAZ À EFFET DE SERRE

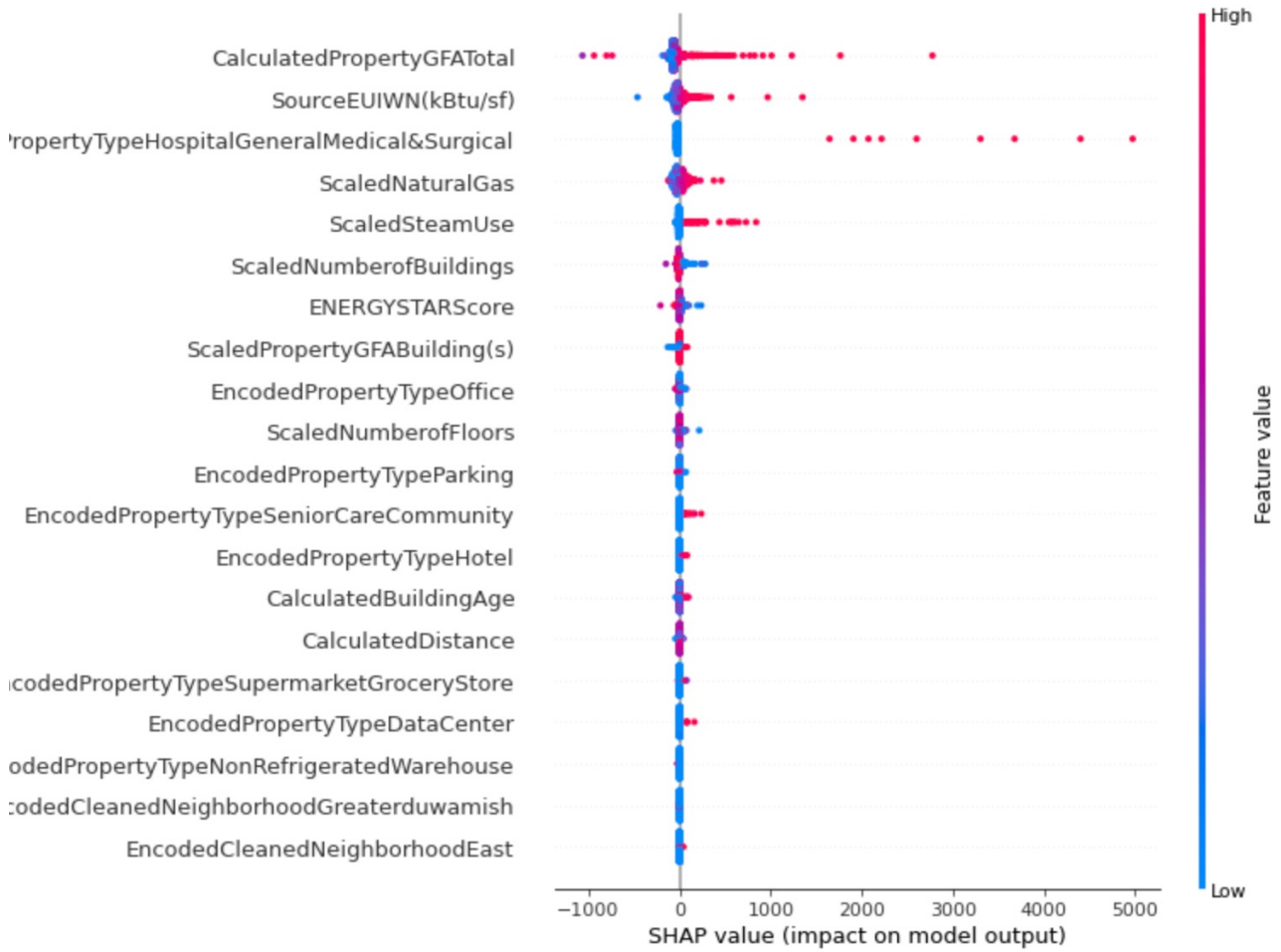
RÉSULTATS

Algorithm	Mean fit time	Std fit time	Mean score time	Std score time	Params	Mean test neg mean absolute error	Std test neg mean absolute error	Rank test neg mean absolute error	Mean test r2	Std test r2	Rank test r2
Mean Dummy Regressor	0.001503	0.000500	0.000846	0.000300	{}	-1.678015e+02	2.672511e+01	1	-2.610612e-02	2.686547e-02	1
Median Dummy Regressor	0.001472	0.000558	0.000652	0.000064	{}	-1.315795e+02	3.500447e+01	1	-8.998440e-02	4.005828e-02	1
Linear Regression	0.012656	0.000133	0.004537	0.000143	{'fit_intercept': False}	-2.888345e+11	3.566367e+11	1	-6.639459e+20	9.734236e+20	1
Linear Regression (subset)	0.004787	0.001560	0.005365	0.003442	{'fit_intercept': True}	-1.569024e+02	1.296976e+01	1	7.926226e-02	1.066833e-01	1
Elastic Net	0.051906	0.025001	0.005625	0.001868	{'alpha': 0.0029836472402833404}	-1.305567e+02	1.489781e+01	36	3.619571e-01	1.561828e-01	1
Support Vector Machine	0.016456	0.000227	0.004458	0.000100	{'C': 1.0, 'epsilon': 2, 'loss': 'squared_epsi...	-1.255647e+02	1.550445e+01	56	3.715127e-01	1.923181e-01	1
Random Forest	1.060939	0.011716	0.017169	0.001369	{'bootstrap': True, 'max_depth': 20, 'min_samp...	-5.660990e+01	2.329665e+01	1	6.777523e-01	1.316636e-01	1
XGBoost Regressor	1.110912	0.006605	0.006897	0.000102	{'colsample_bytree': 1, 'max_depth': 5, 'min_s...	-5.148527e+01	2.327103e+01	5	7.468652e-01	1.429198e-01	1

IMPACT DE L'ENERGY STAR SCORE

Algorithm	Mean fit time	Std fit time	Mean score time	Std score time	Params	Mean test neg mean absolute error	Std test neg mean absolute error	Rank test neg mean absolute error	Mean test r2	Std test r2	Rank test r2
XGBoost Regressor	0.623358	0.003035	0.005272	0.000100	{'n_estimators': 300, 'min_samples_leaf': 5, '...	-59.423369	31.925333	10	0.747189	0.14041	1
StarScore XGBoost Regressor	0.514480	0.004906	0.005467	0.000036	{'n_estimators': 300, 'min_samples_leaf': 2, '...	-58.557476	29.985112	1	0.761383	0.09813	1

IMPACT DE L'ENERGY STAR SCORE



CONCLUSION

- Entrainement et optimisation des modèles
- Preuve de concept
- Axes d'amélioration
  - Stratification des splits
  - Modèles plus avancés