

# **Prédiction de la Consommation d'Energie des Bâtiments de Seattle**

Réalisé par:

**Marie Agathe SECK**

**Diakhou NDAO**

**Pathé DIAGNE**

**Fallou BADJI**

Formatrice :

**Mme Mously DIAW**

# INTRODUCTION



**SEATTLE:** ville de l'Etat de Washington, Nord-Ouest des USA;

**Superficie:** 369,2 km<sup>2</sup> ;

**Habitants :** 737 015 en 2020

## CONTEXTE

La ville de Seattle s'intéresse de près aux émissions des bâtiments non destinés à l'habitation. En 2013, la ville de Seattle a adopté un plan d'action climatique visant à atteindre zéro émission nette de gaz à effet de serre (GES) d'ici 2050.

### Objectif et Mission

Notre équipe s'intéresse à la prédiction de la consommation totale d'énergie des bâtiments non destinés à l'habitation.

De manière spécifique on se chargera de :

- ✓ Réaliser une analyse exploratoire.
- ✓ Tester différents modèles de prédiction afin de répondre au mieux à la problématique.
- ✓ Mettre en place une application pour prédire la consommation d'un bâtiment.

# PLAN



## INTRODUCTION

---

**01**

Analyse exploratoire des données

**02**

Apurement des données

**03**

Modélisation

**04**

Choix du modèle définitif

---

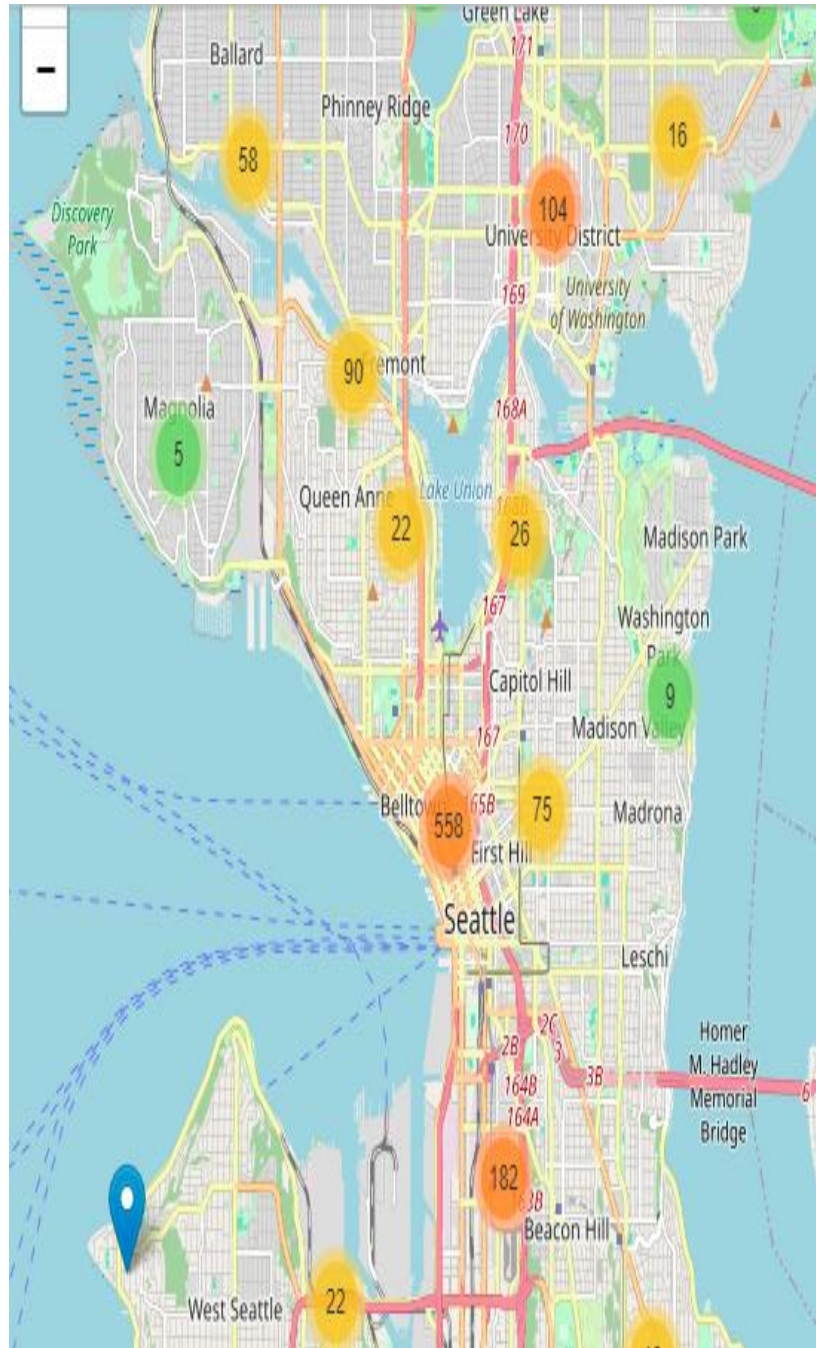
Déploiement et test de l'application de prédiction

**1**

# **Analyse exploratoire**



# Statistiques globales



3376

bâtiments

1540

Bâtiments non destinés à l'Habitation

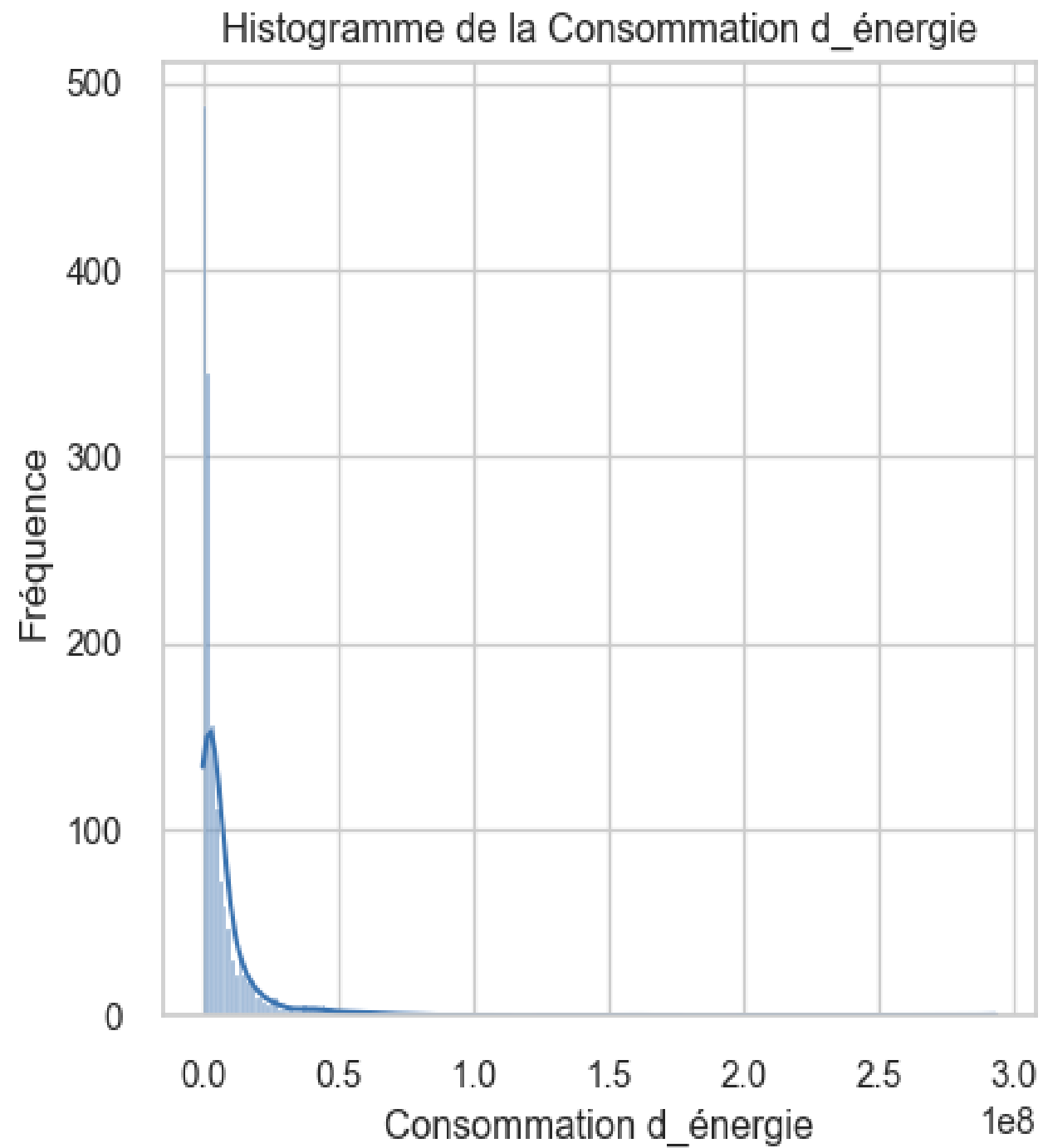
5.403667 KBtu

Consommation moyenne d'énergie

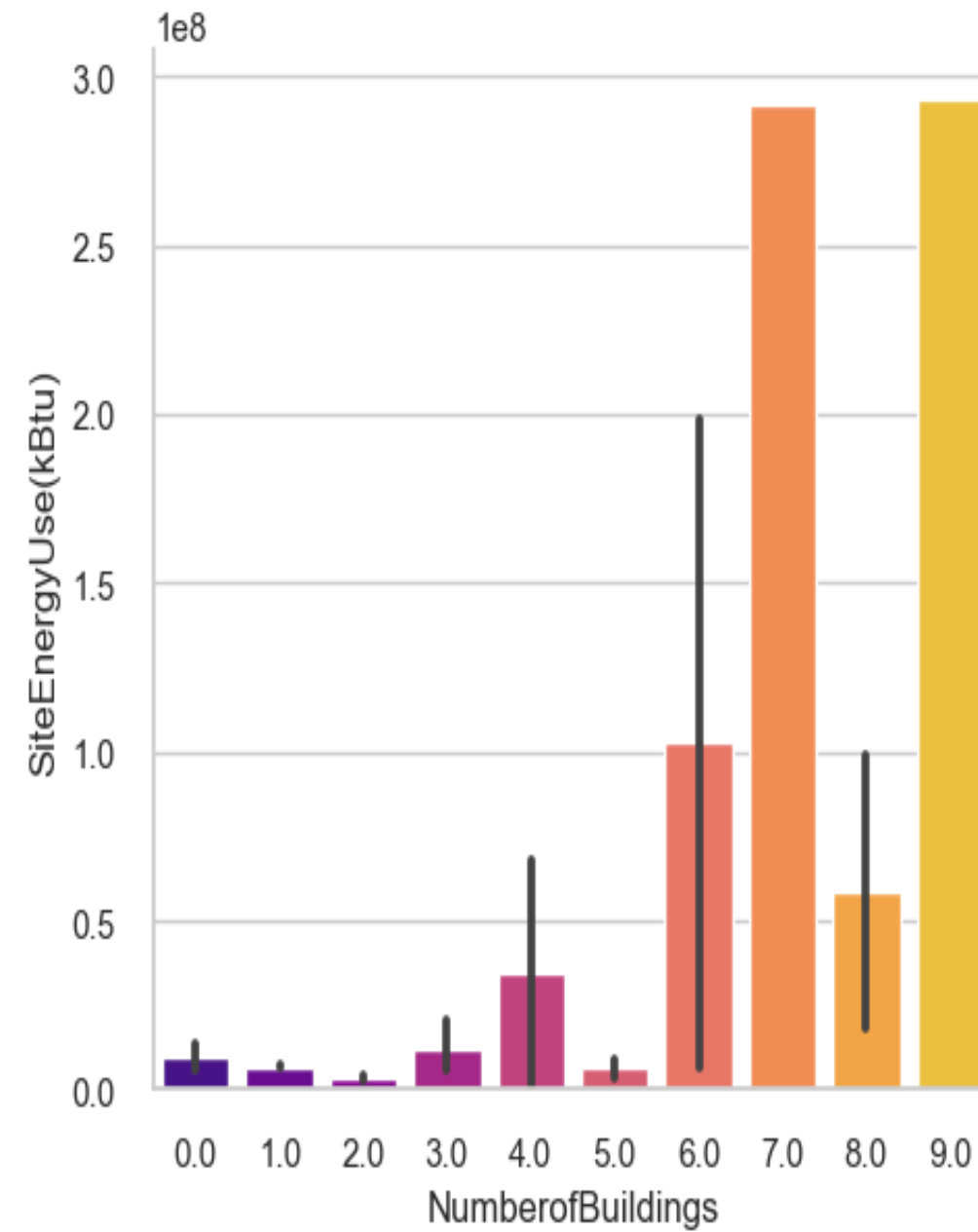
1.175916

Emission moyenne de gaz à effet de serre

# Quelques illustrations

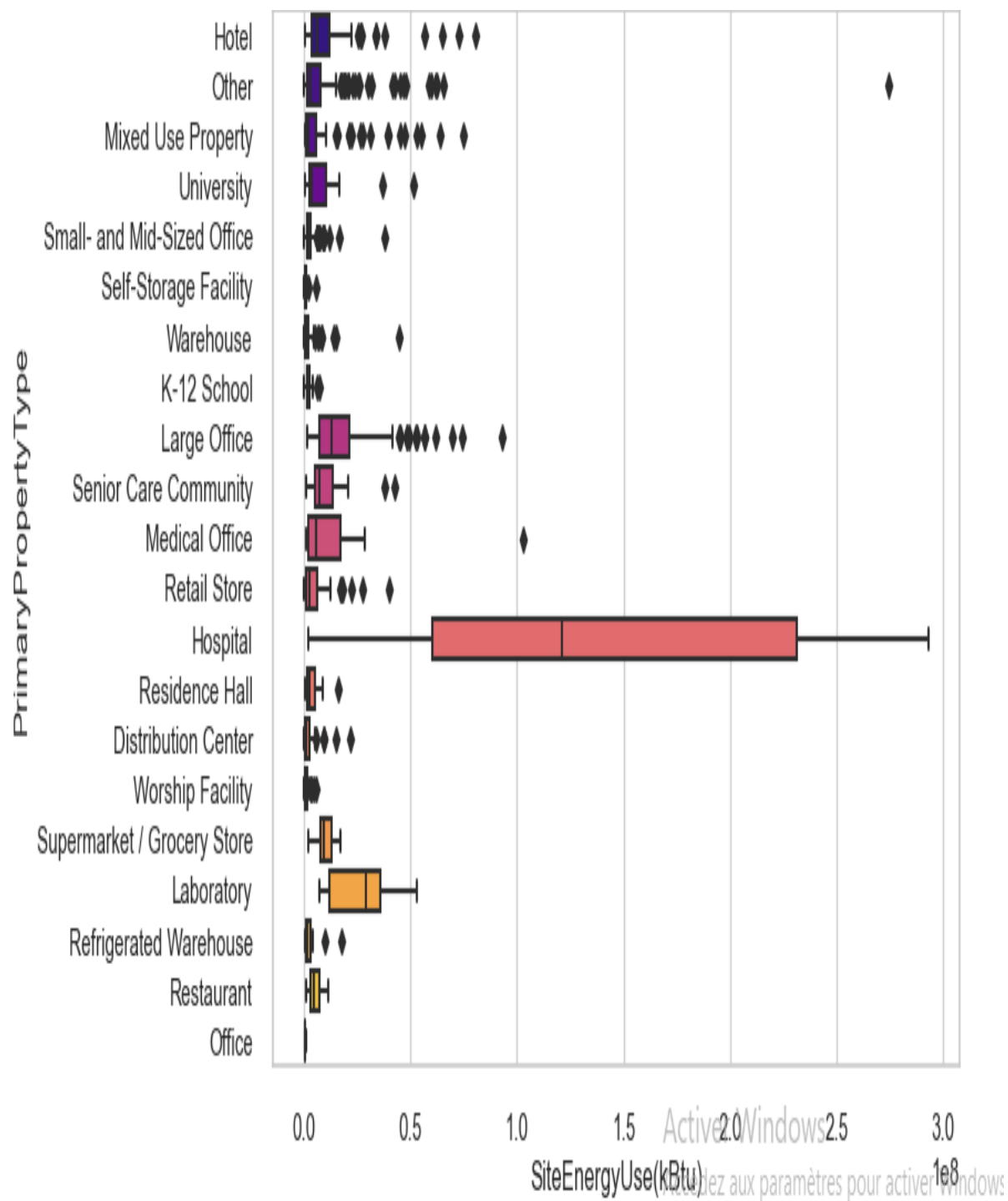


Consommation d'énergie en fonction du nombre de bâtiments

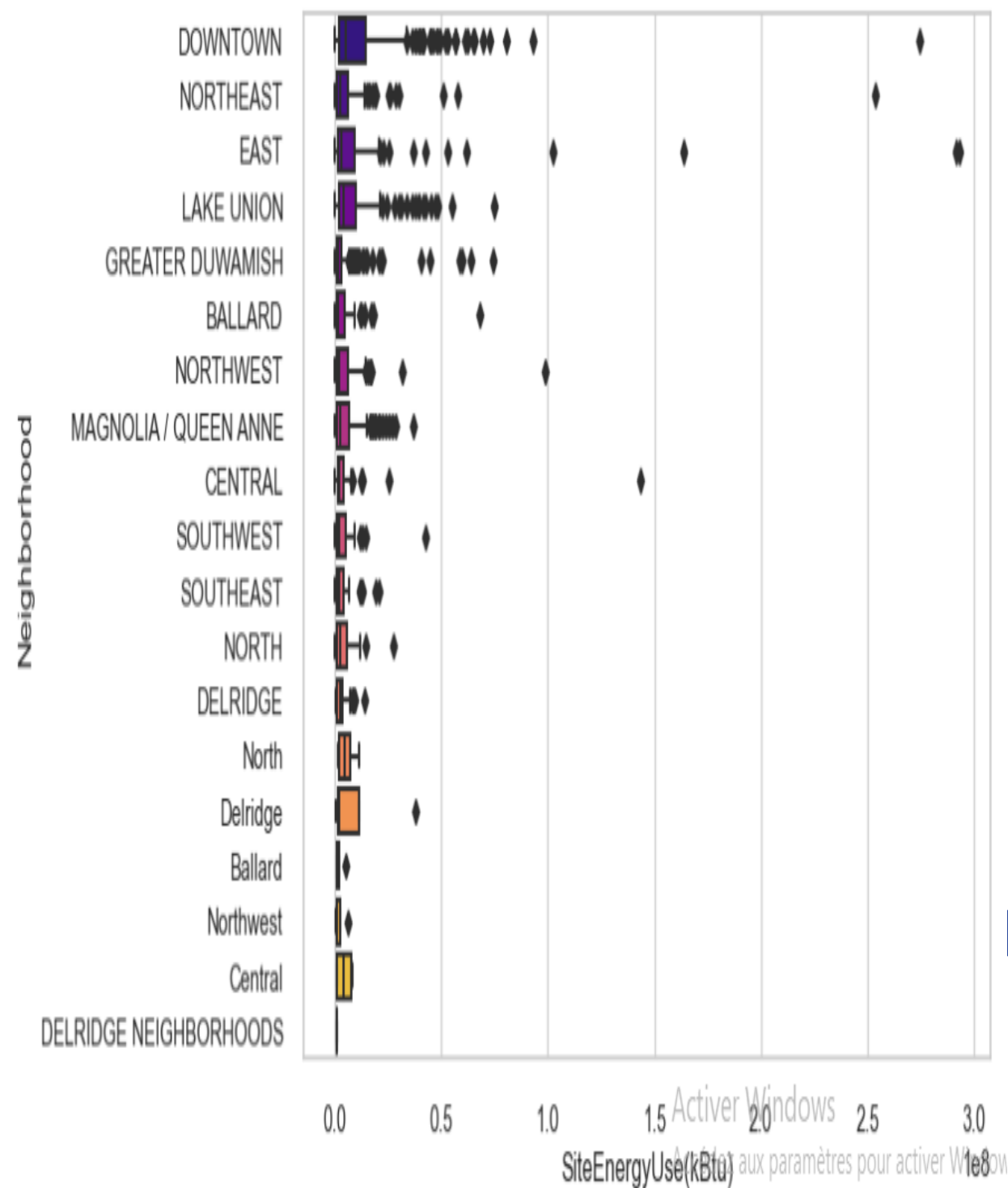


# Quelques illustrations

Boxplot de la consommation annuelle  
énergétique selon l'utilisation principale

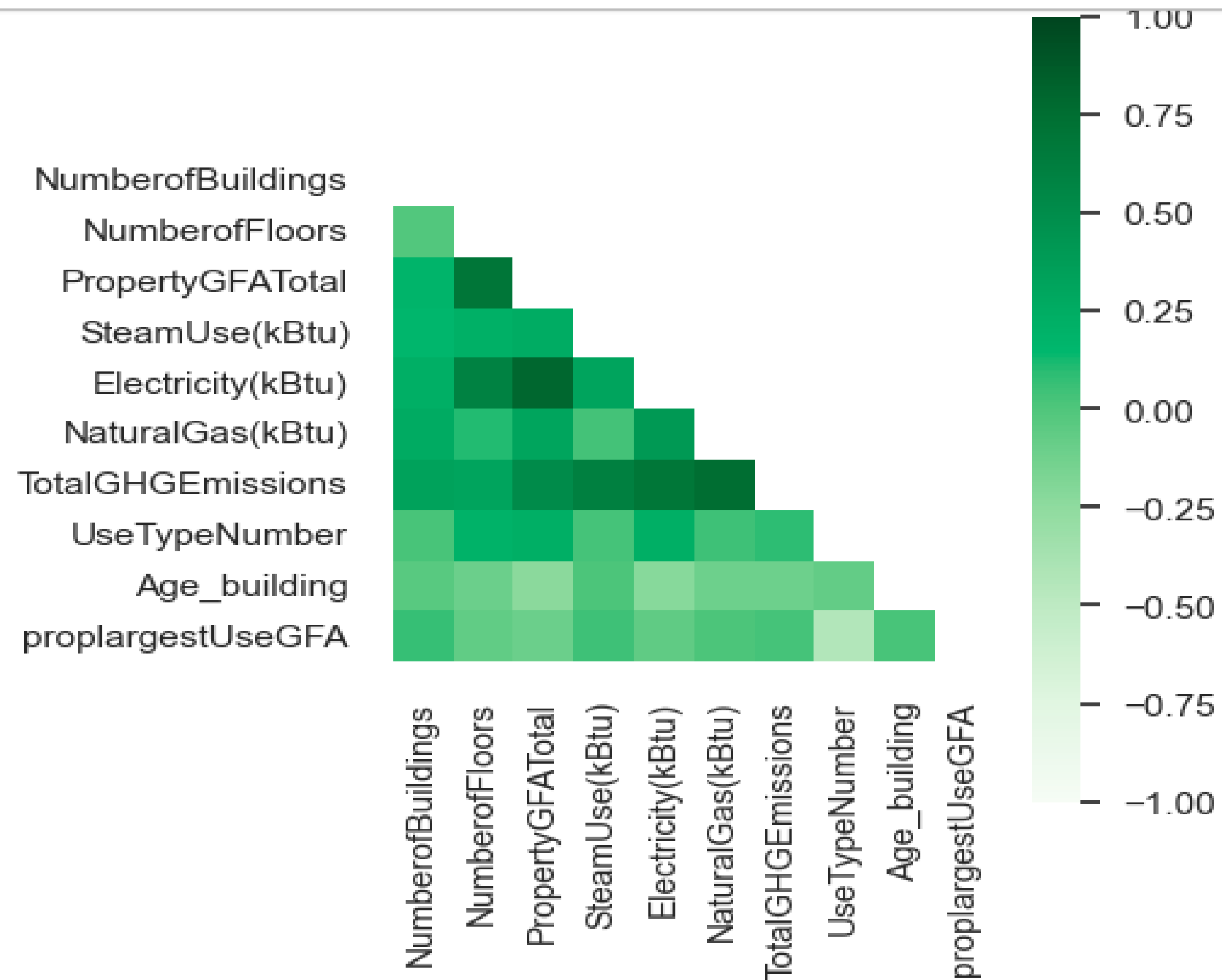


Boxplot de la consommation annuelle  
énergétique selon le Quartier





## Quelques illustrations



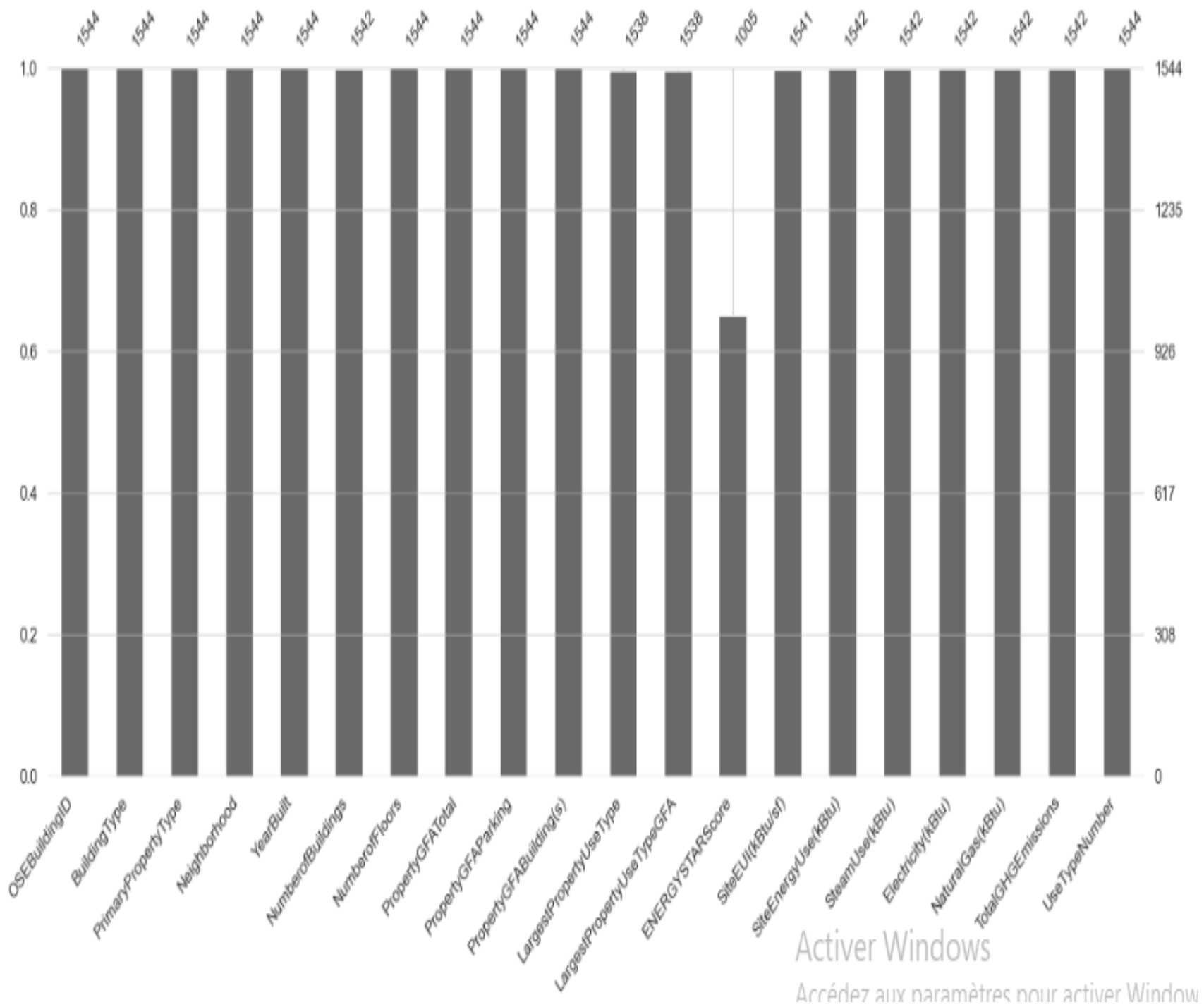


**2**

# **Apurement et transformations**



# Traitement des données manquantes



ENERGYSTARScore	539
LargestPropertyUseType	6
LargestPropertyUseTypeGFA	6
SiteEUI(kBtu/sf)	3
TotalGHGEmissions	2
NaturalGas(kBtu)	2
Electricity(kBtu)	2
NumberOfBuildings	2
SteamUse(kBtu)	2
SiteEnergyUse(kBtu)	2
OSEBuildingID	0
BuildingType	0
PropertyGFABuilding(s)	0
PropertyGFAParking	0
PropertyGFATotal	0
NumberOfFloors	0
YearBuilt	0
Neighborhood	0
PrimaryPropertyType	0
UseTypeNumber	0

## ENERGYSTARScore : imputation par KNN

## Les autres : Imputation par la médiane et le mode

# Incohérences

- nombres de bâtiments (au sein d'un building) nulle
  - valeurs négatives des quantités d'électricité consommées
  - valeurs négatives de la quantité d'émissions de gaz à effet de serre
  - **Traitement :**
    - nombres de bâtiments (imputation par la médiane)
    - Valeur absolue pour les données négatives
- 

## Valeurs aberrantes

Valeurs aberrante : une observation qui se situe à une distance significative des autres valeurs dans un ensemble de données

- **Méthode de Traitement : WINSORISATION**

technique qui consiste à ajuster les valeurs extrêmes d'un ensemble de données à des seuils spécifiques (selon les centiles etc) plutôt que de les supprimer.

# Création et recodage de variables

- **Age\_building** : une variable qui calcule l'âge du batiments
- *Calcul :  $2016 - YearBuilt$*
- **proplargestUseGFA** : Le pourcentage de la surface du LargestProperty sur le total
- *Calcul :  $LargestPropertyUseTypeGFA / PropertyGFATotal$*
- **UseTypeNumber** : nombre de tous les types d'utilisation de la propriété

- 
- **Recodages des modalités de la variable Neighborhood**

**3**

# **Modélisation**



# Spécification des modèles

## La matrice X des variables explicatives

- ☐ PrimaryPropertyType
- ☐ Neighborhood
- ☐ NumberofBuildings
- ☐ NumberofFloors
- ☐ PropertyGFATotal
- ☐ SteamUse(kBtu)
- ☐ Electricity(kBtu)
- ☐ NaturalGas(kBtu)
- ☐ TotalGHGEmissions
- ☐ UseTypeNumber
- ☐ Age\_building
- ☐ proplargestUseGFA
- ☐ ENERGYSTARScore

## La variable à expliquer

- ☐ SiteEnergyUse(kBtu)

## Métriques d'évaluation

- ☐  $R^2$
- ☐ mean\_squared\_error
- ☐ mean\_absolute\_percentage\_error
- ☐ max\_error

## Preprocessor

### **numeric\_transformer**

- ☐ Standardisation

### **categorical\_transformer**

- ☐ OneHotEncoder

# Modèles paramétriques

## Régression linéaire

Formule mathématique:  $y = f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \dots + \beta_r x_r + \varepsilon$

### Explication du modèle:

L'analyse de régression calcule la relation estimée entre une variable dépendante et une ou plusieurs variables explicatives. Elle **permet de modéliser la relation entre les variables choisies et de prévoir des valeurs en fonction du modèle.**

## Elastic-net

Formule mathématique :

### Explication :

C'est une solution au problème de l'overfitting des modèles.

La conséquence est de rétrécir les coefficients (comme dans la régression ridge) et de mettre certains coefficients à zéro (comme dans le LASSO).



# Présentation des résultats

## Sans EnergyStarScore

Scores	Regression linéaire	Elastic Net
Avant hyperparameters tuning		
Training dataset	0.909	0.815
Test dataset	0.849	0.784
Après Hyperparameters Tuning		
Training dataset	0.909	0.879
Test dataset	0.849	0.843

## Avec EnergyStarScore

Scores	Regression linéaire	Elastic Net
Avant hyperparameters tuning		
Training dataset	0.909	0.815
Test dataset	0.849	0.784
Après Hyperparameters Tuning		
Training dataset	0.909	0.878
Test dataset	0.849	0.843

# Méthodes ensemblistes

## Random forest

Le Random Forest combine plusieurs arbres de décision, formés aléatoirement sur des sous-ensembles de données et de caractéristiques, pour améliorer la robustesse du modèle par le vote majoritaire ou la moyenne.

## XGBOOST

C'est une méthode de **Boosting**. Elle fait un assemblage d'arbres décisionnels (weak learners) qui prédisent les résidus et corrige les erreurs des arbres décisionnels précédents.

L'avantage est qu'elle permet de réduire les biais, et rend plus performante la prédiction.

# Méthodes ensemblistes

## LightGBM

Light GBM est un cadre de renforcement de gradient qui utilise un algorithme d'apprentissage basé sur des arbres.

**Light GBM fait croître l'arbre verticalement** tandis que d'autres algorithmes font pousser des arbres horizontalement, ce qui signifie que Light GBM fait pousser l'arbre par **feuille** tandis que l'autre algorithme se développe par niveau. Il choisira la feuille avec une perte de delta maximale pour se développer. Lors de la croissance de la même feuille, l'algorithme par feuille peut réduire plus de perte qu'un algorithme par niveau.

# Présentation des résultats

Sans EnergyStarScore				,	Avec EnergyStarScore			
SCORE	RandomForestRegressor	Xgboost	LightGBM		SCORE	RandomForestRegressor	Xgboost	LightGBM
Avant Hyperparametre Tuning					Avant Hyperparametre Tuning			
Training dataset	0,976	0,999	0,884		Training dataset	0,983	0,999	0,884
Test dataset	0,946	0,993	0,784		Test dataset	0,963	0,992	0,789
Après Hyperparametre Tuning					Après Hyperparametre Tuning			
Training dataset	0,959	0,999	0,999		Training dataset	0,960	0,999	0,962
Test dataset	0,945	0,997	0,991		Test dataset	0,944	0,996	0,952

- Pas d'impact majeur de EnergyStarScore dans la prédiction

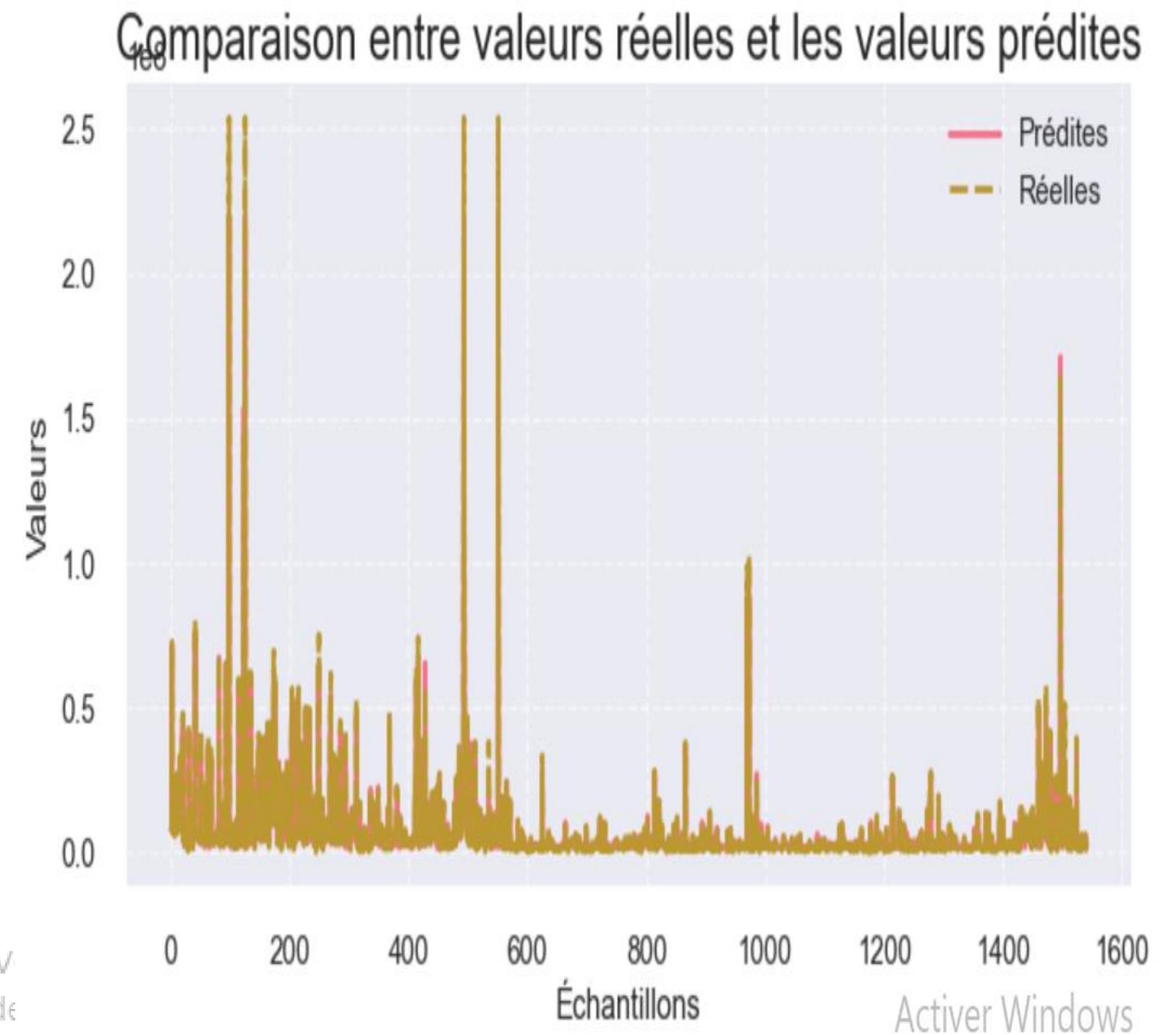
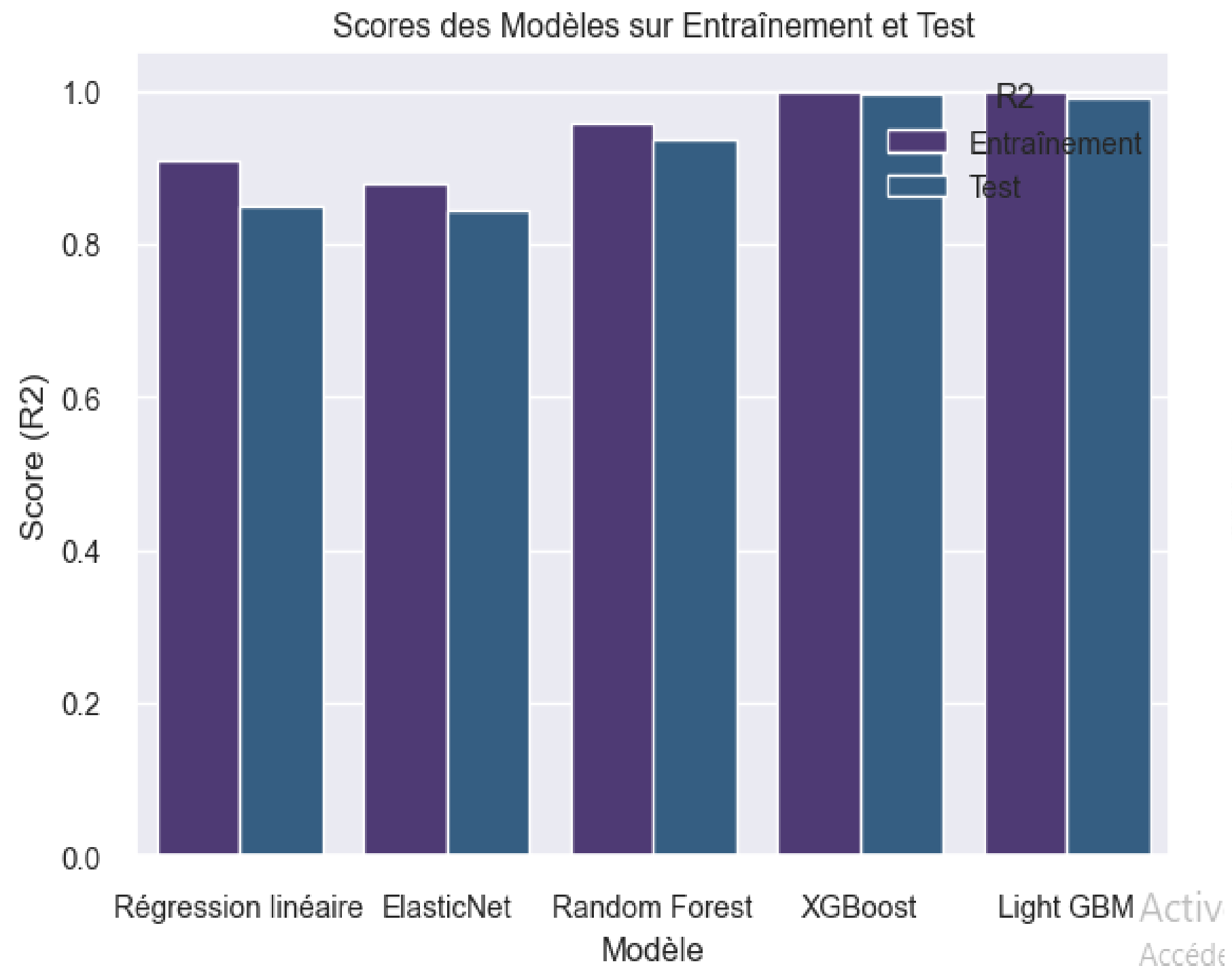
**4**

**Choix définitif**



# Modèle définitif

## XGBOOST retenu



# Déploiement du modèle XGBOOST

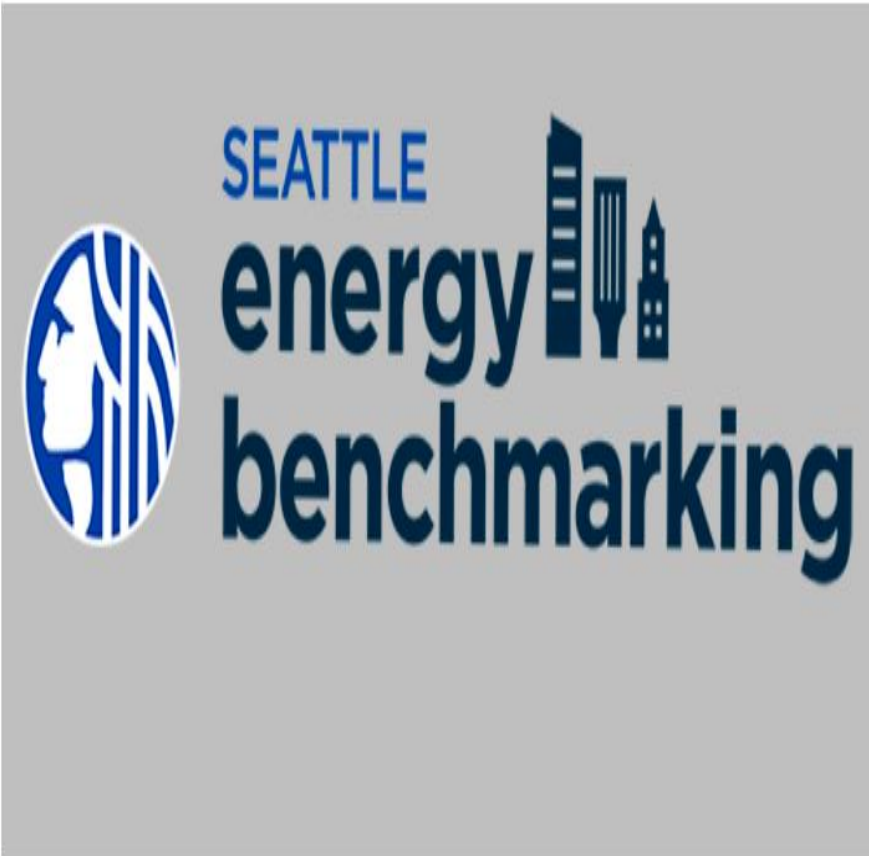
X

Sélectionnez l'utilisation principale de la propriété

Hotel

Sélectionnez le quartier du Bâtiment

DOWNTOWN



Prédiction de la consommation annuelle  
d'énergie des propriétés de la ville de  
SEATTLE

Activer Windows  
Accédez aux paramètres pour activer Windows.

Deploy

X

Sélectionnez l'utilisation principale de la propriété

Hotel

Sélectionnez le quartier du Bâtiment

DOWNTOWN

Deploy

Émissions totales de gaz à effet de serre

265328

-

+

Nombre de types d'utilisation de la propriété

12

-

+

Âge du bâtiment

21

-

+

Pourcentage de la superficie du plus grand sous-bâtiment par rapport à la superficie totale.

65

-

+

Prédiction

Consommation d'énergie annuelle prédite (en kBtu) :

✓

[30229266.]

Activer Windows  
Accédez aux paramètres pour activer Windows.



Merci!