Projet 4 - Openclassrooms Anticipez les besoins en consommation de bâtiments

City of Seattle

Le 17.03.2023

Dabidin Keshika

OPENCLASSROOMS

Plan

1. Contexte de l'étude
2. Présentation des données
3. Analyse et feature engineering des données d'émissions de CO_2 et de consommation d'énergie
4. Présentation des différents modèles utilisés et des résultats obtenus
5. Conclusion

City of Seattle

Introduction

- **Contexte :** Dans la ville de Seattle, nous intéresserons à la consommation et aux émissions en carbone des bâtiments non destinés à l'habitation.

Les relevés de l'année 2016 sont disponibles pour cette étude.

- **Objectifs:**
  - Prédire les émissions de CO_2 et la consommation d'énergie des bâtiments sans les relevés annuels (car coûteux).
  - Évaluer l'intérêt de l'Energy Star Score (compliqué dans la mise en place)
  - Mise en place d'un modèle réutilisable.

City of Seattle

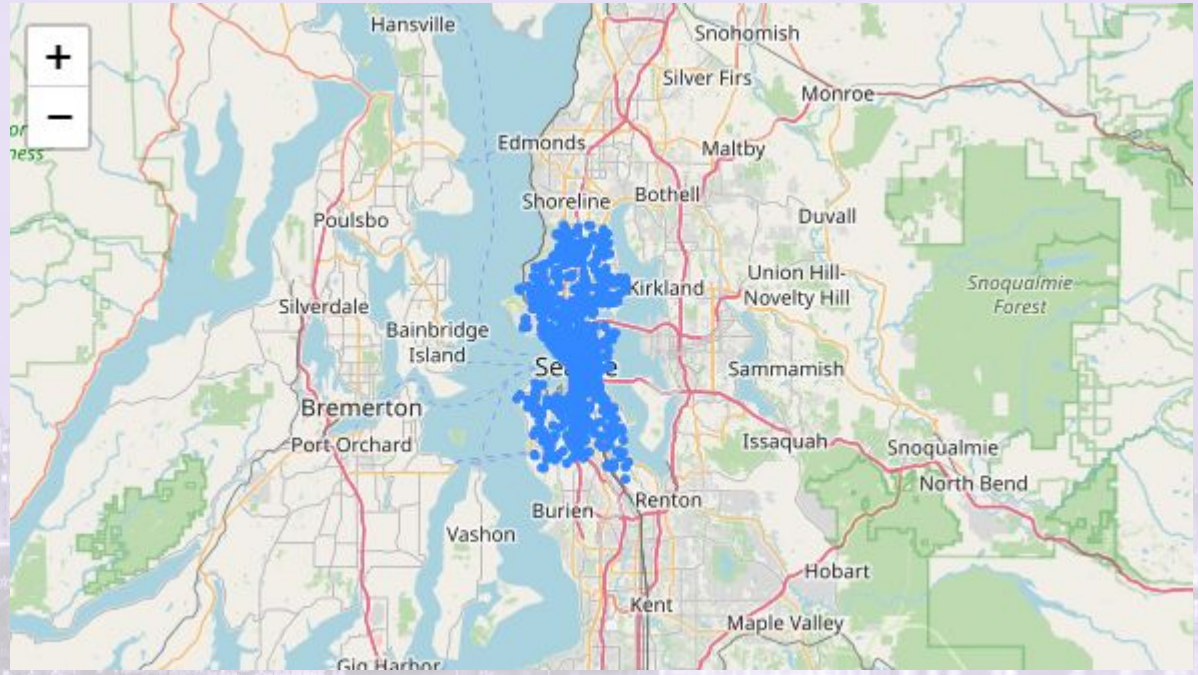
Présentation des données

- **Source :**
https://data.seattle.gov/dataset/2016-Building-Energy-Benchmarking/2bpz-gwpy

- **Résumé :**

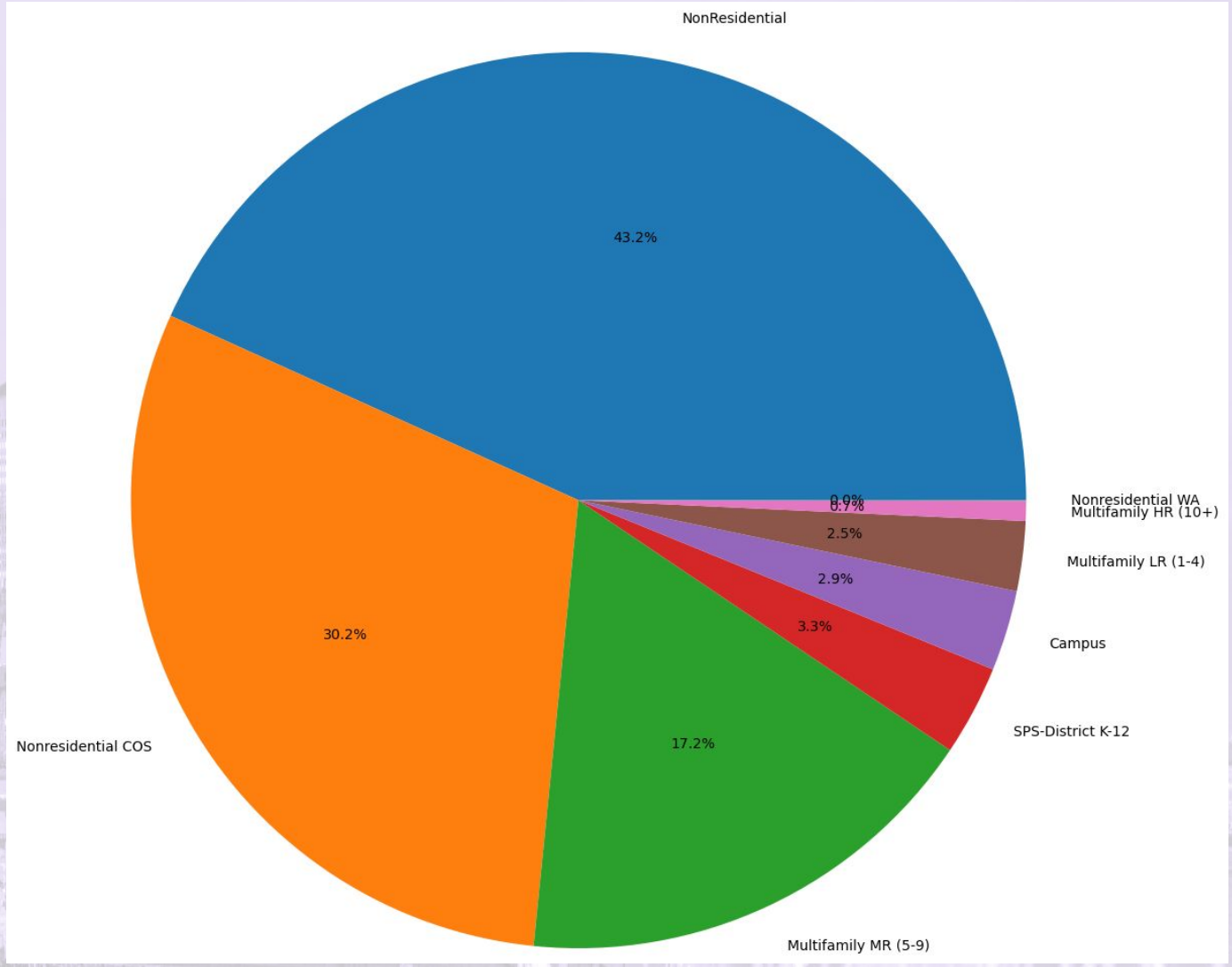
	Nombre lignes	Nombre colonnes	Taux de remplissage moyen (%)	Doublons
fichier	3376	46	87,2	0

Tous les bâtiments sont localisés dans la ville de Seattle.



City of Seattle

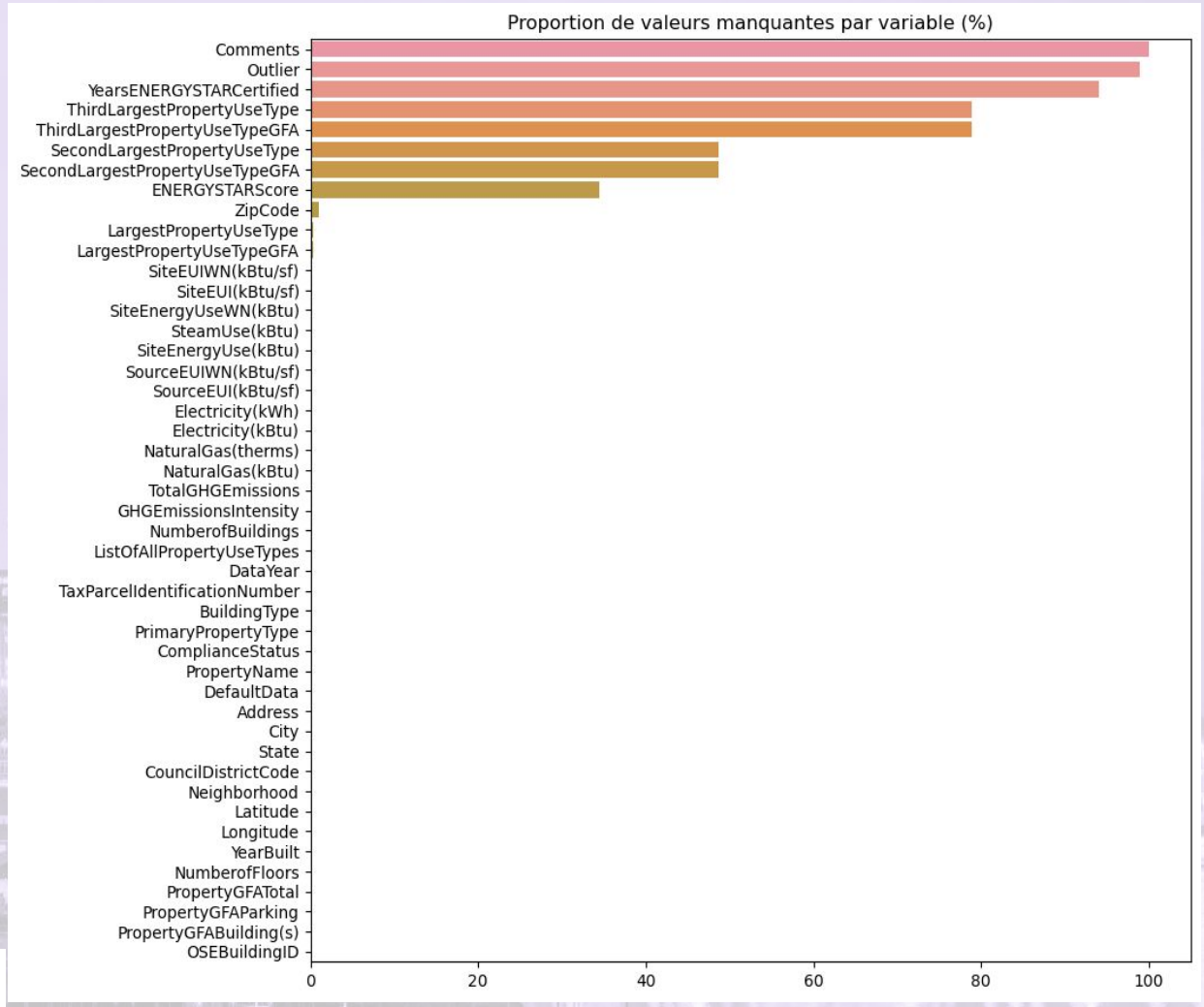
Présentation des données



City of Seattle

Présentation des données

- **Valeurs manquantes :**
  - Suppression des outliers, bâtiments non conformes et données où toutes les valeurs sont manquantes.
  - Pour certaines variables comme Secondary and Third property type GFA, remplacement par 0 si pas de bâtiment.
  - Energy Star Score avec valeurs manquantes - séparation lors du machine learning.



	Nombre lignes	Nombre colonnes	Taux remplissage moyen	Doublons
fichier	1647	44	97.1%	0

City of Seattle

Présentation des données

OSEBuildingID
PropertyName
TaxParcelIdentificationNumber
DataYear
DefaultData
Comments
ComplianceStatus
Outlier
CouncilDistrictCode
Neighborhood
ZipCode
Latitude
Longitude
Address
BuildingType
PrimaryPropertyType
YearBuilt
ListOfAllPropertyUseTypes
LargestPropertyUseType
SecondLargestPropertyUseType
ThirdLargestPropertyUseType
NumberofBuildings
NumberofFloors
PropertyGFATotal
PropertyGFAParking
PropertyGFABuilding(s)
LargestPropertyUseTypeGFA
SecondLargestPropertyUseTypeGFA
ThirdLargestPropertyUseTypeGFA

← Variables d'identification

← Variables sur le type de bâtiment (catégorielles)

← Variables de surface (numériques)

SiteEUI(kBtu/sf)
SiteEUIWN(kBtu/sf)
SourceEUI(kBtu/sf)
SourceEUIWN(kBtu/sf)
SiteEnergyUse(kBtu)
SiteEnergyUseWN(kBtu)
SteamUse(kBtu)
Electricity(kWh)
Electricity(kBtu)
NaturalGas(therms)
NaturalGas(kBtu)
OtherFuelUse(kBtu)
TotalGHGEmissions
GHGEmissionsIntensity
YearsENERGYSTARCertified
ENERGYSTARScore

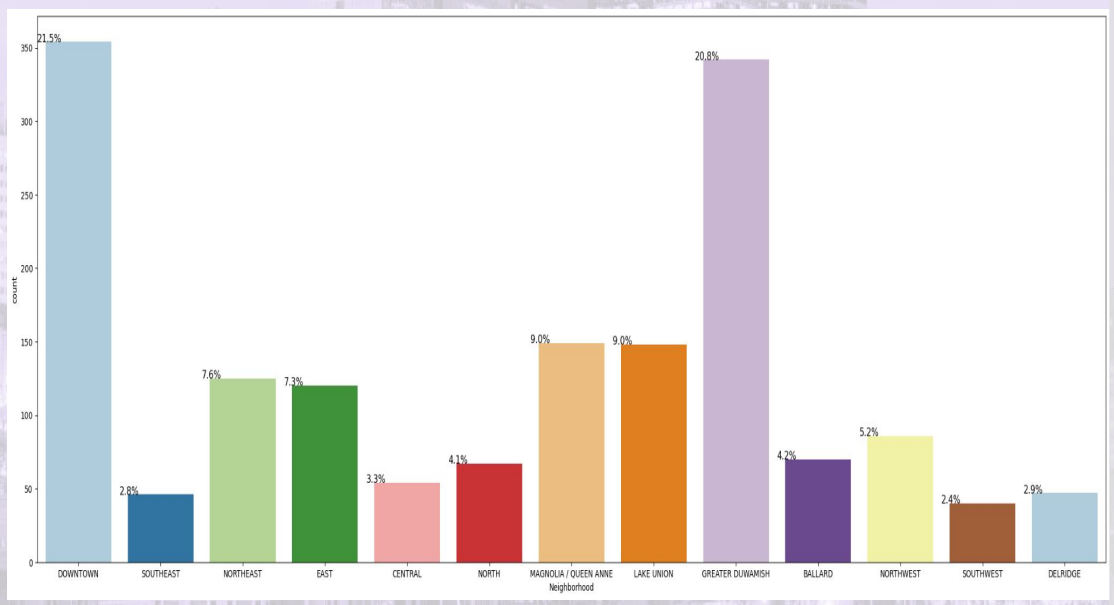
← Variables Cibles

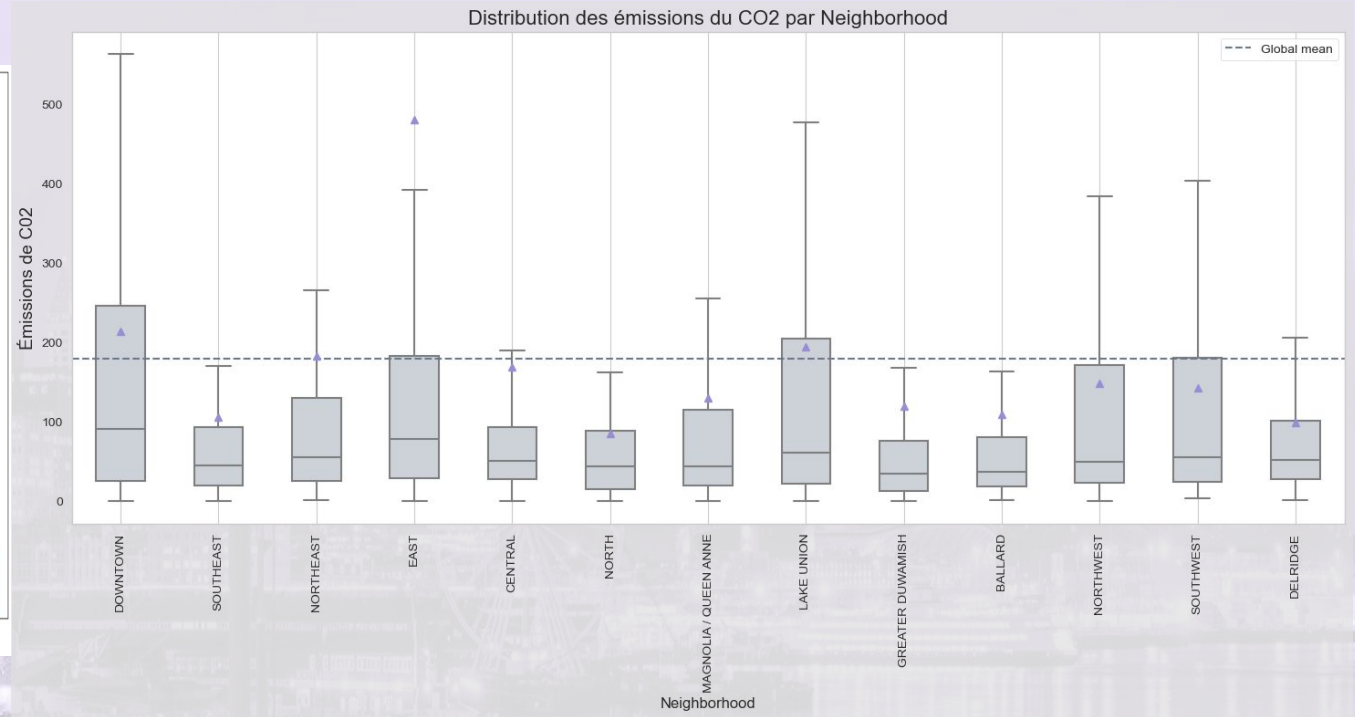
Variable Energy Star Score

City of Seattle

Analyse des données et Feature Engineering

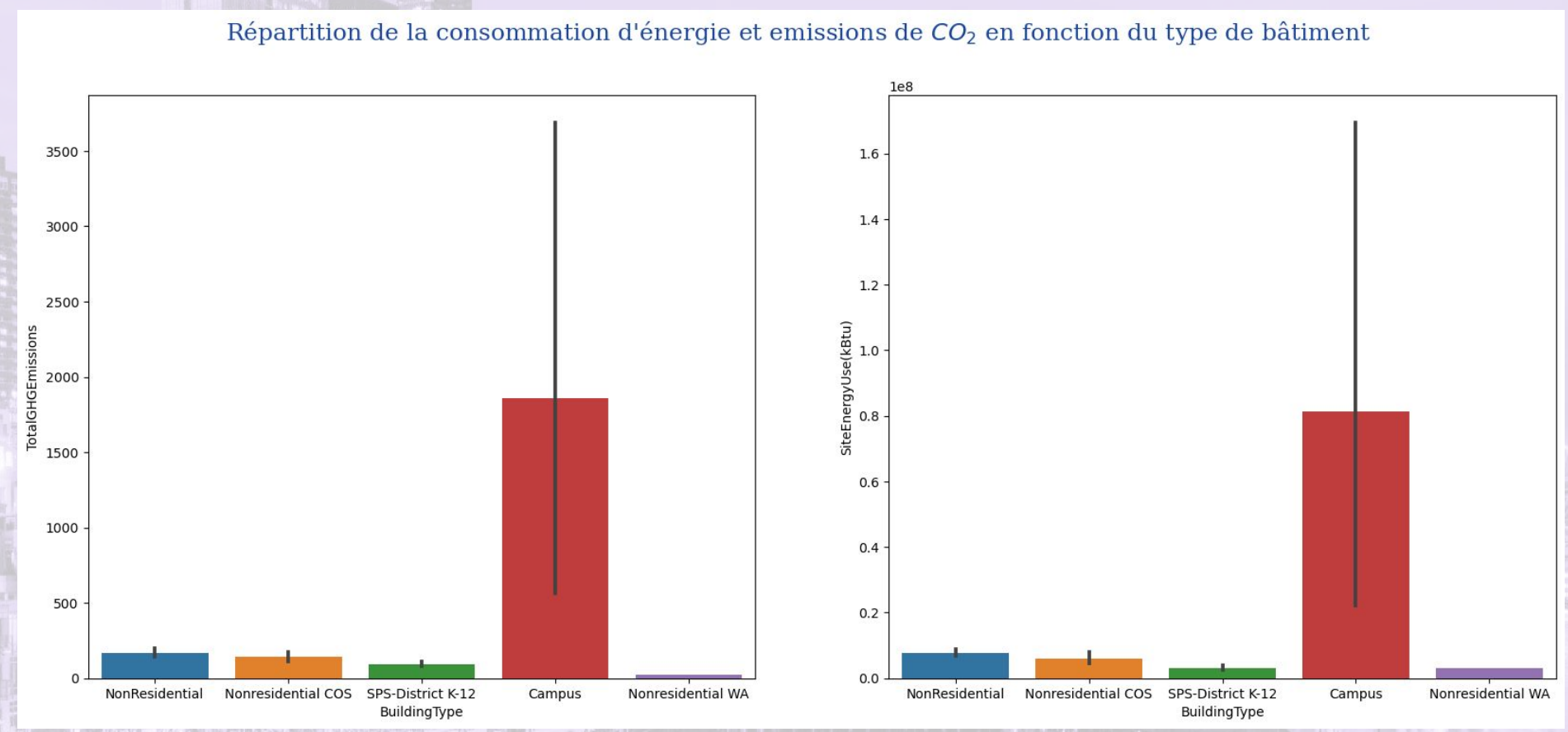
Neighborhood





City of Seattle

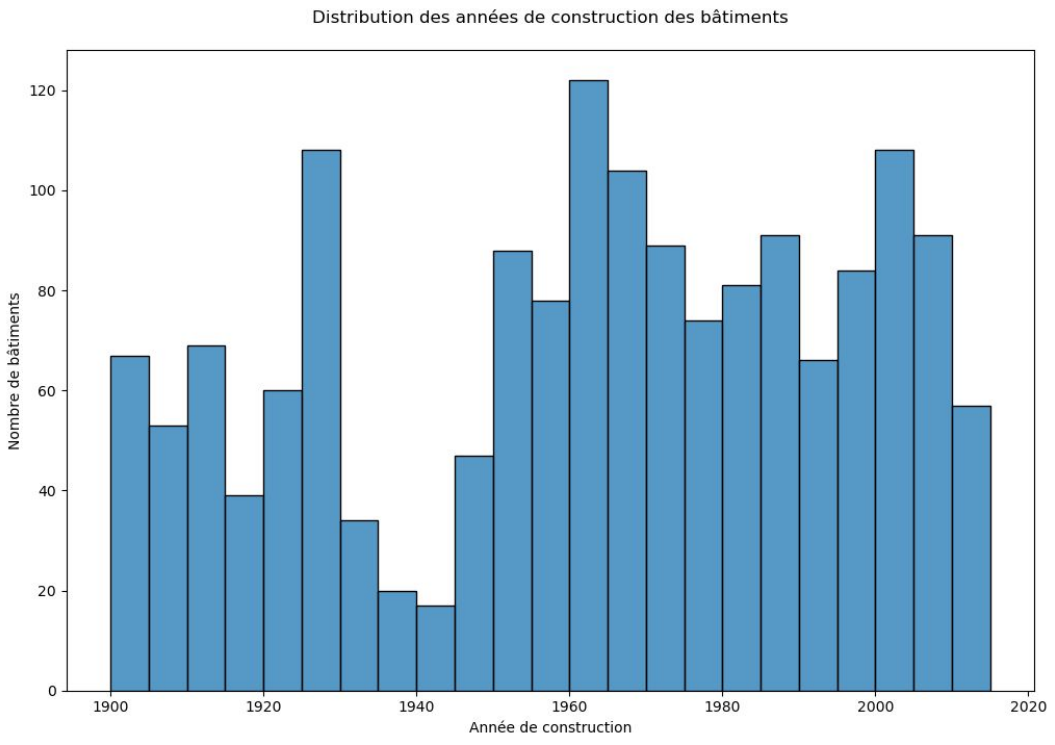
Analyse des variables et Feature Engineering

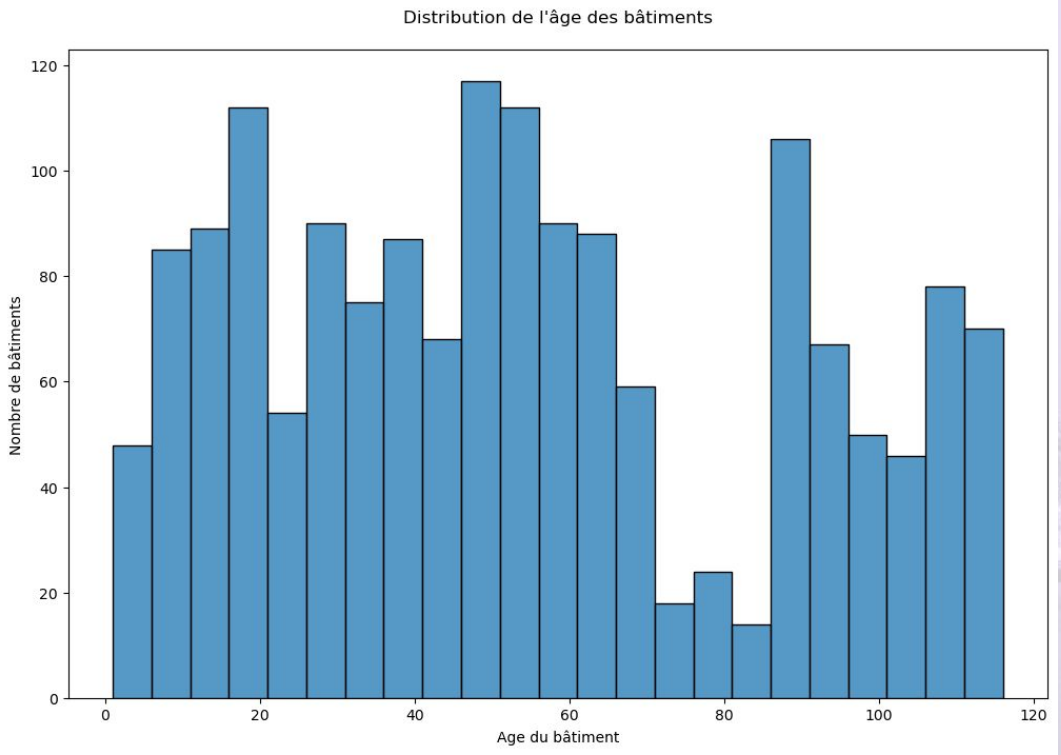


City of Seattle

Analyse des variables et Feature Engineering

Âge des bâtiments

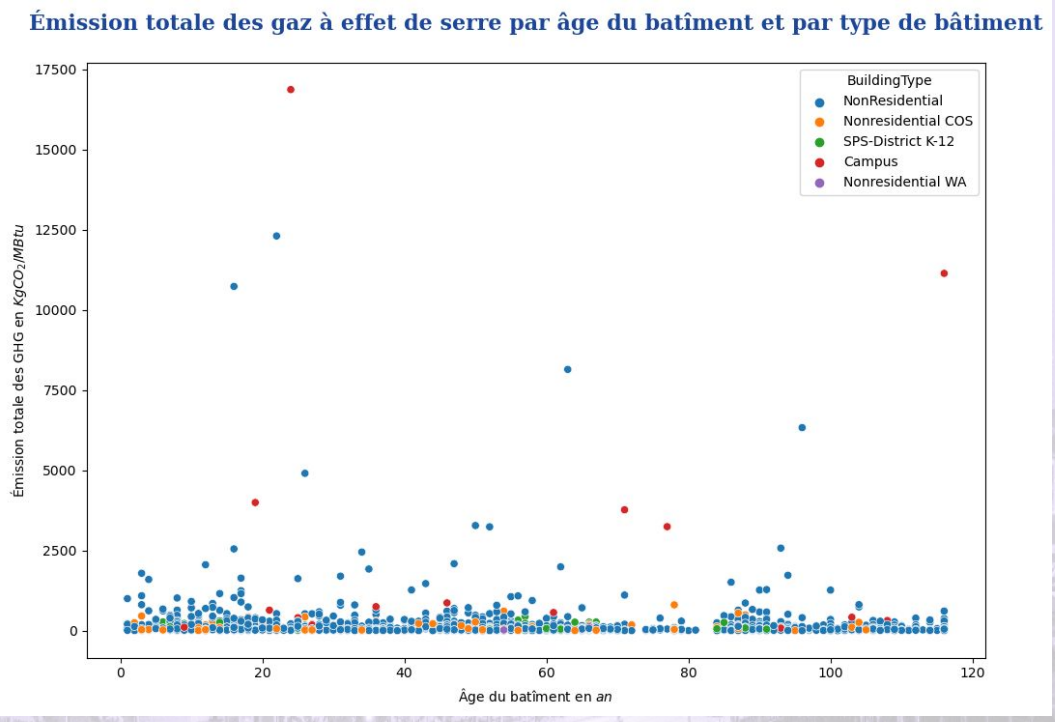


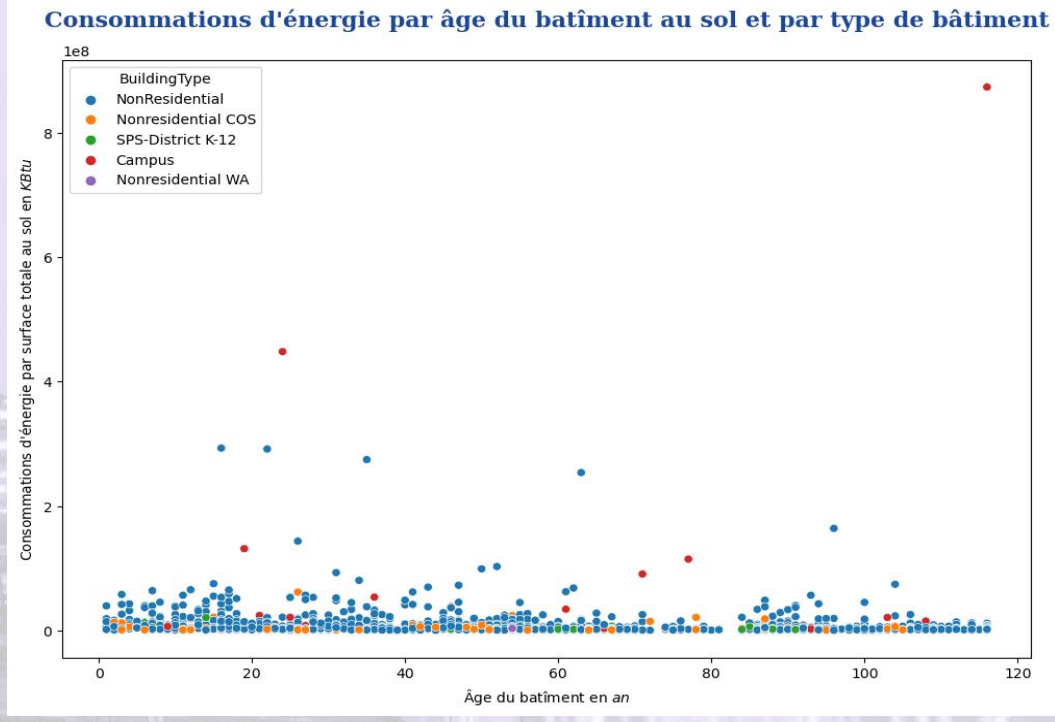


City of Seattle

Analyse des variables et Feature Engineering

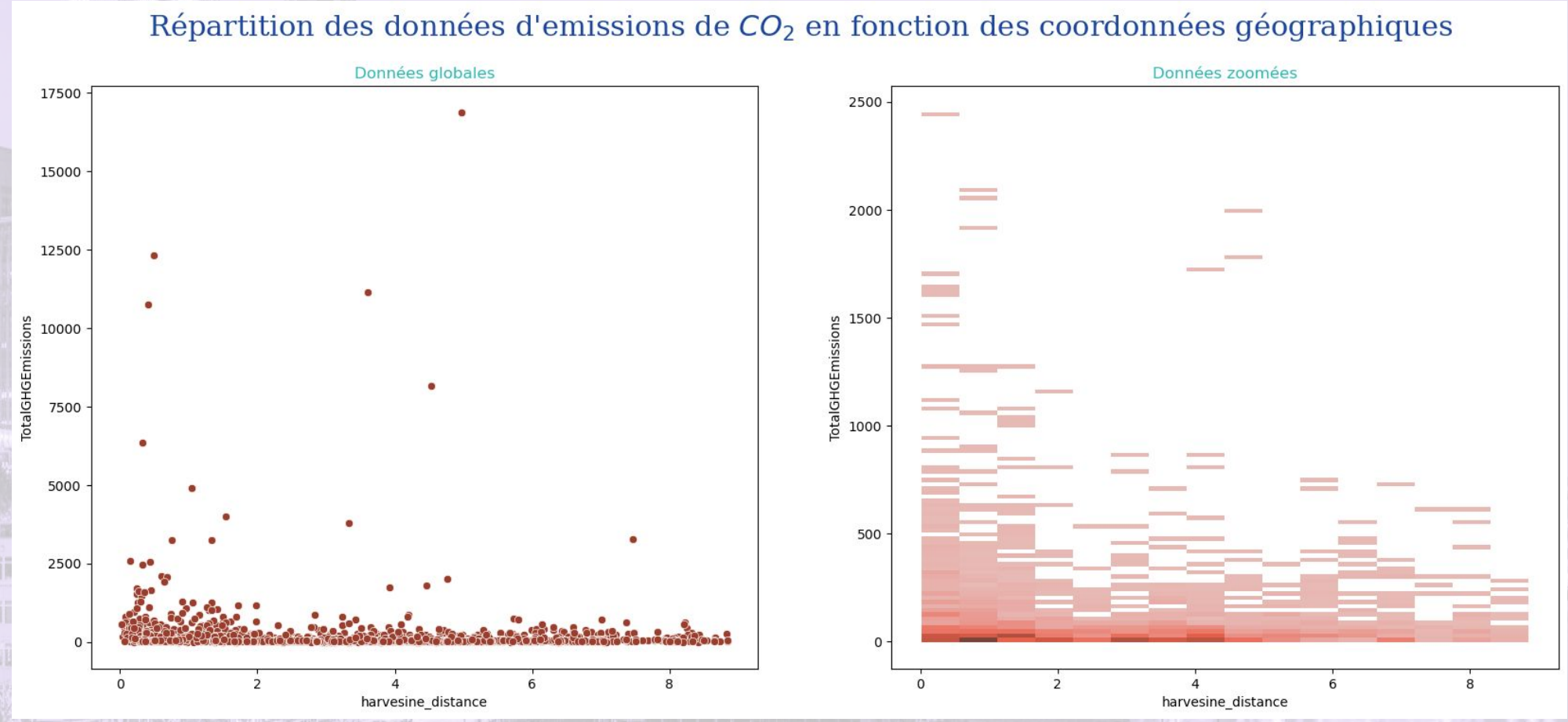
Âge des bâtiments





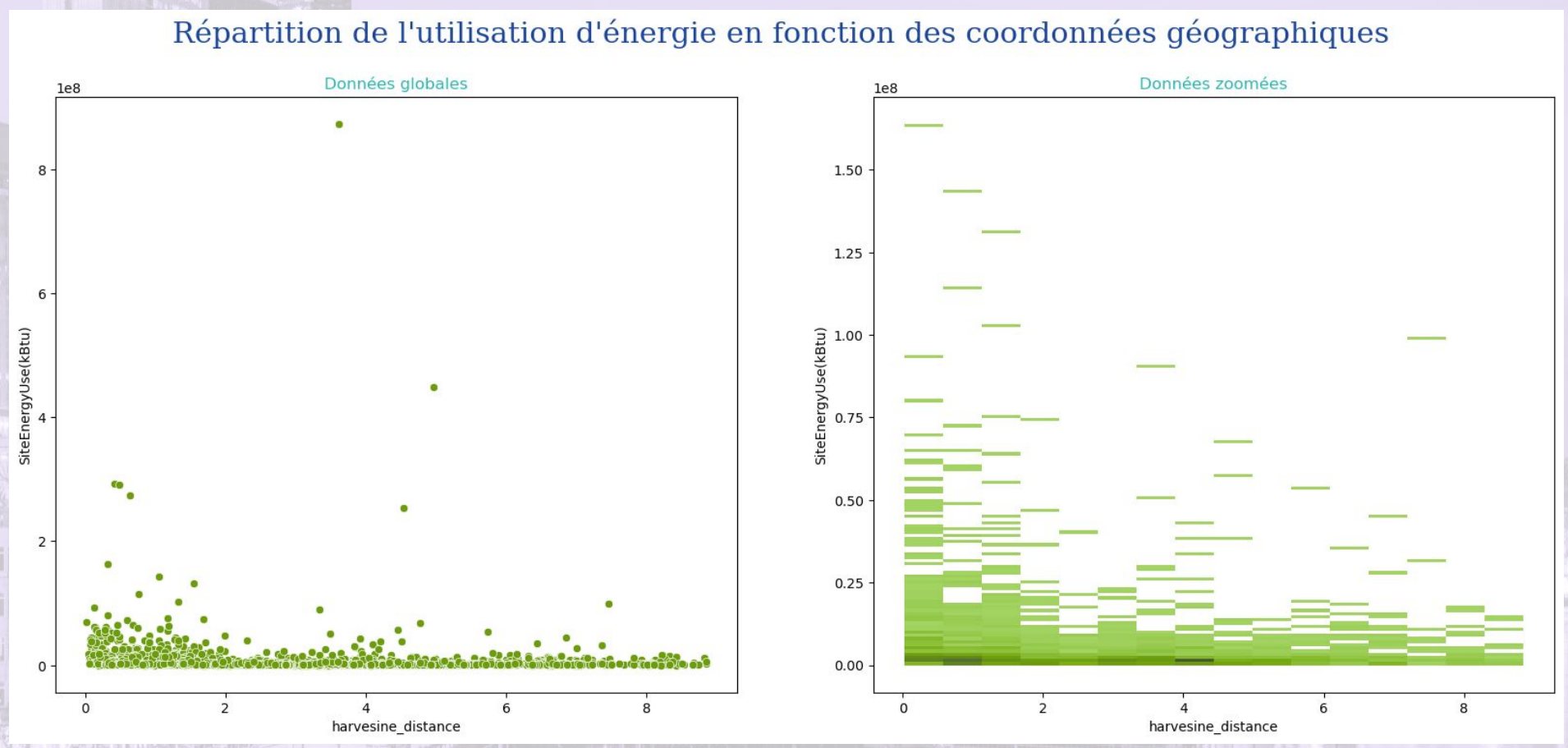
City of Seattle

Analyse des variables - Émission de CO2



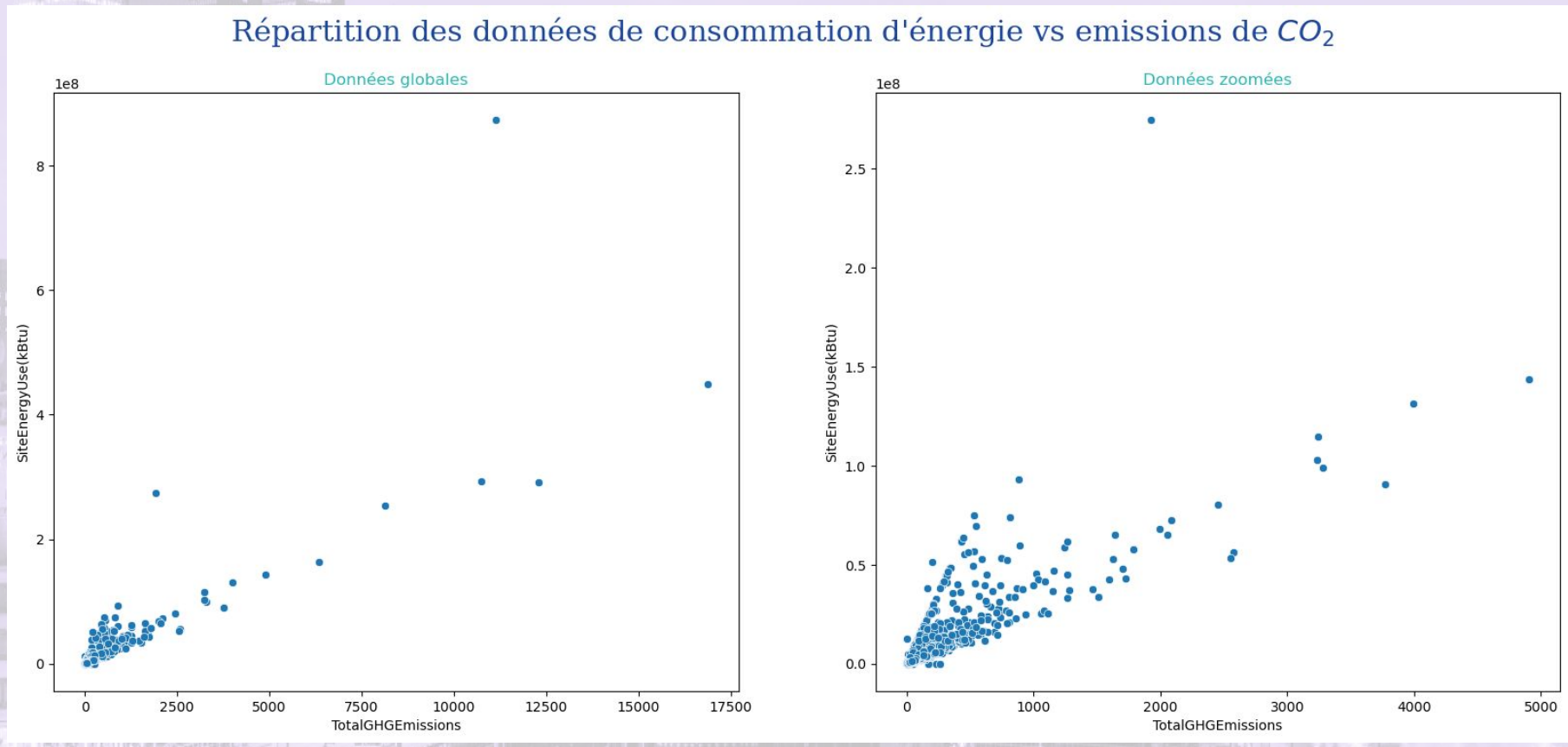
City of Seattle

Analyse des variables - Consommation d’énergie



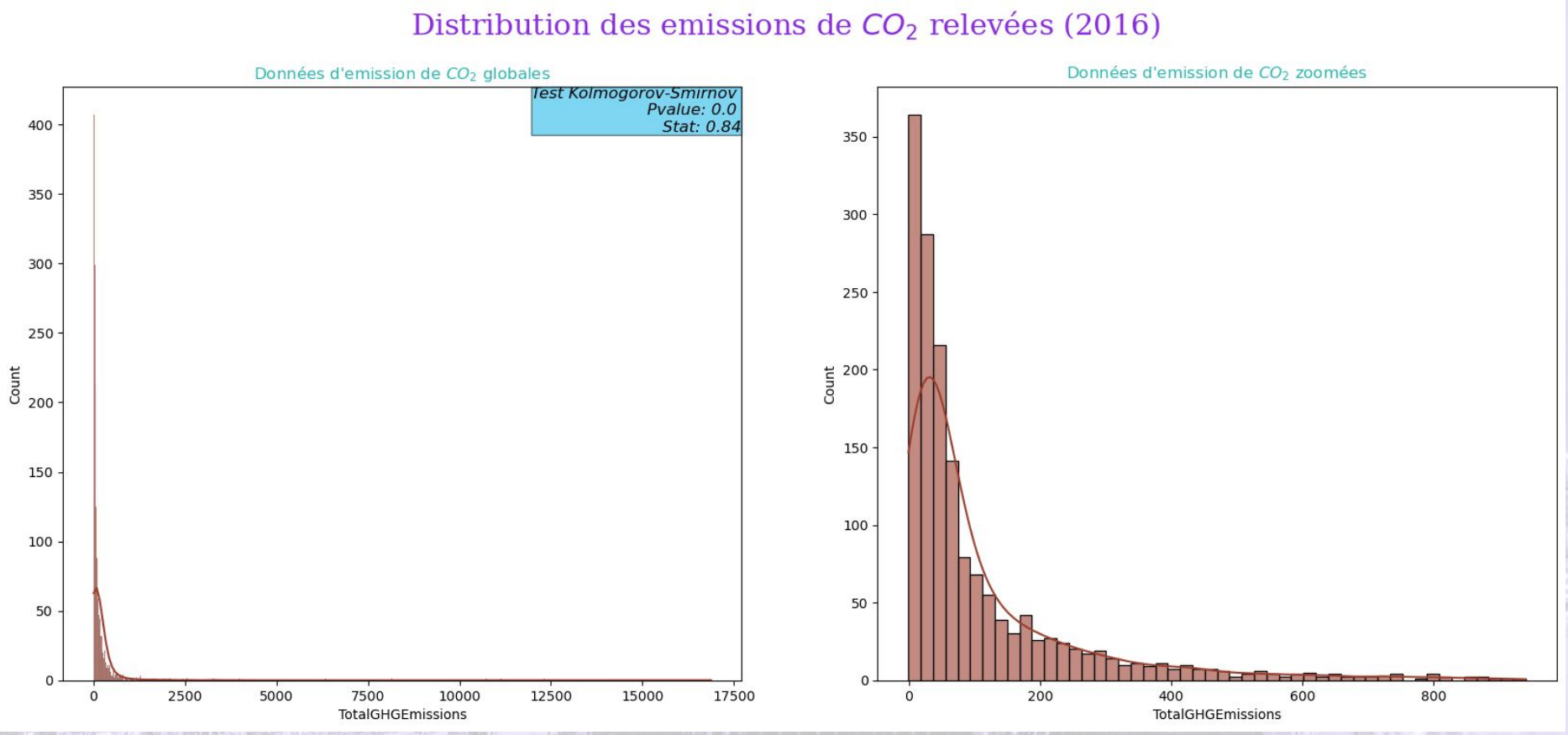
City of Seattle

Analyse des variables et Feature Engineering



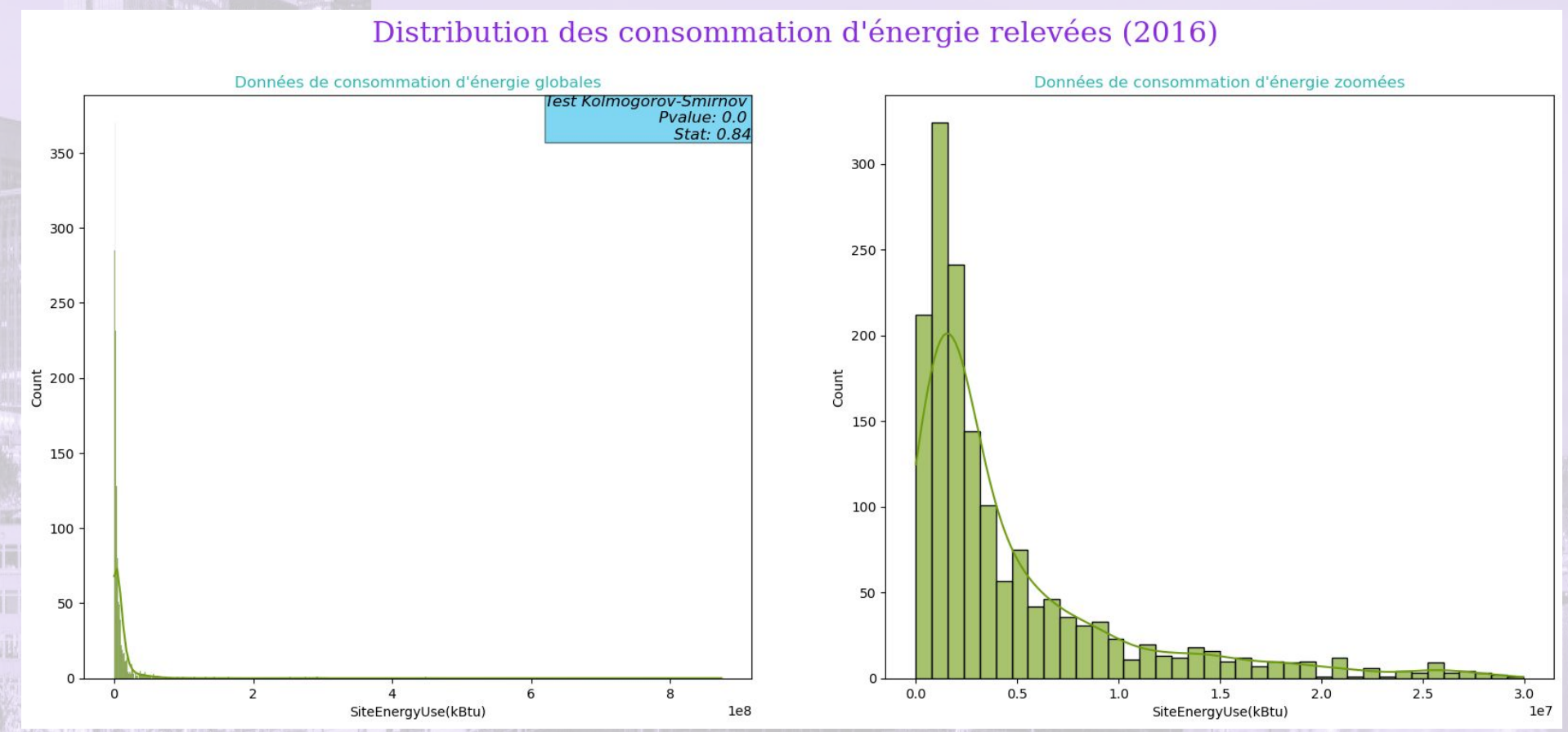
City of Seattle

Analyse des variables - Émission de CO2



City of Seattle

Analyse des variables - Consommation d’énergie



City of Seattle

Analyse des variables et Feature Engineering

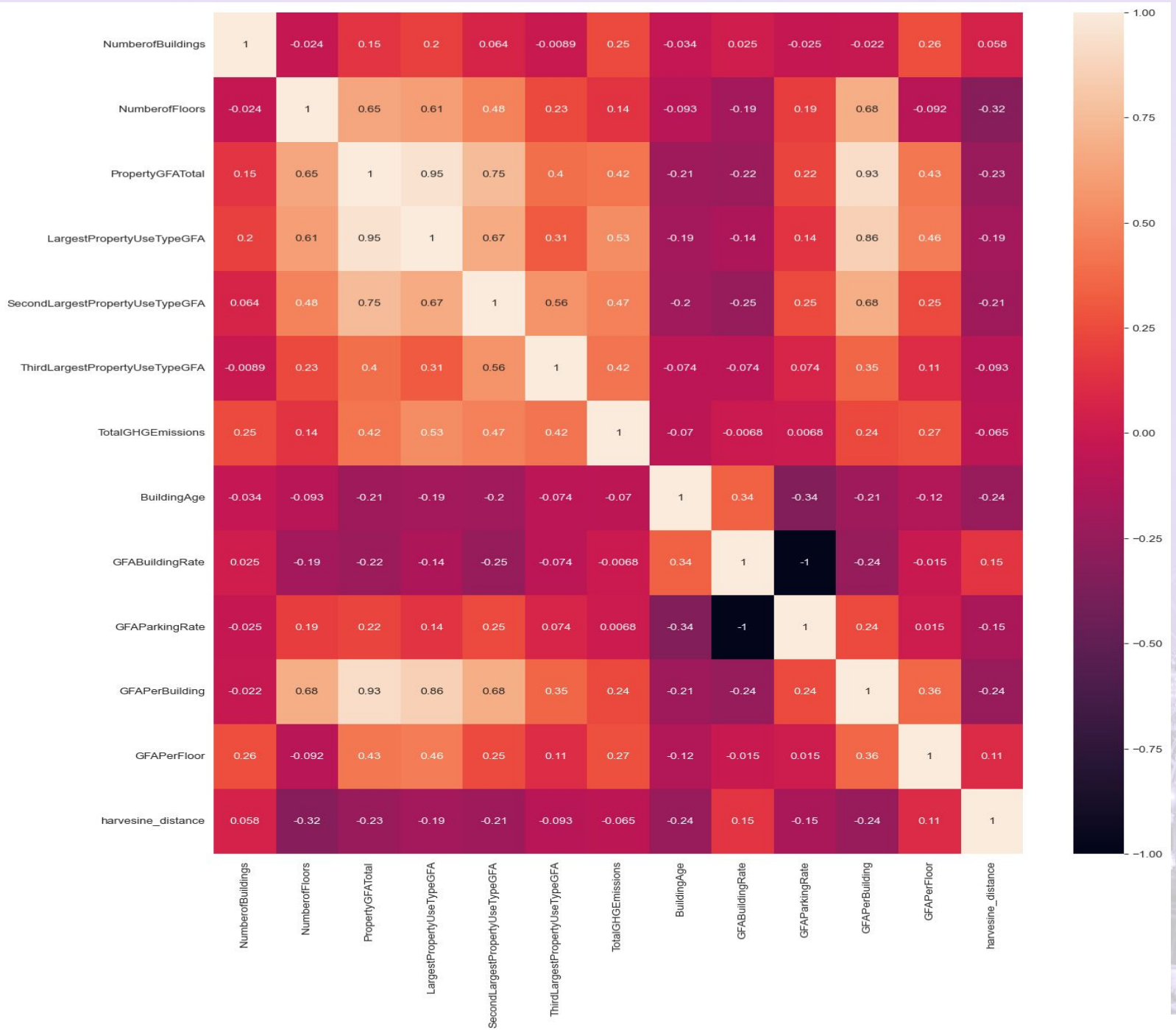
- **Feature Engineering :**
- Calcul des ratios de surface :

  Surface Bâtiment / Nombres de Bâtiments

  Surface Bâtiment / Nombre d'étages

- Suppression des variables de surface parking et surface de bâtiments pour remplacer avec ratio de Surface bâtiment / Surface totale (bâtiment + parking) et Surface Parking / Surface totale, ce qui permet de voir l'influence de la surface du parking.
- Calcul du Haversine Distance.

City of Seattle



Heatmap des corrélations

City of Seattle

Modélisation

- **Prédictions**

  Site Energy Use (site).

  Total GreenHouse Gases (GHG).

- **Intérêt de la variable EnergySTARScore**

  Difficile à calculer.

  Valeurs manquantes.

- **Evaluation des performances du modèle**

  Sur des données non connues.

  Selon les types de bâtiments.

City of Seattle

Approche Modélisation

Étape 1	Étape 2	Étape 3	Étape 4	Étape 5
Nettoyage et filtrage des données	**Séparation des données** en jeu d'entraînement et jeu de validation	**Encodage** des features catégorielles par soit du one hot encoding, soit du ordinal encoding, soit du target encoding **Normalisation** des features catégorielles et numériques	**Apprentissage avec optimisation du modèle** avec: - Calcul des scores en Cross Validation - Influence des hyperparamètres - Importance des variables	**Évaluation** du modèle : - R^2 et Mean Absolute Error - Diagramme des résidus - Temps d'entraînement

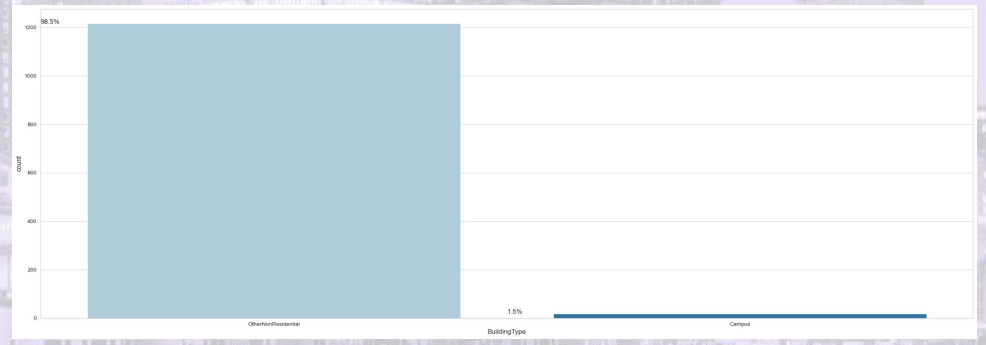
City of Seattle

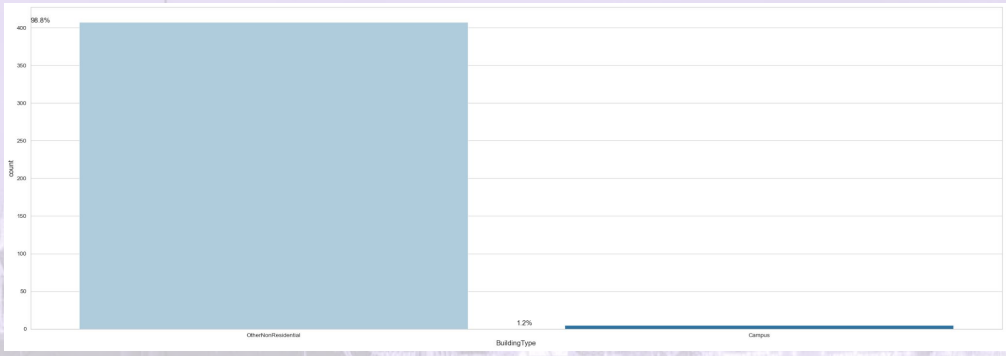
Étape 2 : Séparation des données

Train/test Split:

La méthode traintestsplit permet de séparer le jeu de données en jeu d'entraînement et jeu de validation (jeu test du modèle).

- Shuffle :



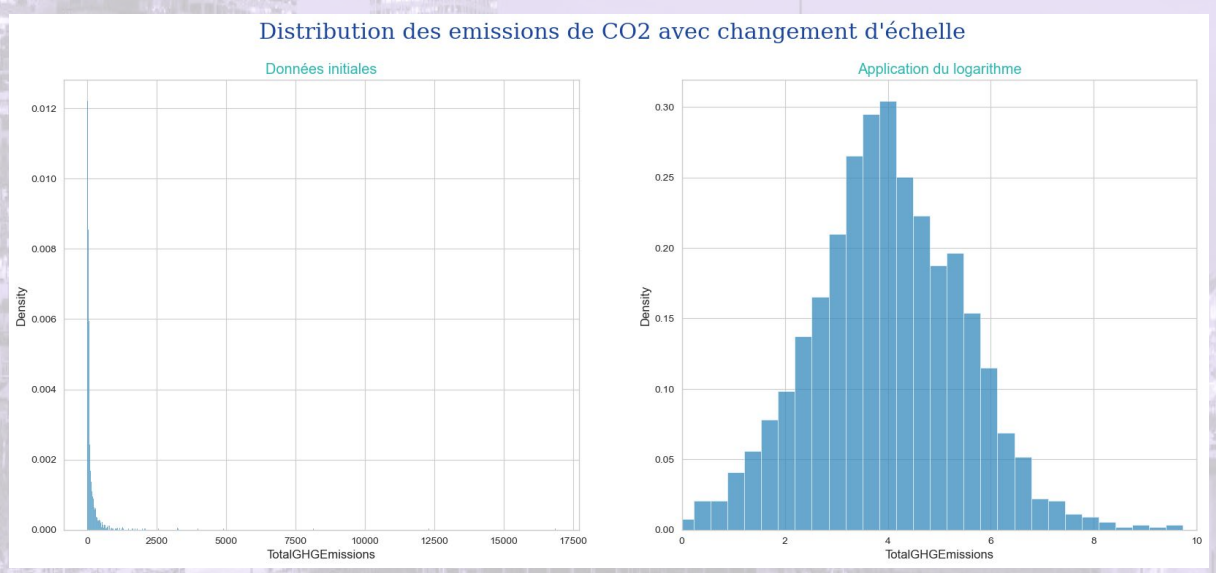


- Jeu d'entraînement : 75%
- Jeu test : 25%

City of Seattle

Étape 3 : Encodage et Numérisation

- **BuildingType:** Campus et Autre type de Résidence - One hot Encoding
- **PrimaryPropertyType et LargestPropertyType - Building Emission/Consumption Type :**
- Tri des bâtiments en fonction de leur type d'émission ou de consommation (création d'un dictionnaire)
- Mettre à jour le dictionnaire régulièrement et s'attendre à recevoir directement la variable Building Emission/Consumption Type
- Ordinal Encoding selon catégorie : High, Medium, Low
- **Neighborhood :** pas de distinction claire - Target Encoding
- **Échelle log :**



- **Normalisation des données**

City of Seattle

Étape 4 : Apprentissage et Optimisation

- **<u>Génération d'un Pipeline incluant :</u>**

- pré-traitement des variables (méthodes des modules d'encodage et de normalisation)

- soit recherche des meilleurs paramètres du modèle s'il s'agit d'une pipeline d'optimisation des hyperparamètres (par exemple avec GridSearchCV)

- soit le fitting du modèle s'il s'agit d'une pipeline d'application du modèle optimisé

City of Seattle

Étape 5 : Évaluation du modèle choisi

- **Mean Absolute Error (MAE) :**

  Intuitif, importance proportionnelle à la valeur des erreurs

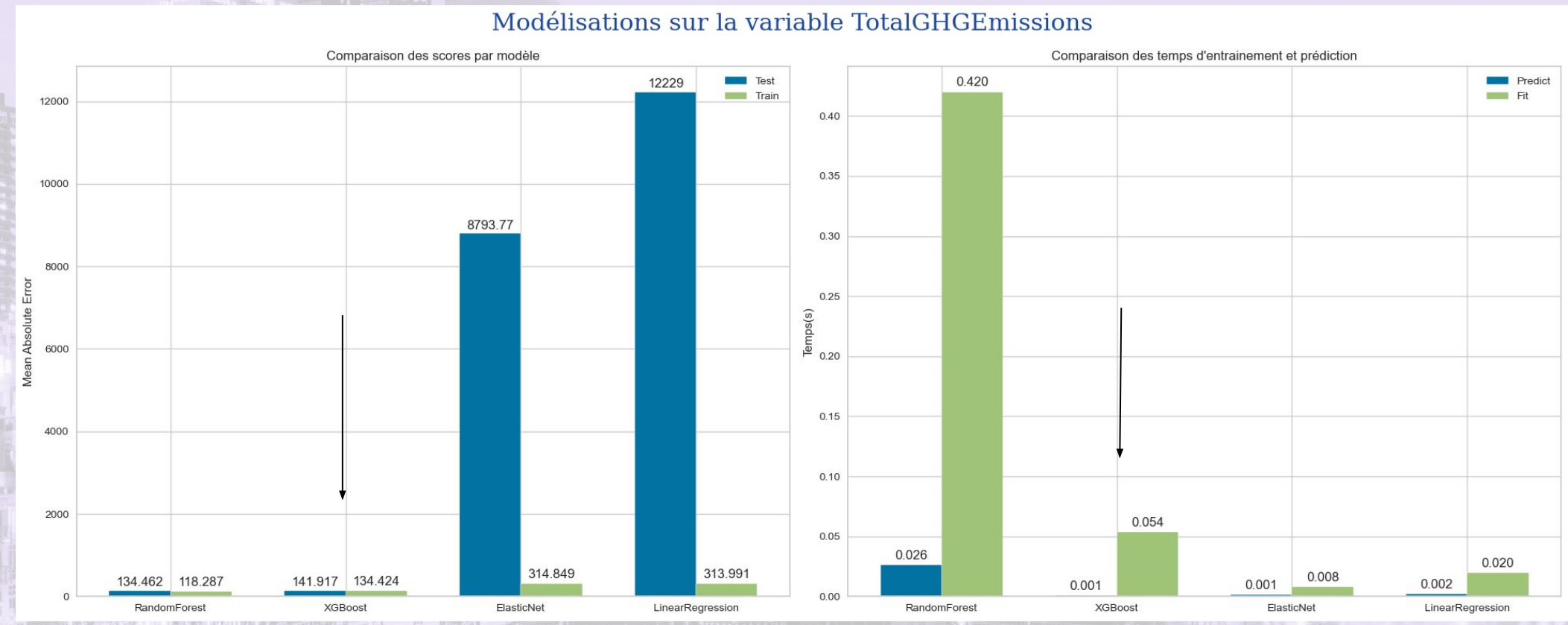
- **R^2 :**

  Évalue la proportion de variance expliquée par le modèle

- **Temps consacré à l'apprentissage**

Étape 5 : Évaluation du modèle choisi

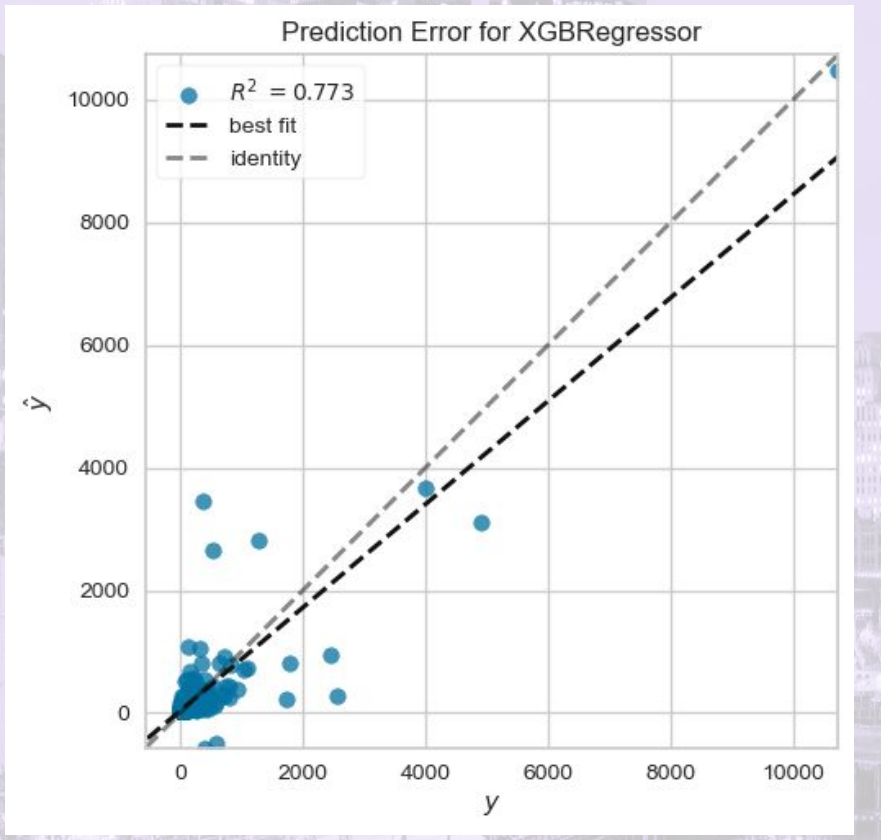
Comparaison des modèles pour les émissions de CO_2 :

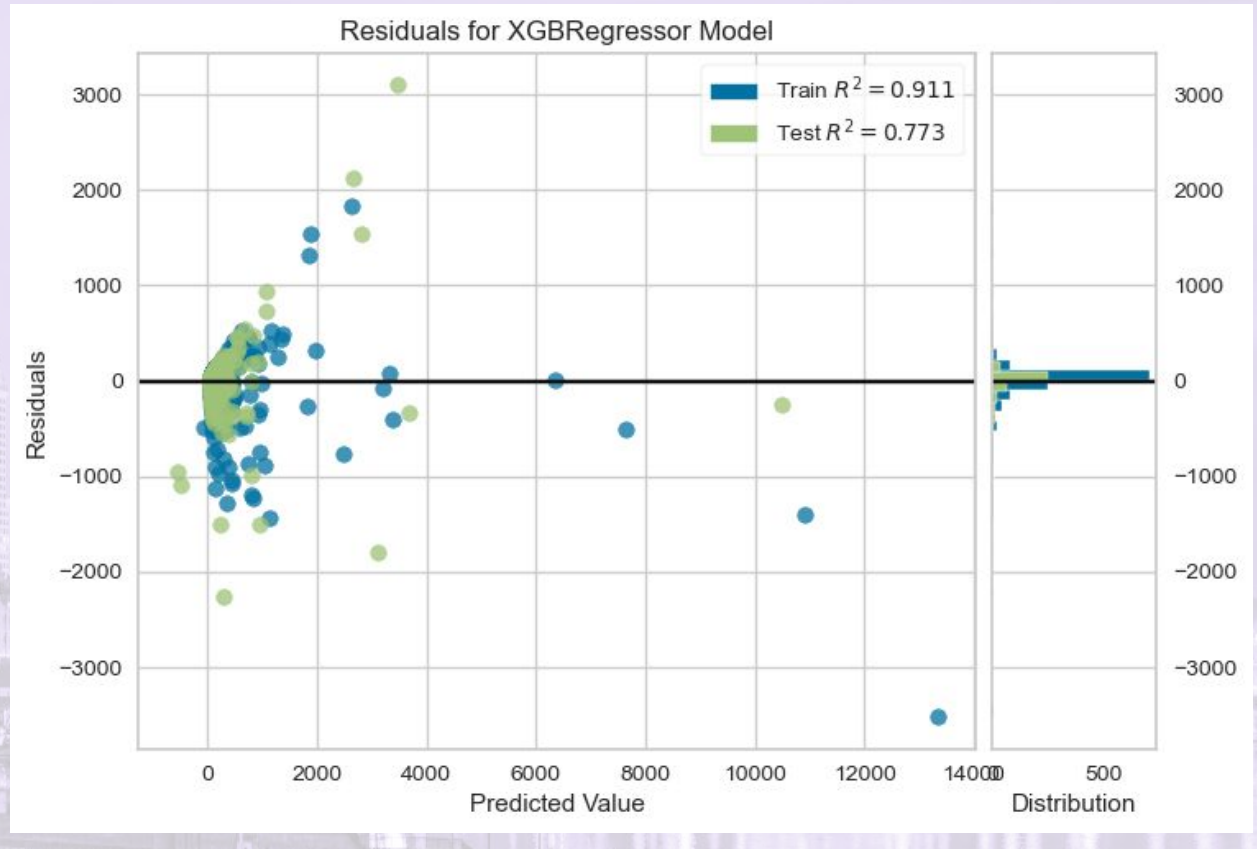


City of Seattle

Modèle XGBoost Optimisé

Pour les émissions de CO_2 :

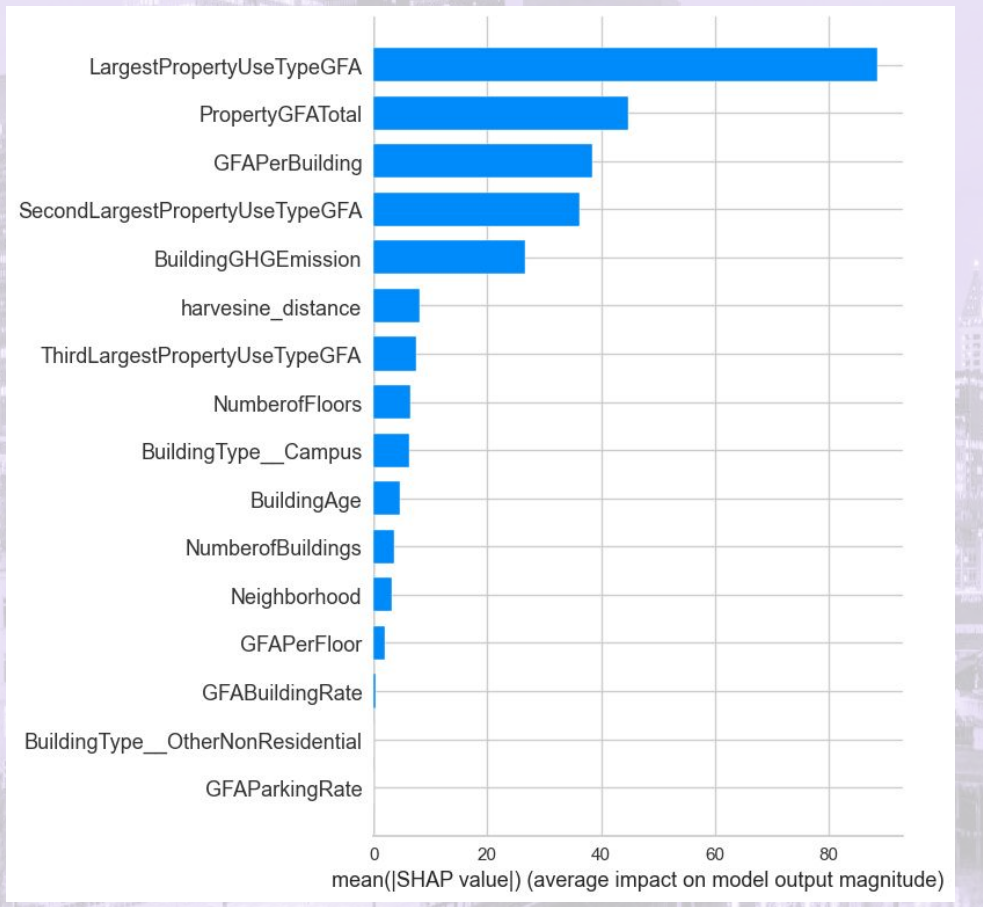


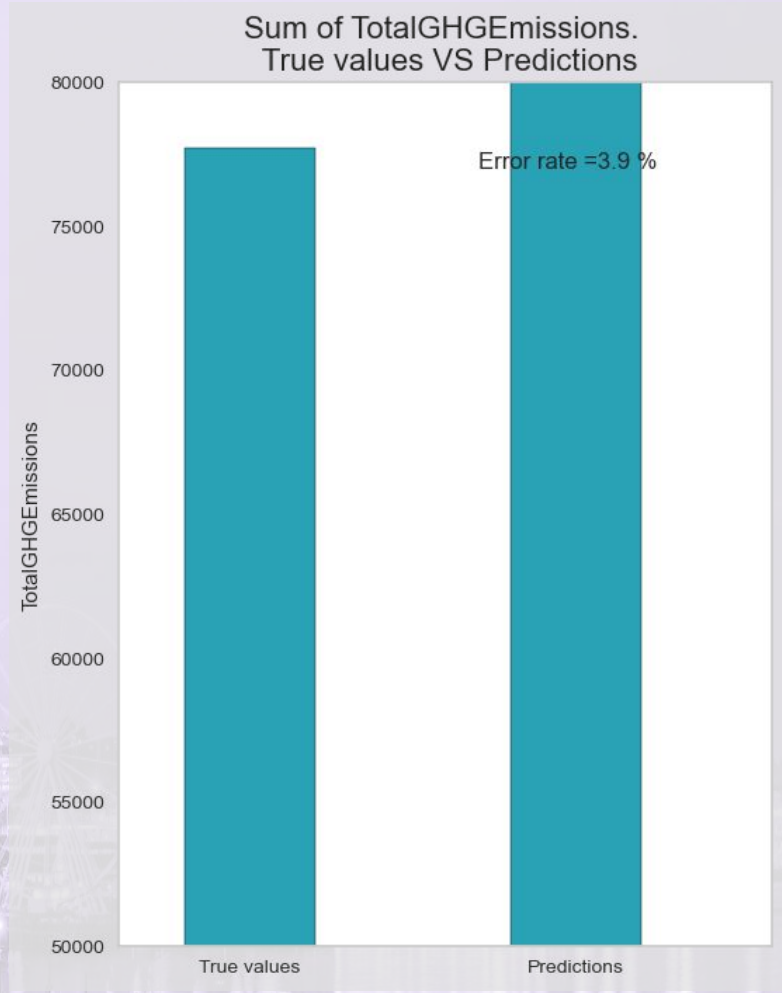


City of Seattle

Modèle XGBoost Optimisé

Pour les émissions de CO_2 :

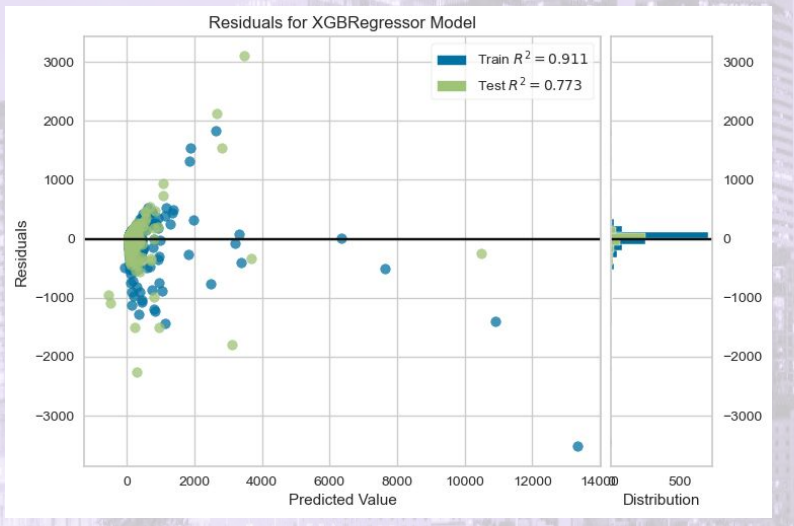


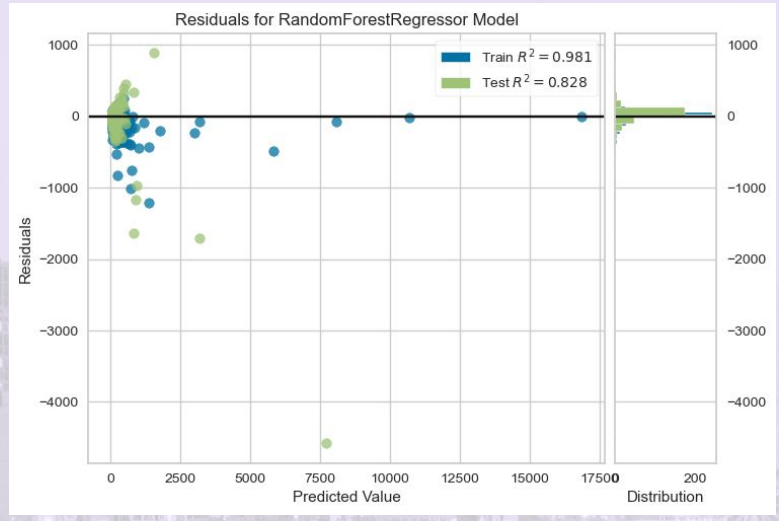


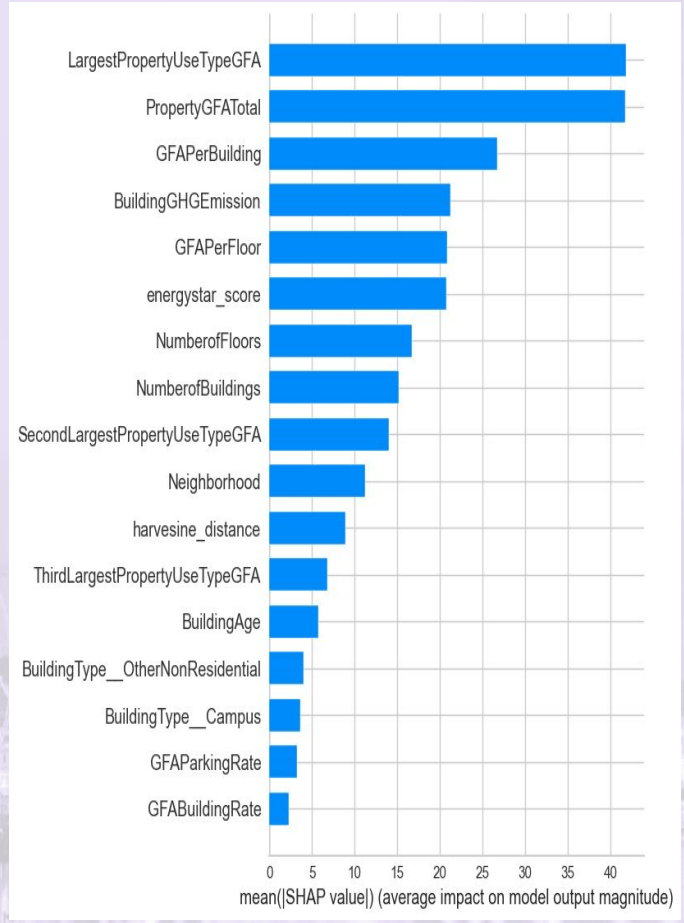
City of Seattle

Influence de EnergyStarScore

Pour les émissions de CO_2 :



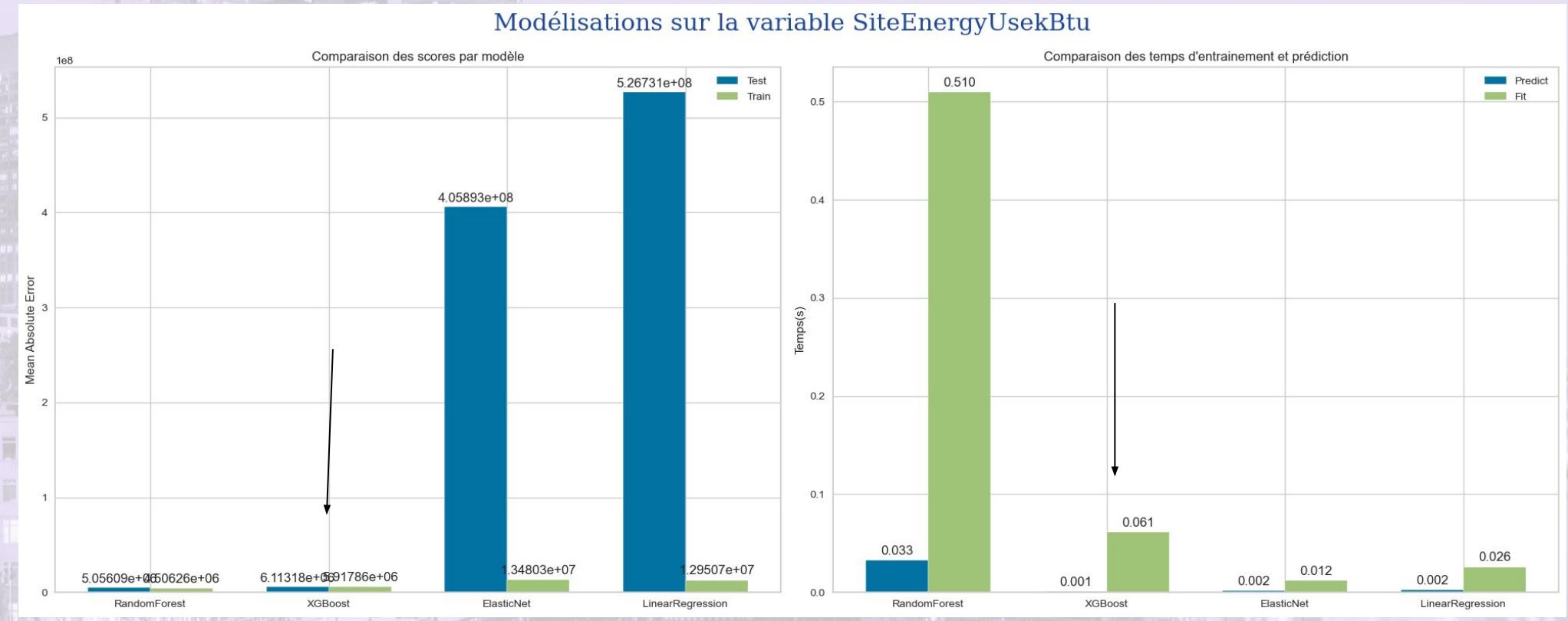




City of Seattle

Étape 5 : Évaluation du modèle choisi

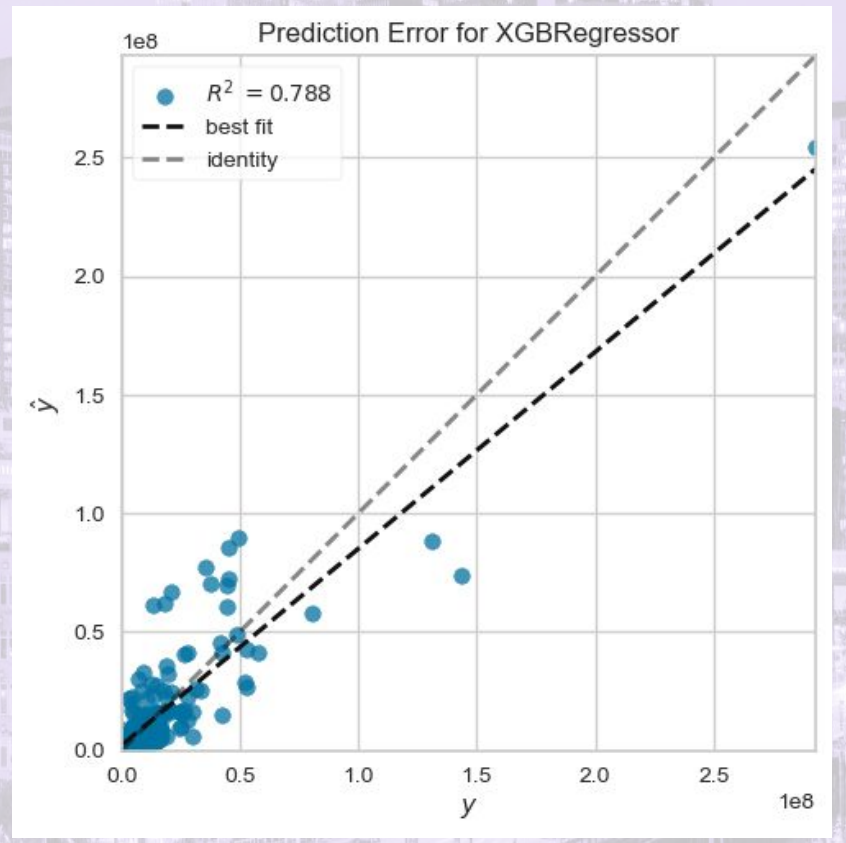
Comparaison des modèles pour la consommation énergétique :



City of Seattle

Modèle XGBoost Optimisé

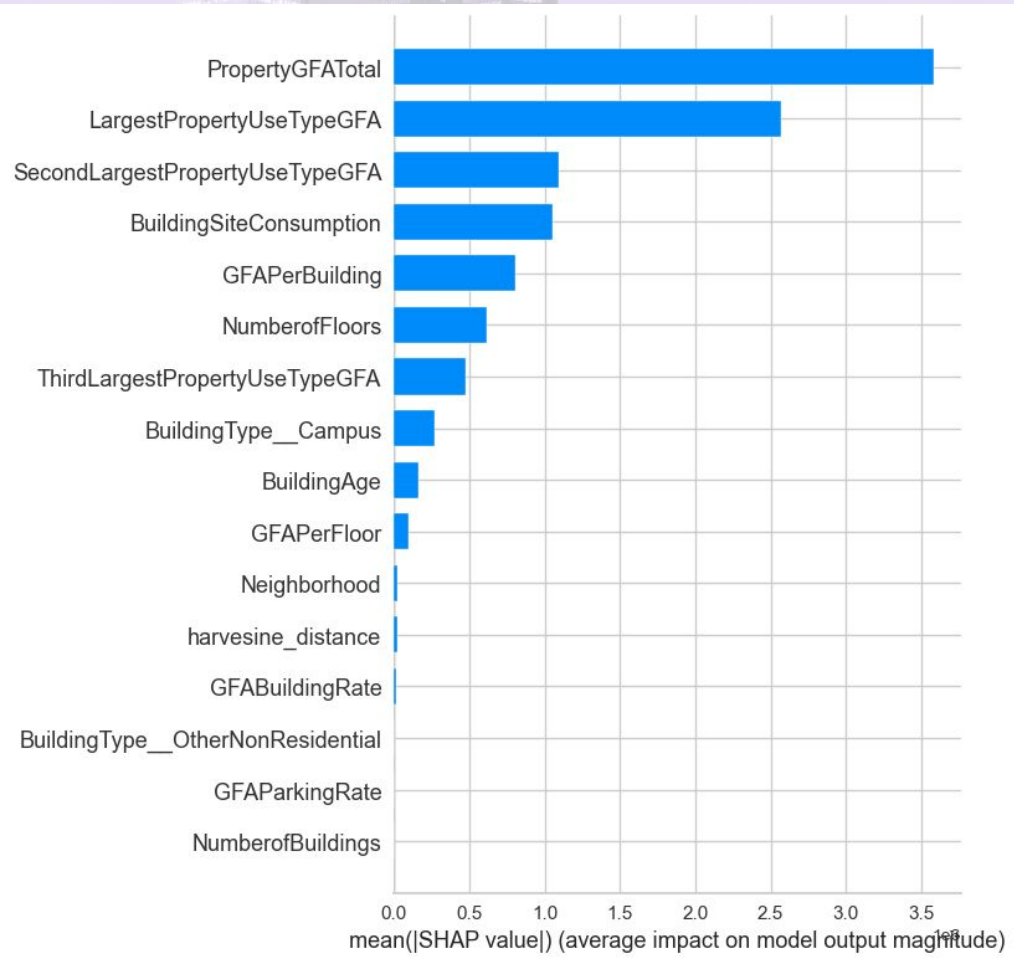
Pour la consommation d'énergie :

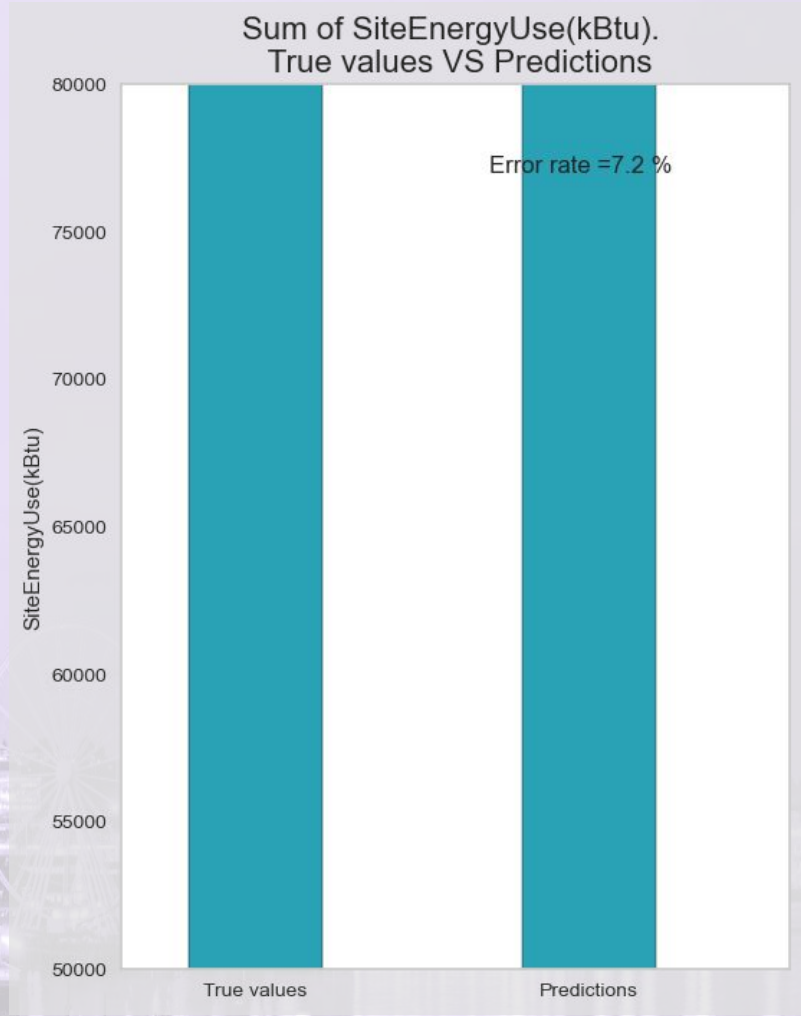


City of Seattle

Modèle XGBoost Optimisé

Pour la consommation d'énergie :

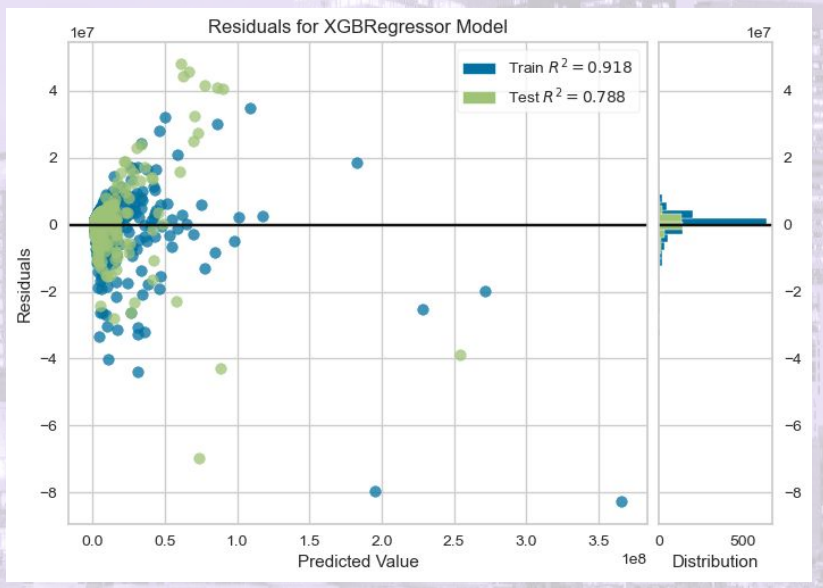


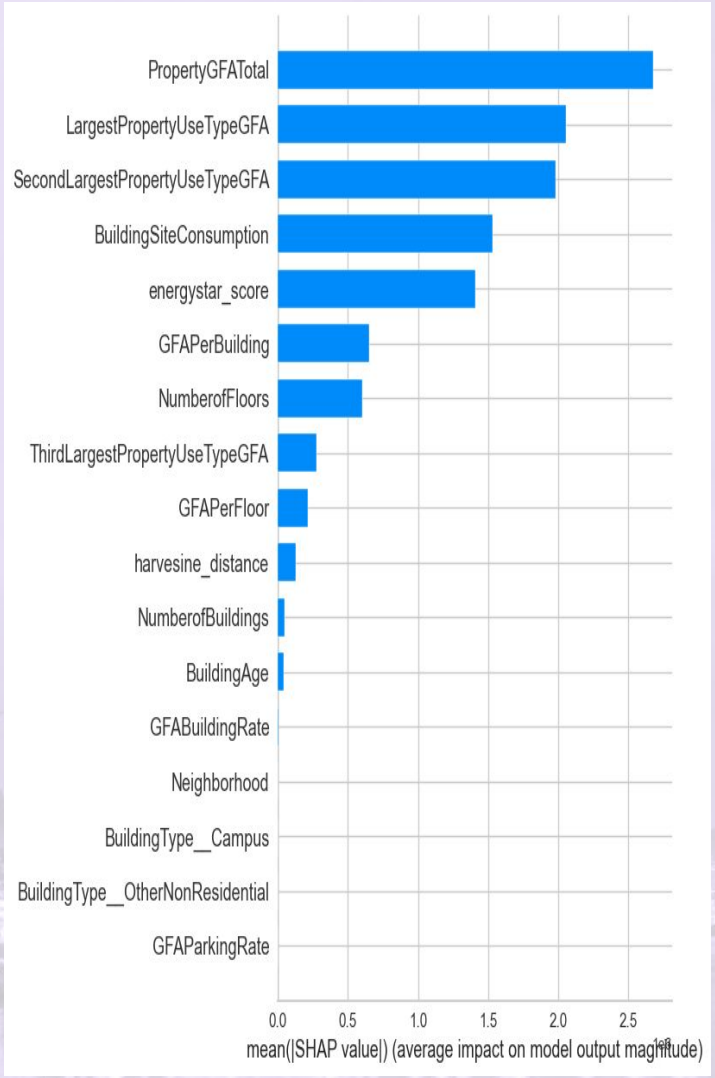


City of Seattle

Modèle XGBoost Optimisé

Pour la consommation d'énergie :





City of Seattle

Conclusions sur le modèle XGBoost

- Modèle beaucoup plus performant que les précédents
- La transformation log. de la cible améliore les scores
- Modèle très délicat à régler
- La sélection de variable n'améliore pas le modèle
- Le modèle pourrait bénéficier d'un échantillon d'apprentissage plus grand

City of Seattle

Conclusion

- **Prédiction de la consommation énergétique**
  - Site Energy Use et Total Greenhouse Gases (GHG) modélisés par 4 modèles différents
  - Résultat optimal obtenu avec XGBoost
  - Total GreenHouse Gases (GHG) modélisable grâce à la corrélation avec Site Energy Use

- **Intérêt de la variable EnergySTARScore**
  - Elle améliore les modèles lorsqu'elle est présente
  - Faible importance pourtant dans le classement des variables

- **Evaluation des performances du modèle**

- Les performances du modèle sont assez comparables à celles obtenues en cross-validation sur le training set
- Les courbes d'apprentissage montrent que les modèles pourraient être améliorés avec davantage de données

City of Seattle

Merci pour votre attention.

OPENCLASSROOMS