Projet 4 :

Anticipez les besoins en consommation électrique de bâtiments de Seattle

Nom : TRABIS

Prénom : Mohamed

Table des matières

1. Introduction
2. Préparation des données
   a) Évaluation et découverte
   b) Nettoyage et validation
3. Analyse exploratoire des données
4. Modèles prédictifs
5. Analyse prédictive de La consommation annuelle d'énergie
6. Analyse prédictive des émissions de CO2
7. Annexe - Application

Introduction

Introduction

- **Contexte** :

  Vous travaillez pour la **ville de Seattle**. Pour atteindre son objectif de ville neutre en émissions de carbone en 2050, votre équipe s'intéresse de près aux émissions des bâtiments non destinés à l'habitation.

- **Problématique de la ville de Seattle :**

  Des relevés minutieux ont été effectués par des agents en 2015 et en 2016. Cependant, ces relevés sont coûteux à obtenir, et à partir de ceux déjà réalisés, **vous voulez tenter de prédire les émissions de CO2 et la consommation totale d'énergie** de bâtiments pour lesquels elles n'ont pas encore été mesurées.

Introduction

- **Mission** :

  Voici un récapitulatif de la mission :

  → Réaliser une courte analyse exploratoire.

  → Tester différents modèles de prédiction afin de répondre au mieux à la problématique.

  Quelques conseils du **project lead**:

  ✓ Déduire des variables plus simples (nature et proportions des sources d'énergie utilisées).

  ✓ Optimiser les performances en appliquant des transformations simples aux variables (normalisation, passage au log, etc.).

  ✓ Mettre en place une évaluation rigoureuse des performances de la régression, et optimiser les hyperparamètres et le choix d'algorithme de ML à l'aide d'une validation croisée.

Préparation des données

Préparation des données- Évaluation et découverte des données

- Les données de consommation énergétique sont à télécharger à cette adresse.

- Des relevés minutieux ont été effectués par des agents en 2015 et en 2016, sous format de deux fichiers CSV :

1. 2015-building-energy-benchmarking.csv
2. 2016-building-energy-benchmarking.csv

- Les fichiers CSV ne sont pas volumineux (3Mo pour les deux fichiers), ils contiennent :

→ *Des relevés d'environ de 3400 bâtiments de la ville de Seattle de 2015 et 2016*

- La structure des données a changé entre 2015 et 2016.

Préparation des données- Nettoyage et validation des données

- Les étapes effectuées pour le nettoyage et la validation des données :

  - Importer les deux fichiers.
  - Vérifier les colonnes des deux fichiers.
  - Mutualiser les colonnes.
  - Détecter et supprimer les valeurs aberrantes.
  - Remplir les valeurs manquantes

Préparation des données- Nettoyage et validation des données

- Nettoyage du fichier de 2015 :

  - Splitter la colonne « ***Location*** » qui contient les données de géolocalisation et adresse des batiments en 6 colonnes : *'Latitude', 'Longitude', 'Address', 'City', 'State'*

  - Renommer les colonnes suivantes :
    - *"Zip Codes": "ZipCode",*
    - *"GHGEmissions(MetricTonsCO2e)":"TotalGHGEmissions",*
    - *"latitude":"Latitude",*
    - *"longitude":"Longitude",*
    - *"city":"City",*
    - *"address":"Address",*
    - *"state":"State",*
    - *"Comment":"Comments",*
    - *"GHGEmissionsIntensity(kgCO2e/ft2)":"GHGEmissionsIntensity"*

Préparation des données- Nettoyage et validation des données

- Supprimer les colonnes ci-dessous, car elles ne sont pas présentent dans le fichiers de 2016 :

  'SPD Beats', '2010 Census Tracts', 'City Council Districts', 'OtherFuelUse(kBtu)', 'zip',

  'Seattle Police Department Micro Community Policing Plan Areas'

- **Remarque :**

  Suite à ce nettoyage nous avons maintenant les même colonnes dans les deux datasets 2015 et 2016.

  Pour mieux analyser les données, il faut effectuer une jointure entre les deux datasets .

  Notre dataset finale contient 6716 lignes et 46 colonnes.

Préparation des données- Nettoyage et validation des données

- Description des colonnes de la base de données :

Colonnes	Description
OSEBuildingID	Un identifiant unique attribué à chaque propriété couverte par la Seattle à des fins de suivi et d'identification.
DataYear	Année de la collection des données
BuildingType	Classification du type de bâtiment de la ville de Seattle.
PrimaryPropertyType	L'utilisation principale d'une propriété (par exemple, un bureau, un magasin de détail). L'usage principal est défini comme une fonction qui représente plus de 50 % d'un bien. Il s'agit du champ Type de propriété - EPA calculé de Portfolio Manager.
PropertyName	Nom officiel ou de propriété commune.
TaxParcelIdentificationNumber	NIP du comté de King de la propriété
Location	Emplacement
CouncilDistrictCode	Propriété District municipal de la ville de Seattle.
Neighborhood	Quartier
YearBuilt	Année au cours de laquelle une propriété a été construite ou a subi une rénovation complète.
NumberofBuildings	Nombre de bâtiments
NumberofFloors	Nombre d'étages
PropertyGFATotal	Superficie totale du bâtiment et du stationnement.
PropertyGFAParking	Espace total en pieds carrés de tous les types de stationnement (entièrement fermé, partiellement fermé et ouvert).
PropertyGFABuilding(s)	Espace au sol total en pieds carrés entre les surfaces extérieures des murs d'enceinte d'un bâtiment. Cela comprend toutes les zones à l'intérieur du ou des bâtiments, telles que l'espace des locataires, les espaces communs, les cages d'escalier, les sous-sols, le stockage, etc.
ListOfAllPropertyUseTypes	Toutes les utilisations de la propriété signalées dans Portfolio Manager
LargestPropertyUseType	La plus grande utilisation d'une propriété (par exemple, bureau, magasin de détail) par GFA
LargestPropertyUseTypeGFA	La surface de plancher brute (GFA) de la plus grande utilisation de la propriété.
SecondLargestPropertyUseType	La deuxième plus grande utilisation d'une propriété (par exemple, bureau, magasin de détail) par GFA
SecondLargestPropertyUseTypeGFA	La troisième plus grande utilisation d'une propriété (par exemple, bureau, magasin de détail) par GFA.
YearsENERGYSTARCertified	Années où la propriété a reçu la certification ENERGY STAR.

Préparation des données- Nettoyage et validation des données

ENERGYSTARScore	Une note de 1 à 100 calculée par l'EPA évalue la performance énergétique globale d'une propriété, sur la base de données nationales pour contrôler les différences entre le climat, les utilisations du bâtiment et les opérations. Un score de 50 représente la médiane nationale.
SiteEUI(kBtu/sf)	La consommation énergétique du site est la quantité annuelle de toute l'énergie consommée par la propriété sur le site, telle qu'elle est indiquée sur les factures de services publics. Le site EUI est mesuré en milliers d'unités thermiques britanniques (kBtu) par pied carré.
SourceEUI(kBtu/sf)	La consommation d'énergie à la source est l'énergie annuelle utilisée pour exploiter la propriété, y compris les pertes de production, de transmission et de distribution. La source EUI est mesurée en milliers d'unités thermiques britanniques (kBtu) par pied carré.
SiteEUIWN(kBtu/sf)	L'énergie du site WN correspond à la consommation d'énergie du site que la propriété aurait consommée dans des conditions météorologiques moyennes sur 30 ans. WN Site EUI est mesuré en milliers d'unités thermiques britanniques (kBtu) par pied carré.
SourceEUIWN(kBtu/sf)	L'énergie de la source WN est la consommation d'énergie de la source que la propriété aurait consommée dans des conditions météorologiques moyennes sur 30 ans.
SiteEnergyUse(kBtu)	La quantité annuelle d'énergie consommée par la propriété à partir de toutes les sources d'énergie.
SiteEnergyUseWN(kBtu)	La quantité annuelle d'énergie consommée par la propriété à partir de toutes les sources d'énergie, ajustée à ce que la propriété aurait consommé dans des conditions météorologiques moyennes sur 30 ans.
SteamUse(kBtu)	La quantité annuelle de vapeur urbaine consommée par la propriété sur place, mesurée en milliers d'unités thermiques britanniques (kBtu).
Electricity(kWh)	La quantité annuelle d'électricité consommée par la propriété sur place, y compris l'électricité achetée au réseau et produite par les systèmes renouvelables sur place, mesurée en kWh.
Electricity(kBtu)	La quantité annuelle d'électricité consommée par la propriété sur place,
NaturalGas(therms)	La quantité annuelle de gaz naturel fourni par les services publics consommée par la propriété, mesurée en therm.
NaturalGas(kBtu)	La quantité annuelle de gaz naturel fourni par les services publics consommée par la propriété, mesurée en milliers d'unités thermiques britanniques (kBtu).
TotalGHGEmissions	La quantité totale d'émissions de gaz à effet de serre, y compris le dioxyde de carbone,
GHGEmissionsIntensity(kgCO2e/ft2)	Émissions totales de gaz à effet de serre divisées par la surface de plancher brute de la propriété, mesurée en kilogrammes d'équivalent de dioxyde de carbone par pied carré. Ce calcul utilise un facteur d'émissions de GES du portefeuille de ressources de production de Seattle City Light
DefaultData	La propriété a utilisé des données par défaut pour au moins une caractéristique de propriété.
ComplianceStatus	Si une propriété a satisfait aux exigences d'analyse comparative énergétique pour l'année de déclaration en cours.
Outlier	Si une propriété est une valeur aberrante élevée ou faible.

Préparation des données- Nettoyage et validation des données

- Les étapes effectuées pour conserver que les bâtiments non destinés à l'habitation :

  - Dans la colonne «***BuildingType***» nous avons les types de bâtiments ci-dessous :

  'NonResidential', 'Nonresidential COS', 'Multifamily MR (5-9)', 'SPS-District K-12', 'Multifamily LR (1-4)', 'Campus', 'Multifamily HR (10+)', 'Nonresidential WA'

  - Suppression des lignes qui contiennent la chaîne de caractère « ***Multifamily*** ».

- **Remarque :**

  Suite à cette suppression la dataset contient 3318 lignes et 46 colonnes, ce qui représente environ 50% des données initiales.

Préparation des données- Nettoyage et validation des données

- Ci-dessous Les statistiques descriptives de la dataset :

	PropertyGFAParking	PropertyGFABuilding(s)	SourceEUI(kBtu/sf)	SourceEUIWN(kBtu/sf)	Electricity(kBtu)	TotalGHGEmissions
count	3318.00	3318.00	3309.00	3309.00	3309.00	3309.00
mean	13303.30	102363.90	175.44	178.66	5636555.58	177.04
std	43596.62	234074.87	180.79	180.63	17409003.50	666.44
min	-2.00	-50550.00	-2.00	-2.10	-115417.00	-0.80
25%	0.00	28507.75	76.20	80.80	723667.00	19.72
50%	0.00	47368.00	131.30	134.80	1623657.00	49.16
75%	0.00	94471.50	204.90	207.80	4878886.00	138.87
max	512608.00	9320156.00	2620.00	2620.00	657074389.00	16870.98

- **Remarque :**

  On constate la présence des valeurs négatives pour des colonnes de consommations et de surfaces.

  Nous allons conserver que les lignes avec des valeurs supérieurs à 0.

Préparation des données- Nettoyage et validation des données

- Détecter et supprimer les valeurs aberrantes :
  - Supprimer les lignes avec les valeurs renseignés dans la colonne «Outlier » puis supprimer cette colonne :

Outlier	% Dataset
Low Outlier	0.64%
High Outlier	0.36%
NaN	99%

  - Supprimer les lignes qui sont inférieurs au dernier centile de la colonne «***Electricity(kBtu)***».

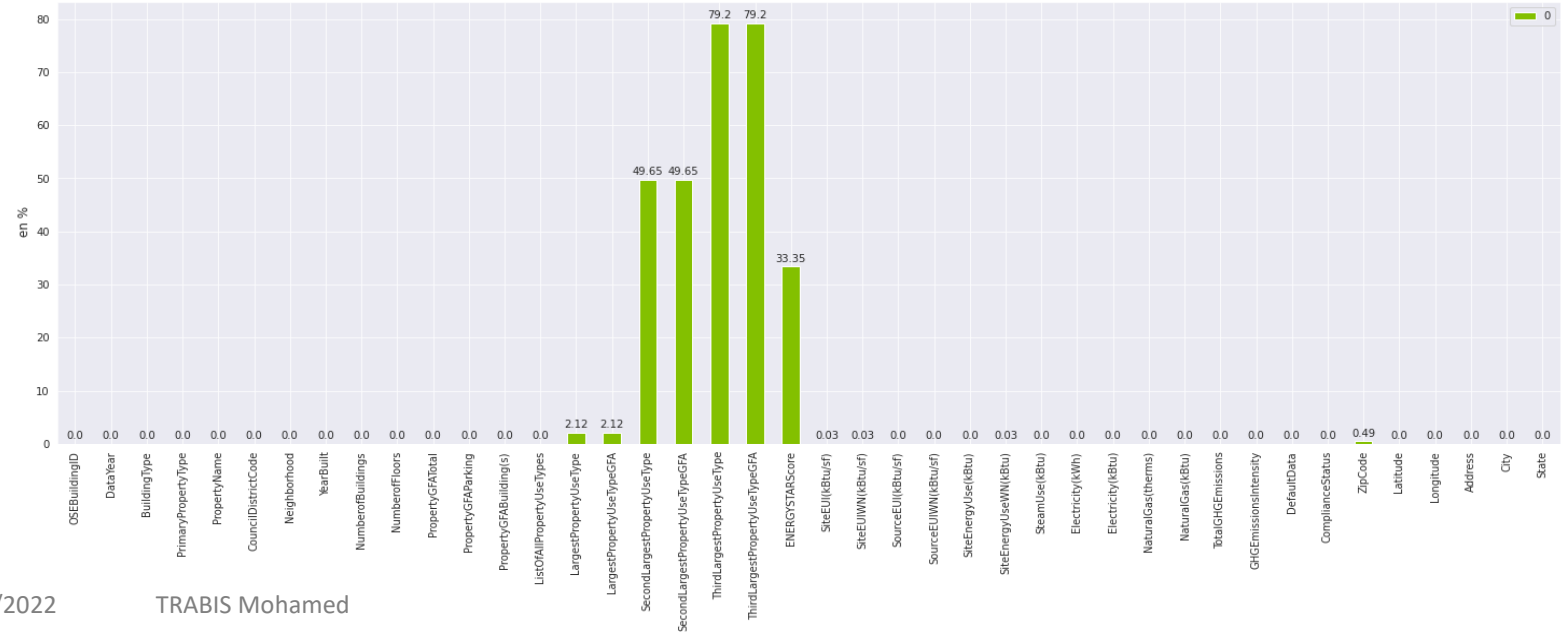
- **<u>Remarque :</u>**

  La dataset contient 3226 lignes et 46 colonnes.

Préparation des données- Nettoyage et validation des données

- Traiter les valeurs manquantes :
  - Remplacer les valeurs manquantes de la colonne «***ListOfAllPropertyUseTypes*** » par les valeurs de la colonne «***PrimaryPropertyType*** »
  - Remplacer les 'NaN' de la colonne ***'NumberofFloors***' par 0.

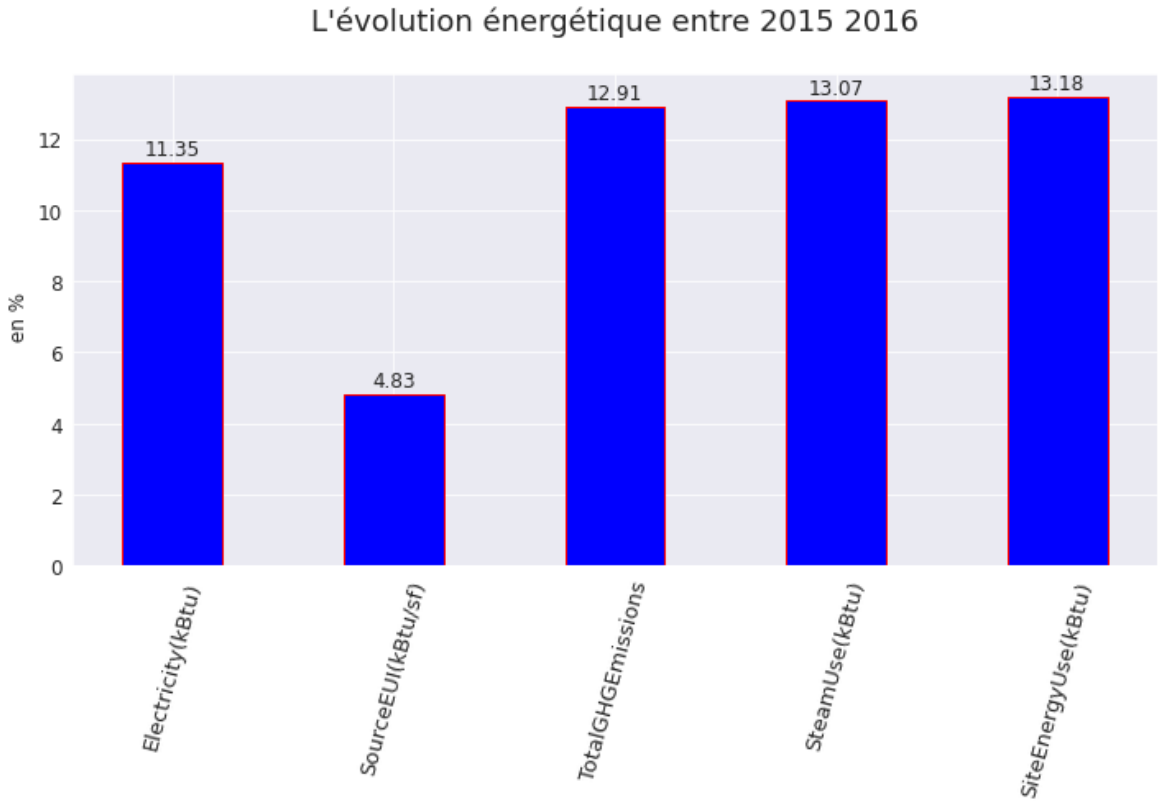
- Graphique des valeurs manquantes :



Analyse exploratoire des données

Analyse exploratoire des données

- Graphique de l'évolution énergétique entre 2015 et 2016 :



- **Remarque :** On constate que La quantité annuelle d'énergie consommée a augmenter de 13% de 2015 à 2016

Analyse exploratoire des données - ENERGY Star

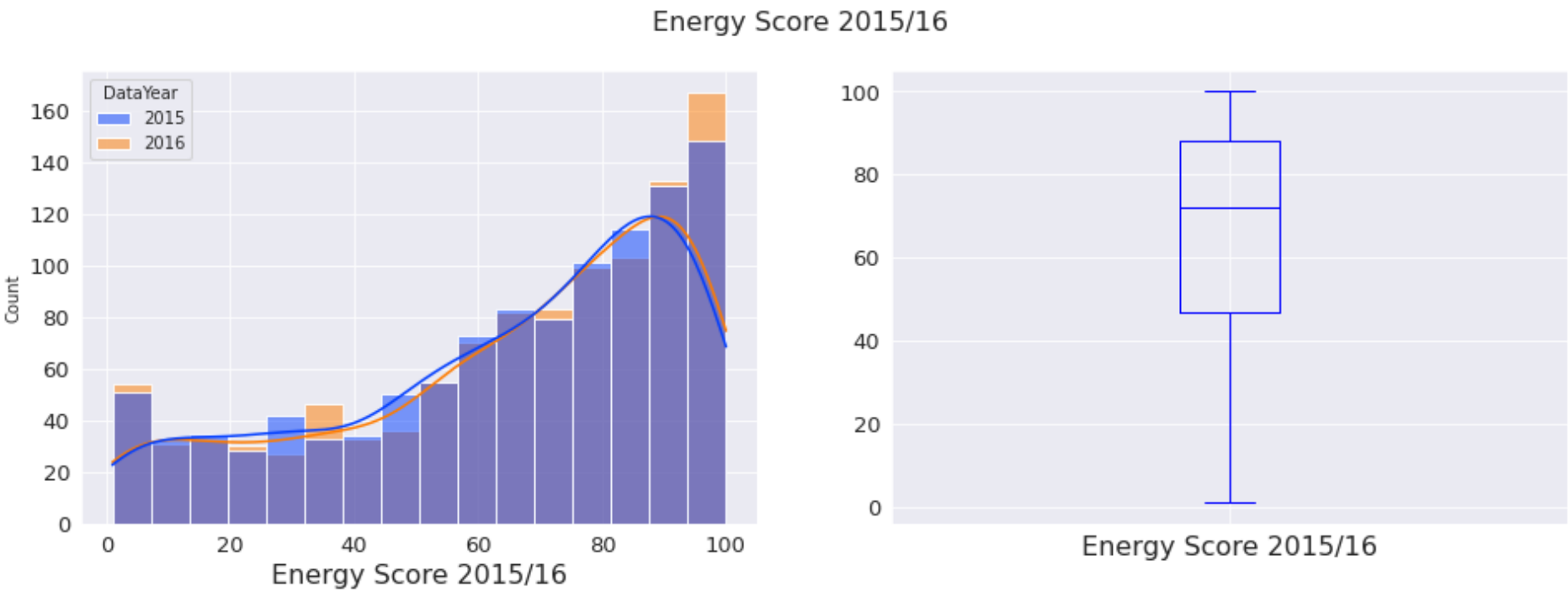
- **Le Score ENERGY STAR** :

  Le score ENERGY STAR (sur une échelle de 1 à 100), est un outil d'évaluation qui vous aide à évaluer les performances de votre propriété par rapport à des bâtiments similaires à l'échelle nationale. Cela vous aidera à identifier les propriétés de votre portefeuille à cibler pour une amélioration ou une reconnaissance. Un score de 50 est la médiane. Ainsi, si votre propriété obtient un score inférieur à 50, cela signifie qu'elle est moins performante que 50 % des propriétés similaires à l'échelle nationale, tandis qu'un score supérieur à 50 signifie qu'elle fonctionne mieux que 50 % par rapport aux autres propriétés. Un score de 75 ou plus signifie qu'elle est la plus performante et qu'elle peut être éligible à la certification ENERGY STAR.

- Le score ENERGY STAR donne un aperçu complet de la performance énergétique de votre propriété. Il évalue les actifs physiques du bâtiment, les opérations et le comportement des occupants, basé sur des données réelles et mesurées.

Analyse exploratoire des données - ENERGY Star

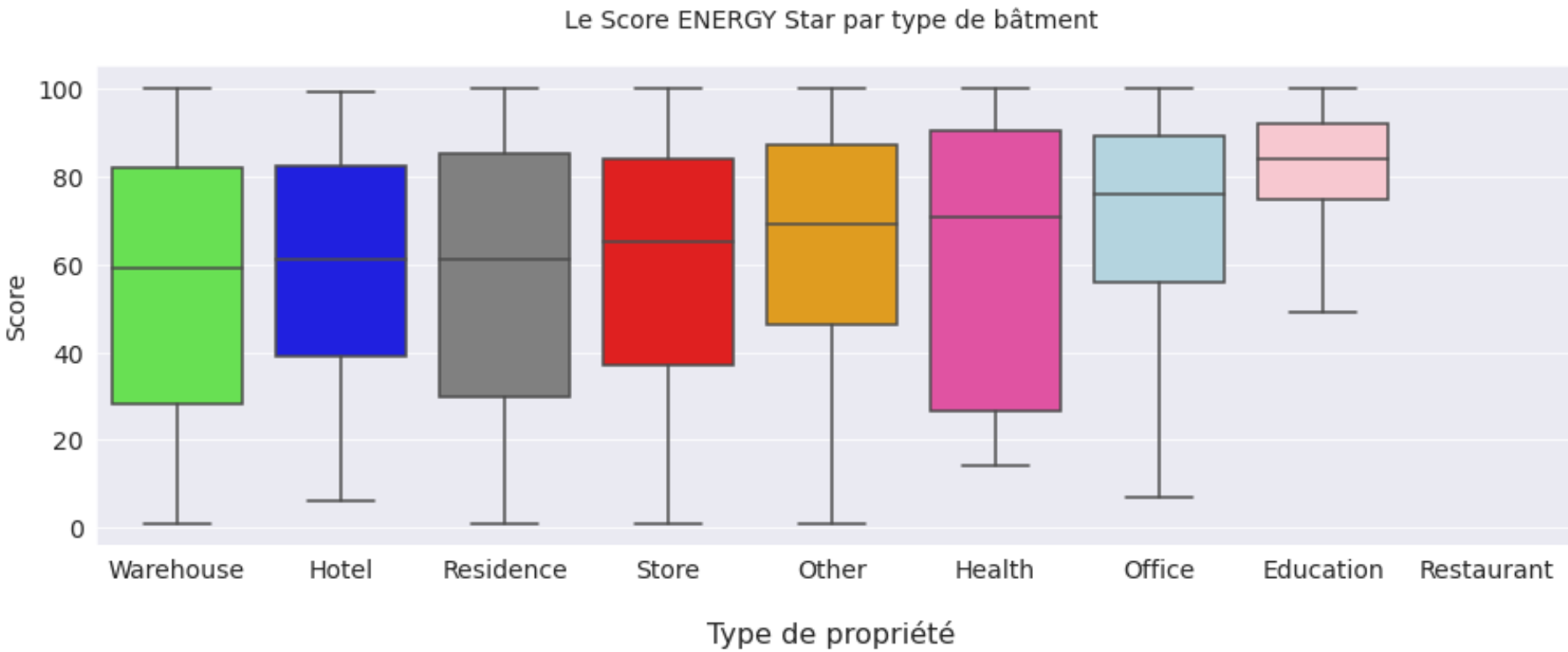
- Ci-dessous le graphique représentant le Score ENERGY Star:



- **Remarque :** On constate que la valeur médiane du score énergétique est d'environ 73 sur 100, on constate ausi une légère amélioration du score en 2016.

Analyse exploratoire des données - ENERGY Star

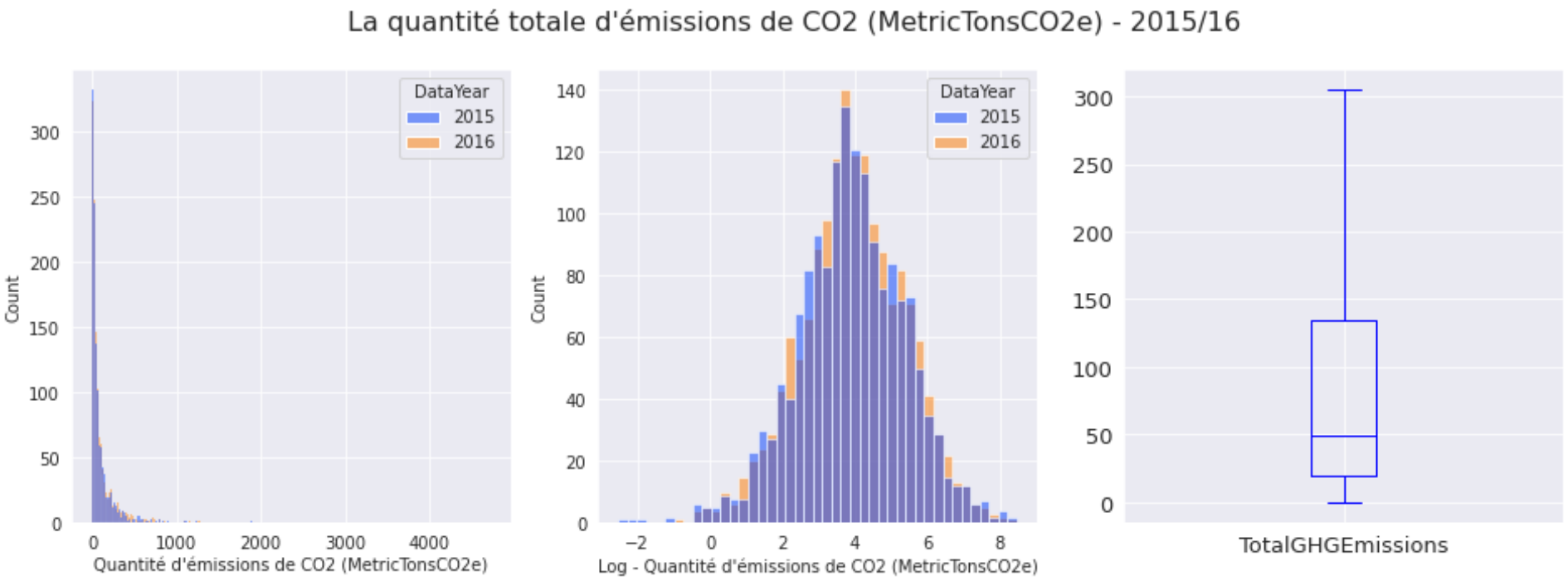
- Le Score ENERGY Star par type de bâtiment:



- **Remarque :** Les bâtiments de type « Education » ont un bon score énergétique(une médiane d'environ 83 sur 100), contrairement aux centres de distribution et stockage.

Analyse exploratoire des données - Emissions de CO2

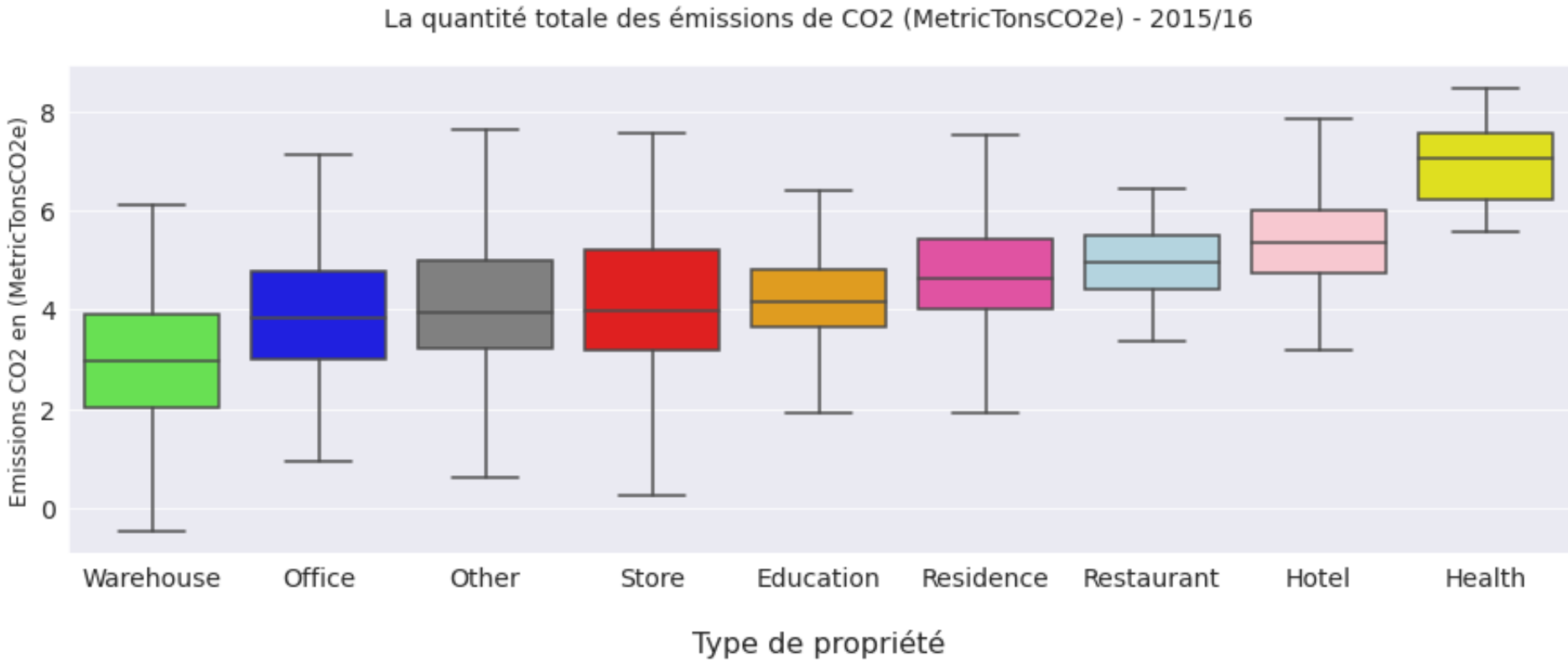
- Graphique des émissions de CO2 (2015/2016):



- **Remarque :** La majorité des bâtiments ont une quantité totale des émissions de CO2 entre 0 et 300 MetricTonsCO2e;

Analyse exploratoire des données - Emissions de CO2

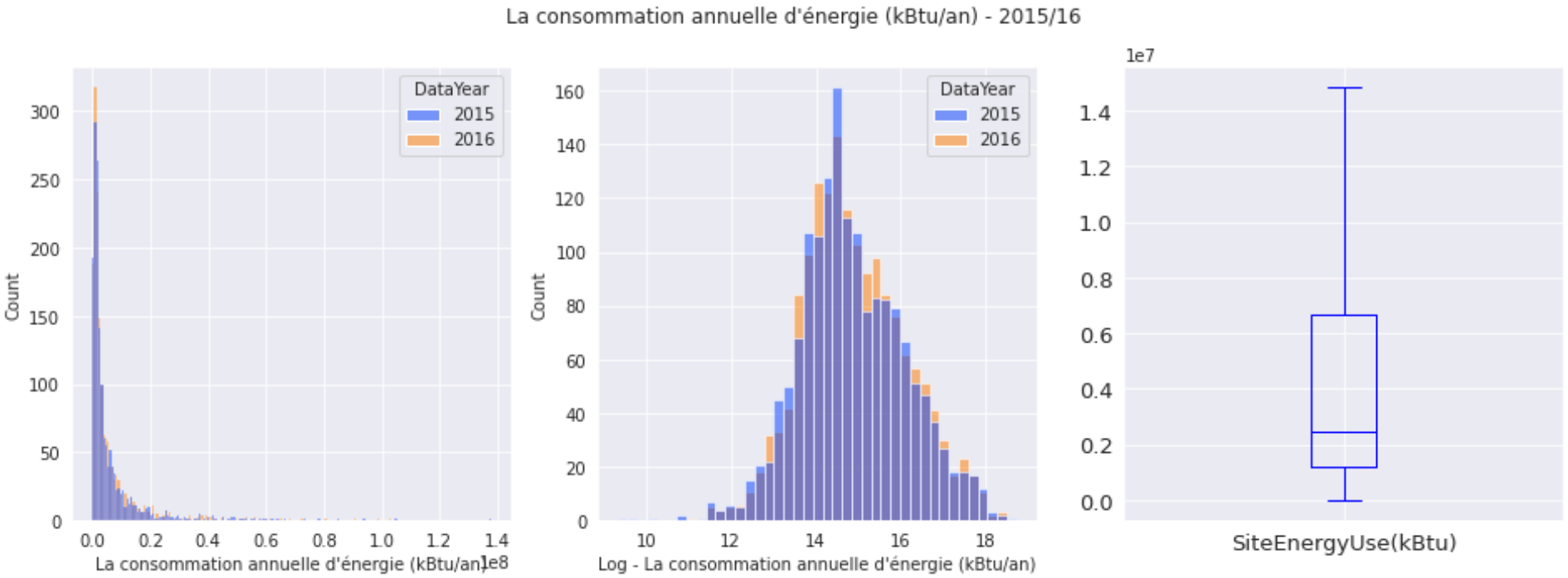
- Les émissions de CO2 par type de bâtiment:



- **Remarque :** Les bâtiments qui émettent une quantité élevée de CO2 sont : les hôtels ,les hôpitaux et les laboratoire

Analyse exploratoire des données - Consommation d'énergie

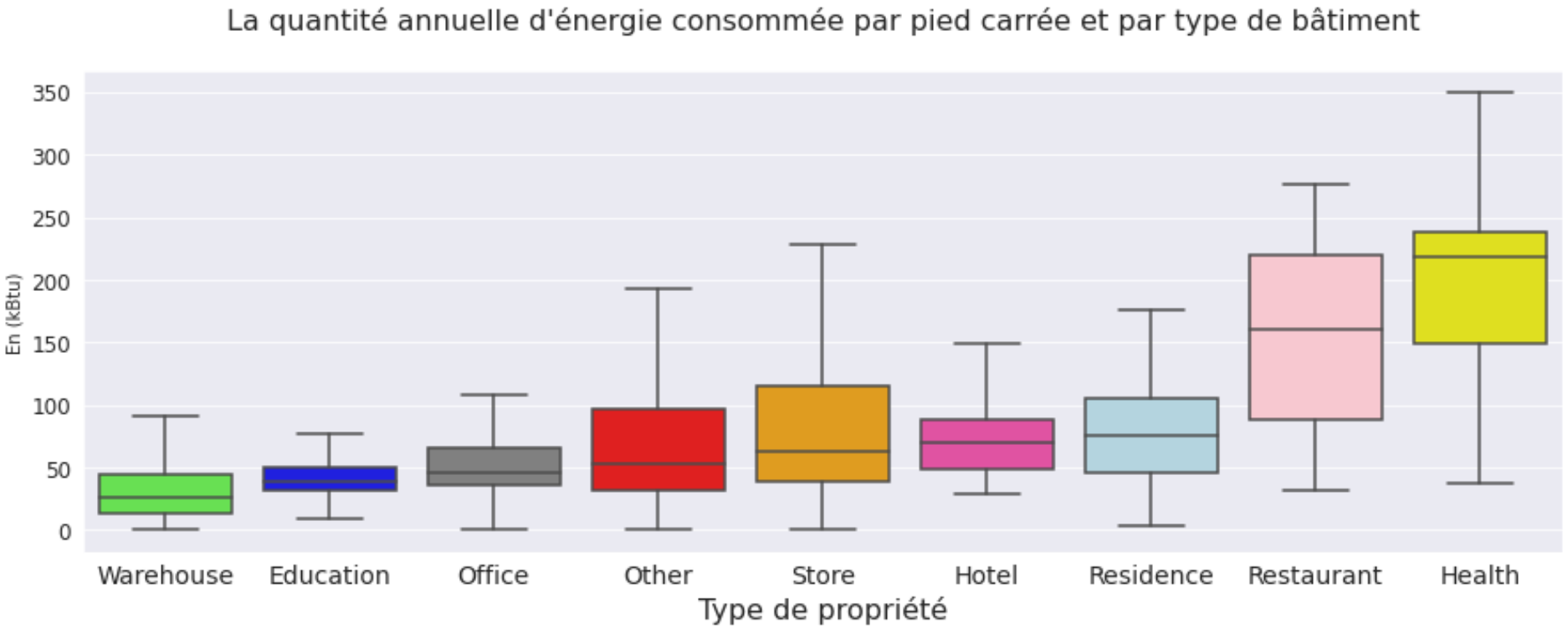
- Graphique de La quantité annuelle d'énergie consommée :



- **Remarque :** La valeur médiane de la consommation annuelle d'énergie est de 0.25le7.

Analyse exploratoire des données - Consommation d'énergie

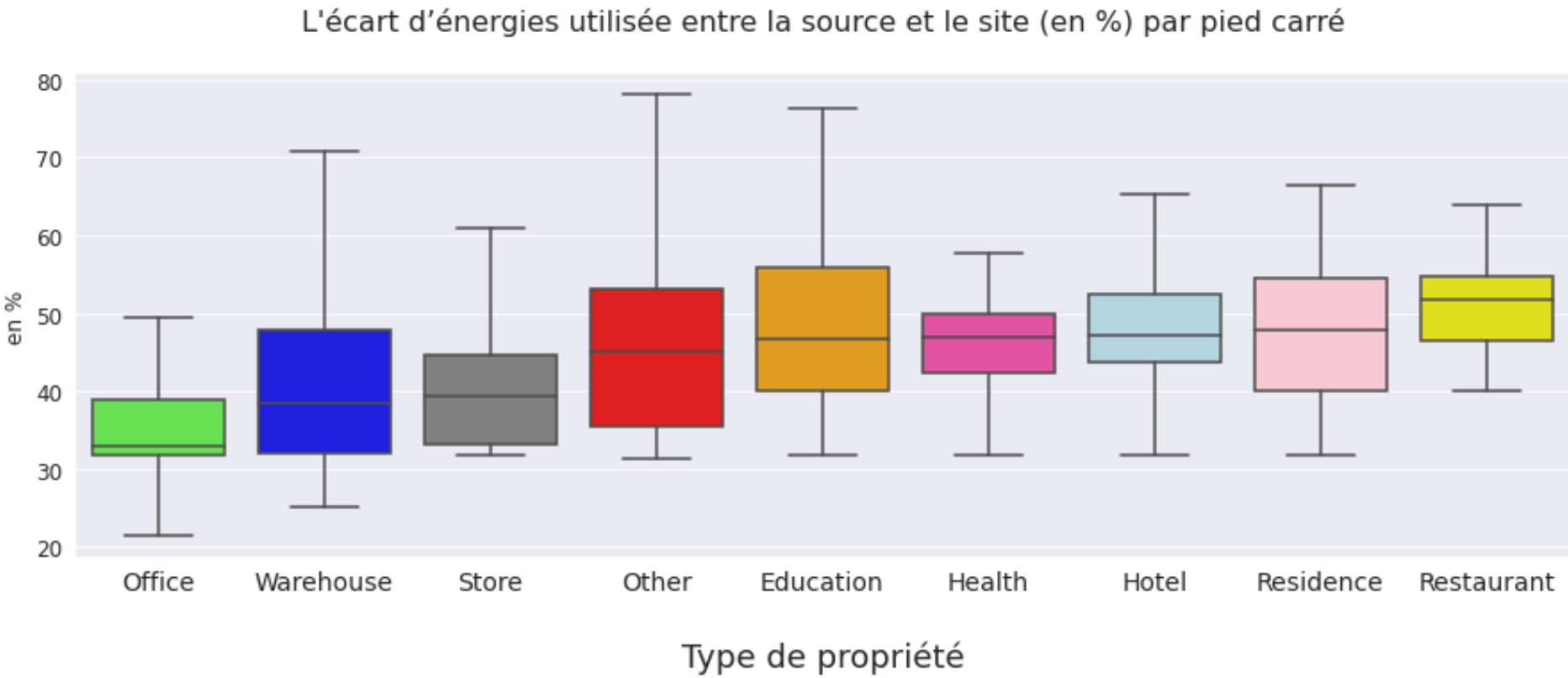
- La quantité annuelle d'énergie consommée par pied carrée et par type de bâtiment:



- **Remarque :** Comme pour les émissions de CO2 les hôtels ,les hôpitaux et les laboratoires sont les plus gros consommateurs d'énergie.

Analyse exploratoire des données – Ecart site / source

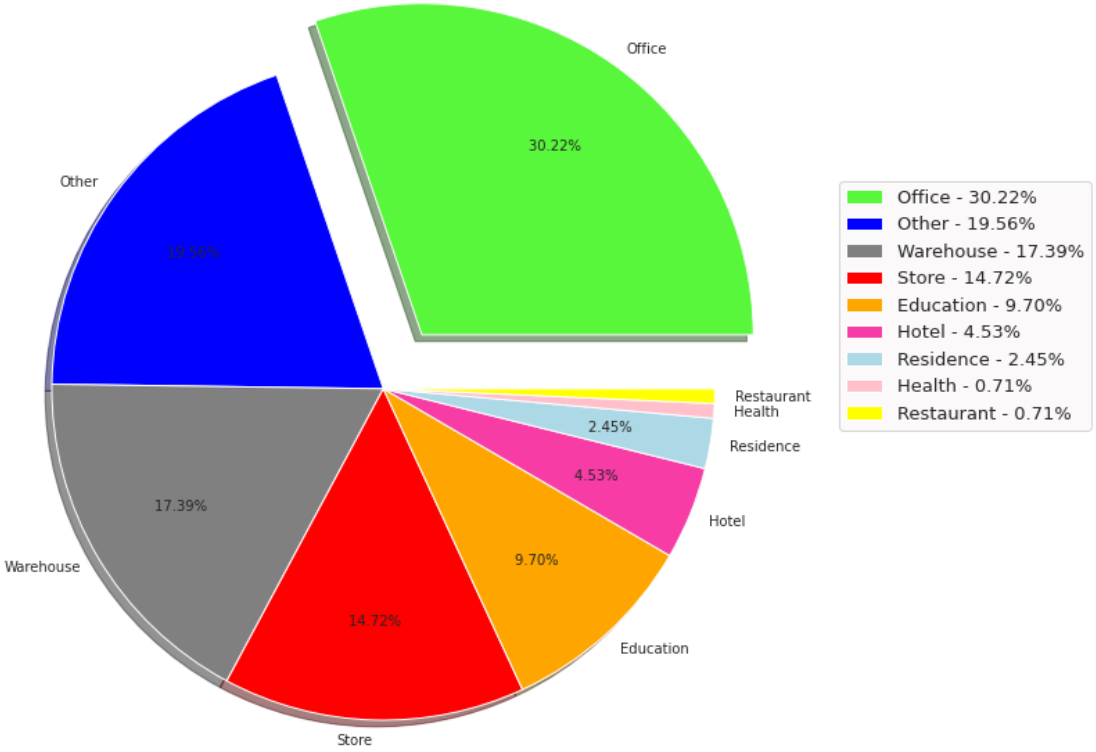
- L’écart d'intensité de consommation d'énergie entre le site et la source :



- **Remarque :** La médiane de l’écart d’intensité de consommation est d’environ de 52% pour la catégorie « Restaurant » et d’environ de 33% pour la catégorie « Bureau ».

Analyse exploratoire des données – Type de bâtiment

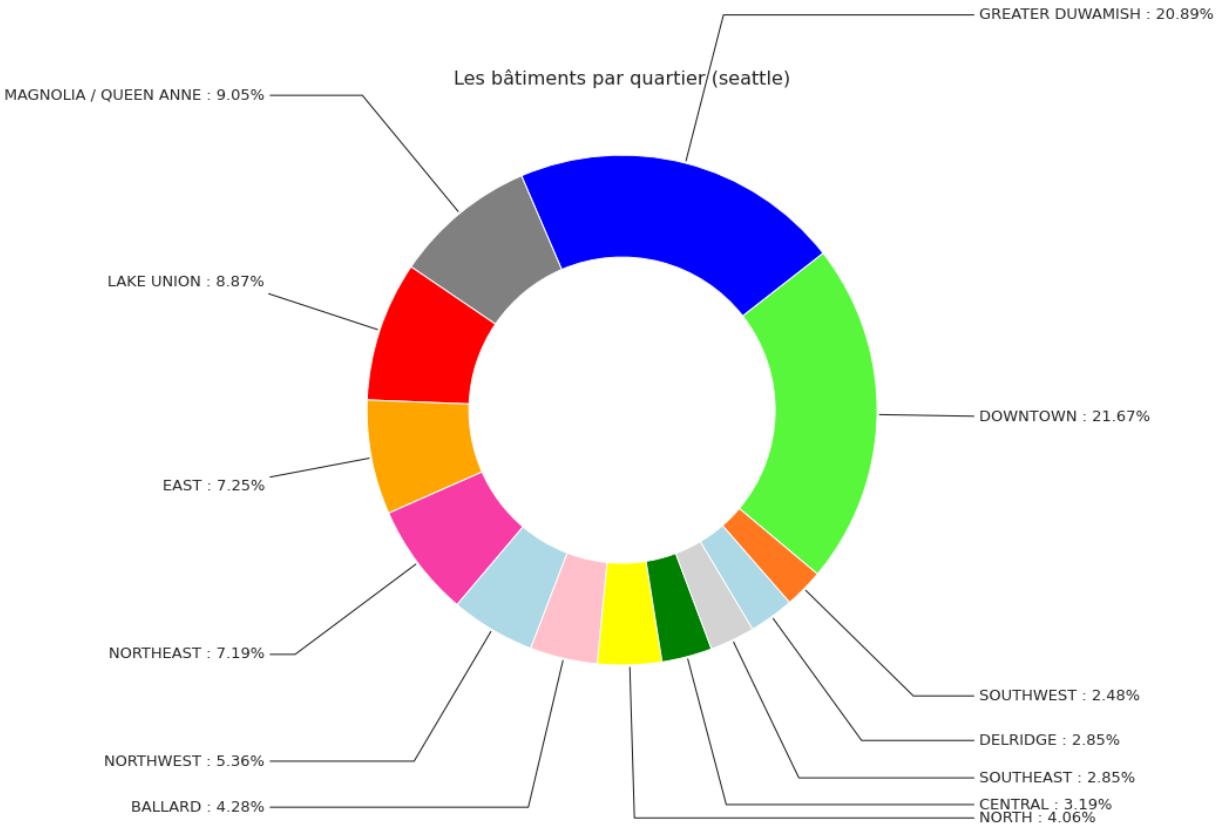
- Graphique des propriétés par type de bâtiment :



- **<u>Remarque :</u>** Les bureaux représentent 30% des bâtiments, les restaurants ne représentent que 0,71%.

Analyse exploratoire des données – Quartier

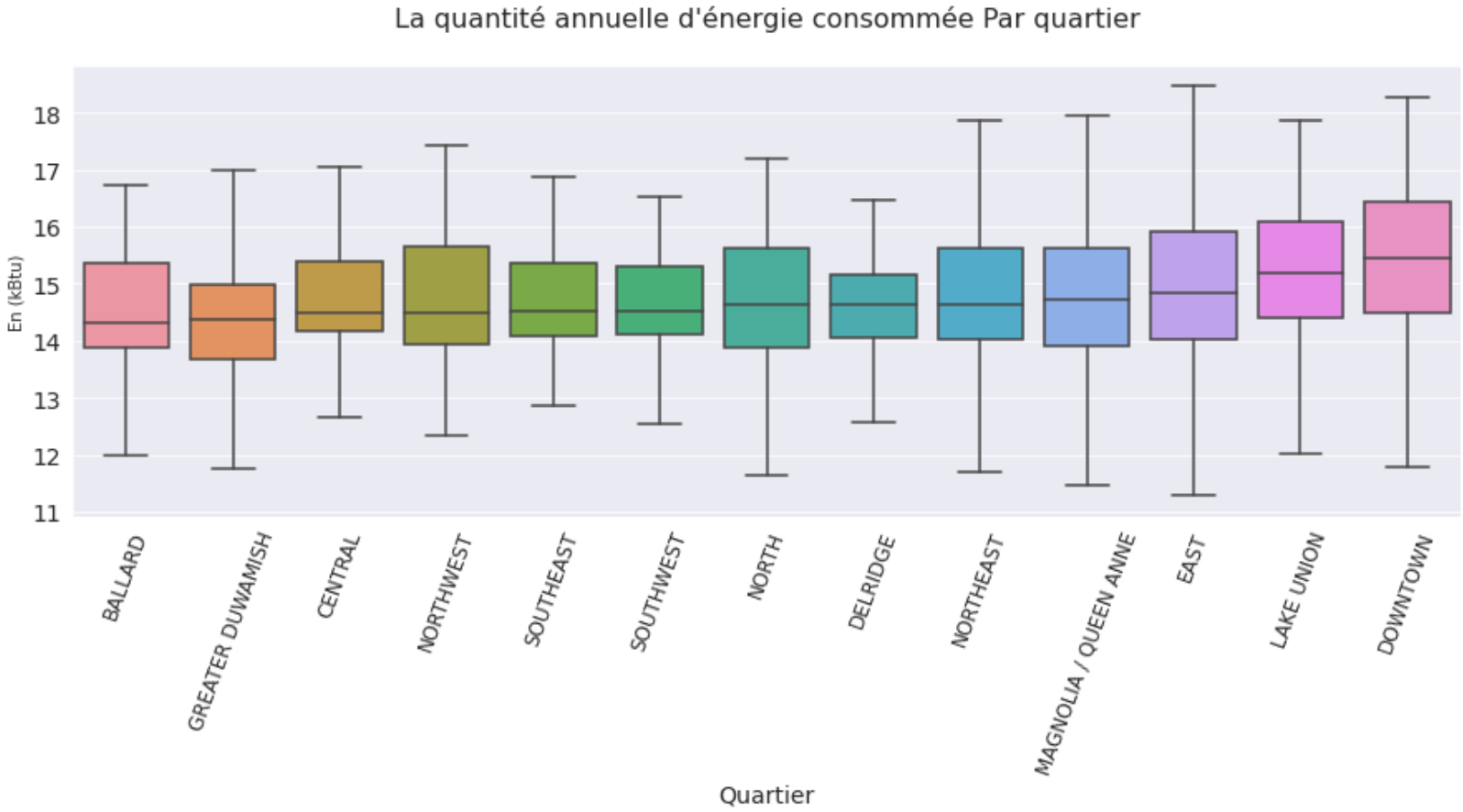
- Graphique des bâtiments par quartier:



- **<u>Remarque :</u>** Le quartier du centre ville compte 21% des propriétés suivi du quartier « Greater Duwamish »

Analyse exploratoire des données – Quartier

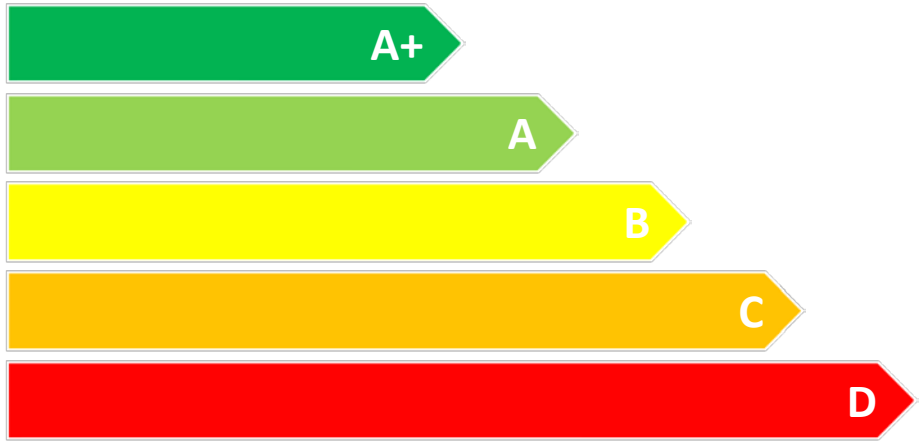
- Graphique de la consommation annuelle des bâtiments par quartier:



- **Remarque :** Le centre ville à la valeur médiane de consommation la plus élevée.

Analyse exploratoire des données – ENERGIE-GRADE

- Pour rendre l' « ENERGYSTAR» visible et facile à comprendre j'ai crée une échelle graphique qui le divise en 5 classes (A+, A, B, C, D), ce système de classification est inspiré de « l'*Étiquette-énergie européenne* ».

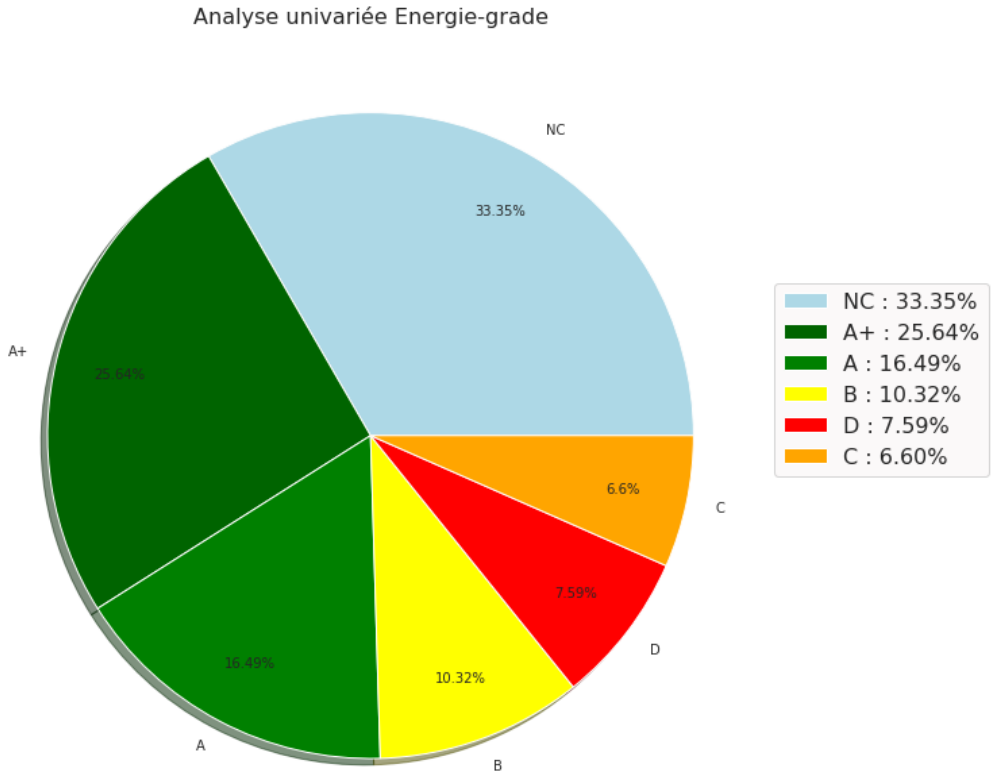


- Ce système de classification dépend du score ENERGY STAR:

ENERGIE-GRADE	A+	A	B	C	D
ENERGYSTAR	81 à 100	61 à 80	41 à 60	21 à 40	1 à 20

Analyse exploratoire des données – ENERGIE-GRADE

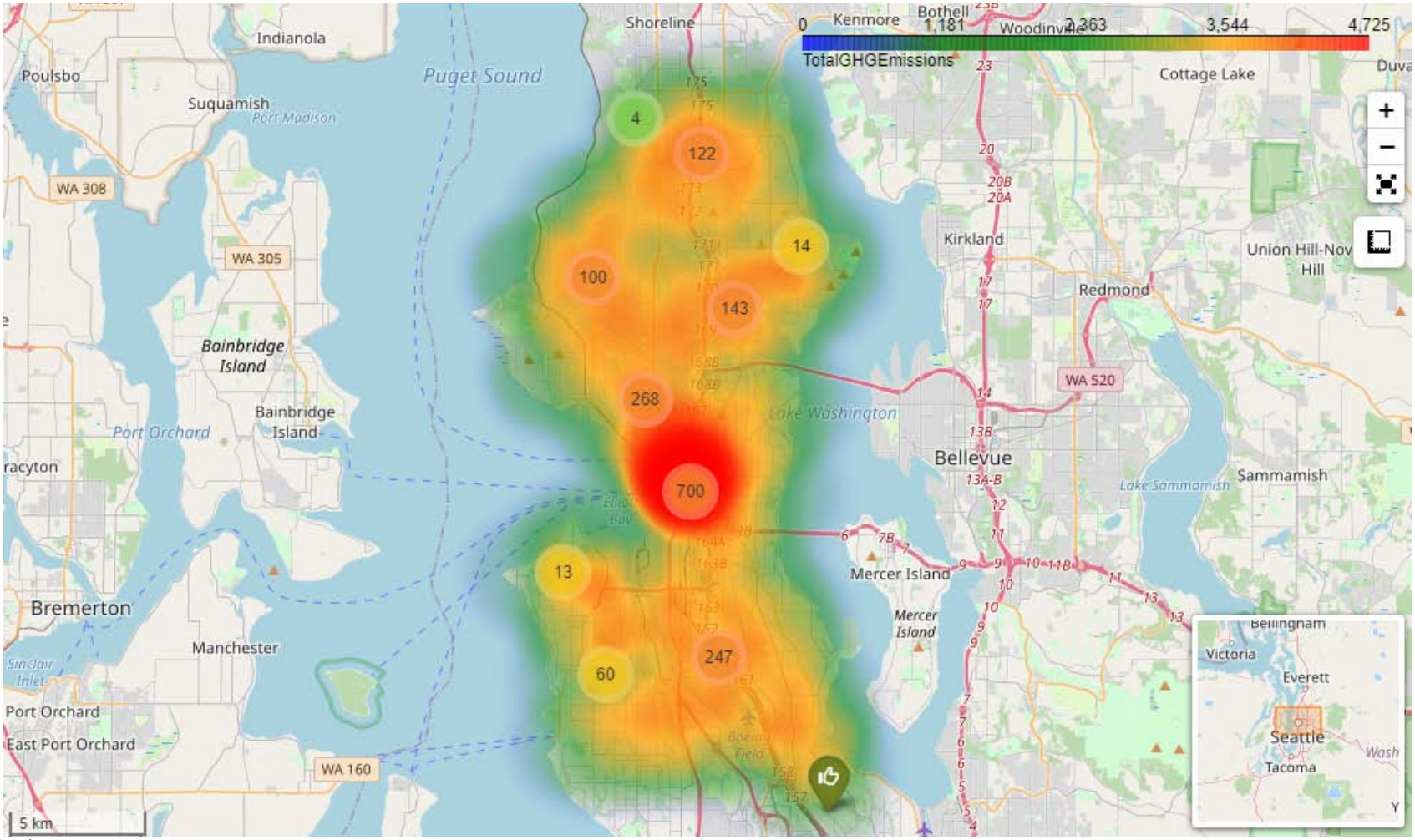
- Graphique de la classification énergétique des bâtiments :



- **Remarque :** 25% des bâtiment ont une classification A+, et seulement 7% ont une classification D

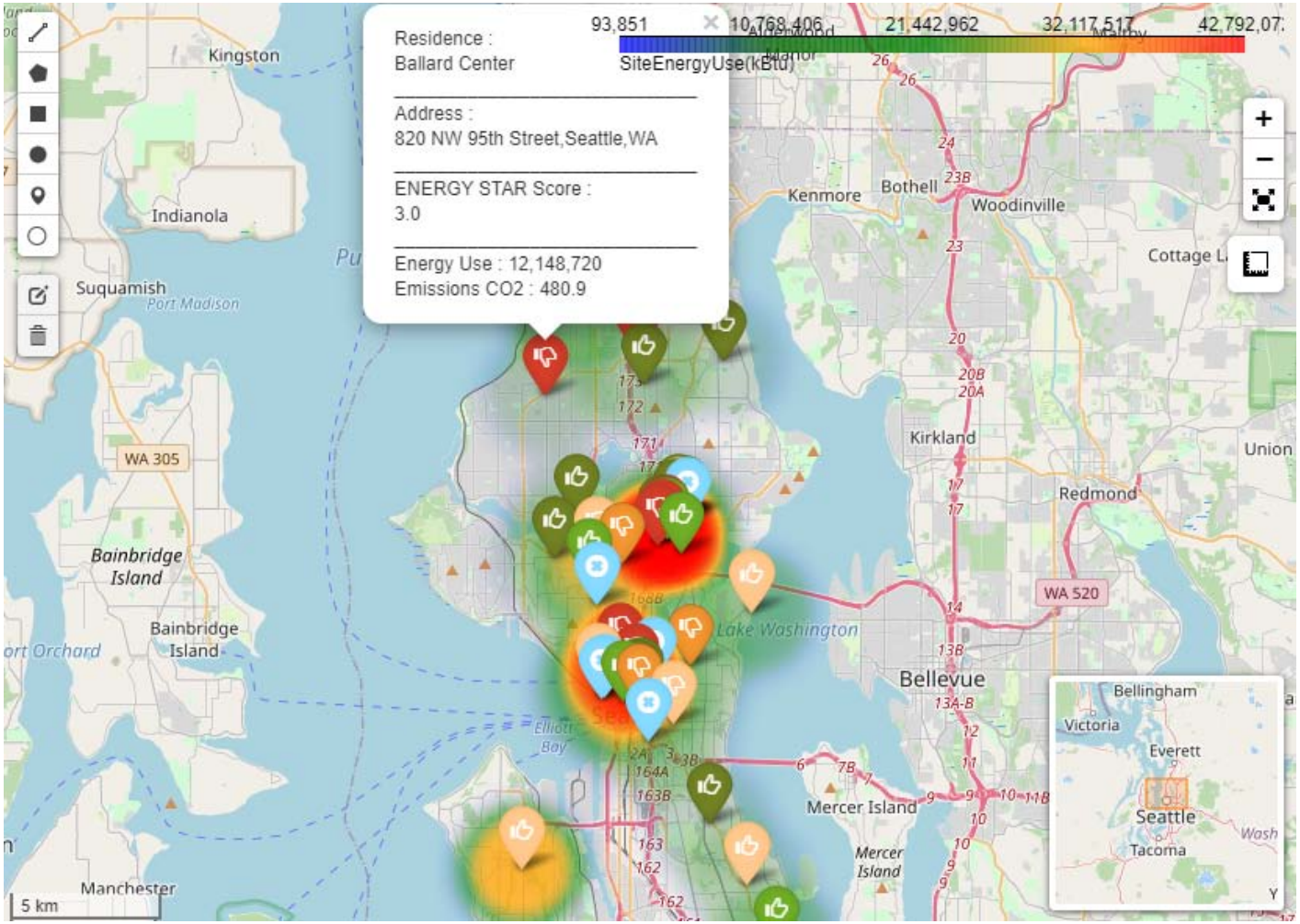
Analyse exploratoire des données – Cartographie

- La cartographie des bâtiments de la ville de ***SEATTLE*** avec la ***Heatmap*** des émissions de CO2 'TotalGHGEmissions'



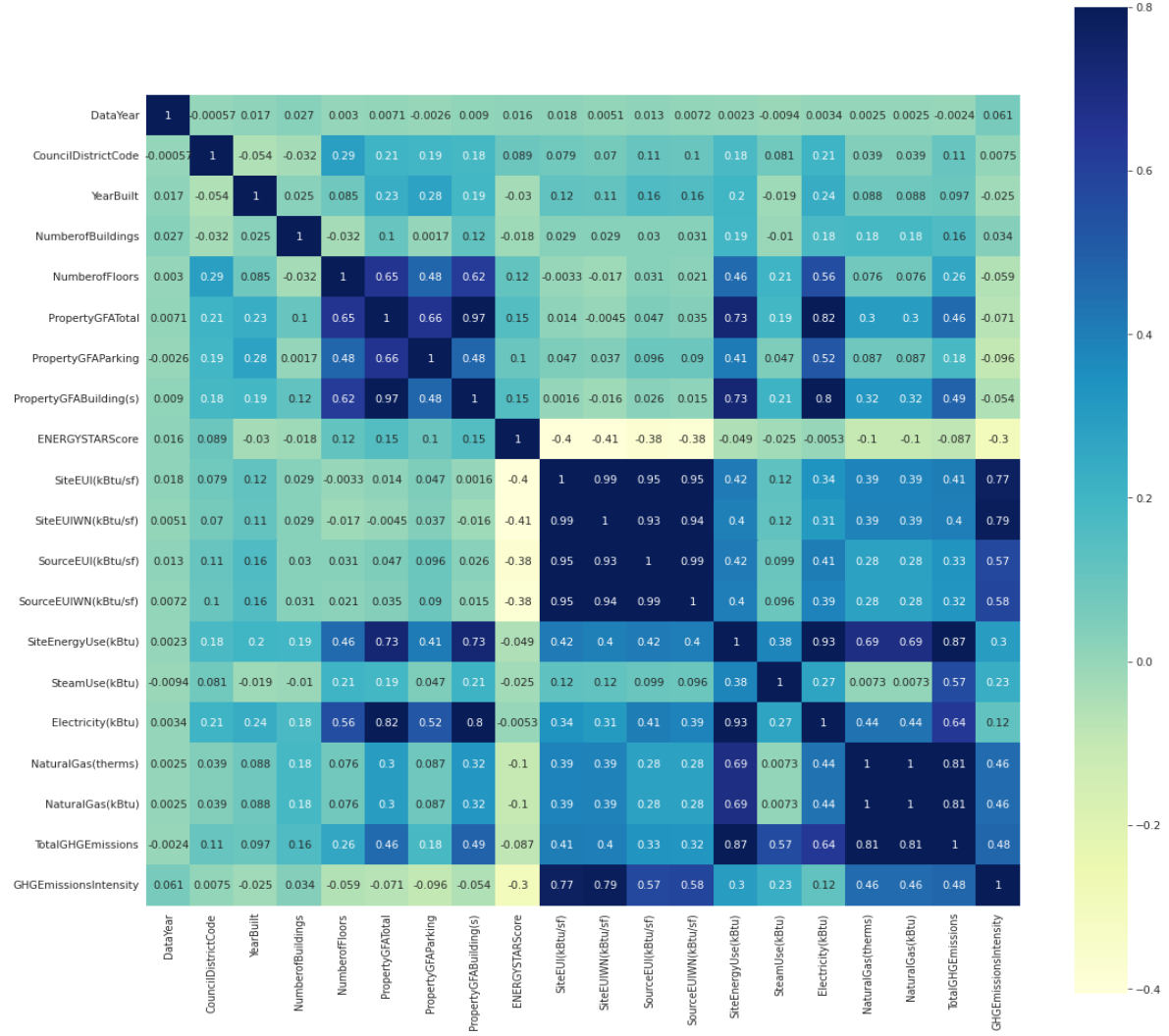
Analyse exploratoire des données – Cartographie

- Un exemple de l'emplacement des bâtiments de la catégorie « Résidence » sur la carte de ***SEATTLE*** avec la ***Heatmap*** de la consommation totale d'énergie :



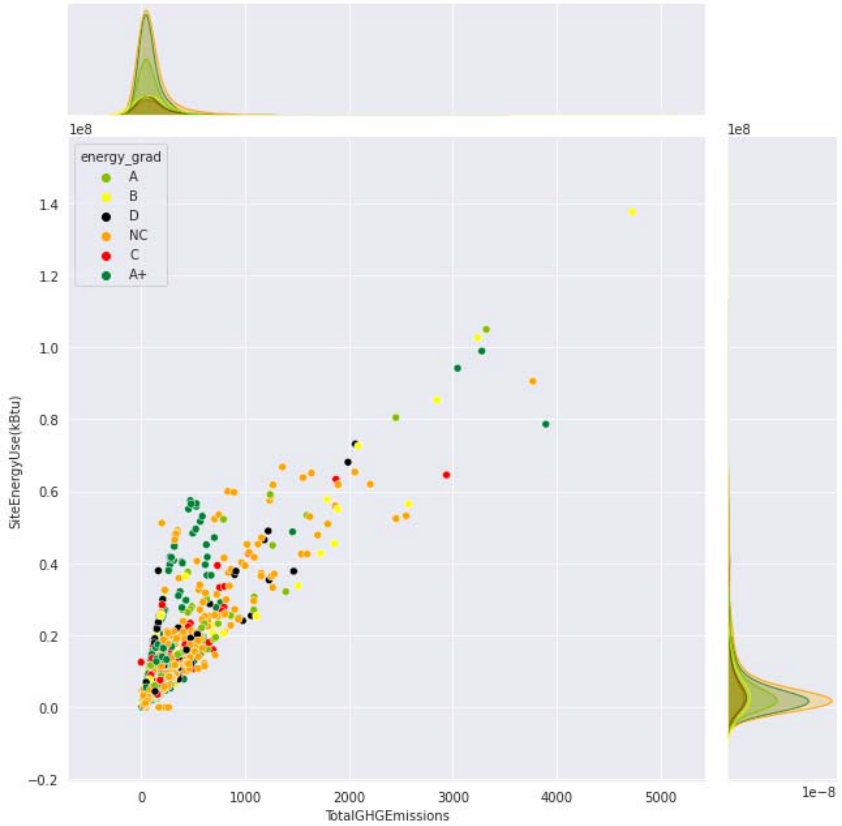
Analyse exploratoire des données – Corrélation

- Ci-dessous la matrice de corrélation 2D entre les variables :



Analyse exploratoire des données – Analyse Bivariée

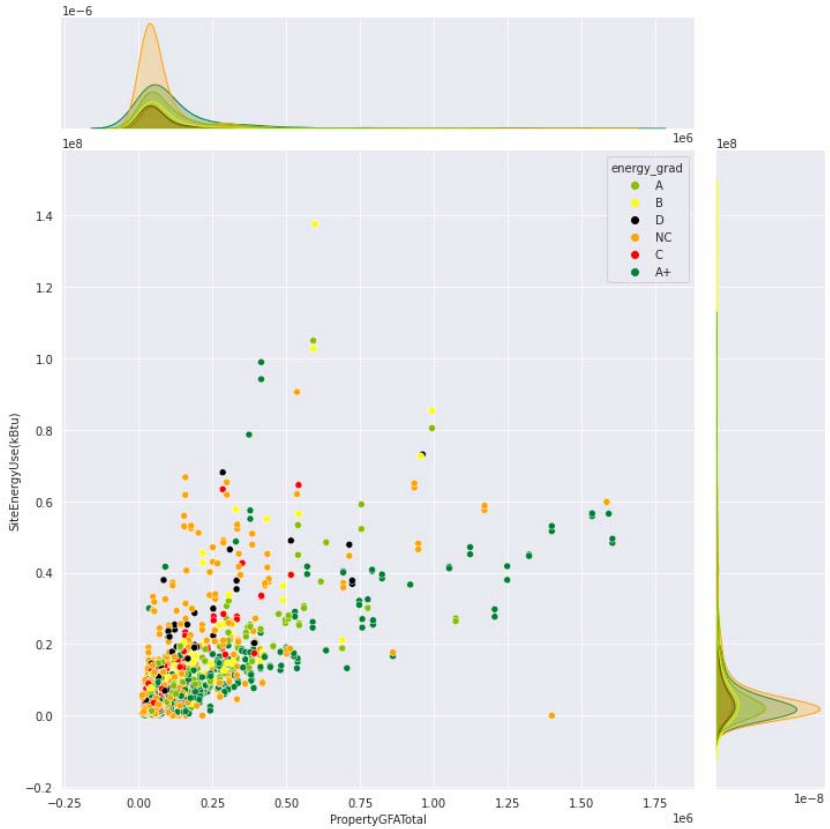
- Graphique pour identifier la relation entre la consommation totale d'énergie et les émission de CO2:



- **Remarque :** Si la consommation d'énergie augmente les émissions de CO2 augmente aussi

Analyse exploratoire des données – Analyse Bivariée

- Graphique pour identifier la relation entre la consommation totale d'énergie et la superficie totale :



- **Remarque :** Si La superficie du bâtiment augmente la consommation d'énergie augmente aussi.

Modèles prédictifs

Modèles prédictifs – Présentation

- **Définition**:

  Une régression a pour objectif d'expliquer une variable Y par le moyen d'une autre variable X. Par exemple, le salaire d'une personne peut être expliqué à travers son niveau universitaire.

- **Les étapes effectuées pour la prédiction des valeurs cibles :**
  - Importation des packages (pandas, numpy, sklearn, xgboost….)
  - Préparation des données
  - Transformation et encodage des données avec le pipeline
  - Exploration des différents modèles prédictifs suivants :
    1. Modèle de régression linéaire
    2. Modèle de régression Ridge
    3. Modèle de régression Lasso
    4. Modèle Random Forest (Forêt d'arbres décisionnels)
    5. Modèle XGBoost (Extreme Gradient Boosting)

Modèles prédictifs – Transformation et encodage

- **Les étapes de transformation et encodage des données avec le pipeline :**

1. Séparer les colonnes numériques des colonnes catégorielles**.**

2. Définir des pipelines numériques et catégorielles, pour effectuer la standardisation et l'encodage des variables :
   a) Standardiser les variables numériques
   b) Encoder les variables catégorielles

3. Assembler les deux pipelines (numérique et catégorielle) dans un transformateur.

Modèles prédictifs – Méthodologie

- **La méthodologie d'exploration pour chaque modèle de régression :**

  - ✓ Définir X (les valeurs explicatives) et y (la valeur cible)
  - ✓ Splitter les données en données d'entrainement (80%), et données de test (20%)
  - ✓ Instancier le modèle
  - ✓ Entraîner le modèle sur les données d'entraînement
  - ✓ Tester le modèle de prédiction sur les données de test
  - ✓ Effectuer une validation croisée avec :
    - **GridSearchCV** pour le réglage des hyperparamètres afin de déterminer les valeurs optimales pour un modèle donné
    - **RandomizedSearchCV** Contrairement à GridSearchCV, toutes les valeurs de paramètres ne sont pas testées avec un temps d'exécution plus faible
  - ✓ Comparer les résultats pour définir le meilleur modèle

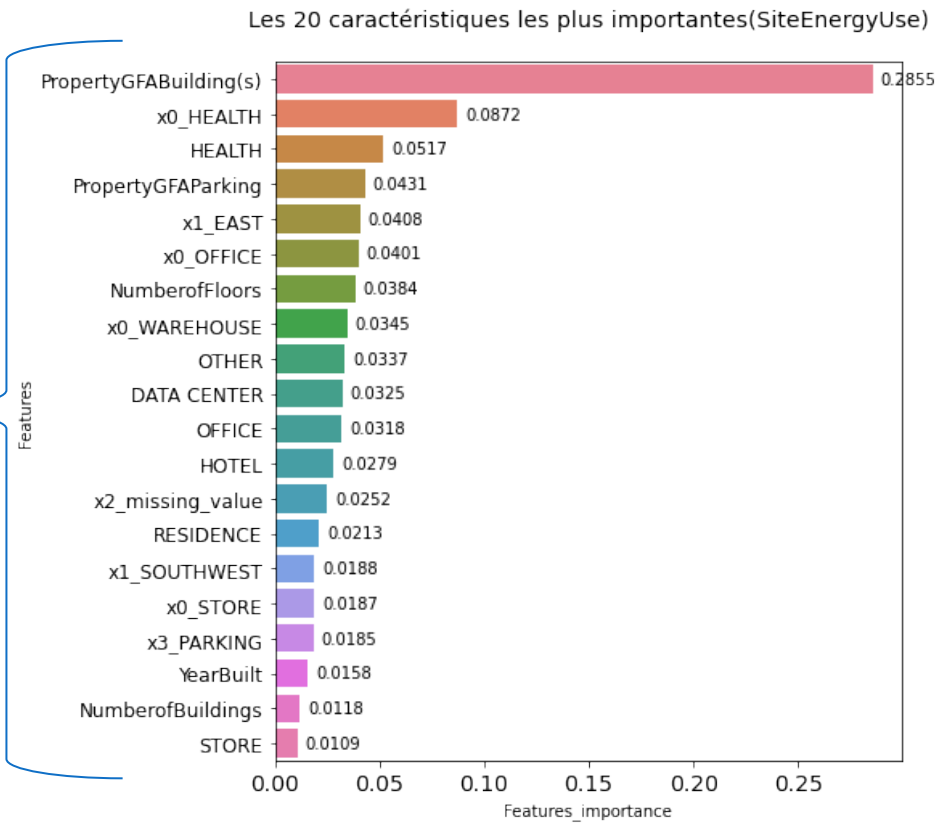
Analyse prédictive de La consommation annuelle d'énergie «SiteEnergyUse»

Analyse prédictive - La consommation totale d'énergie

- Définir X et y :

	Colonnes catégorielles	Colonnes numériques
X	PrimaryPropertyType, Neighborhood, LargestPropertyUseType, SecondLargestPropertyUseType, ThirdLargestPropertyUseType,	YearBuilt, NumberofBuildings,NumberofFloors, PropertyGFAParking, PropertyGFABuilding(s), Count_List, HOTEL, POLICE STATION, OTHER,EDUCATION, HEALTH, OFFICE, COURTHOUSE, AUTOMOBILE DEALERSHIP,WAREHOUSE, STORE, RESIDENCE, MUSEUM, DISTRIBUTION CENTER,PARKING, RESTAURANT, DATA CENTER, CONVENTION CENTER,STRIP MALL, WHOLESALE CLUB/SUPERCENTER,MANUFACTURING/INDUSTRIAL PLANT, LIFESTYLE CENTER, FIRE STATION,PERFORMING ARTS, BANK BRANCH, MOVIE THEATER,PRISON/INCARCERATION,
Y		SiteEnergyUse(kBtu)

Les 20 caractéristiques
les plus importantes (XGBoost) :



Analyse prédictive - La consommation totale d'énergie

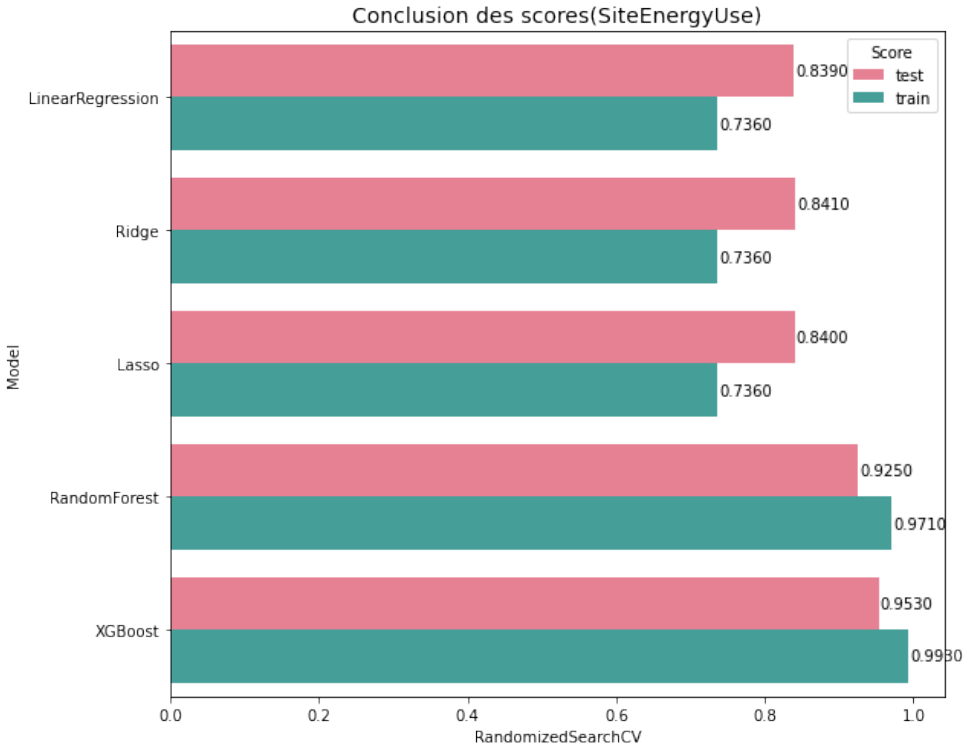
- Ci-dessous les scores « r2 » des différents modèles:

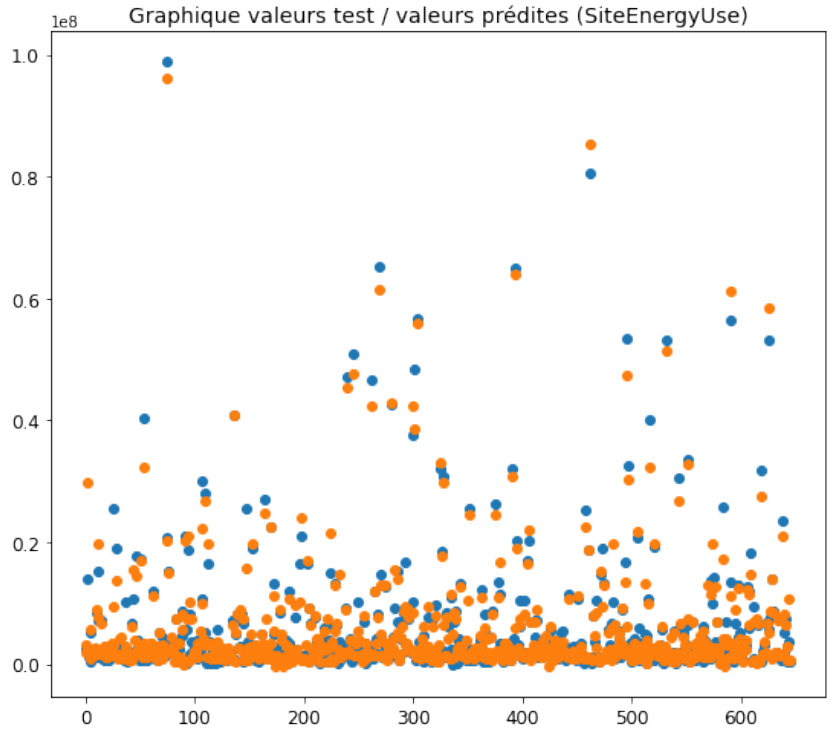
SiteEnergyUse					
Modèle de régression	**RMSE**	**Score r2**	**Modèle sans CV**	**GridSearchCV**	**RandomizedSearchCV**
Régression linéaire	4844313	**Score test**	0.839	0.839	0.839
		Score train	0.736	0.736	0.736
Régression ridge	4824161	**Score test**	0.841	0.843	0.841
		Score train	0.736	0.730	0.736
Régression lasso	4842296	**Score test**	0.840	0.840	0.840
		Score train	0.736	0.736	0.736
Régression Random Forest	2855653	**Score test**	0.925	0.927	0.925
		Score train	0.973	0.972	0.972
XGBoost	2235856	**Score test**	0.879	**0.953**	**0.953**
		Score train	0.883	**0.993**	**0.993**

Analyse prédictive - Conclusion

- **Conclusions :**

  Suite à l'analyse des scores le modèle de régression « ***XGBoost »*** est le modèle le plus performant pour la prédiction de la quantité annuelle d'énergie consommée avec un score « r2 » de 95,3% sur les données de test et 99.3% sur les données d'entrainement, comme le montre le graphique ci-dessous :





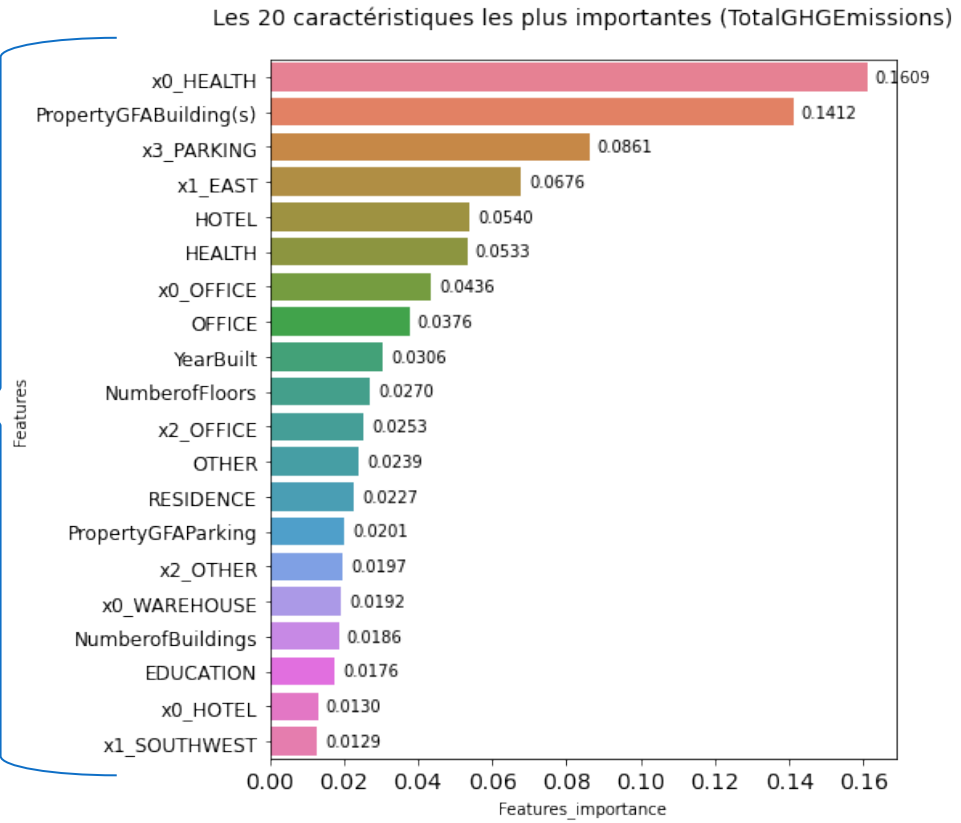
Analyse prédictive des émissions de CO2 «TotalGHGEmissions»

Analyse prédictive - Les émissions de CO2

- Définir X et y :

	Colonnes catégorielles	Colonnes numériques
X	PrimaryPropertyType, Neighborhood, LargestPropertyUseType, SecondLargestPropertyUseType, ThirdLargestPropertyUseType,	YearBuilt, NumberofBuildings,NumberofFloors, PropertyGFAParking, PropertyGFABuilding(s), Count_List, HOTEL, POLICE STATION, OTHER,EDUCATION, HEALTH, OFFICE, COURTHOUSE, AUTOMOBILE DEALERSHIP,WAREHOUSE, STORE, RESIDENCE, MUSEUM, DISTRIBUTION CENTER,PARKING, RESTAURANT, DATA CENTER, CONVENTION CENTER,STRIP MALL, WHOLESALE CLUB/SUPERCENTER,MANUFACTURING/INDUSTRIAL PLANT, LIFESTYLE CENTER, FIRE STATION,PERFORMING ARTS, BANK BRANCH, MOVIE THEATER,PRISON/INCARCERATION,
Y		TotalGHGEmissions

Les 20 caractéristiques
les plus importantes (XGBoost) :



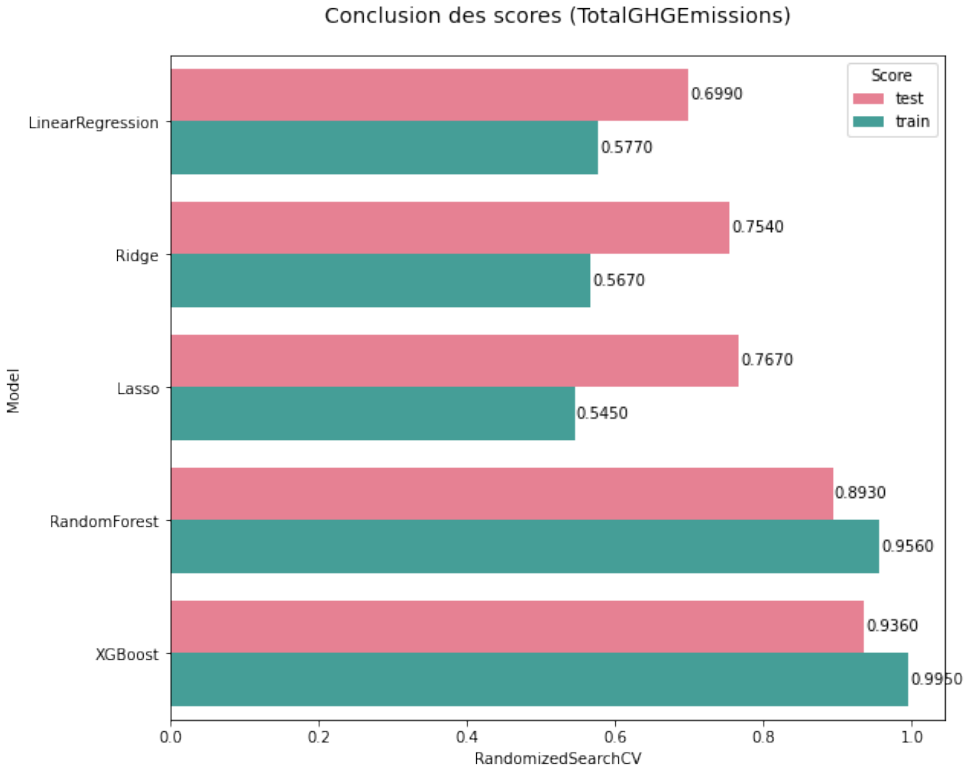
Analyse prédictive - Les émissions de CO2

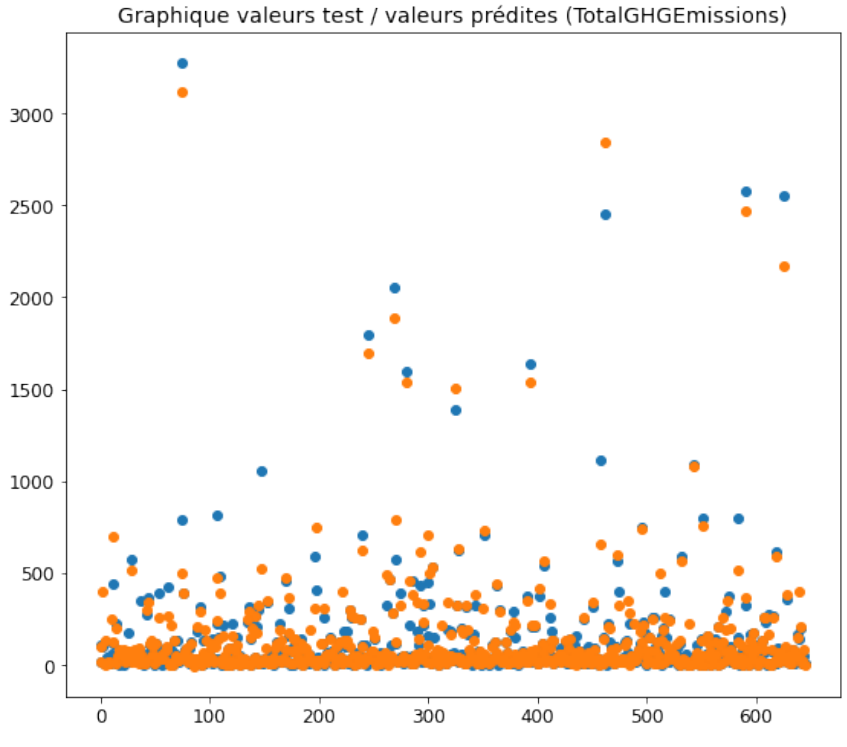
- Ci-dessous les scores « r2 » des différents modèles:

Modèle de régression	RMSE	Score r2	Modèle sans CV	GridSearchCV	RandomizedSearchCV
Régression linéaire	174	Score test	0.699	0.699	0.699
		Score train	0.577	0.577	0.577
Régression ridge	155	Score test	0.754	0.754	0.754
		Score train	0.567	0.567	0.567
Régression lasso	149	Score test	0.767	0.767	0.767
		Score train	0.545	0.545	0.545
Régression Random Forest	97	Score test	0.892	0.891	0.893
		Score train	0.956	0.959	0.956
XGBoost	72	Score test	0.831	0.936	0.936
		Score train	0.867	0.995	0.995

Analyse prédictive - Conclusion

- Ci-dessous le graphique de performance des modèles prédictifs :





- **Remarque :**

  Suite à l'analyse des score, le modèle de régression XGBoost est le modèle le plus performant pour la prédiction des émissions de CO2 avec un score « r2 » de 93,5% sur les données de test et 99.5% sur les données d'entrainement

Analyse prédictive – Modèle XGBoost avec «ENERGY STAR Score»

- Ci-dessous les scores « ***r2*** » du modèle XGBoost avec la variables explicative «***ENERGY STAR Score***» pour évaluer l'impact de cette variable sur le modèle de prédiction de la consommation annuelle d'énergie «***SiteEnergyUse***» et les émissions de CO2 «***TotalGHGEmissions*** » :

	SiteEnergyUse		TotalGHGEmissions	
Modèle de régression	**Score r2**	**RandomizedSearchCV**	**Score r2**	**RandomizedSearchCV**
XGBoost	Score test	0.953	Score test	0.888
	Score train	0.998	Score train	0.998

- **Remarque :**

  On constate que les scores du modèle XGBoost n'ont pas été améliorés malgré l'ajout de cette variable explicative.

  Il n'est pas nécessaire d'intégrer cette variable dans le modèle de prédiction, en sachant que son calcul est très fastidieux.

Annexe – Application

Application - Présentation

- La ville de ***SEATTLE*** veut atteindre son objectif de ville neutre en émissions de CO2 en 2050.

- Mon application (lien) «***SEATTLE-ECO2*** » a pour objectif :

  ✓ Donner une vision énergétique globale de la ville.

  ✓Analyser les données de référence pour identifier les installations qui sont sous-performant.

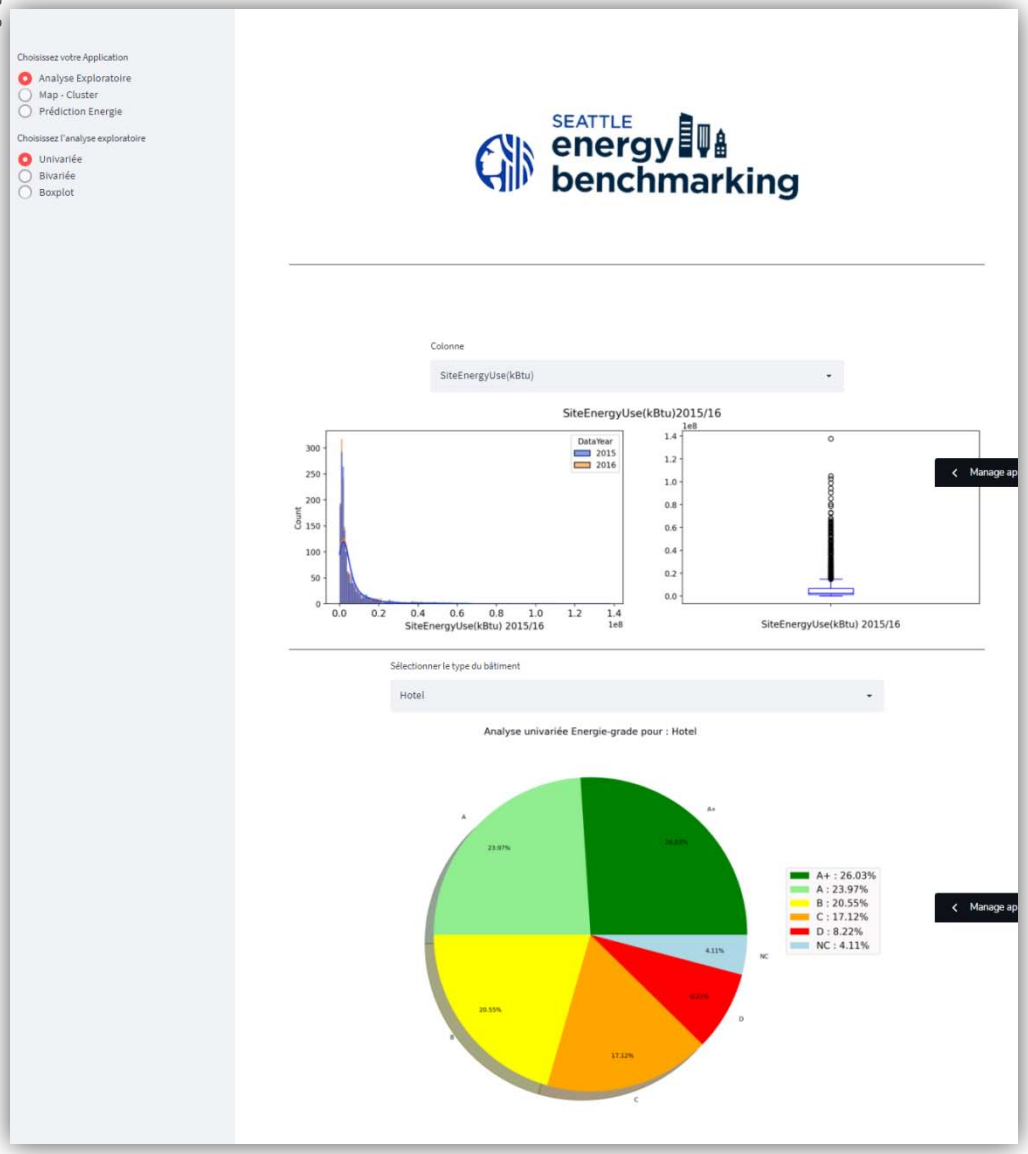
  ✓Aider à créer une approche ciblée pour améliorer le score ***ENERGYSTAR.***

  ✓Permettre de faire une analyse comparatives de la consommation énergétique et les émissions de CO2.

  ✓Évaluer et comparer la performance énergétique des bâtiments.

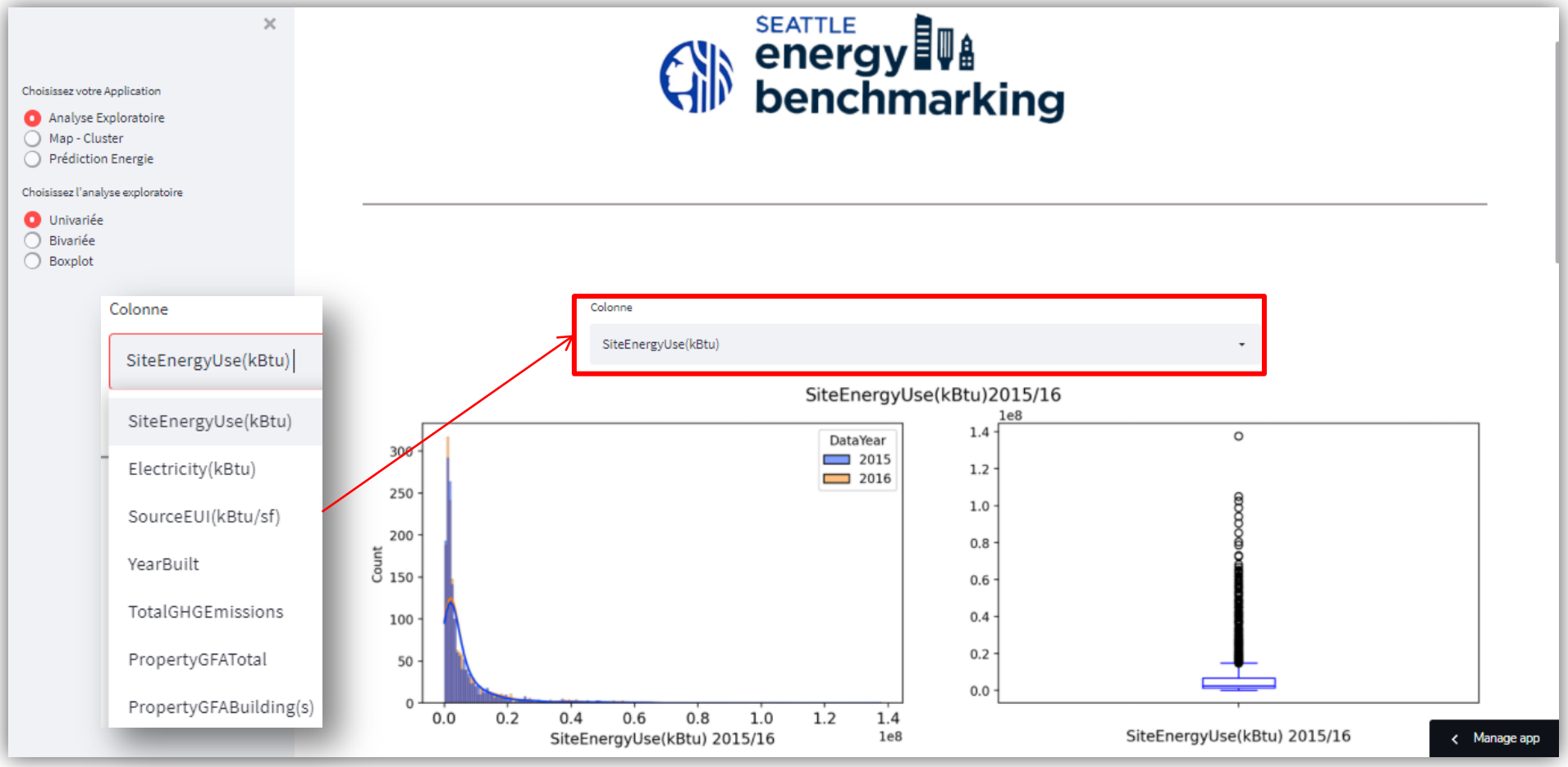
Application - Présentation

- Page d'accueil :



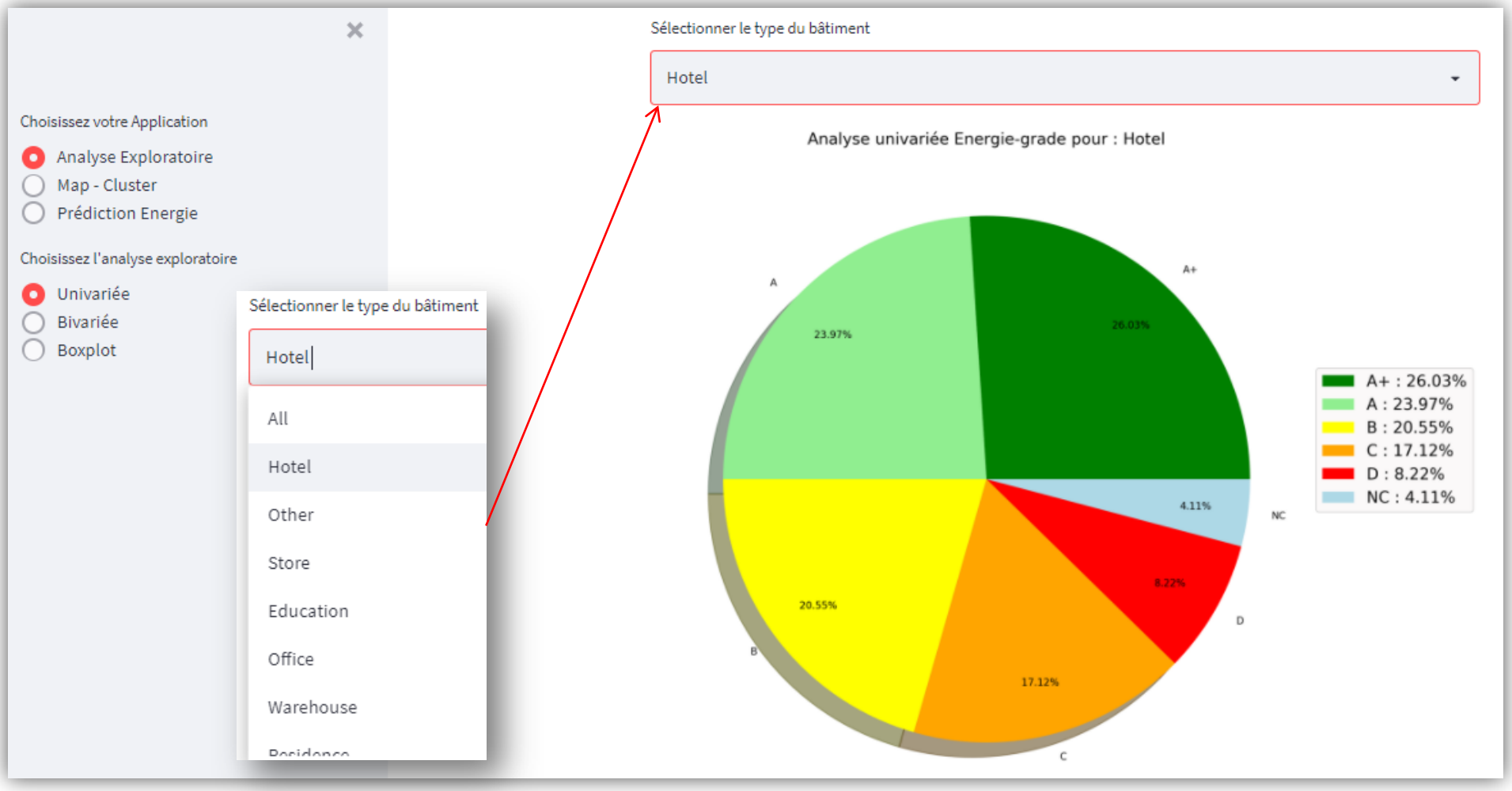
Application - Analyse univariée

- Analyse univariée des variables :



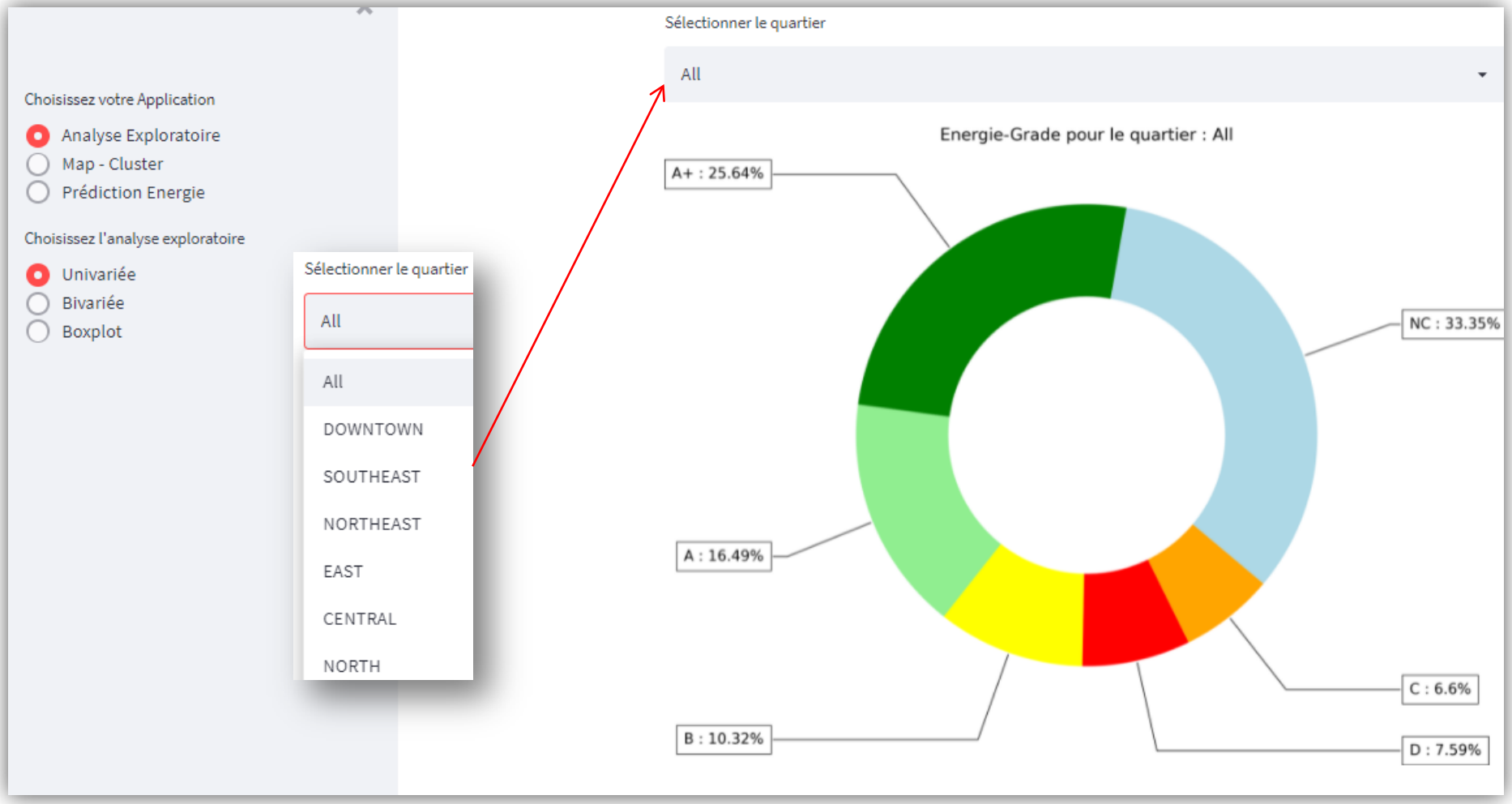
Application - Analyse univariée

- Analyse univariée par type de bâtiment :



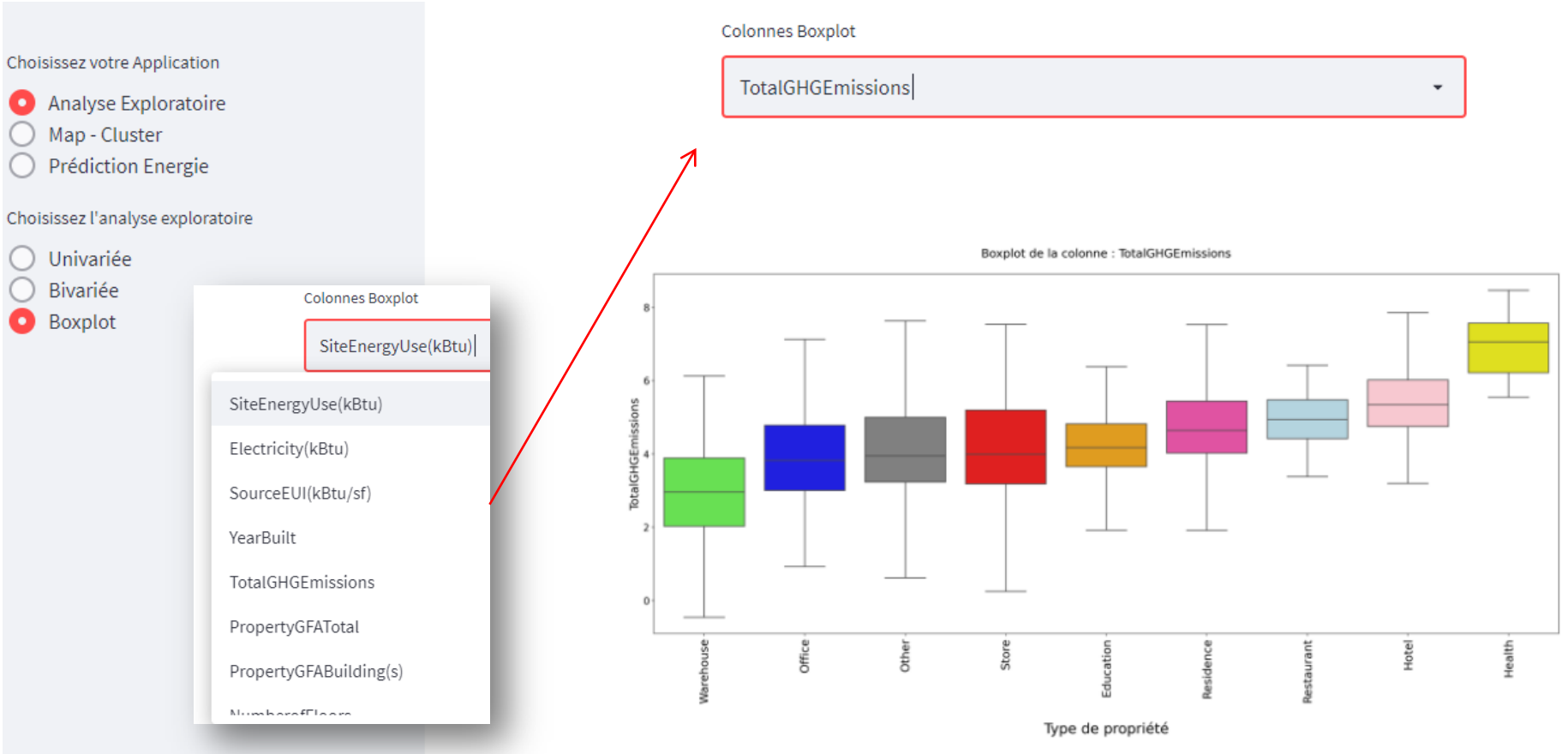
Application - Analyse univariée

- Analyse univariée par quartier :



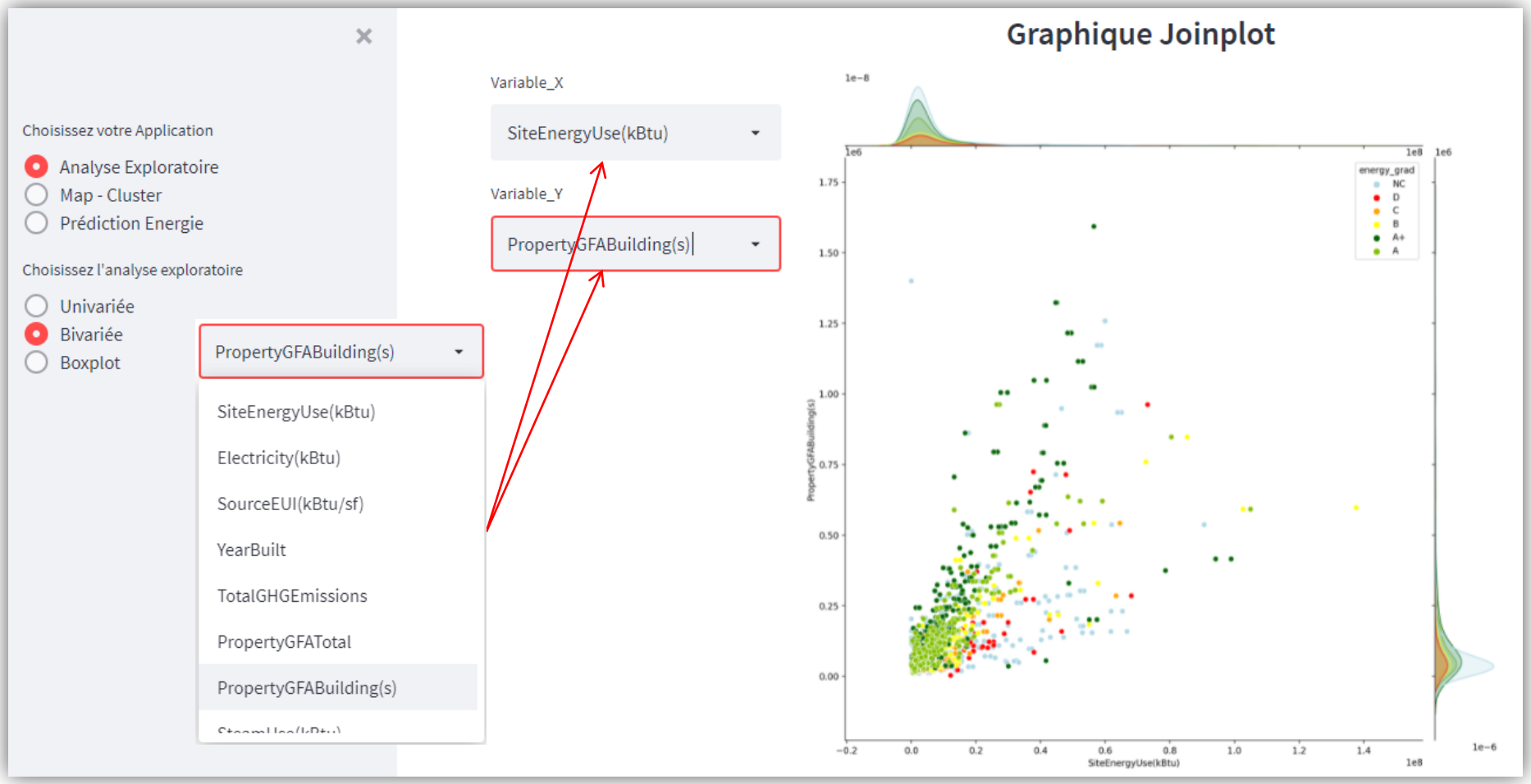
Application - Analyse univariée

- Graphique boxplot par type de bâtient:



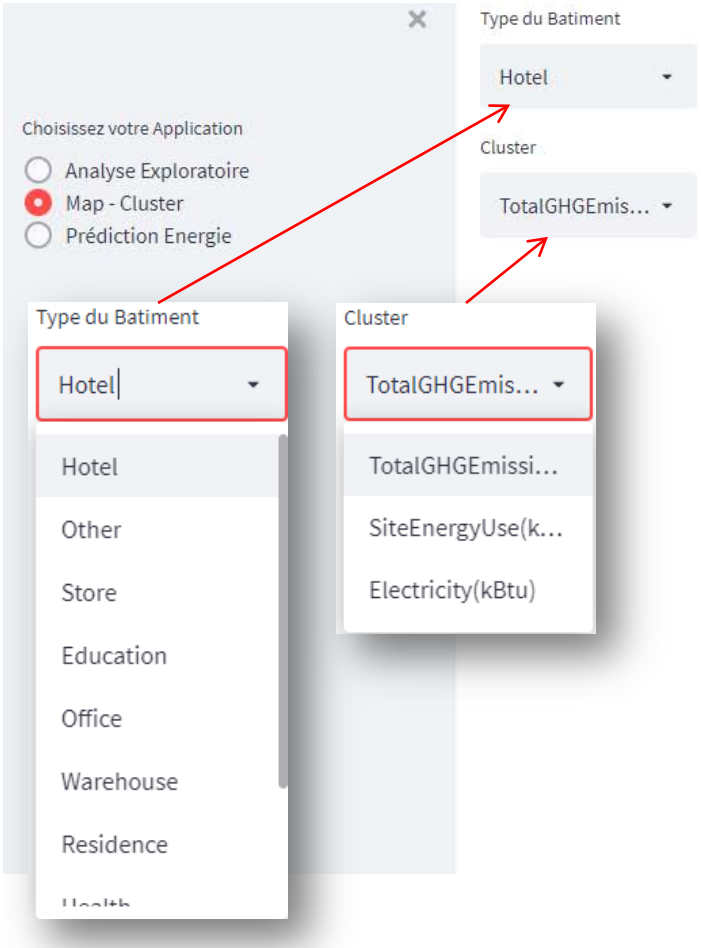
Application - Analyse bivariée

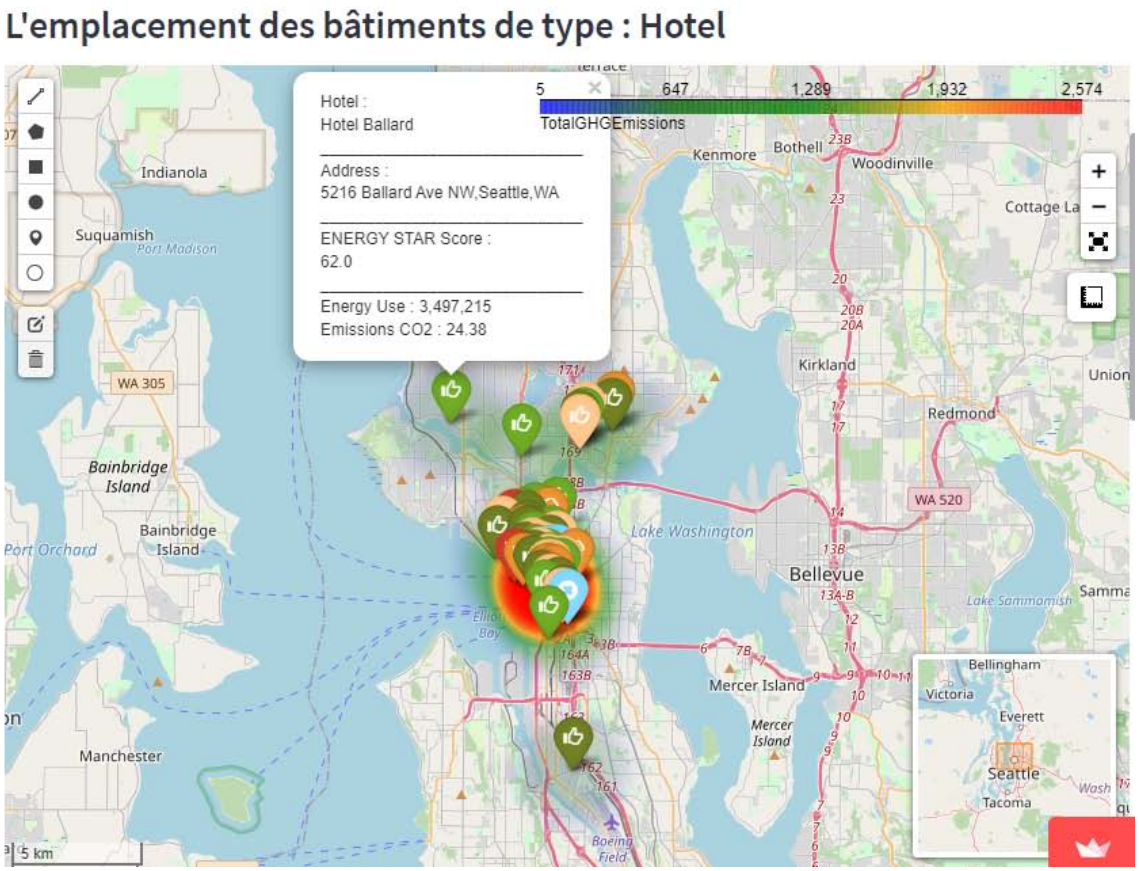
- Graphique pour identifier la relation entre deux variables :



Application – Map / Cluster

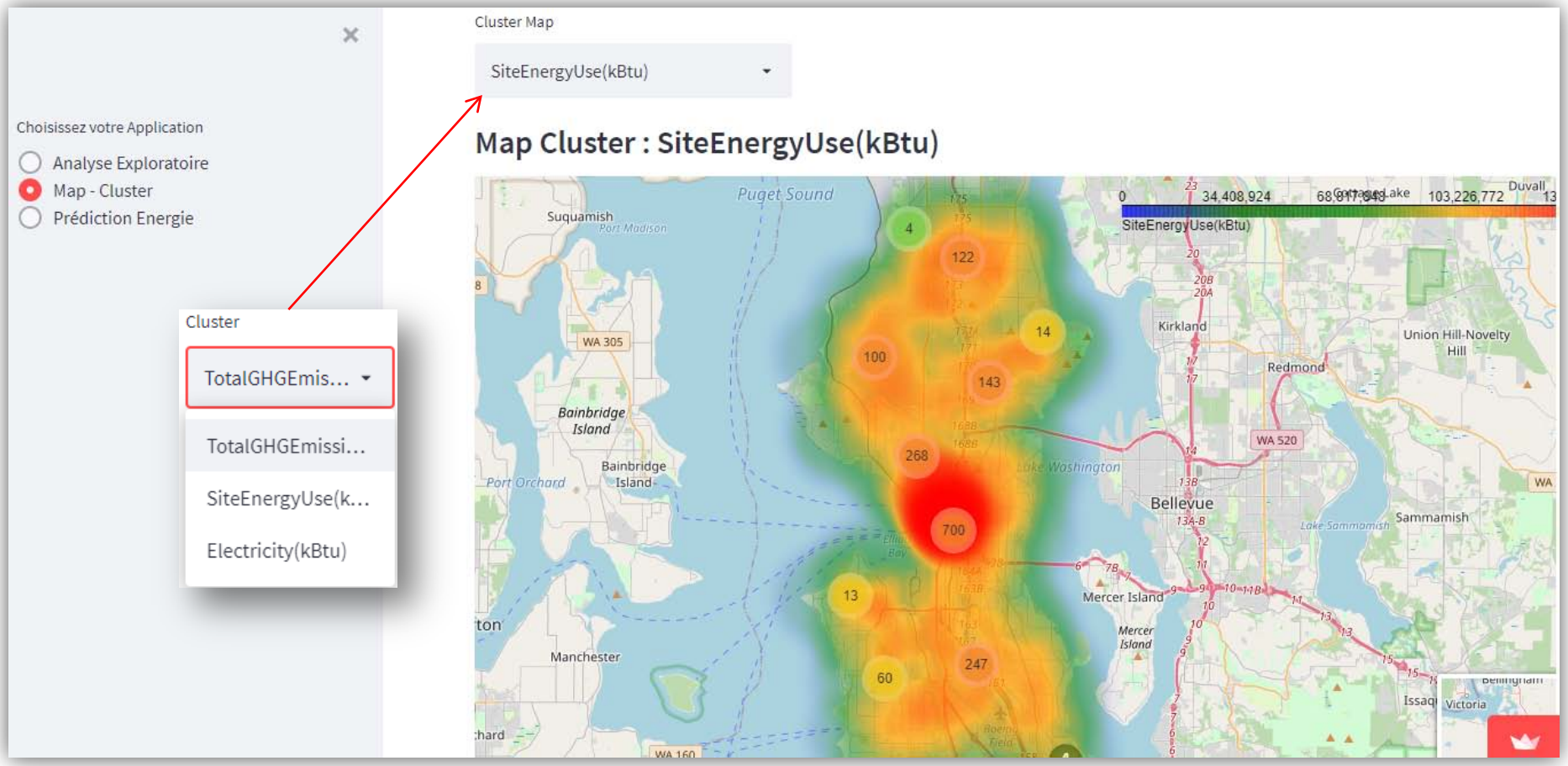
- La cartographie des bâtiments de la ville de ***SEATTLE*** avec la ***Heatmap :***





Application – Map / Cluster

- La cartographie des bâtiments de la ville de ***SEATTLE*** avec la ***Heatmap :***



Application – Prédiction de la consommation annuelle d'énergie

- Prédiction de La quantité annuelle d'énergie consommée «SiteEnergyUse» :