Analyse univariée des données – Émissions de CO_2 (Seattle)

Projet : Prédiction des émissions de gaz à effet de serre
Notebook associé : `02_Analyse_univariee.ipynb`

---

1. Objectif général du notebook

Ce notebook vise à réaliser une **analyse univariée approfondie** du dataset afin de :

- comprendre le comportement individuel des variables,
- détecter les anomalies statistiques,
- préparer les transformations nécessaires pour le Machine Learning.

---

2. Vue globale du dataset

Typologie des variables

Type de variable	Nombre
Numériques (float / int)	Majoritaires
Qualitatives (object / category)	Minoritaires
Booléennes	Très rares

Le dataset est dominé par des variables quantitatives, ce qui justifie une analyse statistique détaillée.

---

3. Analyse de la variable cible – `TotalGHGEmissions`

Statistiques descriptives

Statistique	Valeur
Moyenne	138.76
Médiane	46.01
Minimum	0.40
Maximum	12 307.16
Écart-type	540.27
Skewness	16.77
Kurtosis	329.03

Quartiles

Quartile	Valeur
Q1 (25 %)	19.54
Q2 (50 %)	46.01
Q3 (75 %)	120.88
IQR	101.35

Interprétation

La distribution est **fortement asymétrique à droite**, dominée par quelques bâtiments extrêmement émetteurs.
Une **transformation logarithmique** est indispensable avant modélisation.

---

4. Variables de consommation énergétique

Variables analysées

- Électricité, gaz naturel, vapeur
- Consommation totale du site
- Intensités énergétiques (EUI)

Synthèse des comportements observés

Caractéristique	Observation
Forme des distributions	Très asymétriques
Dispersion	Très élevée
Valeurs extrêmes	Nombreuses
Variables WN vs non-WN	Très fortement corrélées

Les variables normalisées météo (WN) apportent peu d'information supplémentaire.

---

5. Caractéristiques physiques des bâtiments

Variables étudiées

- `PropertyGFATotal`
- `PropertyGFABuilding(s)`
- `PropertyGFAParking`
- `NumberofFloors`
- `NumberofBuildings`

Observations clés

Variable	Pattern observé
Surfaces	Skewed à droite
Taille typique	Petite à moyenne
Très grands bâtiments	Peu nombreux mais dominants
Nombre d'étages	Majorité entre 1 et 5

La surface et la hauteur expliquent en partie les écarts énergétiques.

---

6. Dimension temporelle – `YearBuilt`

Élément	Observation
Étendue temporelle	Large
Bâtiments anciens	Proportion élevée
Homogénéité	Faible

L'âge influence la performance énergétique, mais n'explique pas seul les émissions.

---

7. Performance énergétique – `ENERGYSTARScore`

Statistiques générales

Élément	Observation
Domaine	[0 ; 100]
Dispersion	Modérée
Valeurs manquantes	Nombreuses
Normalité	Non vérifiée

Le score est informatif mais nécessite une **gestion spécifique des NaN**.

---

8. Analyse des variables qualitatives

Variables étudiées

- `BuildingType`
- `PrimaryPropertyType`
- `Neighborhood`
- `LargestPropertyUseType`
- `ListOfAllPropertyUseTypes`

Patterns observés

Variable	Pattern principal
BuildingType	Émissions très variables selon le type
PrimaryPropertyType	Domination du tertiaire
Neighborhood	Forte concentration géographique
PropertyUseTypes	Bâtiments souvent multi-usages

Ces variables sont **fortement structurantes** pour la modélisation.

---

9. Recommandations issues de l'analyse univariée

Problème détecté	Action recommandée
Asymétrie forte	Log-transform
Valeurs extrêmes	Transformation ou robust scaling
Variables redondantes	Sélection
Modalités nombreuses	Regroupement
Valeurs manquantes	Imputation ou suppression

---

10. Rôle du notebook dans le projet

Ce notebook constitue :

- une **base méthodologique**,
- un **outil de diagnostic des données**,
- un **support commun pour l'équipe**.

Il doit être consulté **avant toute analyse bivariée ou modélisation**.

---

11. Conclusion générale

L'analyse univariée a permis :

- une compréhension fine des distributions,
- l'identification de patterns structurels,
- une préparation rigoureuse de la phase de Machine Learning.

Elle constitue une **fondation essentielle** pour la suite du projet.