Analyse univariée – Caractérisation individuelle des variables

Projet : Prédiction des émissions de CO_2 (Seattle)

---

1. Introduction

Ce document présente une synthèse complète du travail réalisé dans le notebook d'analyse univariée. L'objectif est de **caractériser individuellement chaque variable du dataset**, afin de comprendre son comportement statistique, identifier d'éventuelles anomalies et préparer efficacement la phase de modélisation en Machine Learning.

L'analyse univariée constitue une étape fondamentale, car elle permet d'éviter des interprétations erronées lors de l'étude des relations entre variables et de guider les choix méthodologiques ultérieurs.

---

2. Objectifs de l'analyse univariée

Le notebook d'analyse univariée a poursuivi les objectifs suivants :

- Explorer chaque variable indépendamment des autres
- Analyser la forme des distributions (symétrie, asymétrie)
- Identifier la dispersion et la présence de valeurs aberrantes
- Évaluer la qualité des données (valeurs manquantes, variabilité)
- Préparer les transformations statistiques nécessaires
- Effectuer une sélection préliminaire des variables pertinentes

---

3. Vue globale du dataset

3.1 Typologie des variables

Le dataset brut est composé majoritairement de **variables numériques**, complétées par des variables qualitatives décrivant les typologies de bâtiments, leurs usages et leur localisation.

Cette structure est favorable à la modélisation, tout en nécessitant un travail de préparation spécifique pour les variables catégorielles (nettoyage, regroupement, encodage).

---

4. Analyse de la variable cible : `TotalGHGEmissions`

4.1 Caractéristiques statistiques

L'analyse de la variable cible met en évidence :

- Une **moyenne nettement supérieure à la médiane**
- Une **dispersion très élevée**

- Une **asymétrie extrême à droite**
- Une **présence marquée d'outliers**

Ces caractéristiques indiquent que la majorité des bâtiments sont faiblement émetteurs, tandis qu'un nombre très limité concentre des émissions extrêmement élevées.

4.2 Interprétation

La médiane apparaît comme un indicateur plus représentatif que la moyenne pour décrire le niveau typique des émissions.
La distribution est clairement non normale et dominée par des valeurs extrêmes, ce qui constitue une information essentielle pour la suite de l'analyse.

---

5. Analyse des variables de consommation énergétique

Les variables liées à la consommation énergétique (électricité, gaz, vapeur, énergie totale, intensités énergétiques) présentent des comportements similaires :

- Distributions fortement asymétriques à droite
- Forte dispersion entre observations
- Présence fréquente de valeurs extrêmes
- Corrélation très élevée entre les versions normalisées météo (WN) et non normalisées

Ces résultats montrent que les variables WN sont largement redondantes avec leurs équivalents non normalisés, ce qui permet d'envisager une simplification du jeu de variables.

---

6. Analyse des caractéristiques physiques des bâtiments

Les variables décrivant les surfaces et la structure des bâtiments (surfaces totales, surfaces de parking, nombre d'étages, nombre de bâtiments) mettent en évidence :

- Une **forte asymétrie à droite** des surfaces
- La présence de quelques bâtiments de très grande taille
- Une majorité de bâtiments de petite à moyenne dimension
- Un nombre d'étages généralement compris entre 1 et 5

Ces variables jouent un rôle central dans l'explication des niveaux de consommation énergétique et d'émissions.

---

7. Dimension temporelle : `YearBuilt`

L'analyse de l'année de construction révèle :

- Une répartition hétérogène des bâtiments dans le temps
- Une forte proportion de bâtiments anciens
- Une influence indirecte de l'âge sur la performance énergétique

L'âge des bâtiments ne suffit pas à expliquer seul les émissions, mais il constitue un facteur explicatif complémentaire pertinent.

---

8. Performance énergétique : `ENERGYSTARScore`

La variable `ENERGYSTARScore` se caractérise par :

- Une échelle bornée entre 0 et 100
- Une proportion importante de valeurs manquantes (bâtiments non certifiés)
- Des profils énergétiques plus performants pour les bâtiments certifiés

Cette variable fournit une information synthétique sur l'efficacité énergétique, mais son utilisation doit tenir compte des valeurs manquantes.

---

9. Analyse des variables qualitatives

Les variables qualitatives analysées incluent notamment :

- Le type de bâtiment (`BuildingType`)
- L'usage principal (`PrimaryPropertyType`)
- Le quartier (`Neighborhood`)
- Le type d'usage dominant (`LargestPropertyUseType`)

9.1 Patterns observés

L'analyse univariée met en évidence des **patterns clairs** :

- Les émissions de GES varient significativement selon le type et l'usage des bâtiments
- Certaines catégories présentent des niveaux d'émissions typiques plus élevés
- Une forte hétérogénéité existe au sein de chaque catégorie
- Les distributions restent asymétriques, avec quelques bâtiments très fortement émetteurs

Ces résultats confirment que les variables qualitatives jouent un rôle structurant dans l'explication des émissions.

---

10. Justification de l'approche univariée

L'analyse univariée est indispensable car elle permet :

- De comprendre le comportement individuel des variables
- De détecter des anomalies invisibles en analyse bivariée
- De choisir les transformations statistiques adaptées
- D'éliminer précocement certaines variables non informatives
- De faciliter la communication des résultats auprès de publics non techniques

---

11. Recommandations de transformation

À l'issue de l'analyse univariée, les recommandations suivantes ont été formulées :

Situation observée	**Recommandation**
Asymétrie forte (skewness > 2)	Transformation logarithmique
Asymétrie modérée (skewness 1–2)	Transformation racine carrée
Variance quasi nulle	Suppression de la variable
Plus de 70 % de valeurs manquantes	Suppression ou traitement spécifique
Variables catégorielles	Encodage adapté (One-Hot, Label, Target)

12. Conclusion

Ce travail d'analyse univariée a permis de comprendre en profondeur la structure du dataset, de mettre en évidence des patterns statistiques et métier dans les émissions de CO_2, et de préparer de manière rigoureuse la phase de modélisation.
Il constitue une **fondation méthodologique essentielle** pour la construction d'un modèle de prédiction fiable, robuste et interprétable.