Note de Synthèse : Initialisation et Ingestion (Phase 0)

Projet : Seattle Energy Benchmarking

Date : 31 Décembre 2025

Objet : Validation de l'infrastructure technique et diagnostic structurel initial

---

1. État d'avancement et objectifs

La première phase du projet, dédiée à l'initialisation de l'environnement et à l'ingestion des données, est désormais **clôturée avec succès**. Tous les objectifs définis en amont ont été atteints :

- L'infrastructure de configuration via **Hydra** est opérationnelle.
- L'arborescence complète du projet a été déployée automatiquement.
- Le jeu de données source a été récupéré, chargé et audité de manière immuable.
- Le diagnostic a permis de valider l'intégrité du dataset pour les phases suivantes.

2. Conformité aux bonnes pratiques

Le développement du Notebook 00 a été guidé par une exigence de reproductibilité et de clarté logicielle :

- **Modularité :** Utilisation de modules internes (`src.utils`, `src.data`) pour séparer la logique métier de la présentation.
- **Traçabilité :** Mise en place d'un système de logging (`setup_eda_logger`) pour enregistrer chaque étape critique de l'ingestion.
- **Gestion de Configuration :** Externalisation des chemins et paramètres dans des fichiers YAML, facilitant la portabilité du projet.
- **Documentation :** Intégration d'un dictionnaire de données automatisé et traduit, assurant une compréhension métier précise des 46 variables ainsi qu'une classification de ces variables.

3. Diagnostic des Données

L'inspection initiale du dataset (3 376 lignes, 46 colonnes) révèle les points clés suivants :

qualité et intégrité

- **données manquantes :** Une forte concentration de valeurs nulles est observée sur les variables liées aux usages secondaires/tertiaires et aux certifications ENERGY STAR. Ces colonnes devront faire l'objet d'un arbitrage lors de la phase de nettoyage.
- **Incohérences relevées :** Présence résiduelle de valeurs négatives pour les émissions de gaz à effet de serre et de valeurs nulles sur des postes de consommation, nécessitant une distinction entre "absence d'usage" et "donnée manquante".

Analyse des variables cibles

Le projet identifie deux cibles potentielles pour la modélisation :

1. **TotalGHGEmissions** (Volume absolu) : Présente une forte asymétrie à droite et une grande sensibilité aux valeurs extrêmes.
2. **GHGEmissionsIntensity** (Efficacité relative) : Plus stable et pertinente pour comparer des bâtiments de tailles hétérogènes.

La corrélation modérée entre ces deux variables (0,47) suggère qu'elles capturent des réalités physiques distinctes. La stratégie retenue privilégiera l'utilisation de modèles séparés plutôt qu'une approche multi-output, afin de garantir une meilleure interprétabilité des résultats.

4. Identification des redondances

Plusieurs groupes de variables redondantes ont été isolés pour les phases suivantes :

- Doublons d'unités (kWh vs kBtu, Therms vs kBtu).
- Redondances structurelles (Surfaces totales vs somme des surfaces partielles).

5. Conclusion et perspectives

La Phase 0 confirme la viabilité du dataset et la robustesse de l'environnement technique. Le notebook est considéré comme finalisé et presque propre. Les fondations sont prêtes pour entamer la **Phase 1 : Audit de qualité et nettoyage (Data Cleaning)**, où les décisions relatives au traitement des outliers et à l'imputation des valeurs manquantes seront actées.

Livrables validés :

- ☑ Arborescence système prête.
- ☑ Dataset brut sécurisé.
- ☑ Dictionnaire et classification de données généré (`reports/notebook_0/`).
- ☑ Diagnostic de distribution des cibles effectué.