Dictionnaire des variables issues du nettoyage des données

1. Rappel conceptuel : qu'est-ce qu'une variable flag ?

Une variable flag est une variable binaire construite explicitement pour **signaler une situation particulière dans les données**, sans prétendre représenter directement une caractéristique physique ou économique du phénomène étudié.

Un flag n'a pas pour objectif principal d'expliquer la variable cible, mais de :

- tracer une **anomalie de déclaration**,
- conserver la mémoire d'une **correction ou d'une transformation appliquée**,
- distinguer des **cas structurellement non comparables**,
- contrôler un biais potentiel lié à la qualité ou à la granularité des données.

Ces variables peuvent être mobilisées à trois niveaux :

- **audit et diagnostic de la qualité des données**,
- **variables de contrôle** dans un modèle,
- **critères d'exclusion ou de stratification** lors des analyses.

Elles ne doivent pas être interprétées comme des variables structurelles, même si elles sont intégrées à un modèle statistique.

---

2. Variables créées lors du nettoyage

2.1 `IsZeroFloorReported`

Cette variable binaire indique si le nombre d'étages déclaré dans la base brute était nul ou non valide.

- Valeur 1 : `NumberofFloors ≤ 0` dans les données initiales
- Valeur 0 : `NumberofFloors > 0`

Cette variable ne décrit pas un bâtiment sans étage, mais une incohérence ou une absence d'information dans la déclaration initiale.

Cas d'usage

- En phase de nettoyage, elle permet de distinguer les observations pour lesquelles une correction ou une imputation du nombre d'étages a été nécessaire.
- En analyse exploratoire, elle sert à évaluer si les bâtiments présentant une mauvaise qualité déclarative ont un comportement énergétique ou des émissions atypiques.
- Dans un modèle de régression, elle peut être introduite comme **variable de contrôle**, afin d'absorber un effet lié à la qualité de la donnée, sans interprétation causale directe.

Recommandation d'utilisation

- Variable pertinente comme flag d'audit et de contrôle.
- À ne pas interpréter comme une caractéristique physique du bâtiment.
- À ne pas utiliser comme substitut du nombre d'étages.

2.2 `IsAggregatedCampus`

Cette variable binaire identifie les observations correspondant à des entités agrégées de type campus.

- Valeur 1 : `BuildingType = Campus`
- Valeur 0 : autres types de bâtiments

Elle signale que l'observation regroupe potentiellement **plusieurs bâtiments hétérogènes**, partageant une consommation ou une surface agrégée.

Cas d'usage

- En phase de nettoyage, elle permet d'éviter l'application de règles de correction ou d'imputation conçues pour des bâtiments unitaires.
- En analyse descriptive, elle autorise des comparaisons séparées entre bâtiments individuels et entités agrégées.
- Dans un modèle, elle peut être utilisée comme variable de segmentation ou de contrôle pour capter les effets d'échelle et de structure.

Recommandation d'utilisation

- Variable essentielle pour la cohérence méthodologique.
- À inclure systématiquement dans toute analyse mobilisant des variables structurelles (étages, surface, intensité).
- À utiliser avec prudence dans les modèles prédictifs, compte tenu de la non-comparabilité structurelle avec les bâtiments unitaires.

2.3 `IsMixedUse`

Cette variable binaire indique si un bâtiment possède un usage secondaire déclaré, signalant une mixité d'usages.

- Valeur 1 : `SecondLargestPropertyUseType` est renseigné dans les données initiales
- Valeur 0 : `SecondLargestPropertyUseType` est manquant (bâtiment mono-usage)

Cette variable capture l'information structurelle sur la mixité sans conserver les détails lacunaires des usages secondaires.

Cas d'usage

- En phase de nettoyage, elle permet de préserver l'information sur la mixité tout en supprimant les colonnes originales à fort taux de valeurs manquantes.
- En analyse exploratoire, elle aide à distinguer les bâtiments mono-usage des mixtes pour évaluer les impacts sur la consommation énergétique.

- Dans un modèle, elle peut servir de variable explicative ou de contrôle pour capturer les effets de la complexité d'usage.

Recommandation d'utilisation

- Variable utile pour la segmentation des bâtiments par type d'usage.
- À interpréter comme un indicateur de structure plutôt qu'une mesure précise de mixité.
- À combiner avec `PrimaryPropertyType` pour des analyses plus fines.

2.4 `Has_EnergyStarScore`

Cette variable binaire indique si le score ENERGY STAR était renseigné dans les données initiales.

- Valeur 1 : `ENERGYSTARScore` est renseigné
- Valeur 0 : `ENERGYSTARScore` est manquant

Cette variable flagge un pattern de valeurs manquantes potentiellement informatif, lié à un mécanisme MAR (Missing At Random) conditionné par `PrimaryPropertyType`.

Cas d'usage

- En phase de nettoyage, elle permet de capter si l'absence de score constitue un signal prédictif avant imputation.
- En analyse exploratoire, elle sert à évaluer les biais potentiels liés à la non-déclaration volontaire (risque MNAR).
- Dans un modèle, elle peut être utilisée comme variable de contrôle pour absorber les effets liés à la qualité de déclaration.

Recommandation d'utilisation

- Variable pertinente pour l'audit des patterns de valeurs manquantes.
- À créer avant toute imputation de `ENERGYSTARScore`.
- À ne pas interpréter comme une mesure de performance énergétique, mais comme un indicateur de fiabilité déclarative.

2.5 Synthèse des nouvelles variables

Nom de la variable	Type	Nature	Finalité principale
IsZeroFloorReported	Binaire	Flag	Qualité de déclaration
IsAggregatedCampus	Binaire	Flag	Granularité / structure des données
IsMixedUse	Binaire	Flag	Mixité d'usages
Has_EnergyStarScore	Binaire	Flag	Pattern de valeurs manquantes

---

3. Variables supprimées lors du nettoyage

Certaines variables ont été exclues du jeu de données nettoyé pour des raisons méthodologiques clairement identifiées. Ces suppressions visent à améliorer la stabilité des analyses et la lisibilité des modèles.

Les catégories de variables supprimées incluent notamment :

- **Variables à taux de valeurs manquantes extrêmes** : Colonnes avec une densité d'information insuffisante (100% ou >94% de manquants), rendant toute imputation non crédible et alourdissant le pipeline sans valeur ajoutée.
- **Variables redondantes ou à faible valeur discriminante** : Colonnes singleton ou structurellement liées à d'autres, remplacées par des flags pour préserver l'information essentielle.
- **Variables temporaires pour diagnostic** : Utilisées pour validation puis supprimées avant modélisation pour éviter la redondance.

Liste exhaustive des variables supprimées :

Nom de la variable	Raison de suppression	Ordre concerné
Comments	Taux de valeurs manquantes de 100% ; aucune information exploitable.	4
YearsENERGYSTARCertified	Taux de valeurs manquantes de 94,26% ; densité résiduelle insuffisante pour modélisation.	4
Outlier	Taux de valeurs manquantes de 98,99% ; utilisée temporairement pour diagnostic des aberrants, puis supprimée avant entraînement.	5
ThirdLargestPropertyUseType	Taux de valeurs manquantes de 78,18% ; dépendance structurelle avec usages secondaires ; information compensée par `IsMixedUse`.	6
ThirdLargestPropertyUseTypeGFA	Taux de valeurs manquantes de 78,18% ; même justification que ci-dessus.	6
SecondLargestPropertyUseType	Taux de valeurs manquantes de 47,75% ; information transformée en `IsMixedUse` binaire ; suppression pour éviter lacunes.	7
SecondLargestPropertyUseTypeGFA	Taux de valeurs manquantes de 47,75% ; même justification que ci-dessus.	7

La suppression de ces variables n'implique pas une perte d'information substantielle, mais au contraire une **réduction du bruit statistique** et des risques d'instabilité des estimations.

---

4. Positionnement méthodologique

Les variables introduites à ce stade relèvent d'une **logique de traçabilité et de contrôle**, et non d'enrichissement explicatif direct. Leur rôle est de garantir que les décisions de nettoyage, d'imputation ou d'exclusion restent observables, justifiables et reproductibles.

Elles constituent un socle méthodologique indispensable pour :

- documenter les choix de préparation des données,
- sécuriser l'interprétation des résultats,
- renforcer la crédibilité statistique et scientifique de l'analyse.