

Tableau récapitulatif des décisions et actions de nettoyage

Le présent tableau consolide l'ensemble des décisions méthodologiques et opérationnelles prises au cours des sections précédentes. Il constitue le référentiel unique pour l'exécution du pipeline de nettoyage des données du projet Seattle Energy Benchmarking 2016. Chaque ligne correspond à une action atomique, justifiée par les analyses diagnostiques menées et ordonnée selon une séquence logique d'exécution.

Ordre d'exécution du pipeline de nettoyage

Nous avons défini la séquence d'exécution suivante pour garantir l'intégrité de la modélisation :

1. **Section 0** : Filtrage du périmètre.
2. **Section 2** : Audit de cohérence physique.
3. **Section 3** : Détection des outliers statistiques.
4. **Section 1** : Gestion des valeurs manquantes et suppression des colonnes.

Problème de l'imputation avant le split train/test

Imputer les valeurs manquantes **avant** le découpage train/test introduit un biais méthodologique majeur, communément appelé *data leakage*. En effet, les statistiques utilisées pour l'imputation (médiane par usage, médiane globale, etc.) sont alors calculées sur l'ensemble du dataset, y compris sur les observations qui devraient appartenir au jeu de test. Le modèle est donc indirectement exposé à de l'information qu'il ne devrait pas connaître au moment de l'entraînement, ce qui conduit à des performances artificiellement optimistes et non reproductibles.

Ce problème a été traité dans l'architecture du pipeline via le module `section1.py`. La fonction `Section1.run()` dispose d'un argument de configuration `enable_imputation`, activé par défaut (`True`).

- En **phase d'EDA**, l'imputation est activée afin de faciliter l'exploration, la visualisation et l'analyse statistique des données.
- En **phase de modélisation**, l'imputation est désactivée : le split train/test est effectué sur des données non imputées, et l'imputation est ensuite intégrée proprement dans le pipeline de modélisation (apprise uniquement sur le train et appliquée au test).

Justification et choix méthodologique

Nous avons choisi de placer la Section 1 en dernière position car l'imputation des valeurs manquantes ou la suppression préventive de colonnes risquerait de masquer des réalités physiques ou de biaiser les distributions statistiques nécessaires au calcul des bornes IQR. En auditant d'abord la cohérence arithmétique (Section 2) et les déviations extrêmes (Section 3), nous travaillons sur un signal pur, non pollué par des valeurs artificielles injectées lors d'une imputation. Cet ordre permet de qualifier précisément chaque bâtiment avant de décider si son manque de données justifie une exclusion finale.

Nous précisons que dans notre environnement de travail, l'ordre initial importait peu car nos fonctions d'audit ne modifiaient pas le jeu de données brutes. Les algorithmes statistiques utilisés ignorent nativement les valeurs nulles, ce qui garantit que le calcul de nos seuils reste identique, que les valeurs manquantes aient été flagguées ou non. Cette robustesse nous a permis d'isoler les incohérences physiques majeures sans dénaturer la représentativité du parc immobilier de Seattle pour la prédiction des émissions de CO₂.

Remarque méthodologique sur l'imputation par strate

Lors de l'application des diagnostics et des audits de la **Section 1**, nous avons constaté que certains sous-groupes d'usage (`PrimaryPropertyType`) étaient statistiquement sous-représentés ou ne présentaient aucune donnée valide pour certaines variables structurelles (notamment le nombre d'étages).

Face à ces "tranches vides" (détectées par les avertissements de type *Mean of empty slice*), nous avons pris la décision d'implémenter une **imputation en cascade** :

- **Priorité à la spécificité** : L'imputation est d'abord tentée au niveau de la sous-catégorie d'usage pour préserver la cohérence architecturale.
- **Sécurité par palier** : En cas d'absence de référentiel dans la sous-catégorie, le pipeline bascule sur la médiane globale du projet en dernier ressort.
- **Audit de fiabilité** : Cette décision permet de garantir un dataset 100% exploitable par les algorithmes de Machine Learning tout en marquant explicitement (via la colonne `IsZeroFloorReported`) les lignes dont la valeur a été reconstituée.

Remarque méthodologique sur la fiabilisation des intensités énergétiques (EUI)

La validité d'un modèle de prédiction des émissions de CO2 dépend directement de la cohérence interne des variables d'intensité. L'EUI (Energy Use Intensity) constitue l'indicateur pivot du dataset puisqu'il normalise la consommation énergétique par la dimension spatiale du bâtiment. Cependant, l'analyse diagnostique a révélé que les données d'intensité originales étaient inexploitables en l'état pour deux raisons fondamentales. D'une part, la correction préalable des consommations normalisées (**Weather Normalized**) a mécaniquement rendu les intensités d'origine obsolètes. D'autre part, la détection d'observations affichant une intensité nulle malgré une consommation positive signalait une corruption logique du reporting initial (données manquantes ou erreurs de calcul tiers).

Restauration de la cohérence physique (Weather Normalized)

Nous avons identifié une faille critique dans les données d'origine : de nombreuses lignes affichaient une consommation brute strictement positive, alors que les variables normalisées météo (**SiteEnergyUseWN**, **SiteEUIWN** et **SourceEUIWN**) étaient nulles ou manquantes. Physiquement, un bâtiment qui consomme de l'énergie ne peut pas avoir une consommation normalisée nulle. Pour corriger ce biais qui aurait conduit à écarter injustement des bâtiments valides ou à fausser les moyennes, nous avons forcé la copie de la valeur brute vers la variable normalisée dès que cette anomalie était détectée. Cette étape de synchronisation garantit que l'impact climatique est neutralisé sans ignorer la réalité de la consommation énergétique mesurée.

Justification du référentiel de surface

Pour que nos calculs d'intensité soient fiables, nous avons dû déterminer s'il fallait diviser la consommation par la surface totale de la parcelle ou par la surface brute du bâtiment . Nos tests ont conclu que seule la surface bâtie permet de retrouver exactement les valeurs déclarées par la ville de Seattle. En isolant ainsi l'enveloppe thermique, nous évitons de diluer artificiellement la consommation des bâtiments dotés de vastes parkings ou jardins. Ce choix nous permet de démasquer l'inefficacité réelle des structures qui, autrement, auraient paru performantes grâce à l'étendue de leur terrain.

Ce choix de restaurer les valeurs normalisées nous a imposé de recalculer systématiquement les intensités énergétiques (EUI) correspondantes pour maintenir une cohérence totale dans notre dataset. Nous ne pouvions pas nous contenter des ratios déclarés, car ils ne reflétaient plus la réalité physique après notre intervention sur les consommations globales. En recalculant nous-mêmes ces indicateurs, nous garantissons que chaque intensité repose sur les mêmes règles de calcul et sur les surfaces bâties réelles.

Structure du tableau

Le tableau est organisé en colonnes distinctes permettant une traçabilité complète des opérations de nettoyage. La colonne Ordre définit la séquence d'exécution des actions, garantissant que les dépendances entre transformations sont respectées. La colonne Variable cible identifie précisément la ou les variables concernées par l'action. La colonne Action technique spécifie l'opération concrète à implémenter dans le code. La colonne Justification méthodologique explicite le raisonnement analytique qui sous-tend la décision. Enfin, la colonne Validation attendue précise les contrôles qualité à effectuer après exécution.

Section 0

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
-------	----------------	------------------	------------------------------	---------------------

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
1	BuildingType et PrimaryPropertyType	Filtrer et conserver uniquement les observations où BuildingType est différent de Multifamily MR (5-9), Multifamily HR (10+) et Multifamily LR (1-4) et PrimaryPropertyType est vraiment non-residentiel	Le périmètre du projet est strictement limité aux bâtiments non résidentiels. Les bâtiments résidentiels obéissent à des logiques de consommation énergétique structurellement différentes. Leur inclusion introduirait une hétérogénéité incompatible avec l'objectif de modélisation. Cette restriction est conforme au cahier des charges métier et assure la cohérence statistique des analyses ultérieures.	Vérifier que BuildingType ne contient plus aucune des trois modalités résidentielles. Contrôler que la dimension du dataframe a diminué du nombre attendu de lignes résidentielles. Confirmer que toutes les modalités restantes correspondent bien à des usages non résidentiels.
2	BuildingType	Remplacer la modalité Nonresidential WA par NonResidential	L'analyse du bâtiment identifiant 50101 (Burke Museum, classé University dans PrimaryPropertyType) révèle une incohérence taxonomique. La modalité Nonresidential WA ne possède qu'une seule occurrence et son usage correspond à la définition standard de NonResidential. Cette harmonisation élimine une catégorie singleton qui n'apporte aucune valeur discriminante et risque de poser des problèmes lors de l'encodage ou de la validation croisée.	Vérifier que la modalité Nonresidential WA a totalement disparu du jeu de données. Confirmer que le nombre d'occurrences de NonResidential a augmenté d'une unité. S'assurer que le bâtiment 50101 est désormais classé NonResidential.

Section 2

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
--------------	-----------------------	-------------------------	-------------------------------------	----------------------------

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
3	PropertyGFATotal	Supprimer toutes les lignes où PropertyGFATotal est inférieur ou égal à zéro	<p>Une surface totale nulle ou négative constitue une impossibilité physique absolue. Un bâtiment ne peut exister sans occuper d'espace.</p> <p>Cette anomalie révèle soit une erreur de saisie critique, soit une absence totale de données exploitables pour le calcul des intensités énergétiques. La conservation de telles lignes corromprait mécaniquement tous les ratios énergétiques par unité de surface, rendant impossible la comparaison entre bâtiments et faussant les distributions des variables dérivées comme l'EUI (Energy Use Intensity).</p>	<p>Vérifier qu'aucune observation ne présente une valeur de PropertyGFATotal inférieure ou égale à zéro après nettoyage. Confirmer que toutes les valeurs de PropertyGFATotal sont strictement positives.</p> <p>Documenter le nombre exact de lignes supprimées pour cette raison dans les logs de traçabilité.</p>

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
4	SiteEnergyUse(kBtu), TotalGHGEmissions, Electricity(kBtu), NaturalGas(kBtu)	Supprimer toutes les lignes présentant au moins une valeur strictement négative dans ces variables énergétiques	<p>Les variables de consommation énergétique et d'émissions de gaz à effet de serre ne peuvent être négatives par définition physique. Une consommation négative violerait le premier principe de la thermodynamique. Ces valeurs traduisent des erreurs de saisie, des problèmes de conversion d'unités ou des corruptions de données lors de l'export. Leur maintien introduirait un biais systématique dans l'estimation des distributions et pourrait conduire le modèle à apprendre des patterns aberrants.</p> <p>Une seule variable négative suffit à invalider l'ensemble de l'observation car elle signale une défaillance globale du processus de reporting pour ce bâtiment.</p>	<p>Confirmer l'absence totale de valeurs négatives dans les quatre variables énergétiques critiques après nettoyage. Contrôler que les distributions de ces variables ne présentent plus d'anomalies dans les queues de distribution négatives. Documenter précisément quelles variables étaient négatives pour chaque ligne supprimée afin d'identifier d'éventuels patterns systématiques d'erreur.</p>

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
5	SiteEnergyUseWN(kBtu), SiteEUIWN(kBtu/sf), SourceEUIWN(kBtu/sf)	Pour toutes les lignes où la variable brute est strictement positive mais la variable Weather Normalized est nulle ou manquante, copier la valeur brute dans la variable normalisée	Les variables Weather Normalized synchronisées avec leurs équivalents bruts garantissent la cohérence déterministe entre consommation et intensité. Cette correction technique préserve l'exploitabilité des observations en évitant les divisions par zéro lors du recalculation des intensités énergétiques. L'hypothèse sous-jacente est que l'absence de correction météorologique est moins préjudiciable pour la modélisation qu'une valeur nulle aberrante qui corromprait les calculs d'intensité énergétique.	Vérifier que toutes les lignes avec énergie brute positive disposent désormais d'une valeur WN cohérente. Contrôler l'absence de valeurs nulles résiduelles dans les variables Weather Normalized lorsque les variables brutes sont positives. Valider que cette synchronisation n'a pas créé de nouvelles incohérences avec d'autres variables dérivées.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
6	NumberofFloors	Créer une variable binaire IsZeroFloorReported valant 1 si NumberofFloors était initialement inférieur ou égal à zéro et 0 sinon, puis créer une variable binaire IsAggregatedCampus valant 1 si BuildingType est égal à Campus et 0 sinon	L'analyse qualitative de seize bâtiments déclarant zéro étage a révélé deux populations distinctes. D'une part, des structures individuelles massives comme le Grand Hyatt Seattle ou le centre commercial Pacific Place, pour lesquelles le zéro constitue manifestement une erreur de saisie ou une omission lors du reporting. D'autre part, des entités agrégées de type Campus comme l'University of Washington, pour lesquelles le concept d'étage perd son sens car il s'agit d'un ensemble de bâtiments distincts. La création de ces deux flags permet de préserver cette information structurelle tout en facilitant le traitement différencié ultérieur. Le flag IsZeroFloorReported servira d'indicateur de confiance pour les futures imputations, tandis que IsAggregatedCampus permettra au modèle de distinguer les agrégats multi-bâtiments des structures uniques.	Vérifier que IsZeroFloorReported contient exactement le nombre de 1 correspondant au nombre initial de bâtiments avec NumberofFloors inférieur ou égal à zéro. Contrôler que IsAggregatedCampus identifie correctement tous les bâtiments de type Campus. Valider que ces variables sont créées avant toute modification de la variable NumberofFloors elle-même. S'assurer que les deux flags sont de type binaire strict.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
7	NumberofFloors	Pour les lignes où NumberofFloors est inférieur ou égal à zéro, remplacer la valeur par NaN	<p>Pour les bâtiments déclarant zéro étage, la valeur zéro ne peut être maintenue car elle créerait des calculs de densité aberrants (surface infinie par étage) lors des analyses multidimensionnelles et de la détection d'outliers. Le recodage en NaN permet de traiter explicitement cette absence d'information plutôt que de conserver une valeur physiquement incohérente. Cette transformation préserve l'ensemble des autres variables de l'observation, notamment les consommations énergétiques et les surfaces qui restent exploitables. Les valeurs manquantes ainsi créées seront imputées ultérieurement en Section 1 par des méthodes statistiques basées sur la relation entre surface totale et nombre d'étages au sein de chaque catégorie d'usage.</p>	<p>Confirmer que tous les bâtiments avec zéro étage ont été recodés en NaN. Contrôler que le nombre de valeurs manquantes dans NumberofFloors a augmenté exactement du nombre attendu de conversions. Valider que les lignes recodées conservent toutes leurs autres variables intactes.</p>

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
8	SiteEUIWN(kBtu/sf), SourceEUIWN(kBtu/sf)	Recalculer les intensités énergétiques en divisant les consommations normalisées corrigées par PropertyGFABuilding(s) (reviens au même que de le faire uniquement pour les bâtiments avec variables WN modifiée	La correction des consommations Weather Normalized rend les intensités d'origine caduques. Le recalcul est indispensable pour restaurer la cohérence déterministe entre l'énergie et la surface. Le choix de PropertyGFABuilding(s) comme dénominateur, validé par une MAE de 0.00, garantit l'alignement strict avec la méthodologie de Seattle. Cette mise en conformité élimine le biais induit par les surfaces non chauffées (parkings) et fiabilise la détection ultérieure des outliers de performance.	Confirmer que le dénominateur utilisé exclut les surfaces hors-bâti pour maintenir la précision mathématique du ratio. Vérifier l'absence de valeurs infinies suite au recalcul et à leur remplacement par NaN. Contrôler que la corrélation entre les nouvelles intensités et les émissions de CO2 est renforcée. Valider la disparition des incohérences où l'énergie est positive mais l'intensité nulle.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
9	LargestPropertyUseTypeGFA, PropertyGFATotal	Calculer le ratio LargestPropertyUseTypeGFA divisé par PropertyGFATotal puis supprimer toutes les lignes où ce ratio est strictement supérieur à une valeur critique	Un ratio dépassant un certain seuil signifie que la surface déclarée pour l'usage principal excède la surface totale du bâtiment de manière inacceptable. Cette incohérence physique est jugée irréversible et ne peut résulter d'une simple imprécision de mesure. Conserver ces observations fausserait systématiquement le calcul de l'intensité énergétique par unité de surface en sous- estimant artificiellement la consommation réelle. Un dépassement au- delà du seuil introduit un biais trop important pour être compensé par quelque méthode statistique que ce soit.	Vérifier qu'aucune observation ne présente un ratio dépassant le seuil critique après suppression. Calculer le nombre exact de lignes supprimées pour cette raison et documenter leur répartition par Primary.PropertyType. Contrôler que les distributions de LargestPropertyUseTypeGFA et PropertyGFATotal ne présentent plus de cas extrêmes au-delà du seuil critique. Valider que cette suppression est effectuée avant tout calcul d'indicateurs énergétiques dérivés.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
10	Electricity(kBtu), NaturalGas(kBtu), SteamUse(kBtu), SiteEnergyUseWN(kBtu)	Calculer la somme des trois sources énergétiques, puis calculer l'écart relatif absolu entre cette somme et SiteEnergyUseWN(kBtu), enfin supprimer toutes les lignes où cet écart relatif est strictement supérieur à 0.25 (on évite de supprimer trop de bâtiments)	<p>La cohérence énergétique constitue un prérequis fondamental pour la validité des émissions de CO2 calculées. La municipalité de Seattle applique des facteurs d'émission spécifiques à chaque source d'énergie pour obtenir les émissions totales.</p> <p>Si la somme des sources énergétiques détaillées ne correspond pas au total déclaré avec un écart supérieur à dix pourcent, cela signale soit une source d'énergie majeure omise, soit une corruption des données de conversion. Un écart de cette amplitude ne peut être attribué à de simples arrondis lors des conversions d'unités. Conserver ces observations reviendrait à accepter des valeurs de CO2 mathématiquement corrompues, ce qui viderait de sens l'objectif même de prédiction des émissions.</p>	<p>Vérifier qu'aucune observation ne présente un écart énergétique supérieur à dix pourcent après nettoyage. Documenter précisément les bâtiments supprimés avec leurs écarts respectifs. Analyser la distribution des écarts résiduels pour confirmer qu'ils restent dans des plages acceptables. Valider que les observations conservées présentent une cohérence mathématique entre sources et total.</p> <p>Contrôler que cette suppression n'introduit pas de biais sectoriel en vérifiant la représentativité des types de bâtiments restants.</p>

Section 3

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
11	SiteEUIWN(kBtu/sf)	Aligner l'intensité énergétique sur la surface bâtie (PropertyGFABuilding(s))	Cette étape préparatoire garantit que la détection d'outliers s'appuie sur une densité énergétique réelle, évitant que des erreurs de reporting de surface ne biaissent les seuils statistiques.	Validation de la cohérence entre les consommations synchronisées en Section 2 et le référentiel de surface bâti.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
12	Variables de volume (Z-score)	Calculer les Z-scores par <code>PrimaryPropertyType</code> et sommer les dépassements de seuil critique	Permet d'isoler les bâtiments dont les consommations globales sont aberrantes au sein de leur propre catégorie d'usage, sans pénaliser les secteurs naturellement énergivores.	Création de la feature <code>extreme_zscore_count</code> capturant le cumul des anomalies de volume.
13	Variables de structure (IQR)	Appliquer la détection IQR segmentée par usage sur les intensités et les surfaces	L'approche par quartile respecte l'hétérogénéité architecturale du parc. Elle cible spécifiquement les anomalies de ratio (EUI) et de dimensionnement (GFA).	Identification des flags <code>is_iqr_outlier</code> pour les variables critiques de performance et de taille.
14	Toutes variables techniques	Archiver <code>extreme_zscore_count</code> et supprimer les colonnes de calcul individuelles	On transforme une série de diagnostics binaires en un indicateur de fiabilité multidimensionnel conservé pour enrichir le signal du modèle de prédiction.	Nettoyage du dataset final tout en préservant le score d'anomalie agrégé comme feature d'entrée.
15	Lignes filtrées	Exclure les observations cumulant trop de Z-scores ou présentant un IQR critique, hors structures massives	Assure la pureté statistique du dataset en supprimant les corruptions de données, tout en protégeant les hôpitaux ou universités (<code>massive_structures_types</code>) dont la taille justifie l'écart.	Vérification de la réduction du bruit statistique et confirmation de la préservation des mégastuctures légitimes.

Section 1 : Finalisation et Imputations en cascade

Cette section constitue le dernier verrou de qualité avant l'exportation du dataset. Elle transforme les lacunes de données en signaux explicites pour le modèle et assure l'intégrité statistique de l'échantillon.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
16	Outlier et ComplianceStatus	Suppression des lignes marquées ou non conformes.	Centralisation des exclusions : purge des anomalies statistiques détectées en Section 3 et des défauts administratifs (non-conformité).	<code>exclusion_reason</code> documenté ; 100% de conformité administrative.
17	IsMixedUse	Création du flag binaire via <code>SecondLargestPropertyUseType</code> .	Capture la complexité structurelle. On transforme une donnée textuelle lacunaire en un signal binaire robuste pour le modèle.	Suppression des colonnes sources d'usage secondaire après création.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
18	Has_EnergyStarScore	Création du flag de présence avant toute imputation.	L'absence de score (MNAR) est une information en soi. Le flag permet de conserver ce signal avant que la valeur numérique ne soit complétée par la médiane.	Flag présent pour 100% des lignes ; distinction entre score réel et imputé.
19	LargestPropertyUseType	Réparation par <code>PrimaryPropertyType</code> ou étiquette "Unknown".	Étape critique : garantit que chaque bâtiment possède une catégorie pour permettre l'imputation groupée (Étape 20). Traite les cas limites (ex: 4 bâtiments).	Zéro <code>Nan</code> sur la colonne pivot d'usage avant l'imputation numérique.
20	Variables Numériques	Imputation en cascade : Médiane par Usage > Médiane Globale.	Respecte la morphologie du bâti (un hôtel consomme différemment d'un entrepôt). La cascade assure la complétude sans perte de précision.	<code>missing_before</code> vs <code>after</code> cohérent dans l'audit ; absence totale de <code>Nan</code> .
21	NumberofFloors	Imputation en cascade avec arrondi à l'entier (<code>.round().astype(int)</code>).	Assure la cohérence physique des bâtiments : le modèle ne doit pas traiter de "demi-étages" issus d'un calcul de médiane.	Valeurs entières strictes sur l'ensemble de la colonne.
22	Lignes vides (> 30%)	Suppression des observations au taux de vacuité résiduel critique.	Filtre de sécurité final : rejette les bâtiments dont le profil reste trop incertain malgré les tentatives d'imputation, limitant le bruit.	Taux de remplissage final vérifié ; motif "Excessive Missing Data" en audit.
23	Colonnes obsolètes	Drop définitif (ex: ID, Comments, 3rd Use, Outlier).	Réduction de la dimensionnalité. On élimine les variables "bruit" ou administratives pour ne conserver que les prédicteurs et la cible.	Le schéma final correspond exactement à la configuration YAML.