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.

Remarque méthodologique sur l'imputation par strate

Lors de l'application des diagnostics et des audits de la **Section 2**, 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 de la catégorie parente, puis 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).

Justification du recalcul par l'analyse d'erreur

Pour identifier le référentiel de calcul exact et éviter d'introduire un biais de normalisation, une étude comparative basée sur l'Erreur Moyenne Absolue (**MAE**) a été menée. Ce test statistique visait à comparer les valeurs d'intensité déclarées par la ville avec deux modèles de recalcul distincts. Le premier modèle utilisait la surface totale de la parcelle (**PropertyGFATotal**) comme dénominateur, tandis que le second se limitait à la surface brute du bâtiment (**PropertyGFABuilding**).

Les résultats du test MAE ont été déterminants. Le calcul basé sur la surface totale a généré une erreur moyenne de **6.214851**, révélant un décalage systématique entre le bâti réel et la surface déclarée incluant les extérieurs. À l'inverse, le calcul basé sur la surface du bâtiment a produit une MAE de **0.0000**. Cette identité mathématique parfaite a permis de confirmer que la méthodologie officielle de Seattle isole strictement l'enveloppe thermique close. L'utilisation erronée de la surface totale dans le pipeline aurait artificiellement dilué la consommation des bâtiments dotés de vastes parkings ou jardins, masquant ainsi leur inefficacité réelle.

Application de la formule de mise en conformité

L'action de nettoyage a consisté à rétablir la vérité physique du dataset en appliquant systématiquement la formule de normalisation suivante :

Cette équation garantit une cohérence thermodynamique totale : le lien entre la surface, l'énergie et l'intensité est désormais de nature déterministe. Sur le plan de l'apprentissage automatique, ce traitement élimine les signaux contradictoires qui auraient pu induire le modèle en erreur, notamment lors de la phase de détection des valeurs aberrantes par l'écart interquartile (**IQR**). Les outliers ainsi identifiés correspondent désormais à de réelles anomalies de performance énergétique et non à de simples résidus de calculs incohérents.

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 1

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
3	ComplianceStatus	Supprimer toutes les lignes où ComplianceStatus est différent de Compliant	La variable ComplianceStatus indique si un bâtiment a satisfait aux exigences de reporting pour l'année étudiée. Les observations marquées Error - Correct Default Data, Non-Compliant ou Missing Data présentent des défauts de déclaration reconnus par la municipalité de Seattle. Bien que minoritaires, ces lignes comportent un risque de biais systématique dans les données déclarées. Par principe de précaution et alignement avec les standards municipaux, ces observations sont écartées pour garantir la fiabilité du modèle final.	Vérifier que ComplianceStatus ne contient plus que la modalité Compliant. Contrôler le nombre de lignes supprimées et documenter leur proportion initiale. S'assurer que cette suppression n'a pas d'impact disproportionné sur une catégorie particulière de bâtiments.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
4	Comments, YearsENERGYSTARCertified	Supprimer définitivement les colonnes du dataframe	La variable Comments présente un taux de valeurs manquantes de 100 pourcent, ce qui la rend statistiquement inexploitable. La variable YearsENERGYSTARCertified affiche un taux de 94,26 pourcent de valeurs manquantes. La densité d'information résiduelle est insuffisante pour justifier une quelconque tentative d'imputation ou de modélisation. Ces variables n'apportent aucune valeur prédictive et leur maintien alourdirait inutilement le pipeline de traitement.	Confirmer l'absence totale des colonnes Comments et YearsENERGYSTARCertified dans le dataframe nettoyé. Vérifier que le nombre de colonnes a diminué de deux unités. S'assurer qu'aucune référence à ces variables ne subsiste dans les métadonnées ou les logs.
5	Outlier	Conserver temporairement la colonne pour analyse diagnostique puis la supprimer avant l'entraînement du modèle	La variable Outlier contient 98,99 pourcent de valeurs manquantes mais signale les observations identifiées comme atypiques par la municipalité. Cette information peut être précieuse lors de la section d'analyse des valeurs aberrantes pour confronter nos propres détections statistiques aux marqueurs officiels. Cependant, sa quasi-absence de renseignement et son caractère redondant avec nos futures analyses justifient sa suppression avant la phase de modélisation.	Utiliser Outlier dans la section 3 pour valider ou contester les détections automatiques. Après exploitation diagnostique, vérifier la suppression effective de la colonne. Documenter les cas où Outlier était renseigné et confronter ces marquages avec les détections issues des méthodes statistiques appliquées.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
6	ThirdLargestPropertyUseType, ThirdLargestPropertyUseTypeGFA	Supprimer définitivement les deux colonnes du dataframe	Ces variables affichent un taux de 78,18 pourcent de valeurs manquantes, ce qui signifie que moins d'un bâtiment sur cinq possède un troisième usage déclaré. La matrice de corrélation des manquants a révélé une dépendance structurelle entre l'absence de second usage et celle de troisième usage, confirmant une logique en entonnoir. La faible densité d'information et le risque de sur-apprentissage sur une minorité de cas complexes justifient l'exclusion de ces variables. La complexité des bâtiments sera capturée via la variable IsMixedUse créée ultérieurement.	Confirmer la suppression des colonnes ThirdLargestPropertyUseType et ThirdLargestPropertyUseTypeGFA. Vérifier la diminution du nombre de colonnes. S'assurer que l'information de mixité sera compensée par la création de la variable binaire lors du feature engineering.
7	SecondLargestPropertyUseType, SecondLargestPropertyUseTypeGFA	Créer une variable binaire IsMixedUse valant 1 si SecondLargestPropertyUseType est renseigné et 0 sinon, puis supprimer les deux colonnes originales	Avec 47,75 pourcent de valeurs manquantes, ces variables présentent une densité d'information insuffisante pour une exploitation directe. Cependant, la matrice de corrélation a montré une corrélation parfaite de 1 entre l'absence du type d'usage secondaire et celle de sa surface associée, confirmant que l'absence est informative et non accidentelle. Un bâtiment sans second usage est un bâtiment mono-usage. La transformation en indicateur binaire permet de préserver cette information structurelle sans gérer des colonnes lacunaires.	Vérifier que IsMixedUse est créée avec exactement deux modalités (0 et 1). Contrôler que le nombre de 1 correspond au nombre d'observations où SecondLargestPropertyUseType était renseigné. Confirmer la suppression de SecondLargestPropertyUseType et SecondLargestPropertyUseTypeGFA. Valider que la proportion de bâtiments mixtes correspond aux attentes métier.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
8	Toutes les variables numériques avec moins de 1 pourcent de valeurs manquantes	Imputer par la médiane calculée au sein de chaque groupe défini par PrimaryPropertyType	Les variables numériques présentant moins de 1 pourcent de valeurs manquantes ont un impact statistique négligeable sur la distribution globale. L'imputation par médiane groupée respecte la spécificité de chaque usage (un entrepôt n'a pas les mêmes caractéristiques qu'un bureau) tout en étant robuste aux valeurs extrêmes. La médiane est préférée à la moyenne car elle n'est pas affectée par les outliers potentiels au sein de chaque groupe. Cette méthode simple est adaptée aux faibles taux de valeurs manquantes selon les recommandations de la littérature.	Vérifier que toutes les variables numériques concernées n'ont plus aucune valeur manquante après imputation. Contrôler que les médianes imputées sont cohérentes avec les distributions observées dans chaque groupe PrimaryPropertyType. Valider que les distributions avant et après imputation restent statistiquement similaires via des tests de Kolmogorov-Smirnov.
9	Toutes les variables catégorielles avec moins de 1 pourcent de valeurs manquantes	Imputer par la modalité la plus fréquente calculée au sein de chaque groupe défini par PrimaryPropertyType	Pour les variables catégorielles à faible taux de valeurs manquantes, l'imputation par le mode groupé préserve la cohérence sectorielle. Un bâtiment de type Office aura plus de chances de partager les caractéristiques modales d'autres bureaux que celles d'un entrepôt. Cette approche minimise l'erreur d'imputation en s'appuyant sur la proximité métier. Le mode est la statistique centrale naturelle pour les variables qualitatives et son usage est justifié par la faiblesse du taux de valeurs manquantes.	Confirmer l'absence de valeurs manquantes dans les variables catégorielles traitées. Vérifier que les modalités imputées correspondent bien aux modes de chaque groupe PrimaryPropertyType. Contrôler que les fréquences relatives des modalités n'ont pas été artificiellement déformées par l'imputation.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
10	ENERGYSTARScore	Créer une variable binaire Has_EnergyStarScore valant 1 si ENERGYSTARScore est renseigné et 0 sinon	Le test du Chi2 a révélé une p-value de 1,7978 × 10^-180 démontrant que l'absence du ENERGYSTARScore n'est pas aléatoire mais conditionnée par PrimaryPropertyType. Nous sommes en présence d'un mécanisme MAR (Missing At Random). Cependant, il subsiste un risque résiduel de MNAR (Missing Not At Random) si certains propriétaires omettent volontairement de déclarer un score médiocre. La création d'une variable flag permet au modèle de capter si l'absence d'information constitue en soi un signal prédictif de la consommation énergétique. Cette précaution méthodologique préserve l'information contenue dans le pattern de valeurs manquantes.	Vérifier que Has_EnergyStarScore contient exactement deux modalités. Contrôler que le nombre de 0 correspond précisément au nombre de valeurs manquantes initiales de ENERGYSTARScore. Valider que la variable est créée avant toute imputation du score lui-même.
11	ENERGYSTARScore	Imputer les valeurs manquantes par la médiane calculée au sein de chaque groupe défini par PrimaryPropertyType	L'analyse a démontré que PrimaryPropertyType explique significativement mieux la répartition des valeurs manquantes que BuildingType. Le ENERGYSTARScore est une métrique de performance énergétique dont la distribution varie fortement selon le secteur d'activité. Un hôtel n'a pas les mêmes standards qu'un bureau. L'imputation par médiane groupée respecte cette hétérogénéité sectorielle tout en étant robuste aux valeurs extrêmes qui pourraient biaiser une moyenne. La médiane garantit que la valeur imputée représente le bâtiment typique de chaque catégorie.	Confirmer l'absence totale de valeurs manquantes dans ENERGYSTARScore après imputation. Vérifier que les médianes imputées sont cohérentes avec les distributions observées dans chaque groupe PrimaryPropertyType. Comparer les distributions avant et après imputation via des visualisations (boxplots par groupe) et des tests statistiques. S'assurer que les valeurs imputées restent dans l'intervalle [1, 100].

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
12	Lignes avec taux de valeurs manquantes supérieur à 30 pourcent	Supprimer les lignes dont le pourcentage de valeurs manquantes dépasse 30 pourcent	L'analyse par ligne a identifié deux observations présentant plus de 30 pourcent de valeurs manquantes. Ces lignes sont caractérisées par l'absence structurelle des variables énergétiques centrales (consommations, intensités, émissions de CO2). Toute tentative d'imputation serait non crédible d'un point de vue statistique et métier. Leur poids dans l'échantillon est négligeable (deux lignes sur plusieurs milliers) et leur suppression n'affecte ni la représentativité ni la puissance analytique du jeu de données. Leur maintien introduirait un bruit inutile et fragiliserait la robustesse des modèles.	Calculer le taux de valeurs manquantes par ligne avant suppression et confirmer l'identification des lignes à plus de 30 pourcent. Vérifier que précisément deux lignes ont été supprimées. Contrôler que les lignes supprimées étaient bien celles identifiées lors du diagnostic. Valider que le taux maximum de valeurs manquantes par ligne dans le dataset nettoyé est désormais inférieur ou égal à 30 pourcent.
13	Lignes avec taux de valeurs manquantes entre 20 et 30 pourcent	Conserver les lignes dont le taux de valeurs manquantes est compris entre 20 et 30 pourcent	L'analyse a révélé que les deux lignes dans cette tranche conservent une information substantielle. Leur statut ComplianceStatus est Compliant, ce qui signifie que la municipalité les considère comme fiables malgré quelques lacunes. Surtout, les variables critiques pour la modélisation (émissions de CO2, consommations principales) sont renseignées. Les valeurs manquantes sont dispersées sur des variables secondaires qui peuvent être traitées par les stratégies d'imputation déjà définies. Leur suppression serait une perte d'information injustifiée.	Vérifier que les deux lignes identifiées dans cette tranche sont toujours présentes dans le dataset nettoyé. Contrôler que leur ComplianceStatus est bien Compliant. S'assurer que les variables critiques (notamment les émissions de CO2) sont renseignées pour ces observations. Valider que les valeurs manquantes résiduelles seront traitées par les imputations déjà prévues.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
14	Dataset complet	Effectuer un contrôle de cohérence global post-nettoyage	Après l'exécution de toutes les étapes précédentes, il est indispensable de valider la cohérence globale du dataset nettoyé. Ce contrôle transversal permet de détecter d'éventuelles incohérences introduites par la séquence d'opérations ou des effets de bord non anticipés. Il constitue le point de contrôle qualité final avant la phase d'analyse exploratoire approfondie et de modélisation.	Vérifier que le nombre de lignes finales correspond au nombre initial moins les suppressions documentées (résidentiels, non-compliant, lignes à plus de 30 pourcent de NA). Confirmer que le nombre de colonnes correspond au nombre initial moins les suppressions de variables (Comments, YearsENERGYSTARCertified, Outlier, ThirdLargest, SecondLargest) plus les créations (IsMixedUse, Has_EnergyStarScore). Valider qu'aucune variable critique ne présente de valeurs manquantes résiduelles non traitées. Contrôler la cohérence des types de données (numériques, catégorielles, booléennes). Générer un rapport de synthèse listant les dimensions finales, les variables conservées et les transformations appliquées.

Section 2

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
15	PropertyGFATotal	Supprimer toutes les lignes où PropertyGFATotal est inférieur ou égal à zéro	Une surface totale nulle ou négative constitue une impossibilité physique absolue. Un bâtiment ne peut exister sans occuper d'espace. 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).	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. Documenter le nombre exact de lignes supprimées pour cette raison dans les logs de traçabilité.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
16	SiteEnergyUse(kBtu), TotalGHGEmissions, Electricity(kBtu), NaturalGas(kBtu)	Supprimer toutes les lignes présentant au moins une valeur strictement négative dans ces variables énergétiques	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. 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.	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.
17	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
18	NumberofFloors	Pour les lignes où NumberofFloors est inférieur ou égal à zéro ET IsAggregatedCampus vaut 0, remplacer la valeur de NumberofFloors par NaN puis faire l'imputation groupé	Pour les bâtiments individuels massifs identifiés lors de l'analyse qualitative, 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 pourront être imputées ultérieurement par des méthodes statistiques basées sur la relation entre surface totale et nombre d'étages au sein de chaque catégorie d'usage.	Confirmer que tous les bâtiments individuels avec zéro étage ont été recodés en NaN. Vérifier que les bâtiments de type Campus conservent leur valeur zéro d'origine. 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.
20	LargestPropertyUseTypeGFA, PropertyGFATotal	Calculer le ratio LargestPropertyUseTypeGFA divisé par PropertyGFATotal puis supprimer toutes les lignes où ce ratio est strictement supérieur à 1.30	Un ratio dépassant 1.30 signifie que la surface déclarée pour l'usage principal excède la surface totale du bâtiment de plus de trente pourcent. Cette incohérence physique est jugée irrémédiable et ne peut résulter d'une simple imprécision de mesure. L'analyse par violin plot a révélé que ces cas extrêmes se concentrent dans des catégories spécifiques comme les entrepôts ou les écoles du district SPS, souvent marquées avec un ComplianceStatus invalide. Conserver ces observations fausserait systématiquement le calcul de l'intensité énergétique par unité de surface en sous-estimant artificiellement la consommation réelle. Un dépassement de trente pourcent introduit un biais trop important pour être compensé par quelque méthode statistique que ce soit.	Vérifier qu'aucune observation ne présente un ratio supérieur à 1.30 après suppression. Calculer le nombre exact de lignes supprimées pour cette raison et documenter leur répartition par PrimaryPropertyType. 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
21	LargestPropertyUseTypeGFA, PropertyGFATotal, NumberofBuildings	Pour les lignes où le ratio est compris strictement entre 1.05 et 1.30 ET NumberofBuildings est égal à 1 ou manquant, supprimer ces observations	Pour les bâtiments monostructurels simples, un dépassement de la surface d'usage par rapport à la surface totale compris entre cinq et trente pourcent ne peut être justifié par une quelconque complexité architecturale ou chevauchement d'usages. L'analyse du graphique de distribution uniforme a montré que ces cas se situent dans une longue traîne après la zone de tolérance technique, suggérant des erreurs de saisie plutôt que des particularités physiques légitimes. Pour un entrepôt isolé ou un bureau simple, accepter un ratio de 1.20 reviendrait à tolérer une erreur de vingt pourcent sur la variable de normalisation principale, créant ainsi des bâtiments artificiellement performants dans les calculs d'intensité énergétique. Cette exclusion garantit que seules les structures véritablement complexes bénéficient d'une tolérance étendue.	Identifier et supprimer toutes les lignes monostructurelles avec ratio intermédiaire. Vérifier que NumberofBuildings a été correctement utilisé comme critère de discrimination. Documenter le nombre de suppressions et leur répartition par PrimaryPropertyType. Contrôler que les bâtiments multi-structures avec le même niveau de ratio ont été préservés. Valider que la distinction mono versus multi-structures est cohérente avec les définitions métier.
22	LargestPropertyUseTypeGFA, PropertyGFATotal, NumberofBuildings	Pour les lignes où le ratio est compris strictement entre 1.05 et 1.30 ET NumberofBuildings est strictement supérieur à 1, conserver ces observations conditionnellement à une validation ultérieure de leur Energy Use Intensity	Les bâtiments multi-usages ou multi-structures présentent légitimement des chevauchements de surfaces qui peuvent expliquer un dépassement modéré du ratio. Un hôpital avec plusieurs ailes ou un campus universitaire peut déclarer des surfaces fonctionnelles qui s'étendent au-delà du bâti principal strict par inclusion de zones techniques ou de circulations externes. L'analyse par violin plot a montré que pour certaines catégories comme Hospital ou Medical Office, ce dépassement constitue presque la norme plutôt que l'exception, suggérant une convention de déclaration sectorielle. La conservation conditionnelle permet de maintenir la diversité typologique du dataset tout en reportant la validation finale à une analyse de cohérence énergétique. Ces bâtiments seront ultérieurement confrontés aux distributions d'EUI de leur catégorie pour confirmer qu'ils ne présentent pas d'anomalies de performance.	Identifier précisément les bâtiments multi-structures conservés malgré leur ratio intermédiaire. Vérifier que leur NumberofBuildings est effectivement supérieur à un. Documenter leurs caractéristiques pour permettre la validation ultérieure par analyse d'EUI. Marquer ces observations avec un flag de vigilance pour traçabilité. Préparer une analyse comparative de leur intensité énergétique par rapport aux médianes de leurs catégories respectives.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
23	LargestPropertyUseTypeGFA, PropertyGFATotal	Pour les lignes où le ratio est compris entre 1.00 exclu et 1.05 inclus, conserver toutes ces observations sans restriction	La zone de tolérance technique entre un et cinq pourcent correspond à des écarts attribuables aux différences de conventions de mesure entre le Portfolio Manager de l'EPA et les relevés cadastraux de Seattle. L'analyse du graphique de distribution uniforme a révélé une concentration massive de bâtiments juste après la limite physique de 1.00, confirmant qu'il s'agit d'un pattern systématique plutôt que d'erreurs aléatoires. Ces écarts mineurs résultent probablement de l'inclusion ou l'exclusion des murs extérieurs, des cages d'escalier ou des zones communes selon les méthodologies. L'impact sur le calcul de l'Energy Use Intensity est statistiquement négligeable, la marge d'erreur de cinq pourcent étant inférieure à la variabilité naturelle des mesures énergétiques. Cette conservation préserve quatre-vingt-quinze observations qui contiennent par ailleurs des informations énergétiques complètes et fiables.	Confirmer que toutes les observations avec ratio entre 1.00 et 1.05 ont été conservées. Calculer précisément le nombre de bâtiments dans cette tranche et leur répartition par type d'usage. Vérifier que ces observations présentent des distributions énergétiques cohérentes avec le reste de leur catégorie. Valider que l'impact de cette tolérance sur les futures analyses reste effectivement négligeable via des tests de sensibilité.
24	Electricity(kBtu), NaturalGas(kBtu), SteamUse(kBtu), SiteEnergyUse(kBtu)	Calculer la somme des trois sources énergétiques, puis calculer l'écart relatif absolu entre cette somme et SiteEnergyUse(kBtu) divisé par SiteEnergyUse(kBtu), enfin supprimer toutes les lignes où cet écart relatif est strictement supérieur à 0.10	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. 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. L'analyse de distribution a montré que quarante-cinq bâtiments présentent de tels écarts, se concentrant principalement sur les structures de petite à moyenne taille. Conserver ces observations reviendrait à accepter des valeurs de CO2 mathématiquement corrompues, ce qui viderait de sens l'objectif même de prédiction des émissions.	Vérifier qu'aucune observation ne présente un écart énergétique supérieur à dix pourcent après nettoyage. Documenter précisément les quarante-cinq 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. Contrôler que cette suppression n'introduit pas de biais sectoriel en vérifiant la représentativité des types de bâtiments restants.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
25	Electricity(kBtu), NaturalGas(kBtu), SteamUse(kBtu), SiteEnergyUse(kBtu)	Pour les lignes où l'écart énergétique relatif est compris entre 0.05 et 0.10 inclus, conserver ces observations malgré l'incohérence modérée	Les deux bâtiments présentant un écart énergétique compris entre cinq et dix pourcent constituent des cas limites qui méritent d'être préservés pour maintenir la diversité de l'échantillon. Un écart de cette amplitude, bien que supérieur à la tolérance technique stricte, peut résulter de sources d'énergie mineures non détaillées dans les colonnes principales (biomasse, énergie solaire thermique, cogénération) ou de particularités de comptage pour des bâtiments avec des systèmes énergétiques complexes. L'analyse du graphique de cohérence énergétique a montré que ces cas restent statistiquement acceptables pour un modèle de prédiction qui doit être capable de généraliser sur des données réelles imparfaites. Leur suppression réduirait excessivement la taille de l'échantillon sans gain majeur de précision, d'autant que leur ComplianceStatus est Compliant, attestant de la validation municipale de leur déclaration.	Confirmer que les deux observations dans la tranche cinq à dix pourcent ont été conservées. Documenter leurs caractéristiques spécifiques pour permettre une analyse post-modélisation de leur influence. Vérifier que leur conservation n'introduit pas d'outliers dans les distributions d'intensité énergétique. Marquer ces observations avec un flag de vigilance pour traçabilité et analyse de sensibilité ultérieure. Valider que leur ComplianceStatus est effectivement Compliant.
26	YearBuilt	Supprimer toutes les lignes où YearBuilt est strictement inférieur à 1900 ou strictement supérieur à 2015	L'année de construction constitue une variable prédictive majeure des émissions de CO2 car elle capture les évolutions des normes de construction, des matériaux isolants et des standards énergétiques. Seattle ayant été incorporée en tant que ville en 1869, toute date antérieure à 1900 serait suspecte et indiquerait probablement une erreur de saisie ou une confusion avec une autre information historique. De même, l'année de référence du dataset étant 2016 pour le reporting énergétique de l'année 2015, toute construction postérieure à 2015 constitue une impossibilité temporelle. L'analyse diagnostique a confirmé l'absence de tels cas dans le dataset, ce qui valide la qualité du reporting municipal sur cette dimension. Cette règle de validation reste néanmoins inscrite dans le pipeline pour garantir la robustesse du processus de nettoyage en cas de réapplication sur des millésimes ultérieurs ou sur des datasets fusionnés.	Vérifier qu'aucune observation ne présente une année de construction hors bornes après nettoyage. Contrôler que toutes les valeurs de YearBuilt sont comprises dans l'intervalle entre 1900 et 2015 inclus. Valider que cette vérification n'a effectivement éliminé aucune ligne dans le cas présent, confirmant ainsi la cohérence temporelle initiale des données. Documenter cette absence d'anomalie comme point positif de la qualité du dataset source.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
27	Toutes les variables transformées	Recalculer et vérifier la cohérence des variables dérivées après application de toutes les transformations précédentes	Après l'exécution séquentielle de toutes les opérations de nettoyage, il est indispensable de recalculer ou vérifier la cohérence des variables qui pourraient avoir été affectées indirectement par les transformations. Par exemple, si des flags ont été créés ou si des ratios intermédiaires ont été utilisés pour la prise de décision, il faut s'assurer qu'ils ne subsistent pas comme colonnes temporaires dans le dataset final. De même, certaines variables calculées dynamiquement durant le processus (comme le ratio usage sur total ou l'écart énergétique) doivent être supprimées si elles ne font pas partie du schéma de données définitif. Cette étape de nettoyage post-transformation garantit que le dataset final ne contient que les variables légitimes et exploitables pour la modélisation, sans artefacts techniques issus du pipeline de nettoyage.	Lister exhaustivement toutes les colonnes du dataset nettoyé et vérifier qu'aucune variable temporaire de calcul ne subsiste. Supprimer les colonnes de ratio intermédiaires, d'écarts calculés et de flags de traçabilité qui ont servi uniquement à la prise de décision. Confirmer que seules les variables du schéma initial plus les deux flags de feature engineering (IsZeroFloorReported et IsAggregatedCampus) et le flag Has_EnergyStarScore sont présents. Valider que les types de données de toutes les colonnes restantes sont cohérents avec leur nature.
28	Dataset complet	Générer un rapport de traçabilité exhaustif documentant toutes les transformations appliquées avec horodatage et comptages précis	La reproductibilité et l'auditabilité du processus de nettoyage nécessitent une documentation complète de chaque décision et de son impact quantitatif. Ce rapport doit inclure pour chaque type d'anomalie le nombre exact de lignes supprimées, le nombre de lignes conservées après arbitrage, les critères de décision appliqués et les justifications méthodologiques correspondantes. L'horodatage de chaque opération permet de reconstituer l'historique complet des transformations et de détecter d'éventuelles incohérences dans la séquence d'exécution. Les comptages précis garantissent la cohérence arithmétique entre le nombre de lignes initial, les suppressions successives et le nombre de lignes final. Cette traçabilité constitue également une exigence d'explicabilité vis-à-vis des parties prenantes du projet et permet de justifier les choix méthodologiques auprès des experts métier.	Vérifier que le rapport de traçabilité a été généré avec un horodatage valide. Contrôler que la somme des lignes supprimées pour chaque motif plus les lignes conservées correspond exactement au nombre de lignes initial. Valider que chaque type d'anomalie documenté dans le rapport correspond effectivement à une opération exécutée dans le pipeline. S'assurer que les pourcentages de suppression calculés sont arithmétiquement cohérents. Archiver ce rapport dans le répertoire dédié avec une nomenclature claire incluant la date et la version du pipeline.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
29	Dataset complet	Exporter le dataset nettoyé dans un fichier horodaté et exporter séparément le dataset des anomalies détectées avec leurs justifications	La séparation physique entre les données nettoyées exploitables pour la modélisation et les données écartées avec leur documentation constitue une bonne pratique de gestion de données. Le dataset nettoyé doit être exporté dans un format structuré avec un horodatage précis permettant de tracer quelle version a été utilisée pour quelle analyse ultérieure. Parallèlement, le dataset des anomalies doit conserver l'intégralité des lignes écartées enrichies des colonnes de justification créées durant le processus de nettoyage. Ce fichier d'anomalies servira de référence pour les analyses post-mortem, pour la validation des choix de nettoyage avec les experts métier et pour la confrontation avec les prédictions du modèle final. La conservation de ces données écartées permet également de revenir sur certaines décisions si de nouvelles informations métier venaient remettre en question les critères de suppression appliqués.	Confirmer que le fichier du dataset nettoyé a été créé avec un nom incluant un horodatage au format YYYYMMDD_HHMMSS. Vérifier que ce fichier contient exactement le nombre de lignes attendu après toutes les suppressions documentées. Valider que le fichier des anomalies contient toutes les lignes supprimées avec les colonnes de justification complètes. Contrôler que la somme des lignes des deux fichiers exportés correspond au nombre de lignes initial du dataset brut. S'assurer que les deux fichiers sont stockés dans les répertoires appropriés selon l'arborescence du projet.

Section 3

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
30	SiteEnergyUse(kBtu), SiteEnergyUseWN(kBtu), TotalGHGEmissions, SiteEUIWN(kBtu/sf), SourceEUI(kBtu/sf), SourceEUIWN(kBtu/sf), Electricity(kWh), Electricity(kBtu)	Appliquer une transformation logarithmique via np.log1p puis calculer le Z-score sur les valeurs transformées, identifier les observations où la valeur absolue du Z-score est strictement supérieure à 3, créer des variables binaires de flag pour chaque variable analysée	Les distributions de ces variables énergétiques présentent une asymétrie marquée avec de longues traînes qui violent les hypothèses de normalité requises pour les tests paramétriques classiques. La transformation logarithmique log1p compresse l'échelle des valeurs et rapproche les distributions d'une loi normale, rendant la détection par Z-score statistiquement valide. Le seuil de trois écarts-types est fondé sur la règle empirique des 68-95-99.7 pourcent selon laquelle environ 99.7 pourcent des observations d'une distribution normale se trouvent dans cet intervalle. Par conséquent, moins de 0.3 pourcent des points devraient dépasser ce seuil, ce qui en fait des candidats plausibles à l'étiquette d'outliers. Cette méthode permet d'isoler les anomalies globales indépendantes du contexte qui traduisent des erreurs de saisie ou des corruptions de données lors de l'export plutôt que des particularités métier légitimes.	Vérifier que la transformation logarithmique a été appliquée correctement avec np.log1p pour gérer les valeurs nulles. Calculer les Z-scores sur les données transformées et confirmer que le seuil de trois écarts-types a été appliqué symétriquement. Créer une variable binaire de flag pour chaque variable analysée avec une nomenclature cohérente. Documenter précisément le nombre d'outliers détectés pour chaque variable. Générer des visualisations (histogrammes log-transformés et scatter plots) montrant les outliers identifiés. Exporter un fichier CSV contenant toutes les lignes flaggées avec leurs valeurs originales et transformées.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
31	Variables flaggées par Z-score	Créer une variable compteur total_outlier_flags_zscore qui additionne tous les flags créés à l'étape précédente, puis identifier les observations où ce compteur est strictement supérieur ou égal à 3	L'analyse conjointe du tableau récapitulatif et du graphique à barres horizontales a révélé que les bâtiments présentant trois anomalies ou plus constituent des multi-outliers dont l'influence sur un modèle de régression pourrait être disproportionnée. Ces observations ne présentent pas une seule valeur extrême isolée mais un pattern systématique d'anomalies sur plusieurs dimensions énergétiques, ce qui suggère soit une défaillance globale du processus de reporting pour ces bâtiments, soit une spécificité métier tellement marquée qu'elle risque de biaiser l'apprentissage du modèle. La création d'un compteur agrégé permet de prioriser les cas critiques nécessitant une investigation manuelle approfondie avant décision de conservation ou suppression. Cette approche segmentée distingue les outliers univariés acceptables des cas multivariés problématiques.	Confirmer que le compteur total_outlier_flags_zscore a été créé correctement en additionnant tous les flags binaires. Identifier et extraire toutes les observations avec un score supérieur ou égal à trois. Générer un graphique à barres horizontales empilées montrant la composition des anomalies par bâtiment. Créer un tableau détaillé listant ces bâtiments avec leurs caractéristiques principales (PropertyName, PrimaryPropertyType, valeurs des variables énergétiques). Documenter le nombre exact de bâtiments dans cette catégorie critique.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
32	SiteEnergyUse(kBtu), SiteEnergyUseWN(kBtu)	Pour toutes les lignes où SiteEnergyUse(kBtu) est strictement supérieur à zéro ET SiteEnergyUseWN(kBtu) est strictement égal à zéro, remplacer la valeur de SiteEnergyUseWN(kBtu) par la valeur de SiteEnergyUse(kBtu)	L'analyse du graphique de dispersion a révélé un cluster spécifique de bâtiments formant une ligne horizontale à y égale zéro avec des abscisses élevées, indiquant une consommation réelle significative mais une valeur normalisée météorologiquement nulle. Cette configuration est physiquement impossible et traduit un échec systématique de l'algorithme de normalisation Weather Normalized pour ces lignes spécifiques. Le test d'hypothèse par type de bâtiment a confirmé que ces erreurs se concentrent principalement sur les catégories Office et Worship Facility. Plutôt que de supprimer ces observations qui contiennent par ailleurs des informations énergétiques réelles et exploitables, nous choisissons d'imputer la valeur normalisée par la valeur brute. 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.	Identifier précisément les lignes concernées par cette imputation en appliquant le double critère. Documenter le nombre de lignes modifiées et leur répartition par PrimaryPropertyType. Vérifier que les valeurs de SiteEnergyUseWN après imputation correspondent exactement aux valeurs de SiteEnergyUse pour les lignes traitées. Générer un graphique de dispersion avant et après correction pour visualiser la disparition de la ligne horizontale problématique. Valider que cette imputation n'a pas créé de nouvelles incohérences avec d'autres variables dérivées.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
33	**SiteEUIWN(kBtu/sf)**, **SourceEUIWN(kBtu/sf)**	Recalculer les intensités énergétiques en divisant les consommations normalisées corrigées par PropertyGFABuilding(s).	La correction des consommations Weather Normalized à l'étape précédente a rendu 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. 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
34	Variables flaggées par Z-score	Examiner manuellement chaque bâtiment avec total_outlier_flags_zscore supérieur ou égal à 3 en consultant les variables PropertyName, PrimaryPropertyType, LargestPropertyUseType, YearBuilt, NumberofBuildings et toutes les valeurs énergétiques pour déterminer si l'anomalie est légitime (Data Center, Laboratory) ou erronée (valeurs nulles post-correction)	L'analyse typologique a révélé que les multi-outliers ne constituent pas une catégorie homogène mais trois profils distincts nécessitant des traitements différenciés. Le premier profil concerne les infrastructures à très haute intensité énergétique comme le Westin Building et UW Tower qui sont des Data Centers dont la consommation massive est cohérente avec leur usage métier et doit être conservée pour que le modèle apprenne à gérer ces cas extrêmes mais réels. Le deuxième profil rassemble des bâtiments présentant systématiquement des valeurs nulles pour les variables Weather Normalized malgré la correction appliquée, suggérant une défaillance plus profonde nécessitant une suppression. Le troisième profil comprend des entrepôts avec des consommations extrêmement basses mais réelles situées dans la queue inférieure de la distribution et devant être conservés. Cette investigation manuelle permet d'appliquer une décision contextuelle adaptée à chaque cas plutôt qu'une règle aveugle de suppression systématique.	Créer un tableau d'audit listant les bâtiments concernés avec une colonne de décision (Conserver, Supprimer, Investiguer). Documenter pour chaque bâtiment la justification de la décision prise. Valider que les Data Centers et Laboratories identifiés ont été conservés sauf erreur manifeste. Confirmer que les bâtiments présentant des valeurs nulles résiduelles après correction ont été supprimés. Vérifier que les bâtiments à faible consommation mais cohérents avec leur usage ont été maintenus. Exporter ce tableau d'audit pour traçabilité.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
35	NumberofFloors, GHGEmissionsIntensity, SiteEUI(kBtu/sf), NaturalGas(therms), NaturalGas(kBtu), SteamUse(kBtu), PropertyGFATotal, LargestPropertyUseTypeGFA, SecondLargestPropertyUseTypeGFA, ENERGYSTARScore	Appliquer la détection IQR segmentée par PrimaryPropertyType avec un multiplicateur de 3.0 pour identifier les observations hors bornes et créer un flag binaire spécifique par variable.	Ces variables présentent des distributions pour lesquelles la transformation logarithmique est inefficace. Le seuil de 3.0 cible uniquement les anomalies critiques tout en préservant la diversité naturelle du parc immobilier indispensable pour que le modèle apprenne les relations de proportionnalité.	Vérifier la segmentation par usage et le calcul des quartiles par groupe. Documenter les statistiques calculées dans un tableau récapitulatif et générer les visualisations combinant boxplot et stripplot par type de bâtiment.	
36	NumberofFloors	Exclure de la détection IQR toutes les observations où la valeur est égale à zéro ou manquante avant d'appliquer l'algorithme sur le sous-ensemble filtré.	Les valeurs zéro constituent une catégorie à part traitée par feature engineering. Les exclure du calcul des quartiles évite de fausser les bornes et permet d'identifier précisément les gratte-ciels exceptionnels dont la signature énergétique est cohérente.	Créer un masque filtrant les valeurs positives. Vérifier que les zéros et données manquantes ne sont pas traités comme des outliers et valider que les détections correspondent effectivement à des bâtiments de grande hauteur.	
37	Variables flaggées par IQR	La création d'un compteur agrégé pour les flags IQR n'est plus la priorité face à la validation croisée mais sert d'indicateur de cohérence multidimensionnelle.	Contrairement au Z-score, les flags IQR capturent des incohérences sur les intensités relatives et les caractéristiques structurelles. Un cumul de flags permet de distinguer les anomalies isolées des erreurs de déclaration systématiques.	Calculer la distribution du compteur pour identifier les bâtiments avec des scores d'anomalies élevés. Documenter le nombre de bâtiments par tranche de score pour prioriser l'examen des cas multidimensionnels.	
38	PropertyGFATotal, LargestPropertyUseTypeGFA, SecondLargestPropertyUseTypeGFA	Conserver systématiquement tous les outliers de surface supérieurs car ils correspondent à des structures massives comme les universités ou les hôpitaux.	La surface est le prédicteur numéro un du volume de CO2. Conserver les mégastructures est vital pour apprendre les économies d'échelle et éviter de tronquer artificiellement la distribution des tailles de bâtiments dans le modèle.	Identifier les outliers supérieurs pour les variables de surface. Croiser ces identifications avec l'usage du bâtiment pour confirmer la cohérence et documenter leur maintien malgré leur statut statistique.	

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
39 SiteEUI(kBtu/sf), GHGEmissionsIntensity	Supprimer les outliers d'intensité uniquement s'ils ne sont pas justifiés par une surface hors-norme via une logique de validation croisée.	Une intensité délirante sur une surface normale signale une défaillance technique ou une erreur de compteur incompatible avec la thermodynamique. La suppression assainit la variable cible et stabilise les gradients du futur modèle.	Identifier les suspects d'intensité. Appliquer le filtre de suppression si l'outlier n'est pas corrélé à une surface massive. Valider que les valeurs résiduelles respectent des ordres de grandeur physiquement plausibles.	
40 NaturalGas(therms), NaturalGas(kBtu), SteamUse(kBtu)	Conserver l'intégralité des outliers de consommation car ils représentent une hétérogénéité thermique réelle et structurelle du parc.	Supprimer ces points reviendrait à ignorer les bâtiments ayant le plus fort impact sur le bilan carbone global. Ces valeurs extrêmes sont des signaux informatifs que les modèles d'arbres sauront isoler et capturer.	Confirmer la conservation des outliers sur les flux de gaz et de vapeur. Documenter leur contribution au total des émissions pour quantifier leur importance et maintenir la représentativité sur tout le spectre énergétique.	
41 ENERGYSTARScore	Conserver l'intégralité des valeurs incluant les rares outliers car cette variable bornée constitue un signal de performance pur et fiable.	Avec seulement 0.28 pourcent d'outliers, cette variable présente une qualité exceptionnelle. Étant bornée entre 0 et 100, les valeurs extrêmes restent plausibles et ne risquent pas de corrompre les calculs du modèle.	Vérifier que toutes les valeurs du score ont été maintenues sans suppression. Valider la cohérence de la variable comme prédicteur de l'efficacité énergétique pour la prédiction finale des émissions de CO2.	

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
42	DefaultData	Vérifier l'absence de valeurs par défaut dans le dataset en comptant les modalités de cette variable administrative	La variable DefaultData recense les cas où les données énergétiques déclarées ont été remplacées par des valeurs par défaut calculées par la municipalité en l'absence de reporting complet. L'utilisation de données par défaut introduit un biais systématique car ces valeurs sont estimées selon des moyennes sectorielles et ne reflètent pas la performance réelle du bâtiment. Leur présence corromprait la variable cible CO2 en mélangeant des mesures réelles avec des estimations statistiques. L'analyse de cette variable permet de valider que le dataset ne contient que des données effectivement mesurées et déclarées, garantissant ainsi la fiabilité de la relation empirique entre prédicteurs et émissions que le modèle cherchera à apprendre. L'absence de valeurs par défaut constitue un indicateur de qualité majeur attestant de la complétude du processus de reporting énergétique pour les bâtiments du dataset.	Extraire et afficher toutes les modalités distinctes de la variable DefaultData. Compter le nombre d'occurrences pour chaque modalité. Vérifier qu'aucune modalité n'indique l'utilisation de données par défaut. Documenter cette absence comme point positif de qualité du dataset. Si des valeurs par défaut sont détectées, les quantifier par type de bâtiment et déterminer une stratégie de traitement appropriée. Confirmer que la totalité des données énergétiques proviennent de déclarations réelles et non d'estimations.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
43	Toutes les lignes du dataset	Identifier les doublons exacts en comparant toutes les colonnes, identifier les doublons sans ID en excluant OSEBuildingID de la comparaison, identifier les doublons d'adresse en utilisant la colonne Address	L'identification des doublons permet de détecter d'éventuelles erreurs de saisie répétées ou des problèmes d'import de données. Les doublons exacts indiquent une duplication complète de ligne qui n'a aucune justification métier. Les doublons sans ID révèlent des observations identiques sur toutes les caractéristiques mais avec des identifiants différents, suggérant une double saisie du même bâtiment. Les doublons d'adresse sont plus ambigus car une même adresse peut légitimement héberger plusieurs bâtiments distincts dans le cas de campus ou de parcelles multi-structures. L'analyse diagnostique doit donc distinguer les véritables doublons erronés des cas de colocalisation physique de bâtiments indépendants. Cette vérification garantit l'unicité des observations au niveau approprié et évite le sur-apprentissage qui résulterait de la présence de lignes redondantes dans le dataset d'entraînement.	Calculer et documenter le nombre de doublons exacts trouvés. Calculer le nombre de doublons sans ID. Pour les doublons d'adresse, générer un tableau d'audit avec comptage des entrées par adresse, nombre de surfaces distinctes, nombre de types de propriété distincts et écart-type des émissions. Analyser manuellement les adresses présentant plusieurs entrées pour distinguer bâtiments distincts et véritables doublons. Supprimer uniquement les doublons avérés. Conserver les entrées multiples correspondant à des bâtiments physiquement distincts. Documenter la décision pour chaque adresse concernée.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
44	Address	Pour chaque adresse présentant plusieurs entrées, calculer le nombre de surfaces distinctes PropertyGFATotal, le nombre de types de propriété distincts PrimaryPropertyType et l'écart-type des émissions TotalGHGEmissions pour déterminer si les entrées correspondent à des bâtiments distincts ou à des doublons	L'audit d'unicité sur les vingt-neuf cas de doublons d'adresse nécessite une analyse granulaire pour éviter de supprimer à tort des observations légitimes. La présence de surfaces distinctes et de types de propriété variés constitue une preuve forte de l'existence de plusieurs bâtiments physiquement indépendants sur une même parcelle ou adresse administrative. Les écarts-types d'émissions élevés confirment que chaque bâtiment possède sa propre signature énergétique distincte. À l'inverse, une adresse avec plusieurs entrées présentant une seule surface et un seul type de propriété avec un écart-type d'émissions quasi nul suggère un doublon de déclaration pour un même volume physique. Cette analyse permet d'appliquer une décision contextuelle adaptée à chaque situation plutôt qu'une règle aveugle de suppression ou conservation systématique des adresses multiples.	Grouper le dataset par adresse et calculer les agrégations spécifiées. Générer un tableau d'audit avec une ligne par adresse concernée. Identifier les cas où surfaces et types sont distincts (conservation systématique). Identifier les cas où surface et type sont identiques (investigation approfondie). Pour le cas spécifique identifié (100 West Harrison), vérifier les émissions et autres caractéristiques pour décision finale. Documenter la décision pour chaque adresse avec justification. Confirmer que les bâtiments distincts sont maintenus intégralement. Valider que chaque OSEBuildingID unique correspond à une entité fiscale ou énergétique propre.

Ordre	Variable cible	Action technique	Justification méthodologique	Validation attendue
45	Dataset complet après toutes les transformations	Générer un rapport de synthèse final documentant le nombre de lignes initial, le nombre de lignes supprimées par catégorie de nettoyage (sections 1, 2 et 3), le nombre de lignes conservées, le nombre de valeurs imputées par variable, le nombre de variables créées par feature engineering et la liste exhaustive des variables disponibles pour la modélisation	Après l'exécution complète du pipeline de nettoyage couvrant les six sections d'audit, il est indispensable de produire un bilan quantitatif consolidé permettant de mesurer l'impact global des opérations de nettoyage sur la structure et la composition du dataset. Ce rapport de synthèse constitue le pont entre la phase de préparation des données et la phase de modélisation en fournissant aux data scientists les métriques clés sur la qualité et la complétude du dataset nettoyé. Il permet également de vérifier la cohérence arithmétique entre toutes les étapes de transformation et de détecter d'éventuelles incohérences dans le pipeline. Ce document servira de référence pour justifier auprès des parties prenantes les choix de nettoyage et pour contextualiser les performances ultérieures des modèles prédictifs.	Calculer le nombre de lignes du dataset initial avant tout nettoyage. Comptabiliser précisément les suppressions de la section 1 (restriction non-résidentiels, ComplianceStatus). Comptabiliser les suppressions de la section 2 (incohérences physiques, ratios, cohérence énergétique). Comptabiliser les suppressions de la section 3 (outliers Z-score et IQR après investigation). Calculer le taux de rétention global du dataset. Lister toutes les variables créées par feature engineering avec leur définition. Documenter le nombre de valeurs imputées pour chaque variable concernée. Générer un schéma visuel résumant le flux de nettoyage. Exporter ce rapport dans un format structuré avec horodatage pour archivage.