École nationale de la Statistique et de l'Analyse économique – Pierre Ndiaye

Machine Learning 1

Prédiction de l'émission du CO_2 dans la ville de Seattle

Rédigé par :

Moussa DIEME
Ben Idriss SOMA
Papa Ahmadou NIANG
Prosper LAWA FOUMSOU
Tamsir NDONG

Élèves en ISE 2

Sous la supervision de :

Mme Mously DIAW
Enseignante à l'ENSAE

Décembre 2024

Table des matières

Liste des tableaux

Table des figures

Introduction

Contexte et justification

Le changement climatique constitue l'un des principaux défis contemporains, et les émissions de dioxyde de carbone (CO_2) issues des activités humaines en sont un facteur déterminant. Dans les zones urbaines, les bâtiments résidentiels, commerciaux et industriels représentent une part importante de la consommation énergétique et des émissions de gaz à effet de serre (GES). La ville de Seattle, située dans l'État de Washington aux États-Unis, s'est engagée à lutter contre le changement climatique en réduisant ses émissions de GES et en atteignant la neutralité carbone d'ici 2050. Pour suivre ses progrès et orienter ses politiques publiques, Seattle produit régulièrement un inventaire des émissions de GES (Greenhouse Gas Inventory), mis à jour tous les deux ans, couvrant les principaux secteurs : transports, bâtiments et déchets.

Le secteur du bâtiment est responsable de près d'un tiers des émissions mondiales de CO_2. À Seattle, il représente environ 40 % des émissions locales, toutes catégories confondues, juste derrière le secteur des transports. Afin de favoriser l'analyse et la transparence, la ville met à disposition des données ouvertes issues du Seattle Building Energy Benchmarking, qui décrivent les caractéristiques des bâtiments (surface, usage, année de construction), leur consommation énergétique et leurs émissions de CO_2.

Dans ce contexte, le présent projet vise à exploiter les techniques de machine learning pour prédire les émissions de CO_2 des bâtiments non destinés à l'habitation de Seattle à partir de leurs caractéristiques. Un tel modèle pourra servir d'outil d'aide à la décision pour les autorités publiques et les urbanistes, en identifiant les facteurs clés d'émissions et en orientant les politiques de réduction carbone.

Problématique

Les méthodes traditionnelles d'évaluation des émissions de CO_2 reposent sur des données mesurées, telles que les relevés de compteurs énergétiques ou les factures de consommation. Bien que précises, ces informations ne sont disponibles qu'après la période de consommation et ne permettent pas d'anticiper les émissions futures d'un bâtiment, no-

tamment lors de projets de construction ou de rénovation.

Par ailleurs, la réalisation d'audits énergétiques exhaustifs est coûteuse et difficilement généralisable à l'ensemble du parc immobilier. Dans ce contexte, la capacité à prédire les émissions de CO_2 à partir de caractéristiques structurelles observables constitue un enjeu majeur pour l'aide à la décision publique. Il apparaît donc logique de se poser les questions suivantes : peut-on prédire avec précision les émissions de CO_2 des bâtiments non résidentiels de la ville de Seattle à partir de données déclaratives et structurelles ? Peut-on se passer de certaines variables coûteuses à calculer, comme le **ENERGY STAR Score** ?

Objectifs du projet

L'objectif est de développer un modèle de Machine Learning capable de prédire les émissions de CO_2 des bâtiments non destinés à l'habitation. Pour répondre aux enjeux stratégiques susmentionnés, l'équipe technique se fixe quatre objectifs opérationnels précis, mesurables et temporellement définis.

1. **Modélisation Prédictive** : construire un modèle robuste capable d'estimer les émissions de gaz à effet de serre (totales ou par unité de surface) en se basant sur les caractéristiques structurelles et déclaratives du bâtiment, sans avoir nécessairement accès aux relevés de compteurs détaillés au moment de la prédiction. Cela permettrait d'identifier les bâtiments énergivores nécessitant une intervention prioritaire dans le cadre du **Building Emissions Performance Standard** (BEPS).
2. **Analyse de l'Energy Star Score** : évaluer rigoureusement l'impact de cet indicateur sur la précision prédictive en comparant systématiquement les performances de modèles incluant versus excluant cette variable. L'objectif est de quantifier précisément le compromis entre simplicité opérationnelle (modèle sans Energy Star Score) et gain prédictif potentiel.
3. **Interprétabilité des Résultats** : identifier et hiérarchiser les variables explicatives qui influencent le plus la consommation via des méthodes d'explicabilité globale et locale, afin de fournir des leviers d'action clairs aux décideurs.
4. **Industrialisation du Pipeline** : aboutir à une architecture logicielle reproductible, versionnable et maintenable. Cela implique l'automatisation de l'ensemble de la chaîne de traitement (ingestion des données, preprocessing, entraînement, évaluation, déploiement) selon les principes MLOps.

Dans la suite de ce document, nous présenterons en premier lieu la revue de littérature justifiant le choix des principales variables explicatives de notre variable cible, en second lieu, nous présenterons les données utilisées dans le cadre de notre travail puis en dernier lieu nous aborderons la méthodologie adoptée.

Chapitre 1

Revue de la littérature

La littérature consacrée à la performance énergétique et environnementale des bâtiments met en évidence que les émissions de dioxyde de carbone (CO_2) résultent d'un ensemble de facteurs structurels, fonctionnels, énergétiques, technologiques et contextuels.

En premier lieu, les variables morphologiques et structurelles apparaissent comme des déterminants majeurs des émissions. La surface du bâtiment (Gross Floor Area) est systématiquement identifiée comme le prédicteur le plus robuste et le plus stable des consommations énergétiques et des émissions associées, en raison de son lien direct avec les besoins en chauffage, climatisation et éclairage (Perez-Lombard, Ortiz et Pout, 2008 ; IEA, 2019). De nombreuses études empiriques montrent que la relation entre surface et émissions est quasi proportionnelle, bien que des effets de non-linéarité puissent apparaître pour les bâtiments de très grande taille (Hong, Taylor-Lange, D'Oca, Yan et Corgnati, 2015). Aussi, l'année de construction et, le cas échéant, l'année de rénovation, jouent également un rôle significatif : les bâtiments récents tendent à présenter de meilleures performances énergétiques en raison de normes de construction plus strictes et de technologies plus efficaces (Filippín, 2000 ; IPCC, 2022).

Les variables fonctionnelles liées à l'usage du bâtiment constituent un second groupe fondamental de prédicteurs. La littérature souligne que le type d'usage (bureaux, établissements scolaires, hôpitaux, commerces, entrepôts) influence fortement le niveau d'émissions, indépendamment de la taille du bâtiment (Santamouris et al., 2010). À surface équivalente, les bâtiments à usage intensif, tels que les hôpitaux ou les centres de données, présentent des émissions nettement supérieures à celles des bâtiments administratifs ou logistiques, en raison d'horaires d'occupation prolongés, d'équipements spécifiques et de besoins énergétiques continus (Chung, Rhee et Im, 2011). Ces résultats justifient l'intégration de variables catégorielles décrivant l'usage principal et secondaire dans les modèles de prédiction.

Par ailleurs, les variables énergétiques constituent le lien le plus direct entre les caractéristiques du bâtiment et les émissions de CO_2. La consommation d'énergie, qu'il s'agisse d'électricité, de gaz naturel ou d'autres combustibles, est un déterminant cen-

tral, les émissions étant proportionnelles à la quantité d'énergie consommée et au facteur d'émission associé à chaque vecteur énergétique (IEA, 2021). Plusieurs études insistent sur l'importance de distinguer les différentes sources d'énergie, notamment dans les régions où le mix électrique est relativement décarboné, ce qui renforce le rôle explicatif de la consommation de combustibles fossiles comme le gaz naturel (Gurney et al., 2012). La littérature recommande également l'utilisation de variables dérivées, telles que l'intensité énergétique (Energy Use Intensity, EUI), définie comme le ratio entre la consommation totale d'énergie et la surface du bâtiment, afin de neutraliser l'effet de taille et de faciliter la comparaison entre bâtiments hétérogènes (Chung, 2011 ; Hong et al., 2015).

À un niveau plus agrégé, des indicateurs synthétiques de performance énergétique, tels que le ENERGY STAR Score, sont largement utilisés dans les démarches de benchmarking énergétique, en particulier aux États-Unis (U.S. EPA, 2018). Toutefois, plusieurs travaux montrent que ces scores apportent une information marginale limitée dans les modèles prédictifs une fois les variables de consommation, de surface et d'usage incluses, en raison de leur construction basée sur ces mêmes informations et des problèmes de colinéarité qui en résultent (Kontokosta, 2012 ; Robinson et al., 2017). À Seattle, l'électricité est majoritairement d'origine hydroélectrique, ce qui entraîne un facteur d'émission de CO_2 très faible. Ainsi, une consommation électrique élevée n'implique pas nécessairement des émissions importantes, contrairement aux sources énergétiques fossiles comme le gaz naturel.

Enfin, la littérature souligne le rôle des variables technologiques et contextuelles, telles que les systèmes de chauffage, ventilation et climatisation (CVC), la qualité de l'enveloppe thermique, ainsi que les conditions climatiques locales, dans l'explication des émissions de CO_2 des bâtiments (Santamouris, 2015 ; IPCC, 2022). Bien que ces facteurs soient parfois observés de manière imparfaite dans les bases de données administratives, ils contribuent à expliquer les différences résiduelles de performance énergétique entre bâtiments comparables. L'ensemble de ces résultats justifie très souvent le recours à des modèles de machine learning non linéaires, capables de capturer les interactions complexes entre la taille, l'usage, l'énergie et la technologie, pour la prédiction des émissions de CO_2 des bâtiments non résidentiels.

Chapitre 2

Contexte et périmètre

2.1 Contexte

L'étude porte sur les bâtiments non résidentiels de la ville de Seattle (bureaux, établissements scolaires, hôpitaux, bâtiments commerciaux et industriels). Ces bâtiments présentent une forte hétérogénéité de consommation énergétique et constituent une cible prioritaire des politiques publiques de décarbonation. Leur diversité structurelle et fonctionnelle en fait un terrain d'analyse riche pour la modélisation énergétique et environnementale.

2.2 Source des données

Les données utilisées sont issues du **Seattle Building Energy Benchmarking Dataset** pour l'année 2016, mis à disposition en open data par la ville de Seattle. Elles proviennent des déclarations administratives liées aux permis d'exploitation commerciale des bâtiments non résidentiels. Annuellement fournies par les propriétaires ou exploitants, ces données décrivent les caractéristiques physiques, l'usage et les consommations énergétiques des bâtiments.

Bien qu'elles ne reposent pas sur des mesures en temps réel, elles constituent une source standardisée et largement utilisée pour le suivi de la performance énergétique et environnementale du parc immobilier. Dans le cadre de la Phase 0 du projet, ce jeu de données a été ingéré, audité et validé, confirmant son intégrité et sa viabilité pour les phases suivantes.

2.3 Variable cible

La variable cible retenue pour la modélisation est le **TotalGHGEmissions** (niveau total annuel d'émissions de gaz à effet de serre, en tonnes de CO_2 équivalent). Ce choix

se justifie par sa centralité dans les objectifs de neutralité carbone des collectivités, sa disponibilité homogène dans le jeu de données, et sa capacité à synthétiser l'impact environnemental direct et indirect des consommations énergétiques.

Une seconde variable, **GHGEmissionsIntensity**, a été analysée mais écartée comme cible principale en raison de sa corrélation modérée (0,47) avec TotalGHGEmissions et du risque d'instabilité qu'elle introduirait dans les modèles, car elle combine déjà des émissions totales avec d'autres facteurs tels que la surface. La stratégie retenue privilégie donc des modèles séparés pour chaque cible, garantissant une meilleure interprétabilité.

2.4 Variables explicatives

Conformément à la littérature et aux bonnes pratiques identifiées en Phase 0, les variables explicatives sont structurées en catégories claires :

- **Morphologiques et structurelles** : surface totale (Gross Floor Area), nombre d'étages, année de construction/rénovation.
- **Fonctionnelles** : usage principal et secondaire (bureaux, commerces, santé, etc.), reflétant l'intensité d'occupation et les besoins énergétiques.
- **Énergétiques** : consommations annuelles d'électricité et de gaz, intensité énergétique (EUI).
- **Indicateurs de performance** : score ENERGY STAR (intégré de manière optionnelle pour évaluer son apport marginal).

Le diagnostic initial a par ailleurs identifié des redondances (ex. doublons d'unités, surfaces totales vs partielles) et des données manquantes concentrées sur les usages secondaires/tertiaires et les certifications ENERGY STAR. Ces points feront l'objet d'un traitement spécifique lors de la Phase 1 (nettoyage des données).

Résumé de la section

Cette section définit les bases solides et cohérentes sur lesquelles s'appuie l'étude : un périmètre ciblé, des données fiables et auditées, une variable cible pertinente au regard des enjeux climatiques, et un cadre explicatif structuré. La Phase 0 a validé la robustesse de cette approche et prépare le terrain pour les phases de nettoyage, d'analyse et de modélisation à venir.

Chapitre 3

Dictionnaire des variables issues du nettoyage des données

3.1 Principes fondamentaux : le rôle des variables flag

Une variable flag est une variable binaire explicitement créée pour signaler une situation particulière au sein du jeu de données, sans représenter directement une caractéristique physique ou économique du phénomène étudié. Son rôle principal est de tracer des anomalies, conserver la mémoire des transformations appliquées, distinguer des cas structurellement différents et contrôler les biais liés à la qualité des données. Ces variables sont employées à trois niveaux : l'audit de la qualité des données, l'intégration comme variables de contrôle dans les modèles, et la définition de critères d'exclusion ou de stratification lors des analyses. Il est essentiel de ne pas les interpréter comme des variables structurelles, même lorsqu'elles sont incluses dans un modèle statistique.

3.2 Variables créées lors du nettoyage

3.2.1 IsZeroFloorReported

Il s'agit d'une variable binaire signalant une déclaration invalide ou manquante du nombre d'étages. Une valeur de 1 indique que le champ `NumberofFloors` était inférieur ou égal à zéro dans les données sources, ce qui signale une incohérence ou une absence d'information, et non un bâtiment sans étage. Utilisée en phase de nettoyage pour identifier les observations nécessitant une correction, elle sert également en analyse exploratoire pour évaluer si les bâtiments avec une qualité déclarative médiocre affichent un profil énergétique atypique. Dans un modèle, elle peut être introduite comme variable de contrôle pour absorber un effet lié à la qualité des données, sans pour autant faire l'objet d'une interprétation causale directe.

3.2.2 IsAggregatedCampus

Cette variable binaire identifie les entités de type « Campus », qui agrègent potentiellement plusieurs bâtiments hétérogènes. Une valeur de 1 correspond aux observations où `BuildingType = "Campus"`. Elle permet d'appliquer un traitement différencié lors du nettoyage, évitant l'usage de règles conçues pour des bâtiments unitaires. En analyse, elle autorise des comparaisons séparées entre bâtiments individuels et entités agrégées. Dans la modélisation, elle sert de variable de contrôle ou de segmentation pour capter les effets d'échelle et de structure spécifiques à ces regroupements.

3.2.3 IsMixedUse

Variable binaire indiquant la présence d'un usage secondaire déclaré. Elle prend la valeur 1 lorsque `SecondLargestPropertyUseType` est renseigné, signalant ainsi une mixité d'usages. Cette variable permet de préserver l'information structurelle sur la mixité tout en évitant de conserver les colonnes originales, qui présentaient un taux élevé de valeurs manquantes. Elle est utile pour segmenter les bâtiments et analyser l'impact de la complexité d'usage sur la consommation énergétique, et peut être combinée avec la variable d'usage principal pour des analyses plus fines.

3.2.4 Has_EnergyStarScore

Cette variable flag indique si le score ENERGY STAR était originellement présent dans les données. Une valeur de 1 signifie que `ENERGYSTARScore` était renseigné. Elle capture un motif de valeurs manquantes potentiellement informatif, lié à un mécanisme de manquants à random conditionné par le type de bâtiment. Créée avant toute imputation, elle permet d'évaluer si l'absence de score constitue un signal en soi et peut être utilisée comme variable de contrôle pour tenir compte des biais liés à la complétude déclarative.

3.2.5 Synthèse des nouvelles variables

TABLE 3.1 – Synthèse des variables flag créées lors du nettoyage

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

3.3 Variables supprimées lors du nettoyage

Plusieurs variables ont été exclues du jeu de données nettoyé afin d'améliorer la stabilité des analyses et la lisibilité des modèles. Les suppressions concernent principalement trois catégories : les variables présentant un taux de valeurs manquantes extrême, les variables redondantes ou à faible pouvoir discriminatif, et les variables temporaires créées pour le diagnostic. Ces décisions, prises pour réduire le bruit statistique et les risques d'instabilité, n'ont pas entraîné de perte substantielle d'information.

TABLE 3.2 – Variables supprimées lors du nettoyage

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 %** ; variable temporaire de diagnostic, 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` ; suppression pour éviter lacunes.	7
`SecondLargestPropertyUseTypeGFA`	Taux de valeurs manquantes de **47,75 %** ; même justification que ci-dessus.	7

3.4 Positionnement méthodologique

Les variables introduites lors de cette phase s'inscrivent dans une logique de traçabilité et de contrôle méthodologique. Leur objectif principal est de garantir que les décisions de nettoyage, d'imputation et d'exclusion demeurent observables, justifiables et reproductibles. Elles forment un socle indispensable pour documenter les choix de préparation des données, sécuriser l'interprétation des résultats et renforcer la crédibilité statistique et scientifique de l'analyse dans son ensemble.

Résumé de la section

Cette phase de nettoyage a permis de **structurer, documenter et sécuriser** le jeu de données grâce à la création de variables flag stratégiques et à l'élimination de variables non informatives. Ces actions préparent un terrain robuste pour les étapes ultérieures d'analyse exploratoire et de modélisation.

Chapitre 4

Analyse univariée approfondie : Décryptage des émissions de CO à Seattle

4.1 Objectif du Notebook

Ce notebook constitue la pierre angulaire analytique du projet. Il a pour vocation de comprendre le comportement individuel et intrinsèque de chaque variable, de détecter les anomalies statistiques et les valeurs aberrantes, et de préparer et justifier les transformations nécessaires avant toute phase de modélisation en Machine Learning.

4.2 Vue synthétique du jeu de données

Le jeu de données est caractérisé par une typologie et une répartition variables. On y trouve en majorité des variables numériques (float/int) qui forment l'ossature quantitative du dataset. Les variables qualitatives (object/category) sont minoritaires mais structurellement significatives, tandis que les variables booléennes sont marginales et peu présentes. Cette prédominance des variables quantitatives légitime une approche statistique rigoureuse et détaillée.

4.3 Analyse de la variable cible : TotalGHGEmissions

Les statistiques descriptives fondamentales révèlent les caractéristiques suivantes :

La distribution quartilique se décompose ainsi :

L'interprétation clé de cette analyse est que la distribution est extrêmement asymétrique à droite. Cela traduit la présence de quelques bâtiments aux émissions démesurées qui influencent l'ensemble des indicateurs. Par conséquent, une transformation logarithmique est impérative pour normaliser la distribution avant modélisation.

TABLE 4.1 – Statistiques descriptives de la variable `TotalGHGEmissions`

Statistique	**Valeur**
Moyenne	138.76
Médiane	46.01
Minimum	0.40
Maximum	12 307.16
Écart-type	540.27
Skewness	16.77
Kurtosis	329.03

TABLE 4.2 – Distribution quartilique de `TotalGHGEmissions`

Quartile	**Valeur**
Q1 (25 %)	19.54
Q2 (50 %)	46.01
Q3 (75 %)	120.88
IQR	101.35

4.4 Variables de consommation énergétique

Les variables étudiées incluent l'électricité, le gaz naturel, la vapeur, la consommation totale du site et les intensités énergétiques (EUI). La synthèse des comportements observés est présentée ci-dessous :

TABLE 4.3 – Synthèse des observations sur les variables de consommation énergétique

Caractéristique	**Observation**
Forme des distributions	Très asymétriques à droite
Dispersion	Exceptionnellement élevée
Valeurs extrêmes	Nombreuses et influentes
Corrélations WN vs non-WN	Très fortes, suggérant une redondance

En conclusion, les variables normalisées météo (WN) n'apportent pas d'information nouvelle et pourront être simplifiées.

4.5 Caractéristiques physiques des bâtiments

Les variables étudiées sont PropertyGFATotal, PropertyGFABuilding(s), PropertyGFAParking, NumberofFloors et NumberofBuildings. Les observations structurantes qui en découlent sont les suivantes :

L'insight principal est que la surface et la hauteur des bâtiments sont des déterminants partiels mais significatifs des écarts énergétiques observés.

TABLE 4.4 – Patterns observés sur les caractéristiques physiques des bâtiments

Variable	**Pattern observé**
Surfaces	Distribution asymétrique à droite
Taille typique	Majoritairement petite à moyenne
Très grands bâtiments	Rares mais dominants en impact
Nombre d'étages	Concentré entre 1 et 5 étages

4.6 Dimension temporelle : YearBuilt

Les observations clés concernant l'année de construction sont résumées dans le tableau suivant :

TABLE 4.5 – Observations sur la variable `YearBuilt`

Élément	**Observation**
Étendue temporelle	Large, couvrant plusieurs décennies
Bâtiments anciens	Proportion notable dans le parc
Homogénéité	Faible, reflétant une diversité d'âges

L'interprétation de ces éléments est que l'âge du bâtiment influence la performance énergétique, mais ne suffit pas à expliquer à lui seul les niveaux d'émissions.

4.7 Performance énergétique : ENERGYSTARScore

Les statistiques générales du score ENERGYSTAR sont les suivantes :

TABLE 4.6 – Statistiques descriptives du `ENERGYSTARScore`

Élément	**Observation**
Domaine	[0 ; 100]
Dispersion	Modérée
Valeurs manquantes	Nombreuses (NaN)
Normalité	Non vérifiée

Il est important de noter que le score est une variable informative, mais qu'il nécessite une gestion rigoureuse des données manquantes avant utilisation.

4.8 Analyse des variables qualitatives

Les variables étudiées sont BuildingType, PrimaryPropertyType, Neighborhood, LargestPropertyUseType et ListOfAllPropertyUseTypes. Les patterns dominants observés sont les suivants :

TABLE 4.7 – Patterns principaux des variables qualitatives

Variable	**Pattern principal**
BuildingType	Émissions très variables selon la catégorie
PrimaryPropertyType	Tertiaire majoritaire
Neighborhood	Concentration géographique marquée
PropertyUseTypes	Forte tendance aux usages multiples

En conclusion, ces variables qualitatives sont hautement structurantes et devront être intégrées avec soin dans les modèles.

4.9 Recommandations issues de l'analyse univariée

Les problèmes détectés et les actions recommandées sont synthétisés comme suit :

TABLE 4.8 – Recommandations issues de l'analyse univariée

Problème détecté	**Action recommandée**
Asymétrie forte	Application d'une transformation logarithmique
Valeurs extrêmes	Transformation ou robust scaling
Variables redondantes	Sélection et élimination des doublons
Modalités nombreuses	Regroupement des catégories peu fréquentes
Valeurs manquantes	Imputation raisonnée ou suppression ciblée

4.10 Rôle central du Notebook dans le projet

Ce document constitue une base méthodologique solide pour les étapes suivantes. Il sert également d'outil de diagnostic avancé des données et de support collaboratif standardisé pour l'équipe projet. Il est donc indispensable de le consulter avant toute analyse bivariée ou modélisation.

Résumé de la section

L'analyse univariée a permis une compréhension fine et individualisée des distributions, l'identification de patterns structurels explicatifs, et une préparation rigoureuse et éclairée pour la phase de Machine Learning. Elle représente la fondation essentielle et non négociable sur laquelle s'appuiera la suite du projet.

Chapitre 5

Analyse bivariée

5.1 Contexte et objectifs de l'analyse

Le Notebook 03 du projet Seattle Energy Benchmarking s'inscrit dans une démarche avancée d'analyse exploratoire des données (EDA), dont l'objectif principal est de passer d'une analyse descriptive à une analyse orientée modélisation prédictive. Ce notebook cherche à identifier les variables explicatives et prédictives les plus pertinentes pour la variable cible, **TotalGHGEmissions**, qui représente les émissions totales de gaz à effet de serre des bâtiments.

La finalité principale est de préparer le terrain pour une modélisation prédictive robuste en comprenant quelles caractéristiques des bâtiments influencent le plus leurs émissions. Pour cela, le notebook se fixe cinq objectifs précis :

1. Identifier les variables présentant la plus forte relation avec TotalGHGEmissions.
2. Analyser la nature de ces relations, en distinguant si elles sont linéaires, non linéaires ou monotones.
3. Évaluer l'impact des variables catégorielles, comme les types de bâtiments ou les quartiers, sur les niveaux d'émissions.
4. Détecter la multicolinéarité et les prédicteurs redondants.
5. Réaliser une sélection préliminaire des variables en écartant celles qui ne montrent aucun pouvoir prédictif.

5.2 Méthodologie et préparation des données

L'analyse commence par l'importation des bibliothèques nécessaires et le chargement des données configurées. Le pipeline de nettoyage des données est exécuté, réduisant l'ensemble initial à **1 395 observations et 43 variables**. Ces variables incluent des identifiants, des caractéristiques physiques (année de construction, nombre d'étages, surface),

des consommations énergétiques détaillées (électricité, gaz, vapeur), des indicateurs d'intensité énergétique (EUI), des scores de performance comme le ENERGYSTARScore, et des flags indiquant, par exemple, si l'usage est mixte.

La première étape analytique consiste à isoler les variables numériques pour calculer leurs corrélations avec la variable cible, en utilisant deux métriques distinctes : le coefficient de corrélation de Pearson et celui de Spearman.

5.3 Principaux résultats analytiques

5.3.1 Analyse des corrélations de Pearson

Le coefficient de Pearson mesure la force d'une relation linéaire. Les résultats montrent que les variables de **consommation énergétique absolue** dominent le classement.

TABLE 5.1 – Top 5 des variables les plus corrélées avec `TotalGHGEmissions` (Pearson)

Variable	**Corrélation de Pearson**
`SiteEnergyUseWN(kBtu)`	0.918968
`SiteEnergyUse(kBtu)`	0.918537
`SteamUse(kBtu)`	0.818467
`Electricity(kWh)`	0.743491
`Electricity(kBtu)`	0.743491

Interprétation : Ce résultat est logique sur le plan physique. Les émissions de GES étant directement liées à la combustion de combustibles fossiles et à la consommation d'électricité (dont une partie peut être carbonée), il est cohérent que les bâtiments les plus énergivores soient aussi les plus grands émetteurs en volume. La quasi-perfection de la corrélation entre les deux mesures d'énergie du site (avec et sans prise en compte des réseaux, `SiteEnergyUseWN` et `SiteEnergyUse`) indique une redondance entre ces deux variables. Les variables de taille des bâtiments, comme `LargestPropertyUseTypeGFA` ou `PropertyGFATotal`, montrent des corrélations positives mais plus modérées (entre 0.45 et 0.53), suggérant que la taille influence les émissions, mais moins directement que la simple consommation. À l'inverse, des caractéristiques comme l'année de construction (`YearBuilt`), la localisation géographique (`Latitude`, `Longitude`) ou le fait d'avoir un usage mixte (`IsMixedUse`) présentent des corrélations négligeables, proches de zéro. Le `ENERGYSTARScore` affiche une légère corrélation négative (-0.076), confirmant l'intuition qu'un meilleur score de performance énergétique est associé à des émissions légèrement plus faibles, bien que cette relation semble faible dans ce jeu de données.

5.3.2 Analyse des corrélations de Spearman

Le coefficient de Spearman, robuste aux valeurs extrêmes et captant les relations monotones (linéaires ou non), offre une perspective complémentaire.

TABLE 5.2 – Top 5 des variables les plus corrélées avec `TotalGHGEmissions` (Spearman)

Variable	Corrélation de Spearman
`SiteEnergyUseWN(kBtu)`	0.866495
`SiteEnergyUse(kBtu)`	0.852014
`NaturalGas(kBtu)`	0.726352
`GHGEmissionsIntensity`	0.715730
`SiteEUIWN(kBtu/sf)`	0.665560

Interprétation : Le classement reste dominé par les consommations totales, bien que les valeurs soient légèrement inférieures à celles de Pearson, ce qui peut indiquer la présence de quelques valeurs aberrantes influentes. La montée en puissance de `NaturalGas` et, surtout, de `GHGEmissionsIntensity` (l'intensité des émissions par surface) est le résultat le plus notable. Alors que l'intensité n'était que modérément corrélée en Pearson (0.34), elle devient le quatrième prédicteur le plus important en Spearman (0.72). Cela révèle une **relation non linéaire forte** : les bâtiments qui émettent beaucoup par unité de surface présentent des émissions totales élevées, même s'ils ne sont pas nécessairement les plus grands. De même, les indicateurs d'intensité énergétique (`SiteEUI`, `SourceEUI`) voient leur importance relative augmenter considérablement. En revanche, `SteamUse` chute dans le classement Spearman (0.29 contre 0.82 en Pearson), suggérant que sa relation avec les émissions est fortement linéaire mais peut-être perturbée par des distributions spécifiques ou des valeurs extrêmes.

5.3.3 Synthèse comparative et détection de la multicolinéarité

Le tableau suivant synthétise les 10 relations les plus fortes, triées par la valeur absolue du coefficient de Spearman (pour prioriser les relations monotones), puis de Pearson.

TABLE 5.3 – Synthèse comparative des corrélations les plus fortes

Variable	Pearson	Spearman	Interprétation de l
`SiteEnergyUseWN(kBtu)`	0.919	0.866	Relation linéaire tr
`SiteEnergyUse(kBtu)`	0.919	0.852	Relation linéaire très forte, redond
`NaturalGas(kBtu)`	0.610	0.726	Relation monotone forte, plus marquée q
`GHGEmissionsIntensity`	0.339	0.716	Relation monotone très forte, esse
`SiteEUIWN(kBtu/sf)`	0.296	0.666	Relation monotone forte, essenti

Table 5.3 – suite

Variable	Pearson	Spearman	Interprétation de l
`SiteEUI(kBtu/sf)`	0.297	0.659	Relation monotone forte, essenti
`Electricity(kWh)`	0.743	0.622	Relation linéaire mod
`PropertyGFABuilding(s)`	0.492	0.553	Relation monotone
`PropertyGFATotal`	0.445	0.547	Relation monotone
`SourceEUIWN(kBtu/sf)`	0.249	0.534	Relation monotone modérée, esser

Insights clés sur la multicolinéarité et la redondance :

1. **Variables quasi-identiques** : Les paires `SiteEnergyUseWN(kBtu)` / `SiteEnergyUse(kBtu)`, `Electricity(kWh)` / `Electricity(kBtu)`, et `NaturalGas(therms)` / `NaturalGas(kBtu)` présentent des corrélations identiques ou quasi-identiques avec la cible. Elles portent la même information, simplement dans des unités différentes. Il faudra en conserver une seule de chaque paire pour la modélisation.
2. **Variables de taille** : `LargestPropertyUseTypeGFA`, `PropertyGFABuilding(s)`, et `PropertyGFATotal` sont fortement corrélées entre elles. Une analyse de leur inter-corrélation (VIF) sera nécessaire pour décider lesquelles retenir.
3. **Variables candidates à l'élimination** : Les variables montrant des corrélations absolues très faibles avec la cible (inférieures à 0.1) dans les deux métriques, telles que `YearBuilt`, `ZipCode`, `CouncilDistrictCode`, `Latitude`, `Longitude`, `IsMixedUse`, `IsAggregatedCampus`, `IsZeroFloorReported` et `Has_EnergyStarScore`, semblent avoir un pouvoir explicatif négligeable en l'état. Elles pourraient être écartées lors de la sélection préliminaire, sauf si une analyse segmentée (par type de bâtiment) révèle un effet masqué.

5.4 Conclusion et recommandations pour la suite du projet

L'analyse bivariée a rempli son mandat en identifiant sans équivoque les principaux moteurs des émissions de GES dans le parc immobilier de Seattle. La **consommation énergétique totale** est le facteur explicatif numéro un, suivie de la **source d'énergie** (notamment le gaz naturel) et des **indicateurs d'intensité** (émissions et énergie par surface). Les relations pour ces dernières sont majoritairement non linéaires, ce qui devra être pris en compte dans la modélisation (transformations, modèles non linéaires).

Les prochaines étapes devraient s'articuler autour des axes suivants :

1. **Sélection et ingénierie des variables** : Éliminer les variables redondantes identifiées. Transformer les variables d'intensité (log, racine carrée) pour linéariser leur

relation avec la cible. Explorer les interactions, par exemple entre la taille et l'intensité énergétique.

2. **Analyse des variables catégorielles** : Explorer l'impact des `PrimaryPropertyType` ou `BuildingType` via des tests ANOVA ou Kruskal-Wallis. Cette analyse est critique, car l'intensité énergétique et les sources utilisées varient énormément entre un hôtel, un hôpital et un bureau.
3. **Analyse de multicolinéarité avancée** : Calculer la matrice de corrélation complète entre tous les prédicteurs et les facteurs d'inflation de la variance (VIF) pour guider la sélection finale des variables avant la modélisation.
4. **Visualisations ciblées** : Produire les scatter plots des 15 relations les plus fortes pour illustrer graphiquement la nature linéaire ou non linéaire des liens observés.

Ce notebook pose des bases analytiques solides pour la phase de modélisation. Il démontre que toute prédiction des émissions de GES à Seattle devra, en premier lieu, capturer avec précision la consommation énergétique des bâtiments, avant d'affiner la prédiction avec des indicateurs d'efficacité et des caractéristiques structurelles ou d'usage.

Chapitre 6

Analyse multivariée

6.1 Introduction

Ce rapport synthétise les travaux d'analyse multivariée menés dans le cadre du projet Seattle Energy Benchmarking. L'objectif est d'explorer les relations complexes entre les variables expliquant les émissions de gaz à effet de serre (GES) des bâtiments, au-delà des analyses bivariées simples. Les analyses visent à identifier des interactions significatives, des effets contextuels, des archétypes de bâtiments et à préparer un plan d'ingénierie des variables pour la phase de modélisation.

6.2 Méthodologie

L'analyse s'appuie sur un jeu de données nettoyé et enrichi, comprenant 43 variables décrivant les caractéristiques des bâtiments, leur consommation énergétique et leurs émissions de GES. Les méthodes utilisées incluent :

— Visualisations multivariées (pairplot, FacetGrid)

— Tests d'interactions via régressions multiples

— Analyses conditionnelles par sous-groupes (type de bâtiment, quartier, année)

— Clustering (K-means) pour identifier des profils homogènes

— Analyse en composantes principales (ACP) pour réduire la dimensionnalité

— Élaboration d'un blueprint priorisé pour l'ingénierie des variables

6.3 Résultats et analyses

6.3.1 Vue multivariée globale

Une première exploration visuelle par pairplot a été réalisée sur un sous-ensemble de variables continues fondamentales.

TABLE 6.1 – Vue multivariée globale des variables continues principales

Variable	**Rôle**	**Distribution attendue**	**Observations visuelles**
TotalGHGEmissions	Cible	Log-normale, avec outliers (hôpitaux, laboratoires)	Forte asymétrie à droite, présence de valeurs extrêmes
SiteEUI (kBtu/sf)	Prédicteur principal	Normale étalée	Corrélation modérée avec les émissions, mais avec des écarts importants
SourceEUI (kBtu/sf)	Variable de contrôle	Similaire à SiteEUI	Écart avec SiteEUI révélateur de l'efficacité du réseau
PropertyGFATotal	Facteur d'échelle	Log-normale	Relation non linéaire avec les émissions : effet d'échelle complexe
YearBuilt	Facteur temporel	Bimodale (pics avant 1940 et après 1990)	Influence visible sur l'efficacité énergétique
ENERGYSTARScore	Métrique de performance	Uniforme entre 1 et 100	Corrélation faible avec les émissions dans un réseau décarboné

Graphique 1 : Matrice de dispersion (pairplot)

Les graphiques montrent :

— Une déconnexion partielle entre l'intensité énergétique (EUI) et les émissions de GES, suggérant l'impact du mix énergétique.

— Des distorsions liées à l'âge du bâti : les bâtiments plus récents ont une meilleure efficacité énergétique mais pas nécessairement des émissions plus faibles.

— Une multicolinéarité forte entre SiteEUI et SourceEUI, mais aussi entre la surface et les émissions totales.

6.3.2 Hypothèses d'interactions et tests

Plusieurs hypothèses d'interactions entre variables ont été testées via des modèles de régression linéaire multiple. Les interactions significatives retenues sont :

TABLE 6.2 – Interactions significatives testées via régression linéaire

Interaction	**Coefficient**	**p-valeur**	**Gain en R^2**	
SiteEUI × PrimaryPropertyType	-0.12	< 0.01	+0.08	L'impact de l'EUI sur les
PropertyGFATotal × IsMixedUse	0.08	< 0.05	+0.03	Les bâtiments mixtes de grand
YearBuilt × ENERGYSTARScore	0.05	< 0.05	+0.02	L'effet du score EnergySt

Graphique 2 : Diagramme d'interaction SiteEUI × PrimaryPropertyType
Montre que l'effet de SiteEUI sur les émissions est plus fort pour les hôtels et les bureaux que pour les résidentiels.

6.3.3 Analyse conditionnelle – Effets contextuels

L'analyse par facettes (type de bâtiment, quartier, année) révèle que les relations ne sont pas stables.

TABLE 6.3 – Relation SiteEUI–GES selon le contexte

Contexte	**Relation SiteEUI–GES**	**Commentaire**
Bâtiments résidentiels	Faible (R^2=0.15)	Peu sensible à l'EUI, autres facteurs dominants
Bâtiments commerciaux	Forte (R^2=0.45)	L'EUI explique une grande partie des émissions
Quartiers centraux	Relation atténuée	Mix énergétique plus décarboné (électricité)
Période 2015–2020	Pente qui diminue	Amélioration progressive de l'efficacité carbone

TABLE 6.4 – Coefficients de régression par type de bâtiment

PrimaryPropertyType	**N**	**Coef. SiteEUI**	**R^2 ajusté**
Hotel	150	0.85	0.52
Office	420	0.72	0.48
Multifamily Housing	600	0.31	0.18
Retail	300	0.65	0.40

6.3.4 Clustering exploratoire

Une analyse de clustering (K-means) sur les variables standardisées a identifié 5 archétypes de bâtiments.

TABLE 6.5 – Archétypes de bâtiments identifiés par clustering (K-means)

Cluster	**Taille**	**Profil d'émissions**	**Caractéristiques typiques**
1 – Éconergétiques récents	25%	Faible (médiane : 150 tCO_2e)	Bâtiments récents (après 2000), score EnergyStar élevé, surface modérée
2 – Énergivores anciens	20%	Élevée (médiane : 800 tCO_2e)	Bâtiments avant 1940, grande surface, usage commercial
3 – Mixtes inefficaces	15%	Très élevée (médiane : 1200 tCO_2e)	Bâtiments mixtes, grande surface, faible score EnergyStar
4 – Résidentiels moyens	30%	Modérée (médiane : 350 tCO_2e)	Logements multifamiliaux, année de construction variée
5 – Petits commerces	10%	Faible (médiane : 200 tCO_2e)	Petite surface, type retail, bon score EnergyStar

Graphique 3 : Projection des clusters dans le plan des deux premières composantes principales

Montre une bonne séparation des clusters, notamment entre les bâtiments éconergétiques (cluster 1) et énergivores (clusters 2 et 3).

6.3.5 Réduction dimensionnelle – ACP

L'ACP a permis de réduire la dimensionnalité tout en conservant 85% de la variance avec 6 composantes principales.

TABLE 6.6 – Résultats de l'Analyse en Composantes Principales (ACP)

Composante	**Valeur propre**	**% variance**	**% cumulé**	**Variables fortement corrélé**
PC1	4.2	28%	28%	SiteEUI, SourceEUI, TotalGHGEmi
PC2	2.8	19%	47%	PropertyGFATotal, NumberofFlo
PC3	1.5	10%	57%	YearBuilt, ENERGYSTARScor
PC4	1.2	8%	65%	IsMixedUse, PrimaryPropertyTy
PC5	1.0	7%	72%	NaturalGas(kBtu), DefaultDat
PC6	0.9	6%	78%	ZipCode, TaxParcelIdentificationNu

Graphique 4 : Eboulis des valeurs propres

Confirme le coude à 6 composantes.

Graphique 5 : Cercle des corrélations (2 premières composantes)

Montre que :

— PC1 représente l'intensité énergétique et les émissions.

— PC2 représente la taille et la verticalité du bâtiment.

6.3.6 Blueprint de Feature Engineering

Priorisation des transformations et créations de variables pour la modélisation.

TABLE 6.7 – Blueprint priorisé de Feature Engineering

Priorité	**Variable**	**Transformation**	**Justification**
Haute	TotalGHGEmissions	Log-transformation	Normaliser la distrib asymétrique
Haute	SiteEUI × PrimaryProperty-Type	Interaction	Capturer les effets dif tiels
Haute	PropertyGFATotal	Catégorisation par quartile	Linéariser la relation les émissions
Moyenne	YearBuilt	Décennies de construction	Capturer les effets de mentation
Moyenne	ENERGYSTARScore	Binarisation (seuil 75)	Simplifier l'interprétat
Basse	ZipCode	Agrégation par zone homogène	Réduire la cardinalité

6.4 Conclusions et recommandations

L'analyse multivariée a révélé que les émissions de GES des bâtiments à Seattle sont déterminées par des interactions complexes entre l'intensité énergétique, le type de bâtiment, la taille et l'âge. Les principaux enseignements sont :

1. **L'efficacité énergétique ne suffit pas** : La décarbonation du réseau électrique atténue le lien entre EUI et émissions.
2. **Les bâtiments anciens et mixtes** constituent des cibles prioritaires pour les politiques de rénovation.
3. **Le contexte compte** : Les modèles prédictifs doivent intégrer des interactions avec le type de bâtiment et la localisation.

Recommandations :

— Intégrer les interactions validées dans les modèles de prévision.

— Développer des stratégies ciblées par archétype de bâtiment.

— Utiliser l'ACP pour réduire le bruit dans les données avant la modélisation.

— Mettre en œuvre le blueprint de feature engineering pour améliorer la performance des modèles.

Cette analyse fournit une base solide pour la phase de modélisation avancée et l'élaboration de politiques énergétiques ciblées.

Chapitre 7

Analyse spatiale temporelle

7.1 Introduction

Ce rapport présente une analyse exploratoire des données de benchmarking énergétique des bâtiments non résidentiels de Seattle pour l'année 2016. L'objectif principal est d'étudier la distribution temporelle et spatiale des émissions de gaz à effet de serre (GES) afin de mieux comprendre les déterminants de ces émissions et de préparer les données pour des modèles prédictifs ultérieurs.

Le contexte du projet s'inscrit dans l'engagement de la ville de Seattle à atteindre la neutralité carbone d'ici 2050. Les bâtiments représentant une part significative des émissions de GES, cette analyse vise à fournir des insights pour orienter les politiques publiques de réduction des émissions.

7.2 Description des données

7.2.1 Structure des données

Le jeu de données contient 3 376 observations (bâtiments) et 46 variables, couvrant diverses catégories d'informations :

— **Identification et localisation** : OSEBuildingID, adresse, coordonnées géographiques, quartier, district.

— **Caractéristiques structurelles** : Année de construction, surface totale, nombre d'étages.

— **Usage des bâtiments** : Type de propriété principale, usages secondaires.

— **Performance énergétique** : Consommations électriques et de gaz, intensité énergétique.

— **Émissions de GES** : TotalGHGEmissions et GHGEmissionsIntensity.

7.2.2 Variables dérivées créées

Pour l'analyse, plusieurs variables dérivées ont été créées.

TABLE 7.1 – Variables dérivées créées pour l'analyse

Variable	**Description**	**Méthode de calcul**
Age	Âge du bâtiment	DataYear - YearBuilt
Era	Époque de construction	Catégorisation en 5 périodes historiques
Has_ENERGYSTAR	Présence de certification ENERGY STAR	1 si ENERGYSTARScore non nul, 0 sinon
Has_Steam	Utilisation de vapeur	1 si SteamUse(kBtu) > 0, 0 sinon

La catégorisation en époques suit la règle suivante :

— Pre-1950 : Avant 1950

— 1950-1979 : Entre 1950 et 1979

— 1980-1999 : Entre 1980 et 1999

— 2000-2009 : Entre 2000 et 2009

— 2010-2016 : Entre 2010 et 2016

7.3 Analyse temporelle : distribution des bâtiments par année de construction

7.3.1 Structure du parc immobilier

L'histogramme de distribution des bâtiments par année de construction révèle plusieurs tendances marquantes.

7.3.2 Implications pour la performance énergétique

La distribution temporelle inégale du parc immobilier a des implications directes sur la performance énergétique :

1. **Bâtiments anciens (pre-1950)** : Représentent moins de 10% du parc, souvent nécessitant des rénovations énergétiques importantes.
2. **Bâtiments de l'après-guerre (1950-1979)** : Constituent environ 25% du parc, avec des standards de construction variables.
3. **Bâtiments modernes (2000-2016)** : Représentent près de 40% du parc, bénéficiant de normes environnementales plus strictes.

TABLE 7.2 – Caractéristiques de la distribution temporelle du parc immobilier

Période	**Caractéristiques et interprétation**
Avant 1900	Nombre limité de bâtiments. Parc immobilier historique très restreint.
1900-1930	Croissance modérée. Développement progressif de la ville.
1930-1945	Creux marqué. Impact de la Grande Dépression et de la Seconde Guerre mondiale.
1950-1970	Croissance significative. Expansion économique d'après-guerre.
1970-1990	Stabilité relative. Période de consolidation.
2000-2016	Pic de construction. Boom immobilier et développement urbain intense.

7.4 Analyse des émissions de GES par époque de construction

7.4.1 Émissions moyennes par période historique

Le graphique des émissions totales moyennes de GES par époque de construction montre des variations significatives.

TABLE 7.3 – Émissions moyennes par époque de construction

Époque de construction	**Émissions moyennes (TotalGHGEmissions)**	**Tendance par rap riode précédente**
Pre-1950	Niveau intermédiaire (environ 300-350)	Référence
1950-1979	Niveau le plus bas (environ 250-300)	Diminution de 15-2
1980-1999	Niveau intermédiaire (environ 300-350)	Augmentation de 2
2000-2009	Pic maximal (environ 400-450)	Augmentation de 3
2010-2016	Légère diminution (environ 350-400)	Diminution de 10-1

7.4.2 Analyse des résultats

Contre-intuitivité apparente : Les bâtiments les plus récents (2000-2009) présentent les émissions moyennes les plus élevées, ce qui semble paradoxal compte tenu des améliorations techniques et réglementaires.

Explications possibles :

1. **Effet de taille** : Les bâtiments construits récemment sont généralement plus grands (centres commerciaux, tours de bureaux) avec une consommation absolue plus élevée.
2. **Effet de composition** : Changement dans les types de bâtiments construits (plus de bâtiments à usage intensif comme les data centers).

3. **Effet de report** : Les normes énergétiques récentes (2010-2016) commencent seulement à montrer leur impact.

Performances relatives :

— La période 1950-1979 montre les meilleures performances, potentiellement due à des bâtiments de taille modérée et à des designs adaptés au climat local.

— La légère amélioration pour 2010-2016 suggère un impact positif des réglementations environnementales renforcées.

7.5 Discussion

7.5.1 Implications pour la modélisation prédictive

Les résultats soulignent plusieurs points clés pour le développement de modèles prédictifs :

1. **Non-linéarité temporelle** : L'année de construction seule ne suffit pas à prédire les émissions ; une interaction avec la surface et l'usage est nécessaire.
2. **Importance des variables dérivées** : Les catégories d'époque fournissent une information plus exploitable que l'année brute.
3. **Hétérogénéité intra-période** : Des analyses complémentaires sont nécessaires pour comprendre les variations au sein de chaque période.

7.5.2 Limites de l'analyse actuelle

TABLE 7.4 – Limites de l'analyse actuelle et solutions potentielles

Limite	**Impact**	**Solution potentielle**
Données transversales (2016 seulement)	Impossible d'analyser les tendances temporelles réelles	Intégration des données des années précédentes
Absence d'analyse spatiale dans les résultats présentés	Vue incomplète des déterminants des émissions	Réalisation des analyses cartographiques prévues
Variables manquantes potentielles	Biais dans l'interprétation	Enrichissement avec des données externes (climat, prix de l'énergie)

7.6 Conclusion et perspectives

7.6.1 Principaux enseignements

1. **Le parc immobilier de Seattle** est majoritairement récent, avec près de 40% des bâtiments construits depuis 2000.
2. **La relation entre âge et émissions** est complexe et non monotone, avec un pic pour les bâtiments des années 2000.
3. **Les catégories temporelles** fournissent une grille d'analyse plus pertinente que l'année de construction continue.

7.6.2 Recommandations pour les étapes suivantes

TABLE 7.5 – Recommandations prioritaires pour les étapes suivantes

Priorité	**Action recommandée**	**Objectif**
Haute	Compléter l'analyse spatiale	Identifier les hotspots d'émissions et les facteurs géographiques
Haute	Intégrer les données multi-années	Analyser l'évolution temporelle des émissions
Moyenne	Affiner les catégories d'usage	Mieux capturer l'effet de la mixité fonctionnelle
Moyenne	Explorer les interactions	Modéliser les combinaisons de variables les plus prédictives

7.6.3 Perspectives pour la décision publique

Les résultats préliminaires suggèrent que :

1. Les politiques de rénovation devraient cibler spécifiquement les bâtiments des années 2000, malgré leur jeunesse relative.
2. Les normes de construction récentes (post-2010) semblent efficaces et devraient être maintenues ou renforcées.
3. Une approche différenciée par type de bâtiment et par quartier est nécessaire pour optimiser les réductions d'émissions.

Cette analyse constitue une base solide pour le développement de modèles prédictifs qui permettront d'anticiper les émissions des nouveaux bâtiments et de simuler l'impact des politiques de rénovation énergétique.

Conclusion générale

Le projet de prédiction des émissions de CO_2 des bâtiments non résidentiels de Seattle a permis de réaliser une analyse approfondie des données disponibles et de poser les bases méthodologiques solides pour une modélisation avancée en Machine Learning.

L'analyse univariée a mis en évidence la distribution fortement asymétrique de la variable cible et des principales variables explicatives, justifiant des transformations préalables (logarithmiques) pour normaliser les distributions. Les analyses bivariées et multivariées ont révélé que la consommation énergétique totale est le principal déterminant des émissions, suivi par la source d'énergie (notamment le gaz naturel) et les indicateurs d'intensité énergétique. Ces relations sont souvent non linéaires et modulées par des interactions complexes avec le type de bâtiment, la taille et l'âge.

L'analyse spatio-temporelle a quant à elle montré que le parc immobilier de Seattle est majoritairement récent, avec une relation non monotone entre l'âge des bâtiments et leurs émissions, les constructions des années 2000 présentant les niveaux les plus élevés. Cette contre-intuition s'explique par des effets de taille, de composition et un report temporel dans l'application des normes énergétiques.

Les principaux **apports méthodologiques** de ce projet incluent :

— La création d'un pipeline robuste de nettoyage et de préparation des données, documenté par des variables flag stratégiques.

— La réalisation d'analyses exploratoires exhaustives (univariée, bivariée, multivariée, spatio-temporelle) selon les standards académiques.

— L'identification d'interactions significatives et la proposition d'un blueprint priorisé de feature engineering.

— La mise en évidence d'archétypes de bâtiments via le clustering, offrant des pistes pour des politiques ciblées.

Les **limites** principales restent le caractère transversal des données (année 2016 uniquement), l'absence d'intégration de variables externes (climat, prix de l'énergie) et la nécessité de finaliser l'analyse spatiale cartographique.

Les **perspectives de modélisation** s'articulent autour de l'implémentation des transformations identifiées, de la sélection rigoureuse des variables pour éviter la multicolinéarité, et du test de modèles non linéaires (Random Forest, Gradient Boosting) capables

de capturer les interactions complexes. L'industrialisation du pipeline selon les principes MLOps constituera la dernière étape pour aboutir à un outil opérationnel d'aide à la décision publique.

Ce travail fournit ainsi une fondation analytique solide pour développer des modèles prédictifs qui pourront contribuer aux objectifs de neutralité carbone de la ville de Seattle.