

Rendu 1 — Rapport d'exploration, data visualisation & pré-processing des données

1. Introduction au projet

Contexte

► Contexte d'insertion du projet dans votre métier

Le projet s'inscrit dans un contexte de **transition vers un rôle orienté data / IA**, dans lequel la compréhension des émissions de CO₂ des véhicules a une forte résonance métier :

- Pilotage de projets data,
- Capacité à analyser des datasets massifs, hétérogènes, réglementaires.

► Du point de vue technique

Ce rendu 1 mobilise :

- Les outils :
 - GitHub
 - VS Code
 - Anaconda
- Les skills :
 - Des pipelines de **pre-processing avancés**
 - Des analyses statistiques robustes
 - De la coordination d'équipe

► Du point de vue économique

Les enjeux économiques liés aux émissions CO₂ sont majeurs :

- Fiscalité automobile,
- Conformité WLTP,
- Transition énergétique,
- Impact sur la valeur résiduelle des véhicules et les politiques publiques.

La capacité à prédire ou analyser les émissions est **clé** pour des constructeurs, collectivités, flottes et régulateurs.

► Du point de vue scientifique

Les émissions WLTP/NECD dépendent :

- Des paramètres physiques (masse, puissance, cylindrée),
- De l'aérodynamique,
- Du type d'énergie,
- Des technologies embarquées.

Le dataset capte ces dimensions, permettant d'étudier **corrélations, causalités hypothétiques**, relations non-linéaires et interactions.

Objectifs

Objectifs principaux

Analyser les émissions de Co2 des véhicules vendus en Europe pour :

- Identifier les véhicules les plus polluants
- Etudier l'évolution dans le temps
- Comprendre les caractéristiques techniques qui influencent ces émissions
- Prédire l'émission Co2 d'un véhicule à partir de ses caractéristiques physiques

Niveau d'expertise du groupe

Groupe étudiant DataScientest BootCamp MLE. Pas d'expérience passée en codage ni d'expertise automobile.

2. Compréhension & manipulation des données

Cadre

► Jeux de données utilisés

Parmi les deux jeux de données disponibles — l'un français et l'autre européen — nous avons retenu celui offrant la couverture la plus large et la plus représentative. Bien que les données s'étendent de 2010 à 2024, la volumétrie et les contraintes de traitement nous ont conduits à focaliser l'analyse sur l'année 2022. Il recense :

- Caractéristiques techniques des véhicules,
- Consommations WLTP,
- Émissions de CO₂,
- Autonomie électrique,
- Technologies innovantes,
- Type de motorisation, etc.

► Disponibilité

Les données sont présentes sur le site de European Environment Agency

Source exacte : <https://www.eea.europa.eu/data-and-maps/data/co2-cars-emission-20>

► Volumétrie

- >91 Millions de lignes et 40 colonnes sur le dataset 2010-2024
- Choix de limiter l'étude à 2022 : 9.479.544 lignes
- Identification de 19 colonnes non utiles au modèle (cf annexe 1) et suppression des doublons strict :
 - **533.390 lignes**
 - **21 colonnes**
 - Type mixte : numériques, catégorielles, textuelles, encodages techniques.

Pertinence

► Variables les plus pertinentes pour l'objectif

Selon nos analyses, les variables les plus explicatives des émissions WLTP/NEDC sont :

- **Type de carburant (Ft)**
- **Masse (m (kg))**
- **Cylindrée (ec)**
- **Puissance (ep)**
- **Consommation fuel (Fuel consumption)**
- **Autonomie électrique (electric range)**
- **Dimensions du véhicule (W, At1 et At2)**

► Variable cible

- **Ewltp (g/km)** (prioritaire, cohérente avec la normalisation actuelle)
- Ignorer **Enedc (g/km)** (ancienne norme)
(*Ewltp présente presque 0% de Na contre plus de 80% de Na pour Enedc (g/km)*)

► Particularités du jeu de données

- Volumétrie très élevée
- Beaucoup de valeurs manquantes (jusqu'à 90% selon colonne).
- Colonnes techniques extrêmement bruitées.
- Présence d'encodages multiples, doubles espaces, variables textuelles hybrides.
- Hétérogénéité PHEV vs BEV vs ICE déséquilibrée et mal renseignée

► Limitations

- Toutes les variables nécessitent un nettoyage massif et/ou regroupement, voir même un pré-traitement avant split pour mettre en cohérence les colonnes entre elles.

Pre-processing & Feature Engineering

| Colonne | Type | Définition | Nettoyage et traitement |
|----------------|--------------|---|--|
| Country | Catégorielle | Pays d'immatriculation | Aucun nettoyage nécessaire |
| Man | Catégorielle | Marque du constructeur (Renault, Toyota...) | Mettre tout en majuscule, suppression des caractères et espaces en début et fin de chaîne. Remplacer les Man mal saisi par la valeur correcte. Cross-vérification et correction et/ou remplissage de Man Remplacer les valeurs génériques 'AANSS', 'AAIVA' et 'ND' par Na |

| | | | |
|--------|--------------|--|---|
| | | | Remplir les Na de Man par le mode de Man en groupant par Mk puis en groupant par Cn, puis le reste par 'OTHER' |
| Va | Catégorielle | Variante du modèle (niveau moteur/équipement) | Supprimer les espaces avant/après, harmoniser en majuscules, supprimer tous les espaces, supprimer les séparateurs et garder les lettres et chiffres, retirer les mots non techniques (ex : 4x4, 4WD, ...), grouper les occurrences <20 dans OTHER |
| Ve | Catégorielle | Version commerciale | Appliquer majuscule partout, supprimer les espaces avant/après, harmonise en majuscules, supprimer tous les espaces internes, supprimer les séparateurs et garder lettres/chiffres, retirer tout caractère non technique (tout sauf A-Z et 0-9), retirer les mots non techniques ('START', 'ENGINE', 'CODE', 'MODEL', 'TYPE']), grouper toutes les valeurs apparaissant < 20 fois en OTHER, convertir tous les fake NaN en vrai NaN, forcer le type de la colonne en string, remplacer tous les Nan en OTHER, remplacer les valeurs aberrantes en OTHER,fillna en OTHER |
| Mk | Catégorielle | Marque | Nettoyer la colonne Mk, mettre tout en majuscule, supprimer les caractères et espaces en début et fin de chaine Remplacer les Mk mal saisi par la valeur correcte Remplir les Na de Mk par le mode de Mk en groupant par Cn puis le reste par 'OTHER' |
| Cn | Catégorielle | Nom commercial du modèle | Convertir toutes les fausses valeurs manquantes en vrais NaN, Conversion en MAJ + trim, Normaliser les espaces (1 seul), uniformiser les tirets, Supprimer les caractères parasites mais garder LETTRES / CHIFFRES / ESPACES, Normaliser encore une fois les espaces, OTHER pour les vrais NaN, Regrouper les valeurs proches avec un RapidFuzz, nettoyage à la main de 3008 5008 |
| Cr | Catégorielle | Catégorie sous laquelle le véhicule est immatriculé dans le pays | - NaN attribués à la variable unique M1 |
| m (kg) | Quantitative | Masse à vide du véhicule | Remplacer les outliers de m(kg) par la moyenne de la masse du même modèle |

| | | | |
|---------------------|--------------|---|---|
| | | | Remplir les Na de m(kg) par la moyenne de la masse du même modèle puis le reste par restant par la médiane de la masse |
| Ewltp (g/km) | Quantitative | CO₂ WLTP → variable cible | - Supprimer les null nan, renaming de la colonne en target_Co2, suppression de la valeur aberrante MORELO |
| W (mm) | Quantitative | Empattement (distance axes roues) | Imputation de médiane avec la colonne Mk puis médiane globale pour le reste des NA et flags de missings |
| At1 (mm) | Quantitative | Voie avant (largeur entre roues AV) | Imputation de médiane avec la colonne Mk puis médiane globale pour le reste des NA et flags de missings |
| At2 (mm) | Quantitative | Voie arrière | Imputation de médiane avec la colonne Mk puis médiane globale pour le reste des NA et flags de missings |
| Ft | Catégorielle | Type de carburant (diesel, petrol, BEV, PHEV...) | - Harmonisation et Standardisation Textuelle Nous avons commencé par nettoyer les variables textuelles (carburant et alimentation) qui présentaient de nombreuses incohérences de format. Pour cela, nous avons appliqué un nettoyage strict via des expressions régulières (Regex). En supprimant les caractères spéciaux (tirets, slashes), les espaces superflus et en convertissant tout en majuscules, nous avons pu fusionner des doublons sémantiques (par exemple, "Diesel-Electric" et "diesel/electric" sont devenus une seule catégorie unique "DIESELELECTRIC") |
| Fm | Catégorielle | Mode d'énergie (M = microhybride, H = hybride...) | - Cette variable textuelle a été harmonisée pour éliminer les doublons (ex: "diesel-electric" et "DIESELELECTRIC" fusionnés) via un nettoyage strict (majuscules, suppression de caractères spéciaux). Pour réduire le bruit, les catégories rares (< 20 occurrences) ont été regroupées sous le label "OTHER". Les valeurs manquantes ont été comblées par le mode le plus fréquent |
| ec (cm3) | Quantitative | Cylindrée moteur | - Nous avons appliqué une logique physique : les cylindrées des véhicules électriques et à hydrogène ont été forcées à 0, car techniquement inexistantes. Pour les véhicules thermiques avec données manquantes, nous avons imputé la médiane calculée spécifiquement sur le même type de carburant, évitant ainsi une moyenne globale imprécise. |

| | | | |
|----------------------------|--------------|--|--|
| ep (KW) | Quantitative | Puissance moteur | - La puissance étant une caractéristique universelle, aucune valeur nulle n'a été insérée. Les données manquantes ont été traitées par une imputation statistique contextuelle : chaque trou a été remplacé par la médiane de la puissance des véhicules du même groupe de carburant, assurant la cohérence technique (ex: ne pas attribuer la puissance d'une sportive à une citadine). |
| z (Wh/km) | Quantitative | Consommation électrique (BEV/PHEV) | Suppression des petrol avec un z renseigné, création colonne energy_type avec la colonne mode_alimentation_energ_véhicule pour attribution de la mediane par type de carburant, remplacer les nan des véhicules thermiques par 0 |
| IT | Catégorielle | Innovations techniques (codes E...) | Remplacer les codes de IT par 1 si IT non vide et par 0 si vide. Si Erwltp non nul mettre IT sur 1 |
| Erwltp (g/km) | Quantitative | Réduction de CO ₂ lié à IT selon WLTP | Remplacer les Na de Erwltp par 0.0 si IT est 0, et par la médiane si IT est 1 |
| Fuel consumption | Quantitative | Consommation carburant (L/100km) | - Le nettoyage a respecté la nature du moteur : consommation forcée à 0 pour les véhicules "zéro émission" (électrique, hydrogène). Pour les thermiques sans données, nous avons appliqué la médiane de consommation propre à leur type de carburant, garantissant des estimations réalistes. |
| Electric range (km) | Quantitative | Autonomie électrique (BEV/PHEV) | Suppression des petrol avec un z renseigné, création colonne energy_type avec la colonne mode_alimentation_energ_véhicule pour attribution de la mediane par type de carburant, remplacer les nan des véhicules thermiques par 0 |

Visualisations & Statistiques

Relations entre variables

Les corrélations des colonnes retenues avec la variable cible que nous avons identifiée (EWLTP) :

| Corrélation avec target_co2 (EWLTP) | |
|-------------------------------------|-----------------|
| consommation_carburant_l/100km | 0.960853 |
| cylindre_du_moteur_cm3 | 0.728743 |
| IT | 0.457279 |
| Erwltp (g/km) | 0.362407 |
| puissance_du_moteur_kw | 0.232432 |

| | |
|---------------------|------------------|
| W (mm) | 0.231917 |
| At2 (mm) | 0.201127 |
| At1 (mm) | 0.185705 |
| W (mm)_missing | -0.006617 |
| At2 (mm)_missing | -0.007921 |
| At1 (mm)_missing | -0.010266 |
| m (kg) | -0.035630 |
| Electric range (km) | -0.684059 |
| z (Wh/km) | -0.805417 |

Les patterns voitures électriques, thermiques et hybrides ressortent naturellement.

Distribution & outliers

Nous avons documenté :

- Outliers physiques conservés (ex : BEV > 700 km),
- Outliers non physiques supprimés ou réimputés.

Stats confirmant les observations

- Calcul des quantiles Q01 / Q25 / Q50 / Q75 / Q99 pour chaque colonne.
- Inspection de 50 premières lignes post-masquage.

Projection vers la modélisation

Le dataset contient :

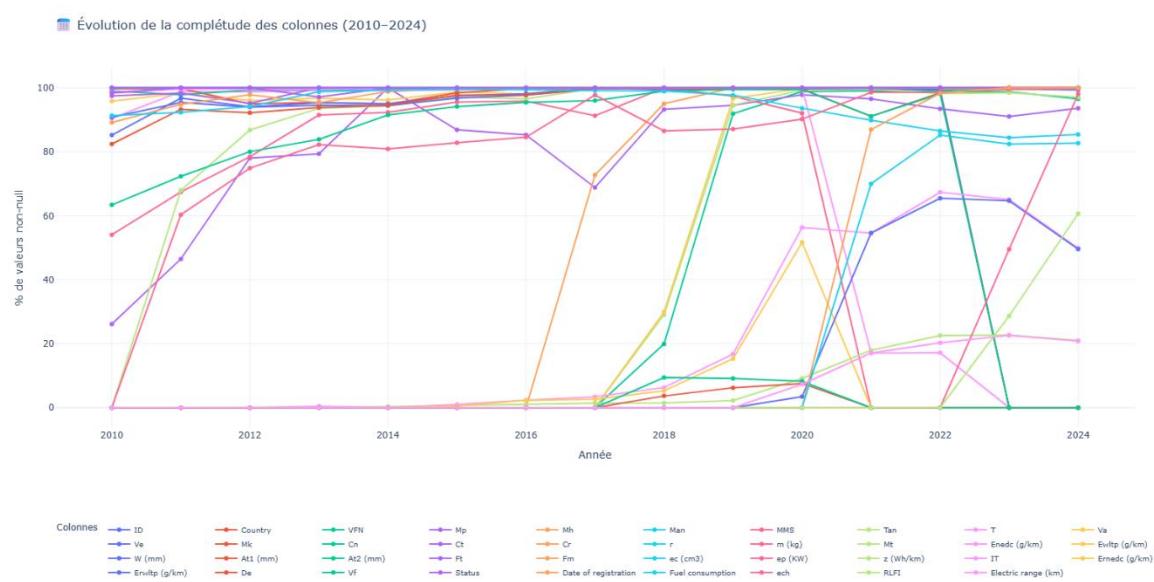
- Variables très explicatives,
- Clusters naturels (thermiques, hybrides, électriques),
- Distributions clairement liées à un phénomène physique → **fort potentiel prédictif**.

Annexe 1 - Tableau des variables supprimées

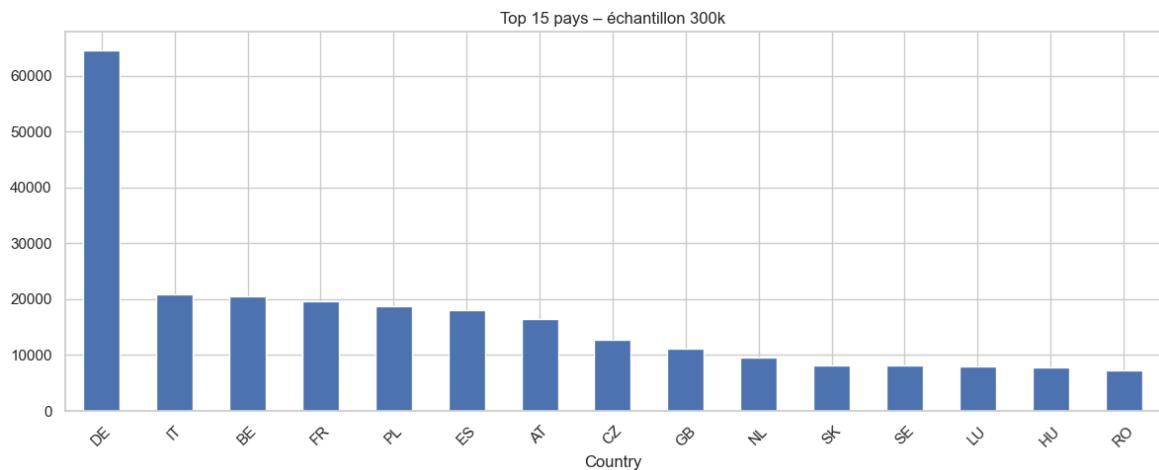
| Colonne | Définition | Raison de suppression |
|---------------|---|--|
| ID | Identifiant unique de l'enregistrement | ID unique sur les tableaux des années séparés. |
| VFN | Identifiant technique unique pour un modèle | On garde la colonne Cn puisque mieux renseigné et plus facilement interprétable que VFN |
| Mp | Groupe de constructeur | Les marques ont été racheté au fur et à mesure des années par des groupes constructeurs différents. Ce qui explique pourquoi pour une même marque (Man), des noms de constructeurs différents apparaissent (Mp). |
| Mh | Nom du constructeur selon la dénomination standard utilisée dans l'UE. | Mh peut être déduite à partir de Man. La colonne Man est plus propre et mieux renseignée. |
| MMS | Nom du constructeur tel qu'il apparaît dans le registre du pays membre (Member State). | Colonne à 100% vide |
| Tan | Numéro d'homologation du type du véhicule | La colonne Tan est purement administrative. Elle n'apporte pas de valeur ajoutée pour nos modèles de prédiction de la consommation de CO2 ou électrique. |
| T | Type véhicule au sens technique | Colonne pas exploitable pour le modèle |
| Cr | Catégorie sous laquelle le véhicule est immatriculé dans le pays | Lié à Ct |
| r | Nombre de rapports | Non pertinent pour la modélisation, ce n'est pas une variable explicative |
| Mt | Masse utilisée dans le protocole d'essai WLTP pour ce véhicule | m et Mt sont ultra corrélés (0.99). m est bien mieux renseignée que Mt. On gardera que m |
| Enedc (g/km) | Émissions de CO2 mesurées selon l'ancien cycle d'essai NEDC | Plus de 80% de Na sur l'année 2022, et n'est plus rempli depuis puisque calculé selon l'ancienne méthode. Ne peut donc être une variable cible. |
| Ernedc (g/km) | Réduction d'émissions CO2 attribuée à ces technologies innovantes selon le cycle NEDC | Colonne à 100% vide |
| De | Facteur de déviation utilisé dans l'homologation | Colonne à 100% vide |
| Vf | Facteur de vérification pour confirmer que les mesures d'émissions respectent les exigences | Colonne à 100% vide |
| Status | Statut du rapport: P = Provisional data, F = Final data. | Valeur fixe, non pertinent pour la modélisation |
| year | Année d'enregistrement des données | Valeur fixe, non pertinent pour la modélisation |

| | | |
|----------------------|---|------------------------------------|
| Date of registration | Date d'immatriculation du véhicule | Non pertinent pour la modélisation |
| ech | Caractère qui indique sous quelles dispositions le véhicule a été homologué | Colonne à 100% vide |
| RLFI | Code d'identification Roadload utilisé lors des essais pour regrouper les véhicules | Colonne à 100% vide |

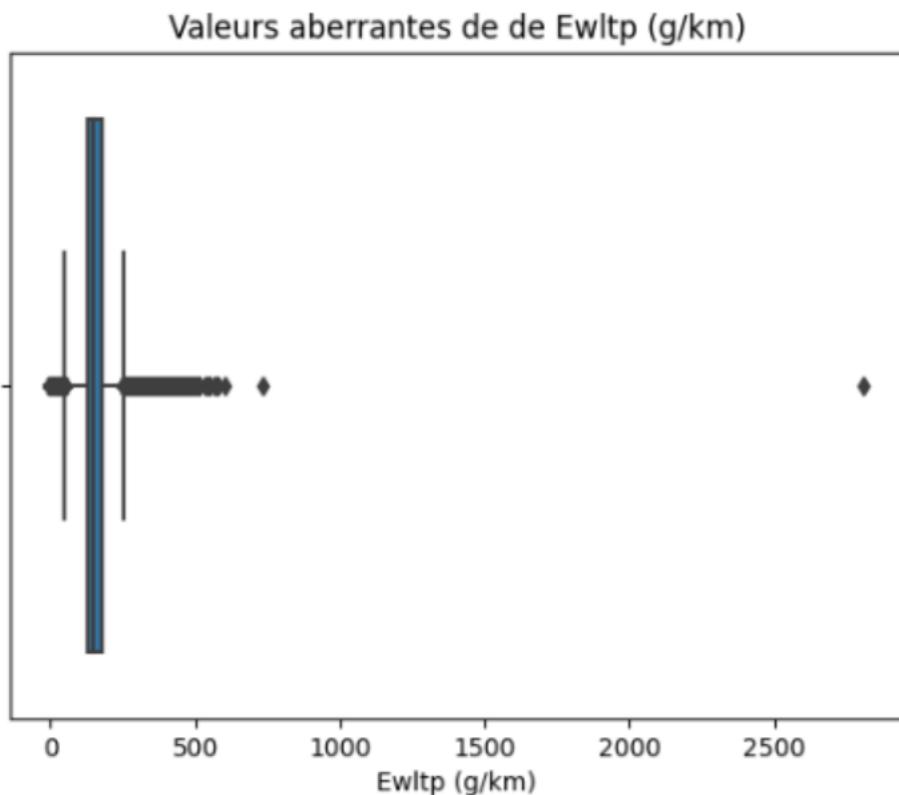
Annexe 2 – Data visualisation



- La complétude des colonnes est globalement élevée mais **certaines variables critiques (Fuel consumption, IT, Electric range, ech, z...)** montrent de fortes disparités temporelles, notamment avant 2018 et autour de 2020–2022.
- Les années récentes (2023–2024) affichent une **stabilisation mais parfois une baisse brutale** pour certaines colonnes, indiquant des ruptures de reporting ou des changements réglementaires.
- Le graphique met en évidence que le **choix de la fenêtre temporelle aura un impact majeur sur la qualité du modèle**, et justifie potentiellement une filtration par années “stables”.

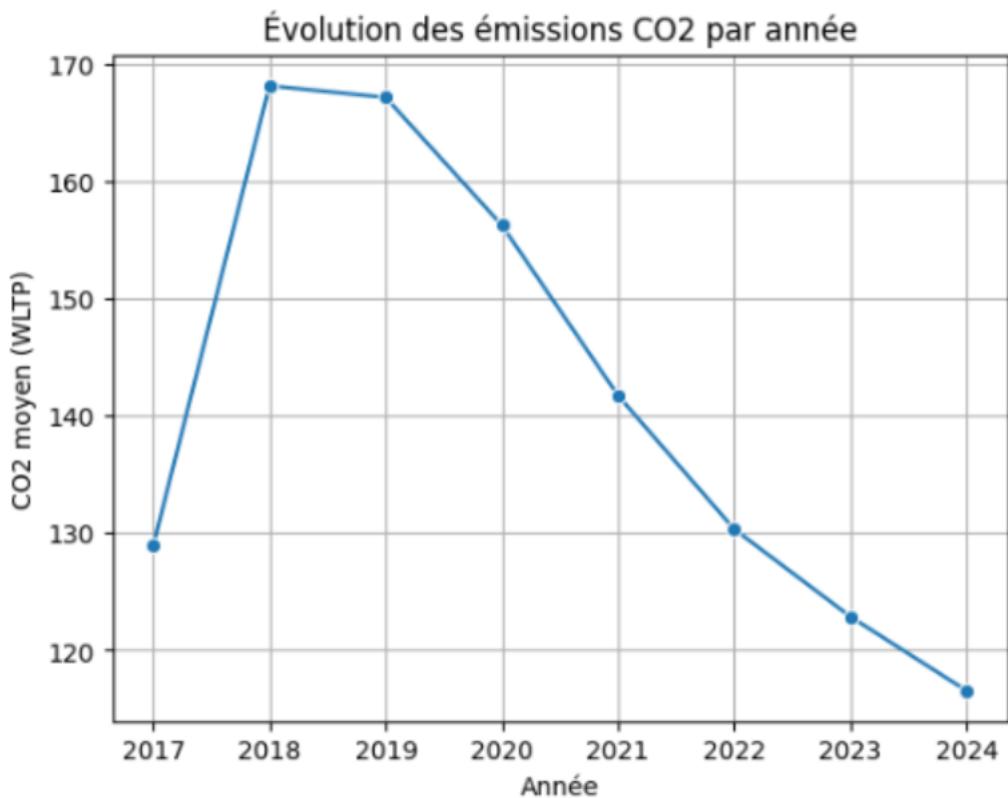


- L'Allemagne domine très largement l'échantillon, avec un volume d'enregistrements trois fois supérieur au deuxième pays.
 - Les autres pays européens (Italie, Belgique, France, Pologne, Espagne...) sont représentés de manière homogène mais nettement plus faible.
 - Cette distribution déséquilibrée implique un risque de **biais pays** qu'il faudra contrôler dans le modèle (OHE + régularisation).
-



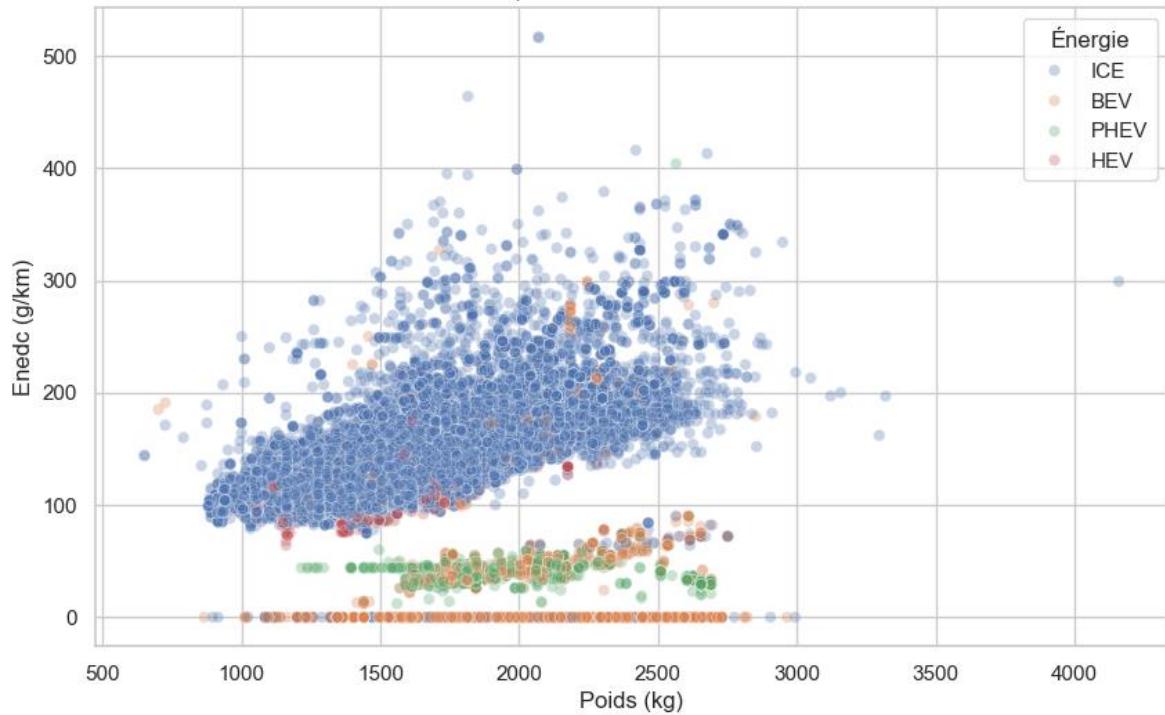
- La présence de valeurs aberrantes à droite qui fausse la distribution et ne permettent pas de tirer des conclusions correctes du boxplot

- Ces valeurs doivent être vérifiées pour corriger les erreurs de saisie pour ne pas fausser l'étude par la suite



- Un déclin continu des valeurs moyennes des émissions CO₂ à partir de 2019 qui peut être lié à l'hybridation et l'électrification des véhicules et à l'optimisation des moteurs thermiques
- La moyenne d'émission CO₂ en 2024 de 116 g/km qu'on s'approche des exigences réglementaires
- La moyenne de 2018 est faussée par la valeur la plus aberrante de WLTP=2810 g/km
- La présence de valeurs aberrantes de WLTP sur les années de 2019 à 2024 d'une moyenne de à peu près 500 g/km fausse les valeurs mais ne fausset pas la tendance globale décroissante des émissions

Relation entre le poids du véhicule et les émissions CO₂

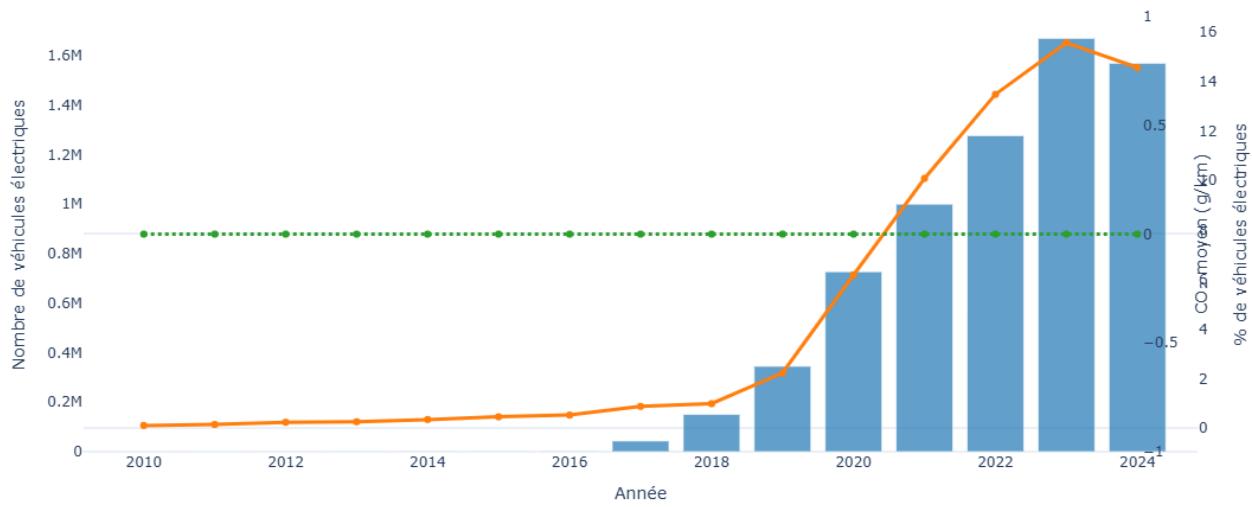


- Pour les motorisations thermiques (ICE), un **lien clair et croissant** apparaît : plus le véhicule est lourd, plus les émissions CO₂ augmentent.
- Les BEV se distinguent par des émissions nulles, tandis que les PHEV/HEV restent dans une zone faible mais légèrement corrélée au poids.
- Le graphique illustre que la **technologie énergétique influence davantage les émissions que le poids seul**, ce qui oriente fortement les choix de modélisation.
- Le graphique illustre également des incohérences dans les colonnes (véhicules électriques qui émettent du Co2 notamment). Un pré-traitement lourd sera à effectuer.

Classification énergie — Résultat provisoire

| Type énergie | Nombre | % du dataset |
|---------------------------------------|-----------|--------------|
| ICE (Thermique) | 3 518 100 | ~80.7% |
| BEV (100% électrique) | 345 039 | ~7.9% |
| PHEV (Hybride rechargeable) | 218 188 | ~5.0% |
| HEV (Hybride non-rechargeable) | 24 225 | ~0.6% |

⚡ Évolution des véhicules électriques et de leurs émissions CO₂ (2010–2024)



■ Nombre de véhicules électriques ■ % de véhicules électriques ■ CO₂ moyen (véhicules électriques)

- Le nombre de véhicules électriques explose à partir de **2019**, atteignant plus d'1,5 million en 2023–2024, signe d'une adoption massive du marché.
- La part des VE suit la même trajectoire et dépasse **15% en 2023**, confirmant une transition énergétique rapide.
- Après un premier traitement, nous retrouvons un CO₂ moyen des VE qui reste logiquement **à zéro sur toute la période**. Ce qui nous permet de retrouver une cohérence du dataset pour ces motorisations.