



Émissions CO₂ des véhicules en Europe

Alexis Decloquement | Tom Burret | Stefan Loubry | Stéphane Lable

nov25_bootcamp_da

Introduction	4
1. Cadre	6
1.1 Jeu de données initial	6
1.2 Décision de remplacer le jeu de données	6
1.3 Création de nouvelles variables	8
2. Limites et contraintes rencontrées	10
2.1 Analyse des solutions envisagées	10
2.2 Justification méthodologique du choix final	12
3. Visualisations et statistiques	12
3.1 Objectifs de la phase d'exploration et de visualisation	12
3.2 Etude des véhicules thermiques : essence et diesel	13
3.2.1 Création d'un DataFrame propre aux véhicules thermiques	13
3.2.2 Identification et nettoyage des outliers	14
3.2.3 Visualisations statistiques du lien entre les caractéristiques techniques des véhicules thermiques (essence et diesel) et leurs émissions de CO ₂	16
Matrice de corrélation	16
Graphiques en nuage de points avec droite de régression linéaire	17
3.2.4 Essence vs Diesel ?	19
3.2.5 En résumé	19
3.3 Etude des véhicules électriques	20
3.3.1 Création d'un DataFrame propre aux véhicules thermiques	20
3.3.2 Identification et nettoyage des outliers	20
3.3.3 Visualisations statistiques du lien entre les caractéristiques techniques des véhicules électriques	22
Matrice de corrélation	22
Graphiques en nuage de points avec droite de régression linéaire	24
3.3.4 En résumé	25
3.4 Etude des véhicules hybrides : essence ou diesel / électrique	26
3.4.1 Création d'un DataFrame propre aux véhicules hybrides	26
3.4.2 Identification et nettoyage des outliers	27
3.4.3 Visualisations statistiques du lien entre les caractéristiques techniques des véhicules hybrides	29
Matrice de corrélation	29
Graphiques en nuage de points avec droite de régression linéaire	30
3.4.4 En résumé	32
3.5 Etude des véhicules aux motorisations alternatives	33

3.5.1 Création d'un DataFrame propre aux véhicules à motorisations alternatives	33
3.5.2 Identification et nettoyage des outliers	34
3.5.3 Visualisations statistiques du lien entre les caractéristiques techniques des véhicules à motorisations alternatives.	35
Matrice de corrélation	35
Graphiques en nuage de points avec droite de régression linéaire	36
3.5.4 En résumé	37
3.6 Composition du parc automobile	38
3.6.1 Les étiquettes CO ₂	38
3.6.2 Les quartiles de poids	41
3.6.3 Les quartiles de puissance	43
3.7 Comparatif du parc automobile au sein de l'UE	46
3.8 Problématique	
	50
4. Modélisation	51
4.1 Introduction : Looker Studio	51
4.2 Intégration des données et contraintes de l'outil	52
4.3 Choix du thème et mises en page	54
4.4 Détail de la présentation Looker Studio	55
PAGE 1 : Problématique	55
PAGE 2 : Homologations en Europe	56
PAGE 3 : Véhicules thermiques	59
PAGE 4 : Étiquettes CO ₂	62
PAGE 5 : Électrique et efficience	65
PAGE 6 : France versus [Norvège]	
	68
5. Conclusion	70
6. Pour aller plus loin : Modélisation Machine Learning	72
7. Annexes	85
7.1 Code final	85
7.2 Lien Looker Studio	85
7.3 Lien PowerPoint	85

Introduction

Au cours des dernières années, les émissions de CO₂ des véhicules particuliers dans l'Union européenne ont suivi une tendance globale à la baisse. Cette évolution s'explique principalement par le durcissement des normes environnementales, la mise en place de politiques publiques incitatives et les progrès technologiques du secteur automobile¹.

Un tournant majeur a été l'introduction du **WLTP (Worldwide Harmonised Light Vehicles Test Procedure)** en 2017, devenu obligatoire en 2018, remplaçant l'ancien cycle **NEDC (New European Driving Cycle)** en vigueur depuis 1980, jugé peu représentatif des conditions réelles de conduite².

Ce test européen vise à mesurer et standardiser la consommation de carburant, les émissions de CO₂ et de polluants des véhicules dans des conditions plus proches de la conduite réelle. Il repose sur des cycles de conduite plus longs, des vitesses plus élevées et une prise en compte plus fine des équipements du véhicule.

Des trajectoires contrastées selon les pays :

Tous les pays européens n'évoluent pas au même rythme. Certains se démarquent par une réduction plus rapide des émissions moyennes par véhicule, généralement grâce à :

- Une adoption précoce des véhicules électriques et hybrides ;
- Une fiscalité pénalisant les motorisations thermiques ;
- Des investissements importants dans les infrastructures de recharge.

Selon **Eurostat**, en **2023**, la part combinée des véhicules électriques et hybrides dans les immatriculations neuves dépasse **50 %** dans plusieurs pays européens, avec des pics à **78 % en Finlande** et plus de **60 % en Suède et aux Pays-Bas**³.

À l'inverse, certains pays restent plus dépendants des motorisations thermiques, ce qui freine la baisse globale de leurs émissions moyennes par véhicule.

¹ Commission européenne : [Transport decarbonisation](#)

² Commission européenne : [WLTP](#)

³ Eurostat : [New car registrations](#)

Impact de l'émergence des véhicules électriques et hybrides :

L'essor des véhicules électriques constitue un levier majeur de réduction des émissions de CO₂ :

- Ils affichent **zéro émission de CO₂ à l'usage** ;
- Leur part croissante dans les immatriculations neuves fait mécaniquement baisser les émissions moyennes ;
- L'effet est particulièrement visible dans les pays où leur adoption est rapide et soutenue.

D'après l'**ACEA (Association des constructeurs européens d'automobiles)**, les **véhicules 100 % électriques (BEV)** représentaient **14,6 % des immatriculations neuves en Union européenne en 2023⁴**, contre moins de 10 % en 2021.

Par ailleurs, l'**Agence européenne pour l'environnement (EEA)** indique que plus de **2,4 millions de voitures électriques** ont été immatriculées en 2024 dans l'UE, contre environ **1 million en 2020**, illustrant la forte accélération de l'électrification du parc automobile⁵.

Cependant, l'impact global sur les émissions de CO₂ dépend fortement :

- De la **vitesse de renouvellement du parc automobile**, un véhicule thermique pouvant rester en circulation plus de 10 à 15 ans ;
- De la **part réelle des véhicules électriques dans les nouvelles immatriculations**, qui varie fortement selon les pays.

D'un point de vue analytique, la baisse observée ne résulte pas d'un facteur unique, mais d'une **combinaison de leviers** :

- Réglementation ;
- Évolution de l'offre automobile ;
- Comportements des consommateurs.

Les données mettent ainsi en évidence une **transition engagée mais inégale**, avec des pays moteurs accélérant la décarbonation du transport routier, tandis que d'autres accusent encore un retard.

⁴ ACEA : [New car registrations](#)

⁵ EEA : [New registrations of electric cars in Europe](#)

1. Cadre

1.1 Jeu de données initial

Jeu de données provenant du site <https://co2cars.apps.eea.europa.eu/> pour l'année 2024 recensement des infos des véhicules neufs homologations en 2024 dans toutes l'Europe.

Données brutes : 10,779,681 lignes pour 40 colonnes pour 3.31Gb

# ID	Country	VFN	Mp	Mh	Man	MMS	Tan	T	Va	Ve	Mk	Cn
148534149	NL	IP-C519_2022_00007-WF0-1	FORD	FORD WERKE GMBH	FORD WERKE GMBH	Missing value	e13*2007/46*1911*17	DEH	R0DE1PX	SAPCENASIBS	FORD	FOCUS
148534150	NL	IP-ZKU_REDUTG61-VF3	STELLANTIS	AUTOMOBILES PEUGEOT	AUTOMOBILES PEUGEOT	Missing value	e2*2007/46*0532*21	V	S	ZKUZ-37B0GN	OPEL	VIVARO
148534151	NL	IP-2023_225HD-YV1-1	VOLVO CARS POLESTAR SUZUKI	VOLVO	VOLVO CAR CORPORATION	Missing value	e4*2007/46*1315*21	Z	ZSH4	ZSH4VD07	VOLVO	S60
148534152	NL	IP-C519_2022_00007-WF0	FORD	FORD WERKE GMBH	FORD WERKE GMBH	Missing value	e13*2007/46*1911*18	DEH	R0DE1PX	SAPCENASIBS	FORD	FOCUS
148534153	NL	IP-0165-JT1	SUBARU-MAZDA-TOYOTA	TOYOTA MOTOR CORPORATION	TOYOTA MOTOR CORPORATION	Missing value	e6*2018/85*0006*0*03	AB7UP)	KGB70(H)	KGB70L-AHMNKW(2D)	TOYOTA	TOYOTA AVX GO X
148534154	NL	IP-FJB1MDP748_000-UU1-0	RENAULT-NISSAN-MITSUBISHI	DACIA	AUTOMOBILE DACIA SA	Missing value	e19*2007/46*0026*22	DIF	BEV	MD6UA4GMS2V0	DACIA	SANDERO
148534155	NL	IP-MQB7SZ_A0_1062-WWW	VOLKSWAGEN	VOLKSWAGEN AG	VOLKSWAGEN AG	Missing value	e1*2018/85*0030*2*04	CT	ACDUCBX0	FD6FD6DD002PR4WB0	VOLKSWAGEN	TIGUAN
148534156	NL	IP-MQB7SZ_A0_1062-WWW	VOLKSWAGEN	VOLKSWAGEN AG	VOLKSWAGEN AG	Missing value	e13*2018/85*0014*0*08	CS	ACDUS8	FMSFM5DF0154WEA1	VOLKSWAGEN	TAIGO
148534157	NL	IP-2023_416EMULH827-YV1-1	VOLVO CARS POLESTAR SUZUKI	VOLVO	VOLVO CAR CORPORATION	Missing value	e9*2007/46*3146*18	X	XKEH	XKEHRL07	VOLVO	C40
148534158	NL	IP-2023_246HD-YV1-1	VOLVO CARS POLESTAR SUZUKI	VOLVO	VOLVO CAR CORPORATION	Missing value	e4*2007/46*1220*22	U	UZH4	UZHV4D07	VOLVO	XC60
148534159	NL	IP-2023_416EMULH827-YV1-1	VOLVO CARS POLESTAR SUZUKI	VOLVO	VOLVO CAR CORPORATION	Missing value	e9*2007/46*0684*28	Z	2ZEM	2ZEM1L07	VOLVO	EX30
148534160	NL	IP-JBB1N4PDB1A_000-VF1	RENAULT-NISSAN-MITSUBISHI	RENAULT	RENAULT SAS	Missing value	e2*2007/46*0684*28	RJB	HH2	N44WB22A500B	RENAULT	CAPTUR E-TECH I
148534161	NL	IP-FD-INSP0010_000-UU1	RENAULT-NISSAN-MITSUBISHI	DACIA	AUTOMOBILE DACIA SA	Missing value	e19*2007/46*0026*23	DIF	RHS	N64WA2P5M500B	DACIA	JOGGER
148534162	NL	IP-2021_225KF-YV1-1	VOLVO CARS POLESTAR SUZUKI	VOLVO	VOLVO CAR CORPORATION	Missing value	e4*2007/46*1315*21	Z	ZSK8	ZSK8VR07	VOLVO	S60
148534163	NL	IP-09_3142Z-SV1	TESLA	TESLA INC	TESLA INC	Missing value	e4*2018/85*0013*13	005	YTCR	PJB35573X	TESLA	MODEL Y
148534164	NL	IP-0931280-KMH-1	HYUNDAI MOTOR EUROPE	HYUNDAI	HYUNDAI MOTOR COMPANY	Missing value	e9*2018/85*1105*4*04	NE	FSE62	E11A11	HYUNDAI	IONIQ5
148534165	NL	IP-0931273-KNA-1	KIA	KIA	KIA CORPORATION	Missing value	e9*2018/85*1124*1*00	SG2	CSPI11	D61AY1	KIA	NIRO
148534166	NL	IP-MQB37AS_B1_1941-TMB	VOLKSWAGEN	SKODA	SKODA AUTO AS	Missing value	e8*2018/85*0017*03	PS	ACDUCBX0	FD6FD6DD0024WE0551A	SKODA	KODIAQ
148534167	NL	IP-HPX_EDC_5B86-ZAC	STELLANTIS	STELLANTIS EUROPE	STELLANTIS EUROPE SPA	Missing value	e3*2018/85*0007*06	FH1	2AG172AA3	1E6ANE1158	FIAT	600
148534168	NL	IP-MQB7ZZ_A0_0529-VSS-1	VOLKSWAGEN	SEAT	SEAT SA	Missing value	e9*2007/46*3134*42	KJ	BDLAC	FMSFM5DF0084B1AAF	SEAT	IBIZA
148534169	NL	IP-ZKX_REDUT761-YAR-0	SUBARU-MAZDA-TOYOTA	TOYOTA	TOYOTA MOTOR EUROPE NV SA	Missing value	e2*2007/46*0537*19	V	Z	ZKZC-P780N(1V)	TOYOTA	PROACE TAXI
148534170	NL	IP-2023_0407-WTK-1	MERCEDES-BENZ AG	MERCEDES-BENZ AG	MERCEDES-BENZ AG	Missing value	e1*2018/85*0001*7*09	R2CS	H05D0	CZA050A	MERCEDES-BEN	C 300 E
148534171	NL	IP-BAT1MP74A_001-VF1-1	RENAULT-NISSAN-MITSUBISHI	RENAULT	RENAULT SAS	Missing value	e2*2007/46*0684*24	RJB	HE2	M66UALLA5300	RENAULT	CAPTUR
148534172	NL	IP-HNE_MB6D54ZC-V83	STELLANTIS	STELLANTIS AUTO	STELLANTIS AUTO SAS	Missing value	e2*2007/46*0639*28	U	P	HNEK-X1T500	PEUGEOT	208
148534173	NL	IP-JAB1N8H0010_000-VF1	RENAULT-NISSAN-MITSUBISHI	RENAULT	RENAULT SAS	Missing value	e2*2007/46*0676*28	RJA	BH2	N84WA3VA5308	RENAULT	CLIO E-TECH HYB

10,779,681 rows x 40 cols 25 per page < < Page 1 of 431188 > >

'ID', 'Country', 'VFN', 'Mp', 'Mh', 'Man', 'MMS', 'Tan', 'T', 'Va', 'Ve', 'Mk', 'Cn', 'Ct', 'Cr', 'r', 'm (kg)', 'Mt', 'Enedc (g/km)', 'Ewltp (g/km)', 'W (mm)', 'At1 (mm)', 'At2 (mm)', 'Ft', 'Fm', 'ec (cm3)', 'ep (KW)', 'z (Wh/km)', 'IT', 'Ernedc (g/km)', 'Erwltp (g/km)', 'De', 'Vf', 'Status', 'year', 'Date of registration', 'Fuel consumption ', 'ech', 'RLF1', 'Electric range (km)'

Gestion des données manquantes :

Création d'une fonction permettant l'analyse des valeurs manquantes d'un DataFrame. Le but étant de parcourir chaque colonne, identifier celles contenant des valeurs manquantes et afficher, pour chacune, le nombre et le pourcentage de valeurs manquantes. Enfin, il indique le nombre total et la proportion de colonnes concernées, afin d'évaluer rapidement la qualité et la complétude des données avant le prétraitement.

Sur les 40 colonnes initiales 30 ont des valeurs manquantes soit 75% du dataset.

1.2 Décision de remplacer le jeu de données

A la suite de l'exploration des variables plusieurs anomalies ont été détectées :

- Noms de constructeurs mal orthographiés ;

- Noms des variables contenant des espaces :



```
1 data.columns.str.strip()
```

- Noms de constructeurs en format numérique ;
- Noms des pays abrégés ;
- Variables numériques avec des valeurs manquantes pouvant impacter les analyses – Remplacement par des 0

Définition d'une liste de colonnes jugées non pertinentes ou redondantes pour l'analyse, puis les suppressions de celles-ci. L'objectif étant d'alléger le dataset tout en conservant les variables utiles pour l'analyse et l'exploration.

Liste des colonnes supprimées :

Enedc (g/km), W (mm), At1 (mm), At2 (mm), Ernedc (g/km), De, Vf, Ct, Cr, r, MMS, Mh, VFN, Tan, Ve, Status, RLFT, ech, Va, T, Mp, Man, Fm, ec (cm3), IT, Erwltp (g/km), year, Date of registration.

Définition d'un dictionnaire de correspondance pour renommer certaines colonnes du dataset avec des plus explicites et lisibles. Permettant ainsi l'harmonisation des variables pertinentes, comme le poids, les constructeurs, les émissions de CO2, le type de carburants, la puissance, et la consommation électrique, mais aussi la compréhension lors de l'analyse exploratoire.

Création d'une fonction permettant le nettoyage des noms de constructeurs. Définition des règles de correspondance entre des préfixes abrégés et des noms complets de constructeurs. Pour chaque valeur de la colonne *Constructeur*, la fonction remplace les noms abrégés ou incomplets par un nom standardisé, tout en conservant les valeurs manquantes. L'objectif est d'éviter les doublons et les incohérences.

Code permettant de standardiser manuellement, au fil de l'exploration des noms de constructeurs non vus jusque là. Ajout dans un dictionnaire de l'orthographe le plus pertinent pour un constructeur donné.

Idem pour les abréviations des pays, convertis en nom complet pour éviter les incompréhensions lors des analyses.

Harmonisation des types de carburants pour une lecture plus précise.

1.3 Création de nouvelles variables

Variable Étiquettes CO₂ :

Création d'une nouvelle variable catégorielle appelée *Etiquette_CO2* en classant les émissions de CO₂ (WLTP) en intervalles prédéfinis, allant de la classe **A** (véhicules très peu émetteurs) à la classe **G** (véhicules très fortement émetteurs). Chaque véhicule est ainsi associé à une étiquette en fonction de son niveau d'émission.

Cette transformation est particulièrement utile pour :

- réaliser des **graphiques plus lisibles** (comparaisons par classes plutôt que valeurs continues),
- **comparer des groupes de véhicules** entre pays, constructeurs ou motorisations,
- faciliter une **lecture non technique** des résultats, notamment pour un public non expert,
- rapprocher l'analyse des **logiques réglementaires et des labels environnementaux**.

Variable WLTP_Poids :

Calcul des **quartiles** du poids des véhicules (**WLTP_poids**), puis crée une nouvelle variable catégorielle qui classe chaque véhicule selon son niveau de poids. Les véhicules sont ainsi répartis en quatre groupes : **léger, moyen, lourd et très lourd**, en fonction de leur position dans la distribution des poids.

Cette transformation est utile pour :

- simplifier l'analyse d'une variable continue en classes interprétables,
- réaliser des comparaisons plus pertinentes entre catégories de véhicules,
- produire des graphiques plus lisibles (boxplots, barplots, comparaisons par groupes),
- analyser l'impact du poids sur les émissions de CO₂ de manière plus pédagogique.

Variables Puissance kW :

Calcul des quartiles de la puissance moteur (**Puissance_KW**), puis classe chaque véhicule dans une catégorie de puissance : **faible, moyenne, puissante ou très puissante**, selon sa position dans la distribution des puissances.

Suppression des incohérences métier :

Certains véhicules ont des caractéristiques techniques contradictoires avec leur type de motorisation.

Plus précisément :

- Suppression des véhicules **thermiques ou hybrides** ayant une **consommation de carburant nulle**, ce qui est irréaliste.
- Suppression des véhicules **électriques ou hybrides** ayant une **autonomie électrique nulle**.
- Suppression des véhicules **électriques ou hybrides** dont la **consommation électrique est nulle**.

Dataset après nettoyage et ajout des nouvelles variables :

Constructeur	Model	# WLTP ...	# Co2_Emission(WLTP)	# Type_Carburant	# Puissance_KW	# Consommation_Wh/km	# Fuel consumption	# Electric range (km)	# Pays	# Etiquette_CO2	# Poids_Quartile	# PuissanceKW_Quartile	# Cout(€)
FORD	FOCUS	1523.0	121.0	Essence	92.0	0.0	5.3	0.0	Netherlands	C	Q2,Moyen	Q2,Moyen	230.0
OPEL	VIVARO	2454.0	0.0	Electric	100.0	244.0	0.0	343.0	Netherlands	A	Q4,Très_lourd	Q2,Moyen	N/C
VOLVO	S60	2176.0	16.0	Hybride_Essence	186.0	163.0	0.7	93.0	Netherlands	A	Q4,Très_lourd	Q4,Très_Puissant	N/C
FORD	FOCUS	1549.0	123.0	Essence	92.0	0.0	5.4	0.0	Netherlands	C	Q2,Moyen	Q2,Moyen	260.0
TOYOTA	TOYOTA AYGO X	1088.0	108.0	Essence	53.0	0.0	4.8	0.0	Netherlands	B	Q1,Léger	Q1,Faible	N/C
DACIA	SANDERO	1262.0	125.0	Essence	81.0	0.0	5.5	0.0	Netherlands	C	Q1,Léger	Q2,Moyen	310.0
VOLKSWAGEN	TIGUAN	2033.0	101.0	Hybride_Essence	110.0	178.0	0.4	117.0	Netherlands	A	Q4,Très_lourd	Q3,Puissant	N/C
VOLKSWAGEN	TAIGO	1335.0	125.0	Essence	70.0	0.0	5.6	0.0	Netherlands	C	Q1,Léger	Q1,Faible	310.0
VOLVO	C40	2293.0	0.0	Electric	185.0	166.0	0.0	570.0	Netherlands	A	Q4,Très_lourd	Q4,Très_Puissant	N/C
VOLVO	XC60	2272.0	22.0	Hybride_Essence	186.0	182.0	0.9	82.0	Netherlands	A	Q4,Très_lourd	Q4,Très_Puissant	N/C
VOLVO	EX30	1937.0	0.0	Electric	200.0	172.0	0.0	334.0	Netherlands	A	Q4,Très_lourd	Q4,Très_Puissant	N/C
RENAULT	CAPTUR E-TECH HYBRID	1555.0	105.0	Essence	69.0	0.0	4.5	0.0	Netherlands	B	Q2,Moyen	Q1,Faible	N/C
DACIA	JOGGER	1522.0	101.0	Essence	69.0	0.0	4.7	0.0	Netherlands	B	Q2,Moyen	Q1,Faible	N/C
VOLVO	S60	1826.0	139.0	Essence	145.0	0.0	6.1	0.0	Netherlands	C	Q3,lourd	Q4,Très_Puissant	1386.0
TESLA	MODEL Y	2095.0	0.0	Electric	235.0	157.0	0.0	455.0	Netherlands	A	Q4,Très_lourd	Q4,Très_Puissant	N/C
HYUNDAI	IONIQ5	2165.0	0.0	Electric	168.0	170.0	0.0	507.0	Netherlands	A	Q4,Très_lourd	Q4,Très_Puissant	N/C
KIA	NIRO	1565.0	100.0	Essence	77.0	0.0	4.4	0.0	Netherlands	A	Q2,Moyen	Q2,Moyen	N/C
SKODA	KODIAQ	2061.0	10.0	Hybride_Essence	110.0	176.0	0.5	114.0	Netherlands	A	Q4,Très_lourd	Q3,Puissant	N/C
FIAT	600	1479.0	105.0	Essence	74.0	0.0	4.9	0.0	Netherlands	B	Q2,Moyen	Q1,Faible	N/C
SEAT	IBIZA	1243.0	120.0	Essence	70.0	0.0	5.3	0.0	Netherlands	B	Q1,Léger	Q1,Faible	210.0
TOYOTA	PROACE TAXI	2489.0	0.0	Electric	1000.0	273.0	0.0	310.0	Netherlands	A	Q4,Très_lourd	Q2,Moyen	N/C
MERCEDES-BENZ	C 300 E	2305.0	14.0	Hybride_Essence	150.0	191.0	0.6	110.0	Netherlands	A	Q4,Très_lourd	Q4,Très_Puissant	N/C
RENAULT	CAPTUR	1383.0	131.0	Essence	67.0	0.0	5.8	0.0	Netherlands	C	Q2,Moyen	Q1,Faible	650.0
PEUGEOT	208	1249.0	113.0	Essence	74.0	0.0	5.1	0.0	Netherlands	B	Q1,Léger	Q1,Faible	50.0
RENAULT	CLIO E-TECH HYBRID	1456.0	97.0	Essence	69.0	0.0	4.3	0.0	Netherlands	A	Q2,Moyen	Q1,Faible	N/C

10,117,232 rows x 15 cols 25 per page

<< < Page 1 of 404690 > >>

...

2. Limites et contraintes rencontrées

Dans le cadre de ce projet d'analyse des émissions de CO₂ des véhicules en Europe, le jeu de données initial comportait **10 779 681 lignes pour 40 colonnes**, représentant un volume total d'environ **3,31 Go**.

Un tel volume de données est parfaitement exploitable dans des environnements de calcul adaptés (bases de données relationnelles, moteurs distribués, cloud), mais pose des **contraintes importantes** lorsqu'il s'agit de l'intégrer dans un outil de visualisation tel que **Looker Studio**.

En effet, Looker Studio impose une **limite stricte de 100 Mo par fichier CSV importé**, ce qui rend impossible l'utilisation directe du dataset complet, même après un nettoyage préalable des colonnes. Cette contrainte technique a donc nécessité une **réflexion méthodologique** afin de trouver une solution permettant de conserver la valeur analytique des données tout en respectant les limitations de l'outil.

2.1 Analyse des solutions envisagées

Deux solutions principales ont été étudiées :

Solution 1 : Utilisation d'une base de données BigQuery

La première option consistait à importer le dataset complet dans **BigQuery**, puis à connecter Looker Studio directement à cette base de données via des requêtes SQL. Cette approche présentait plusieurs avantages :

- Possibilité de conserver **l'intégralité du jeu de données** ;
- Requêtes performantes sur de très grands volumes ;
- Filtrage et agrégation directement côté base.

Cependant, cette solution a été **écartée** pour des raisons principalement économiques. L'utilisation de BigQuery engendre des **coûts liés au stockage et aux requêtes**, estimés à environ **300 €** pour ce projet, ce qui n'était pas compatible avec le cadre académique et budgétaire de l'étude.

Solution 2 : Échantillonnage du jeu de données

La seconde solution retenue reposait sur la création d'un **échantillon représentatif** du dataset complet. Cette méthode consiste à sélectionner un sous-ensemble de lignes de manière aléatoire, tout en conservant une taille suffisante pour garantir la robustesse statistique des analyses. L'échantillonnage a été réalisé à l'aide de la fonction suivante :

```
sample = data.sample(n=900000, random_state=42)
```

Dans cette expression :

- `n = 900000` correspond au nombre de lignes conservées ;
- `random_state = 42` fixe la graine aléatoire afin de garantir la **reproductibilité** des résultats.

Le seuil de **900 000 lignes** a été défini comme un compromis optimal :

- Suffisamment élevé pour préserver la richesse et la diversité des données ;
- Suffisamment réduit pour respecter la limite de taille imposée par Looker Studio après conversion et export.

Impact de l'échantillonnage sur la qualité des données

L'un des principaux enjeux de l'échantillonnage est le **risque de perte de représentativité**. En réduisant le nombre d'observations, il est légitime de s'interroger sur la capacité de l'échantillon à refléter fidèlement les caractéristiques du jeu de données initial.

Dans ce projet, cet impact est limité grâce à l'application de la **loi des grands nombres**, un principe fondamental en statistiques. Cette loi stipule que lorsque la taille d'un échantillon augmente, les statistiques calculées à partir de celui-ci (moyennes, proportions, distributions) tendent à se rapprocher des valeurs réelles de la population.

Autrement dit, plus un échantillon est volumineux, plus il devient **stable et représentatif**. Avec **900 000 lignes**, soit environ **8,3 % du dataset initial**, l'échantillon reste très large d'un point de vue statistique. Cette taille permet de conserver des distributions proches de celles observées dans le jeu de données complet, notamment pour les variables majeures telles que :

- Le constructeur ;
- Le type de carburant ;
- Le pays d'immatriculation ;
- Les émissions moyennes de CO₂ ;
- Le poids et la puissance des véhicules.

Dataset après le sample de 900 000 lignes :

Constructeur	Modèle	# WTP_poids	# Co2_Emission(WTP)	# Type_Carburant	# Puissance_kw	# Consommation_kw/h/km	# Fuel consumption	# Electric range (km)	# Pays	# Etiquette_CO2	# Poids_Quartile	# Puissance_kw_Quartile	# Cont(E)
RENAULT	MEGANE	1672.0	123.0	Diesel	85.0	0.0	7.3	0.0	Romania	C	Q3_Lourd	Q2_Moyen	260.0
RENAULT	CLIO	1290.0	121.0	Essence	67.0	0.0	5.4	0.0	Italy	C	Q1_Léger	Q1_Faible	230.0
PEUGEOT	2008	1438.0	132.0	Diesel	96.0	0.0	5.0	0.0	Italy	C	Q2_Moyen	Q2_Moyen	240.0
TOYOTA	TOYOTA YARIS	1280.0	87.0	Essence	68.0	0.0	3.8	0.0	France	A	Q1_Faible	Q1_Faible	0.0
DACIA	SANDERO	1215.0	120.0	Essence	49.0	0.0	5.3	0.0	France	B	Q1_Léger	Q1_Faible	210.0
PEUGEOT	3008	1797.0	125.0	Essence	100.0	0.0	5.5	0.0	France	C	Q3_Lourd	Q2_Moyen	310.0
KIA	SPORTAGE	2681.0	25.0	Hybride_Essence	132.0	16.0	1.1	66.0	Sweden	A	Q4_Très_Jourd	Q4_Très_Puissant	0.0
PEUGEOT	208	1610.0	0.0	Electric	100.0	155.0	0.0	362.0	France	A	Q3_Lourd	Q2_Moyen	0.0
NISSAN	PRIMASTAR	2240.0	186.0	Diesel	110.0	0.0	7.1	0.0	Spain	E	Q4_Très_Jourd	Q3_Puissant	48901.0
DACIA	SANDERO	1285.0	106.0	Autre	74.0	0.0	6.6	0.0	France	B	Q1_Léger	Q1_Faible	0.0
KIA	EV3	1896.0	0.0	Electric	150.0	149.0	0.0	435.0	Sweden	A	Q4_Très_Jourd	Q4_Très_Puissant	0.0
OMODA	OMODAS	1599.0	168.0	Essence	108.0	0.0	7.1	0.0	Poland	E	Q3_Lourd	Q3_Puissant	11803.0
MITSUBISHI	ASX	1729.0	30.0	Hybride_Essence	68.0	133.0	1.3	48.0	Germany	A	Q3_Lourd	Q1_Faible	0.0
DACIA	SANDERO	1236.0	119.0	Essence	67.0	0.0	5.3	0.0	France	B	Q1_Léger	Q1_Faible	190.0
DACIA	DUSTER	1440.0	126.0	Autre	74.0	0.0	6.5	0.0	Italy	C	Q2_Moyen	Q1_Faible	330.0
FORD	KUGA	1671.0	154.0	Essence	137.0	0.0	6.8	0.0	Croatia	D	Q3_Lourd	Q4_Très_Puissant	4026.0
CUPRA	FORMENTOR	1593.0	155.0	Essence	110.0	0.0	6.9	0.0	Germany	D	Q3_Lourd	Q3_Puissant	4279.0
SKODA	SUPERB	1809.0	130.0	Essence	110.0	0.0	5.7	0.0	Germany	C	Q3_Lourd	Q3_Puissant	540.0
PEUGEOT	208	1249.0	116.0	Essence	74.0	0.0	5.2	0.0	France	B	Q1_Léger	Q1_Faible	125.0
SUZUKI	SX4	1515.0	131.0	Essence	75.0	0.0	5.8	0.0	Czech Republic	C	Q2_Moyen	Q2_Moyen	650.0
MITSUBISHI	MITSUBISHI SPACE STAR	1020.0	112.0	Essence	52.0	0.0	4.9	0.0	Netherlands	B	Q1_Léger	Q1_Faible	0.0
RENAULT	CLIO E-TECH HYBRID	1470.0	300.0	Essence	69.0	0.0	4.3	0.0	France	A	Q2_Moyen	Q1_Faible	0.0
SKODA	MARCO	1507.0	137.0	Essence	110.0	0.0	6.0	0.0	Czech Republic	C	Q3_Puissant	Q3_Puissant	1100.0
OPEL	ASTRA SPORTS TOURER	1590.0	175.0	Essence	56.0	0.0	3.5	0.0	Poland	B	Q1_Léger	Q2_Moyen	190.0
DACIA	SPRING	1109.0	0.0	Electric	16.0	145.0	0.0	220.0	Spain	A	Q1_Léger	Q1_Faible	0.0

2.2 Justification méthodologique du choix final

Le choix de l'échantillonnage s'inscrit donc dans une **démarche pragmatique et méthodologiquement justifiée**, prenant en compte à la fois :

- Les **contraintes techniques** de l'outil de visualisation ;
- Les **limites économiques** du projet ;
- Les **exigences statistiques** liées à la qualité de l'analyse.

Dans le cadre de ce projet, l'objectif principal étant l'analyse des **tendances globales**, des **comparaisons entre pays** et des **effets de l'électrification du parc automobile**, l'échantillon retenu offre un niveau de précision largement suffisant. Il permet de produire des visualisations fiables et interprétables, tout en assurant des temps de chargement raisonnables et une intégration fluide dans Looker Studio.

En conclusion, la décision de travailler sur un **échantillon de 900 000 lignes** constitue un **compromis équilibré entre faisabilité technique et rigueur analytique**, garantissant la pertinence des résultats sans compromettre la qualité statistique des analyses menées.

3. Visualisations et statistiques

3.1 Objectifs de la phase d'exploration et de visualisation

Une fois la phase de pré-processing réalisée, nous avons pu conduire une analyse exploratoire approfondie de l'ensemble des variables retenues, ainsi que de leurs

interactions, dans le but d'évaluer leur influence sur les émissions de CO₂ des véhicules étudiés. Cette approche visait à mettre en évidence des corrélations, des relations statistiques et des tendances globales, permettant de faire émerger une problématique propre à notre jeu de données.

Les questions de recherche s'articulaient autour des axes suivants :

- évaluer l'impact relatif du poids, de la puissance et de la consommation de carburant sur les émissions de CO₂ des véhicules à motorisation thermique et hybride ;
- analyser l'influence du poids sur la consommation énergétique et l'autonomie des véhicules électriques ;
- comprendre la composition du parc automobile en fonction des émissions de CO₂, du poids des véhicules et du type de motorisation.

Afin d'optimiser l'organisation du travail et d'assurer une analyse cohérente, les explorations ont été réparties par type de motorisation :

- véhicules à motorisation exclusivement thermique (essence et diesel) ;
- véhicules à motorisation exclusivement électrique ;
- véhicules hybrides (essence ou diesel / électrique) ;
- véhicules à motorisations alternatives (ex : hydrogène ou GPL), quantitativement minoritaires. Nous les avons regroupées sous le label “Autre”.

Enfin, une analyse comparative à l'échelle des pays de l'Union européenne a été menée afin d'identifier d'éventuelles différences dans la composition du parc automobile. Cette analyse visait à évaluer l'influence potentielle des cadres réglementaires nationaux sur la structure et la répartition des motorisations au sein du parc automobile européen.

3.2 Etude des véhicules thermiques : essence et diesel

3.2.1 Crédation d'un DataFrame propre aux véhicules thermiques

Dans un premier temps, nous avons constitué un dataframe regroupant uniquement les véhicules thermiques (essence et diesel). Cette opération sera répétée pour chacune des motorisations précédemment identifiées. Pour les véhicules thermiques, les variables relatives à la consommation et à l'autonomie électrique, non pertinentes dans ce contexte, ont été supprimées.

3.2.2 Identification et nettoyage des outliers

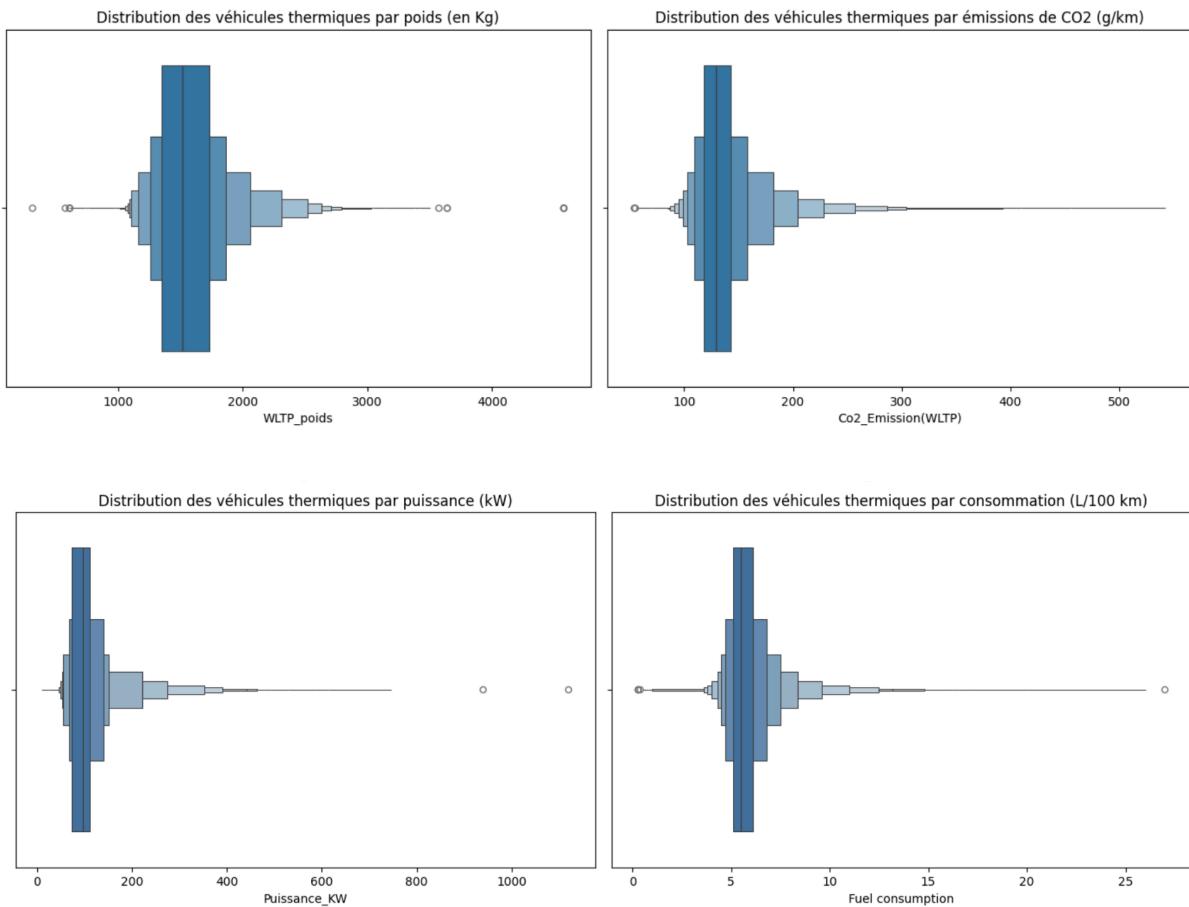
Dans un second temps, nous avons dressé la liste des variables numériques à étudier en affichant une première description de leur distribution respective :

- le poids (“WLTP_poids”),
- les émissions de CO₂ (“Co2_Emission(WLTP”)),
- la puissance (“Puissance_KW”),
- la consommation de carburant (“Fuel consumption”).

	WLTP_poids	Co2_Emission(WLTP)	Puissance_KW	Fuel consumption
count	7.752305e+06	7.752305e+06	7.752305e+06	7.752305e+06
mean	1.563257e+03	1.343820e+02	1.020938e+02	5.753485e+00
std	3.054901e+02	2.886568e+01	4.838072e+01	1.226213e+00
min	3.170000e+02	5.500000e+01	1.000000e+01	3.000000e-01
25%	1.355000e+03	1.180000e+02	7.300000e+01	5.100000e+00
50%	1.520000e+03	1.290000e+02	9.600000e+01	5.500000e+00
75%	1.734000e+03	1.430000e+02	1.100000e+02	6.100000e+00
max	4.574000e+03	5.430000e+02	1.120000e+03	2.700000e+01

Tableau descriptif de la distribution des variables pour les véhicules thermiques

Dans un troisième temps, nous avons étudié les distributions plus en détail de ces variables afin d'identifier la présence de potentiels outliers grâce à des graphiques en boîtes à moustaches.



Analyse des boîtes à moustache et nettoyage des outliers

Poids : plusieurs valeurs aberrantes sont observées au-delà de 3 500 kg. Après vérification à l'aide de sources externes, il apparaît probable qu'il s'agisse d'erreurs de mesure ou de saisie. Ces observations ont donc été supprimées du jeu de données.

Émissions de CO₂ : certaines valeurs inférieures à 50 g de CO₂ émis par kilomètre ont été identifiées comme aberrantes et ont été éliminées.

Puissance : quelques véhicules présentent des puissances proches, voire supérieures, à 1 000 kW. Une analyse plus approfondie montre qu'il s'agit principalement de véhicules de type *supercars* (par exemple, la marque Koenigsegg). Ces valeurs sont extrêmes, mais cohérentes avec la réalité, elles ont alors été conservées.

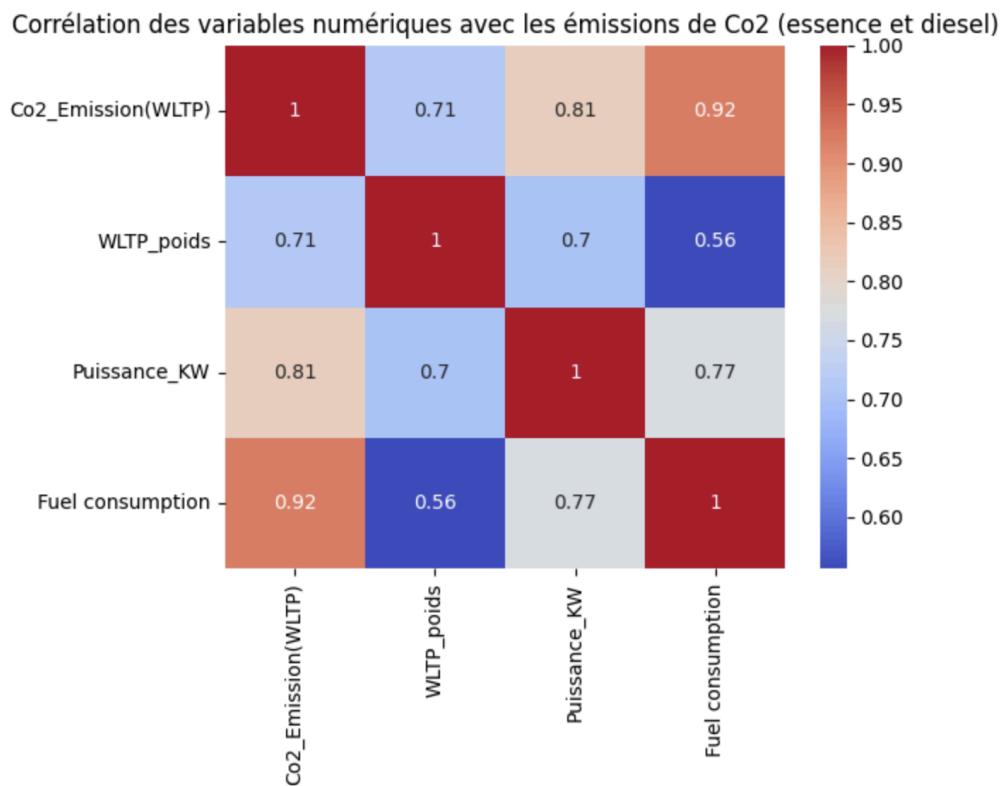
Consommation de carburant : des valeurs aberrantes ont été détectées au-delà de 25 L/100 km et en-deçà de 3 L/100 km. Afin de préserver la qualité et la robustesse des prochaines analyses, ces observations ont été supprimées.

À l'issue de ce processus de nettoyage, le jeu de données est désormais prêt pour une analyse approfondie des différentes variables.

3.2.3 Visualisations statistiques du lien entre les caractéristiques techniques des véhicules thermiques (essence et diesel) et leurs émissions de CO₂

Matrice de corrélation

Dans un premier temps, une matrice de corrélation, représentée sous la forme d'une *heatmap*, a été construite afin de mettre en évidence de manière explicite les interactions entre les variables de poids, de consommation et de puissance, ainsi que leur relation avec les émissions de CO₂.



Observations

À partir de la **heatmap de corrélation**, nous observons une très forte corrélation positive entre les émissions de CO₂ et la consommation de carburant ($r = 0,92$), la puissance du moteur ($r = 0,81$), ainsi que le poids des véhicules analysés ($r = 0,71$).

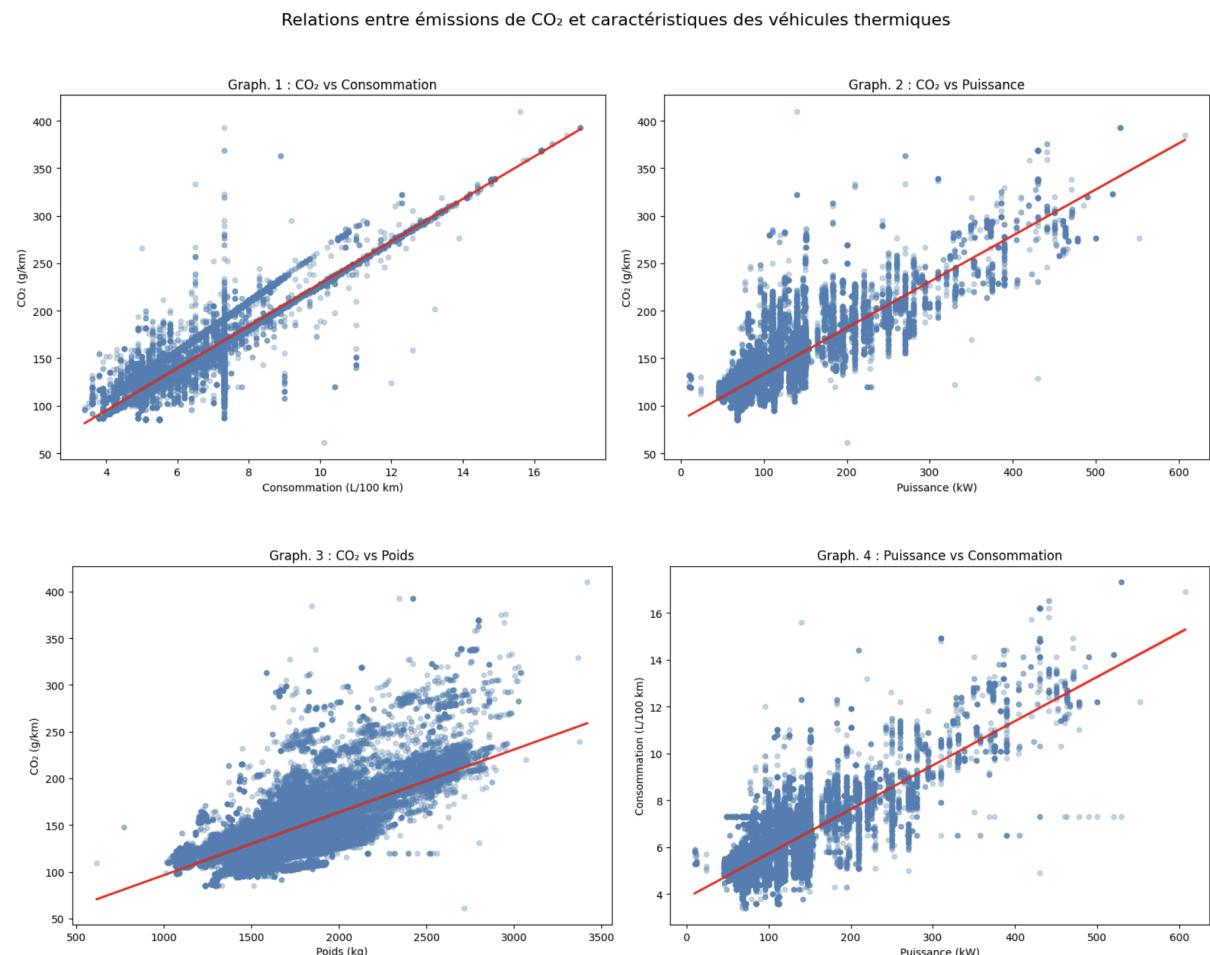
D'autres corrélations positives significatives émergent également, notamment entre la puissance du moteur et la consommation de carburant ($r = 0,77$), entre le poids et la puissance du véhicule ($r = 0,70$), ainsi qu'entre le poids et la consommation de carburant ($r = 0,56$).

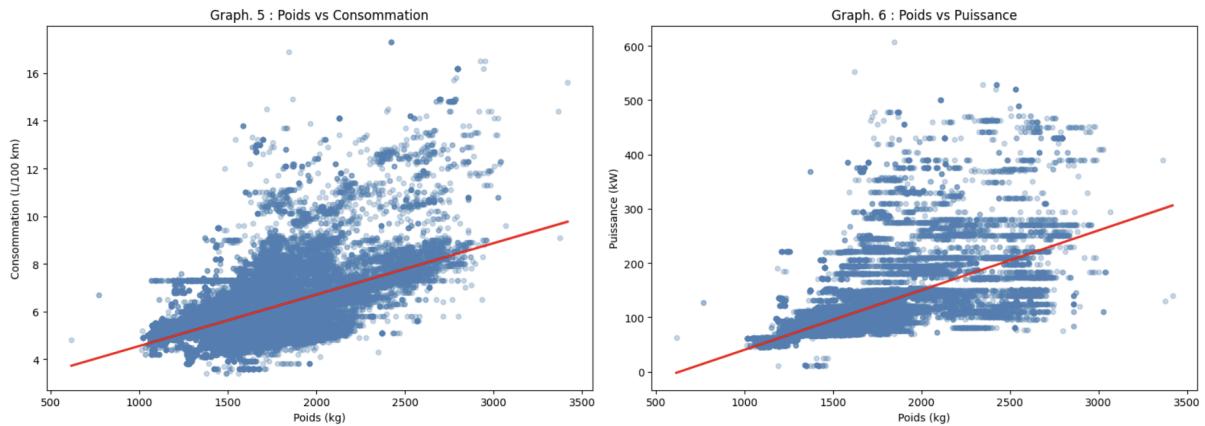
En synthèse, ces résultats indiquent que plus un véhicule thermique est lourd, plus il nécessite une motorisation puissante, ce qui se traduit par une augmentation de la consommation de carburant et, par conséquent, des émissions de CO₂.

Graphiques en nuage de points avec droite de régression linéaire

Afin de poursuivre l'analyse exploratoire, nous avons réalisé un ensemble de graphiques en nuages de points entre les variables présentant les corrélations les plus fortes. L'objectif était d'évaluer l'existence d'une relation linéaire entre ces variables.

À noter que, pour des raisons de performance et de capacité de calcul sur Google Colab, le jeu de données a été échantillonné à 100 000 véhicules.





Observations

Graphique 1 – CO₂ vs Consommation

À partir des nuages de points et de leurs droites de régression linéaire, nous observons que les émissions de CO₂ augmentent de manière positive et fortement linéaire avec la consommation des véhicules thermiques. Cette relation confirme le lien direct entre consommation de carburant et émissions polluantes.

Graphique 2 – CO₂ vs Puissance

La puissance du moteur présente également une corrélation importante avec les émissions de CO₂. La relation reste globalement linéaire, bien qu'une dispersion plus marquée des observations soit observable, en particulier pour les véhicules les plus puissants.

Graphique 3 – CO₂ vs Poids

Le poids des véhicules est fortement corrélé aux émissions de CO₂. La relation linéaire est présente, mais moins prononcée que pour la consommation. Une dispersion accrue apparaît pour les véhicules les plus lourds, dont les émissions de CO₂ tendent à augmenter plus fortement.

Graphique 4 – Puissance vs Consommation

La consommation de carburant est corrélée à la puissance du moteur de manière relativement linéaire. Toutefois, une dispersion plus importante est observée pour les puissances supérieures à 300 kW, indiquant que la consommation des véhicules les plus puissants augmente de façon plus marquée.

Graphique 5 – Poids vs Consommation

Une corrélation positive est observée entre le poids et la consommation des véhicules étudiés. Néanmoins, la consommation tend à croître beaucoup plus rapidement à partir d'un certain seuil de masse, situé autour de 1 500 kg.

Graphique 6 – Poids vs Puissance

Une corrélation entre le poids et la puissance des véhicules est également mise en évidence, mais la relation apparaît moins linéaire. En effet, au-delà d'environ 1 500 kg, la puissance augmente plus rapidement que le poids, suggérant une tendance à la surmotorisation. Ce phénomène peut notamment être relié à la part croissante des SUV (*Sport Utility Vehicles*) sur le marché automobile : des véhicules

généralement plus lourds, spacieux et puissants, souvent utilisés pour des déplacements quotidiens majoritairement urbains. Comme le souligne le site *Autocar Professionals*⁶, les SUV représentaient environ 54 % des ventes de voitures particulières neuves en Europe sur l'année 2024, soit près de 6,92 millions de véhicules sur un total d'environ 12,9 millions d'immatriculations.

3.2.4 Essence vs Diesel ?

Il convient de noter que des analyses similaires de corrélation et de linéarité ont aussi été menées séparément pour les véhicules essence et diesel. Dans un souci de concision, seules les observations mettant en évidence des tendances communes à l'ensemble des variables propres à ces deux types de motorisation sont présentées dans ce rapport.

La principale différence observée réside dans le fait que les véhicules diesel sont, en moyenne, plus lourds que les véhicules essence. Pour cette raison, ils sont également équipés de motorisations plus puissantes — à l'exception des véhicules sportifs à essence — et présentent une consommation de carburant légèrement inférieure. En revanche, leurs émissions moyennes de CO₂ sont plus élevées.

Ce résultat s'explique par le meilleur rendement énergétique des moteurs diesel par rapport aux moteurs essence. Toutefois, ces moteurs étant majoritairement installés sur des véhicules plus lourds, plus volumineux et plus puissants, leurs émissions de CO₂ par kilomètre restent, en moyenne, supérieures. L'efficience accrue du moteur diesel ne suffit donc pas à compenser l'impact du poids et des besoins énergétiques accrus de ces véhicules à l'échelle du parc automobile.

Par ailleurs, les motorisations diesel font l'objet de critiques importantes en raison de leurs émissions de NOx (oxydes d'azote) et de particules fines, reconnues pour leurs effets nocifs sur la santé humaine. Elles demeurent néanmoins privilégiées pour les véhicules lourds et puissants, pour lesquels les gains en consommation de carburant et en émissions de CO₂ sont jugés significatifs, malgré ces externalités négatives.

3.2.5 En résumé

Cette étude exploratoire menée sur le panel de véhicules essence et diesel a permis de dégager plusieurs enseignements majeurs :

- Les émissions de CO₂ évoluent positivement avec la consommation de carburant, le poids et la puissance des véhicules.

⁶ Autocar Professionals : [Europe sees record SUV market share at 54% of 12.9 million PV sales in CY2024](#)

- La puissance du moteur est positivement corrélée à la consommation de carburant, de même que le poids des véhicules est corrélé à leur puissance.
- La relation entre les émissions de CO₂ et la consommation de carburant est particulièrement linéaire.
- Les émissions de CO₂ augmentent plus fortement à mesure que le poids des véhicules croît.
- La puissance des véhicules tend également à augmenter de manière plus marquée avec le poids, un phénomène largement attribuable à la part croissante des véhicules de type SUV.

En conclusion, l'analyse met en évidence un parc automobile thermique caractérisé par des véhicules de plus en plus lourds et puissants. Cette surmotorisation contribue significativement à une hausse de la consommation de carburant et, par conséquent, à l'augmentation des émissions de CO₂.

3.3 Etude des véhicules électriques

3.3.1 Création d'un DataFrame propre aux véhicules thermiques

À l'instar des véhicules à motorisation thermique (essence et diesel), un dataframe distinct a été élaboré pour les véhicules électriques. Ce dernier conserve uniquement les variables pertinentes pour ce type de motorisation, à savoir la consommation énergétique exprimée en Wh/km ainsi que l'autonomie électrique ("electric range"). En revanche, les variables propres aux véhicules thermiques, telles que la consommation de carburant et les émissions de CO₂, ont été exclues. Cette distinction se justifie par le fait qu'un véhicule électrique ne produit aucune émission directe de CO₂ lors de son utilisation (aucune combustion de matière fossile).

3.3.2 Identification et nettoyage des outliers

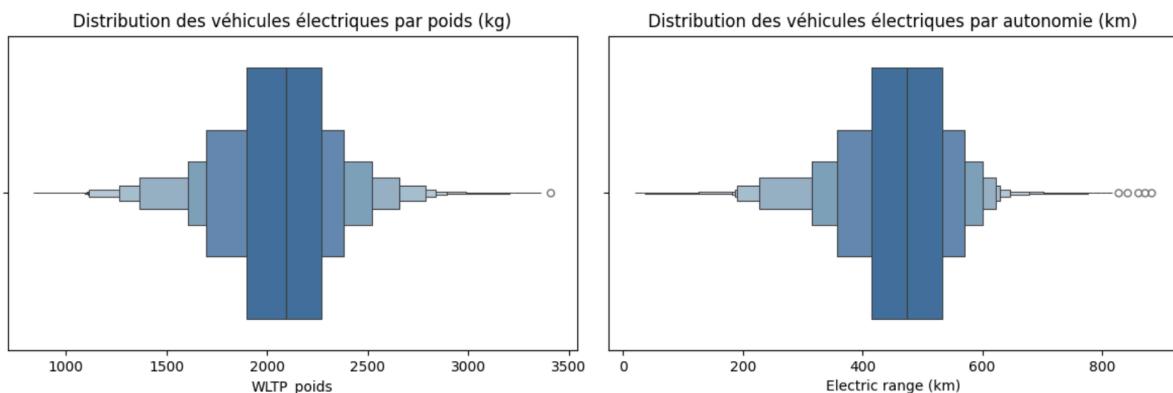
Ensuite, nous avons dressé la liste des variables numériques à étudier en affichant une première description de leur distribution respective :

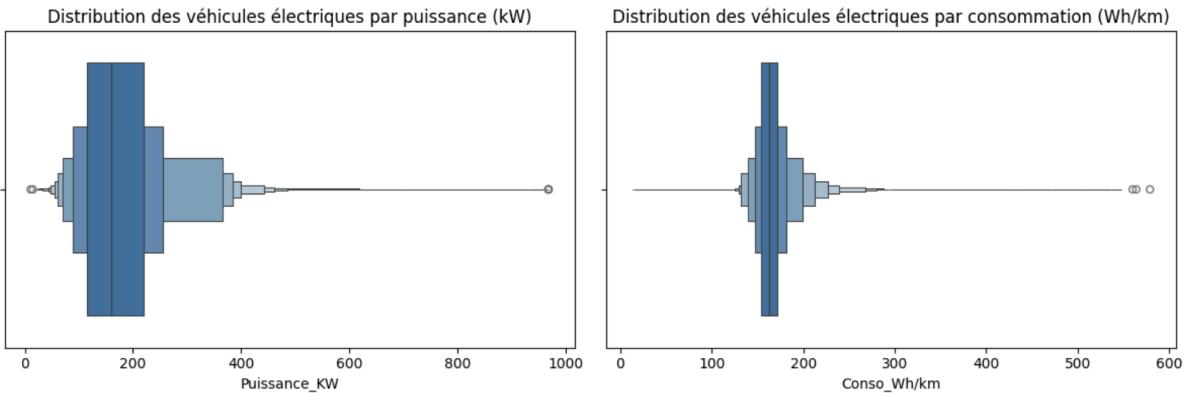
- le poids ("WLTP_poids"),
- l'autonomie ("Electric range (km)"),
- la puissance ("Puissance_KW"),
- la consommation électrique ("Conso_Wh/km").

	WLTP_poids	Electric range (km)	Puissance_KW	Conso_Wh/km
count	1.358432e+06	1.358432e+06	1.358432e+06	1.358432e+06
mean	2.066572e+03	4.680911e+02	1.807157e+02	1.652528e+02
std	3.135290e+02	9.683302e+01	8.731034e+01	2.049026e+01
min	8.410000e+02	1.900000e+01	1.100000e+01	1.300000e+01
25%	1.897000e+03	4.140000e+02	1.150000e+02	1.540000e+02
50%	2.093000e+03	4.740000e+02	1.600000e+02	1.630000e+02
75%	2.270000e+03	5.330000e+02	2.200000e+02	1.720000e+02
max	3.410000e+03	8.830000e+02	9.680000e+02	5.790000e+02

Tableau descriptif de la distribution des variables pour les véhicules électriques

Dans la continuité de l'analyse menée pour les véhicules à motorisation thermique, les distributions de ces variables ont été examinées de manière approfondie afin d'identifier la présence de valeurs aberrantes ou extrêmes (*outliers*), à l'aide de graphiques en boîtes à moustaches.





Analyse des boîtes à moustache et nettoyage des outliers

Poids : Quelques valeurs aberrantes sont observées pour des poids proches de 3 500 kg. Après vérification, ces valeurs correspondent néanmoins à des véhicules existants et sont donc considérées comme plausibles, bien que extrêmes. Elles ont, par conséquent, été conservées dans le jeu de données.

Autonomie : La distribution met en évidence certaines valeurs extrêmes supérieures à 800 km. Bien que ces autonomies soient élevées et relativement marginales, elles demeurent cohérentes avec les performances des modèles de véhicules électriques les plus récents. Ces observations ont donc été maintenues.

Puissance : Cette variable présente le plus grand nombre de valeurs extrêmes, correspondant à des véhicules à la puissance anormalement faible ou, à l'inverse, excessivement élevée. Dans le premier cas, ces anomalies semblent résulter de confusions entre la puissance du moteur et celle de la batterie. Les valeurs inférieures à 30 kW ont ainsi été considérées comme aberrantes et supprimées.

En revanche, les valeurs de puissance très élevées semblent associées à des véhicules spécifiquement préparés ou à des prototypes. Nous les avons donc conservées.

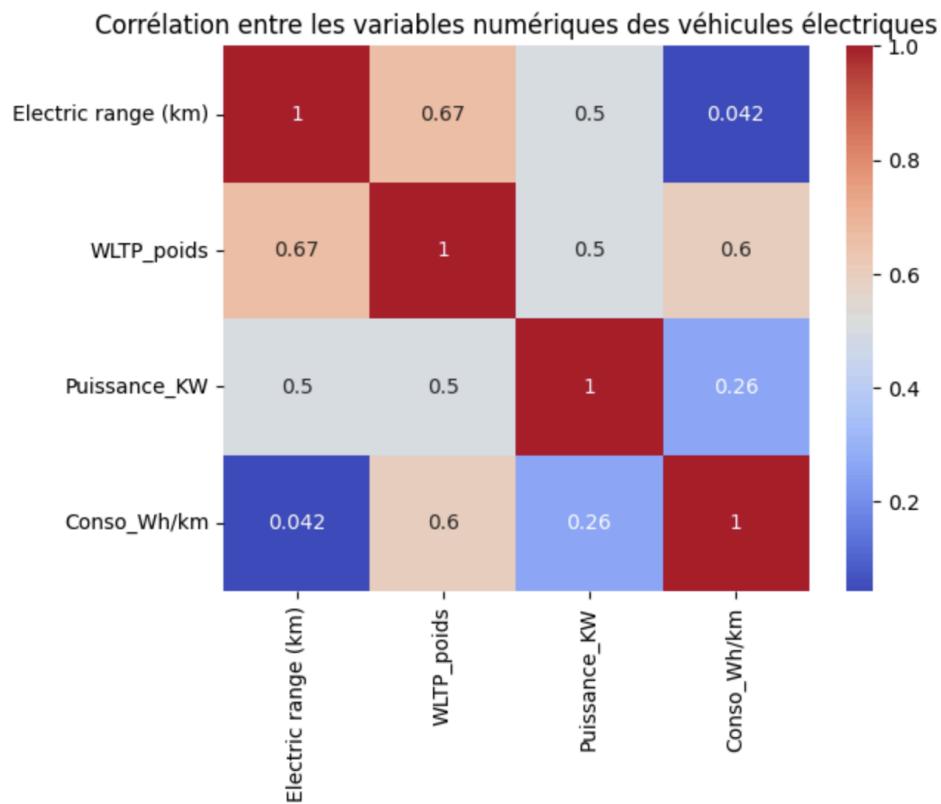
Consommation électrique : Des valeurs aberrantes supérieures à 550 Wh/km ont été observées. Celles-ci ont été jugées non plausibles et ont, par conséquent, été retirées du jeu de données.

3.3.3 Visualisations statistiques du lien entre les caractéristiques techniques des véhicules électriques

Matrice de corrélation

Dans un premier temps, une matrice de corrélation, représentée sous la forme d'une *heatmap*, a été construite afin de mettre en évidence de manière claire les

interactions entre les variables d'autonomie, de poids, de puissance et de consommation électrique.



Observations

L'analyse de la *heatmap* met en évidence **une corrélation positive marquée entre le poids des véhicules et plusieurs variables clés, notamment l'autonomie ($r = 0,67$), la consommation ($r = 0,60$) et la puissance ($r = 0,50$).**

Une autre corrélation positive notable est observée entre la puissance du moteur et l'autonomie des véhicules ($r = 0,50$).

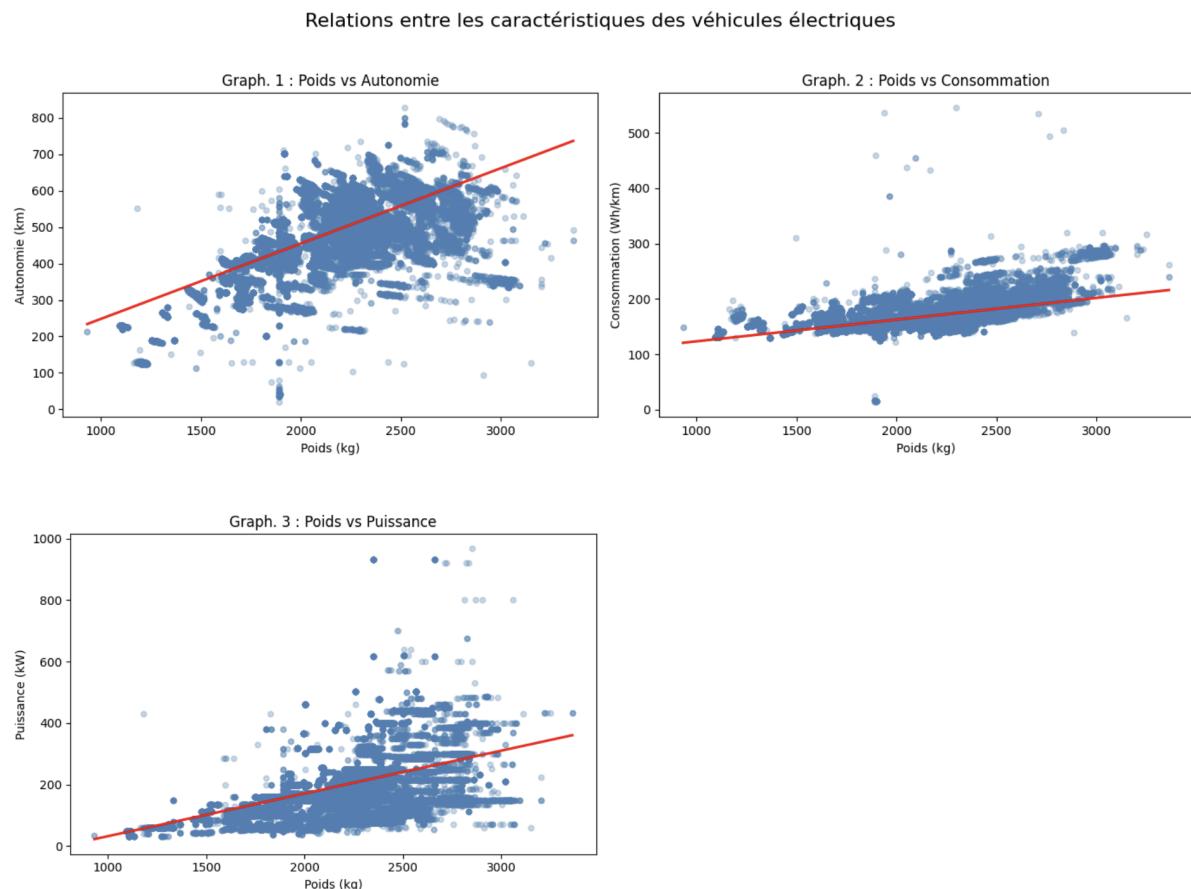
En revanche, de manière plus inattendue, la consommation électrique apparaît faiblement corrélée à l'autonomie, avec un coefficient de corrélation proche de zéro ($r = 0,042$), suggérant l'absence de relation linéaire significative entre ces deux variables. Ce résultat suggère l'existence d'une forte hétérogénéité dans la capacité des batteries au sein du parc automobile électrique. Ces disparités sont susceptibles de masquer l'effet direct de la consommation énergétique sur l'autonomie.

En synthèse, ces résultats indiquent la nécessité d'analyser plus finement l'influence du poids des véhicules sur leur autonomie, leur consommation et leur puissance, afin de mieux comprendre les mécanismes sous-jacents observés.

Graphiques en nuage de points avec droite de régression linéaire

Pour poursuivre l'analyse exploratoire, des graphiques en nuages de points ont été réalisés pour les trois variables présentant une corrélation modérée à forte avec le poids. L'objectif de cette visualisation était de mettre en évidence d'éventuelles relations linéaires entre ces variables.

Il convient de noter que, pour rendre ces calculs et visualisations réalisables sur Google Colab, un échantillon de 100 000 véhicules a été sélectionné à partir du jeu de données complet.



Observations

Graphique 1 - Poids vs Autonomie

L'analyse du nuage de points montre une dispersion importante des valeurs autour de la droite de régression linéaire. L'accroissement du poids lié à des batteries plus volumineuses pourrait théoriquement améliorer l'autonomie. Cependant, ce graphique révèle une réalité plus nuancée : l'augmentation du poids total des véhicules se fait le plus souvent au détriment de l'autonomie.

Graphique 2 - Poids vs Consommation

Le graphique indique que la consommation électrique augmente plus rapidement

que le poids des véhicules. Autrement dit, l'accroissement du poids se traduit par une augmentation proportionnellement plus forte de la consommation électrique. Cette dynamique explique pourquoi le poids n'a qu'un effet limité sur l'autonomie : bien que les batteries plus grandes puissent théoriquement améliorer l'autonomie, l'augmentation concomitante de la consommation liée au poids des véhicules réduit ces gains.

Graphique 3 - Poids vs Puissance

L'observation du nuage de points montre que la dispersion de la puissance augmente avec le poids. Un nombre significatif de véhicules lourds présentent des puissances élevées, en particulier au-delà de 2 000 kg. Cette tendance est similaire à celle observée pour les véhicules thermiques et reflète le profil typique des SUV, qui combinent un poids important et une puissance élevée.

3.3.4 En résumé

L'analyse exploratoire de notre panel de véhicules électriques met en évidence plusieurs phénomènes caractéristiques de ce type de motorisation.

Intuitivement, on pourrait penser qu'une augmentation de la taille des batteries se traduirait par une amélioration de l'autonomie. Cependant, dans la pratique, la conception générale des véhicules électriques et l'accroissement concomitant du poids tendent souvent à réduire l'autonomie plutôt qu'à l'améliorer.

Par ailleurs, le poids accru des véhicules entraîne une augmentation de la consommation électrique, ce qui explique souvent la limitation de l'autonomie malgré des batteries plus grandes.

Enfin, comme pour les véhicules thermiques, la puissance des véhicules tend à augmenter avec le poids. Cette observation, couplée à des recherches complémentaires, révèle un effet marqué des SUV dans le segment électrique. Selon le site spécialisé InsideEV⁷, parmi les cinq véhicules électriques les plus vendus en Europe en 2024, quatre sont des SUV ou des crossovers (Tesla Model Y, Volvo EX30, Skoda Enyaq et VW ID.4), ce qui illustre la prédominance de cette catégorie.

En somme, bien que les véhicules électriques représentent un atout stratégique pour la transition énergétique en raison de leur absence d'émissions directes de CO₂, leur efficience énergétique reste globalement limitée. Ces observations suggèrent que les véhicules électriques sont souvent conçus pour privilégier le style, le confort et la puissance, plutôt que l'optimisation de la consommation énergétique.

⁷ InsideEV : [The 20 Best-Selling EVs Of 2024 In Europe](#)

3.4 Etude des véhicules hybrides : essence ou diesel / électrique

3.4.1 Création d'un DataFrame propre aux véhicules hybrides

Tout comme pour les véhicules thermiques et électriques, nous avons créé un DataFrame unique aux motorisations hybrides en gardant seulement les véhicules dont le type de motorisation correspondait à 'hybride essence' ou 'hybride diesel'.

Dans un premier temps, nous avons procédé à une vérification que les valeurs du DataFrame modifié sont bien propres. Nous avons donc vérifié qu'il n'y avait pas la présence de valeurs manquantes ou de doublons ou valeurs nulles.

Ensuite, nous avons procédé à la sélection des variables étudiées. Pour les véhicules hybrides quasiment toutes les variables quantitatives dans le DataFrame modifiée nous intéressent. Effectivement, pour cause d'une double motorisation thermique électrique les variables Conso_Wh/km et Electric range (km) sont des variables qui nous servent dans notre analyse ainsi que celle que l'on a précédemment étudiées pour les véhicules thermiques.

Cependant après la sélection de ces variables et la création d'un DataFrame pour les véhicules hybrides, nous avons remarqué des valeurs minimums à zéro, ce qui n'est pas logique pour notre analyse. Nous avons alors procédé au retrait de ces valeurs égales à zéro pour les variables 'Conso_Wh/km' , 'Fuel consumption' et 'Electric range (km)'. Nous decisions de supprimer ces valeurs dans notre analyse car nous considérons qu'une valeur de zéro dans une de ces colonnes est fausse et pas représentative d'un véhicule réel, logiquement une voiture hybride devrait émettre du CO₂, avoir une consommation de carburant au dessus de 0 et avoir une autonomie de batterie au dessus de 0km.

Ce qui nous donne la description de la distribution des variables quantitatives dans le tableau suivant :

	WLTP_poids	Co2_Emission(WLTP)	Puissance_KW	Conso_Wh/km	Fuel consumption	Electric range (km)
count	681863.000000	681863.000000	681863.000000	681863.000000	681863.000000	681863.000000
mean	2196.486737	25.943566	153.267038	193.817516	1.393335	78.299012
std	315.834672	15.217446	61.683223	41.585016	1.459109	24.173898
min	1084.000000	6.000000	15.000000	14.000000	0.100000	11.000000
25%	1949.000000	18.000000	110.000000	168.000000	0.800000	61.000000
50%	2159.000000	24.000000	135.000000	184.000000	1.100000	74.000000
75%	2384.000000	32.000000	186.000000	210.000000	1.400000	91.000000
max	3960.000000	404.000000	607.000000	430.000000	26.000000	702.000000

Tableau descriptif de la distribution des variables quantitatives du DataFrame véhicules Hybrides

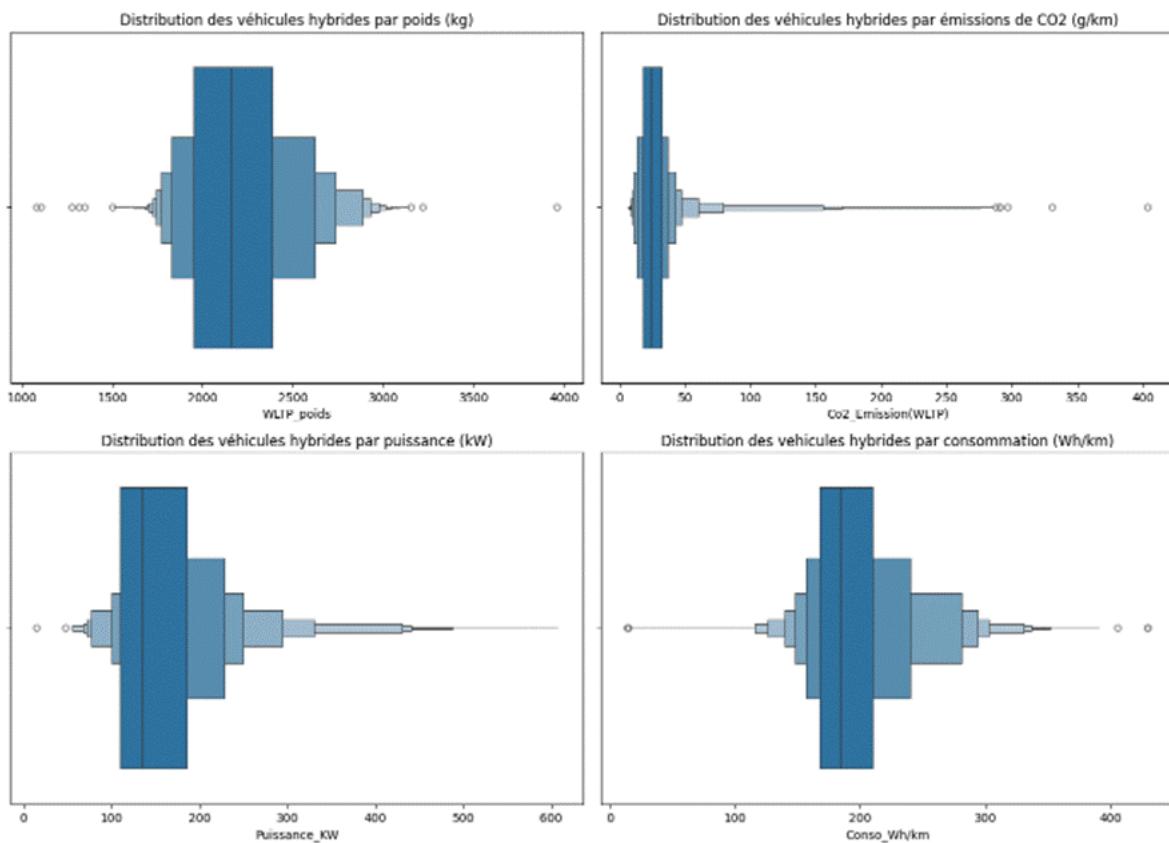
3.4.2 Identification et nettoyage des outliers

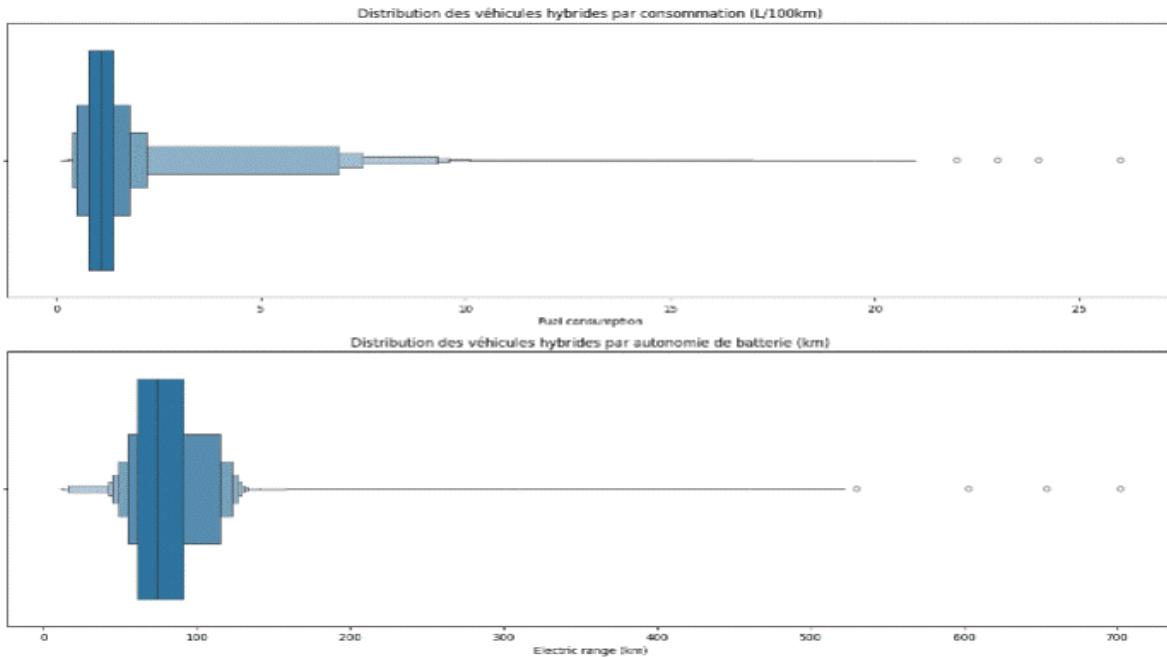
Pour résumer les variables quantitatives étudiées dans ce data frame sont les suivantes :

- WLTP_Poids
- CO2_Emmision (WLTP)
- Puissance_Kw
- Conso_Wh/km
- Fuel consumption
- Electric range (km)

Dans la continuité de notre analyse nous avons choisi de pousser plus loin l'analyse de cette distribution à travers la création de graphiques en boîtes à moustaches afin de faciliter la recherche et l'élimination potentielle de valeurs extrêmes ou aberrantes.

Graphiques en boîtes à moustaches pour variables étudiées :





Analyse des boîtes à moustache et nettoyage des outliers

Poids : Nous remarquons grâce au graphique la présence de valeurs extrêmes en dessous de 1500 kg et au-dessus de 3200 kg. Plusieurs voitures sont bien au-dessus de la valeur indiquée par le DataFrame. Pour les valeurs en dessous de 1500 kg, elles représentent bien des valeurs aberrantes à l'exception de la Dacia Spring qui fait bien ce poids là. Pour les valeurs au-dessus de 3200 kg, ce sont des valeurs aberrantes à retirer.

Émissions de CO₂ : Les valeurs extrêmes pour les véhicules hybrides dont les émissions de CO₂ sont hautes (supérieur à 250), mais semblent être des valeurs réelles (voiture de sport) que nous garderons.

Puissance (kW) : Nous remarquons la présence d'outliers pour des puissances inférieures à 50 KW. Après de rapides recherches internet, nous supprimons une des valeurs et gardons la Dacia Spring qui a bien une puissance de 48 kW.

Consommation (Wh/km) : Nous observons la présence de valeurs extrêmes du fait de valeurs en dessous de 15 Wh/km et au-dessus de 400 Wh/km. La valeur en dessous de 15 Wh/km représente bien un véhicule réel, tandis que les véhicules au-dessus de 400 Wh/km sont aberrants.

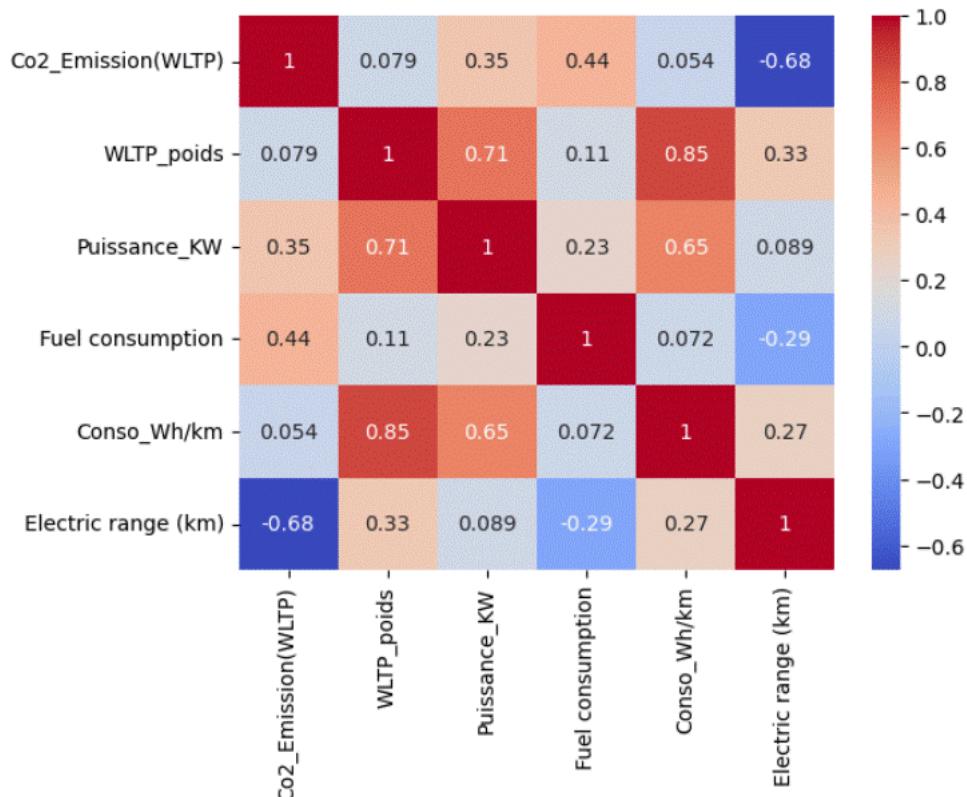
Consommation (L / 100 km) : Nous remarquons la présence de valeurs extrêmes à plus de 20 L /100km. Ce sont bien des valeurs aberrantes à supprimer.

Autonomie de batterie : Nous constatons la présence de valeurs extrêmes au-dessus de 500 km d'autonomie de batterie. Ce sont des valeurs aberrantes à retirer du jeu de données.

3.4.3 Visualisations statistiques du lien entre les caractéristiques techniques des véhicules hybrides

Matrice de corrélation

Corrélation entre les variables numériques des véhicules hybrides :



Observations

A partir de la *heatmap* ci-dessus, nous remarquons plusieurs corrélations entre les variables numériques des véhicules à motorisations hybrides. Les corrélations qui ressortent le plus sont celles entre la Conso_Wh/km et WLTP_Poids ($r=0.85$), entre CO2_Emission et Electric range (km) ($r=-0.68$), la Puissance_KW et WLTP_Poids de ($r=0.71$).

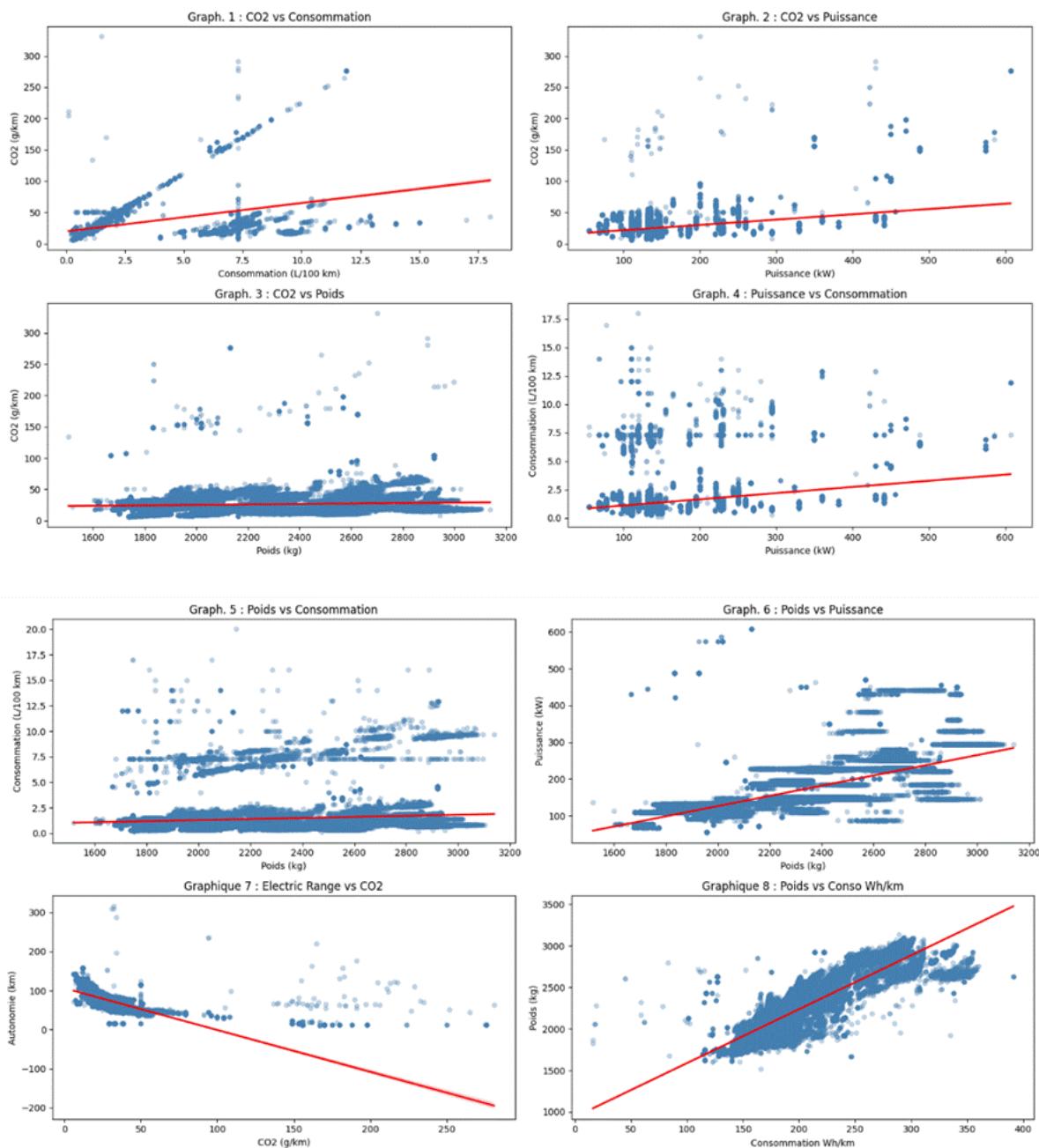
Nous constatons donc deux corrélations positives très fortes et une corrélation inverse. Cela indique un fort lien positif entre les variables Conso_Wh/km et WLTP_poids et entre Puissance_KW et WLTP_Poids; et aussi un lien négatif entre CO2_Emission et Electric Range (km).

Nous observons aussi une corrélation très faible qui paraît étonnante vis-à-vis de nos analyses précédentes notamment l'effet du poids sur les émissions de CO₂ ($r=0.079$)

Graphiques en nuage de points avec droite de régression linéaire

Pour pousser notre analyse plus loin nous avons réalisé des graphiques à nuage de points avec droite de régression linéaire. Il convient encore de noter que, pour rendre ces calculs et visualisations réalisables sur Google Colab, un échantillon de 100 000 véhicules a été sélectionné à partir du jeu de données complet.

Relations entre émissions de CO₂ et caractéristiques des véhicules hybrides



Graphique 1 - CO₂ vs Consommation l/100km (r = 0.44)

Dans le graphique à nuage de points, nous constatons un éparpillement des données autour de la droite de régression linéaire, ce qui explique la corrélation moyenne entre les deux variables. Plus précisément, on constate que beaucoup de véhicules de l'échantillon ont une consommation L/100km égale à 7.5. Il est possible de remarquer deux tendances, certains points du graphique semblent croître de façon parfaitement linéaire entre CO₂ et consommation (une augmentation de consommation L/100km peut entraîner une augmentation du CO₂ émis) et d'autres points indiquent qu'une augmentation de la consommation L/100km entraîne une moindre augmentation des émissions de CO₂ qui reste entre 0 et 50 g/km ou proche de la moyenne du tableau descriptif.

Graphique 2 - CO₂ vs Puissance (r = 0.35)

Dans ce graphique à nuage de points associé à la relation entre les émissions de CO₂ et la puissance en kW, nous remarquons qu'une augmentation de la puissance peut entraîner une légère hausse dans les émissions de CO₂, mais cela reste très faible. Effectivement cela est due à la moyenne très faible d'émissions de CO₂ pour les véhicules hybrides de 25.94 g/km. Nous constatons aussi certains points qui ont des émissions de CO₂ bien au-dessus de la moyenne, mais cela n'est pas spécifique uniquement à de fortes puissances, il est alors difficile d'en tirer des conclusions.

Graphique 3 - CO₂ vs Poids (r = 0.079)

Dans ce graphique à nuage de points associé à la relation entre les émissions de CO₂ et le poids, il ressort qu'une augmentation du poids n'a quasiment aucun effet sur la quantité de CO₂ émise. Cela peut être expliqué par la présence du moteur électrique supplémentaire qui augmente le poids des véhicules mais qui ne rejette pas de CO₂ quand il est utilisé.

Graphique 4 - Puissance vs Consommation (r = 0.23)

Dans ce graphique à nuage de points qui explore la relation entre la puissance et la consommation de carburant, nous observons que les valeurs sont très éparpillées autour de la droite de régression linéaire. Certains véhicules à faible puissance consomment beaucoup de carburant, tandis que les véhicules plus puissants ne consomment pas, ou que légèrement, plus de carburant.

Graphique 5 - Poids vs Consommation (r = 0.11)

Dans ce graphique à nuage de point qui visualise la relation entre le poids et la consommation de carburant, nous remarquons que le poids a peu d'influence sur la consommation de carburant. Effectivement dû à la faible consommation en essence des hybrides (1.39 L/100km en moyenne), l'effet du poids sur la consommation de carburant ne se remarque quasiment pas, à l'inverse des voitures thermiques où une corrélation positive a été constatée.

Graphique 6 - Poids vs Puissance (r = 0.71)

Dans ce graphique à nuage de point qui visualise la relation entre le poids et la puissance des véhicules on remarque une tendance similaire au véhicules

électriques et thermiques, qui est que le poids des véhicules semble augmenter avec la puissance des véhicules. On peut donc dire qu'on a des véhicules qui deviennent de plus en plus puissants plus ils sont lourds. On peut associer cela à l'effet SUV sur les véhicules qui augmente en taille (et par conséquent sont plus lourds) et en puissance.

Graphique 7 - Autonomie vs CO₂ (r = -0.68)

Dans ce graphique à nuage de point qui représente la relation entre l'autonomie de batterie et les émissions de CO₂, nous observons un effet inverse de l'autonomie de la batterie sur les émissions de CO₂. Effectivement il y a une tendance qui indique que plus une voiture a une grande autonomie, moins elle va émettre de CO₂. Nous constatons aussi une tendance que plus une voiture émet de CO₂, plus elle aura une autonomie de batterie proche de zéro. L'importance de l'autonomie de batterie pour les réductions d'émissions de CO₂ est ainsi mise en valeur .

Graphique 8 - Poids vs Consommation Wh/km (r = 0.85)

Dans le graphique de nuage de points qui représente la relation entre le poids et la consommation électrique de la batterie en Wh/km, il apparaît une corrélation positive entre le poids et la consommation électrique des véhicules qui semble augmenter plus le poids des véhicules est lourd. Nous retrouvons de nouveau une tendance vers des véhicules très lourds et très puissants comme précédemment. De plus, en vue de ces données, nous pouvons en déduire que les véhicules plus lourds sont moins efficaces d'un point de vue énergétique, car ils vident leur batterie plus rapidement que des voitures plus légères.

3.4.4 En résumé

Suite à nos observations lors de l'exploration des données relatives aux véhicules hybrides, nous remarquons quelques points clés :

- Les véhicules hybrides sont en moyenne très lourds (près de 2200 kg). Cette moyenne très haute est bien au-dessus de la moyenne de poids des véhicules thermiques et se rapproche plus de la moyenne de poids des véhicules électriques.
- La consommation de carburant est très basse pour les véhicules hybrides et est moyennement corrélée aux émissions de CO₂.
- Les émissions de CO₂ sont inversément corrélées à l'autonomie de batterie ce qui montre l'effet positif en termes de réduction d'émissions de CO₂ de l'électrification des véhicules.
- Nous constatons aussi un effet SUV comme chez les véhicules électriques en vue des très fortes puissances en KW et très haute consommation en Wh/km qui sont corrélés de façon positive et forte au poids du véhicule.

- Ces véhicules lourds sont moins efficaces d'un point vue énergétique remettant en question l'impact des véhicules hybrides sur d'autres paramètres que pûrement le CO₂.

3.5 Etude des véhicules aux motorisations alternatives

3.5.1 Création d'un DataFrame propre aux véhicules à motorisations alternatives

Suivant la même méthode que pour les sections précédentes, nous avons décidé de créer un DataFrame unique pour les véhicules à motorisations alternatives. Ici, les véhicules à motorisations alternatives regroupent les véhicules qui utilisent des carburants autres que l'essence, le diesel ou l'électrique. Nous avons donc sélectionné tous les types de carburant appartenant à 'Autre' dans notre DataFrame modifié. Comme précédemment, nous vérifions que le DataFrame que nous venons de créer ne représente pas de doublons, valeurs manquantes ou de valeurs nulles.

Nous avons choisi de garder les variables numériques suivantes à étudier : le poids, les émissions de CO₂, la puissance en kw et la consommation de carburant. En effet, la consommation en Wh/km ne s'applique pas à ces véhicules, ni l'autonomie de la batterie électrique.

Pour le DataFrame nouvellement créé, nous avons remarqué la présence de valeurs minimum à zéro pour les colonnes "CO2_Emissions (WLTP)" et "Fuel Consumption". Ces dernières ont été supprimées avant la réalisation d'une étude descriptive des variables numériques donnant le tableau suivant:

	WLTP_poids	Co2_Emission(WLTP)	Puissance_KW	Fuel consumption
count	318065.000000	318065.000000	318065.000000	318065.000000
mean	1367.733944	120.053429	72.978885	6.419676
std	95.165664	13.863628	10.982156	0.853541
min	1089.000000	98.000000	46.000000	0.700000
25%	1304.000000	113.000000	67.000000	5.800000
50%	1317.000000	116.000000	74.000000	6.100000
75%	1451.000000	126.000000	74.000000	7.100000
max	2482.000000	241.000000	137.000000	17.000000

Tableau descriptif de la distribution des variables quantitatives du DataFrame véhicules motorisations alternatives :

Cependant il est important de noter que à travers la suppression des 0 dans le dataframme 'Autre' nous avons retiré tous les véhicules à motorisations de type hydrogène de notre analyse car elle ne rejette pas de CO₂ pendant leur utilisation. On peut négliger pour l'instant cette catégorie de véhicules en vue des grands axes de notre analyse et du temps accordé, mais une analyse plus poussée sans regroupement des motorisations alternatives peut être envisageable.

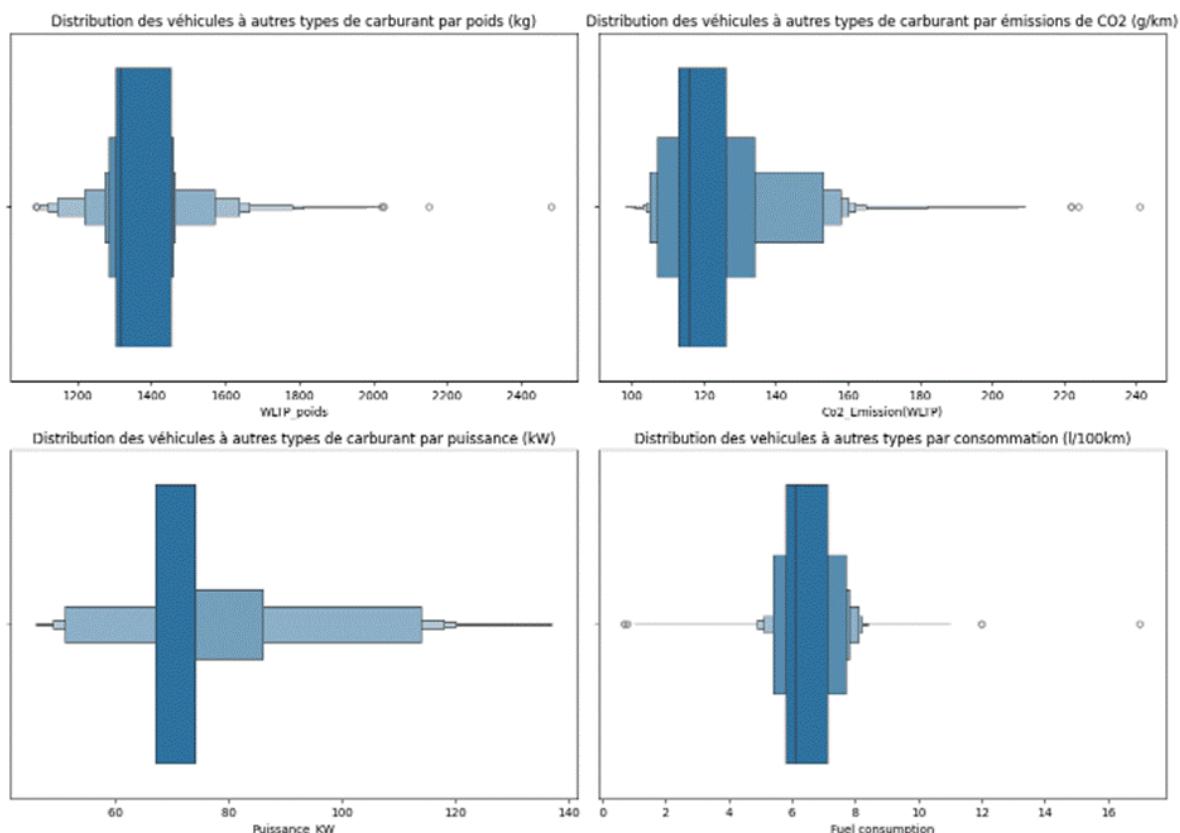
3.5.2 Identification et nettoyage des outliers

Pour résumer les variables quantitatives étudiées dans ce DataFrame sont les suivantes :

- WLTP_Poids
- CO2_Emmision (WLTP)
- Puissance_Kw
- Fuel consumption

Dans la continuité de notre analyse nous avons choisi de pousser plus loin l'analyse de cette distribution à travers la création de graphiques en boîtes à moustaches afin de faciliter la recherche et l'élimination potentielle de valeurs extrêmes ou aberrantes pour le DataFrame représentant les véhicules à motorisations alternatives.

Graphiques en boîtes à moustaches pour les variables étudiées :



Analyse des graphique en boîtes à moustache et nettoyage des outliers

Poids : nous constatons la présence de valeurs extrêmes à moins de 1100 kg et à plus de 2000 kg. Les valeurs à moins de 1100 kg représentent bien des valeurs réelles que nous garderons, tandis que celles à plus de 2000 kg représentent des valeurs aberrantes à supprimer.

Émissions de CO₂ : nous remarquons la présence d'outliers au-dessus de 220 g/km de CO₂. Il y a une valeur aberrante qui est la Peugeot boxer ID (157084324) le reste des données représente bien des valeurs réelles.

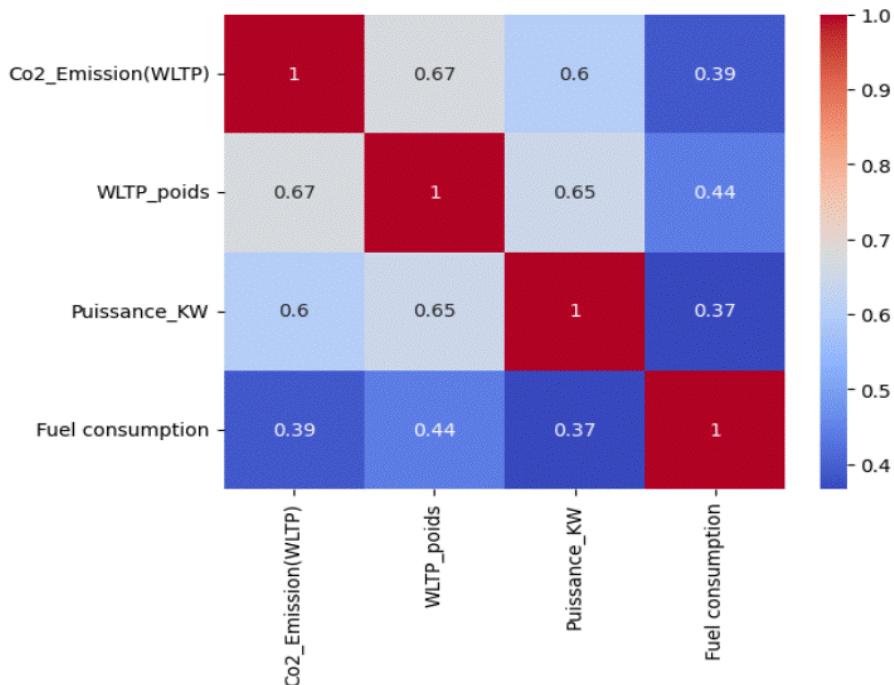
Puissance (kW) : pas de présence d'outlier.

Consommation (L/100km) : nous remarquons la présence d'outliers à moins de 2 l/100 km et à plus ou égal à 12 l/100km. Dans les deux cas, ce sont des valeurs aberrantes à supprimer.

3.5.3 Visualisations statistiques du lien entre les caractéristiques techniques des véhicules à motorisations alternatives.

Matrice de corrélation

Matrice de corrélation des variables numériques pour voitures à motorisations alternatives



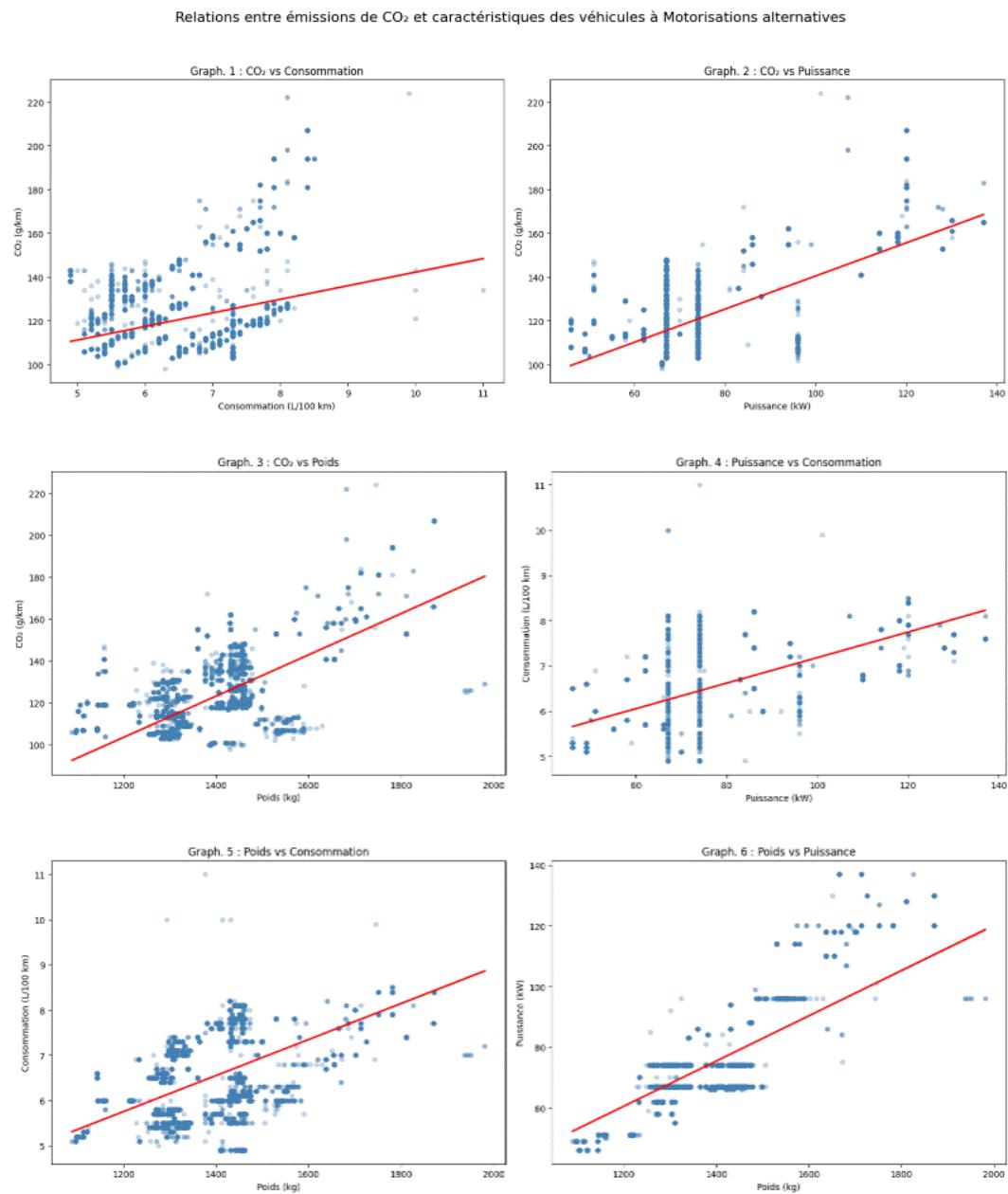
Observations

A la lecture de cette matrice, nous remarquons plusieurs corrélations positives entre les variables numériques. Les plus fortes étant les relations entre le poids et les

émissions de CO₂ **r=0.67**, le poids et la puissance Kw **r= 0.65** et la puissance et les émissions de CO₂ **r= 0.6**.

Graphiques en nuage de points avec droite de régression linéaire

Pour poursuivre notre analyse nous avons créé les graphiques en nuages de points avec droite de régression linéaire. Encore une fois nous rappelons que ces graphiques ont été créés avec un sample de la base de données complète pour faciliter les calculs sur Google Colab



Observations

Graphique 1 - CO₂ vs Consommation (r = 0.39)

Dans le graphique à nuage de points qui représente la relation entre les émissions de CO₂ et la consommation de carburant, nous constatons une légère corrélation entre les émissions de CO₂ et la consommation. Cela semble indiquer que plus la consommation augmente, plus les émissions de CO₂ augmentent aussi. Cependant, nous remarquons que certaines valeurs à basse consommation ont une valeur d'émission de CO₂ élevée et inversement des véhicules à haute consommation ont des valeurs moyennes d'émissions de CO₂ faibles, expliquant la faible corrélation. Ici on ne retrouve pas la forte corrélation entre émissions de CO₂ et consommation L/100km trouver chez les véhicules thermiques. Cela peut être due au mélange de différentes motorisations regrouper dans le dataframe 'autre'

Graphique 2 - CO₂ vs Puissance (r = 0.6)

Dans le graphique, nous constatons que plus la puissance semble augmenter plus les émissions de CO₂ augmentent en fonction. Nous observons aussi différentes valeurs d'émissions de CO₂ pour des véhicules avec une puissance identique ce qui explique que la corrélation n'est pas plus forte.

Graphique 3 - CO₂ vs Poids (r = 0.67)

Ici, nous constatons une corrélation positive entre poids et émissions de CO₂, mais nous pouvons voir que pour certains véhicules très lourds 2000 kg ont une valeur d'émission de CO₂ proche de la moyenne de 120 g/km ce qui atténue la force de la corrélation.

Graphique 4 - Puissance vs Consommation (r = 0.37)

Nous observons une faible corrélation positive pour cause des différentes valeurs de consommation associées à une unique valeur de puissance.

Graphique 5 - Poids vs Consommation (r = 0.44)

Nous remarquons que les véhicules ont une tendance à consommer plus quand le poids augmente. Cependant entre 1200 kg et 1500 kg beaucoup de véhicules ont une consommation inférieure à la moyenne de 6.41 l/100km.

Graphique 6 - Poids vs Puissance (r = 0.65)

Nous constatons que plus le poids augmente, plus la puissance est grande. Néanmoins, du fait de plusieurs valeurs de puissance identiques à différentes valeurs de poids, la corrélation n'est pas plus forte.

3.5.4 En résumé

Suite à nos observations lors de l'exploration des données relatives aux véhicules avec d'autres types de motorisations, nous remarquons quelques points clés :

- Ce sont les véhicules les plus légers avec une moyenne de 1367 kg par véhicule du DataFrame.
- Avec 318065 lignes, c'est l'une des catégories de véhicules la moins représentée donc elle est quantitativement négligeable vis-à-vis des autres types de véhicules étudiés.
- Contrairement aux véhicules thermiques, nous ne retrouvons pas une forte corrélation entre les émissions de CO₂ et la consommation L/100km.
- Nous observons que le poids et la puissance sont les indicateurs majeurs d'émissions de CO₂ en vue de leur forte corrélation positive (0.67 pour le poids et 0.6 pour la puissance respectivement au CO₂).

3.6 Composition du parc automobile

Maintenant que nous disposions d'une compréhension plus précise des caractéristiques des véhicules et de leurs interactions, il est apparu pertinent d'adopter une vision plus globale de la composition du parc automobile.

Pour ce faire, trois nouvelles variables ont été créées afin de faciliter la classification des véhicules et d'améliorer la visualisation des données : des **étiquettes CO₂** et des **quartiles de poids et des quartiles de puissance**.

3.6.1 Les étiquettes CO₂

Ces étiquettes permettent d'évaluer et de représenter de manière simple les émissions de CO₂ associées à chaque véhicule de notre panel.

Depuis 2006, les constructeurs automobiles ont l'obligation d'indiquer le niveau d'émissions de CO₂ des véhicules neufs vendus, exprimé en grammes par kilomètre parcouru (g/km). En France, cette exigence se matérialise par l'étiquette énergétique et CO₂, qui informe notamment sur la quantité de CO₂ émise par kilomètre pour chaque véhicule.

Sept classes distinctes existent, allant de A (véhicule le moins polluant) à G (véhicule le plus polluant) :

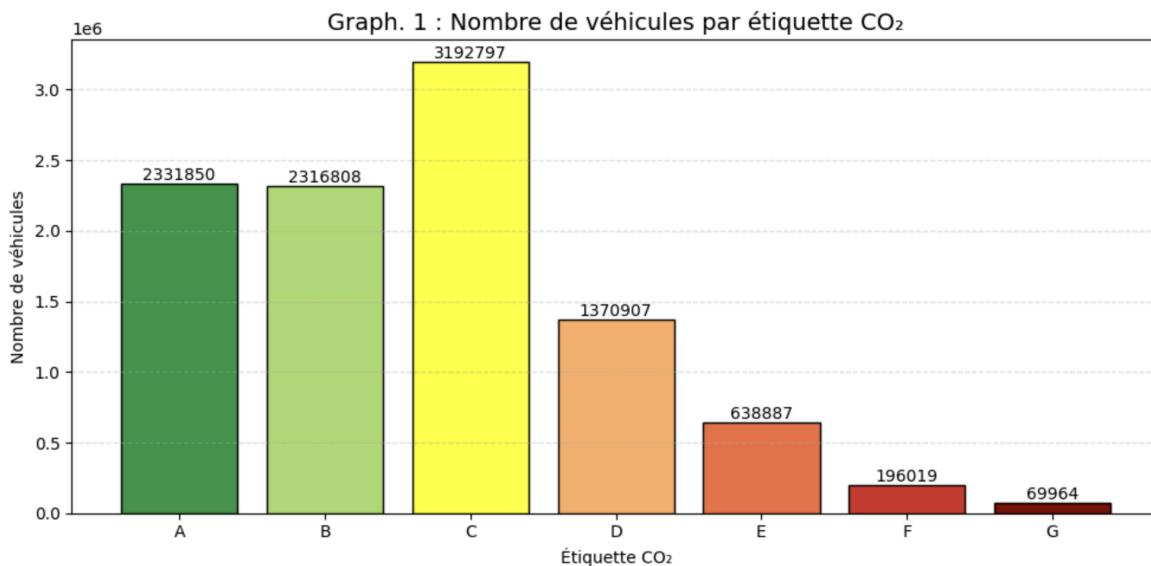
- **A** : ≤ 100 g/km
- **B** : 101–120 g/km
- **C** : 121–140 g/km
- **D** : 141–160 g/km

- **E** : 161–200 g/km
- **F** : 201–250 g/km
- **G** : > 250 g/km

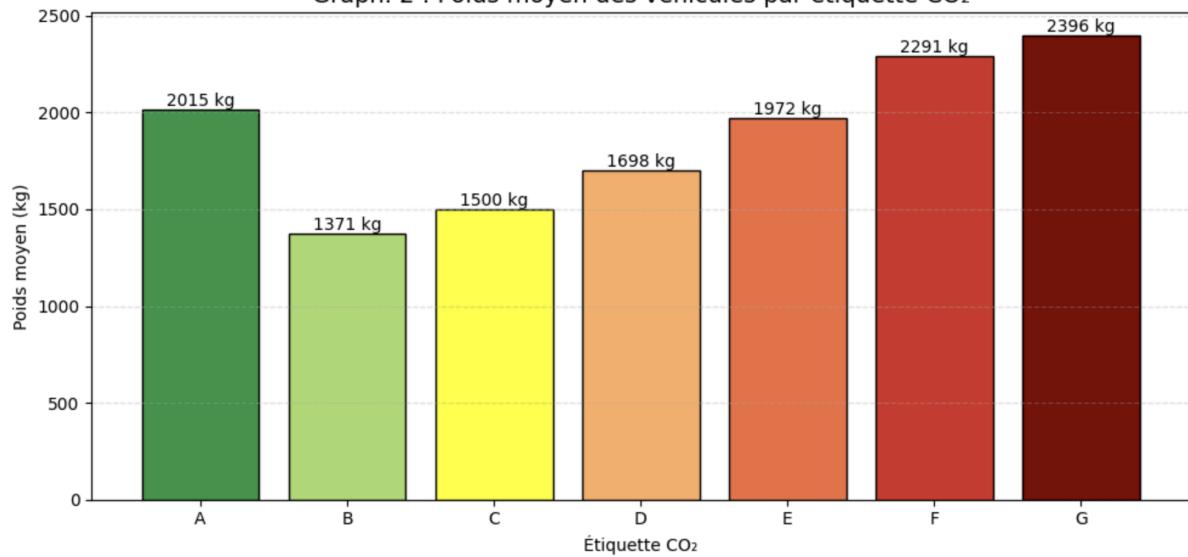
Il convient de noter que cette classification reflète uniquement les émissions directes de CO₂ lors de l'usage des véhicules. Par conséquent, les véhicules électriques, neutres en émissions directes, sont systématiquement classés A. En revanche, l'étiquette ne renseigne pas sur les émissions de CO₂ et les impacts environnementaux liés aux différentes étapes du cycle de vie du véhicule.

Malgré cette limitation, l'étiquette constitue un outil de comparaison pertinent pour les consommateurs. Elle a également des implications financières, via des bonus pour les véhicules les plus propres et des malus pour les véhicules les plus polluants et/ou les plus lourds. Ces critères et les montants associés varient chaque année. À titre indicatif, en 2025 en France, le malus débute à 50 € pour un véhicule émettant 113 g/km et peut atteindre 70 000 € pour un véhicule émettant 192 g/km.

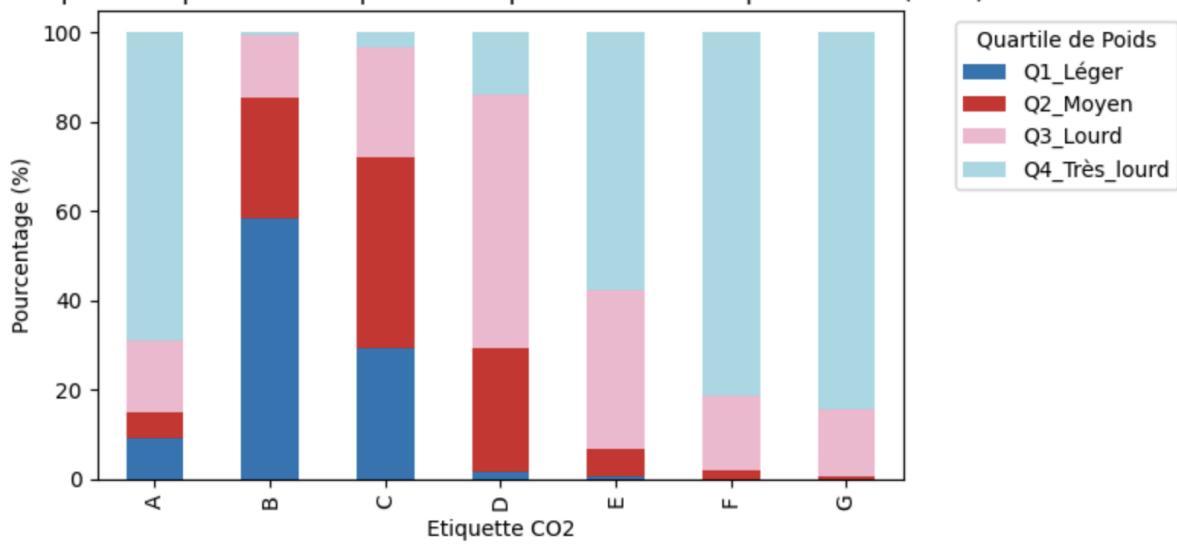
Afin de mieux appréhender la composition de notre panel de véhicules selon cette catégorisation, nous avons réalisé plusieurs graphiques de distribution. Bien que la norme des étiquettes soit propre à la France, nous avons appliqué cette classification à l'ensemble de l'échantillon européen afin d'en tirer des enseignements globaux sur les émissions de CO₂.



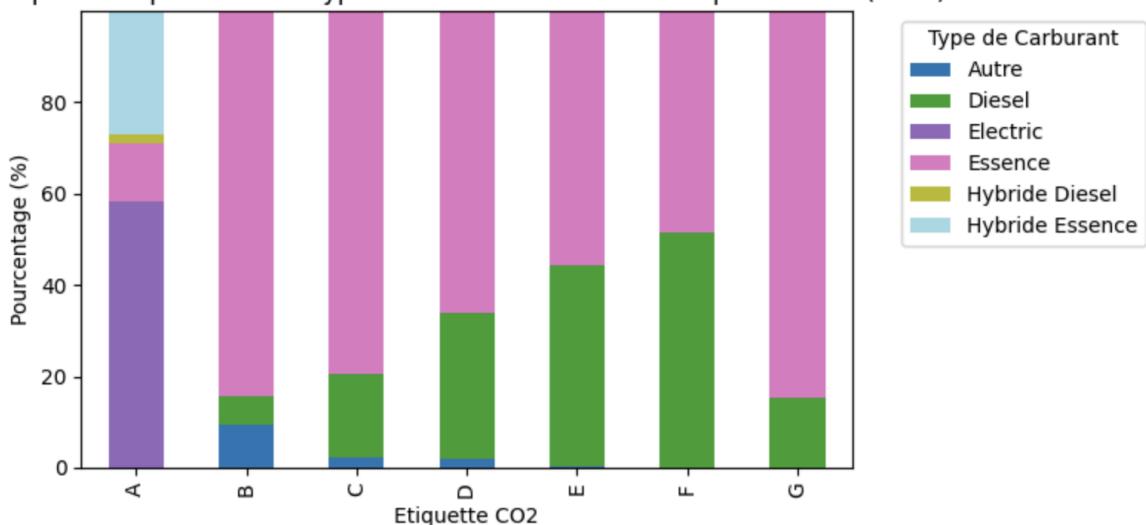
Graph. 2 : Poids moyen des véhicules par étiquette CO₂



Graph. 3 : Répartition des quartiles de poids selon les étiquettes CO₂ (en %)



Graph. 4 : Répartition des types de carburant selon les étiquettes CO₂ (en %)



Observations

Graphique 1 – Nombre de véhicules par étiquette CO₂

La catégorie la plus représentée est la C, suivie des catégories A et B. Cette répartition peut suggérer que le parc automobile européen tend à devenir progressivement plus propre en termes d'émissions de CO₂.

Graphique 2 – Poids moyen des véhicules par étiquette CO₂

Une observation notable se dégage : les véhicules deviennent globalement plus légers à mesure que leur étiquette CO₂ s'améliore, de G à B. En revanche, les véhicules classés A présentent, en moyenne, un poids supérieur à ceux de la catégorie E, ce qui constitue une exception intéressante.

Graphique 3 – Répartition des quartiles de poids selon les étiquettes CO₂

Ce graphique révèle que la catégorie A est majoritairement composée de modèles lourds : environ 70 % des véhicules appartiennent au quartile des véhicules très lourds, et 15 % au quartile des véhicules classés comme lourds.

Graphique 4 – Répartition des types de carburant selon les étiquettes CO₂

Les véhicules bénéficiant de l'étiquette A sont majoritairement des véhicules électriques, représentant environ 60 % du total, suivis par les hybrides essence/diesel, qui constituent environ 20 %.

Ces résultats soulignent l'importance d'approfondir l'analyse du poids en fonction du type de motorisation, afin de mieux comprendre l'interaction entre étiquette CO₂, type de carburant et caractéristiques physiques des véhicules.

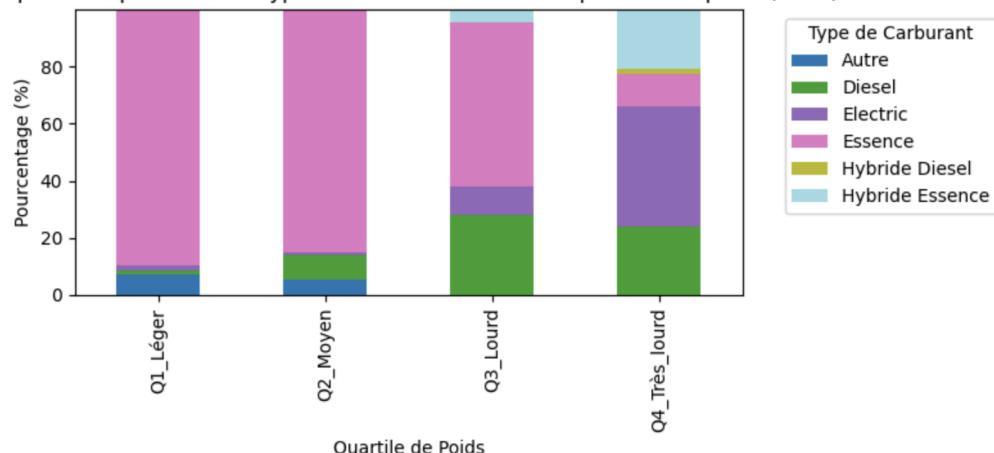
3.6.2 Les quartiles de poids

Lors de nos premières analyses, nous avons constaté que le poids des véhicules joue un rôle déterminant dans les émissions de CO₂ des véhicules thermiques, ainsi

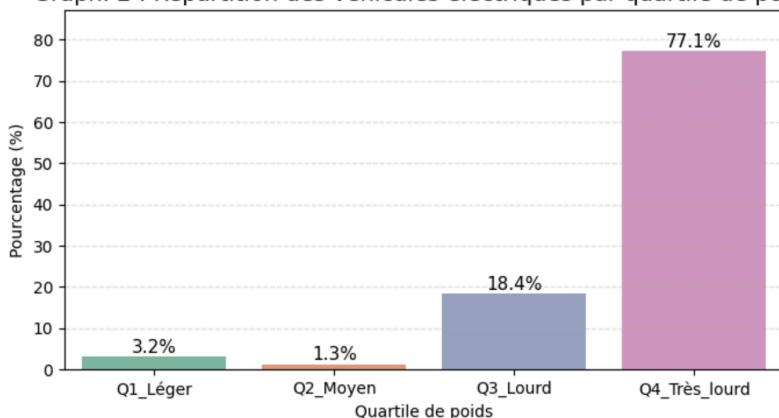
que dans la consommation et la diminution d'autonomie des véhicules électriques. Il est donc apparu pertinent d'étudier la distribution du parc automobile en fonction du poids, en tenant compte du type de motorisation.

Pour ce faire, le jeu de données a été réparti en quatre classes de poids équitablement définies, afin de faciliter la visualisation et l'interprétation des résultats.

Graph. 1 : Répartition des types de carburant selon les quartiles de poids (en %)



Graph. 2 : Répartition des véhicules électriques par quartile de poids



Observations

Graphique 1 – Répartition des types de carburant par quartile de poids

L'analyse révèle que les véhicules les plus légers appartiennent majoritairement à la catégorie essence, représentant environ 90 % des véhicules de ce quartile. À l'inverse, les véhicules très lourds sont principalement électriques, représentant environ 50 % de ce quartile.

Graphique 2 – Répartition des véhicules électriques par quartile de poids

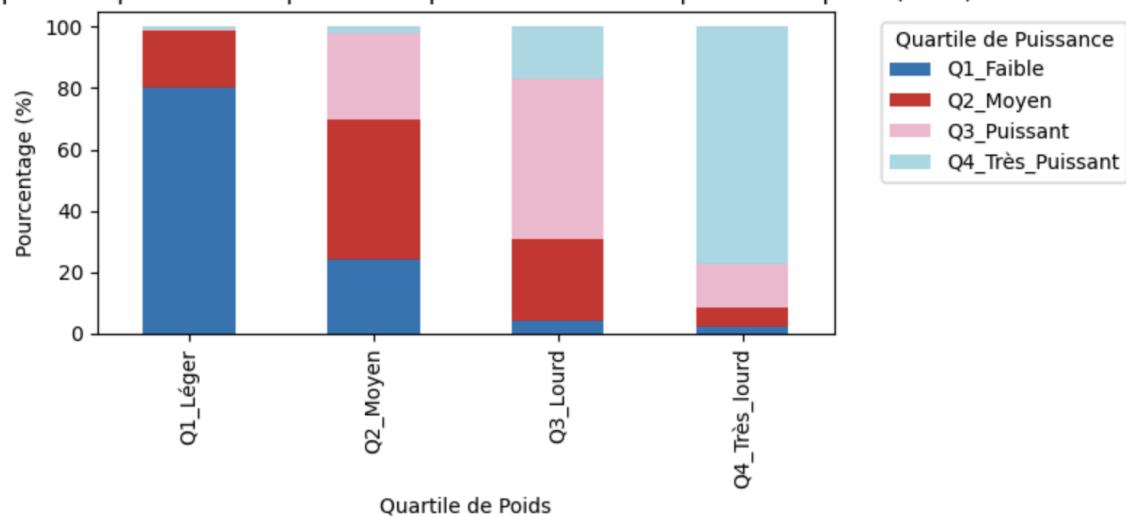
Les véhicules électriques sont surreprésentés dans les quartiles Q3 (lourds) et Q4 (très lourds). Plusieurs facteurs expliquent ce phénomène. Premièrement, le poids

des batteries est significatif : selon Beev⁸, il varie en moyenne entre 250 kg pour les citadines et compactes, et plus de 500 kg pour certaines grandes Tesla, utilitaires et camions. Deuxièmement, comme mentionné précédemment, les SUV représentent environ la moitié du parc électrique, ce qui contribue également à augmenter le poids moyen des véhicules de cette catégorie.

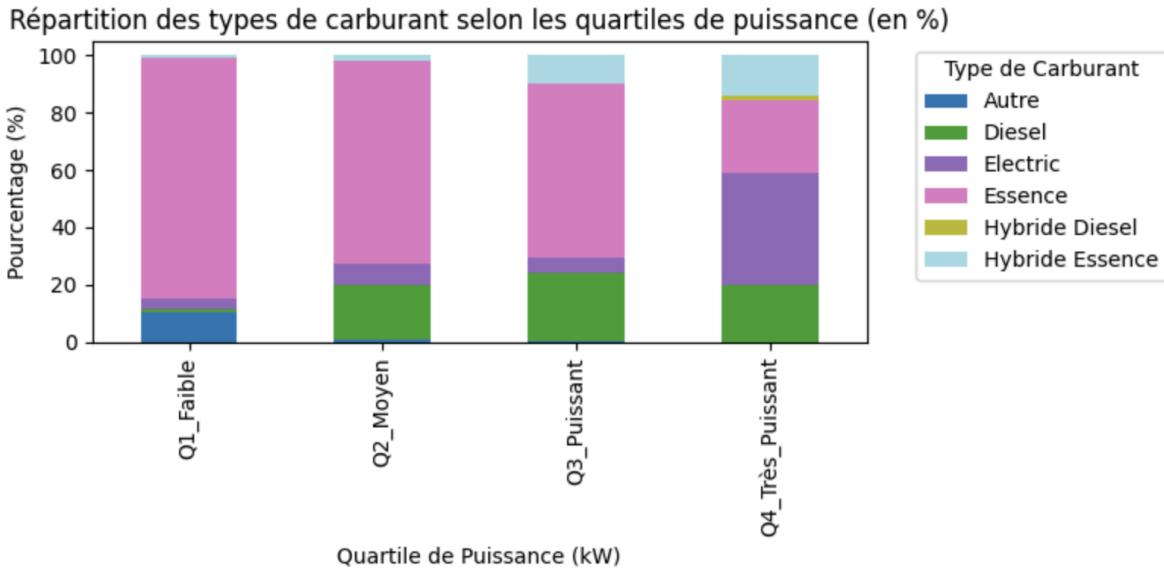
3.6.3 Les quartiles de puissance

Nous avons également approfondi notre analyse en considérant la puissance des véhicules. L'utilisation des quartiles préalablement établis permet de produire des visualisations claires et exploitables, facilitant l'interprétation des tendances au sein du parc automobile.

Graph. 1 : Répartition des quartiles de puissance selon les quartiles de poids (en %)



⁸ Beev : [The weight of an electric battery for an electric car](#)



Observations

Graphique 1 - Répartition des quartiles de puissance selon les quartiles de poids

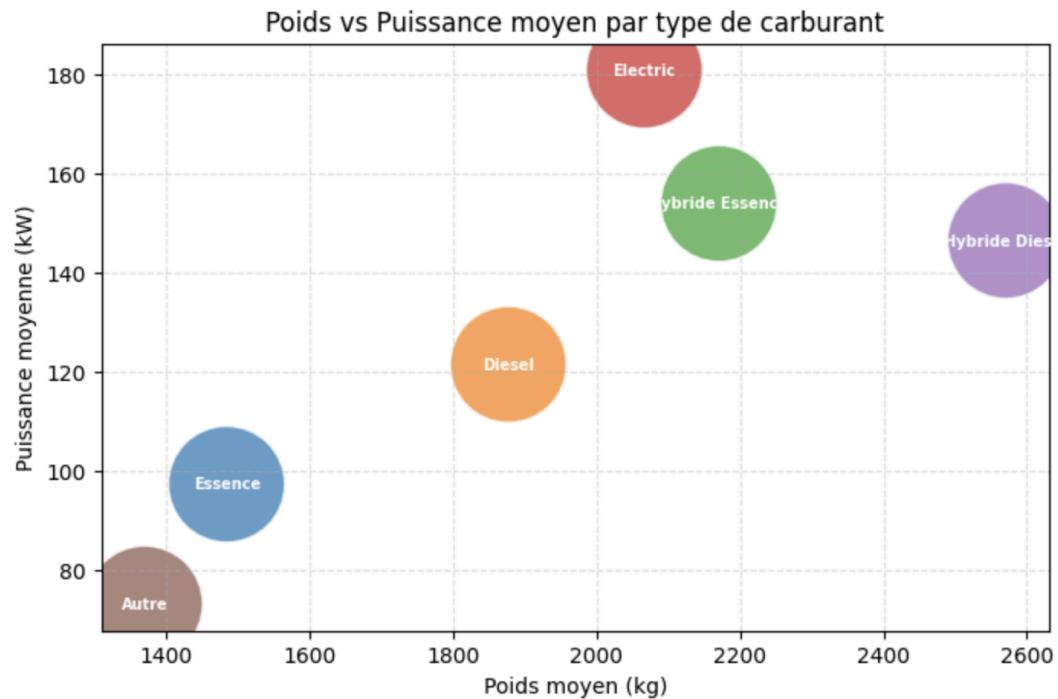
L'analyse met en évidence une tendance claire : plus un véhicule est lourd, plus sa puissance tend à être élevée. Par exemple, environ 80 % des véhicules classés dans le quartile « très puissants » appartiennent également au quartile « très lourds ». Cependant, les quartiles de poids « moyen » et « lourd » présentent une distribution de puissance plus homogène que les quartiles « légers » et « très lourds ».

Cette observation est appuyée par le précédent nuage de points représentant le poids en fonction de la puissance, où l'on constate que la puissance augmente plus rapidement que le poids à partir de 1 500 kg. Cela suggère une tendance à la surmotorisation pour les véhicules lourds et très lourds, caractéristique typique des SUV et des modèles à forte puissance.

Graphique 2 - Répartition des types de carburant selon les quartiles de puissance

Ce graphique confirme une fois de plus les tendances observées dans les analyses précédentes, à savoir la forte représentation des véhicules électriques parmi les véhicules « très puissants ».

Enfin, pour approfondir l'analyse, un graphique à bulles a été réalisé, mettant en relation la puissance (kW) et le poids des véhicules en fonction de leur type de carburant. L'objectif de cette visualisation était de mettre en évidence l'effet SUV, caractérisé par des véhicules combinant un poids élevé et une puissance moteur importante, souvent supérieure aux besoins réels d'utilisation.



Observations

Ce graphique confirme le phénomène des SUV au sein du parc des véhicules électriques, phénomène plus prononcé que pour les véhicules essence. De manière analogue, les véhicules hybrides présentent également cet effet, avec un poids moyen encore plus élevé, soulignant la tendance générale à la surmotorisation et au poids accru dans ces catégories.

3.7 Comparatif du parc automobile au sein de l'UE

Enfin, nous avons aussi réalisé quelques visuels pour avoir une idée plus large de la composition du parc automobile Européen.

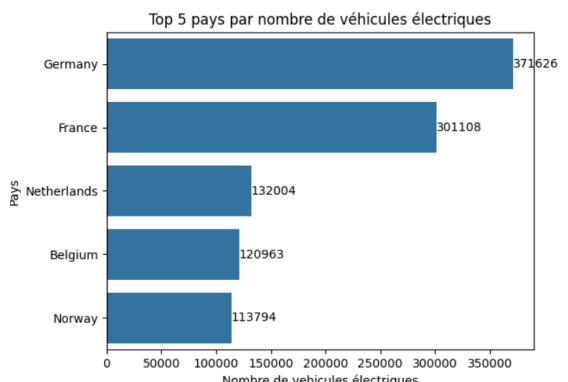
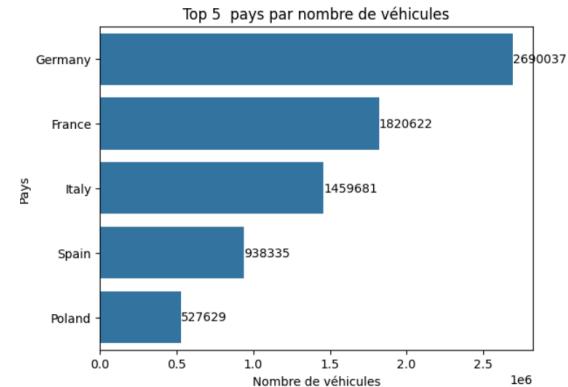
Premièrement, il est intéressant de voir quels pays représentent le plus de véhicules homologués en Europe. Nous avons donc procédé à la création d'un graphique à barres horizontale pour représenter un top 5 des pays ayant le plus de véhicules homologués.

On peut voir dans le graphique à droite que L'Allemagne représente le pays avec le plus de véhicules homologués avec 2 690 037 véhicules, suivi par la France et l'Italie.

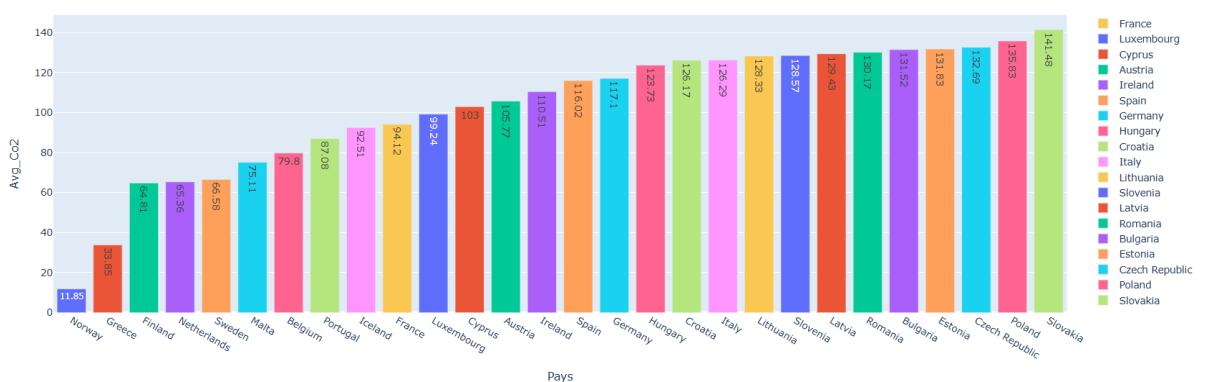
En vue de nos analyses précédentes, il est important aussi de voir quels pays ont le plus de véhicules à motorisations électriques. Dans le graphique de droite on peut voir que l'Allemagne reste en haut de la liste, ainsi que la France en deuxième.

La différence entre le graphique précédent et celui-ci se situe dans les trois dernières positions. On voit que les Pays Bas, la Belgique et la Norvège occupent les dernières places avec près de 120 000 véhicules électriques homologués chacun.

Ainsi pour étudier les effets qu'un passage aux véhicules électriques peut avoir sur un pays, il serait intéressant de porter notre analyse sur l'un des pays avec un nombre important de voitures électriques.



Moyenne d'émissions de CO₂ des véhicules par pays

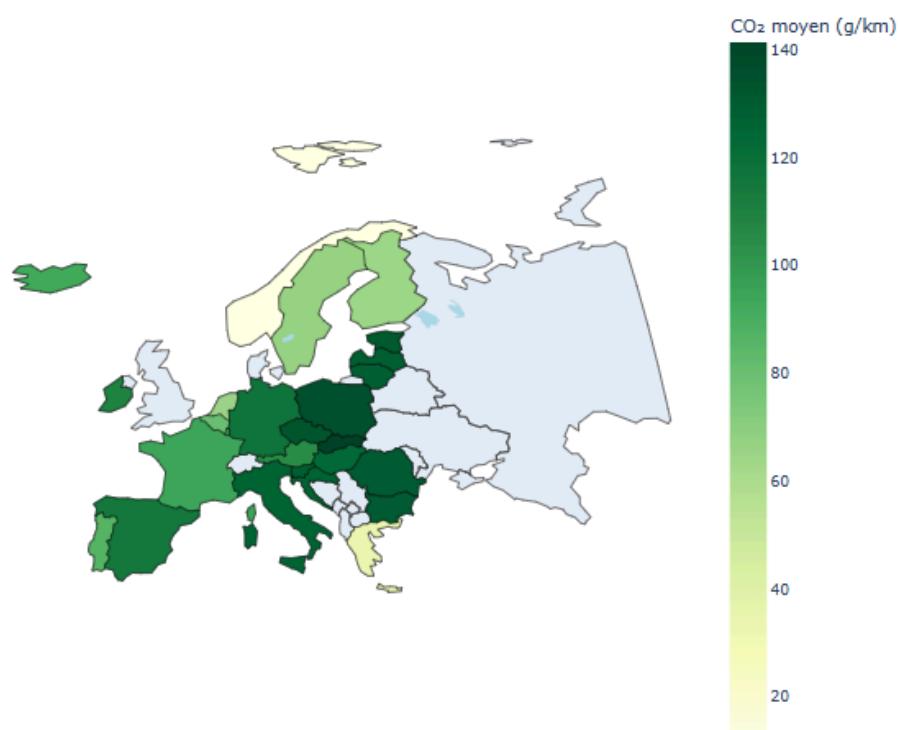


Au vu de notre sujet, le graphique du dessus permet une visualisation de la moyenne des émissions de CO₂ pour chaque pays.

Nous constatons que la Norvège est l'un des pays avec la moyenne la plus basse avec 11.85 g/km de CO₂ en moyenne. Les pays mentionnés précédemment avec le plus grand nombre de véhicules homologués comme la France ou l'Allemagne ont des moyennes plus hautes. La France avec 94.12 g/km de CO₂ et l'Allemagne avec 117.1 g/km de CO₂.

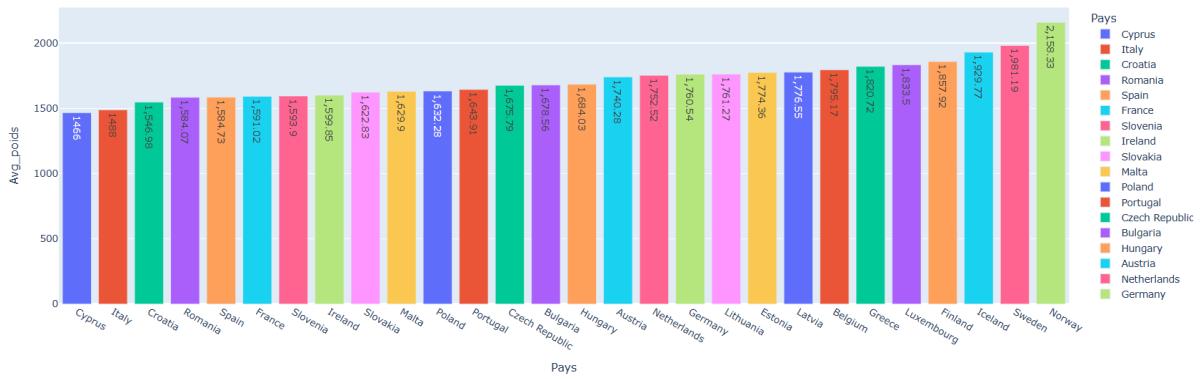
Une carte représentative de la moyenne d'émissions de CO₂ par pays a aussi été créée pour faciliter la comparaison de la moyenne d'émissions de CO₂ à l'aide d'un code couleur comme on peut le voir en dessous.

Émissions moyennes de CO₂



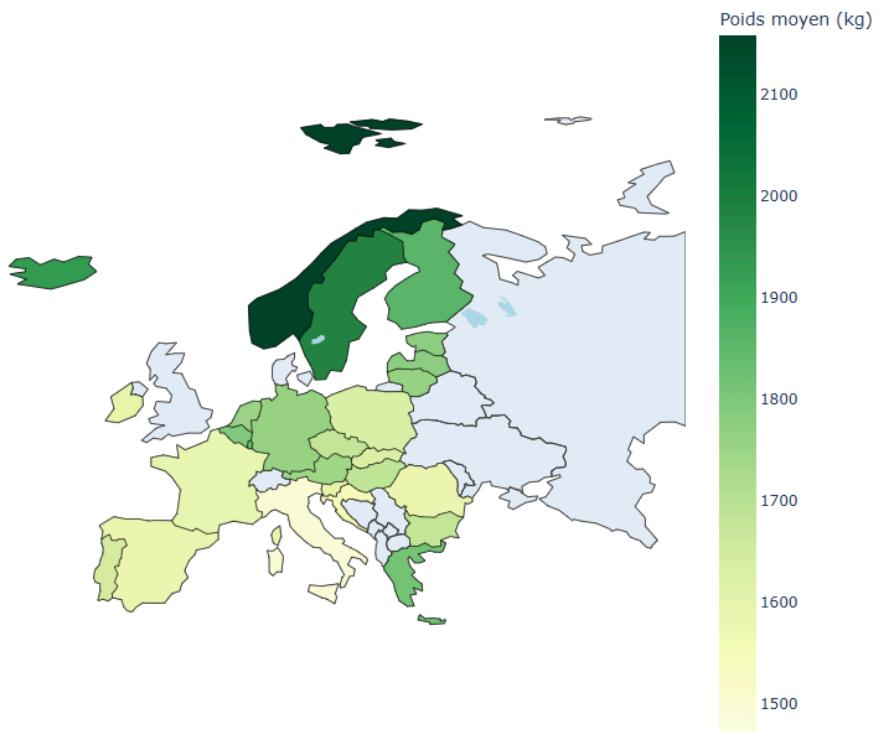
Nous observons donc une forte différence entre un pays comme la Norvège et les pays faisant partie du top 5 en termes de véhicules homologués. Cependant, la France et l'Allemagne possèdent plus de véhicules électriques que la Norvège. Il faut donc creuser plus loin, à l'aide d'autres graphiques comparatifs, pour avoir des indices sur la disparité entre ces moyennes. On peut aussi se demander si un plus grand parc de véhicules représente un frein à l'électrification du parc automobile pour cause d'une nécessité d'infrastructure plus importante.

Moyennes de poids WLTP par pays:



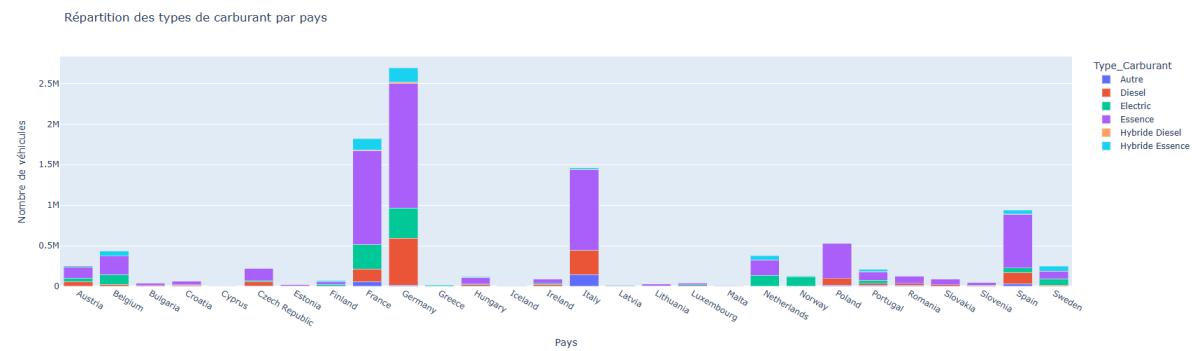
Un autre aspect de notre analyse a porté sur le poids des véhicules. Dans le graphique à barres ci-dessus, nous observons que le poids moyen diverge très peu (autour de 1600 kg par pays), sauf dans le cas de la Norvège qui se démarque par un poids moyen de véhicule de 2158 kg. On retrouve donc le pays avec la meilleure moyenne en termes d'émissions de CO₂ à avoir la plus grande moyenne de poids de véhicules, donc les véhicules les plus lourds en Europe. La carte choroplète en dessous permet une visualisation plus simple du graphique à barres.

Poids moyen kg



A l'aide du graphique situé ci-dessous, nous pouvons voir la répartition des types de carburant par pays. Nous remarquons tout de suite la prédominance de véhicules thermiques pour la France et pour l'Allemagne, ce qui explique la disparité mentionnée précédemment entre les moyennes d'émissions de CO₂, car il y a une

plus grande proportions de véhicules à essence, ce qui entraîne une plus grosse moyenne d'émissions de CO₂.



Dans le cas de la Norvège, nous constatons une forte proportion de véhicules électriques et très peu de voitures à motorisation thermique expliquant ainsi la moyenne très basse d'émissions de CO₂. Un cas atypique est celui de l'Italie qui a plus de véhicules avec d'autres types de motorisations que de véhicules électriques. Une analyse complémentaire sur l'électrification versus l'utilisation d'autres types motorisations pourrait éventuellement être faite.

Pour résumer, nous constatons de grandes divergences lorsque nous étudions les émissions moyennes de CO₂ par pays. Cela peut s'expliquer par la répartition du type de carburant dans ce pays et pas forcément par le nombre de véhicules homologués. Cela explique le paradoxe qui veut que les pays ayant le plus grand nombre de véhicules à faibles émissions de CO₂ ne sont pas forcément ceux qui performent le mieux en termes d'émissions moyennes de CO₂.

Nous observons que le poids est aussi lié au type de carburant utilisé, car les pays qui ont les moyennes les plus élevées sont les pays avec une plus grande répartition de véhicules à motorisation électrique.

3.8 Problématique

Au fil de nos explorations et des premières visualisations, une problématique centrale a progressivement émergé. Celle-ci vise à questionner les limites d'une approche exclusivement fondée sur les émissions de CO₂, qui peut apparaître réductrice, voire trompeuse, dans le cadre de la transition énergétique, au regard de l'influence exercée par d'autres variables, et en particulier le poids des véhicules. Le cas spécifique des véhicules électriques, neutres en émissions directes de CO₂, mais caractérisés par une augmentation significative de leur masse, a notamment contribué à nourrir cette réflexion.

Dans cette perspective, le fil conducteur du dashboard interactif développé sur Looker Studio peut être résumé par la problématique suivante :

Une approche centrée uniquement sur les émissions de CO₂ est-elle suffisante pour piloter efficacement la transition énergétique du secteur automobile ?

4. Modélisation

4.1 Introduction : Looker Studio

- Lien [Looker Studio](#)

Après avoir entrepris l'exploration des données, les avoir nettoyées, dégagé de premières analyses et implémenté l'ajout de nouvelles variables (étiquettes CO₂, quartiles de poids, quartiles de puissance, etc.), il nous faut répondre à la problématique nouvellement définie :

Une approche centrée uniquement sur les émissions de CO₂ est-elle suffisante pour piloter efficacement la transition énergétique du secteur automobile ?

L'**objectif de ce dashboard Looker Studio est le suivant** : permettre à un décideur politique d'avoir une vue globale de ce qui se fait en matière d'homologations de véhicules, thermiques comme électriques, en Europe et en France. Il s'agit ainsi d'offrir un focus sur la pertinence des nouvelles législations à venir en France⁹, sur la base d'analyses de données précises. D'un point de vue méthodologique, nous avons choisi un déroulement en entonnoir, chaque page Looker apportant des informations clés à notre démonstration :

- **Page 1** : Problématique ;
- **Page 2** : Vision globale avec un état des lieux des homologations en Europe ;
- **Page 3** : Constat de l'impact du poids des véhicules sur les homologations des véhicules thermiques, en France comme en Europe ;
- **Page 4** : Analyse des étiquettes CO₂ françaises sur les homologations, avec la possibilité de mesurer ce qu'il en aurait été si un tel système était implanté dans d'autres pays européens ;
- **Page 5** : Questionnement autour des caractéristiques de poids, de puissance et d'autonomie des véhicules électriques homologués ;
- **Page 6** : Enfin, toujours concernant les véhicules électriques, réflexion sur les différentes natures de véhicules électriques et sur l'hégémonie des SUV électriques, notamment en Europe.

⁹ Fin de l'exonération du malus poids pour les véhicules électriques, courant 2026 : "Un véhicule dont la source d'énergie est exclusivement l'électricité est exonéré de la taxe sur la masse en ordre de marche jusqu'au 30 juin 2026 inclus. L'exonération est déterminée en fonction de la date de 1^{re} immatriculation du véhicule. À partir du 1^{er} juillet 2026, pour un véhicule dont la source d'énergie est exclusivement l'électricité, la masse en ordre de marche fait l'objet d'un abattement de 600 kilogrammes." <https://www.service-public.gouv.fr/particuliers/vosdroits/F35950>

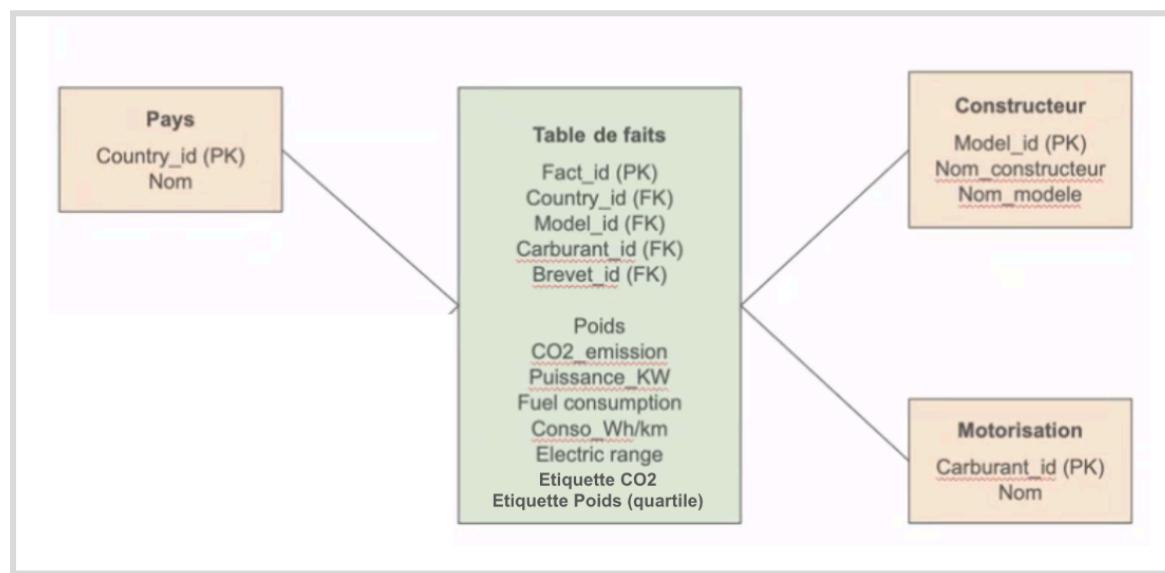
Le dashboard, bien que principalement axé sur le cas français, permettra, via des sélecteurs dédiés, d'élargir si besoin l'analyse et d'accéder aux différentes informations présentes dans la base de données.

4.2 Intégration des données et contraintes de l'outil

Comme évoqué précédemment, plusieurs optimisations de la base ont été réalisées afin de faciliter l'implémentation dans Looker Studio et de garantir des performances suffisantes : restriction aux variables pertinentes et utilisées, recours à un échantillonnage, ainsi que création de variables catégorielles afin de mieux visualiser les informations dans Looker Studio.

La base de données finale, « CO₂ data v3 », est donc utilisée après avoir validé, d'un point de vue technique, que l'utilisation de plusieurs bases selon un schéma relationnel en étoile n'aurait pas été pertinente en termes de performances dans Looker Studio.

Schéma relationnel en étoile comprenant la table de faits ainsi que les catégories Motorisation, Constructeur et Pays.



En termes de poids, le fichier « CO₂ data v3 » nous a permis de disposer de suffisamment de données pour conserver la pertinence des analyses, tout en gardant une marge suffisante en cas d'ajout de nouvelles propriétés ou de champs calculés.

Fichier principal utilisé dans Looker Studio.

Nom	Type de connecteur	Type
CO2 data v3	Importation de fichiers CSV	Intégrée

Après l'implémentation des variables dans Looker Studio, nous nous sommes assurés que les agrégations par défaut étaient correctes et permettaient un usage pertinent dans nos différents graphiques.

Liste des dimensions dans Looker Studio.

Champ ↑	/	Type	Agrégation par défaut	
Dimensions (16)				
Co2_Emission(WLTP)	:	123 Nombre	Somme	
Conso_Wh/km	:	123 Nombre	Somme	
Constructeur	:	ABC Texte	Aucun	
Cout(€)	:	123 Nombre	Somme	
Electric range (km)	:	123 Nombre	Somme	
Etiquette_CO2	:	ABC Texte	Aucun	
France versus Europe	fx :	ABC Texte	Aucun	
Fuel consumption	:	123 Nombre	Somme	
ID	:	123 Nombre	Somme	
Model	:	ABC Texte	Aucun	
Pays	:	🌐 Pays	Aucun	
Poids_Quartile	:	ABC Texte	Aucun	
Puissance_KW	:	123 Nombre	Somme	
PuissanceKW_Quartile	:	ABC Texte	Aucun	
Type_Carburant	:	ABC Texte	Aucun	
WLTP_poids	:	123 Nombre	Somme	

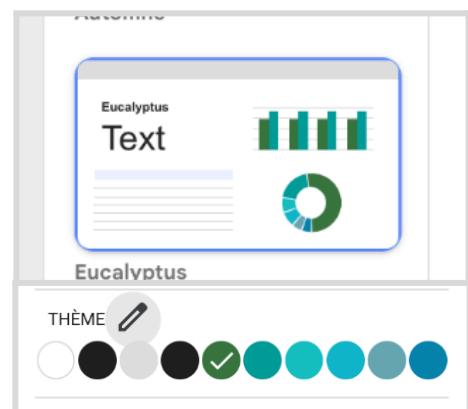
Par ailleurs, l'usage de Looker Studio nous a permis de mener une seconde phase exploratoire grâce à la facilité de visualisation. Nous avons ainsi pu effectuer plusieurs tests, explorer différentes pistes et retenir les plus pertinentes. Finalement, voici l'usage retenu des différents filtres, qui seront ensuite détaillés dans les différentes pages.

Détail des filtres utilisés dans Looker Studio. Après plusieurs tests, plusieurs filtres n'ont pas été utilisés, l'utilisation de sélecteurs ayant été favorisée pour plus d'interaction.

Nom	Utilisé(s) dans le rapport	Description
Electrique	6 graphiques	Inclure Type_Carburant Égal(e) à (=) Electric
France	19 graphiques	Inclure Pays Égal(e) à (=) France
Essence	12 graphiques	Inclure Type_Carburant Égal(e) à (=) Essence

4.3 Choix du thème et mises en page

Afin de garantir une lecture fluide et cohérente des graphiques, tout en respectant une uniformité visuelle au sein de nos rapports, nous avons choisi d'utiliser un thème par défaut de Looker Studio, **Eucalyptus**. Ce choix permet de respecter les règles d'accessibilité et de visualisation, notamment grâce à l'utilisation de couleurs principales harmonisées et de couleurs d'opposition suffisamment contrastées pour assurer une bonne lisibilité.



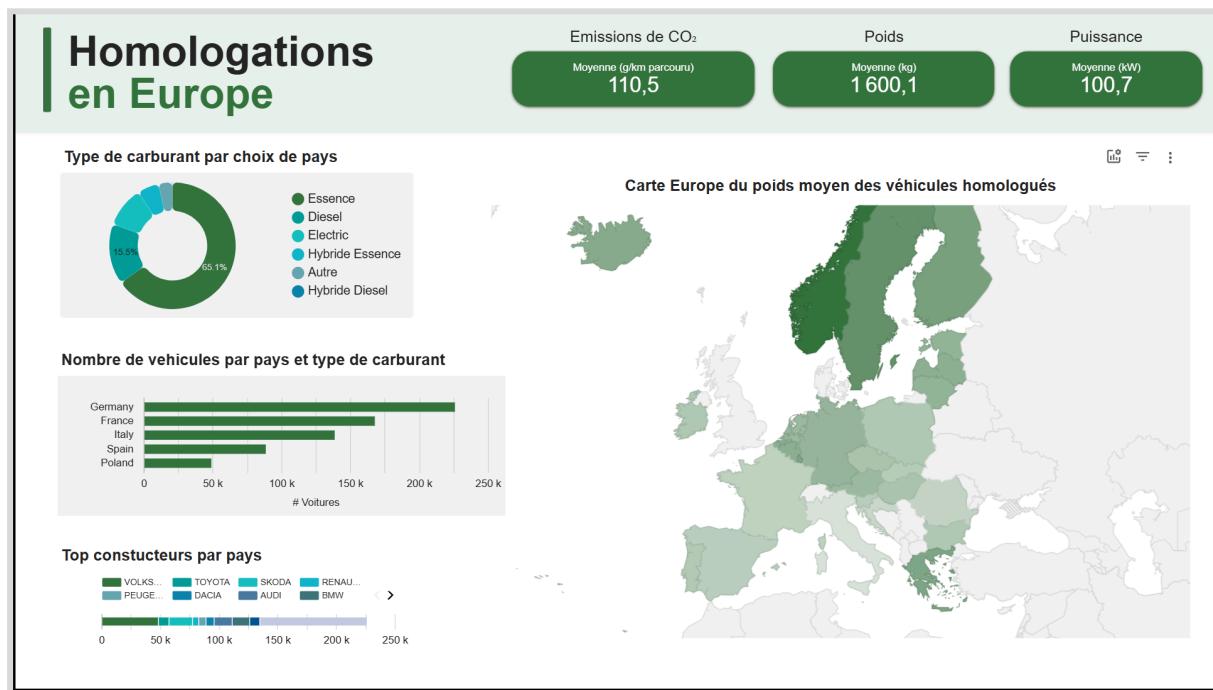
4.4 Détail de la présentation Looker Studio

PAGE 1 : Problématique

The screenshot shows a presentation slide with a light gray background. In the top-left corner, there is a small logo consisting of a teal square with a white icon and the text "DataScientest". Below the logo is a large, semi-transparent circular graphic centered on the slide. Inside this circle, the word "Problématique" is written in bold green capital letters. Below the circle, a question is presented in black text:
Une approche centrée uniquement sur les émissions de CO₂ est-elle suffisante pour piloter efficacement la transition énergétique du secteur automobile ?

At the bottom of the slide, there is a footer with the names of the creators: "Alexis Decloquement | Tom Burret | Stefan Loubry | Stéphane Lable" and the file name "nov25_bootcamp_da".

PAGE 2 : Homologations en Europe



Objectif de la page

L'objectif de cette page est d'apporter, dans un premier temps, une vision globale des différentes variables de la base de données. Première étape de l'analyse en entonnoir, elle permet de comprendre rapidement l'étendue des données et, via un ensemble de sélecteurs, d'y accéder selon l'information recherchée.

Indicateurs clés affichés

- Émissions de CO₂ ;
- Poids des véhicules ;
- Puissance ;
- Type de carburant ;
- Pays ;
- Constructeur.

Détail des graphiques

Bandeau supérieur : Il permet d'obtenir une première lecture des émissions moyennes de CO₂, du poids moyen des véhicules ainsi que de leur puissance moyenne.

Type de carburant par choix de pays : Ce diagramme en anneau permet de visualiser la répartition des types de carburant, en fonction des filtres appliqués via les autres graphiques. Il met en évidence que l'essence est majoritaire, avec près de 65 % du total des homologations.

Nombre de véhicules par pays et type de carburant : Ce diagramme en barres horizontales permet d'avoir une vision des volumes de véhicules homologués par pays, et de comparer rapidement les écarts observés d'un pays à l'autre, en fonction des filtres appliqués via les autres graphiques.

Top constructeurs par pays : Ce diagramme en barres empilées horizontales permet d'obtenir le détail des principaux constructeurs par pays, en fonction des filtres appliqués via les autres graphiques.

Carte Europe : Enfin, la carte de l'Europe permet d'avoir rapidement une vision du poids moyen des véhicules homologués en 2024 à l'échelle européenne. La carte est également impactée par les filtres de carburant ou encore de constructeur.

Filtres et sélecteurs

Sur cette page, il est possible de cliquer sur chacun des pays de la carte afin de filtrer l'ensemble des autres graphiques. L'analyse peut également être affinée en sélectionnant un type de carburant. Enfin, le clic sur un constructeur permet d'analyser la répartition de ce constructeur selon les différents types de carburant et selon le pays sélectionné.

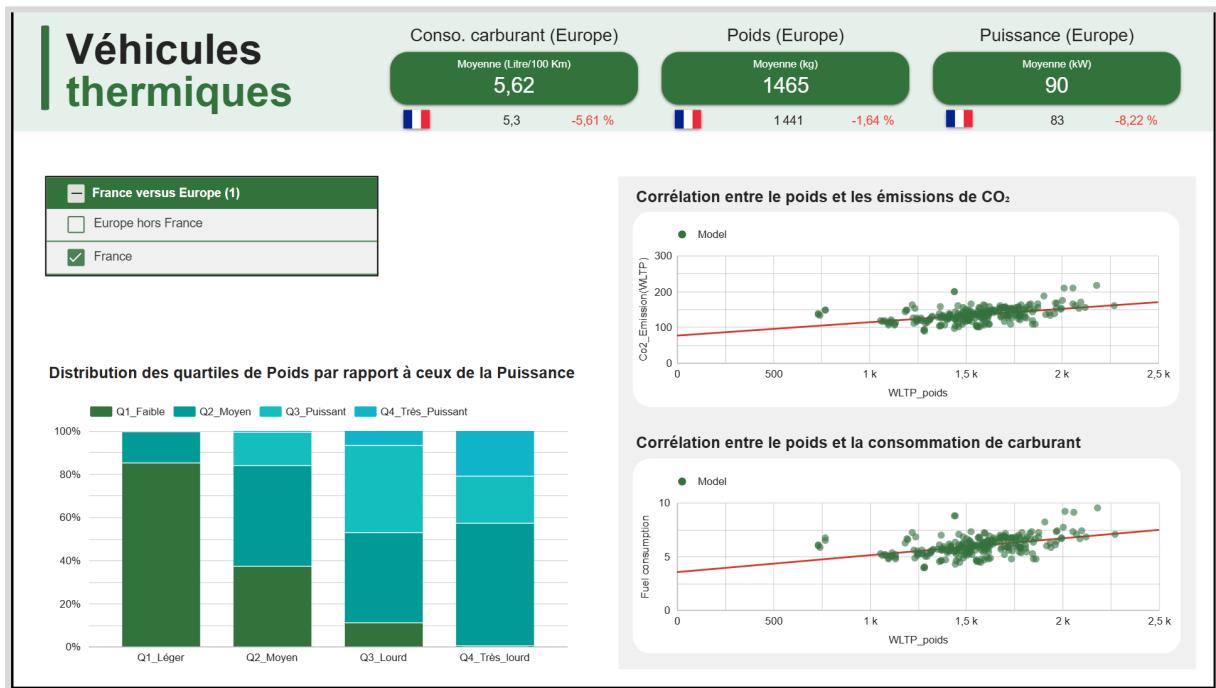
L'objectif est d'offrir un accès global à la quasi-totalité des informations qui seront ensuite détaillées, les pages suivantes faisant office de focus sur les différentes thématiques abordées.

Key insights

- Les données des véhicules homologués diffèrent d'un pays à l'autre, laissant apparaître plusieurs profils de pays. Certains restent majoritairement orientés vers l'essence, tandis que d'autres présentent une part plus importante d'homologations de véhicules électriques. Il existe également des disparités en termes de poids, puissance ou encore consommation CO₂ des véhicules selon les pays ;

- Le poids, la puissance et les émissions de CO₂ semblent suivre des tendances différentes selon les types de carburant, tendances qui seront approfondies dans les pages suivantes du dashboard Looker ;
- Enfin, l'essence demeure majoritaire comme type de carburant à l'échelle du continent européen.

PAGE 3 : Véhicules thermiques



Objectif de la page

L'objectif de la page « Véhicules thermiques » est de mettre en avant la corrélation entre le poids des véhicules, leurs émissions de CO₂ et leur consommation de carburant.

Indicateurs clés affichés

- Consommation de carburant ;
- Poids ;
- Puissance ;
- Pays ;
- Quartiles de poids ;
- Quartiles de puissance ;
- Émissions de CO₂.

Détail des graphiques

Bandeau supérieur : les indicateurs clés présentés sont les moyennes de la consommation de carburant, le poids et la puissance. Ces données sont affichées à l'échelle européenne, avec un comparatif spécifique pour la France. On observe que la France se situe légèrement en dessous de la moyenne européenne, tout en restant sur une tendance similaire.

Distribution des quartiles de poids par rapport à ceux de la puissance : à gauche de la page, il est possible d'analyser un diagramme en barres empilées verticales représentant la distribution des quartiles de poids en relation avec ceux de la puissance. Cette répartition met clairement en évidence que les véhicules appartenant au quartile Q4 (très lourds) sont nettement plus puissants que ceux des quartiles Q1 (légers) et Q2 (moyens).

Corrélation entre le poids et les émissions de CO₂ : à droite de la page, des nuages de points accompagnés d'une droite de régression linéaire permettent d'établir une corrélation forte entre le poids et les émissions de CO₂.

Corrélation entre le poids et la consommation de carburant : la même observation s'applique, une relation claire ressort de ces nuages de points, également accompagnés de leur droite de régression linéaire.

Filtres et sélecteurs

En plus d'apporter une vision précise des homologations sur le territoire français, l'objectif second de cette page est de situer la France par rapport à l'Europe. C'est pourquoi un comparatif est proposé dans le bandeau supérieur. Pour les graphiques situés sous ce bandeau, un filtre par défaut sur la France est appliqué au niveau du pays. Il permet d'analyser la distribution des quartiles de poids par rapport à la puissance pour la France, ainsi que les corrélations entre le poids et les émissions de CO₂, et entre le poids et la consommation de carburant.

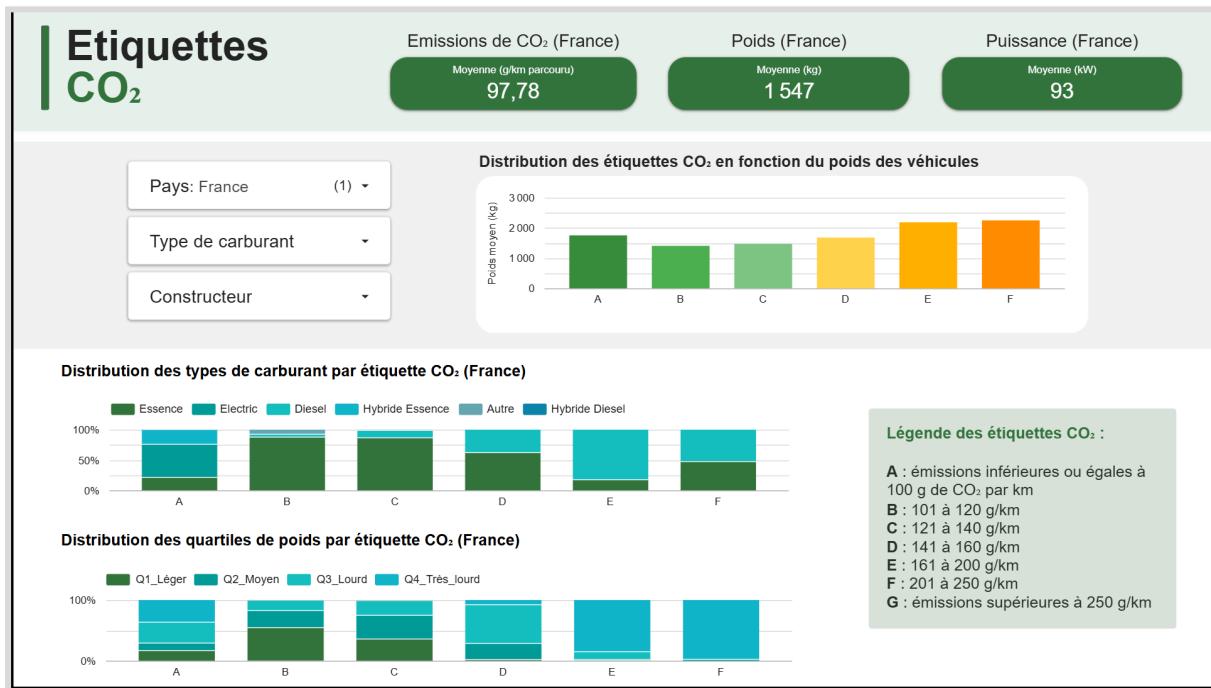
Bien que l'analyse soit principalement centrée sur le territoire français, le filtre « Europe » permet de confirmer qu'à l'échelle européenne, les distributions, bien que non strictement identiques, suivent les mêmes tendances.

Key insights

- La France affiche des valeurs inférieures à la moyenne européenne en consommation de carburant, en poids et en puissance des véhicules thermiques;
- Corrélation positive entre le poids et les émissions de CO₂ : plus un véhicule est lourd, plus ses émissions augmentent ;

- Lien direct entre le poids et la consommation de carburant, confirmé par une tendance linéaire croissante ;
- Les véhicules légers sont majoritairement associés à des puissances plus faibles, tandis que les véhicules lourds concentrent les puissances élevées ;
- Les différences observées s'expliquent par la structure du parc automobile, les choix technologiques et les politiques nationales.

PAGE 4 : Étiquettes CO₂



Objectif de la page

L'objectif de la page est de mettre en avant une nouvelle donnée ajoutée à la base, à savoir la distribution des étiquettes CO₂. Cette donnée, française, permet d'analyser la répartition des véhicules du parc automobile français selon les différentes classes d'étiquettes, de A à F.

Indicateurs clés affichés

- Émissions de CO₂ ;
- Poids ;
- Puissance ;
- Pays ;
- Constructeur ;
- Étiquette CO₂ ;
- Quartiles de poids.

Détail des graphiques

Bandeau supérieur : Dans le bandeau supérieur, les moyennes des variables émissions de CO₂, poids et puissance pour la France sont mises en avant.

Distribution des étiquettes CO₂ en fonction du poids des véhicules : le bandeau gris principal présente un histogramme vertical illustrant la distribution des étiquettes CO₂ en fonction du poids des véhicules. Une première observation notable est que les véhicules classés en étiquette A sont, en moyenne, plus lourds que ceux classés en étiquette B ou C, ce qui semble en inadéquation avec les constats établis dans les pages précédentes. Comme nous le verrons ensuite, ce résultat s'explique par la présence des véhicules électriques.

Distribution des types de carburant par étiquette CO₂ (France) : en observant dans ce diagramme en barres empilées, par exemple, la composition des véhicules classés en étiquette A, on constate qu'une part importante est constituée de véhicules électriques. Ces véhicules, en moyenne plus lourds, n'émettent pas de CO₂ à l'usage et peuvent ainsi biaiser la lecture et la pertinence de l'utilisation des étiquettes CO₂.

Distribution des quartiles de poids par étiquette CO₂ (France) : sur ce second diagramme en barres empilées, on obtient une confirmation de l'observation précédente. Les véhicules classés en étiquette A présentent une distribution de poids bien différente de celle des véhicules B et C, avec une surreprésentation des quartiles Q3 (lourds) et Q4 (très lourds), principalement liée aux véhicules électriques.

Filtres et sélecteurs

Trois sélecteurs sont proposés sur cette page : pays, type de carburant et constructeur. Par défaut, le sélecteur pays est positionné sur la France, tandis que les sélecteurs « type de carburant » et « constructeur » sont laissés sans restriction. Ils permettent néanmoins d'affiner l'analyse selon le carburant ou le constructeur sélectionné. Par ailleurs, en modifiant le pays, même si le système d'étiquettes n'est pas mis en place hors de France, le graphique de distribution permet de visualiser comment seraient répartis les véhicules des autres pays.

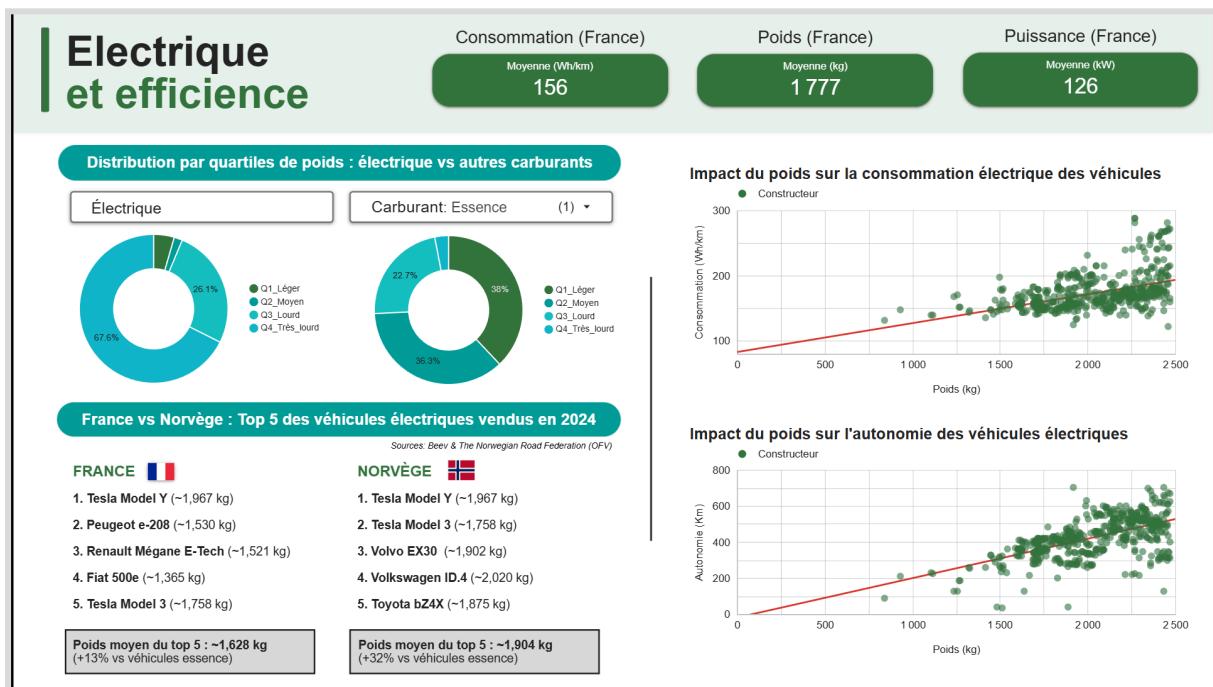
Key insights

- Les émissions moyennes en France restent inférieures à 100 gCO₂/km, correspondant majoritairement aux étiquettes A et B ;
- Relation directe entre le poids du véhicule et l'étiquette CO₂ : plus le poids augmente, plus l'étiquette se dégrade (de A à F) ;
- Les véhicules des classes A à C sont majoritairement plus légers, tandis que

les classes E et F concentrent les véhicules lourds et très lourds ;

- Les étiquettes les plus performantes (A et B) sont fortement associées aux motorisations électriques et hybrides, tandis que les classes les plus élevées restent dominées par les véhicules thermiques ;
- Le poids apparaît comme un facteur structurant des émissions, indépendamment du constructeur, confirmant son rôle clé dans la classification CO₂.

PAGE 5 : Électrique et efficience



Objectif de la page

Maintenant qu'un état des lieux des véhicules thermiques ainsi que des étiquettes CO₂ en France a été réalisé, un focus spécifique sur les véhicules électriques permet de faire ressortir plusieurs conclusions en lien avec la nature des véhicules, selon leur poids, leur autonomie et leur consommation.

Indicateurs clés affichés

- Consommation ;
- Poids ;
- Puissance ;
- Consommation électrique ;
- Autonomie ;
- Quartiles de poids ;
- Type de carburant ;
- Top 5 des véhicules vendus en 2024 (données externes).

Détail des graphiques

Bandeau supérieur : la lecture des variables présentes dans le bandeau supérieur (moyenne de la consommation en Wh/km, du poids et de la puissance) permet d'obtenir une première photographie des véhicules électriques et de constater que leur poids et leur puissance sont en moyenne supérieurs aux thermiques.

Distribution par quartiles de poids : électrique vs autres carburants : sur ce diagramme en anneau, la distribution des quartiles de poids des véhicules électriques apparaît nettement plus élevée que pour les autres types de carburant, avec une forte représentation des quartiles Q3 (lourds) et Q4 (très lourds).

Impact du poids sur la consommation électrique des véhicules : les nuages de points, permettant de mesurer l'impact du poids sur la consommation électrique des véhicules électriques, mettent en évidence une corrélation plus faible que celle observée pour les véhicules thermiques. Un véhicule électrique plus lourd, bien qu'il puisse disposer d'une batterie offrant une autonomie théorique plus importante, tend à consommer davantage en raison de son poids.

Impact du poids sur l'autonomie des véhicules électriques : ces autres nuages de points montrent que l'autonomie des véhicules électriques n'est pas garantie par leur poids. Les points apparaissent plus dispersés autour de la droite de régression linéaire, indiquant qu'un véhicule électrique plus lourd ne garantit pas nécessairement une autonomie nettement supérieure.

France vs Norvège : Top 5 des véhicules électriques vendus en 2024 : afin de compléter les analyses issues de la base de données, un top 5 des véhicules électriques vendus en 2024 permet d'obtenir une première approche des types de véhicules commercialisés et de leur poids approximatif. Un premier constat montre que les véhicules vendus en Norvège, par exemple, sont sensiblement plus lourds que ceux vendus en France. Il existe donc à la fois une différence de caractéristiques entre les véhicules électriques et les autres types de motorisation, mais également au sein même de la catégorie électrique. On observe notamment que le parc français est clairement plus léger que le parc norvégien, avec une moyenne du top 5 de 1 628 kg contre 1 904 kg pour la Norvège.

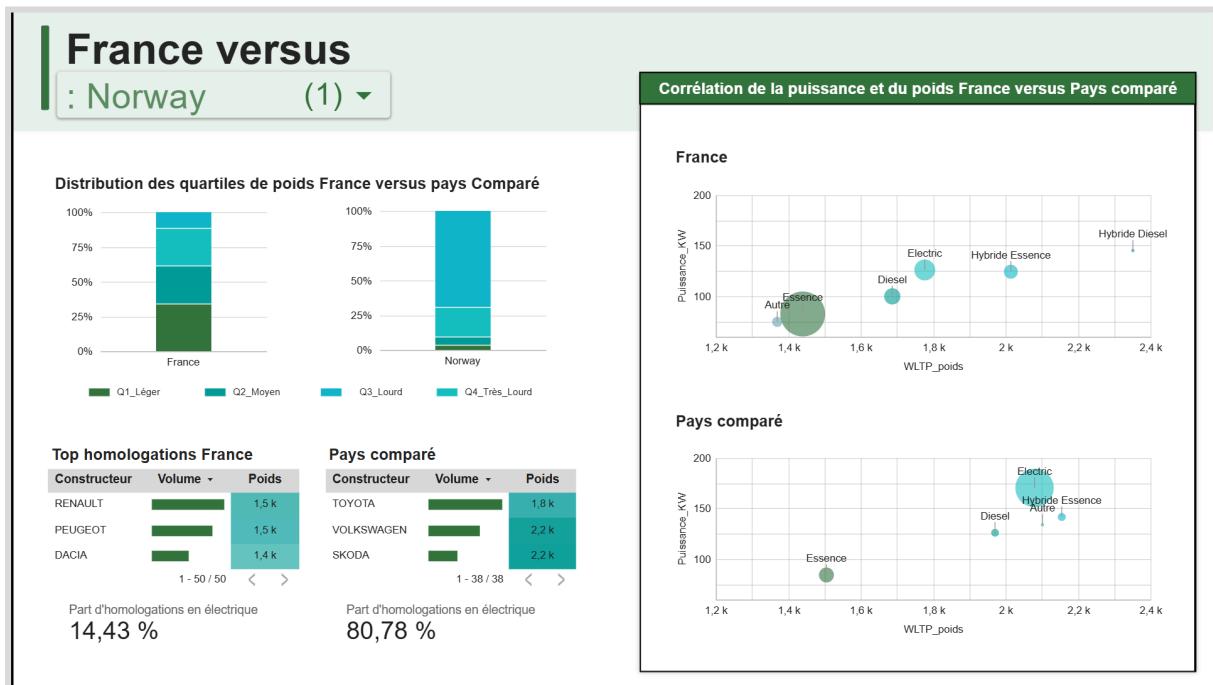
Filtres et sélecteurs

Sur cette page, un seul sélecteur est proposé et permet de comparer un type de carburant à l'électrique. Par défaut, le diagramme en anneau de distribution des quartiles de poids situé à gauche est fixé sur l'électrique et ne peut pas être modifié. Celui de droite permet de sélectionner les autres types de carburant pour la comparaison.

Key insights

- Le poids impacte directement la consommation électrique : plus un véhicule est lourd, plus la consommation (en Wh/km) augmente ;
- Un lien négatif est observé entre le poids et l'autonomie : les véhicules lourds affichent une autonomie plus variable et globalement moins efficiente ;
- Les véhicules électriques sont majoritairement concentrés dans les quartiles de poids élevés, contrairement aux motorisations thermiques ;
- Le poids reste un facteur clé de performance énergétique, y compris pour les motorisations électriques.

PAGE 6 : France versus [Norvège]



Objectif de la page

L'objectif de cette page est de proposer un comparateur par pays entre la France et un autre pays au choix, sur des variables de distribution des quartiles de poids, de corrélation entre la puissance et le poids, ainsi qu'un top des homologations France versus pays comparé, avec une mise en avant spécifique pour les homologations électriques.

Indicateurs clés affichés

- Pays ;
- Quartiles de poids ;
- Constructeur ;
- Poids ;
- Part de véhicules électriques ;
- Puissance.

Détail des graphiques

Distribution des quartiles de poids : deux histogrammes en barre portant sur les quartiles de poids, avec une légende commune, permettent de comparer la distribution des quartiles de poids en France et dans le pays comparé. Par défaut, on observe qu'en Norvège les véhicules proposés sont nettement plus lourds. Cela s'explique d'une part par la forte proportion de véhicules électriques dans les homologations norvégiennes et, d'autre part, par le fait que les véhicules électriques homologués en Norvège sont particulièrement lourds et puissants (type SUV).

Top homologations France versus Pays Comparé : en dessous, les tops d'homologation de la France et du pays comparé permettent d'obtenir une première lecture des constructeurs dominants ainsi que des poids moyens des véhicules homologués et de comparer les différentes natures de parc automobile. Afin d'apporter un contexte supplémentaire, la part d'homologations électriques est également mise en avant.

Corrélation de la puissance et du poids, France versus pays comparé : enfin, sur le panneau de droite, constitué de nuages de points accompagnés de leurs droites de régression linéaire, les corrélations entre le poids et la puissance sont présentées selon le type de carburant. Le premier graphique se concentre sur la France, tandis que le second porte sur le pays comparé. La taille des bulles permet par ailleurs de visualiser le nombre de véhicules homologués (record count) par type de carburant.

Filtres et sélecteurs

Un sélecteur pays est présent dans le bandeau supérieur et positionné par défaut sur la Norvège. Il permet de compléter l'analyse menée sur la page précédente entre la France et la Norvège. Afin d'apporter davantage d'interactivité, il est possible de modifier le pays sélectionné et d'adapter ainsi l'ensemble des graphiques liés au pays comparé.

Key insights

- Les véhicules homologués en Norvège sont en moyenne plus lourds et plus puissants, notamment pour les motorisations électriques ;
- Les constructeurs dominants diffèrent : la France privilégie des marques généralistes et des modèles compacts, tandis que la Norvège accueille davantage de modèles premium et de SUV ;
- La corrélation poids/puissance est plus marquée en Norvège, traduisant un positionnement de marché orienté vers des véhicules plus puissants.

5. Conclusion

Suite à ce travail exploratoire, statistique et de visualisation portant sur un jeu de données de plusieurs millions de véhicules homologués dans l'Union européenne en 2024, il apparaît que le **parc automobile tend à devenir plus propre en termes d'émissions de CO₂**, avec une part croissante des véhicules électriques dans les immatriculations. Dans certains pays, cette évolution est particulièrement marquée, notamment la Norvège dont les ventes de véhicules électriques représentent 89% des ventes totales de véhicules neufs en 2024 selon l'Argus¹⁰, ce qui en fait un cas extrême en Europe.

Selon une étude¹¹ de 2025 de l'*International Council on Clean Transportation* (ICCT), les émissions de gaz à effet de serre sur l'ensemble du cycle de vie d'un véhicule électrique à batterie vendu en Europe sont estimées à **environ 73 % inférieures** à celles d'un véhicule à essence équivalent, cela même en prenant en compte la production de la batterie et la fabrication du véhicule. Cette réduction s'explique par la combinaison d'une électricité de plus en plus décarbonée en Europe et de l'absence de rejet de CO₂ à l'usage pour les véhicules électriques.

Néanmoins, il est nécessaire de reconnaître que les véhicules électriques présentent également **des externalités environnementales** significatives tout au long de leur cycle de vie. La fabrication des batteries repose sur l'extraction et le traitement de minéraux critiques, qui sont énergivores et associés à des impacts sociaux et environnementaux forts. Sans oublier qu'Eni Plenitude¹² nous rappelle à juste titre que la production électrique utilisée pour la recharge peut encore intégrer des sources fossiles selon les pays, bien que la tendance à l'échelle européenne soit à une décarbonation progressive.

La seule prise en compte des émissions de CO₂ en phase d'utilisation n'est donc pas suffisante pour évaluer l'efficacité environnementale globale d'un véhicule. Une approche plus complète doit intégrer **l'efficience énergétique et les caractéristiques physiques des véhicules**, notamment le poids. Nos analyses ont mis en évidence un effet de masse important : le poids croissant des véhicules, notamment du fait de la taille des batteries et de la popularité des SUV, tend à réduire les gains énergétiques attendus de l'électrification.

Pour orienter efficacement la transition énergétique du secteur automobile, il est essentiel de coupler des objectifs de réduction des émissions de CO₂ avec des incitations à produire et à adopter des véhicules **plus légers et plus efficaces**. Cela implique de résoudre ce que nous avons nommé « l'effet SUV », observé tant pour les véhicules thermiques que pour les véhicules électriques. Autrement dit, la commercialisation de véhicules qui privilégient souvent le confort, la puissance et le style au détriment de l'efficience énergétique.

¹⁰ L'Argus : [Émissions de CO₂. Les pools ont permis aux constructeurs d'atteindre leurs objectifs en 2024](#)

¹¹ International Council on Clean Transportation (ICCT) : [Life-cycle greenhouse gas emissions from passenger cars in the european union in 2025](#)

¹² Eni Plenitude : [Bilan carbone voiture électrique vs essence](#)

Sur le plan réglementaire, des initiatives récentes visent à intégrer des critères supplémentaires aux seules émissions de CO₂ dans les politiques publiques. En France, par exemple, les barèmes de malus¹³ évoluent pour renforcer l'encadrement du poids des véhicules et encourager à des profils plus légers. Toutefois, ces mesures comportent encore des exemptions et ne sont pas suffisamment strictes pour garantir un changement profond des pratiques de conception et de consommation.

En somme, pour qu'un marché de l'automobile réellement « vert » puisse véritablement voir le jour, les politiques publiques doivent aller plus loin que la seule prise en compte des émissions CO₂ à l'usage, et intégrer des critères d'efficience énergétique, de poids et de cycle de vie complet dans leurs réglementations et leurs dispositifs d'incitation.

¹³ ecologie.gouv.fr : [Fiscalité environnementale relative aux véhicules](#)

6. Pour aller plus loin : Modélisation Machine Learning

Dans le contexte du durcissement des réglementations environnementales et de la nécessité de réduire les émissions de gaz à effet de serre, la capacité à estimer précisément les émissions de CO₂ des véhicules constitue un enjeu majeur pour les acteurs du secteur automobile et les décideurs publics. Les émissions de CO₂ dépendent de nombreuses caractéristiques techniques telles que la consommation de carburant, le poids du véhicule, la puissance du moteur ou encore la technologie de motorisation, rendant leur modélisation non triviale.

L'objectif de cette partie est de mettre en place une approche de Machine Learning supervisée afin de prédire les émissions de CO₂ (norme WLTP) à partir des caractéristiques techniques des véhicules. Il s'agit d'un problème de régression, dans lequel la variable cible, donc celle que nous souhaitons prédire, est continue. Les véhicules entièrement électriques ont été exclus de l'analyse afin d'éviter un biais lié aux émissions nulles à l'usage et de se concentrer sur les motorisations thermiques et hybrides.

La problématique centrale de cette modélisation peut ainsi être formulée comme suit : *Dans quelle mesure les caractéristiques techniques d'un véhicule thermique permettent-elles de prédire de manière fiable et généralisable ses émissions de CO₂ ?*

Pour répondre à cette question, un pipeline de modélisation complet a été mis en œuvre, incluant le prétraitement des données, la sélection des variables pertinentes, l'entraînement d'un modèle de régression interprétable, ainsi qu'une évaluation rigoureuse des performances à l'aide de la validation croisée et de métriques adaptées. Cette démarche vise à garantir des résultats robustes, interprétables et exploitables, tout en évitant les phénomènes de sur-apprentissage.

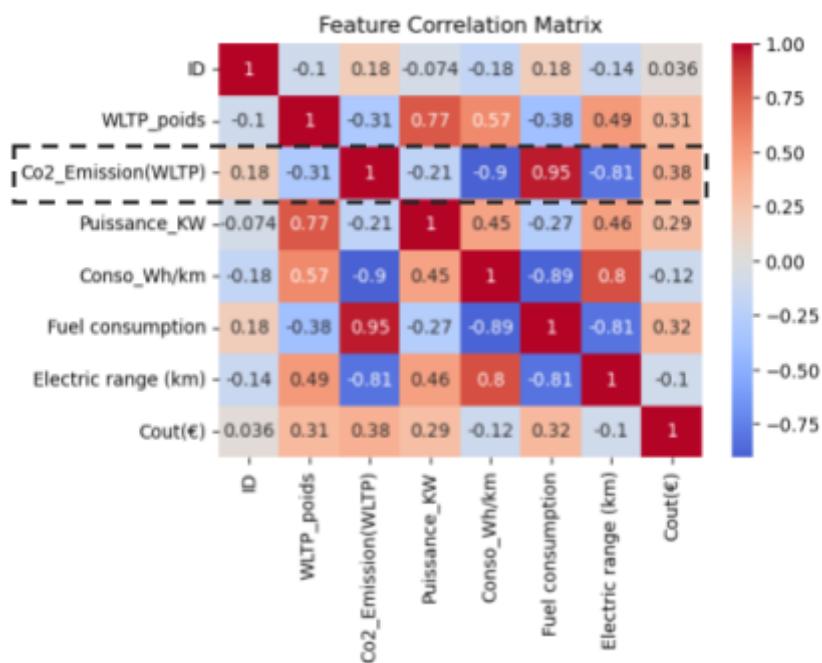
Problématique & cadre:

- Objectif : **prédire les émissions de CO₂ (WLTP)** à partir des caractéristiques techniques d'un véhicule
- Type de problème : **régression supervisée**
- Véhicules électriques exclus afin d'éviter un biais lié aux émissions nulles à l'usage. Ainsi que les Types de carburants "Autre" dont l'émission de Co2 est à 0.

Sélection des variables:

- Conservation des variables à **forte valeur explicative** :
 - poids (WLTP_poids) ;
 - puissance (Puissance_KW) ;
 - consommation de carburant ;
 - type de carburant ;
 - constructeur ;
 - pays.
- Suppression des variables redondantes ou non pertinentes pour la prédiction.

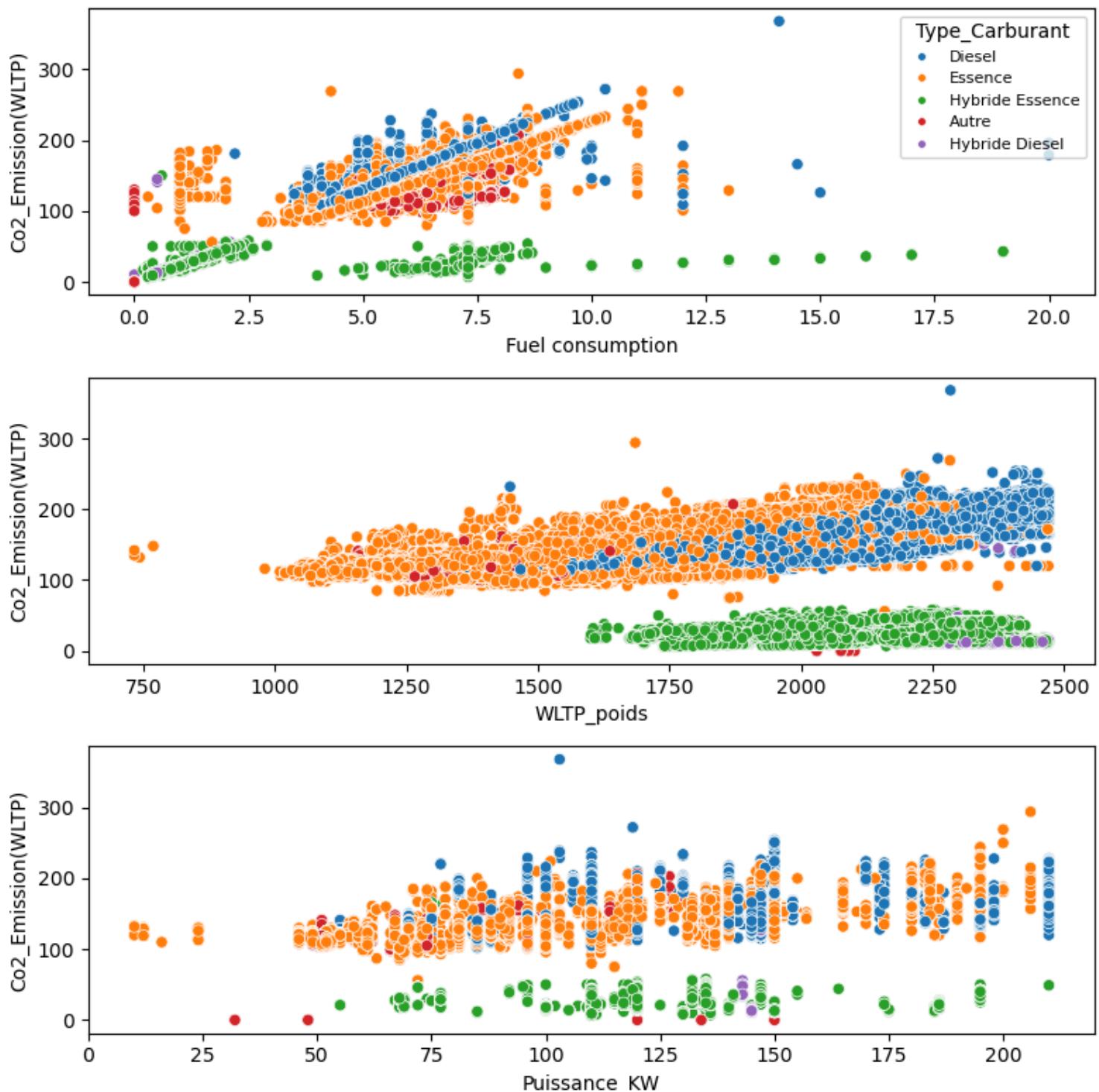
Analyse exploratoire & corrélations:



- Calcul de la **matrice de corrélation** sur les variables numériques.
- La matrice de corrélation met en évidence des relations très fortes entre les émissions de CO₂ et certaines variables clés, notamment la consommation de carburant et les indicateurs liés à l'électrification. Les corrélations négatives

observées s'expliquent par la coexistence de différentes technologies de motorisation dans le jeu de données et ne traduisent pas une incohérence des données.

- Justification de l'utilisation d'un **modèle multivarié**.



L'analyse exploratoire des données met en évidence des relations fortes et cohérentes entre les caractéristiques des véhicules et leurs émissions de CO₂ (WLTP). En particulier, la consommation de carburant présente une relation quasi linéaire avec les émissions de CO₂, confirmée à la fois par les visualisations et par une corrélation très élevée. Cette relation s'explique par le lien physique direct entre la consommation énergétique d'un véhicule et ses émissions.

D'autres variables, telles que le poids (WLTP) et la puissance du moteur, montrent une influence plus indirecte et plus diffuse sur les émissions. Leur effet dépend fortement de la technologie de motorisation. En effet, les véhicules hybrides et électrifiés, généralement plus lourds en raison de la présence de batteries, présentent néanmoins des niveaux d'émissions de CO₂ plus faibles. Cette coexistence de différentes technologies introduit des effets de composition dans le jeu de données, expliquant certaines corrélations globales contre-intuitives observées lors de l'analyse bivariée.

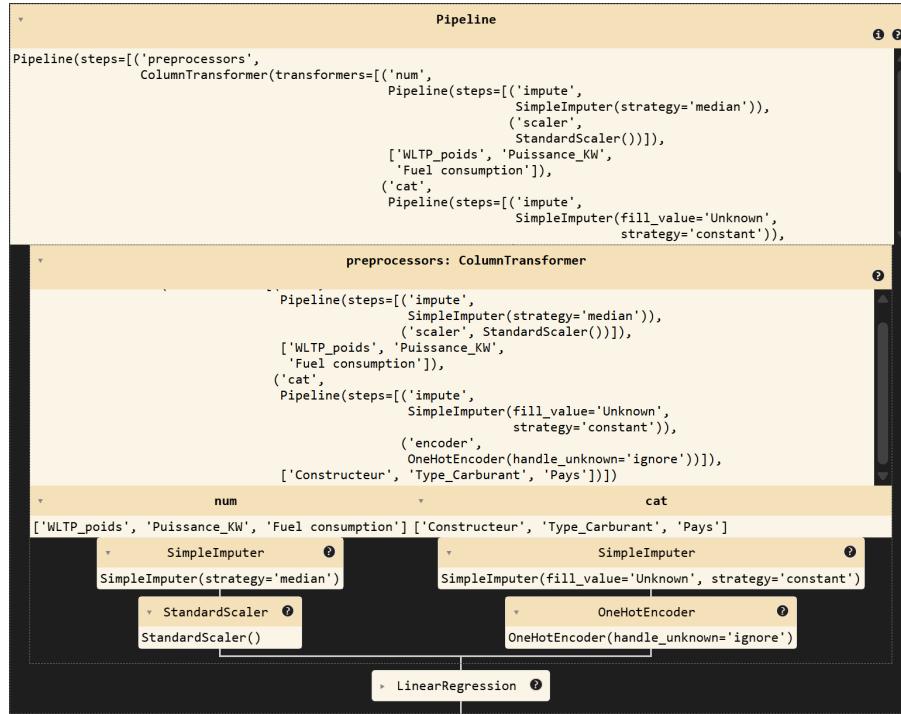
Prétraitement des données (Pipeline)

- Séparation des variables :
 - Numériques (ColumnTransformer)
 - Catégorielles (ColumnTransformer)
- **Variables numériques :**
 - Imputation par la médiane (robuste aux valeurs extrêmes)
 - SimpleImputer(strategy = 'median')
 - Standardisation pour homogénéiser les échelles
 - StandardScaler()
- **Variables catégorielles :**
 - Imputation par la valeur la plus fréquente
 - SimpleImputer(strategy = 'constant')
 - Encodage One-Hot-Encoder avec gestion des catégories inconnues
- Utilisation d'un **Pipeline + ColumnTransformer** pour :

- Éviter les fuites de données
- Garantir la reproductibilité
- Automatiser le preprocessing

Choix du modèle

- Utilisation d'une **régression linéaire** :
- formule : $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_9x_9 + \epsilon$
 - ou :
 - **y** : émissions de CO₂ (variable cible)
 - **x₁,x₂,...,x₉** : **variables explicatives**
(consommation, poids, puissance, type de carburant, constructeur, etc.).
 - **$\beta_1,...,\beta_9$** : **coefficients estimés par le modèle**, qui mesurent l'influence de chaque variable sur le CO₂, toutes choses égales par ailleurs.
 - **β_0** : **intercept**, (correspondant avec standardisation) à la **valeur moyenne du CO₂ prédite par le modèle**.
 - **ϵ /epsilon** : **erreur résiduelle**, c'est-à-dire la part du CO₂ non expliquée par les variables du modèle (bruit)
 - Modèle simple et interprétable
 - Bonne base de référence (baseline)
 - Adapté à un projet d'analyse décisionnelle
- Permet de comprendre l'impact des variables sur les émissions de CO₂



Évaluation du modèle

- **Validation croisée :**

Entraîne et évalue le modèle plusieurs fois sur des partitions/plis (**folds**) de données différentes afin d'obtenir une estimation plus robuste.

- **K-Fold ($k = 10$) :**

Découpe le jeu de données en k sous-ensembles, utilisés alternativement pour l'entraînement et la validation du modèle.

- Métriques utilisées :

- **R² pour la capacité explicative :**

Permet de vérifier si le modèle parvient à capter les relations entre les variables explicatives et la variable cible. Plus la valeur est élevée, plus le modèle traduit une meilleure capacité de prédiction.

- **MSE (Mean Square Error) pour l'erreur de prédiction :**

Plus la MSE est faible, plus les prédictions du modèle sont proches des valeurs réelles, ce qui indique une meilleure précision des prédictions.

- Comparaison des scores :
 - Train vs test
 - Détection de l'overfitting ou underfitting

model	LinearRegression
R ² - Test_score	0.929
R ² - Train_score	0.929
MSE - Test_score	73.461
MSE - Train_score	73.393
time_sec	270.30879282951355
RMSE	8.57
MAE	5.32

Interprétation des résultats :

Le **coefficent de détermination (R²)** mesure la capacité du modèle à expliquer la variabilité de la variable cible à partir des variables explicatives. Une valeur proche de 1 indique que le modèle parvient à capturer une grande partie de la structure présente dans les données. Dans notre analyse, les scores obtenus en "entraînement" et en "validation" sont identiques ($R^2 = 0,928$), ce qui montre que le modèle explique de manière stable et cohérente la relation entre les caractéristiques des véhicules et leurs émissions de CO₂, sans perte de performance sur des données non vues.

L'**erreur quadratique moyenne (MSE)** quantifie l'écart moyen entre les valeurs prédites par le modèle et les valeurs réelles observées. Dans le cadre de notre analyse, les valeurs de MSE observées sont très proches entre l'échantillon d'entraînement (73,461) et l'échantillon test (73,393), ce qui indique que la précision des prédictions reste constante lorsque le modèle est appliqué à des données nouvelles. La **Root Mean Squared Error (RMSE)** est d'environ **8,57 g de CO₂ par kilomètre**.

La RMSE permet de mesurer l'erreur moyenne du modèle tout en pénalisant davantage les écarts importants entre les valeurs prédites et les valeurs réelles, ce qui en fait un bon indicateur de la robustesse globale du modèle.

En complément, la **Mean Absolute Error (MAE)** est égale à **5,32 g de CO₂ par kilomètre**. Cette métrique, plus intuitive, représente l'erreur moyenne absolue commise par le modèle et offre une lecture directe et facilement interprétable de la précision des prédictions.

Ainsi, la **RMSE** est particulièrement adaptée pour détecter et pénaliser les erreurs importantes, tandis que la **MAE** est privilégiée pour communiquer de manière simple et concrète sur la performance moyenne du modèle.

La similitude des scores entre les jeux d'entraînement et de test montre que le modèle ne sur-apprend pas les données d'apprentissage et généralise correctement. Ces résultats suggèrent que la régression linéaire est bien adaptée au problème étudié et que les relations entre les variables explicatives et les émissions de CO₂ sont suffisamment linéaires et structurées pour être capturées efficacement par ce type de modèle.

Vérification des résultats à l'aide de la permutation de la target.

- Estimation de data leakage (fuite de données)

```
● ● ●
1 y_shuffled = np.random.permutation(target)
2
3 r2_shuff = cross_val_score(model_LR, data, y_shuffled, cv=kf, scoring="r2", n_jobs=-1).mean()
4 print("R2 avec y mélangé :", r2_shuff)
```

Résultat :

R2 avec y mélangé : -0.00011529047304144057

Interprétation :

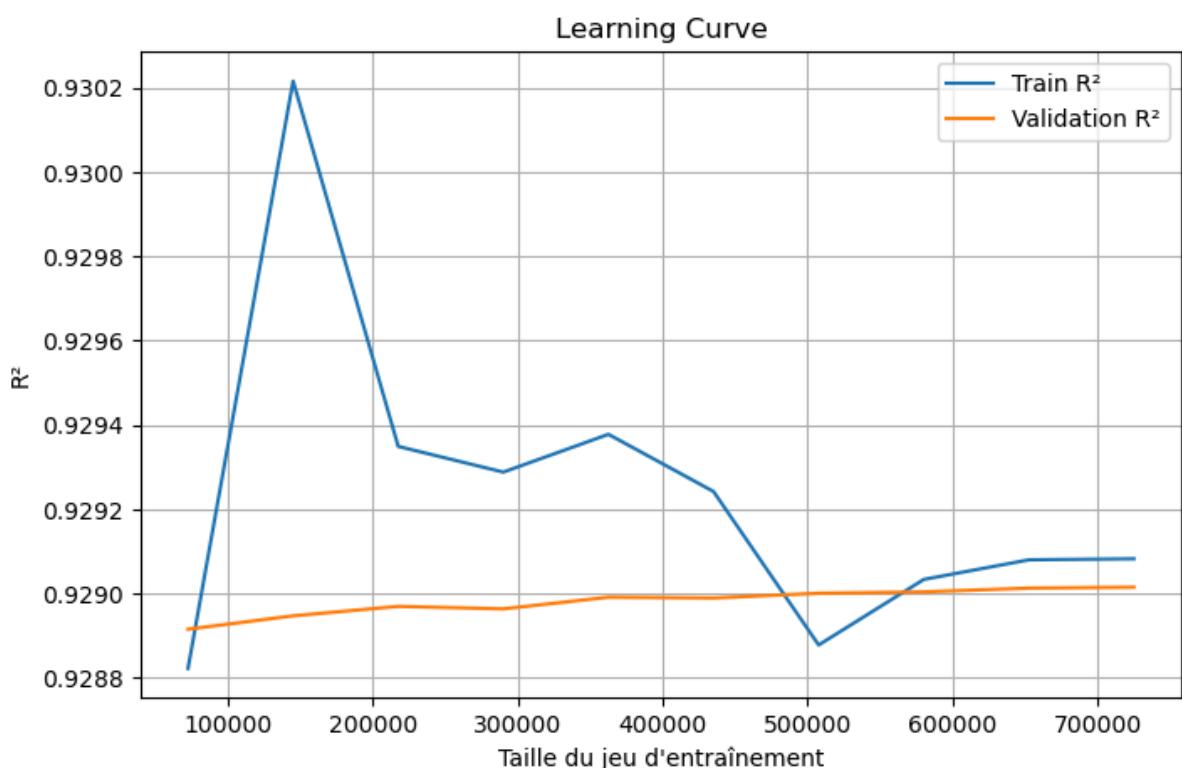
Un test de permutation de la variable cible a été réalisé afin de vérifier l'absence de fuite d'information. Après mélange aléatoire des valeurs de la cible, les performances du modèle chutent vers un R² proche de zéro, confirmant que les bonnes performances initiales reposent sur des relations réelles entre les variables.

La permutation aléatoire de la variable cible conserve la distribution globale des valeurs (minimum, maximum, moyenne, variance ainsi que le nombre d'observations), mais détruit complètement les correspondances entre les variables explicatives et la cible. Toute relation statistique entre les données d'entrée et la variable à prédire est ainsi supprimée, ce qui permet de vérifier que les

performances du modèle reposent sur une relation réelle et non sur une fuite d'information.

Analyse de la learning curve

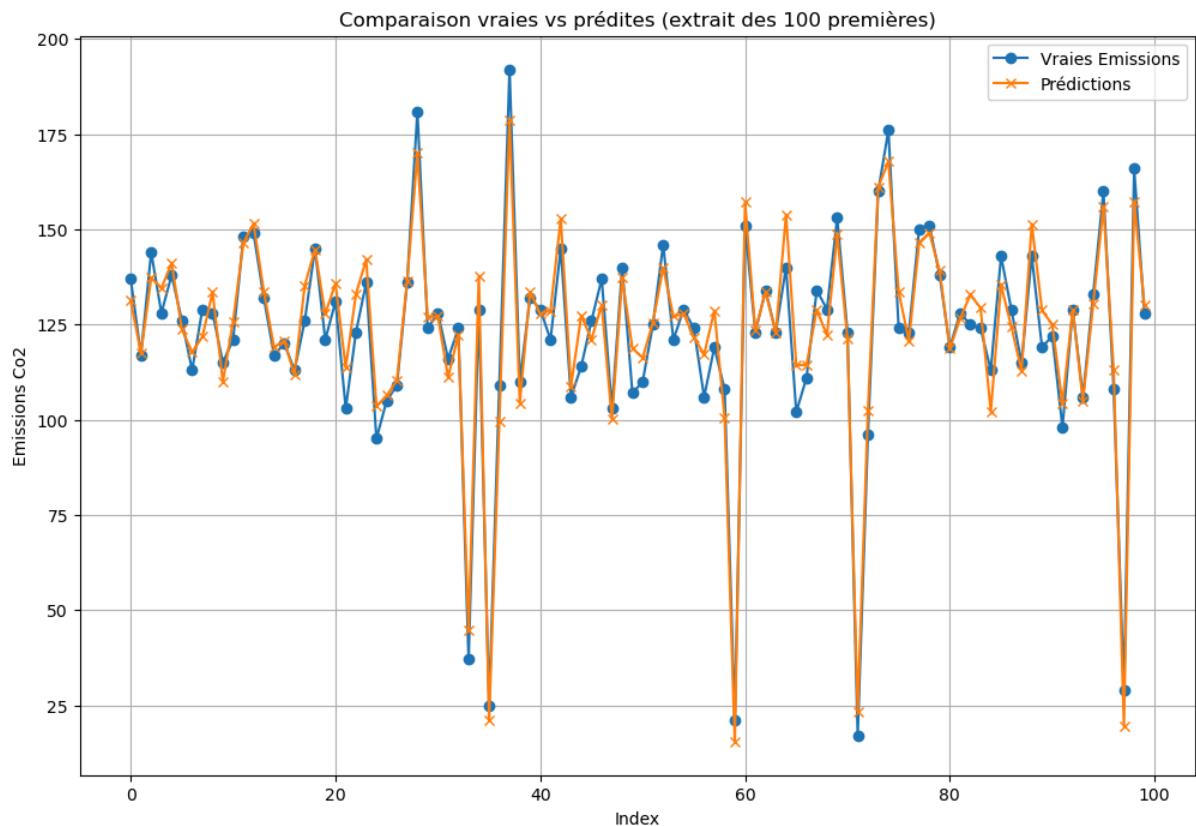
- Étude de l'évolution des performances en fonction de la taille du jeu d'entraînement
- **Observation :**
 - Convergence des scores train et validation
 - Stabilité du modèle sur de grands volumes de données
- Confirme une **bonne généralisation** du modèle



Analyse des prédictions

- Comparaison visuelle :
 - Valeurs réelles vs valeurs prédites
 - Analyse sur un sous-ensemble d'observations

- Les prédictions suivent **globalement bien la tendance** des valeurs réelles
- Bonne capacité du modèle à reproduire les émissions de CO₂
- Les différences résiduelles sont cohérentes avec l'erreur moyenne observée (MSE)



Résultat du modèle de Régression Linéaire :

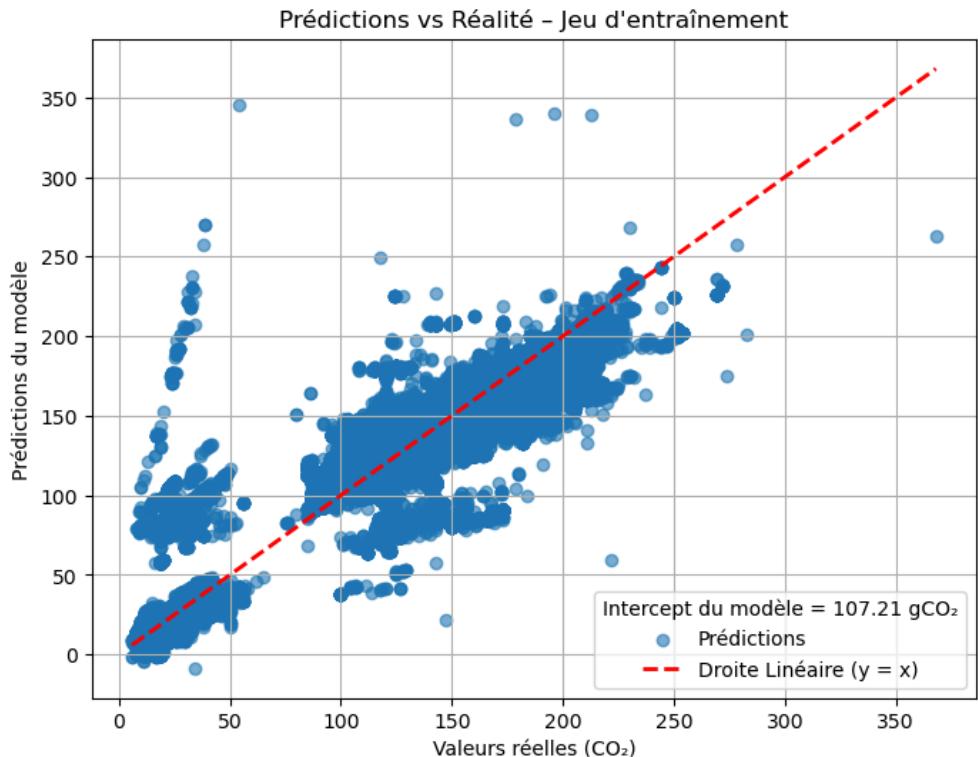
- Graphique prédictions vs réalité :

Le nuage de points montre une forte concordance entre les valeurs réelles et prédictives, confirmant la capacité du modèle à capturer la relation entre les caractéristiques des véhicules et leurs émissions de CO₂.

Le modèle a tendance à :

- **Lisser** les valeurs extrêmes
- Sous-estimer légèrement les très fortes émissions

- Surestimer certaines très faibles émissions
- Comportement classique d'un modèle linéaire



Interprétation

Moyenne des émission de CO₂ du dataset = 110.51

Valeur de l'intercept = 107.21

Dans le cadre d'un modèle utilisant des variables standardisées, l'intercept correspond à la valeur moyenne des émissions de CO₂ prédites par le modèle, c'est-à-dire la prédition associée à un véhicule présentant des caractéristiques moyennes.

Variables et coefficients :

Constructeur	95.32296161950177
Type Carburant	69.96013252349151
Pays	44.778770706932264
Fuel consumption	17.753523981612013
WLTP_poids	4.315332954224411
Puissance_KW	3.278276697864734

La variable « **Constructeur** » présente une importance globale élevée dans le modèle.

Cette importance ne traduit pas un effet direct, mais reflète des différences systématiques entre constructeurs liées à des choix technologiques, des stratégies de conception ou des caractéristiques non directement observées. Le constructeur agit ainsi comme une variable capturant des effets résiduels non expliqués par les variables techniques.

Exemple des différents impacts en fonction des constructeurs :

- différences de **technologie moteur**
- rendement **réel** des moteurs
- stratégies **d'optimisation WLTP**
- boîtes de vitesses
- **aérodynamique**
- calibration moteur

La variable “**Fuel consumption**” présente une importance globale plus modérée car son effet est capturé de manière directe par le modèle à travers un unique coefficient. Contrairement aux variables catégorielles à nombreuses modalités, son influence n'est pas amplifiée par un mécanisme d'agrégation, ce qui explique son classement relatif tout en confirmant son rôle central dans la prédiction des émissions de CO₂.

Conclusion méthodologique

- Le modèle permet une **prédiction fiable des émissions de CO₂**
- Les performances élevées s'expliquent par :
 - La qualité du preprocessing

- La taille du dataset
- Le choix d'un modèle adapté
- Cette approche constitue une **base solide** pour des modèles plus complexes (Random Forest Regressor, Lasso, LassoCV, ElasticNet...)

L'utilisation d'un modèle de régression multivariée permet de prendre en compte simultanément l'ensemble des variables explicatives ainsi que leurs interactions implicites. Les performances obtenues sont élevées et stables, avec des scores R² similaires entre les jeux d'entraînement et de validation croisée. Les learning curves confirment une bonne capacité de généralisation du modèle et l'absence de sur-apprentissage, les performances convergeant rapidement lorsque la taille du jeu d'entraînement augmente.

L'analyse des coefficients montre que le modèle intègre correctement non seulement les variables techniques directes (consommation, poids, puissance), mais également des variables induites telles que le constructeur, qui capturent des effets technologiques ou structurels non explicitement mesurés.

Ainsi, le modèle permet de prédire de manière fiable et cohérente les émissions de CO₂ à partir des caractéristiques d'un véhicule.

7. Annexes

7.1 Code final

- [Preprocessing](#)
- [Etude exploratoire : thermique, électrique, composition du parc automobile](#)
- [Etude exploratoire : hybride, autres motorisations et carte européenne](#)
- [Machine Learning](#)

7.2 Lien Looker Studio

- [Looker Studio](#)

7.3 Lien PowerPoint

- [Présentation PowerPoint](#)