



PREDICTION DES EMISSIONS DE CO2 SELON LES CARACTERISTIQUES DES VEHICULES

Contexte	2
But du projet	2
Choix des données	3
Information sur le jeu de données	4
Donnée cible	4
Description des données.....	4
Exploration des données.....	5
Exploration technique	5
Exploration métier	6
Visualisation des données.....	7
Répartition des véhicules par pays:	7
Répartition des véhicules par type de carburant.....	7
Comparaison des émissions de CO2 par type de carburant	8
Comparaison des émissions de CO2 par fabricant.....	9
Relation entre puissance moteur et émissions de CO2	10
Focus sur des relations entre variables explicatives et cible.....	11
Corrélation entre variables	12
Interprétation des innovations technologiques	16
Manipulation des données	17
Suppression des colonnes non retenues	17
Colonnes non pertinentes.....	17
Colonnes colinéaires	17
Suppression des lignes non retenues.....	18
Sélection de l'énergie concernée par notre étude.....	18
Traitement des valeurs manquantes.....	18
Traitement des outliers	18
Enrichissement des données	25
Créations de nouvelles variables.....	25
Scaling	26

Contexte

Dans le cadre de la formation suivante :

- Métier : Data scientist
- Organisme : DataScientest
- Modalité : Formation continue
- Calendrier : De Septembre 2024 à Juin 2025

Polina Quignon, Vincent Guillemot et Denis Froment réalisent le projet :

PREDICTION DES EMISSIONS DE CO2 SELON LES CARACTERISTIQUES DES VEHICULES.

Nous remercions Eliott Douieb, notre chef de projet DataScientest, pour ses conseils et encouragements au long de cette première phase du projet.

But du projet

Il s'agit de prédire les émissions de CO2 probables de nouveaux véhicules.

Ce projet se basera sur des bases de données de caractéristiques de véhicules.

On peut donc mentionner les cas d'usages suivants à considérer :

- Un nouveau véhicule va être produit (ou n'a pas encore été testé), à quelle émission de CO2 peut-on s'attendre pour ce véhicule ?
- Un constructeur veut créer un véhicule ayant une caractéristique donnée, quels paramètres/contraintes doit-il viser pour émettre le moins de CO2 possible ?

Choix des données

L'émission de CO2 suscite l'intérêt public, notamment en vue de l'action pour le climat.

Aussi, plusieurs sources de données incluant des caractéristiques de véhicules et des résultats de tests d'émission de CO2 sont trouvables sur internet.

Notre choix s'est porté sur la base :

EEA geospatial data catalog, Monitoring of CO2 emissions from passenger cars

Raisons de notre choix :

- Cette base contient la donnée cible que nous cherchions
- Cette base possède de nombreux enregistrements aussi bien en nombre de véhicules qu'en nombre de caractéristiques par véhicule
- Cette base est actuelle puisqu'il s'agit de l'exhaustivité des ventes annuelles de véhicules en Europe, publiée peu avant le début de notre projet.

Une alternative intéressante en termes de caractéristiques exposées était disponible sur data.gouv.fr mais moins actuelle et moins riche.

Source des données : <http://co2cars.apps.eea.europa.eu/>

Dictionnaire de données : <https://sdi.eea.europa.eu/catalogue/srv/api/records/87fd2bce-6ad5-46d8-af41-f27cf2e45a8/attachments/Table-definition.xlsx>

L'exploration des données les plus récentes (2023), présenté dans les chapitres suivants, nous a permis de confirmer que le panel en termes de marques, modèles, et type d'énergie était suffisamment riche pour permettre une bonne exploitation.

La comparaison avec les données 2022 nous montre en outre que des données quantitatives (distances entre roues) ont été omises en 2023 mais étaient présentes en 2022. Aussi, une comparaison des données manquantes nous indique que les données de 2022 sont plus complètes :

N° colonne	Nom	Nom détaillé	Taux de NA br. 2023	Taux de NA br. 2022
1	ID	ID	0.0%	0.0%
2	Country	Country	0.0%	0.0%
3	VFN	Vehicle family identification number	1.2%	1.3%
4	Mp	Pool	8.9%	6.5%
5	Mh	Manufacturer name (EU standard)	0.0%	0.0%
6	Man	Manufacturer name (OEM declaration)	0.0%	0.0%
7	MMS	Manufacturer name (MS registry denomination)	100.0%	100.0%
8	Tan	Type approval number	0.3%	0.2%
9	T	Type	0.1%	0.0%
10	Va	Variant	0.3%	0.2%
11	Ve	Version	0.4%	0.4%
12	Mk	Make	0.0%	0.0%
13	Cn	Commercial name	0.0%	0.9%
14	Ot	Category of the vehicle type approved	0.1%	0.2%
15	Cr	Category of the vehicle registered	0.0%	0.0%
16	r	Total new registrations	0.0%	0.0%
17	m (kg)	Mass in running order (kg)	0.0%	0.0%
18	Mt	WLTP test mass	1.5%	1.8%
19	Enedc (g/km)	Specific CO2 Emissions g/Km (NEDC)	100.0%	82.9%
20	Ewltip (g/km)	Specific CO2 Emissons in g/km (WLTP)	0.1%	0.1%
21	W (mm)	Wheel base in mm	100.0%	0.4%
22	A1t1 (mm)	Axle width steering axle in mm	100.0%	1.6%
23	A1t2 (mm)	Axle width other axle in mm	100.0%	1.9%
24	Ft	Fuel type	0.0%	0.0%
25	Fm	Fuel mode	0.0%	0.0%
26	ec (cm3)	Engine capacity in cm3	15.6%	13.5%
27	ep (kW)	Engine power in kW	0.5%	1.4%
28	z (Wh/km)	Electric energy consumption in Wh/km	77.3%	77.4%
29	It	Innovative technology	34.9%	32.7%
30	Emedc (g/km)	Emissions reduction through innovative technologies in g/km	100.0%	100.0%
31	Envltip (g/km)	Emissions reduction through innovative technologies in g/km (35.3%	34.5%
32	De	Deviation factor	100.0%	100.0%
33	Vf	Verification factor	100.0%	100.0%
34	Status	P = Provisional data, F = Final data.	0.0%	0.0%
35	year	Registration year	0.0%	0.0%
36	Date of registration	Date of registration	0.0%	1.7%
37	Fuel consumption	Fuel consumption	17.6%	14.7%
38	ech		50.4%	100.0%
39	RLFI	Roadload (Matrix) family's identifier	71.3%	100.0%
40	Electric range (km)		77.4%	79.7%

Nous choisissons donc les données de 2022 comme base du projet.

Le dictionnaire de données étant le même que pour 2023, une application pour cette année-là serait également possible plus tard.

figure: Comparaisons des données manquantes 2022 vs 2023

Information sur le jeu de données

Donnée cible

Nous devons prédire l'émission de CO2 en g/km des véhicules.

Nous identifions dans le jeu de données la variable cible comme étant :

Ewlt (g/km)

en effet, il s'agit du résultat de la procédure d'essai harmonisée mondiale pour les véhicules légers, connue sous le nom de WLTP (Worldwide harmonized Light vehicles Test Procedure).

Description des données

Nous avons enrichi une version locale du dictionnaire des données, avec des éléments qualifiant leur contenu (format, taux de remplissage, modalités, explications métier, appréciation de la pertinence pour prédiction d'émission CO2).

Ce tableau est partagé pour les besoins de notre équipe sur un drive sharepoint.

Voici un extrait visuel :

colonne	Nom	Nom détaillé	Description	Disponibilité de la	Type informatique	Taux NA des NA	Gesti des NA	Distribution des valeurs / Modalités	Remarques sur la colonne	Remarques colonne Denis
6 Man	Manufacturer name (OEM declaration)	Manufacturer name (OEM declaration)	object	0%	'VOLKSWAGEN AG' 'STELLANTIS EUROPE SPA' 'BAYERISCHE MOTOREN WERKE AG' 'SKODA AUTO AS' 'TESLA INC' 'MERCEDES-BENZ AG' 'MAZDA MOT					
7 MMS	Manufacturer name (IMS registry denomination)	Manufacturer name (IMS registry denomination)	float64	100%						
8 Tan	Type approval number	Type approval number	object	0%	'E13*2007/46*1845*26' 'E3*2007/46*0373*33' 'E13*2007/46*1845*27' 'E1*2007/46*2063*05' 'E1*2018/858*00004*12' 'E1*2007/46*1682*15'					
9 T	Type	Type	object	0%	'A1' '356' 'EMUL2E' 'E2' 'FMX' '3T' '005' 'MEK' 'NU' 'D1'					
10 Va	Variant	Variant	object	0%	'DXBXW0AC4' 'HXS12' 'DLAAZ0AE2' '11D1' '4ACK1EBL1GX1' '81B1R' 'ACDGBBX01' 'Y7CR' 'MKC' 'ACDXDBX0'					
11 Ve	Version	Version	object	0%	'F0MF06C90154B01CANN012GA0' '0AW40000' 'DAE1G120101ASA' 'IAW500BT' 'INF6FD6DD0014B1STI					
12 Mk	Make	Make	object	0%	'VOLKSWAGEN VW' 'FIAT' 'SKODA' 'TESLA' 'MERCEDES-BENZ' 'MAZDA' 'CUPRA' 'VOLVO' 'KIA'					
13 Cn	Commercial name	Commercial name	object	0%	'T-ROC' 'FIAT TIPO' 'COOPER SE' 'ID4 GTX 220 KW' 'COUNTRYMAN COOPER S ALL4' 'SUPERB' 'MODEL Y' 'CITAN TOURER/T-CLASS' 'KAROO' 'MAZDA2'					
14 Ct	Category of the vehicle type approved	Category of the vehicle type approved	object	0.3%	'M1' 'M1G' 'N2' 'N1G' 'N1' ''N2G'					
15 Cr	Category of the vehicle registered	Category of the vehicle registered	object	0%						
16 r	Total new registrations	Total new registrations	int64	0%						
17 m (kg)	Mass in running order (kg)	Mass in running order (kg)	float64	0%						
18 Mt	WLTP test mass	WLTP test mass	float64	2%						
19 Enedc (g/km)	Specific CO2 Emissions g/Km (NEDC)	Specific CO2 Emissions g/Km (NEDC)	float64	100%						
20 Ewlt (g/km)	Specific CO2 Emissions in g/km (WLTP)	Specific CO2 Emissions in g/km (WLTP)	float64	0%					vide	cible
21 W (mm)	Wheel base in mm	Wheel base in mm	float64	100%						
22 A1t (mm)	Axle width steering axle in mm	Axle width steering axle in mm	float64	100%						
23 A2t (mm)	Axle width other axle in mm	Axle width other axle in mm	float64	100%						
24 Ft	Fuel type	Fuel type	object	0%	'petrol' 'electric' 'petrol/electric' 'diesel' 'diesel/electric' 'lpg' 'ng' 'hydrogen' 'e85'					
25 Fm	Fuel mode	Fuel mode	object	0%	M' 'H' 'E' 'P' 'B' 'F' (Mono-fuel, H:off-charging hybrid, Electric-pure, Plug-in-hybrid, Bi-fuel, Flex-fuel)					
26 ec (cm3)	Engine capacity in cm3	Engine capacity in cm3	float64	16%						
27 ep (kW)	Engine power in kW	Engine power in kW	float64	0%						
28 z (Wh/km)	Electric energy consumption in Wh/km	Electric energy consumption in Wh/km	float64	77%	numérique					
29 IT	Innovative technology	Innovative technology	object	35%	'e13 29' 'e3 32' '' e24 29 37' 'e9 29' 'e8 29 37' 'e13 37' 'e2 29 37' 'e9 32 37' 'e1 29'					intéressant. Attention 11 véhi
30 Ernedc (g/km)	Emissions reduction through innovative technologies in g/km	Emissions reduction through innovative technologies in g/km	float64	100%						
31 Erwltp (g/km)	Emissions reduction through innovative technologies in g/km (WLTP)	Emissions reduction through innovative technologies in g/km (WLTP)	float64	35%						
32 De	Deviation factor	Deviation factor	float64	100%						
33 Vf	Verification factor	Verification factor	float64	100%						
34 Status	P = Provisional data, F = Final data.	P = Provisional data, F = Final data.	object	0%	'F'					
35 year	Registration year	Registration year	int64	0%	toujours 2023 pour ce fichier :)					
36 Date of registration	Date of registration	Date of registration	object	0%	'2023-03-14' '2023-01-27' '2023-05-15' '2023-11-10' '2023-08-10' '2023-12-13' '2023-03-22' '2023-06-02' '2023-11-06' '2023-04-25'					long seule modalité : "2023"

figure: Extrait du dictionnaire des données enrichi d'informations par l'équipe

Exploration des données

Exploration technique

- Plusieurs colonnes sont vides ou bien n'ont qu'une unique valeur sur toute la base.
- Nous avons coloré en gris les variables toujours vides et en marron celles qui n'ont qu'une valeur dans le tableau ci-dessous :

N° colonne	Nom	Nom détaillé	Anomalies	Type informatique	Taux NA
1 ID	ID			int64	0%
2 Country	Country			object	0%
3 VFN	Vehicle family identification number			object	1%
4 Mp	Pool			object	9%
5 Mh	Manufacturer name (EU standard)			object	0%
6 Man	Manufacturer name (OEM declaration)			object	0%
7 MMS	Manufacturer name (MS registry denomination)	informations déjà dans 'Man'		float64	100%
8 Tan	Type approval number			object	0%
9 T	Type			object	0%
10 Va	Variant			object	0%
11 Ve	Version			object	0%
12 Mk	Make			object	0%
13 Cn	Commercial name			object	0%
14 Ct	Category of the vehicle type approved			object	0.1%
15 Cr	Category of the vehicle registered			object	0%
16 r	Total new registrations	toujours "1"		int64	0%
17 m (kg)	Mass in running order (kg)			float64	0%
18 Mt	WLTP test mass			float64	2%
19 Enedc (g/km)	Specific CO ₂ Emissions g/Km (NEDC)	norme obsolète		float64	100%
20 Ewltp (g/km)	Specific CO ₂ Emissions in g/km (WLTP)			float64	0%
21 W (mm)	Wheel base in mm			float64	100%
22 At1 (mm)	Axle width steering axle in mm			float64	100%
23 At2 (mm)	Axle width other axle in mm			float64	100%
24 Ft	Fuel type			object	0%
25 Fm	Fuel mode			object	0%
26 ec (cm ³)	Engine capacity in cm ³			float64	16%
27 ep (KW)	Engine power in KW			float64	0%
28 z (Wh/km)	Electric energy consumption in Wh/km			float64	77%
29 IT	Innovative technology			object	35%
30 Ernedc (g/km)	Emissions reduction through innovative technologies in g/km			float64	100%
31 Erwltp (g/km)	Emissions reduction through innovative technologies in g/km (WLTP)			float64	35%
32 De	Deviation factor			float64	100%
33 Vf	Verification factor			float64	100%
34 Status	P = Provisional data, F = Final data.	toujours "F"		object	0%
35 year	Registration year	toujours "2023"		int64	0%
36 Date of registration	Date of registration			object	0%
37 Fuel consumption	Fuel consumption			float64	18%
38 ech				object	50%
39 RLFI	Roadload (Matrix) family's identifier			object	71%
40 Electric range (km)				float64	77%

Figure: variables mono-modales et vides mises en évidence

Conclusion : Ces variables vides et mono-modales ne sont pas exploitables, nous les supprimerons avant d'exploiter les données.

- Variables avec valeurs manquantes

Dans la colonne "taux de NA" de notre tableau, on voit que certaines valeurs sont manquantes, ceci peut très bien être justifié par le métier. Exemple : pas d'autonomie électrique pour un véhicule pur essence.

Nous tiendrons compte dans l'analyse. A priori inutile d'essayer de remplir ces variables.

Exploration métier

- Doublons : Il y a autant de lignes que de ventes de véhicules en Europe. On a donc le même modèle exact de véhicule qui figure n fois dans la base.

=> Cela permet de voir l'impact total des ventes sur l'émission de CO2 des véhicules.

Nous nommerons cet ensemble la “base des ventes” pour les visualisations.

=> Cela sera inutile dans l'analyse des caractéristiques influant sur l'émission de CO2, on gardera donc 1 seul enregistrement par modèle.

Nous nommerons cet ensemble la “base des modèles” pour les visualisations.

- ID : Cette colonne n'apporte pas d'information métier et devra aussi être supprimée pour ne pas influencer les modèles de prédiction.
- Explication des valeurs : nous complétons la colonne “Description” avec des recherches métier pour mieux comprendre les valeurs.

N° colonne	Nom	Nom détaillé	Description
1 ID	ID		numéro d'identification
2 Country	Country		Pays
3 VFN	Vehicle family identification number		Identifiant donné par le constructeur
4 Mp	Pool		Groupe du fabricant
5 Mh	Manufacturer name (EU standard)		
6 Man	Manufacturer name (OEM declaration)		
7 MMS	Manufacturer name (MS registry denomination)		
8 Tan	Type approval number		Identifiant donné par le constructeur
9 T	Type		Information donnée par le constructeur
10 Va	Variant		Information donnée par le constructeur
11 Ve	Version		Information donnée par le constructeur
12 Mk	Make		Marque
13 Cn	Commercial name		Nom commercial du modèle
14 Ct	Category of the vehicle type approved		pour transports de matière dangereuse
15 Cr	Category of the vehicle registered		M1=transport de personnes, N1=marchandises
16 r	Total new registrations		
17 m (kg)	Mass in running order (kg)		masse pouvant rouler
18 Mt	WLTP test mass		masse pour test normé
19 Enedc (g/km)	Specific CO2 Emissions g/Km (NEDC)		
20 Ewltp (g/km)	Specific CO2 Emissions in g/km (WLTP)		norme officielle
21 W (mm)	Wheel base in mm		intercentres roues AR et AV
22 At1 (mm)	Axle width steering axle in mm		distance entre roues AV
23 At2 (mm)	Axle width other axle in mm		distance entre roues AR
24 Ft	Fuel type		type de carburant
25 Fm	Fuel mode		monofuel, hybrid, plug-in...
26 ec (cm3)	Engine capacity in cm3		cylindrée
27 ep (KW)	Engine power in KW		puissance
28 z (Wh/km)	Electric energy consumption in Wh/km		consommation électrique
29 IT	Innovative technology		code d'innovation utilisée
30 Ernedc (g/km)	Emissions reduction through innovative technologies		norme obsolète
31 Erwltp (g/km)	Emissions reduction through innovative technologies		g de CO2/km épargné grâce à cette innovation
32 De	Deviation factor		
33 Vf	Verification factor		
34 Status	P = Provisional data, F = Final data.		
35 year	Registration year		
36 Date of registration	Date of registration		
37 Fuel consumption	Fuel consumption		consommation de carburant
38 ech			Norme euros d'émissions de polluants
39 RLFI	Roadload (Matrix) family's identifier		~ charge routière
40 Electric range (km)			autonomie électrique

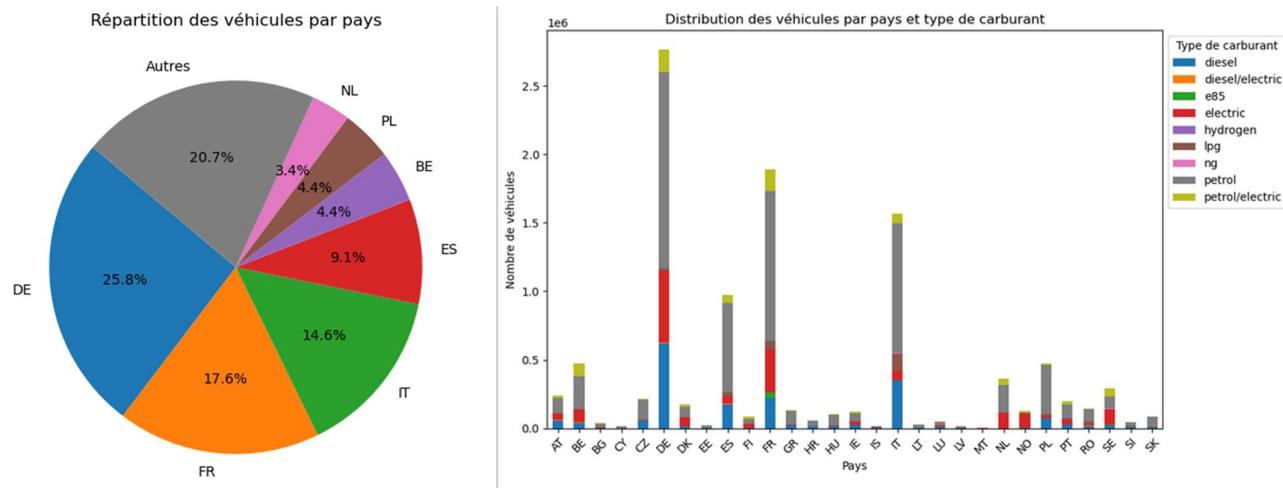
Visualisation des données

Il s'agit de plus de 10 millions de lignes, il est donc nécessaire de prendre la mesure des informations contenues.

Lors de la phase d'exploration des données, nous convenons d'étudier l'ensemble des données disponibles. Lors des phases ultérieures, nous éliminerons probablement des catégories moins pertinentes, par exemple des moteurs diesels que les constructeurs produiront de moins en moins.

Voici les visualisations notables qui permettent de comprendre le contenu :

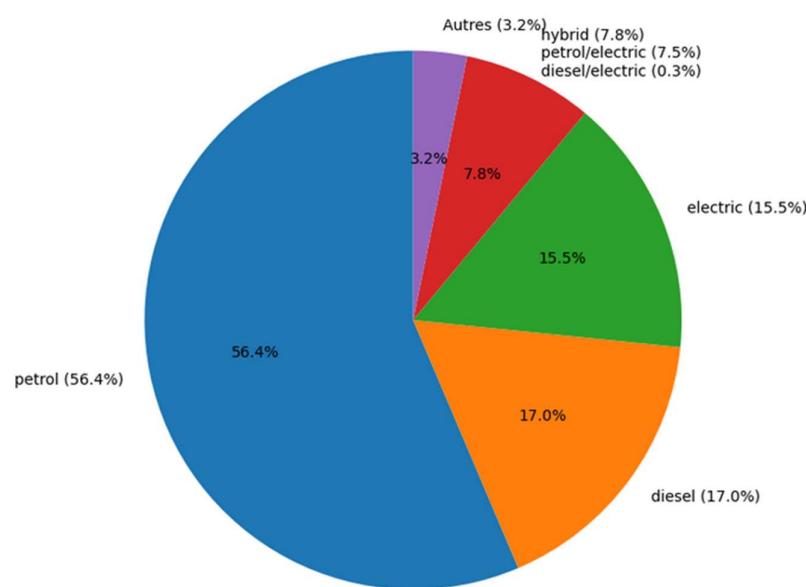
Répartition des véhicules par pays :



Source: base des ventes

- On voit que les véhicules sont répartis sur beaucoup de pays, donc c'est exhaustif et bien réparti selon les populations des pays.
- La répartition des types de carburant montre que le type "essence" est le plus représenté. Cela nous intéresse car ce type de véhicule à forte émission de CO₂ sera encore fabriqué dans les années à venir.

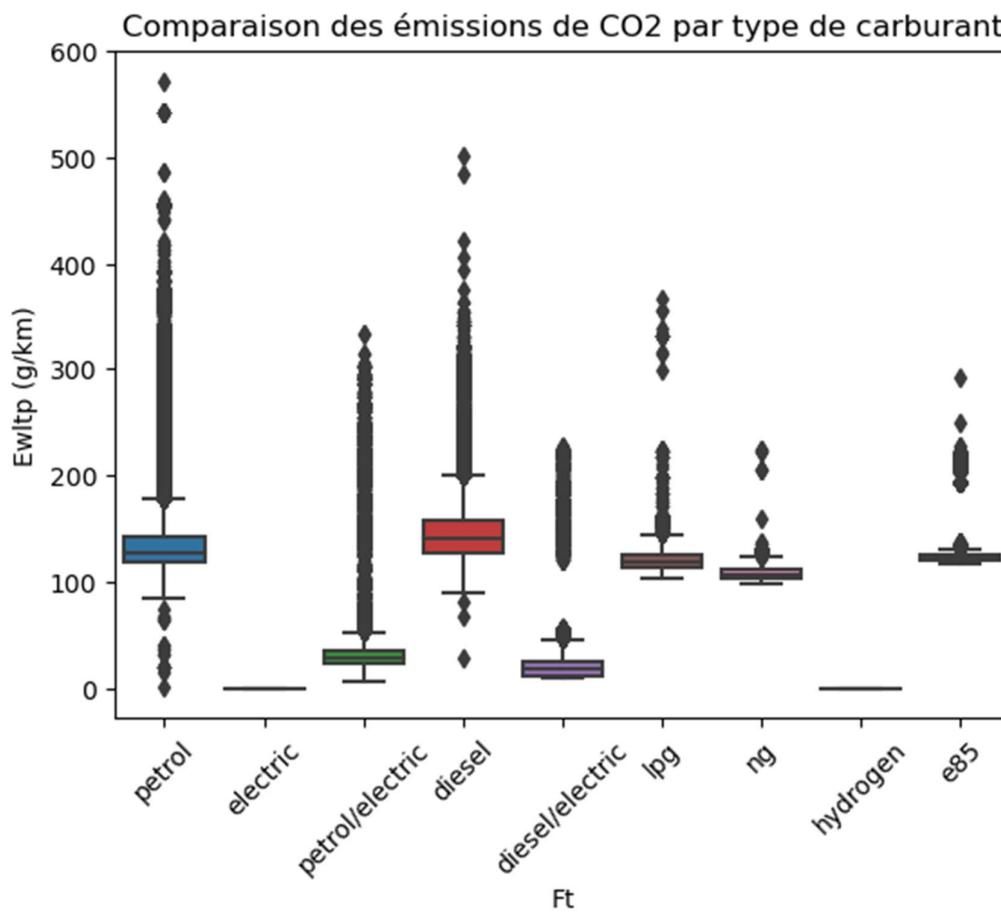
Répartition des véhicules par type de carburant



Source: base des ventes

- On voit ici la répartition qui confirme la prépondérance de l'essence tous pays confondus.

Comparaison des émissions de CO2 par type de carburant



Source: base des ventes

Objectif du boxplot : comparer la répartition des émissions de CO2 pour différents types de carburants ou motorisations et détecter les outliers.

Variation des émissions entre les types de carburant :

- Les véhicules à essence (“petrol”) et diesel ont des émissions relativement élevées
- Les véhicules électriques n'émettent pas de CO2
- Les motorisations hybrides (petrol/electric, diesel/electric) ont des émissions plus faibles en moyenne que leurs équivalents non hybrides (petrol et diesel)
- Les carburants lpg (gaz de pétrole liquéfié), ng (gaz naturel) et e85 (bioéthanol) montrent une réduction des émissions par rapport aux carburants traditionnels, mais restent au-dessus des véhicules hybrides

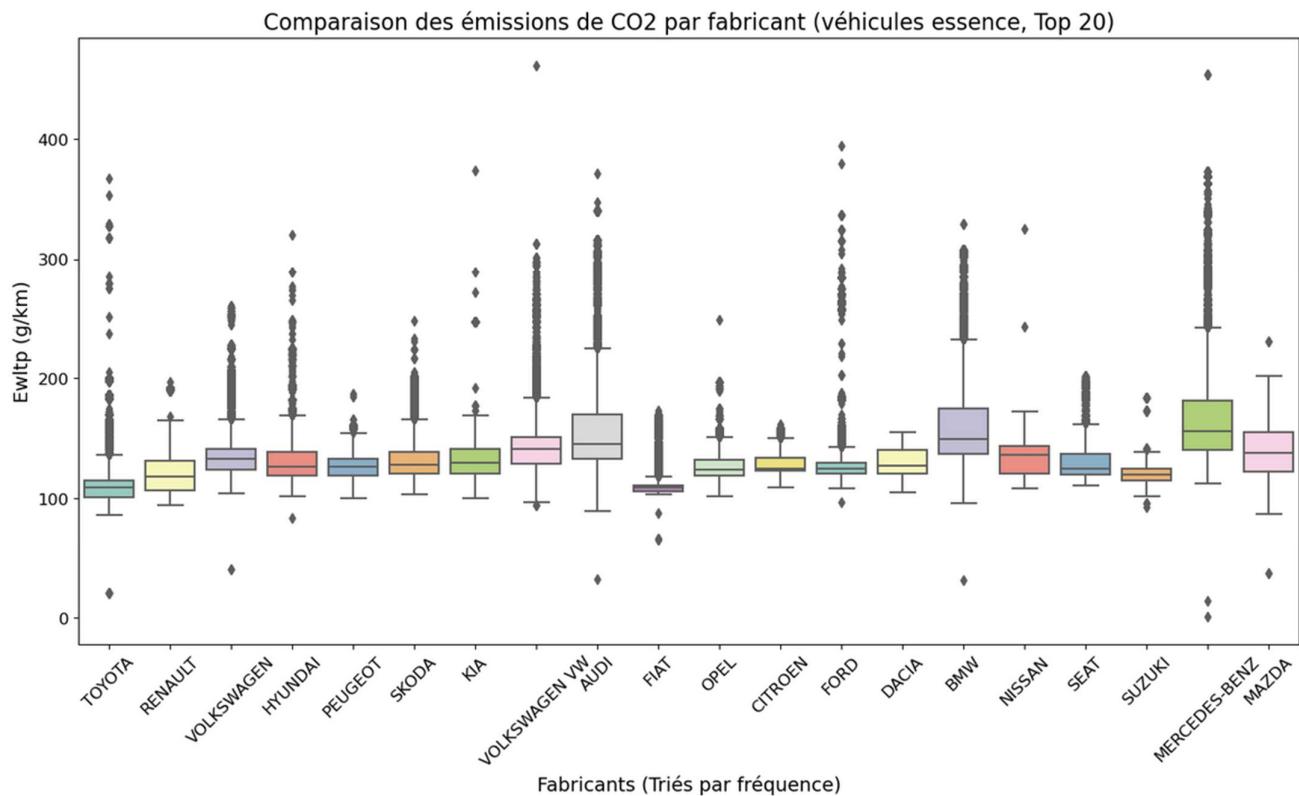
Étendue des émissions :

- Pour petrol et diesel, les distributions montrent une large plage d'émissions, avec des valeurs maximales dépassant 500 g/km pour certains modèles

Outliers :

- Beaucoup d'outliers sont visibles pour les carburants petrol et diesel, indiquant la présence de véhicules particulièrement polluants.
- On voit aussi des aberrations, par exemple des essence (petrol) à émission zéro, il faudra s'occuper de cette donnée.
- Un groupe de véhicules hybride diesel+electrique semble se différencier de la masse puisqu'il y a un groupe de points bien au-delà du 3ème quartile.

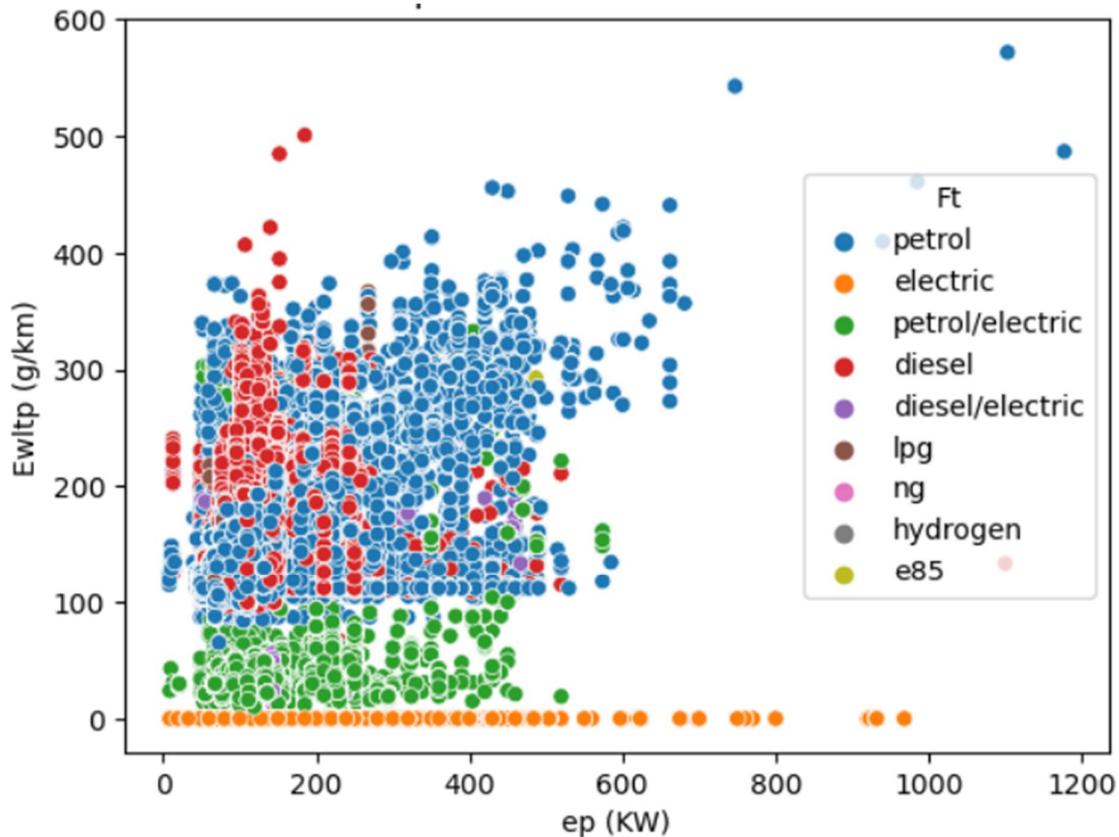
Comparaison des émissions de CO2 par fabricant



Source: base des ventes

- Nous avons une grande représentativité de fabricants.
- Certains fabricants sont spécialistes de grosses émissions (VW, BMW, Mercedes), et certains sont plus représentés dans les faibles émissions (Suzuki, Toyota, Fiat). Ces fabricants ne ciblent probablement pas la même catégorie d'acheteurs ou bien privilégient des énergies différentes (Toyota spécialiste hybride).
- Les aberrations repérées lors des émissions de CO2 par type de carburant sont bien entendu présentes sur ce graphe aussi.

Relation entre puissance moteur et émissions de CO2



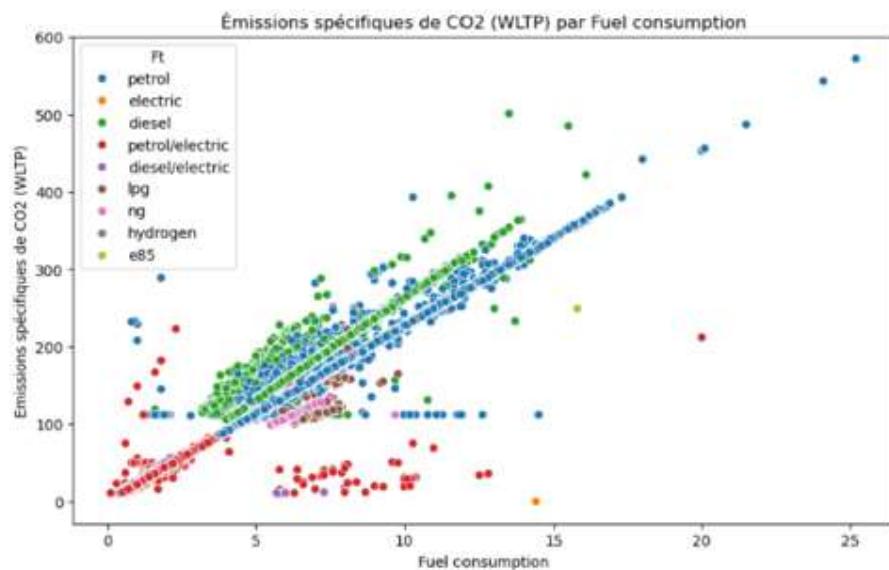
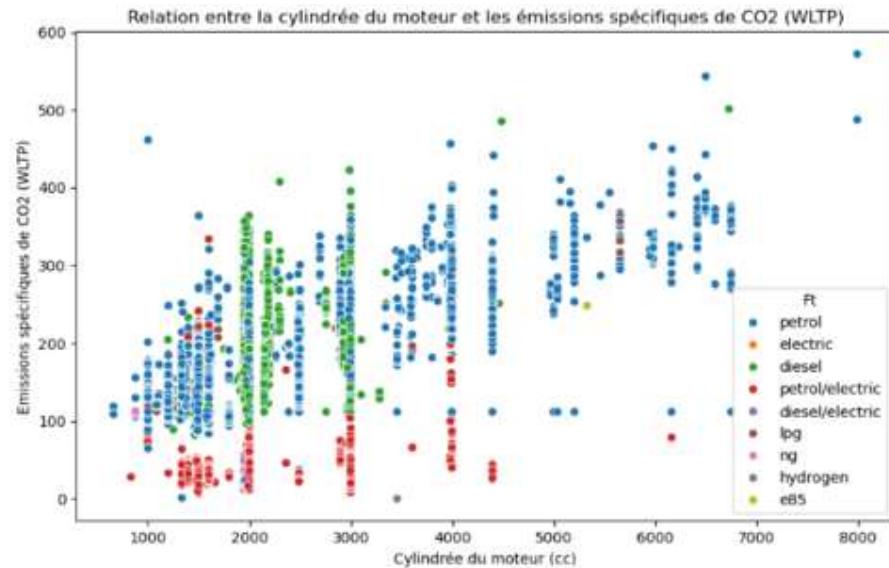
Source: base des ventes

Ce nuage de points (scatter plot) représente les émissions de CO2 (Ewltp) en fonction de la puissance moteur (ep). Chaque point est coloré selon le type de carburant ou motorisation (Ft).

- On y voit des répartitions différentes caractéristiques aux énergies utilisées. Sans surprise, les électriques sont répartis sur la ligne du zéro.
- Il est à noter que les répartitions se recouvrent, certains véhicules essence pur émettent donc plus de CO2 au kilomètre que certains véhicules hybrides par exemple.

Nota: ce scatterplot est bien entendu identique si l'on prend la base des modèles.

Focus sur des relations entre variables explicatives et cible

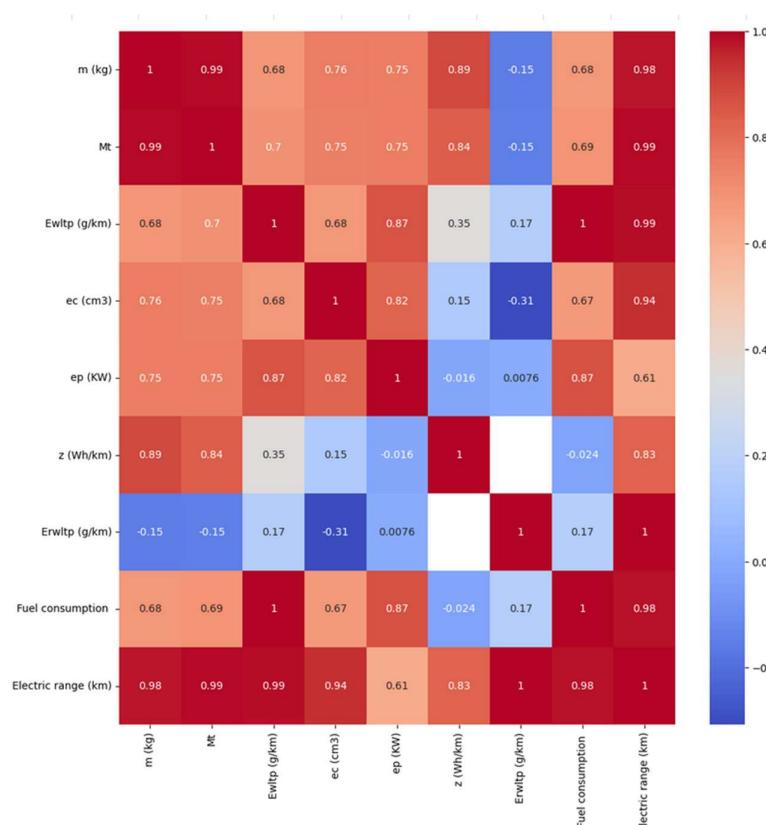


Des tendances sont nettes, et particulièrement deux tendances sont linéaires entre consommation de carburant et émissions de CO₂, caractéristiques respectivement des modèles diesel et essence.

Corrélation entre variables

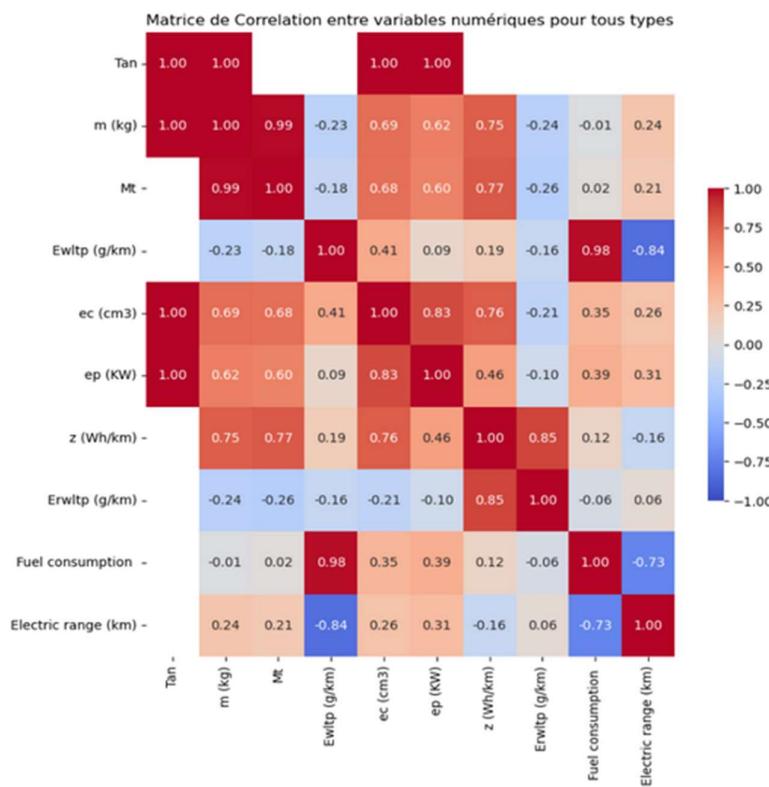
Nous nous intéressons aux corrélations entre les variables numériques.

Nous regardons d'abord **tous véhicules confondus** sur la base des ventes :



Source: base des ventes

Puis regardons sur la base des modèles :



Source: base des modèles

Nous concluons que les corrélations sur la base des ventes sont biaisées, certains véhicules étant présents des milliers de fois plus que d'autres.

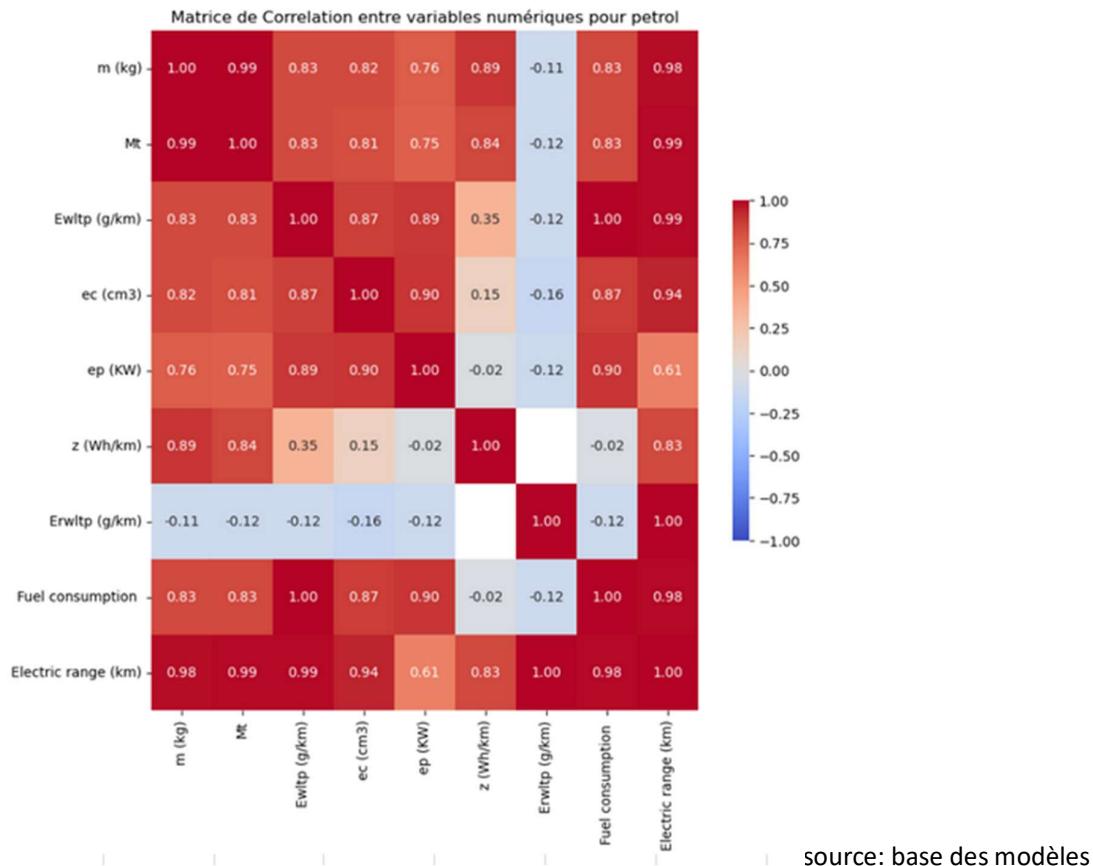
La base des modèles sera notre référence pour établir les relations entre caractéristiques des véhicules et variable cible.

Dans ce dernier tableau, regardons plus précisément les corrélations entre variables quantitatives explicatives et notre variable cible "Ewltp (g/km)" :

Toutes énergies confondues, on note que l'émission de CO2 est :

- Corrélée à
 - La consommation de carburant (Fuel consumption),
 - La cylindrée (ec...),
 - La consommation électrique (z) => probablement pour les hybrides
- Inversement corrélée à
 - La masse (Mt ou m) => ceci est un point surprenant au 1er abord. Mais une différenciation par type d'énergie s'impose.
 - L'autonomie électrique => plus on peut rouler en électrique moins on émet (pour hybrides), c'est discutable avec le poids des batteries, mais probablement vrai pour une plug-in hybride rechargeée.

Focus sur essence uniquement



source: base des modèles

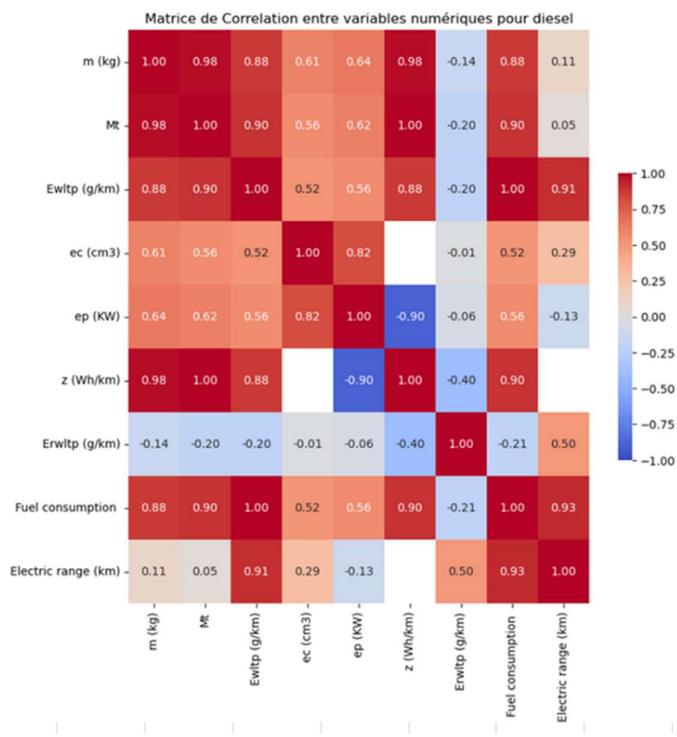
Pour l'émission de CO₂ : Les rapports à la cylindrée, puissance, consommation sont + nets encore.

La théorie de la masse semble confirmer que tous véhicules confondus, c'est trompeur, les hybrides sont + lourds et émettent moins. Il faut donc bien considérer les corrélations entre valeurs numériques en différenciant par type d'énergie.

On note ici une anomalie dans les données : en type essence pure, certains constructeurs ont donné des valeurs de consommation électrique. Cela concerne 11 véhicules seulement, nous devront décider lors de la phase de feature engineering que faire avec cette variable minoritairement représentée.

On note une forte corrélation entre la puissance moteur (ep) et l'émission de CO₂, sans surprise.

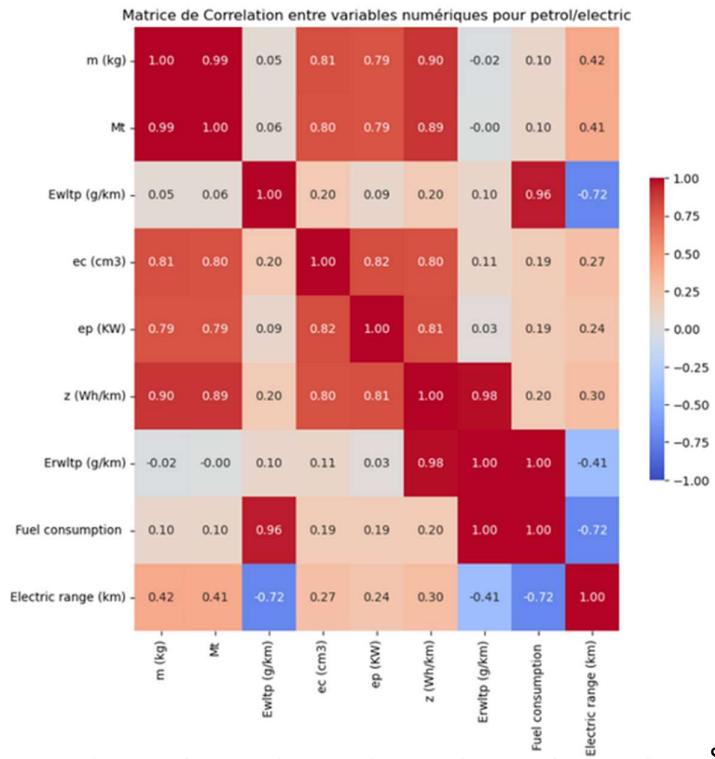
Focus sur diesel uniquement



Source: base des modèles

Pour l'émission de CO₂ : Les rapports à la cylindrée, puissance, masse sont de même type que pour essence, mais avec des nuances sensibles dans les valeurs.

Focus sur hybride essence+électrique



Source: base des modèles

Pour l'émission de CO₂ : La cylindrée est fortement corrélée. La masse est totalement décorrélée.

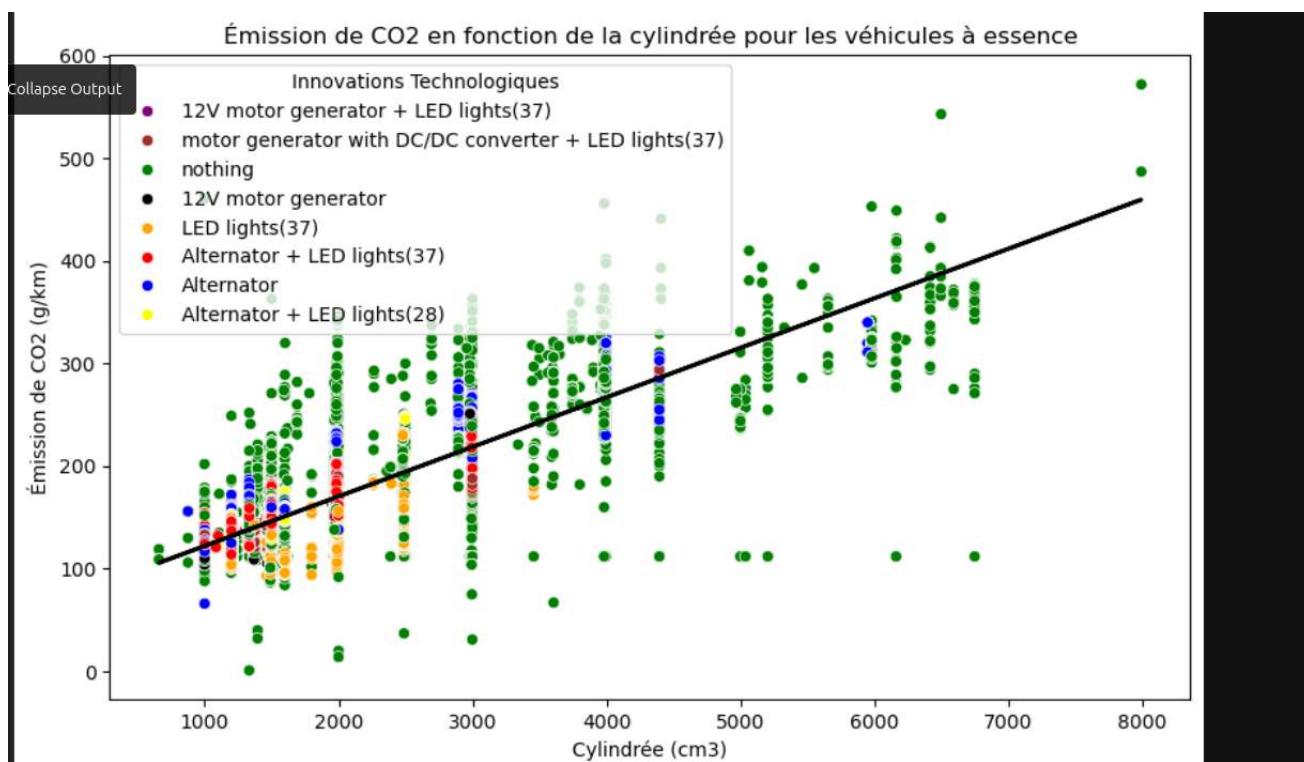
Interprétation des innovations technologiques

Dans la colonne « IT » (Innovative technology), on trouve des modalités de type : « e1 37 29 e14 28 », avec :

- $e<x>$: le code d'une autorité locale (pays x) ayant certifié l'innovation embarquée dans le véhicule, à savoir :
e13' = Luxembourg
e3' = Italy
e24' = Ireland
e9' = Spain
e8' = Czech Republic
e13' = Luxembourg again
e2' = France
e1' = Germany
- xy : le code d'une innovation technologique, à savoir :
28 LED lights
29 Alternator
32 motor generator with DC/DC converter
33 12V motor generator
35 LED lights
37 LED lights
38 Smart diesel fuel heater

(Source : <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A32013R0195>)

Une étude de la distribution des émissions de CO2 en fonction de la cylindrée avec différenciation selon les innovations technologiques utilisées donne :



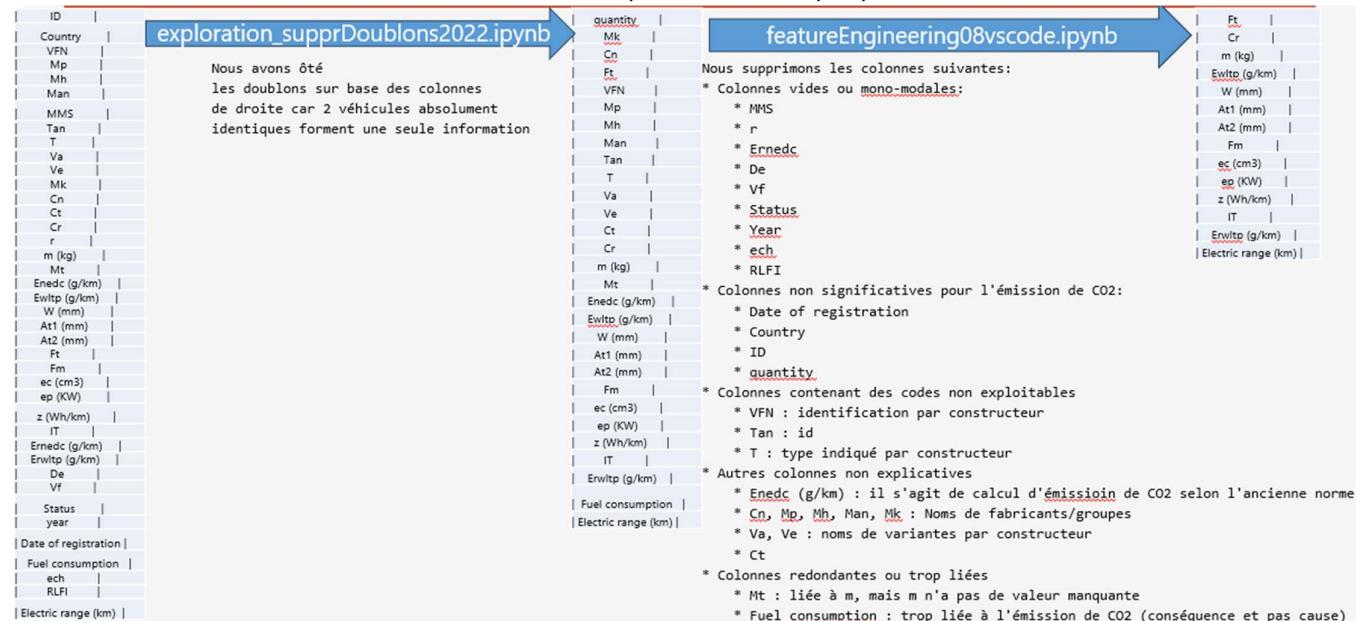
Ce n'est pas extraordinaire, mais on peut penser que les innovations LED light(37), et combinaison motor generator with DC/DC converter+LED lights se trouvent souvent sous la ligne de régression. 12V motor generator aussi. Attention, on a tracé ici une régression linéaire. Or c'est peut-être pas linéaire, et on n'a pas filtré les outliers avant ;)
Méfions-nous aussi: dans la littérature que j'ai pu trouver, ils semblent dire que l'alternateur c'est très impactant pour émission de CO2...

Manipulation des données

Suppression des colonnes non retenues

Colonnes non pertinentes

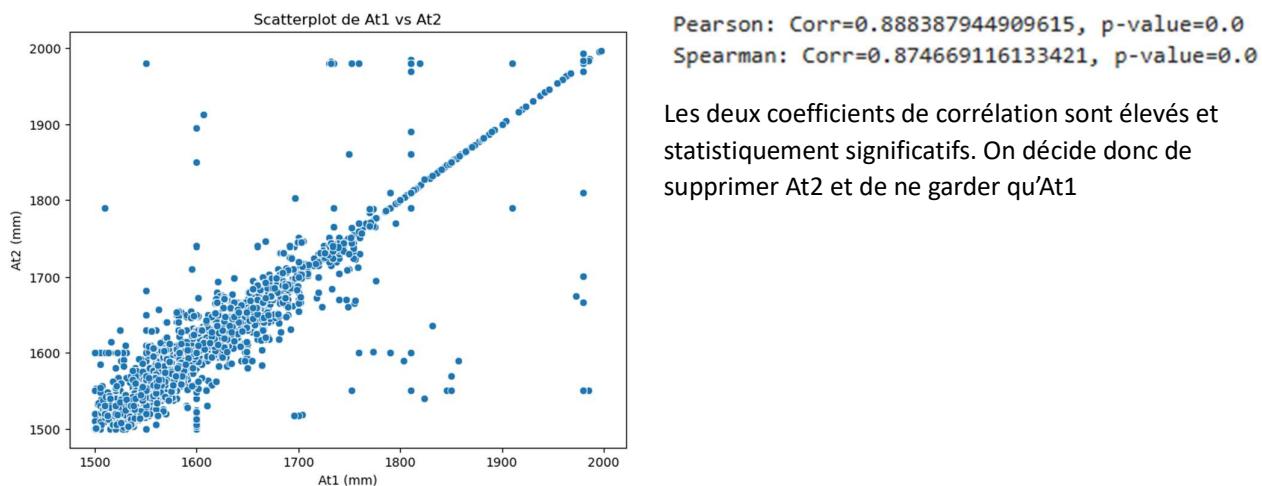
Pour notre but d'étudier l'émission de CO₂, nous avons procédé en 2 étapes pour sélectionner les colonnes :



Colonnes colinéaires

Focus sur At1 et At2: en ayant fait un pairplot on s'est aperçu de la relation très linéaire entre les deux variables.

Le graphique ci-dessous l'illustre bien mais nous avons aussi voulu le confirmer statistiquement avec des tests de corrélation dont les résultats sont également précisés ci-après



Suppression des lignes non retenues

Sélection de l'énergie concernée par notre étude

- Nous avons décidé de faire porter l'étude sur les énergies ['petrol', 'diesel', 'petrol/electric', 'diesel/electric'], car les autres types d'énergie sont soit à émission de CO2 nulle, soit des marchés de niche.
Nous supprimons donc toutes les autres lignes.

Traitement des valeurs manquantes

Valeurs d'essieu ['W (mm)', 'At1 (mm)', 'At2 (mm)']

- Nous supprimons les lignes pour lesquelles il manque au moins une valeur. Cela ne représente que 3% de l'ensemble, et ces valeurs nous semblent importantes pour notre sujet.

Cylindrées ['ec (cm3)']

- Nous supprimons 25 lignes dont la cylindrée est manquante

Autonomie électrique [' Electric range (km)']

- Nous supprimons 17 000 lignes pour des hybrides ayant ici des valeurs manquantes. En effet nous pensons que ces valeurs sont précieuses pour les futurs calculs des modèles. Et nous ne pouvons pas remplir ces 17 000 valeurs.

Traitement des outliers

• Masse ['m (kg)']

- à l'examen des valeurs extrêmes, nous corrigons le poids de certains modèles qui présentent $m > 3900 \text{ Kg}$
- Le reste (notamment les valeurs les plus basses) ne choque pas au vu des modèles concerné

• Émission de CO2 ['Ewltp (g/km)']

- L'émission de CO2 étant différente par type d'énergie, on examine les outliers par énergie

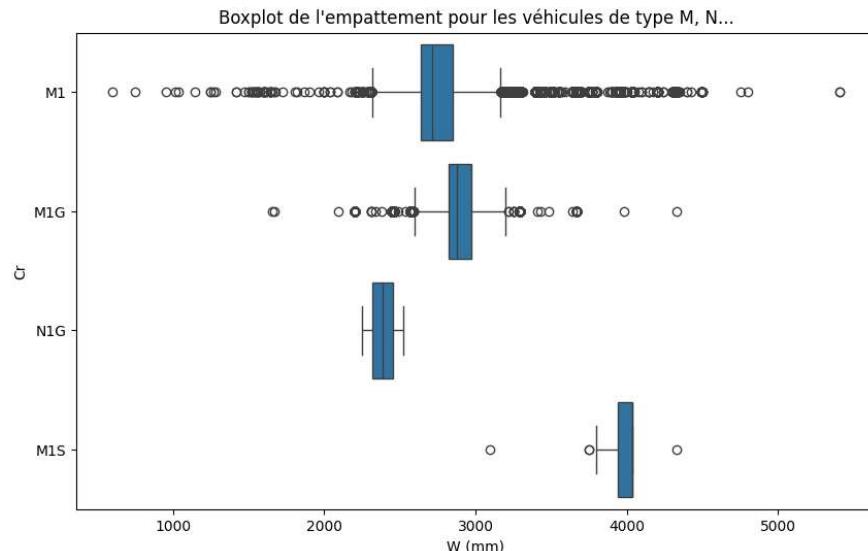
• Essence:

- On vérifie les outliers bas. Conclusion : Tous les ['Ewltp (g/km)'] < 100 sont en réalité des hybrides => on les reclasse
- Ceux entre 100 et 110 sont souvent des hybrides => on tagge des hybrides ou essence prouvés via recherche, et on effectue une prédition de classification par RandomForest pour classifier les autres. On obtient ainsi une répartition plausible sur cette range

- Diesel: Visiblement beaucoup mieux saisis, ils ne nécessitent pas de correction
- Hybrides 'petrol/electric' et 'diesel/electric' : de même

- Empattement : ['W (mm)']**

L'empattement (distance entre les milieux des essieux) montre une distribution différenciée selon le type du véhicule (particulier ou petit transporteur).



Un regard sur les outliers au sens classique ($<Q1-1.5 \times IQR, >Q3+1.5 \times IQR$) nous montre beaucoup trop d'outliers, dont on voit vite que beaucoup sont justifiés car les véhicules sont très différents.

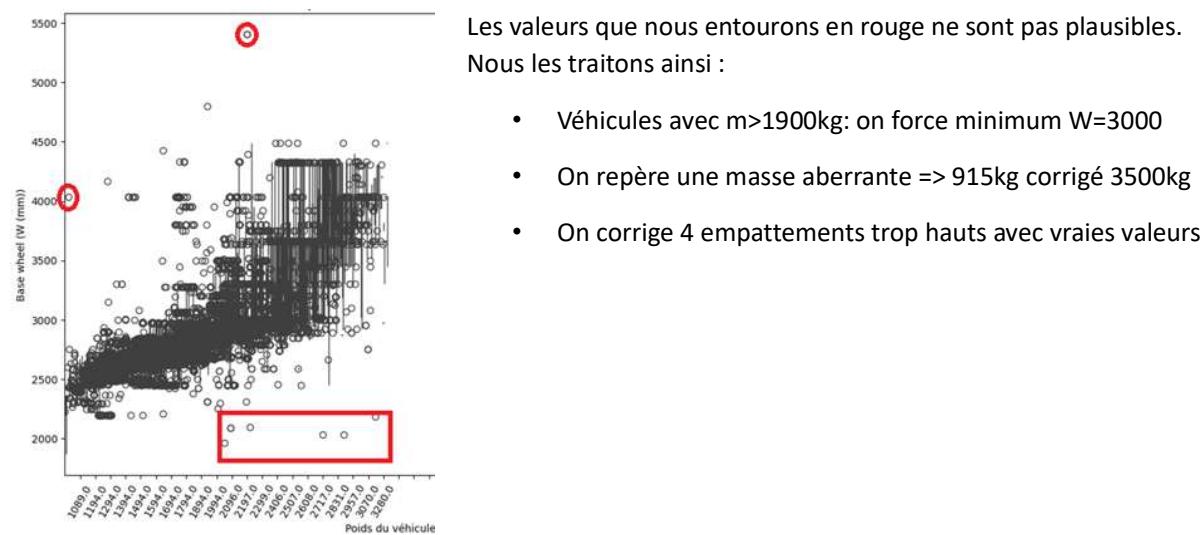
Par contre, si entre 1.5m et 4.5m les valeurs sont plausibles du point de vue "métier", nous devons examiner les lignes qui sont hors de ce range.

Nous constatons et corrigons dans ces lignes :

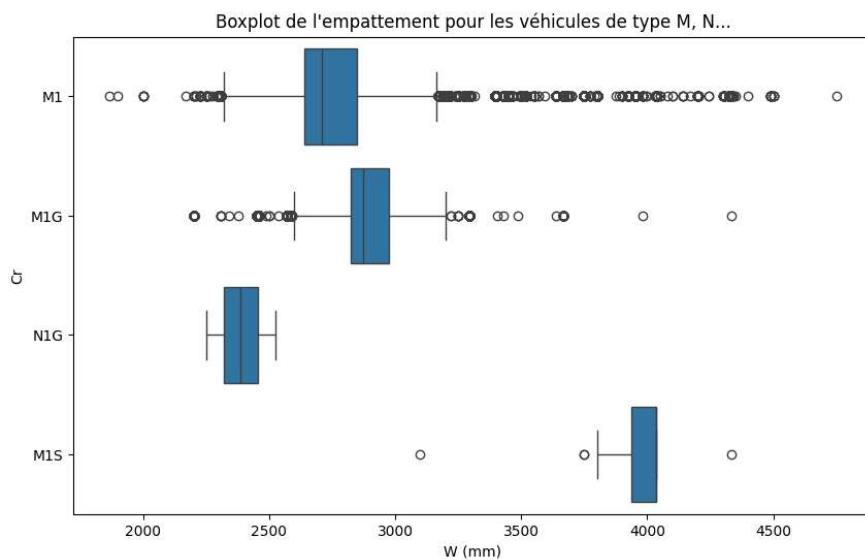
- Des chiffres 'oubliés' (954 au lieu de 3954...)
- Des « empattement divisés par 2 » saisi au lieu de l'empattement complet
- Des chiffres remplis aléatoirement !

De plus, en listant les empattements les plus petits, on constate que c'est à partir de 1867mm que des valeurs sont conformes aux données constructeur. Aussi, nous forçons W=2500 pour ceux qui ont W<1867.

Après traitement de ces extrêmes, l'empattement étant naturellement lié à la masse d'un véhicule, nous utilisons un scatterplot du rapport masse / empattement pour repérer des anomalies :



Nous vérifions les distributions corrigées en réitérant le boxplot initial :



Le range et les outliers sont maintenant plausibles

- Voies ['At1 (mm)'] et ['At2 (mm)']**

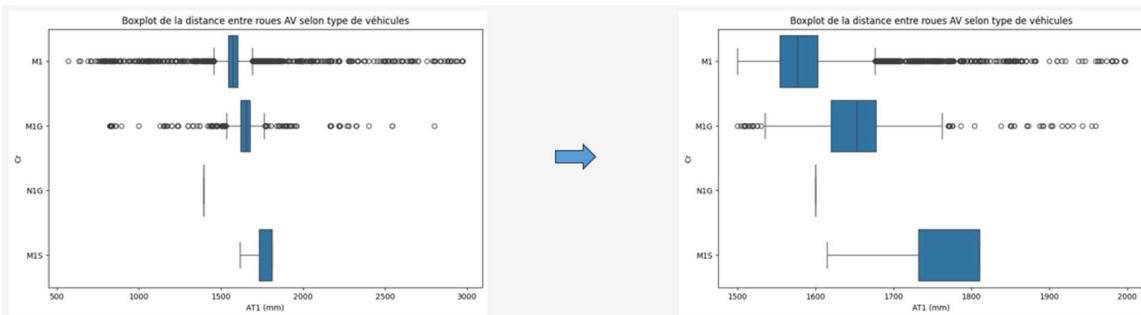
(la voie est l'espacement entre 2 roues du même essieu)

Dans le dataset initial :

- Chiffres 'oubliés'
- « Essieu / 2 » saisi au lieu du complet
- Chiffres remplis en dépit du bon sens !

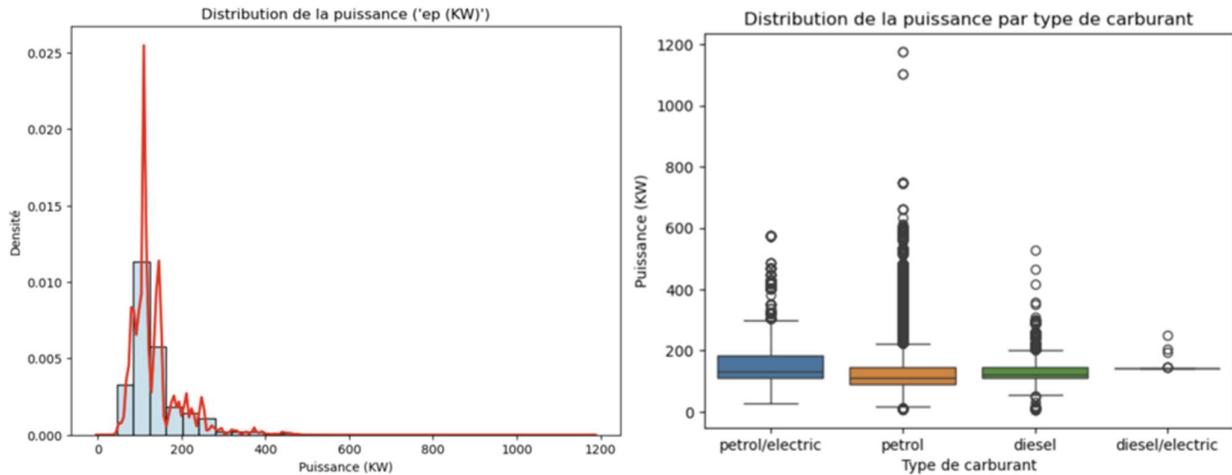
Nos actions:

- On corrige avec ce qu'on trouve en données constructeur
- On met $Atn = 1600$ pour ceux qui ont $Atn < 1500$
- On s'appuie sur la législation : $Atn = 1550$ si $Atn > 1998$



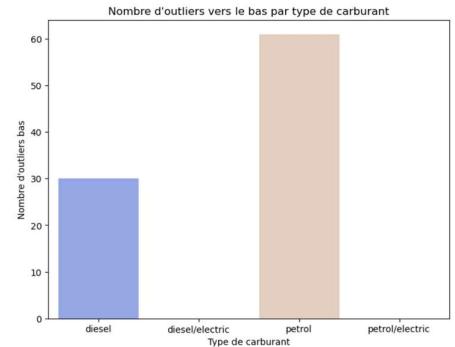
- **Puissance ['ep (KW)']**

Nous voyons sur la distribution générale des données ainsi que sur le box plot qu'il y a des extrêmes à regarder de plus près.

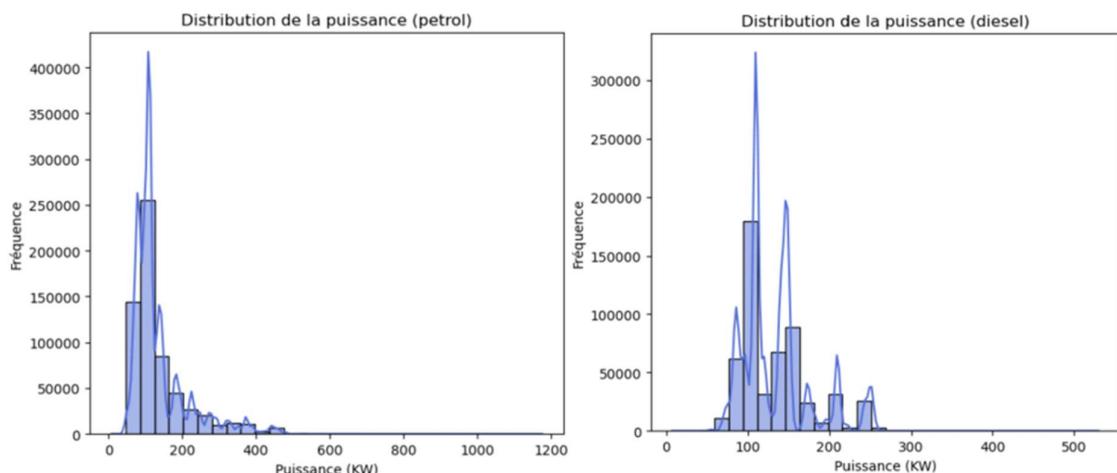


En première lecture on pouvait penser que des puissances aussi élevées n'étaient pas crédibles et que c'étaient des erreurs de saisies. Après avoir regardé plusieurs des lignes avec les maximums par type de carburant, nous voyons finalement que ces modèles de voitures existent vraiment et que leur puissance est celle indiquée. Par contre, pour les valeurs faibles, outliers vers le bas, il s'agit très certainement d'erreurs de saisie. Le jeu de données ayant seulement autour de 90 valeurs vers le bas, on décide de supprimer ces données.

Ft	total_count	outlier_count	outlier_percentage
diesel	532588	60689	11,4
diesel/electric	13681	7	0,05
petrol	618391	76601	12,4
petrol/electric	174781	1238	0,7



Les distributions des puissances par type de carburant semblent assez normales pour les moteurs diesel, essence et hybrides essence / électriques. On voit qu'on a par contre un maximum de valeurs concentrées autour des mêmes valeurs de puissance pour les hybrides diesel / électriques. On confirme avec le tableau ci-après résumant quelques données statistiques pour chaque type de carburant.



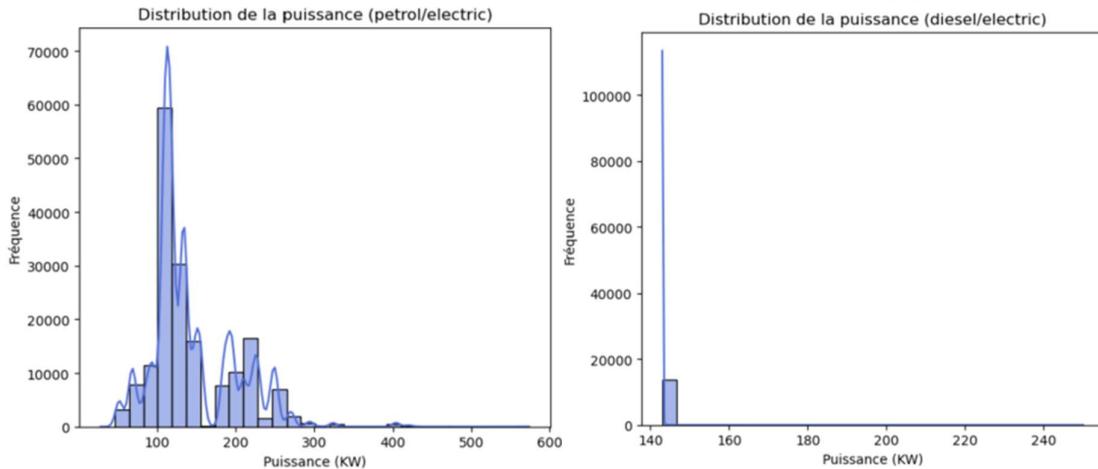
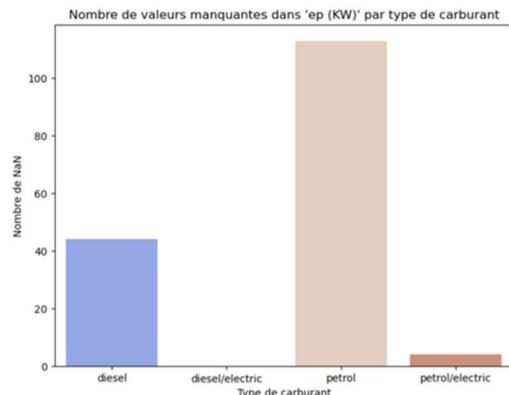


Tableau synthétique des données statistiques des puissances pour les 4 types de carburants :

Ft	count	mean	std	min	25%	50%	75%	max
diesel	532588	134,3	42,8	7	110	120	147	530
diesel/electric	13681	143,0	1,2	143	143	143	143	250
petrol	618391	136,5	77,1	8	92	110	145	1177
petrol/electric	174781	144,2	53,9	28	110	132	186	574

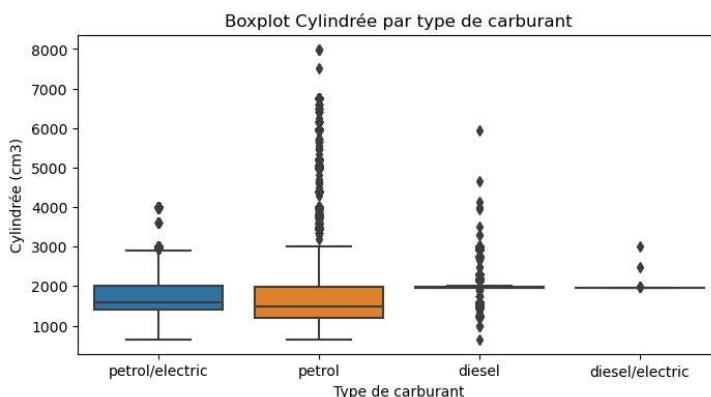
En ce qui concerne le traitement des NaN (tableau ci-après), on voit que le nombre de lignes concernées est faible, on décide donc de les supprimer.

Ft	ep (KW)
diesel	Nombre de NaN
	0,83%
diesel/electric	Nombre de NaN
	0,00%
petrol	Nombre de NaN
	1,83%
petrol/electric	Nombre de NaN
	0,23%

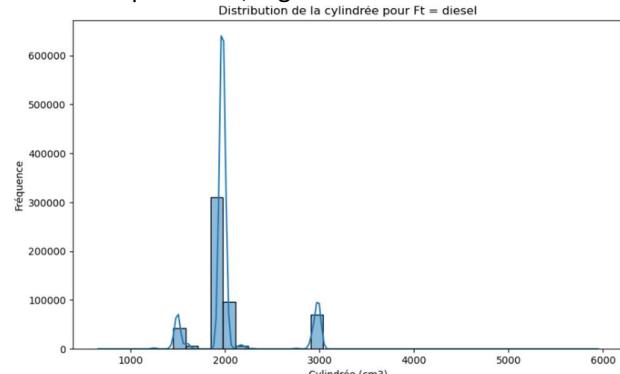
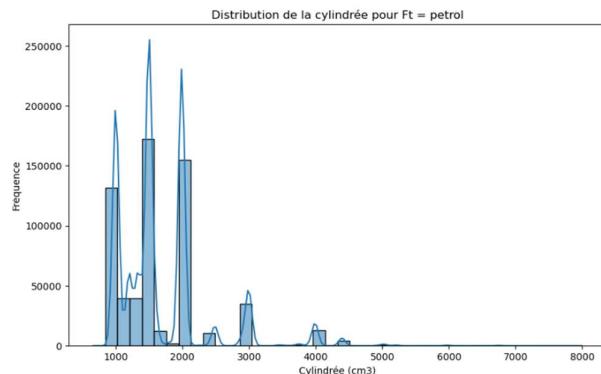


- **Cylindrées ['ec (cm3)']**

On réalise un boxplot pour la variable Cylindrée en distinguant les Fuel types afin d'analyser la répartition des valeurs et de mettre en évidence de potentiels outliers.



Nous vérifions les outliers : ils sont nombreux. Pour savoir s'ils sont plausibles, regardons la distribution.

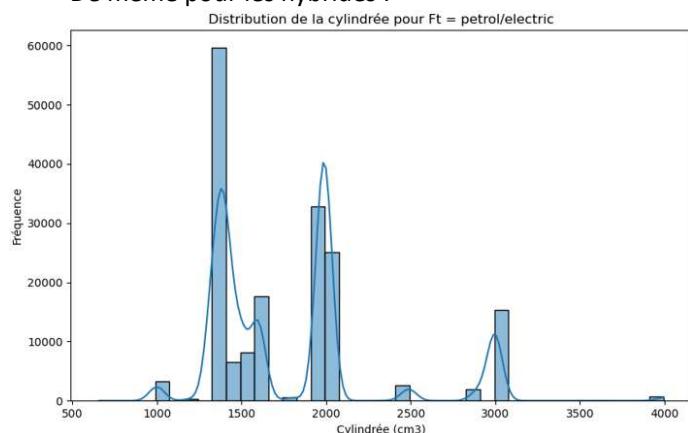


En réalité, les distributions sont multimodales et sont typiques d'utilisation du véhicule :

Moteur diesel

- Petits moteurs (voitures économiques, citadines) :
 - De 1 000 cm³ à 1 600 cm³
- Moteurs moyens (berlines, SUV de taille moyenne) :
 - De 1 600 cm³ à 2 500 cm³
- Moteurs puissants (SUV, véhicules utilitaires, grands berlines) :
 - De 2 500 cm³ à 3 000 cm³
- Moteurs très puissants (grands SUV, véhicules utilitaires lourds) :
 - Plus de 3 000 cm³ (ex : moteurs V6, V8)

De même pour les hybrides :



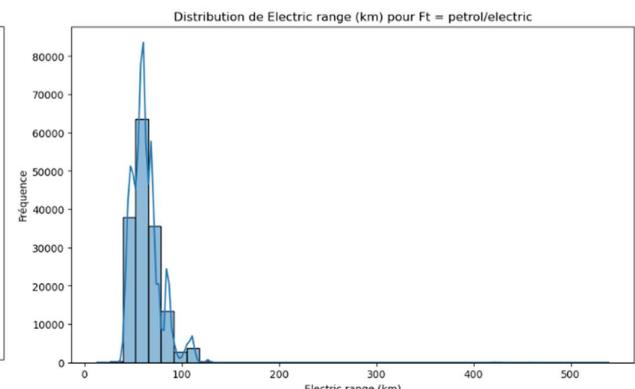
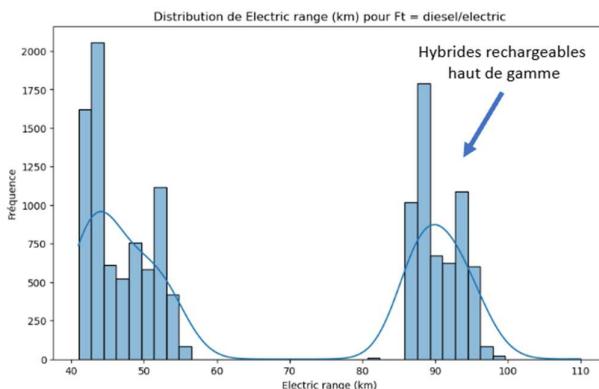
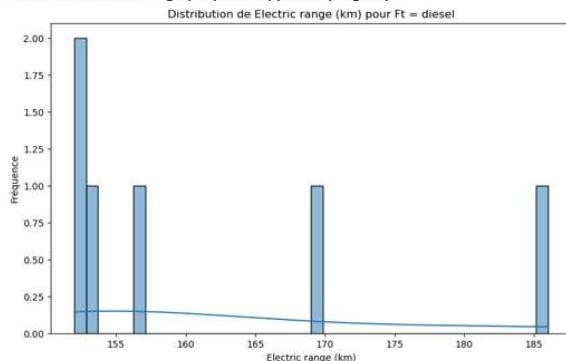
Il s'agissait donc d'un piège, il ne faut pas supprimer tous les outliers qui sont logiques au vu des distributions.

- **Autonomie électrique [' Electric range (km)']**

Certains véhicules diesel purs ont une valeur pour l'autonomie électrique, nous la mettons à zéro pour ne pas perturber les modèles, ainsi tous les carburants fossiles purs ont zéro en autonomie électrique.

Ft		Electric range (km)		
diesel	Nombre de NaN	532 626	6 lignes avec Electric range (km)>0 >> NAN	NAN >>> 0
	Pourcentage de NaN	100,00		
diesel/electric	Nombre de NaN	2	2 lignes avec NAN à supprimer	
	Pourcentage de NaN	0,01		
petrol	Nombre de NaN	618 504	OK	Nan>>0
	Pourcentage de NaN	100,00		
petrol/electric	Nombre de NaN	17 204	à supprimer	
	Pourcentage de NaN	9,84		

Diesel avec Electric range (km) >0 à supprimer (6 lignes)



Enrichissement des données

Créations de nouvelles variables

- **Innovative technology ["IT"]**

Cette variable est de type objet (chaîne de caractère), et composée de champs séparés par des espaces : code de pays certificateur, puis code (nombre à deux chiffres) de l'innovation. Nous décidons d'extraire les codes d'innovation dans de nouvelles colonnes comme suit :

Index	IT	IT28	IT29	IT32	IT33	IT35	IT37	IT38	IT39
	initial	nouveau							
7280	e5 32 e5 37	0	0	1	0	0	1	0	0
9176	e5 32 e5 37	0	0	1	0	0	1	0	0
10339	e5 37 e9 32	0	0	1	0	0	1	0	0
10396	e5 37 e9 32	0	0	1	0	0	1	0	0
10403	e13 29 37	0	1	0	0	0	1	0	0

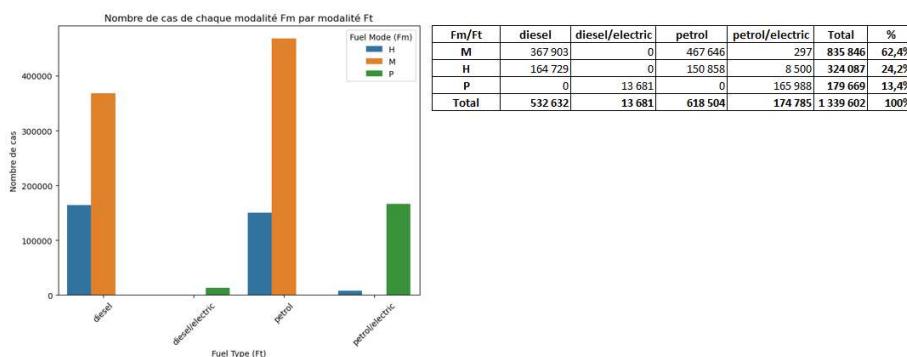
Ainsi, les modèles pourront se baser sur les ITnn pour déduire d'éventuelles influences

- **Le type d'énergie ['Ft']**

Nous dichotomisons aussi ce type, donnant ainsi lieu à 4 colonnes au lieu d'une.

- **Fuel mode (Fm) - variable nominative**

On regarde la répartition du nombre de lignes par modalité des variables Fuel mode et Fuel type



Ensuite, on procède à l'encodage de la variable Fuel mode en utilisant le One Hot Encoding. Ainsi, on obtient 3 nouvelles colonnes : Fm_H, Fm_M et Fm_P.

Afin d'avoir des valeurs numériques, ces variables sont remplies avec 0 ou 1.

Après ces transformations et avant le scaling, une nouvelle élimination des doublons est effectuée (car des colonnes ont été supprimées dans les étapes récentes).

Scaling

Afin d'être considérées correctement par les modèles de prédiction, les valeurs doivent être ramenées dans une moyenne et variance d'ordre habituel.

Ceci se fait avec des scalers scikit-learn. Le choix des scalers dépend de la distribution des données.

Nous vérifions les notes: Visuellement, seule la variable "m (kg)" pourrait sembler de distribution normale.

Nous testons selon Anderson-Darling (il y a plus de 5000 données, ainsi Shapiro-Wilk n'était pas conseillé). Le test d'Anderson-Darling invalide l'hypothèse d'une distribution normale.

Donc aucune distribution n'est de type normal. Nous procérons selon le scaler : robust scaling.

Deux colonnes toutefois contiennent en grande majorité des zéros: il s'agit des valeurs électrique (consommation, autonomie) pour les véhicules à énergie fossile pure. Pour celles-ci nous utilisons un MinMaxScaler afin de garder les zéros à zéro et avoir toutes les valeurs positives dans le range 0-1.

Notons qu'une facilité de programmation est mise en place afin de pouvoir basculer vers l'un ou l'autre modèle si nous voulons par la suite expérimenter plusieurs scalers selon les colonnes, pour fine-tuner les hyper-paramètres de nos modèles de prédiction.

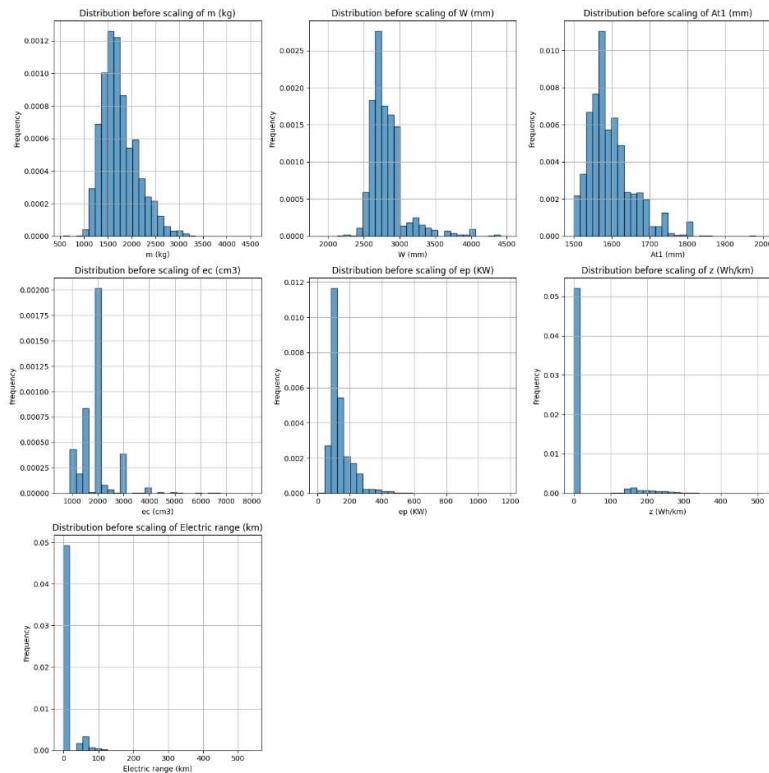
```
standard_cols = []
robust_cols = ["m (kg)", "W (mm)", "At1 (mm)", "ec (cm3)", "ep (KW)"]
min_max_cols = ["z (Wh/km)", "Electric range (km)"]
```

Nous excluons bien entendu la variable cible du scaling.

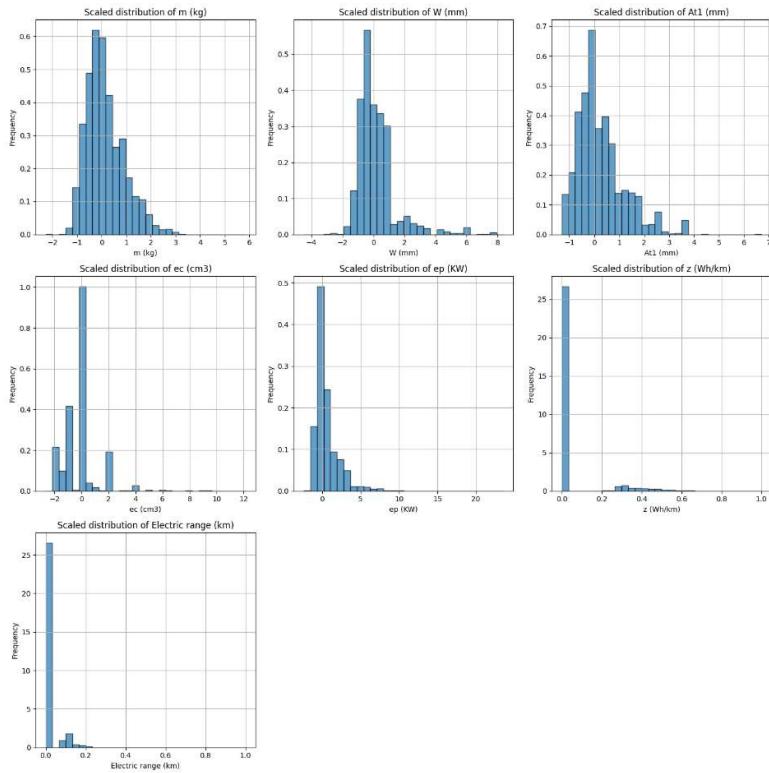
Nous sauvegardons les scalers utilisés, dans le répertoire models, afin de pouvoir nous en servir lorsque nous voudrons prédire l'émission de CO₂ d'un nouveau modèle, ou répondre à d'hypothétiques questions d'un constructeur sur des prévisions.

Nous consignons page suivante les représentations graphiques des distributions des variables explicatives avant et après scaling.

Distribution avant scaling:



Distribution après scaling:



Pour conclure cette première partie du projet, nous pouvons dire que nous disposons après une intense préparation, d'environ 120 000 lignes différentes et en avons réservé 80% pour entraîner bientôt les modèles à prédire notre valeur cible.