# Emission de CO2 par les véhicules

Exploration des données, data visualisation et

preprocessing des données


Marc BASSELIER
Thierry GONCALVES-NOVO
Tanguy LALOUELLE
Louise RIGAL

# Historique des révisions

*Summary of revisions*

| Révision<br>Révision | Noms<br>Names | Date | Description |
|---|---|---|---|
| 0000 | Thierry GONCALVES-NOVO | 27/07/2024 | Version initiale et ajout de la rédaction de Louise RIGAL |
| 0001 | L'équipe | 31/07/2024 | 1ère version complétée avec 5 visualisations |
| 0002 | Marc | 30/08/2024 | Rajout du preprocessing |
| 0003 | | | |
| 0004 | | | |
| 0005 | | | |
| 0006 | | | |

# Sommaire

*Table of contents*



# Table des illustrations

*Table of illustration*

# Table des tableaux

*Table of board*

# 1. Introduction au projet

## 1.1 Contexte

Le transport routier contribue à environ un cinquième des émissions totales de l'Union européenne (UE) de dioxyde de carbone (CO2), le principal gaz à effet de serre (GES), dont 75 % proviennent des voitures particulières. Le secteur des transports est le seul secteur majeur de l'UE où les émissions de GES continuent d'augmenter.

Les émissions de CO2 des voitures de tourisme sont mesurées dans le cadre du test de certification des véhicules, qui est basé sur le nouveau cycle de conduite européen (NEDC), et est également appelé test NEDC.

La consommation de carburant des véhicules est directement dérivée de la mesure des émissions de dioxyde de carbone (CO2), d'hydrocarbures (HC) et d'oxyde de carbone (CO) effectuées lors des tests de certification, en tenant compte du bilan carbone des gaz d'échappement. Les véhicules modernes conformes aux normes européennes (Euro5 et Euro6) ont des niveaux d'émissions de CO et de HC faibles à l'échappement (contribuant à environ 1 % de la consommation de carburant). En d'autres termes, les émissions de CO2 peuvent être considérées comme proportionnelles au carburant consommé pendant le fonctionnement du véhicule.

L'écart croissant entre la consommation de carburant en conditions réelles et celle des véhicules homologués, ainsi que la difficulté d'évaluer l'effet réel des technologies de réduction du CO2, ont conduit l'UE à revoir la procédure d'homologation des voitures particulières et des véhicules utilitaires légers, ce qui a abouti à l'introduction de la nouvelle procédure d'essai des véhicules utilitaires légers harmonisée à l'échelle mondiale (WLTP). Cette nouvelle procédure est utilisée pour l'évaluation des émissions, y compris le CO2, dans le cadre de la réception par type des véhicules utilitaires légers depuis le 1er septembre 2017. Toutefois, les objectifs en matière de CO2 continuent d'être évalués par rapport aux valeurs de CO2 de la NEDC.

L'Union Européenne a fixé à l'ensemble des constructeurs automobiles pour objectif une réduction des émissions moyennes de CO2 concernant l'immatriculation des voitures neuves. À partir de 2035, toutes les nouvelles voitures qui arriveront sur le marché de l'UE devraient être à zéro émission de CO2. Ces règles n'affectent pas les voitures existantes.

Du fait du changement de procédure d'essai à partir du 1er septembre 2017, une période de transition a été tolérée pour le passage progressivement du test NEDC au test WLTP.
Des objectifs intermédiaires ont été fixées :

**Tableau 1 - Objectifs de réductions de l'émissions de CO2 de l'UE**

| Période | Emission de CO2 (g/km) | Type de test |
|---|---|---|
| 2015-2020 | 130 | NEDC |
| 2020-2024 | 95 | NEDC |
| 2021-2024 | 119 | WLTP |
| 2025-2029 | 93.6 | WLTP |
| 2030-2034 | 49.5 | WLTP |
| 2035 | 0 | WLTP |

## 1.2 Objectifs

L'objectif principal de notre projet est d'identifier les différents facteurs et caractéristiques techniques jouant un rôle dans la pollution émise par les véhicules. Prédire à l'avance la pollution de certains types de véhicules est une information cruciale pour opérer une décarbonation de l'industrie automobile.

Aucun d'entre nous présente une expertise sur le sujet.

Il existe plusieurs projets concernant les émissions de CO2 émises par les véhicules, menés à partir d'une base de données mise à disposition au Canada.
Par ailleurs, Olivier Viollet a publié un projet de machine learning sur la base de données Car Labelling de l'ADEME en exploitant les données 2014 en code R (https://rpubs.com/Olivier_Viollet_2019/CO2_emissions).

À notre connaissance, il n'y a pas de projet de machine learning en python et les bibliothèques associées sur la base de données Labelling de l'ADEME et encore moins sur la base de données mise à disposition par la commission européenne (EEA).

# 2 Compréhension et manipulation des données

Nous disposons de deux sources de données en accès libre :

- Les données mises à disposition par l'European Environment Agency (EEA)
  https://www.eea.europa.eu/data-and-maps/data/co2-cars-emission

- Les données mises à disposition par le gouvernement français par l'intermédiaire de l'Agence de l'environnement et de la maîtrise de l'énergie (ADEME)
  Emissions de CO2 et de polluants des véhicules commercialisés en France - data.gouv.fr

## 2.1 Cadre

### 2.1.1 Description du jeu de données de l'EEA

Le règlement (UE) n° 2019/631 impose aux pays d'enregistrer des informations pour chaque nouvelle voiture particulière immatriculée sur leur territoire.

Chaque année, chaque État membre soumet à la Commission toutes les informations relatives à ses nouvelles immatriculations. En particulier, les informations suivantes sont requises pour chaque nouvelle voiture particulière immatriculée :

- nom du constructeur,
- numéro d'homologation,
- type, variante, version, marque et nom commercial,
- émissions spécifiques de CO2 (protocoles NEDC et WLTP),
- masses du véhicule, empattement, largeur des voies,
- capacité et puissance du moteur,
- type et mode de carburant,
- éco-innovations et consommation d'électricité.

Sur le site de l'EEA, on retrouve plusieurs jeux de données portant sur l'ensemble des véhicules enregistrés dans l'Union Européenne (UE-27 et Royaume-Uni) entre 2010 et 2023.

Les jeux de données de l'European Environment Agency sont notablement plus lourds que ceux réunissant exclusivement les données françaises :

- Chaque année, plusieurs centaines de milliers de véhicules sont répertoriés par l'Union européenne. À titre d'exemple, le dataset de l'année 2014 enregistre plus de 400 000 entrées au niveau européen, pour environ 55 000 à l'échelle de la France.

- On retrouve également quelques variables supplémentaires par rapport aux jeux de données français : l'empattement (mm), la capacité moteur (cm3), etc.

### 2.1.2 Description du jeu de données de l'ADEME

Les jeux de données mis à disposition par l'ADEME concernent les caractéristiques techniques des véhicules commercialisés en France entre 2001 et 2014, ainsi que les consommations de carburant, les émissions de CO2 et les émissions de polluants dans l'air.

Depuis 2001, l'ADEME acquiert tous les ans ces données auprès de l'Union Technique de l'Automobile du motocycle et du Cycle UTAC (en charge de l'homologation des véhicules avant leur mise en vente) en accord avec le ministère du développement durable.

Pour chaque véhicule les données d'origine (transmises par l'UTAC) sont les suivantes :
- les consommations de carburant,
- les émissions de dioxyde de carbone (CO2),
- les émissions des polluants de l'air (réglementés dans le cadre de la norme Euro),
- l'ensemble des caractéristiques techniques des véhicules (gammes, marques, modèles, n° de CNIT, type d'énergie ...).

Sur le site Car Labelling, L'ADEME complète ces données avec les informations suivantes :
- les valeurs du bonus malus et de l'étiquette Classe Energie - CO2 (qui varient en fonction de la réglementation issue de la Loi de Finance et de ses décrets),
- les résultats d'expertise tels que le coût annuel de la consommation de carburant sur 15 000 km.

Elle établit également des classements pour distinguer les véhicules « les plus propres en CO2 et les plus économes en énergie » (Palmarès).
L'ADEME publie chaque année un guide officiel « Véhicules particuliers neufs : consommations conventionnelles de carburant et émissions de CO2 ».

Les jeux de données mis à disposition par l'ADEME sont certes peu récents (données disponibles jusqu'en 2014), mais ils fournissent un nombre d'observations intéressant (plus de 16 000 entrées pour la compilation des années 2012 à 2015) et des variables très pertinentes pour notre analyse.

### 2.1.3 Volumétries et choix

La volumétrie des jeux de données sur lesquels nous nous sommes penchés est très variable.

Les jeux de données récents de l'European Environment agency sont lourds et prennent en général beaucoup de temps à être téléchargés. À titre d'exemple, le dataset publié pour l'année 2023 pèse 2,28 GB, celui répertoriant les voitures enregistrées pour l'année 2019, plus d'1 GB… Nous ne sommes pas parvenus à lire certains jeux de données européens (RAM insuffisante, encoding différents…).

À partir de ces constats, nous nous sommes concentrés sur les jeux de données français et européens de l'année 2014, pour une première exploration.

Après une observation approfondie nous avons décidé de conserver uniquement les jeux de données français, fournis par l'ADEME. Afin d'avoir en notre possession un large nombre d'observations, nous avons procédé à la concaténation des jeux de données français des années 2012 à 2015.

Plusieurs raisons ont motivées ce choix :

- Une raison principale : la consommation de carburant est absente du jeu de l'EEA en 2014. Cette variable nous semble pourtant cruciale pour notre analyse, car très corrélées aux émissions de CO2 et de polluants.
- La taille des jeux de données français concaténés, bien qu'évidemment inférieure à celle du jeu de données européen, reste satisfaisante pour mener une analyse avec 160826 entrées.

## 2.2 Composition du jeu de données et pertinence des variables

Note : afin d'augmenter la quantité de données disponibles, le jeu de données utilisé regroupe le jeu de données de l'ADEME de 2012 à 2015.

**Tableau 2 - Résumé concis des informations du jeu de données**

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | Marque | 160826 non-null | object |
| 1 | Modèle dossier | 160826 non-null | object |
| 2 | Modèle UTAC | 160826 non-null | object |
| 3 | Désignation commerciale | 160826 non-null | object |
| 4 | CNIT | 160826 non-null | object |
| 5 | Type Variante Version (TVV) | 160826 non-null | object |
| 6 | Carburant | 160826 non-null | object |
| 7 | Hybride | 160826 non-null | object |
| 8 | Puissance administrative | 160826 non-null | int64 |
| 9 | Puissance maximale (kW) | 160770 non-null | float64 |
| 10 | Boîte de vitesse | 160826 non-null | object |
| 11 | Consommation urbaine (l/100km) | 160588 non-null | float64 |
| 12 | Consommation extra-urbaine (l/100km) | 160588 non-null | float64 |
| 13 | Consommation mixte (l/100km) | 160667 non-null | float64 |
| 14 | CO2 (g/km) | 160667 non-null | float64 |
| 15 | CO type I (g/km) | 159943 non-null | float64 |
| 16 | HC (g/km) | 37430 non-null | float64 |
| 17 | NOX (g/km) | 159943 non-null | float64 |
| 18 | HC+NOX (g/km) | 122688 non-null | float64 |
| 19 | Particules (g/km) | 150599 non-null | float64 |
| 20 | masse vide euro min (kg) | 160826 non-null | int64 |
| 21 | masse vide euro max (kg) | 160826 non-null | int64 |
| 22 | Champ V9 | 160392 non-null | object |
| 23 | Date de mise à jour | 68977 non-null | object |
| 24 | Carrosserie | 139946 non-null | object |
| 25 | gamme | 139946 non-null | object |

### 2.2.1 Pertinence des variables

Plusieurs variables retiennent notre attention :

- Variable cible : les émissions de CO2 (en g/km)
- Variables pertinentes dans notre analyse :
  - Puissance moteur (kW),
  - Consommations de carburant (l/100km) : urbaine, extra urbaine et mixte,
  - Résultat d'essai de CO type I,
  - Résultats d'essai sur différentes particules polluantes : HC (Hydrocarbures imbrûlés), nox, hcnox, ptcl (particules),
  - Motorisation : type de carburant utilisé (essence, gazole, …),
  - Masse du véhicule : en ordre de marche maximal et minimal.

### 2.2.2 Particularité du jeu de données

On observe quelques particularités dans notre jeu de données :

Une surreprésentation de certaines marques de véhicules, notamment Mercedes avec 123926 occurrences. Un rééchantillonnage pourrait être pertinent pour la suite.

**Tableau 3 - Proportion (en %) de valeurs manquantes dans le jeu de données français**

| | | | |
|---|---|---|---|
| Marque | 0.000000 | Consommation mixte (l/100km) | 0.098865 |
| Modèle dossier | 0.000000 | CO2 (g/km) | 0.098865 |
| Modèle UTAC | 0.000000 | CO type I (g/km) | 0.549041 |
| Désignation commerciale | 0.000000 | HC (g/km) | 76.726400 |
| CNIT | 0.000000 | NOX (g/km) | 0.549041 |
| Type Variante Version (TVV) | 0.000000 | HC+NOX (g/km) | 23.713827 |
| Carburant | 0.000000 | Particules (g/km) | 6.359046 |
| Hybride | 0.000000 | masse vide euro min (kg) | 0.000000 |
| Puissance administrative | 0.000000 | masse vide euro max (kg) | 0.000000 |
| Puissance maximale (kW) | 0.034820 | Champ V9 | 0.269857 |
| Boîte de vitesse | 0.000000 | Date de mise à jour | 57.110791 |
| Consommation urbaine (l/100km) | 0.147986 | Carrosserie | 12.982975 |
| Consommation extra-urbaine (l/100km) | 0.147986 | gamme | 12.982975 |

Le jeu de données comporte deux variables qui présentent un grand nombre de valeurs manquantes (16 : HC (g/km) et 23 : Date de mise à jour). Ces variables seront donc supprimées. La présence d'un quantité non négligeable de valeurs manquantes dans les variables Carrosserie et gamme est provoquée par l'inclusion du jeu de données de 2015 (ces variables y sont absentes).

Afin de pouvoir déterminer plus facilement les variables à cibler, il est possible de créer une *heatmap*, une figure affichant le degré de corrélation entre les variables numériques d'un jeu de données. Un intérêt particulier sera donné aux variables ayant un fort degré de corrélation (le plus éloigné de 0) avec la variable cible : CO2 (g/km).

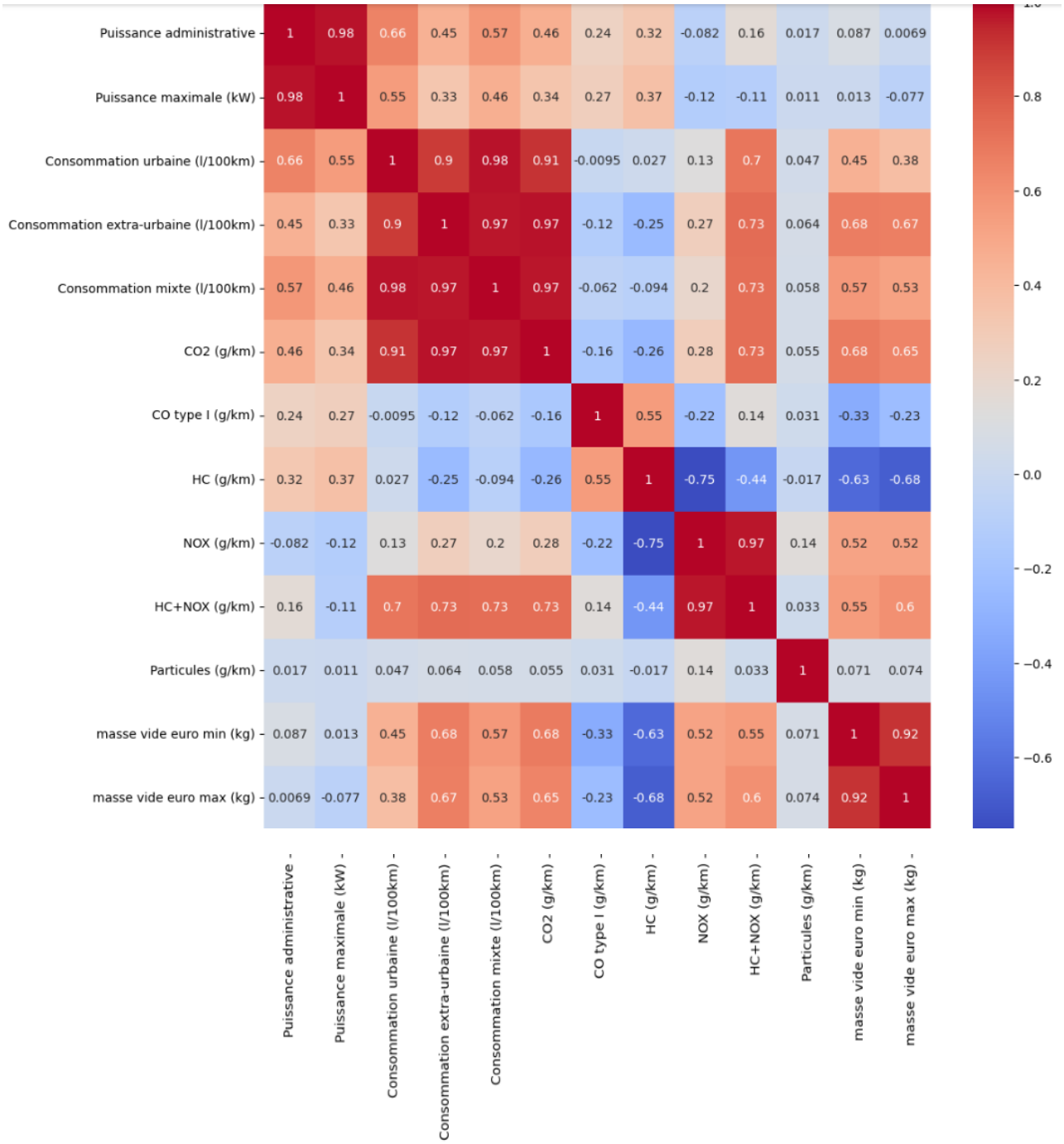


**Figure 1 - Matrice de corrélation (Heatmap) du jeu de données**

Plusieurs variables sont corrélées avec la variable cible, notamment une, avec un degré de corrélation très élevé (0.97) : la Consommation mixte (l/100km), qui, comme son nom l'indique, donne la consommation en carburant du véhicule en litre pour 100 km (urbaine et extra-urbaine).

# 3 Visualisation

## 3.1 Émissions de CO2 par rapport à la consommation

En raison de sa très forte corrélation avec la variable cible, il est intéressant de visualiser la relation entre ces deux variables sous la forme d'un nuage de points.

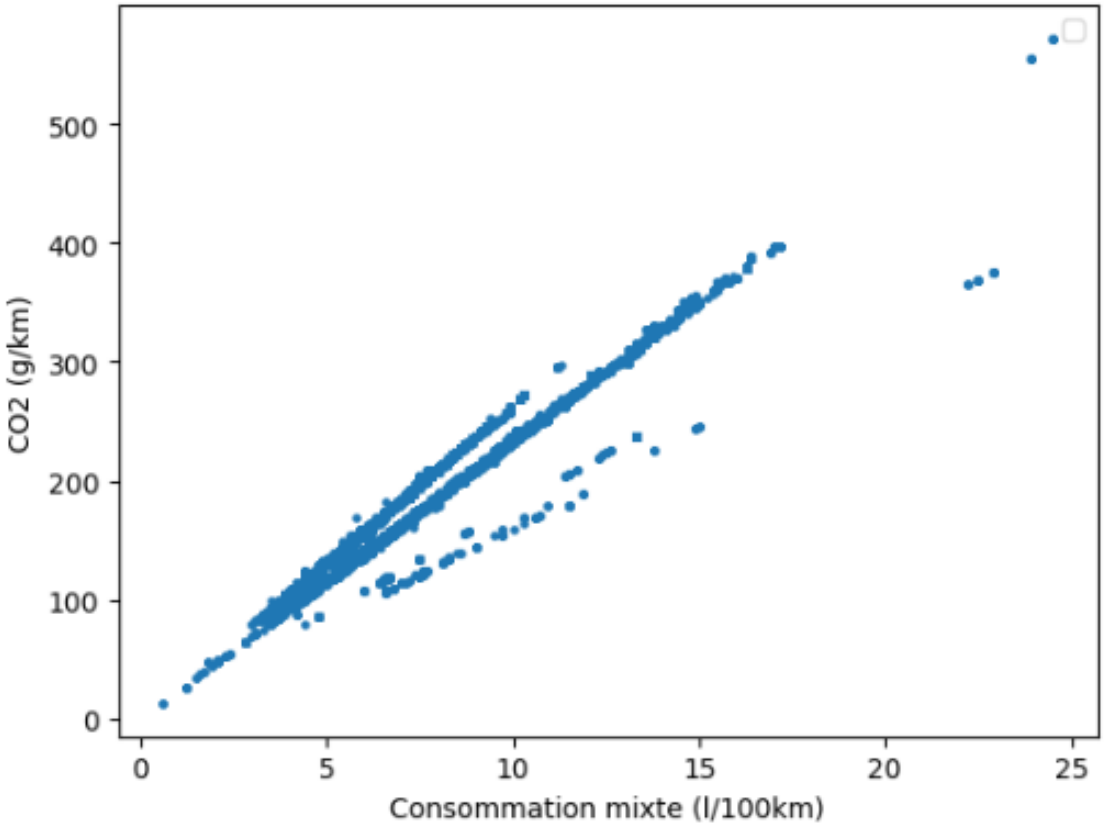


Figure 2 - Émission de CO2 (g/km) en fonction de la consommation mixte (l/100km)

Comme attendu, les points se regroupent de façon linéaire, ce qui signifie que cette variable nous sera utile pour prédire les émissions. Toutefois, plusieurs droites semblent se dessiner, plutôt qu'une. Cela indique donc qu'une variable supplémentaire affecte les résultats, certainement une variable catégorielle.

Parmi les variables catégorielles disponibles, l'une d'elle (Carburant) indique la motorisation du véhicule (essence, gazole, GPL, …). Les carburants étant de composition différentes, les émissions provoquées par la consommation d'un litre de carburant ne sera pas forcément la même.

Dans ce jeu de données, les véhicules hybrides ont un label différent des véhicules non-hybrides, même s' ils utilisent le même carburant. Par exemple : **ES** désigne les véhicules **essence non-hybrides**, tandis que **EH** désigne un véhicule **essence hybride**. Les labels ont donc été regroupés afin de réduire leur nombre.

Le graphique ci-dessous reprend donc les mêmes variables que précédemment, cette fois en prenant en compte la motorisation du véhicule :

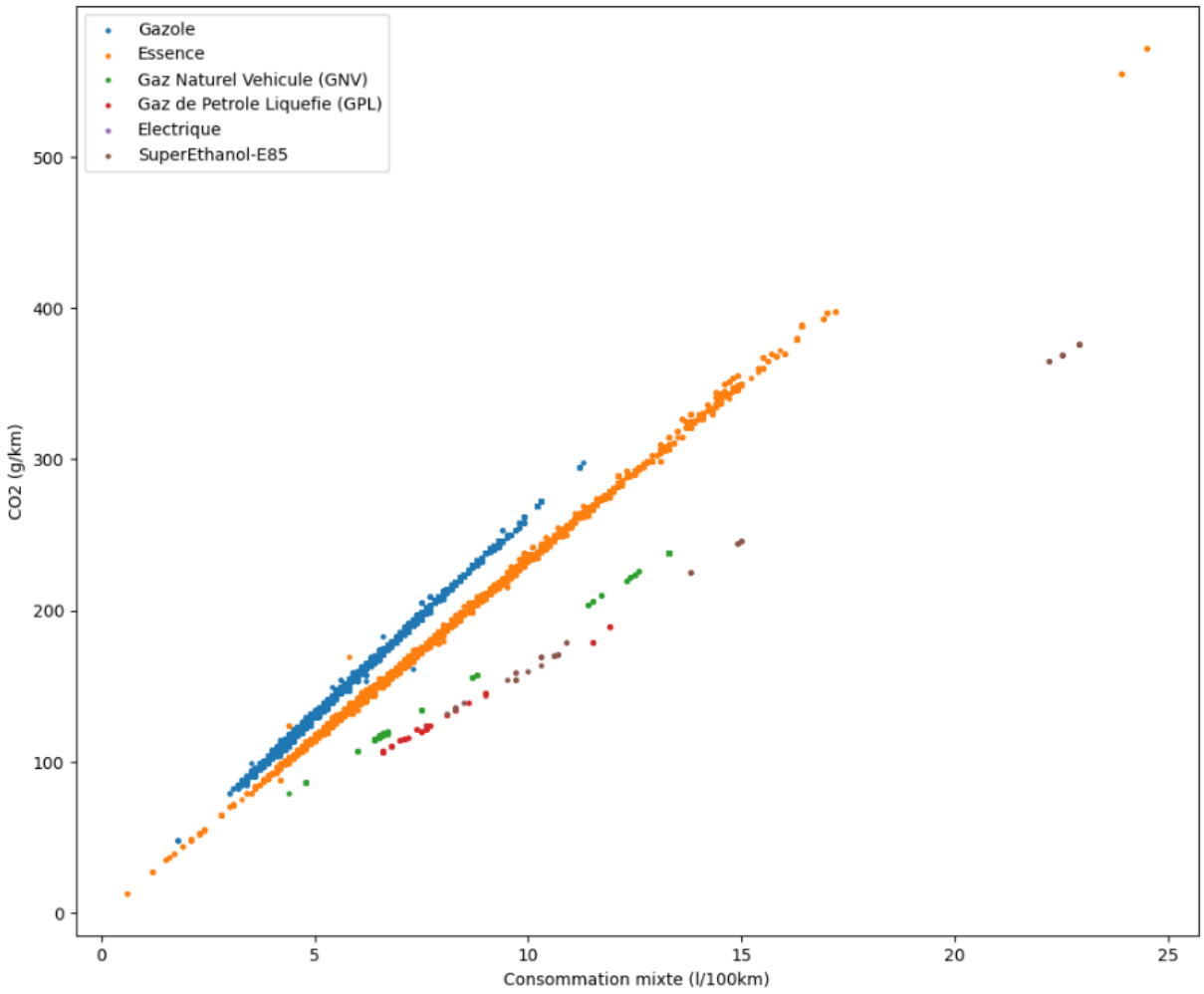


**Figure 3 - Émission de CO2 (g/km) en fonction de la consommation mixte (l/100km) et du type de carburant**

Comme nous pouvons le voir, le type de carburant utilisé a aussi un impact sur les émissions de CO2 du véhicule, il sera donc utile de garder cette variable pour prédire les émissions des véhicules.

## 3.2 Répartition des différents types de carburants des véhicules

Comme indiqué dans le graphique précédent, le type de carburant a une influence sur les émissions de CO2. Toutefois, la faible présence sur le graphique de véhicules utilisant un carburant autre que l'essence ou le gazole indique un potentiel déséquilibre dans le jeu de données. Il est donc intéressant de déterminer la proportion de chaque type de motorisation présentes dans le jeu de données.

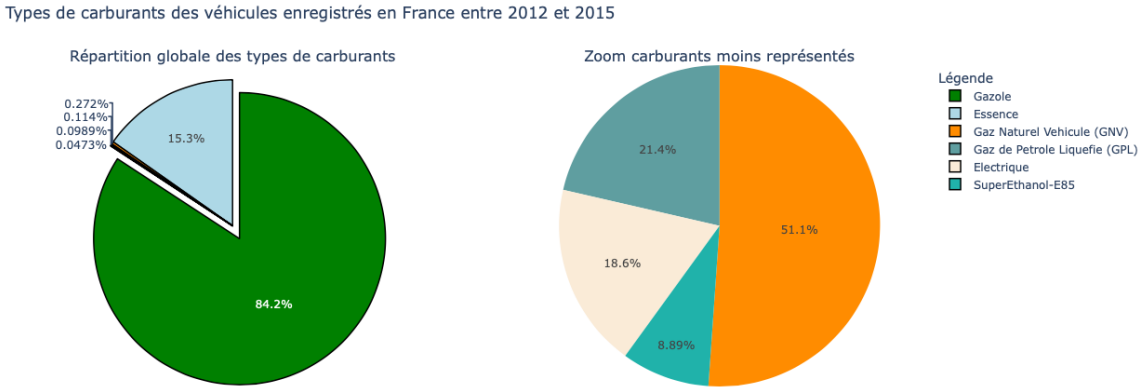


**Figure 4 - Proportion de chaque type de motorisation**

Comme nous pouvons le voir dans le graphique ci-dessus, il y a une très forte représentation de véhicules utilisant du gazole comme carburant (84,2%), et dans une bien moindre mesure, les véhicules essence (15.3%). Les quatre autres motorisations ne représentent au final qu'un total de 0.5% du jeu de données restant.

Ce déséquilibre entre les types de carburants présents peut entraîner un biais qui peut fausser les prédictions d'émission des véhicules utilisant ces types de carburants sous-représentés.

## 3.3 Emission de CO2 (g/km) en fonction des différents carburants

Le graphique ci-dessous doit nous permettre de vérifier la distribution des valeurs d'émissions des véhicules selon le type de carburant utilisé, afin, entre autre, de faire apparaître d'éventuelles valeurs aberrantes :

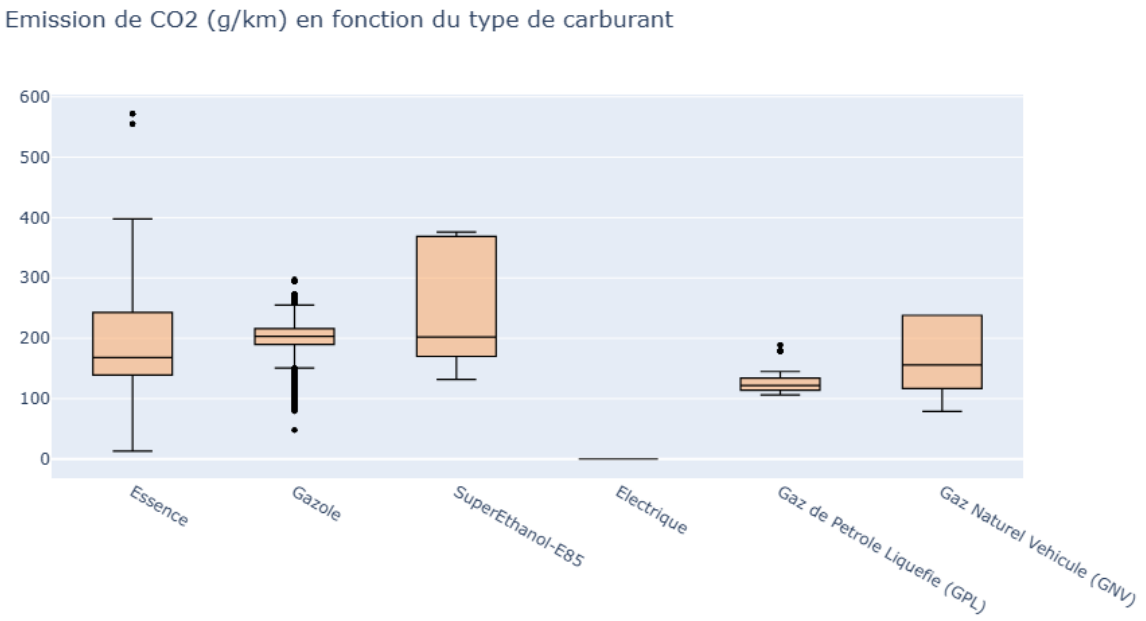


**Figure 5 - Boîte à moustaches (Box plot) de l'émission de CO2 (g/km) en fonction du type de carburant**

La présence de points hors des boîtes (notamment dans la catégorie gazole) indique la présence de valeurs éloignées du reste des autres valeurs. Toutefois, leurs écarts ne semblent pas significatifs, ce qui signifie que ces valeurs, bien que extrêmes, restent valables et peuvent donc être gardées dans le jeu de données.

# 4 Preprocessing des données

## 4.1 Rappel des conclusions sur le jeu de données choisi

Le jeu de données retenu est constitué de 160 826 observations.

La répartition des données est déséquilibrée au niveau des marques de véhicules et des types de carburant. Les marques Mercedes et Volkswagen sont surreprésentées. Le gazole représente 84.2% et l'essence 15.3% de la distribution de carburant. Les autres types de carburant représentent seulement 0.5%, et incluent les véhicules électriques qui ont une émission de CO2 nulle.

La matrice de corrélation ainsi que les différentes visualisations montrent une forte corrélation entre l'émission de CO2 et la consommation mixte de carburant, avec une dépendance selon le type de carburant.

## 4.2 Méthodologie suivie

Si le prepocessing des données a pu évoluer au cours de l'avancement du projet, notamment au niveau de la sélection des données à utiliser, nous avons suivi une ligne directrice claire de simplicité qui nous a guidé dans les choix effectués.

La méthodologie suivie s'explique par le caractère extrêmement concret de la problématique d'émission de CO2 par un véhicule motorisé. La plupart des données de notre base font références à des concepts familiers pour tout le monde, et plus important encore, ces concepts accessibles semblent expliquer notre variable cible. Il nous a semblé important de conserver cette simplicité et lisibilité pour permettre d'obtenir un modèle explicatif et permettant d'agir en amont et en aval de prises de décision pour les acteurs potentiels (constructeur, utilisateur, administration réglementaire ou lobbying écologique…). Cette méthodologie est essentiellement constitué des deux principes suivants :

- Limiter si possible le nombre de variables du jeu de données sans compromettre la qualité des résultats de modélisation
- Ne conserver si possible que des variables largement accessibles et compréhensibles (pour un utilisateur qui n'aurait pas accès à l'ensemble des données moteur, par exemple)

Notons enfin que le preprocessing des données n'est pas une étape séparée du reste de la modélisation et que nous avons ainsi en équipe finit par simplifier le nombre de variables processées au fur et à mesure que nous avons constaté que les résultats de certains membres de l'équipe sur un nombre de variables restreint restaient équivalents à ceux sur un nombre de variables plus larges.

## 4.3 Sélection des données

La sélection finale de données a porté sur les 4 variables suivantes :

- Puissance administrative
- Consommation mixte (l/100km)
- Masse vide euro min (kg)
- Carburant

Nous avons supprimé de la base initiale toutes les données relatives aux résultats d'essai sur particules polluantes (CO, HC, nox, hcnox, ptcl) dont la disponibilité est a priori concomitante à la disponibilité de notre donnée cible d'émission de particule CO2, et donc peu légitime pour l'utilité d'un modèle de prédiction.

Pour les autres champs de données (pour rappel - cf 2.2.1 - : la puissance moteur, la consommations de carburant, la motorisation et le type de carburant utilisé), nous n'avons conservé qu'une variable par domaine. Ce choix est justifié a priori par la très forte corrélation entre les différentes variables disponibles pour chacun de ces domaines couverts. Il est validé a posteriori par les résultats obtenus.

Pour chacun de ces domaines, nous avons conservé la variable la plus explicative et/ou la plus accessible. Les variables choisies sont également celles qui contiennent le moins de

valeurs manquantes et/ou le moins d'occurrences uniques. On peut ainsi conjecturer que ces variables sont plus lisibles dans leurs catégories respectives.

Le détail des choix est décrit ci-dessous :

- consommation mixte plutôt que consommation extra-urbaine ou consommation urbaine : cela semble la variable qui décrit le mieux l'ensemble de la catégorie. Néanmoins, il convient ici de souligner que les modèles produits sont peu sensibles à ce choix, ce qui permet leur utilisation même pour une étude qui ne se concentrerait que sur les véhicules à usage urbain.
- puissance administrative plutôt que puissance maximale : 68 occurrences contre 290, ne contient aucune donnée manquante, et est a priori plus facilement accessible ;
- masse vide min plutôt que masse vide max : un tout petit peu moins d'occurrences uniques.

Il est important de noter que nous n'avons pas eu besoin de conserver les données catégorielles relatives à la description du véhicule (marque, modèle, carrosserie, gamme) ni la date de mise à jour (liée a priori à la date de lancement du véhicule).

La date pourrait être pertinente (on peut imaginer que les technologies progressent en matière de pollution dans le temps à caractéristiques moteur équivalentes), mais elle n'a pas eu d'incidence sur notre modèle. Cela est probablement dû à la période relativement courte de 4 ans étudiée ici.

En revanche, il semble qu'aucun constructeur n'ait réellement d'avantage compétitif en matière de pollution au CO2, et, de manière encore plus pertinente, que la catégorie du véhicule (gamme, carrosserie) n'a pas d'incidence significative et que leur impact est déjà intégré dans notre jeu de donnée (par les données de consommation, puissance et poids).

Nous avons donc supprimé toutes ces données dans les modèles décrits dans ce rapport. A noter pour finir que ce faisant, nous avons conservé une sur-représentation des marques Mercedes et Volkswagen dans le jeu de données. Cela n'est pas préjudiciable (les résultats sont équivalents là encore quelque soit la méthode choisie) et cela s'explique notamment par la grande diversité des modèles de ces 2 constructeurs vis-à-vis de nos variables, ce qui évite les effets de surentraînement ou de concentration sur les modèles produits. Les conserver nous semblait ainsi être le choix le plus objectif et cohérent pour ce travail.

## 4.4 Nettoyage et préparation des données

Le nettoyage des données est effectué en supprimant simplement les valeurs manquantes.
La table passe ainsi de 160 826 entrées à 160 667, ce qui ne représente que 0.1% de données manquantes !

Notre seule variable catégorielle 'Carburant' comporte 13 occurences. Mais l'analyse précise de ces occurences permet de les regrouper en 5 catégories: Essence et Gazole, qui représentent la majorité des entrées, Gaz Naturel, Gaz de Pétrole Liquéfié (GPL) et Super-Ethanol-E85. Les données ont été renommées en conséquence, puis la variable a été transformée en variables d'état (OneHotEncoder fait simplement avec l'instruction pd.get_dummies). Et pour simplifier l'analyse des modèles, nous avons modifier les catégories de carburant en liste numérique (avec LabelEncoder).

Ensuite, la variable cible CO2 a été conservée comme telle, tandis que les 3 autres variables numériques ont été normalisées par l'application de la fonction StandardScaler, compte tenu des différentes échelles de ces métriques (entre 800 et 3000 pour le poids et entre 2 et 80 pour la puissance administrative aux extrêmes).

Enfin, les données ont été séparées en un échantillon d'entraînement et un échantillon de test de 20% - en utilisant le paramétrage de la génération de variables aléatoires afin de pouvoir comparer les différents modèles étudiées sur des bases comparables.

## 5 Conclusion

Nous avons eu deux jeux de données à notre disposition. Un premier dataset regroupant les données collectées par le gouvernement français par l'intermédiaire de l'Agence de l'environnement et de la maîtrise de l'énergie (ADEME). Un deuxième jeu rassemblant les données mises à disposition par l'European Environment Agency (EEA).

La volumétrie ainsi que l'absence notable de la variable associée à la consommation de carburant dans le jeu de données européen, nous conduit à privilégier la source données de l'ADEME.
Le jeu de données retenu est constitué par les données disponibles en France entre 2012 et 2015, représentant 160 826 observations.

Nous avons conservé la quasi-intégralité de ces entrées, en revanche, nous avons fortement restreint le nombre de variables explicatives dans le but premier d'obtenir un modèle facilement utilisable à une majorité d'acteurs concernés. Ce choix s'est également fait et révélé a posteriori, au fur et à mesure de l'analyse des modèles générées.

Au terme de l'exploration de données, il s'avère que notre projet d'un point de vue Machine Learning est un problème de régression.