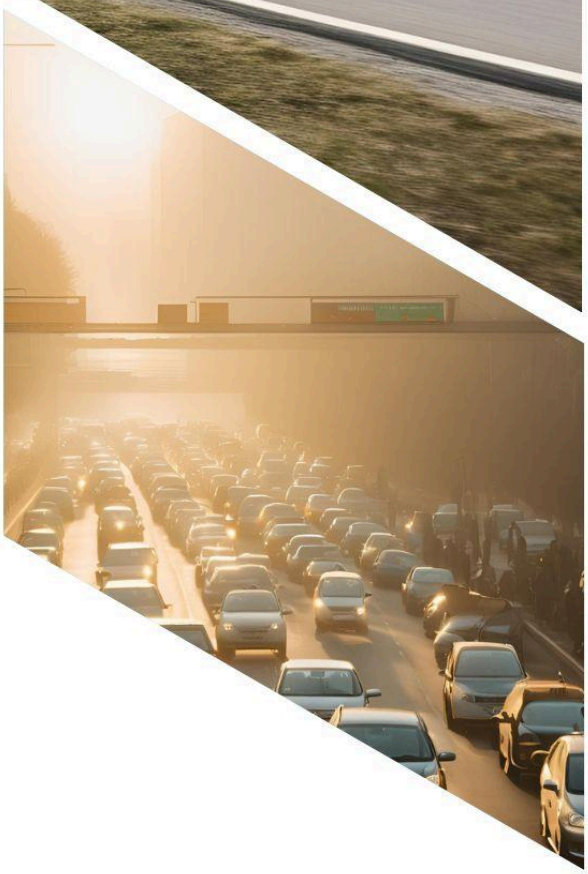


EMISSIONS DE CO2 DES VÉHICULES

RAPPORT FINAL

DATA ANALYST

FORMAT CONTINU
OCTOBRE 2024

ALEXANDRE CLAUDON
GREGORY CODRON
EREL HUARD

DataScientest

Table des matières

Introduction au projet

I. Contexte

Le secteur du transport est aujourd'hui l'un des principaux contributeurs aux émissions de dioxyde de carbone (CO_2), un gaz à effet de serre responsable du réchauffement climatique.

Selon l'Agence européenne pour l'environnement, il serait responsable de près de 25% des émissions mondiales de CO_2 [1] . Face à l'urgence climatique et aux réglementations de plus en plus strictes mises en place par les gouvernements, la réduction des émissions de véhicules est devenue un enjeu majeur pour l'industrie automobile et les politiques publiques.

Dans ce contexte, il est essentiel de comprendre les facteurs influençant ces émissions afin d'identifier les leviers permettant de les réduire. L'analyse des différentes caractéristiques des véhicules permet non seulement de mieux comprendre leur rôle, mais aussi d'anticiper les émissions des futurs modèles de voitures grâce à la modélisation prédictive.

II. Objectifs

L'objectif de notre projet est donc d'étudier les émissions de CO2 des véhicules en s'appuyant sur des jeux de données contenant les caractéristiques techniques de nombreux véhicules. A travers une analyse exploratoire et l'application de modèle de prédiction, nous chercherons à identifier les principaux facteurs influençant les émissions de CO2 et à proposer des pistes pour une mobilité plus durable.

[1] European Environment Agency : https://www.eea.europa.eu/fr/signaux/signaux-2022/articles/il-est-grand-temps-de

Compréhension et manipulation des données

I. Cadre

Nous disposons de deux jeux de données pour atteindre les objectifs de notre projet:

➢ **Émissions de CO_2 et de polluants des véhicules commercialisés en France**

Dernière mise à jour = mars 2014

Ces données sont collectées par l'ADEME 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) et sont disponibles librement sur le site data.gouv.fr. Il comprend pour chaque véhicule des caractéristiques telles que : le type de carburant, la consommation, le poids du véhicule , sa puissance, sa cylindrée et ses émissions de CO_2.

Il est composé de 55 044 lignes et 26 colonnes

➢ **Émissions de CO_2 des voitures particulières en Europe (2014)**

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

Ce jeu de données reprend tous les enregistrements des voitures immatriculées dans l'Union Européenne. Il est mis à disposition par l'Agence Européenne de l'Environnement. Les informations sont enregistrées par chaque Etat membre. Il comprend notamment: le nom du constructeur, les poids des et dimensions des véhicules, la cylindrée, la puissance du moteur, le type de carburant.

Il est composé de 417 938 lignes et 26 colonnes.

II. Pertinence

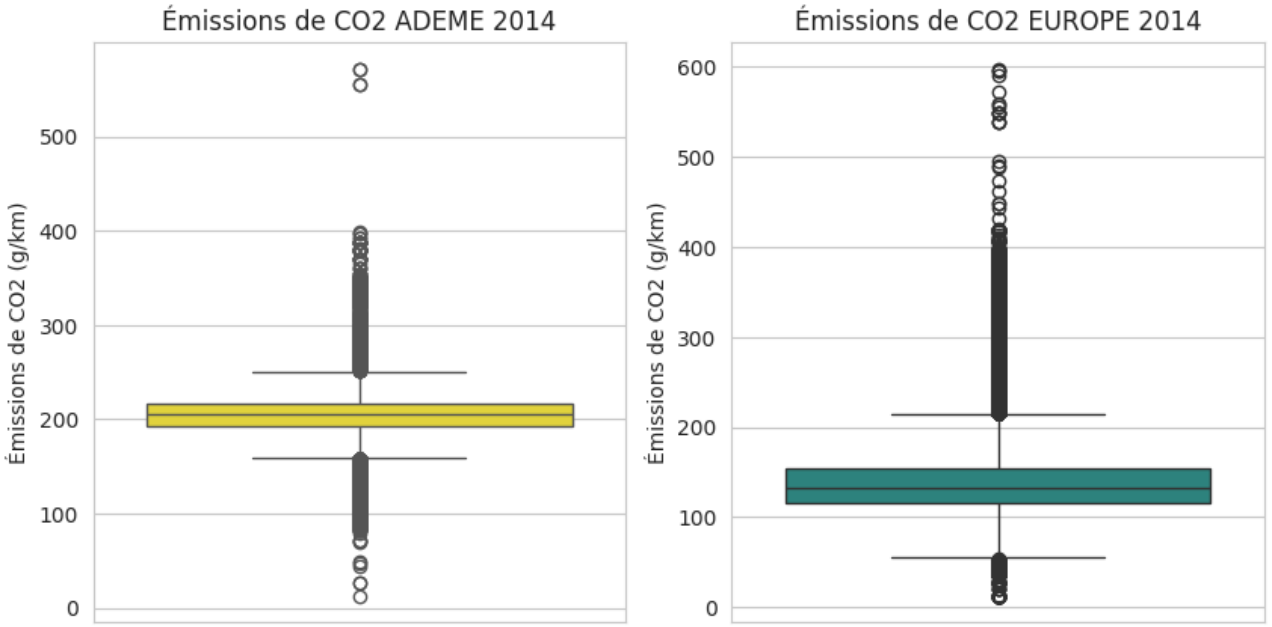
Selon notre objectif, les variables les plus pertinentes sont:

- **La puissance du moteur** : la capacité du moteur à produire de l'énergie. Elle est exprimée en kW ;
- **La cylindrée** : le volume total des cylindres dans le moteur. Elle est exprimée en cm3 ;
- **Le type de carburant** : les principaux types sont essence, diesel, électrique, hybride, GPL (gaz de pétrole liquéfié) et GNV (gaz naturel pour véhicule).

- **La consommation du véhicule :** cette valeur mesure l'efficacité énergétique du véhicule. Elle est exprimée en litres par 100km (l/100km) ;
- **La masse du véhicule** correspond au poids du véhicule à vide. Elle est exprimée en kilos (kg) ;
- **La carrosserie** définit le type de structure et de design du véhicule(Berline, break, coupé, berline,...)
- **Le constructeur :** la marque automobile qui produit le véhicule (ex. Renault, Peugeot, Volkswagen, Mercedes, etc.). Certains constructeurs sont plus avancés en matière de réduction des émissions et d'innovation écologique que d'autres.

Notre variable cible étant les **émissions de CO_2**. Elle est exprimée en gramme par kilomètre (g/km)

A. Distribution de la variable cible



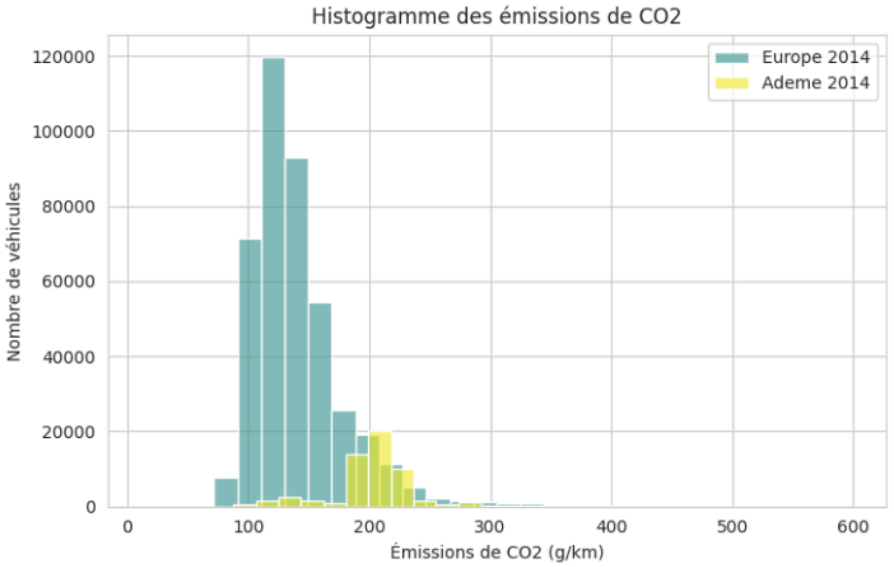
Le graphique ci-dessus montre la distribution de la variable CO_2 sur les deux dataset. On remarque que la distribution de la variable est légèrement différente entre les deux jeux de données (ADEME 2014 et EUROPE 2014).

Dans le jeu de données de l'ADEME, la médiane des émissions de CO_2 est de 205 g/km. La majorité des valeurs sont concentrées dans une plage restreinte autour de cette médiane. On observe plusieurs outliers en dessous de 150 g/km et au dessus de 250 g/km. Ces valeurs extrêmes pourraient représenter des véhicules très économes ou très polluants.

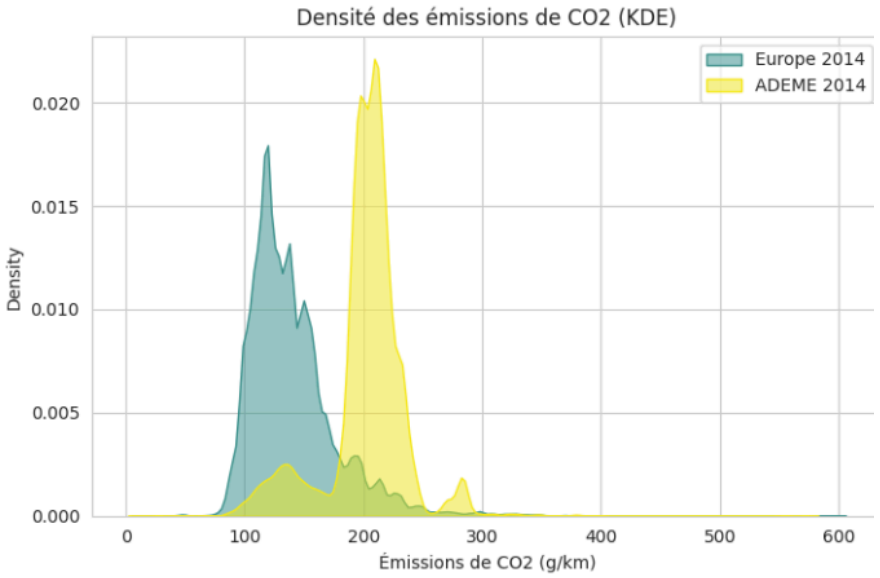
Dans le jeu de données de l'Europe, la médiane des émissions de CO_2 est plus basse et se situe autour de 132 g/km. La boite est également étroite ce qui indique une faible variabilité des valeurs.

Les deux jeux de données présentent des valeurs extrêmes. Celles-ci sont plus nombreuses dans le jeux de données de l'Europe

Pour aller plus loin dans l'analyse de la distribution de notre variable cible, nous allons analyser les deux graphiques ci-dessous.

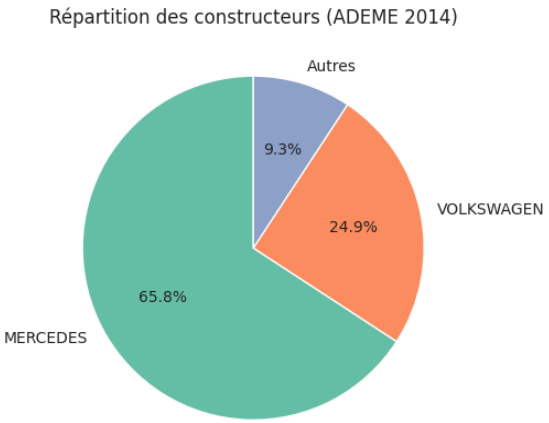


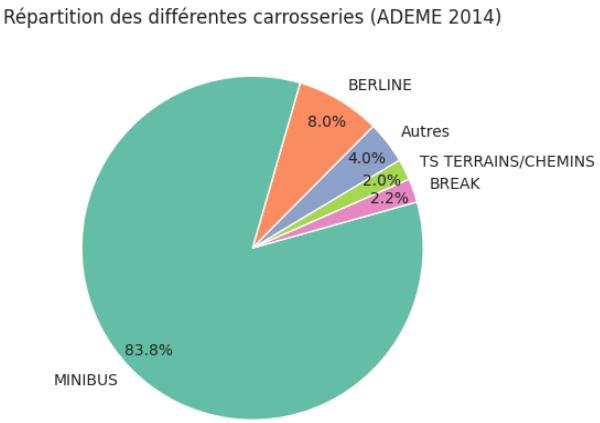
L'histogramme ci-dessus montre la répartition des émissions de CO2 dans les 2 jeux de données Ademe 2014 en jaune et Europe 2014 en bleu-vert. On remarque que globalement les émissions sont plus élevées dans le jeu de données de l'Ademe avec un pic autour de 200 g/km. Les véhicules de cette base de données semblent être en moyenne plus polluants que ceux de la base Europe 2014.



Le graphique ci-dessus représente la densité des émissions de CO_2 pour chaque dataset. Il permet de mieux visualiser la distribution sans être affecté par les tailles d'échantillons différentes. Ces deux courbes confirment ce que nous avons pu observer précédemment. La distribution des émissions de CO_2 dans le jeux de données de l'Ademe est plus concentrée et possède un pic très marqué autour de 200g/km. Cela suggère que les véhicules de cette base sont plus homogènes et ont, en moyenne, des émissions plus élevées.

B. Répartition des entrées par constructeur et carrosserie





- **Répartition des constructeurs**

Lorsque l'on regarde les constructeurs présents dans le jeu de données de l'Ademe, on remarque une très forte représentation des véhicules Mercedes (65.8%) ainsi que Volkswagen(24.9%). Les autres constructeurs sont regroupés dans la catégorie “Autres” et ne représentent que 9,3%.

Cette répartition des constructeurs peut potentiellement influencer les résultats si l’on cherche à généraliser l’analyse à l’ensemble du marché automobile.

- **Répartition des carrosseries**

Les minibus dominent avec 83.8%. Les berlines représentent 8% du total. Les autres carrosseries sont très peu représentées.

Ce déséquilibre peut également influencer les prédictions d'émissions des autres véhicules.

III. Pre-processing et feature engineering

Un prétraitement minutieux des données est essentiel pour garantir la qualité des résultats de notre modèle de prédiction. Voici les principales étapes de transformation appliquées.

A. Sur le dataset Ademe 2014

- **Modification des virgules en points pour les variables numériques**

Dans un environnement Python les virgules ne sont pas reconnues correctement comme séparateurs décimaux, ce qui aurait entraîné des erreurs lors du calcul et de l’analyse. Afin que ces variables soient correctement interprétées nous avons remplacé toutes les virgules par des points.

- **Conversion des variables numériques en float**

Une partie des variables numériques était de type object. Nous les avons converties en type float afin de pouvoir traiter ces données comme des variables continues dans les étapes suivantes du projet.

- **Suppression des lignes contenant des valeurs manquantes pour les émissions de CO_2**

L’objectif de notre projet étant de prédire les émissions de CO_2 des véhicules, cette variable est notre variable cible. Nous avons décidé de supprimer toutes les lignes pour lesquelles les émissions de CO_2 étaient manquantes car elles ne nous permettent pas de créer un

modèle de prédiction fiable. Cela élimine également les véhicules électriques qui n'ont pas d'émissions de CO_2 lors de leur utilisation.

- **Suppression des doublons**

La présence de doublons pourrait biaiser les résultats et entraîner un surapprentissage du modèle. Nous avons donc supprimer toutes les lignes en doubles pour s'assurer que chaque véhicule est représenté qu'une seule fois.

- **Traitement des valeurs manquantes (NaN)**

Le dataset présente plusieurs variables avec des valeurs manquantes. Le choix de la méthode pour traiter ces valeurs dépend de la nature de chaque variable.

  - ➢ Pour les variables liées à la **consommation de carburant** ('conso_urb','conso_exurb','conso_mixte') et aux **émissions de polluants** ('co_typ_1','nox','hcnox','ptcl'), nous avons remplacé les valeurs manquantes par la médiane. Ces variables sont sujettes à des outliers qui peuvent fortement influencer les moyennes. La médiane est moins sensible aux valeurs extrêmes et permet de mieux représenter la distribution des données.
  - ➢ Pour les variables 'masse_ordma_min' et 'masse_ordma_max' , qui représentent **les masses des véhicules**, nous avons remplacé les valeurs manquantes par la moyenne car elles semblent moins susceptibles d'être influencées par des valeurs extrêmes.

- **Suppression des colonnes**

Les colonnes 'hc' et 'date_maj' sont des colonnes qui n'apportent aucune valeur ajoutée à notre analyse et contiennent un trop grand nombre de valeurs manquantes. Nous avons donc décidé de supprimer ces colonnes.

B. Sur le dataset Europe 2014

- **Renommer des types de carburant**

Les types de carburants sont renseignés de plusieurs façon dans le dataset ce qui peut entraîner des incohérences dans l'analyse. Nous avons donc renommé les types de carburants.

- **Suppression des doublons**

Pour garantir une distribution représentative des véhicules, nous avons supprimé les doublons afin de ne conserver qu'une seule occurrence de chaque véhicule.

- **Suppression des colonnes inutiles**

Certaines colonnes contenaient un trop grand nombre de valeurs manquantes, ce qui les rendait inutilisables pour notre analyse. Nous avons donc supprimé les colonnes suivantes: 'z (Wh/km)','IT' et 'Er (g/km)'

- **Suppression des lignes contenant des valeurs manquantes et/ou un 0 pour les émissions de CO_2**

Pour les mêmes raisons que pour le dataset de l'ADEME, nous décidons de supprimer les lignes dont les émissions ne sont pas renseignées. Cela permet également d'écarter les véhicules électriques.

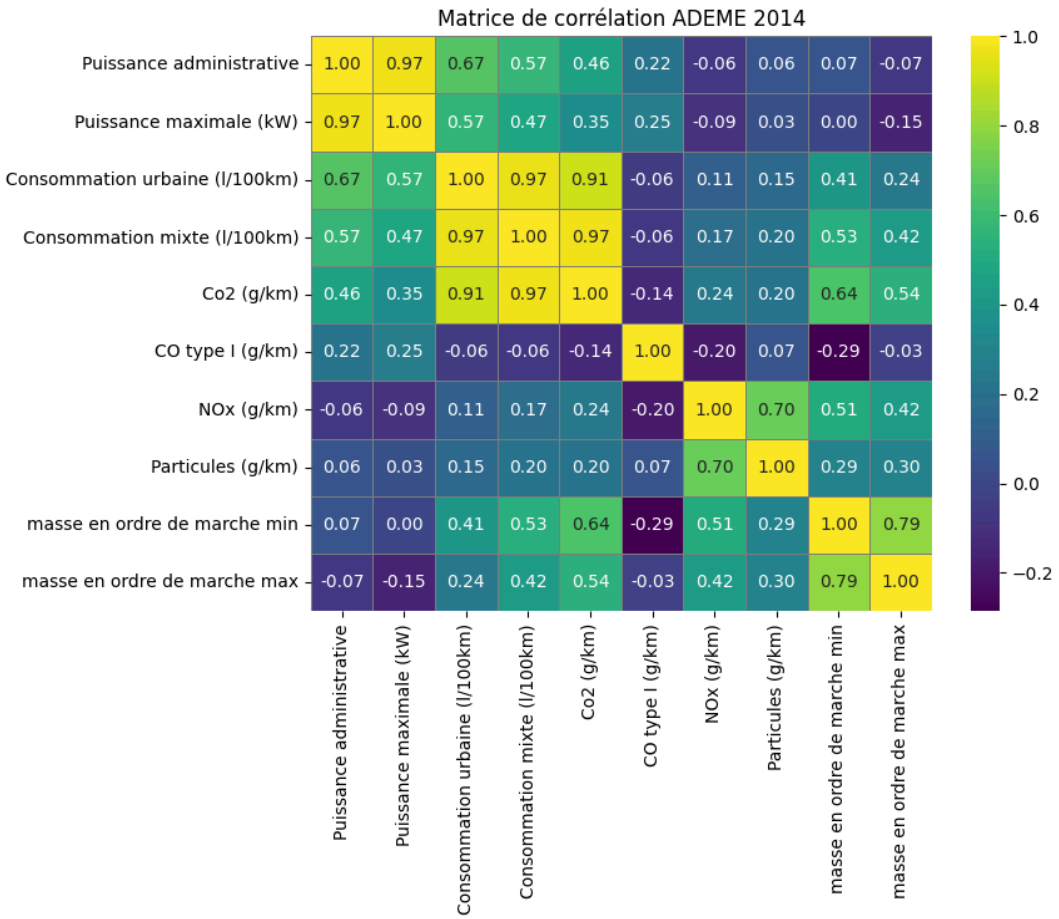
- **Suppression des lignes avec une valeur manquantes pour le type de carburant**

Le type de carburant joue un rôle important dans les émissions de CO_2. Afin de ne pas introduire d'incertitudes dans le modèle, nous avons supprimé les lignes où cette information était manquante.

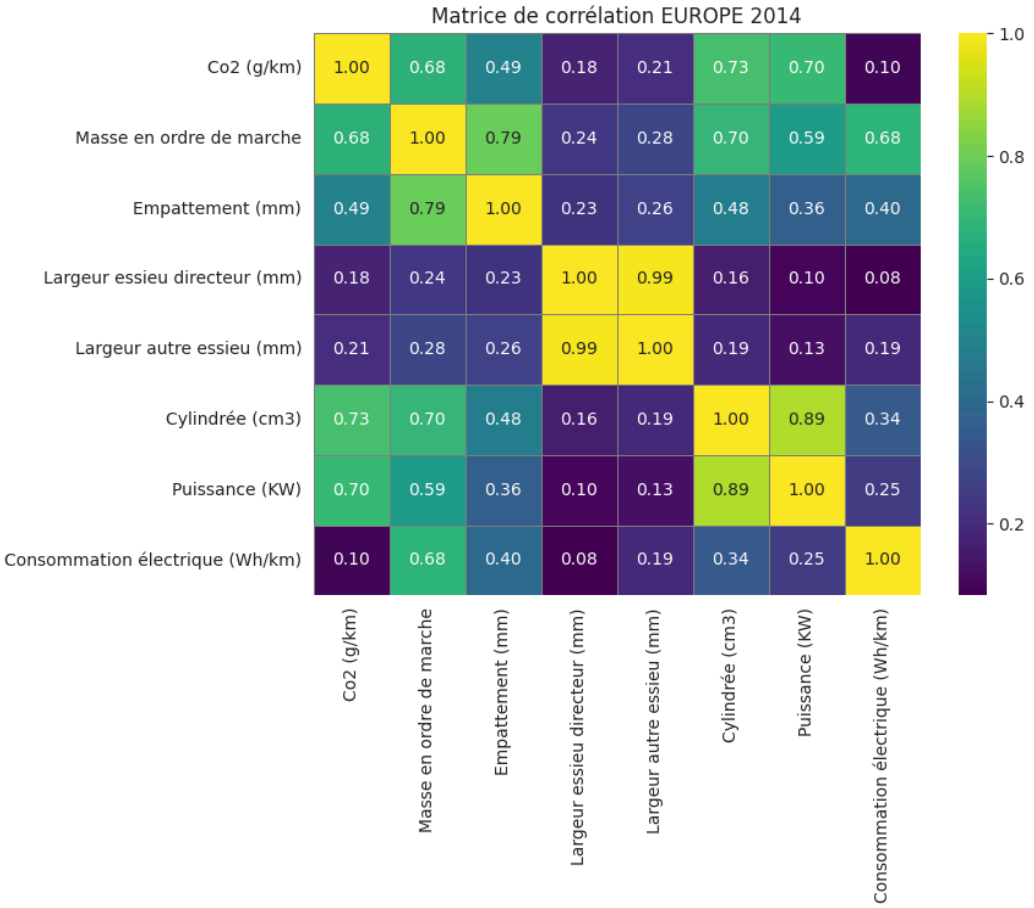
IV. Visualisation et statistiques

A. Heatmap

Afin de visualiser rapidement les relations entre les variables de nos 2 jeux de données, nous avons décidé de créer des heatmap. Celles-ci nous permettent d'identifier quelles variables sont fortement liées ou indépendantes.



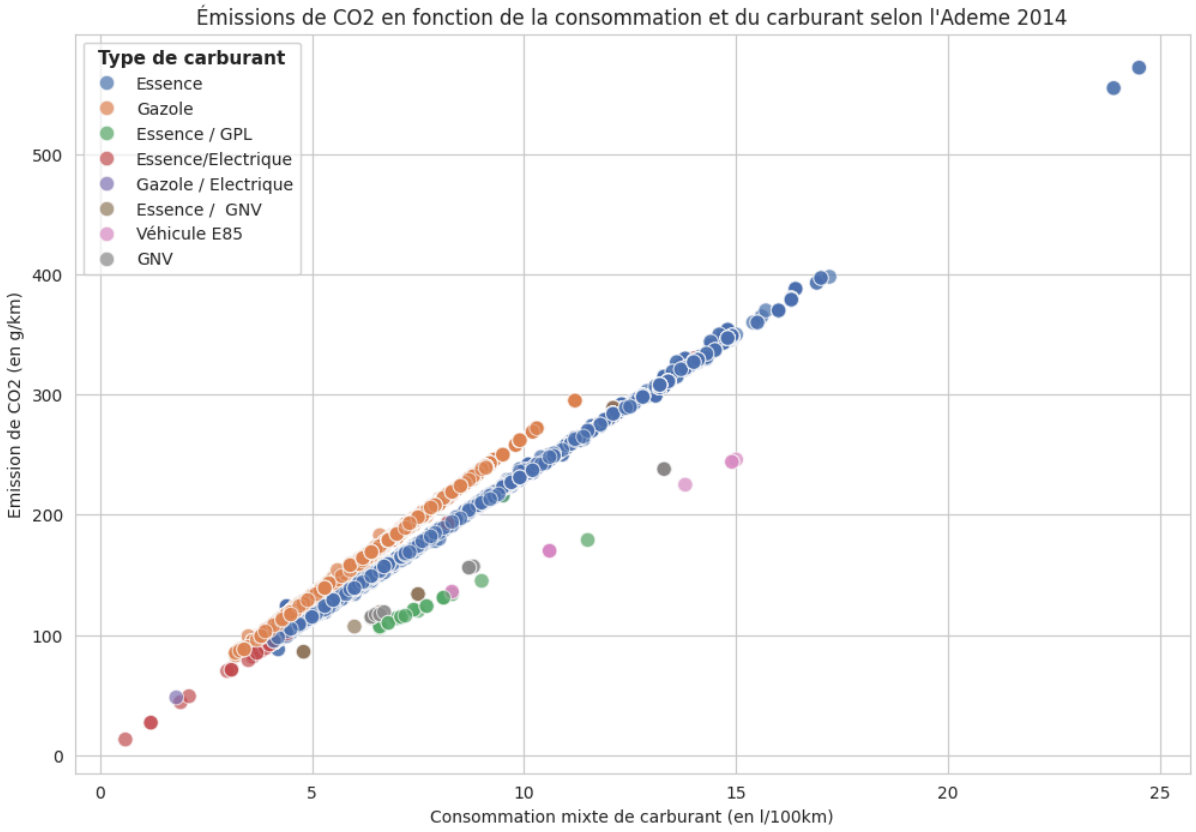
La matrice de corrélation ci-dessus est issue du jeu de données de l'ADEME 2014. On remarque une forte corrélation entre les émissions de CO_2 et la consommation des véhicules (<0.97). Plus un véhicule consomme de carburant, plus celui-ci émet de CO_2. On note également, un lien entre la masse du véhicule et les émissions de CO_2 (entre 0.54 et 0.64). Plus un véhicule est lourd, plus il consomme et donc émet de CO_2.



La heatmap ci-dessus montre le lien entre les données du dataset Europe 2014. Elle nous permet de compléter l'analyse réalisée plus haut avec le dataset de l'Ademe 2014. Ici, nous pouvons observer une corrélation forte des émissions de CO_2 avec la cylindrée (0.73) et la puissance du véhicule (0.70). Plus un moteur est puissant et de grande cylindrée, plus il émet de CO_2. Nous pouvons également observer la même corrélation que dans le jeu de donnée de l'ADEME entre la masse du véhicule et les émissions de CO_2(0.68).

Pour compléter, nous remarquons une corrélation importante entre la masse du véhicule et l'empattement (0.79) et entre la masse et la cylindrée (0.70).

B. Émissions de CO_2 en fonction de la consommation et du type de carburant



Le graphique ci-dessus montre la corrélation entre la consommation mixte de carburant (en l/100km) et les émissions de CO_2 selon le dataset de l'ADEME 2014. Les points sont colorés en fonction du type de carburant.

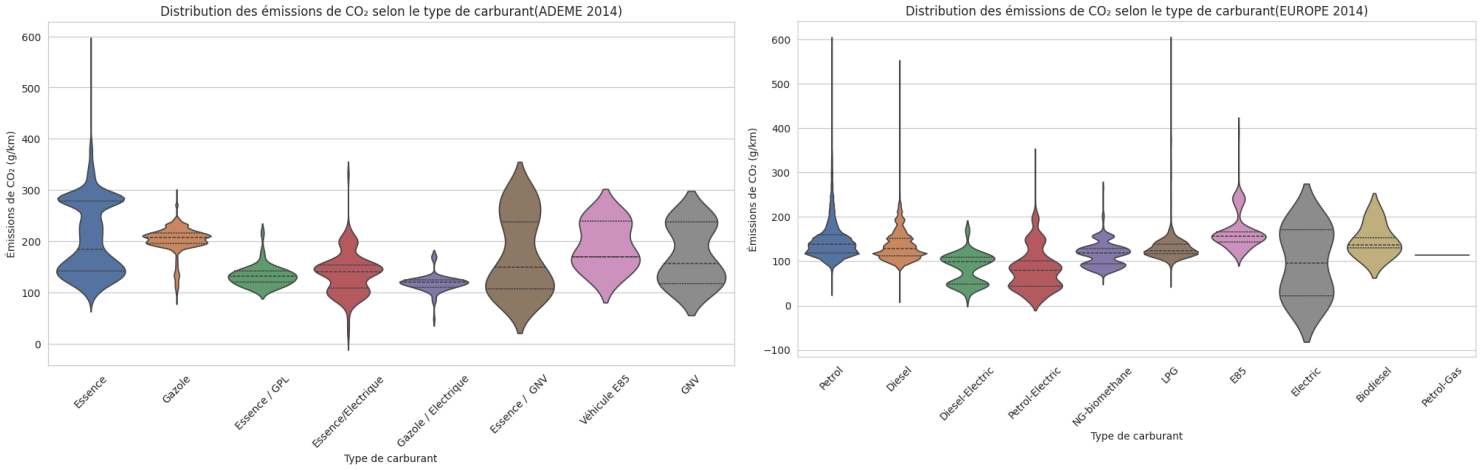
On remarque une **corrélation linéaire forte entre la consommation de carburant et les émissions de CO_2**. Cette corrélation est attendue puisque la combustion du carburant entraîne directement la production de CO_2.

Les véhicules essence semblent émettre moins de CO_2 que les véhicules diesels pour une même consommation.

Le jeu de données de l'EUROPE ne contient pas d'informations sur les consommations des véhicules. Nous ne pouvons pas effectuer de comparaison similaire avec ce deuxième jeu de données.

C. Émissions de CO_2 selon le type de carburant

➢ **Violin plot des émissions de CO_2 selon le type de carburant**



Les deux graphiques ci-dessus montrent la distribution des émissions de CO_2 selon le type de carburant. Celui de gauche est réalisé avec les données du dataset de l'Ademe et celui de droite avec le dataset de l'Agence Européenne de l'environnement. La largeur de chaque "violon" indique la densité des données,ainsi plus la partie est large et plus les valeurs sont concentrées à cet endroit.

Sur les deux graphiques, les tendances semblent similaires même si on remarque que les données semblent plus dispersées sur le jeu de données de l'Agence Européenne que sur celui de l'Ademe L'essence et le diesel sont les carburants les plus représentés.

Essence:

- **Ademe**: Distribution bimodale avec deux pics distincts : autour des 150 g/km et 280 g/km, présence de valeurs extrêmes jusqu'à 600 g/km
- **Europe:** Émissions autour de 150 g/km, présence de valeurs extrêmes jusqu'à 600 g/km

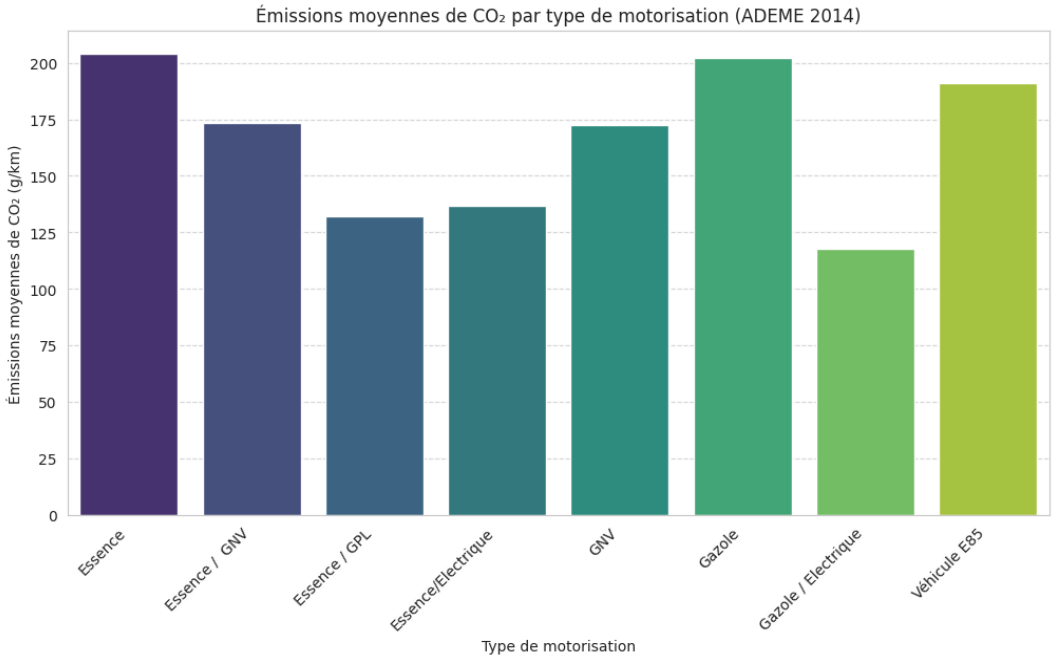
Diesel:

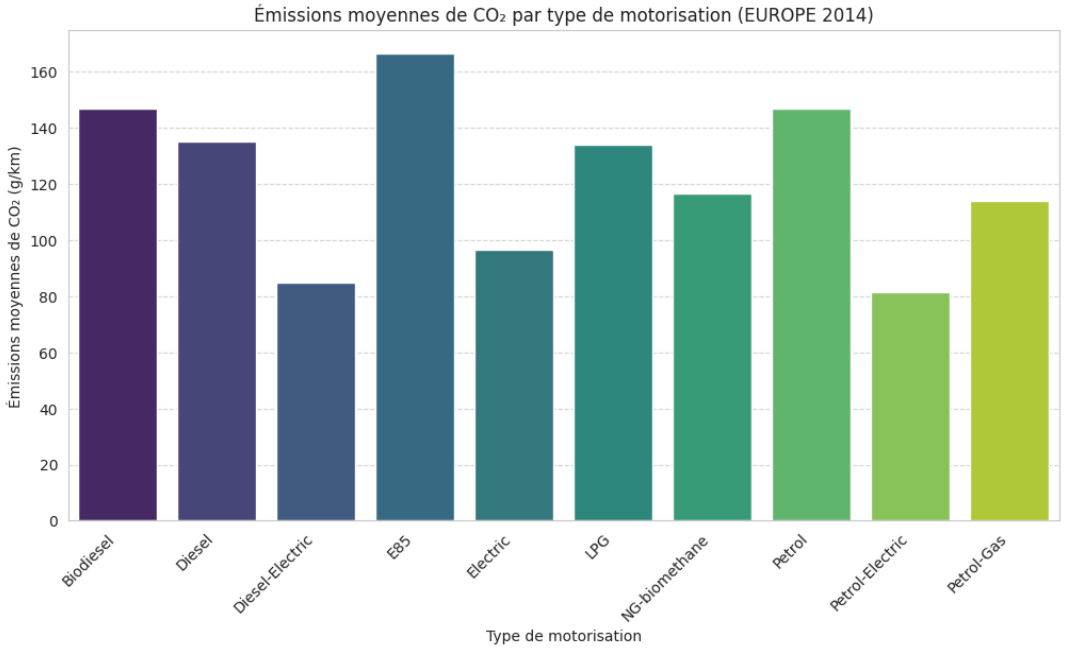
- **Ademe**: Émissions autour de 200 g/km. Moins de dispersion que pour l'essence mais présence également de valeurs élevées (300 g/km)
- **Europe:** Émissions autour de 150 g/km, présence de valeurs extrême jusqu'à 550 g/km

Hybrides et énergies alternatives:

Les véhicules hybrides ont des émissions plus faibles et moins dispersées.

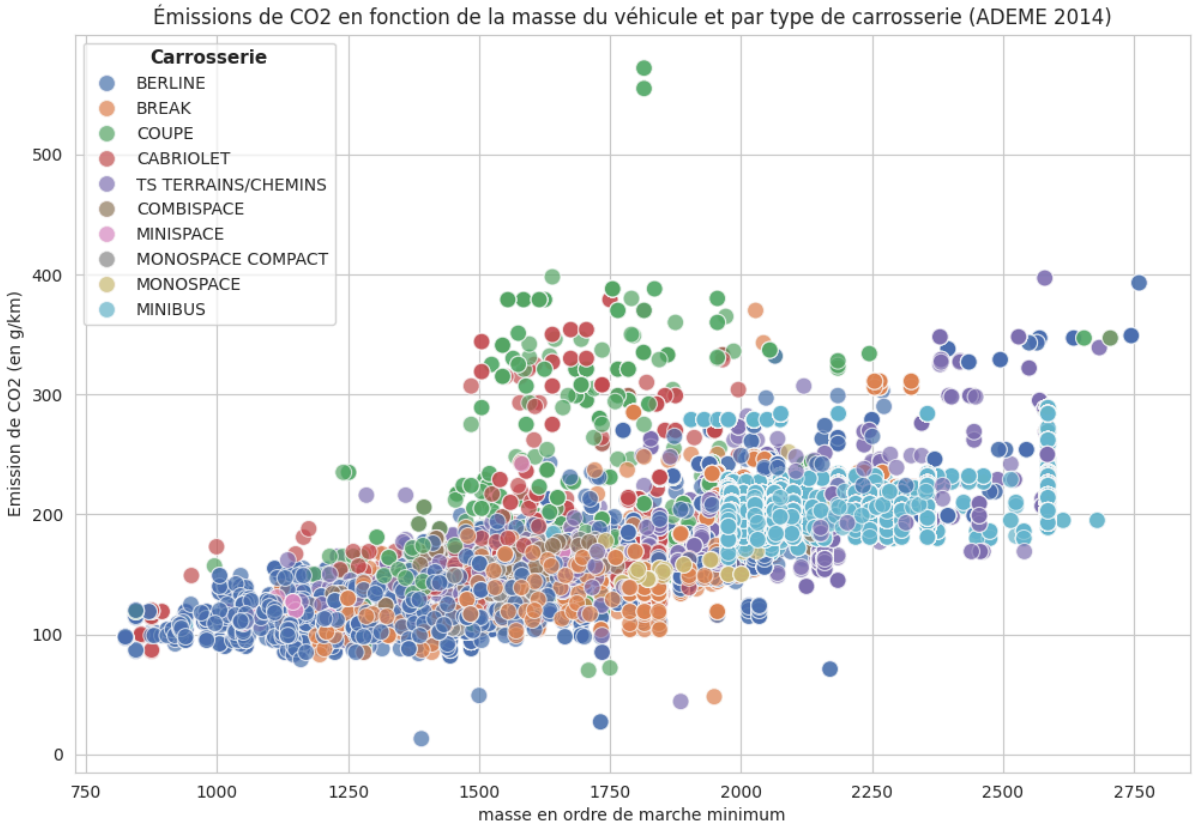
➢ **Barplot des moyennes d'émissions de CO_2 par type de carburant**





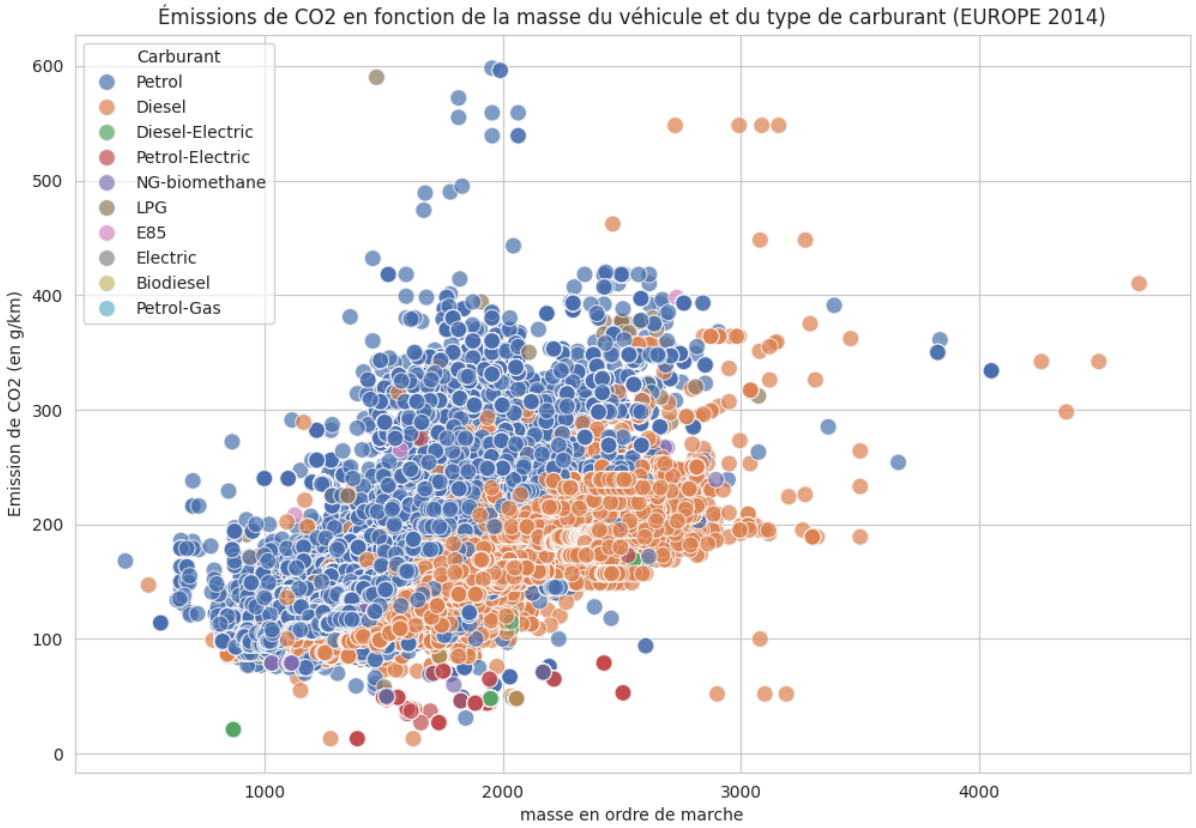
Les graphiques ci-dessous nous montrent la moyenne des émissions pour chaque type de carburant dans chacun des deux dataset analysés. On remarque que l'essence a en moyenne les émissions les plus élevées par rapport aux autres carburants. Les véhicules hybrides ou avec des carburants alternatifs semblent émettre moins de CO_2. Ce qui confirme ce que l'on a pu observer avec les graphiques précédents.

D. Émissions de CO_2 en fonction de la masse du véhicule



La heatmap affichée plus haut indiquait une corrélation entre les émissions de CO_2 et la masse du véhicule. Le graphique ci-dessus confirme que de manière générale plus un véhicule est lourd plus celui-ci émet de CO_2. En effet, un véhicule lourd aura tendance à consommer plus d'énergie pour se déplacer.

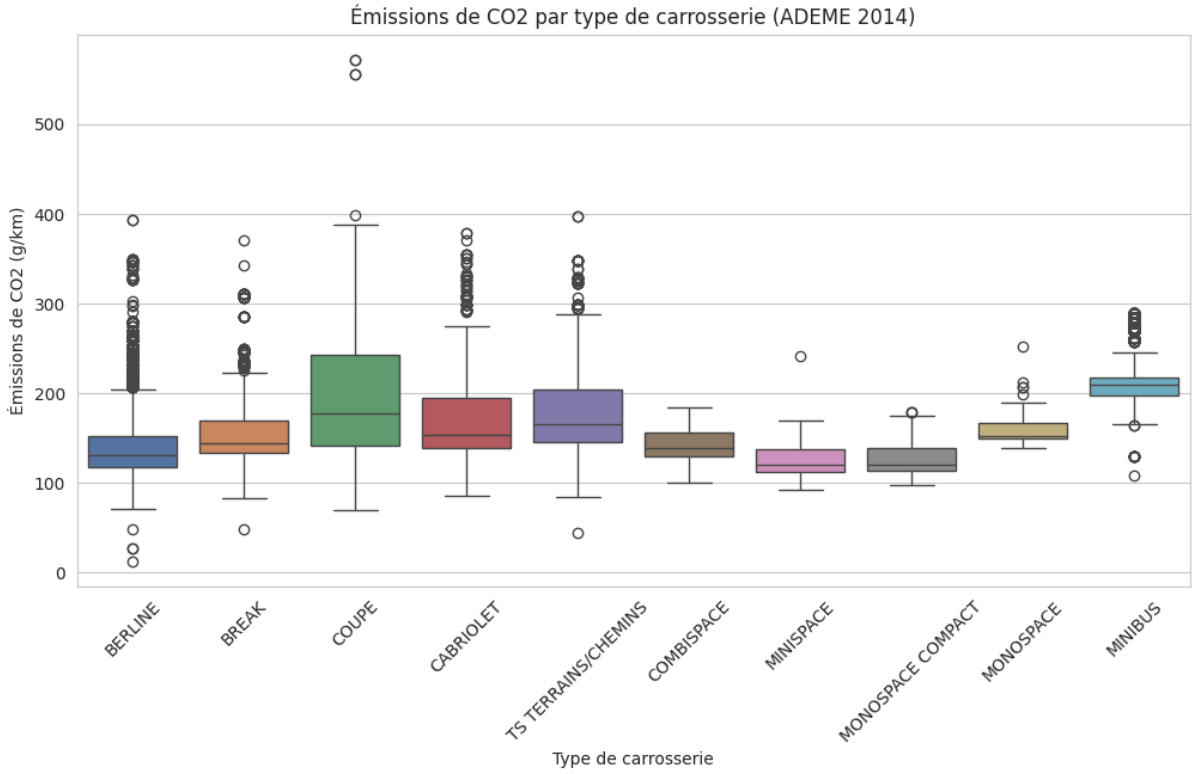
Toutefois, pour une même masse, on observe une variabilité importante des émissions. Cela peut s'expliquer par différents facteurs tels que la motorisation, le type de carburant, l'aérodynamisme…Cette variabilité se remarque notamment pour les véhicules de type coupé et cabriolet. Ces types de véhicules peuvent présenter un large éventail de motorisation allant du cabriolet puissant ou du coupé sportif au modèle plus urbains.



Ce graphique illustre les émissions de CO2 (g/km) en fonction de la masse du véhicule avec un code couleur correspondant au type de carburant. Il est réalisé avec le jeu de données EUROPE 2014.

Nous remarquons la même corrélation entre la masse du véhicule et les émissions que dans le graphique précédent avec le dataset de l'ADEME 2014. La dispersion des points montre qu'à masse égale, les émissions peuvent varier en fonction du type de carburant.

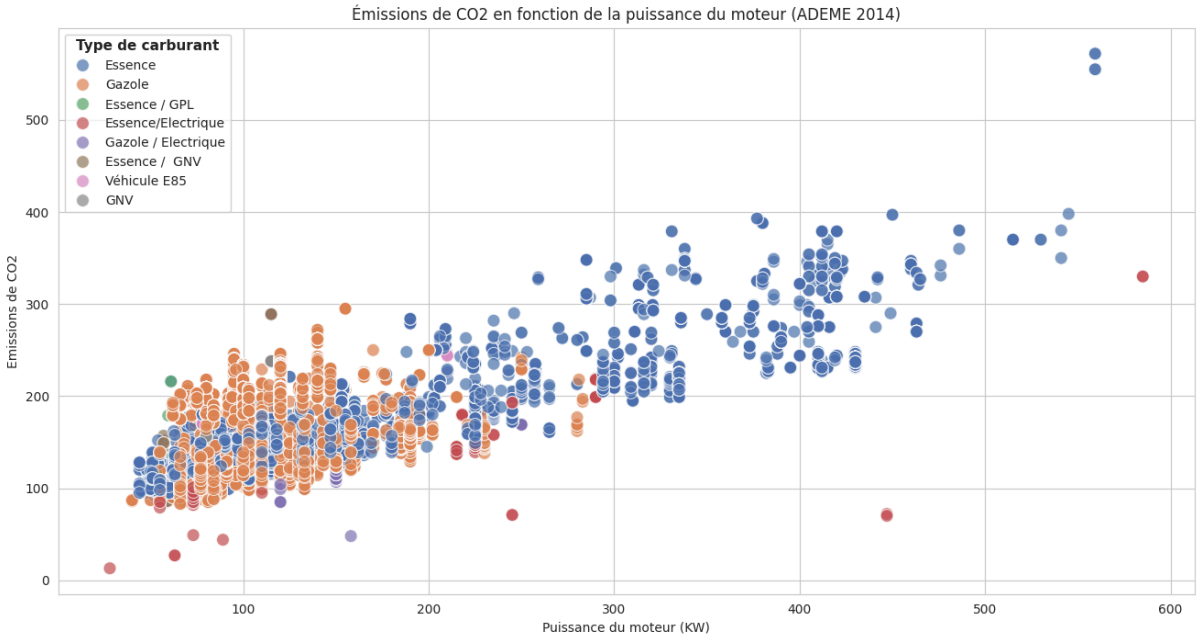
Les véhicules diesel et essence sont très présents sur l'ensemble de la gamme de poids. Le diesel semble légèrement plus bas que l'essence en termes d'émissions pour un même poids. On remarque également que certains véhicule légers ont des émissions importantes et à contrario des véhicules lourds avec des émissions plus faibles existent

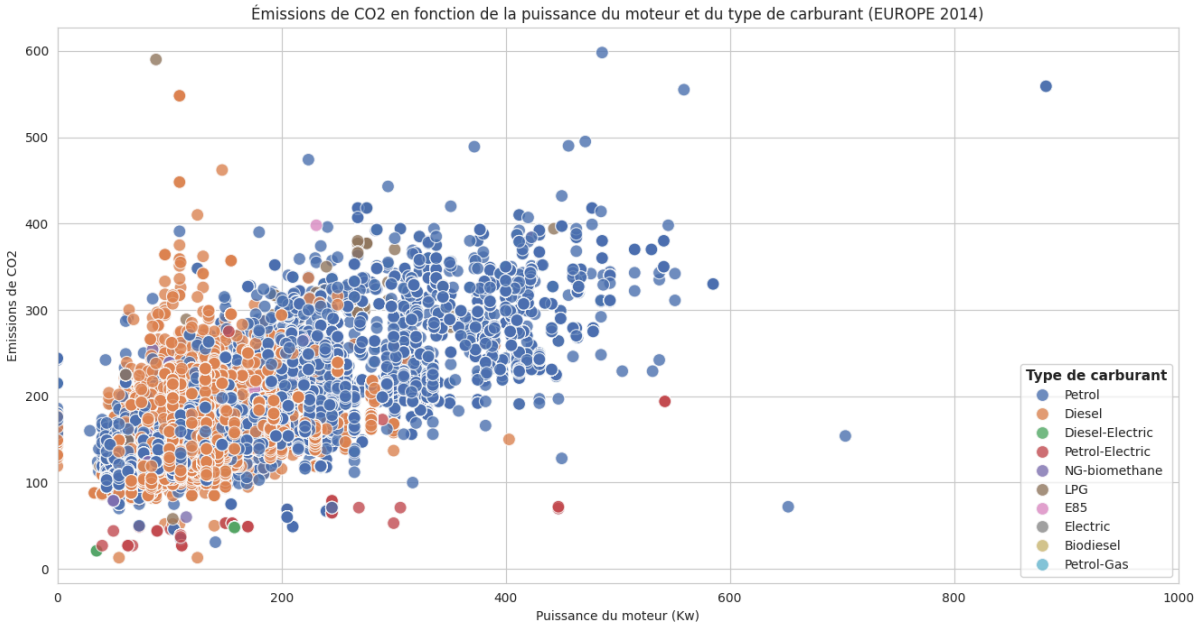


Le boxplot ci-dessus nous permet de voir la distribution des émissions de CO_2 selon le type de carrosserie. Les coupés, les cabriolets et les Ts terrains/chemins ont en moyenne des émissions plus élevées. Les véhicules plus compact tels que les minispace ou monospace compact semblent avoir des émissions plus faibles.

La variabilité des émissions selon la carrosserie suggère que **le modèle de véhicule joue un rôle important dans les émissions de CO_2**.

E. Émissions de CO_2 en fonction de la puissance et de la capacité du moteur

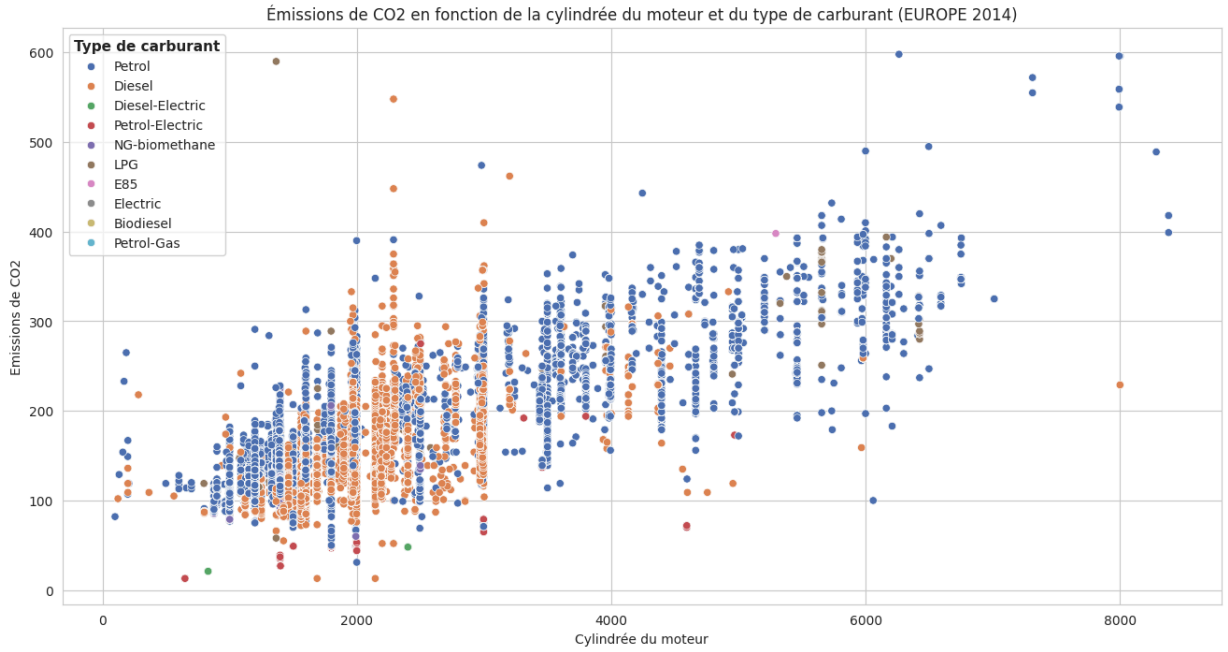




Les deux graphiques ci-dessus illustrent la relation entre les émissions de CO2 et la puissance du moteur avec une distinction par type de carburant. Nous pouvons comparer les deux dataset.

Nous remarquons sur les deux graphiques une corrélation positive entre la puissance du moteur et les émissions de CO_2: **plus la puissance augmente et plus les émissions sont élevées**. Les véhicules essences et diesel dominent en nombre et montrent une large dispersion des émissions.

F. Émissions de CO_2 en fonction de la cylindrée du moteur et du type de carburant



Ce graphique représente les émissions de CO_2 en fonction de la cylindrée du moteur pour différents types de carburants avec le jeu de données Europe 2014.

Globalement, plus la cylindrée du moteur est grande, plus les émissions de CO_2 augmentent. Cependant, on peut observer une dispersion des émissions sur une même cylindrée.

On remarque également que les moteurs diesel (en orange sur le graphique) sont nombreux dans les faibles et moyennes cylindrées. Les moteurs essence sont majoritaires dans les cylindrées plus élevées(>3000 cm3) et montrent des émissions plus élevées.

Modélisation des émissions de CO_2 des véhicules

L'analyse des deux jeux de données, celui de l'Ademe et celui de l'Agence Européenne de l'Environnement a permis de mettre en lumière plusieurs facteurs influençant les émissions de CO_2. Les résultats montrent une forte corrélation entre les émissions de CO_2 et des variables telles que la consommation de carburant, la puissance du moteur, la cylindrée et la masse du véhicule. Ces éléments seront importants pour comprendre et anticiper les émissions des futurs modèles de voitures.

Pour la suite de notre projet, nous allons étudier différents modèles de machine learning afin d'évaluer leurs capacités à effectuer des prédictions sur les émissions de CO_2 des futurs véhicules.

V. Modélisation sur le jeu de données de l'ADEME

A. Sélection des variables explicatives

Le choix des variables explicatives repose sur leur disponibilité en amont de la conception du véhicule, leur pertinence métier et sur leur capacité à prédire les émissions de CO_2.

Nous avons conservé les variables suivantes:

- **Carrosserie :** Cette variable regroupe les différents types de carrosseries des véhicules tels que minibus, break, berline, coupé, cabriolet, monospace… Elle influence l'aérodynamisme, la masse du véhicule et donc les émissions de CO_2
- **Gamme:** Cette variable reflète la position commerciale du véhicule (économique, moyenne-inférieure, moyenne supérieur, luxe…). Ce positionnement peut être associé à une différence de puissance ou de motorisation ce qui a un impact significatif sur les émissions de CO_2.
- **Type de carburan**t (*'cod_cbr'*) : le type de carburant a une influence directe sur le niveau d'émission de CO2 d'un véhicule.
- **Marque** (*'lib_mrq'*): cette variable permet de voir les différences de stratégie technologique entre les constructeurs.
- **Puissance administrative** (*'puiss_admin_98'*) et **puissance maximale** (en kW) (*'puiss_max'*): Ces deux variables donnent une estimation des performances du véhicule. une puissance élevée donne souvent des émissions de CO_2 importantes.
- **Masse en ordre de marche min** (*'masse_ordma_min*') et **masse en ordre de marche max** (*'masse_ordma_max'*): Il y a une forte corrélation entre le poids du véhicule et la consommation d'énergie nécessaire pour son utilisation.
- **Hybride:** Cette variable binaire indique si le véhicule possède une motorisation hybride ou non. Les véhicules hybrides ont généralement des émissions de CO_2 plus faibles.

Nous avons fait le choix d'écarter les variables liées à la consommation du véhicule car nous avons pu observer précédemment qu'elles étaient fortement corrélées avec notre variable cible (CO_2). Les inclure dans notre modèle pourrait introduire un biais important en faussant l'apprentissage du modèle. De plus, ces informations ne sont pas connues avant la conception du véhicule or notre objectif est de prédire les émissions futures à partir des caractéristiques disponibles en amont de la conception.

B. Préparation des données pour la modélisation

Pour pouvoir réaliser notre modélisation, il nous a fallu préparer en amont nos données.

- **Préparation et séparation du jeu de données**

Après avoir défini notre variable cible comme étant CO_2, nous avons créées trois types de variables:

- Variables catégorielles classiques: (Carrosserie, gamme, cod_cbr, lib_mrq)
- Une variable binaire spécifique: (hybride)
- des variables numériques (puiss_admin_98, puiss_max, masse_ordma_min, masse_ordma_max)

Nous avons ensuite séparé notre jeu de données en deux sous-ensembles: 80% pour l'entraînement et 20% pour le test.

- **Encodage et normalisation**

Les variables catégorielles ont été encodées via **OneHotEncoder** qui permet de convertir chaque modalité en une colonne binaire indépendante.

Les variables numériques ont été traitées en deux étapes:

- Imputation des valeurs manquantes à l'aide de la moyenne via **SimpleImputer** afin d'éviter la perte d'information.
- Standardisation avec **StandardScaler** afin d'harmoniser les échelles de valeurs entre les différentes variables.

- **Mise en place d'une pipeline**

Afin de bien structurer notre projet, nous avons décidé de créer un **pipeline** avec scikit-learn. Cette approche présente plusieurs avantages:

- Cela nous permet de gagner en lisibilité en regroupant toutes les étapes de traitement en un seul objet

- Nous évitons les fuites de données en s’assurant que le prétraitement soit uniquement appliqué sur les données d'entraînement
- Nous pouvons tester plus facilement différents modèles sans avoir à réécrire tout notre code.

C. Résultats des modélisations

Nous avons testé plusieurs modèles sur le jeu de données de l’ADEME, nous décidons de nous concentrer, ici sur les 4 plus performants. Le critère principal de sélection a été le score sur le jeu de test complété par des mesures d’erreurs tel que le RMSE (Root Mean Squared error) ou le MAE (Mean Absolute Error) ainsi que la différence entre le score sur le jeu d'entraînement et le score sur le jeu de test pour s’assurer que le modèle se généralise bien.

Nous avons exploré les modèles suivant:

- Linear Regression,
- Ridge,
- Lasso,
- ElasticNet,
- KNeighbors Regressor,
- Decision Tree Regressor,
- Random Forest Regressor,
- Gradient Boosting Regressor.

	Score train	Score test	MSE	RMSE	MAE	diff train test
LinearRegression	82,9 %	83,3 %	198,16	14,077	10,805	0,4 %
LinearRegression	74,7 %	75,2 %	293,35	17,127	12,397	0,5 %
RandomForestRegressor(random_state=42)	95,4 %	94,5 %	64,49	8,031	5,558	-0,9 %
Ridge(alpha=1,0)	82,9 %	83,2 %	198,18	14,078	10,806	0,3 %
Lasso(alpha=0,1)	80,6 %	81,0 %	224,93	14,998	11,305	0,4 %
ElasticNet(alpha=0,1, l1_ratio=0,5)	77,3 %	77,7 %	263,53	16,234	12,164	0,4 %
KneighborsRegressor(n_neighbors=3)	93,2 %	92,5 %	88,99	9,433	6,27	-0,7 %
KneighborsRegressor(n_neighbors=4)	93,3 %	92,3 %	91,4	9,56	6,209	-1,0 %
KneighborsRegressor(n_neighbors=5)	93,3 %	92,3 %	91,06	9,543	6,229	-1,0 %
DecisionTreeRegressor(random_state=42)	95,5 %	94,4 %	65,78	8,111	5,565	-1,1 %
GradientBoostingRegressor(n_estimators=100, learning_rate=0,1, random_state=42)	91,5 %	91,7 %	97,72	9,886	7,471	0,2 %
GradientBoostingRegressor(n_estimators=500, learning_rate=0,1, random_state=42,max_depth=4)	94,8 %	94,4 %	66,12	8,131	6,099	-0,4 %
500, learning_rate = 0,2, max_depth = 7, min_samples_leaf = 1, min_samples_split = 5, subsample = 0,8, random_state=42)	95,4 %	94,7 %	62,53	7,908	5,575	-0,7 %

Résultats des différentes modélisation réalisées sur le jeu de l'ADEME

Nous avons décidé de nous concentrer sur les 4 modèles les plus performant à savoir:

1. Random Forest Regressor,
2. Decision Tree Regressor,
3. Kneighbors Regressor et
4. Gradient Boosting Regressor.

Le modèle Decision Tree Regressor est un modèle d'arbre de décision. Ce modèle est simple et rapide à entraîner.

Ce modèle obtient de très bon résultats avec un score test élevé. La différence entre le score test et train est faible ce qui indique que le modèle semble bien généraliser.

- R^2= 0.944
- RMSE = 8.11
- MAE=5.57
- Différence train/test = -1.1%

2. Random Forest Regressor

Le Random Forest Regressor est un ensemble d'arbres de décision. Il permet une prédiction plus robuste que le Decision Tree Regressor avec moins de risque de surapprentissage. Ici, la différence entre le score test et le score train est plus faible qu'avec le modèle précédent ce qui montre une bonne généralisation.

- R^2= 0.945
- RMSE= 8.03
- MAE=5.56
- Différence train/test =-0.9%

3. KNeighborsRegressor

Le modèle des K plus proches voisins prédit la valeur cible en prenant la moyenne des k véhicules les plus similaires.

Les scores obtenus sur le jeu de test sont relativement bons et on remarque que le modèle généralise bien. Nous remarquons toutefois, qu'avec l'augmentation de n_neighbors, les scores ne s'améliorent pas vraiment et les temps de calcul s'allongent, ce qui limite son usage pour les grands jeux de données.

n_neighbors	R^2	RMSE	MAE	Différence train/test
3	0.932	9.433	6.27	-0.7%
4	0.933	9.56	6.209	-1.0%
5	0.933	9.54	6.229	-1.0%

4. Gradient Boosting Regressor

Ce modèle construit les arbres de manière séquentielle. Chaque nouvel arbre essaie de corriger les erreurs du précédent. Ce modèle est plus complexe à mettre en place car il y a beaucoup d'hyper-paramètres à régler. Il est également plus lent à entraîner mais il montre de très bonnes performances.

Nous avons cherché à optimiser les performances de notre modèle en testant plusieurs combinaisons de paramètres en utilisant **GridSearchCV.** . Le tableau ci-dessous montre les résultats obtenus:

N° modèle	Paramètres	R^2	RMSE	MAE	Différence train/test
1	n_estimators=100, learning_rate=0,1, random_state=42	0.917	9.88	7.471	0.2%
2	n_estimators=500, learning_rate=0,1, random_state=42, max_depth=4	0.944	8.131	6.099	-0.4%
3	n_estimators=500, learning_rate = 0,2, max_depth = 7, min_samples_leaf = 1, min_samples_split = 5, subsample = 0,8, random_state=42	0.947	7.908	5.575	-0.7%

Le modèle 1 offre déjà de très bonnes performances, mais en augmentant le nombre d'estimateurs et en limitant la profondeur des arbres dans le modèle 2 on observe une nette amélioration des performances. Le score sur le jeu de test est plus élevé (0.944 contre 0.917) et la diminution du RMSE et du MAE montre une meilleure précision des prédictions.

Enfin, le modèle 3 est plus complexe mais a de meilleures performances au global. Le score sur le jeu de test est très proche de celui du modèle 2 mais le RMSE et le MAE sont encore plus bas.

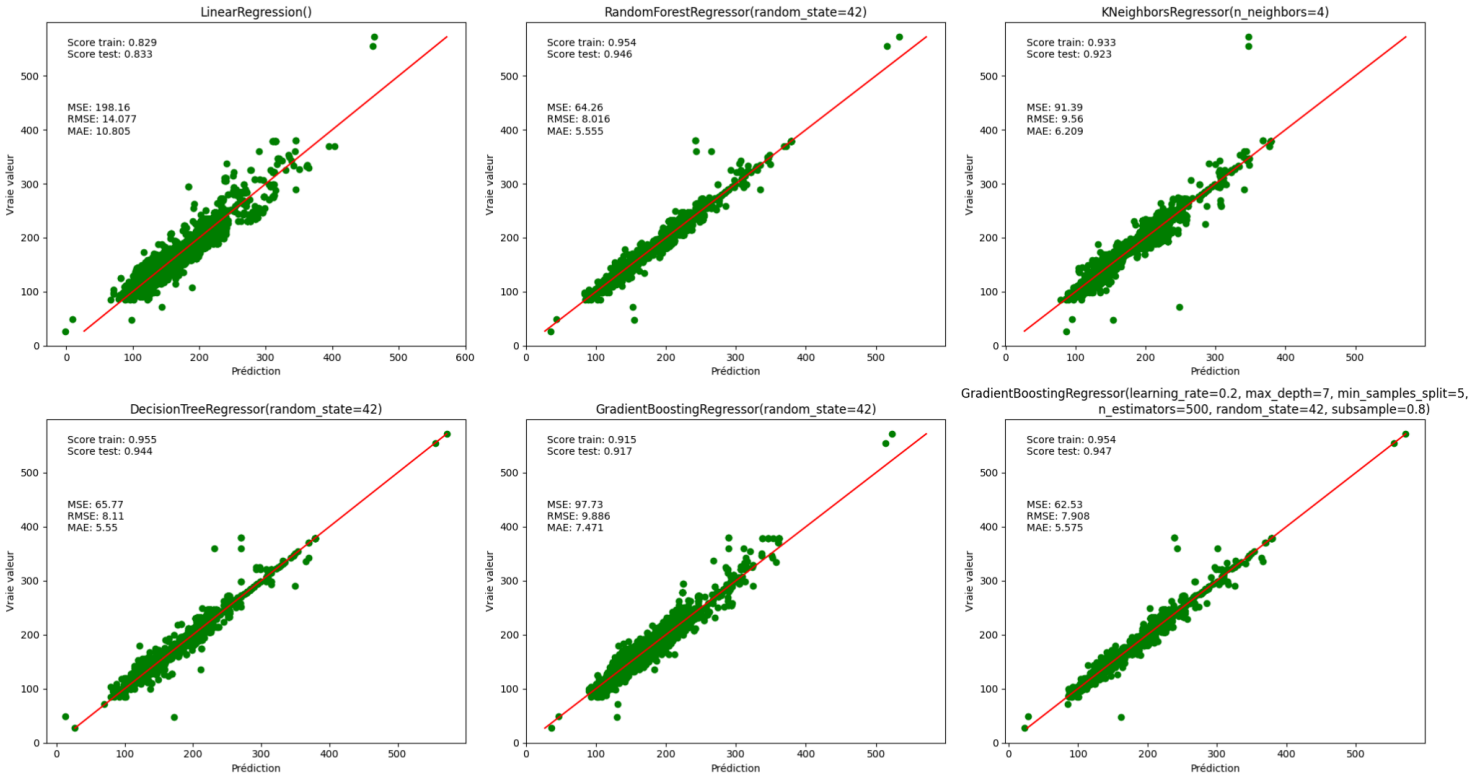
On remarque ainsi que l'ajout de paramètres à notre modèle le rend plus performant mais augmente également la différence entre le jeu d'entraînement et le jeu de test.

Le modèle 3 du GradientBoostingRegressor est le meilleur modèle qu'on ait pu tester sur notre jeu de données.

II. Modélisation sur le jeu de données Europe

Malgré nos essais nous n'avons pas réussi à fusionner le jeu de données mis à disposition par l'Agence Européenne de l'Environnement avec celui de l'ADEME en raison

Comparaison des modèles de régression pour ademe2014



principalement d'un manque de variables communes suffisantes pour effectuer une jointure cohérente.

Dans un souci de comparaison et d'enrichissement des perspectives nous avons tout de même souhaité exploiter ce deuxième jeu de données. Celui-ci présente un intérêt particulier car il contient des variables techniques qui ne sont pas présentes dans le jeu de l'ADEME telles que l'empattement ou la largeur des essieux qui pourraient également influencer les émissions de CO_2.

A. Sélection des variables explicatives

Le choix des variables repose sur les mêmes conditions que pour le jeu de données de l'Ademe, à savoir leur disponibilité en amont de la conception, leur pertinence métier et leur capacité à prédire les émissions de CO_2.

Pour ce jeu de données nous avons donc également conservé les variables suivantes: **marque** (Mk), **Type de carburant** (Ft), **Masse en ordre de marche** (m (kg)), **Puissance du moteur** (ep (KW)) auxquelles nous avons ajoutons les variables ci-dessous qui ne sont pas présentes dans le jeu de données de l'ADEME:

- **Empattement** (w (mm)): correspond à la distance entre les essieux avant et arrière du véhicule. c'est un bon indicateur de la taille globale du véhicule.
- **Largeur de l'essieu directeu**r (at1 (mm)) et **largeur d'un autre essieu** ('at2 (mm)): Ces variables indiquent la largeur entre les roues d'un même essieu. Elles permettent d'avoir des informations supplémentaires sur le gabarit du véhicule.
- **Cylindrée** (ec (cm3)):Cette variable représente le volume total des cylindres du moteur. La cylindrée a un impact direct sur la puissance du moteur.

Nous avons fait le choix d'écarter la variable **Type** ('T'). Le type est un code alphanumérique regroupant tous les véhicules ayant les mêmes caractéristiques. Il s'agit donc d'un code standardisé qui englobe certaines informations techniques. Cette variable ne peut pas être utilisée de manière fiable dans notre modélisation car elle ne permet pas une généralisation. Le type peut être spécifique à un contexte donné (une année, un marché…) et il n'existe pas pour les futurs véhicules qui ne sont pas encore commercialisés. De plus, l'intégration de cette variable peut masquer l'effet réel des autres variables explicatives.

B. Préparation des données pour la modélisation

La préparation des données et la modélisation est similaire à ce que nous avons pu faire sur le jeu de données de l'ADEME. Nous avons également structuré notre modélisation à l'aide d'une pipeline.

- **Préparation et séparation du jeu de données**

Nous avons défini 2 types de variables:

- Variables catégorielles (Mk, Ft);
- Variables numériques ('m (kg)','w (mm)','at1 (mm)','at2 (mm)','ec (cm3)','ep (KW)')

Nous avons ensuite séparé notre jeu de données en deux sous-ensembles: 80% pour l'entraînement et 20% pour le test.

- **Encodage et normalisation**

Les variables catégorielles ont été encodées via **OneHotEncoder** et les variables numériques ont été imputées à l'aide de la moyenne avec **SimpleImputer** et normalisées avec **StandardScaler.**

C. Résultats des modélisations

Dans cette partie nous présentons les résultats des modélisations réalisées sur le jeu de données de l'Agence Européenne de l'Environnement. Afin de pouvoir comparer nos résultats avec le jeu de l'ADEME nous nous concentrons sur les mêmes modèles.

Les différents modèles ont été évalués sur l'ensemble de test à l'aide de plusieurs métriques, le score R^2, le RMSE et le MAE.

1. Decision Tree Regressor

- R^2=0.963
- RMSE =7.546
- MAE=2.396
- Différence train/test= -2.7%

Le Decision Tree Regressor montre un très bon résultat. 92,6% de la variance des émissions de CO^2 est expliquée par le modèle. Les erreurs sont relativement faibles et montrent que les prédictions sont très proches des vraies valeurs. Le modèle ne semble pas souffrir d'un surapprentissage important et généralise bien.

2. Random Forest Regressor

III. R^2= 0.972
IV. RMSE = 6.325
V. MAE= 2.39
VI. Différence train/test= -1.5%

Ce modèle est celui qui montre les meilleures performances globales sur le jeu de données de l'Europe. 97.1% de la variance des émissions de CO_2 est expliquée par le modèle. Les erreurs sont également faibles. La différence entre les performances d'entraînement et de test montre que le modèle généralise très bien. Le Random Forest est plus précis que le Decision Tree.

3. KNeighborsRegressor

n_neighbors	R^2	RMSE	MAE	Différence train/test
3	0.963	7.302	2.851	-1.3%

Le KNeighbors Regressor est légèrement moins performant que les deux modélisations précédentes mais montre tout de même de bons résultats et une bonne capacité de généralisation.

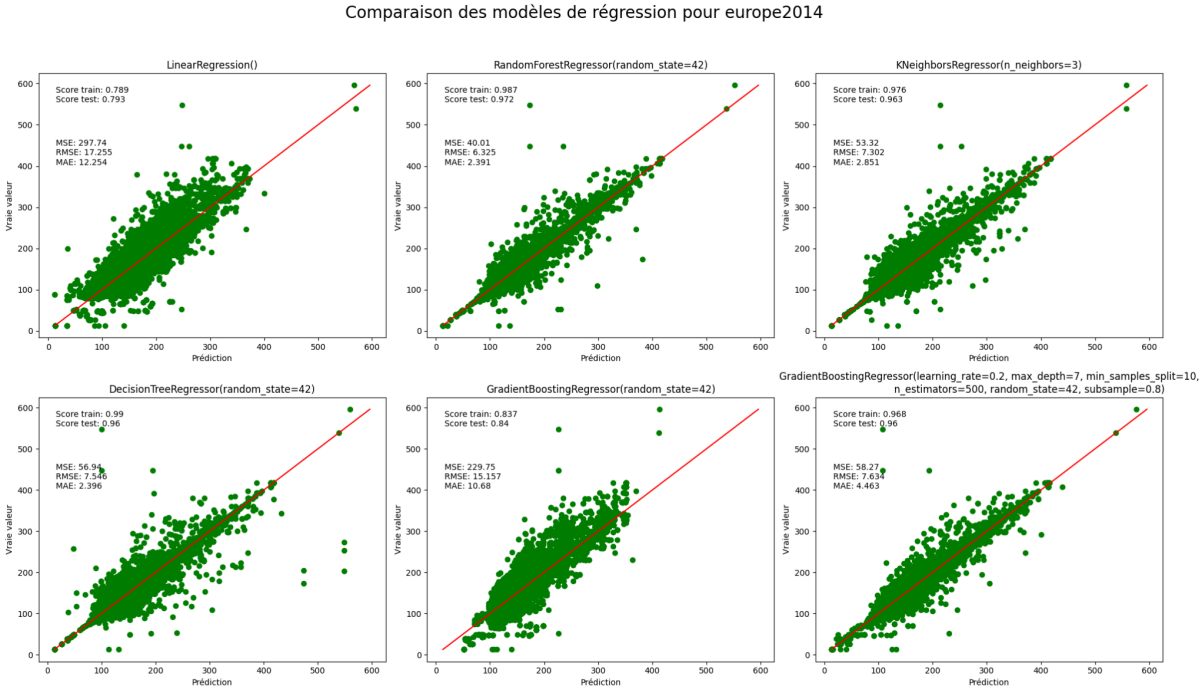
Cependant, son temps de calcul est assez important. Nous n'avons pas pu essayer avec plus 3 voisins, le modèle devant recalculer les distances pour chaque observation, augmenter ce nombre aurait rendu le temps de calcul trop long. Ce temps de calcul élevé le rend moins adapté pour une application à grande échelle.

4. Gradient Boosting Regressor

Pour le Gradient Boosting Regressor, nous avons testé les mêmes modèles que pour l'ADEME. Les résultats obtenus sont présentés dans le tableau ci-dessous:

N° modèle	Paramètres	R^2	RMSE	MAE	Différence train/test
1	n_estimators=100, learning_rate=0,1, random_state=42	0.837	15.157	10.68	0.3%
2	n_estimators=500, learning_rate=0,1, random_state=42, max_depth=4	0.917	10.957	7.341	0.1%
3	n_estimators=500, learning_rate = 0,2, max_depth = 7, min_samples_leaf = 1, min_samples_split = 5, subsample = 0,8, random_state=42	0.96	7.634	4.463	- 0.8%

On remarque, qu'en augmentant le nombre de paramètres, les résultats s'améliorent. Ainsi, le modèle n°3 fournit les meilleures performances. Cependant le temps de calcul pour Gradient Boosting Regressor est très long avec le jeu de données de l'Europe pouvant rendre son application à grande échelle plus compliqué que le Random Forest Regressor.



III. Comparaison des résultats sur chaque jeu de données

A. Performance des modèles sur chaque dataset

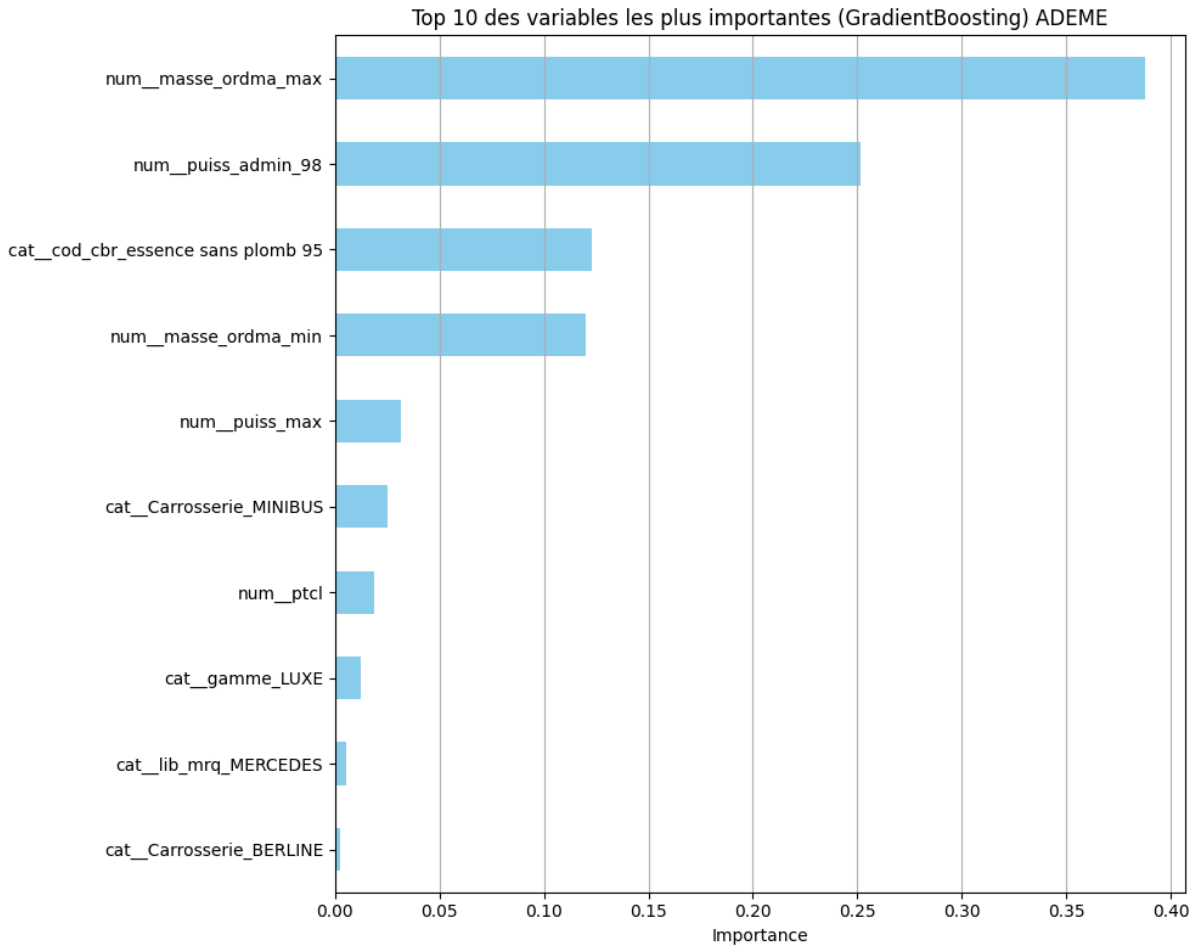
Le tableau ci-dessous nous permet d'avoir une vision d'ensemble des performances des différents modèles sur nos deux jeux de données.

Modèle	R^2 ADEME	RMSE ADEME	MAE ADEME	R^2Europe	RMSE Europe	MAE Europe
Decision Tree Regressor	0.944	8.11	5.57	0.963	7.546	2.396
Random Forest Regressor	0.945	8.03	5.56	0.972	6.325	2.391
Kneigbhors Regressor (k=3)	0.932	9.433	6.27	0.963	7.302	2.85
Gradient Boosting Regressor (n°3)	0.947	7.908	5.575	0.96	7.634	4.463

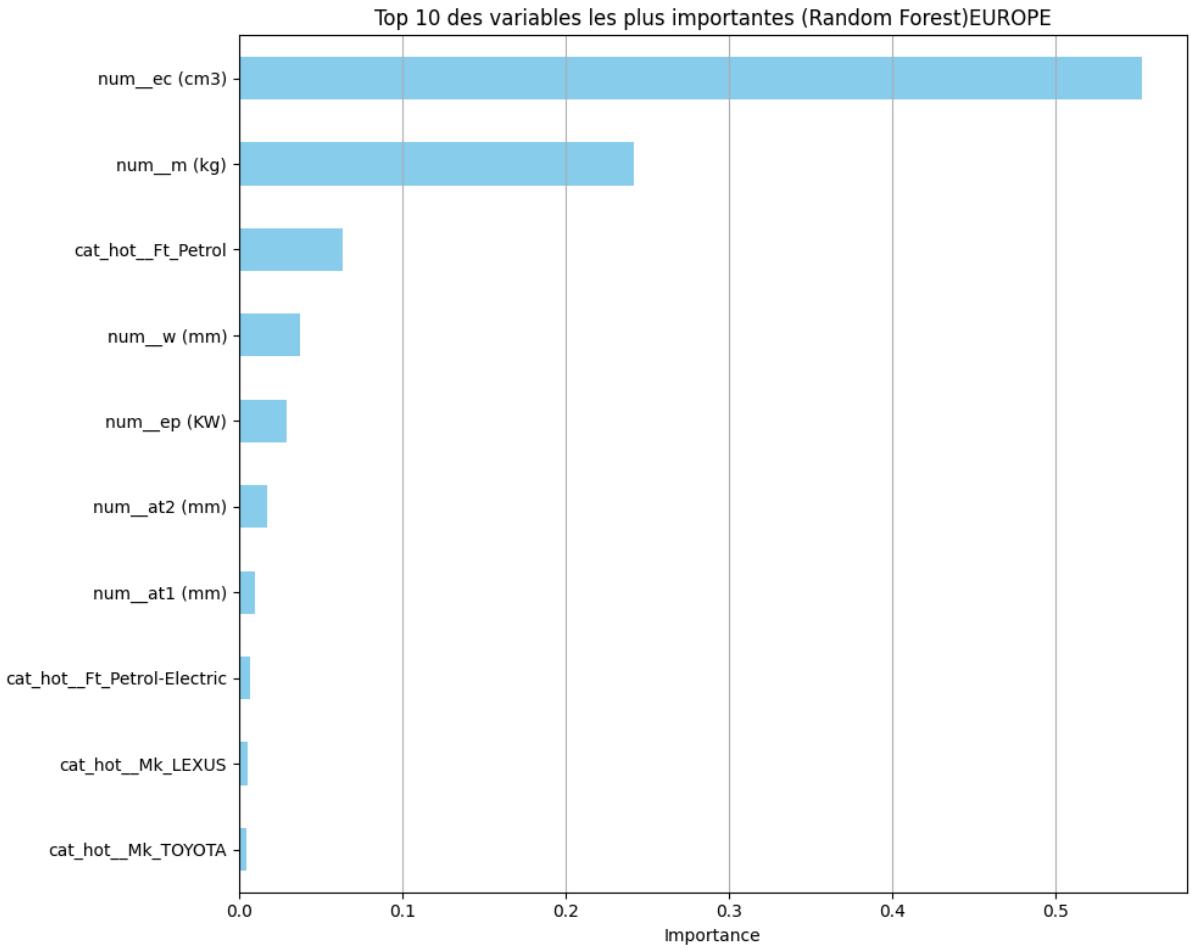
Nous pouvons voir que nous avons pu obtenir des résultats très satisfaisants sur les deux jeux de données avec des scores sur le jeu de test supérieur à 0.93.

B. Importances des variables

L'analyse de l'importance des variables dans les prédictions va nous aider à mieux comprendre nos résultats.



Lorsque l'on regarde l'importance des variables dans les prédictions du jeu de données de l'ADEME, nous remarquons que la variable qui arrive en tête est la masse en ordre de marche maximal (39%). Ainsi le poids d'un véhicule est un facteur déterminant des émissions de CO_2. La deuxième variable qui à de l'importance est la puissance administrative du véhicule (25%). On remarque donc que les émissions de CO_2 sont principalement influencées par des caractéristiques techniques du véhicule (Poids, puissance, carburant). Les caractéristiques telles que la marque, la carrosserie ou encore la gamme jouent un rôle mineur dans la prédiction.



Le graphique ci-dessus nous montre l'importance des variables dans le modèle Random Forest sur le jeu Europe. Nous remarquons que la variable qui a le plus d'importance est la cylindrée (cm3) (55%). La cylindrée est liée à la taille du moteur et a une influence importante sur la consommation de carburant et donc les émissions de CO2. La deuxième variable est le poids du véhicule et la troisième variable est l'essence. Enfin ,les variables concernants les dimensions techniques (largeur, empattement avant/arrière) ont une importance moindre mais peuvent permettre d'affiner la prédiction

Conclusion

L'objectif de notre projet était d'identifier les facteurs influençant les émissions de CO2 des véhicules en utilisant différents modèles de prédiction sur deux jeux de données: un issu de l'ADEME et l'autre issu de l'Agence Européenne de l'Environnement. Chacun offrait des caractéristiques propres tant en termes de variables disponibles que de qualité des données. Ils montraient toutefois une certaine complémentarité qui nous a poussé à réaliser notre analyse sur les deux jeux de données en parallèle.

Parmi les modèles testés, le Gradient Boosting Regressor offre les meilleures performances globales avec un score allant jusqu'à 0.965 sur le jeu de données de l'ADEME et 0.925 sur le jeu de l'Agence Européenne. Le Random Forest Regressor a également montré de très bonnes performances avec des temps de calcul moins longs.

L'analyse de l'importance des variables a montré que pour les deux jeux de données se sont la masse, la puissance, la cylindrée ou encore le type de carburant des véhicules qui prédominent.

Limites de l'étude

L'impossibilité de fusionner les deux bases de données nous a conduit à réaliser deux analyses en parallèle sans possibilité de généraliser nos résultats sur un jeu de données commun. Certaines variables existent dans une base mais pas dans l'autre ce qui nous empêche de croiser les indicateurs. Enfin, chaque base a nécessité, une analyse, une préparation, une modélisation distincts complexifiant l'interprétation globale

Perspectives d'amélioration

Pour la suite de ce projet, plusieurs pistes pourraient être explorées:

- Déséquilibre dans le dataset de l'ADEME: les minibus représentent 83% des véhicules de ce jeu de données ce qui peut introduire un biais dans les résultats de notre modélisation. Il serait donc pertinent de chercher des solutions pour limiter ce biais comme la réalisation d'échantillonnage ou d'essayer d'enrichir le jeu de données en ajoutant des observations issues d'autres types de véhicules et ainsi réduire le déséquilibre actuel.
- Fusionner les deux bases de données: Cela permettrait d'avoir un jeu de données plus riche et donc une vision plus complète des véhicules et nous permettrait de tirer parti de leurs complémentarités. Les modèles gagneraient en apprentissage et pourraient ainsi mieux généraliser.
- Amélioration du Gradient Boosting Regressor: Ce modèle s'est montré très performant, cependant il est très sensible au choix de ses hyperparamètres. On

pourrait donc essayer d'améliorer davantage ses performance en cherchant à optimiser aux mieux ces paramètres

Bilan

Ce projet mené en groupe a été à la fois une expérience enrichissante et exigeante. L'un des principaux défis a été la coordination à distance: il n'est pas toujours simple d'allier le rythme des sprints au travail collaboratif. Nous avons réussi à bien communiquer, dans un climat serein et constructif ce qui nous a permis d'obtenir de bons résultats.

Sur le plan technique, nous avons pu consolider nos connaissances en nettoyage, exploration et préparation des données, mais aussi découvrir et approfondir des notions plus complexes qui ont nécessité des investissements personnels importants, mais ont été très formateurs.