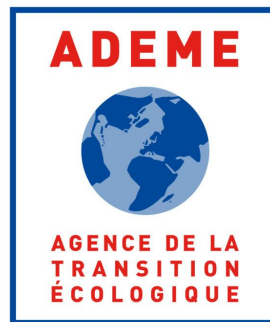


Prédiction des émissions de CO₂ des véhicules



Sabine Lohier
sb.lohier@gmail.com



Objectifs :

- Réaliser une analyse exploratoire des données
- Bâtir une modélisation prédictive des émissions de CO_2 des véhicules

Les données :

- Liste des références des véhicules commercialisés en France de 2001 à 2015 ainsi que leurs caractéristiques techniques, administratives, leurs émissions en polluants et gaz à effet de serre (CO , CO_2 , NOx ..)
- Provenance : ADEME

Données labellisées et y_i continus → **Modèle de régression - apprentissage supervisé**
Mise en place d'un modèle explicable → **Modèle de régression linéaire**



La base contient 55 044 lignes et 26 colonnes :

- 13 variables catégorielles
- 13 variables numériques

Suppression {

- **Doublons** : 7 doublons purs
- **Target manquante** : 34 occurrences, véhicules 100 % électriques

➤ **Séparation de la base en données d'entraînement (75%) et en données de test (25%)**





Variables catégorielles :

- Identification et données administratives
 - *Marque, désignation du modèle et désignation commerciale* lib_mrqlib_mod_dosslib_mod
dscom
 - *Code National d'Identification du Type du véhicule* cnit *et Type - Variante- Version* tvv
- Caractéristiques techniques :
 - *Type de carburant* cod_cbr, *hybridité* hybride
 - *Type de transmission et nombre de vitesses* typ_boite_nb_rapp
 - *Type de gamme* gamme *et de carrosserie* Carrosserie
- Caractéristiques environnementales :
 - *Norme Euro* champ_v9
- Autre : *Date de mise à jour* date_maj





Variables catégorielles :

- Identification et données administratives
 - *Marque, désignation du modèle et désignation commerciale* lib_mrqlib_mod_dosslib_mod
dscom
 - *Code National d'Identification du Type du véhicule* cnit *et Type - Variante- Version* tvv
- Caractéristiques techniques :
 - *Type de carburant* cod_cbr, *hybridité* hybride
 - *Type de transmission et nombre de vitesses* typ_boite_nb_rapp
 - *Type de gamme* gamme *et de carrosserie* Carrosserie
- Caractéristiques environnementales :
 - *Norme Euro* champ_v9
- Autre : *Date de mise à jour* date_maj

➔ 2 variables présentent des valeurs manquantes soit 7.2%





Variables quantitatives:

➤ Caractéristiques techniques :

- *Puissance administrative* `puiss_admin_98`, *puissance maximale* `puiss_max`
- *Masse en ordre de marche minimale* `masse_ordma_min` *et maximale* `masse_ordma_max`

➤ Consommations :

- *en milieu urbain* `conso_urb`, *extra urbain* `conso_exurb` *et mixte* `conso_mixte`

➤ Caractéristiques environnementales :

- *Émissions des polluants* : CO `co_typ_1`, *hydrocarbures imbrûlés* `hc`, *Particules* `ptcl`, *NOx* `nox`
`hcnnox`
- *Émissions en CO2* `co2`





Variables quantitatives:

➤ Caractéristiques techniques :

- *Puissance administrative* `puiss_admin_98`, *puissance maximale* `puiss_max`
- *Masse en ordre de marche minimale* `masse_ordma_min` *et maximale* `masse_ordma_max`

➤ Consommations :

- *en milieu urbain* `conso_urb`, *extra urbain* `conso_exurb` *et mixte* `conso_mixte`

➤ Caractéristiques environnementales :

- *Émissions des polluants : CO* `co_typ_1`, *hydrocarbures imbrûlés* `hc`, *Particules* `ptcl`, *NOx* `nox`
`hcnnox`
- *Émissions en CO2* `co2`

➔ **7 variables présentent des valeurs manquantes soit 8.1%**



Suppressions des variables

Id. et données admin.	<ul style="list-style-type: none">■ Marque, désignation du modèle et désignation commerciale lib_mrq lib_mod_doss lib_mod_dscom■ Code National d'Identification du Type du véhicule enit et Type Variante Version tvv
Caract. techniques	<ul style="list-style-type: none">■ Type de carburant cod_cbr, hybridité hybride■ Type de transmission et nombre de vitesses typ_boite_nb_rapp■ Type de gamme gamme et de carrosserie Carrosserie■ Puissance administrative puiss_admin_90, puissance maximale puiss_max■ Masse minimale masse_ordma_min et maximale masse_ordma_max en ordre de marche
Conso.	<ul style="list-style-type: none">■ Consommation en carburant en milieu urbain conso_urb, extra urbain conso_exurb et mixte conso_mixte
Caract. env.	<ul style="list-style-type: none">■ Norme Euro champ_v9■ Émissions des polluants : CO co_typ_1, hydrocarbures imbrûlés hc, Particules ptel, NOx nox henox■ Émissions en CO2 co2
Autre	<ul style="list-style-type: none">■ Date de mise à jour date_maj



➤ Le type de carburant `cod_cbr` :

9 modalités incluant 2 aspects fonctionnels des véhicules :

- Le carburant fossile utilisé,
- Leur nature hybride ou non.

Carburant	Total des véhicules	% des véhicules
DIESEL	37114	90.0%
ESSENCE	4 068	9.8%
AUTRE	70	0.2%

➤ Le type de carburant `cod_cbr` :

9 modalités incluant 3 aspects fonctionnels des véhicules :

- Le carburant fossile utilisé,
- Leur nature hybride ou non

➔ 90 % des véhicules fonctionnent au gasoil.

➤ La nature hybride `hybridité` :

Hybride	Total des véhicules	% des véhicules
non	40 971	99.3%
oui	281	0.7%

➤ **Le type de carburant** `cod_cbr` :

9 modalités incluant 3 aspects fonctionnels des véhicules :

- Le carburant fossile utilisé,
- Leur nature hybride ou non

➡ 90 % des véhicules fonctionnent au gasoil.

➤ **La nature hybride** `hybride` :

➡ 99% des véhicules ont un moteur hybride.

➤ **Le type de transmission et le nombre de vitesses** `typ_boite_nb_rapp` :

14 modalités constituées:

- du mode de transmission (“A” pour automatique, “M” pour manuelle ...)
- Suivi par le nombre de vitesses

Transmission	Total des véhicules	% des véhicules
MANUELLE	24 616	59.7%
AUTOMATIQUE	16 325	39.6%
AUTRE	311	0.7%



➤ La gamme gamme:

6 catégories désignant la montée en gamme

→ 88.5 % des véhicules sont de gamme **moyenne inférieure à moyenne supérieure**

type de gamme	Total des véhicules	% des véhicules
ECONOMIQUE	175	0.5%
INFÉRIEURE	791	1.9%
MOY-INFER	25 157	60.9%
MOY-SUPER	11 371	27.6%
SUPÉRIEURE	1 098	2.7%
LUXE	2 660	6.4%



➤ La gamme gamme:

6 catégories désignant la montée en gamme.

→ 88.5 % des véhicules sont de gamme **moyenne inférieure à moyenne supérieure**

➤ La carrosserie Carrosserie : 10 modalités

type de carrosserie	Total des véhicules	% des véhicules
MINISPACE	99	0.2%
MONOSPACE COMPACT	244	0.6%
COMBISPACE	178	0.4%
BERLINE	3 286	8.0%
COUPÉ	649	1.6%
CABRIOLET	395	1%
BREAK	881	2.1%
MONOSPACE	43	0.1%
TS TERRAINS/CHEMINS	819	2.0%
MINIBUS	34 658	84.0%

➤ La gamme gamme:

6 catégories désignant la montée en gamme.

➔ 88.5 % des véhicules sont de gamme **moyenne inférieure à moyenne supérieure**

➤ La carrosserie Carrosserie : 10 modalités

➔ 84 % des véhicules sont des ‘minibus’ et seulement 8% des ‘berlines’

type de carrosserie	Total des véhicules	% des véhicules
MINISPACE	99	0.2%
MONOSPACE COMPACT	244	0.6%
COMBISPACE	178	0.4%
BERLINE	3 286	8.0%
COUPÉ	649	1.6%
CABRIOLET	395	1%
BREAK	881	2.1%
MONOSPACE	43	0.1%
TS TERRAINS/CHEMINS	819	2.0%
MINIBUS	34 658	84.0%

➤ La gamme gamme:

6 catégories désignant la montée en gamme.

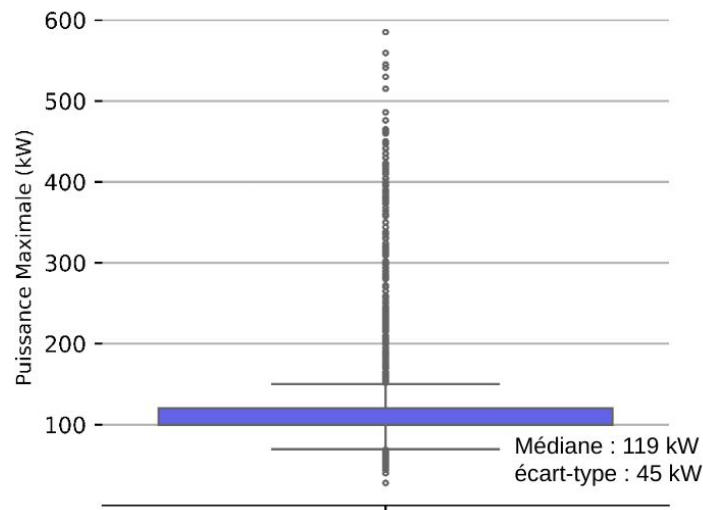
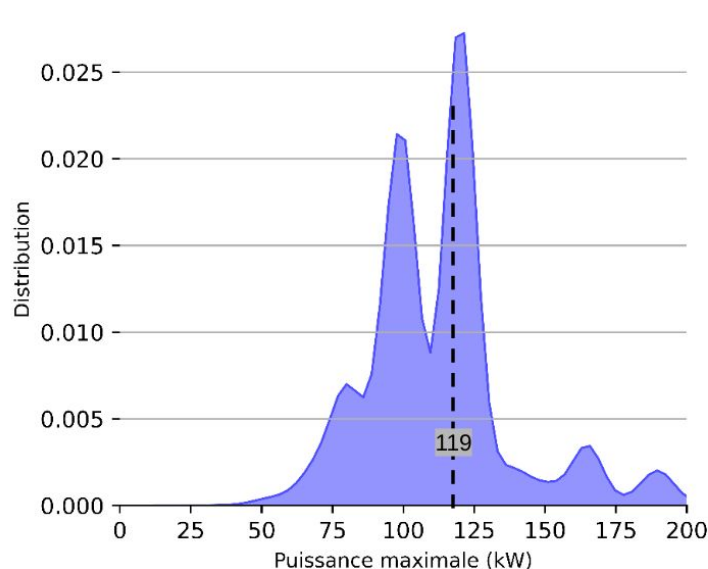
➔ 88.5 % des véhicules sont de gamme **moyenne inférieure à moyenne supérieure**

➤ La carrosserie Carrosserie : 10 modalités

➔ 84 % des véhicules sont des ‘minibus’ et seulement 8% des ‘berlines’

type de carrosserie	MINISPACE	MONOSPACE COMPACT	COMBISPACE	BERLINE	COUPE	CABRIOLET	BREAK	MONOSPACE	TS TERRAINS	MINIBUS	total
DIESEL	0.1%	0.5%	0.4%	4.4%	0.4%	0.3%	1.5%	0.1%	1.3%	91.0%	100%
ESSENCE	1.3%	1.7%	0.7%	39.9%	12.5%	7.0%	8.2%	0.1%	7.7%	20.9%	100%

➤ La puissance maximale `puiss_max`:

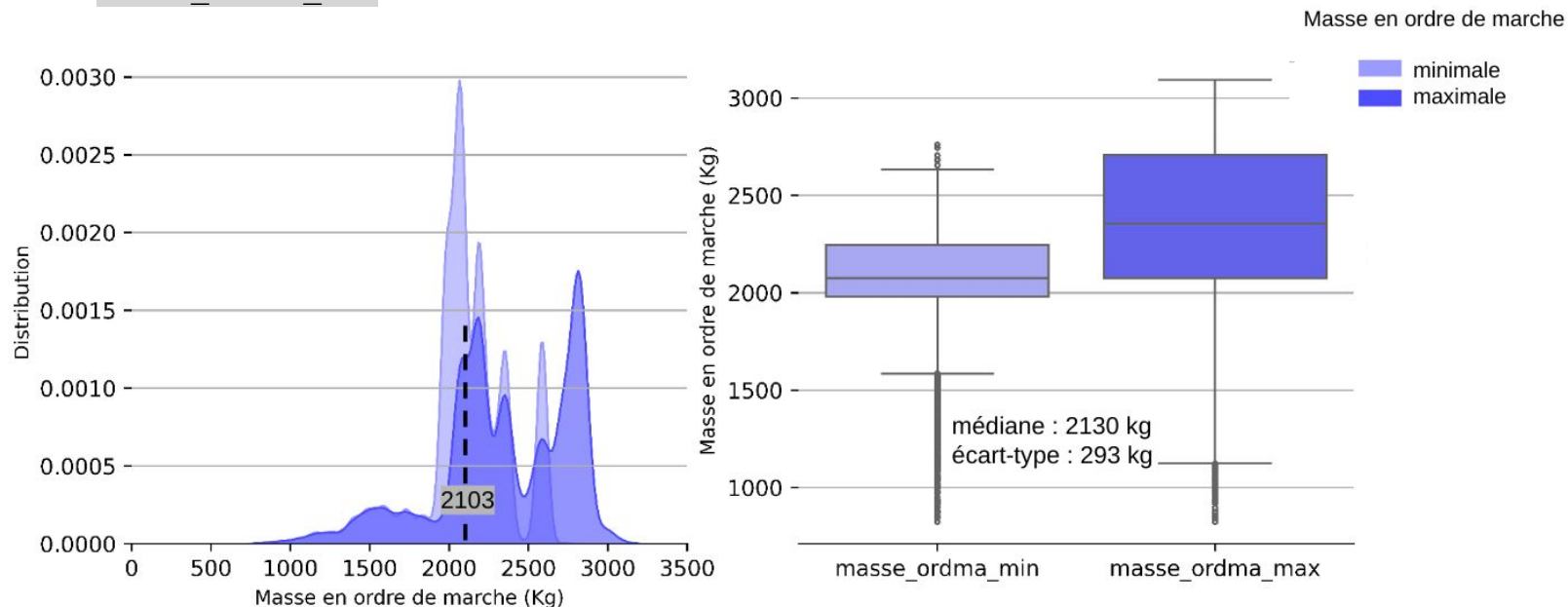


Distribution de la puissance maximale des véhicules.

Outliers

- `puiss_max` > 500 kW (11 observations)
- `puiss_max` < 40 kW (1 observation)

➤ La masse `masse_ordma_min`:

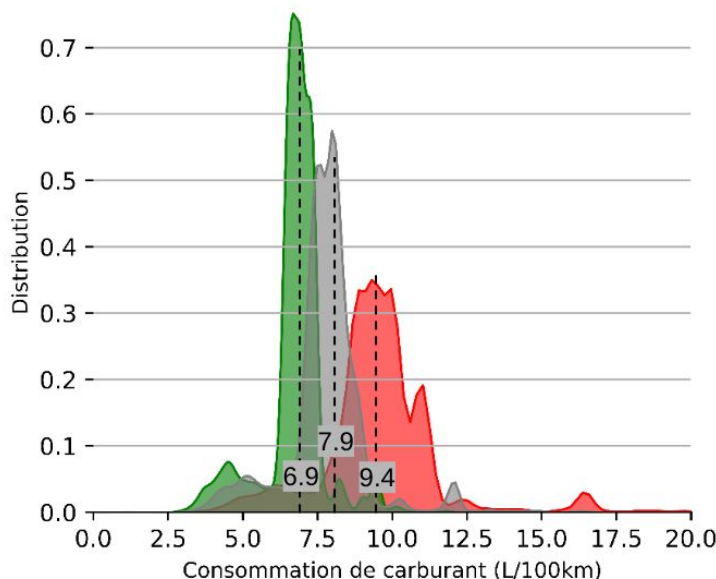


Distribution de la masse minimale et maximale en ordre de marche des véhicules.

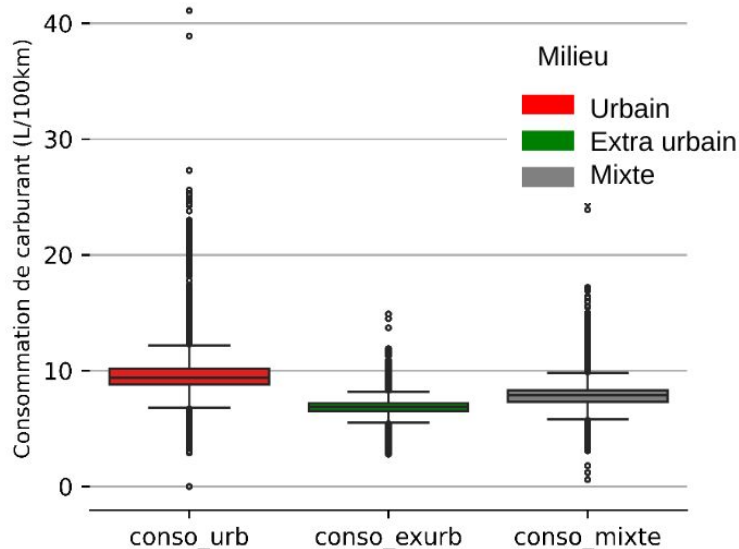
- Masse minimale médiane élevée : due à une forte représentativité des véhicules lourds

Note : Seule la masse minimale est retenue pour la suite de l'analyse

➤ Les consommations `conso_urb`, `conso_exurb`, `conso_mixte`



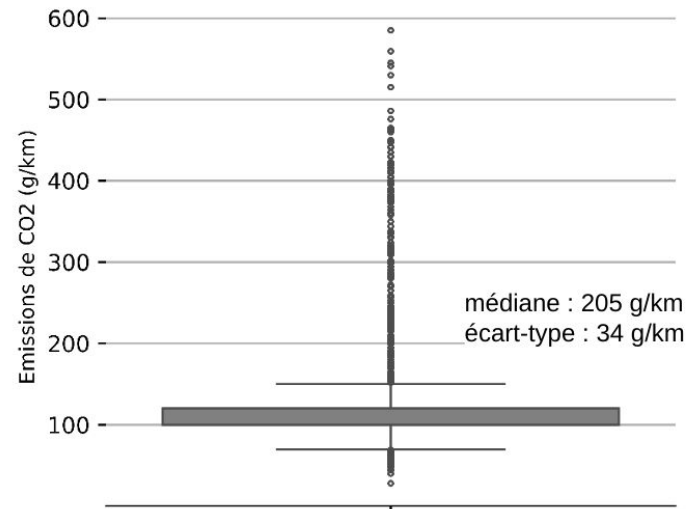
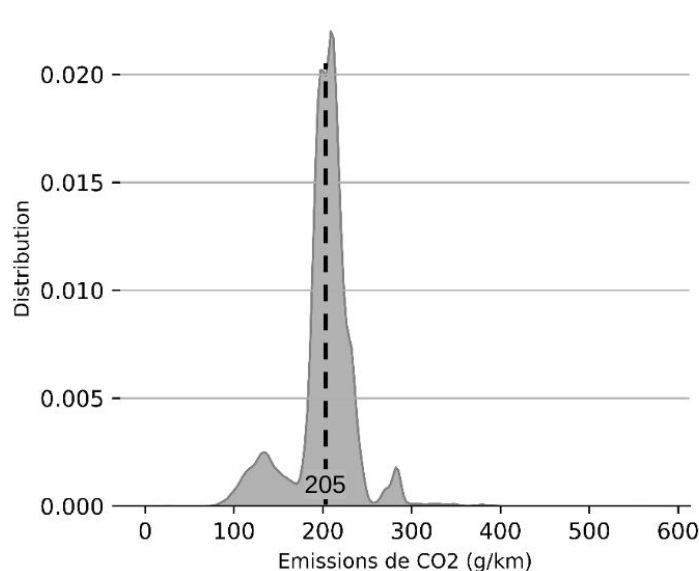
Distribution de la consommation urbaine, extra urbaine, mixte de carburant (en L/100km)



- Outliers
- `conso_mixte` > 20 L/100km (2 observations)
 - `conso_mixte` < 3 L/100km (4 observations)

Note : Seule la consommation mixte est retenue pour la suite de l'analyse

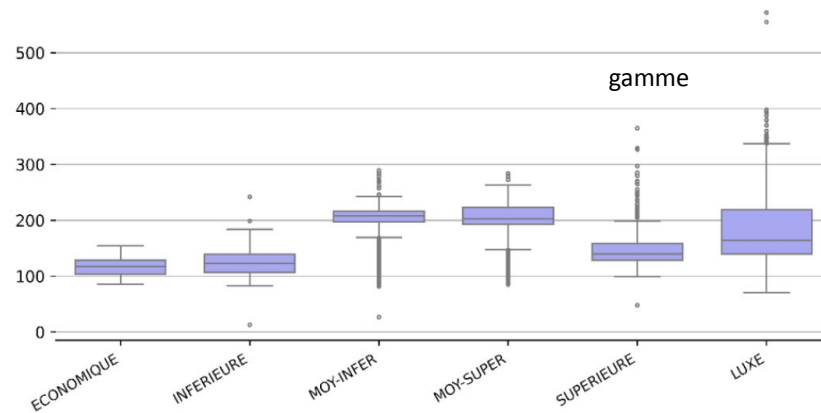
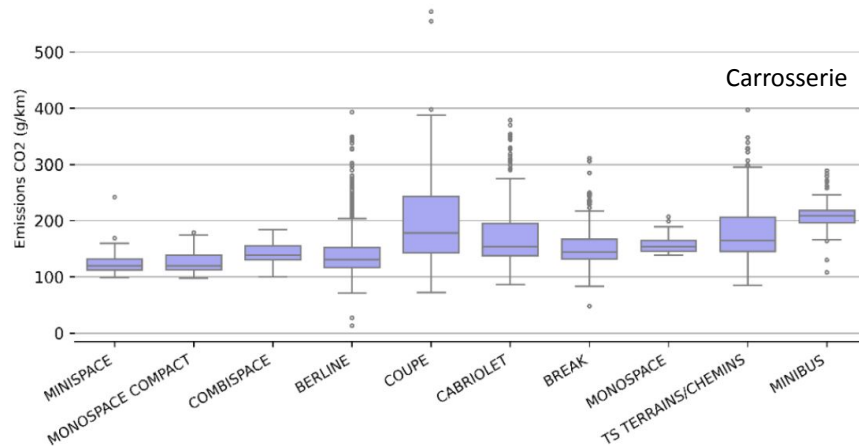
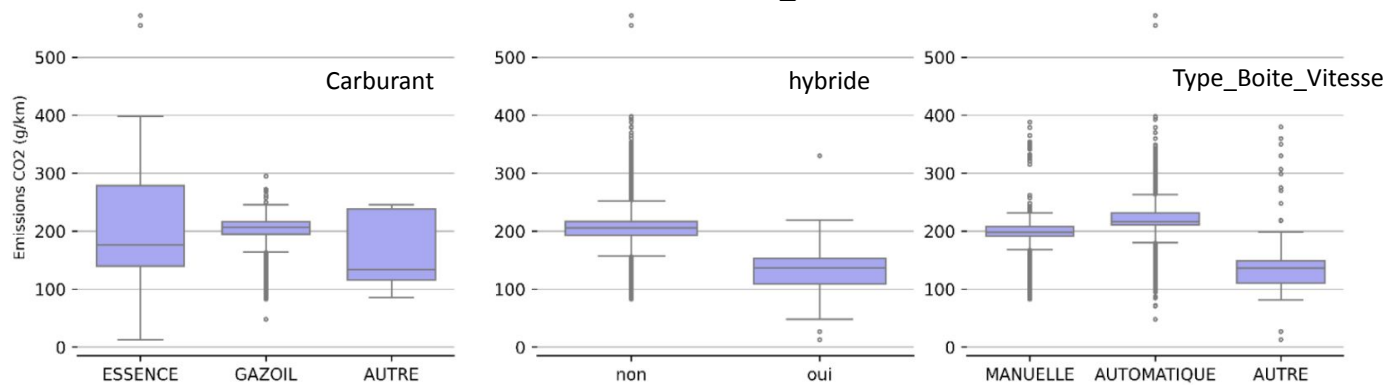
➤ Les émissions de co2 `co2`:



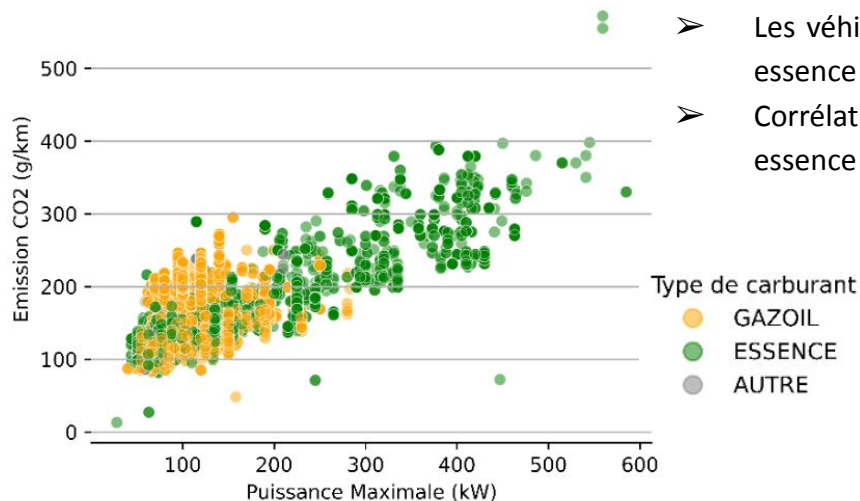
Distribution des émissions de CO2 des véhicules.

- Outliers
- `co2` > 500 g/km (2 observations)
 - `co2` < 40 g/km (3 observations)

➤ Émissions de CO2 vs les modalités des variables catégorielles :

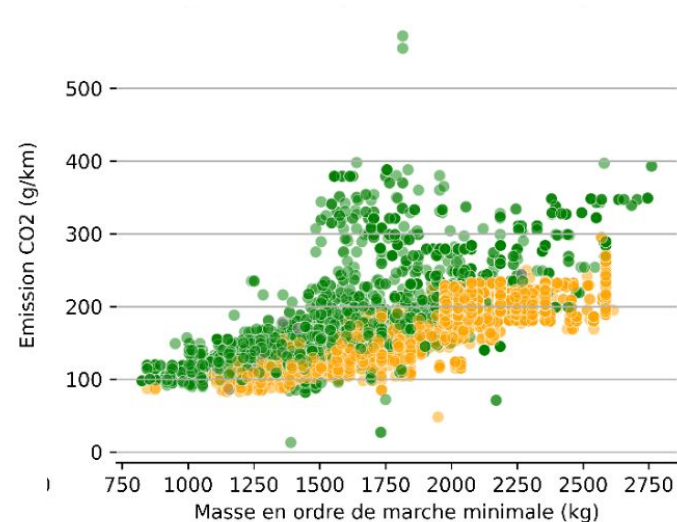


➤ Emissions de CO2 vs variables quantitatives :

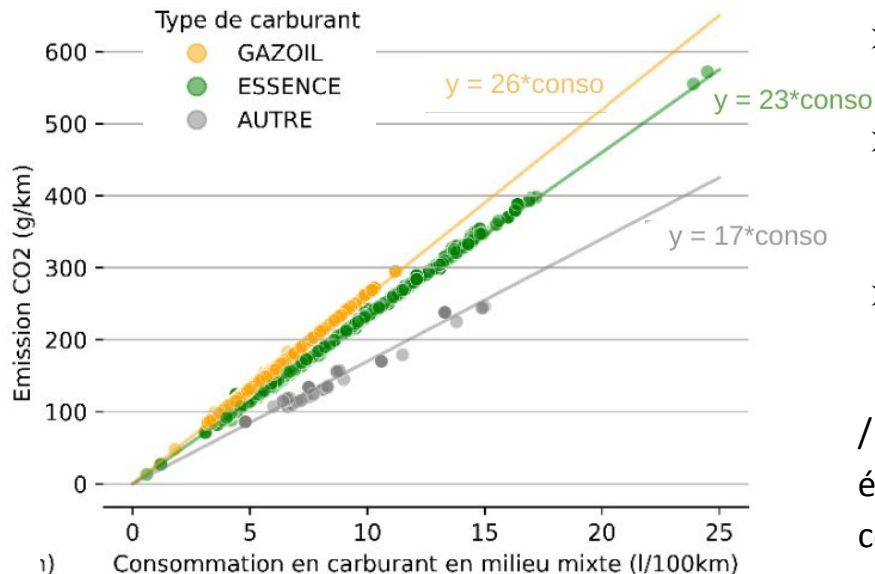


- Les véhicules dont `puiss_max > 300 kW` ne sont que des véhicules à moteur essence
- Corrélation entre émissions de CO2 et la puissance max pour les véhicules essence

- Bonne corrélation en les émissions de CO2 et la masse
- A masse égale, les véhicules essence émettent plus de CO2 que les véhicules diesel
- 2 groupes d'outliers :
 - 2 outliers pour lesquels $CO_2 > 500$ g/km
 - 7 outliers pour lesquels $CO_2 < 80$ g/km (dont 3 déjà définis précédemment + 4 véhicules hybrides rechargeables)



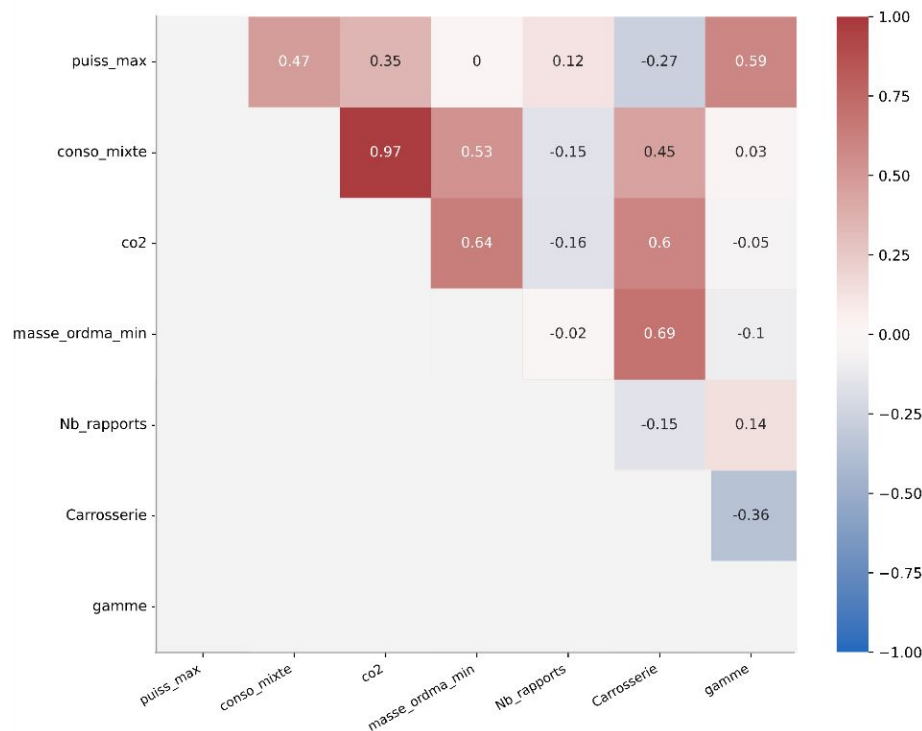
➤ Émissions de CO2 vs variables quantitatives :



- Les véhicules **diesel** consomment moins que les véhicules **essence**
- Excellente corrélation entre la consommation de carburant et les émissions de CO2 selon les différents carburants utilisés
- À consommation équivalente, les véhicules à moteurs diesel ont tendance à émettre légèrement plus de CO2

/!\ Les observations approchent au plus près la relation entre les émissions de CO2 et la consommation de carburant en considérant une combustion parfaite du carburant

➤ Matrice de Corrélation



➤ Les émissions de CO2 sont très fortement corrélées à la consommation mixte (coef. = 0.97), suivi par la masse (coef = 0.53), le type de carrosserie, et plus faiblement à la puissance maximale (coef = 0.35).

➤ On observe de fortes corrélations entre variables explicatives, par exemple entre la consommation et la masse (0.53), la puissance (0.47) et la carrosserie (0.45).

➤ Sélection des variables explicatives :

- *Type de carburant* `cod_cbr`
- *hybridité* `hybride`
- *Type de transmission et nombre de vitesses*
`typ_boite_nb_rapp`
- *Type de gamme* `gamme`
- *Type de carrosserie* `Carrosserie`
- *puissance maximale* `puiss_max`
- *Masse minimale* `masse_ordma_min`
- *Consommation mixte* `conso_mixte`

➤ Ajout de la variable `conso_mixte` comme un hyperparamètre

➤ Séparation des données en set d'entraînement (70%) et de validation (30%)

➤ Choix du modèle : LinearRegression() de scikit-learn.

Variables catégorielles

➤ Traitement des variables catégorielles :

- *Type de carburant* `cod_cbr` : regroupement des catégories en 3 modalités principales : “ESSENCE”, “GAZOIL”, “AUTRE”
- *Type de transmission et nombre de vitesses* `typ_boite_nb_rapp` : scindé en 2 variables distinctes :
 - `typ_boite_Vitesse` : ayant 3 modalités : “MANUELLE”, “AUTOMATIQUE”, “AUTRE”
 - `Nb_Rapport`

➤ Valeur manquante : Pas de valeur manquante au sein des variables catégorielles sélectionnées

➤ Encodage :

- *Type de gamme* `gamme` : Encodage ordinal manuel selon l'ordre de montée en gamme
 - *Type de carrosserie* `Carrosserie` : Encodage ordinal manuel selon l'ordre croissant de masse médiane
 - *Type de carburant* `cod_cbr`
 - *hybridité* `hybride`
 - *Type de transmission* `typ_boite_Vitesse`
 - *Nombre de vitesses* `Nb_Rapport`
- } Encodage ordinal via ordinalEncoder



Variables numériques

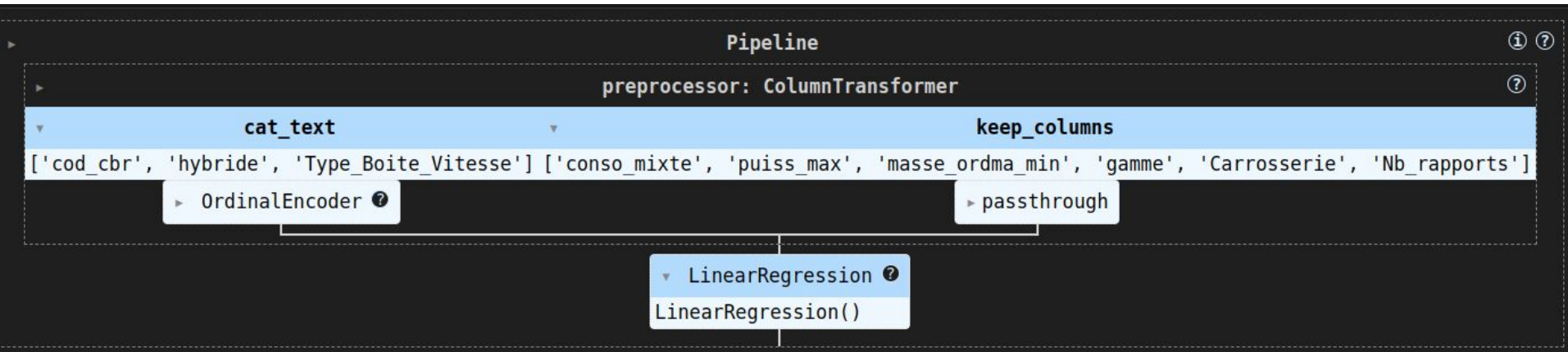
➤ Outliers :

- `puiss_max` : Imputations de valeurs plafonds :
 - `puiss_max` = 500 kW si `puiss_max` > 500 kW (11 observations)
 - `puiss_max` = 40 kW si `puiss_max` < 40 kW (1 observation)
- `conso_mixte` : Imputations de valeurs plafonds :
 - `conso_mixte` = 20 l/100km si `conso_mixte` > 20 l/100km (2 observations)
 - `conso_mixte` = 3 l/100km si `conso_mixte` < 3 l/100km (4 observations)
- `co2` : Calcul des émissions de CO2 en fonction de la `conso_mixte` et du carburant selon l'équation de combustion parfaite

- **Valeurs Manquantes** : Pas de valeur manquante au sein des variables quantitatives sélectionnées
- **Normalisation** : Il n'est pas requis de standardiser les variables numériques dans une même échelle avec l'algo **LinearRegression()** de scikit-Learn.



➤ Pipeline de transformation :



8 variables explicatives en entrée du modèle + 1 hyperparamètre (conso_mixte)

➤ Modèle 1 (avec `conso_mixte`)

Métrique	Jeu de train	Jeu de validation
MAE	1.34	1.36
RMSE	3.2	3.42
R2	0.9911	0.9896



Equation des émissions de CO2 :

$$\begin{aligned} \text{co2} = & -28.39 + 23.69 * \text{conso_mixte} + 19.73 * \text{cod_cbr} \\ & + 0.0031 * \text{masse_ordma_min} \\ & + 0.77 * \text{Carrosserie} \\ & + 0.0 * \text{puiss_max} - 0.35 * \text{Nb_rapports} \\ & - 0.44 * \text{gamme} - 0.74 * \text{hybride} \\ & - 0.78 * \text{Type_Boite_Vitesse} \end{aligned}$$

- Très bonnes métriques en considérant la valeur médiane de la target réelle (205g/km)
- Le biais du modèle est très faible
- Pas de sur-apprentissage
- La target est principalement déterminée par 3 variables : `conso_mixte` , `cod_cbr` et `masse_ordma_min`
- L'apport des autres variables est à la marge

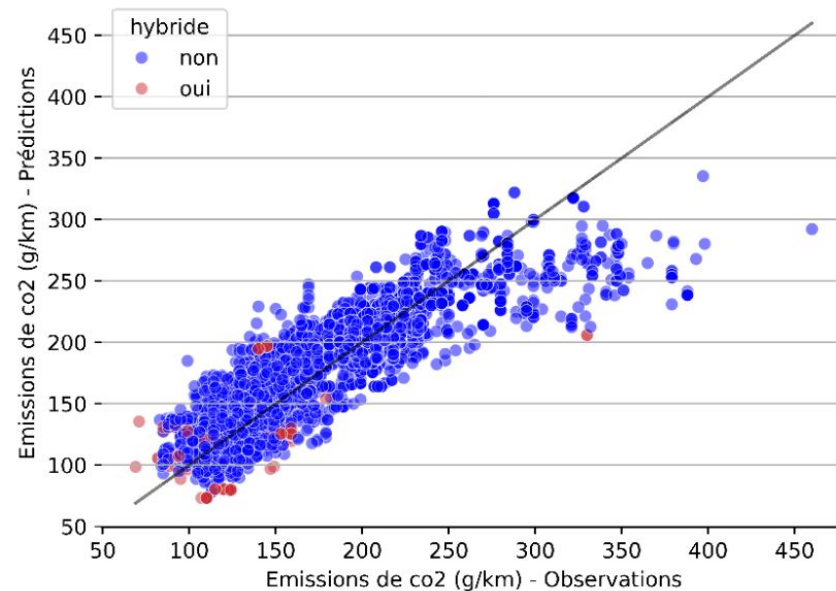
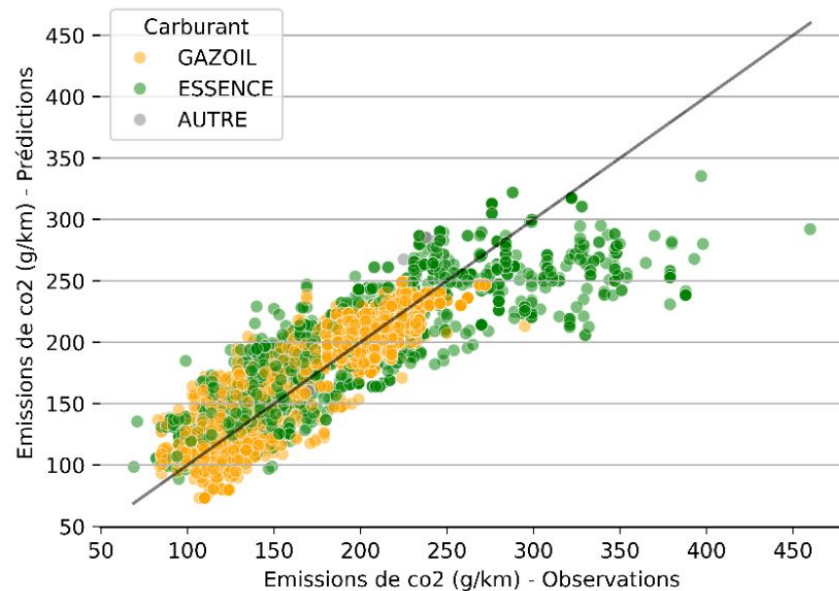


➤ **Modèle 2 (sans conso_mixte)**

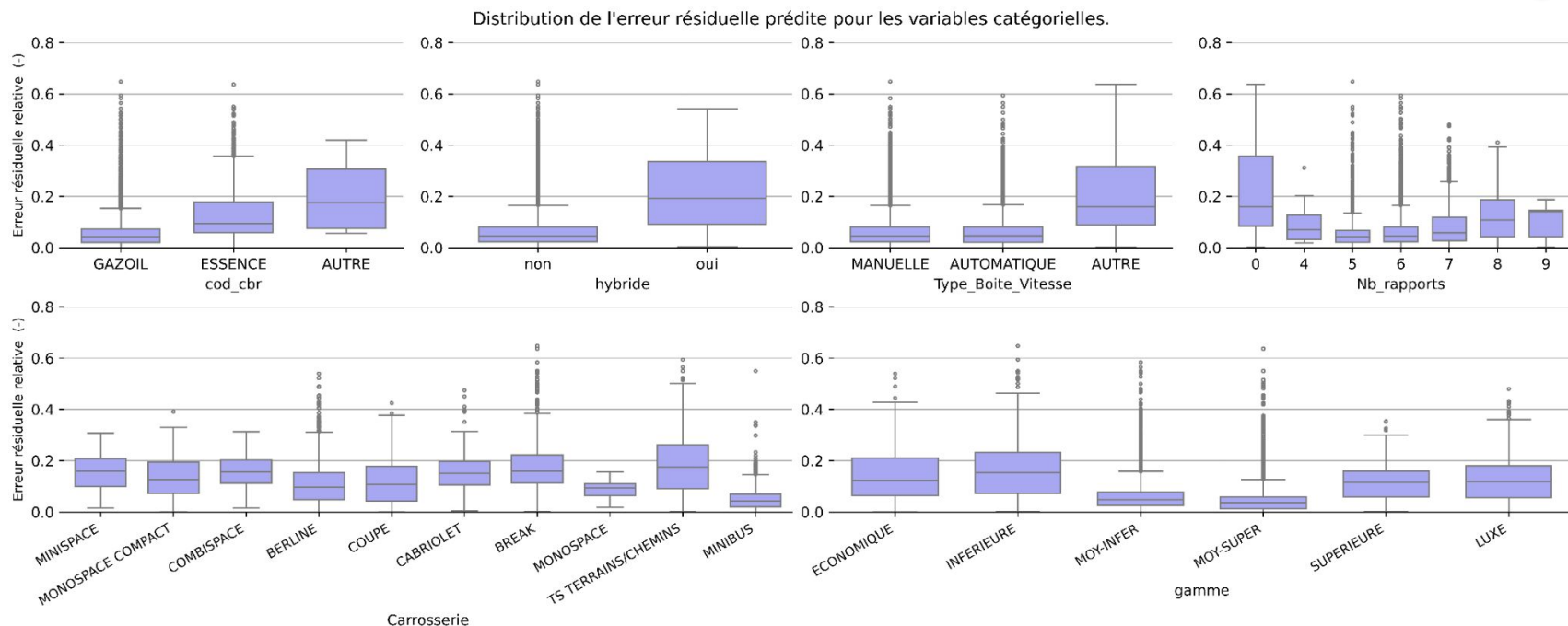
Métrique	Jeu de train	Jeu de validation	➤ Equation des émissions de CO2 :
MAE	11.89	11.8	$\begin{aligned} \text{co2} = & 99.89 + 0.039 * \text{masse_ordma_min} + 0.33 * \text{puiss_max} \\ & + 8.75 * \text{Carrosserie} \\ & - 23.89 * \text{cod_cbr} - 67.21 * \text{hybride} - 6.27 * \text{gamme} \\ & - 5.90 * \text{Type_Boite_Vitesse} - 5.38 * \text{Nb_rapports} \end{aligned}$
RMSE	16.05	16.04	
R2	0.776	0.7716	

- Les MAE et RMSE restent relativement correctes en considérant la valeur médiane des observations (205g/km)
- Modèle relativement bien ajusté aux données explicatives
- Pas de sur-apprentissage
- Redistribution des paramètres associés aux variables explicatives en supprimant la variables `conso_mixte` :
 - Leur poids sont tous à la hausse
 - Avec un apport significatif
 - en positif : des variables `masse_ordma_min`, `puiss_max`, `Carrosserie`
 - en négatif : des variables `cod_cbr`, `hybride`

➤ Modèle 2 (sans `conso_mixte`)



- Pour les observations où les émissions de CO2 < 250 g/km, le modèle surestime légèrement les émissions de CO2
- Pour les observations où les émissions de CO2 > 250 g/km, le modèle sous estime significativement les émissions de CO2
- Erreur résiduelle entre les prédictions et les observations plus importante pour les véhicules essence ainsi que les véhicules hybrides.



- Le modèle capture mieux les véhicules à moteur diesel et non hybride ...
- ... et plus généralement les catégories qui sont suffisamment assez représentées pour être apprises par le modèle.

➤ Modèle 2 - Features polynomiales de second ordre :

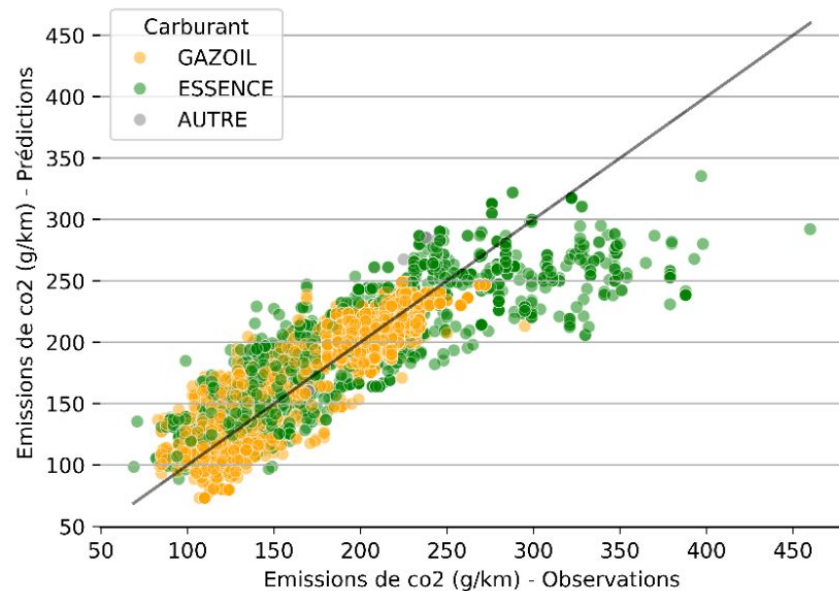
Métrique	Jeu de train	Jeu de validation	Jeu de train	Jeu de validation
MAE	11.89	11.8	9.26	9.17
RMSE	16.05	16.04	12.93	13.07
R2	0.776	0.7716	0.8538	0.8485

➤ Equation des émissions de CO2 :

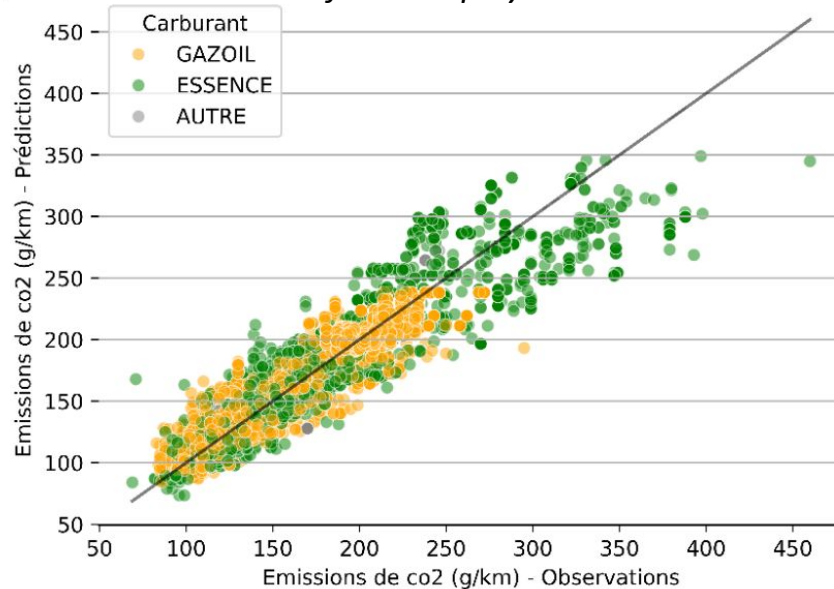
$$\begin{aligned}
 \text{co2} = & -39.20 + 0.18 * \text{masse_ordma_min} \\
 & + 0.71 * \text{puiss_max} \\
 & + 54.63 * \text{cod_cbr} - 76.40 * \text{hybride} \\
 & \dots \\
 & - 0.35 * \text{puiss_max} * \text{hybride} \\
 & - 0.02 * \text{masse_ordma_min} * \text{hybride} \\
 & - 76.40 * \text{hybride}^2 \\
 & - 156.20 * \text{Type_Boite_Vitesse} * \text{hybride}
 \end{aligned}$$

- Les MAE, RMSE et coefficient de détermination sont nettement améliorés
- Pas de sur-apprentissage
- Contribution significative
 - des features simples : `masse_ordma_min` , `puiss_max` , `cod_cbr` , `hybride`
 - des combinaisons polynomiales impliquant la variable `hybride` (en négatif)

➤ Modèle 2 - Features polynomiales de second ordre :



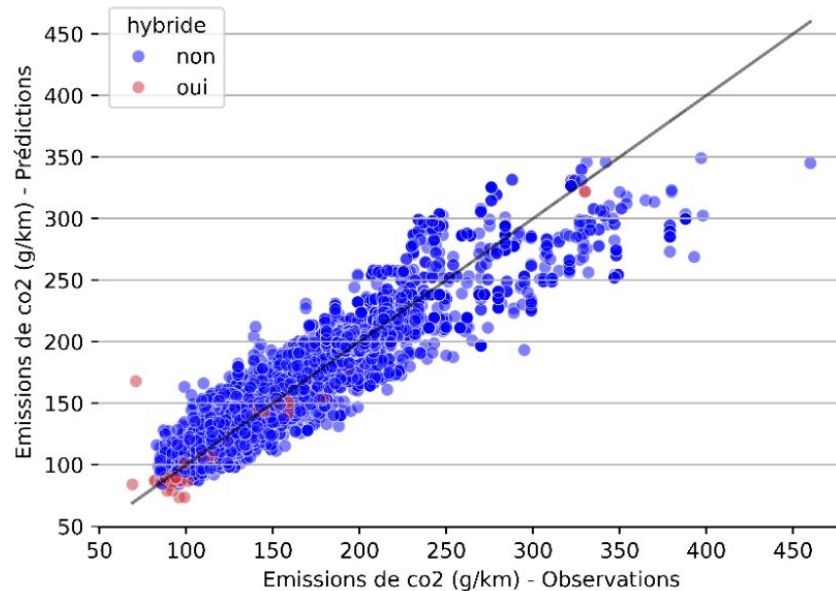
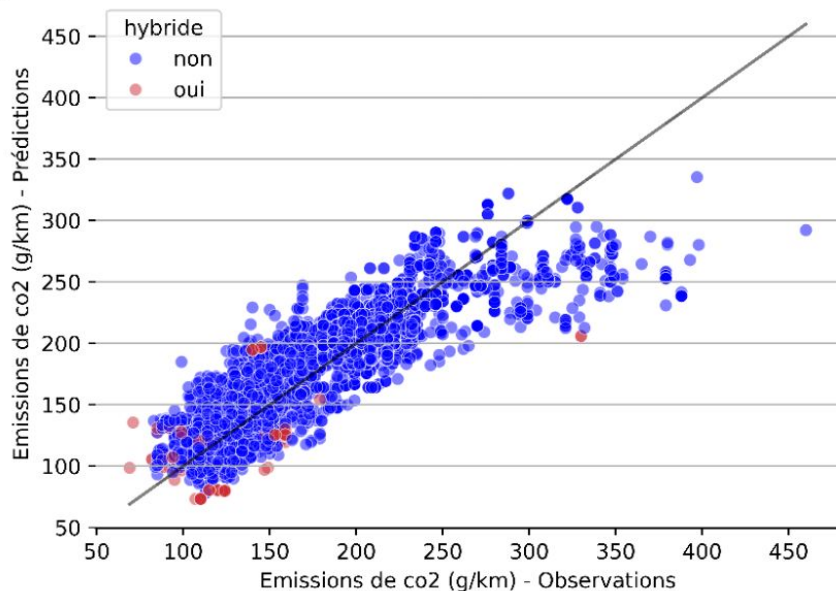
avec features polynomiales



- Corrections des surestimations des émissions de CO2 pour les observations dont la target réelle est < 250 g/km
- Corrections sous-estimations des émissions de CO2 pour les observations dont la target réelle est > 250 g/km
- Meilleures prédictions pour les véhicules hybrides.

➤ Modèle 2 - Features polynomiales de second ordre :

avec features polynomiales



- Corrections des surestimations des émissions de CO2 pour les observations dont la target réelle est < 250 g/km
- Corrections sous-estimations des émissions de CO2 pour les observations dont la target réelle est > 250 g/km
- Meilleures prédictions pour les véhicules hybrides.



➤ **Modèle 1 (avec conso_mixte)**

Métrique	Jeu de train	Jeu de validation	Jeu de test
MAE	1.34	1.36	1.32
RMSE	3.2	3.42	2.81
R2	0.9911	0.9896	0.9932

➤ **Modèle 2 - Features polynomiales de second ordre**

Métrique	Jeu de train	Jeu de validation	Jeu de test
MAE	9.26	9.17	9.32
RMSE	12.93	13.07	13.19
R2	0.8538	0.8485	0.8517

➤ Les 2 modèles possèdent une bonne capacité de généralisation aux données test avec des métriques très similaires à celles du test d'entraînement et de validation



L'analyse exploratoire de la base de données provenant de l'ADEME a révélé de **forts déséquilibres au sein des classes des différentes variables catégorielles** :

- 90.0% des véhicules enregistrés en base fonctionnent au gazoil en régime moteur, contre 9.9 % pour essence.
- Plus de 99% de nos observations sont constituées de véhicules non hybrides
- L'essentiel des véhicules sont des minibus (84%).

Selon les analyses multivariées :

- Les émissions de CO2 sont très fortement corrélées à la consommation de carburant en conduite mixte (avec un coefficient de 0.97), à la masse, au type de carrosserie, et plus faiblement à la puissance maximale (coefficient de corrélation de 0.35).

Création d'un **modèle prédictif de régression linéaire** basé sur **8 variables explicatives + 1** (consommation mixte - hyperparamètre) :

- En incluant la consommation mixte :
 - Excellentes prédictions d'émission de CO2 caractérisées par une MAE de 1.3 g/km, une RMSE 3.2 g/km sur le set d'entraînement
 - Très bonne généralisation aux données de test avec une MAE de 1.3 g/km, une RMSE de 2.8 g/km et un R2 de 0.99.
- En excluant la consommation mixte :
 - Relativement bonnes prédictions améliorées par l'ajout de feature polynomiales d'ordre 2, les MAE et RMSE sont de 9 et 13 g/km
 - Bonne généralisation aux données de test avec une MAE de 9.3 g/km, une RMSE de 13.2 g/km et un R2 de 0.85.