



Jynaldo Jeannot
Yoan Jsem

KAGGLE CHALLENGE

Estimate CO2 emissions from cars in Europe



SOMMAIRE

01

Mise En Contexte

02

Analyse Exploratoire

03

Preprocessing

04

Modèles Utilisés

05

Résultats

06

Axes d'Amélioration





01 - MISE EN CONTEXTE

Objectif

Prédire les émissions de CO₂ d'une voiture à partir de ses caractéristiques.

Métrique

On utilisera comme métrique d'évaluation de notre modèle la Mean Absolute Error (MAE)



01 - MISE EN CONTEXTE

Target

Nous sommes dans un problème de régression avec
une cible: "**Ewltp (g/km)**"

Données

Pour ce challenge nous disposons de 2 jeux de
données: train.csv avec la target et un jeu de
données test.csv sans la target que nos devons
prédirer

02 - ANALYSE EXPLORATOIRE

7M5

Observations

37

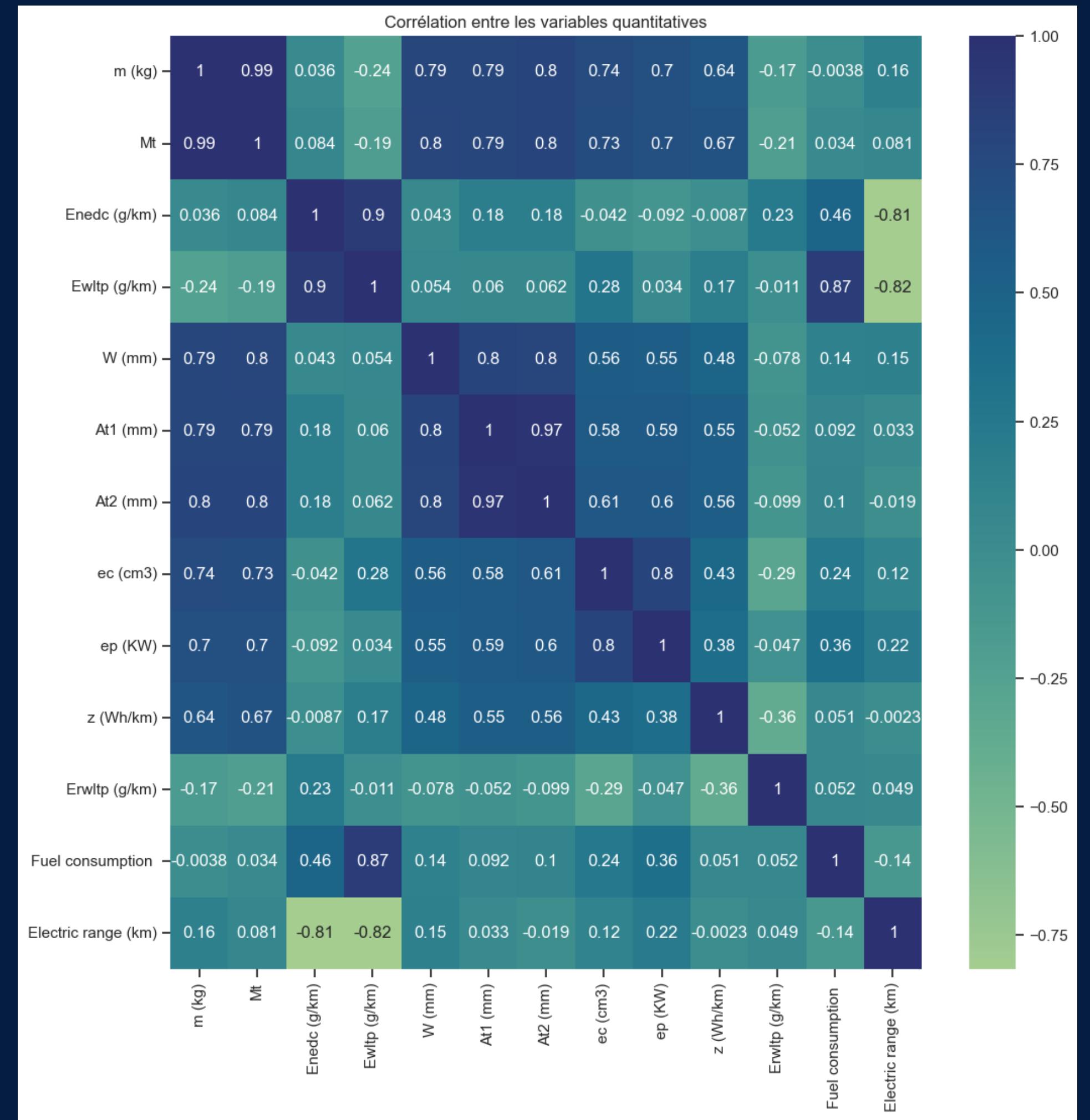
Colonnes



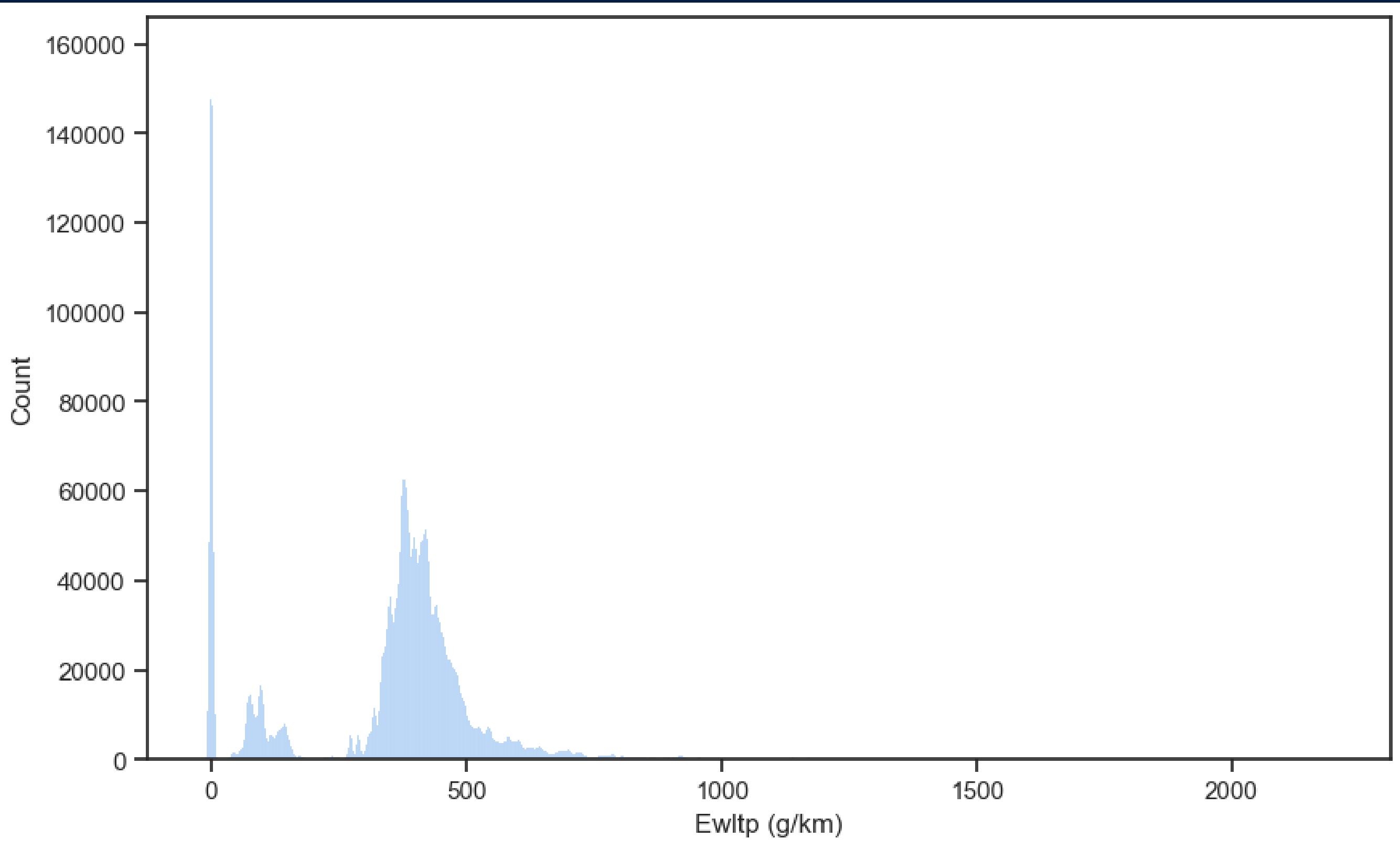
Valeurs uniques et valeurs manquantes

Nombres de valeurs uniques	
Mp	10
Fm	6
Ct	5
Cr	3
r	1
Status	1
Ernedc (g/km)	0
De	0
Vf	0
MMS	0

Colonne	pourcentage manquant	nombre
MMS	100.00	7571649
Ernedc (g/km)	100.00	7571649
De	100.00	7571649
Vf	100.00	7571649
Enedc (g/km)	83.84	6348010
Electric range (km)	82.96	6281247
z (Wh/km)	77.98	5904329
Erwltp (g/km)	46.48	3519145
IT	37.78	2860870
Fuel consumption	23.51	1779861
ec (cm3)	13.51	1022765
Mt	11.10	840236
VFN	8.61	652232
Mp	6.41	485564
At2 (mm)	2.34	177298
At1 (mm)	2.19	165685
Date of registration	1.70	129058
Cn	1.52	115232
Ve	0.42	32077

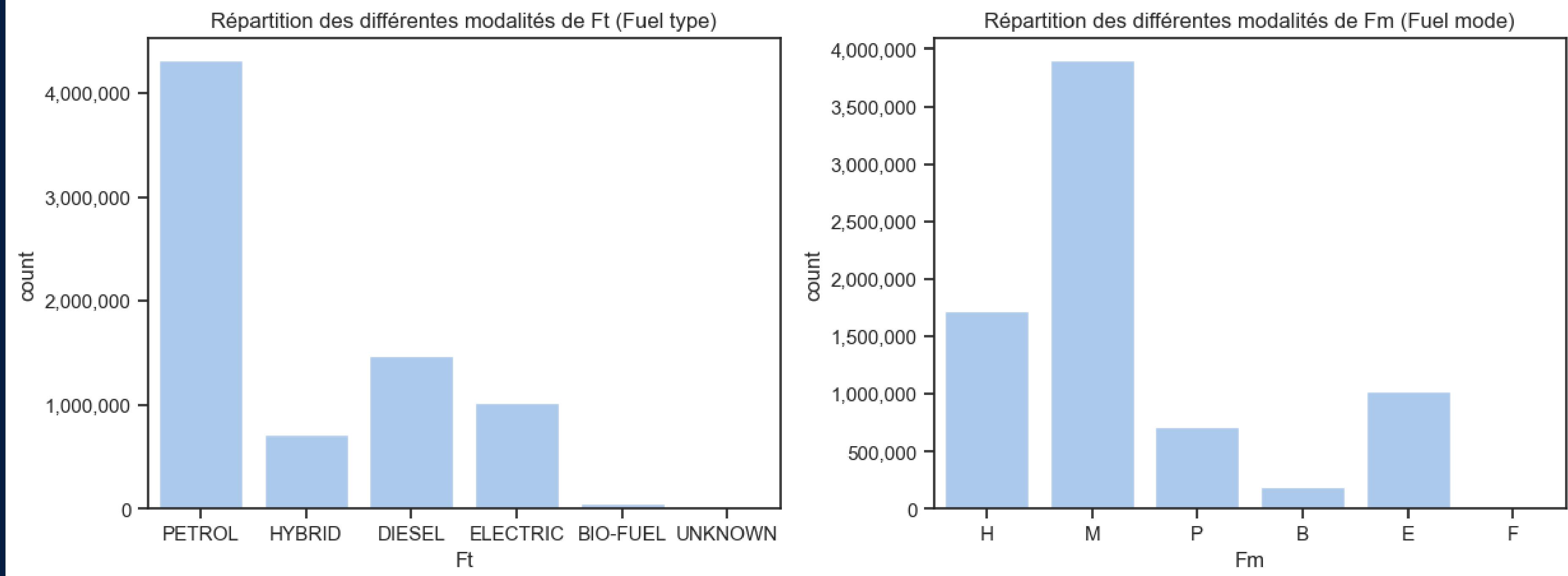


Target Overview



Ewltp (g/km)	
count	7571649.00
mean	340.28
std	183.04
min	-15.31
25%	317.60
50%	386.60
75%	438.21
max	2200.96

Analyse de Fuel Type (Ft) et Fuel Mode (Fm)



Analyse de Fuel Type (Ft) et Fuel Mode (Fm)

Ft	BIO-FUEL	DIESEL	ELECTRIC	HYBRID	PETROL
Fm					
B	1596	0	0	0	194948
E	0	0	1021754	0	0
F	11337	0	0	0	0
H	19920	298550	0	0	1405851
M	13889	1174086	0	0	2714016
P	0	0	0	715679	0

03 - PREPROCESSING

1- Récupération d'Observations

Récupérations d'observations de Fm en croisant les informations avec Ft qui n'a aucune valeurs manquantes

Si la voiture n'est pas électrique ou hybride il paraît logique de ne pas avoir d'autonomie électrique. On assigne la valeur 0 dans ce cas

Si la voiture est électrique il paraît logique de ne pas avoir de consommation d'essence. On assigne la valeur 0 dans ce cas

Si la voiture n'est pas électrique ou hybride il paraît logique de ne pas avoir de consommation électrique. On assigne la valeur 0 dans ce cas



Fm

Electric
range
(km)

Fuel
Consumption

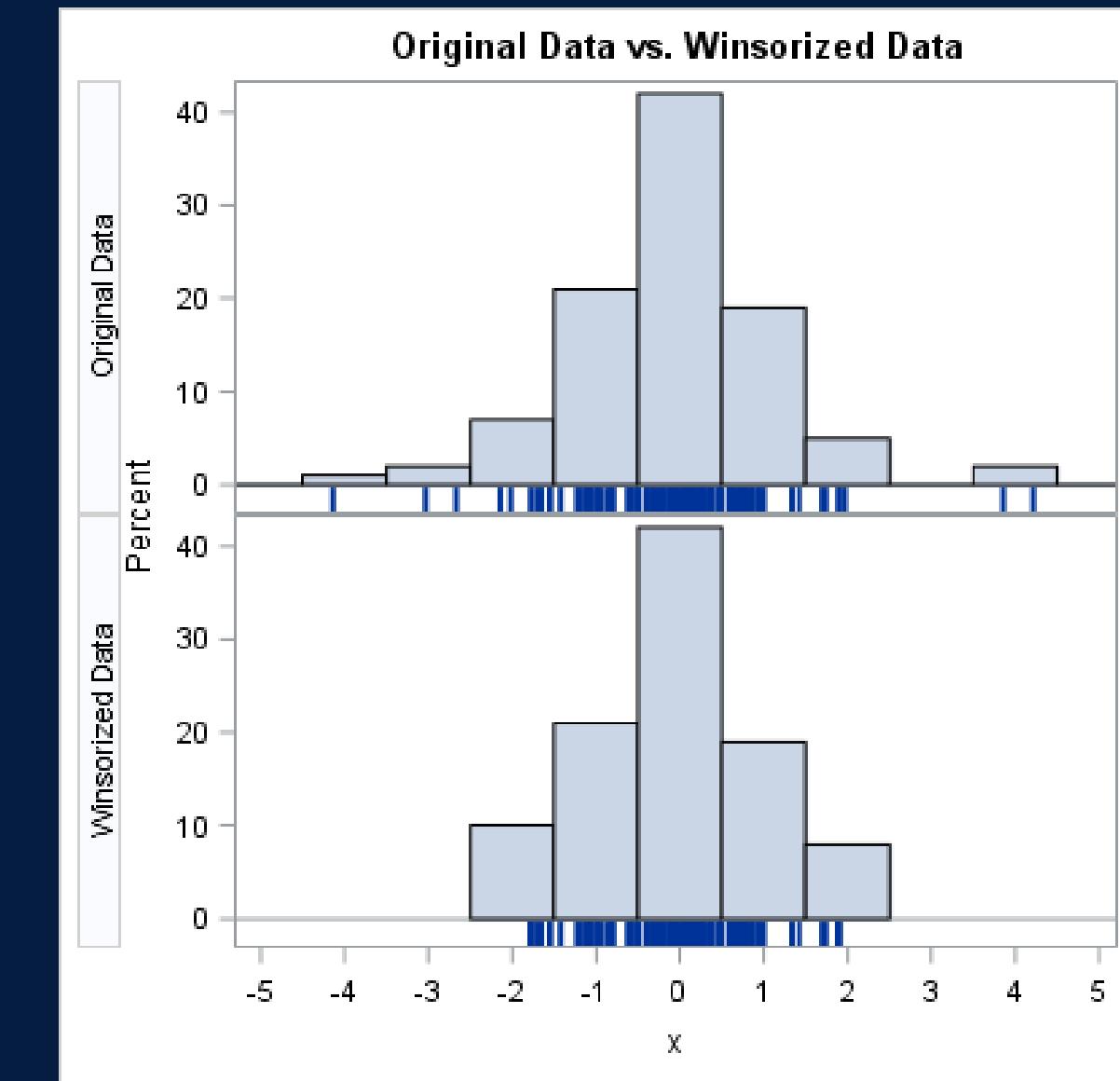
z
(Wh/k
m)

2 - Supprimer les colonnes inutiles

- Supprimer les colonnes avec 1 seul valeur unique (aucune info) ou 0 valeur unique (que des NaN)
- Supprimer les colonnes avec **+ de 50%** de NaN
- Supprimer les colonnes Date (pour train et test) et ID (seulement pour train)

3 - Traitement des outliers

Winsorization des outliers



4 - Imputation des valeurs manquantes (quantitatives)

$$\text{Coefficient de variation (col)} = \frac{\text{Var (col)}}{\text{Moyenne (col)}}$$

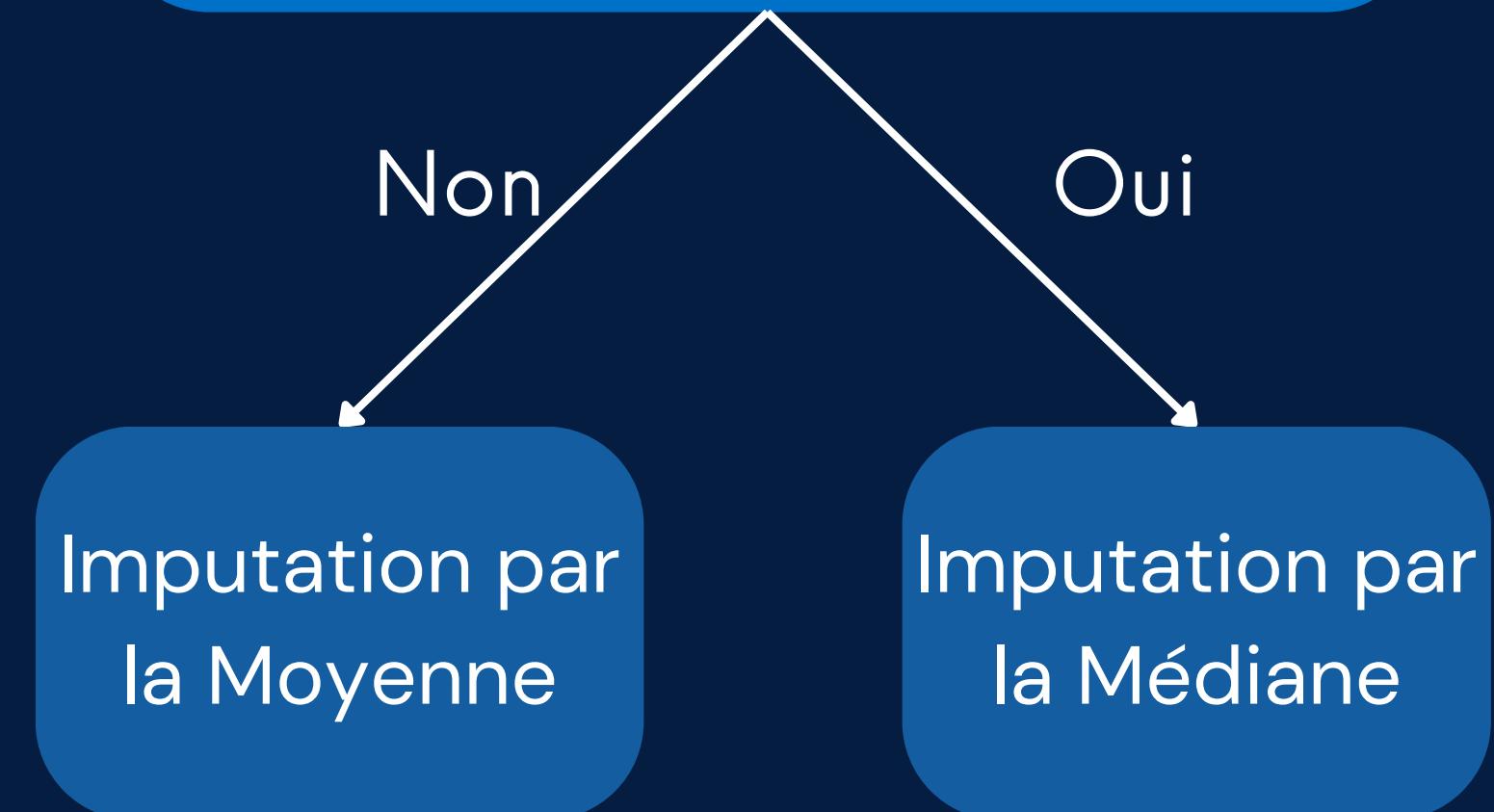
4.5 - Imputation des valeurs manquantes (qualitatives)

Imputation par le mode

5 - Encoding

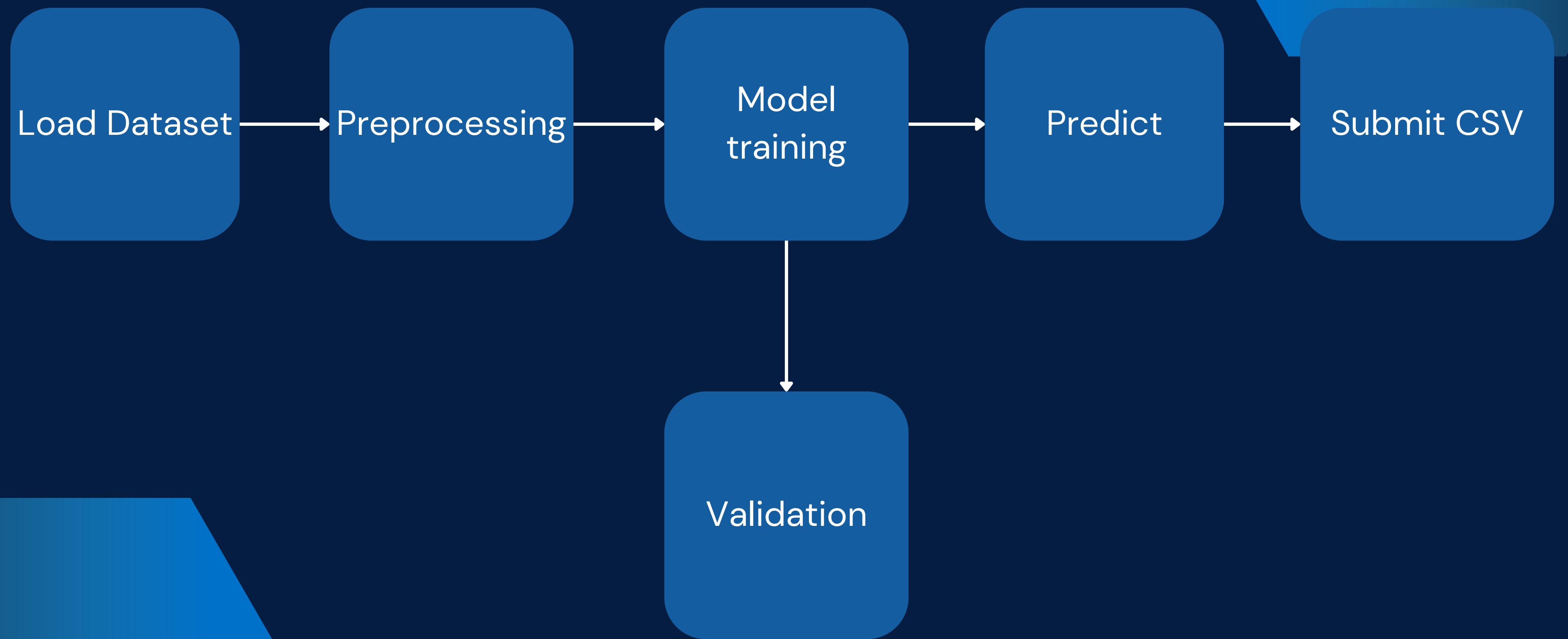
- Count Encoder ($n\text{unique} \geq 15$)
- One Hot Encoder ($n\text{unique} < 15$)

Coefficient de variation > 0.15



04 - PROCESSUS ET MODÈLES

1 - General Processing



2 - Baseline Model: Régression linéaire

- Prendre seulement les variables numériques
- Imputation par la moyenne

MAE: 43.28

3 - Random Forest / Bagging (Decision Tree)

- Prendre toutes les variables récupérables
- Encoding Ordinal

Afin de prendre en compte les relations non linéaires

4 - Catboost

- Prendre toutes les variables récupérables
- Encoding Ordinal

Passer au modèles de type boosting, plus performants



5 - Xgboost

- Preprocessing présenté plus tôt
- Hyper paramètres par Grid Search

6 - K-fold Xgboost

Afin de réduire le risque d'overfitting et stabiliser le modèle.

On pose k=15

05 - RÉSULTATS

Meilleurs modèles:

- Xgboost
- K-fold Xgboost



06 - AXES D'AMÉLIORATION

- Trouver un Feature Engineering qui puisse apporter de l'information (nécessite de la connaissance métier)
- Améliorer la récupération d'observations, aller plus en profondeur à l'aide des variables (comme ce qu'on a pu faire avec Ft et Fm)
- Mieux hyperparamétriser nos modèles (Grid Search plus grand encore)



Jynaldo Jeannot
Yoan Jsem

Merci pour votre attention

Ressources



[GitHub Repository](#)

