

Data Mining Project

Wandaloo cars

Réalisé par : OUSSAKEL Khadija

Encadré par : EL ASRI Ikram

Data visualization

Data preprocessing

Data exploration.

Model building prediction

Plan

Introduction

Prétraitement des données

Visualisation et exploration des données

Construction d'un modèle de prédiction

Conclusion

Introduction

La prédiction du prix des voitures par data mining est utilisée par les constructeurs automobiles, les concessionnaires et les sites de vente de voitures pour estimer les prix de vente futurs, et pour aider à la prise de décision dans les négociations de prix . Elle peut également être utilisée par les acheteurs de voitures pour déterminer si un prix de vente est raisonnable par rapport à la valeur de marché estimée de la voiture.



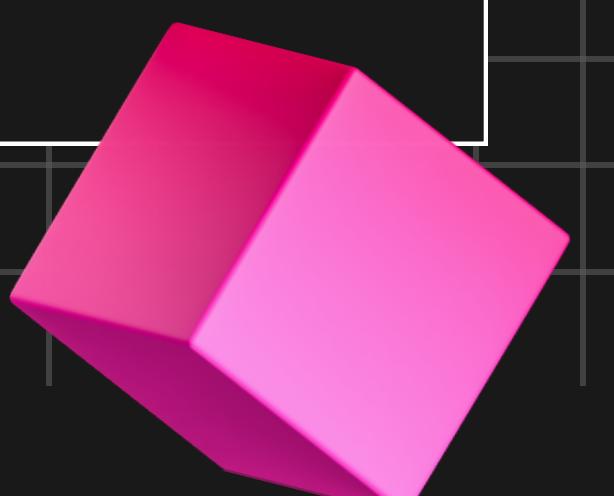
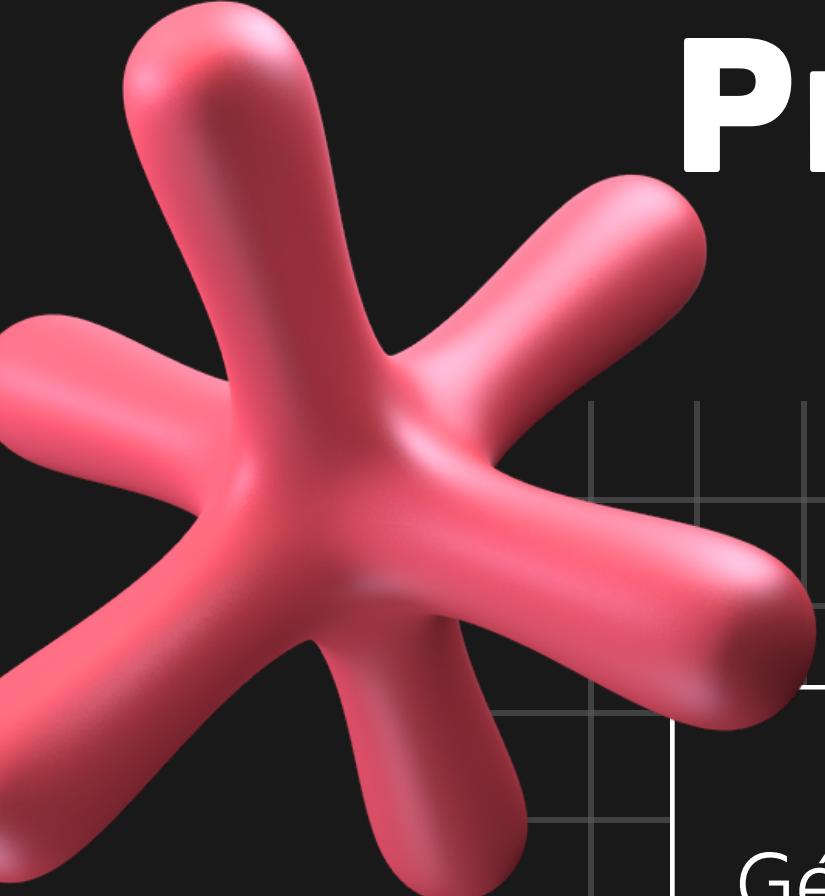
Site Wandaloo



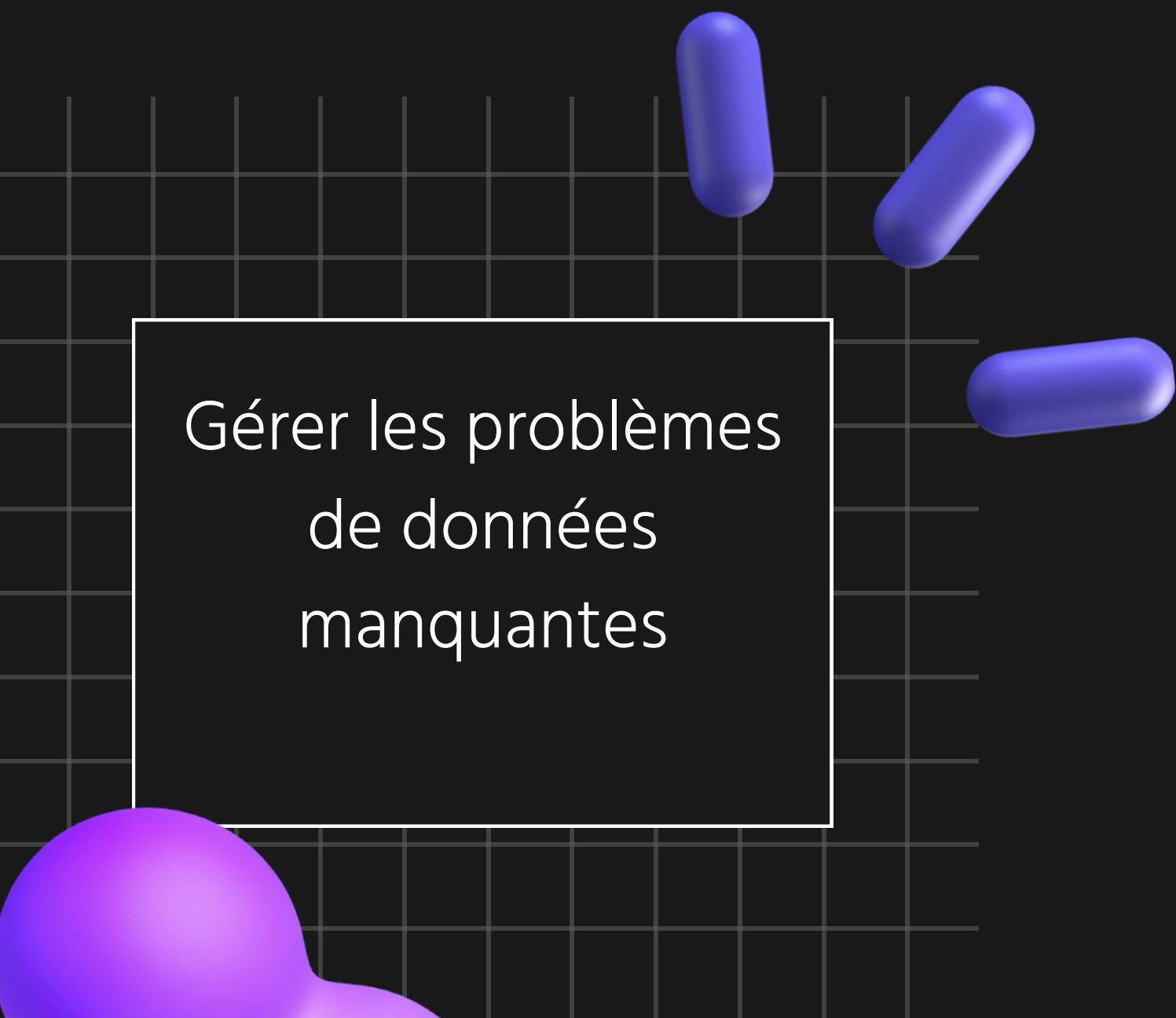
Wandaloo.com est un site web dédié au monde de l'automobile et de la mobilité.

Il propose une grande variété de contenus sur les voitures, les motos, les scooters, les vélos électriques et les autres moyens de transport.

Prétraitement des données

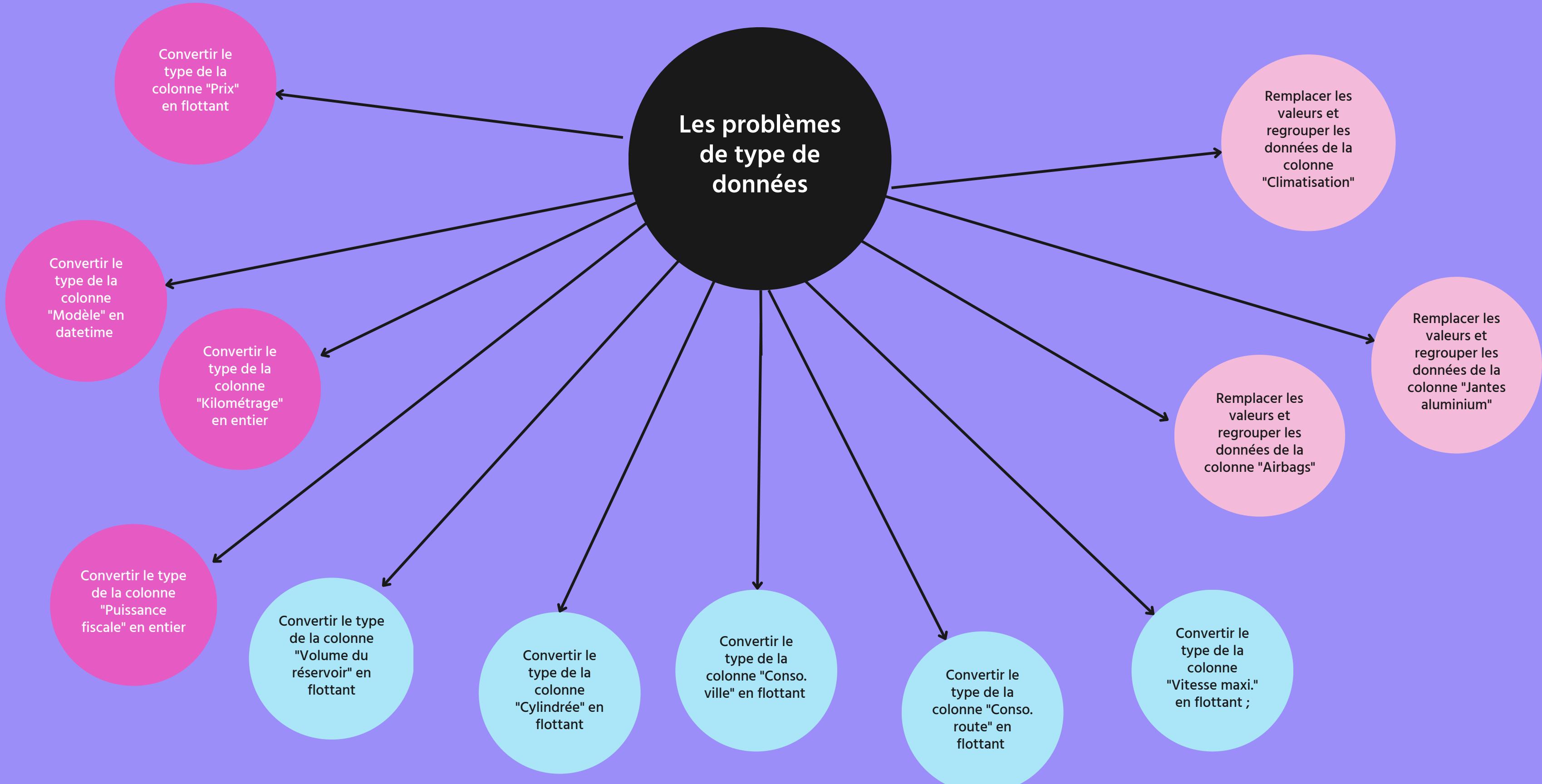


Gérer les problèmes
de type de données



Gérer les problèmes
de données
manquantes

Les problèmes de type de données



1

On supprime les colonnes nommées "Architecture", "Cylindrée en cm³", "Conso. ville en l/100km", "Conso. route en l/100km", "Vitesse maxi. en km/h" et "Volume du réservoir en Litre" du dataframe wandaloo_cars qui ont beaucoup de valeurs manquantes

3

Sur le fichier jupyter Notebook, le traitement de tous les données manquantes a été réalisé

Les problèmes de données manquantes

Traitement des données manquantes de la colonne 'Version'

Traitement des données manquantes des colonnes 'Main', 'Modèle' et 'Kilométrage'

Traitement des données manquantes de la colonne 'Transmission'

Traitement des données manquantes de la colonne 'Carburant'

Traitement des données manquantes de la colonne 'Puissance fiscale en cv'

Traitement des données manquantes de la colonne 'Couleur extérieure'

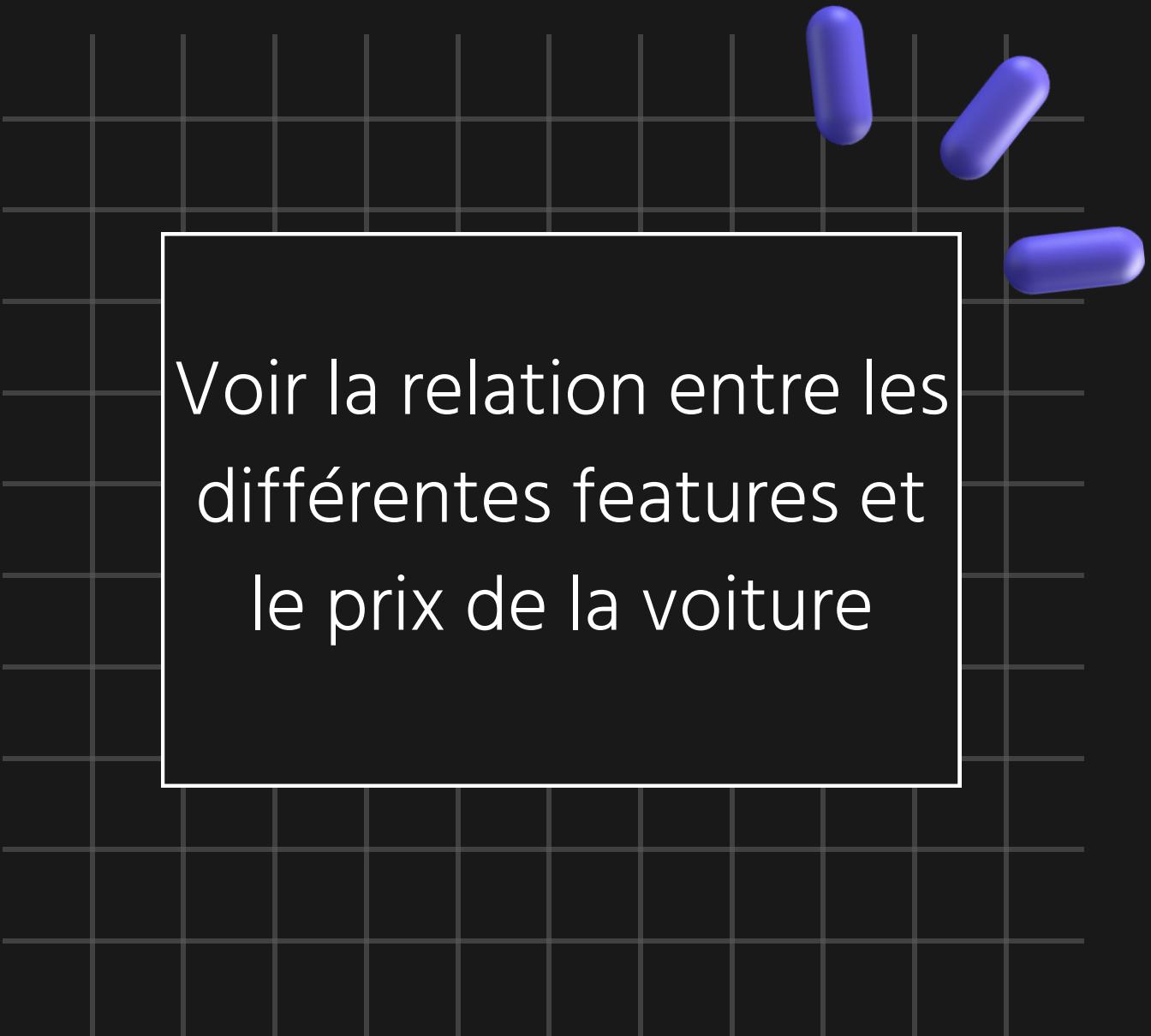
Traitement des données manquantes de la colonne 'Etat du véhicule'

Traitement des données manquantes de la colonne 'Climatisation'

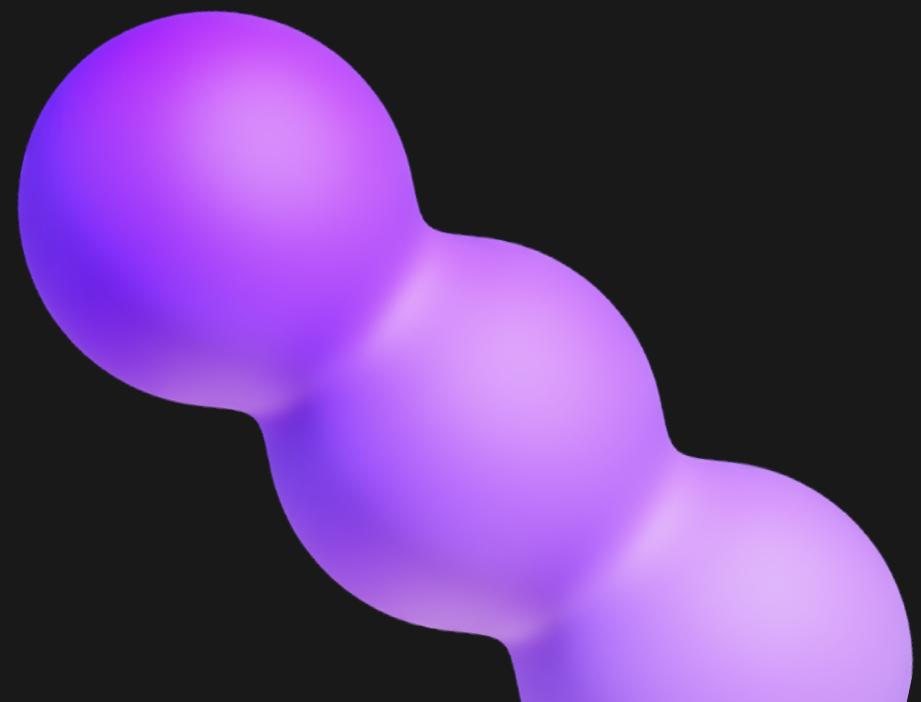
2

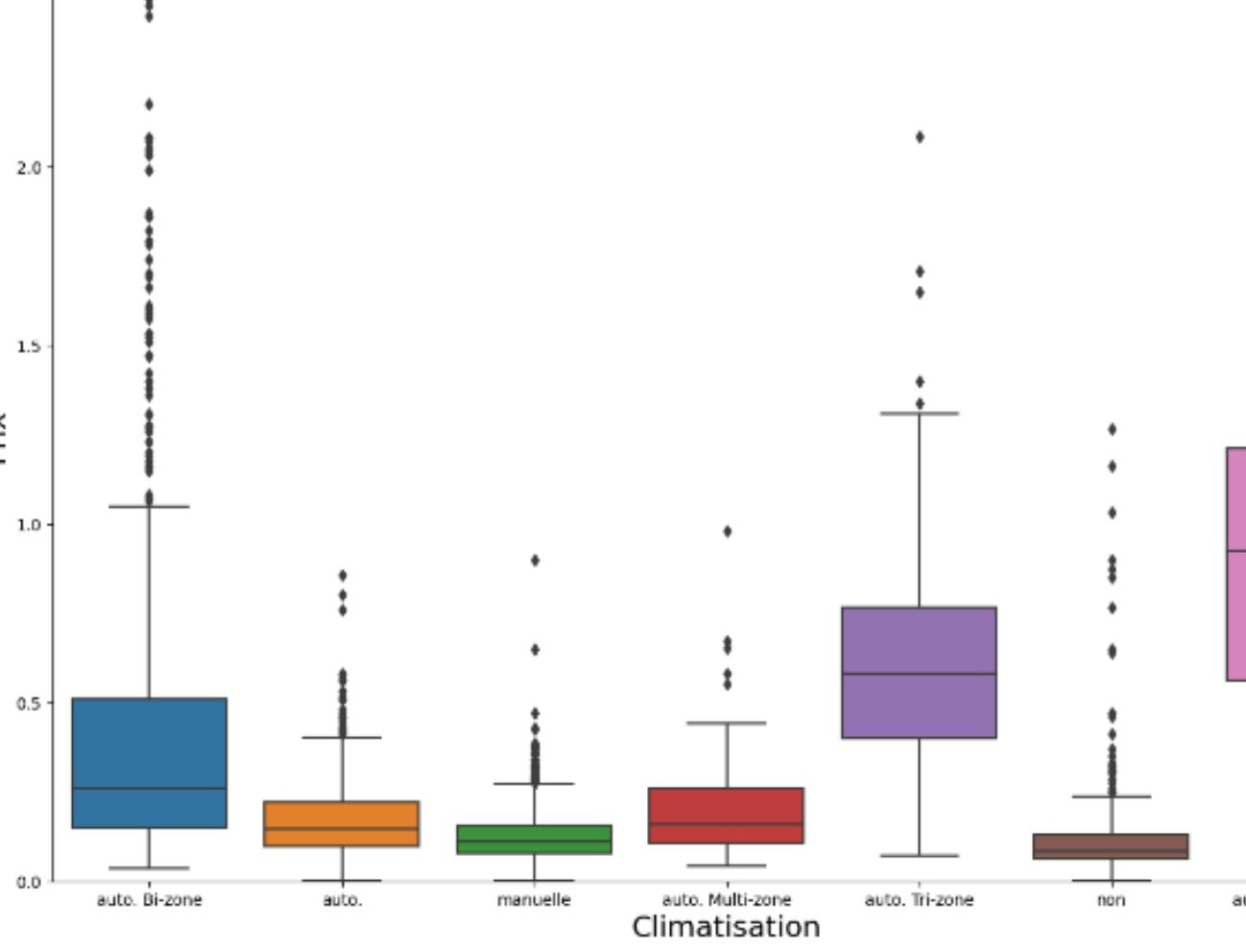
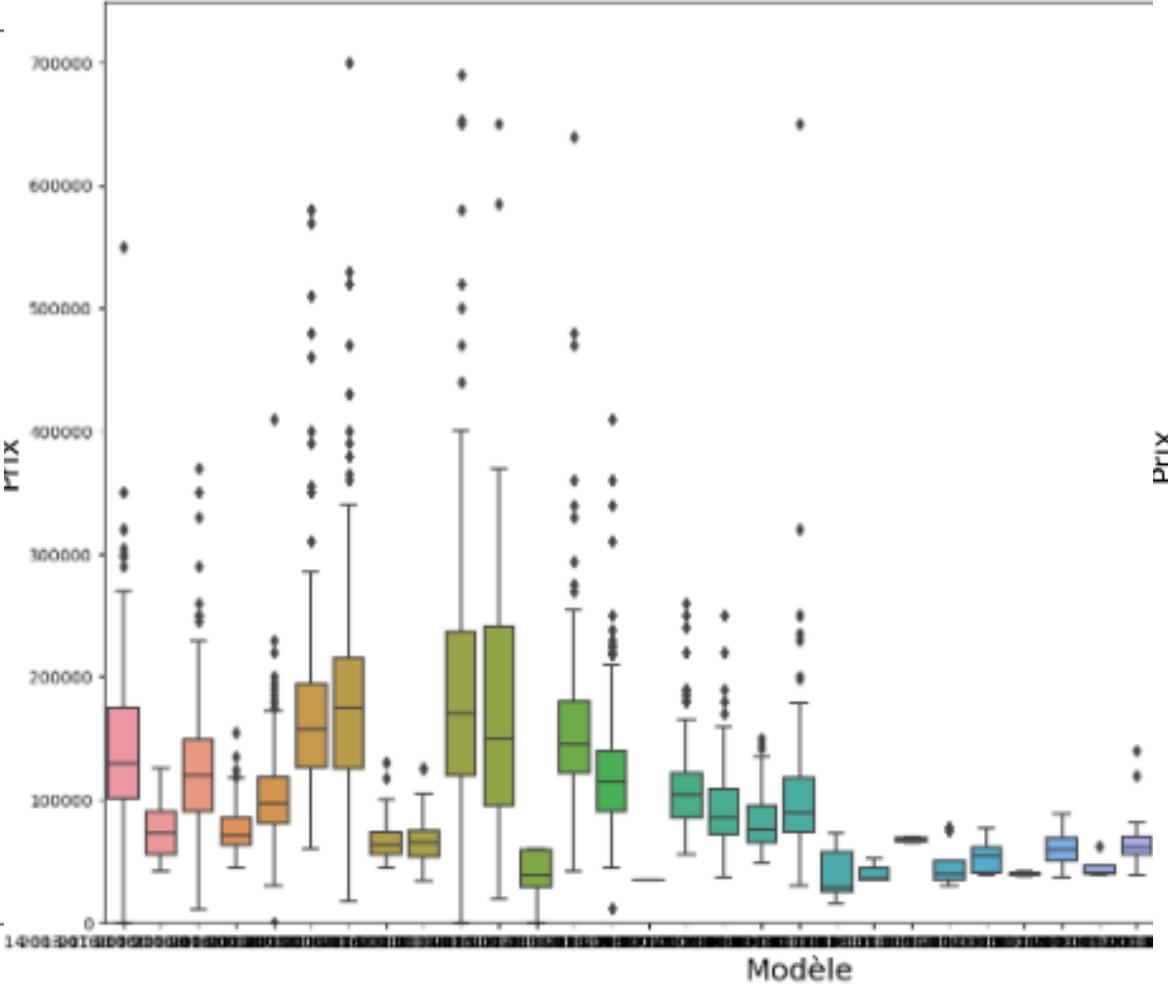
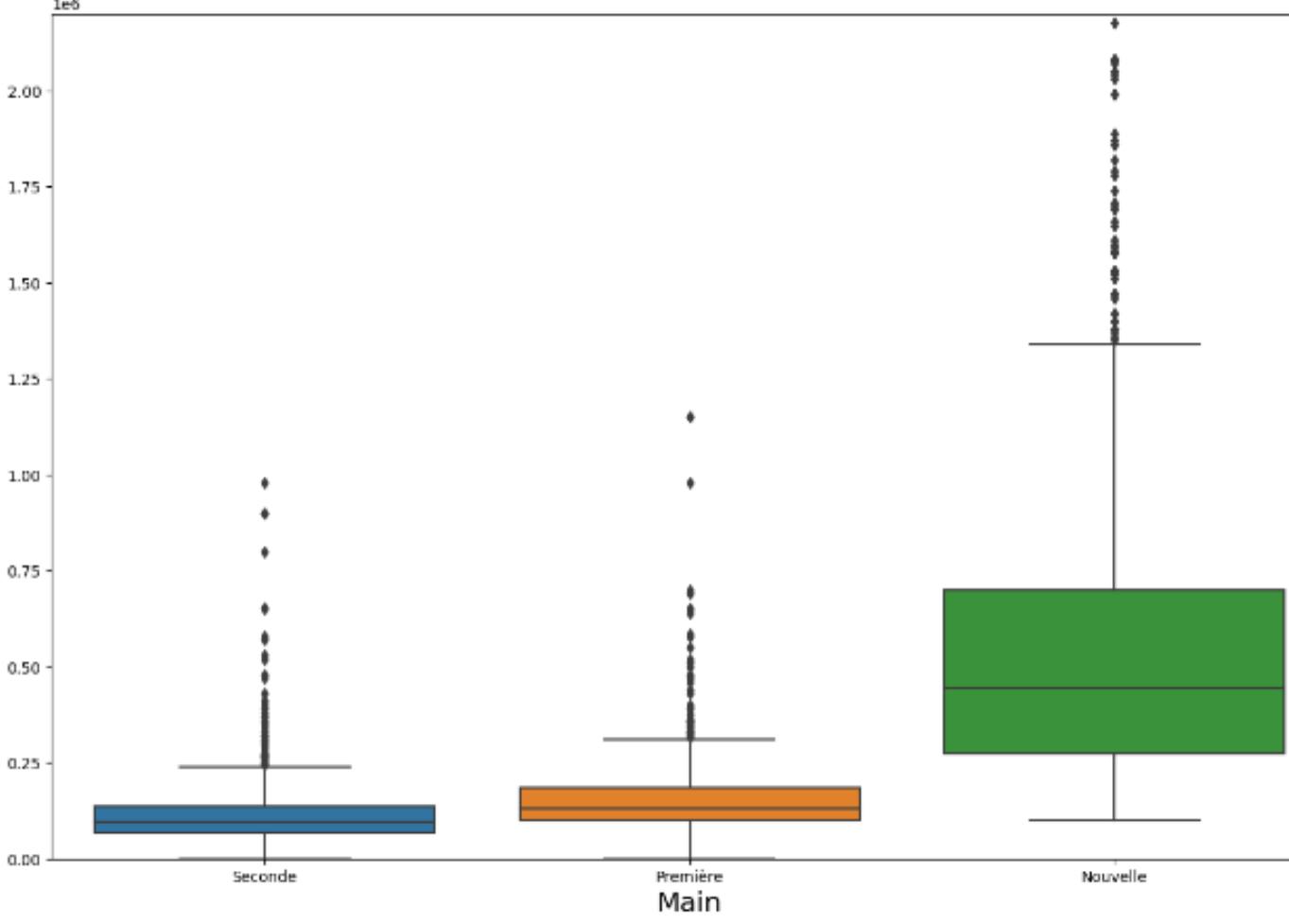
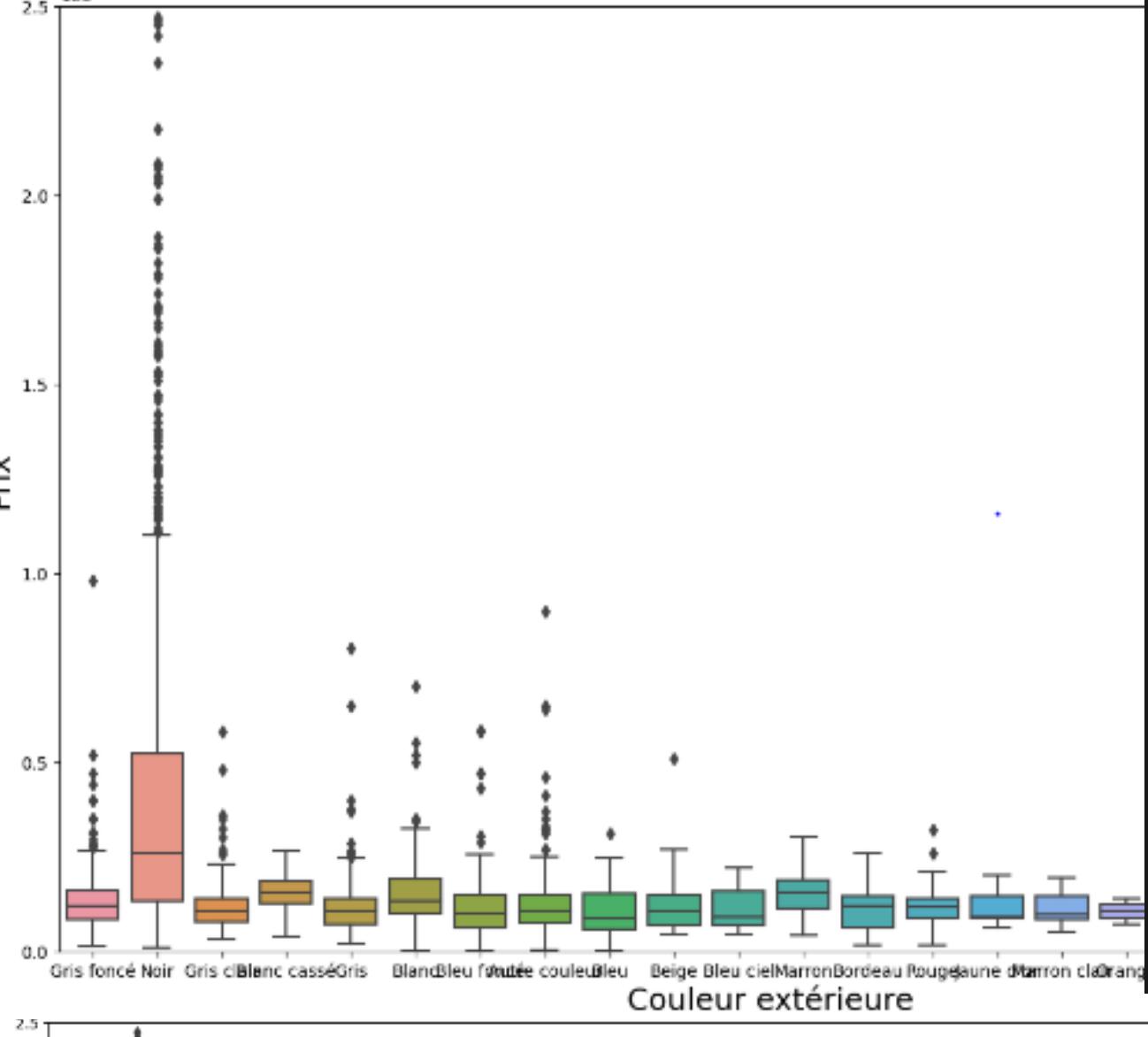
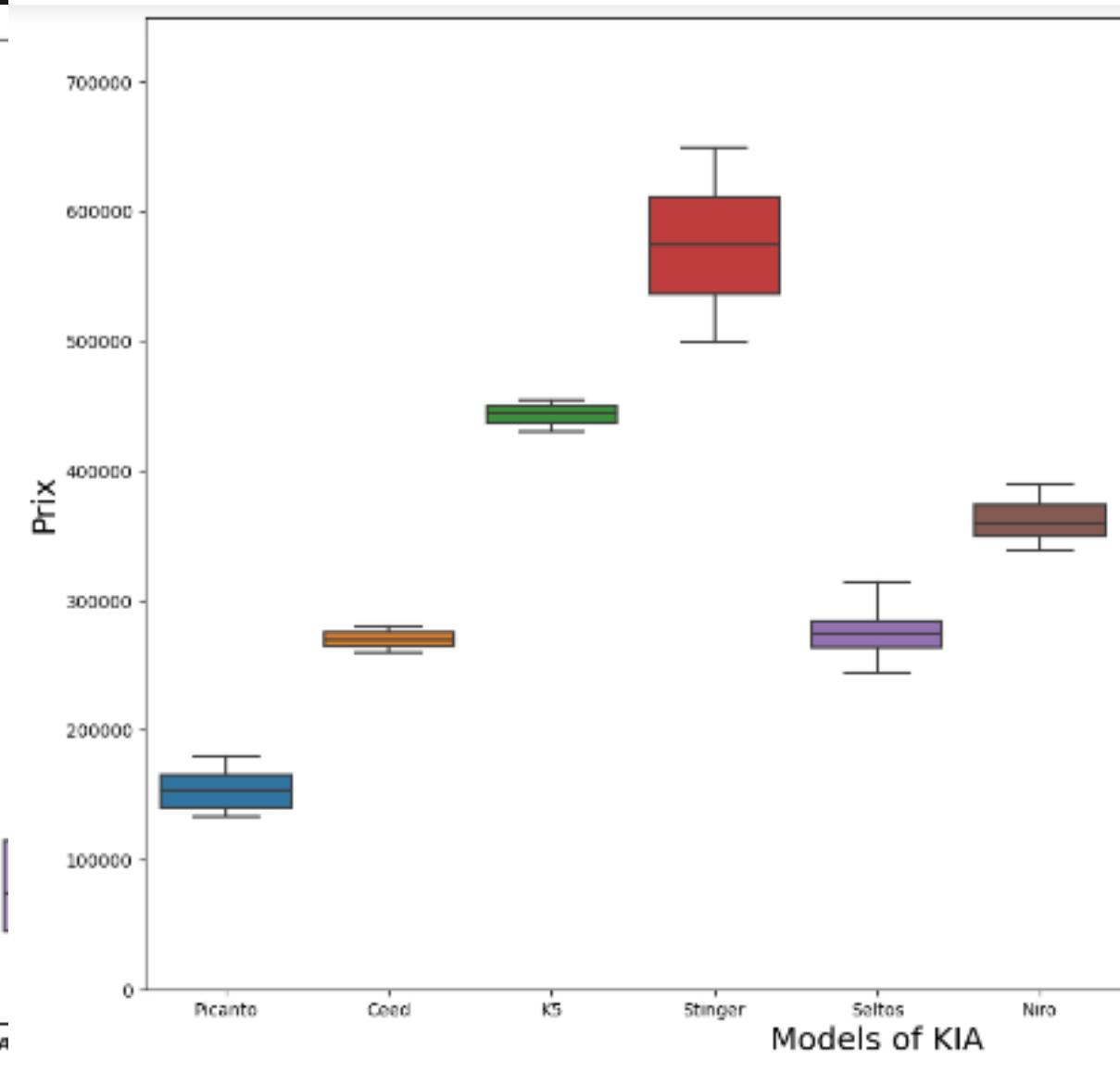
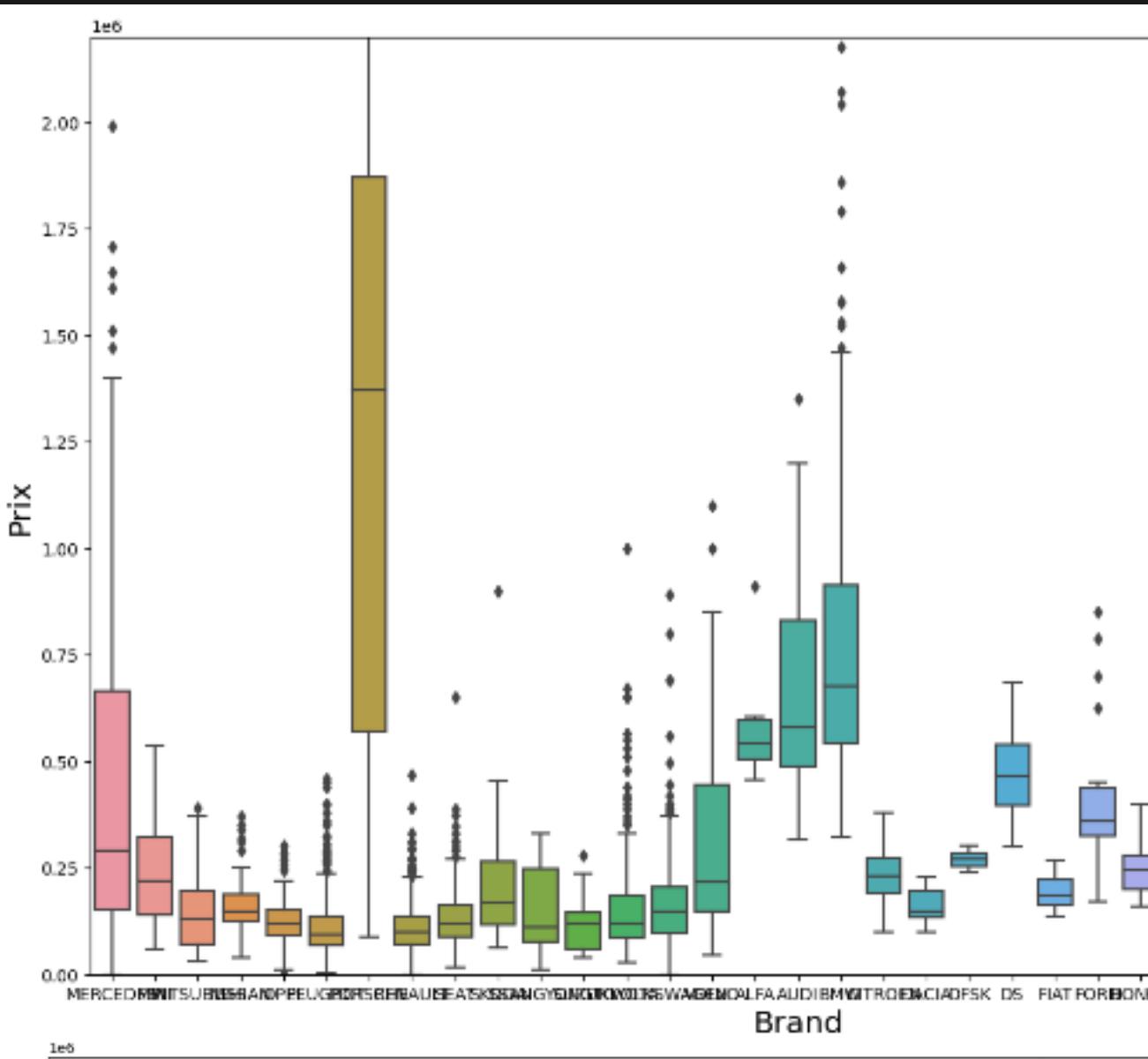
Les lignes qui n'ont pas de valeur Kilométrage n'ont aucune valeur dans les colonnes Main ou Modèle. Cela indique que ces lignes représentent de nouveaux éléments

Visualisation et exploration des données

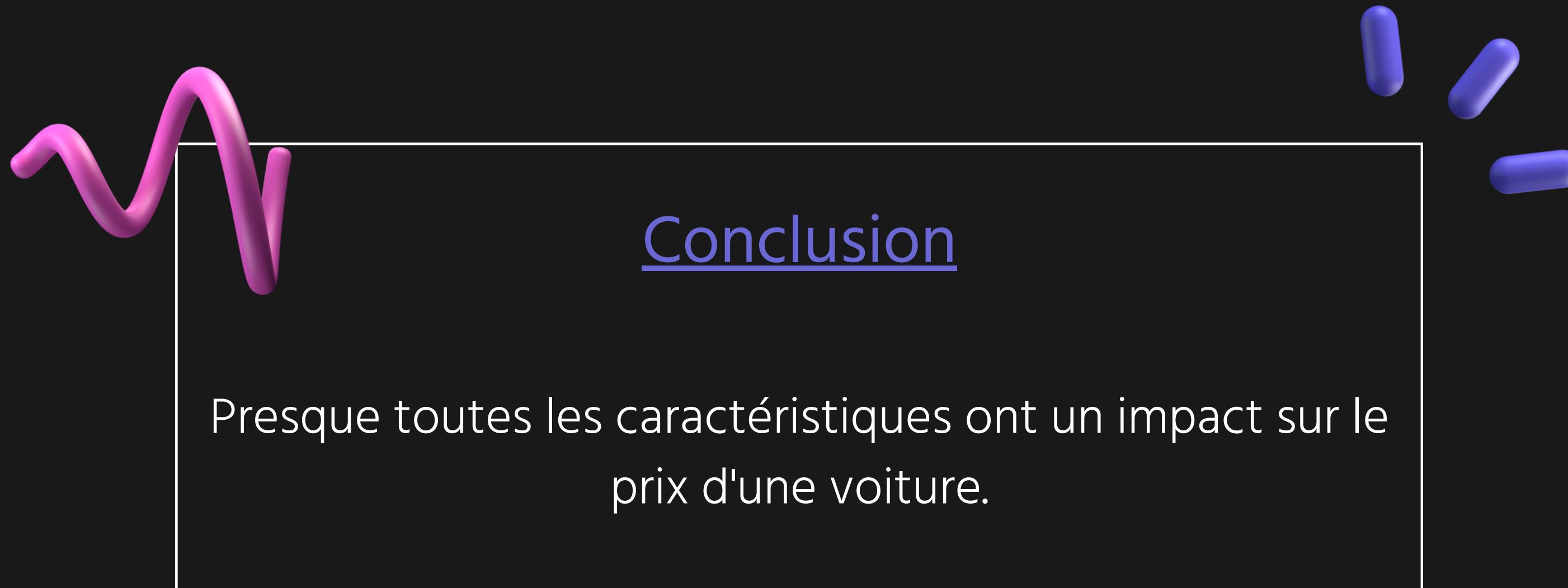


Voir la relation entre les différentes features et le prix de la voiture





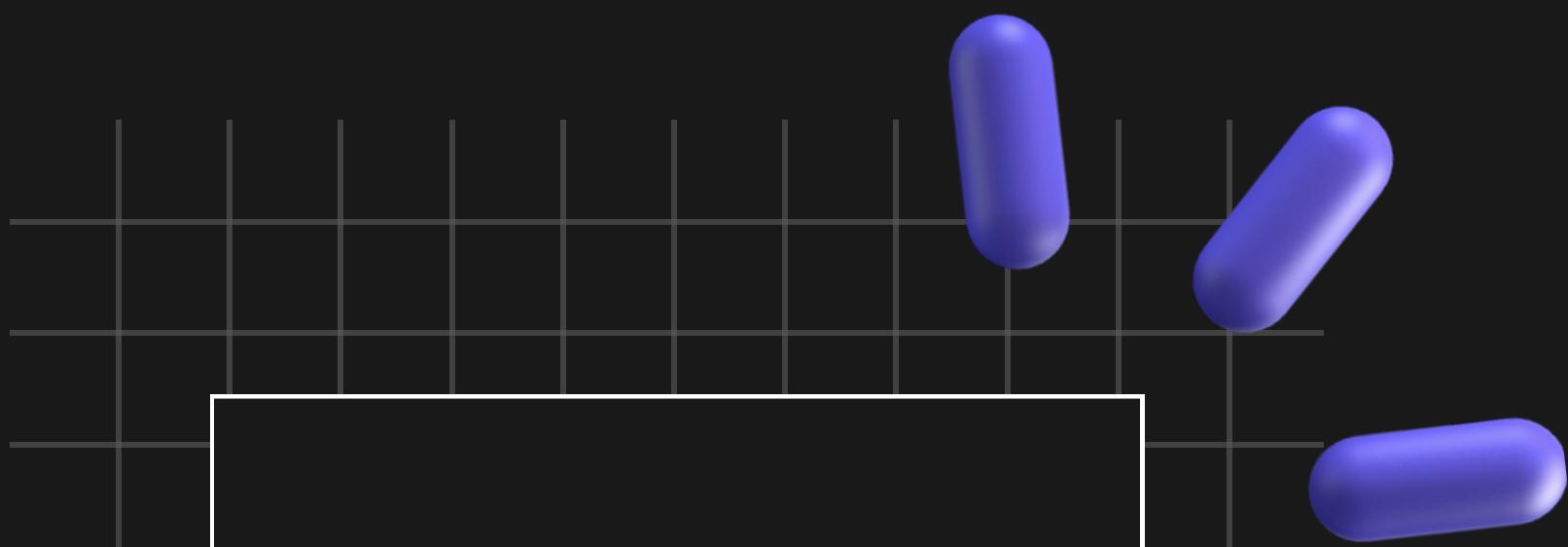
Visualisation et exploration des données



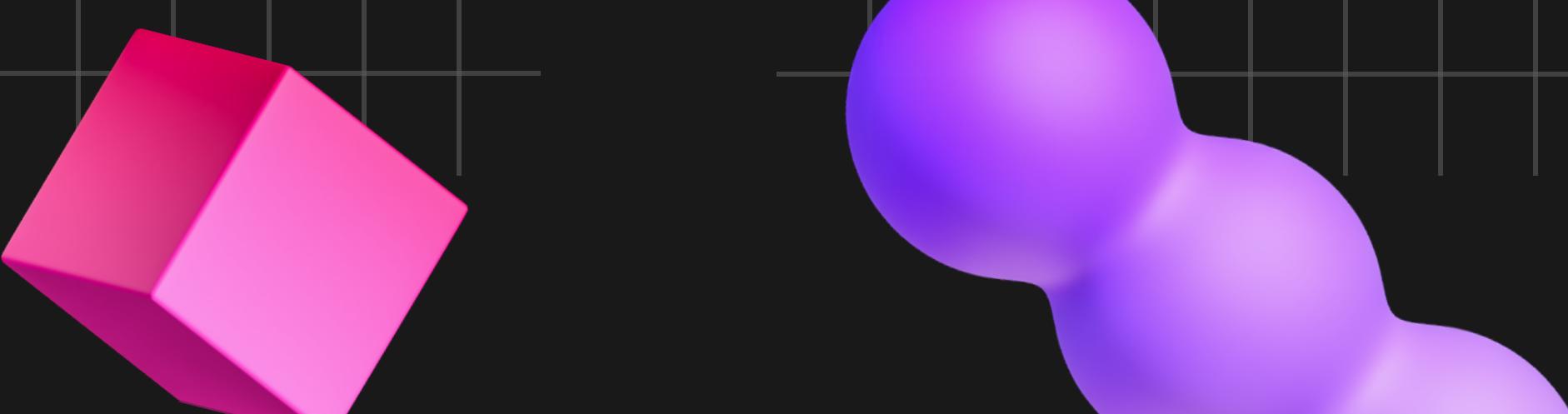
Construction d'un modèle de prédiction



Construire un modèle de prédiction pour prédire le prix des voitures en utilisant les caractéristiques restantes de la préparation des données.



Appliquer une sélection de fonctionnalités à l'ensemble de données nettoyé et reconstruire le modèle.



Construction d'un modèle de prédiction

Régression linéaire

score de performance
d'entraînement:
0.9883229191784942

score de performance de
test:
0.8483309846841154

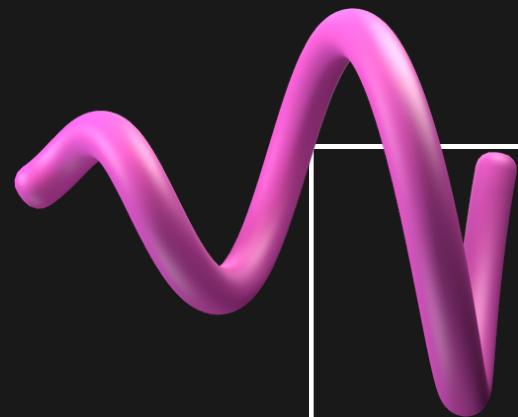
Construction d'un modèle de prédiction

Après sélection de fonctionnalités

score de performance
d'entraînement:
0.9819446586936844

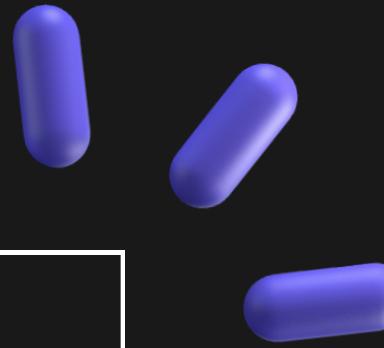
score de performance
de test:
0.8984241708303494

Construction d'un modèle de prédition

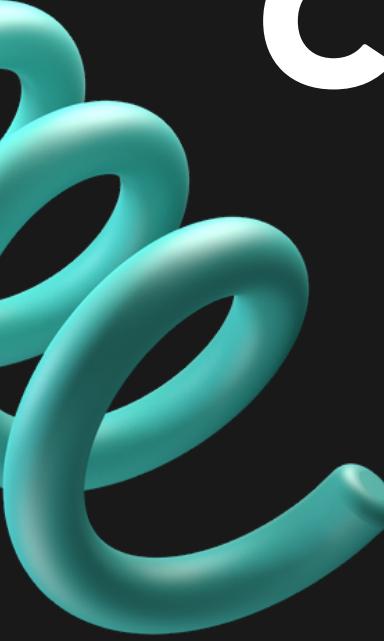


Conclusion

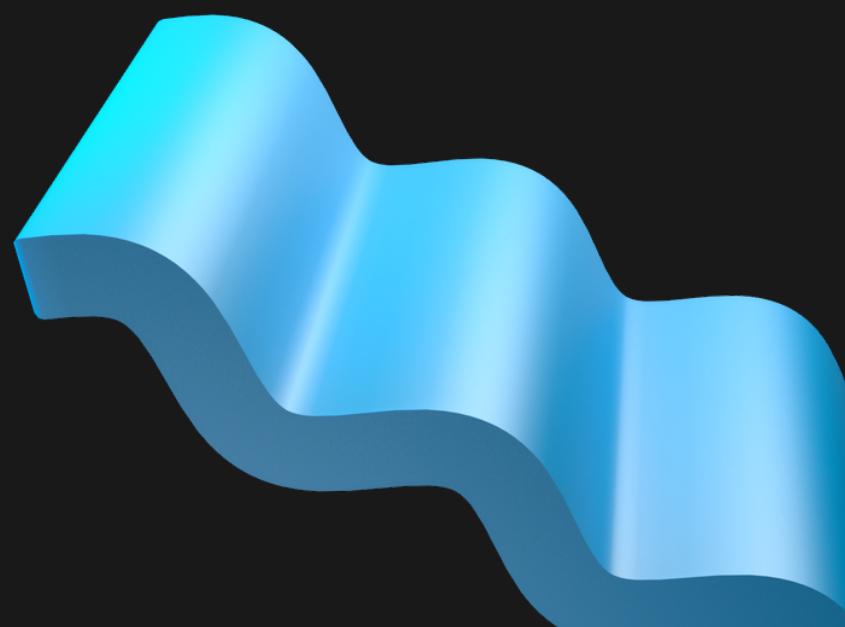
En évaluant et comparant les deux modèles, on conclut que le modèle avec sélection de fonctionnalités a mieux performé sur les données Test.



Conclusion



En conclusion, ce projet visait à prétraiter et analyser un ensemble de données extraites de Wandaloo.com sur les voitures, dans le but ultime de construire un modèle de prédiction pour estimer les prix des voitures. Tout au long du projet, nous avons diagnostiqué les données et identifié des problèmes de nettoyage tels que des valeurs manquantes, des doublons et des problèmes structurels. Nous avons ensuite visualisé et exploré les données.



Enfin, nous avons construit deux modèles de prédiction en utilisant les données prétraitées : l'un utilisant toutes les fonctionnalités et l'autre avec une sélection de fonctionnalités. Nous avons évalué et comparé les deux modèles, concluant que le modèle avec sélection de fonctionnalités a mieux performé.