كلية العلوم
FACULTÉ DES SCIENCES

جامعة مولاي إسماعيل
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵎⵓⵍⴰⵢ ⵉⵙⵎⴰⵄⵉⵍ
UNIVERSITÉ MOULAY ISMAÏL

Master

Intelligence Artificielle et Analyse des Données

Projet Concessionnaire Automobile

Data Mining

Préparé par :

AHANSAL Salah Eddine
EL KADAH Rachid
EL HAMDI Brahim

Professeur :

Mr. E. Zemmouri

1. Introduction

Dans le domaine de l'analyse des données, les informations relatives aux clients et aux véhicules jouent un rôle crucial pour mieux comprendre le comportement des utilisateurs et optimiser les stratégies commerciales. Les immatriculations des véhicules, combinées aux données sociodémographiques des clients, offrent une opportunité unique de segmenter la clientèle et de développer des modèles prédictifs. Ce projet s'inscrit dans cette optique, visant à tirer parti des données pour générer des insights exploitables.

Objectifs du projet :

a. **Analyse des données** : Explorer et comprendre les données issues des clients, du catalogue et des immatriculations tout en effectuant un nettoyage rigoureux pour garantir leur qualité et leur cohérence.
b. **Identification des facteurs clés** : Exploiter les corrélations entre les différentes variables afin d'identifier les facteurs les plus influents sur les résultats observés.
c. **Segmentation des données** : Appliquer l'algorithme de clustering K-Means pour identifier des groupes homogènes au sein des données.
d. **Modélisation prédictive** : Utiliser des modèles supervisés tels que les SVM, les arbres de décision et Random forest pour classifier et prédire l'appartenance à un groupe ou un comportement spécifique.

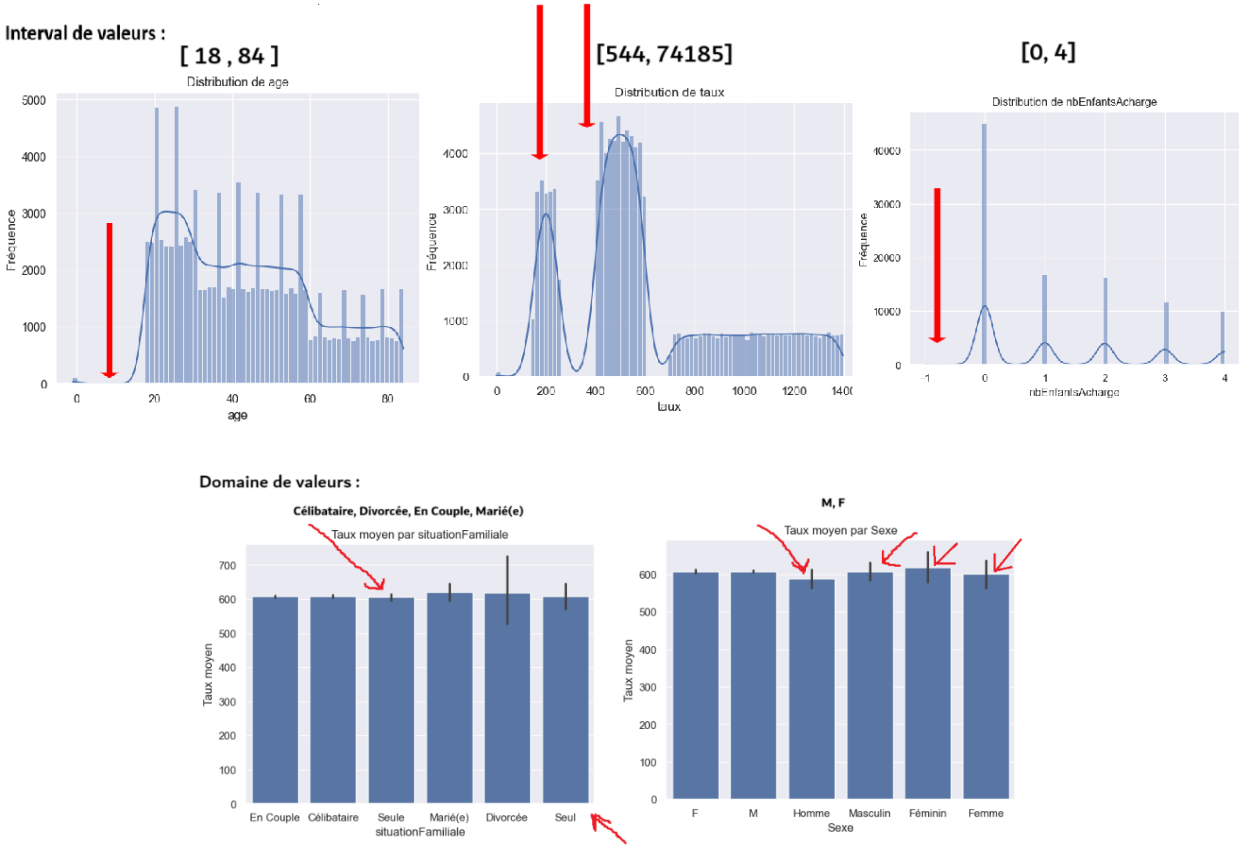
Problématique :

Aider le concessionnaire automobile à déterminer la catégorie de la voiture en exploitant les données collectées lors de la phase d'avant-vente.

2. Analyse exploratoire des données

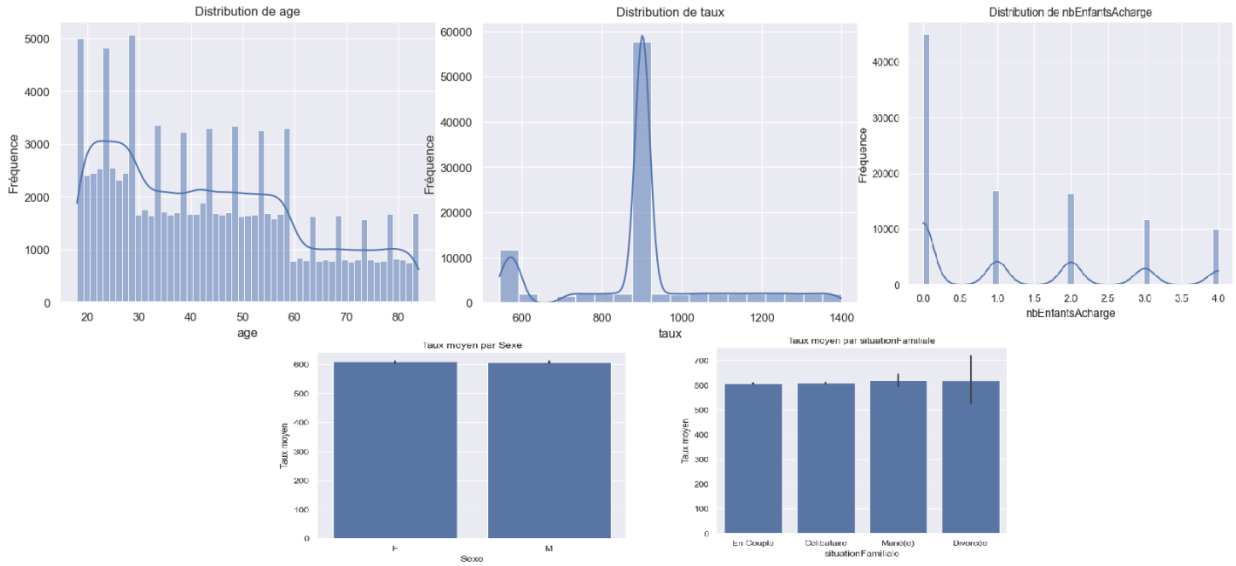
PARTIE 1 : CLIENTS

Tout d'abord, il est nécessaire de **nettoyer les données** en traitant les valeurs manquantes ou incorrectes, comme les erreurs de saisie. Pour gérer les valeurs manquantes, **les colonnes numériques** ont été complétées en remplaçant les valeurs manquantes par **la médiane**, tandis que pour **les colonnes catégorielles**, les valeurs manquantes ont été remplacées par **la modalité** la plus fréquente. Après cette étape, des valeurs incorrectes, situées hors des domaines définis par le dictionnaire de données, ont été identifiées :

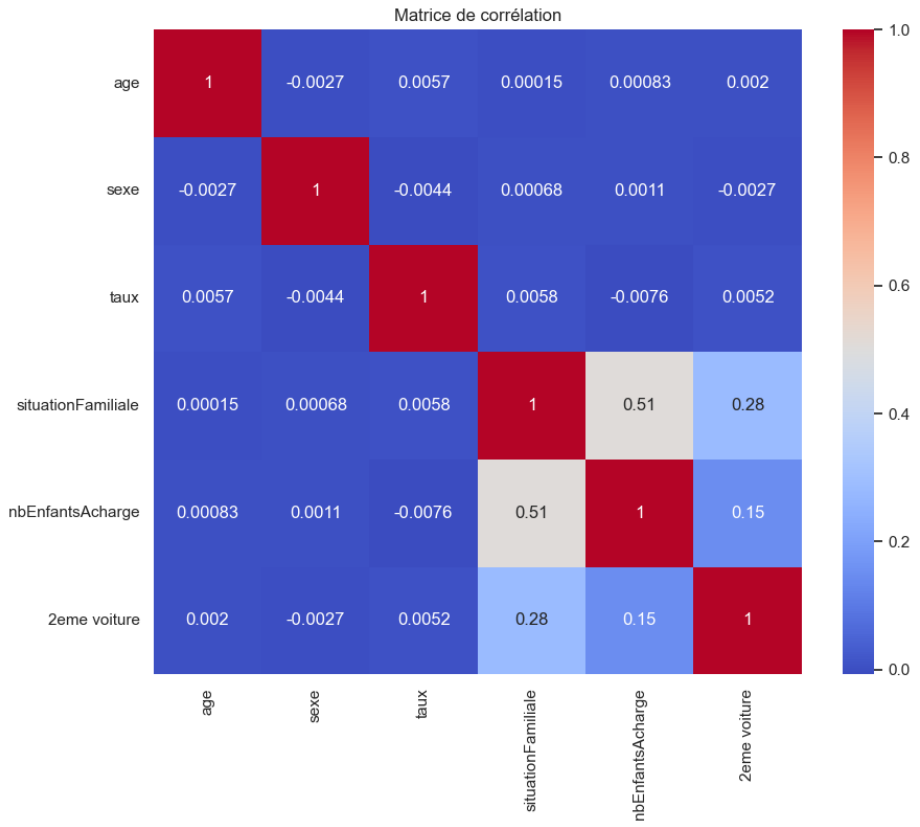


En parallèle du traitement **des valeurs manquantes**, nous avons corrigé les valeurs **incohérentes** ou **incorrectes** dans les données. Les données hors domaine ont été supprimées pour les colonnes telles que ***age*** et ***nbEnfantsAcharge***, car elles représentaient moins de **5%** des enregistrements. Cependant, pour la colonne ***taux***, qui comportait **55,82%** de valeurs incorrectes, une approche différente a été adoptée. **La moyenne** des valeurs valides (respectant les contraintes) a été calculée et utilisée pour remplacer les valeurs incorrectes.

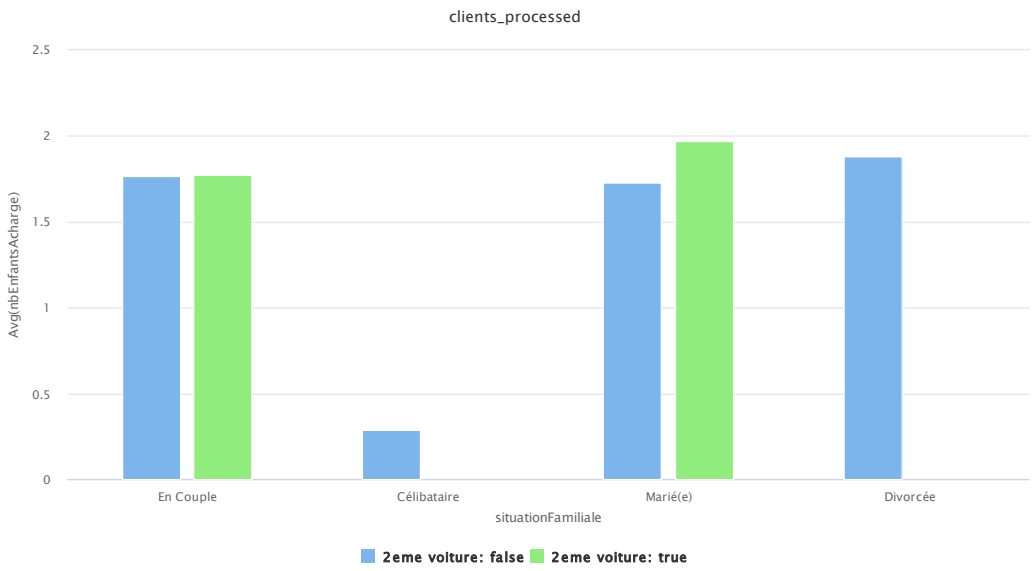
Pour les colonnes catégorielles, nous avons standardisé les données en remplaçant les valeurs par leurs **synonymes** pour garantir une cohérence. Par exemple, dans la colonne ***sexe***, les valeurs comme "Homme" et "Masculin" ont été uniformisées en "M", tandis que "Femme" et "Féminin" ont été converties en "F". De même, dans la colonne ***situationFamiliale***, des valeurs telles que "Seule" et "Seul" ont été remplacées par "Célibataire". Cette étape a permis d'harmoniser les données catégorielles tout en respectant leur sémantique.



Nous avons analysé les corrélations entre les attributs des clients, comme l'âge, le taux, la situation familiale, le nombre d'enfants à charge et la possession d'une deuxième voiture. Cela nous a permis d'identifier les relations importantes entre ces variables :

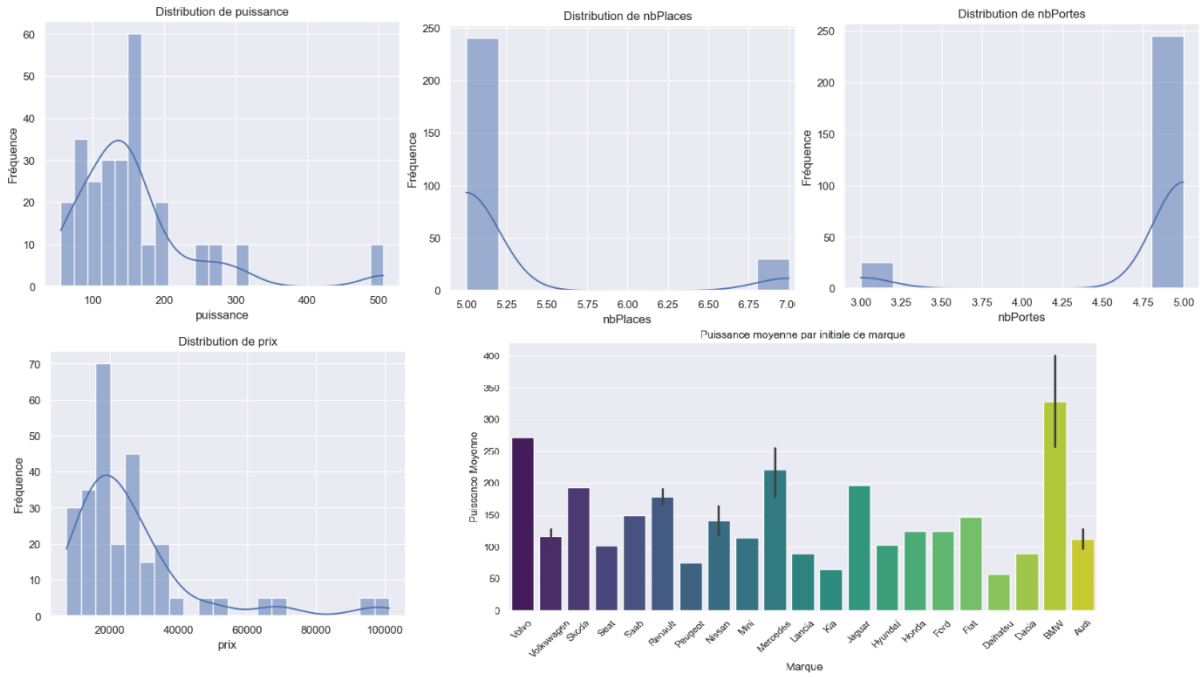


Les corrélations les plus fortes sont observées entre le nombre d'enfants, la situation familiale et la possession d'une deuxième voiture. Après avoir analysé ces colonnes, nous avons constaté que la moyenne du nombre d'enfants est très faible chez les célibataires, tandis que les personnes en couple et mariées sont celles qui possèdent le plus souvent une deuxième voiture, avec également un nombre d'enfants plus élevé.

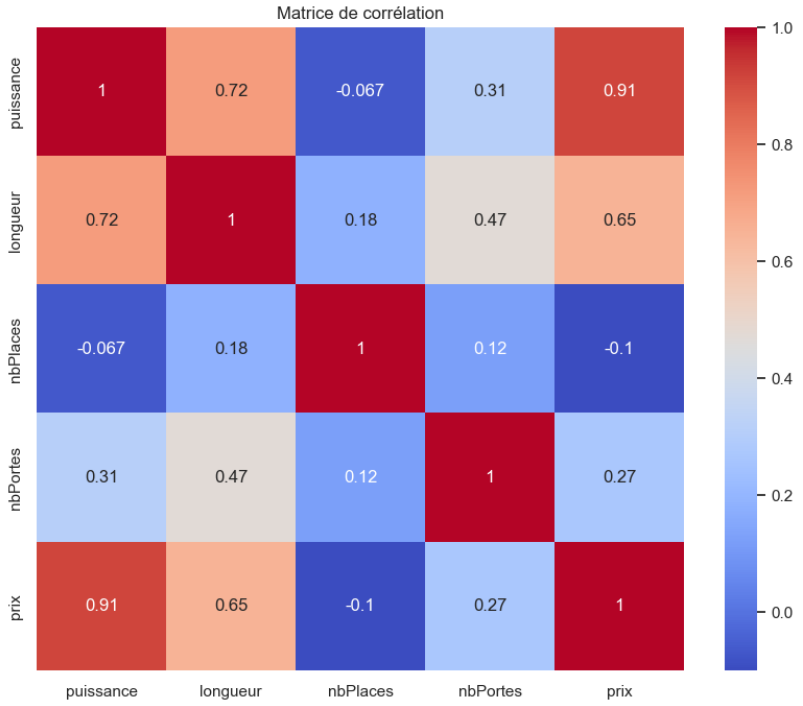


PARTIE 2 : CATALOGUE

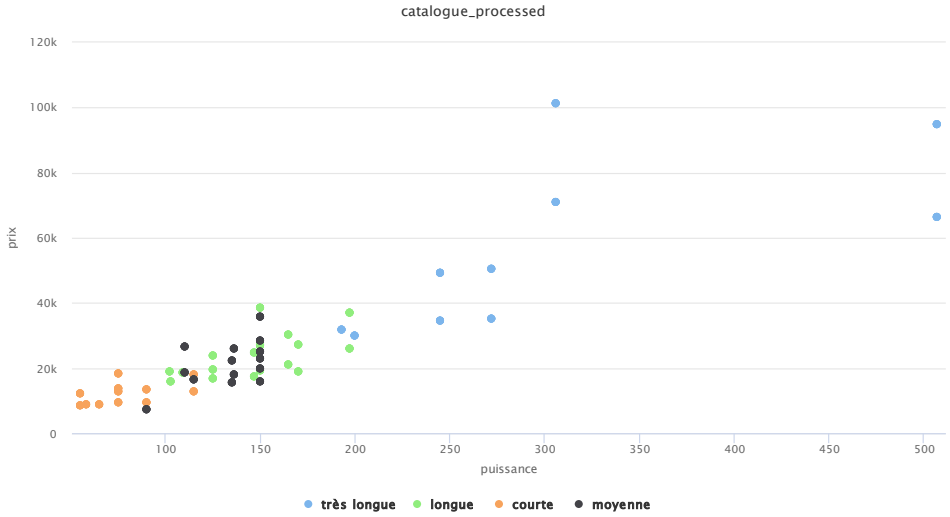
Concernant le catalogue, nous avons constaté que la distribution des données respecte bien les domaines de valeurs définis. Par conséquent, la distribution des attributs est cohérente avec les attentes :



Remarque : À partir de ce point, notre travail s'est concentré exclusivement sur les voitures neuves :



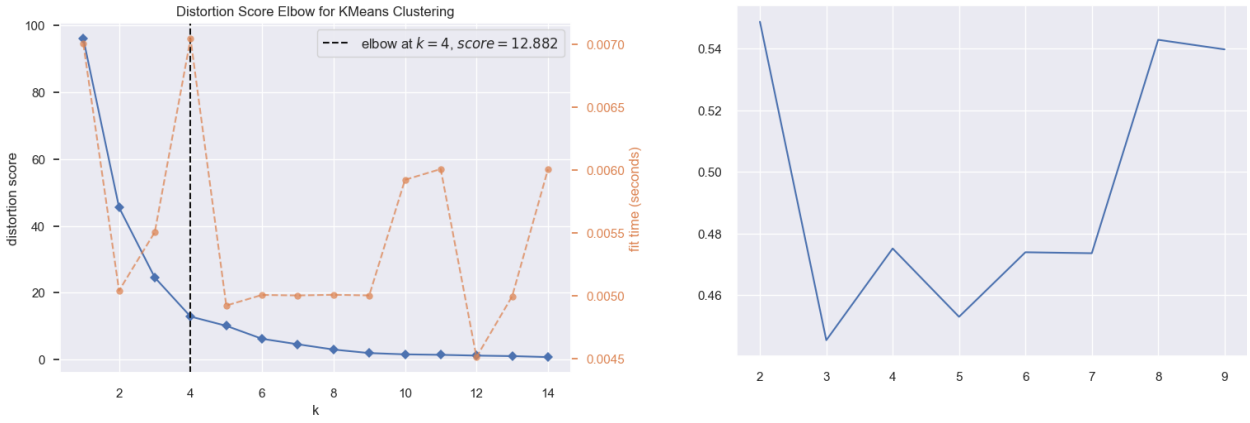
Les corrélations les plus fortes sont observées entre **la puissance**, **le prix** et **la longueur**. Après avoir analysé ces colonnes, nous avons constaté que **l'augmentation** de la puissance s'accompagne d'une **hausse** du prix. De plus, chaque fois que la longueur **augmente**, la puissance et le prix **augmentent** également.



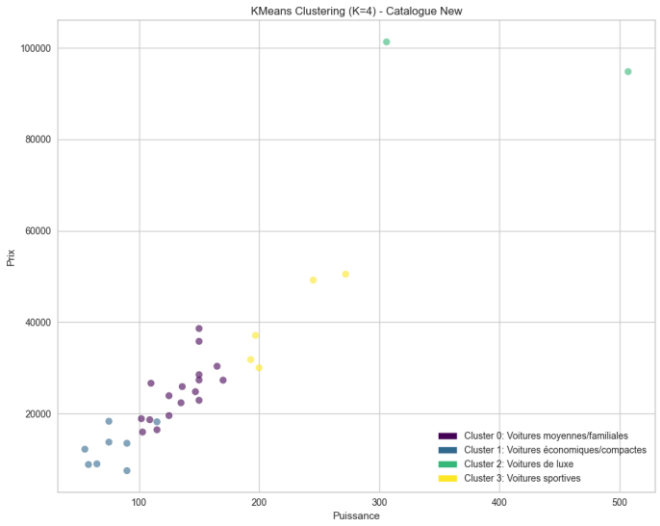
PARTIE 3 : IMMATRICULATIONS

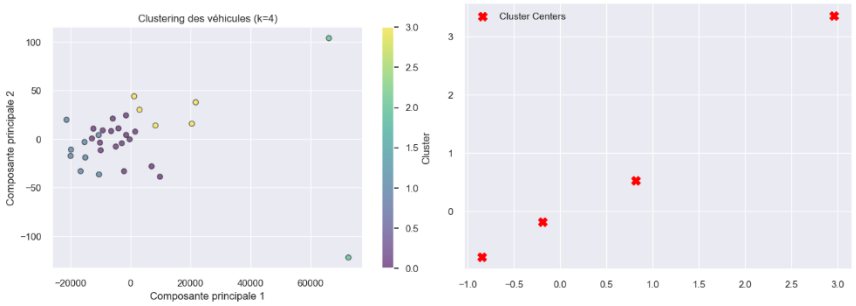
En ce qui concerne les immatriculations, nous avons observé que la distribution des données suit les mêmes règles que celles de la table du catalogue. De plus, la matrice de corrélation est presque identique.

3. *Clustering des données du catalogue*



D'après le graphique de gauche présente **la méthode du coude** utilisée pour déterminer le nombre optimal de clusters (k) dans l'algorithme **KMeans**. Dans le graphique, on observe que la distorsion diminue rapidement pour des valeurs de k entre 2 et 4, ce qui indique une forte amélioration dans la qualité de la modélisation. Cependant, au-delà de **k=4**, la diminution de la distorsion devient beaucoup moins marquée, formant ce qu'on appelle **un coude**, donc **k=4** est un choix optimal en termes de compromis entre la complexité du modèle et la qualité des clusters. En combinant cela avec le graphique de droite, **le coefficient de silhouette** pour **k=4** reste acceptable, bien que k=2 montre un meilleur score de silhouette. Cependant, le compromis entre la réduction de la distorsion et la séparation des clusters suggère que **k=4** peut être le choix optimal dans cette situation.

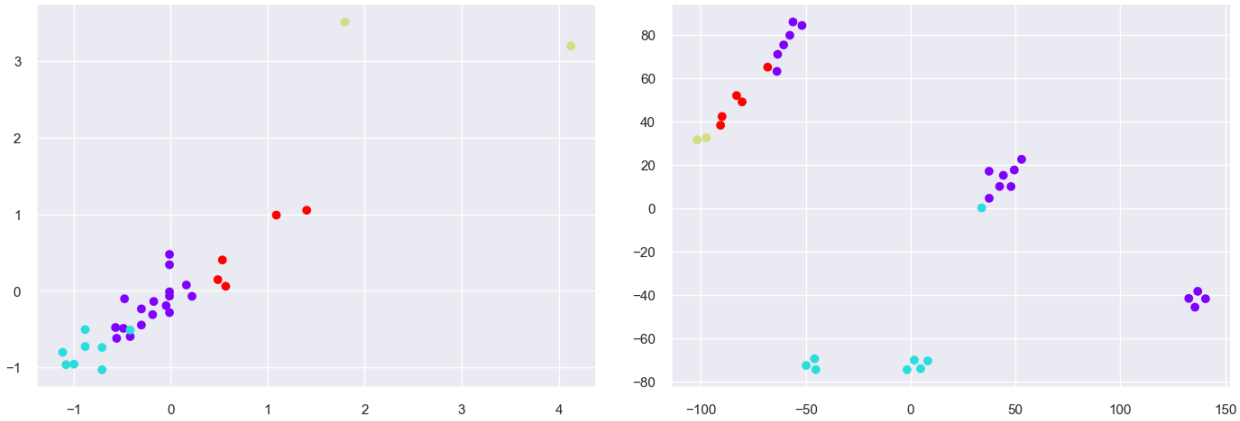




Ces graphiques présentent une analyse visuelle détaillée des résultats du clustering **KMeans** avec k=4, basé uniquement sur les caractéristiques principales : '**puissance**', '**longueur**' et '**prix**'. Les véhicules sont segmentés en quatre groupes distincts, selon les catégories définies :

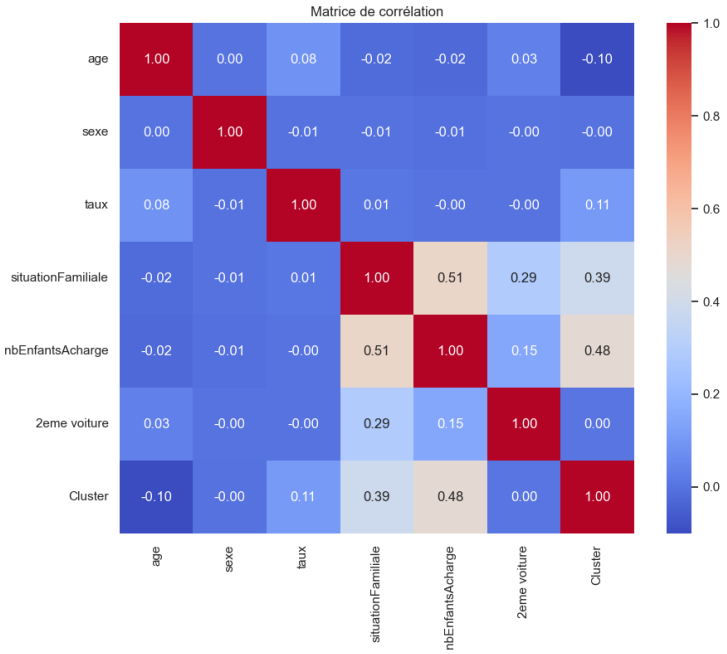
- **Cluster 0 :** Voitures moyennes/familiales,
- **Cluster 1 :** Voitures économiques/compactes,
- **Cluster 2 :** Voitures de luxe,
- **Cluster 3 :** Voitures sportives.

Ces visualisations permettent d'apprécier comment l'algorithme a identifié des patterns significatifs dans les données pour regrouper les véhicules selon leurs attributs spécifiques



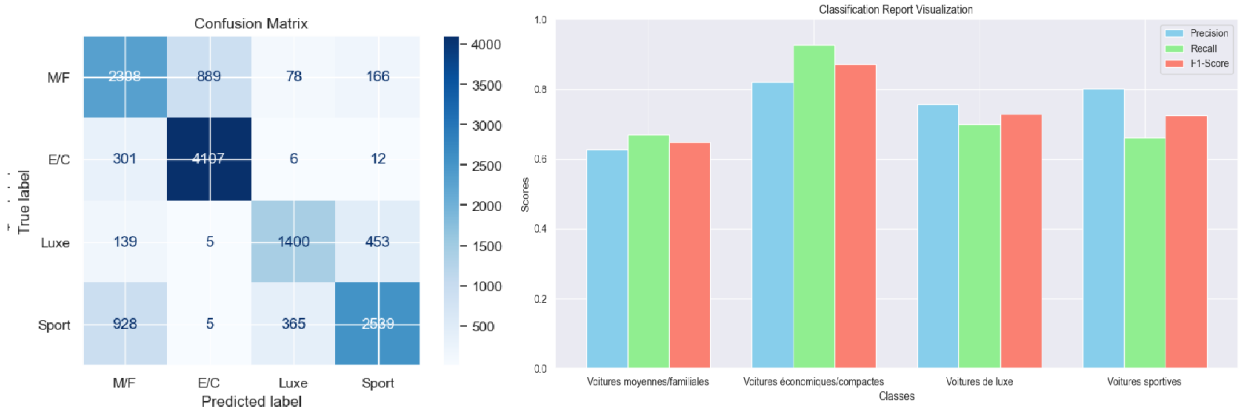
Les deux graphiques montrent que l'algorithme K-Means a bien segmenté les données en quatre clusters distincts. La projection PCA (à gauche) permet de visualiser comment les clusters se forment dans un espace réduit, tandis que la deuxième projection (à droite) montre une séparation encore plus nette des clusters. Ces résultats confirment la capacité de l'algorithme à regrouper les véhicules en fonction de leurs caractéristiques principales.

4. Classification



Pour l'analyse de la corrélation dans la nouvelle table résultant de la fusion des données **clients** et **immatriculations**, il est crucial de se concentrer sur les attributs des **clients** et les **clusters**, car ce sont ces informations qui serviront à prédire **la catégorie de véhicule**. Lors de cette **fusion**, les caractéristiques des clients sont associées aux données des véhicules, créant ainsi une base de données plus riche pour effectuer les suggestions. Cependant, l'attribut ***"sexe"*** a été écarté de l'analyse, car sa corrélation avec les clusters de véhicules est jugée **faible**. En se focalisant sur des attributs plus pertinents tels que l'***âge***, le ***taux***, la ***situation familiale*** et la présence d'une ***deuxième voiture***, nous garantissons une meilleure précision dans la prédiction de la catégorie de véhicule.

Le modèle choisi pour la prédiction est le **Random Forest**, qui a montré une excellente performance avec **Accuracy** de **75.57%**



5. Application

Dans cette étude, nous avons conçu une interface utilisateur intuitive pour faciliter le processus de recommandation de véhicules.

La figure ci-dessous illustre cette interface graphique conviviale, pensée pour permettre à l'utilisateur de renseigner efficacement ses informations personnelles. Elle intègre plusieurs champs spécifiques destinés à collecter les données essentielles pour proposer des recommandations personnalisées.

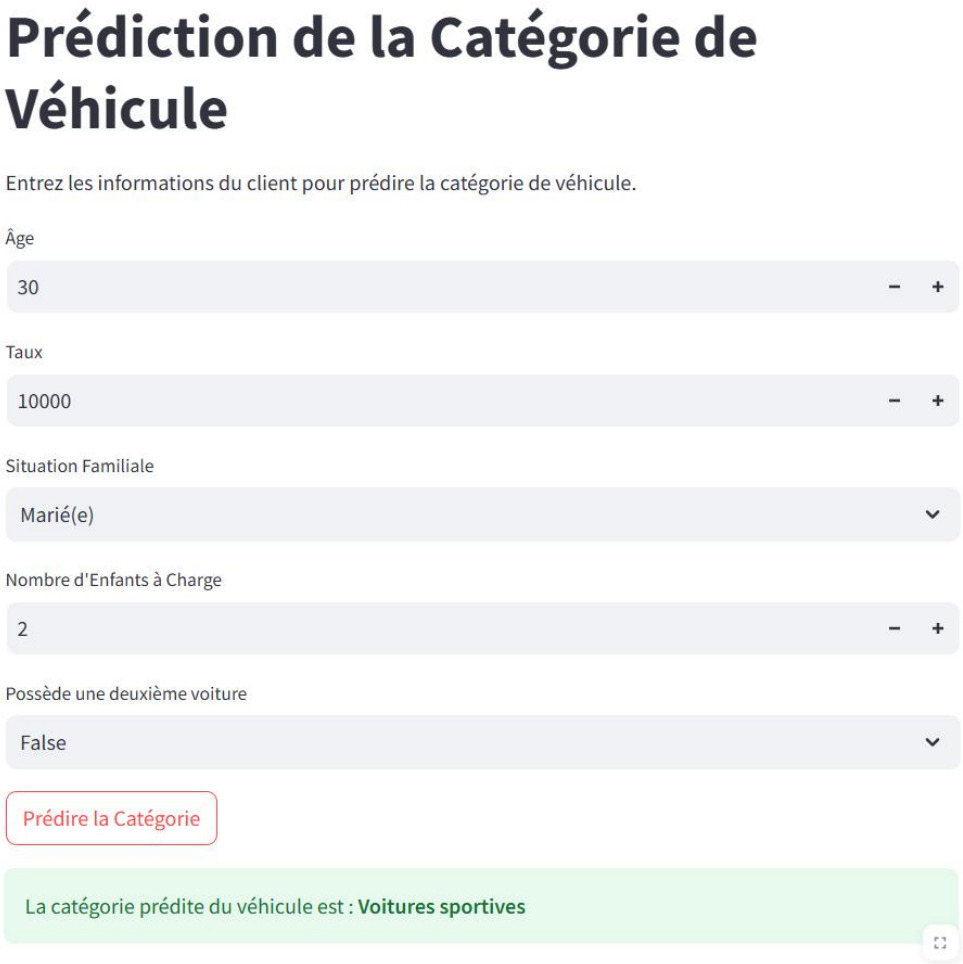


Illustration de Voitures sportives