# Projet Big Data Analytics : Analyse de la Clientèle d'un Concessionnaire Automobile pour la Recommandation de Modèles de Véhicules

INALIARIJAONA Ony Mirana

RAKOTOARIMANANA Manoa Iharivola

RAKOTOARISON Tojo Fandresena Flavien

RAKOTOARIVONY Tendry Hery ny Aina

**Mai 2024 – 09 Juin 2024**

**Enseignant Encadreur :**

Nicolas PASQUIER

# Analyse de données avec des outils de Machine Learning (R)

Dans le cadre de ce projet, plusieurs choix stratégiques ont été faits pour optimiser la gestion et l'analyse des données.

Nous avons opté pour **le langage R** en raison de ses puissantes capacités en analyse statistique, en machine learning, et en visualisation. R offre une vaste bibliothèque de packages spécialisés, ce qui le rend particulièrement adapté à des projets d'analyse de données de grande envergure.

Pour établir des connexions robustes et efficaces aux bases de données SQL, nous avons utilisé les packages **ODBC** et **DBI**. Ces outils nous ont permis d'extraire les données nécessaires de manière fluide et sécurisée. La centralisation des données dans Hive, facilite l'analyse des données à grande échelle avec des requêtes SQL.

L'exploration préliminaire des données a été réalisée à l'aide de statistiques descriptives et de visualisations. Cette étape est cruciale pour comprendre la distribution des variables et identifier les anomalies potentielles, ce qui permet de préparer les données de manière optimale pour les étapes ultérieures de modélisation.

Ces choix méthodologiques et technologiques ont été guidés par la nécessité d'assurer une analyse précise et efficiente des données.

# 1. Connexion aux Sources de Données :

Pour accéder aux données nécessaires à notre analyse, nous avons utilisé R pour établir une connexion aux tables externes présentes dans Hive, comme décrit dans le script **« 1-DriverConnection.R ».** Ce script utilise les bibliothèques « RJDBC », « DBI », « rJava » et « odbc » pour se connecter à Hive et importer les données directement dans R.

Cette configuration nous a permis de centraliser les données dans Hive et de les importer directement dans R pour une analyse efficace.

# 2. Analyse exploratoire des données :

L'exploration des données est une étape cruciale dans le processus d'analyse, car elle permet de comprendre la structure, la qualité et les caractéristiques des données avant d'appliquer des techniques de modélisation avancées. Cette phase vise à identifier les anomalies, les valeurs manquantes et les tendances générales dans les jeux de données. Voici comment nous avons procédé pour explorer les données de notre projet (**2-DataExploration.R).**

### a. Chargement des Données :

Nous avons commencé par charger les bibliothèques nécessaires et les données à partir des tables Hive dans notre environnement R.

### b. Inspection Initiale des Datasets :

Une fois les données chargées, nous avons effectué une inspection initiale pour comprendre la structure et le contenu des jeux de données. Cette inspection comprend l'utilisation de fonctions telles que `str()`, `names()` et `summary()` pour obtenir un aperçu des données.

### c. Nettoyage et Préparation des Données :

Lors de l'exploration des données, nous avons identifié certaines anomalies et valeurs incorrectes que nous avons corrigées pour assurer la qualité des données avant l'analyse. Les étapes de nettoyage incluent la mise à jour des colonnes avec des valeurs incorrectes, la correction des âges et taux négatifs, et le nettoyage des valeurs de sexe et de situation familiale.

Voici un extrait des principales opérations de nettoyage effectuées :

```
# Mettre à jour la colonne 'marketing.situation familiale'
marketing$situation_familiale <- tolower(marketing$situation_familiale)
marketing$situation_familiale <- with(marketing, ifelse(situation_familiale
=="c?libataire", "celibataire",situation_familiale))
marketing$situation_familiale <- with(marketing, ifelse(situation_familiale
=="célibataire", "celibataire",situation_familiale))


# concernant l'age nous avons des anomalies avec des ages negatifs. Pour
corriger cette anomalie nous allons remplacer les ages negatifs par la mediane
: 41
clients$age <- with(clients, ifelse(age < 0, 41 ,age))
# de meme pour le taux avec une mediane à 521
clients$taux <- with(clients, ifelse(taux < 0, 521 ,taux))

#pour le sexe du client
# Supprimer les caracteres speciaux
clients$sexe <- gsub("[^A-Za-z]", "", clients$sexe)

# Mettre en majuscules
clients$sexe <- toupper(clients$sexe)

# Supprimer les espaces vides
clients$sexe <- gsub("\\s+", "", clients$sexe)

# Remplacer les valeurs
clients$sexe <- ifelse(clients$sexe %in% c("MASCULIN", "HOMME"), "M",
clients$sexe)
clients$sexe <- ifelse(clients$sexe %in% c("FEMININ", "FEMME"), "F",
clients$sexe)
clients$sexe <- ifelse(clients$sexe %in% c("FEMININ", "FEMME"), "F",
clients$sexe)
```

### d. Visualisation des Données :

Pour mieux comprendre la distribution et les caractéristiques des données, nous avons créé des visualisations à l'aide de la bibliothèque ggplot2. Nous avons réalisé les principales visualisations suivantes : un histogramme des âges des clients montrant une répartition uniforme avec une majorité moins de 50 ans, un barplot des marques de voitures indiquant que certaines marques sont plus populaires que d'autres, un boxplot de la puissance en fonction de la longueur révélant que les voitures plus longues tendent à avoir plus de puissance, et un scatter plot de la relation entre puissance et prix montrant que les voitures plus longues et très longues sont plus puissantes et plus chères, tandis que les voitures courtes sont moins puissantes et moins chères.

La heatmap des corrélations entre les variables numériques révèle des relations importantes entre les variables numériques. Nous avons observé une forte corrélation positive entre la puissance et le prix des voitures, ainsi qu'entre les rejets de CO2 et le coût énergétique. Une très forte corrélation positive a été notée entre MalusBonus et RejetsCO2, suggérant que ces deux variables sont presque interchangeables en termes de valeur informative. En revanche, il y a une faible corrélation négative entre le nombre de places et le prix, et une corrélation positive modérée entre le nombre de portes et le prix, ainsi qu'entre la puissance et le nombre de portes. Ces observations nous aident à réduire la redondance et à identifier les variables importantes pour le clustering.

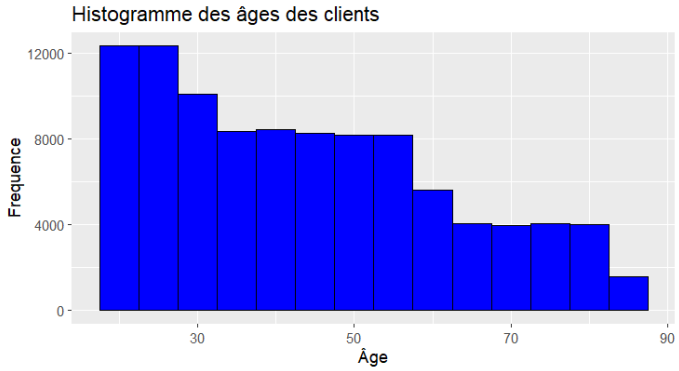


**Figure1: Histogramme des âges des clients**

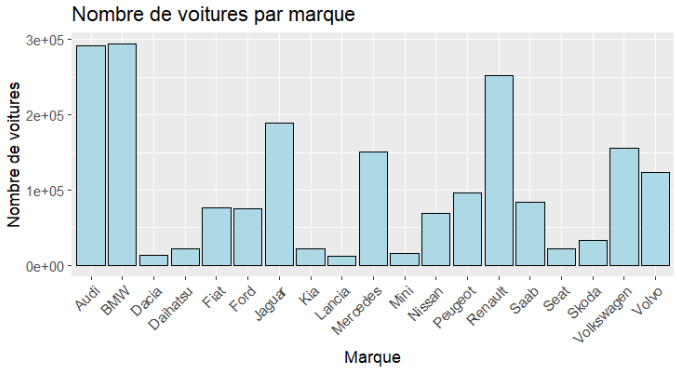


**Figure 1: Barplot du nombre de voitures par marque**

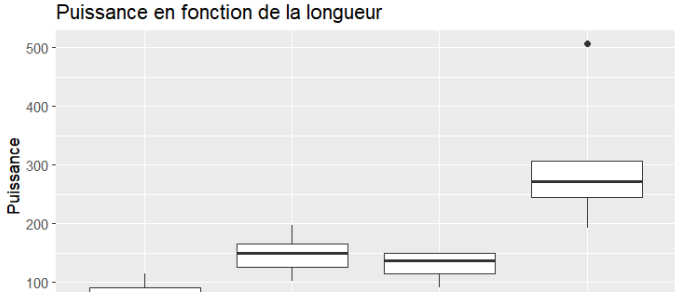


**Figure 3: Boxplot de la puissance en fonction de la longueur**

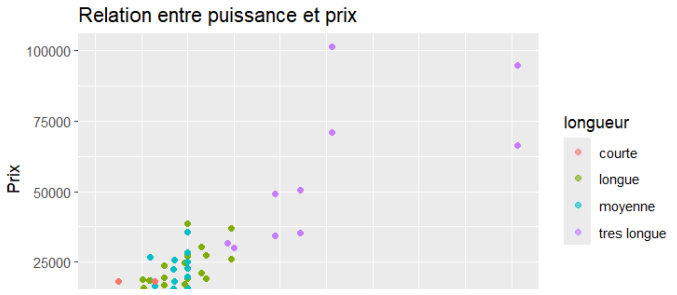


**Figure 4: Scatter plot de la relation entre puissance et prix**

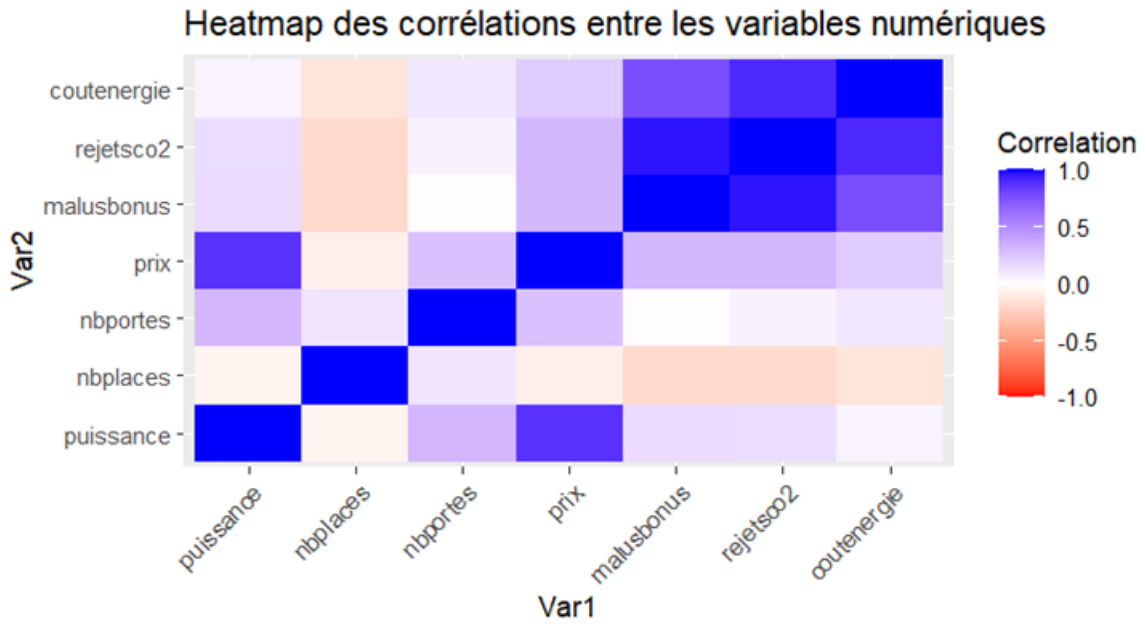


**Figure 1: Heatmap des corrélations entre les variables numériques**

Pour plus de détails sur les étapes spécifiques de chargement, inspection, visualisation et nettoyage des données, vous pouvez vous référer au script complet **« 2-DataExploration.R` ».**

Ces étapes nous ont permis d'identifier et de corriger les anomalies dans les données, garantissant ainsi leur qualité et leur fiabilité pour les analyses ultérieures.

# 3. Identification des catégories de véhicules :

L'objectif de cette section est de catégoriser les véhicules en différentes catégories pertinentes afin de mieux comprendre les segments de marché et d'optimiser les analyses ultérieures. Le fichier **« 3-categoriesVehicules.R »** décrit le processus de classification des véhicules en fonction de plusieurs critères clés. Voici une explication détaillée du contenu et des étapes suivies dans ce script.

### a. Introduction à la Catégorisation des Véhicules

Nous avons utilisé une approche de clustering hiérarchique pour définir les catégories de véhicules basées sur des critères tels que la puissance, la longueur, le nombre de places et de portes, le malus/bonus, les rejets de CO2, le coût énergétique, et le prix. L'objectif est de regrouper les véhicules en segments homogènes pour des analyses plus ciblées.

### b. Principales Étapes du Processus

- **Préparation des Données :** Les données pertinentes ont été sélectionnées et normalisées pour préparer le clustering.
- **Clustering Hiérarchique :** Un clustering hiérarchique a été effectué en utilisant la méthode `ward.D` pour regrouper les véhicules en clusters.
- **Analyse des Clusters :** Les clusters ont été analysés pour déterminer les caractéristiques dominantes de chaque groupe.

- **Définition des Catégories :** Les clusters ont été mappés aux catégories spécifiques telles que citadine, sportive, familiale, confort, longue, luxe, et écologique, en fonction des caractéristiques analysées.

### c. Résultats du Clustering

Les véhicules ont été catégorisés en fonction des caractéristiques dominantes des clusters. Nous avons appliqué nos observations sur la visualisation des données pour définir ces catégories.

Voici un résumé des catégories définies :

- **Sportive :** Véhicules avec une puissance élevée.
- **Luxe :** Véhicules avec un prix moyen élevé.
- **Familiale :** Véhicules avec un nombre de places élevé.
- **Citadine :** Véhicules avec une puissance moyenne faible.
- **Écologique :** Véhicules avec de faibles rejets de CO2.
- **Confort** : Véhicules avec plus de trois portes.

# 4. Application des catégories de véhicules définies aux données des Immatriculations :

Après avoir défini les catégories de véhicules, ces catégories ont été appliquées aux données des immatriculations pour enrichir les informations disponibles. Cela a permis de relier chaque immatriculation à une catégorie de véhicule spécifique, facilitant ainsi une analyse plus détaillée des segments de marché. Les étapes détaillées de ce processus sont également décrites dans le fichier **« 3-categoriesVehicules.R ».**

```
# Convertir toutes les valeurs de la colonne "marque" en minuscules pour la
fusion insensible à la casse
immatriculation$marque <- tolower(immatriculation$marque)
catalogue$marque <- tolower(catalogue$marque)
catalogue$couleur <- tolower(catalogue$couleur)
immatriculation$couleur <- tolower(immatriculation$couleur)

# Fusionner les données
immatrCatalog <- merge(x = immatriculation, by = c(
"marque","nom","puissance", "longueur", "nbplaces", "nbportes","couleur",
"prix"), y = catalogue )
immatrCatalog <- unique(immatrCatalog)
View(immatrCatalog)
```

# 5. Fusion des données Clients et Immatriculations :

Ensuite, nous avons fusionné les données des clients avec les données des immatriculations enrichies. Cela nous a permis de créer un ensemble de données complet, intégrant les informations clients et véhicules, ce qui est essentiel pour une analyse cohérente et complète. Ce processus est détaillé dans le fichier **« 3-categoriesVehicules.R »**.

```
clientsImmat <- merge(x = clients, by = c( "immatriculation"), y =
immatrCatalog )
clientsImmat <- unique(clientsImmat)

# Vérifier le nombre d'éléments dans chaque catégorie
print(table(catalogue$categorie))
print(table(immatrCatalog$categorie))
print(table(clientsImmat$categorie))
```

Cette fusion des données a facilité une analyse détaillée des préférences et des comportements des clients en fonction des catégories de véhicules. En ayant un ensemble de données intégrant les informations sur les clients et les véhicules, nous avons pu effectuer des analyses plus précises et pertinentes pour le marché cible.

# 6. Création d'un modèle de classification supervisée pour la prédiction de la catégorie de véhicules à partir de la fusion des données clients et immatriculations :

Dans cette section, nous avons appliqué plusieurs modèles de classification supervisée pour prédire les catégories de véhicules en fonction des caractéristiques des clients. Les scripts **« 4-ClassificationTree.r »**, **« 5-ClassificationRandomForest.R »**, **« 6-ClassificationNeuralNetworks.R »**, **« 7-ClassificationNaiveBayes.R »**, **«8-ClassificationKNN.R »** décrivent l'utilisation de différentes techniques de classification.

## a. Choix des Colonnes Utilisées

Lors de la mise en place des modèles de classification, nous avons sélectionné les colonnes suivantes :

- **Classe à Prédire :** La catégorie des véhicules (`categorie`).
- **Variables Prédictives :** Âge (`age`), Sexe (`sexe`), Taux (`taux`), Situation familiale (`situation_familiale`), Nombre d'enfants à charge (`nbr_enfant`), Deuxième voiture (`voiture_2`).

- **Variables Ignorées :** Immatriculation (`immatriculation`), Marque (`marque`), Nom (`nom`), Puissance (`puissance`), Longueur (`longueur`), Nombre de places (`nbplaces`), Nombre de portes (`nbportes`), Couleur (`couleur`), Occasion (`occasion`), Prix (`prix`).

Cette sélection a été guidée par l'objectif de prédire la catégorie de véhicule en fonction des caractéristiques des clients.

## b. Étapes Communes aux Modèles de Classification

Pour chaque modèle de classification, les étapes suivantes ont été suivies :

1. **Préparation des Données** : Division des données en ensembles d'entraînement et de test.
2. **Construction du Modèle :** Utilisation des bibliothèques appropriées pour construire le modèle.
3. **Prédiction** : Application du modèle aux données de test pour effectuer des prédictions.

## c. Spécificités des Modèles

### 1. Arbre de Décision (`4-ClassificationTree.r`) :

L'arbre de décision est un modèle simple et interprétable qui divise les données en segments basés sur les variables prédictives. Le script 4-ClassificationTree.r utilise les bibliothèques rpart et caret pour construire et évaluer l'arbre de décision. Le modèle est construit en utilisant la fonction rpart, et les prédictions sont effectuées avec la fonction predict.

### 2. Forêt Aléatoire (`5-ClassificationRandomForest.R`) :

La forêt aléatoire améliore la précision en combinant plusieurs arbres de décision. Le script 5-ClassificationRandomForest.R utilise les bibliothèques randomForest et caret. Le modèle est construit en utilisant la fonction randomForest, et les prédictions sont faites en utilisant predict.

### 3. Réseaux de Neurones (`6-ClassificationNeuralNetworks.R`) :

Les réseaux de neurones capturent des relations complexes entre les variables grâce à une architecture de réseau. Le script 6-ClassificationNeuralNetworks.R utilise la bibliothèque neuralnet pour construire et évaluer le modèle. Le modèle est construit en utilisant la fonction neuralnet, et les prédictions sont obtenues avec la fonction compute.

### 4. Naive Bayes (`7-ClassificationNaiveBayes.R`) :

Naive Bayes utilise des probabilités pour classer les données de manière rapide et efficace. Le script 7-ClassificationNaiveBayes.R utilise la bibliothèque e1071 pour construire et évaluer le modèle. Le modèle est construit en utilisant la fonction naiveBayes, et les prédictions sont effectuées avec predict.

### 5. K-Nearest Neighbors (KNN) (`8-ClassificationKNN.R`):

KNN classe les données en fonction de la proximité des points de données dans l'espace des caractéristiques. Le script 8-ClassificationKNN.R utilise la bibliothèque class pour appliquer l'algorithme KNN. Les prédictions sont faites en utilisant la fonction knn.

### d. Comparatif des Résultats

Pour mieux visualiser et comparer les performances des différents modèles, nous avons résumé les résultats obtenus dans le tableau ci-dessous :

| Modèles | Tests | Prédiction |
|---|---|---|
| Arbre de Décision C5.0 | taux_C51 | 0.745079 |
| | taux_C52 | 0.745814 |
| | taux_C53 | 0.745079 |
| | taux_C54 | 0.745814 |
| K-Nearest Neighbors | taux_knn1 | 0.672660 |
| | taux_knn2 | 0.696153 |
| | taux_knn3 | 0.673061 |
| | taux_knn4 | 0.696087 |
| Naive Bayes | taux_nb1 | 0.704408 |
| | taux_nb2 | 0.704441 |
| Réseaux de Neurones | taux_nn1 | 0.719112 |
| | taux_nn2 | 0.706480 |
| | taux_nn3 | 0.666310 |
| | taux_nn4 | 0.718611 |
| Forêt Aléatoire | taux_rf1 | 0.744912 |
| | taux_rf2 | 0.745179 |
| | taux_rf3 | 0.735421 |
| | taux_rf4 | 0.735220 |
| Arbre de Décision rpart | taux_rp1 | 0.742606 |
| | taux_rp2 | 0.742606 |
| | taux_rp3 | 0.742606 |
| | taux_rp4 | 0.742606 |
| Arbre de Décision tree | taux_tr1 | 0.742606 |
| | taux_tr2 | 0.742606 |

Pour plus de détails sur la mise en œuvre et l'évaluation de chaque modèle, vous pouvez vous référer aux scripts complets respectifs.

Après avoir évalué les différents modèles de classification, nous avons constaté que l'algorithme C5.0 a obtenu le meilleur taux de succès. Nous avons donc utilisé C5.0 pour prédire les catégories des données marketing, car il a démontré la précision la plus élevée parmi tous les modèles testés.

# 7. Application du modèle de prédiction aux données Marketing

Dans cette section, nous avons appliqué le modèle de prédiction aux données marketing afin de prédire les catégories des véhicules pour les clients potentiels. Nous avons utilisé le modèle C5.0, qui a montré le meilleur taux de succès lors de notre évaluation.

- **Prétraitement des Données Marketing :**

Avant d'appliquer le modèle, nous avons prétraité les données marketing pour nous assurer qu'elles étaient dans le bon format pour la prédiction.

```
# Prétraitement des données marketing
marketing$sexe <- as.factor(marketing$sexe)
marketing$situation_familiale <- as.factor(marketing$situation_familiale)
```

- **Application du Modèle C5.0 :**

Nous avons utilisé le modèle C5.0 pour prédire les catégories des véhicules en fonction des caractéristiques des clients présents dans les données marketing.

```
# Utilisation du modèle C5.0 avec le meilleur taux de succès pour prédire les
catégories des données marketing
marketing_predictions <- predict(tree_C54, marketing)
```

- **Ajout des Prédictions aux Données Marketing :**

Les prédictions obtenues ont été ajoutées aux données marketing, ce qui nous a permis de voir quelle catégorie de véhicule était attribuée à chaque client potentiel.

```
# Ajout des prédictions aux données marketing
marketing$predicted_categorie <- marketing_predictions

# Affichage des résultats
print(marketing)
```

- **Résultats :**

Les résultats montrent les catégories de véhicules prédites pour chaque client potentiel dans les données marketing. Cela nous permet de cibler plus efficacement les clients avec des offres de véhicules qui correspondent à leurs caractéristiques et préférences.

| | marketing_result.id | marketing_result.age | marketing_result.sexe | marketing_result.taux | marketing_result.situation_familiale | marketing_result.nbr_enfant | marketing_result.voiture_2 | marketing_result.predicted_categorie |
|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 26 | F | 420 | en couple | 3 | TRUE | confort |
| 2 | 20 | 59 | M | 748 | en couple | 0 | TRUE | confort |
| 3 | 15 | 60 | M | 524 | en couple | 0 | TRUE | confort |
| 4 | 2 | 35 | M | 223 | celibataire | 0 | FALSE | confort |
| 5 | 8 | 43 | F | 431 | celibataire | 0 | FALSE | confort |
| 6 | 12 | 55 | M | 588 | celibataire | 0 | FALSE | confort |
| 7 | 13 | 19 | F | 212 | celibataire | 0 | FALSE | confort |
| 8 | 6 | 27 | F | 153 | en couple | 2 | FALSE | confort |
| 9 | 10 | 22 | M | 154 | en couple | 1 | FALSE | confort |
| 10 | 1 | 21 | F | 1396 | celibataire | 0 | FALSE | confort |
| 11 | 7 | 59 | F | 572 | en couple | 2 | FALSE | confort |
| 12 | 16 | 22 | M | 411 | en couple | 3 | TRUE | confort |
| 13 | 17 | 58 | M | 1192 | en couple | 0 | FALSE | confort |
| 14 | 3 | 48 | M | 401 | celibataire | 0 | FALSE | confort |
| 15 | 9 | 64 | M | 559 | celibataire | 0 | FALSE | confort |
| 16 | 14 | 34 | F | 1112 | en couple | 0 | FALSE | confort |
| 17 | 18 | 54 | F | 452 | en couple | 3 | TRUE | confort |
| 18 | 5 | 80 | M | 530 | en couple | 3 | FALSE | luxe |
| 19 | 19 | 35 | M | 589 | celibataire | 0 | FALSE | confort |
| 20 | 11 | 79 | F | 981 | en couple | 2 | FALSE | confort |

**Figure : Résultat prédiction marketing**

# Conclusion

Le projet d'analyse de la clientèle d'un concessionnaire automobile pour la recommandation de modèles de véhicules a été une réussite grâce à l'utilisation de R pour l'analyse de données et le machine learning. En intégrant efficacement les données depuis Hive, nous avons pu explorer et nettoyer les données de manière approfondie.

Nos analyses ont permis de catégoriser les véhicules en segments pertinents et de comprendre les préférences des clients. La fusion des données clients et immatriculations a enrichi notre compréhension des comportements d'achat. Parmi les modèles de classification testés, l'algorithme C5.0 s'est révélé le plus performant pour la prédiction des catégories de véhicules.

En appliquant ce modèle aux données marketing, nous avons pu cibler efficacement les clients potentiels avec des recommandations personnalisées, améliorant ainsi l'expérience client et les stratégies de vente. Ce projet souligne l'importance de l'analyse de données pour des décisions commerciales éclairées et des recommandations précises.