Rapport de Méthode d'analyse

Yous Ines M1 DSS

Introduction :

Le but de cette étude est de réaliser une classification en groupes (cluster) et d'interpréter les résultats à l'aide de tests de comparaison de moyenne. Un cluster est un regroupement d'individus partageant des caractéristiques similaires, basé sur une mesure de distance ou de similarité entre les observations. En effet, cette démarche permet d'identifier des regroupements naturels ou des tendances au sein des données, tout en mettant en évidence les facteurs qui distinguent ces groupes. Le clustering aide à analyser les profils caractéristiques des différentes classes obtenues.

Nous analyserons un ensemble de données portant sur 32 modèles d'automobiles, en tenant compte de leur consommation de carburant ainsi que leurs performances.

Pour ce faire nous aurons les 11 variables suivantes : Mgp , Cyl , Disp , Hp , Drat, Wt, Qsec , Vs, Am, Gear, Carb

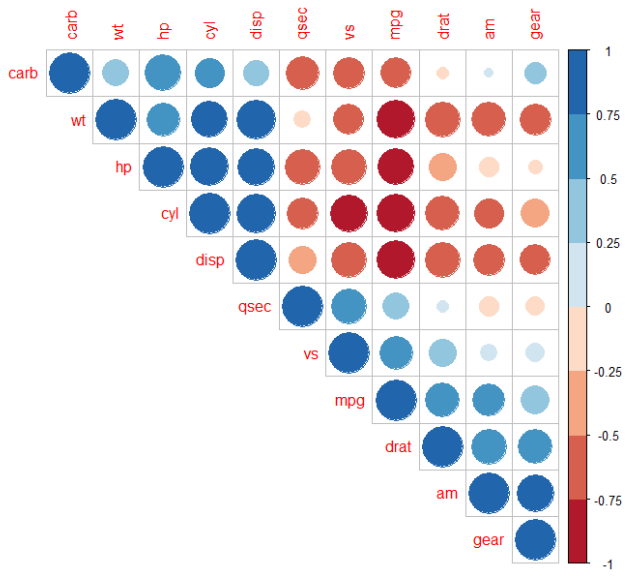
1. Corrélation

Une matrice de corrélation sert à analyser les liens entre plusieurs variables quantitatives en même temps. Elle est utilisée pour identifier si certaines variables sont fortement ou faiblement liées. En effet, si deux variables ont une corrélation proche de 1 ou -1, cela indique une forte relation linéaire entre elles, tandis qu'une corrélation proche de 0 suggère qu'il n'y a pas de lien significatif. Pour que ces liens soient fiables, il est important de vérifier leur significativité statistique à l'aide de tests appropriés (comme le test de significativité des coefficients de corrélation). Cela permet d'éviter les conclusions erronées basées sur des corrélations qui pourraient être dues au hasard.

Les corrélogrammes sont des outils visuels pratiques pour représenter ces relations. Par exemple, en utilisant la fonction **corrplot** dans R, on peut afficher les coefficients de corrélation sous forme de cercles colorés dont la taille et la teinte indiquent l'intensité et la direction des liens entre les variables.

Pour éviter les biais, il est recommandé de prendre en compte uniquement les corrélations significatives. Cela peut être fait en complétant l'analyse graphique par la vérification des p-values associées aux coefficients de corrélation. Cette démarche permet de confirmer que les relations observées ne sont dues au hasard.

Dans le corrélogramme ci-dessus, on observe la matrice de corrélation d'un ensemble de variables. Par exemple :



- Une forte corrélation négative (cercle rouge foncé) est visible entre les variables Cyl (nombre de cylindres) et Mpg (miles par gallon). Cela signifie que lorsque le nombre de cylindres augmente, la consommation de carburant en miles par gallon diminue.
- À l'inverse, une forte corrélation positive (cercle bleu foncé) est observée entre Cyl et Disp (cylindrée du moteur), indiquant que ces deux variables augmentent ensemble.

Ce type de visualisation facilite l'identification des relations importantes, comme celles-ci, et aide à se concentrer uniquement sur les corrélations pertinentes après avoir testé leur significativité.

2. ACP (Analyse en Composantes Principales)

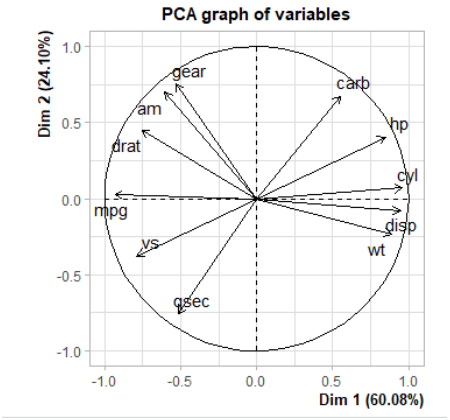
Cependant, lorsqu'il y a un grand nombre de variables, la matrice de corrélation peut devenir difficile à interpréter en raison de la complexité des relations entre les variables. Pour simplifier l'analyse tout en conservant l'essentiel des informations, on peut utiliser une méthode comme l'ACP. Cette approche permet de réduire la dimensionnalité des données en identifiant des combinaisons linéaires de variables (appelées composantes principales) qui expliquent la majeure partie de la variabilité des données. L'ACP est particulièrement utile pour visualiser les relations globales et identifier les variables les plus influentes dans un jeu de données.

Les composantes principales sont des combinaisons linéaires des variables initiales, construites de manière à maximiser la variance expliquée. Elles servent à visualiser les relations globales et à identifier les variables les plus influentes dans un jeu de données.

La règle de Kaiser est utilisée pour sélectionner les composantes principales en ne conservant que celles qui ont des valeurs propres supérieures à 1. Ici, cela concerne les Composantes 1 et 2

Concernant le graphe des variables, cette image illustre un espace réduit à deux dimensions obtenu grâce à une analyse en composantes principales (ACP). Les flèches indiquent la direction et l'intensité des contributions des variables sur les deux axes principaux, Dim 1 (60,08%) et Dim 2 (24,10%). Les variables proches l'une de l'autre (par exemple, cyl, disp et wt) sont fortement corrélées, tandis que celles opposées, comme mpg et cyl, sont en opposition. Ce graphe permet d'interpréter les relations entre les variables et d'identifier celles qui influencent le plus chaque dimension.

```
> resultat$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1  6.60840025             60.0763659                          60.07637
comp 2  2.65046789             24.0951627                          84.17153
comp 3  0.62719727              5.7017934                          89.87332
comp 4  0.26959744              2.4508858                          92.32421
comp 5  0.22345110              2.0313737                          94.35558
comp 6  0.21159612              1.9236011                          96.27918
comp 7  0.13526199              1.2296544                          97.50884
comp 8  0.12290143              1.1172858                          98.62612
comp 9  0.07704665              0.7004241                          99.32655
comp 10 0.05203544              0.4730495                          99.79960
comp 11 0.02204441              0.2004037                         100.00000
```



Ensuite la somme des Pourcentages de variance des composantes choisit doit être supérieur à 80%, c'est la règle des 80% d'information ou d'inertie cumulée. Ici, les 2 composantes regroupent 84,17% de l'information (60,08 + 24,09).

Grâce à ces méthodes nous sommes passé de 11 à 2 Dimensions.

Donc pour la Dim 1puis la Dim 2, nous prenons les contributions les plus importantes par rapport à la contribution moyenne.

Dimension 1 :

Tout d'abord on calcule la contribution moyenne = totalité de l'information/ le nombre de variables (100/11 = 9,09%). Donc on sélectionne toutes les variables qui ont une contribution supérieure à 9,09%. La somme des résultats au-dessus de 9,09% = 72,94%

Variables actives

Show 20 entries Search:

Variable	Coord	Contrib	Cos2	Cor
cyl	0.961	13.98	0.924	0.961
disp	0.946	13.56	0.896	0.946
mpg	-0.932	13.14	0.869	-0.932
wt	0.890	11.98	0.792	0.89
hp	0.848	10.89	0.720	0.848
vs	-0.788	9.39	0.621	-0.788
drat	-0.756	8.65	0.572	-0.756
am	-0.604	5.52	0.365	-0.604
carb	0.550	4.58	0.303	0.55
gear	-0.532	4.28	0.283	-0.532
qsec	-0.515	4.02	0.266	-0.515

Showing 1 to 11 of 11 entries Previous 1 Next

On aura donc les variables Cyl, Disp, Wt et Hp, qui se projettent en positif et les variables Mpg et Vs, qui se projettent en négatif. Cette disposition indique une opposition entre les 2 groupes.

Ensuite le coefficient de corrélation au carré (cos^2) permet d'évaluer la qualité de représentation des variables sur cette composante. Par exemple, les variables Cyl, Disp, Wt présentent des valeurs élevées de cos^2 de 0,924, 0,886 et 0,898, ce qui montrent qu'elles sont fortement bien représentées sur cette dimension. Les variables ayant un cos^2 faible, comme Gear (0,282), contribuent moins à cette composante.

Ainsi, l'analyse du cos^2 aide à identifier les variables qui influencent le plus la composante, guidant une interprétation plus précise des résultats.

Les individus situés à droite sur Dim.1 (comme 24) ont des moteurs puissants, de grande cylindrée, mais consomment beaucoup (faible mpg).

Dimension 2 :

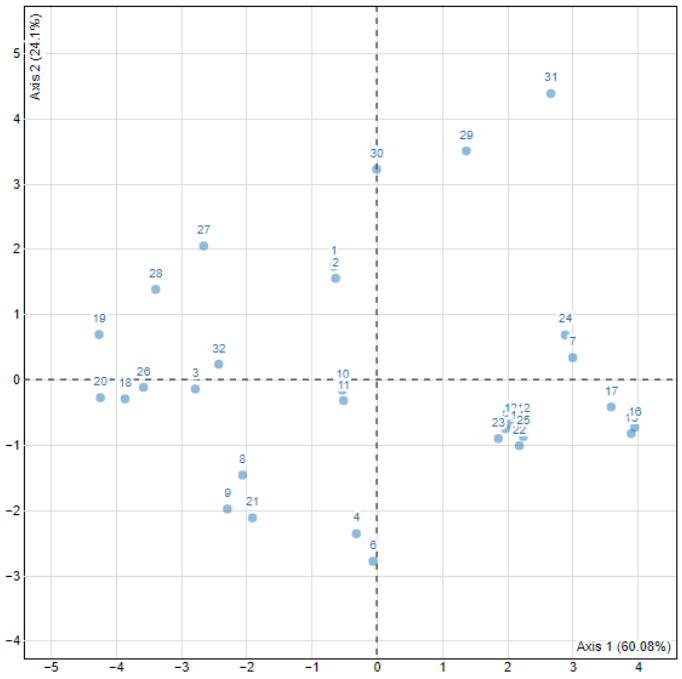
Tout d'abord on sélectionne toutes les variables qui ont une contribution supérieure à 9,09%. La somme des résultats au-dessus de 9,09% = 78,39%

Variables actives

Show 20 entries Search:

Variable	Coord	Contrib	Cos2	Cor
qsec	-0.754	21.47	0.569	-0.754
gear	0.753	21.38	0.567	0.753
am	0.699	18.44	0.489	0.699
carb	0.673	17.10	0.453	0.673
drat	0.447	7.55	0.200	0.447
hp	0.405	6.19	0.164	0.405
vs	-0.377	5.37	0.142	-0.377
wt	-0.233	2.05	0.054	-0.233
disp	-0.080	0.24	0.006	-0.08
cyl	0.071	0.19	0.005	0.071
mpg	0.026	0.03	0.001	0.026

On aura donc les variables Gear, Am et Carb qui se projettent en positif et la variable Qsec qui se projette en négatif. Donc d'un côté les caractéristiques mécaniques des voitures avec Gear, Am et Carb et de l'autre la vitesse d'accélération avec Qsec



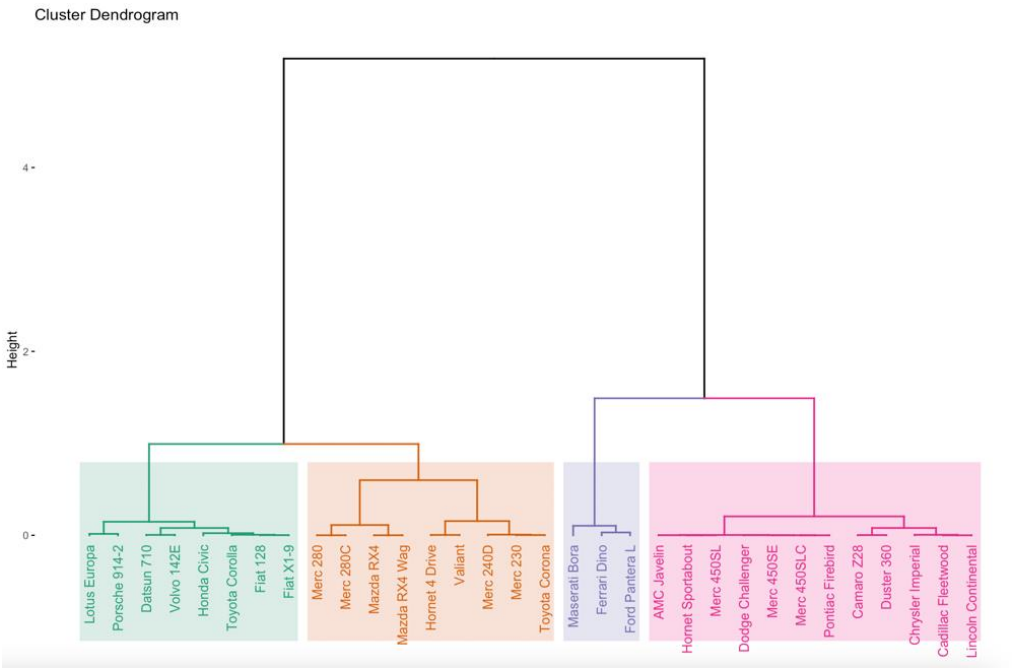
Ce graphique représente la projection des individus dans le plan factoriel issu de l'ACP. Plus les points sont proches, plus leurs profils sont similaires en termes des variables analysées.

On peut observer des regroupements suggérant des sous-groupes d'individus partageant des caractéristiques communes. Enfin, l'opposition entre certains points, répartis de part et d'autre des axes, montre des différences marquées entre leurs profils.

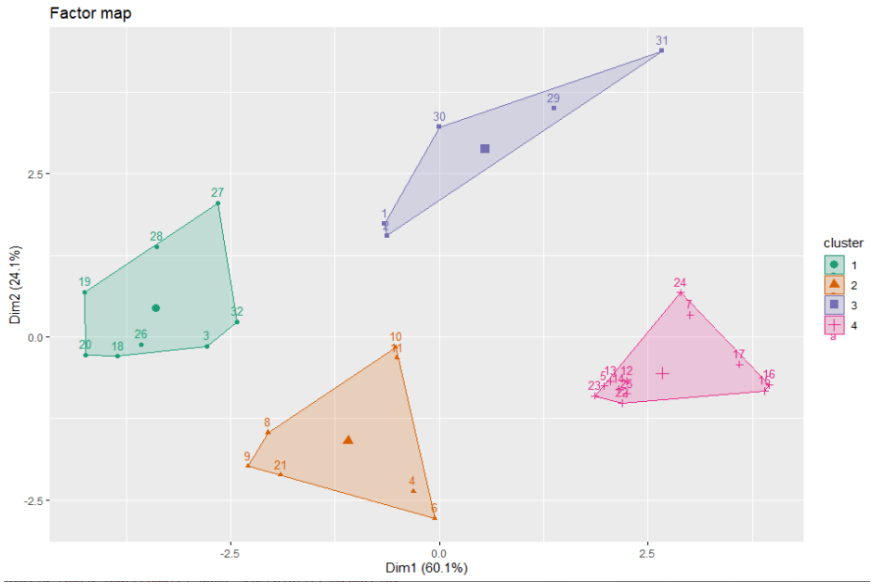
3. Classification hiérarchique

La classification hiérarchique est une méthode d'analyse non supervisée qui permet de regrouper des unités en sous-groupes selon leurs similarités (= Cluster). Elle peut être ascendante, où chaque élément commence comme une classe et est progressivement fusionné avec les classes voisines, ou descendante, où l'on commence avec une seule classe globale et on la divise en sous-groupes.

Une approche mixte combine la méthode k-means (pour créer un nombre initial de classes) et la classification hiérarchique pour affiner ces groupes. Le k-means est une méthode de clustering non supervisée qui regroupe les observations autour de centres de gravité. Par la suite, la classification hiérarchique permet de réduire la complexité et d'obtenir des regroupements plus précis.



Le dendrogramme permet de visualiser la hiérarchie des clusters, en montrant comment les individus sont successivement regroupés en fonction de leur similarité. Chaque branche du dendrogramme représente une fusion entre deux groupes d'individus similaires, et la hauteur à laquelle les branches se rejoignent reflète la distance entre ces groupes. Cela permet d'identifier les niveaux de similarité et de déterminer combien de groupes sont pertinents pour l'analyse.



La carte factorielle, quant à elle, projette les individus dans un espace à deux dimensions, généralement basé sur les deux premières composantes principales de l'ACP. Cela permet d'avoir une vision plus claire de la distribution des clusters dans cet espace réduit. Dans cette carte, les individus se regroupent en fonction de leurs caractéristiques, et la proximité entre les points indique des profils similaires, tandis que les points plus éloignés suggèrent des différences importantes entre les individus.

- Le Cluster vert avec des voitures comme la Porsche 914-2, regroupe des véhicules économes et mécaniquement bien équipés.
- Le Cluster orange avec des voitures comme la Toyota Corona, représente des voitures économes mais peu performantes
- Le Cluster violet avec des voitures comme tels que la Maserati Bora regroupe des modèles puissants, technologiquement avancés, mais lourds et peu économes en carburant.
- Le Cluster rose avec des voitures comme la Hornet Sportabout rassemble des voitures combinant puissance et accélération.

4. Conclusion :

Cette étude a permis d'analyser un ensemble de données sur les modèles automobiles en utilisant la corrélation, l'ACP et la classification hiérarchique. Ces méthodes ont révélé des relations importantes entre les variables, simplifié la structure des données et identifié quatre groupes distincts de véhicules.

Les résultats montrent la pertinence de ces approches statistiques pour mettre en évidence des profils caractéristiques et des tendances au sein des données. Cette analyse confirme l'intérêt de combiner exploration visuelle, tests statistiques et techniques de classification pour une meilleure compréhension des données.