Méthode d'Analyse de Données

Perrine Warter

I. Introduction de notre étude.

L'objectif de cette étude est de construire des clusters et de savoir les interpréter avec les tests de comparaison de moyenne.

Cela va permettre de révéler des groupes naturels ou des modèles dans des jeux de données, en facilitant l'interprétation de leurs relations et d'identifier les facteurs qui différencient leurs groupes.

En résumé, faire des clusters permet de voir les profils des classes.

Pour cette étude, nous allons analyser des données de 32 automobiles (modèles 1973-1974) en fonction de leur consommation de carburant et 10 aspects de la conception et des performances.

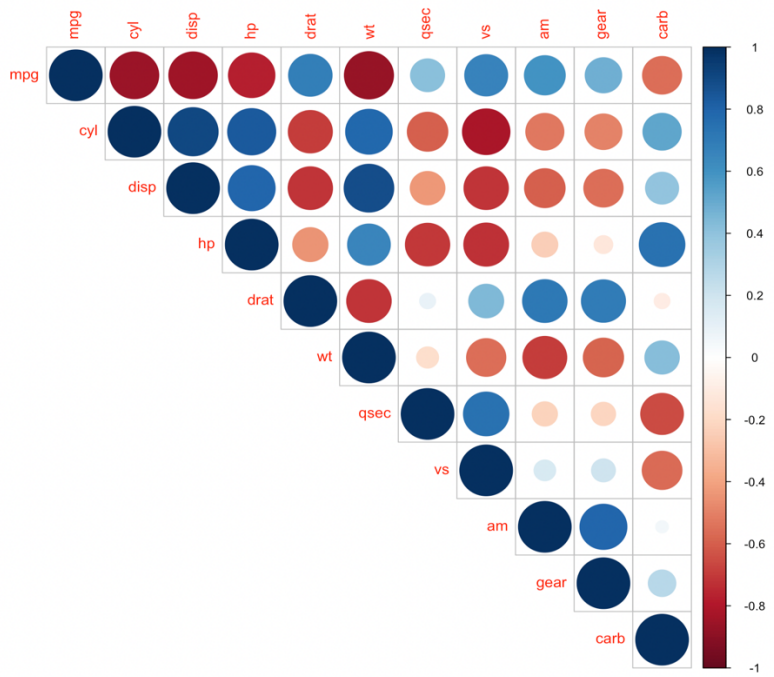
Nos 11 variables sont donc :

- Mgp : Miles par gallon
- Cyl : nombre de cylindres
- Disp : cylindrée
- Hp : puissance brute en chevaux
- Drat : rapport de l'essieu arrière
- Wt : poids
- Qsec : temps sur un quart de mile
- Vs : Disposition en ligne ou en V du moteur
- Am : transmission (0 = automatique, 1 = manuelle)
- Gear : nombre de vitesses avant
- Carb : Nombre de carburateurs

II. Tests de corrélation.

La première étape va être d'analyser les corrélations entre les variables. En effet, une corrélation forte entre variables peut fausser les résultats des clusters car des informations similaires peuvent être surreprésentées.

Le coefficient de corrélation se calcule entre deux variables quantitatives continue.
On dit qu'il y a corrélation linéaire entre deux variables (deux séries statistiques à observées sur les éléments d'une même population lorsque les variations des deux variables se produisent dans le même sens (corrélation positive) ou lorsque les variations sont de sens contraire (corrélation négative)



Ce corrélogramme représente le graphique de notre matrice de corrélation. Il met en évidence les variables les plus corrélées. Pour étudier les coefficients de corrélation nous allons analyser la couleur et la grandeur des cercles.
Tous les cercles bleus représentent une corrélation positive et les rouges une corrélation négative.
Cependant nous voulons savoir à partir de quelle valeur on parle de corrélation significative. On utilise donc une matrice de corrélation qui nous donne les coefficients de corrélation entre chaque variable (1er tableau) et les p values correspondants aux niveaux de significativité des corrélations. Ces tableaux nous donne la nuance de significativité

```
> rcorr(M)
       mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
mpg   1.00 -0.99 -0.99 -0.96  0.94 -0.99  0.71  0.93  0.83  0.77 -0.80
cyl  -0.99  1.00  0.99  0.97 -0.92  0.97 -0.77 -0.96 -0.78 -0.74  0.82
disp -0.99  0.99  1.00  0.94 -0.95  0.99 -0.69 -0.93 -0.84 -0.80  0.76
hp   -0.96  0.97  0.94  1.00 -0.82  0.90 -0.88 -0.98 -0.63 -0.56  0.92
drat  0.94 -0.92 -0.95 -0.82  1.00 -0.97  0.47  0.79  0.94  0.92 -0.57
wt   -0.99  0.97  0.99  0.90 -0.97  1.00 -0.59 -0.87 -0.90 -0.85  0.70
qsec  0.71 -0.77 -0.69 -0.88  0.47 -0.59  1.00  0.90  0.20  0.14 -0.95
vs    0.93 -0.96 -0.93 -0.98  0.79 -0.87  0.90  1.00  0.59  0.54 -0.91
am    0.83 -0.78 -0.84 -0.63  0.94 -0.90  0.20  0.59  1.00  0.98 -0.34
gear  0.77 -0.74 -0.80 -0.56  0.92 -0.85  0.14  0.54  0.98  1.00 -0.24
carb -0.80  0.82  0.76  0.92 -0.57  0.70 -0.95 -0.91 -0.34 -0.24  1.00

n= 11


P
     mpg    cyl    disp   hp     drat   wt     qsec   vs     am     gear   carb
mpg         0.0000 0.0000 0.0000 0.0000 0.0000 0.0147 0.0000 0.0017 0.0058 0.0032
cyl  0.0000        0.0000 0.0000 0.0000 0.0000 0.0059 0.0000 0.0044 0.0099 0.0020
disp 0.0000 0.0000        0.0000 0.0000 0.0000 0.0183 0.0000 0.0011 0.0030 0.0070
hp   0.0000 0.0000 0.0000        0.0020 0.0002 0.0004 0.0000 0.0359 0.0704 0.0000
drat 0.0000 0.0000 0.0000 0.0020        0.0000 0.1426 0.0035 0.0000 0.0000 0.0700
wt   0.0000 0.0000 0.0000 0.0002 0.0000        0.0563 0.0005 0.0001 0.0008 0.0171
qsec 0.0147 0.0059 0.0183 0.0004 0.1426 0.0563        0.0002 0.5474 0.6815 0.0000
vs   0.0000 0.0000 0.0000 0.0000 0.0035 0.0005 0.0002        0.0539 0.0839 0.0000
am   0.0017 0.0044 0.0011 0.0359 0.0000 0.0001 0.5474 0.0539        0.0000 0.3056
gear 0.0058 0.0099 0.0030 0.0704 0.0000 0.0008 0.6815 0.0839 0.0000        0.4849
carb 0.0032 0.0020 0.0070 0.0000 0.0700 0.0171 0.0000 0.0000 0.3056 0.4849
```

Tout d'abord nous posons les hypothèses.
L'hypothèse nulle (H0) suppose que le coefficient de corrélation est nul, indiquant ainsi l'absence de corrélation linéaire entre les variables.
Nous nous basons sur un risque α = 5%.
Tant que la p-valeur est inférieure au risque α on rejette l'hypothèse H0.
Nous observons que pour la majorité de nos variables nous rejetons l'hypothèse nulle H0.
Ainsi, nos données sont fortement corrélées.

Pour répondre à notre objectif qui est la construction de clusters nous utilisons la procédure HCPC (Hierarchical Clustering on Principal Component) combinant un ACP et un clustering hiérarchique. Cela va permettre de segmenter nos données en clusters.

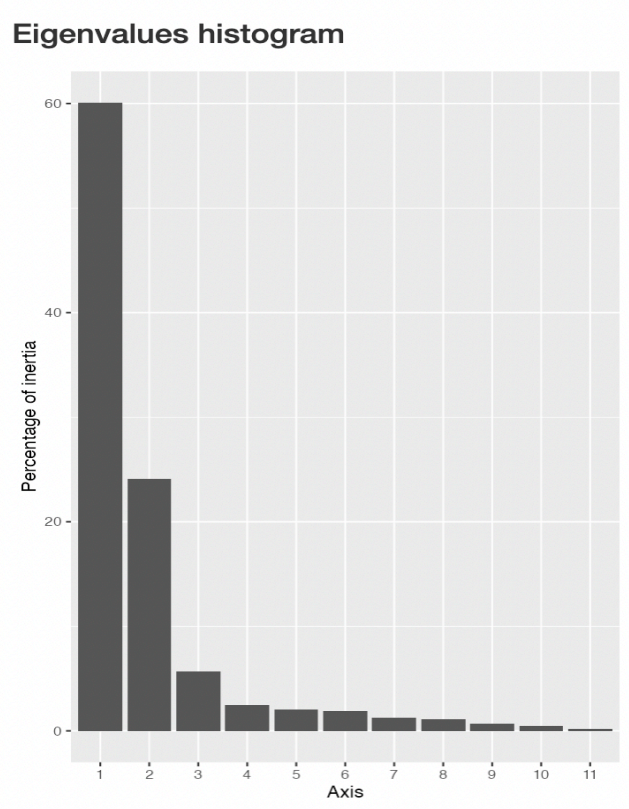
III. Analyse en Composantes Principales (ACP)

En effet, nous devons d'abord utiliser l'ACP : Analyse en Composantes Principales qui va pouvoir condenser l'information, celle-ci est particulièrement utile pour réduire la dimensionnalité lorsque les variables sont corrélées. Elle va combiner ces variables pour créer des composantes principales et ainsi expliquer la variance de manière optimale tout en conservant autant que possible les informations essentielles.

Cela met en évidence quelles variables contribuent le plus aux nouvelles dimensions, permettant ainsi de prioriser les facteurs les plus influents.

Pour se faire nous utilisons la règle de Kaiser, celle-ci permet de sélectionner les composantes principales en ne conservant que celles qui ont des valeurs propres supérieures à 1

	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



Nous obtenons que les 2 premières composantes qui ont des valeurs propres supérieures à 1.
Ces 2 composantes regroupent 84,16% de l'information de notre base de données respectant aussi la règle des 80% d'information ou d'inertie cumulée.

Pour la suite, nous travaillerons sur ces 2 dimensions, l'une après l'autre.

Pour la première dimension, nous relevons les contributions les plus importantes par rapport à la contribution moyenne.
La contribution moyenne représente la totalité de l'information 100% divisée par le nombre de variables (100%11 = 9,09).
Donc nous sélectionnons toutes les variables ayant une contribution supérieure à 9,09%

Axe

Axis 1 (60.08%)

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

La première composante (tendance principale de 60,08% de l'information totale) issue de l'ACP normée est expliquée à hauteur de 72,94 % par les variables CYL, DISP, WT, HP qui se

projettent en positif et les variables MPG, VS qui se projettent en négatif. Du fait de cette projection, cette tendance oppose le groupe (CYL, DISP, WT et HP) au groupe (MPG et VS). Cette dimension regroupe donc des informations sur la performance du moteur, la puissance et l'efficacité énergétique.

Nous analysons ensuite la deuxième dimension qui regroupe 24,09% de l'information totale.

Axe

Axis 2 (24.1%)

Variables actives

Show 20 entries

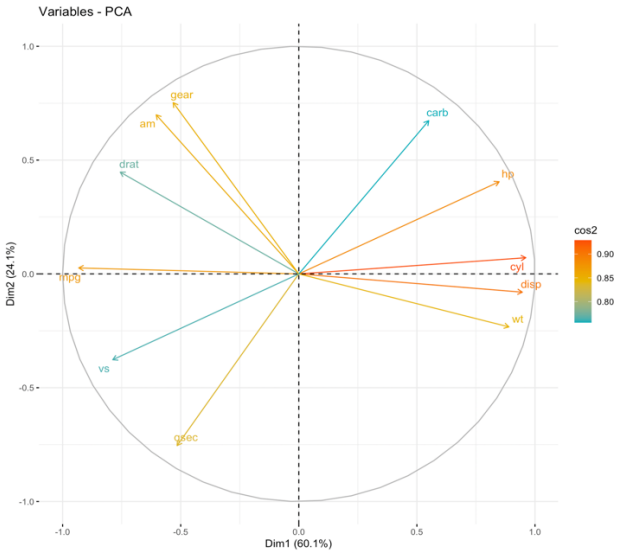
Search:

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

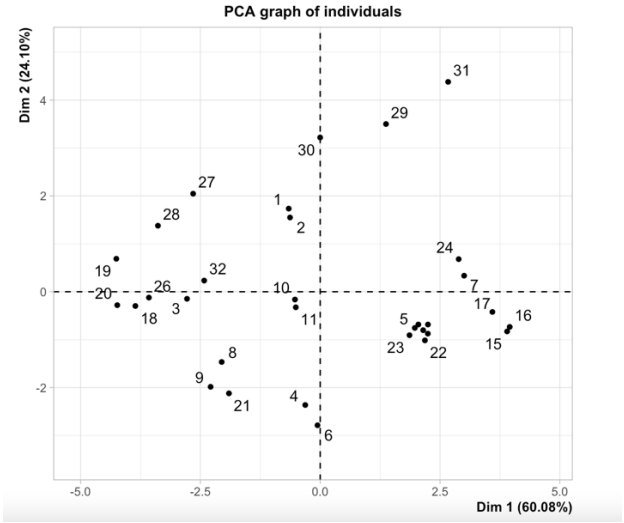
Showing 1 to 11 of 11 entries

Previous 1 Next

Cette seconde composante est expliquée à hauteur de 78,39% par les variables GEAR, AM et CARB qui se projettent en positif et la variable QSEC qui se projette en négatif.
Cette opposition met en lumière une tendance distincte : d'un côté, les caractéristiques mécaniques telles que le nombre de vitesses (GEAR), le type de boîte de vitesses (AM), et le nombre de carburateurs (CARB), et de l'autre, la vitesse d'accélération sur un quart de mile (QSEC).
Cette dimension regroupe donc les informations liées au type de transmission, au design mécanique et à la réactivité du véhicule.



Nous pouvons ensuite visualiser distinctement nos deux dimensions et leurs variables respectives grâce au cercle des corrélations ainsi que les variables fortement corrélées entre elles (positivement ou négativement). Par exemple AM et GEAR qui paraissent très corrélées positivement ou bien GEAR et QSEC qui semblent corrélées négativement mettant en opposition la complexité mécanique et l'accélération du véhicule.



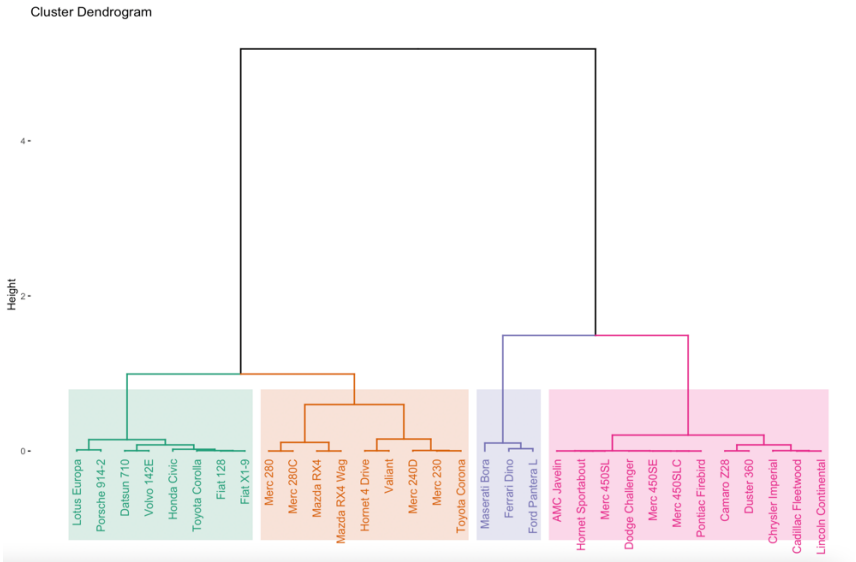
Enfin nous pouvons visualiser chaque individu par rapport à sa répartition selon les dimensions principales.

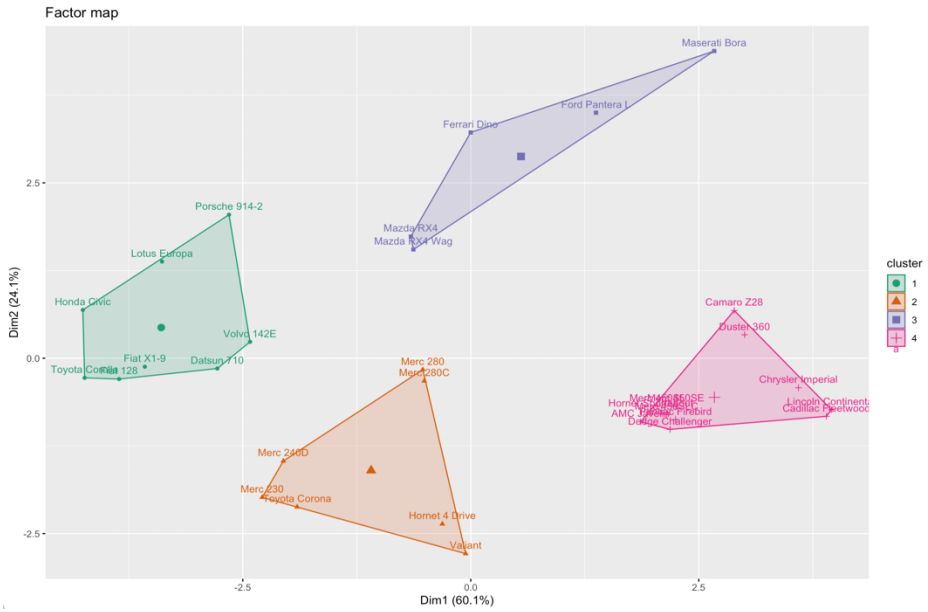
IV. Clustering hiérarchique.

Une fois les dimensions réduites, les méthodes de clustering comme les classifications hiérarchiques peuvent être appliquées.
Elles vont permettre de regrouper en quelques classes les unités d'une base de données statistique en sous-groupes. Il s'agit de méthodes d'analyse non supervisée.

Il existe la classification ascendante hiérarchique, la classification descendante hiérarchique(segmentation) et la classification mixte hiérarchique (on mélange deux formes de classification).

Pour finaliser notre étude qui est la construction de clusters nous utilisons la classification ascendante hiérarchique qui va regrouper les unités qui se ressemble.





Ce dendrogramme (1er schéma) et la carte factorielle des clusters (2ème schéma) montrent les regroupements entre individus. Nous pouvons voir 4 groupes bien séparés qui représente nos clusters et leurs positions dans l'espace des dimensions principales.

V. Conclusion.

Ainsi, le cluster violet (dimension 1 et 2 positives) contenant la Ferrari Dino, la Ford Pantera L, La maserati Bora etc regroupe les véhicules puissants, mécaniquement sophistiqués mais lourd et peu économes.
Le groupe vert (dimension 1 négative et 2 positive) avec la Honda Civic ou la Porsche 914-2 contient des véhicules économes et bien équipés mécaniquement.

Le cluster orange (dimension 1 négative et 2 négatives) regroupant la Merc 280, 230 ou la Toyota Corona représente des modèles économes mais peu performant

Enfin, le cluster rose (dimension 1 positive et la 2 négative) avec la Camaro Z28, la Duster 360 contient les véhicules combinant la puissance et la vitesse d'accélération, cela pourrait regrouper les voitures de sport.