Analyse exploratoire
et
Modèle pour la Data Science

Elèves :

Maxime ERDEM
Milan CORREGGIO
Mathieu LE CAM

Sommaire

I Analyse exploratoire

1) ACP

Dans ce chapitre, nous allons étudier un set de données concernant les produits vendus par un grossiste. Il comprend les dépenses annuelles sur diverses catégories de produit.
Nous allons effectuer une ACP afin d'étudier le comportement des clients, c'est-à-dire observer ce qu'ils achètent le plus souvent et les catégoriser.

a. Processus

Après avoir exécuté la commande M=PCAshiny(Data) on obtient l'information suivante :

```
Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
Variance               2.645   1.703   0.740   0.564   0.286   0.063
% of var.             44.083  28.376  12.334   9.396   4.761   1.050
Cumulative % of var.  44.083  72.459  84.794  94.189  98.950 100.000

Individuals (the 10 first)
                     Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
1                  | 0.795 |  0.193  0.003  0.059 | -0.305  0.012  0.147 |  0.141  0.006  0.031 |
2                  | 0.753 |  0.434  0.016  0.333 | -0.328  0.014  0.190 | -0.319  0.031  0.179 |
3                  | 2.332 |  0.811  0.057  0.121 |  0.815  0.089  0.122 | -1.523  0.713  0.427 |
4                  | 1.133 | -0.779  0.052  0.472 |  0.653  0.057  0.332 | -0.163  0.008  0.021 |
5                  | 1.577 |  0.166  0.002  0.011 |  1.271  0.216  0.650 | -0.066  0.001  0.002 |
6                  | 0.736 | -0.156  0.002  0.045 | -0.295  0.012  0.161 | -0.148  0.007  0.040 |
7                  | 0.738 | -0.335  0.010  0.206 | -0.525  0.037  0.506 |  0.303  0.028  0.169 |
8                  | 0.623 |  0.141  0.002  0.051 | -0.231  0.007  0.137 | -0.390  0.047  0.392 |
9                  | 0.884 | -0.517  0.023  0.343 | -0.659  0.058  0.557 | -0.183  0.010  0.043 |
10                 | 1.782 |  1.592  0.218  0.799 | -0.741  0.073  0.173 | -0.210  0.014  0.014 |

Variables
                    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
Fresh             | 0.070  0.184  0.005 |  0.689 27.871  0.475 |  0.699 65.976  0.488 |
Milk              | 0.887 29.715  0.786 |  0.109  0.692  0.012 | -0.052  0.365  0.003 |
Grocery           | 0.942 33.554  0.887 | -0.191  2.134  0.036 |  0.093  1.175  0.009 |
Frozen            | 0.083  0.262  0.007 |  0.798 37.366  0.636 | -0.153  3.182  0.024 |
Detergents_Paper  | 0.892 30.101  0.796 | -0.333  6.514  0.111 |  0.117  1.855  0.014 |
Delicassen        | 0.404  6.184  0.164 |  0.658 25.422  0.433 | -0.451 27.448  0.203 |
```

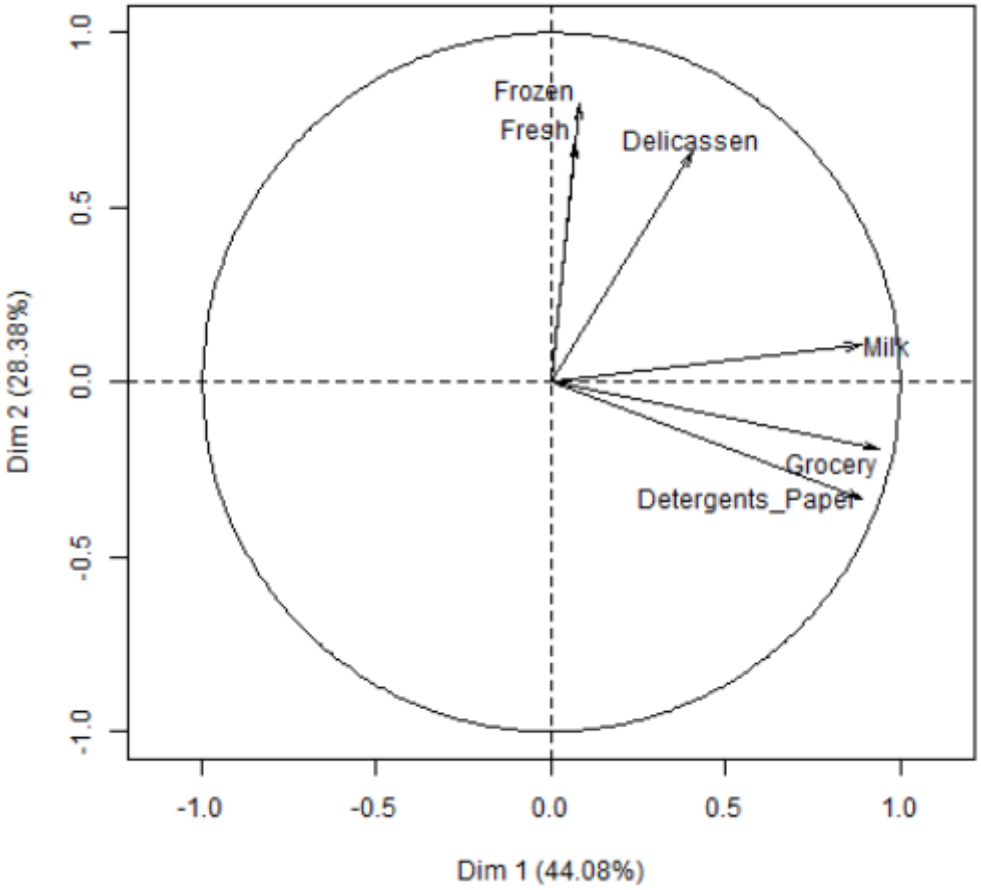
On peut observer les valeurs propres, le pourcentage de variance exprimé pour chaque axe ainsi que le pourcentage d'information qu'on a (cumulative % of var) si l'on combine les axes.

Le but est de choisir le nombre d'axes pertinents. Pour ce faire nous disposons de deux règles :

- Le critère de Kaiser (on prend tous les axes où les valeurs propres sont supérieures ou égales à 1, ici DIM1 et DIM2).
- La règle du coude, qui consiste à afficher la courbe de l'évolution de la proportion d'information contenue dans chaque axe et de sélectionner les axes situés avant le décrochage de la courbe)

  En utilisant le critère de Kaiser, on retient 2 axes (DIM 1 et DIM2) qui représente environ 73% de l'inertie totale.

b. Cercle de corrélation

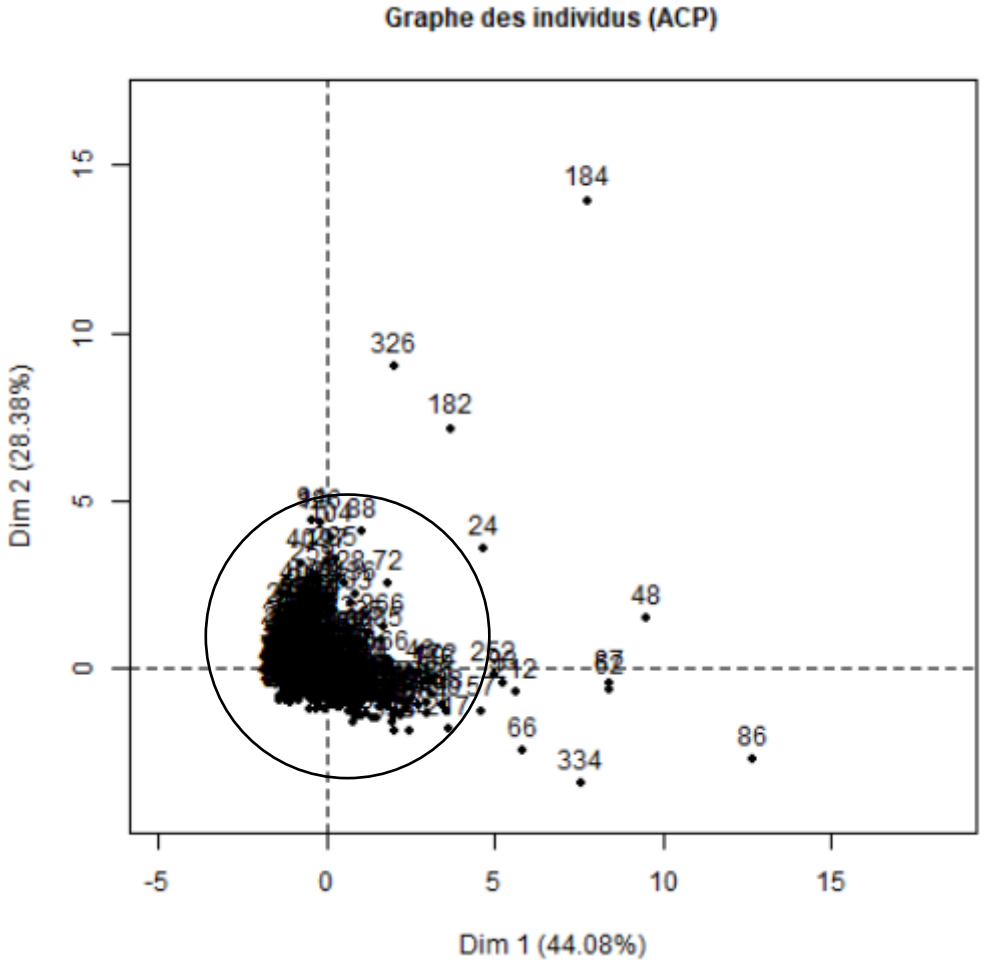


En analysant le cercle de corrélation obtenu par notre étude, on peut effectuer certaines observations.

On remarque que les variables Frozen and Fresh sont fortement corrélées et portées par l'axe 2.

De même on observe que les variables Detergents_Paper and Grocery sont également corrélées.

c. Graphe des individus



En analysant le graphe des individus, on remarque que la plupart des individus ont un comportement similaire car ils sont proches les uns des autres dans le graphique (nuage de point important). On note toutefois que certains individus se démarquent des autres car ils ont un comportement d'acheteurs différent et achètent uniquement certains produits.

Nous pouvons donc classer les individus en trois catégories :

- Les individus qui consomment beaucoup de produit laitiers, détergents et épiceries.
- Les individus qui consomment beaucoup de produit frais, surgelés et hors d'œuvres.
- Les individus qui consomment à la fois des produits laitiers, détergents, épiceries et surgelés, frais, hors d'œuvres.

2) AFC

Dans ce chapitre, nous allons étudier le recensement des arbres dans la région parisienne. Pour ce faire nous utiliserons le tableau de correspondance exprimant les domanialités (en ligne) en fonction d'ensembles de région (en colonne). Ces deux variables sont qualitatives.

Par la suite, nous analyserons le graphe de l'AFC.

	A	B	C	D	E	F	G	H	I	J	K	L
1	DOMANIALITE	BOIS DE BOULOGN	BOIS DE VINCENNE	HAUTS-DE-SEINE	PARIS 10E ARRDT	PARIS 11E ARRDT	PARIS 12E ARRDT	PARIS 13E ARRDT	PARIS 14E ARRDT	PARIS 15E ARRDT	PARIS 16E ARRDT	PARIS 17E ARRDT
2	Alignement	4205	5794	0	2665	4483	7942	11561	7118	8474	11012	635
3	CIMETIERE	0	0	5425	0	0	10	0	1245	157	321	79
4	DAC	0	0	0	0	0	0	9	0	0	0	
5	DASCO	0	0	0	235	360	515	944	320	717	361	42
6	DASES	0	0	0	0	0	4	0	0	1	0	
7	DFPE	0	0	0	53	100	111	314	91	131	36	10
8	DJS	2	0	0	104	112	535	999	279	356	439	22
9	Jardin	1	4970	0	491	928	3389	3639	2738	7918	3523	351
10	PERIPHERIQUE	0	0	0	0	0	600	568	128	276	1748	46
11	PRIVE	0	0	0	118	77	1064	796	2533	945	2048	19

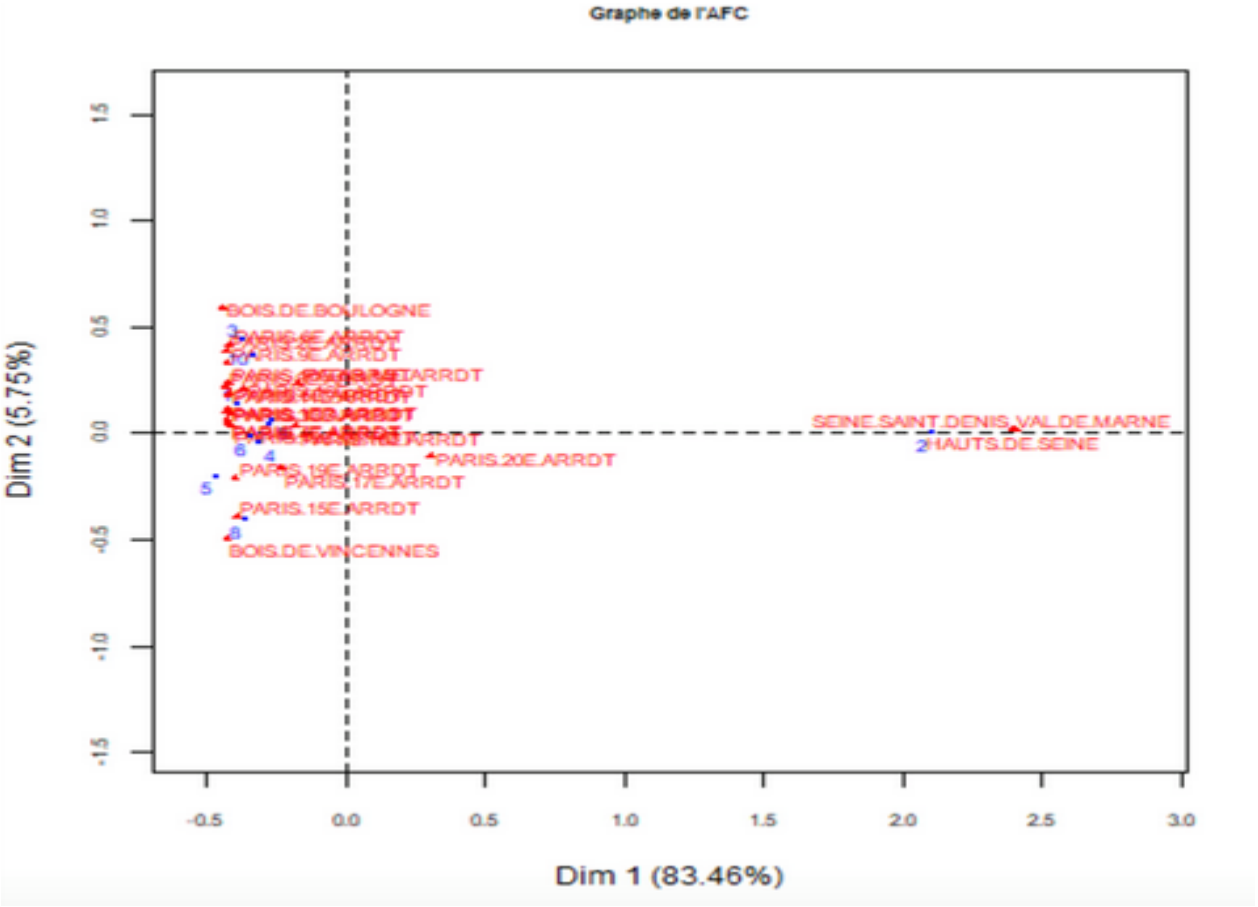
Tableau de correspondances

	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	PARIS 1ER ARRDT	PARIS 20E ARRDT	PARIS 2E ARRDT	PARIS 3E ARRDT	PARIS 4E ARRDT	PARIS 5E ARRDT	PARIS 6E ARRDT	PARIS 7E ARRDT	PARIS 8E ARRDT	PARIS 9E ARRDT	SEINE-SAINT-DENIS	VAL-DE-MARNE
2	1116	5415	477	897	2014	1715	1387	6230	5974	935	0	0
3	0	4355	0	0	0	0	0	0	0	0	11769	7622
4	0	1	0	0	0	0	0	0	0	2	0	0
5	21	889	43	67	121	180	47	63	37	95	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0
7	0	136	1	14	17	29	7	4	5	19	0	0
8	0	649	0	0	9	7	0	20	0	0	0	30
9	329	3911	39	286	675	614	376	2597	1416	150	0	1
10	0	955	0	0	0	0	0	0	0	0	0	0
11	54	963	10	28	127	413	513	967	424	96	0	0

On observe que toutes les régions géographiques ne disposent pas toutes de chaque type de domanialités

Par exemple, la domanialité « cimetière » compte beaucoup d'arbres dans le Haut de Seine, Val de Marne et Seine Saint-Denis

Examinons maintenant ce que nous apprend l'analyse des correspondances sur le tableau :



Indiquons simplement que la proximité entre une ligne et une colonne suggère une association privilégiée entre cette ligne et cette colonne.

Ainsi Seine Saint-Denis, Val de Marne et Haut de Seine tombent exactement au même endroit que cimetière.

Ce qui correspond bien à l'association exclusive que nous avions notée dans notre interprétation directe.

Le premier axe (DIM 1) met en évidence une opposition entre cimetière et Seine Saint-Denis, Val de Marne, Haut de Seine d'une part et tous les autres points d'autre part.

Si l'on regarde le deuxième axe, on a une opposition entre d'une part Bois de Vincennes (jardin), Paris 15 (jardin) etc. et d'autre part Paris 9 (privée) etc.

On peut dire par exemple que privée et jardin sont plus proches entre eux qu'ils ne sont de cimetière.

3) ACM

a. Introduction concernant la base de données utilisée

Nous disposons d'une base de données d'automobiles vendues neuves en 1985. Chaque individu, c'est-à-dire un modèle d'automobile est décrit par 25 variables. Parmi ces variables, on trouve le nom de la marque du véhicule, de nombreuses variables quantitatives ou qualitatives décrivant les aspects techniques du véhicule (dimension, type, motorisation, consommation…) mais aussi le prix et une mesure du facteur de risque qu'implique un tel véhicule (sur une échelle d'entiers allant de -3 à 3, 3 correspondant au risque le plus élevé).

On peut noter à propos de cette base de données :

- Il y a très peu de valeurs manquantes (moins de 10 individus sur 205)
- La base de données représente le marché automobile mondial de 1985. On retrouve donc des véhicules en provenance de tous les pays mais avec les spécificités de cette époque. Nous trouvons donc certaines marques aujourd'hui disparues et certains types de voiture n'existaient pas encore ou étaient très peu développés (SUVs, monospaces) sont absents. On note également qu'à première vue, le prix des véhicules neufs semble anormalement bas mais il faut considérer qu'à cette époque, le pouvoir d'achat disponible avec un dollar était bien plus élevé qu'aujourd'hui (30 ans d'inflation ayant eu cours depuis).
- Un défaut de cette base de données est qu'elle ne comporte pas de variable désignant le modèle du véhicule (seulement sa marque). Il est donc assez difficile de distinguer deux modèles de la gamme d'un même constructeur.

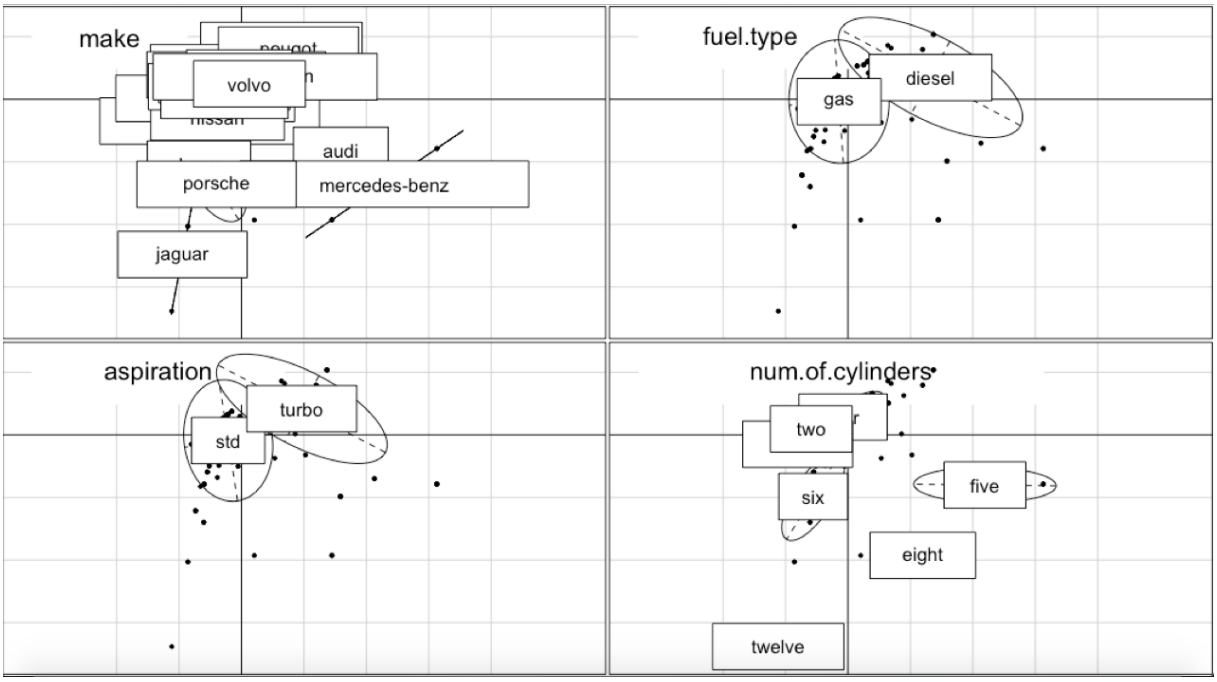
Certains individus (moins d'une dizaine) comportaient quelques valeurs manquantes signifiées par des « ? », nous les avons donc simplement éliminés.

De plus, certaines données était au format « factor » que nous avons donc converties en valeurs numériques.

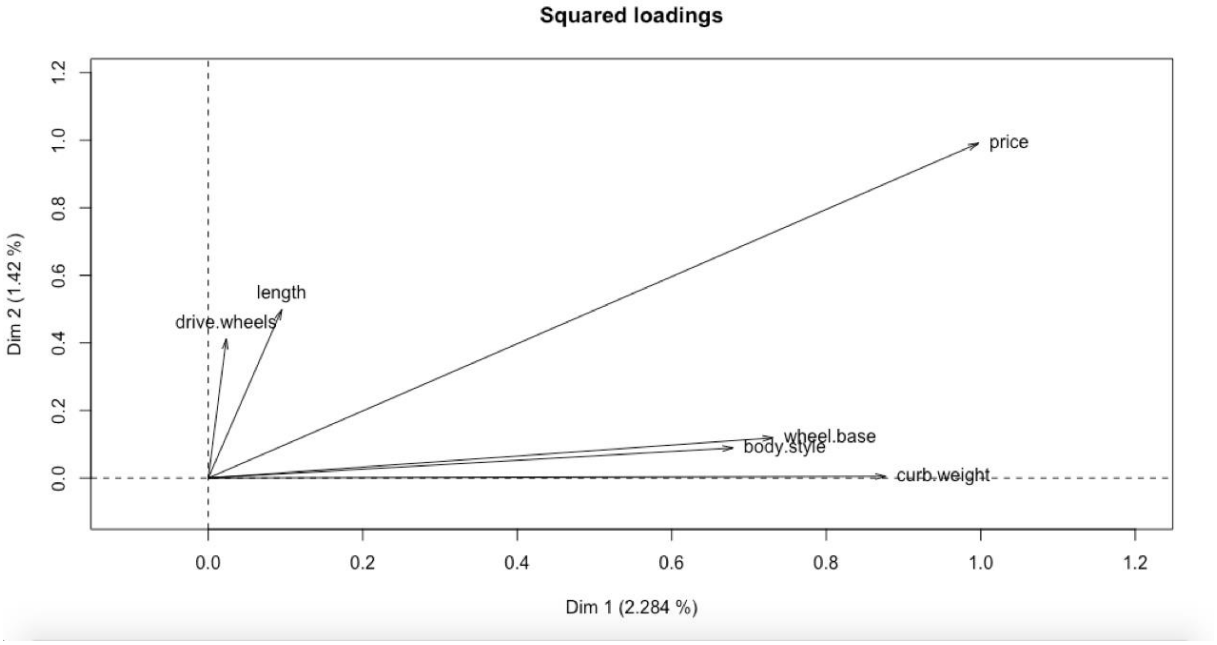
b. ACM, ACP mixte

Notre set de données est sur les caractéristiques des voitures et leur prix. Nous cherchons à savoirs sur les données que nous avons de quoi le prix d'une voiture dépend.

Après avoir analysé les 4 variables qui nous semblaient les plus intéressantes nous avons eu la figure ci-dessous :



Cela nous a permis de faire 2 constats.
Nous remarquons que la marque de la voiture et le nombre de cylindre sont corrélés, plus on a de cylindres, plus la voiture est luxueuse.
De plus, nous remarquons que l'aspiration de type turbo et le diesel ont presque les mêmes ellipses de corrélation. En effet, cela s'explique par le fait qu'en 1985, le turbo était une technologie nouvelle qui était surtout, voir uniquement employée sur les moteurs diesels (et non pas essence comme c'est le cas aujourd'hui)



Comme le montre le graphique ci-dessus nous remarquons que le prix des voitures dépend de nombreuses.

Cependant le poids très faible de chaque axe (aux alentours de 2%) nous amène a nous questionner sur la pertinence de ce modèle.

II Modèles de données

1) Subset selection et modèles de Shrinkage

a. Subset selection

Dans ce chapitre, nous utiliserons la même base de données que lors du chapitre précédent.

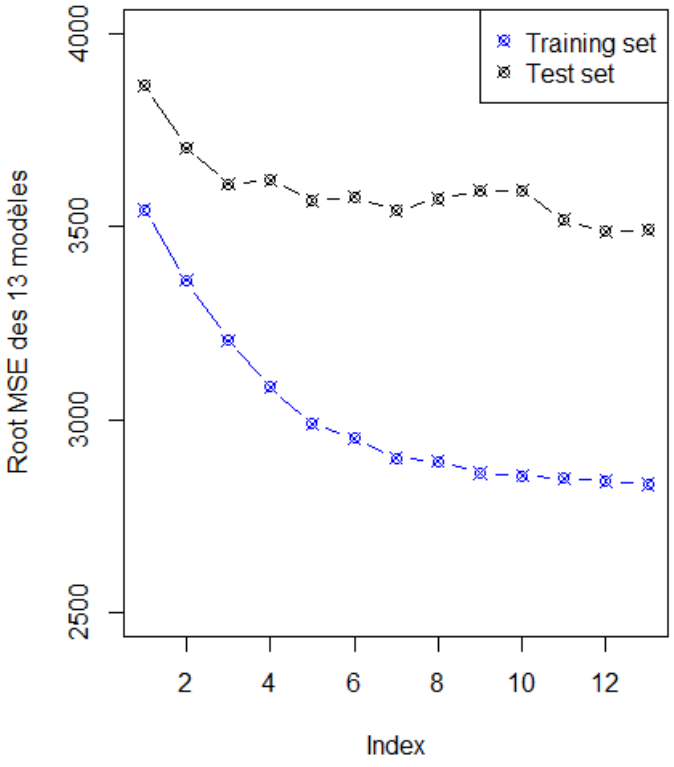
Nous allons tenter de prédire le prix des voitures en utilisant l'ensemble des 13 autres variables quantitatives mises à notre disposition.

Le nombre de variables étant imposant, on va chercher à sélectionner les plus importantes afin de simplifier le modèle, le rendre plus facilement interprétable et réduire les temps de calculs.

A partir de notre échantillon d'origine composé de 205 individus, on construit un échantillon de test de 140 individus.

En utilisant la méthode de la best subset selection sur notre ensemble de test, on va déterminer quelle est la meilleure combinaison de variables à utiliser parmi les 13 à notre disposition pour expliquer la variable prix. On emploiera pour cela la méthode « forward stepwise » qui, détermine 13 modèles ayant chacun un nombre de variables différents.

On calcule par la suite la RMSE entre le modèle et le résultat réel pour chacun des 13 modèles à la fois sur l'échantillon d'entrainement et sur l'échantillon de test

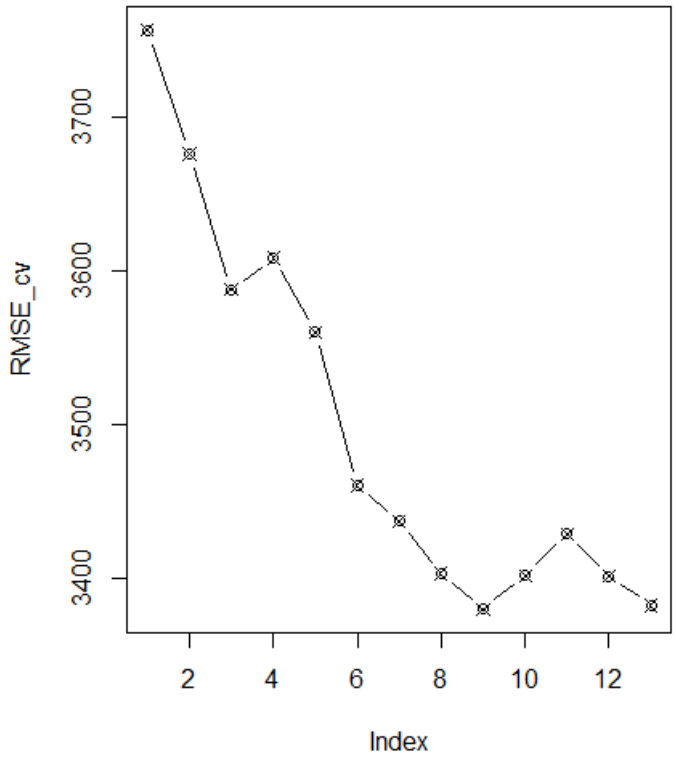


On observe que la RMSE reste toujours plus élevée sur l'échantillon de test que sur l'échantillon de training. L'ajout de variables au modèle améliore son efficacité sur le set d'entraînement de manière significative jusqu'à un nombre de 7 variables. Cette amélioration est présente sur l'échantillon de test mais de manière plus limitée et essentiellement marquée jusqu'à l'ajout d'une cinquième variable. Toutefois, bien que l'amélioration du modèle soit quasi-nulle après 5 variables, il n'y a pas de sur-apprentissage de présent (ou du moins pas de manière significative).

b. Cross validation

On applique maintenant la même technique du forward stepwise mais cette fois-ci en cross-validation sur notre set de données. Le set de données est divisé en 10 parties de tailles égales qui serviront tout à tour de set de test et on calcule la moyenne des 10 MSE déterminées.

Le graphique suivant affiche la valeur moyenne des RMSE pour chaque modèle de 1 à 13 variables :



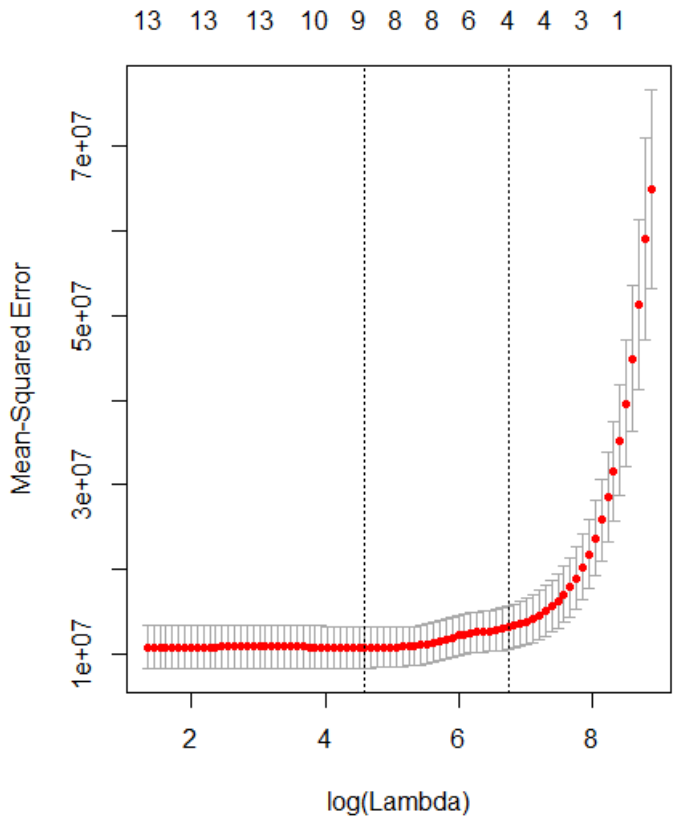
On observe que le modèle le plus efficace est celui utilisant 9 variables. Les modèles ayant un plus grand nombre de variables n'apportent pas de gain en performance et peuvent même conduire à un léger sur-apprentissage.

c. Le modèle de Shrinkage LASSO

On va maintenant chercher à déterminer le modèle le plus performant par la technique du LASSO. Avec cette technique, on cherche à éviter que les coefficients des paramètres du modèle soient trop grands en les pénalisants afin d'éviter tout sur-apprentissage.

En utilisant la fonction cv.glmnet, on va effectuer une cross validation sur un set de modèles LASSO pour un certain nombre de valeurs différentes de lambda. Lambda étant le « Tuning Parameter » dont la valeur influence directement la pénalisation de la taille des coefficients. On obtient donc une centaine de modèles LASSO avec des lambda différents.

On affiche dans un graphique l'évolution de la MSE en fonction de Lambda (ou plutôt du logarithme de lambda pour des raisons de lisibilité) :



On trouve que le log(lambda) minimum vaut 4.59.

```
lasso_model_cv$glmnet.fit$beta[,numero_du_best_model]
     wheel.base           length            width           height      curb.weight      engine.size             bore           stroke compression.ratio
      0.0000000        0.0000000      507.1546595      210.8084844        0.6889787      120.7467786        0.0000000    -2175.7562449      183.5887165
     horsepower         peak.rpm         city.mpg      highway.mpg
     38.6578039        1.6925630        0.0000000        0.0000000
```

Les résultats du modèle de Shrinkage LASSO affiché ci-dessus nous permettent de déterminer l'importance de chaque variable dans la prédiction du prix des véhicules.

2) Arbres de régression et de classification

a. Arbre de régression

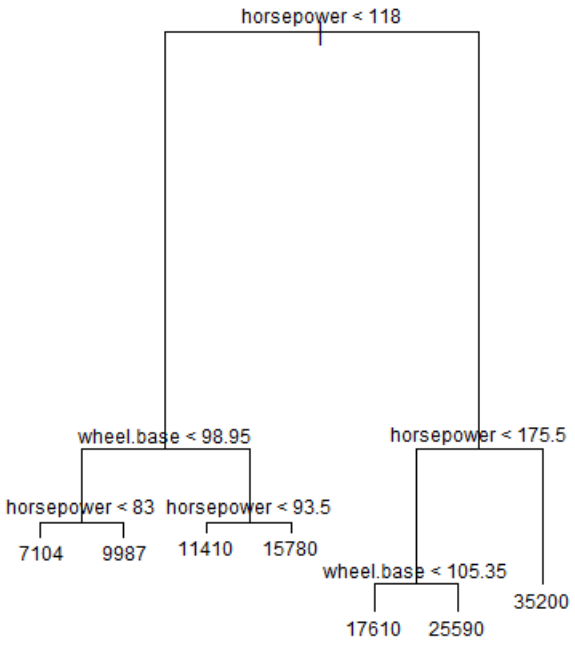
Toujours sur le même set de données des véhicules, nous allons mettre en application la méthode de l'arbre de régression pour prédire le prix d'un véhicule en fonction de sa puissance et de son empattement.

Pour rappel :

- La puissance (horsepower) délivrée par le moteur influe directement sur les performances du véhicule en termes d'accélération et de vitesse de pointe.
- L'empattement (wheel.base) désigne la distance entre l'essieu avant et l'essieu arrière. Les voitures se voulant luxueuses et imposantes tendent généralement à avoir un empattement le plus important possible afin d'offrir de l'espace et du confort à leurs passagers

La variable du prix, de la puissance et de l'empattement sont toutes trois des variables quantitatives continues, ce qui permet donc de les utiliser pour effectuer un arbre de régression.

En utilisant la fonction tree(), on obtient l'arbre de régression linéaire suivant :

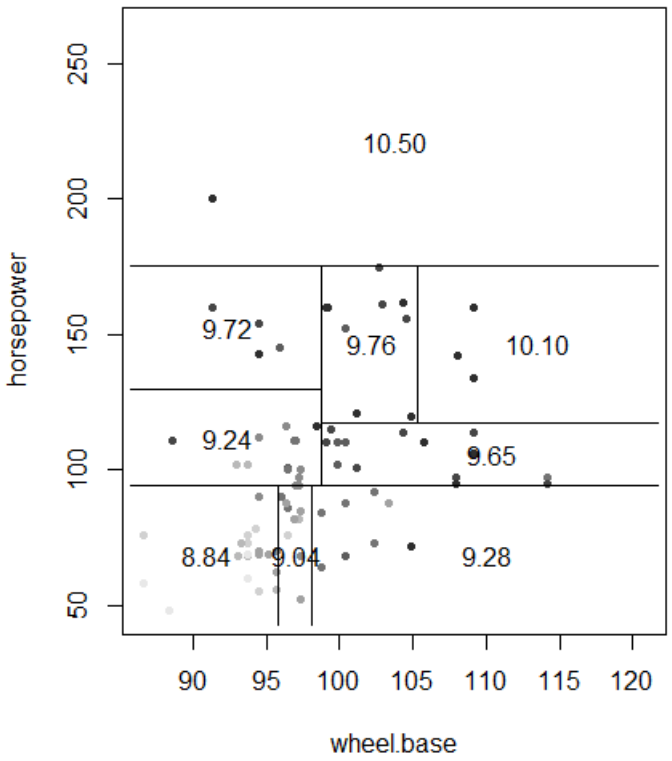


On obtient ainsi sept classes qui reflètent des résultats auxquels on aurait pu penser intuitivement, la classe de véhicules les moins chers est effectivement celle pour laquelle la longueur et la puissance du véhicule sont les plus faibles. De même, les véhicules les plus chers sont ceux ayant les puissances et les longueurs d'empattement les plus élevées.

On note également que c'est principalement la puissance qui détermine le prix des voitures, la longueur de l'empattement n'intervenant que dans un second temps.

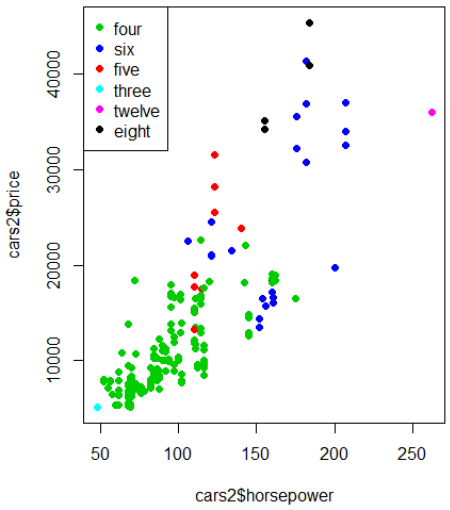
b. Observation de classes :

On affiche dans un graphique la répartition des individus en fonctions de leur puissance et de leur longueur de d'empattement. Les individus sont également représentés dans une variance de gris plus ou moins foncée selon leur prix. Enfin on partitionne le graphique en différentes sections pour lesquelles on aura la probabilité qu'un individu se trouve dans chacune d'elles.



Dans ce graphique, les frontières des partitions correspondent aux critères de décisions utilisés dans l'arbre de régression précédent.

Il est également intéressant d'observer graphiquement le prix des véhicules en fonction de leur puissance tout en affichant une variable qualitative pour chaque individu. Ainsi on peut observer le nombre de cylindres ou le type de carrosserie des individus :



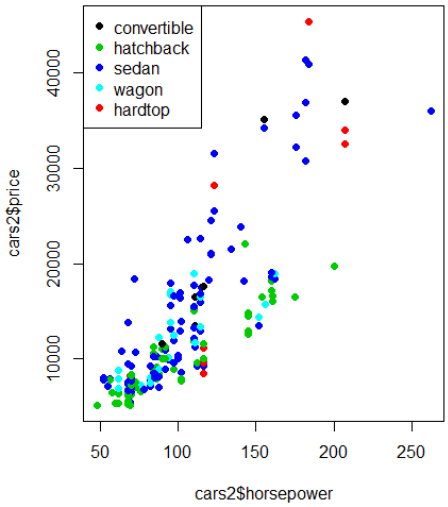
Comme on pouvait s'y attendre, plus le nombre de cylindres du moteur est élevé, plus la puissance et le prix du véhicule le sont également.

Cependant, on observe que certains moteurs 4 cylindres (en vert) sont aussi puissants que des moteurs à 6 cylindres en bleu.

Nous disposons d'une variable qualitative nommée « aspiration » qui décrit le type d'admission en air du moteur des véhicules. En effet, les moteurs peuvent comporter un turbo-compresseur qui permet d'augmenter la quantité d'air présent dans le moteur et donc de développer plus de puissance tout en conservant une taille de moteur équivalente.

Si l'on affiche le même graphique en colorisant les individus cette fois en fonction de s'ils possèdent un turbo (rouge) ou non (noir), on constate que les moteurs quatre cylindres aussi puissants que les six cylindres le sont car ils possèdent un turbo alors que les six cylindres non.

Il est aussi intéressant d'observer la puissance et le prix du véhicule en fonction de son type de carrosserie.



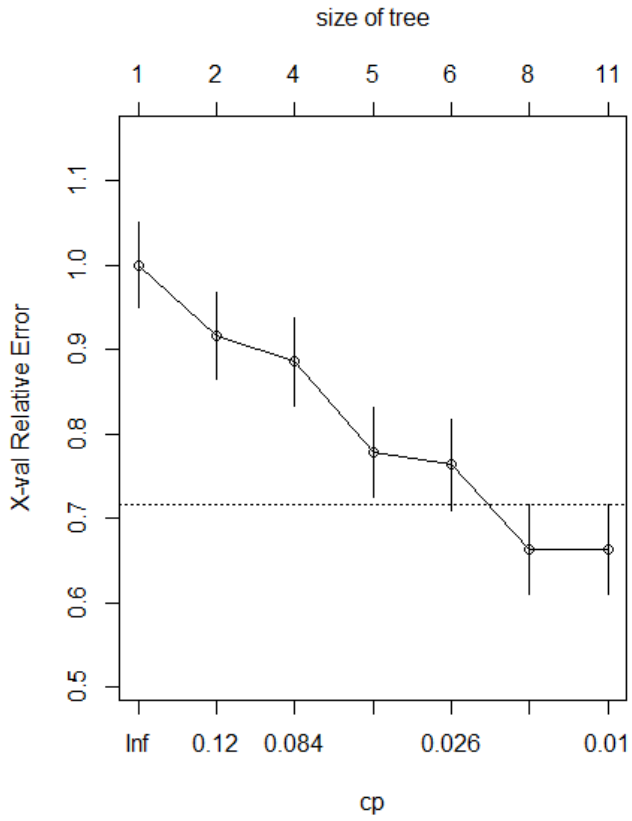
On observe alors que les véhicules de type « Convertible » (cabriolets) et « Hard-top » (coupés) se situent souvent parmi les véhicules les plus puissants et/ou coûteux. A l'inverse, les carrosseries « Hatchback » (compactes à hayon) et « Wagon » (breaks) sont le plus souvent peu coûteuses car ce sont des véhicules populaires. Enfin, les véhicules « sedan », c'est-à-dire les berlines, ont des prix et puissances très variés car il peut à la fois s'agir de berlines abordables et familiales mais aussi de berlines de luxe et de prestige.

c. Arbre de classification avec l'algorithme CART

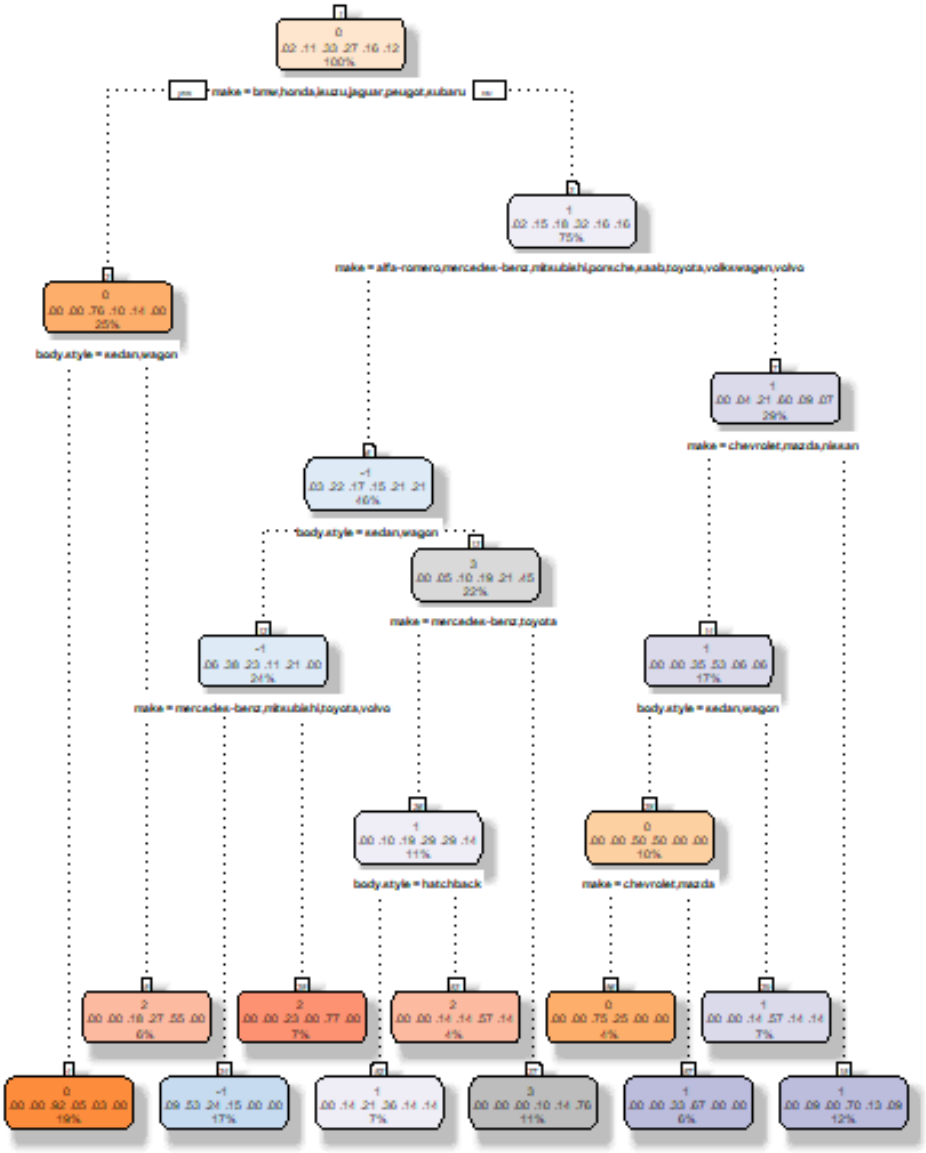
Nous allons maintenant réaliser une classification en utilisant des variables qualitatives. On va classifier les véhicules selon leur indice de risque (« symboling »). (Pour rappel, il s'agit d'une mesure prenant une valeur entière de -3 à 3, 3 étant attribué aux véhicules dits les plus risqués). On prendra comme variables explicatives la marque du véhicule (« make ») ainsi que le type de carrosserie (body.style) toutes deux variables qualitatives également.

On utilise la fonction rpart() afin de réaliser l'arbre de classification en employant l'algorithme CART.

On calcule le coefficient de Gini pour chaque nœud de notre arbre. Celui-ci voit bien sa valeur baisser lorsque l'indice du nœud augmente.



On affiche maintenant la représentation graphique de l'arbre de classification déterminé grâce à l'algorithme CART.



Les éléments de décision sont indiqués en utilisant une police très petite sur cette représentation. Il est donc préférable d'exécuter le programme sous R pour une meilleure lisibilité.

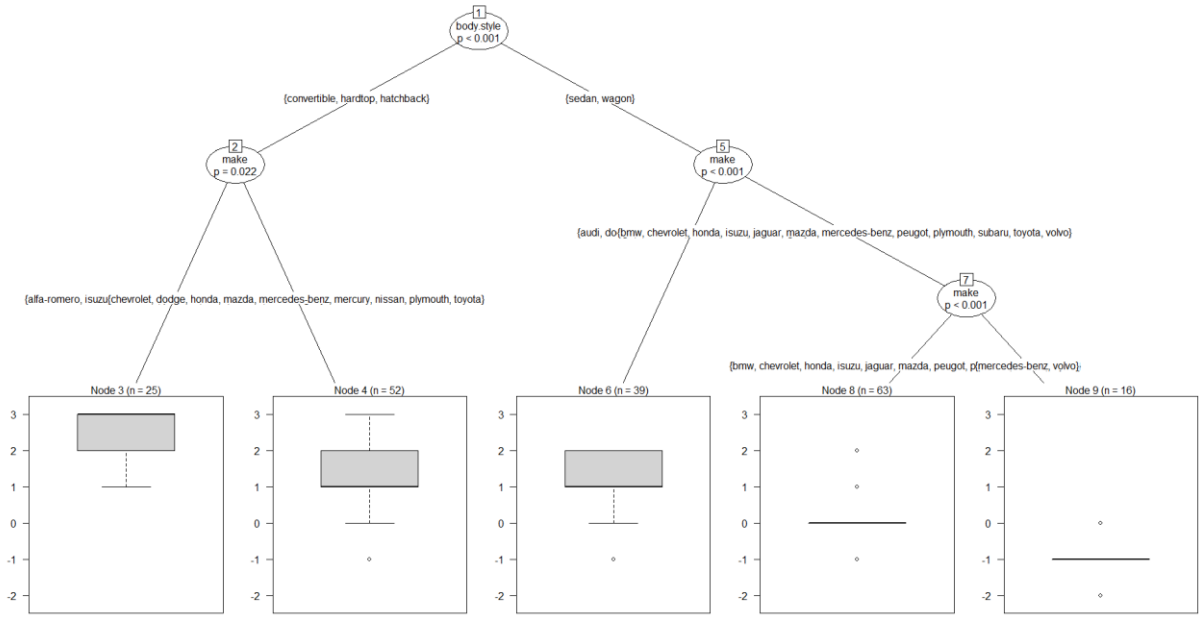
On remarque que la décision pour la classification se fait dans un premier temps sur la marque des véhicules.

Par la suite, le critère de décision employé peut être également le type de carrosserie. Dans ce cas alors, le critère consiste le plus souvent à déterminer s'il s'agit de carrosseries « hatchback » ou « wagon » puisque celles-ci correspondent à des véhicules peu coûteux (comme vu précédemment) et donc qui on la plupart du temps un indice de risque faible.

On peut également déterminer que certaines marques induisent un risque faible comme Volvo par exemple.

d. Arbre de classification avec l'algorithme CTREE

En utilisant la méthode CTREE pour produire un arbre de classification basé sur les mêmes variables que précédemment, on obtient l'arbre suivant :



On obtient cette fois-ci cinq classes différentes. Pour chaque classe on a une représentation de la répartition statistique des individus au sein de la classe en fonction de leur facteur de risque.

Le partitionnement s'effectue cette fois d'abord sur le type de carrosserie puis sur la marque du véhicule.

On note que ce sont principalement les véhicules de type convertible, hardtop et hatchback qui vont conduire à une classification avec un fort risque. Cela est dû au fait que ces carrosseries correspondent ou peuvent correspondre à des véhicules sportifs.

Les berlines (« sedan »), bien que nous ayons vu qu'elles peuvent disposer de moteurs puissants, conduisent principalement à un risque moins élevé. En effet, cela peut s'expliquer par le fait qu'un gros moteur sur ces véhicules parfois imposants est plutôt destiné à fournir un certain agrément de conduite plutôt qu'à permettre un usage sportif dangereux. Les véhicules de type break « wagon » ont quant à eux vocation à être familiaux et à autoriser un volume de chargement élevé et sont donc des véhicules conduisant à un risque faible.

Au sein même d'un type de carrosserie, la marque du véhicule a également une influence sur le risque encouru par le véhicule. En effet, une marque, grâce à son image, attire certains clients plus que d'autres et ayant des styles de conduite différents plus ou moins risqué. De même, pour un type de carrosserie similaire, certains constructeurs offrent leurs modèles avec des motorisations plus puissantes et donc plus risquées que d'autres.

3) Text mining

Pour ce chapitre, nous allons abandonner notre set de données de véhicules et opter pour un set de données comportant les paroles de plusieurs chanteurs. Nous avons choisi de nous occuper d'une chanteuse en particulier car les autres chansons de ce set de données étaient soient en trop petit nombre soient corrompues.

L'objectif de cette étude est de déterminer les termes que la chanteuse américaine Beyonce emploie le plus et si c'est en accord avec ce à quoi on aurait effectivement pu imaginer.

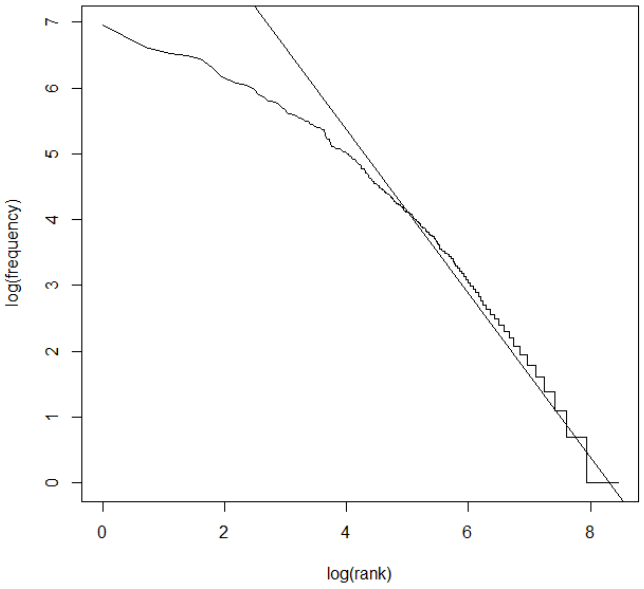


Figure 1

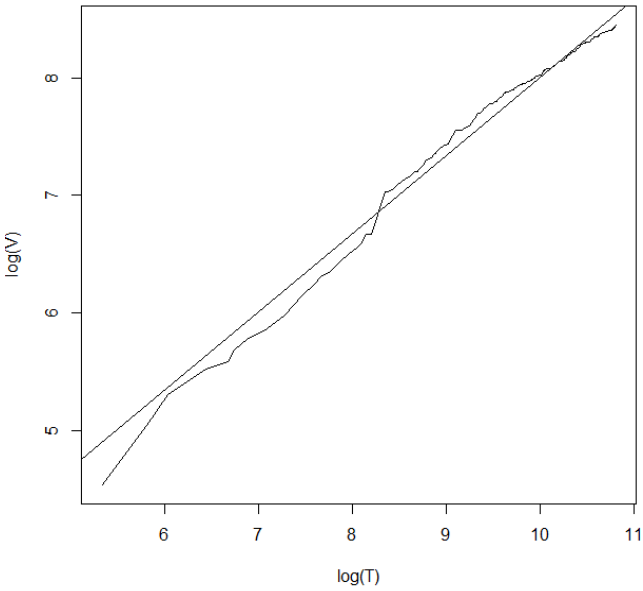


Figure 2

La figure 1 représente les résultats du test de Zipf alors que la figure 2 représente la loi de Heaps.

La figure 1 montre que le nombre de nouveaux mots décroit au fur et à mesure de l'analyse des textes. La figure 2 montre que le nombre de mot distincts augmente au fur et à mesure de l'analyse. Ce résultat est cohérent car plus on a de mot plus ils ont de chances d'être distinct et moins on a de chance d'en trouver de nouveau.

Résultats :

Nous remarquons que le dictionnaire n'utilise que des mots anglais puisque Beyonce est américaine. Les mots les plus employés correspondent à l'idée qu'on se fait des chansons de Beyoncé qui parlent d'amour et de sentiments