ProjetClustering

Karidjatou Diaby Libasse Mboup Saliou Ndao Deffa Ndiaye

19/04/2021

Chargement données

```
data <- read.csv("D:/desktop/Lyon 2/M1/Clustering/Cars.csv", row.names = "Lib_Model")
head(data)
```

```
##                          X Continent   Type Cod_DriveTrain  MSRP Invoice
## 3.5 RL 4dr               1      Asia  Sedan              1 43755   39014
## 3.5 RL w/Navigation 4dr  2      Asia  Sedan              1   461     411
## MDX                      3      Asia    SUV              3 36945   33337
## NSX coupe 2dr manual S   4      Asia Sports              2 89765   79978
## RSX Type S 2dr           5      Asia  Sedan              1  2382   21761
## TL 4dr                   6      Asia  Sedan              1 33195   30299
##                          EngineSize Cylinders Horsepower MPG_City MPG_Highway
## 3.5 RL 4dr                      3.5         6        225       18          24
## 3.5 RL w/Navigation 4dr         3.5         6        225       18          24
## MDX                             3.5         6        265       17          23
## NSX coupe 2dr manual S          3.2         6        290       17          24
## RSX Type S 2dr                  2.0         4        200       24          31
## TL 4dr                          3.2         6        270       20          28
##                          Weight Wheelbase Length  Lib_Make
## 3.5 RL 4dr                 3880       115    197     Acura
## 3.5 RL w/Navigation 4dr    3893       115    197      Audi
## MDX                        4451       106    189       BMW
## NSX coupe 2dr manual S     3153       100    174     Buick
## RSX Type S 2dr             2778       101    172  Cadillac
## TL 4dr                     3575       108    186 Chevrolet
```

population : Lib_Model : Le modèle des voitures . Les variables sont les critères des voitures.

- variables quantitatives : cod_model, MSRP, Invoice, Enginesize, horsepower, mpg_city, mpg_highway, Weight, Wheelbase, Lenght
- variables qualitatives : Continent, Type, Cod_DriveTrain ,Cylinders, Lib_Model, cod_make, Lib_Make
- variables actives : MSRP, Invoice, EngineSize, Horsepower, MPG_City, MPG_Highway, Weight, Wheelbase, Length
- variables supplémentaires : continent, type, Cylinders, Lib_Make : marques

variable à supprimer : cod_model et Cod_DriveTrain (non pertinentes)

```
data <-data[, -1] #supp cod_model
data <-data[, -3] #supp Cod_DriveTrain

head(data)
```

```
##                          Continent   Type  MSRP Invoice EngineSize Cylinders
## 3.5 RL 4dr                    Asia  Sedan 43755   39014        3.5         6
## 3.5 RL w/Navigation 4dr       Asia  Sedan   461     411        3.5         6
## MDX                           Asia    SUV 36945   33337        3.5         6
## NSX coupe 2dr manual S        Asia Sports 89765   79978        3.2         6
## RSX Type S 2dr                Asia  Sedan  2382   21761        2.0         4
## TL 4dr                        Asia  Sedan 33195   30299        3.2         6
##                          Horsepower MPG_City MPG_Highway Weight Wheelbase Length
## 3.5 RL 4dr                      225       18          24   3880       115    197
## 3.5 RL w/Navigation 4dr         225       18          24   3893       115    197
## MDX                             265       17          23   4451       106    189
## NSX coupe 2dr manual S          290       17          24   3153       100    174
## RSX Type S 2dr                  200       24          31   2778       101    172
## TL 4dr                          270       20          28   3575       108    186
##                           Lib_Make
## 3.5 RL 4dr                   Acura
## 3.5 RL w/Navigation 4dr       Audi
## MDX                            BMW
## NSX coupe 2dr manual S       Buick
## RSX Type S 2dr            Cadillac
## TL 4dr                   Chevrolet
```

vérifions si le fichier contient des doublons

```
#duplicated(data)
```

le fichier ne contient donc pas de doublons

vérifions s'il exite des NA et suprimons les

```
#is.na(data)
```

Est-ce que toutes les voitures se ressemblent ?

Si tel n'est pas le cas, quelles voitures se ressemblent, quelles voitures diffèrent ?

Est-ce que certaines caractéristiques sont liées ?

Quels critères peuvent expliquer la taille de la voiture (EngineSize) ?

————————————————————-REALISATION DE L'ACP————————————————————-

```
colnames(data)
```

```
##  [1] "Continent"   "Type"        "MSRP"        "Invoice"     "EngineSize"
##  [6] "Cylinders"   "Horsepower"  "MPG_City"    "MPG_Highway" "Weight"
## [11] "Wheelbase"   "Length"      "Lib_Make"
```

```
v.active <- subset(data, select=c(MSRP, Invoice, EngineSize, Horsepower, MPG_City, MPG_Highway, Weight,
res.pca <- PCA(v.active, scale.unit = TRUE, ncp = 5, graph = FALSE)
```

pour les variables supplémentaires ont crée une nouvelle base de données avec toutes les variables quantitatives et la variable qualitative supplémentaire

```
print(res.pca)
```

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 425 individuals, described by 9 variables
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$call"            "summary statistics"
## 12 "$call$centre"     "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"      "weights for the individuals"
## 15 "$call$col.w"      "weights for the variables"
```

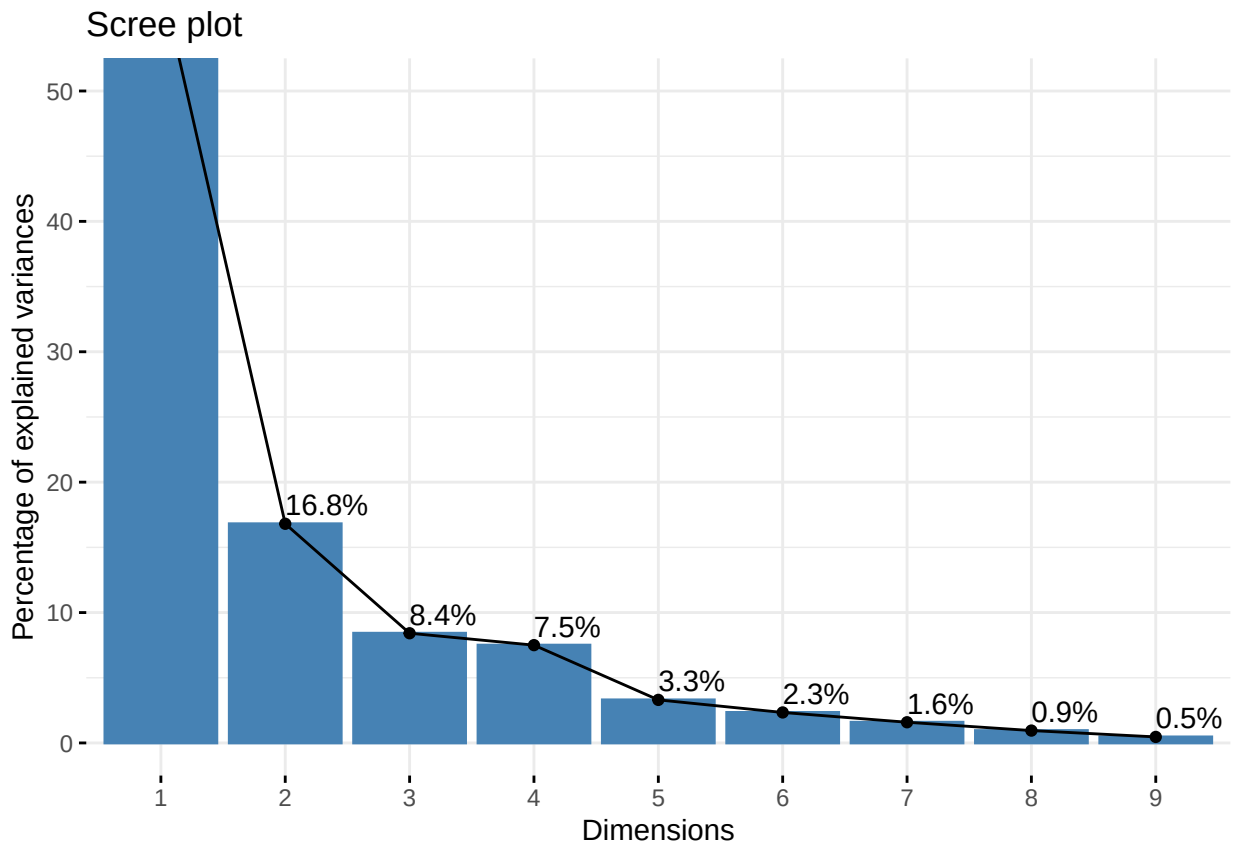
```
eig.val <- get_eigenvalue(res.pca) # get permet d'extraire de la variable res.pca les resultats qui cor
eig.val
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1 5.27721520       58.6357245                    58.63572
## Dim.2 1.51271571       16.8079523                    75.44368
## Dim.3 0.75725265        8.4139184                    83.85760
## Dim.4 0.67508026        7.5008918                    91.35849
## Dim.5 0.29726564        3.3029516                    94.66144
## Dim.6 0.21048772        2.3387525                    97.00019
## Dim.7 0.14288964        1.5876627                    98.58785
## Dim.8 0.08540636        0.9489596                    99.53681
## Dim.9 0.04168681        0.4631868                   100.00000
```

Grace au critère de keither on decide du nombre de valeur propre qu'on retient : les axes dont les valeurs propres sont supp à la valeurs propres moyennes On a 9 valeurs propres valeur propre moyenne dans un acp normé = 1 (100%) inertie / nb valeur propre (P/P= 1) 100/9 11,11%, donc on retient les 2 premiers axe en se concentrant plus sur le 1er car il comporte le plus d'inertie

visualisation

```
plot1<-fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 50)) #confirme ce qu'on a dit plus haut : 2 axes
plot1
```

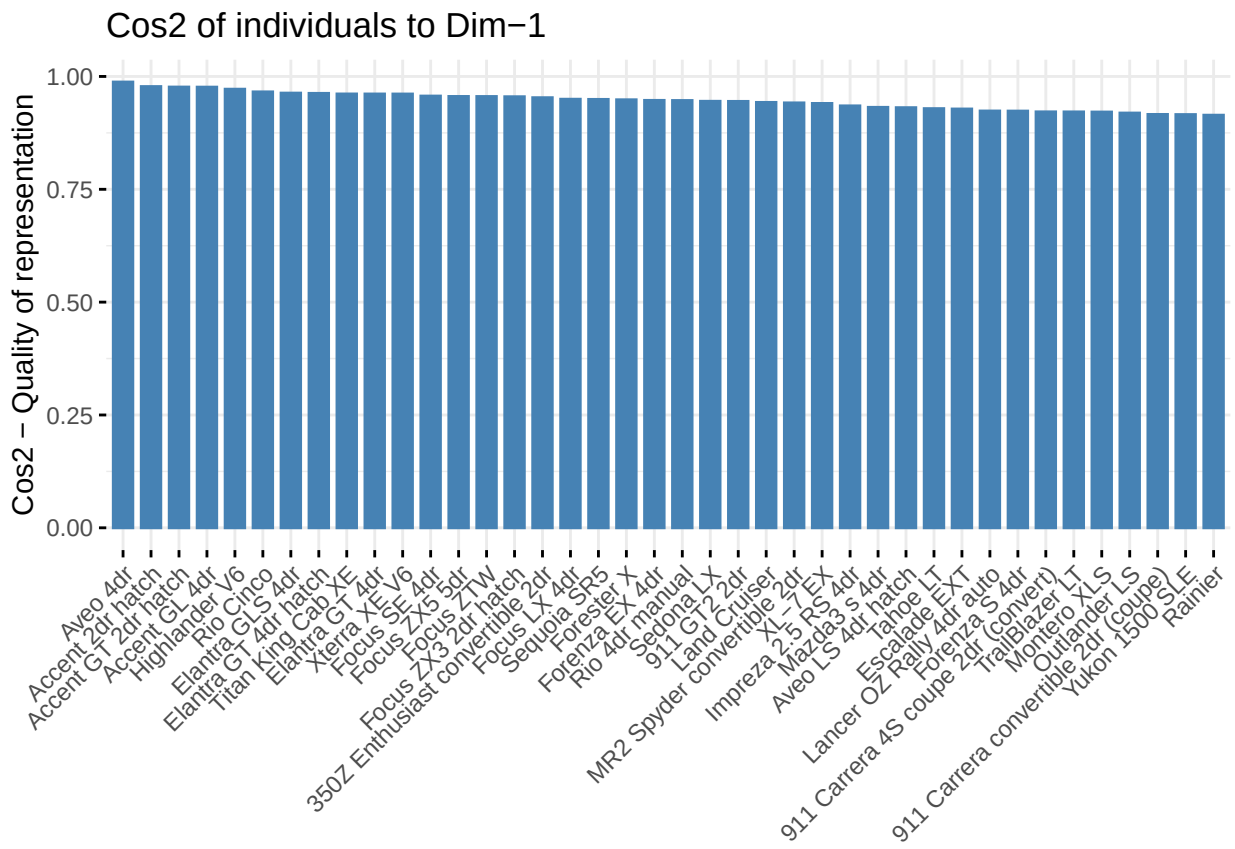



```
ind <- get_pca_ind(res.pca) #résultat pour les individus
print(ind) #coordonnées, cosinus2 et contribution, nous regardons les variables qui ont un cos2 et une
```

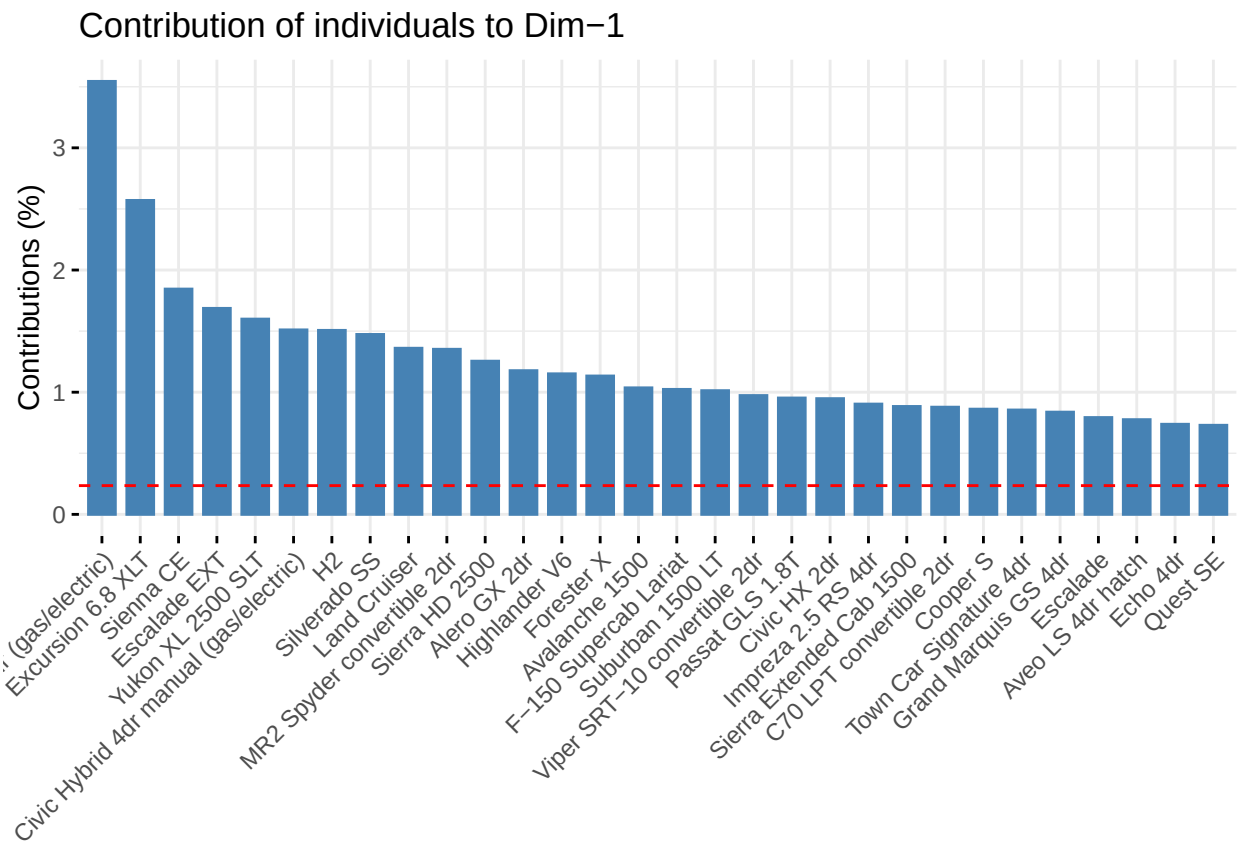
```
## Principal Component Analysis Results for individuals
##  ===================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the individuals"
## 2 "$cos2"    "Cos2 for the individuals"
## 3 "$contrib" "contributions of the individuals"
```

```
#interpretation sur l'axe1
coord<-ind$coord[,1]
contrib<-ind$contrib[,1]
cos2<-ind$cos2[,1]

plot2<- fviz_cos2(res.pca, choice = "ind", top=40)
plot2
```




```
plot3<-fviz_contrib(res.pca, choice = "ind", axes = 1,top=30)
plot3
```




```
#tableau avec pour chaque individus on aura coordonnées, cosinus2 et contribution
display<-cbind(coord,contrib,cos2)
head(display)
```

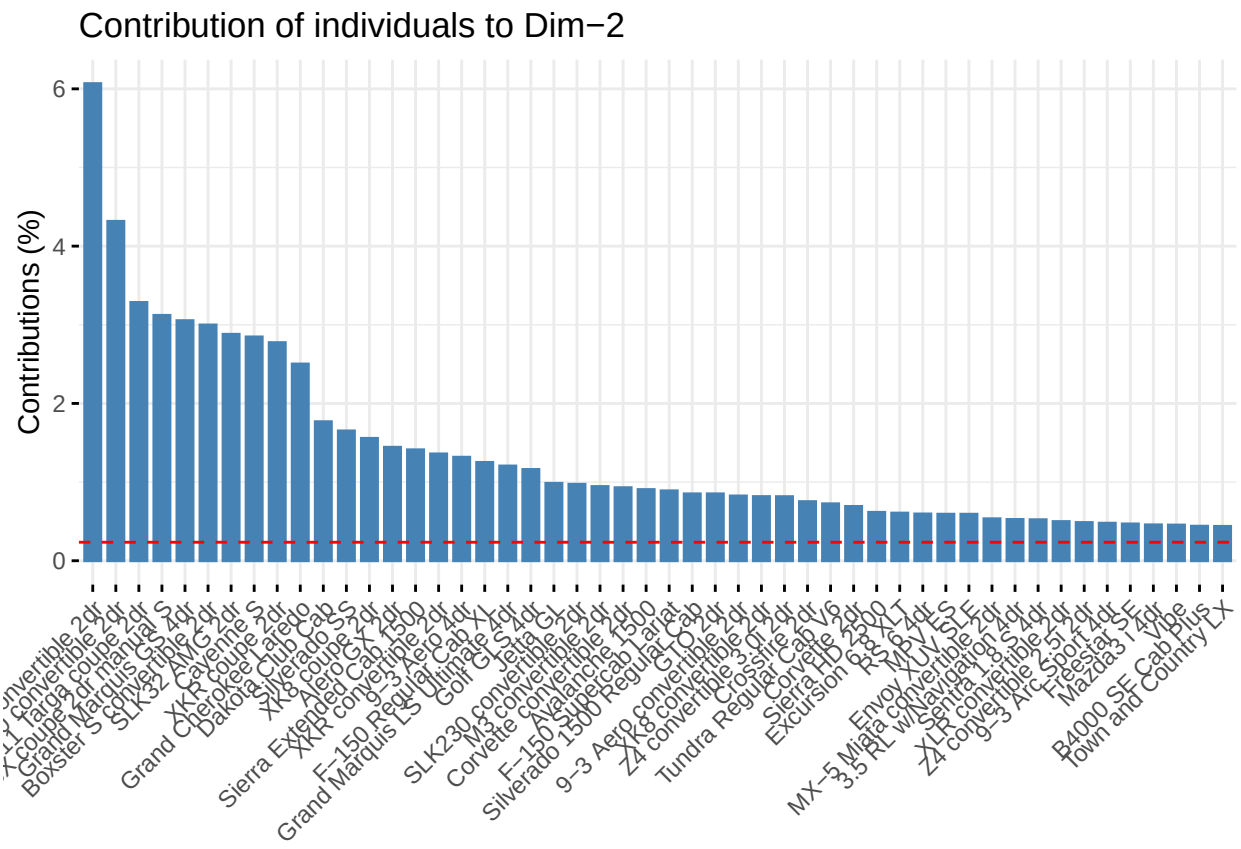
```
##                              coord    contrib       cos2
## 3.5 RL 4dr                1.4982401 0.10008502 0.57439894
## 3.5 RL w/Navigation 4dr   0.6858737 0.02097464 0.09459010
## MDX                       1.4807468 0.09776151 0.59160050
## NSX coupe 2dr manual S    1.1824141 0.06233694 0.05449181
## RSX Type S 2dr           -2.2565959 0.22704599 0.86369079
## TL 4dr                    0.3406968 0.00517538 0.10156042
```

Voitures qui caractérisent le coté positif de l'axe 1 : Excursion 6.8 XLT, Escalade EXT, Yukon XL 2500 SLT. Ces voitures ont les mm caractéristiques, par contre elles sont opposé à celles qui sont dans le coté negatif Voitures qui caractérisent le coté négatif de l'axe 1 : Insight 2dr (gas/electric), Land Cruiser, MR2 Spyder convertible 2dr

```
#interpretation sur l'axe2

coord<-ind$coord[,2]
contrib<-ind$contrib[,2]
cos2<-ind$cos2[,2]

plot4<-fviz_contrib(res.pca, choice = "ind", axes = 2,top=50)
plot4
```




```
#tableau avec pour chaque individus on aura coordonnées, cosinus2 et contribution
display<-cbind(coord,contrib,cos2)
head(display)
```

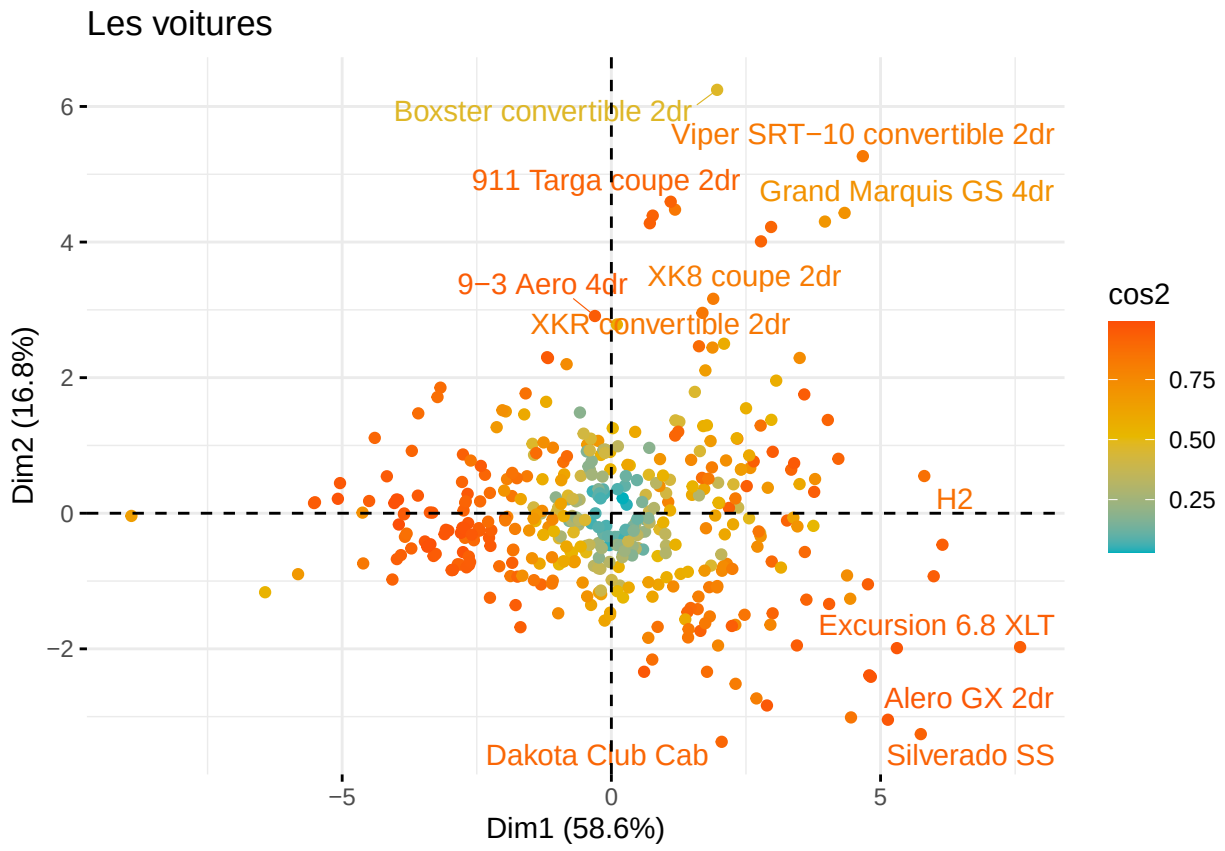
```
##                               coord    contrib       cos2
## 3.5 RL 4dr                0.2712774 0.01144671 0.01883120
## 3.5 RL w/Navigation 4dr -1.8375309 0.52519802 0.67893258
## MDX                       0.7870894 0.09636113 0.16715334
## NSX coupe 2dr manual S    4.4786144 3.11990310 0.78177074
## RSX Type S 2dr            0.2996511 0.01396644 0.01522939
## TL 4dr                    0.7065436 0.07764826 0.43678373
```

Voitures qui caractérisent le coté positif de l'axe 2 : 911 Targa coupe 2dr, NSX coupe 2dr manual S, Boxster convertible 2dr, Viper SRT-10 convertible 2dr Ces voitures ont les mm caractéristiques, par contre elles sont opposé à celles qui sont dans le coté negatif Voitures qui caractérisent le coté négatif de l'axe 2 : Dakota Club Cab, (Tundra Regular Cab V6)

Représentation graphique sur les deux premiers axes : en fonction du cosinus2

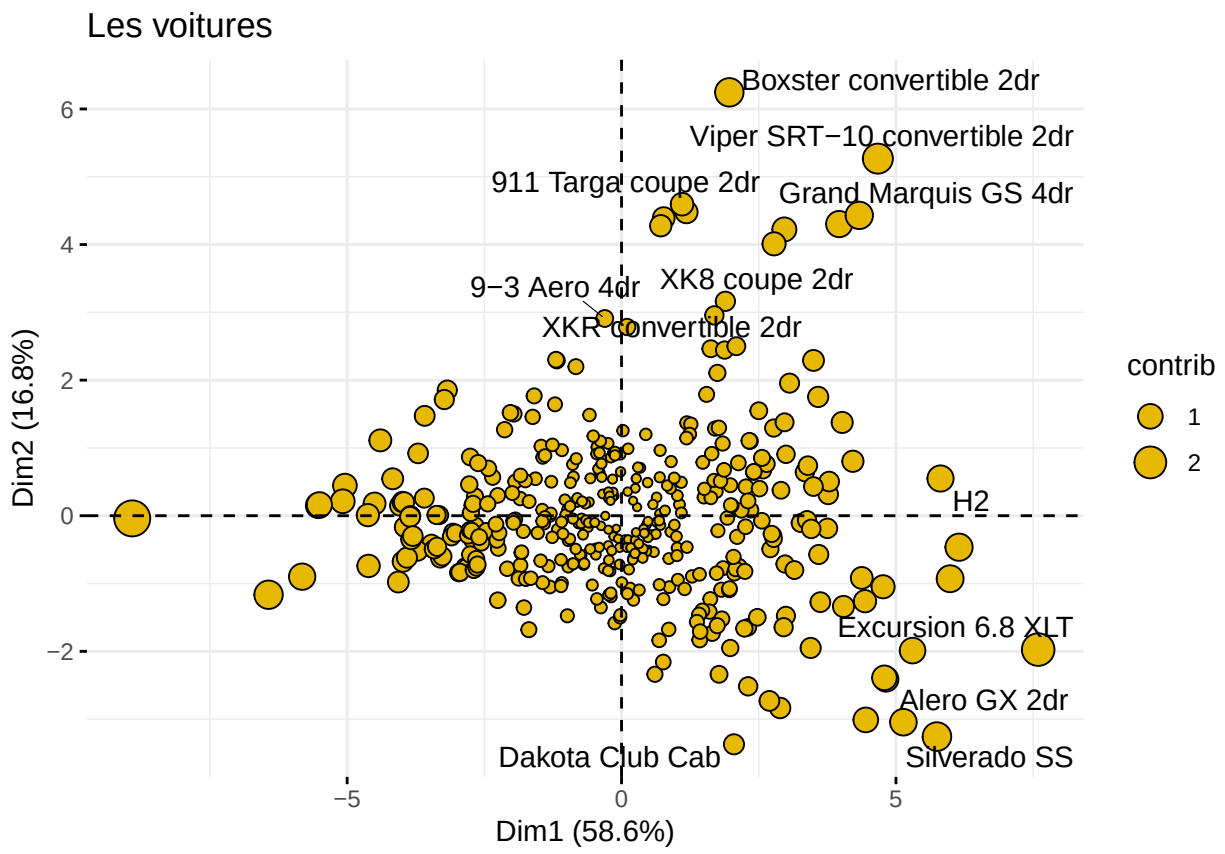
```
plot5<-fviz_pca_ind(res.pca, col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             title="Les voitures",
             repel = TRUE)
plot5
```

```
## Warning: ggrepel: 413 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
plot6<-fviz_pca_ind (res.pca, pointsize = "contrib",
             pointshape = 21, fill = "#E7B800",title="Les voitures",
             repel = TRUE)
plot6
```

```
## Warning: ggrepel: 413 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Boxter convertible 2dr, Viper SRT A0 convertible 2dr, XK8 coupe 2dr voitures

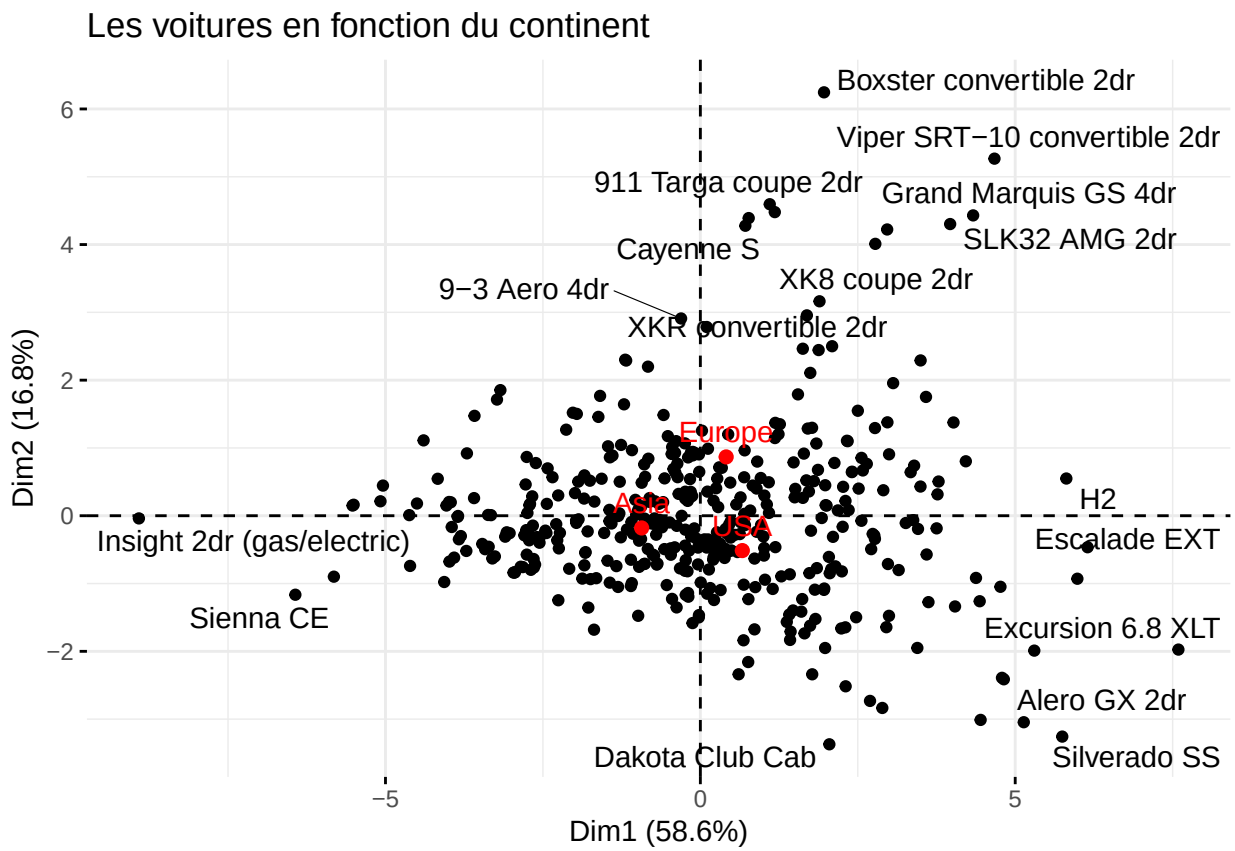
Ajout des varaibles supplémentaires

Continents

```
v.continent<-subset(data, select=c(MSRP, Invoice, EngineSize, Horsepower, MPG_City, MPG_Highway, Weight
res.pca <- PCA(v.continent, quali.sup = 10, graph=FALSE) #acp avec la variable supp continent


plot7 <- fviz_pca_ind(res.pca, repel = TRUE,title="Les voitures en fonction du continent")#Visualisatio
plot7 <- fviz_add(plot7, res.pca$quali.sup$coord, color = "red") #ajout de la variable aux graphiques d
plot7
```

```
## Warning: ggrepel: 408 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



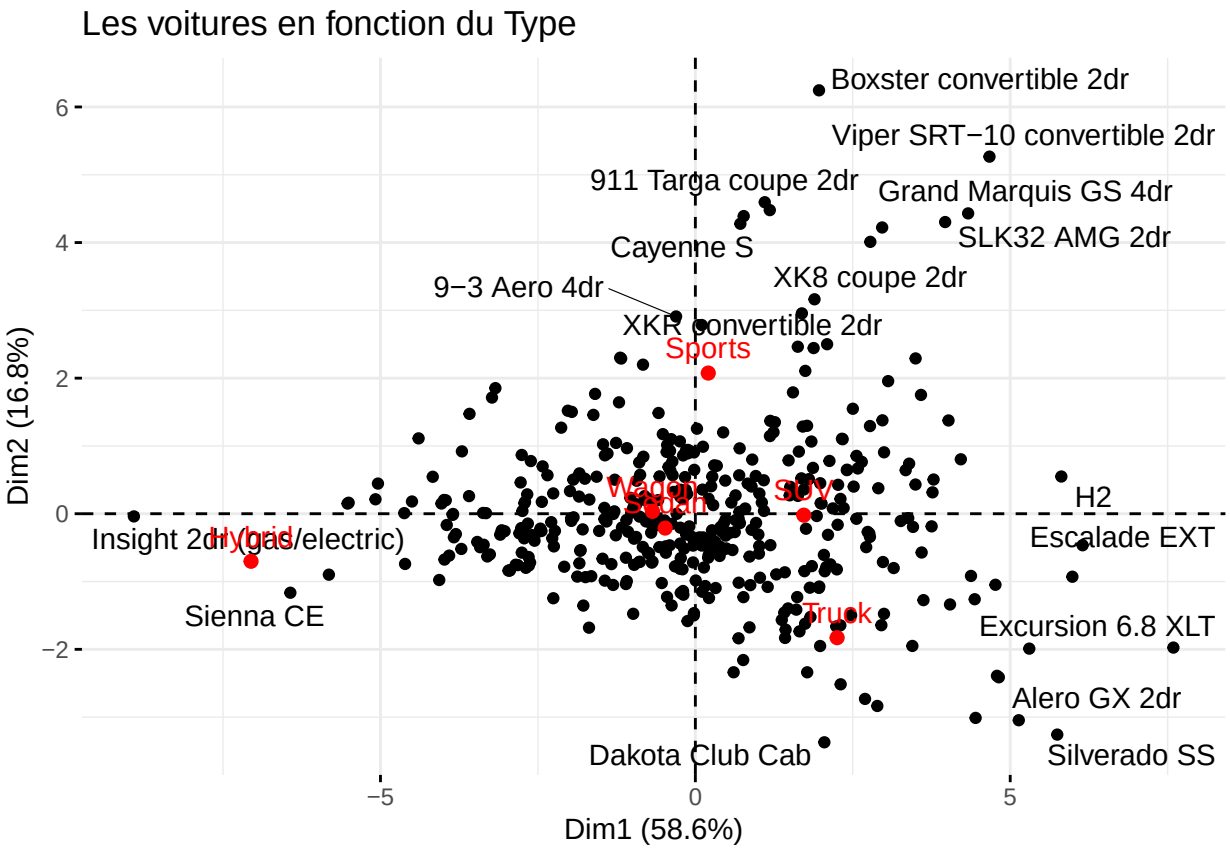
En ajoutant les continents on ne peut pas vraiment dire que cela change bcp de chose, en effet les 3 continents sont au milieu de l'axe, de ce fait nous pouvons conclure que les caractéristiques des véhicules n'ont pas d'influence sur le continent ou elles sont fabriqués, les 3 points ne se démarquent pas, on a pas de critères plus élevé d'un continent à un autre.

Type

```
v.type<-subset(data, select=c(MSRP, Invoice, EngineSize, Horsepower, MPG_City, MPG_Highway, Weight, Whe
res.pca <- PCA(v.type, quali.sup = 10, graph=FALSE) #acp avec la variable supp type

plot8 <- fviz_pca_ind(res.pca, repel = TRUE,title="Les voitures en fonction du Type")
plot8 <- fviz_add(plot8, res.pca$quali.sup$coord, color = "red")
plot8
```

```
## Warning: ggrepel: 408 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



C'est plus interessant quand on ajoute la variable supplementaire Type. il y a du mouvement, par exemble on voit que Hybrid rpz le coté négatif de l'axe 2. HYbrid et SUV sont projeté à l'opposer l'un de l'autre, donc leurs caractéristiques sont tres differentes. On voit aussi que plusieurs type de vehicule rpz des caracteristiques moyennes, pas de demarcation particuliere, elles ressemblent à la moyenne des voitures : Wagon, Sedan. Comme pour hybrid, Truck et Sport se demarquent également.

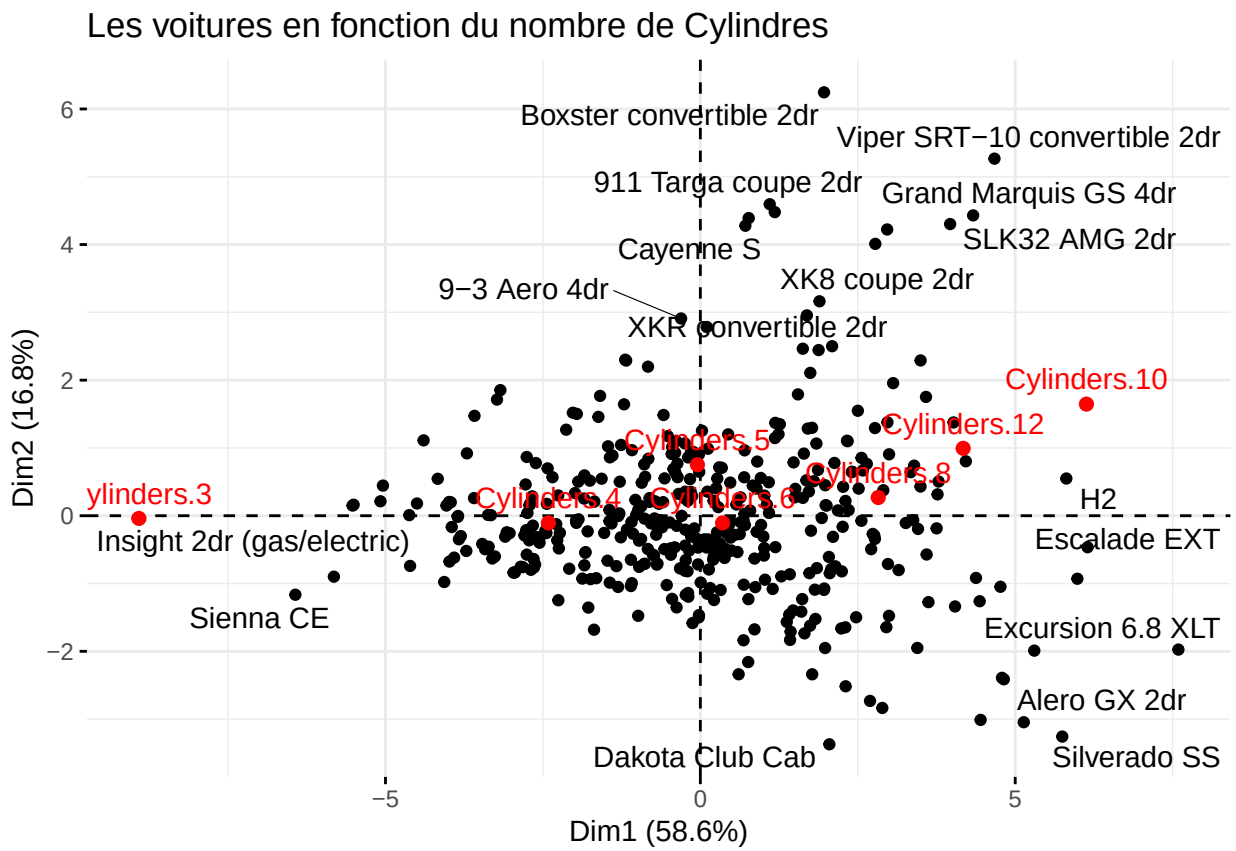
Cylindres

```
v.Cylinders<-subset(data, select=c(MSRP, Invoice, EngineSize, Horsepower, MPG_City, MPG_Highway, Weight
res.pca <- PCA(v.Cylinders, quali.sup = 10, graph=FALSE) #acp avec le variable supp Cylinders
```

```
## Warning in PCA(v.Cylinders, quali.sup = 10, graph = FALSE): Missing values are
## imputed by the mean of the variable: you should use the imputePCA function of
## the missMDA package
```

```
plot9 <- fviz_pca_ind(res.pca, repel = TRUE,title="Les voitures en fonction du nombre de Cylindres")
plot9 <- fviz_add(plot9, res.pca$quali.sup$coord, color = "red")
plot9
```

```
## Warning: ggrepel: 408 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



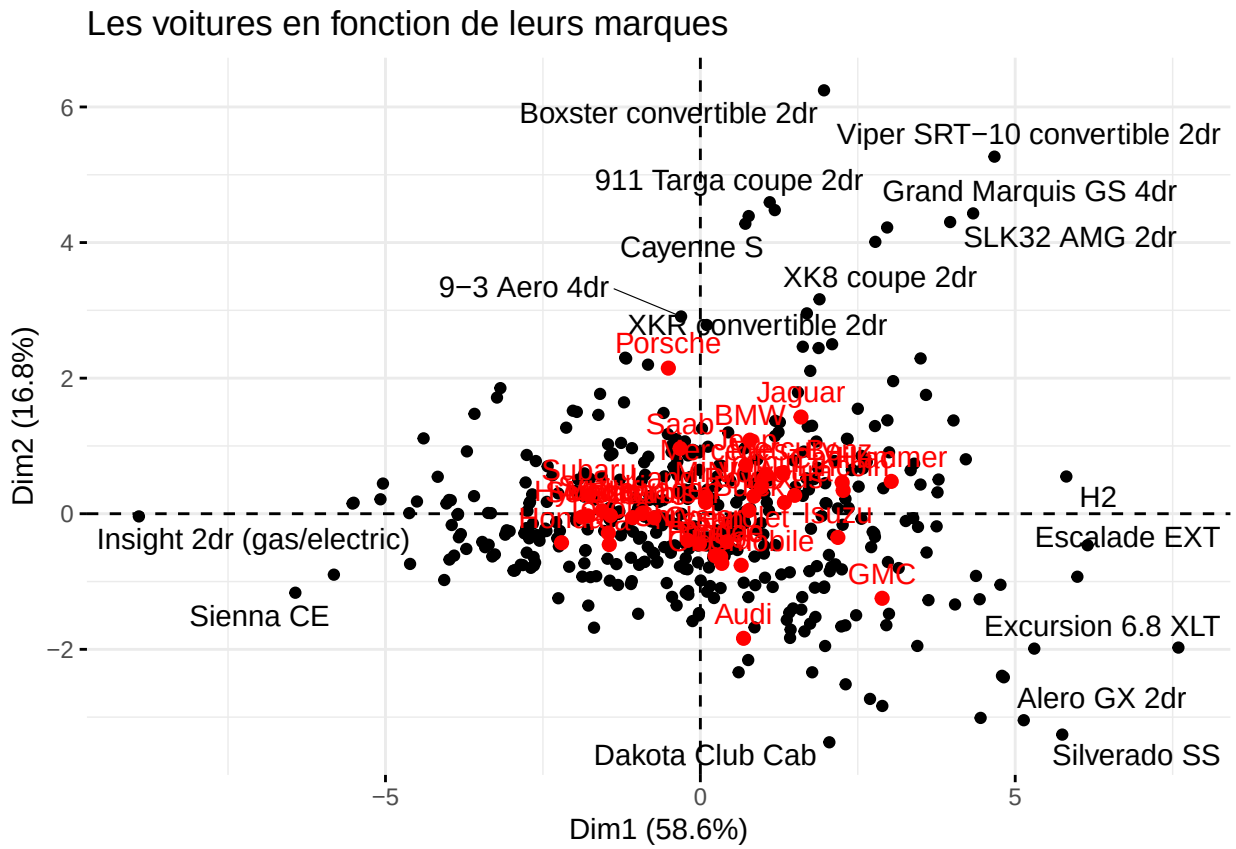
En moyenne les voitures sont équipé de 4 à 8 cylindres (les voitures a 4, 5, 6, 8 cylindres), celles si ont des caractéristiques qui ne les démarquent pas des autres vehicules, c'est dans la moyenne des voitures, ce qui est intéressant c'est que les voitures à 3 cylindre et 10 et 12 cylindre se demarquent et ces dernieres sont opposées les unes des autres (cylindre 3 carac coté neg axe 2 et cylindre 10,12 coté pos) il doit y avoir une explication, on peut s'interesser au carac de la H2 pour voir Les voitures a 3 cylindres et 10 cylindres ont des caracterisques tres tres opposées, on peut penser qu'il y a un bon et mauvais effet à avoir bcp et peu de cylindre, (rechercher une interpretation)

Modèle

```
v.Lib_Make<-subset(data, select=c(MSRP, Invoice, EngineSize, Horsepower, MPG_City, MPG_Highway, Weight,
res.pca <- PCA(v.Lib_Make, quali.sup = 10, graph=FALSE) #acp avec la variables supp Lib_Make

plot10 <- fviz_pca_ind(res.pca, repel = TRUE,title="Les voitures en fonction de leurs marques")
plot10 <- fviz_add(plot10, res.pca$quali.sup$coord, color = "red")
plot10
```

```
## Warning: ggrepel: 408 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Il y a une avalanche de voiture au centre de l'axe, donc au niveau des caractérisques les modeles en generale ne se distingue pas trop. ca veut dire que les modeles proposent des voitures avec des caracteristiques diverse et varié (pas qu'un seul mm type, plusieurs gamme) on le voit bien car pratiquement toutes les marques sont au centre de l'axe, quelques une se demarquent comme porshe, jagua audi et gmc, ces marquent proposent des modèles de voitures qui ont les mm caracteristiques, peut etre le prix, ou le poids ou encore la taille...' point moyen donc les caractéristiques ne sont pas haute ni basse

Analyse des resultat sur les variables

Interpretation sur axe 1

```
#on reprend les variables actives
v.active <- subset(data, select=c(MSRP, Invoice, EngineSize, Horsepower, MPG_City, MPG_Highway, Weight,
res.pca <- PCA(v.active, scale.unit = TRUE, ncp = 5, graph = FALSE)


var <-get_pca_var(res.pca)
coord<-var$coord[,1]
contrib<-var$contrib[,1]
cos2<-var$cos2[,1]
display<-cbind(coord,contrib,cos2)
head(display)
```

```
##                 coord   contrib      cos2
## MSRP         0.3549837  2.387877 0.1260134
## Invoice      0.4987462  4.713618 0.2487477
## EngineSize   0.9117878 15.753706 0.8313570
```

```
## Horsepower   0.8113552 12.474332 0.6582973
## MPG_City    -0.8680735 14.279342 0.7535516
## MPG_Highway -0.8656744 14.200523 0.7493922
```

acp normé : coordonné = coef de correlation entre variable et axe, juste en regardant la valeur de la coordonée on peut avoir une idée sur les variables qui sont les plus corrélés avec l'axe variable contribué a l'axe et seront bien représenté par l'axe colonne coordonnée suffit pour faire l'interprétation ! dans le cadre de ACP NORMEE on voit bien dans display qu'il y a une concordance entre les contrib et les cos2

Les critères EngineSize et Weight vont caractériser le coté positif de l'axe 1 autrement dit, le coté positif de l'axe 1 sera projeté par des voitures qui ont de grosses valeurs en EngineSize et Weight. Inversement le coté neg de l'axe 1 sera projeté des voitures qui ont de faibles valeurs en EngineSize et Weigh

Les critères MPG_City et MPG_Highway vont caractériser le coté négatif de l'axe 1 le coté neg de l'axe 1 sera projeté par des voitures qui ont de grosses valeurs en MPG_City et MPG_Highway donc dans l'axe 1 les voitures qui ont un bon score en EngineSize et Weight ont un faible score en MPG_City et MPG_Highway

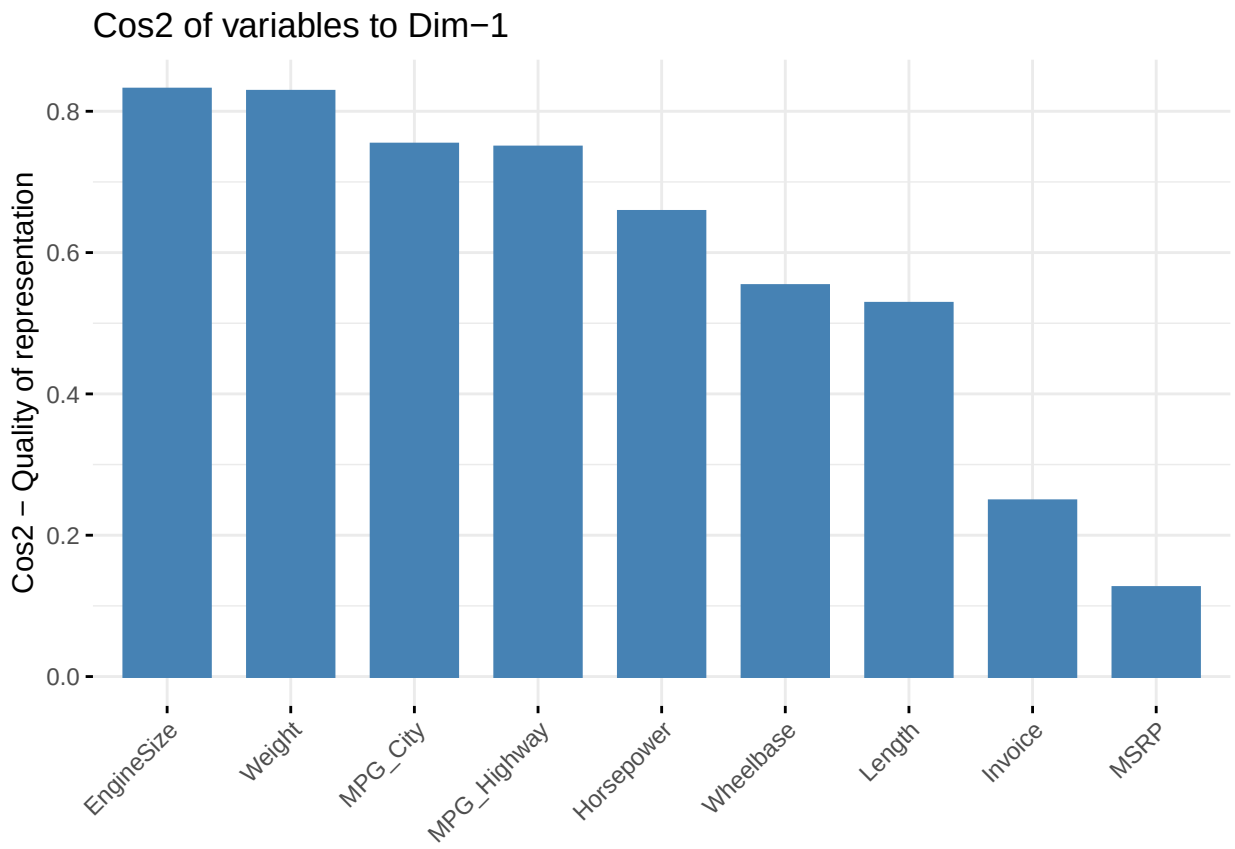
Interpretation sur axe 2

```
var <-get_pca_var(res.pca)
coord<-var$coord[,2]
contrib<-var$contrib[,2]
cos2<-var$cos2[,2]
display<-cbind(coord,contrib,cos2)
head(display)
```

```
##                  coord     contrib        cos2
## MSRP         0.57335353 21.73139781 0.328734268
## Invoice      0.60712648 24.36694217 0.368602562
## EngineSize   0.03717895  0.09137699 0.001382274
## Horsepower   0.40371090 10.77416552 0.162982494
## MPG_City    -0.08117907  0.43564306 0.006590041
## MPG_Highway -0.04217350  0.11757686 0.001778604
```

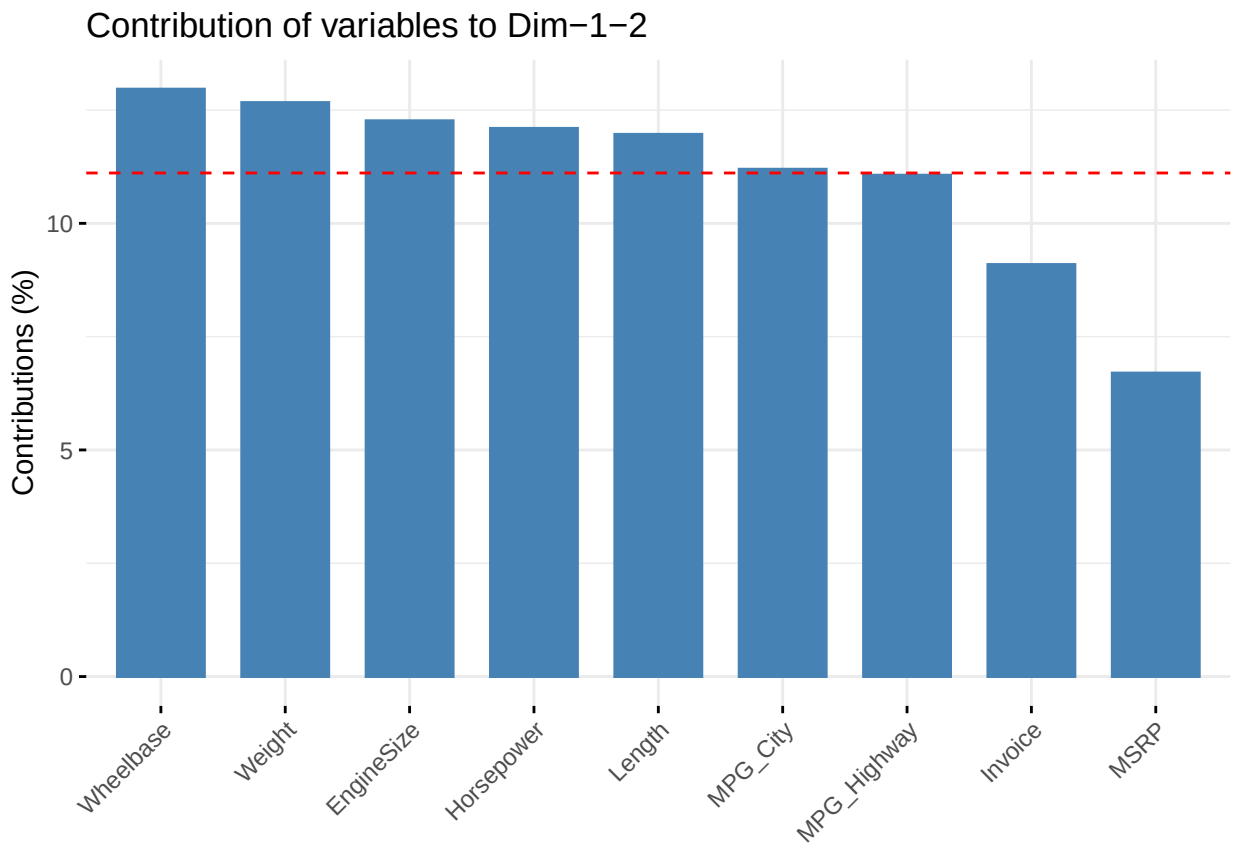
Coté positif de l'axe 2 est caractérisé par Invoice et MSRP (logique les prix se correspondent) des voitures qui coutes cheres coté neg de l'axe 2 est caractérisé par Wheelbase et Length (logique wheelbase sera proportionnelle à la longueur de la voiture) des voitures assez longues donc les voitures qui coutent cheres ne sont pas forcement longues etant donnée que les variables s'opposent dans les coté neg on va trouver des voitures qui coutent pas trop chere et dans le coté positif on va retrouver des voitures qui ne sont pas tres longues

```
plot11<-fviz_cos2(res.pca, choice = "var") #graph du cos2 des var sur le 1er axe
plot11
```



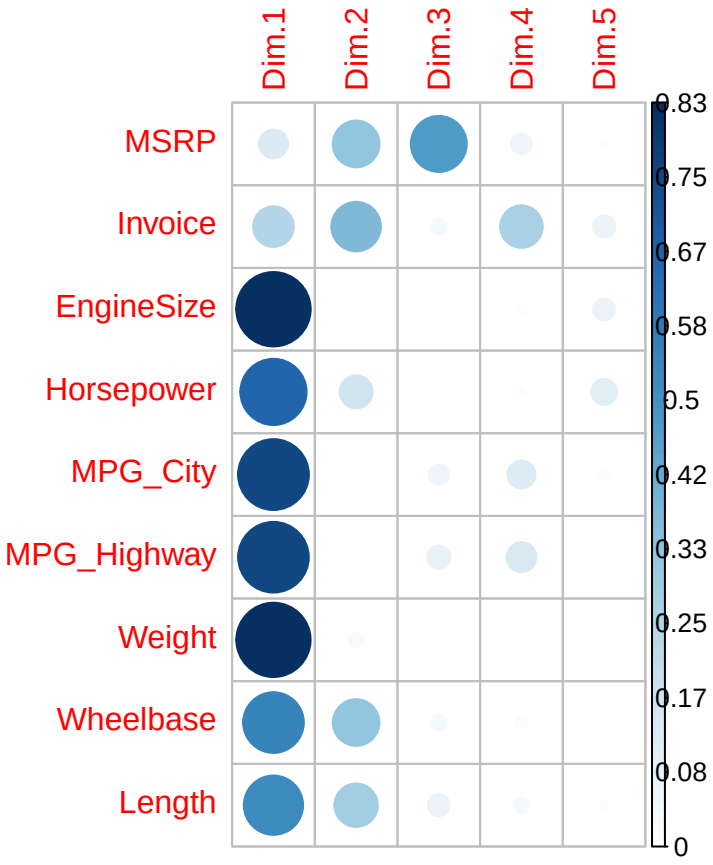

```
#on retrouve EngineSize et Weight qui caracterise bien le coté pos de l'axe 1

plot12<-fviz_contrib(res.pca, choice = "var", axes = 1:2) #graph du cos2 des var cum sur les 2 premiers
plot12
```



beaucoup de variables caractérisent les 2 premiers axes : wheelbase, weight, engizesize, horsepower, lenght

```
corrplot(var$cos2, is.corr=FALSE)
```

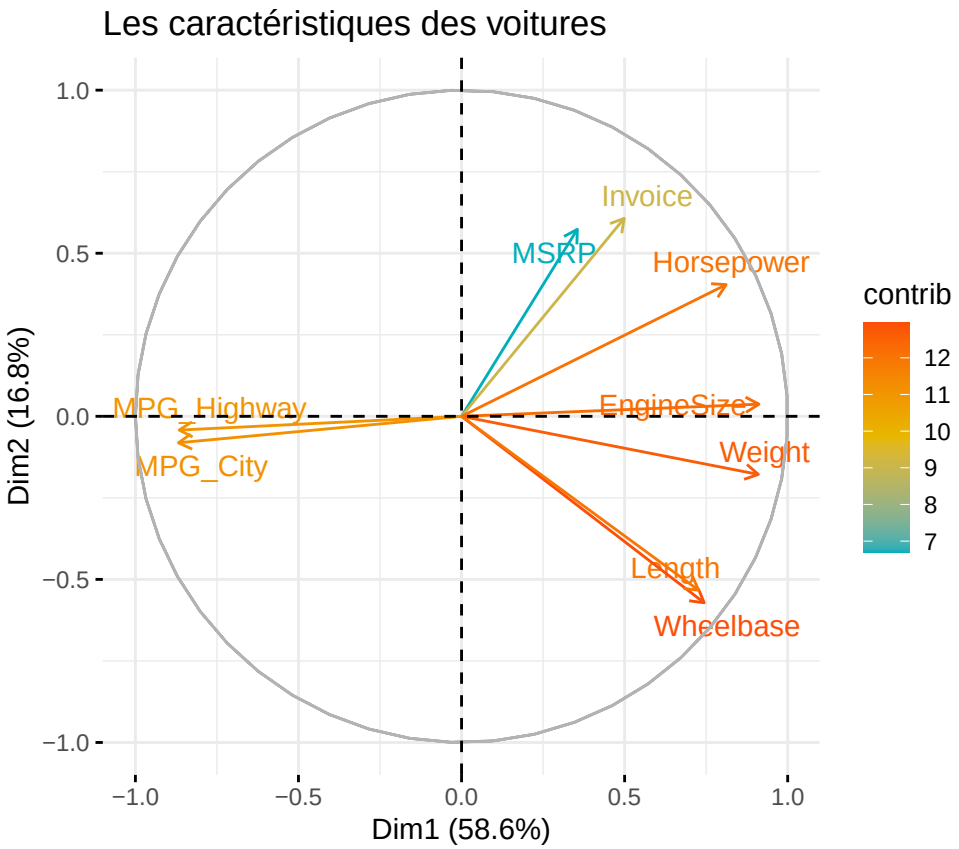


on a pas le signe de la corrélation(caracté du coté pos ou neg ?) on voit bien ce qu'on a constaté tout à lheure, EngineSize et Weight caractérise bcp axe 1

representation graphiques des variables sur le premier plan factoriel

visialisation d'une acp pour les variables

```
plot13<-fviz_pca_var(res.pca, col.var = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
title="Les caractéristiques des voitures",
repel = TRUE)
plot13
```



ce plan factoriel est tres interessant, il vient affirmer tout ce qu'on a precedemment évoqué

Interprétation des distances entre les variables, à savoir que moins il y a de distance plus les critères se ressemblent

Graphique où les points (variables) sont projetés (faiscaux) R les rpz par des droites avec fleches mais ce qui ns interesse cest la pointes cosinus2 de l'angle entre 2 variables = coeff de correlation entre ces deux variables, alors si cos2 faible le coef corélation sera fort donc les variables seront liée linéairement donc si l'une a un score haut l'autre aussi MSRP et wheelbase forment presque un angle droit, donc cos = 90° soit coef de corr proche de zero alors ces deux variables ne sont pas corrélé les points MPG hightway et MPG city sont tres rapproché donc fortement corrélé et par opposition si le cos2 se rapproche de 180° alors correlation neg = bon score = mauvais score

————————————————-*REALISATION DE L'AFC*————————————————————-

Dans cette partie, nous effectuons des croisement de variables catégorielles.Lors d'un croisement de deux variables catégorielles, On cherche à savoir si l'une apporte de l'informations sur l'autre et inversement en appréhendant la relation de differente maniére.

Croisement la variable Type et Continent pour répondre aux questions suivantes:

- Quelle est la distribution des types de voitures selon les continents(profils ligne).
- Est-ce que cette distribution est différente d'un continent à un autre(distance entre les profils).
- Les types de voitures sont-ils différents selon les continents(profils colonne).
- Certains continents ont-ils une préférences pour un certains type de voitures? ou le type de voitures attirent-ils certains continents.

Pour répondre à ces question, nous allons suivre les étapes suivantes:

- On commence par construire un tableau de contingence qui croise le continent contre le type de voitures.
- Proposer une représentation graphique de ce tableau de contingence.
- Enlever les modalités d'effectifs trop faible.
- Verifier la dépendance ou liason des deux variables avec un test de khi2.

La **question statistique** que nous allons nous poser c'est est-ce que le type de voiture dépend du type du continent ? ou est-ce qu'il existe une certaine préférence des type de voitures vis à vis des continents?

L'interet de l'AFC c'est aussi d'identifier les associations entre les modalités ligne et colonnes. Par exemple est ce que les voitures de types Sedan attirent plus les individus du continent Asie.

Statistique descriptive

```
summary(data)
```

```
##   Continent       Type           MSRP            Invoice         EngineSize
##  Asia  :158   Hybrid:  3   Min.   :    22   Min.   :   194   Min.   :1.300
##  Europe:120   Sedan :262   1st Qu.:  2699   1st Qu.: 16291   1st Qu.:2.400
##  USA   :147   Sports: 49   Median : 16385   Median : 23336   Median :3.000
##               SUV   : 59   Mean   : 19194   Mean   : 25970   Mean   :3.202
##               Truck : 24   3rd Qu.: 29995   3rd Qu.: 32902   3rd Qu.:3.900
##               Wagon : 28   Max.   :192465   Max.   :117854   Max.   :8.300
##
##    Cylinders        Horsepower       MPG_City      MPG_Highway
##  Min.   : 3.000   Min.   : 73.0   Min.   :10.00   Min.   :12.00
##  1st Qu.: 4.000   1st Qu.:165.0   1st Qu.:17.00   1st Qu.:24.00
##  Median : 6.000   Median :210.0   Median :19.00   Median :26.00
##  Mean   : 5.813   Mean   :215.9   Mean   :20.07   Mean   :26.85
##  3rd Qu.: 6.000   3rd Qu.:255.0   3rd Qu.:21.00   3rd Qu.:29.00
##  Max.   :12.000   Max.   :500.0   Max.   :60.00   Max.   :66.00
##  NA's   :2
##      Weight       Wheelbase         Length               Lib_Make
##  Min.   :1850   Min.   : 89.0   Min.   :143.0   Toyota       : 29
##  1st Qu.:3105   1st Qu.:103.0   1st Qu.:178.0   Chevrolet    : 28
##  Median :3473   Median :107.0   Median :187.0   Mercedes-Benz: 25
##  Mean   :3577   Mean   :108.2   Mean   :186.4   Ford         : 24
##  3rd Qu.:3977   3rd Qu.:112.0   3rd Qu.:194.0   Honda        : 18
##  Max.   :7190   Max.   :144.0   Max.   :238.0   Nissan       : 18
##                                                 (Other)      :283
```

Transformation de la variable cylindre en factor

```
cols <- c(6)
data[,cols] <- lapply(data[,cols] , factor)
```

```
## Warning in `[<-.data.frame`(`*tmp*`, , cols, value = list(structure(1L, .Label =
## "6", class = "factor"), : provided 425 variables to replace 1 variables
```

```
#cars
```

```
table(data$Type)
```

```
##
## Hybrid  Sedan Sports    SUV  Truck  Wagon
##      3    262     49     59     24     28
```

```
#duplicated(cars$Cod_Model)
```

```
table(data$Continent)
```

```
##
##   Asia Europe    USA
##    158    120    147
```

Construction du tableau de contingence

```
contingence<-table(data$Continent,data$Type)
contingence
```

```
##
##          Hybrid Sedan Sports SUV Truck Wagon
##   Asia        3    94     17  25     8    11
##   Europe      0    78     23   9     0    10
##   USA         0    90      9  25    16     7
```

Dans ce tableau de contingence, on a des effectifs trop faible. Ce qui peut poser probléme lors du chi2 qui se veut des effectifs soient supérieur > 5. Enlévons la modalité **Hybrid** dont **l'effectif est faible** du tableau de contingence.

```
contingence<-table(data$Continent,data$Type, exclude = "Hybrid")
contingence
```

```
##
##          Sedan Sports SUV Truck Wagon
##   Asia      94     17  25     8    11
##   Europe    78     23   9     0    10
##   USA       90      9  25    16     7
```

Ainsi on va se retrouver, avec 03 profils ligne et 05 profils colonnes par la suite.

- **Calcul des profil lignes**(ou distribution conditionnelle par rapport à la modalité d'une autre variable)

Cela se fait avec la fonction **lprop(contingence,digits=1)** du package **questionr**

```
##
##            Sedan Sports SUV   Truck Wagon Total
##   Asia      60.6  11.0   16.1   5.2   7.1 100.0
##   Europe    65.0  19.2    7.5   0.0   8.3 100.0
##   USA       61.2   6.1   17.0  10.9   4.8 100.0
##   Ensemble  62.1  11.6   14.0   5.7   6.6 100.0
```

Analyse des profils ligne ou proportion en ligne du tableau de contingence.

- Le premier profil ligne représente **La distribution du type de voiture selon le continent Asie**.Pour le continent Asie, 60.6% des voitures sont de type Sedan, 11.0% de type sport, 16.1% de type SUV, 5.2% de type Truck et 7.1% de type Wagon. On peut suivre le meme raisonnement pour le continent Europe et USA.
- Ensemble (en ligne) représente **le profil moyen des ligne** c'est à dire la moyenne pondérée des profils lignes. C'est aussi la distribution du type de voitures sans distinction du continent.
- Pour les profils colonnes, on va avoir la distribution des modéle de voiture des continents pour chaque type.

-Ensemble (en colonne): Distribution des modéles de chaque continent qu'elle que soit le type. C'est le **profil moyen des colonnes**.

Les deux variables sont elles indépendantes

```
##
##  Pearson's Chi-squared test
##
## data:  contingence
## X-squared = 30.312, df = 8, p-value = 0.0001861
```

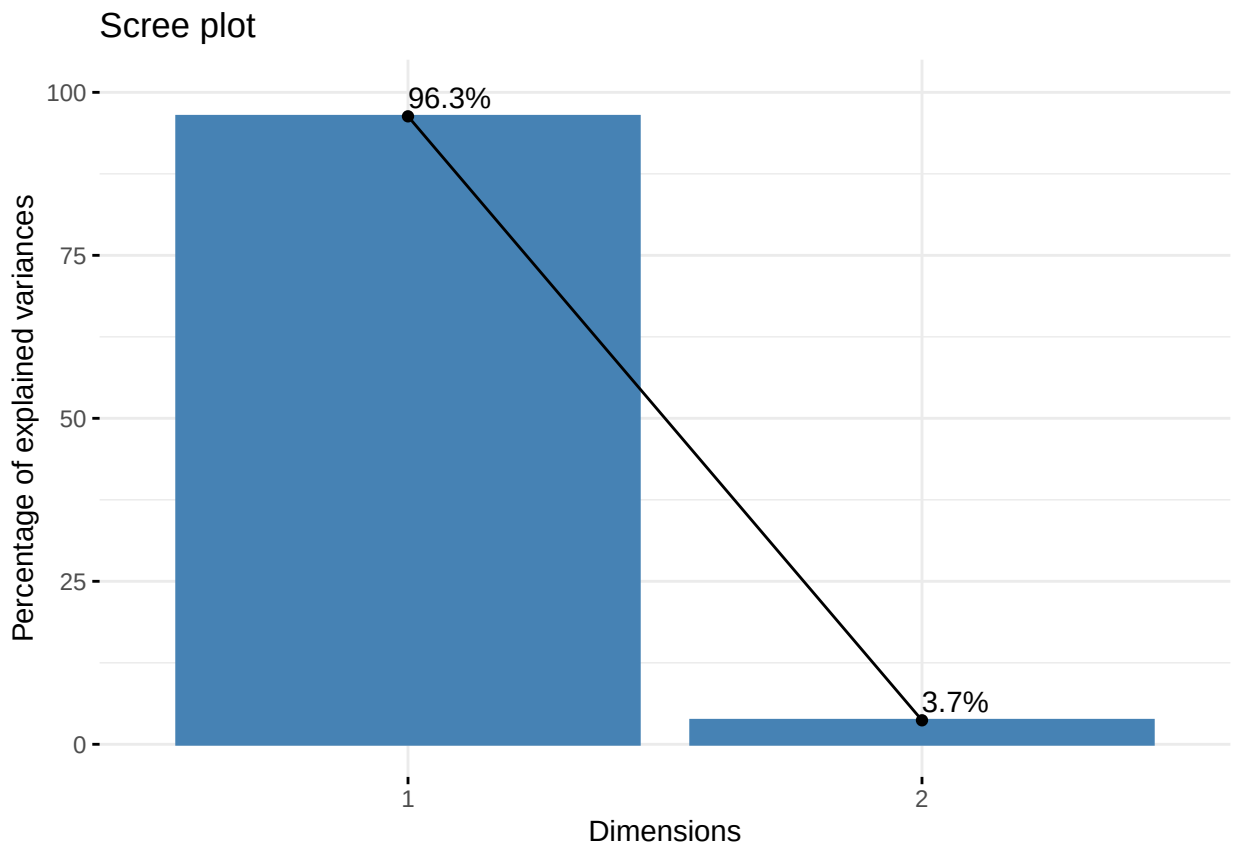
La p-value etant trés petite (<0.05), donc il existe une liaison entre le continent et et le type de voitures. Autrement dit Les individus n'ont pas la meme appétence vis a vis des types de voitures d'un continent à l'autre.

Realisation d'un AFC pour expliquer le lien entre les deux variables

La fonction CA prend en entrée le tableau de contingene et non le tableau individus variables.

Choix du nombre d'axes

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1 0.069185066        96.318619                    96.31862
## Dim.2 0.002644313         3.681381                   100.00000
```



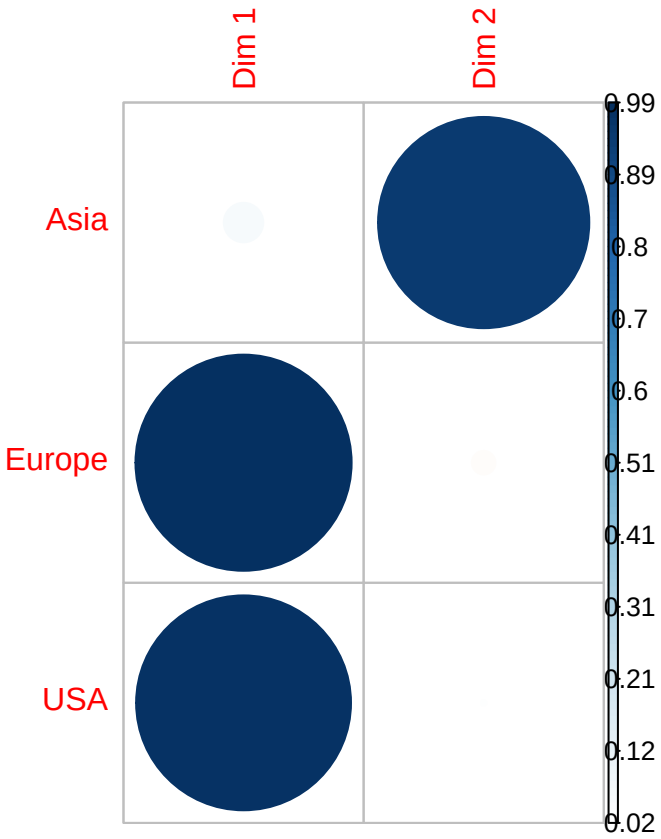
96,3% de l'inertie du nuage est représenté par l'axe 1, ce qui est beaucoup. On va donc interpreter seulement le premier axe. Donc à elle seul on va pouvoir expliqué les écarts à la situation de non indépendance des deux variables. Donnons ainsi une interprétation sémantique à l'axe 1.

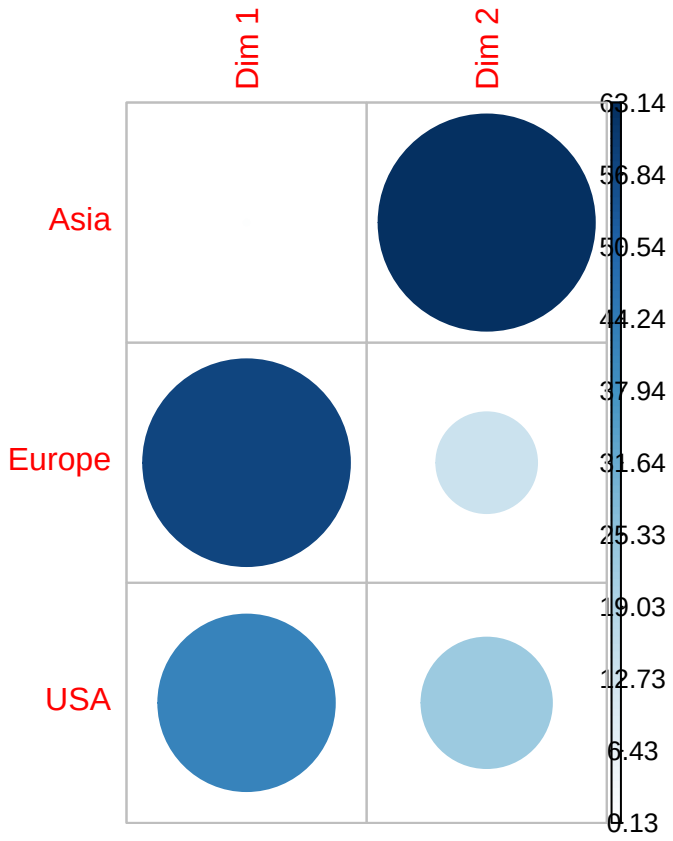
Interpretation sémantique de l'axe 1 avec les profils lignes

```
##              coord    contrib       cos2
## Asia    0.01591084  0.1343981 0.05275685
## Europe -0.37484242 57.7502548 0.99094048
## USA     0.28921707 42.1153470 0.97950963
```

On va regarder les profils qui ont une forte contribution et cos2 élevé sur l'axe 1.

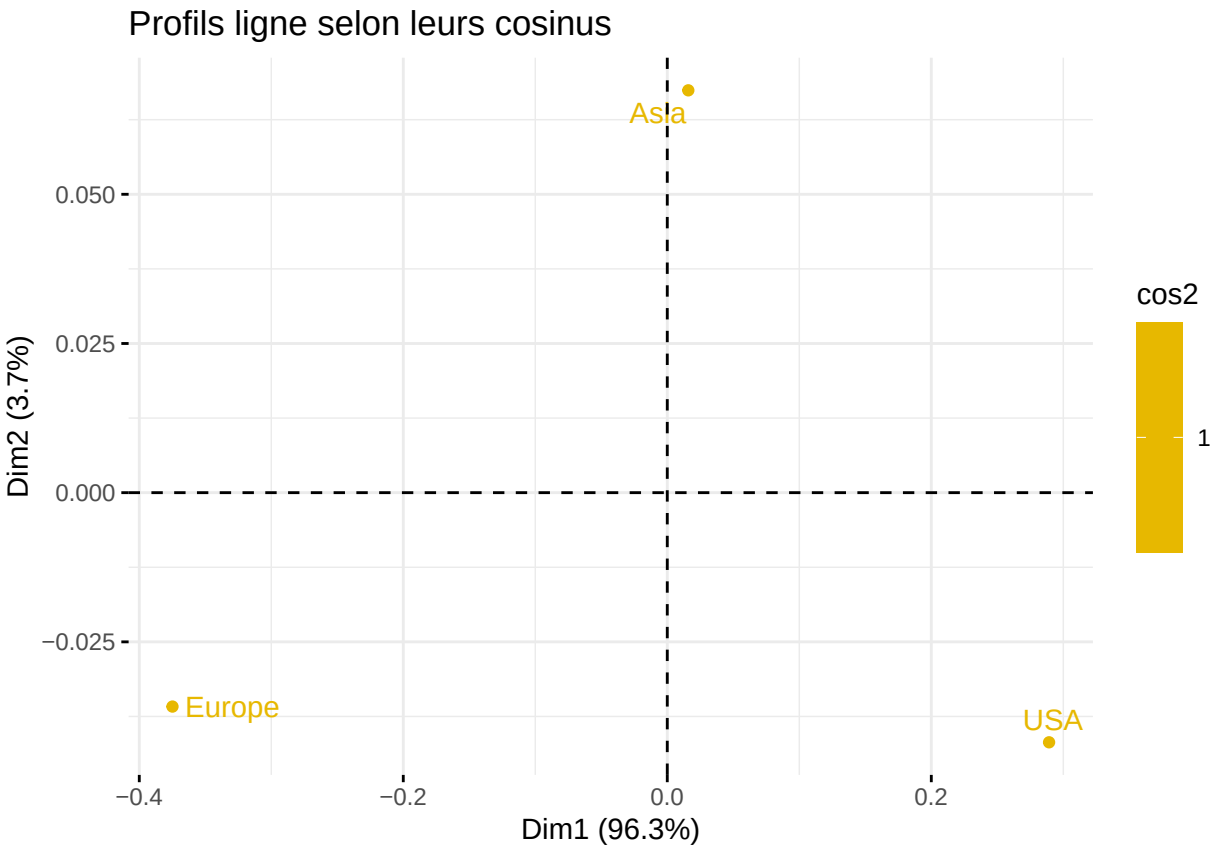
le profil des modéle "USA" caractérise le coté positif de l'axe 1. et le profil des modéles "Europeens" caractérise le coté négatif de l'axe 1. La distribution du type de voiture des **USA** n'est pas la meme que celle des **Européens**. Voilà ce que l'axe 1 permet mettre en évidence.

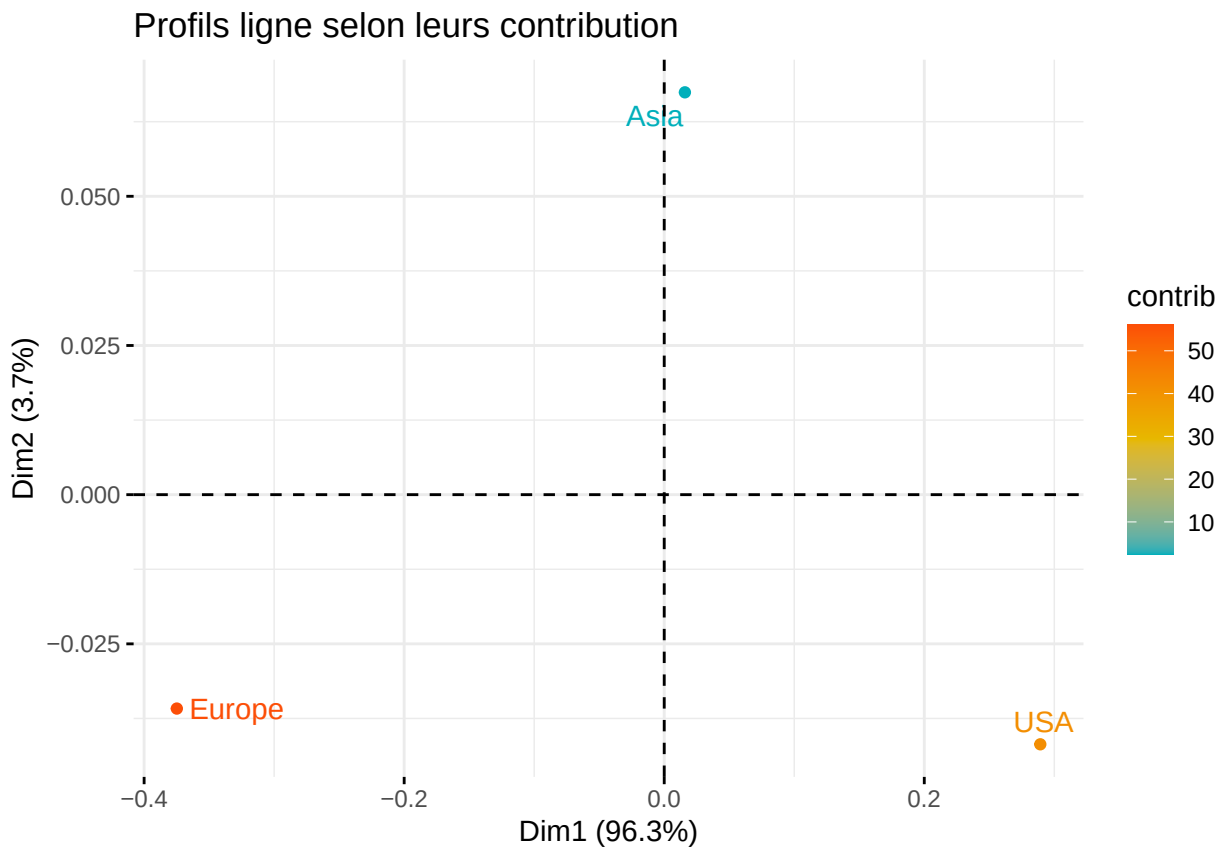

Dim 1
Dim 2
Asia
Europe
USA
0.99
0.89
0.8
0.7
0.6
0.51
0.41
0.31
0.21
0.12
0.02



- On peut raisonner sur ces petit tableau des cos2 et des contribution ,mais ca ne suffit pas parce que
- Ici on ne va pas connaitre si le profils est projetté du coté positif ou négatif de l'axe. Seul le profil "Europe" et "USA" est représenté sur le premier axe qu'il soit en terme de contribution ou cos2. Ce qui consolide une fois de plus notre choix.

Construction d'une représentation graphique des profils lignes sur les deux premiers axes


Profils ligne selon leurs cosinus
Asia
Europe
USA
0.050
0.025
0.000
−0.025
−0.4
−0.2
0.0
0.2
Dim2 (3.7%)
Dim1 (96.3%)
cos2
1



On voit bien sur l'axe 1 le profils des modéle européens s'oppose au profils des modéles USA

- Qu'est ce qu'on peut dire des profils des modéles "Asiatique"

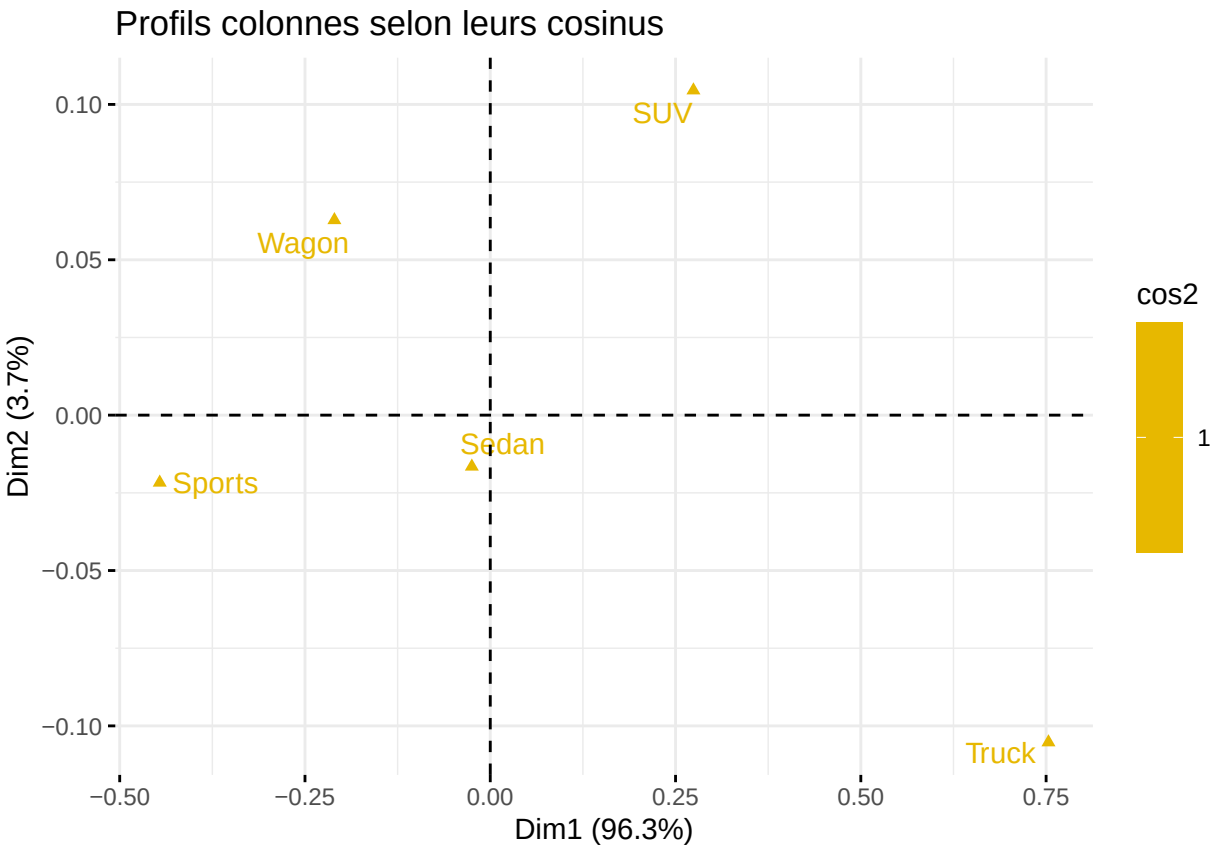
Le profils des modéle Asiatique sur l'axe 1 (est projetté prés du centre de gravité ou de l'origine des axes qui représente le profil moyen). Donc ca veut dire que les Asiatique ont une distribution du type de voiture qui ressemble à la distribution du type dans tous le jeux de données

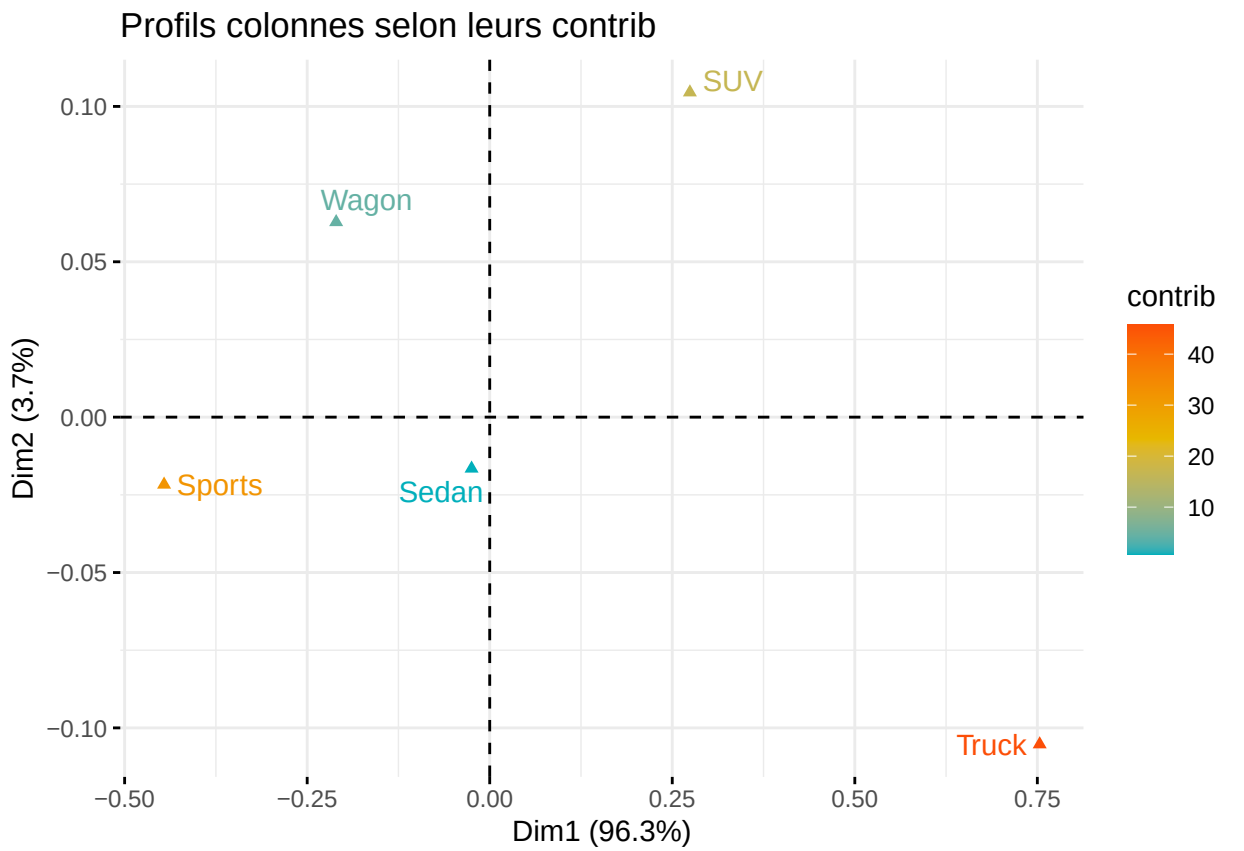
Interpretation sémantique de l'axe 1 avec les profils colonnes

```
##              coord    contrib      cos2
## Sedan  -0.02485058  0.5541784 0.6928499
## Sports -0.44597426 33.3803477 0.9976374
## SUV     0.27415880 15.1890841 0.8730715
## Truck   0.75320150 46.6346563 0.9808357
## Wagon  -0.21030766  4.2417336 0.9181845
```

Le profil des voitures de type Sports est projetés du coté négatifs par oposition au profil des modéles de type Truck projettés du coté positif de l'axe 1. Cela veut dire que la distribution de des modéles de voitures par continent n'est pas la meme pour le type sport que pour le type Truck. La distribution des modéles de voiture par continent n'est pas la méme selon le type.

Construction d'une représentation graphique des profils colonnes sur les deux premiers axes


Profils colonnes selon leurs cosinus
0.10
0.05
0.00
−0.05
−0.10
Dim2 (3.7%)
SUV
Wagon
Sedan
Sports
Truck
−0.50
−0.25
0.00
0.25
0.50
0.75
Dim1 (96.3%)
cos2
1



Au regard du graphique, nous voyons bien les oppositions que nous venons de dire entre les profils de type sport et celui de type Truck. Ce qui signifie que le type Sedan à une distribution du selon les continent qui ressemble à la distribution du continent dans tous les jeux de données.

————————————————————*-REALISATION DE LA CLASSIFICATION*————————————————————-

Dans cette partie nous allons utiliser factoMiner pour effectuer une CAH Dans factoMiner, la CAH est réalisés sur des objets résultants d'une analyse factorielle L'idée est de realiser à priori une PCA et d'utiliser par la suite des objets résultants de l'analyse factorielle pour realiser la CAH Car nous somme en présence de variables quantitatives, donc on réalise l'ACP avant CAH

Pour cettte partie nous ne prendrerons pas en compte la variable lib-Make car elle ne sera pas pertinente pour notre étude

deplacement des colonnes Cylinders et Lib_Make pour mettre les variables qualitatives d'un côté et les quanti de l'autre

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

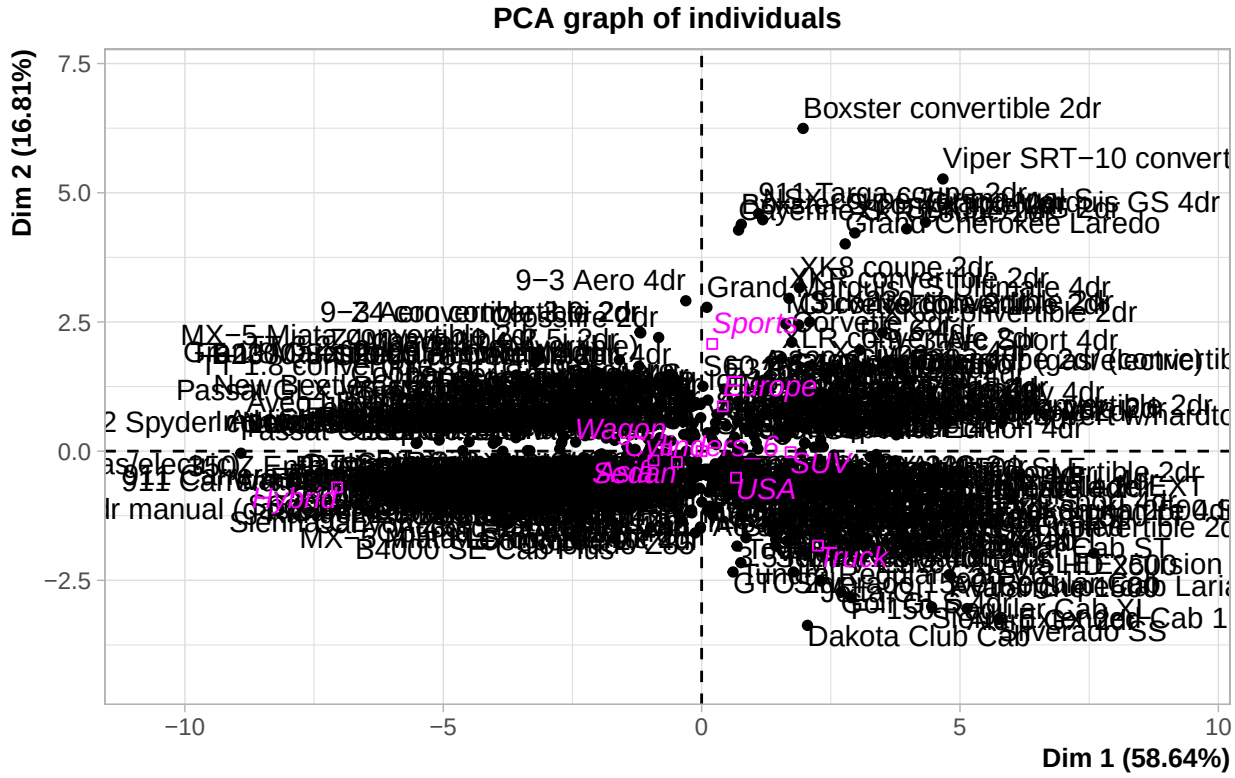
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

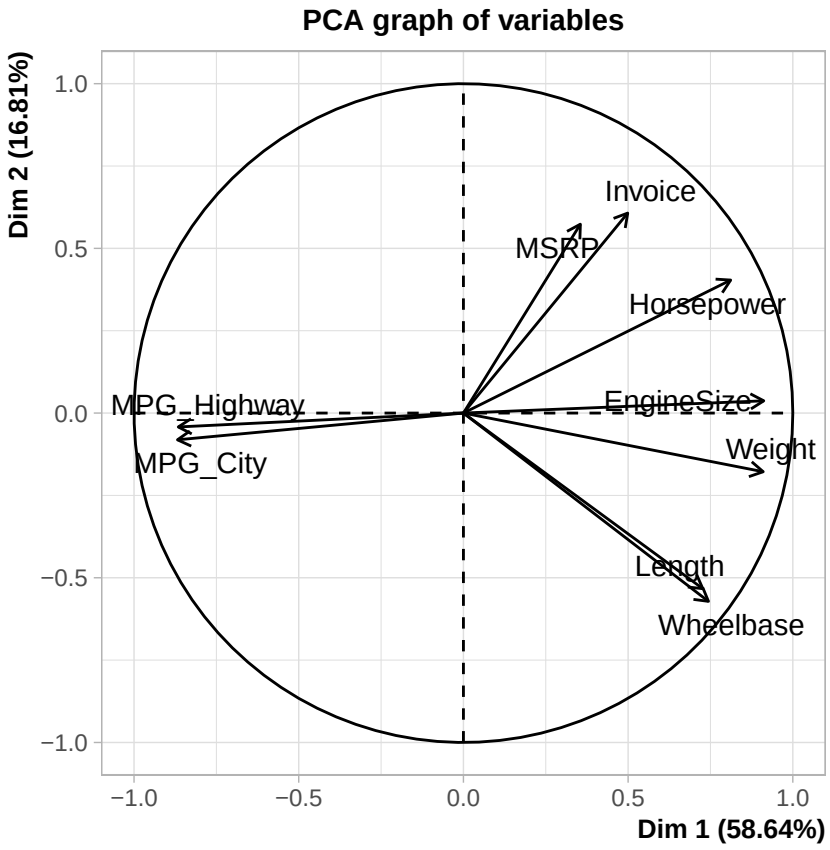
```
data<- data%>% relocate(Cylinders, .after=Type)
data <-data[1:12]
```

Nous sommes donc en présence de 9 variables quantitatives(1 à 3) et 3 variables catégorielles(4 à 12) qui seront des supplémentaires pour l'ACP. Les variables qualitatives vont nous permettre de caractériser la classe obtenue en découpage lors de la classification

Réalisation d'ACP normée pour la classification Ascendante hiérarchique On peut faire la classification en utilisant seulement les 5 ou 8 composants et non l'ensemble des donnéees. Si on veut utiliser l'ensemble des composantes de l'ACP dans la fonction PCA specifier l'argument ncp=Inf. Ainsi toute l'information est conservée. Ce qui revient à construire la classification sur toutes les donnéees. Cependant ca peut etre interessant de ne pas utiliser toutes les composantes de l'ACP car justement les dernières composantes peuvent représenter du bruit, et donc les enlever va permettre de stabiliser les resultats de la classification. Néanmoins on est pas obligé de conserver uniquement les 5 premiers composantes. En observant les valeurs propres on se rend compte qu'avec les 8 premiers composants on garde 95% de l'info par exemple.

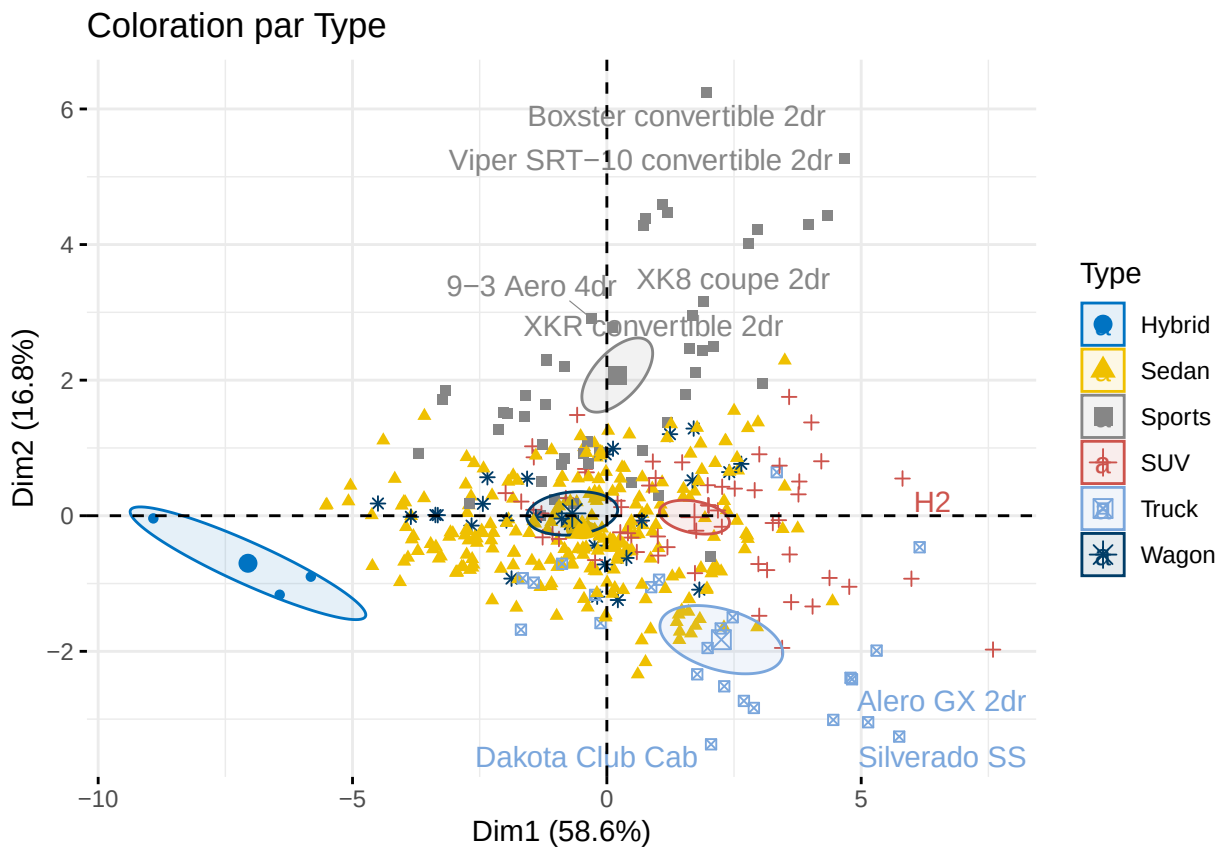
```
# scale.unit=TRUE permet de standardiser les variable-> ACP Normée
res.pca <- PCA(data, scale.unit=TRUE, quali.sup=1:3, ncp=5, graph = TRUE)
```





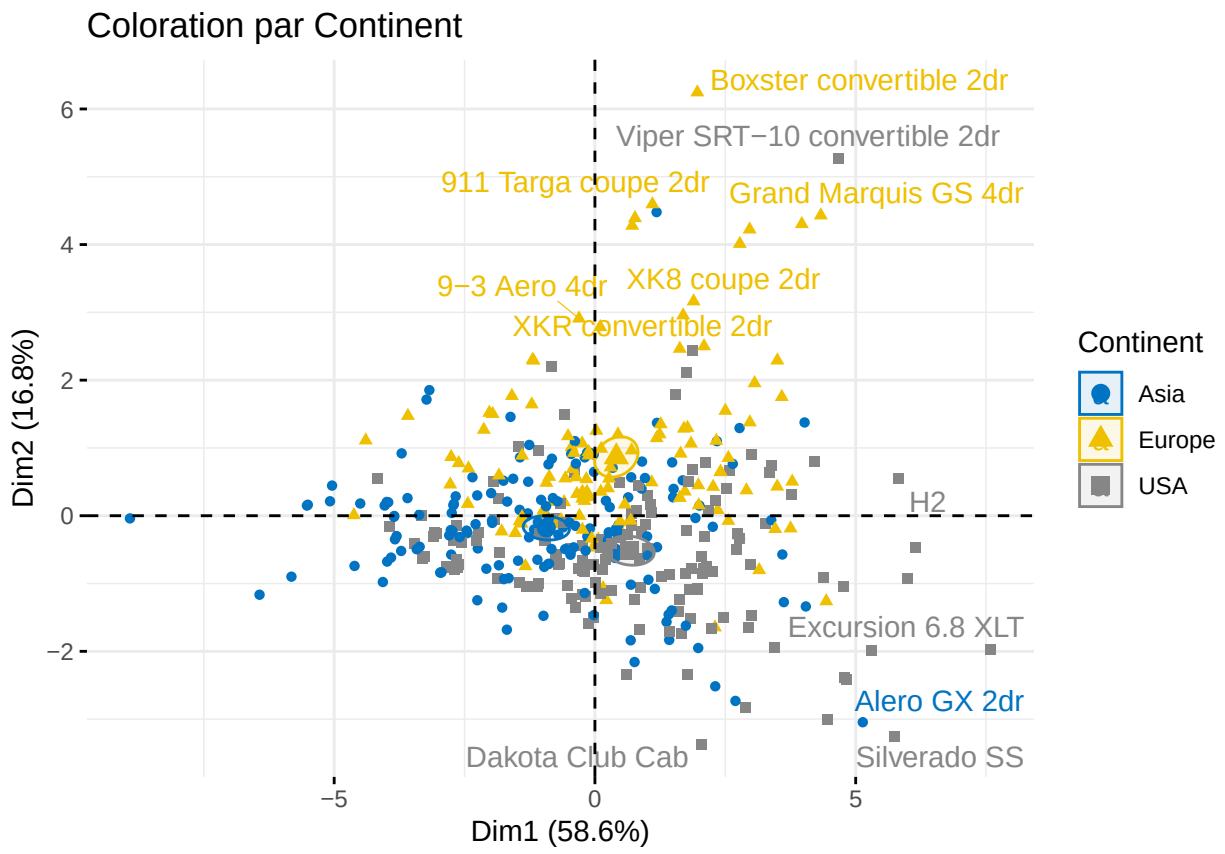

```
# Coloration par la quali Type
plot19 <- fviz_pca_ind(res.pca, habillage = 2,
             addEllipses =TRUE, ellipse.type = "confidence",
             palette = "jco", title="Coloration par Type",
             repel = TRUE)
plot19
```

```
## Warning: ggrepel: 416 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
# Coloration par la quanli Continent
plot20 <- fviz_pca_ind(res.pca, habillage = 1,
                       addEllipses =TRUE, ellipse.type = "confidence",
                       palette = "jco",title="Coloration par Continent", repel = TRUE)
plot20
```

```
## Warning: ggrepel: 413 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



En faisant cela on obtient un arbre hiérachique qui nous est proposé, avec les gains d'inertie à droite quand on passe d'une classe à une autre. Ces derniers nous permettent de voir combien de classes on peut conserver pour faire un découpage. On place la croix sur le niveau de compure pour avoir la projection des points sur le plan factoriel. Avec une coloration des individus selon le niveau d'appartenance de classes. On a egalement un arbre hiérachique qui est déssiné en trois dimension sur le plan de l'acp(resume l'info sur les deux dimension) Et puis l'arbre hierachique(lui représente l'info sur 5 dimensions plus que les 2 dimensions) et puis les couleurs (le regroupement en classe).

```
#Réalisation de la CAH avec HCPS
#res.hcps <- HCPC(res.pca)
```

Il nous retourne 4 objets interessants contenant les résultats.

```
#names(res.hcps)
```

Retourne un dataframe avec le jeu de données initiale et la colonne classe (clust) construite et ajoutée

```
#clusters <- res.hcps$data.clust
```

Description du découpage en classe obtenu par les variabes:

D'une part par les variables qualitatives, d'autre part par les quantitatives. Cependant on aura pas de resultats pour la variable qualitative, si elle est absente sur nos données ou ne permet pas de decrire notre variable de classe(clust)

1. *Description par les variables qualitatives:*

On constate un test de chi2 entre notre variable de classe et les variables qualitatives Cylinders,Type, Continent très significatif(P-value <0.05) Donc d'un point de vue global ces dernières sont liées au découpage en classe construite. On peut ensuite regarder modalité par modalité pour essayer d'expliquer ces liaisons. Regardons si notre variable de classe est liée à certaines modalité des variables cylinder, Type ou Continenent Au regard des résultats obtenus, on voit bien qu'il y une liason entre les modalités de clust et celles des variables catégorielles 69% des cylinders de type Cylinders_4 sont sont dans la classe 1. 99% des cylinders de la classe 1 sont de type cylinders_4. Donc les cylindres 4 sont sur représentés dans cette classe. 35% des continents (Asie) sont dans la classe 1. 58% des continents de la classe 1 sont des modèles asiatiques 100% des Types(=Hybrid) sont dans la classe 1. 100% des cylinders de type cylinders_5 sont dans la classe 2. 100% des cylinders de type cylinders_12 sont dans la classe 3. On peut appliquer ce même raisonnement pour trouver la distribution des modalités de chaque variable catégorielle sur notre découpage en classe.

2. Description de clust par les variables quantitatives.

Au regard du rapport de corrélation entre la variable quantitative et la variable de classe (mesure utilisée en analyse de variance pour voir la liaison entre quali et quanti) On constate que les variables quantitatives sont très liées à la variable qualitative clust.

un valeur positive (respectivement négatif) de v.test indique que les individus de la classe correspondante prennent des valeurs supérieures (respectivement inférieurs) à la moyenne de la classe en question.

```
#desc.var <-res.hcps$desc.var
```

Description par les axes factoriels:

dimensions = variables quantitaives. Clust est trÃ©s liÃ©s avec les axes factorielles d'aprÃ©s les p-values.

description par la première dimension de l'acp:liaison forte La dimension 1 prend des valeurs plus forte que la moyenne pour les individus de la classe 1 c'est à dire que Les indivdus de la classe 1 ont des coordonnéees significativement plus élevés queles autres

```
#res.hcps$desc.axes
```

description par les individus: Permet de voir les paragons de la classe paragons de la classe: individu plus proche du centre de gravité de la classe Le modèle de voiture le plus eloigné de toute les autres classes(centre de gravité) avec dist

```
#res.hcps$desc.ind
```

La methodes des k-means est utilisé pour consolider les classes.

```
#plot21 <-fviz_pca_ind(res.pca,
             #geom.ind = "point", # Montre les points seulement (mais pas le "text")
             #col.ind = clusters$clust, # colorer by groups
             #palette = c("#00AFBB", "#E7B800", "#FC4E07"),
             #addEllipses = TRUE, # Ellipses de concentration,
             #title = "ACP: Modèle 1",
              #subtitle = "Coloration par groupe",

             #legend.title = "Clust",
             #mean.point = FALSE
#)
#plot21
```

Coloration par groupe d'individus

```
#plot22 <- fviz_pca_ind(res.pca, geom.ind = "point",
#             col.ind = clusters$clust, # color by groups
#             palette = c("#00AFBB", "#E7B800", "#FC4E07"),
#             addEllipses = TRUE, ellipse.type = "convex",
#             title = "ACP: Modèle 2",
#              subtitle = "Coloration par groupe",
#             legend.title = "Clust"
#)
#plot22

#ind.p <- fviz_pca_ind(res.pca, geom = "point", col.ind = clusters$clust)
#plot23 <- ggpubr::ggpar(ind.p,
#              title = "ACP: Modèle 3",
#              subtitle = "Coloration par groupe",
#              caption = "Source: factoextra",
#              xlab = "PC1", ylab = "PC2",
#              legend.title = "Clust", legend.position = "top",
#              ggtheme = theme_gray(), palette = "jco"
#)
#plot23

#plot24<-fviz_pca_biplot(res.pca,
#                # Individus
#                geom.ind = "point",
#                fill.ind = clusters$clust, col.ind = "black",
#                pointshape = 21, pointsize = 2,
#                palette = "jco",
#                addEllipses = TRUE,
#                title = "synthèse de L'ACP et de la classification",
#                # Variables
#                alpha.var ="contrib", col.var = "contrib",
#                gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
#                legend.title = list(fill = "Clust", color = "Contrib",
#                                    alpha = "Contrib"))
#plot24
```

```
matrice = data[, 4:12]
data <- scale(matrice)
```

———————————————————————————- *K-MEANS* ———————————————————————

Approche : Affecter l'individu à la classe dont le barycentre est le plus proche

k -means avec les données centrées et réduites avec les données centrées et réduites fonction kmeans dans le package "stats". center = nb de groupes demandés

nstart = nb d'essais avec différents individus de départ

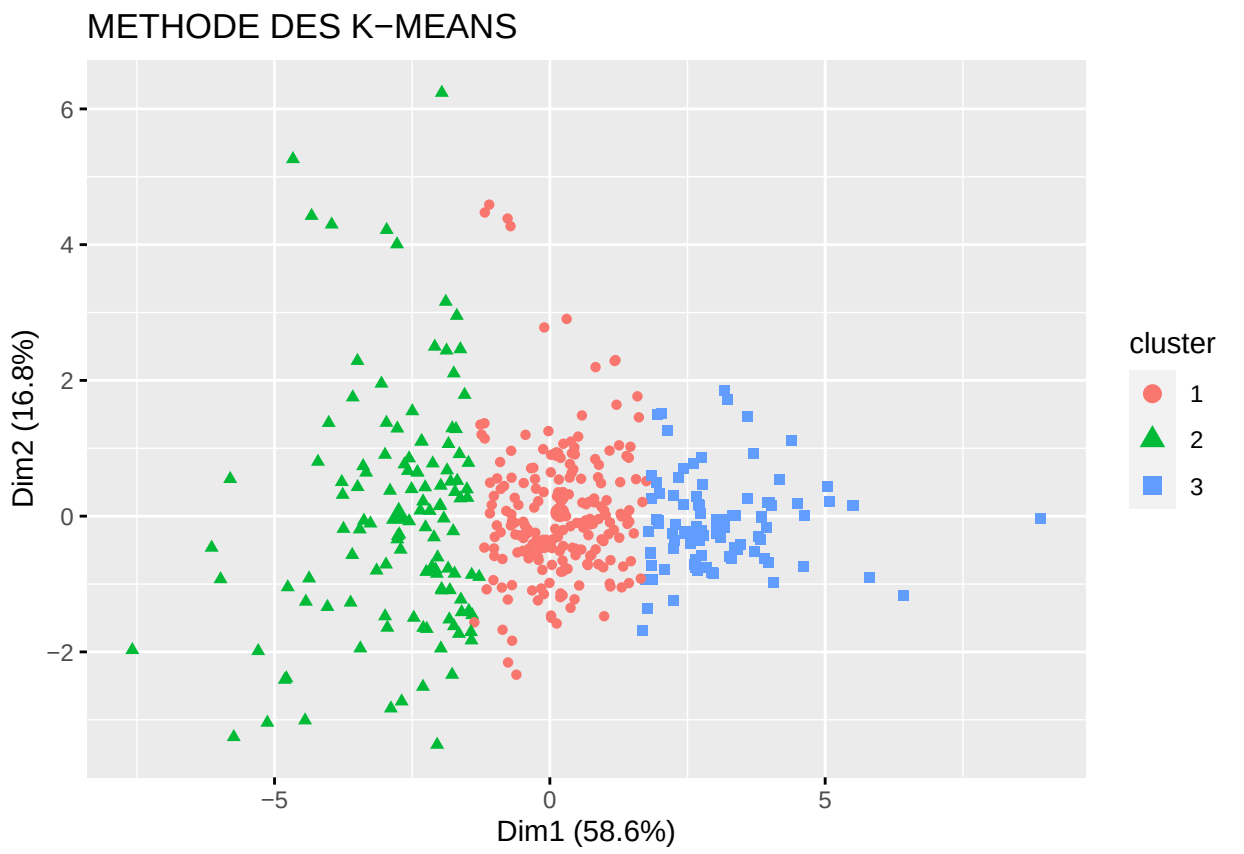
```
res.kmeans <- kmeans(data,centers=3,nstart=10)
#print(res.kmeans)
```

```
#res.kmeans$cluster
```

```
plot28<-fviz_cluster(res.kmeans, data = data, geom = "point",title="METHODE DES K-MEANS",ellipse.type =
```

```
## Warning: argument title is deprecated; please use main instead.
```

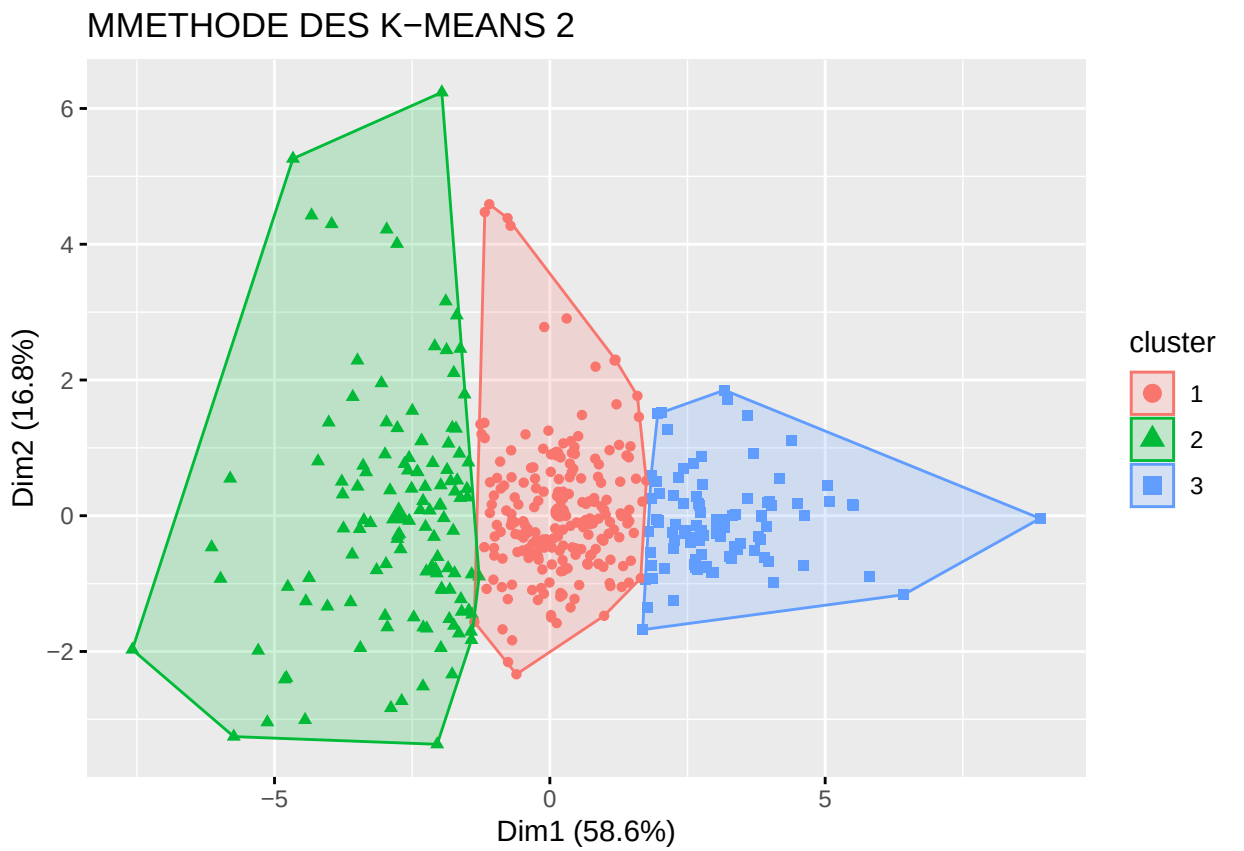
```
plot28
```




```
plot27<-fviz_cluster(res.kmeans, data = data, geom = "point",,title="MMETHODE DES K-MEANS 2",ellipse.ty
```

```
## Warning: argument title is deprecated; please use main instead.
```

```
plot27
```



Avec le graphique2 n a des classes bien homogènes ce quiveut dire que les individus sont bien classifier ; les groupes d'individus sont distincts

————————————————————————- *METHODE PAM* ————————————————

Ici on va faire de la Réallocation solide c'est à dire déplacer des individus d'un groupe à l'autre afin d'obtenir une meilleure partition On veut une matrice numerique qu'avec les quanti

```
matrice = data[, 1:9]
```

Centrer et réduire la matrice

```
data <- scale(matrice)
```

La méthode PAM proprement dite L'idée est de trouver des objets représentatifs médoïdes (par isolation aux barycentres) dans les classes (au lieu de la moyenne).Elle permet également d'améliorer les k-means en intégrant les points atypiques (outliers)
Le médoïde est un point qui existe réellement en composantes principales (coordonnées) et a la particularité d'avoir une distance minimale par rapport aux autres points du groupe . On fixe le K=3 obtenu depuis la CAH
metric : mesure de dissimilarité à utiliser entre les individus ; généralement la distance euclidienne ou de manhattan stand si vrai va standardiser avant de chercher à calculer les dissimilarités . D'ailleurs on peut ignorer ce paramètre déja fait plus haut

```
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 3.6.3
```

```
res.pam <- pam(data,3,metric="euclidean",stand = FALSE)
#Nprint(res.pam)
```

Nous fournit les médoides obtenus avec leurs mesures

```
print(res.pam$medoids)
```

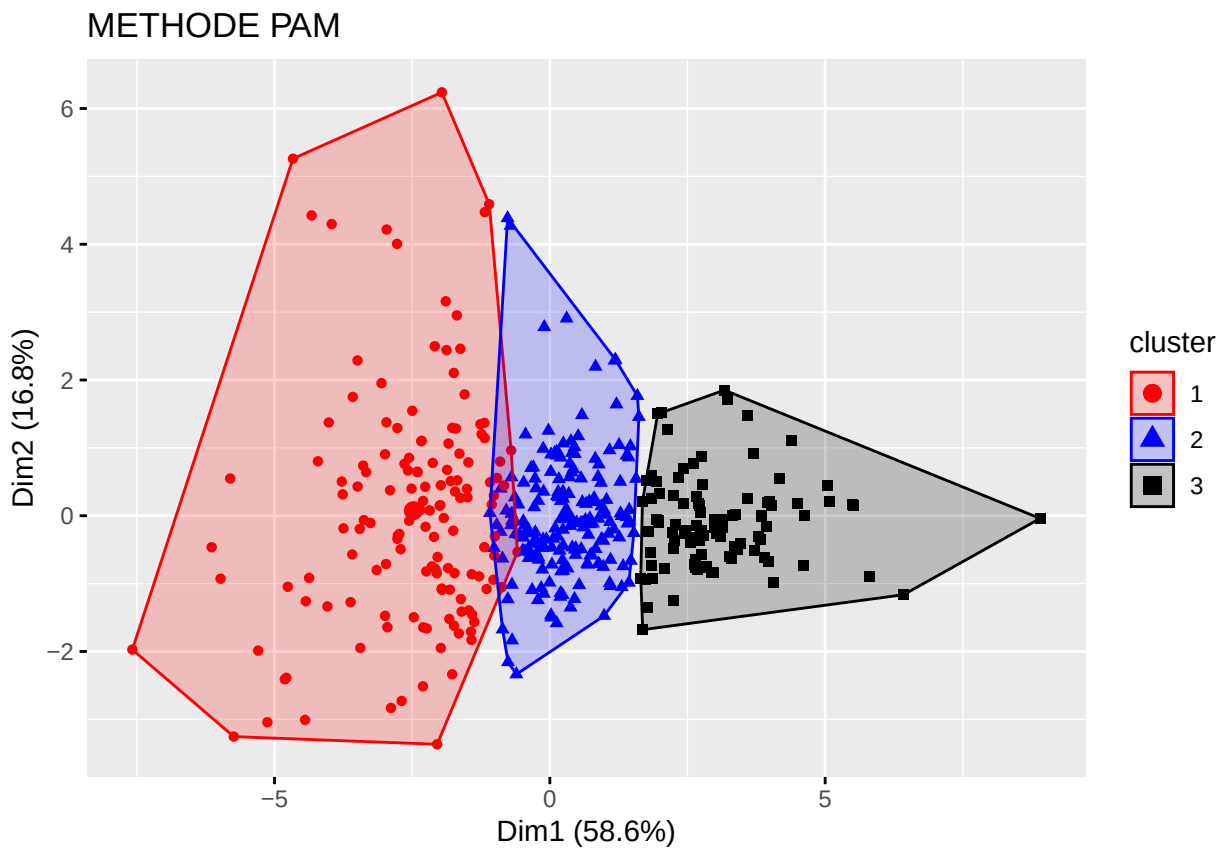
```
##                              MSRP    Invoice EngineSize  Horsepower   MPG_City
## TrailBlazer LT           0.5438587  0.0882757  0.8989091  0.82086877 -0.7748847
## Camry Solara SLE V6 2dr  0.1754428 -0.3303581 -0.1820709 -0.08178626  0.1774750
## Elantra GLS 4dr         -0.2623451 -0.7718137 -1.0828876 -1.08165031  1.1298348
##                          MPG_Highway     Weight  Wheelbase     Length
## TrailBlazer LT           -1.0179457  1.1177942  0.5812799  0.3907194
## Camry Solara SLE V6 2dr   0.3731376 -0.3698068 -0.1389649  0.1824555
## Elantra GLS 4dr           1.2425647 -1.2407582 -0.6191280 -0.5811788
```

Visualisation

```
plot25<-fviz_cluster(res.pam,data=data,palette=c("red","blue","black"),geom="point",ellipse.type = "conv
```

```
## Warning: argument title is deprecated; please use main instead.
```

```
plot25
```



Silhouette : Degré d'appartenance à sa classe d'un individu ou de score

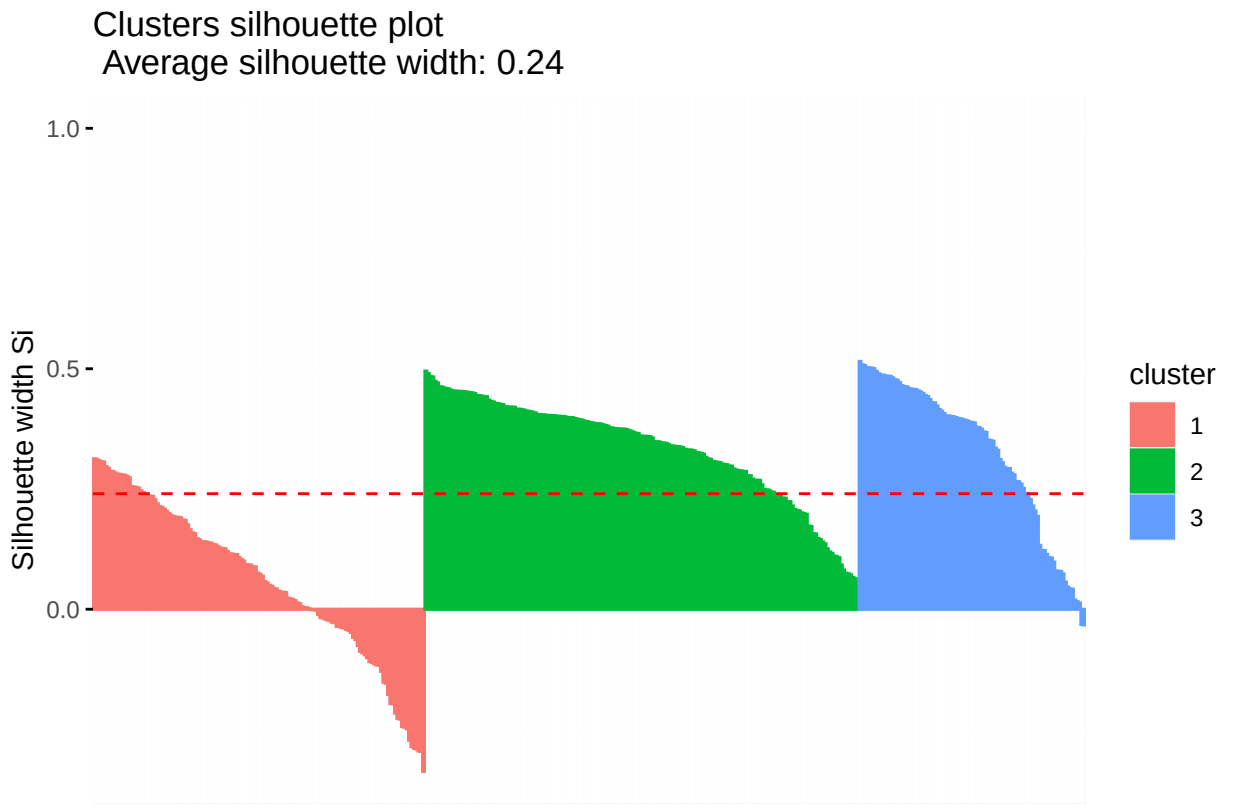
On regarde l'homogenéité des clusters ; parxemple la classe 3 on voit que l'essentiel individus ont de bonnes valeurs de silhouette

```
sil = silhouette(res.pam$cluster,dist(matrice))
```

```
plot26<-fviz_silhouette(sil)
```

```
##   cluster size ave.sil.width
## 1       1  142          0.06
## 2       2  186          0.33
## 3       3   97          0.33
```

```
plot26
```



On pourrait améliorer la capacité à traiter les grandes bases de PAM en travaillant sur des échantillons (approche CLARA)