Analyse Factorielle Discriminante - Correction

```
rm(list=ls())
library("tidyverse")
library("FactoMineR") #pour effectuer l'ACP
library("factoextra") #pour extraire et visualiser les résultats issus de FactoMineR
library("ade4")
library("corrplot")
```

1 Insectes

```
load("insectes.rda")
head(insectes)
```

```
##    X1  X2 X3  X4 X5  X6 type
## 1 191 131 53 150 15 104    A
## 2 185 134 50 147 13 105    A
## 3 200 137 52 144 14 102    A
## 4 173 127 50 144 16  97    A
## 5 171 118 49 153 13 106    A
## 6 160 118 47 140 15  99    A
```

```
str(insectes)
```

```
## 'data.frame':    74 obs. of  7 variables:
##  $ X1  : num  191 185 200 173 171 160 188 186 174 163 ...
##  $ X2  : num  131 134 137 127 118 118 134 129 131 115 ...
##  $ X3  : num  53 50 52 50 49 47 54 51 52 47 ...
##  $ X4  : num  150 147 144 144 153 140 151 143 144 142 ...
##  $ X5  : num  15 13 14 16 13 15 14 14 14 15 ...
##  $ X6  : num  104 105 102 97 106 99 98 110 116 95 ...
##  $ type: Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
```

```
dim(insectes)
```

```
## [1] 74  7
```

1.1 Quelques analyses descriptives

1.1.1 Moyenne et écart-type globaux

```r
# Moyenne
insectes %>% summarise_if(is.numeric,mean)
```

```
##         X1       X2       X3       X4       X5       X6
## 1 177.2568 123.9595 50.35135 134.8108 12.98649 95.37838
```

```r
# Ecart-type
insectes %>% summarise_if(is.numeric,sd)
```

```
##         X1       X2       X3       X4       X5       X6
## 1 29.41254 8.481146 2.751998 10.35093 2.142162 14.30461
```

1.1.2 Moyenne et écart-type par groupe

```r
# Moyenne
insectes %>% group_by(type) %>% summarise_if(is.numeric,mean)
```

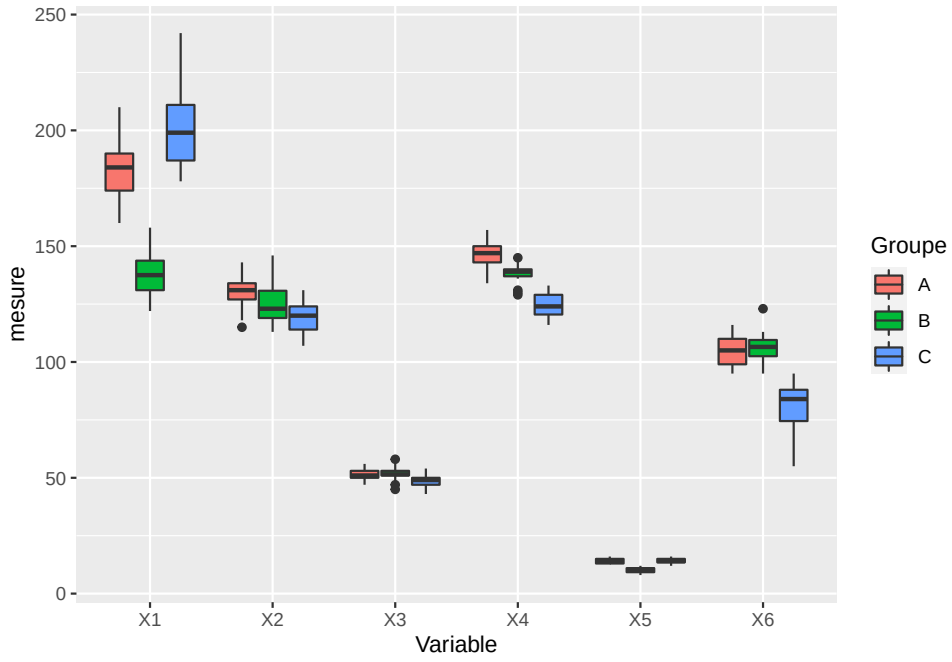
```
## # A tibble: 3 x 7
##   type     X1    X2    X3    X4    X5    X6
## * <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 A      183.  130.  51.2  146.  14.1  105.
## 2 B      138.  125.  51.6  138.  10.1  107.
## 3 C      201   119.  48.9  125.  14.3   81
```

```r
# Ecart-type
insectes %>% group_by(type) %>% summarise_if(is.numeric,sd)
```

```
## # A tibble: 3 x 7
##   type     X1    X2    X3    X4    X5    X6
## * <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 A     12.1   7.16  2.23  5.63 0.889  6.18
## 2 B      9.34  8.55  2.84  4.14 0.971  5.85
## 3 C     14.9   6.65  2.35  4.62 1.10   8.93
```

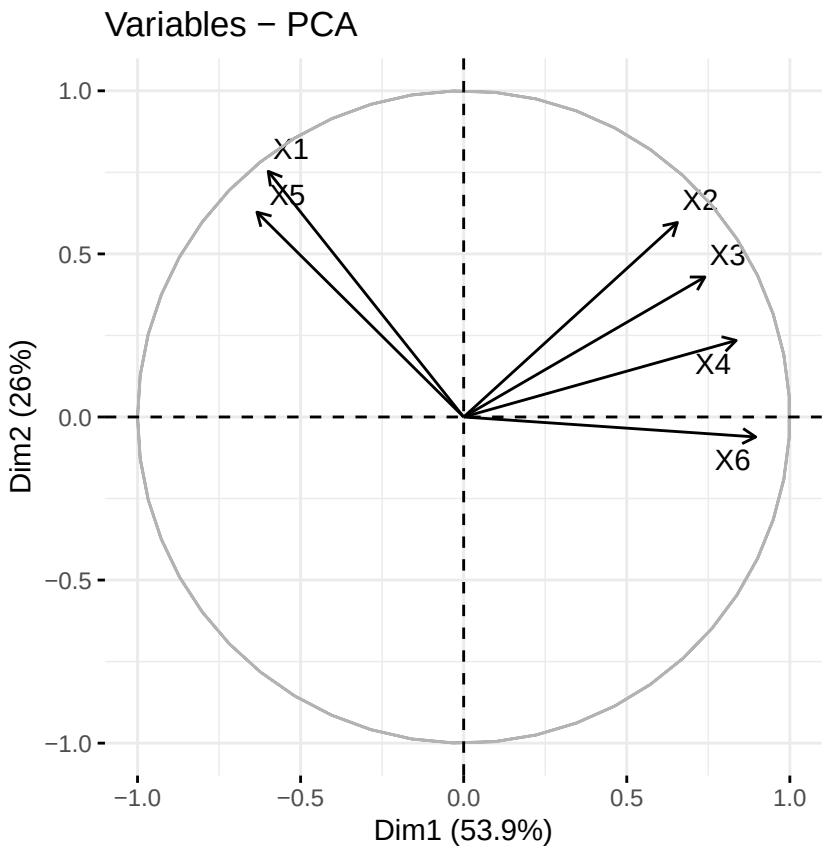
1.1.3 Distribution empirique des variables quantitatives selon type

```r
# réorganisation du jeu de données
insectes.G <- insectes %>% dplyr::select_if(is.numeric) %>%
  gather(key='Variable',value='mesure') %>% mutate(Groupe=rep(insectes$type,6))
ggplot(insectes.G, aes(x=Variable, y=mesure,fill=Groupe)) + geom_boxplot()
```

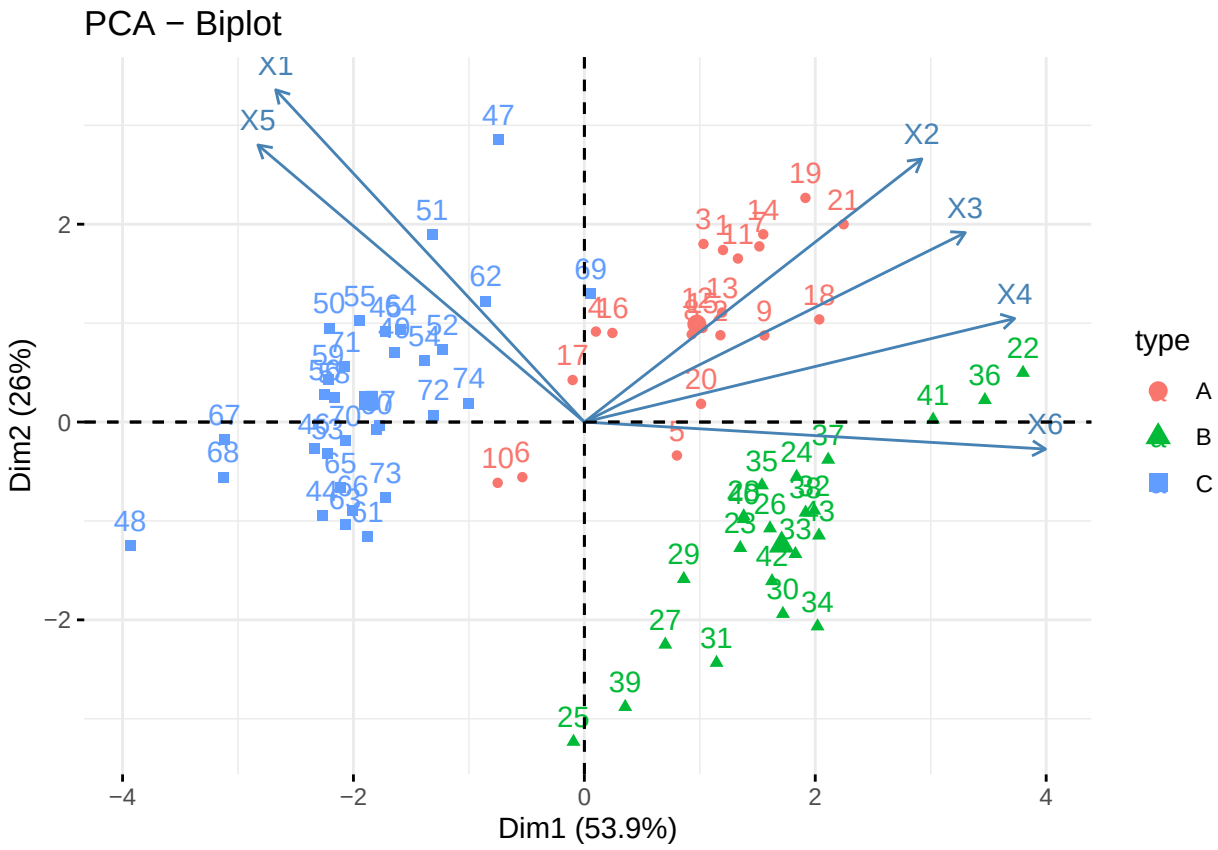


1.1.4 ACP

```
res.pca=PCA(insectes,scale.unit = TRUE,quali.sup = 7,graph=FALSE)
fviz_pca_var(res.pca,axes=c(1,2),repel = TRUE,labelsize=4)
```



```
fviz_pca_biplot(res.pca,axes=c(1,2),habillage="type",labelsize=4)
```



1.1.5 AFD

Code de l'AFD

```
X <- insectes %>% dplyr::select_if(is.numeric)
y <- insectes$type
insectes.dis = discrimin(dudi.pca(X, scan = FALSE),y, scan = FALSE)
```

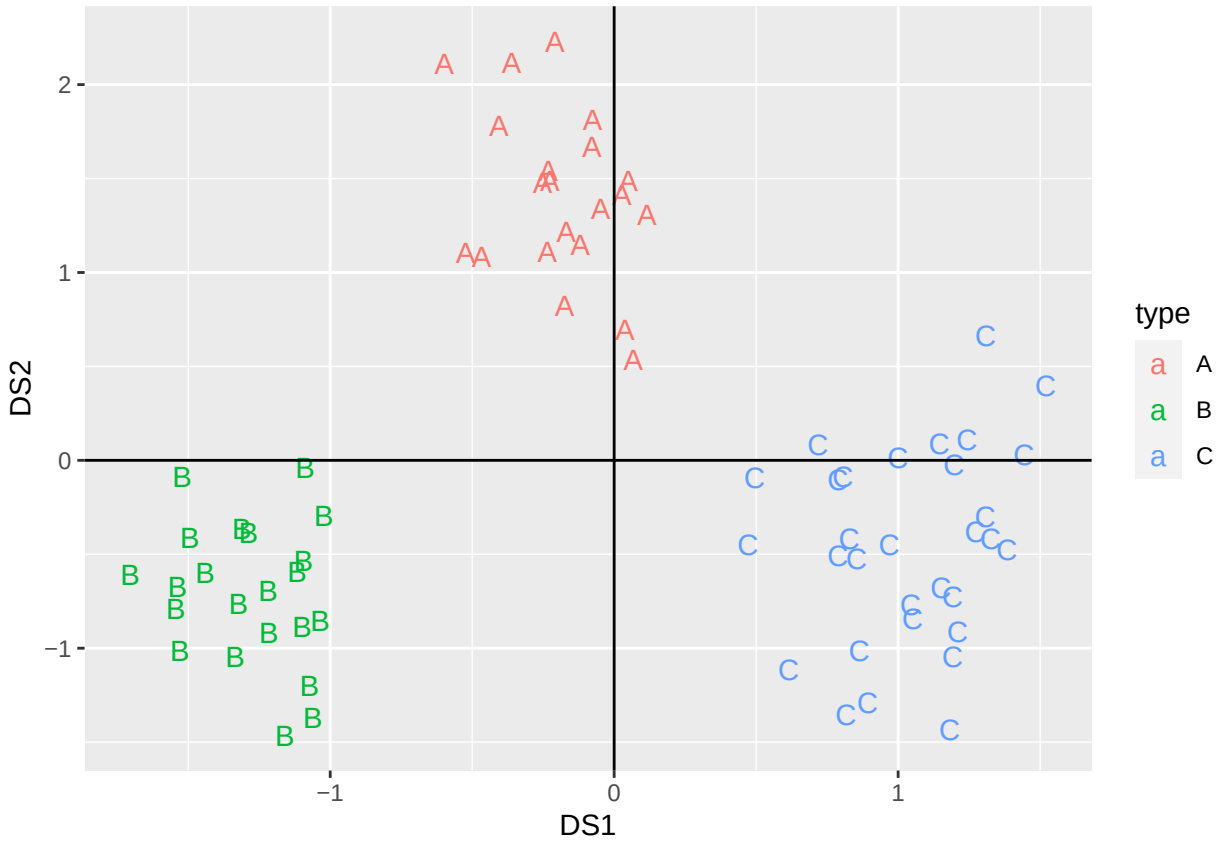
Valeurs propres

```
insectes.dis$eig
```

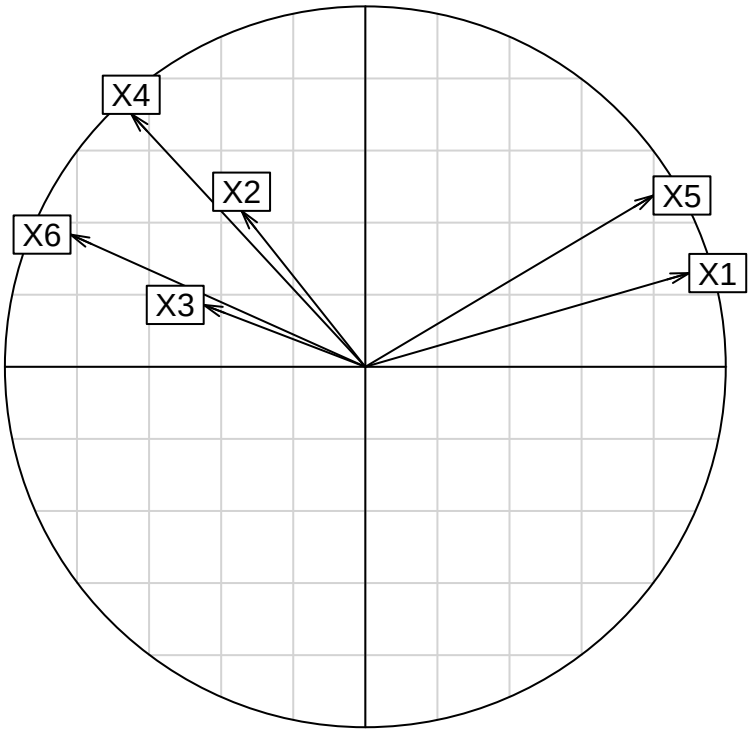
```
## [1] 0.9467500 0.7952981
```

Graphe des individus

```
coord <- insectes.dis$li %>% mutate(type=y)
ggplot(coord,aes(x=DS1,y=DS2,label =type))+
  geom_text(aes(color=type))+geom_hline(yintercept=0)+geom_vline(xintercept = 0)
```



Graphe des variables



Vecteurs propres

```
insectes.dis$fa
```

```
##            DS1         DS2
## X1  0.63804555  0.18391025
## X2 -0.11754375  0.16170590
## X3 -0.08839895 -0.36661997
## X4 -0.15770018  0.95973381
## X5  0.14414920  0.51362282
## X6 -0.15745716  0.07519586
```

Centre de gravité des 3 groupes

```
knitr::kable(aggregate(X, by=list(y),mean),digits=1)
```

Group.1	X1	X2	X3	X4	X5	X6
A	183.1	129.6	51.2	146.2	14.1	104.9
B	138.2	125.1	51.6	138.3	10.1	106.6
C	201.0	119.3	48.9	124.6	14.3	81.0

Coordonnées des centres de gravités projetés

```
insectes.dis$gc
```

```
##          DS1        DS2
## A -0.1846207  1.4066089
## B -1.2818668 -0.7067343
## C  1.0347776 -0.4513107
```

2 Les iris

```
data(iris)
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
```

```
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
dim(iris)
```

```
## [1] 150   5
```

2.1 Quelques analyses descriptives

2.1.1 Moyenne et écart-type globaux

```
# Moyenne
iris %>% summarise_if(is.numeric,mean)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     5.843333    3.057333        3.758    1.199333
```

```
# Ecart-type
iris %>% summarise_if(is.numeric,sd)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1    0.8280661   0.4358663     1.765298   0.7622377
```

2.1.2 Moyenne et écart-type par groupe

```
# Moyenne
iris %>% group_by(Species) %>% summarise_if(is.numeric,mean)
```

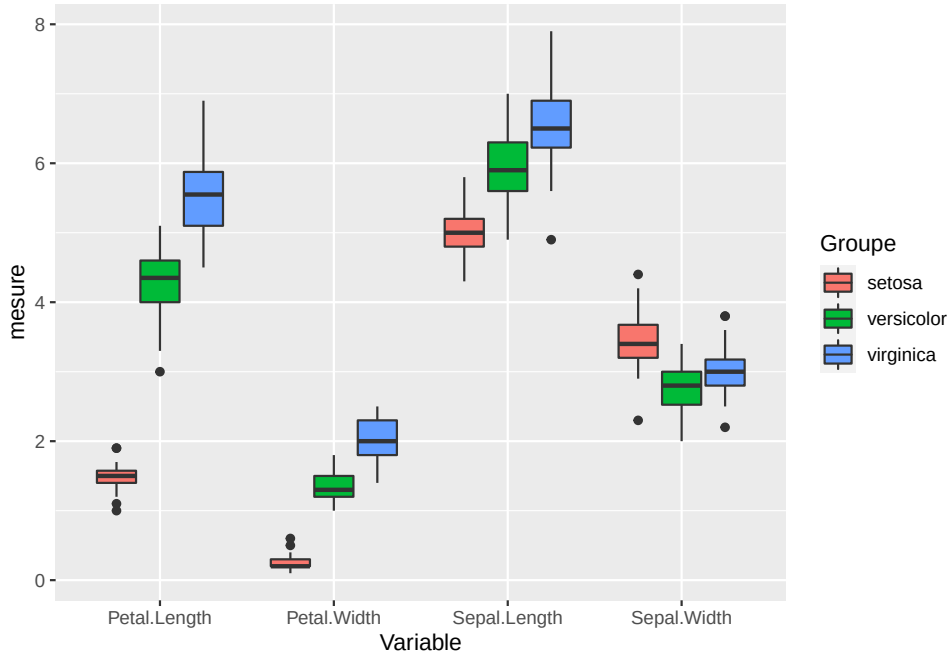
```
## # A tibble: 3 x 5
##   Species    Sepal.Length Sepal.Width Petal.Length Petal.Width
## * <fct>             <dbl>       <dbl>        <dbl>       <dbl>
## 1 setosa             5.01        3.43         1.46       0.246
## 2 versicolor         5.94        2.77         4.26       1.33
## 3 virginica          6.59        2.97         5.55       2.03
```

```
# Ecart-type
iris %>% group_by(Species) %>% summarise_if(is.numeric,sd)
```

```
## # A tibble: 3 x 5
##   Species    Sepal.Length Sepal.Width Petal.Length Petal.Width
## * <fct>             <dbl>       <dbl>        <dbl>       <dbl>
## 1 setosa            0.352       0.379        0.174       0.105
## 2 versicolor        0.516       0.314        0.470       0.198
## 3 virginica         0.636       0.322        0.552       0.275
```

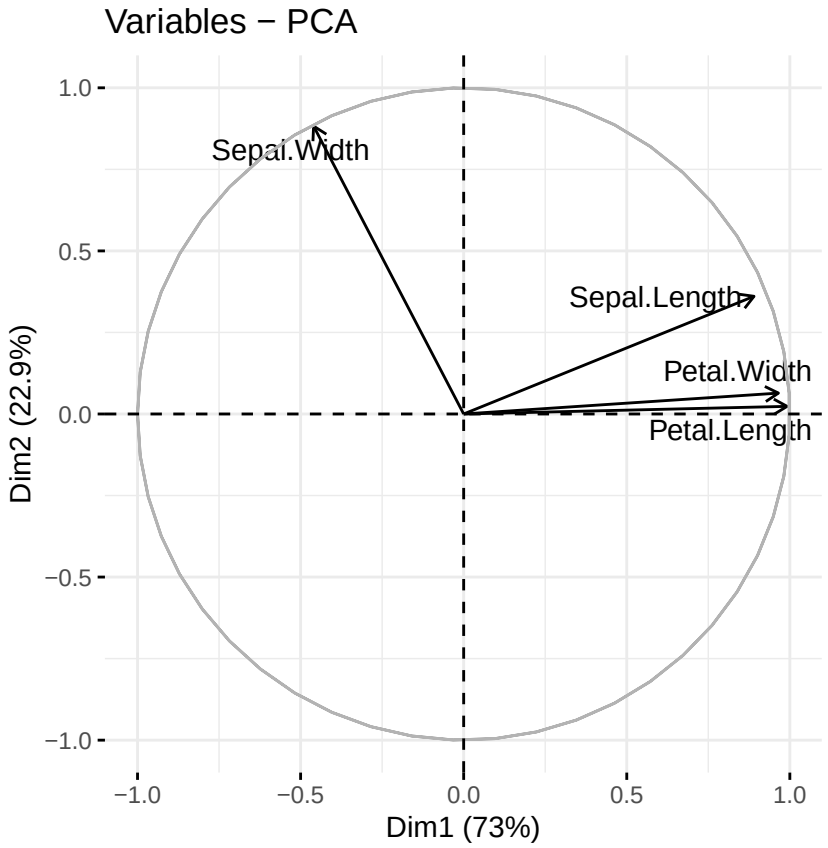
2.1.3 Distribution empirique des variables quantitatives selon `Species`

```r
# réorganisation du jeu de données
iris.G <- iris %>% dplyr::select_if(is.numeric) %>%
  gather(key='Variable',value='mesure') %>% mutate(Groupe=rep(iris$Species,4))
ggplot(iris.G , aes(x=Variable, y=mesure,fill=Groupe)) + geom_boxplot()
```

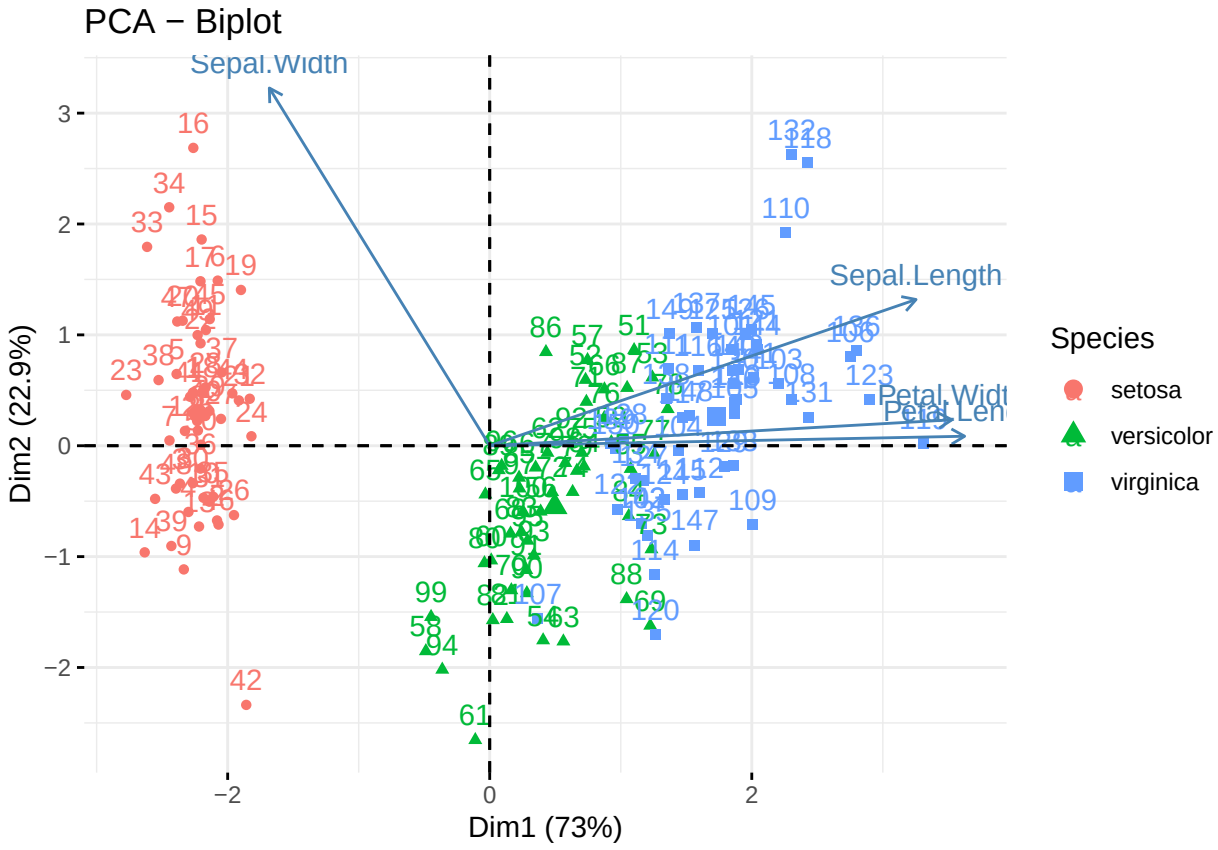


2.1.4 ACP

```r
res.pca=PCA(iris,scale.unit = TRUE,quali.sup = 5,graph=FALSE)
fviz_pca_var(res.pca,axes=c(1,2),repel = TRUE,labelsize=4)
```




```
fviz_pca_biplot(res.pca,axes=c(1,2),habillage="Species",labelsize=4)
```



2.1.5 AFD

Code de l'AFD

```
X <- iris %>% dplyr::select_if(is.numeric)
y <- iris$Species
iris.dis = discrimin(dudi.pca(X, scan = FALSE),y, scan = FALSE)
```

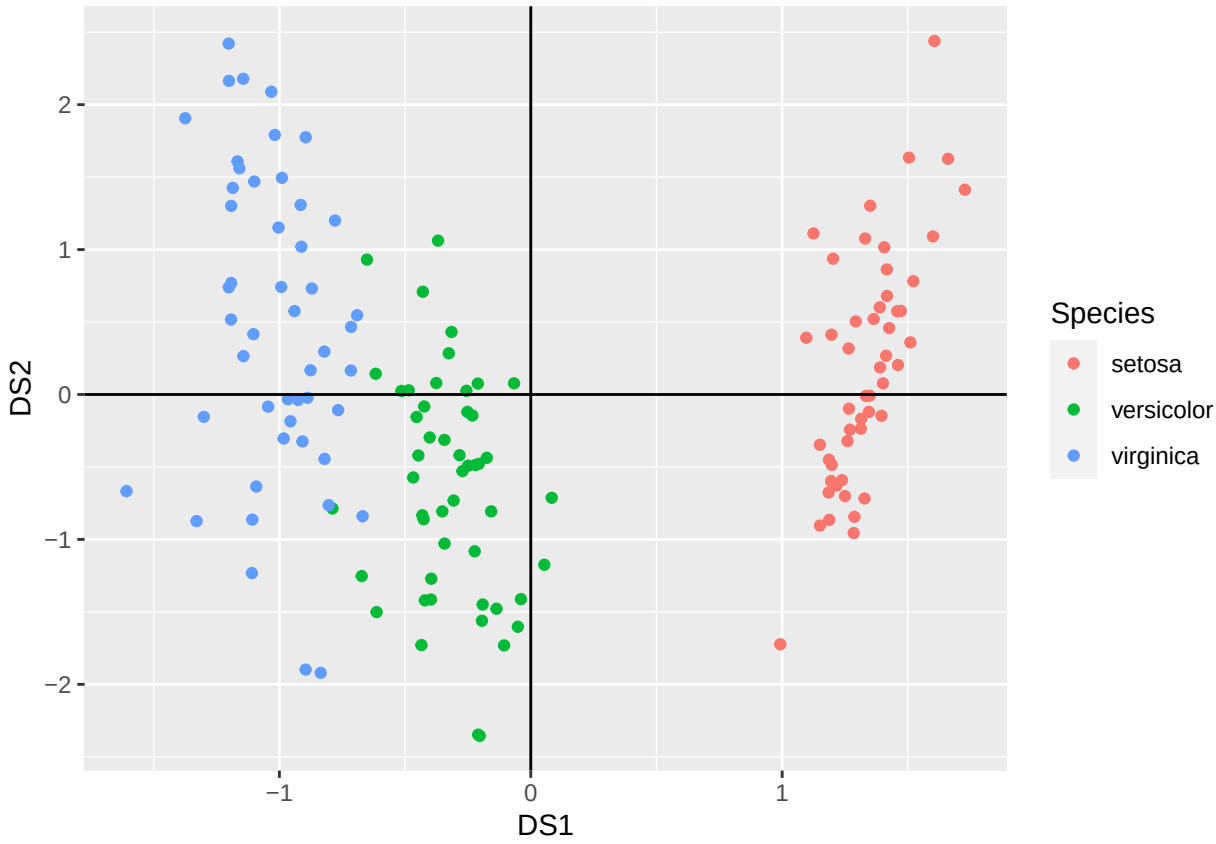
Valeurs propres

```
iris.dis$eig
```

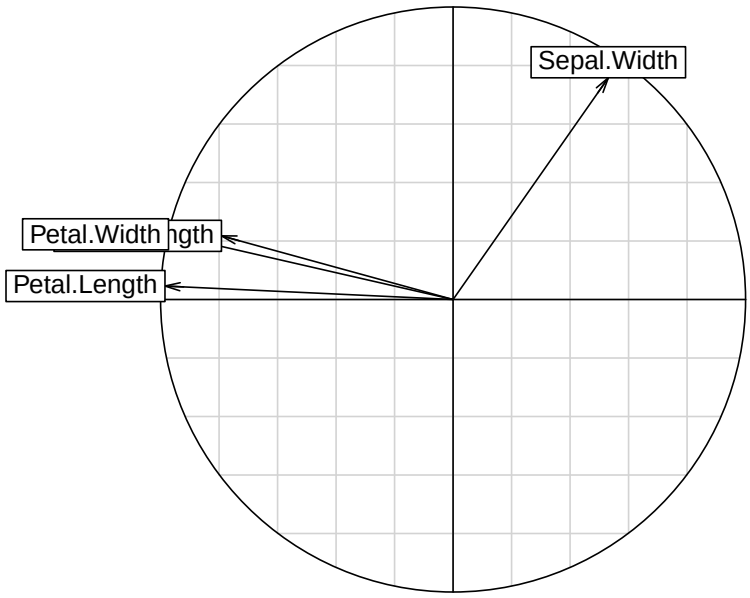
```
## [1] 0.9698722 0.2220266
```

Graphe des individus

```
coord <- iris.dis$li %>% mutate(Species=y)
ggplot(coord,aes(x=DS1,y=DS2,color =Species))+
  geom_point()+geom_hline(yintercept=0)+geom_vline(xintercept = 0)
```



Graphe des variables



Vecteurs propres

```
iris.dis$fa
```

```
##                    DS1          DS2
```

```
## Sepal.Length  0.1200150  0.01772302
## Sepal.Width   0.1168775  0.83778380
## Petal.Length -0.6790443 -1.46087856
## Petal.Width  -0.3743571  1.92176982
```

Centre de gravité des 3 groupes

```
knitr::kable(aggregate(X, by=list(y),mean),digits=1)
```

Group.1	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	5.0	3.4	1.5	0.2
versicolor	5.9	2.8	4.3	1.3
virginica	6.6	3.0	5.6	2.0

Coordonnées des centres de gravités projetés

```
iris.dis$gc
```

```
##                   DS1        DS2
## setosa      1.3338850  0.1916798
## versicolor -0.3199966 -0.6485461
## virginica  -1.0138884  0.4568662
```

Vecteurs propres: coefficients des combinaisons linéaires pour la construction des variables discriminantes

```
iris.dis$va
```

```
##                     CS1        CS2
## Sepal.Length -0.7918878 0.21759312
## Sepal.Width   0.5307590 0.75798931
## Petal.Length -0.9849513 0.04603709
## Petal.Width  -0.9728120 0.22290236
```

Conclusion:

- l'axe 1 a un pouvoir très discriminant et discrimine les 3 espèces, particulièrement l'espèce setosa des deux autres. Cette discrimination se fait sur les variables `Sepal.Length`, `Petal.Length` et `Petal.Width`: les Virginica ont plutôt des longueurs de sépales, des longueurs et des largeurs de pétales importantes. Les Setosa possèdent à l'inverse des longueurs de sépales, des longueurs et des largeurs de pétales réduites. Les Versicolor occupent une position intermédiaire.
- l'axe 2 ne discrimine pas les espèces.
- La variable `Sepal.Width` sépare mais moins nettement (1ère bissectrice) l'espèce setosa et des deux autres.
- Pour la première variable discriminante s_1, les mesures sur les pétales s'opposenrt aux mesures effectuées sur les sépales. Pour la seconde variable discriminante, la largeur des pétales et la longueurs des sépales s'opposent à la longueur des pétales (c'est le signe qui compte).

3 Banque

```
banque = read.table("Banque.csv", header = TRUE, dec = ",",sep=";")
head(banque)
```

```
##     solde mdecouv ncompte memprunt   mdepot mretrait nbenf age
## 1  245.00 1139.67       0   129.57    31.50     0.00     0  22
## 2 2326.07    0.00       2  3810.98 63516.26  2330.18     0  45
## 3  188.41 1503.14       1     0.00    60.98   785.39     0  19
## 4 1256.25  227.96       9 32012.20  2439.03 14246.30     0  62
## 5  946.65  305.66       9 17225.61 11432.93  9291.94     0  36
## 6 1047.41  487.69       7 30487.80  9527.44  5688.58     0  31
##                   csp code
## 1              autre    6
## 2 artisan-commercant    1
## 3              autre    6
## 4           retraite    5
## 5            ouvrier    4
## 6            ouvrier    4
```

```
str(banque)
```

```
## 'data.frame':    500 obs. of  10 variables:
##  $ solde   : num  245 2326 188 1256 947 ...
##  $ mdecouv : num  1140 0 1503 228 306 ...
##  $ ncompte : int  0 2 1 9 9 7 1 5 1 1 ...
##  $ memprunt: num  130 3811 0 32012 17226 ...
##  $ mdepot  : num  31.5 63516.3 61 2439 11432.9 ...
##  $ mretrait: num  0 2330 785 14246 9292 ...
##  $ nbenf   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age     : int  22 45 19 62 36 31 25 28 24 36 ...
##  $ csp     : chr  "autre" "artisan-commercant" "autre" "retraite" ...
##  $ code    : int  6 1 6 5 4 4 6 4 6 6 ...
```

```
banque$code <- as.factor(banque$code)
dim(banque)
```

```
## [1] 500  10
```

3.1 Quelques analyses descriptives

3.1.1 Moyenne par groupe

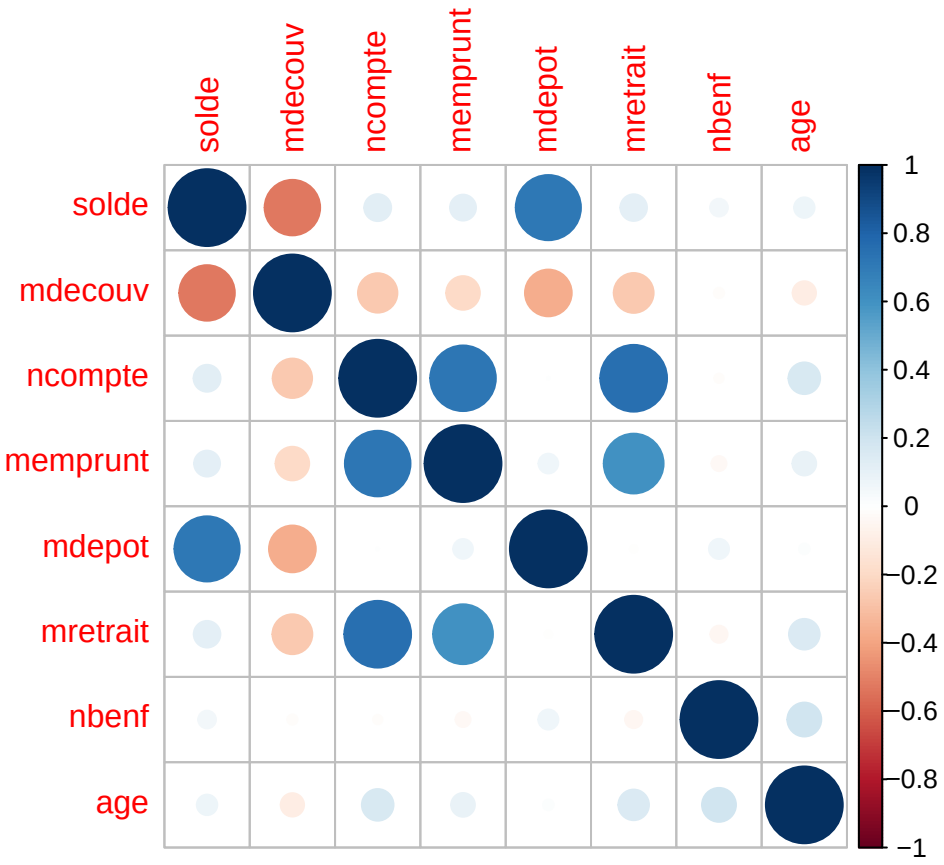
```
banque %>% group_by(csp) %>% summarise_if(is.numeric,mean)
```

```
## # A tibble: 6 x 9
##   csp                solde mdecouv ncompte memprunt mdepot mretrait nbenf   age
## * <chr>              <dbl>   <dbl>   <dbl>    <dbl>  <dbl>    <dbl> <dbl> <dbl>
```

```
## 1 artisan-commercant 2437.    326.    2.24    9615. 55552.    3086. 0.684  38.1
## 2 autre               430.    836.    1.59    2173.  1087.    1522. 0.602  35.2
## 3 cadre              2989.    107.    3.02   10176. 69903.    3762. 0.425  37.6
## 4 employe            1475.    320.    3.87   13211. 17473.    4917. 0.506  37.1
## 5 ouvrier             941.    537.    2.32    6115.  4908.    3246. 0.509  34.8
## 6 retraite           1369.    468.    3.13    9020. 12528.    4350. 0.526  67.1
```

3.1.2 Corrélation entre variables

```
correlation <- banque  %>% dplyr::select_if(is.numeric) %>%cor(.)
correlation %>% corrplot
```

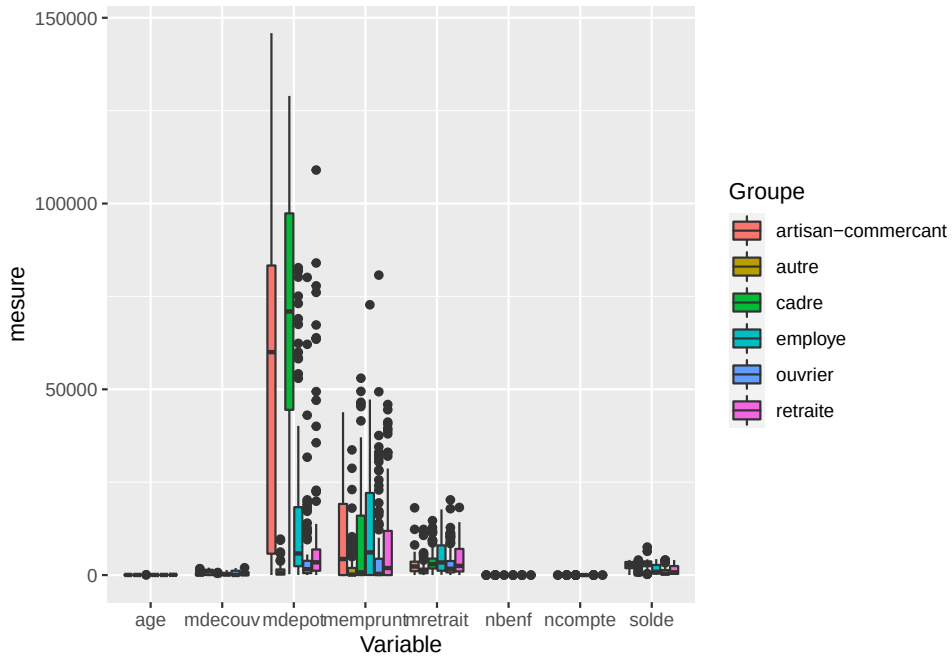


On observe des corrélations logiques entre certaines variables:

- variable `solde`: positive avec `mdepot` et négative entre avec `mdecouv`
- variable `mretrait`: positive avec `ncompte` et `nemprunt`

3.1.3 Distribution empirique des variables quantitatives selon `csp`

```
# réorganisation du jeu de données
banque.G <- banque  %>% dplyr::select_if(is.numeric) %>%
  gather(key='Variable',value='mesure') %>% mutate(Groupe=rep(banque$csp,8))
ggplot(banque.G , aes(x=Variable, y=mesure,fill=Groupe)) + geom_boxplot()
```



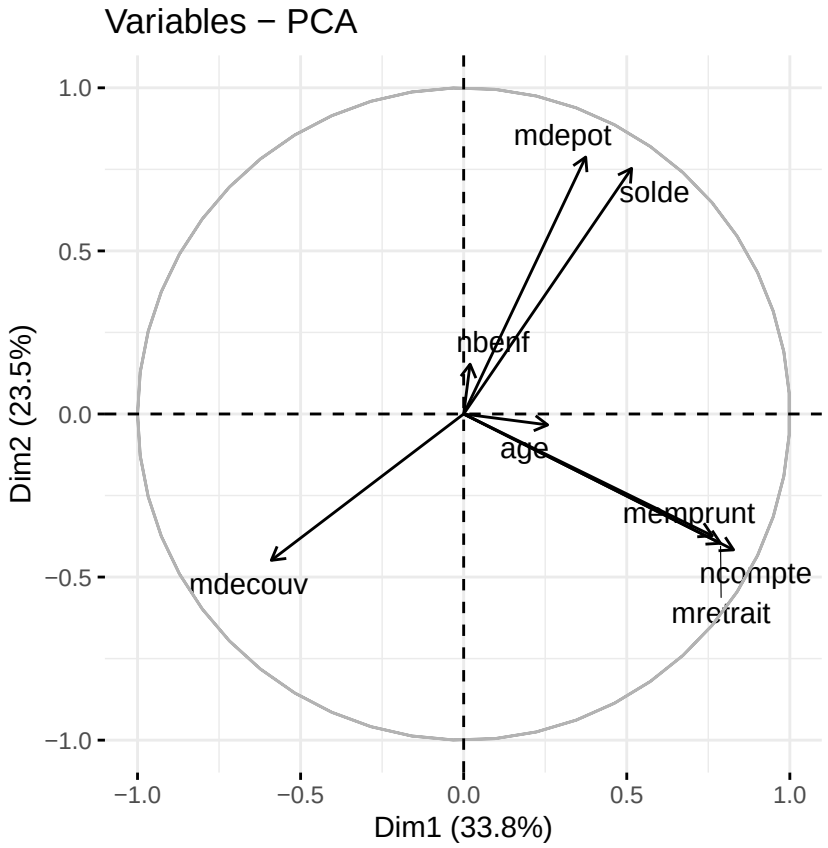
On observe des différences entre groupes pour les variables `mdepot`, `memprunt`, `mretrait`, et très légèrement `solde`.

3.1.4 ACP

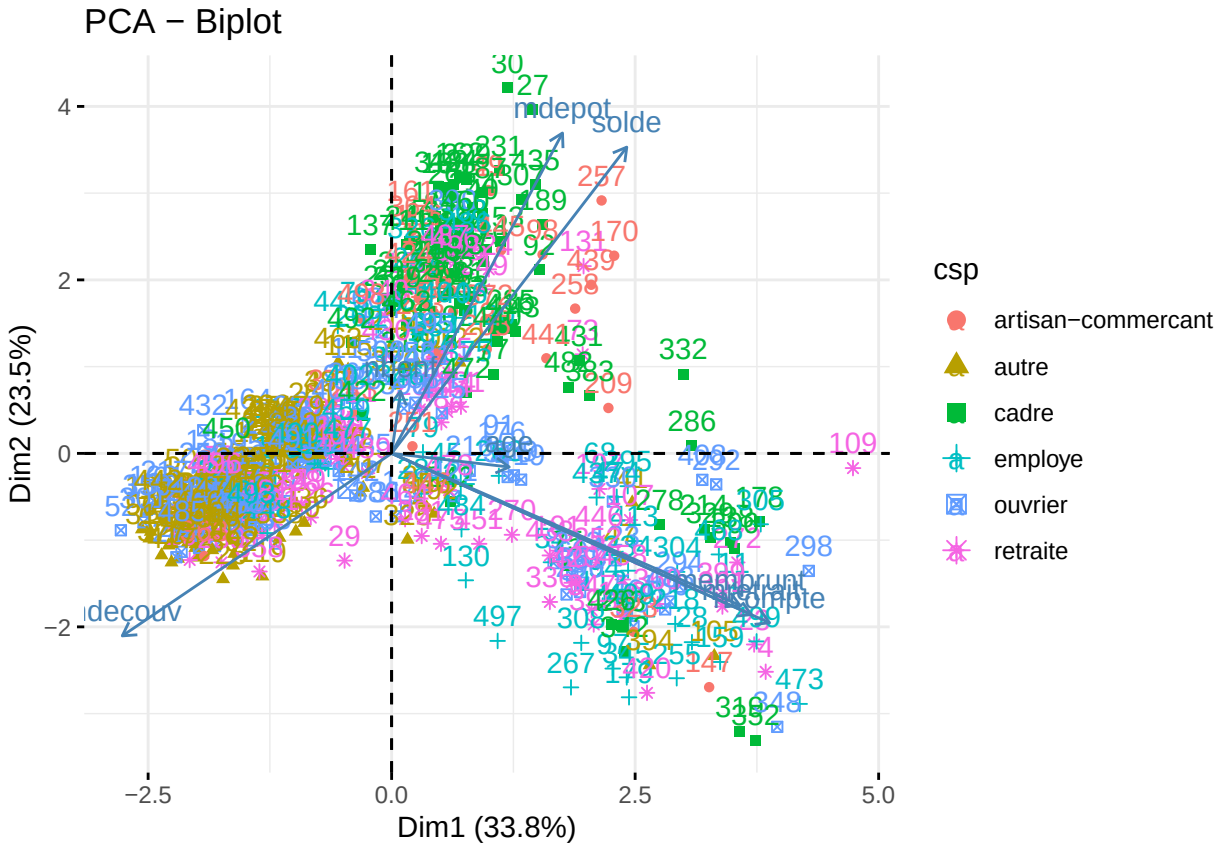
```
res.pca=PCA(banque,scale.unit = TRUE,quali.sup = 9:10,graph=FALSE)
```

Premier plan

```
fviz_pca_var(res.pca,axes=c(1,2),repel = TRUE,labelsize=4)
```



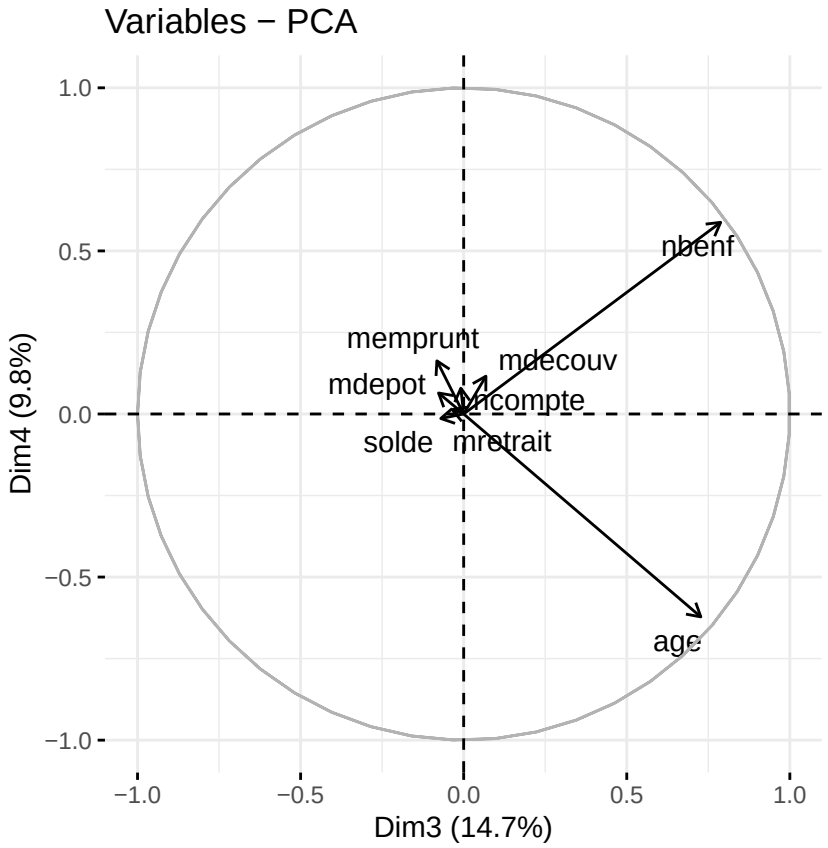

```
fviz_pca_biplot(res.pca,axes=c(1,2),habillage="csp",labelsize=4)
```



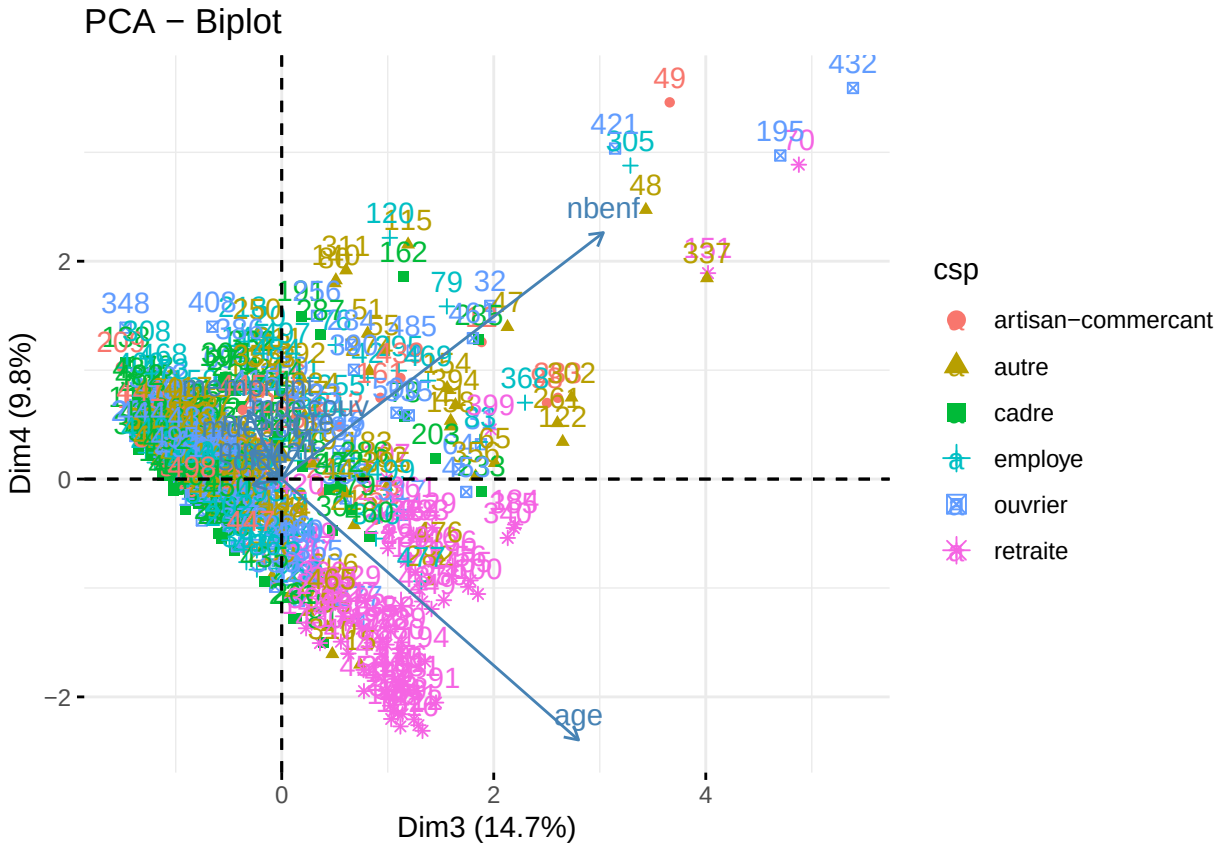
Les variables bien représentées sont celles pour lesquelles on avait une corrélation avec une autre (les variables `nbenf` et `âge` sont mal représentées sur ce plan principal). Il est difficle de distinguer les groupes sur ce plan. Seules les catégories des cadres et artisans se distinguent.

Deuxième plan

```
fviz_pca_var(res.pca,axes=c(3,4),repel = TRUE,labelsize=4)
```




```
fviz_pca_biplot(res.pca,axes=c(3,4),habillage="csp",labelsize=4)
```



Seules les variables `nbenf` et `âge` sont bien représentées sur ce deuxième plan principal. Des groupes (qui ne correspondent pas aux catégories socio-professionnelles) se séparent dans la direction `nbenf`. Le groupe des retraités se distingue des autres (selon la variable âge‘, résultat attendu).

3.1.5 AFD

Code de l'AFD

```
X <- banque %>% dplyr::select_if(is.numeric)
y <- as.factor(banque$csp)
banque.dis = discrimin(dudi.pca(X, scan = FALSE),y, scan = FALSE)
```

Valeurs propres

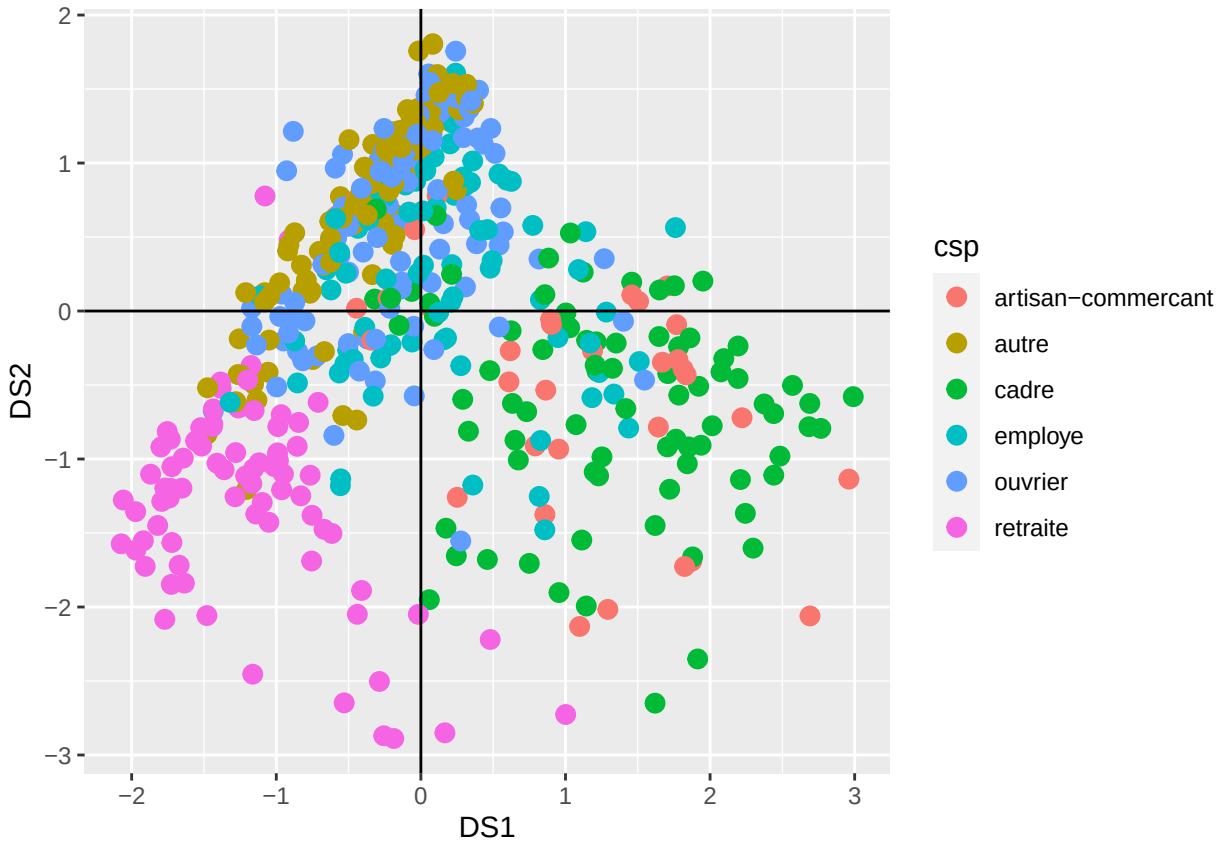
```
banque.dis$eig
```

```
## [1] 0.599093572 0.533430564 0.107382952 0.018786013 0.009446124
```

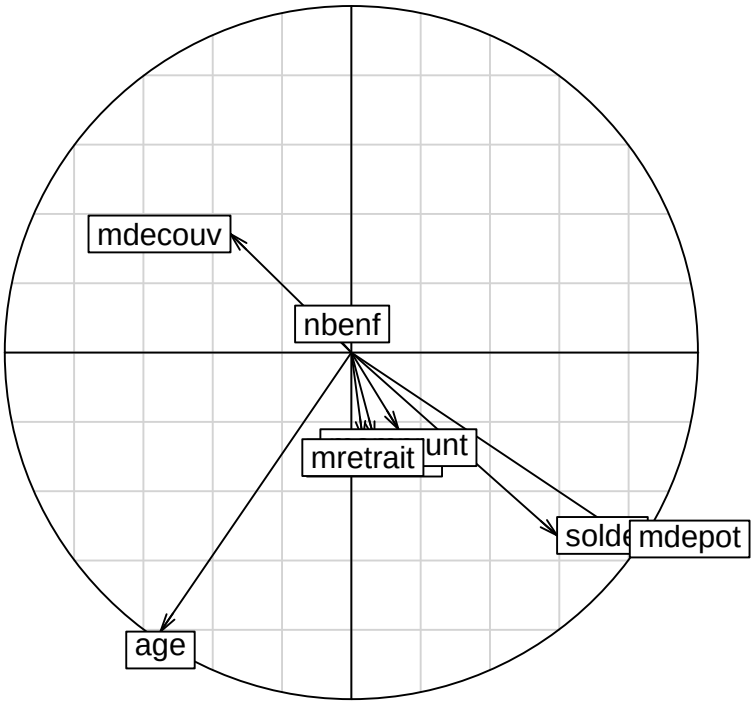
Les eux premiers axes sont discriminants.

Graphe des individus

```
coord <- banque.dis$li %>% mutate(csp=y)
ggplot(coord,aes(x=DS1,y=DS2,color =csp))+
  geom_point(size=3)+geom_hline(yintercept=0)+geom_vline(xintercept = 0)
```



Graphe des variables



Centre de gravité des 3 groupes

```
knitr::kable(aggregate(X, by=list(y),mean),digits=1)
```

Group.1	solde	mdecouv	ncompte	memprunt	mdepot	mretrait	nbenf	age
artisan-commercant	2436.7	326.3	2.2	9614.7	55551.6	3085.9	0.7	38.1
autre	430.5	835.8	1.6	2172.7	1087.5	1521.6	0.6	35.2
cadre	2988.7	107.0	3.0	10175.6	69903.2	3761.8	0.4	37.6
employe	1474.7	320.2	3.9	13211.3	17472.9	4916.7	0.5	37.1
ouvrier	940.9	537.0	2.3	6114.9	4907.6	3245.7	0.5	34.8
retraite	1369.3	468.0	3.1	9019.7	12527.8	4350.4	0.5	67.1

Conclusion:

- La variable `age` discrimine les retraités des autres
- les variables `solde` et `mdepot` discrimine les groupes des artisans et cadres des autres avec des valeurs plus élevées
- L'axe 1 sépare les groupes des cadres et des artisan-commerçant aux retraités
- l'axe 2 sépare ces 3 groupes des autres