TP ACP - Correction

Analyse des données

Master ISEFAR - M1

```
rm(list=ls())
library("knitr") #pour avoir un format table dans les sorties
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("FactoMineR") #pour effectuer l'ACP
library("factoextra") #pour extraire et visualiser les résultats issus de FactoMineR
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library("corrplot") #pour avoir une représentation des corrélations
```

```
## corrplot 0.84 loaded
```

```
library("MASS") #pour obtenir le jeu de données
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library("ppcor") #pour obtenir les corrélations partielles
```

1 Criminalités aux USA

1.1 Données et premières analyses

```
data(USArrests)
dim(USArrests)
```

```
## [1] 50  4
```

```
str(USArrests)
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
##  $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
##  $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
##  $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

```
kable(head(USArrests))
```

	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7

```
#Statistiques simples
summary(USArrests)
```

```
##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

```
USArrests %>% summarise_all(mean)
```

```
##   Murder Assault UrbanPop   Rape
## 1  7.788  170.76    65.54 21.232
```

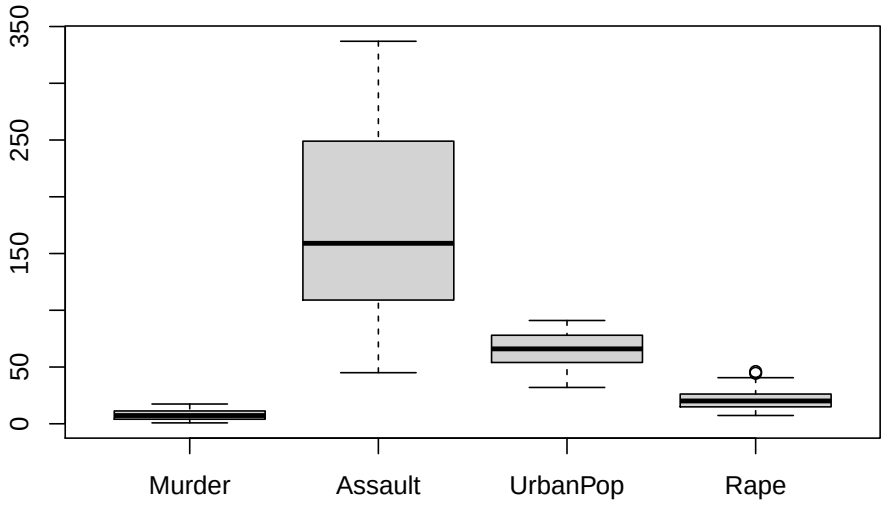
```
USArrests %>% summarise_all(var)
```

```
##     Murder  Assault UrbanPop     Rape
## 1 18.97047 6945.166 209.5188 87.72916
```

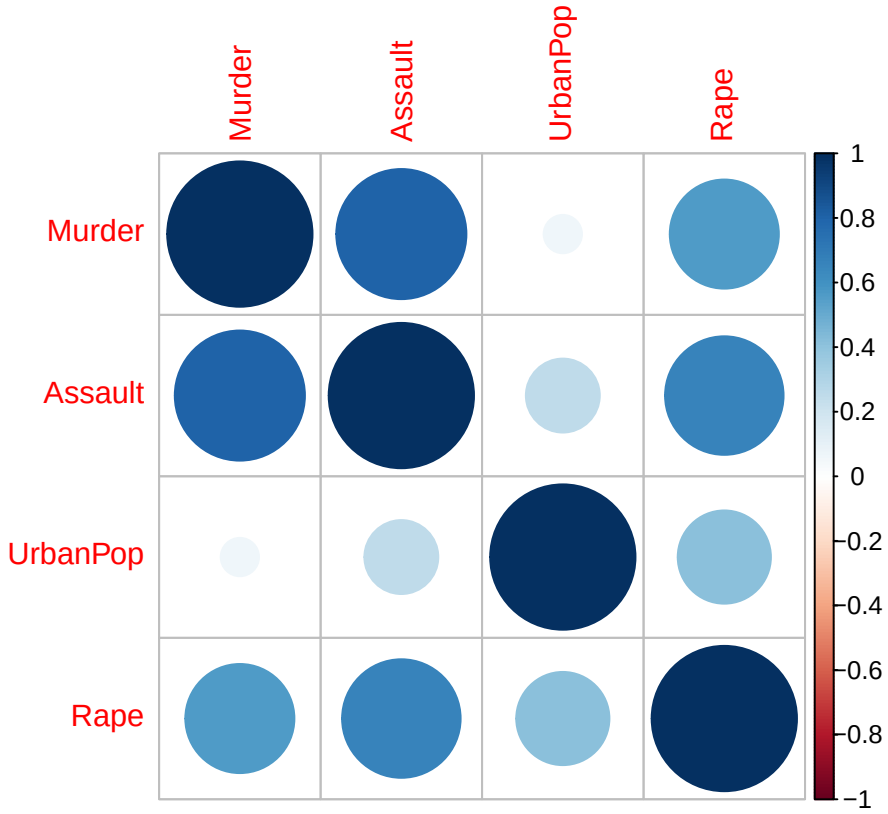
```r
correlation <- USArrests %>% cor()
kable(correlation,digits=3)
```

	Murder	Assault	UrbanPop	Rape
Murder	1.000	0.802	0.070	0.564
Assault	0.802	1.000	0.259	0.665
UrbanPop	0.070	0.259	1.000	0.411
Rape	0.564	0.665	0.411	1.000

```r
# Distribution des variables
USArrests %>% boxplot()
```




```r
#Corrélations
correlation %>% corrplot
```



1.2 ACP normée

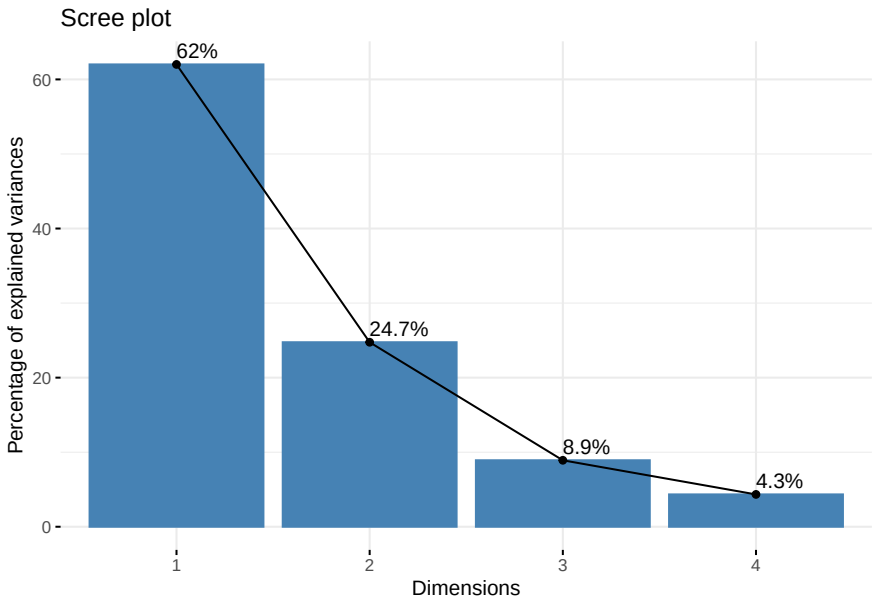
```
res.pca=PCA(USArrests,scale.unit = TRUE,ncp = 4,graph=FALSE)
```

1.2.1 Valeurs propres et choix du nombre d'axes

```
kable(res.pca$eig)
```

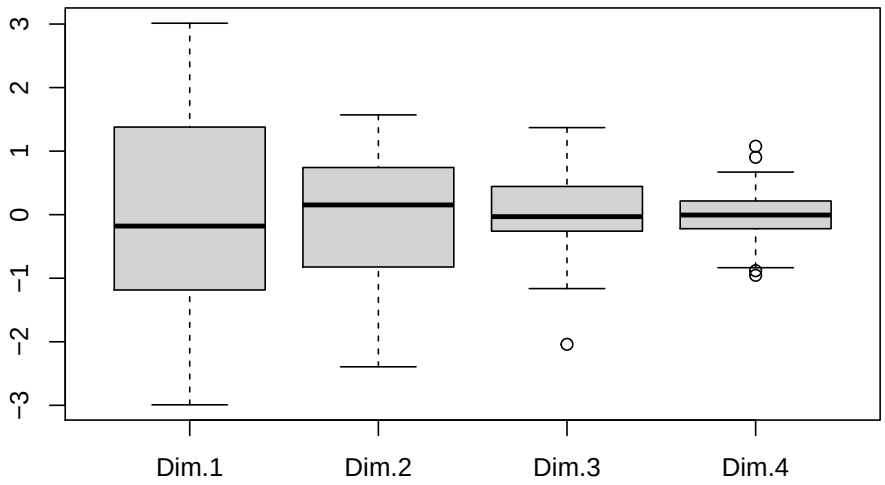
	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	2.4802416	62.006039	62.00604
comp 2	0.9897652	24.744129	86.75017
comp 3	0.3565632	8.914079	95.66425
comp 4	0.1734301	4.335752	100.00000

```
# représentation des valeurs propres
fviz_eig(res.pca, addlabels = TRUE)
```



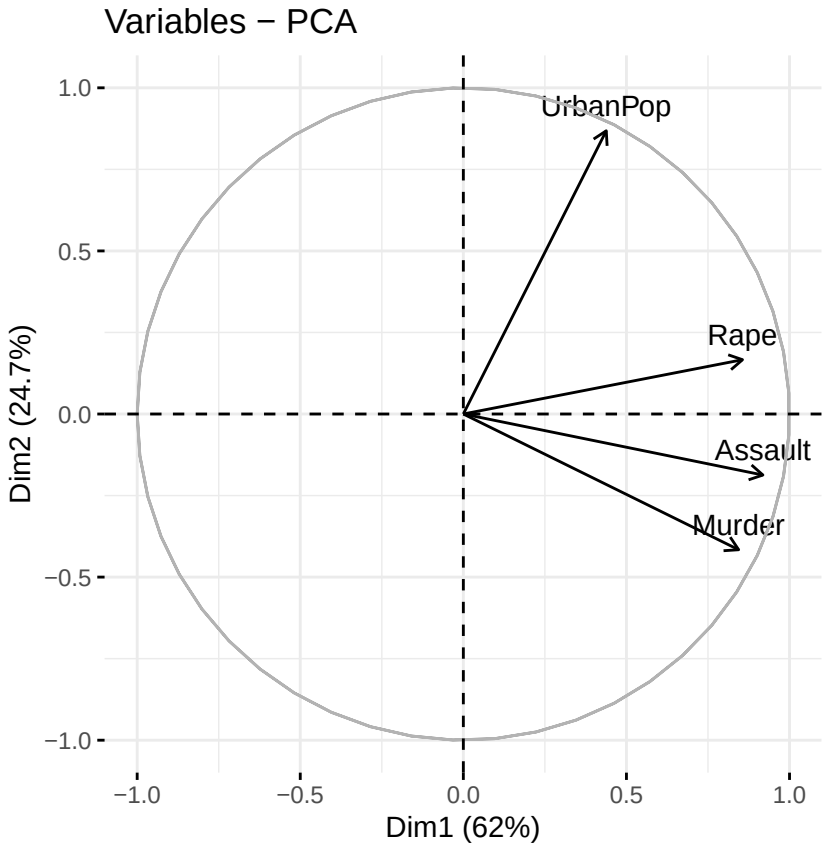
1.2.2 Distribution des nouvelles variables

```
boxplot(res.pca$ind$coord)
```



1.2.3 Variables: cercle des corrélations

```
fviz_pca_var(res.pca,axes=c(1,2))
```



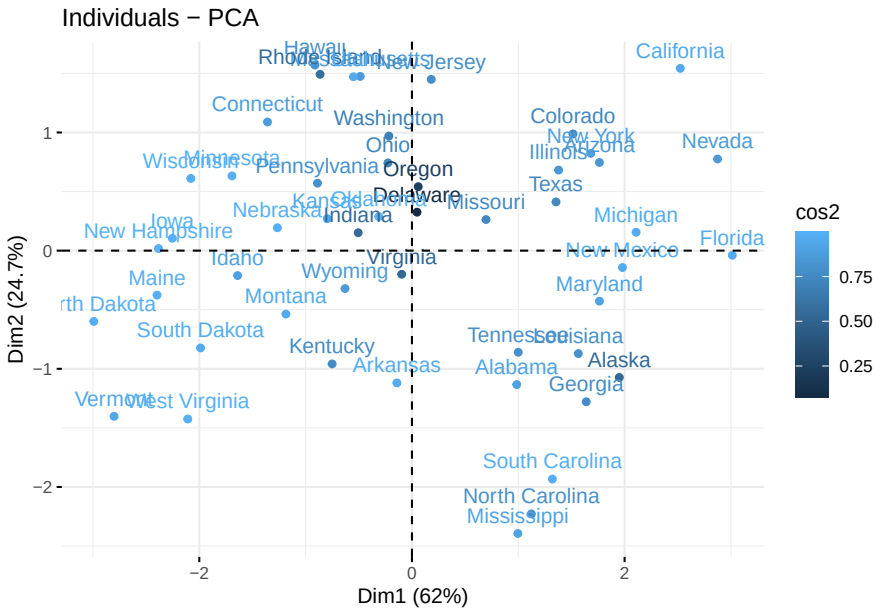
1.2.4 Individus: graphe du nuage de points

```
#Qualité de la représentation des individus sur le plan principal
cos2 = rowSums(res.pca$ind$cos2[, 1:2])
cos2
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##     0.91048430     0.53227351     0.83970870     0.96475028     0.94849191
##       Colorado    Connecticut       Delaware        Florida        Georgia
##     0.73165027     0.87633617     0.07710394     0.96370152     0.77199362
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##     0.80139578     0.89619805     0.83314875     0.54513636     0.99464747
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##     0.94237506     0.76694959     0.79627187     0.98108029     0.90707462
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##     0.85653573     0.96573155     0.99150833     0.91866630     0.74147659
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##     0.95690665     0.98139589     0.85921251     0.99956798     0.76828519
##     New Mexico       New York North Carolina   North Dakota           Ohio
##     0.96356269     0.89452258     0.78984028     0.98359912     0.72742723
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##     0.99810905     0.23989841     0.79398168     0.56878843     0.98075009
##   South Dakota      Tennessee          Texas           Utah        Vermont
##     0.96587407     0.79028567     0.75366622     0.96367294     0.93088171
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
```

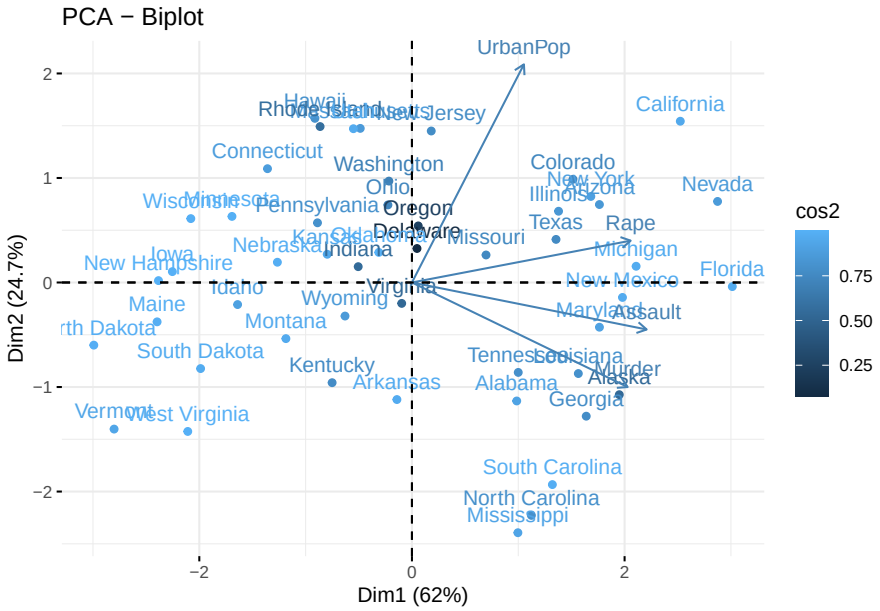
```
##       0.52319302       0.69228695       0.99563736       0.98880950       0.85350110
```

```
#Représentation du nuage des individus sur le plan principal
fviz_pca_ind (res.pca, col.ind = "cos2",axes=c(1,2)) #dégradé de couleur selon la représentativité
```



1.2.5 Biplot

```
fviz_pca_biplot(res.pca,axes=c(1,2), col.ind = "cos2")
```

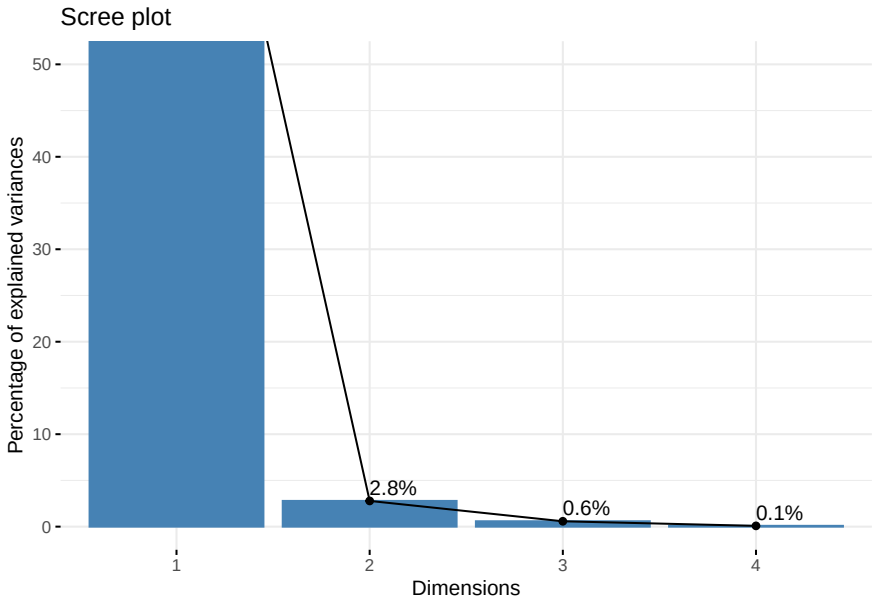


1.3 ACP simple

```
res.pca.simple=PCA(USArrests,scale.unit = FALSE,ncp = 4,graph=FALSE)
```

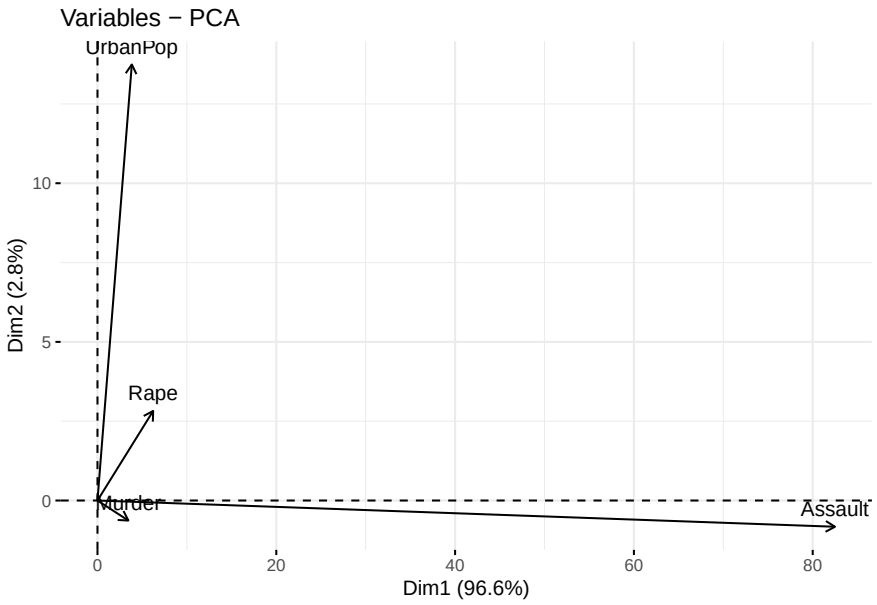
1.3.1 Valeurs propres et choix du nombre d'axes

```
fviz_eig(res.pca.simple, addlabels = TRUE, ylim = c(0, 50))
```



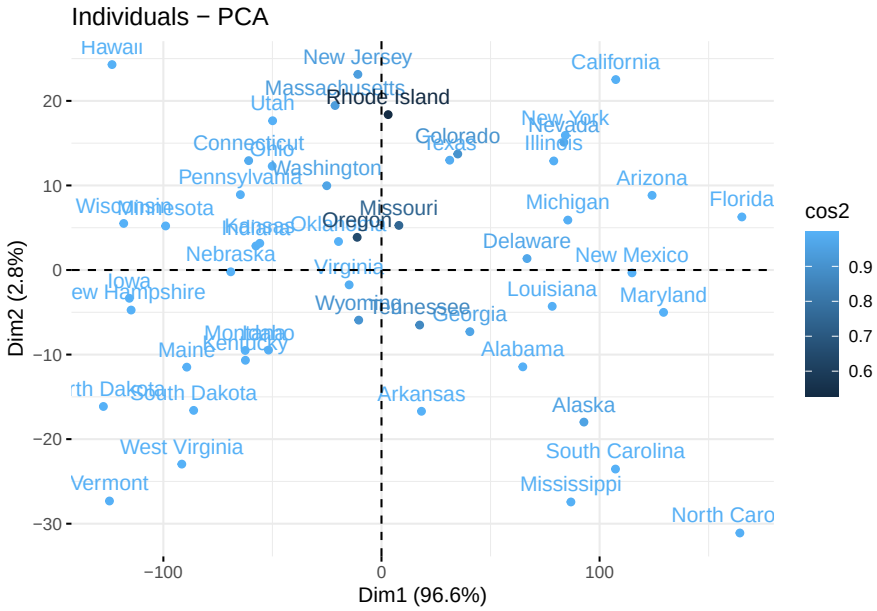
1.3.2 Variables: cercle des corrélations

```
fviz_pca_var(res.pca.simple,axes=c(1,2))
```



1.3.3 Individus: graphe du nuage de points

```
fviz_pca_ind (res.pca.simple, col.ind = "cos2",axes=c(1,2))
```



2 Fertilité et indicateurs socio-économiques en Suisse

2.1 Données et premières analyses

```r
#Jeu de données
data(swiss)
head(kable(head(swiss)))
```

```
## [1] "|             | Fertility| Agriculture| Examination| Education| Catholic| Infant.Mortality|"
## [2] "|:------------|---------:|-----------:|-----------:|---------:|--------:|----------------:|"
## [3] "|Courtelary   |      80.2|        17.0|          15|        12|     9.96|             22.2|"
## [4] "|Delemont     |      83.1|        45.1|           6|         9|    84.84|             22.2|"
## [5] "|Franches-Mnt |      92.5|        39.7|           5|         5|    93.40|             20.2|"
## [6] "|Moutier      |      85.8|        36.5|          12|         7|    33.77|             20.3|"
```

```r
dim(swiss)
```

```
## [1] 47  6
```

```r
str(swiss)
```

```
## 'data.frame':    47 obs. of  6 variables:
##  $ Fertility       : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...
##  $ Agriculture     : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...
##  $ Examination     : int  15 6 5 12 17 9 16 14 12 16 ...
##  $ Education       : int  12 9 5 7 15 7 7 8 7 13 ...
##  $ Catholic        : num  9.96 84.84 93.4 33.77 5.16 ...
##  $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...
```

```r
# Statistiques simples
summary(swiss)
```

```
##    Fertility      Agriculture     Examination      Education
##  Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
##  1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
##  Median :70.40   Median :54.10   Median :16.00   Median : 8.00
##  Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98
##  3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
##  Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
##     Catholic       Infant.Mortality
##  Min.   :  2.150   Min.   :10.80
##  1st Qu.:  5.195   1st Qu.:18.15
##  Median : 15.140   Median :20.00
##  Mean   : 41.144   Mean   :19.94
##  3rd Qu.: 93.125   3rd Qu.:21.70
##  Max.   :100.000   Max.   :26.60
```

```r
swiss %>% summarise_all(mean)
```

```
##   Fertility Agriculture Examination Education Catholic Infant.Mortality
## 1  70.14255    50.65957    16.48936  10.97872 41.14383         19.94255
```

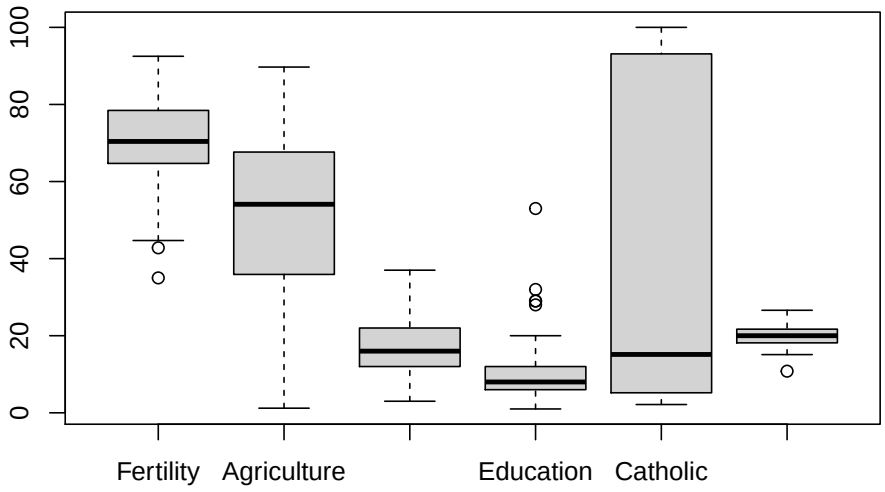
```r
swiss %>% summarise_all(var)
```

```
##   Fertility Agriculture Examination Education Catholic Infant.Mortality
## 1  156.0425    515.7994    63.64662  92.45606 1739.295         8.483802
```

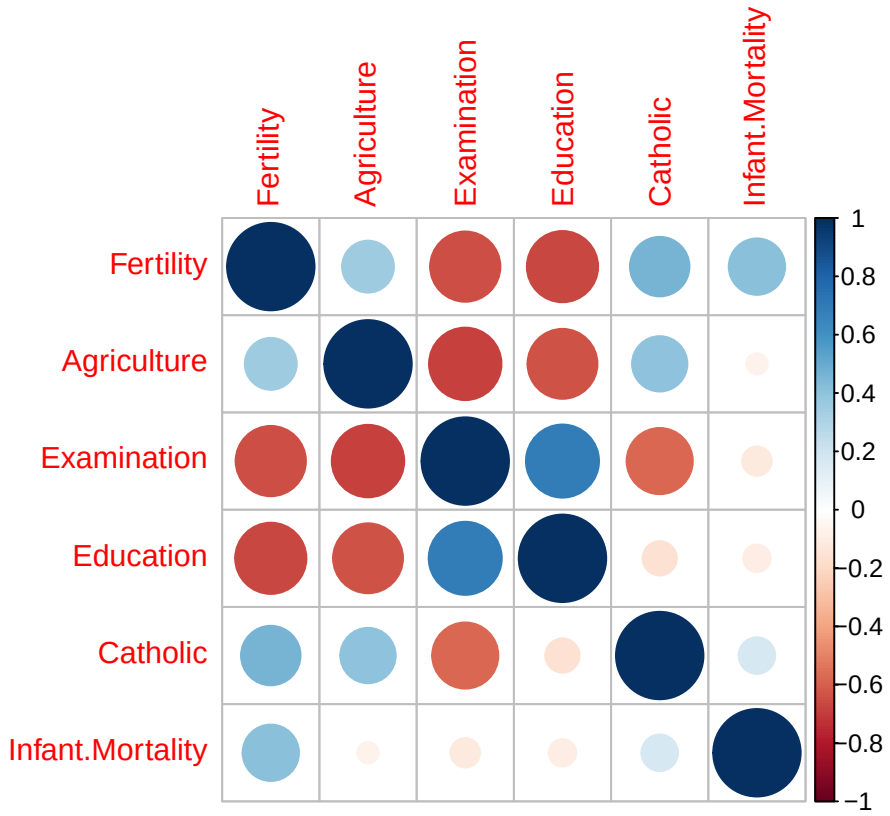
```r
correlation <- swiss %>% cor()
kable(correlation,digits=3)
```

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Fertility	1.000	0.353	-0.646	-0.664	0.464	0.417
Agriculture	0.353	1.000	-0.687	-0.640	0.401	-0.061
Examination	-0.646	-0.687	1.000	0.698	-0.573	-0.114
Education	-0.664	-0.640	0.698	1.000	-0.154	-0.099
Catholic	0.464	0.401	-0.573	-0.154	1.000	0.175
Infant.Mortality	0.417	-0.061	-0.114	-0.099	0.175	1.000

```r
# Distribution des variables
swiss %>% boxplot()
```




```r
#Corrélations
correlation %>% corrplot
```

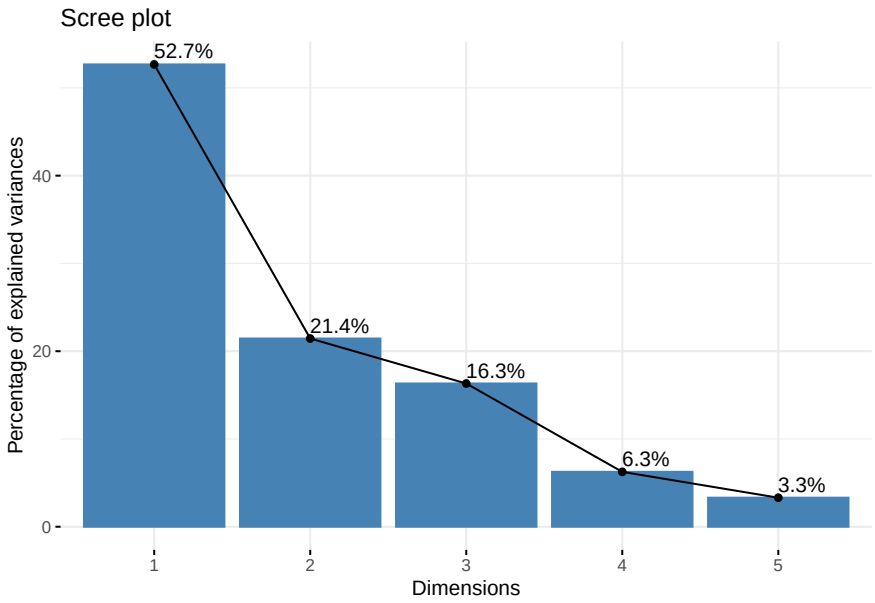


2.2 ACP normée sans la variable fertilité

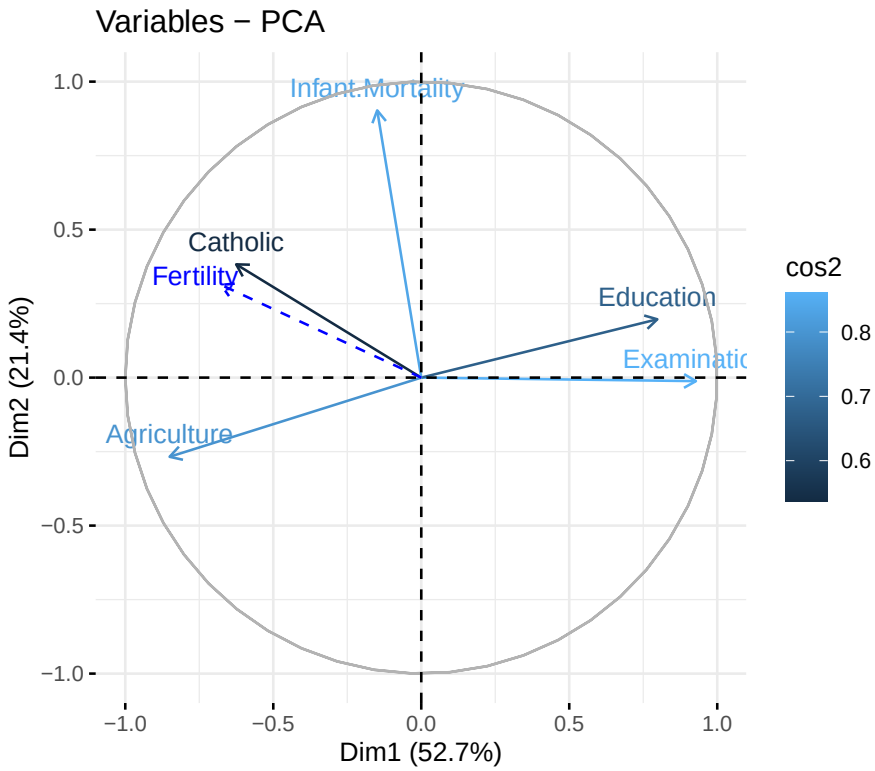
```
res.pca=PCA(swiss,scale.unit = TRUE,ncp = 5,graph=FALSE,quanti.sup = 1)

# Valeurs propres
kable(res.pca$eig)
```

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	2.6335008	52.670015	52.67002
comp 2	1.0722340	21.444681	74.11470
comp 3	0.8160316	16.320632	90.43533
comp 4	0.3127902	6.255805	96.69113
comp 5	0.1654433	3.308867	100.00000

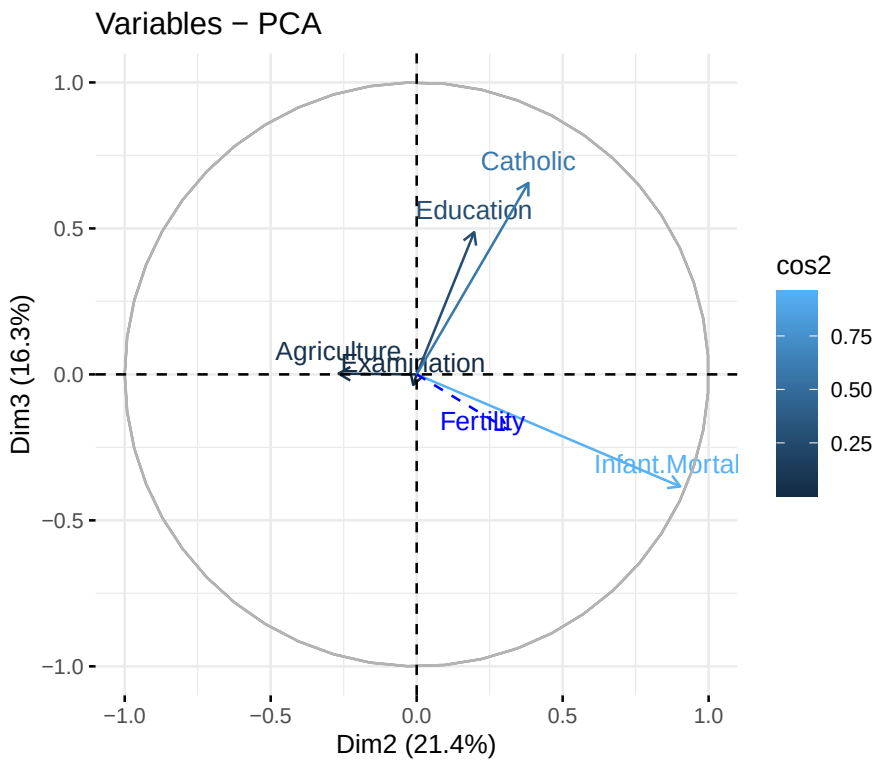
```
fviz_eig(res.pca, addlabels = TRUE)
```




```
# Variables
fviz_pca_var(res.pca,axes=c(1,2),col.var="cos2")
```

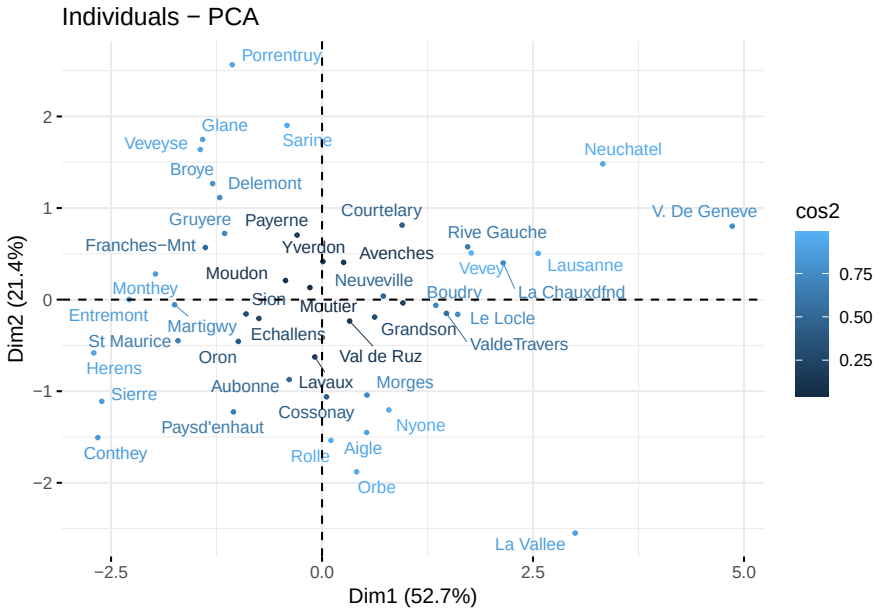


```
fviz_pca_var(res.pca,axes=c(2,3),col.var="cos2")
```




```
# Individus
fviz_pca_ind (res.pca,axes=c(1,2),col.ind = "cos2", repel = TRUE,pointsize = 0.7, labelsize = 3)
```

```
## Warning: ggrepel: 1 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



```r
#Qualité de la représentation des individus sur le plan principal
cos2 = rowSums(res.pca$ind$cos2[, 1:2])
cos2
```

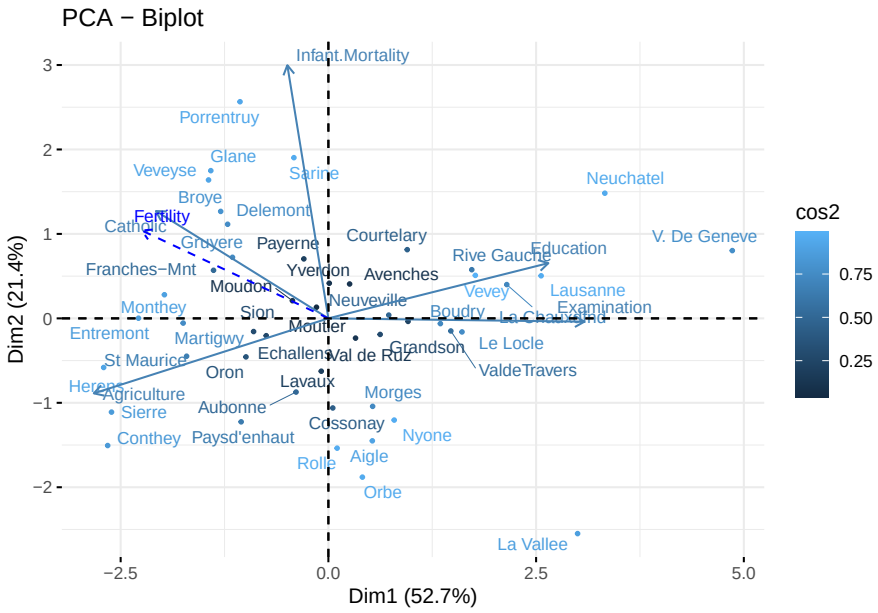
```
##   Courtelary     Delemont Franches-Mnt      Moutier   Neuveville   Porrentruy
##   0.44949346   0.75185513   0.51183875   0.04017159   0.48291837   0.92730492
##        Broye        Glane      Gruyere       Sarine      Veveyse        Aigle
##   0.79751413   0.90642932   0.73462198   0.95204009   0.91733899   0.90120308
##      Aubonne     Avenches     Cossonay    Echallens     Grandson     Lausanne
##   0.50453077   0.11432969   0.43098306   0.26850494   0.28425170   0.96016173
##    La Vallee       Lavaux       Morges       Moudon        Nyone         Orbe
##   0.87381662   0.19992839   0.73037591   0.09774078   0.96120734   0.95107287
##         Oron      Payerne Paysd'enhaut        Rolle        Vevey      Yverdon
##   0.36374083   0.20400551   0.63104766   0.96016397   0.99233920   0.10520609
##      Conthey    Entremont       Herens     Martigwy      Monthey   St Maurice
##   0.83900136   0.86229476   0.91202462   0.74570876   0.88794514   0.65812116
##       Sierre        Sion       Boudry La Chauxdfnd     Le Locle    Neuchatel
##   0.84651256   0.30035312   0.72333599   0.71778856   0.75196527   0.94630996
##   Val de Ruz ValdeTravers V. De Geneve  Rive Droite  Rive Gauche
##   0.12271272   0.54801511   0.76900899   0.22789629   0.61940284
```

```r
#Contribution des individus sur le plan principal
contrib = rowSums(res.pca$ind$contrib[, 1:2])
contrib
```

```
##   Courtelary     Delemont Franches-Mnt      Moutier   Neuveville   Porrentruy
##   2.04082746   3.65067218   2.18567831   0.05105712   0.42981309  13.97102753
##        Broye        Glane      Gruyere       Sarine      Veveyse        Aigle
##   4.53907710   7.68630425   2.11274988   7.32306864   7.00885152   4.40112509
##      Aubonne     Avenches     Cossonay    Echallens     Grandson     Lausanne
##   1.63365832   0.38136434   2.23693750   0.53556182   0.38566967   5.80076362
##    La Vallee       Lavaux       Morges       Moudon        Nyone         Orbe
##  20.15922646   0.78303213   2.38461130   0.23731404   3.38613887   7.14917557
##         Oron      Payerne Paysd'enhaut        Rolle        Vevey      Yverdon
##   1.20773382   1.05406302   3.86990603   4.69737109   3.04198982   0.34336416
##      Conthey    Entremont       Herens     Martigwy      Monthey   St Maurice
##  10.19988742   4.21985933   6.57946733   2.47325480   3.30046452   2.75110072
##       Sierre        Sion       Boudry La Chauxdfnd     Le Locle    Neuchatel
##   7.94862205   0.70278473   1.47611036   4.04186245   2.14091644  13.30216211
##   Val de Ruz ValdeTravers V. De Geneve  Rive Droite  Rive Gauche
##   0.19497803   1.79942862  20.37068777   0.74574260   3.06453698
```

```r
# Les deux
fviz_pca_biplot (res.pca,axes=c(1,2),col.ind = "cos2", repel = TRUE,pointsize = 0.7, labelsize = 3)
```

```
## Warning: ggrepel: 1 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



3 Crabes

3.1 Données et premières analyses

```
data(crabs)
kable(head(crabs))
```

sp	sex	index	FL	RW	CL	CW	BD
B	M	1	8.1	6.7	16.1	19.0	7.0
B	M	2	8.8	7.7	18.1	20.8	7.4
B	M	3	9.2	7.8	19.0	22.4	7.7
B	M	4	9.6	7.9	20.1	23.1	8.2
B	M	5	9.8	8.0	20.3	23.0	8.2
B	M	6	10.8	9.0	23.0	26.5	9.8

```
crabs <- crabs %>% dplyr::select(-index)
kable(head(crabs))
```

sp	sex	FL	RW	CL	CW	BD
B	M	8.1	6.7	16.1	19.0	7.0
B	M	8.8	7.7	18.1	20.8	7.4
B	M	9.2	7.8	19.0	22.4	7.7
B	M	9.6	7.9	20.1	23.1	8.2
B	M	9.8	8.0	20.3	23.0	8.2
B	M	10.8	9.0	23.0	26.5	9.8

```r
dim(crabs)
```

```
## [1] 200   7
```

```r
str(crabs)
```

```
## 'data.frame':    200 obs. of  7 variables:
##  $ sp : Factor w/ 2 levels "B","O": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex: Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ FL : num  8.1 8.8 9.2 9.6 9.8 10.8 11.1 11.6 11.8 11.8 ...
##  $ RW : num  6.7 7.7 7.8 7.9 8 9 9.9 9.1 9.6 10.5 ...
##  $ CL : num  16.1 18.1 19 20.1 20.3 23 23.8 24.5 24.2 25.2 ...
##  $ CW : num  19 20.8 22.4 23.1 23 26.5 27.1 28.4 27.8 29.3 ...
##  $ BD : num  7 7.4 7.7 8.2 8.2 9.8 9.8 10.4 9.7 10.3 ...
```

```r
#Statistiques simples
summary(crabs)
```

```
##  sp      sex           FL              RW              CL
##  B:100   F:100   Min.   : 7.20   Min.   : 6.50   Min.   :14.70
##  O:100   M:100   1st Qu.:12.90   1st Qu.:11.00   1st Qu.:27.27
##                  Median :15.55   Median :12.80   Median :32.10
##                  Mean   :15.58   Mean   :12.74   Mean   :32.11
##                  3rd Qu.:18.05   3rd Qu.:14.30   3rd Qu.:37.23
##                  Max.   :23.10   Max.   :20.20   Max.   :47.60
##        CW              BD
##  Min.   :17.10   Min.   : 6.10
##  1st Qu.:31.50   1st Qu.:11.40
##  Median :36.80   Median :13.90
##  Mean   :36.41   Mean   :14.03
##  3rd Qu.:42.00   3rd Qu.:16.60
##  Max.   :54.60   Max.   :21.60
```

```r
crabs %>% select_if(is.numeric) %>% summarise_all(mean)
```

```
##       FL      RW      CL      CW      BD
## 1 15.583 12.7385 32.1055 36.4145 14.0305
```

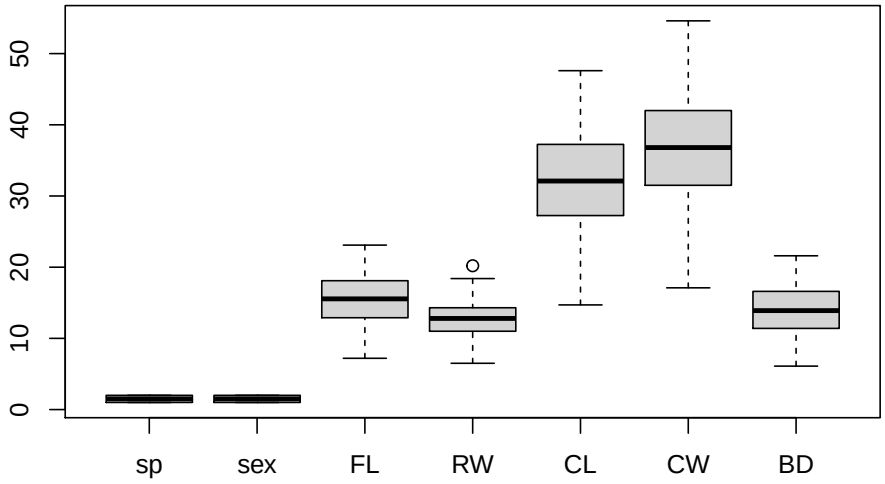
```r
crabs %>% select_if(is.numeric) %>% summarise_all(var)
```

```
##        FL       RW       CL       CW       BD
## 1 12.2173 6.622078 50.67992 61.96768 11.72907
```

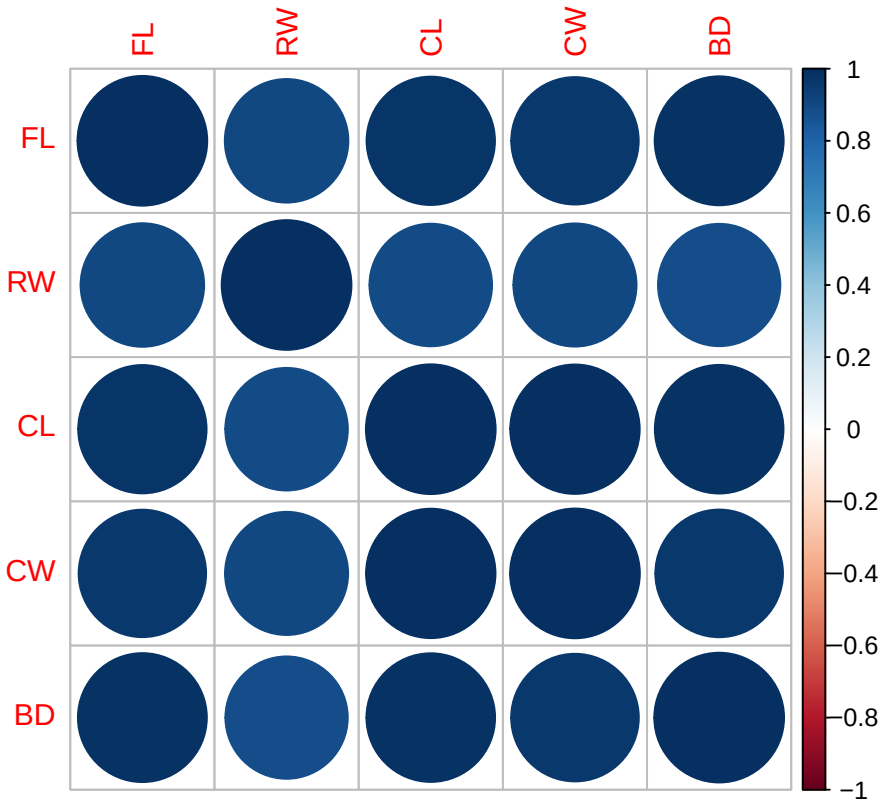
```r
correlation <- crabs %>% select_if(is.numeric) %>% cor()
kable(correlation,digits=2)
```

	FL	RW	CL	CW	BD
FL	1.00	0.91	0.98	0.96	0.99
RW	0.91	1.00	0.89	0.90	0.89
CL	0.98	0.89	1.00	1.00	0.98
CW	0.96	0.90	1.00	1.00	0.97
BD	0.99	0.89	0.98	0.97	1.00

```
# Distribution des variables
crabs %>% boxplot()
```




```
#Corrélations
correlation %>% corrplot
```



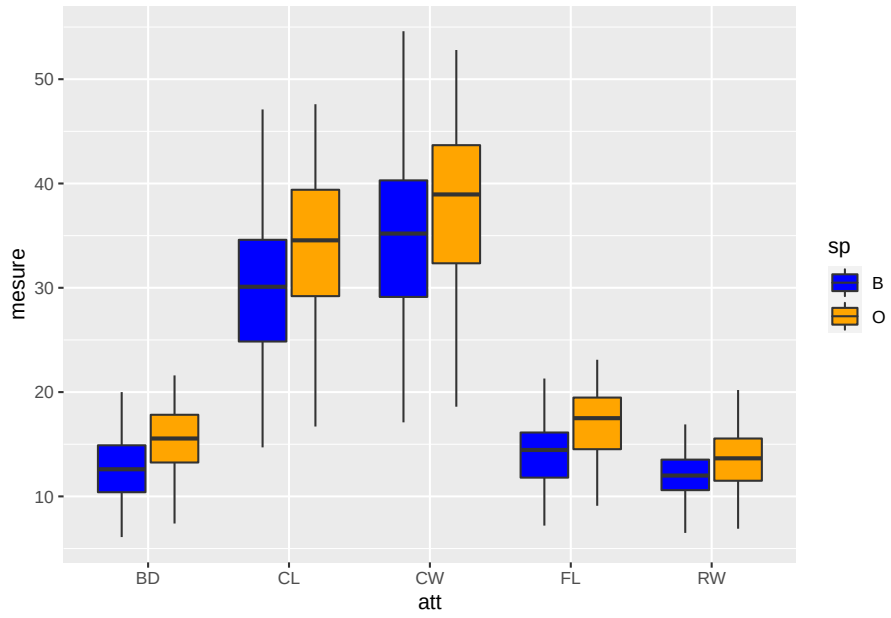
3.2 Distribution des variables en fonction des variables qualitatives

On réorganise le jeu de données

```
crabs.G <- crabs  %>% group_by(sp,sex) %>% gather(key='att',value='mesure',-sp,-sex)
```

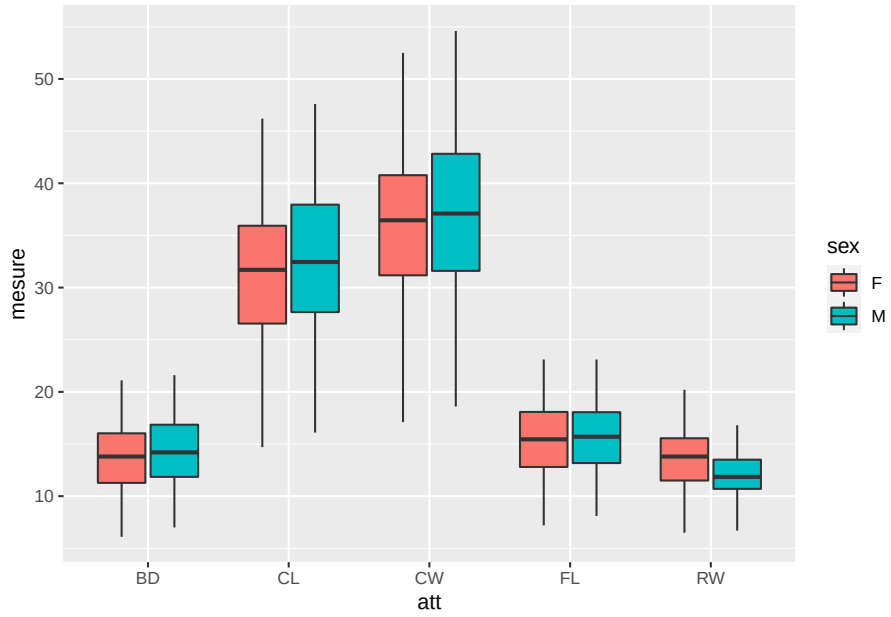
Distribution des variables par espèce

```
ggplot(crabs.G,aes(x=att , y=mesure, fill=sp))+
  geom_boxplot() +
  scale_fill_manual(values=c("blue","orange"))
```



Distribution des variables par sex

```
ggplot(crabs.G,aes(x=att , y=mesure, fill=sex))+geom_boxplot()
```



3.3 ACP

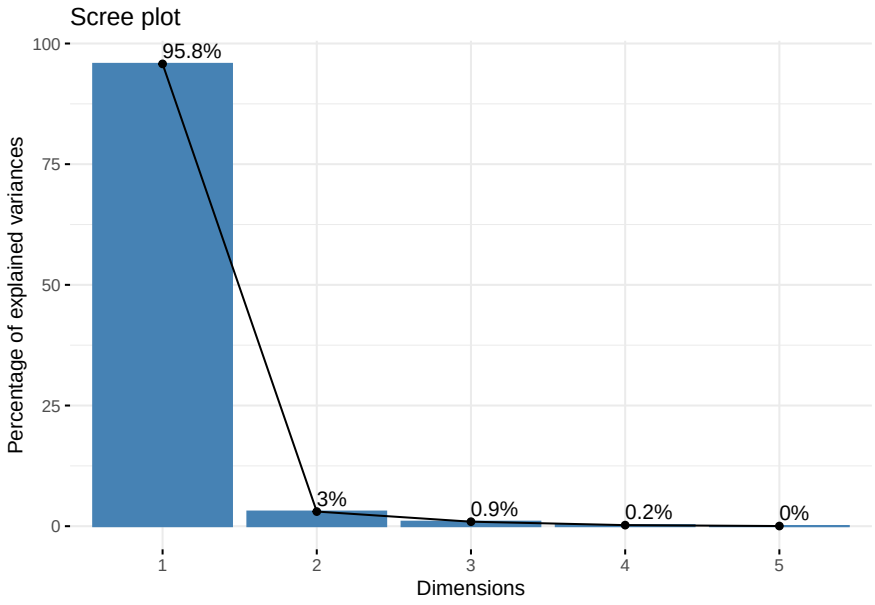
```
res.pca=PCA(crabs,scale.unit = TRUE,ncp = 5,quali.sup = 1:2,graph=FALSE)
# les variables supplémentaires sont intégrées au graphe mais ne sont pas
# prises en compte pour l'ACP
```

3.3.1 Valeurs propres et choix du nombre d'axes

```
kable(res.pca$eig)
```

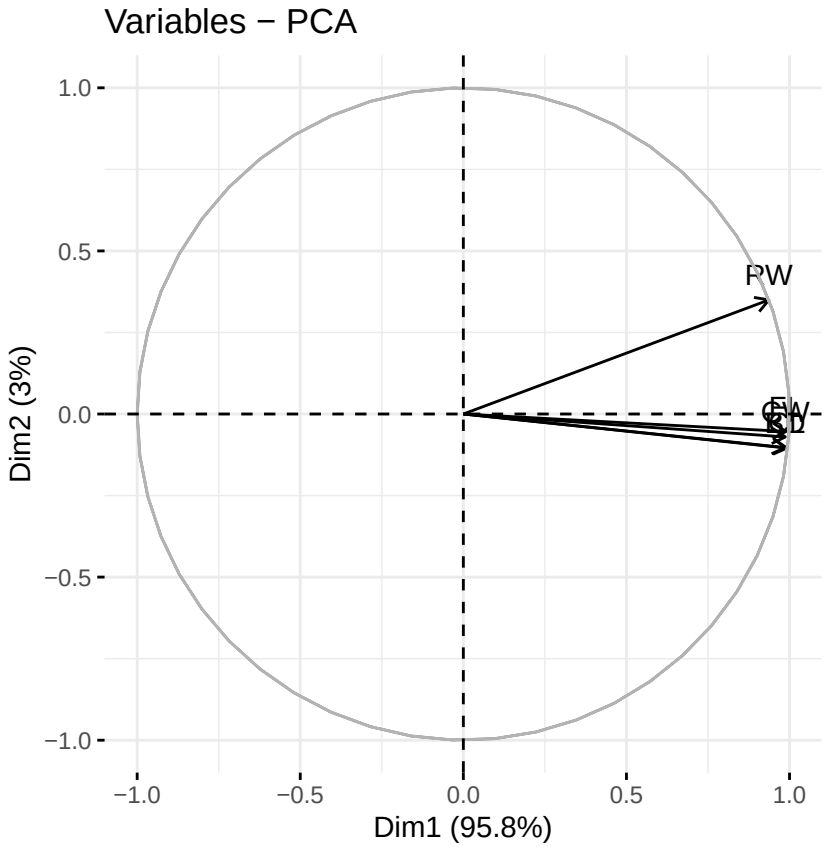
	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.7888348	95.7766957	95.77670
comp 2	0.1516852	3.0337041	98.81040
comp 3	0.0466330	0.9326595	99.74306
comp 4	0.0111354	0.2227071	99.96577
comp 5	0.0017117	0.0342336	100.00000

```
fviz_eig(res.pca, addlabels = TRUE)
```

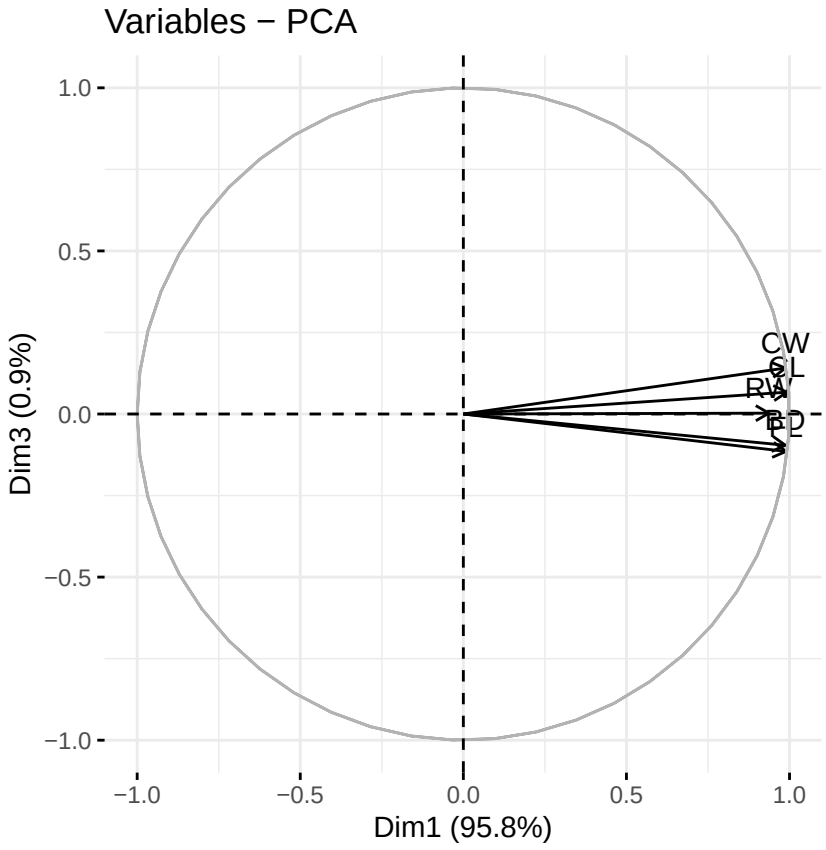


3.3.2 Variables: cercle des corrélations

```
fviz_pca_var(res.pca,axes=c(1,2))
```

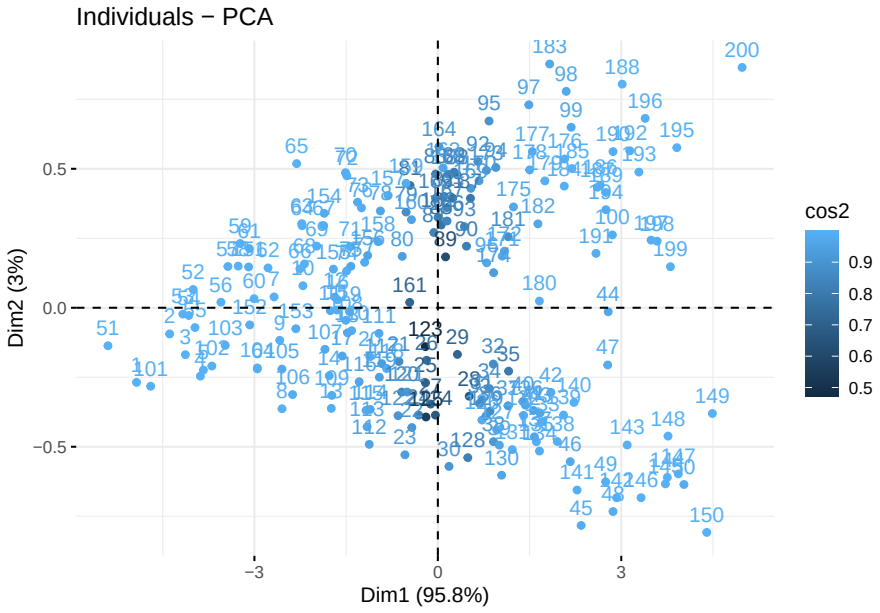



```
fviz_pca_var(res.pca,axes=c(1,3))
```

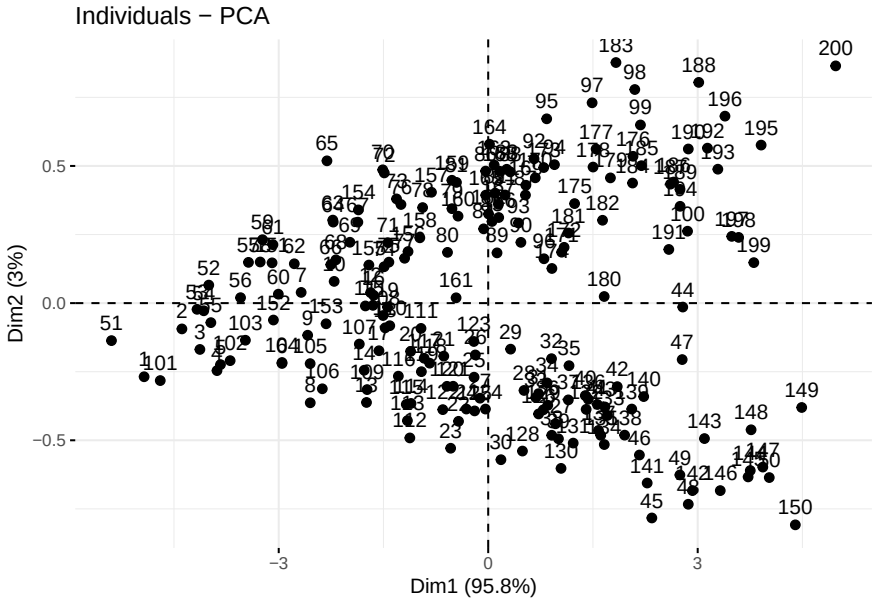


3.3.3 Individus: graphe du nuage de points

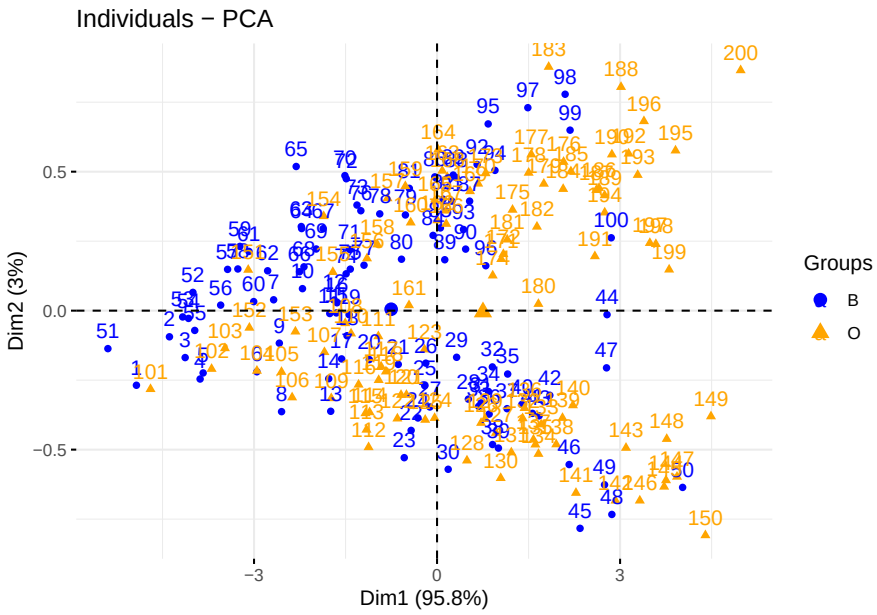
```
fviz_pca_ind (res.pca, col.ind = "cos2",axes=c(1,2)) #dégradé de couleur selon la représentativité
```



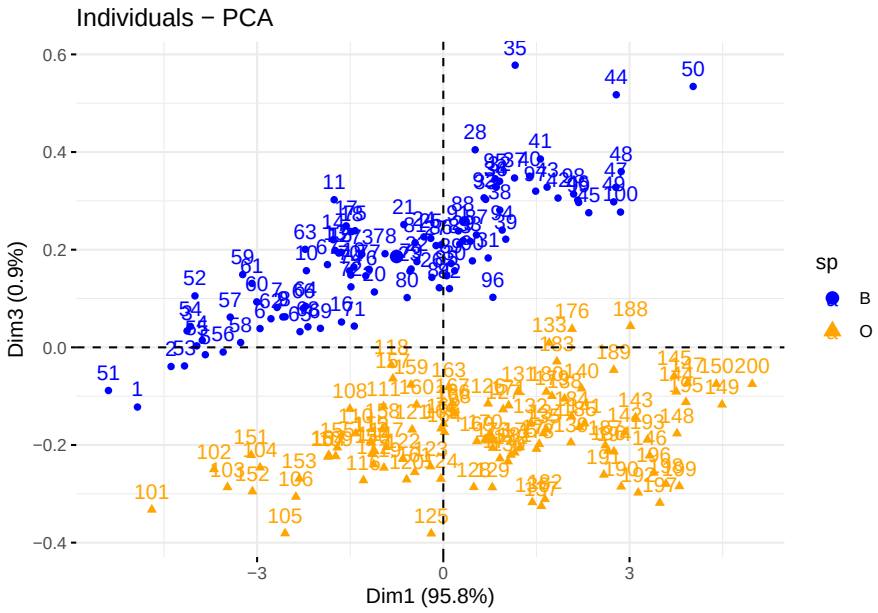
```
fviz_pca_ind (res.pca, ghabillage = "sex",pointsize = 2,axes=c(1,2)) #couleur selon le sexe
```




```
fviz_pca_ind (res.pca, habillage = crabs$sp,axes=c(1,2),palette=c("blue","orange")) #couleur selon l'es
```

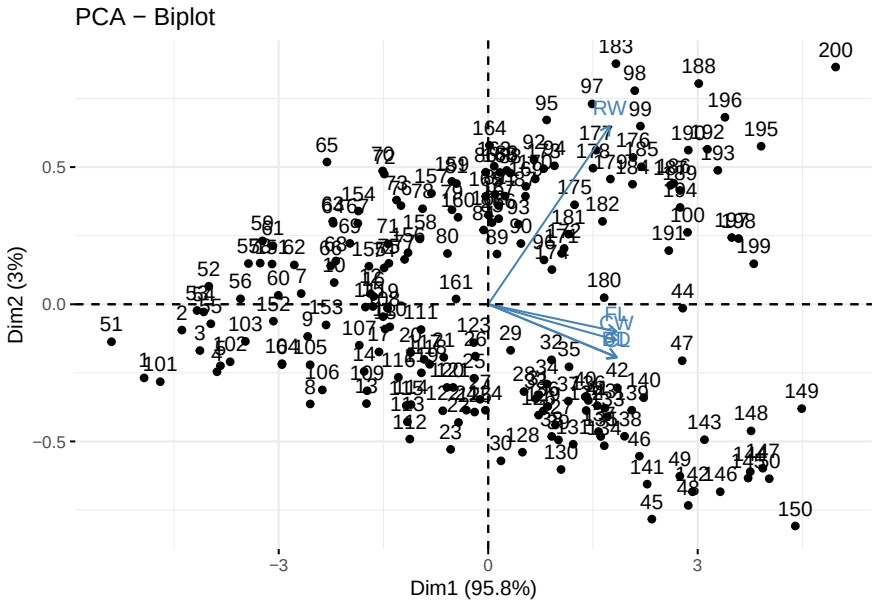



```
fviz_pca_ind (res.pca, habillage = "sp",axes=c(1,3),palette=c("blue","orange")) #couleur selon l'espèce
```

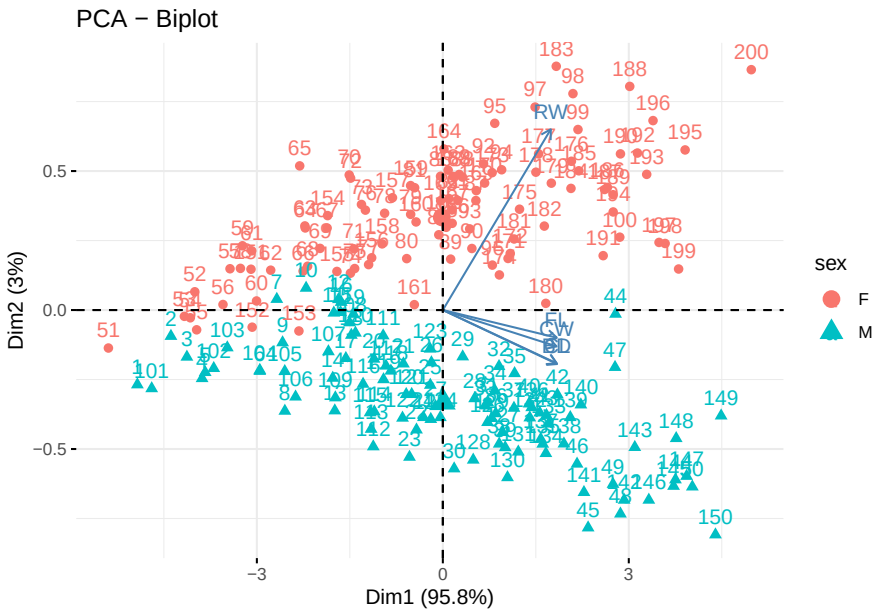


3.3.4 Biplot

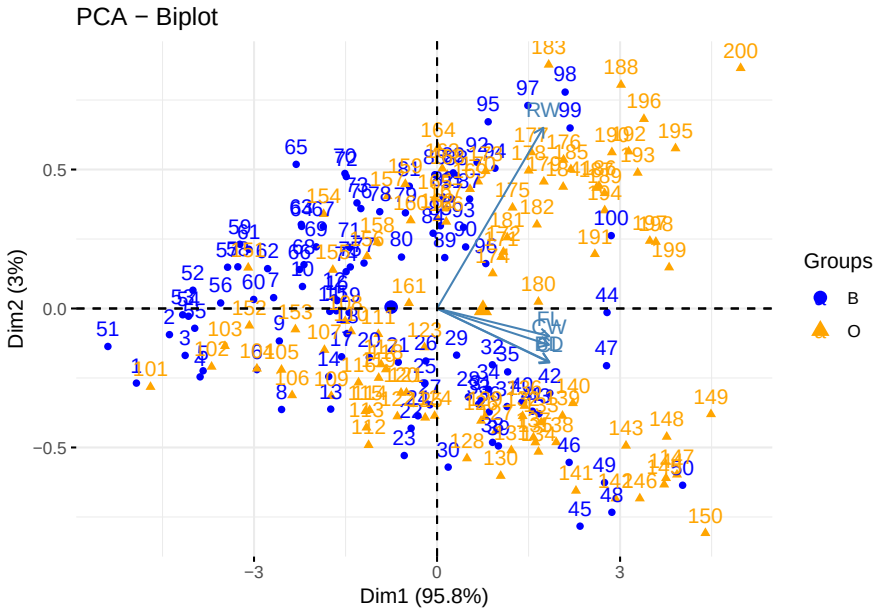
```
fviz_pca_biplot(res.pca,axes=c(1,2))
```



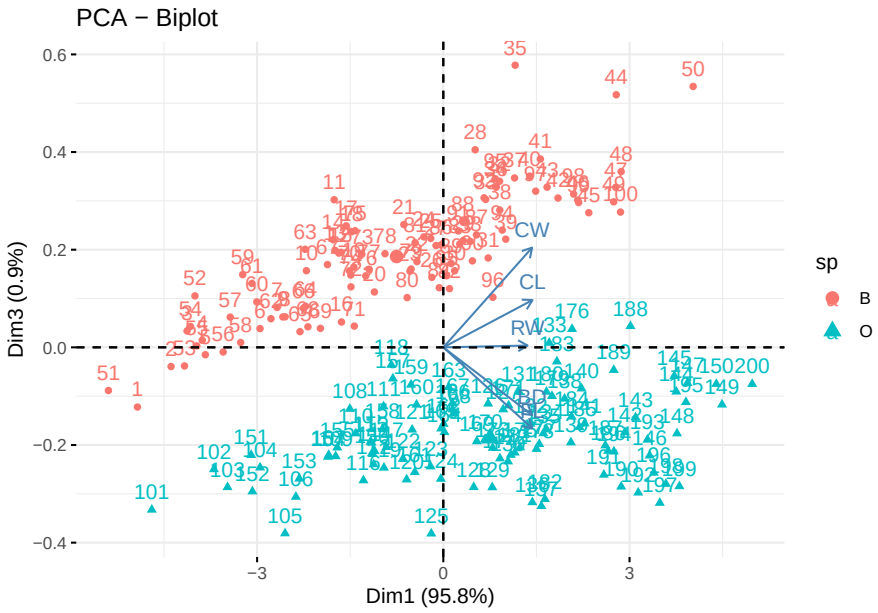
```
fviz_pca_biplot (res.pca, habillage = "sex",pointsize = 2,axes=c(1,2)) #couleur selon le sexe
```




```
fviz_pca_biplot (res.pca, habillage = crabs$sp,axes=c(1,2),palette=c("blue","orange")) #couleur selon l
```




```
fviz_pca_biplot (res.pca, habillage = "sp",axes=c(1,3),palette=c("blue","orange")) #couleur selon l'esp
```

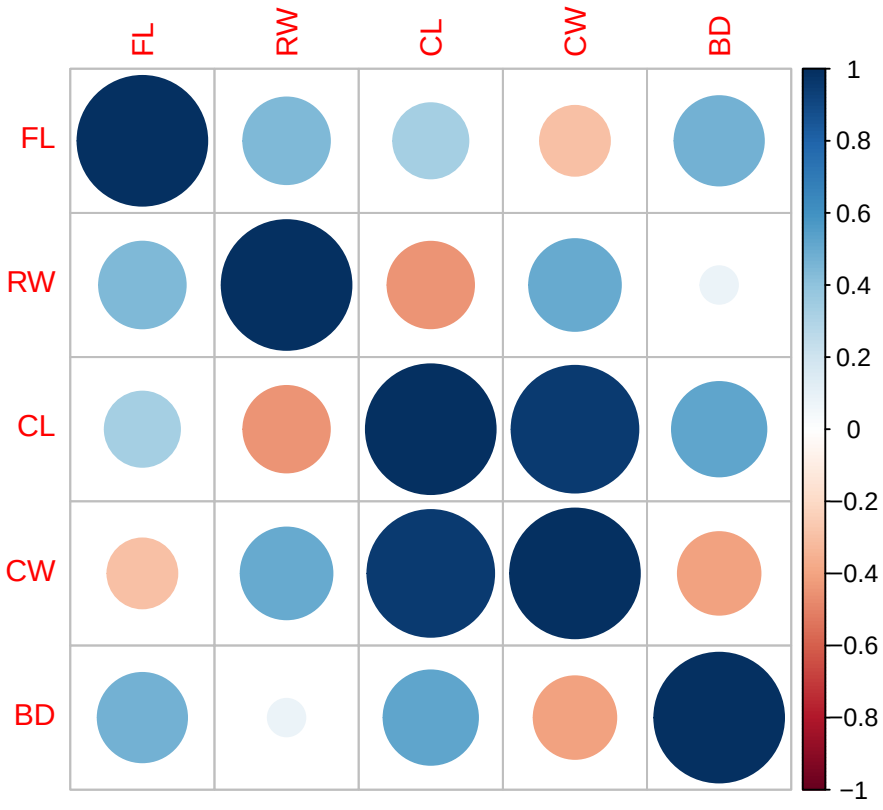


3.3.5 Etude des corrélations partielles

```
#Calcul des corrélations partielles
corrp <- crabs %>% select_if(is.numeric)%>% pcor()
kable(corrp$estimate,digits=2)
```

	FL	RW	CL	CW	BD
FL	1.00	0.45	0.34	-0.29	0.47
RW	0.45	1.00	-0.44	0.50	0.08
CL	0.34	-0.44	1.00	0.95	0.53
CW	-0.29	0.50	0.95	1.00	-0.41
BD	0.47	0.08	0.53	-0.41	1.00

```
#Visualisation
corrplot(corrp$estimate)
```


FL
RW
CL
CW
BD
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1