projet ACP

Salma Bouslama

4 mai 2017

I-INTRODUCTION

Dans ce projet, on va entamer ? une petite ?tude concernant un diagnostique de sant? comportant r?sultats diversifi?. Il est r?alis? sur un ?chantillon de patients de taille 80 personnes ayant de diff?rents ?ges.
Le but de ce projet est d'en tirer quelques interpr?tations utiles entre les diff?rentes r?sultats du diagnostique.

II-OBJECTIF

L'objectif de ce projet est d'?tudier la base ? her ? ainsi que de r?aliser son ACP: Voir la corr?lation du taux de cholest?rol ? chol ? avec les autres variables. Par exemple ? quoi est-il li?, d?pend t-il d'un ?ge pr?cis ?

III-NOM DE LA BASE DE DONNEE: "HER"

IV-DESCRIPTION

Les donn?es proviennent du D?partement am?ricain de la sant? et des services humains, Centre national des statistiques de sant?, troisi?me enqu?te nationale pour l'examen de la sant? et de l'alimentation (Health Exam Results en anglais, d'o? les initiales HER pour le dossier). Ces donn?es font partie des "datasets" utilis?s comme exemples dans l'ouvrageBiostatistics for the Biological and Health Sciences de Marc TRIOLA et Mario TRIOLA, que nous avons eu l'honneur de traduire pour les ?ditions Pearson.

variable:

1 / IDEN est un identificateur de ligne,
2 / SEXE est cod? 0 pour Homme et 1 pour Femme,
3 / AGE est en ann?es,
4 / TAILLE est la taille (cm),
5 / POIDS est le poids (kg),
6 / TTAILLE est le tour de taille (cm),
7 / POULS est le taux de battements (pulsations par minute),
8 / SYS est la pression sanguine systolique (mmHg),
9 / DIA est la pression sanguine diastolique (mmHg),
10 / CHOL est le taux de cholest?rol (mg),
11 / IMC est l'indice de masse corporelle (kg/mˆ2),
12 / JMBG est la longueur de la jambe gauche (cm),
13 / COUD est la largeur du coude (cm),
14 / POIGN est la largeur du poignet (cm),
15 / BRAS est la circonf?rence du bras (cm).
16 / classement des ages par intervalle :***"jeune" "adulte" "vieux"***

V-STATISTIQUE DESCRIPTIVE

1)importation de la base de donn?e:

```r
setwd("C:/Users/USER/Documents/R/salma-bouslama.github.io/projetacp")
her=read.csv("TP_projet1.csv",header=T,sep=";")
```

2)r?sum? statistique:

```r
str(her)
```

```
## 'data.frame':    80 obs. of  16 variables:
##  $ Num    : Factor w/ 80 levels "I0001","I0002",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ sexe   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age    : int  58 22 32 31 28 46 41 56 20 54 ...
##  $ taille : num  180 168 182 174 172 ...
##  $ poids  : num  76.7 65.4 81.3 79.7 69.2 75.7 61.2 91.4 79.5 63 ...
##  $ ttaille: num  90.6 78.1 96.5 87.7 87.1 ...
##  $ pouls  : int  68 64 88 72 64 72 60 88 76 60 ...
##  $ sys    : int  125 107 126 110 110 107 113 126 137 110 ...
##  $ dia    : int  78 54 81 68 66 83 71 72 85 71 ...
##  $ chol   : int  522 127 740 49 230 316 590 466 121 578 ...
##  $ imc    : num  23.8 23.2 24.6 26.2 23.5 24.5 21.5 31.4 26.4 22.7 ...
##  $ jmbg   : num  42.5 40.2 44.4 42.8 40 47.3 43.4 40.1 42.1 36 ...
##  $ coud   : num  7.7 7.6 7.3 7.5 7.1 7.1 6.5 7.5 7.5 6.9 ...
##  $ poign  : num  6.4 6.2 5.8 5.9 6 5.8 5.2 5.6 5.5 5.5 ...
##  $ bras   : num  31.9 31 32.7 33.4 30.1 30.5 27.6 38 32 29.3 ...
##  $ age1   : Factor w/ 4 levels "","adulte","jeune",..: 4 3 2 3 3 2 2 4 3 4 ...
```

```r
summary(her)
```

```
##       Num          sexe          age            taille
##  I0001  : 1   Min.   :0.0   Min.   :12.00   Min.   :144.8
##  I0002  : 1   1st Qu.:0.0   1st Qu.:23.75   1st Qu.:160.2
##  I0003  : 1   Median :0.5   Median :32.00   Median :168.0
##  I0004  : 1   Mean   :0.5   Mean   :34.35   Mean   :167.0
##  I0005  : 1   3rd Qu.:1.0   3rd Qu.:42.50   3rd Qu.:173.5
##  I0006  : 1   Max.   :1.0   Max.   :73.00   Max.   :193.5
##  (Other):74
##      poids           ttaille           pouls             sys
##  Min.   : 42.80   Min.   : 66.70   Min.   : 56.00   Min.   : 89.0
##  1st Qu.: 61.20   1st Qu.: 76.72   1st Qu.: 64.00   1st Qu.:107.0
##  Median : 73.00   Median : 87.70   Median : 72.00   Median :113.0
##  Mean   : 72.29   Mean   : 88.16   Mean   : 72.85   Mean   :114.8
##  3rd Qu.: 81.38   3rd Qu.: 97.33   3rd Qu.: 80.00   3rd Qu.:124.0
##  Max.   :116.10   Max.   :126.50   Max.   :124.00   Max.   :181.0
##
##       dia              chol             imc             jmbg
##  Min.   : 41.00   Min.   :   2.0   Min.   :17.70   Min.   :27.00
##  1st Qu.: 64.00   1st Qu.: 126.8   1st Qu.:22.52   1st Qu.:38.92
##  Median : 71.00   Median : 264.5   Median :25.35   Median :40.85
##  Mean   : 70.33   Mean   : 318.1   Mean   :25.87   Mean   :40.72
##  3rd Qu.: 79.00   3rd Qu.: 450.8   3rd Qu.:28.55   3rd Qu.:42.95
##  Max.   :102.00   Max.   :1252.0   Max.   :44.90   Max.   :48.60
##
##       coud           poign            bras           age1
##  Min.   :5.400   Min.   :4.200   Min.   :23.00         : 1
##  1st Qu.:6.300   1st Qu.:5.075   1st Qu.:27.75   adulte:30
##  Median :6.900   Median :5.400   Median :31.50   jeune :38
```

```
##  Mean   :6.835   Mean   :5.434   Mean   :31.27   vieux :11
##  3rd Qu.:7.400   3rd Qu.:5.800   3rd Qu.:34.00
##  Max.   :8.300   Max.   :6.700   Max.   :43.80
##
```

VI-PRETRAITEMENT

enl?vement des variables inutiles comme le sexe et le num?ro du patient et l'age en valeur r?el

```
her<-her[,c(-1,-2)]
her<-her[,-1]
```

VII-ETUDE DE LA QUALITE DE L'ACP

chargement des librairies

```
library(FactoMineR)
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

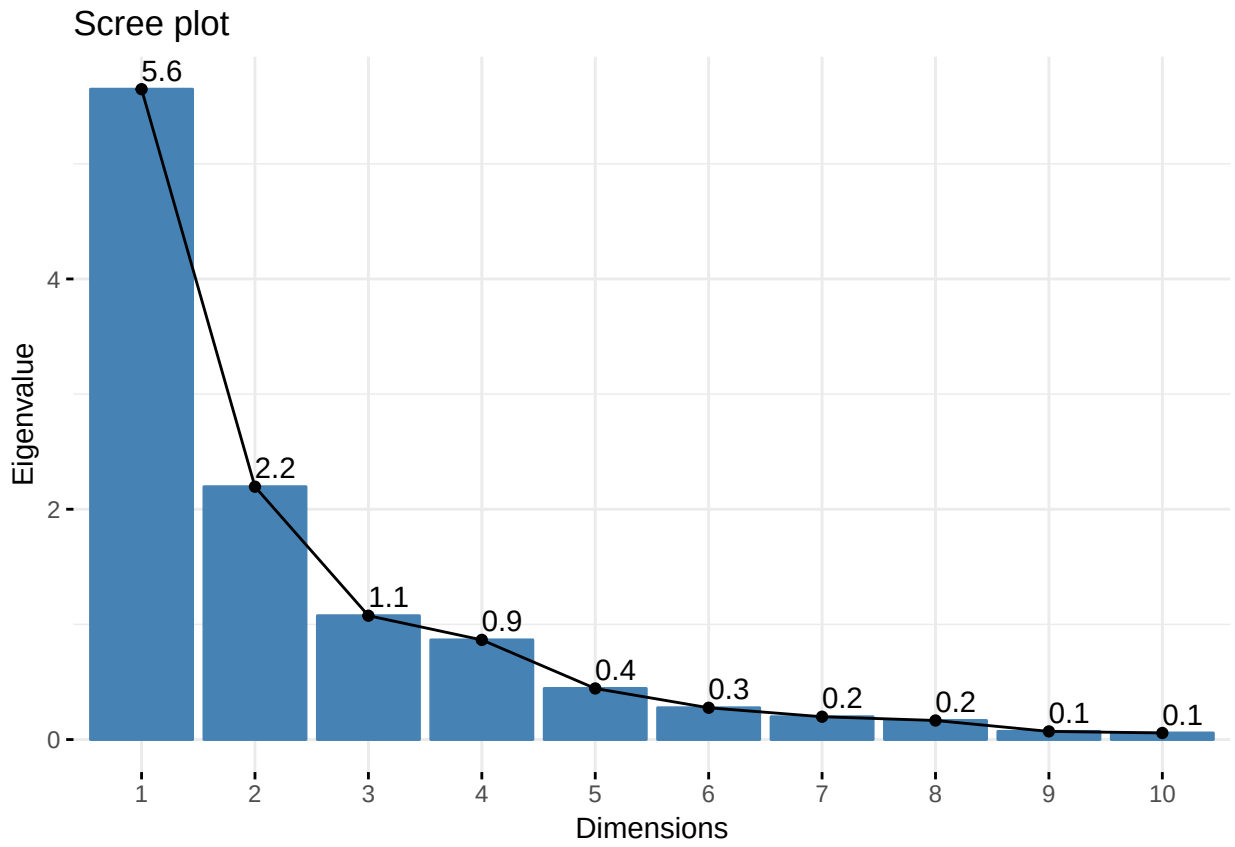
```
library(ggplot2)
```

analyse des valeurs propres

```
acp<-PCA(her,scale.unit = TRUE,quali.sup =13,quanti.sup = 7,ncp = 3,graph = FALSE)
acp$eig
```

```
##          eigenvalue percentage of variance
## comp 1  5.647976896            51.34524451
## comp 2  2.195763632            19.96148756
## comp 3  1.075510052             9.77736411
## comp 4  0.865325331             7.86659392
## comp 5  0.443623602             4.03294184
## comp 6  0.276030225             2.50936569
## comp 7  0.198202235             1.80183850
## comp 8  0.165662571             1.50602337
## comp 9  0.070778548             0.64344135
## comp 10 0.057224364             0.52022149
## comp 11 0.003902542             0.03547766
##         cumulative percentage of variance
## comp 1                           51.34524
## comp 2                           71.30673
## comp 3                           81.08410
## comp 4                           88.95069
## comp 5                           92.98363
## comp 6                           95.49300
## comp 7                           97.29484
## comp 8                           98.80086
## comp 9                           99.44430
## comp 10                          99.96452
## comp 11                         100.00000
```

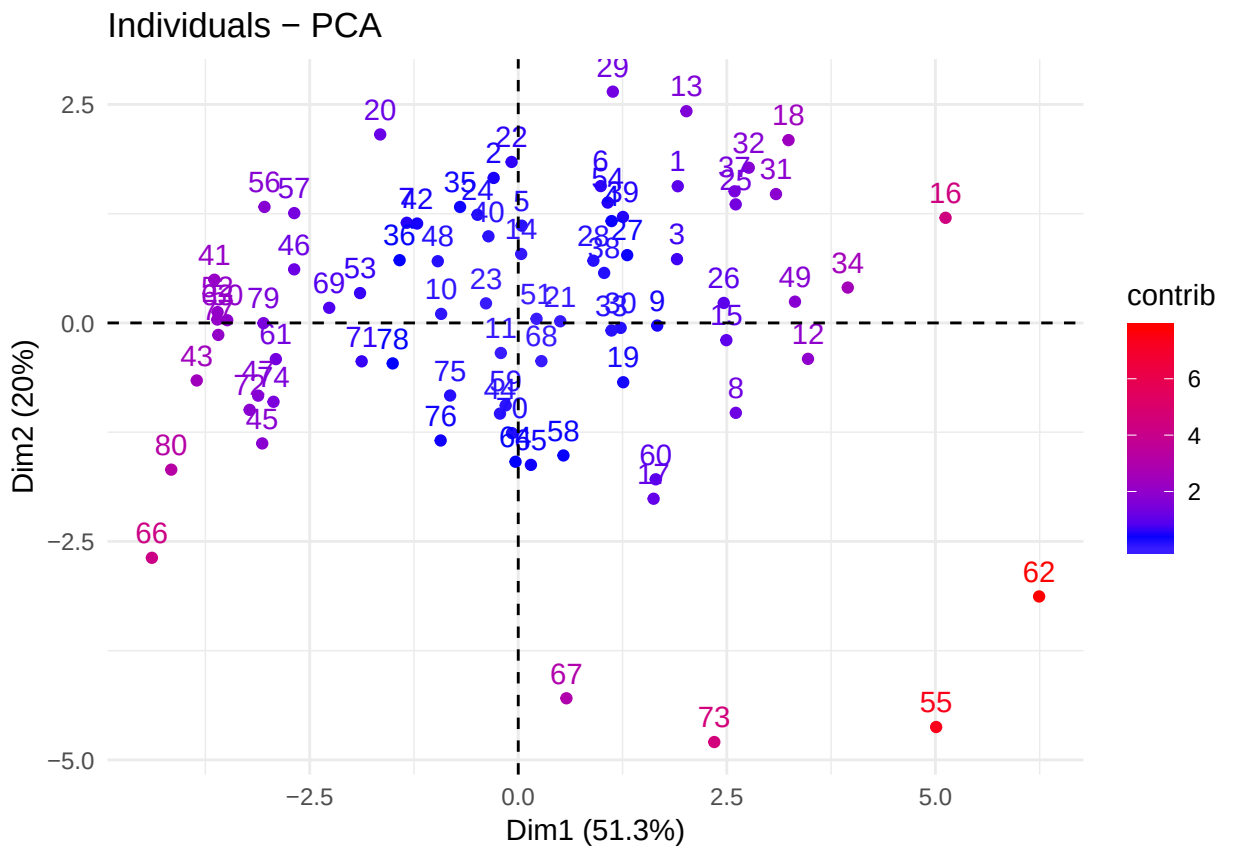
```
fviz_screeplot(acp,choice="eigenvalue",addlabels=TRUE)
```



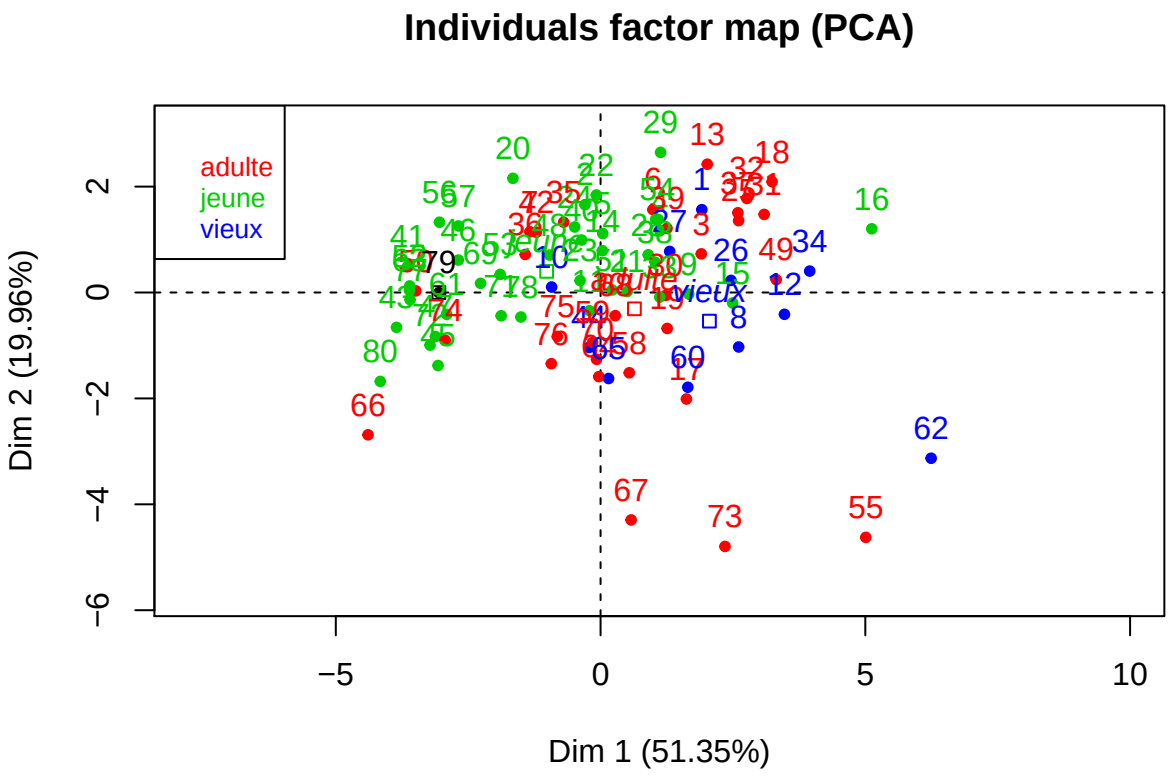
->on peut v?rifier en premier lieu que les trois premi?res valeurs propres sont sup?rieures ? 1, mais on peut se satisfaire des deux premiers axes puisqu'ils repr?sentent 71.3% de l'information disponible,donc l'exploration des donn?es va se reposer sur ces trois axes.
En deuxi?me lieu,il y a la pr?sence d'un effet de coude en se basant sur l'?boulis des valeurs propres.

Nuage des points des individus:

```
fviz_pca_ind(acp, col.ind="contrib")+
  scale_color_gradient2(low="white", mid="blue", high="red",midpoint=0.4) + theme_minimal()
```

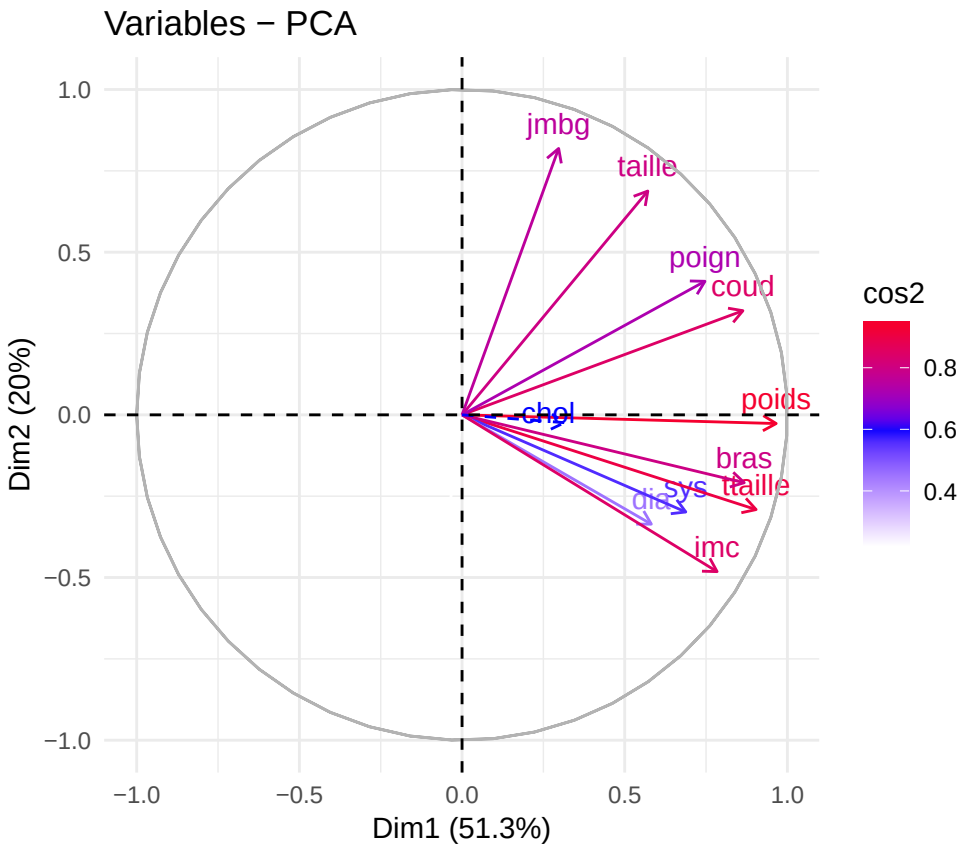



```
plot(acp,habillage =13, choix = "ind")
```



cercle de corr?lation: selon le cos^2

```
fviz_pca_var(acp, col.var="cos2")+
  scale_color_gradient2(low="white", mid="blue", high="red",midpoint=0.6) + theme_minimal()
```

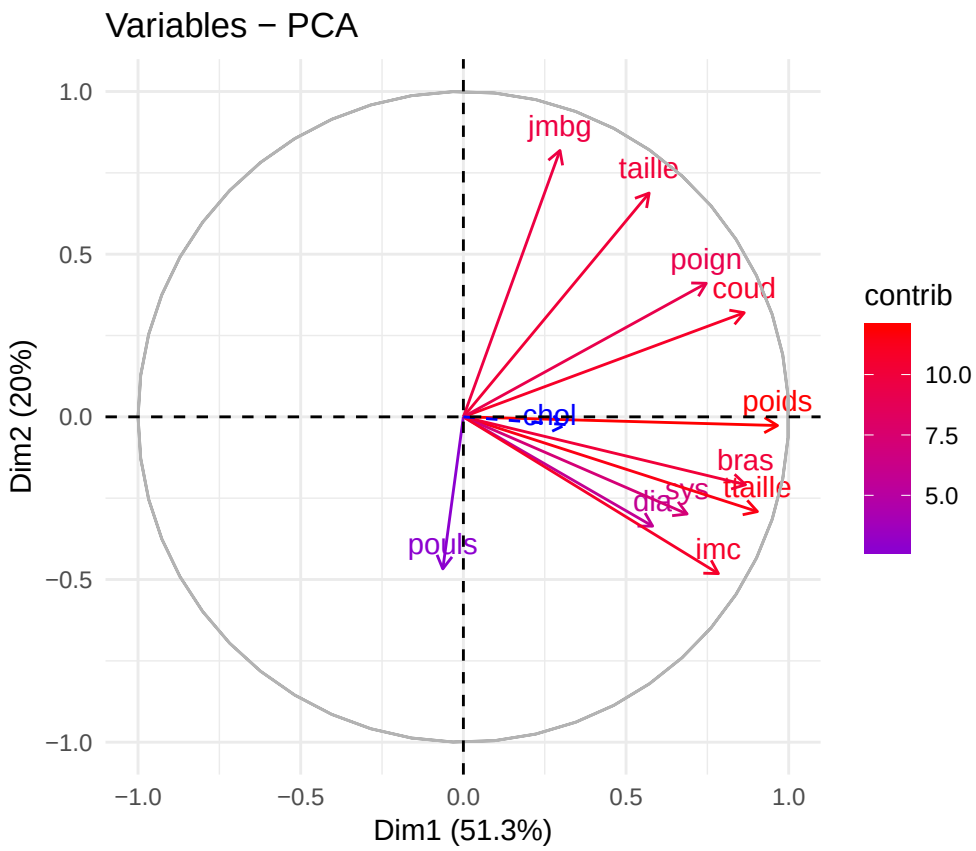


cercle de corr?lation selon la contribution:

```
acp$var$contrib
```

```
##                Dim.1      Dim.2        Dim.3
## taille    5.77160690 21.5447458  9.398350492
## poids    16.50706174  0.0324871  1.216812599
## ttaille  14.46337749  3.8591092  0.849448972
## pouls     0.07090979  9.9427279  9.299326711
## sys       8.37292463  4.0581343 22.003782585
## dia       5.99049184  5.1376389 32.763123209
## imc      10.87932326 10.5728480 11.100195718
## jmbg      1.56086869 30.5306708  0.056901972
## coud     13.19415859  4.6662296  0.386299492
## poign     9.86948437  7.6735401  0.000635961
## bras     13.31979271  1.9818683 12.925122288
```

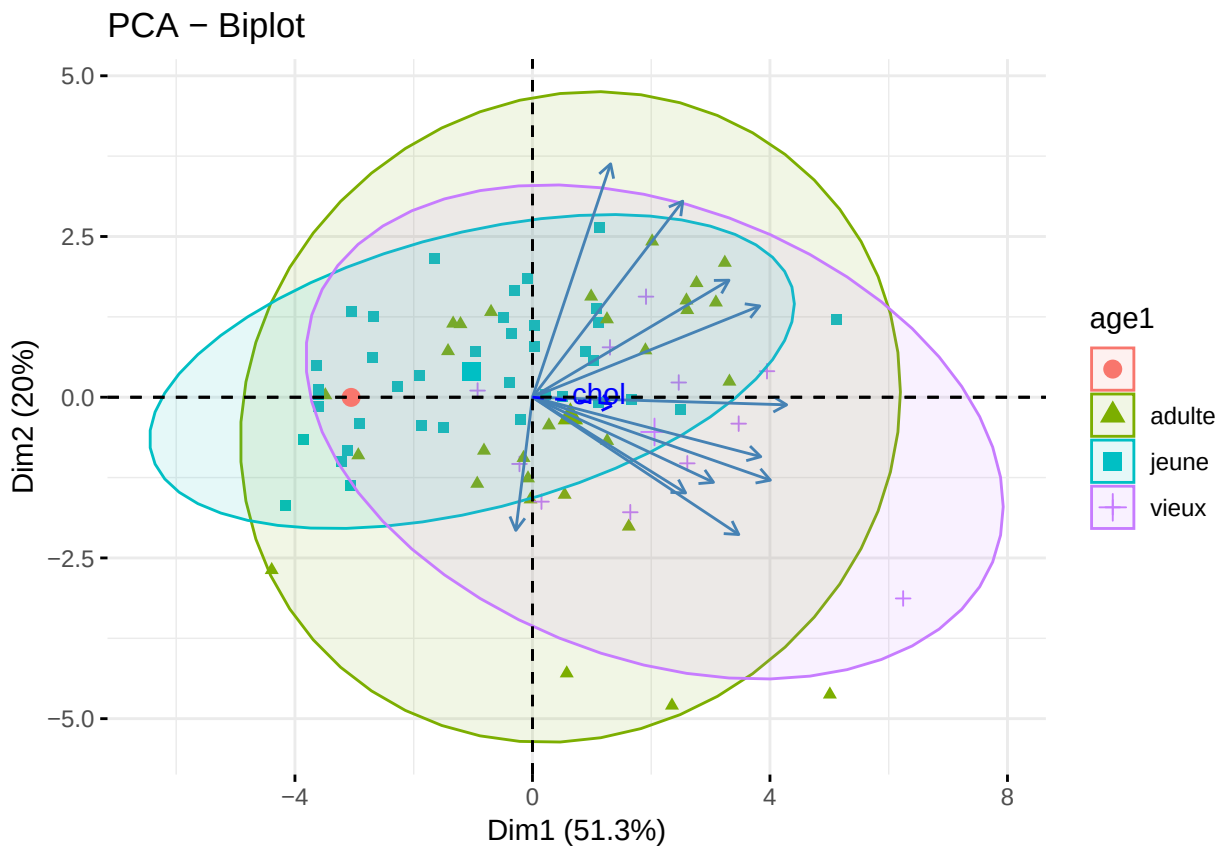
```
fviz_pca_var(acp, col.var="contrib")+
  scale_color_gradient2(low="white", mid="blue", high="red",midpoint=0.6) + theme_minimal()
```



Effectuer un habillage selon la variable cat?gor?ielle et tracer les ?llipses de confiance:

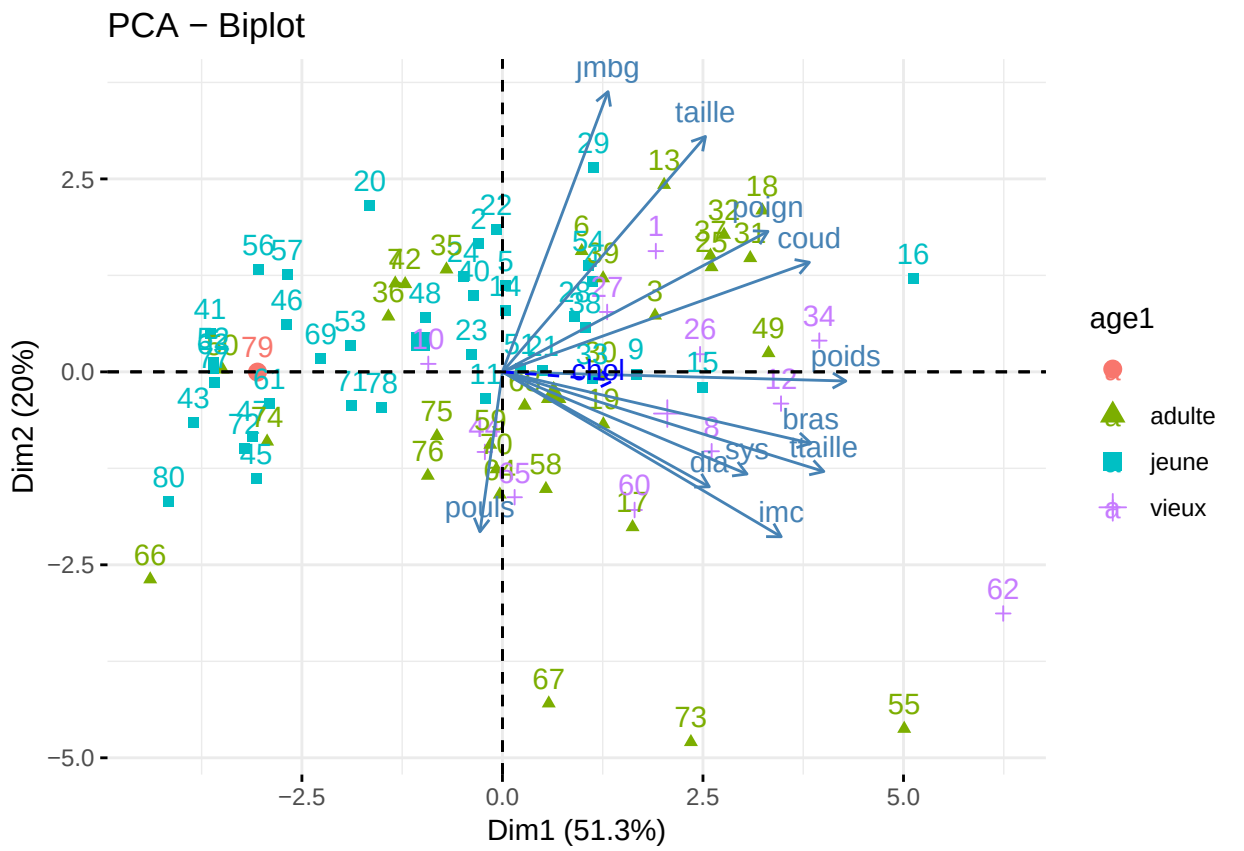
```
fviz_pca(acp,habillage = 13,label = TRUE ,addEllipses = TRUE)
```

```
## Too few points to calculate an ellipse
```



effectuer une superposition du plot de variables et d'individus:

```
fviz_pca(acp,habillage = 13)
```



VIII-INTERPRETATION

-En jetant un coup d'oil sur le cercle de corr?lation et dans le but de d?terminer la bonne repr?sentation des variables, on v?rifie que les deux axes principaux contiennent respectivement 51.3% et 20% de l'information totale.
-Les variables poids, ttaille, bras et coude sont les indicateurs caract?ristiques de la recherche scientifique contribuent le plus au premier axe avec des valeurs respectives de 16.5, 14.46 , 13.31 et 13,19.

En exploitant les r?sultats obtenu ? travers le cercle de corr?lation, on trouve que :
-On remarque que toutes les variables sont plus ou moins d?pendante sauf les battements du cour.
-Les variables : cholest?rol et poids sont fortement corr?l?s : r?sultat attendu puisque l'augmentation de poids est un vrai indicateur d'ob?sit?. En effet, scientifiquement, les personnes maigres risquent moins d'acqu?rir une hypercholest?rol?mie et inversement. -Les variables: cholest?rol, tour de taille "ttaille",la pression sanguine systolique "sys" et diastolique "dia" sont plus ou moins corr?l?s: cela est bien ?vident scientifiquement car les contractions et rel?chement du cour peuvent ?tre des cons?quences d'une hypercholest?rol?mie. Ainsi que pour le tour de taille qui augmentera en fonction du surpoids de la personne.
-La variable cholest?rol n'a aucun effet sur les battement du cour donc elle sont ind?pendantes vue aussi qu'il sont pas corr?l?s (plus que perpendiculaires).

En exploitant les r?sultats obtenu ? travers le nuage de point, on trouve que :
-dans cet ?chantillon, l'effectif des jeunes est le plus ?lev? tandis que les vieux repr?sentent le nombre le plus faible.
-On constate que les jeunes sont les moins touch?s par ce diagnostique puisque leur classe d'?ge est inversement corr?l?e avec les variables poids, cholest?rol. . . Mais ils ont tendance ? avoir une grande taille, des jambes plus langues plus que les deux autres cat?gories.

-Les adultes se r?partissent d'une mani?re r?gul?rent en formant une ellipse sous forme d'un cercle situ?e au milieu.
-Alors que les vieillards bien qu'ils soient une minoritaire dans cet ?chantillon, ont tendance ? avoir une hypercholest?rol?mie, une pression sanguine plus importante.

CONCLUSION

L'objectif de cette ?tude ?tait d'explorer et d'exploiter un jeu de donn?e dont la variable ? expliquer est "chol" (cholest?rol) en fonction d'autres variables explicatives.

le r?sultat nous a permis de classer le cholest?rol pour les personnes les plus ?ges. Aussi, il peut ?tre engendr? par le gain d'un poids suppl?mentaire au cours du temps.
Mais, n'oublions par que m?me d'autres maladies peuvent causer une hypercholest?rol?mie ou bien existe t-il un facteur biologique qui favorise le terrain de la maladie?