ACP Analyse en Composantes Principales : PCA (Principal Component Analysis)

DJEBALI Wissam

25 février 2018

ACP ou PCA

Packages R : FactoMineR, factoextra, corrplot

En résumé, l'analyse en composantes principales permet:

_ d'identifier des "profils cachés" dans un jeu de données,

_ de réduire les dimensions des données en enlevant la redondance des données,

_ d'identifier les variables corrélées

```
data(BigMac2003)
mcdo<-BigMac2003
write.csv(mcdo,"./BigMac2003.csv")
# Visualisation des données
head(mcdo)
```

```
##           BigMac Bread Rice FoodIndex  Bus Apt TeachGI TeachNI TaxRate
## Amsterdam     16     9    9      65.9 2.00 890    34.3    20.5 40.2332
## Athens        21    12   19      63.5 0.61 620    19.5    15.9 18.4615
## Auckland      19    19    9      55.4 1.57 780    22.0    16.1 26.8182
## Bangkok       50    42   25      46.4 0.47 120     4.2     4.0  4.7619
## Barcelona     22    19   10      62.9 0.91 590    25.5    20.1 21.1765
## Basel         15     7    7      98.4 2.34 930    78.5    57.6 26.6242
##           TeachHours
## Amsterdam         39
## Athens            29
## Auckland          40
## Bangkok           35
## Barcelona         39
## Basel             35
```

Standardisation des données

Dans l'analyse en composantes principales, les variables sont souvent normalisées. Ceci est particulièrement recommandé lorsque les variables sont mesurées dans différentes unités (par exemple: kilogrammes, kilomètres, centimètres, ...); sinon, le résultat de l'ACP obtenue sera fortement affecté.

L'objectif est de rendre les variables comparables. Généralement, les variables sont normalisées de manière à ce qu'elles aient au final i) un écart type égal à un et ii) une moyenne égale à zéro.

Techniquement, l'approche consiste à transformer les données en soustrayant à chaque valeur une valeur de référence (la moyenne de la variable) et en la divisant par l'écart type. A l'issue de cette transformation les données obtenues sont dites données centrées-réduites. L'ACP appliquée à ces données transformées est appelée ACP normée.

La standardisation des données est une approche beaucoup utilisée dans le contexte de l'analyse des données d'expression de gènes avant les analyses de type PCA et de clustering.

Lors de la normalisation des variables, les données peuvent être transformées comme suit:

$$z_{ij} = \frac{x_{ij} - mean(x_j)}{sd(x_j)}$$

Où $mean(x_j) = \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ est la moyenne des valeurs de la variable x_j, et $sd(x_j) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$ est l'écart type (SD).

La fonction scale() peut être utilisée pour normaliser les données.

Calculer l'ACP sur les individus/variables actifs avec PCA()

PCA(X, scale.unit = TRUE, ncp = 5, graph = TRUE)

_X: jeu de données de type data frame. Les lignes sont des individus et les colonnes sont des variables numériques

_scale.unit: une valeur logique. Si TRUE, les données sont standardisées/normalisées avant l'analyse.

_ncp: nombre de dimensions conservées dans les résultats finaux.

_graph: une valeur logique. Si TRUE un graphique est affiché.

```
res.pca <- PCA(mcdo, graph = FALSE)
# les variables sont normalisés et centrées

# Résultats de PCA()
print(res.pca)
```

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 69 individuals, described by 10 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"
```

Valeurs propres / Variances

Les valeurs propres (eigenvalues en anglais) mesurent la quantité de variance expliquée par chaque axe principal. Les valeurs propres sont grandes pour les premiers axes et petits pour les axes suivants. Autrement

dit, les premiers axes correspondent aux directions portant la quantité maximale de variation contenue dans le jeu de données.

Nous examinons les valeurs propres pour déterminer le nombre de composantes principales à prendre en considération. Les valeurs propres et la proportion de variances (i.e. information) retenues par les composantes principales peuvent être extraites à l'aide de la fonction get_eigenvalue() [package factoextra].

```
eig.val <- get_eigenvalue(res.pca)
eig.val
```

```
##         eigenvalue variance.percent cumulative.variance.percent
## Dim.1  5.359028761      53.59028761                     53.59029
## Dim.2  1.319156815      13.19156815                     66.78186
## Dim.3  0.822845693       8.22845693                     75.01031
## Dim.4  0.789640640       7.89640640                     82.90672
## Dim.5  0.522552735       5.22552735                     88.13225
## Dim.6  0.459832363       4.59832363                     92.73057
## Dim.7  0.333313576       3.33313576                     96.06371
## Dim.8  0.265617436       2.65617436                     98.71988
## Dim.9  0.124297498       1.24297498                     99.96286
## Dim.10 0.003714483       0.03714483                    100.00000
```

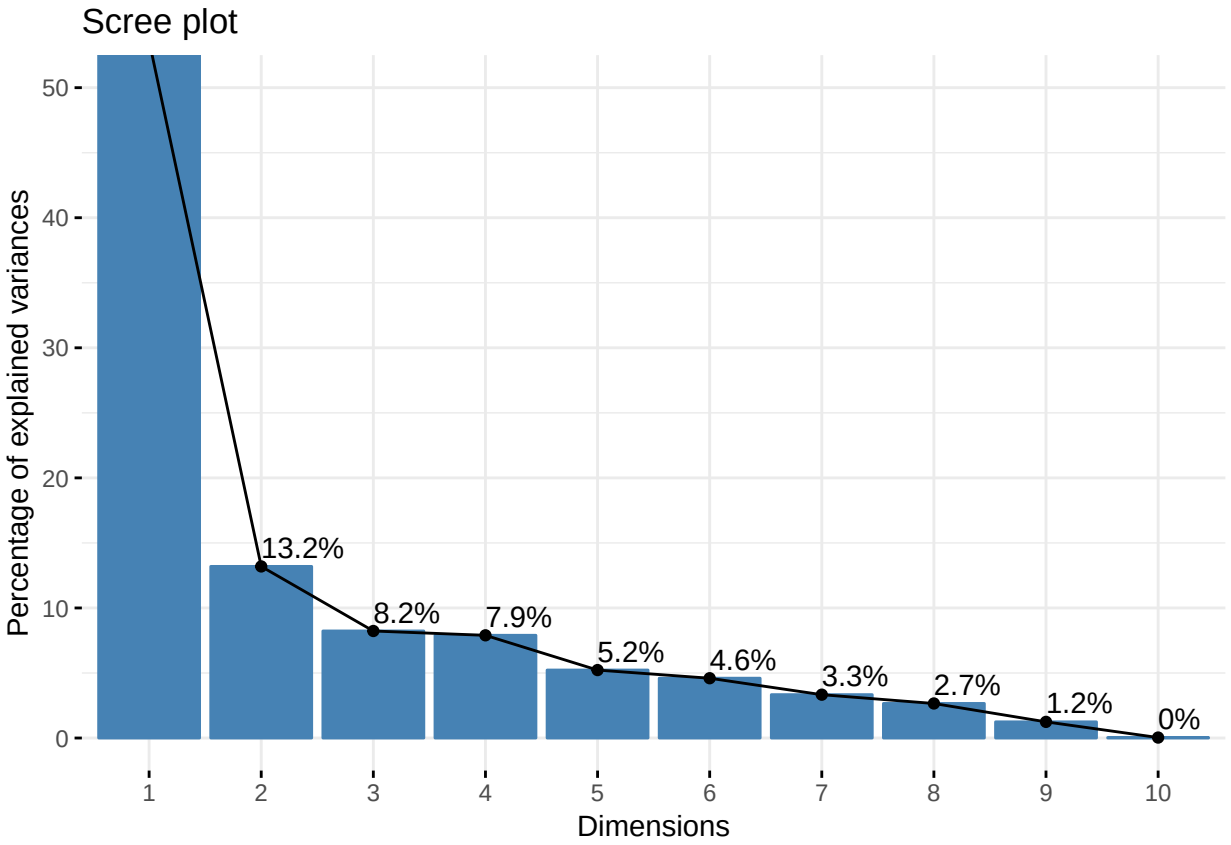
La proportion de variance expliquée par chaque valeur propre est donnée dans la deuxième colonne. Le pourcentage cumulé expliqué est obtenu en ajoutant les proportions successives de variances expliquées.

Les valeurs propres peuvent être utilisées pour déterminer le nombre d'axes principaux à conserver après l'ACP (Kaiser 1961):

Une valeur propre > 1 indique que la composante principale (PC) concernée représente plus de variance par rapport à une seule variable d'origine, lorsque les données sont standardisées. Ceci est généralement utilisé comme seuil à partir duquel les PC sont conservés. A noter que cela ne s'applique que lorsque les données sont normalisées.

Vous pouvez également limiter le nombre d'axes à un nombre qui représente une certaine fraction de la variance totale. Par exemple, si vous êtes satisfaits avec 70% de la variance totale expliquée, utilisez le nombre d'axes pour y parvenir.

```
# Déterminer le nb de composantes principales(=le nb d'axes)
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 50))
```



Du graphique ci-dessus, nous pourrions vouloir nous arrêter à la 5e composante principale. 88% des informations (variances) contenues dans les données sont conservées par les cinq premières composantes principales.

Graphique des variables

Résultats pour les variables

Les résultats pour les variables actives (coordonnées, corrélation entre variables et les axes, cosinus-carré et contributions)

```
var <- get_pca_var(res.pca)
var
```

```
## Principal Component Analysis Results for variables
##  ===================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the variables"
## 2 "$cor"     "Correlations between variables and dimensions"
## 3 "$cos2"    "Cos2 for the variables"
## 4 "$contrib" "contributions of the variables"
```

```
# Coordonnées
head(var$coord)
```

```
##               Dim.1      Dim.2       Dim.3       Dim.4      Dim.5
## BigMac    -0.8025374  0.1901109  0.32542712  0.05873442 -0.2010152
## Bread     -0.6716422  0.2110211  0.20314974 -0.47503773  0.4642116
```

```
## Rice      -0.6558269  0.2693011  0.58983732  0.12225909 -0.2046040
## FoodIndex  0.7938670  0.4041140  0.17811536 -0.24856737  0.1359610
## Bus        0.7850539 -0.1579953  0.25671339  0.16819364  0.1703081
## Apt        0.7690471  0.2304157 -0.04137238 -0.20531810 -0.1724731
```

```
# Cos2: qualité de répresentation
head(var$cos2)
```

```
##               Dim.1      Dim.2       Dim.3       Dim.4      Dim.5
## BigMac    0.6440663 0.03614216 0.105902813 0.003449732 0.04040712
## Bread     0.4511032 0.04452991 0.041269818 0.225660849 0.21549243
## Rice      0.4301089 0.07252309 0.347908062 0.014947286 0.04186281
## FoodIndex 0.6302248 0.16330816 0.031725080 0.061785740 0.01848541
## Bus       0.6163097 0.02496253 0.065901765 0.028289101 0.02900483
## Apt       0.5914335 0.05309141 0.001711674 0.042155520 0.02974697
```

```
# Contributions aux composantes principales
head(var$contrib)
```

```
##               Dim.1     Dim.2      Dim.3      Dim.4     Dim.5
## BigMac    12.018339  2.739793 12.8703126  0.4368737  7.732640
## Bread      8.417629  3.375634  5.0154990 28.5776641 41.238408
## Rice       8.025874  5.497685 42.2810820  1.8929226  8.011212
## FoodIndex 11.760057 12.379738  3.8555321  7.8245390  3.537520
## Bus       11.500399  1.892310  8.0090066  3.5825285  5.550604
## Apt       11.036206  4.024648  0.2080188  5.3385702  5.692625
```

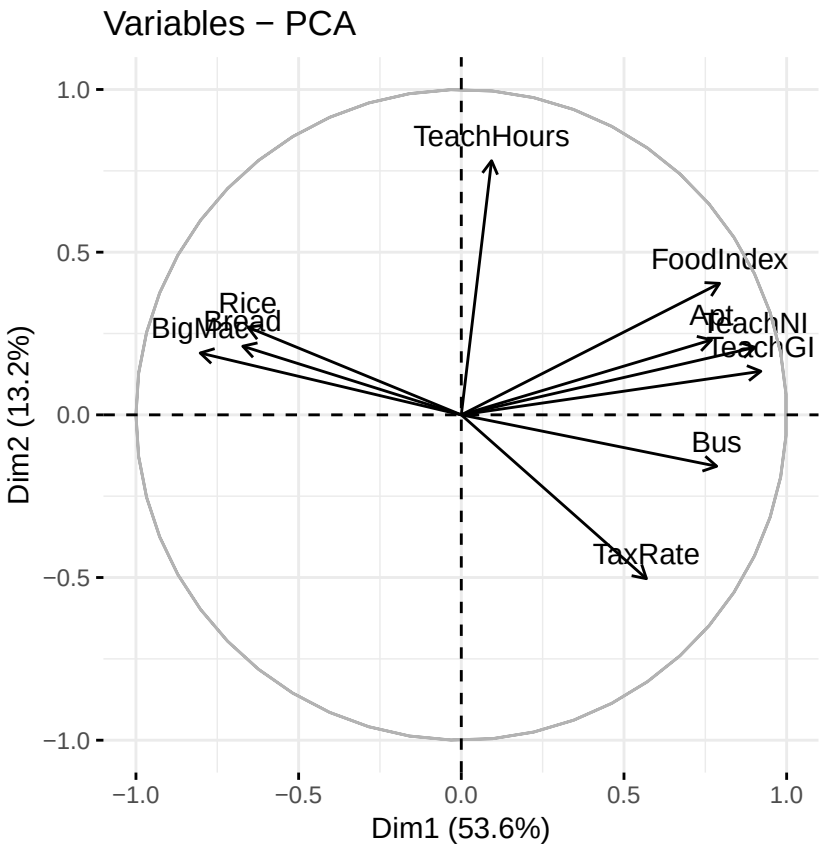
Cercle de corrélation

La corrélation entre une variable et une composante principale (PC) est utilisée comme coordonnées de la variable sur la composante principale. La représentation des variables diffère de celle des observations: les observations sont représentées par leurs projections, mais les variables sont représentées par leurs corrélations (Abdi and Williams 2010).

```
# Coordonnées des variables
head(var$coord, 4)
```

```
##               Dim.1     Dim.2     Dim.3       Dim.4      Dim.5
## BigMac    -0.8025374 0.1901109 0.3254271  0.05873442 -0.2010152
## Bread     -0.6716422 0.2110211 0.2031497 -0.47503773  0.4642116
## Rice      -0.6558269 0.2693011 0.5898373  0.12225909 -0.2046040
## FoodIndex  0.7938670 0.4041140 0.1781154 -0.24856737  0.1359610
```

```
# Cercle de corrélation des variables
fviz_pca_var(res.pca, col.var = "black")
```



_ Les variables positivement corrélées sont regroupées.

_ Les variables négativement corrélées sont positionnées sur les côtés opposés de l'origine du graphique (quadrants opposés).

_ La distance entre les variables et l'origine mesure la qualité de représentation des variables. Les variables qui sont loin de l'origine sont bien représentées par l'ACP

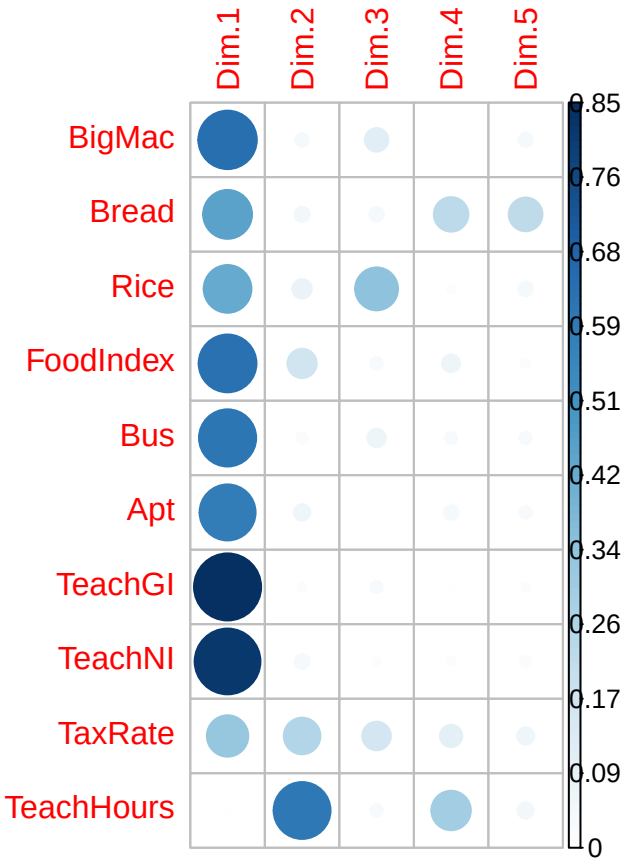
Qualité de représentation

La qualité de représentation des variables sur la carte de l'ACP s'appelle cos2 (cosinus carré) .

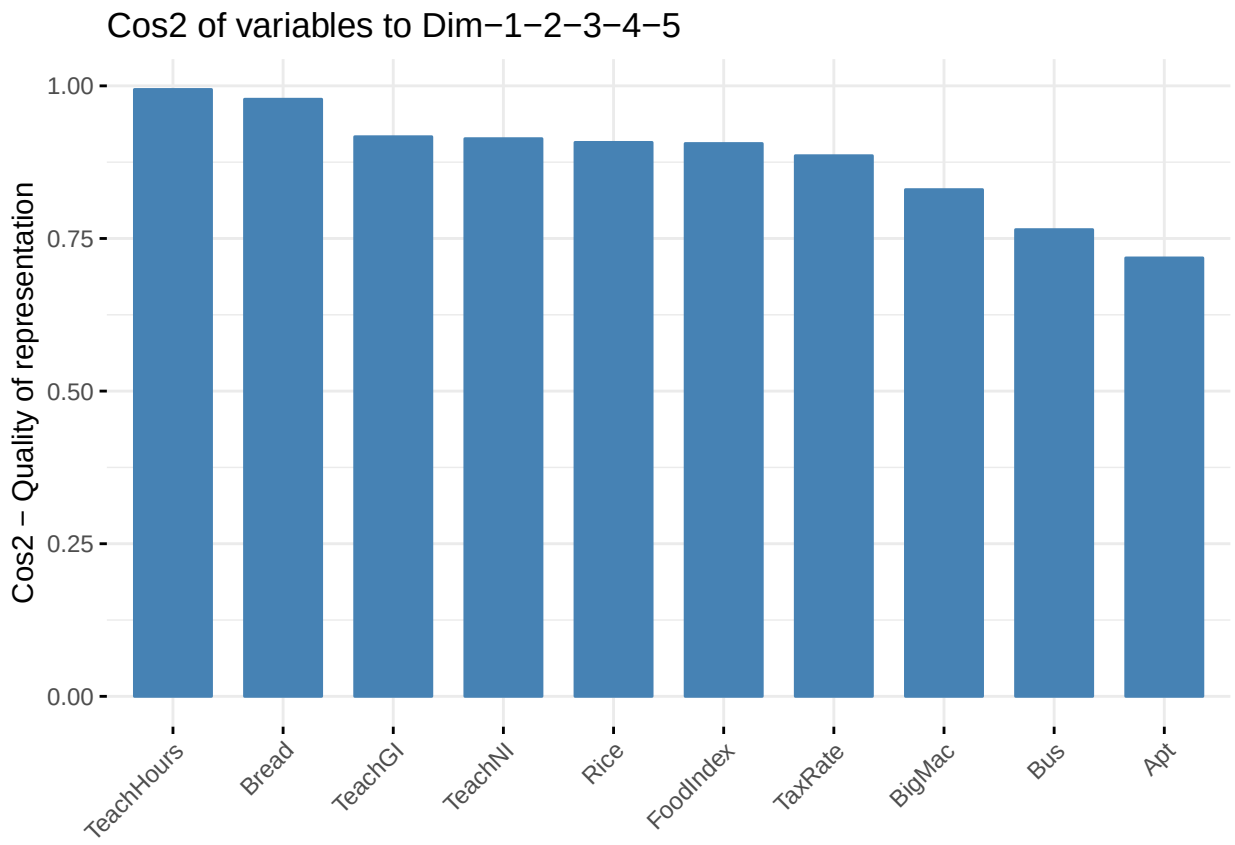
```
head(var$cos2, 4)
```

```
##               Dim.1      Dim.2      Dim.3       Dim.4      Dim.5
## BigMac    0.6440663 0.03614216 0.10590281 0.003449732 0.04040712
## Bread     0.4511032 0.04452991 0.04126982 0.225660849 0.21549243
## Rice      0.4301089 0.07252309 0.34790806 0.014947286 0.04186281
## FoodIndex 0.6302248 0.16330816 0.03172508 0.061785740 0.01848541
```

```
# Visualisation du cos2 des variables sur toutes les dim
corrplot(var$cos2, is.corr=FALSE)
```




```
# Cos2 total des variables sur Dim.1 et Dim.2
fviz_cos2(res.pca, choice = "var", axes = 1:5)
```



_ Un cos2 élevé indique une bonne représentation de la variable sur les axes principaux en considération. Dans ce cas, la variable est positionnée à proximité de la circonférence du cercle de corrélation.

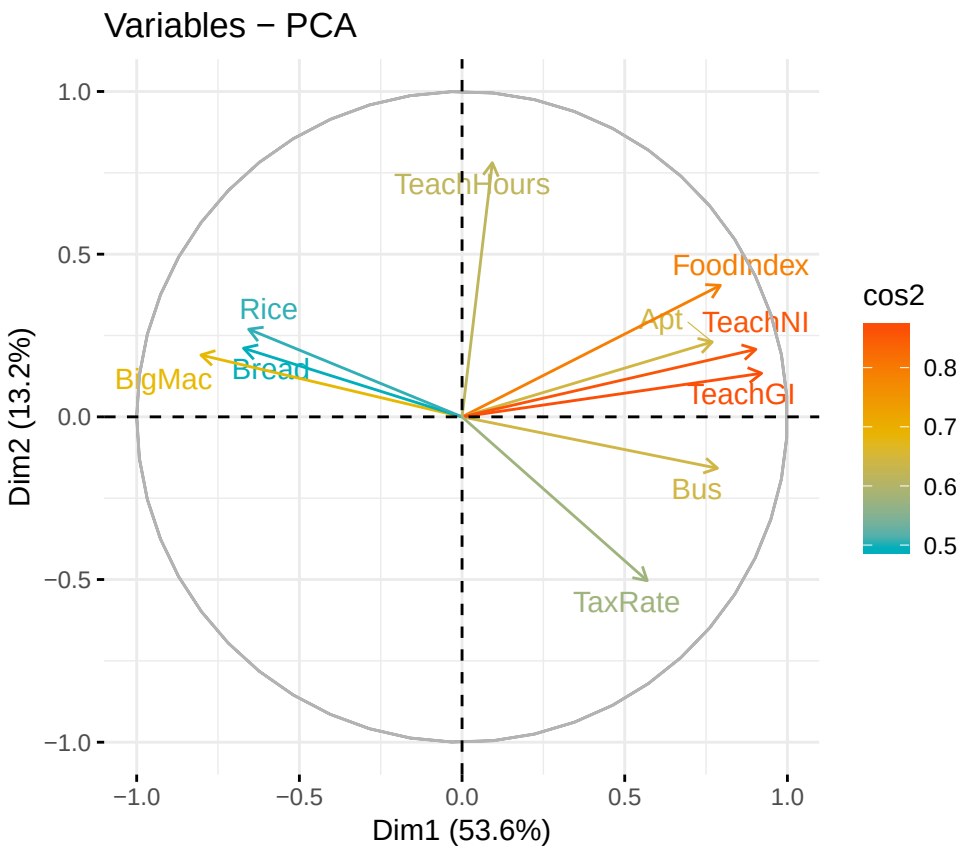
_ Un faible cos2 indique que la variable n'est pas parfaitement représentée par les axes principaux. Dans ce cas, la variable est proche du centre du cercle.

En résumé

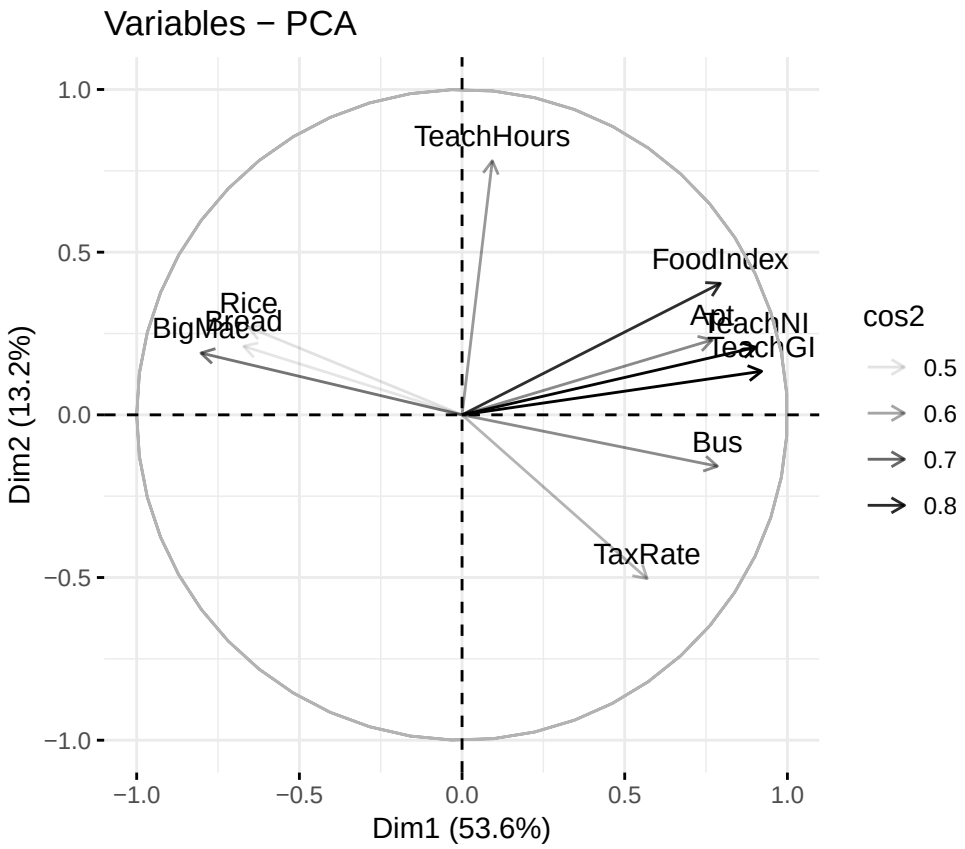
_ Les valeurs de cos2 sont utilisées pour estimer la qualité de la représentation Plus une variable est proche du cercle de corrélation, meilleure est sa représentation sur la carte de l'ACP (et elle est plus importante pour interpréter les composantes principales en considération)

_ Les variables qui sont proche du centre du graphique sont moins importantes pour les premières composantes.

```
# Colorer en fonction du cos2: qualité de représentation
fviz_pca_var(res.pca, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Évite le chevauchement de texte
             )
```




```
# Changer la transparence en fonction du cos2
fviz_pca_var(res.pca, alpha.var = "cos2")
```

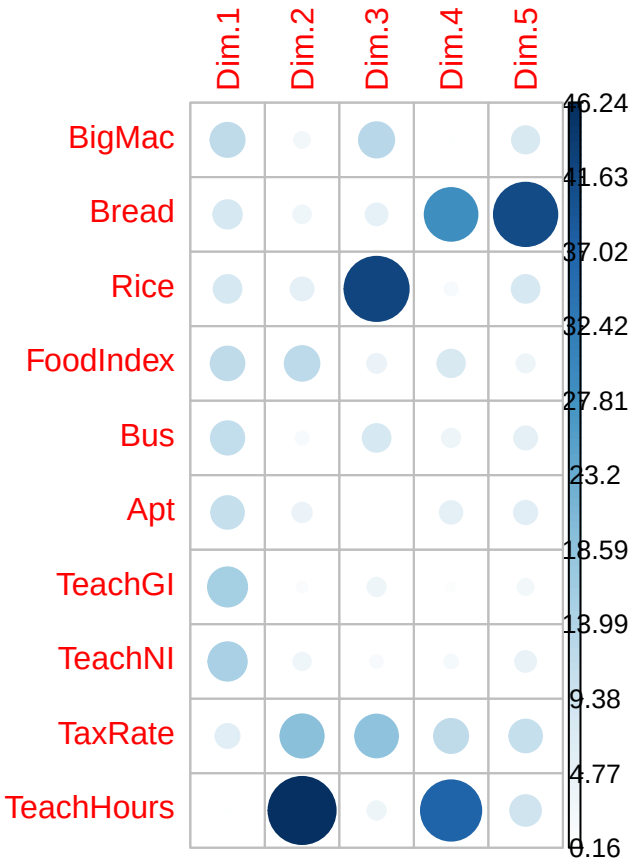


Contributions des variables aux axes principaux

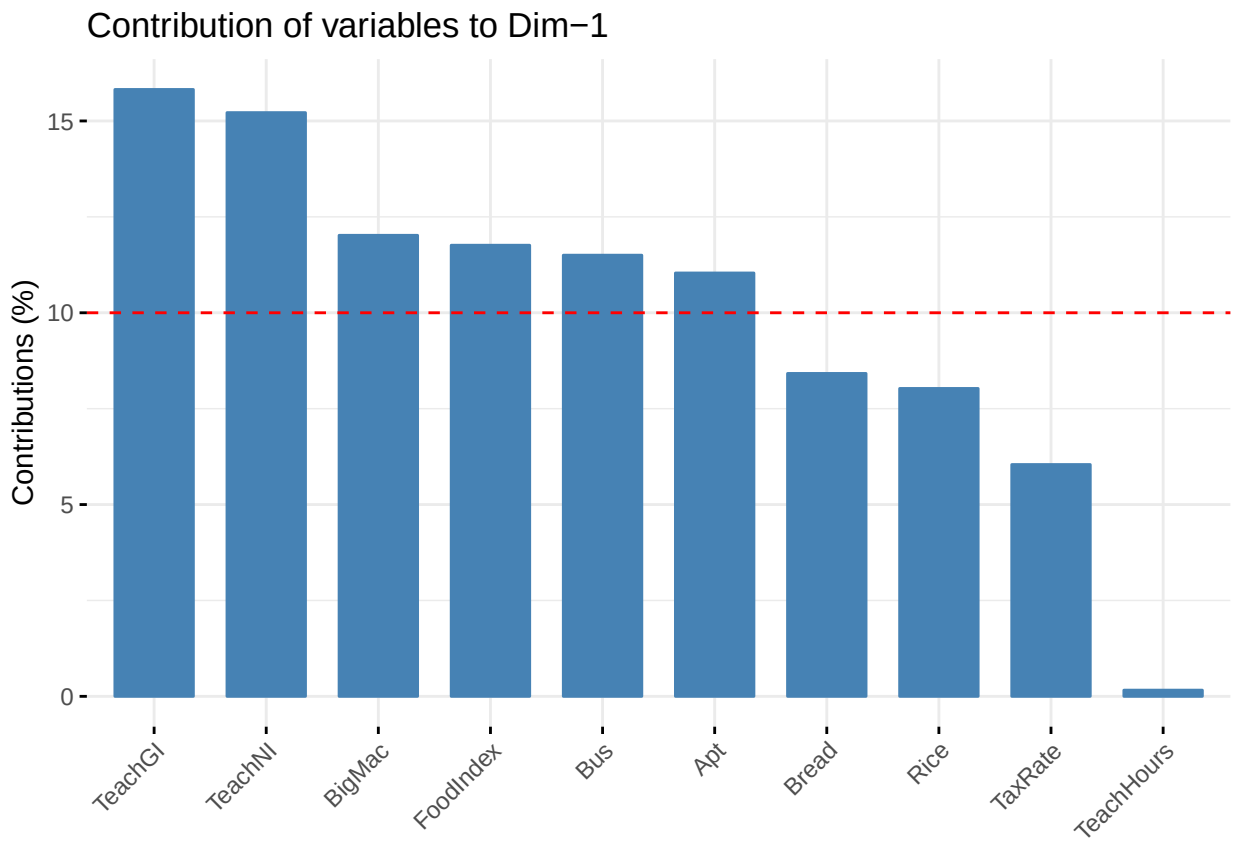
```
# Contribution des variables
head(var$contrib, 4)
```

```
##                Dim.1     Dim.2     Dim.3      Dim.4     Dim.5
## BigMac    12.018339  2.739793 12.870313  0.4368737  7.732640
## Bread      8.417629  3.375634  5.015499 28.5776641 41.238408
## Rice       8.025874  5.497685 42.281082  1.8929226  8.011212
## FoodIndex 11.760057 12.379738  3.855532  7.8245390  3.537520
```

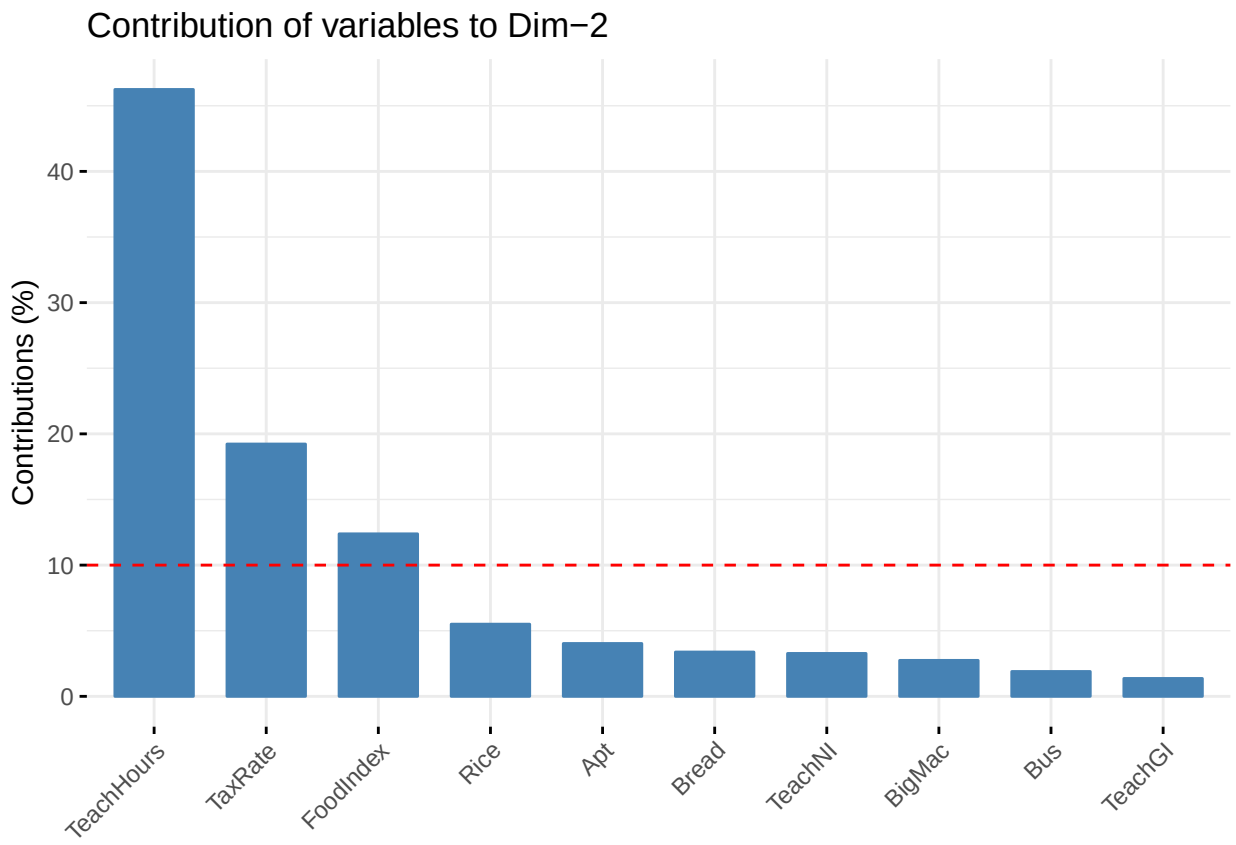
```
# Plus la valeur de la contribution est importante, plus la variable contribue à la composante principa
corrplot(var$contrib, is.corr=FALSE)
```



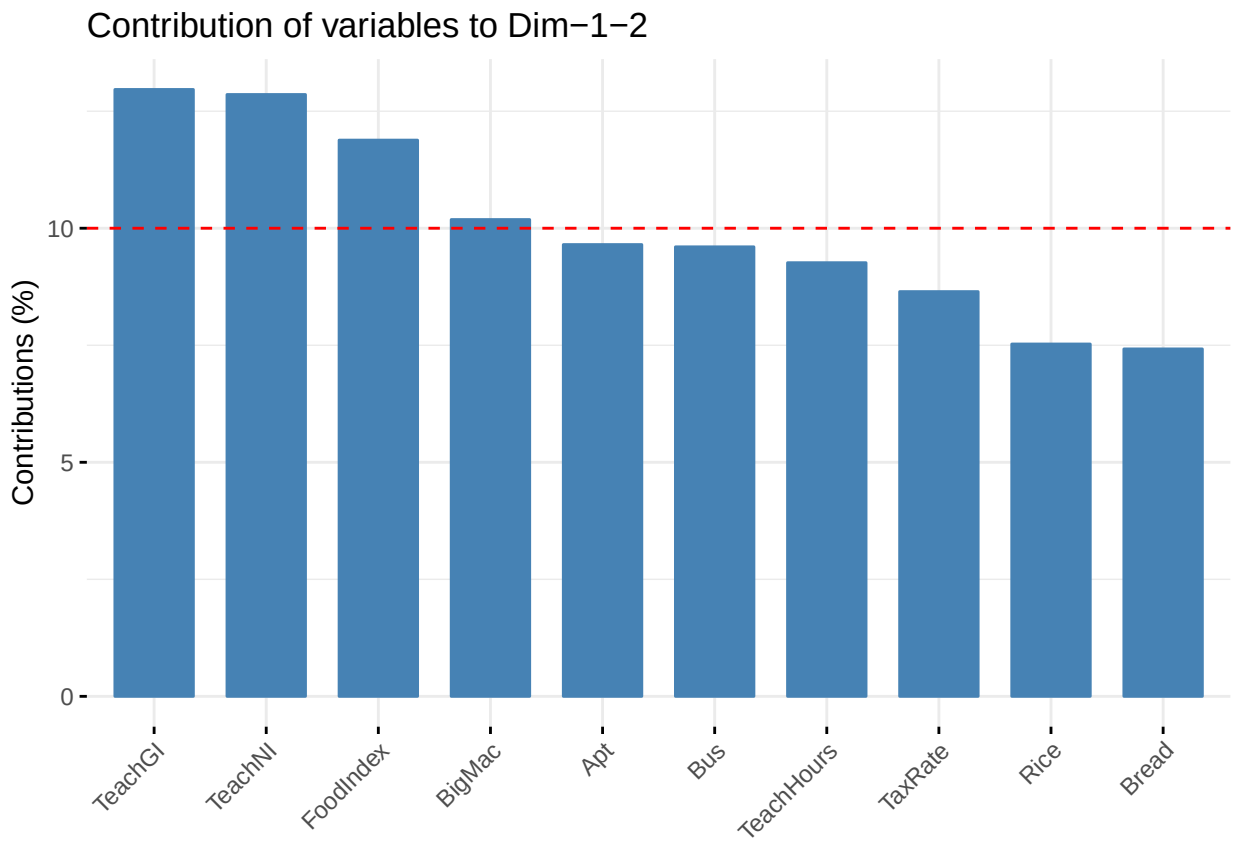

```
# Contributions des variables à PC1
fviz_contrib(res.pca, choice = "var", axes = 1, top = 10)
```




```
# Contributions des variables à PC2
fviz_contrib(res.pca, choice = "var", axes = 2, top = 10)
```




```
# Contribution totale à PC1 et PC2
fviz_contrib(res.pca, choice = "var", axes = 1:2, top = 10)
```



La ligne en pointillé rouge, sur le graphique ci-dessus, indique la contribution moyenne attendue. Si la contribution des variables était uniforme, la valeur attendue serait 1/length(variables) = 1/10 = 10%. Pour une composante donnée, une variable avec une contribution supérieure à ce seuil pourrait être considérée comme importante pour contribuer à la composante.

Notez que la contribution totale d'une variable donnée, pour expliquer la variance retenue par deux composantes principales, disons PC1 et PC2, est calculée comme

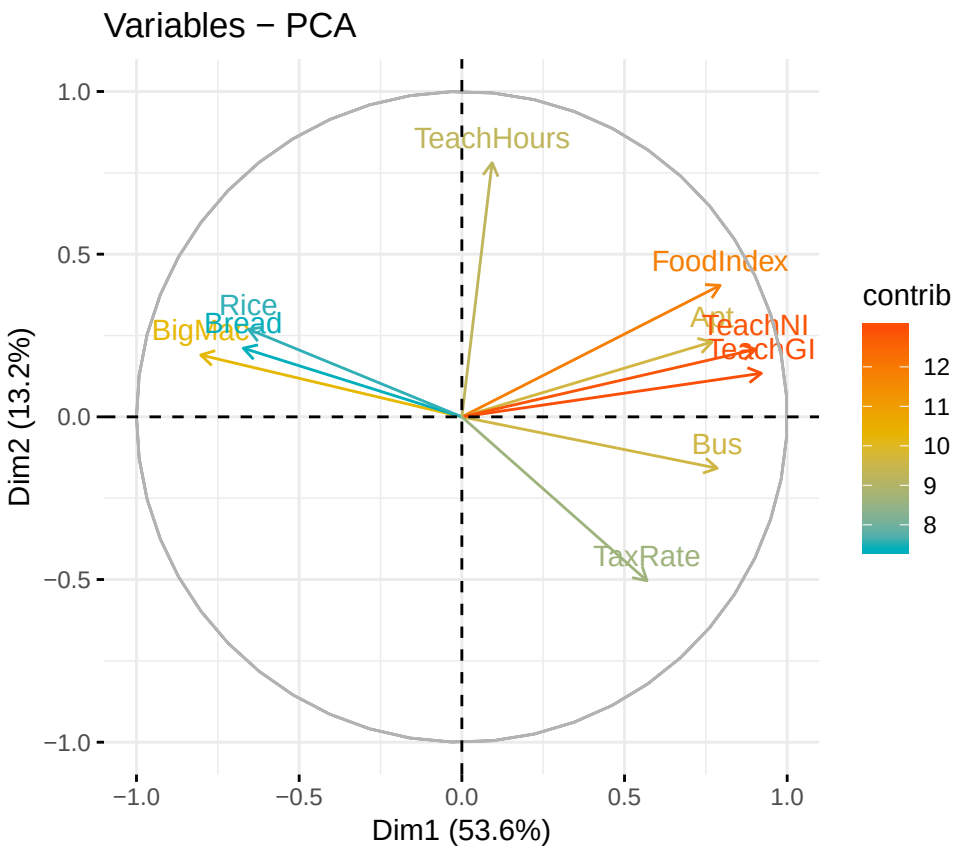
$contrib = \frac{(C_1 \times Eig_1)+(C_2 \times Eig_2)}{Eig_1+Eig_2}$, où

C_1 et C_2 sont les contributions de la variable aux axes PC1 et PC2, respectivement Eig_1 et $Eig-2$ sont les valeurs propres de PC1 et PC2, respectivement. Rappelons que les valeurs propres mesurent la quantité de variation retenue par chaque PC. Dans ce cas, la contribution moyenne attendue (seuil) est calculée comme suit:

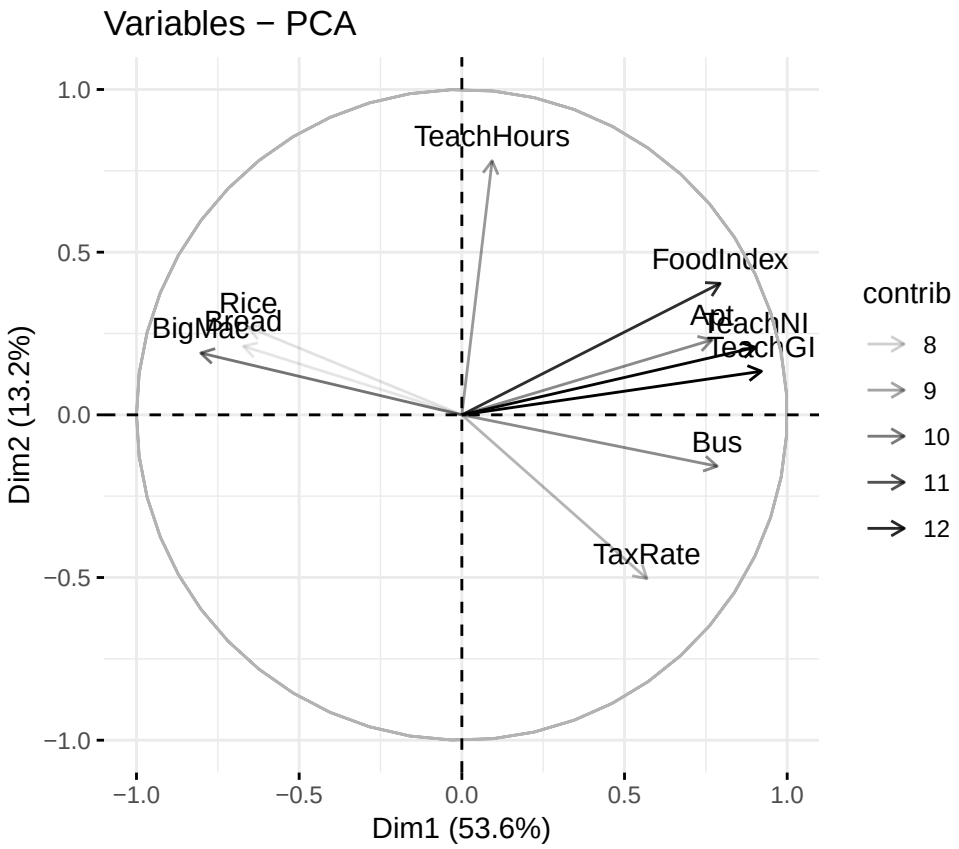
Comme mentionné ci-dessus, si les contributions des 10 variables étaient uniformes, la contribution moyenne attendue pour une PC donnée serait 1/10 = 10%. La contribution moyenne attendue d'une variable pour PC1 et PC2 est:

$\frac{(10 \times Eig_1)+(10 \times Eig_2)}{Eig_1+Eig_2}$

```
# Mise en évidence des variables les + contributives aux axes
fviz_pca_var(res.pca, col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07")
             )
```




```
# Changez la transparence en fonction de contrib
fviz_pca_var(res.pca, alpha.var = "contrib")
```



Description des dimensions

```
res.desc <- dimdesc(res.pca, axes = c(1,2), proba = 0.05)
# Description de la dimension 1
res.desc$Dim.1
```

```
## $quanti
##           correlation      p.value
## TeachGI     0.9207725 4.273670e-29
## TeachNI     0.9030451 2.765618e-26
## FoodIndex   0.7938670 4.072193e-16
## Bus         0.7850539 1.418747e-15
## Apt         0.7690471 1.186822e-14
## TaxRate     0.5690419 3.366465e-07
## Rice       -0.6558269 9.583780e-10
## Bread      -0.6716422 2.665182e-10
## BigMac     -0.8025374 1.122717e-16
```

```
# Description de la dimension 2
res.desc$Dim.2
```

```
## $quanti
##            correlation      p.value
## TeachHours   0.7810112 2.466922e-15
## FoodIndex    0.4041140 5.739652e-04
## Rice         0.2693011 2.524519e-02
```

```
## TaxRate     -0.5035727 1.030711e-05
```

$quanti représente les résultats pour les variables quantitatives. Notez que les variables sont triées en fonction de la p-value de la corrélation.

Graphique des individus

Résultats sur les individus

```
ind <- get_pca_ind(res.pca)
ind
```

```
## 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"
```

```
# Coordonnées des individus
head(ind$coord)
```

```
##                Dim.1        Dim.2      Dim.3      Dim.4      Dim.5
## Amsterdam  2.15731679 -0.930459730  0.3550917  1.1299565  0.5717154
## Athens     0.02775273 -0.769677026 -0.4627027 -0.4729525 -0.7749484
## Auckland   0.85266583 -0.400127954 -0.4689999  0.6422029  0.5375634
## Bangkok   -2.50964121  0.242125756 -0.6292139 -0.6167142  0.1556054
## Barcelona  0.53039715 -0.006763666 -0.6886907  0.1769032  0.1350589
## Basel      4.47802645  0.404243024  0.9224088 -0.5415714 -0.6925279
```

```
# Qualité des individus
head(ind$cos2)
```

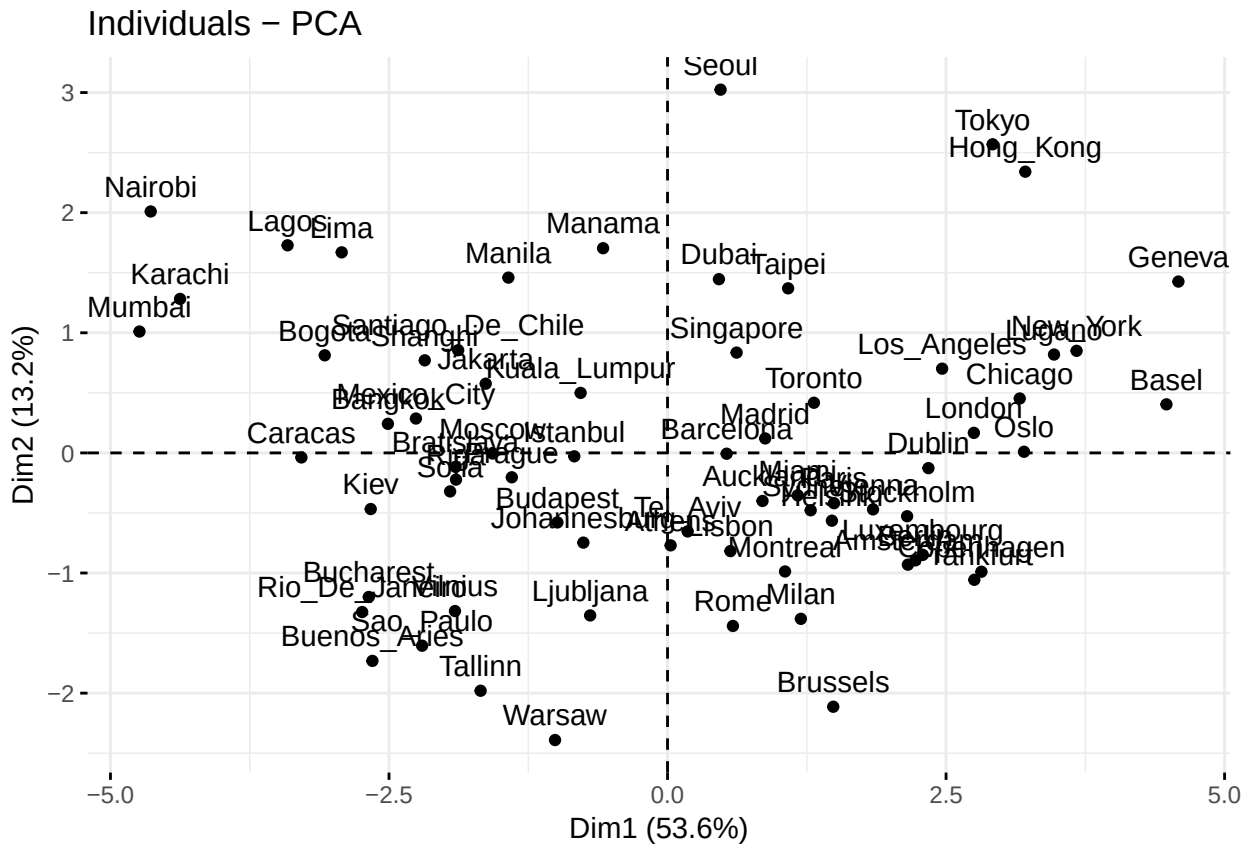
```
##                 Dim.1        Dim.2      Dim.3      Dim.4       Dim.5
## Amsterdam 0.6257611098 1.164061e-01 0.01695359 0.17167386 0.043948136
## Athens    0.0003315998 2.550468e-01 0.09217367 0.09630258 0.258552278
## Auckland  0.3687354432 8.119980e-02 0.11155850 0.20917110 0.146560354
## Bangkok   0.7788293272 7.249382e-03 0.04895710 0.04703130 0.002994113
## Barcelona 0.2524133823 4.104632e-05 0.42555774 0.02807899 0.016366518
## Basel     0.7819042085 6.371847e-03 0.03317624 0.01143646 0.018700566
```

```
# Contributions des individus
head(ind$contrib)
```

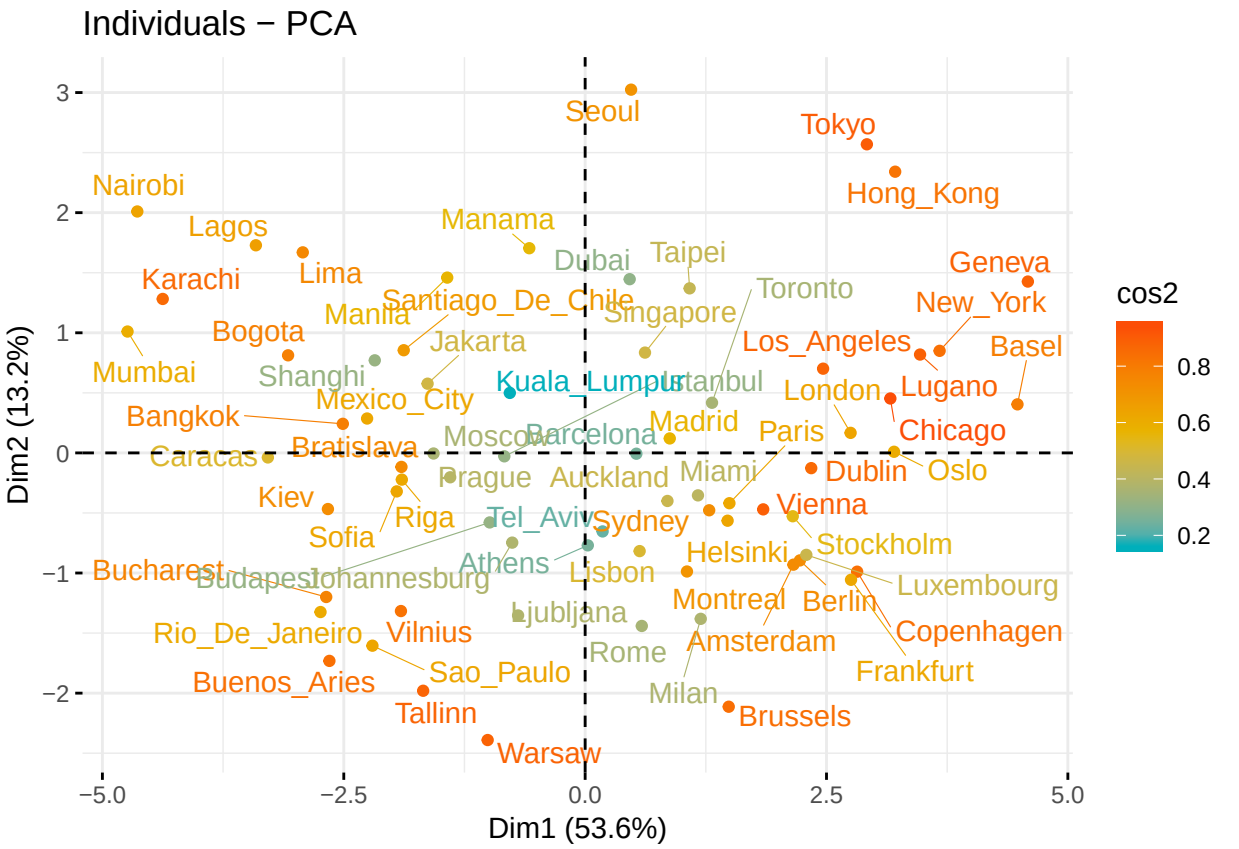
```
##                 Dim.1        Dim.2     Dim.3      Dim.4      Dim.5
## Amsterdam 1.2586143162 9.511514e-01 0.2220821 2.34339157 0.90652653
## Athens    0.0002082937 6.508359e-01 0.3770826 0.41054097 1.66558312
## Auckland  0.1966176683 1.758945e-01 0.3874164 0.75694780 0.80145681
## Bangkok   1.7032880382 6.440750e-02 0.6973153 0.69805457 0.06715369
## Barcelona 0.0760794188 5.025957e-05 0.8353741 0.05743715 0.05059028
## Basel     5.4229815951 1.795310e-01 1.4985782 0.53831046 1.33013370
```

Graphique: qualité et contribution

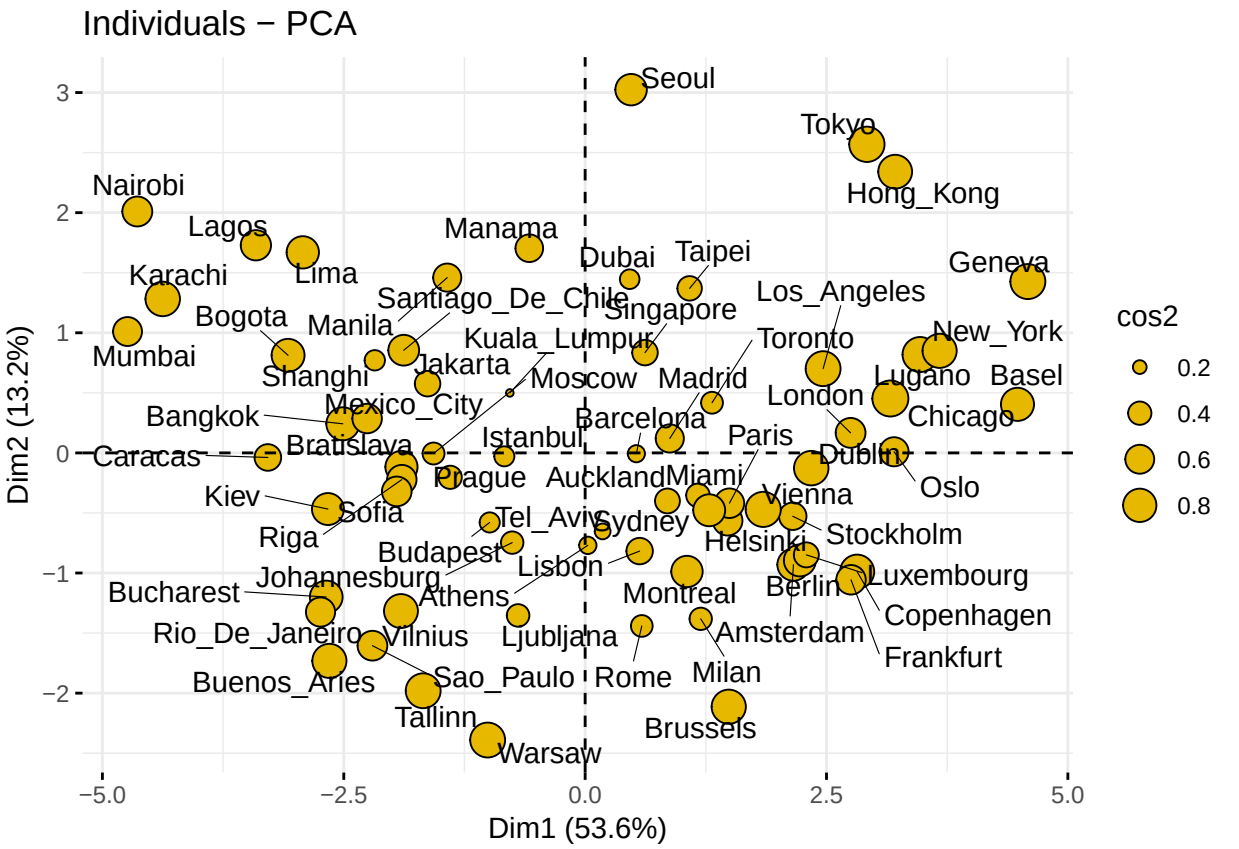
```
# Graphe des individus
fviz_pca_ind (res.pca)
```



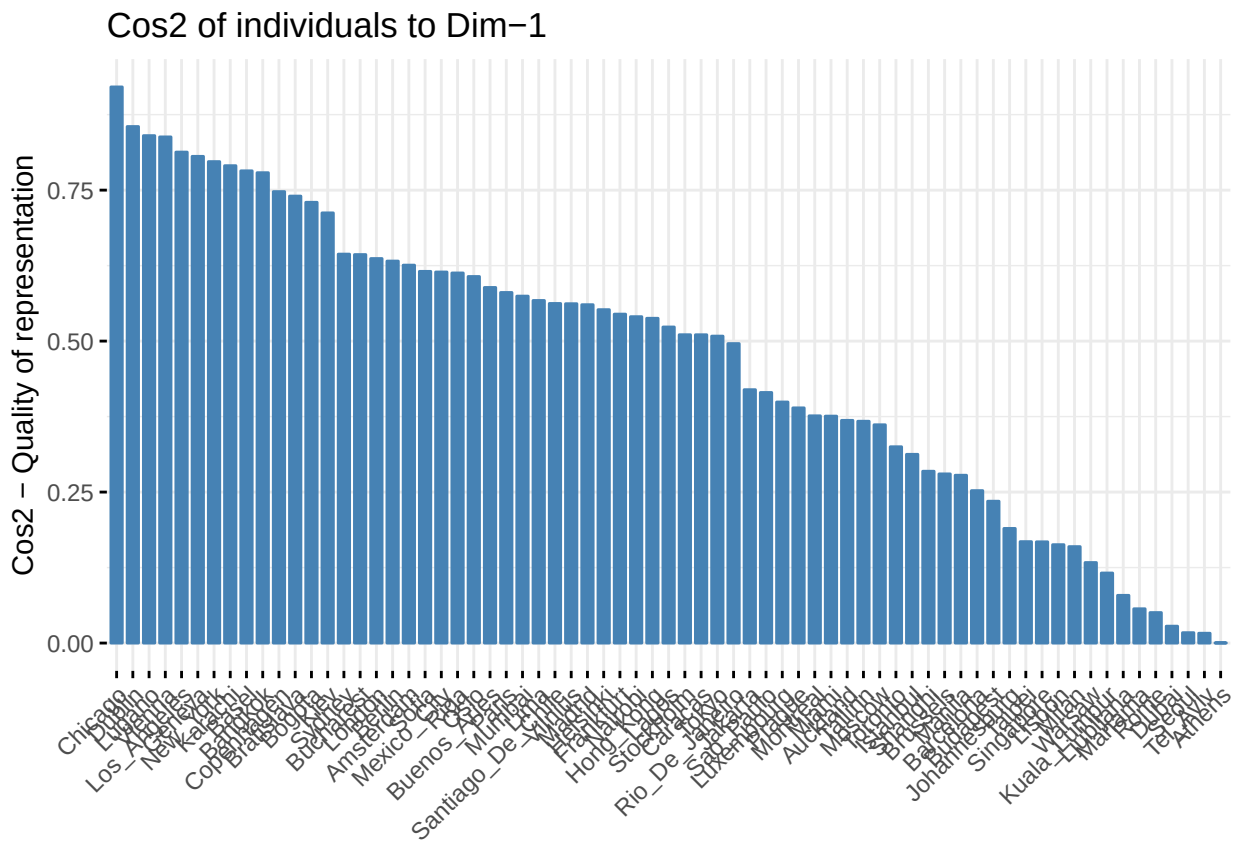

```
# Graphe des individus avec couleurs
fviz_pca_ind (res.pca, col.ind = "cos2",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE # Évite le chevauchement de texte
              )
```



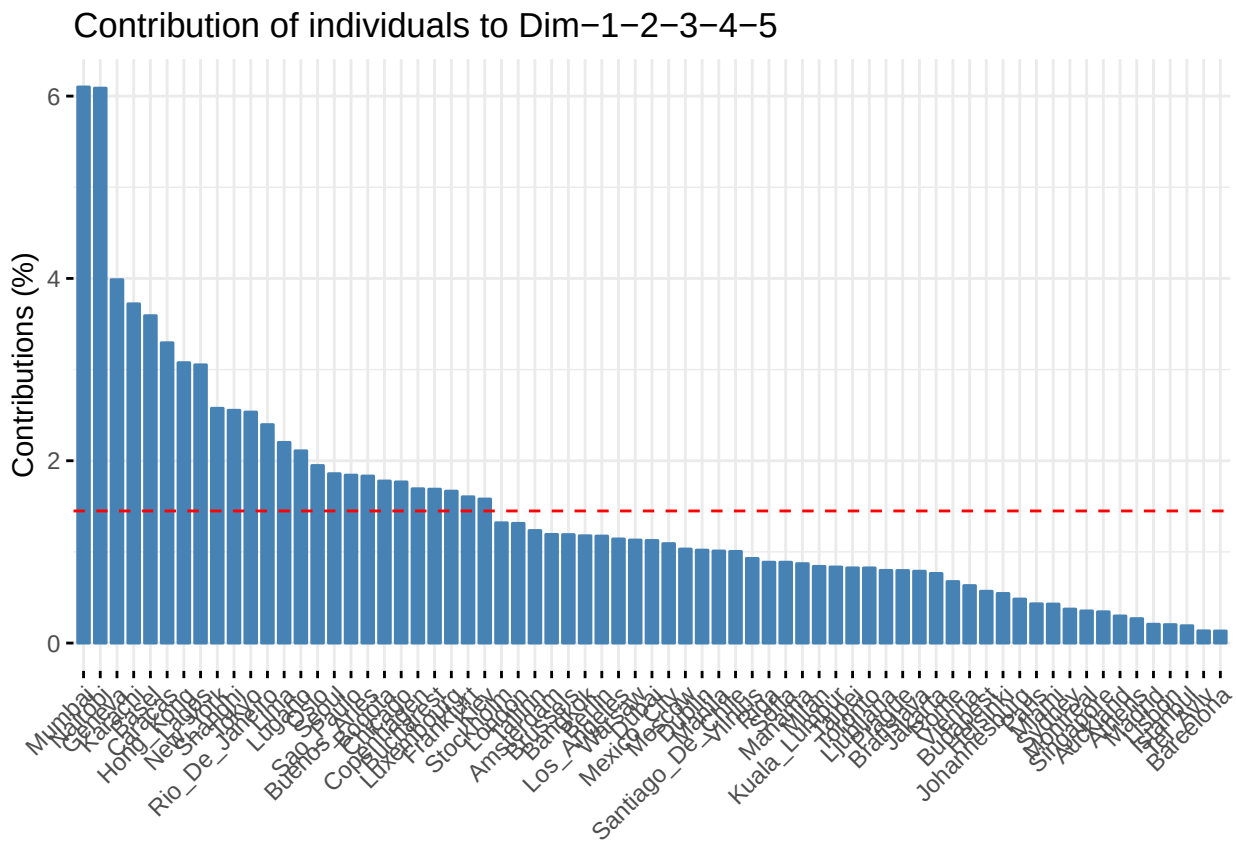

```
# Graphe des individus avec taille des points en fonction du cos2 des individus
fviz_pca_ind (res.pca, pointsize = "cos2",
              pointshape = 21, fill = "#E7B800",
              repel = TRUE # Évite le chevauchement de texte
              )
```




```r
# bar plot de la qualité de représentation (cos2) des individus
fviz_cos2(res.pca, choice = "ind")
```




```
# Contribution totale sur PC1 et PC2
fviz_contrib(res.pca, choice = "ind", axes = 1:5)
```



Colorier par les individus par groupes

Exemple avec les données Iris

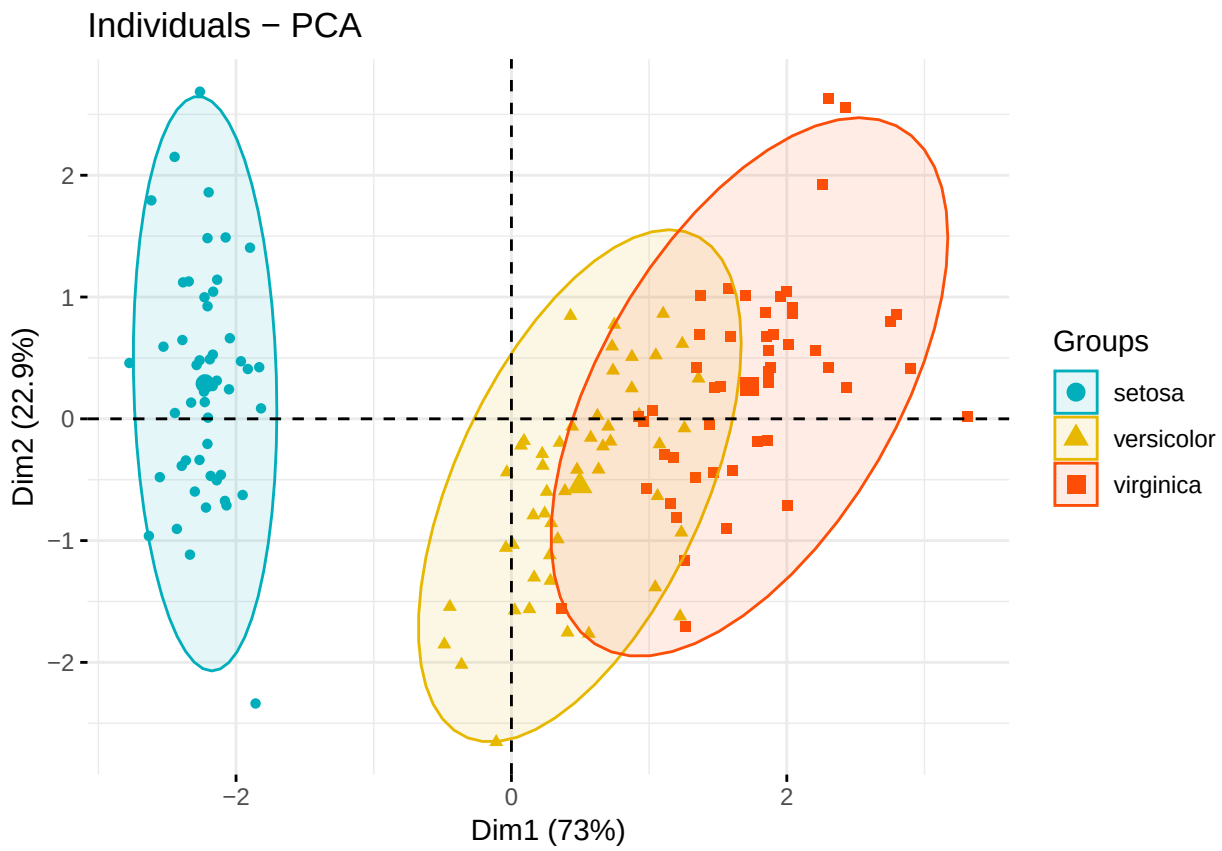
```
head(iris, 3)
```

```
##   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
```

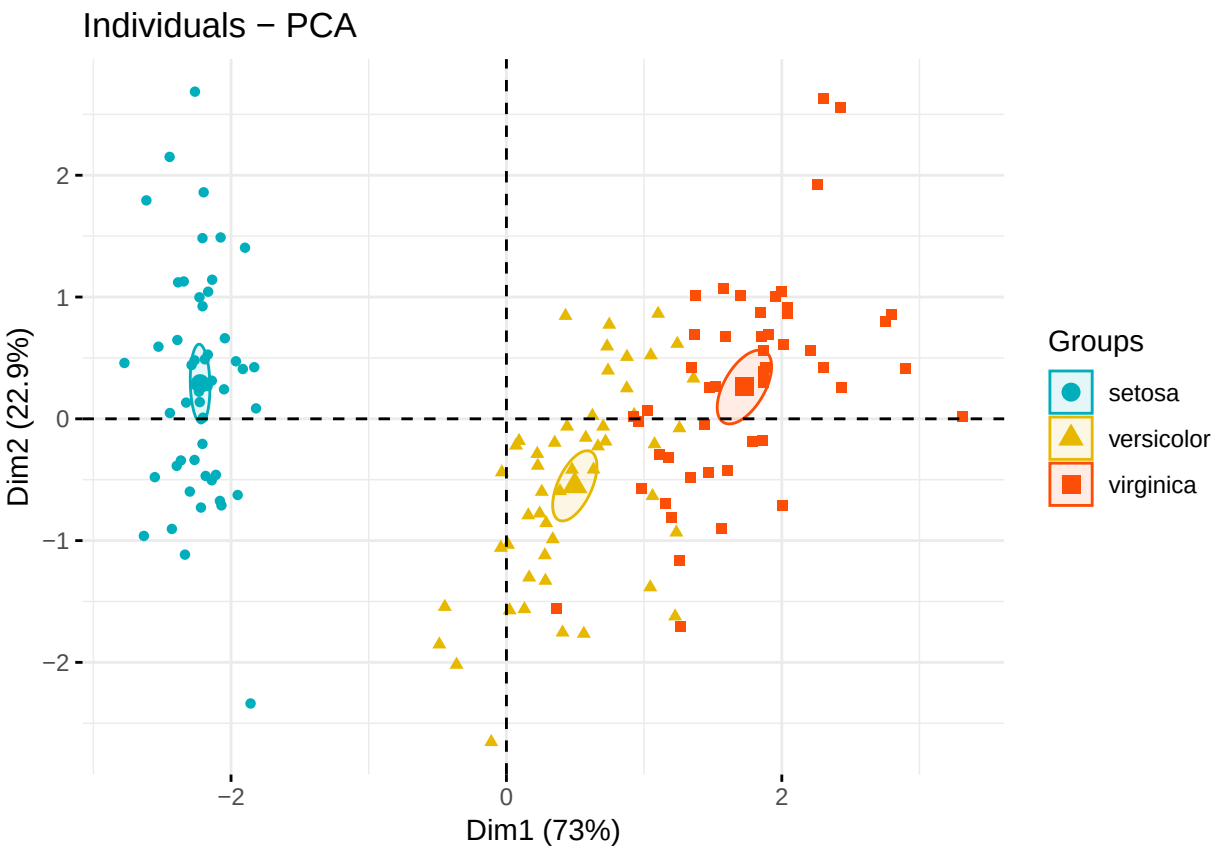
```
# La variable Species (index = 5) est supprimée
# avant l'ACP
iris.pca <- PCA(iris [, - 5], graph = FALSE)

# Dans le code R ci-dessous: l'argument habillage ou col.ind peut être utilisé pour spécifier la variab
#
# Pour ajouter une ellipse de concentration autour de chaque groupe, spécifiez l'argument addEllipses =

fviz_pca_ind(iris.pca,
             geom.ind = "point", # Montre les points seulement (mais pas le "text")
             col.ind = iris$Species, # colorer by groups
             palette = c("#00AFBB", "#E7B800", "#FC4E07"),
             addEllipses = TRUE, # Ellipses de concentration
             legend.title = "Groups"
             )
```




```
# supprimer le point moyen des groupes (centre de gravité), spécifiez l'argument mean.point = FALSE.
#
# ellipses de confiance au lieu des ellipses de concentration, utilisez ellipse.type = "confidence".
# Ajoutez des ellipses de confiance
fviz_pca_ind(iris.pca, geom.ind = "point", col.ind = iris$Species,
             palette = c("#00AFBB", "#E7B800", "#FC4E07"),
             addEllipses = TRUE, ellipse.type = "confidence",
             legend.title = "Groups"
             )
```


Individuals − PCA
Dim2 (22.9%)
Dim1 (73%)
2
1
0
−1
−2
−2
0
2
Groups
setosa
versicolor
virginica