PCA

Max Pereira

14/06/2020

PCA - Principal Component Analysis

A Análise de Componentes Principais (ACP) ou Principal Component Analysis (PCA) é um procedimento matemático que utiliza uma transformação ortogonal (ortogonalização de vetores) para converter um conjunto de observações de variáveis possivelmente correlacionadas num conjunto de valores de variáveis linearmente não correlacionadas chamadas de componentes principais

Carregando os pacotes

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

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

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

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

Dataset

```
##           X100m Long.jump Shot.put High.jump X400m X110m.hurdle Discus
## SEBRLE    11.04      7.58    14.83      2.07 49.81        14.69  43.75
## CLAY      10.76      7.40    14.26      1.86 49.37        14.05  50.72
## BERNARD   11.02      7.23    14.25      1.92 48.93        14.99  40.87
## YURKOV    11.34      7.09    15.19      2.10 50.42        15.31  46.26
## ZSIVOCZKY 11.13      7.30    13.48      2.01 48.62        14.17  45.67
## McMULLEN  10.83      7.31    13.76      2.13 49.91        14.38  44.41
##           Pole.vault Javeline X1500m Rank Points Competition
## SEBRLE          5.02    63.19  291.7    1   8217    Decastar
## CLAY            4.92    60.15  301.5    2   8122    Decastar
## BERNARD         5.32    62.77  280.1    4   8067    Decastar
## YURKOV          4.72    63.44  276.4    5   8036    Decastar
## ZSIVOCZKY       4.42    55.37  268.0    7   8004    Decastar
## McMULLEN        4.42    56.37  285.1    8   7995    Decastar
```

Terminologia do PCA

Indivíduos Ativos: usados durante o processo de PCA (linhas 1:23) Indivíduos suplementares: as coordenadas desses indivíduos serão estimadas usando as informações do PCA e dos parâmetros obtidos dos indivíduos e variáveis ativas Variáveis ativas: variáveis que são usadas no processo de PCA (colunas 1:10) Variáveis suplementares: as coordenadas dessas variáveis serão estimadas

```r
decathlon2.ativo <- decathlon2[1:23, 1:10]
head(decathlon2.ativo[,1:6],4)
```

```
##         X100m Long.jump Shot.put High.jump X400m X110m.hurdle
## SEBRLE  11.04      7.58    14.83      2.07 49.81        14.69
## CLAY    10.76      7.40    14.26      1.86 49.37        14.05
## BERNARD 11.02      7.23    14.25      1.92 48.93        14.99
## YURKOV  11.34      7.09    15.19      2.10 50.42        15.31
```

Nomralização dos dados

O objetivo é colocar todas as variáveis em uma mesma escala Obter desvio padrão igual a 1 e média igual a zero. Os dados podem ser transformados da seguinte forma: (xi - media(x))/sd(x) A função scale() pode ser usada para essa normalização. A função PCA() [FactoMineR] normaliza os dados automaticamente durante o PCA.

```r
res.PCA <- PCA(decathlon2.ativo, graph=FALSE)
print(res.PCA)
```

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 23 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"
```

Eigenvalues / Variâncias

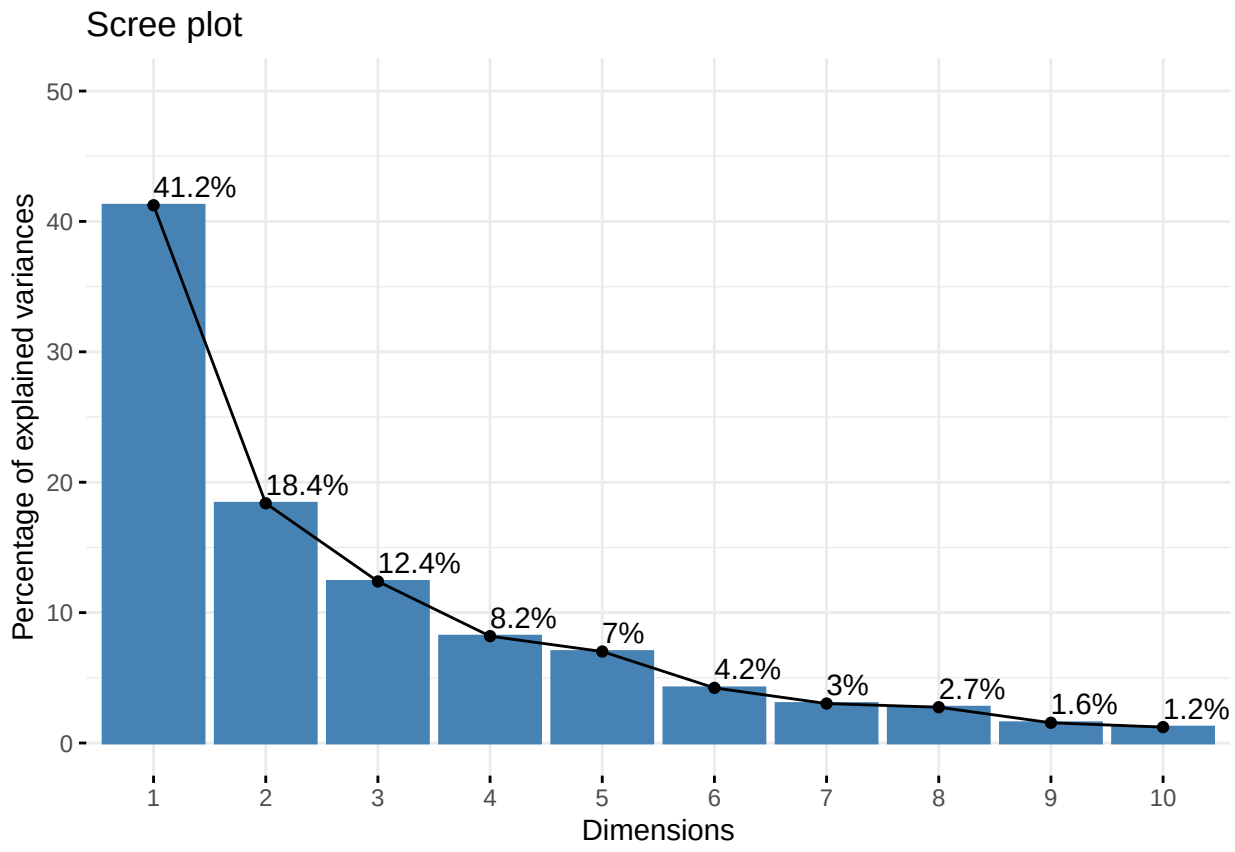
Os valores (eigenvalues) medem o total de variação em cada componente principal. Os valores são maiores para os primeiros PCs e menores para os subsequentes. Ou seja, os primeiros PCs correspondem as direções com o total máximo de variação no conjunto de dados. A análise dos valores (eigenvalues) é feita para determinar o número dos componentes principais que serão considerados. Um eigenvalue > 1 indica uma variância maior (apenas para dados normalizados) Também é possível limitar o número de PCs por uma fração da variância total, ou seja, se 70% é satisfatório use o número de componentes que atingem esse valor.

```r
eig_val <- get_eigenvalue(res.PCA)
eig_val
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1   4.1242133        41.242133                    41.24213
## Dim.2   1.8385309        18.385309                    59.62744
## Dim.3   1.2391403        12.391403                    72.01885
## Dim.4   0.8194402         8.194402                    80.21325
## Dim.5   0.7015528         7.015528                    87.22878
## Dim.6   0.4228828         4.228828                    91.45760
## Dim.7   0.3025817         3.025817                    94.48342
## Dim.8   0.2744700         2.744700                    97.22812
## Dim.9   0.1552169         1.552169                    98.78029
## Dim.10  0.1219710         1.219710                   100.00000
```

Gráfico Scree

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



Resultados

```
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"
```

```
head(var$coord)
```

```
##                  Dim.1       Dim.2      Dim.3       Dim.4      Dim.5
## X100m        -0.8506257 -0.17939806  0.3015564  0.03357320 -0.1944440
## Long.jump     0.7941806  0.28085695 -0.1905465 -0.11538956  0.2331567
## Shot.put      0.7339127  0.08540412  0.5175978  0.12846837 -0.2488129
## High.jump     0.6100840 -0.46521415  0.3300852  0.14455012  0.4027002
## X400m        -0.7016034  0.29017826  0.2835329  0.43082552  0.1039085
## X110m.hurdle -0.7641252 -0.02474081  0.4488873 -0.01689589  0.2242200
```

```
head(var$cos2)
```

```
##                  Dim.1        Dim.2      Dim.3        Dim.4      Dim.5
## X100m        0.7235641 0.0321836641 0.09093628 0.0011271597 0.03780845
## Long.jump    0.6307229 0.0788806285 0.03630798 0.0133147506 0.05436203
## Shot.put     0.5386279 0.0072938636 0.26790749 0.0165041211 0.06190783
## High.jump    0.3722025 0.2164242070 0.10895622 0.0208947375 0.16216747
## X400m        0.4922473 0.0842034209 0.08039091 0.1856106269 0.01079698
## X110m.hurdle 0.5838873 0.0006121077 0.20149984 0.0002854712 0.05027463
```

```
head(var$contrib)
```

```
##                  Dim.1      Dim.2     Dim.3       Dim.4     Dim.5
## X100m        17.544293  1.7505098  7.338659  0.13755240  5.389252
## Long.jump    15.293168  4.2904162  2.930094  1.62485936  7.748815
## Shot.put     13.060137  0.3967224 21.620432  2.01407269  8.824401
## High.jump     9.024811 11.7715838  8.792888  2.54987951 23.115504
## X400m        11.935544  4.5799296  6.487636 22.65090599  1.539012
## X110m.hurdle 14.157544  0.0332933 16.261261  0.03483735  7.166193
```

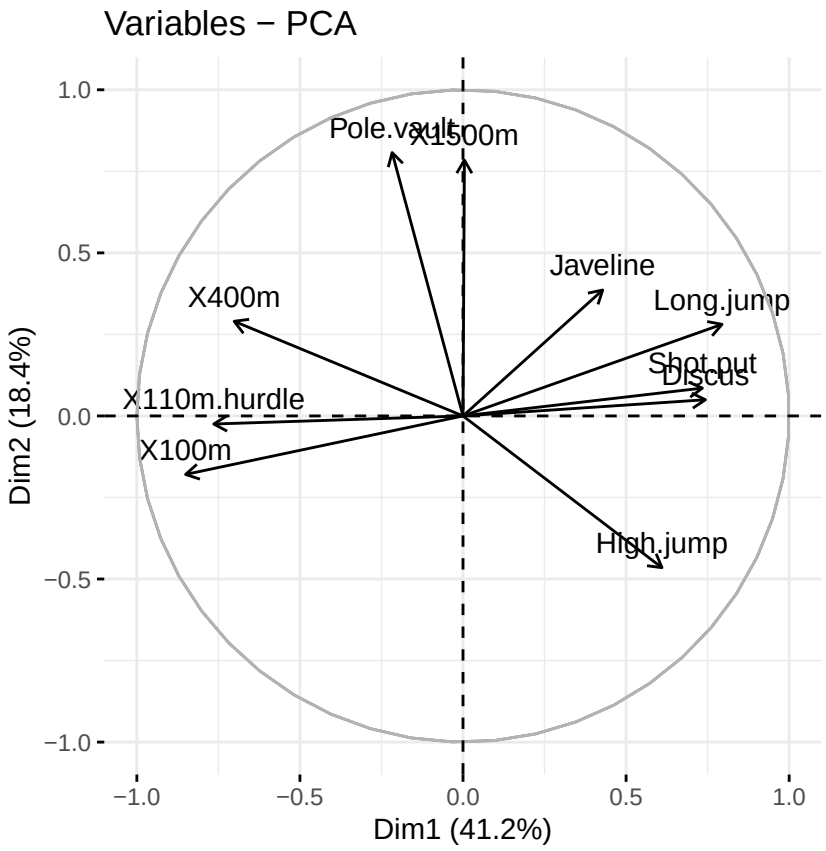
Círculo de correlação

Variáveis correlacionadas positivamente são agrupadas juntas. Variáveis correlacionadas negativamente são posicionadas nos lados opostos (quadrantes opostos). A distância entre as variáveis e a origem mede a qualidade das variáveis no mapa de fator. As variáveis que estão longe da origem são melhores representadas no mapa de fator.

```
head(var$cor)
```

```
##                  Dim.1       Dim.2      Dim.3       Dim.4      Dim.5
## X100m        -0.8506257 -0.17939806  0.3015564  0.03357320 -0.1944440
## Long.jump     0.7941806  0.28085695 -0.1905465 -0.11538956  0.2331567
## Shot.put      0.7339127  0.08540412  0.5175978  0.12846837 -0.2488129
## High.jump     0.6100840 -0.46521415  0.3300852  0.14455012  0.4027002
## X400m        -0.7016034  0.29017826  0.2835329  0.43082552  0.1039085
## X110m.hurdle -0.7641252 -0.02474081  0.4488873 -0.01689589  0.2242200
```

```
fviz_pca_var(res.PCA, col.var = "black")
```



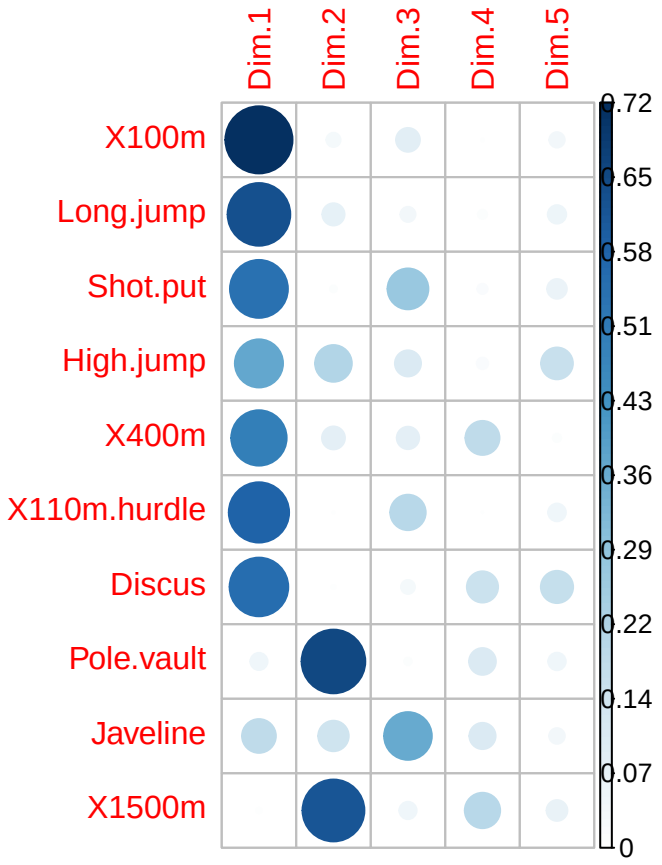
Qualidade da representação

A qualidade da representação das variáveis no mapa de fator é chamado de cos2 (cosseno quadrático)

```
head(var$cos2)
```

```
##                  Dim.1        Dim.2      Dim.3        Dim.4      Dim.5
## X100m        0.7235641 0.0321836641 0.09093628 0.0011271597 0.03780845
## Long.jump    0.6307229 0.0788806285 0.03630798 0.0133147506 0.05436203
## Shot.put     0.5386279 0.0072938636 0.26790749 0.0165041211 0.06190783
## High.jump    0.3722025 0.2164242070 0.10895622 0.0208947375 0.16216747
## X400m        0.4922473 0.0842034209 0.08039091 0.1856106269 0.01079698
## X110m.hurdle 0.5838873 0.0006121077 0.20149984 0.0002854712 0.05027463
```

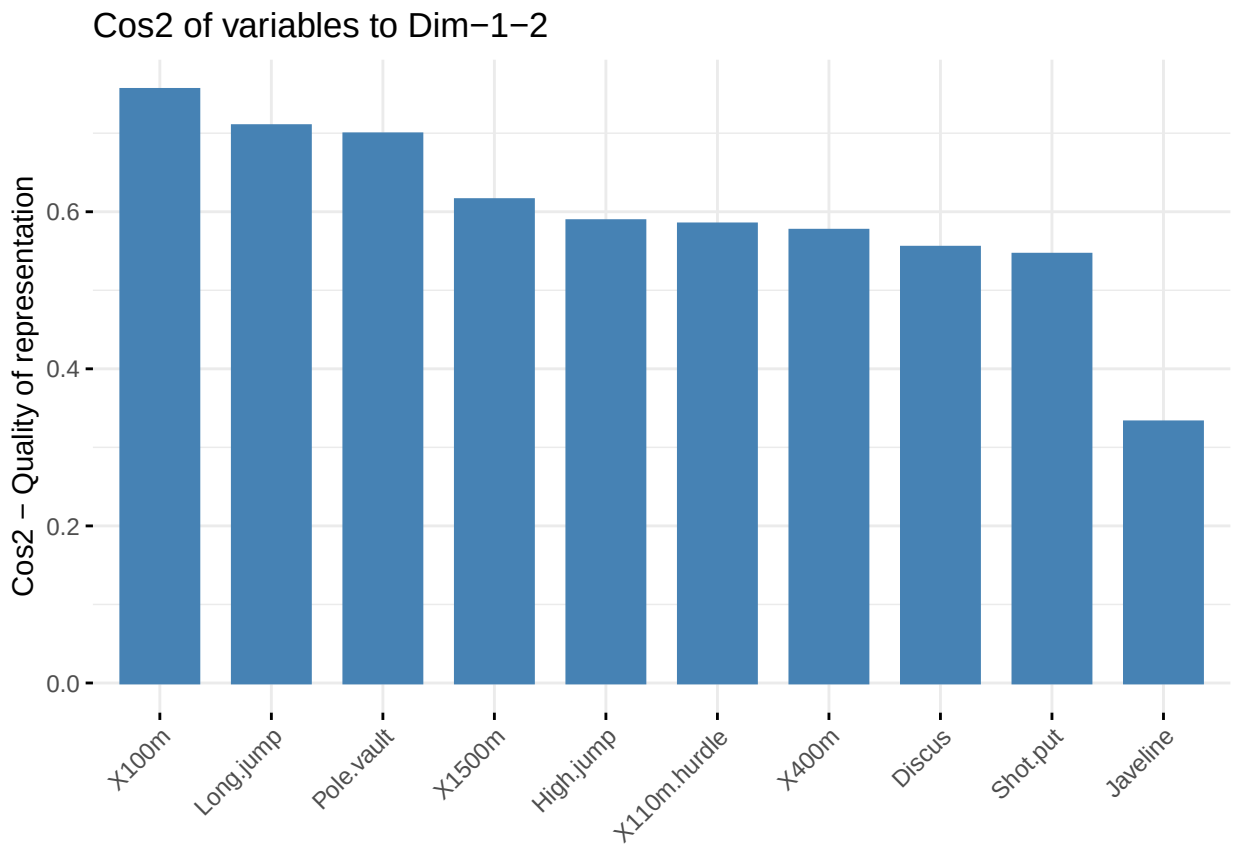
```
corrplot(var$cos2, is.corr = FALSE)
```



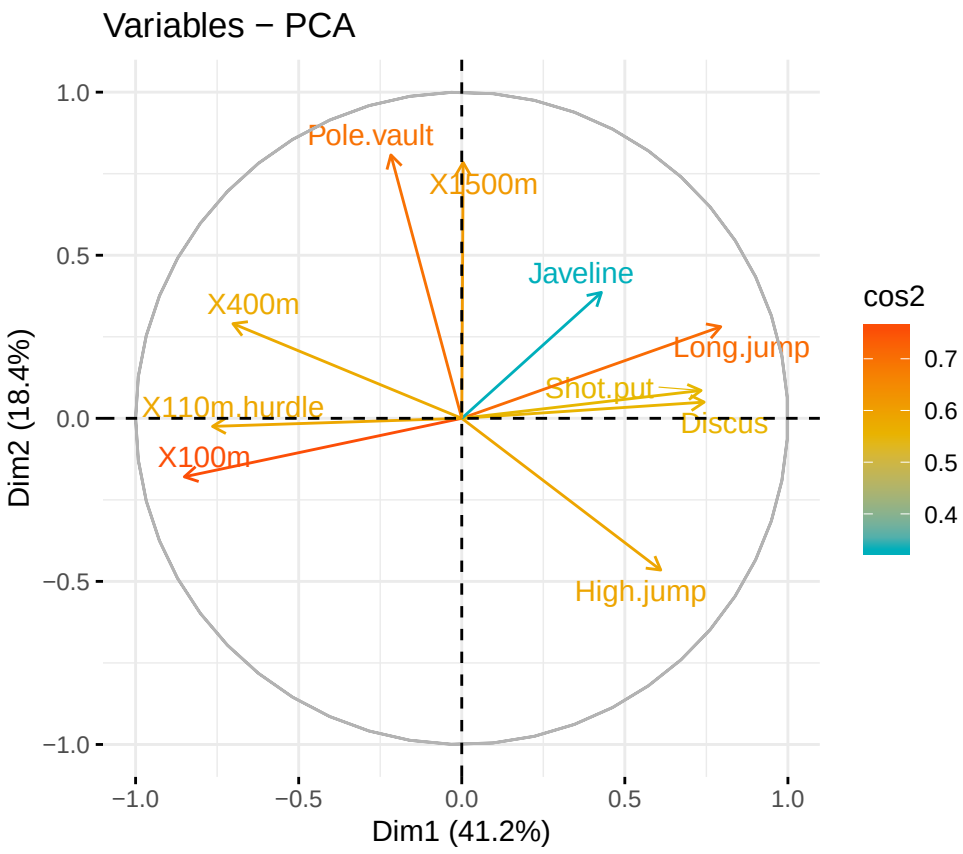
Total do cos2 nas dimensões Dim.1 e Dim.2

Um valor alto para o cos2 indica uma boa representação da variável no componente principal. Nesse caso, a variável é posicionada perto da circunferência do círculo de correlação. Um valor baixo para o cos2 indica que a variável não é perfeitamente representada pelos PCs. Nesse caso, a variável está perto do centro do círculo. Para uma determinada variável, a soma dos cos2 para todos os PCs é igual a um. Se a variável é perfeitamente representada por somente dois PCs (Dim.1 e Dim.2), a soma do cos2, nesses dois PCs, é igual a um. Nesse caso, as variáveis serão posicionadas no Cículo de correlações. Para algumas variáveis, mais do que 2 componentes podem ser necessários para representar perfeitamente os dados. Nesse caso, as variáveis são posicionadas dentro do círculo de correlações.

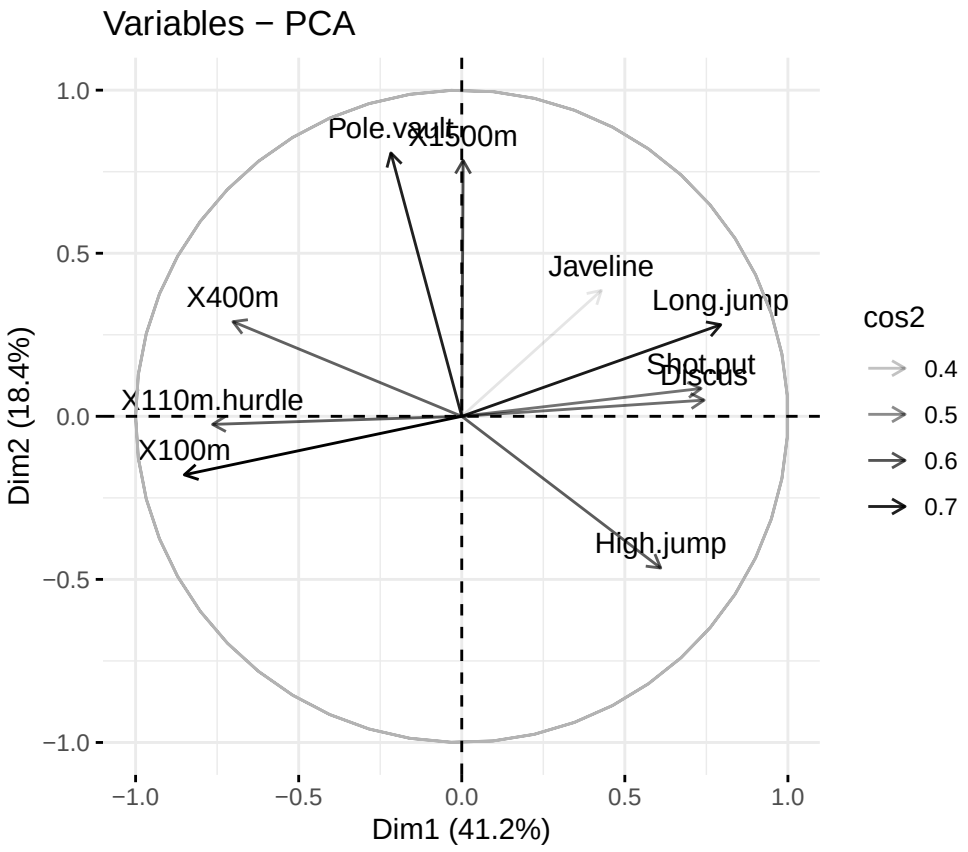
```
fviz_cos2(res.PCA, choice = "var", axes = 1:2)
```




```
fviz_pca_var(res.PCA, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Evita sobreposição dos textos
             )
```




```
fviz_pca_var(res.PCA, alpha.var = "cos2")
```



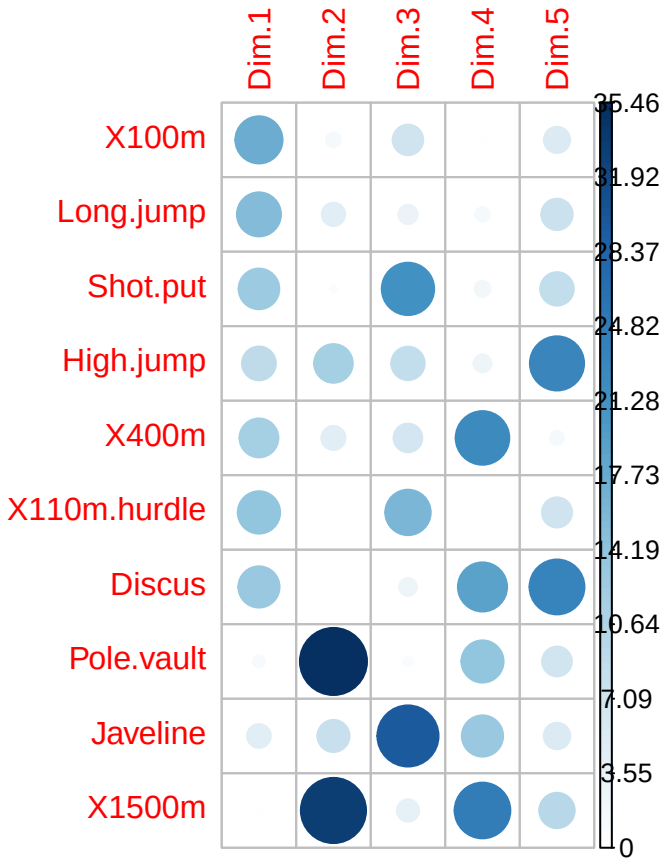
Contribuições das variáveis para os PCs (expressas em percentual)

As variáveis que são correlacionadas com o PC1 (Dim.1) e PC2 (Dim.2) são as mais importantes para explicar a variabilidade no conjunto de dados. As variáveis que não estão correlaciondas com qualquer PC ou correlaciondas com as últimas dimensões são variáveis com baixa contribuição e podem ser removidas para simplificar a análise.

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

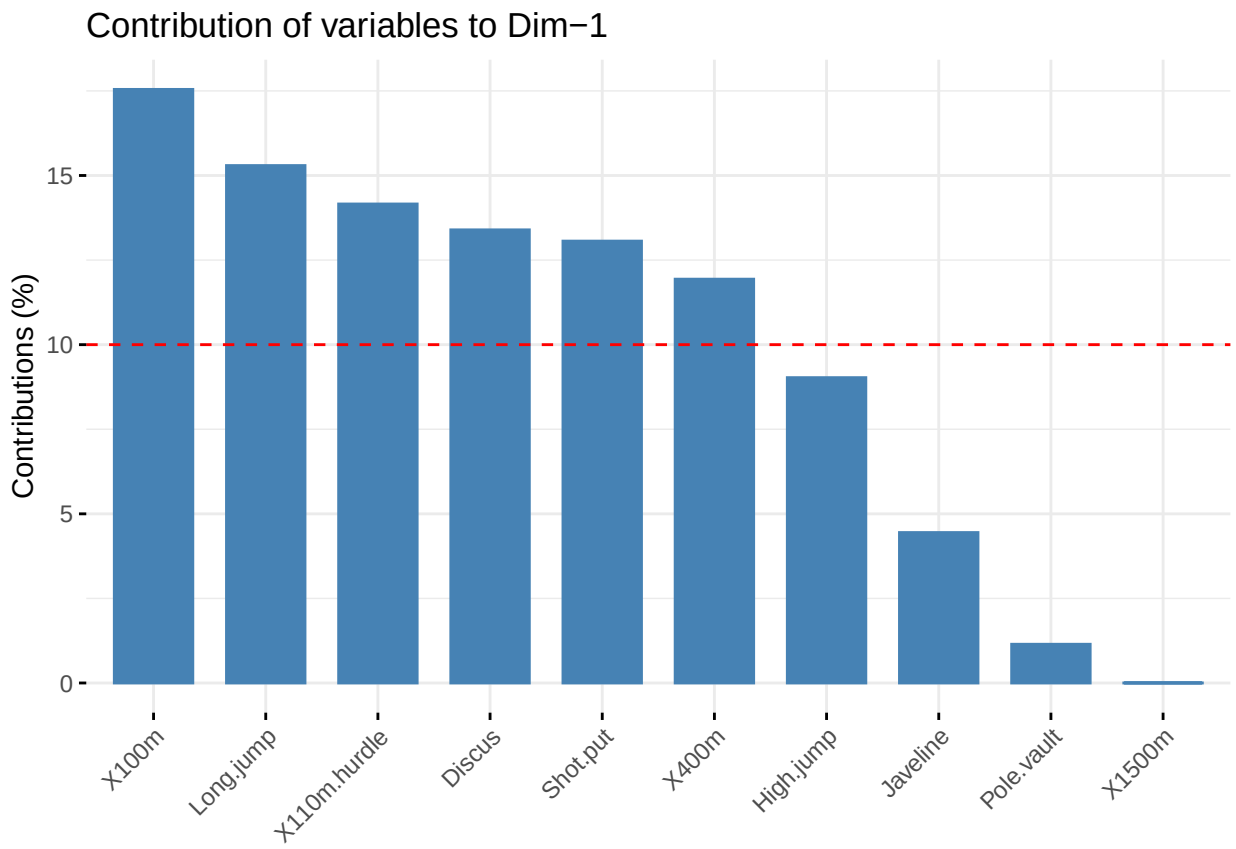
```
##               Dim.1      Dim.2     Dim.3     Dim.4     Dim.5
## X100m     17.544293  1.7505098  7.338659 0.1375524  5.389252
## Long.jump 15.293168  4.2904162  2.930094 1.6248594  7.748815
## Shot.put  13.060137  0.3967224 21.620432 2.0140727  8.824401
## High.jump  9.024811 11.7715838  8.792888 2.5498795 23.115504
```

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



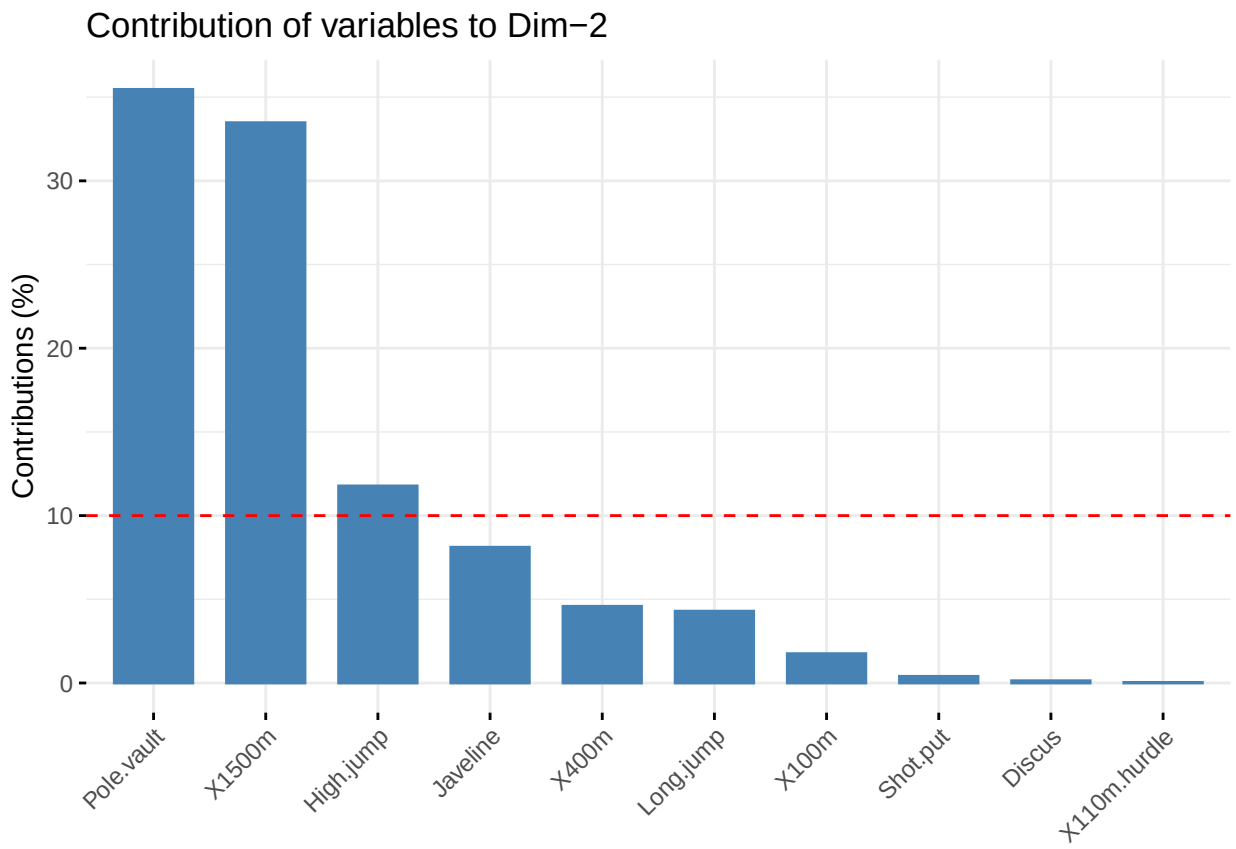
Contribuições das variáveis para o PC1

```
fviz_contrib(res.PCA, choice = "var", axes = 1, top = 10)
```



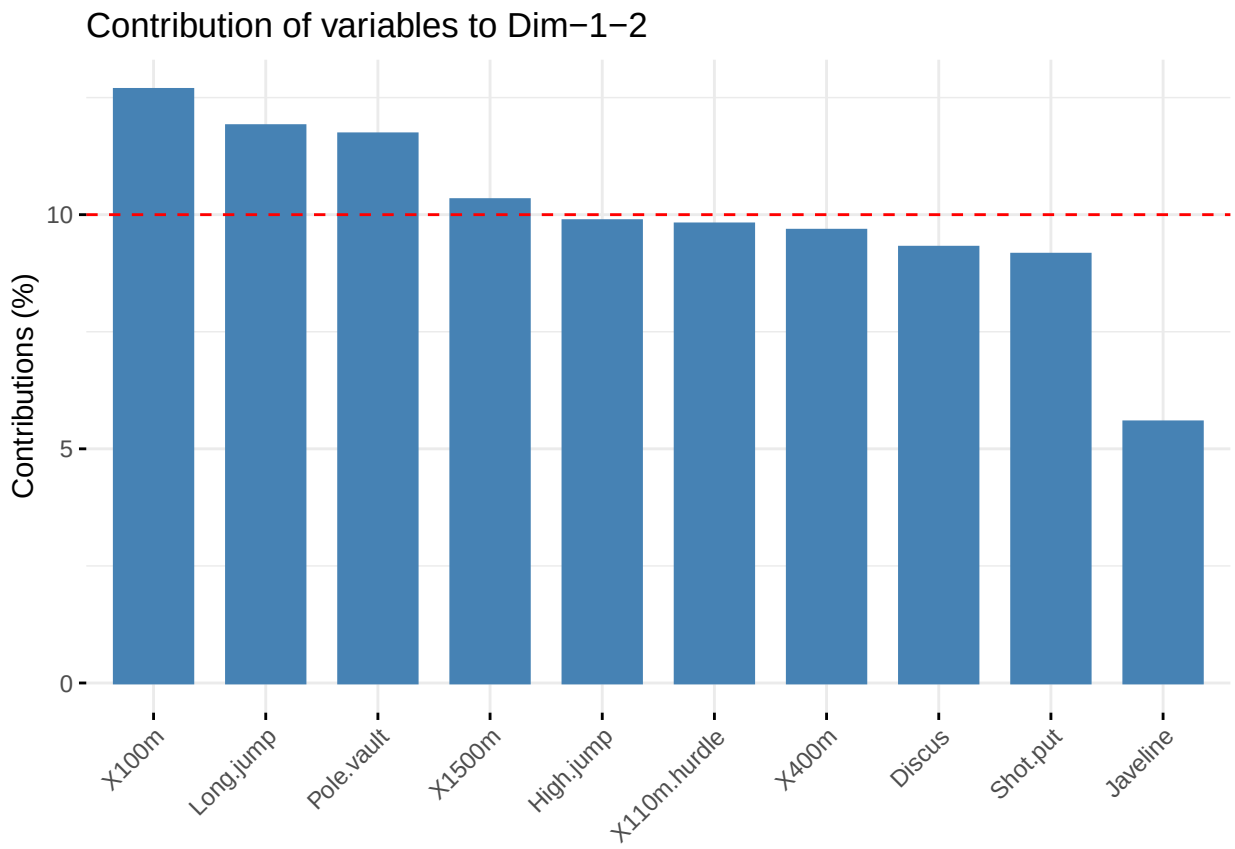
Contribuições das variáveis para o PC2

```
fviz_contrib(res.PCA, choice = "var", axes = 2, top = 10)
```



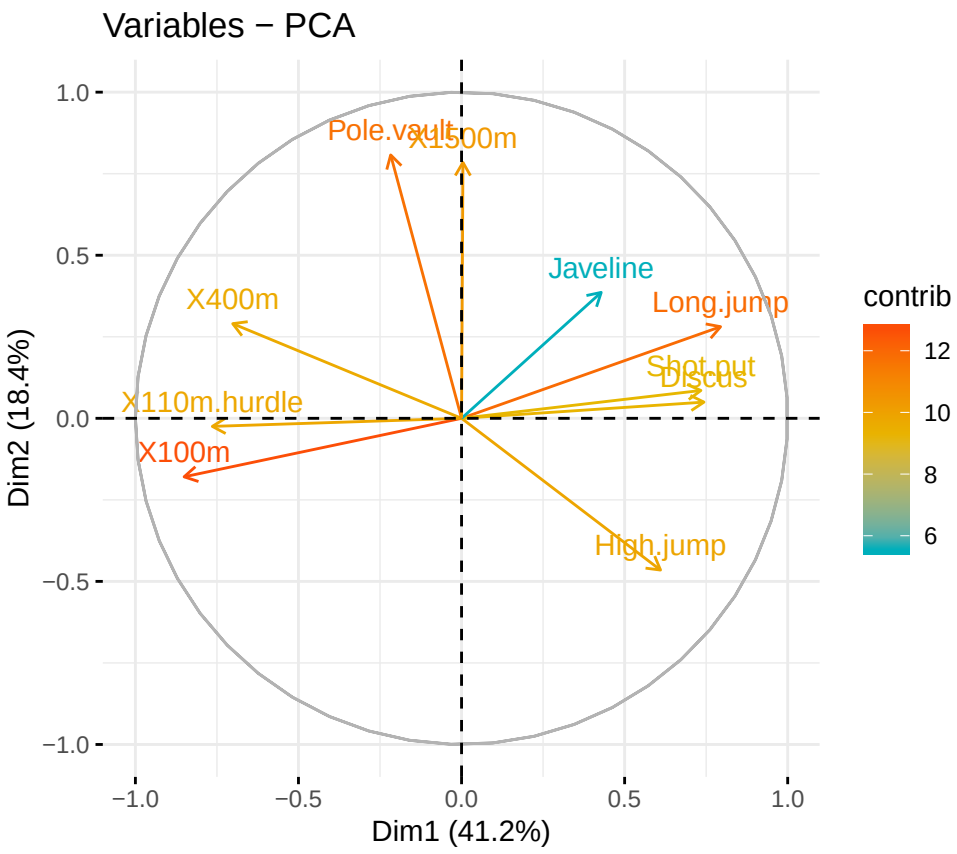
A contribuição total para o PC1 e PC2

```
fviz_contrib(res.PCA, choice = "var", axes=1:2, top = 10)
```

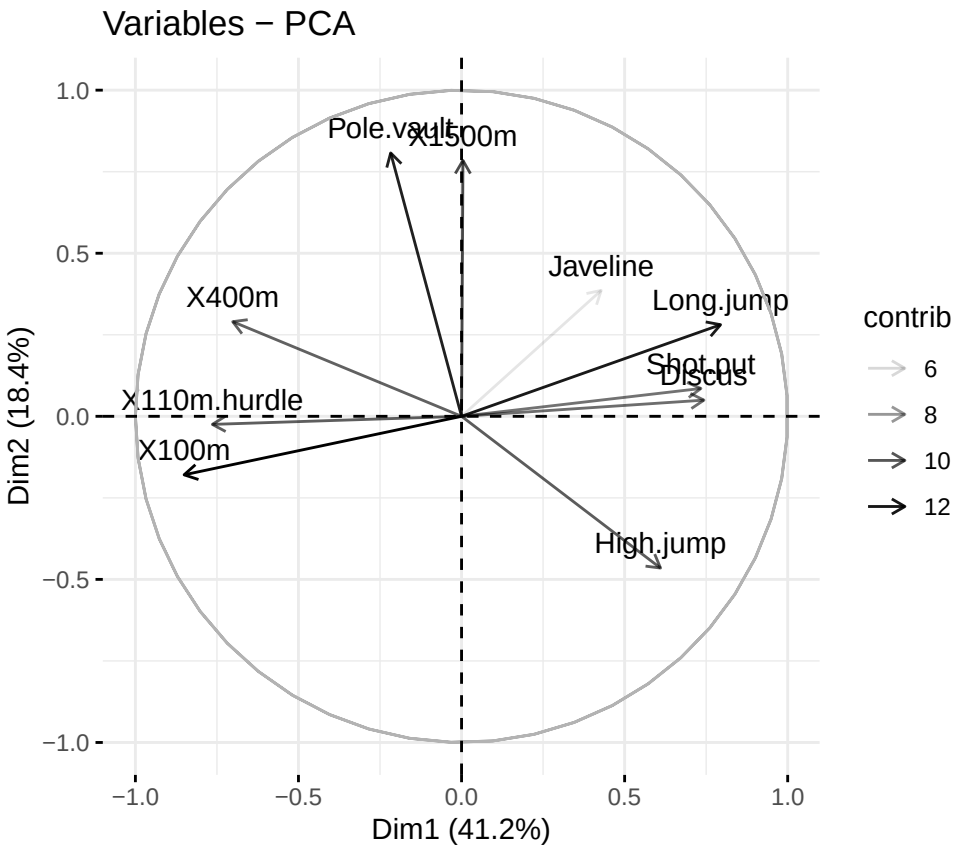


As variáveis mais importantes podem ser destacadas no gráfico de correlações

```
fviz_pca_var(res.PCA, col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"))
```




```
fviz_pca_var(res.PCA, alpha.var = "contrib")
```



Descrição da dimensão

Pode ser usado para identificar as variáveis mais significantes associadas a um dado componente principal.

```
res.desc <- dimdesc(res.PCA, axes=c(1,2), proba=0.05)
res.desc$Dim.1
```

```
## $quanti
##              correlation      p.value
## Long.jump      0.7941806 6.059893e-06
## Discus         0.7432090 4.842563e-05
## Shot.put       0.7339127 6.723102e-05
## High.jump      0.6100840 1.993677e-03
## Javeline       0.4282266 4.149192e-02
## X400m         -0.7016034 1.910387e-04
## X110m.hurdle  -0.7641252 2.195812e-05
## X100m         -0.8506257 2.727129e-07
##
## attr(,"class")
## [1] "condes" "list "
```

```
res.desc$Dim.2
```

```
## $quanti
```

```
##            correlation      p.value
## Pole.vault   0.8074511 3.205016e-06
## X1500m       0.7844802 9.384747e-06
## High.jump   -0.4652142 2.529390e-02
##
## attr(,"class")
## [1] "condes" "list "
```

Resultados

```
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"
```

```
head(ind$coord)
```

```
##                Dim.1      Dim.2      Dim.3       Dim.4       Dim.5
## SEBRLE     0.1955047  1.5890567  0.6424912  0.08389652  1.16829387
## CLAY       0.8078795  2.4748137 -1.3873827  1.29838232 -0.82498206
## BERNARD   -1.3591340  1.6480950  0.2005584 -1.96409420  0.08419345
## YURKOV    -0.8889532 -0.4426067  2.5295843  0.71290837  0.40782264
## ZSIVOCZKY -0.1081216 -2.0688377 -1.3342591 -0.10152796 -0.20145217
## McMULLEN   0.1212195 -1.0139102 -0.8625170  1.34164291  1.62151286
```

```
head(ind$cos2)
```

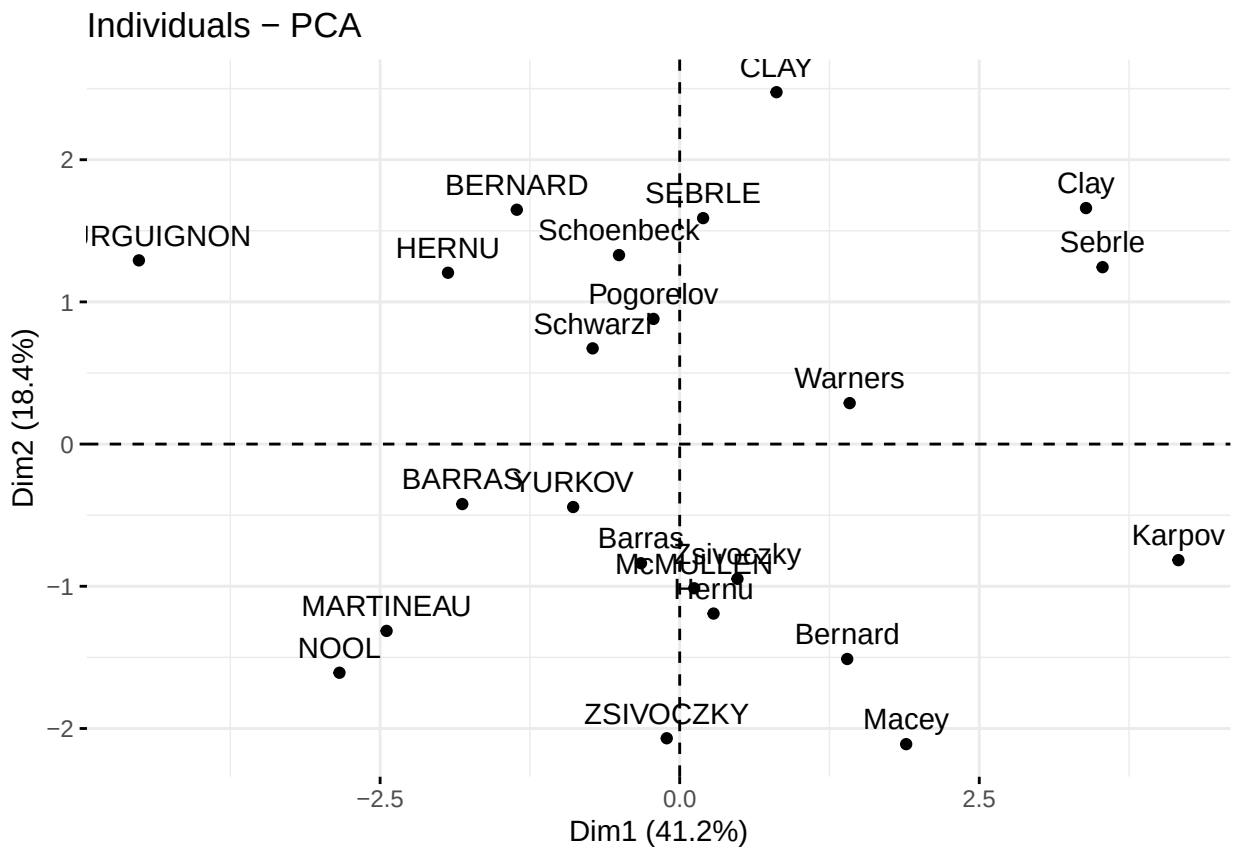
```
##                 Dim.1      Dim.2       Dim.3       Dim.4        Dim.5
## SEBRLE    0.007530179 0.49747323 0.081325232 0.001386688 0.2689026575
## CLAY      0.048701249 0.45701660 0.143628117 0.125791741 0.0507850580
## BERNARD   0.197199804 0.28996555 0.004294015 0.411819183 0.0007567259
## YURKOV    0.096109800 0.02382571 0.778230322 0.061812637 0.0202279796
## ZSIVOCZKY 0.001574385 0.57641944 0.239754152 0.001388216 0.0054654972
## McMULLEN  0.002175437 0.15219499 0.110137872 0.266486530 0.3892621478
```

```
head(ind$contrib)
```

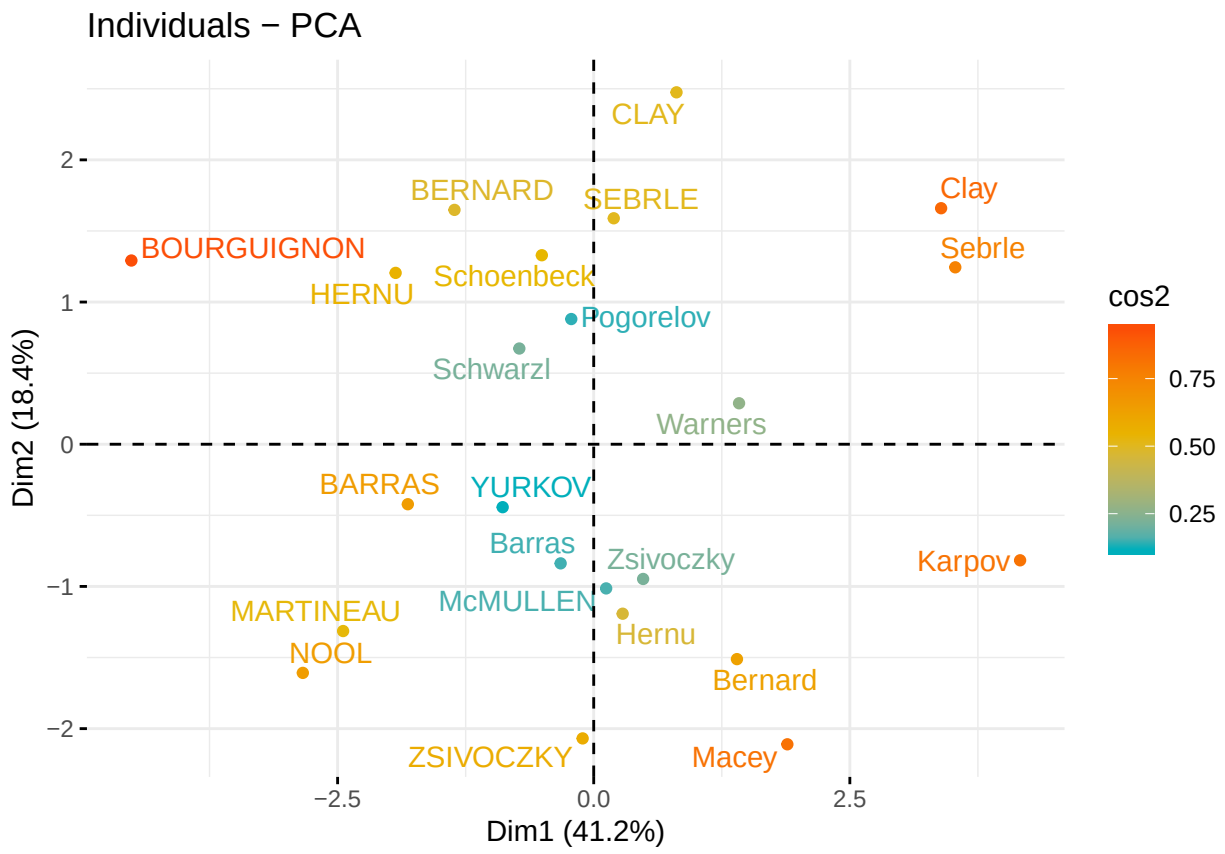
```
##                Dim.1      Dim.2      Dim.3       Dim.4       Dim.5
## SEBRLE    0.04029447  5.9714533  1.4483919  0.03734589  8.45894063
## CLAY      0.68805664 14.4839248  6.7537381  8.94458283  4.21794385
## BERNARD   1.94740183  6.4234107  0.1411345 20.46819433  0.04393073
## YURKOV    0.83308415  0.4632733 22.4517396  2.69663605  1.03075263
## ZSIVOCZKY 0.01232413 10.1217143  6.2464325  0.05469230  0.25151025
## McMULLEN  0.01549089  2.4310854  2.6102794  9.55055888 16.29493304
```

Qualidade e contribuição (plot)

```
fviz_pca_ind(res.PCA)
```

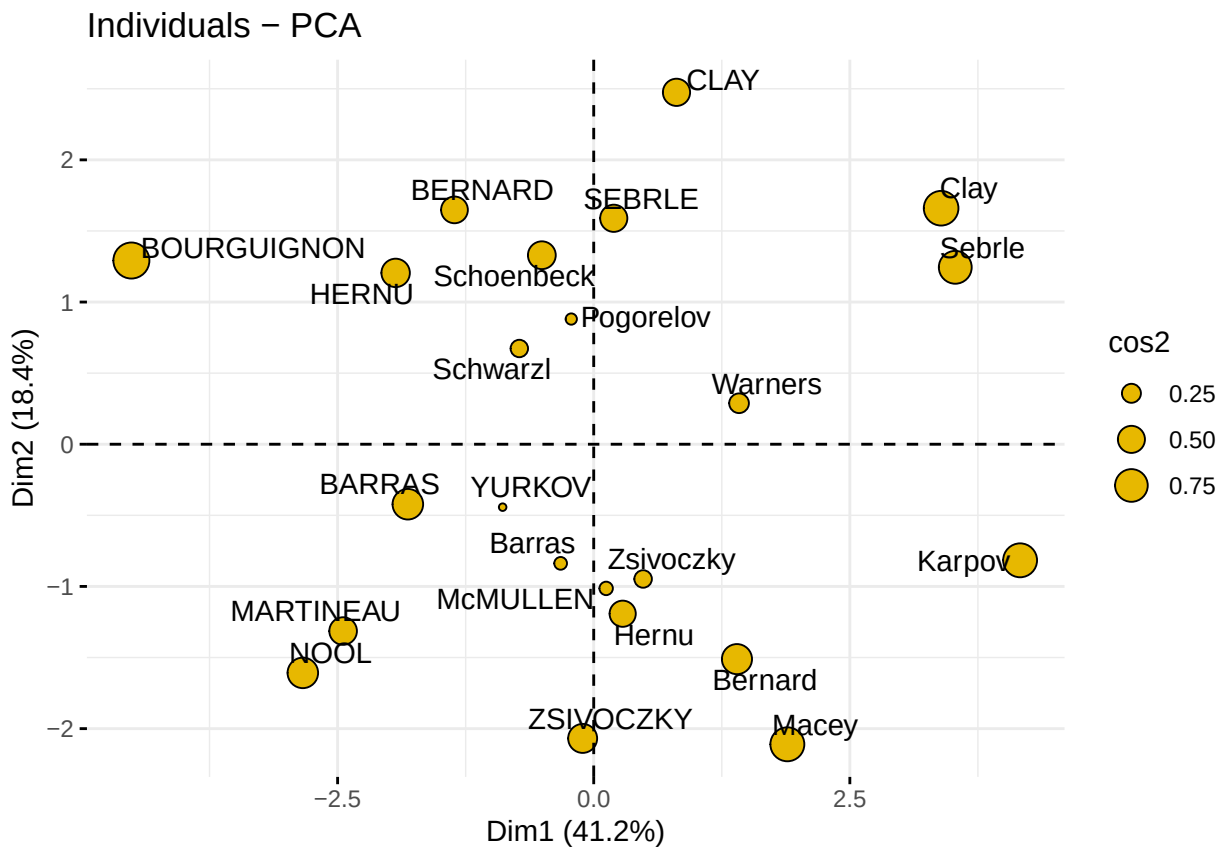



```
fviz_pca_ind(res.PCA, col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```

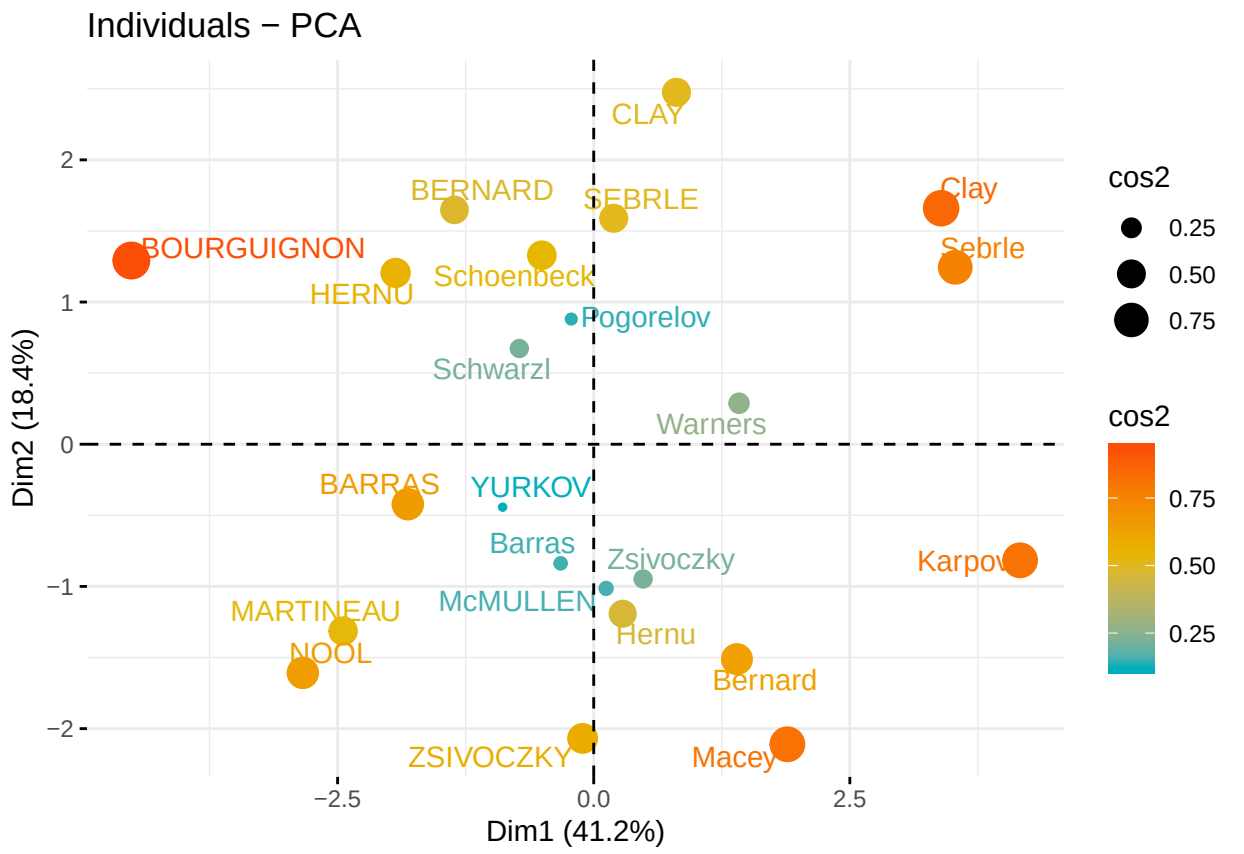


Pontos

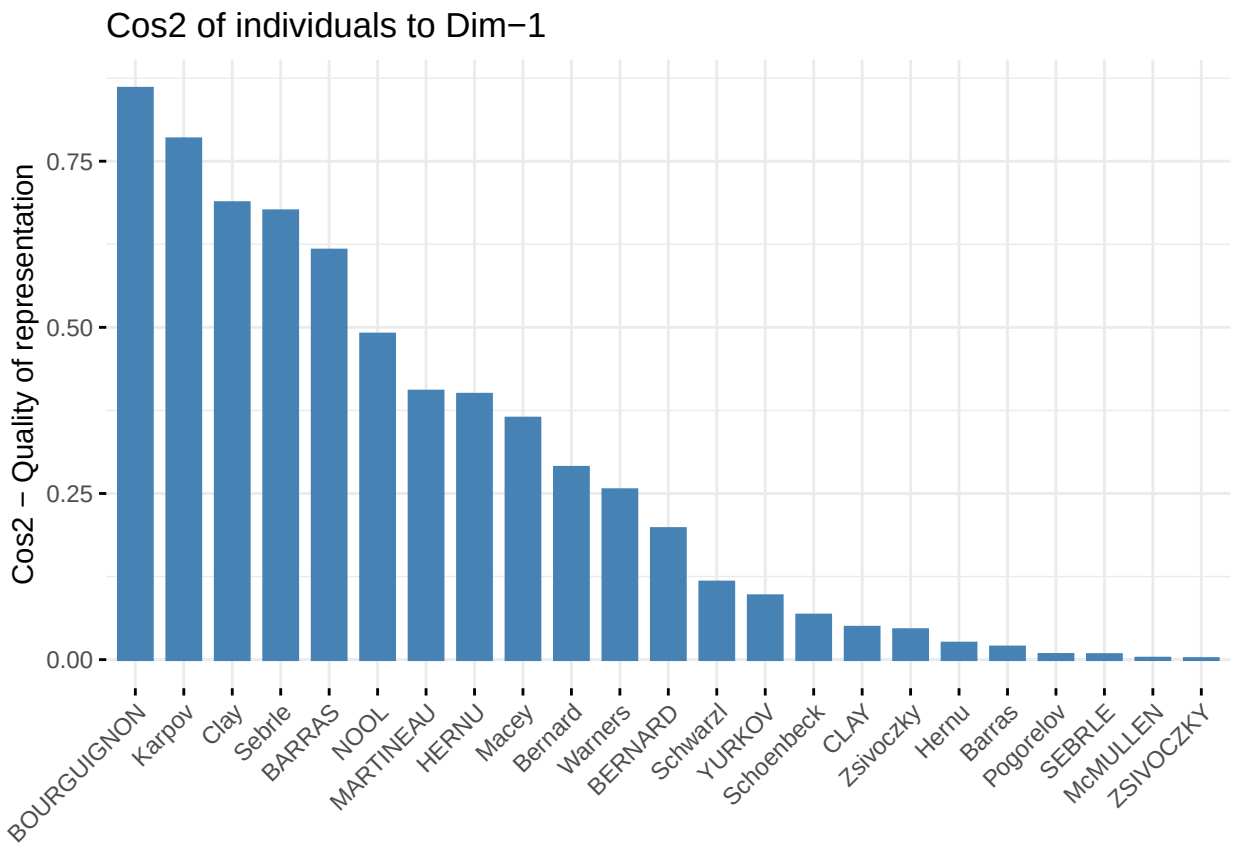
```
fviz_pca_ind(res.PCA, pointsize = "cos2",
             pointshape = 21, fill="#E7B800",
             repel = TRUE)
```



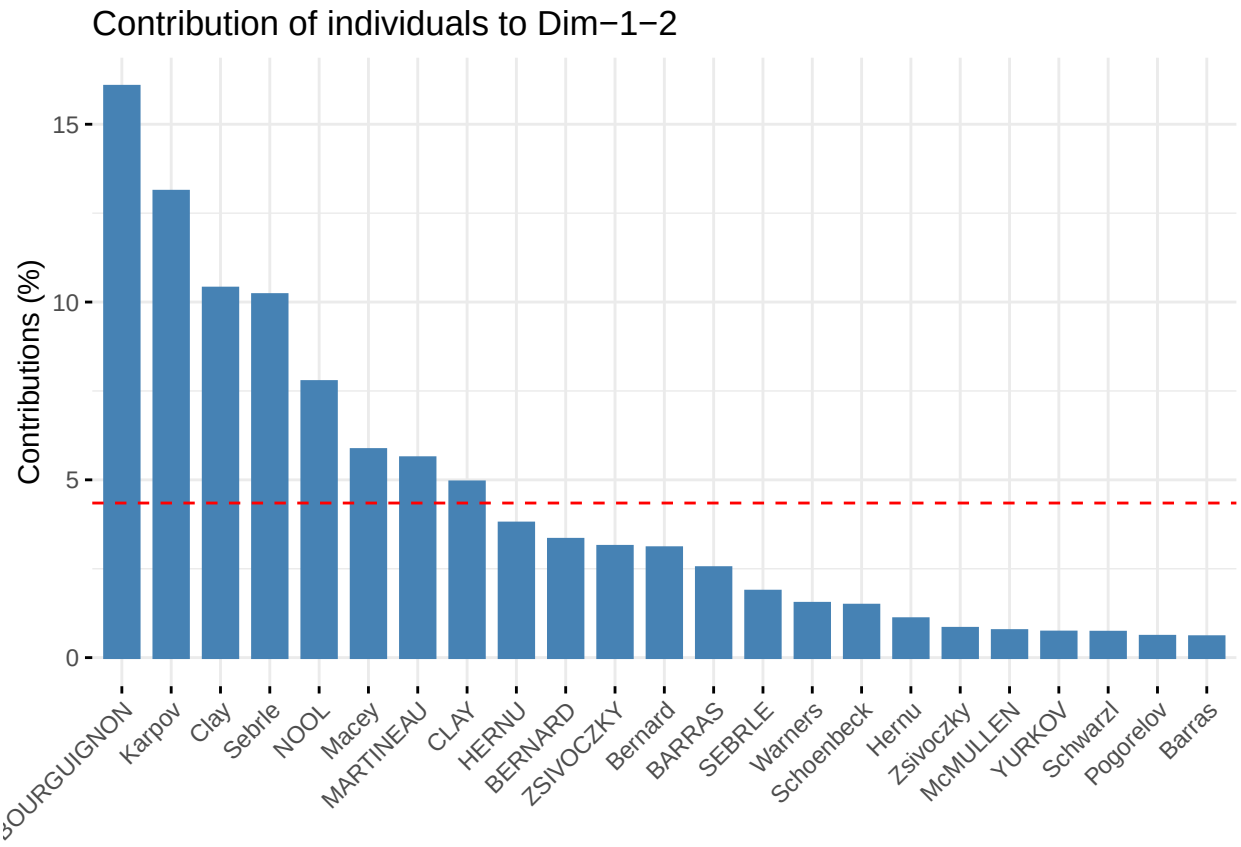

```
fviz_pca_ind(res.PCA, col.ind = "cos2", pointsize = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```




```
fviz_cos2(res.PCA, choice = "ind")
```




```
fviz_contrib(res.PCA, choice = "ind", axes = 1:2)
```


Contribution of individuals to Dim−1−2
Contributions (%)
15
10
5
0
OURGUIGNON
Karpov
Clay
Sebrle
NOOL
Macey
MARTINEAU
CLAY
HERNU
BERNARD
ZSIVOCZKY
Bernard
BARRAS
SEBRLE
Warners
Schoenbeck
Hernu
Zsivoczky
McMULLEN
YURKOV
Schwarzl
Pogorelov
Barras