PRINCIPAL COMPONENT ANALYSIS

CHARLES WIREDU

2/20/2022

```
setwd("~/Desktop/Data science Assignments")

###install required packages for PCA
#install.packages("FactoMineR")
#install.packages("factoextra")
library("FactoMineR")
library("factoextra")
```

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

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

```
#install.packages("ggplot2")
library("ggplot2")
```

```
##load dataset and subsetting the first column
library(readr)
Samples_1 <- read_csv("~/Desktop/Samples-1.csv")
```

```
## Rows: 6 Columns: 8
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Sites
## dbl (7): speciesA, speciesB, speciesC, speciesD, speciesE, speciesF, speciesG
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(Samples_1)
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/Resources/
## modules/R_de.so'' had status 1
```

```
new_sample1 <- data.frame(Samples_1[,-1], row.names= Samples_1$Sites)
View(new_sample1)
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/Resources/
## modules/R_de.so'' had status 1
```

```
head(new_sample1[,1:6], 4)
```

```
##       speciesA speciesB speciesC speciesD speciesE speciesF
## site1       46       80       66       24       95       75
## site2       38       86       44       51       91       83
## site3       16       89       65       78      123      135
## site4       81       15       31       97       17       34
```

Data Description

The data used for principal component analysis consist of both different types of species and the sites. The dataset has 7 columns and 6 rows.The data set used had no missing values , therefore a perfect choice for principal component analysis. Below is a brief description of the data set using the R statistical software.

```
summary(new_sample1 )
```

```
##     speciesA        speciesB        speciesC        speciesD    
##  Min.   :16.00   Min.   : 5.00   Min.   :31.00   Min.   :24.00  
##  1st Qu.:37.25   1st Qu.:15.25   1st Qu.:49.25   1st Qu.:57.75  
##  Median :42.00   Median :48.00   Median :65.50   Median :86.00  
##  Mean   :51.17   Mean   :48.50   Mean   :66.50   Mean   :73.17  
##  3rd Qu.:72.25   3rd Qu.:84.50   3rd Qu.:88.50   3rd Qu.:94.75  
##  Max.   :89.00   Max.   :89.00   Max.   :97.00   Max.   :97.00  
##     speciesE         speciesF         speciesG     
##  Min.   :  2.00   Min.   : 34.00   Min.   : 30.00  
##  1st Qu.: 20.50   1st Qu.: 75.00   1st Qu.: 79.75  
##  Median : 61.00   Median : 79.00   Median : 86.00  
##  Mean   : 59.83   Mean   : 83.33   Mean   : 90.83  
##  3rd Qu.: 94.00   3rd Qu.: 94.25   3rd Qu.:104.25  
##  Max.   :123.00   Max.   :135.00   Max.   :155.00  
```

The code below computes principal component analysis on the active individuals and variables. The output of the function is a list, including the following components

```
res.pca <- PCA(new_sample1,scale.unit= TRUE, graph = FALSE)
print(res.pca)
```

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 6 individuals, described by 7 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"
```

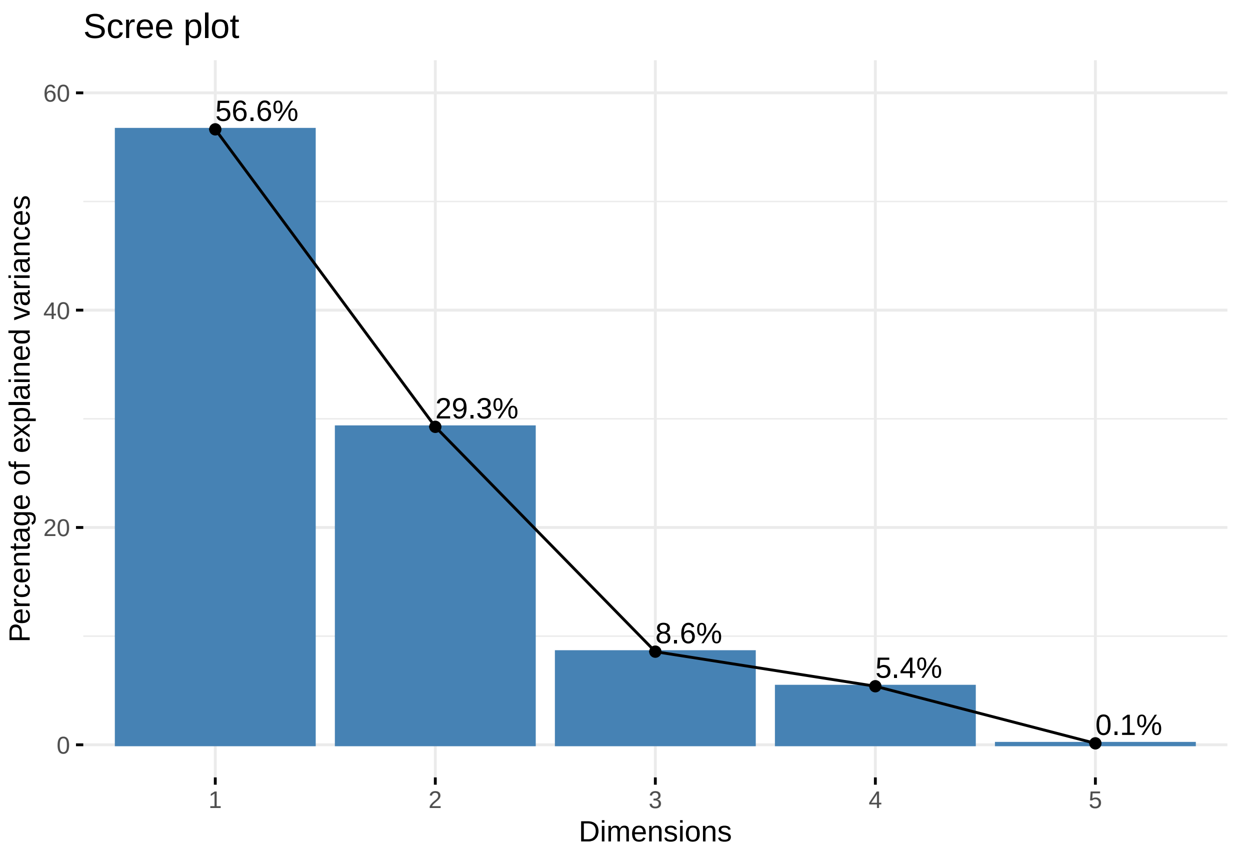
Extracting the eigenvalues/variance of principal components The eigenvalues measures the amount of variation retained by each principal components.They are usually large for the first PCs. That is, the first PCs correspond to the directions with maximum amount of variation in the data set.The sum of all the eigenvalues give a total a little above 7.0. The proportion of variation explained by each eigenvalue is given in the second column. For example 3.96 divided by 6.97 equals 56.64% of the variation explained by the first eigenvalue. The cumulative percentages explained is obtained by adding the successive proportions of variation explained to obtain the running total.For example 85.90% is explained by the two eigenvalues together.The function below illustrate this,

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

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1 3.964859352       56.6408479                    56.64085
## Dim.2 2.048397556       29.2628222                    85.90367
## Dim.3 0.599987080        8.5712440                    94.47491
## Dim.4 0.377301937        5.3900277                    99.86494
## Dim.5 0.009454076        0.1350582                   100.00000
```

The function below reports the scree plot which is an alternative method to determine the number of principal components.This is a plot eigenvalues ordered from largest to the smallest. The number of components is determined at the point,beyond which other eigenvalues are relatively small. Below is an illustration of the scree plot;

```
fviz_eig(res.pca,addlabels = TRUE,ylim= c(0,60))
```



The use of the function below in principal component analysis list all the matrices containing all the results for the active variables(coordinates,correlation,squared cosine and 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"
```

The different components can be assessed as follows:

```
##coordinates
head(var$coord)
```

```
##                Dim.1       Dim.2        Dim.3       Dim.4        Dim.5
## speciesA -0.85923096 -0.01075314 -0.169120077  0.48205532  0.025051742
## speciesB  0.92561001 -0.36725561 -0.003195971  0.08104185 -0.042325249
## speciesC -0.04812317  0.85415477 -0.470162856 -0.21657288  0.012118959
## speciesD -0.60591717  0.54074830  0.576607174 -0.08881910  0.009538841
## speciesE  0.96101189 -0.25397146  0.068339245 -0.02478600  0.081670367
## speciesF  0.73812254  0.63668651  0.074816996  0.21016789 -0.006105393
```

```
##cor: correlation between variable and dimension
head(var$cor)
```

```
##                Dim.1       Dim.2        Dim.3       Dim.4        Dim.5
## speciesA -0.85923096 -0.01075314 -0.169120077  0.48205532  0.025051742
## speciesB  0.92561001 -0.36725561 -0.003195971  0.08104185 -0.042325249
## speciesC -0.04812317  0.85415477 -0.470162856 -0.21657288  0.012118959
## speciesD -0.60591717  0.54074830  0.576607174 -0.08881910  0.009538841
## speciesE  0.96101189 -0.25397146  0.068339245 -0.02478600  0.081670367
## speciesF  0.73812254  0.63668651  0.074816996  0.21016789 -0.006105393
```

```
## cos2; quality on the factor map
head(var$cos2)
```

```
##                Dim.1      Dim.2        Dim.3        Dim.4        Dim.5
## speciesA 0.738277847 0.00011563 2.860160e-02 0.2323773331 6.275898e-04
## speciesB 0.856753892 0.13487669 1.021423e-05 0.0065677818 1.791427e-03
## speciesC 0.002315839 0.72958037 2.210531e-01 0.0469038128 1.468692e-04
## speciesD 0.367135619 0.29240873 3.324758e-01 0.0078888318 9.098949e-05
## speciesE 0.923543849 0.06450150 4.670252e-03 0.0006143457 6.670049e-03
## speciesF 0.544824887 0.40536971 5.597583e-03 0.0441705417 3.727583e-05
```

```
##contribution to principal components
head(var$contrib)
```

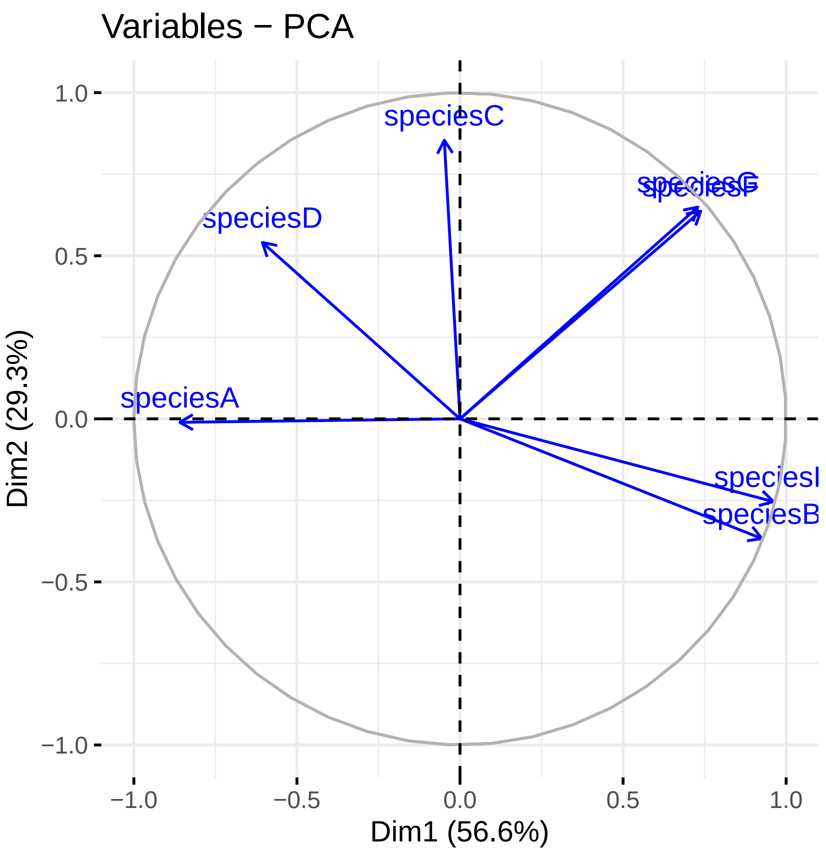
```
##                Dim.1        Dim.2        Dim.3     Dim.4      Dim.5
## speciesA 18.62053055  0.005644899  4.767036054 61.589223  6.6382988
## speciesB 21.60868309  6.584497463  0.001702408  1.740723 18.9487241
## speciesC  0.05840912 35.617127419 36.842978567 12.431373  1.5535012
## speciesD  9.25973878 14.274998802 55.413832109  2.090854  0.9624366
## speciesE 23.29323103  3.148876258  0.778392161  0.162826 70.5520999
## speciesF 13.74134208 19.789601479  0.932950561 11.706948  0.3942832
```

```
head(var$coord,3)
```

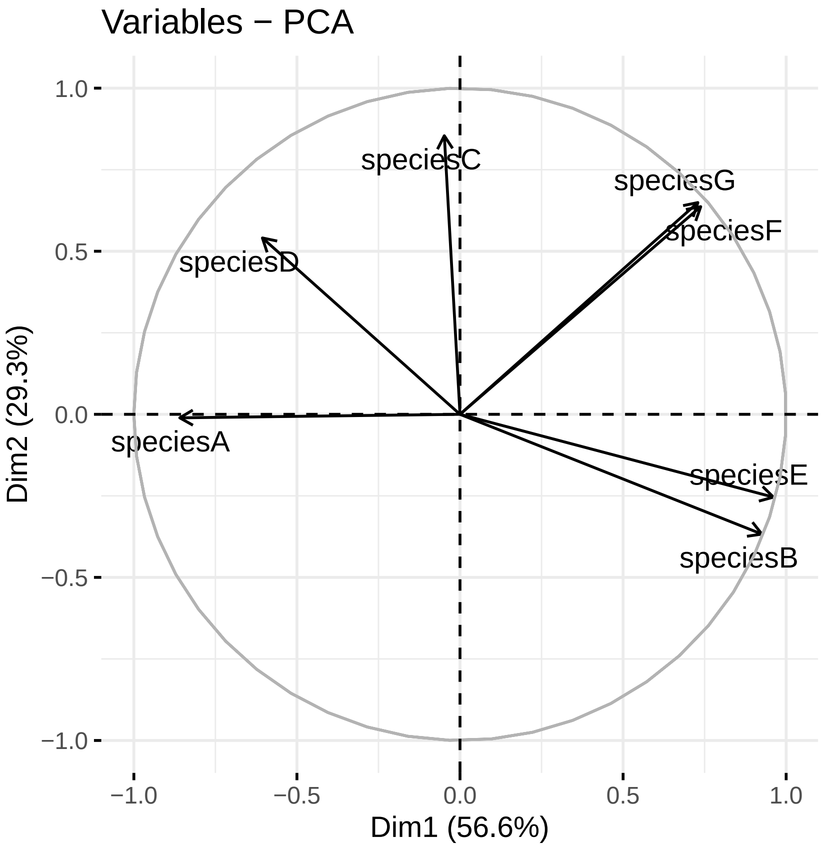
```
##                Dim.1       Dim.2        Dim.3       Dim.4       Dim.5
## speciesA -0.85923096 -0.01075314 -0.169120077  0.48205532  0.02505174
## speciesB  0.92561001 -0.36725561 -0.003195971  0.08104185 -0.04232525
## speciesC -0.04812317  0.85415477 -0.470162856 -0.21657288  0.01211896
```

Plotting of variables is demonstrated by the functions below;

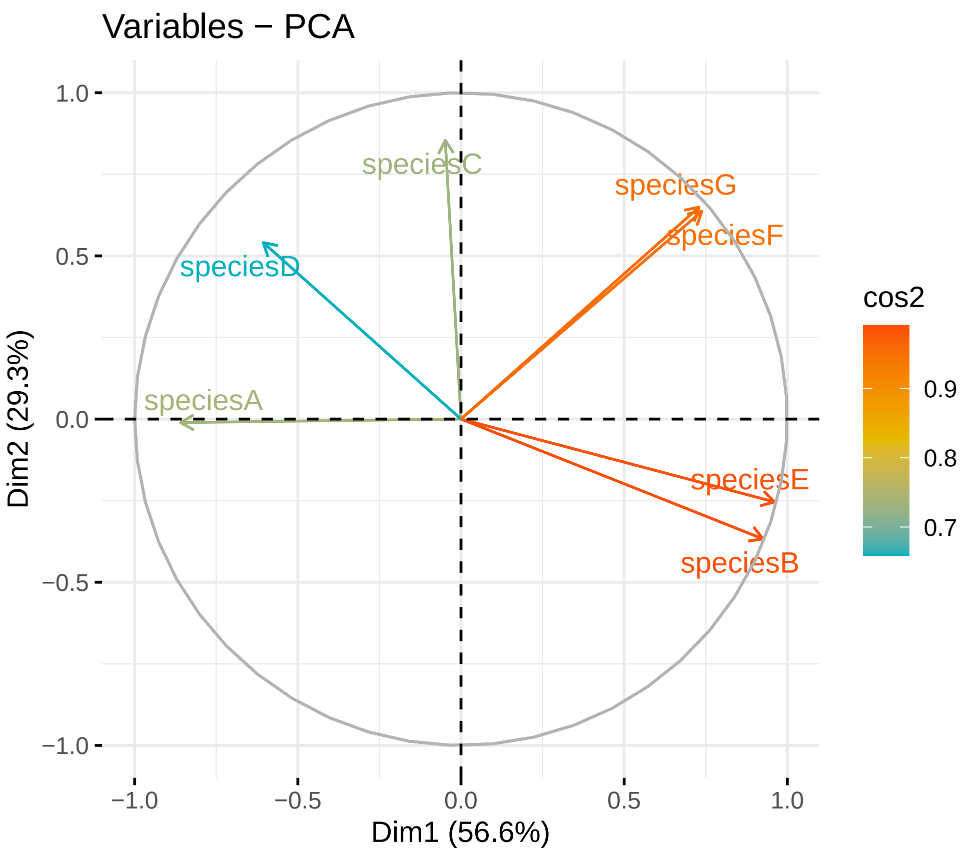
```
plot1<-fviz_pca_var(res.pca,col.var= "blue")
plot1
```



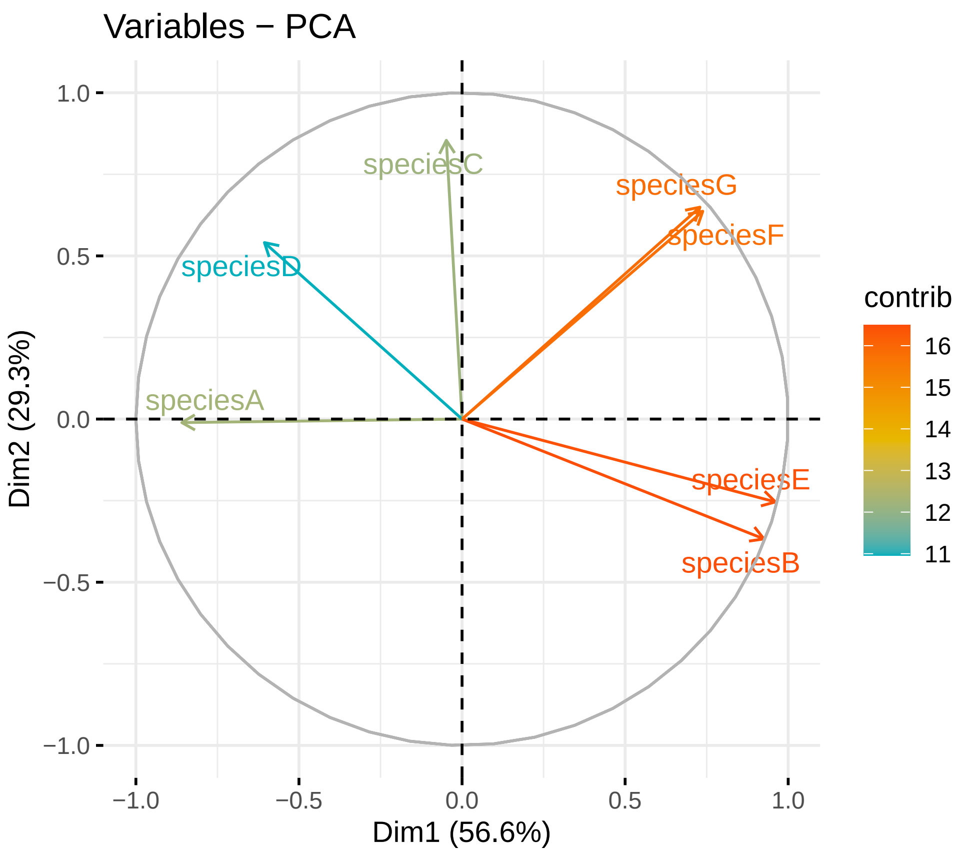
```
plot2<-fviz_pca_var(res.pca,pointsize="cos2", pointshape= 21,repel="True",fill="black")
plot2
```



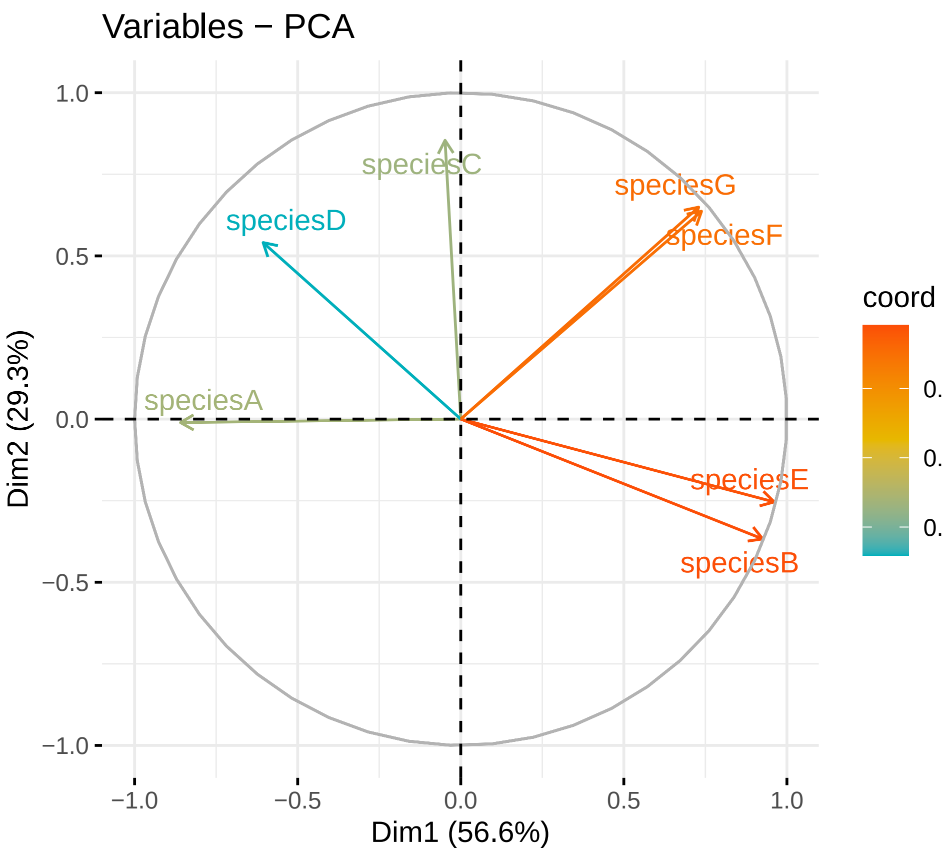
```
plot3<-fviz_pca_var(res.pca,pointsize="cos2", col.var = "cos2",gradient.cols= c("#00AFBB","#E7
                    repel= TRUE)
plot3
```



```
fviz_pca_var(res.pca,pointsize="contrib", col.var = "contrib",gradient.cols= c("#00AFBB","#E7B8
             repel= TRUE)
```



```
fviz_pca_var(res.pca,pointsize="coord", col.var = "coord",gradient.cols= c("#00AFBB","#E7B800"
             repel= TRUE)
```



The plots above also known as variable correlation plot shows the relationships between all variables.It can be interpreted as positively correlated variables are grouped together whiles negatively correlated variables are positioned on opposite sides of he plot. The quality of the variables is measured by the distance between the variables and the origin.

Plotting of individuals- quality and contribution. The function below list all the individual component coupled with their plots

```
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)# coordinates of individuals
```

```
##              Dim.1     Dim.2      Dim.3        Dim.4       Dim.5
## site1  1.1951413 -1.320730 -1.2905781 -0.002962978  0.09781942
## site2  1.2876107 -1.254958  0.1164269  0.136953784 -0.18805503
## site3  2.9904266  1.051379  0.9465344  0.147367970  0.07863657
## site4 -2.8184606 -1.609609  0.8631057  0.133677885  0.06469264
## site5 -0.9447908  1.140238 -0.1202938 -1.236919052 -0.02401635
## site6 -1.7099272  1.993680 -0.5151951  0.821882392 -0.02907725
```

```
head(ind$cor)# correlation between individuals and dimension
```

```
## NULL
```

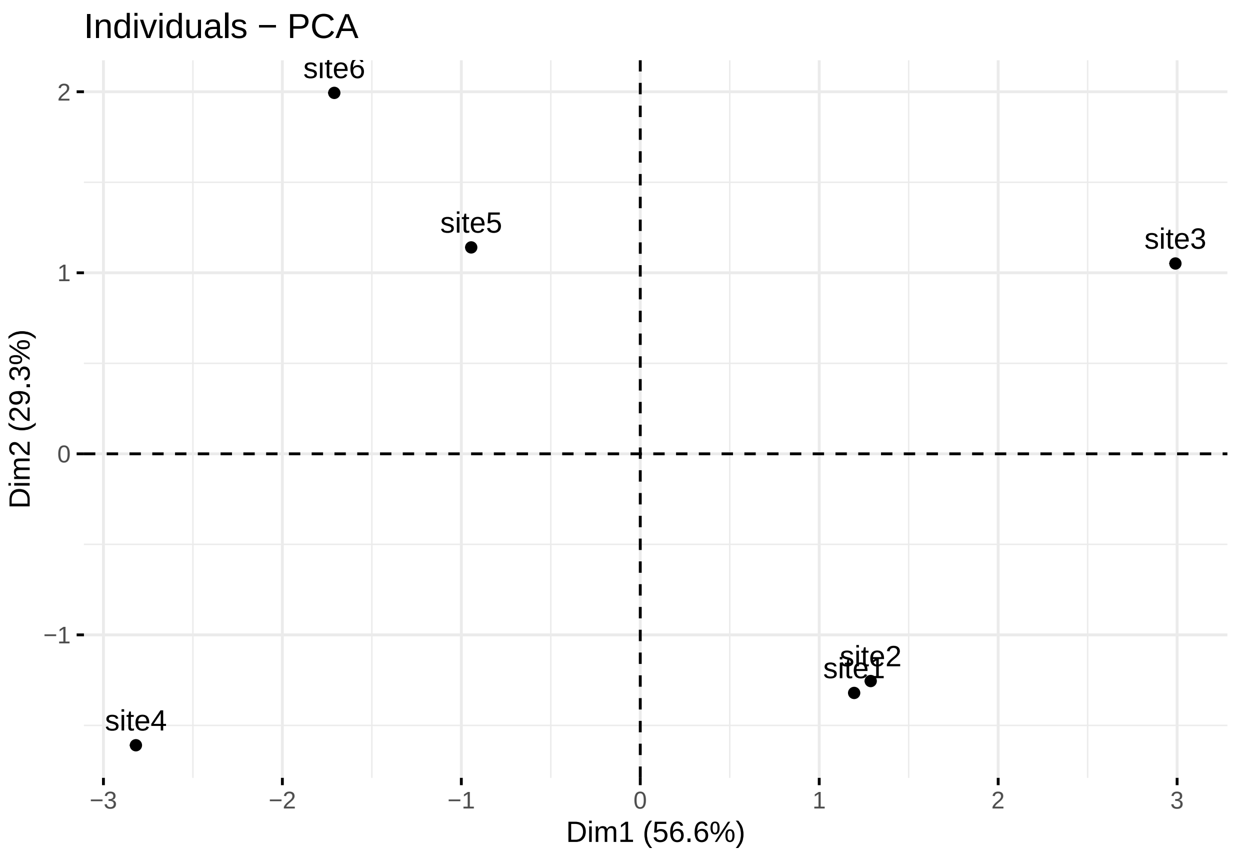
```
head(ind$cos2)#quality of individuals
```

```
##           Dim.1     Dim.2       Dim.3        Dim.4        Dim.5
## site1 0.2946378 0.3598139 0.343572684 1.810952e-06 0.0019737866
## site2 0.5023248 0.4771706 0.004106977 5.682814e-03 0.0107148307
## site3 0.8150521 0.1007482 0.081656707 1.979362e-03 0.0005635964
## site4 0.7028866 0.2292461 0.065915766 1.581177e-03 0.0003703148
## site5 0.2388122 0.3478376 0.003871432 4.093245e-01 0.0001543118
## site6 0.3729225 0.5069604 0.033853748 8.615545e-02 0.0001078376
```

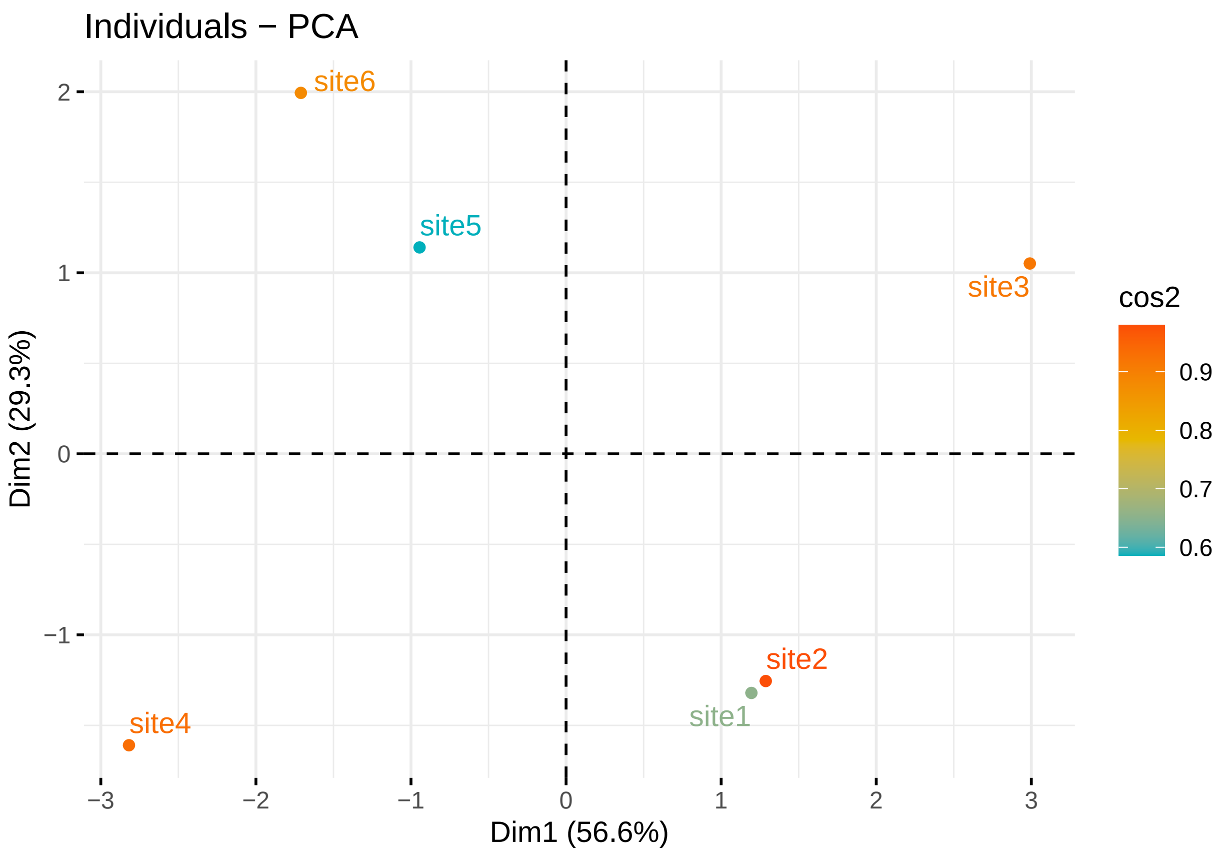
```
head(ind$contrib)#contribution of the individuals
```

```
##           Dim.1     Dim.2      Dim.3        Dim.4     Dim.5
## site1  6.004259 14.192615 46.2674395 3.878079e-04 16.868632
## site2  6.969316 12.814238  0.3765423 8.285291e-01 62.344708
## site3 37.591292  8.993997 24.8874054 9.593253e-01 10.901313
## site4 33.392190 21.080222 20.6935414 7.893668e-01  7.378013
## site5  3.752255 10.578542  0.4019697 6.758375e+01  1.016819
## site6 12.290688 32.340386  7.3731017 2.983864e+01  1.490515
```

```
graph1<-fviz_pca_ind(res.pca)
graph1
```

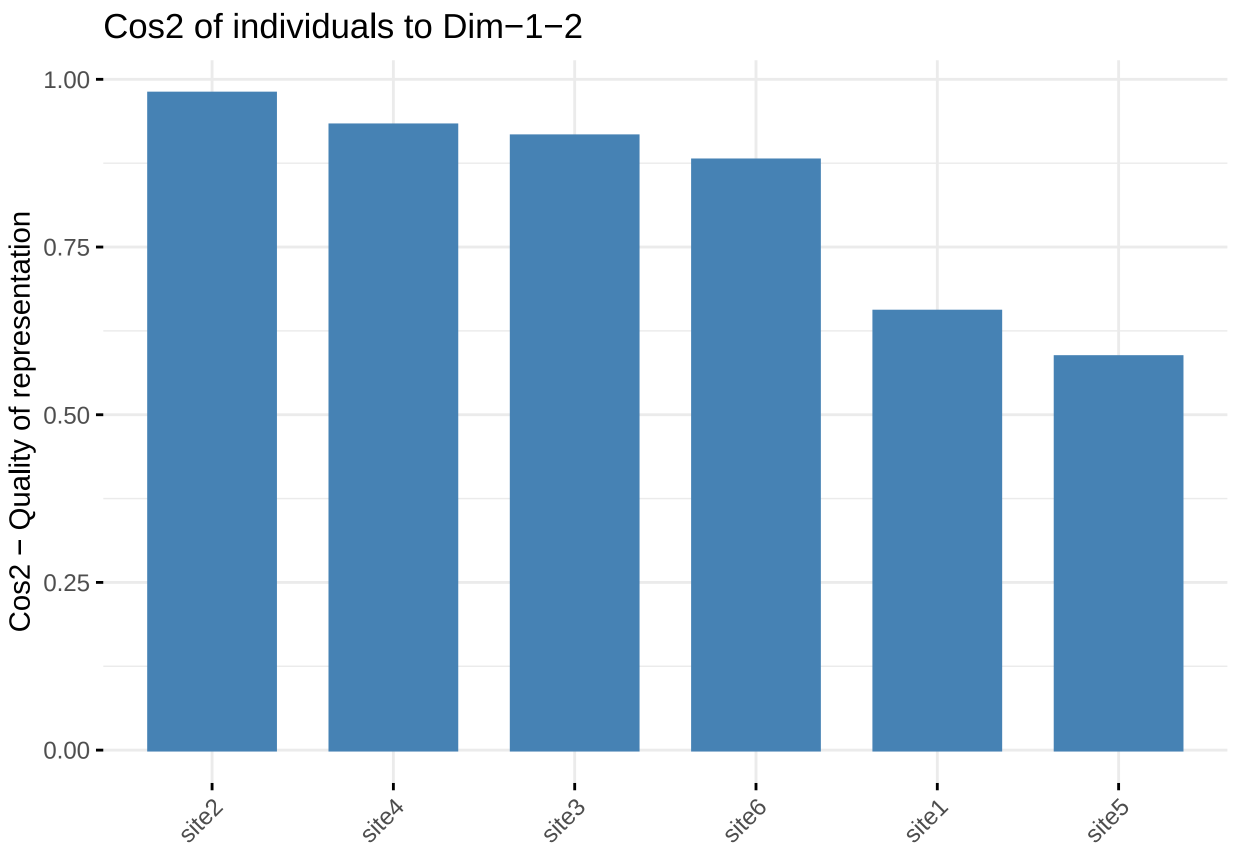


```
graph2<-fviz_pca_ind(res.pca, col.ind = "cos2",gradient.cols= c("#00AFBB","#E7B800", "#FC4E07"
graph2
```



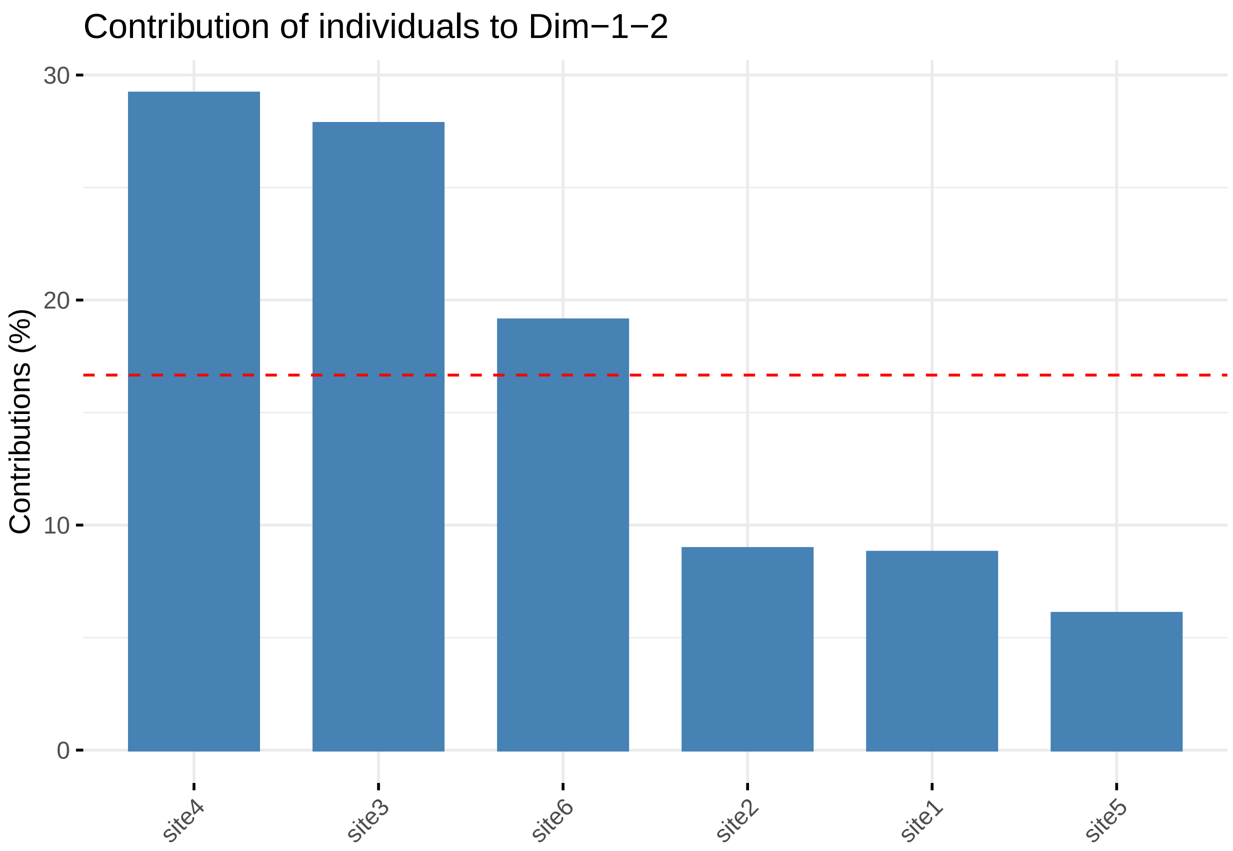
Plot of quality representation of cos2 is illustrated by the function below;

```
fviz_cos2(res.pca, choice= "ind", axes =1:2)
```



Plot of contribution of individuals to first two dimensions

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

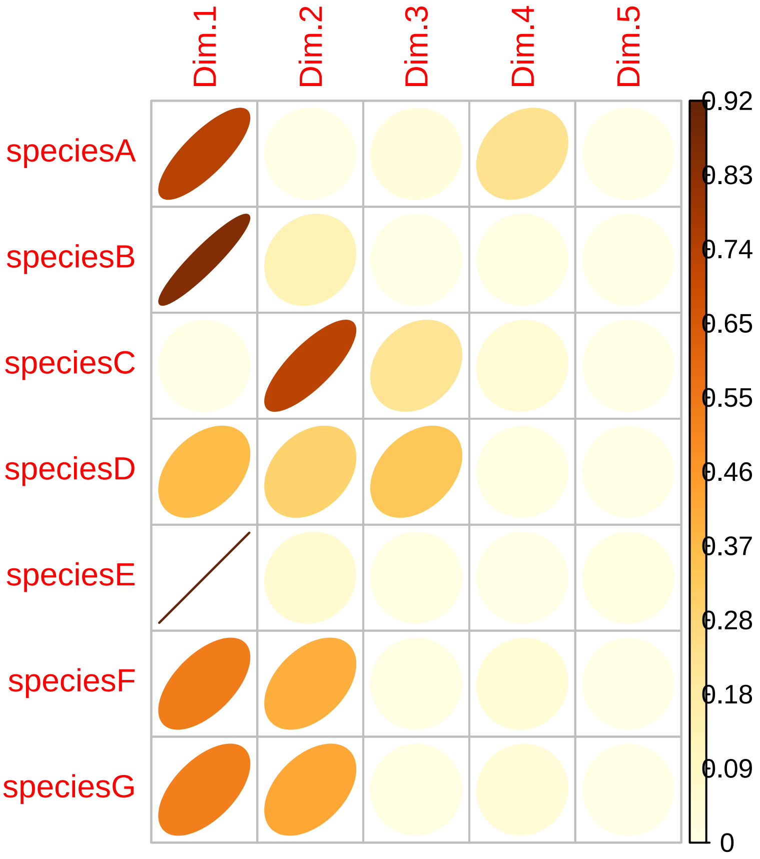


Correlation matrix graph. The function below illustrates correlation matrix;

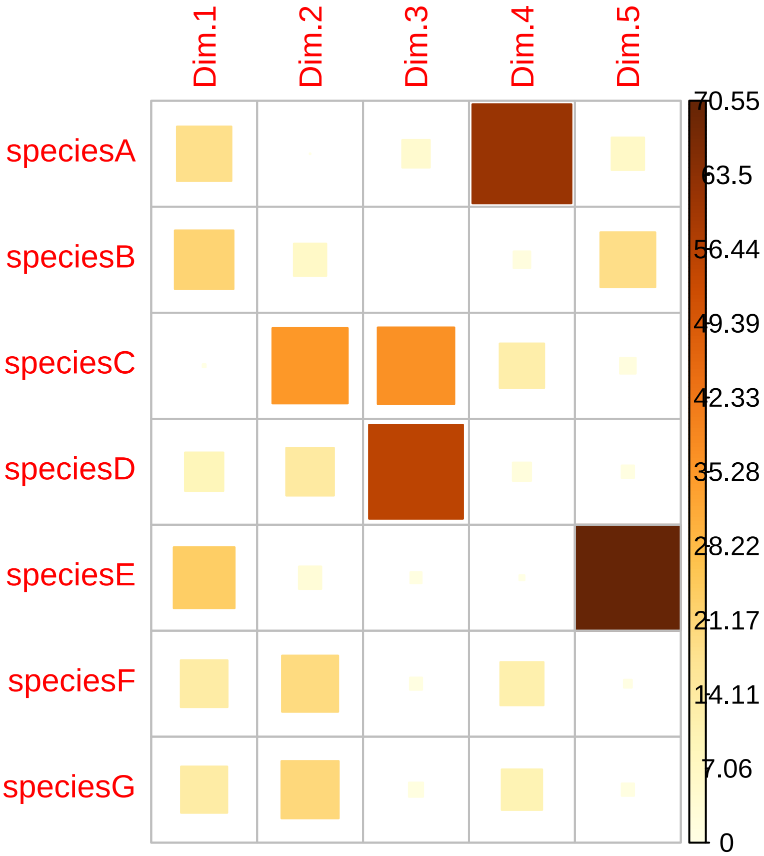
```
#install.packages("corrplot")
library("corrplot")
```

```
## corrplot 0.92 loaded
```

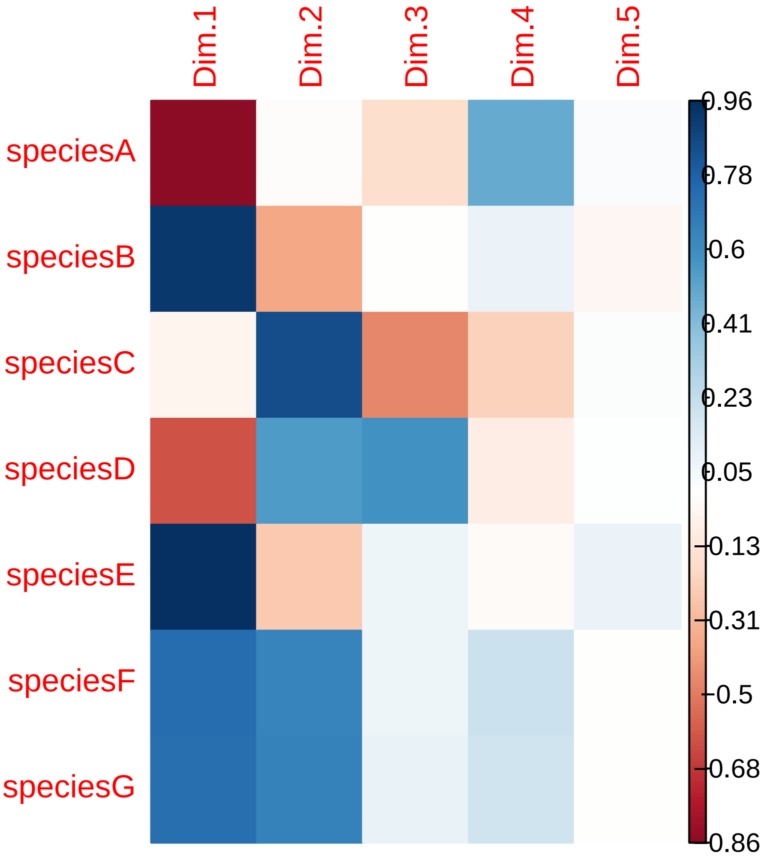
```
corrplot(var$cos2, method= 'ellipse',is.corr=FALSE)
```



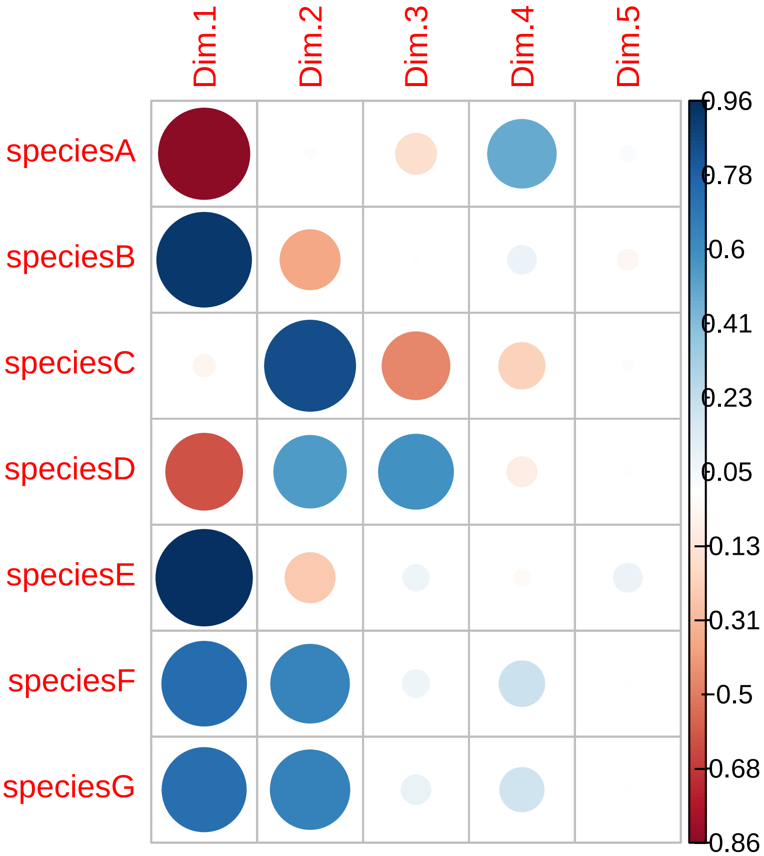
```
corrplot(var$contrib,method= 'square', is.corr=FALSE)
```



```
corrplot(var$coord,method= 'color', is.corr=FALSE)
```



```
corrplot(var$cor,method= 'circle', is.corr=FALSE)
```


Dim.1
Dim.2
Dim.3
Dim.4
Dim.5
speciesA
speciesB
speciesC
speciesD
speciesE
speciesF
speciesG
0.96
0.78
0.6
0.41
0.23
0.05
-0.13
-0.31
-0.5
-0.68
-0.86