DATA MINING COURSE

TEYI KODJO JEROME SEDOWO

2025-03-19

# Package

**Here, we’ll use FactoMineR for the analysis and factoextra for ggplot2-based visualization**

library(FactoMineR)  
library(factoextra)  
library(readr)  
library(corrplot)

# Import the data set

data <- read\_delim("C:/Users/Jérôme/Desktop/TEYI\_KODJO\_JEROME\_SEDOWO/ETUDE/AIMS\_SENEGAL\_2024-2025/Review phase Courses/Block 4/Data Mining and Big data/Tutorial\_1/ACP\_eaux.txt",   
 delim = "\t", escape\_double = FALSE,   
 col\_types = cols(CA = col\_number(), MG = col\_number(),   
 `NA` = col\_number(), K = col\_number(),   
 SUL = col\_number(), NO3 = col\_number(),   
 HCO3 = col\_number(), CL = col\_number()),   
 trim\_ws = TRUE)  
#View(data)

data\_numeric <- data[6:13]  
head(data\_numeric)

## # A tibble: 6 × 8  
## CA MG `NA` K SUL NO3 HCO3 CL  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 78 24 5 1 10 3.8 357 4.5  
## 2 48 11 34 1 16 4 183 50   
## 3 71 5.5 11.2 3.2 5 1 250 20   
## 4 89 31 17 2 47 0 360 28   
## 5 4.1 1.7 2.7 0.9 1.1 0.8 25.8 0.9  
## 6 85 80 385 65 25 1.9 1350 285

# Exploratory analysis

Here we print a simple statistics for continous variables

summary(data\_numeric)

## CA MG NA K   
## Min. : 1.2 Min. : 0.20 Min. : 0.80 Min. : 0.00   
## 1st Qu.: 36.0 1st Qu.: 5.50 1st Qu.: 5.00 1st Qu.: 0.90   
## Median : 63.0 Median :12.00 Median : 9.10 Median : 2.00   
## Mean :102.5 Mean :25.86 Mean : 93.85 Mean : 11.09   
## 3rd Qu.:116.0 3rd Qu.:31.50 3rd Qu.: 36.00 3rd Qu.: 6.00   
## Max. :528.0 Max. :95.00 Max. :968.00 Max. :130.00   
## SUL NO3 HCO3 CL   
## Min. : 1.1 Min. : 0.000 Min. : 4.9 Min. : 0.30   
## 1st Qu.: 9.0 1st Qu.: 0.450 1st Qu.: 154.0 1st Qu.: 3.50   
## Median : 16.0 Median : 1.500 Median : 236.0 Median : 14.20   
## Mean : 135.7 Mean : 3.834 Mean : 442.2 Mean : 52.47   
## 3rd Qu.: 43.0 3rd Qu.: 4.000 3rd Qu.: 360.0 3rd Qu.: 38.00   
## Max. :1371.0 Max. :35.600 Max. :3380.5 Max. :982.00

cor(data\_numeric)

## CA MG NA K SUL NO3  
## CA 1.00000000 0.7027224 0.11794153 0.12535483 0.91309695 -0.06344287  
## MG 0.70272239 1.0000000 0.60756895 0.66113238 0.60546334 -0.21238801  
## NA 0.11794153 0.6075689 1.00000000 0.83656419 0.06429603 -0.11624022  
## K 0.12535483 0.6611324 0.83656419 1.00000000 -0.02515575 -0.16592834  
## SUL 0.91309695 0.6054633 0.06429603 -0.02515575 1.00000000 -0.15650372  
## NO3 -0.06344287 -0.2123880 -0.11624022 -0.16592834 -0.15650372 1.00000000  
## HCO3 0.13494940 0.6197724 0.85621354 0.88156811 -0.06913651 -0.06039047  
## CL 0.27640957 0.4812610 0.58752083 0.40043988 0.31781920 -0.12017032  
## HCO3 CL  
## CA 0.13494940 0.2764096  
## MG 0.61977235 0.4812610  
## NA 0.85621354 0.5875208  
## K 0.88156811 0.4004399  
## SUL -0.06913651 0.3178192  
## NO3 -0.06039047 -0.1201703  
## HCO3 1.00000000 0.1906228  
## CL 0.19062285 1.0000000

# Data Standardization

By default PCA() in **FactoMinR** standardizes the data automatically during the PCA. So we will not standardize the data manually before the PCA

resul\_pca <- PCA(data\_numeric, graph = FALSE)  
print(resul\_pca)

## \*\*Results for the Principal Component Analysis (PCA)\*\*  
## The analysis was performed on 57 individuals, described by 8 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"

*This is many information found in many different lists and matrices.*

# Visualization and Interpretation

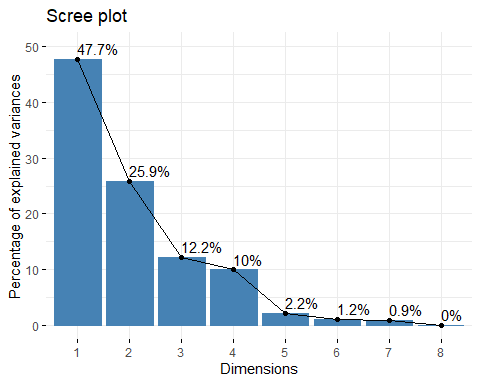
### Eigenvalues

eig\_value <- get\_eigenvalue(resul\_pca)  
eig\_value

## eigenvalue variance.percent cumulative.variance.percent  
## Dim.1 3.8167747447 47.709684309 47.70968  
## Dim.2 2.0680904354 25.851130442 73.56081  
## Dim.3 0.9728158313 12.160197892 85.72101  
## Dim.4 0.7962420036 9.953025045 95.67404  
## Dim.5 0.1792036107 2.240045134 97.91408  
## Dim.6 0.0924269941 1.155337427 99.06942  
## Dim.7 0.0740850743 0.926063429 99.99548  
## Dim.8 0.0003613058 0.004516322 100.00000

### Visualisation and Interpretation

fviz\_eig(resul\_pca, addlabels = TRUE, ylim = c(0,50))



From yhe plot above, we might want to stop at the third principal component. 68% of the information contained in the data are retained by the first Three principal components.

## Graph of variables

var <- get\_pca\_var(resul\_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"

### Coordinates of variables

var$coord

## Dim.1 Dim.2 Dim.3 Dim.4 Dim.5  
## CA 0.5496004 0.77641500 0.170495237 -0.17809784 0.005109279  
## MG 0.9104573 0.25440564 0.036883459 -0.14815849 -0.196246861  
## NA 0.8551621 -0.41427700 0.033126250 0.15439228 0.262944434  
## K 0.8354674 -0.45847406 -0.005010044 -0.10636587 -0.190470398  
## SUL 0.4496677 0.86757992 0.031460991 -0.02949458 0.141085221  
## NO3 -0.2337948 -0.09000400 0.958423890 0.13060377 -0.026621009  
## HCO3 0.7840386 -0.49889576 0.129768176 -0.31140014 0.102477046  
## CL 0.6203998 0.09503392 -0.069702512 0.76975595 -0.064270252

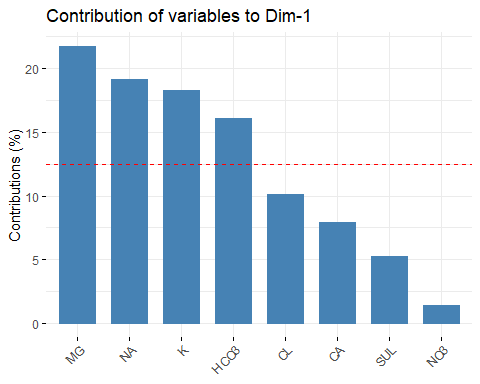
## Contribution of the variables

var$contrib

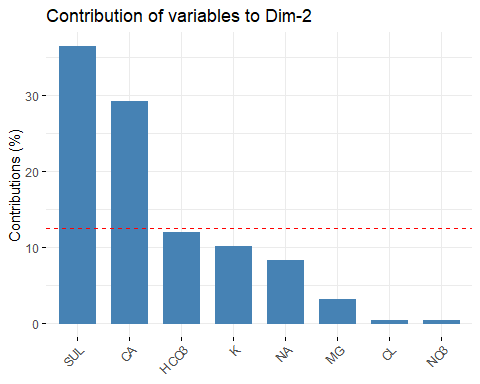
## Dim.1 Dim.2 Dim.3 Dim.4 Dim.5  
## CA 7.914028 29.1486404 2.988091369 3.9835677 0.01456708  
## MG 21.718142 3.1295648 0.139840403 2.7568176 21.49110176  
## NA 19.160213 8.2987394 0.112801250 2.9936849 38.58168659  
## K 18.287845 10.1638912 0.002580194 1.4208869 20.24455434  
## SUL 5.297695 36.3956481 0.101745256 0.1092545 11.10749910  
## NO3 1.432099 0.3917004 94.424486502 2.1422313 0.39545973  
## HCO3 16.105655 12.0351111 1.731034686 12.1784641 5.86011908  
## CL 10.084322 0.4367046 0.499420338 74.4150930 2.30501232

## Contribution of Variables to PC1

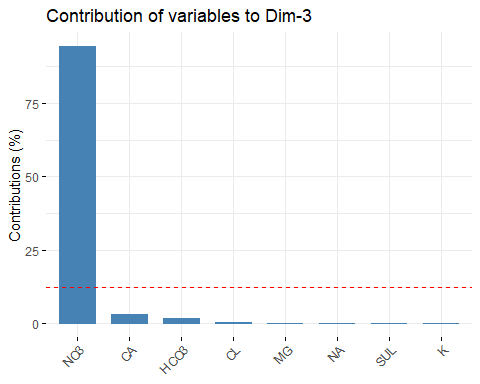
fviz\_contrib(resul\_pca, choice = "var", axes = 1)



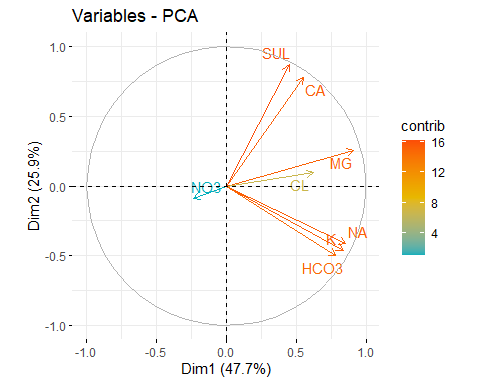
fviz\_contrib(resul\_pca, choice = "var", axes = 2)



fviz\_contrib(resul\_pca, choice = "var", axes = 3)

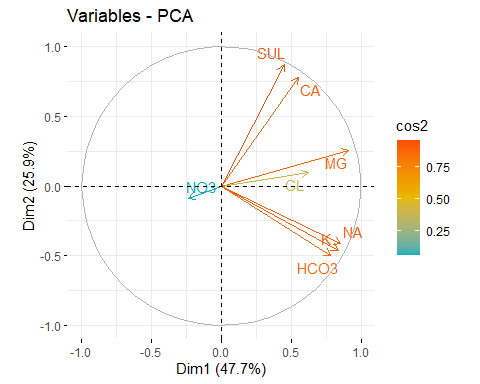


fviz\_pca\_var(resul\_pca, col.var = "contrib",  
 gradient.cols = c("#00AFBB","#E7B800","#FC4E07"),  
 repel = TRUE)



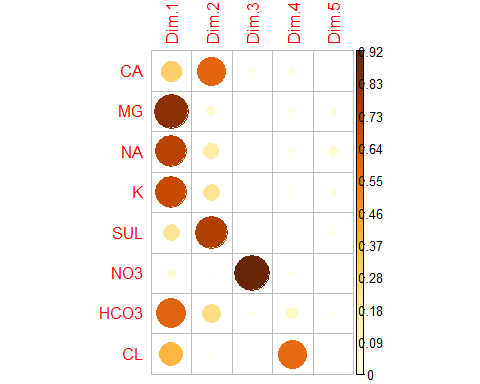
### Correlation circle

fviz\_pca\_var(resul\_pca, col.var = "cos2",  
 gradient.cols = c("#00AFBB","#E7B800","#FC4E07"),  
 repel = TRUE)

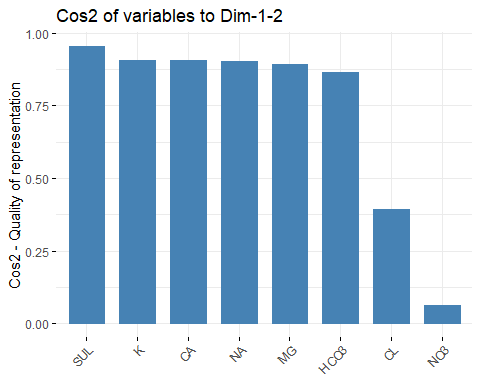


## Quality of representation

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



fviz\_cos2(resul\_pca, choice = "var", axes = 1:2)



## Dimension description

*Description of dimension 1*

res.desc <- dimdesc(resul\_pca, axes = c(1,2), proba = 0.05)  
res.desc$Dim.1

##   
## Link between the variable and the continuous variables (R-square)  
## =================================================================================  
## correlation p.value  
## MG 0.9104573 9.573425e-23  
## NA 0.8551621 2.510398e-17  
## K 0.8354674 6.376416e-16  
## HCO3 0.7840386 5.492255e-13  
## CL 0.6203998 2.640912e-07  
## CA 0.5496004 9.518680e-06  
## SUL 0.4496677 4.495443e-04

*Description of dimension 2*

res.desc <- dimdesc(resul\_pca, axes = c(1,2), proba = 0.05)  
res.desc$Dim.2

##   
## Link between the variable and the continuous variables (R-square)  
## =================================================================================  
## correlation p.value  
## SUL 0.8675799 2.528799e-18  
## CA 0.7764150 1.279229e-12  
## NA -0.4142770 1.356729e-03  
## K -0.4584741 3.350678e-04  
## HCO3 -0.4988958 7.814521e-05

# Graph of individuals

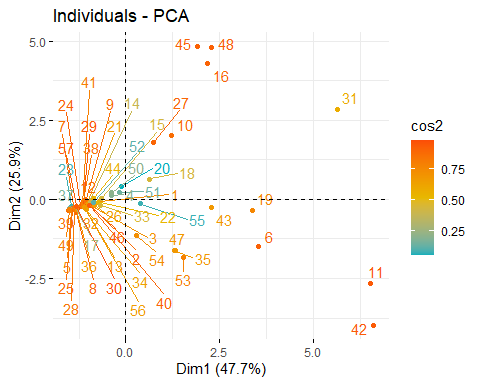
ind <- get\_pca\_ind(resul\_pca)  
ind

## 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"

# ind$coord  
# ind$cos2  
# ind$contrib

# Quality of contribution

fviz\_pca\_ind(resul\_pca, col.ind = "cos2",  
 gradient.cols = c("#00AFBB","#E7B800","#FC4E07"),  
 repel = TRUE)



fviz\_pca\_ind(resul\_pca, pointsize = "cos2",  
 pointshape = 21, fill = "#E7B800",  
 repel = TRUE)

