# Homework - 2

Group 4

5/19/2020

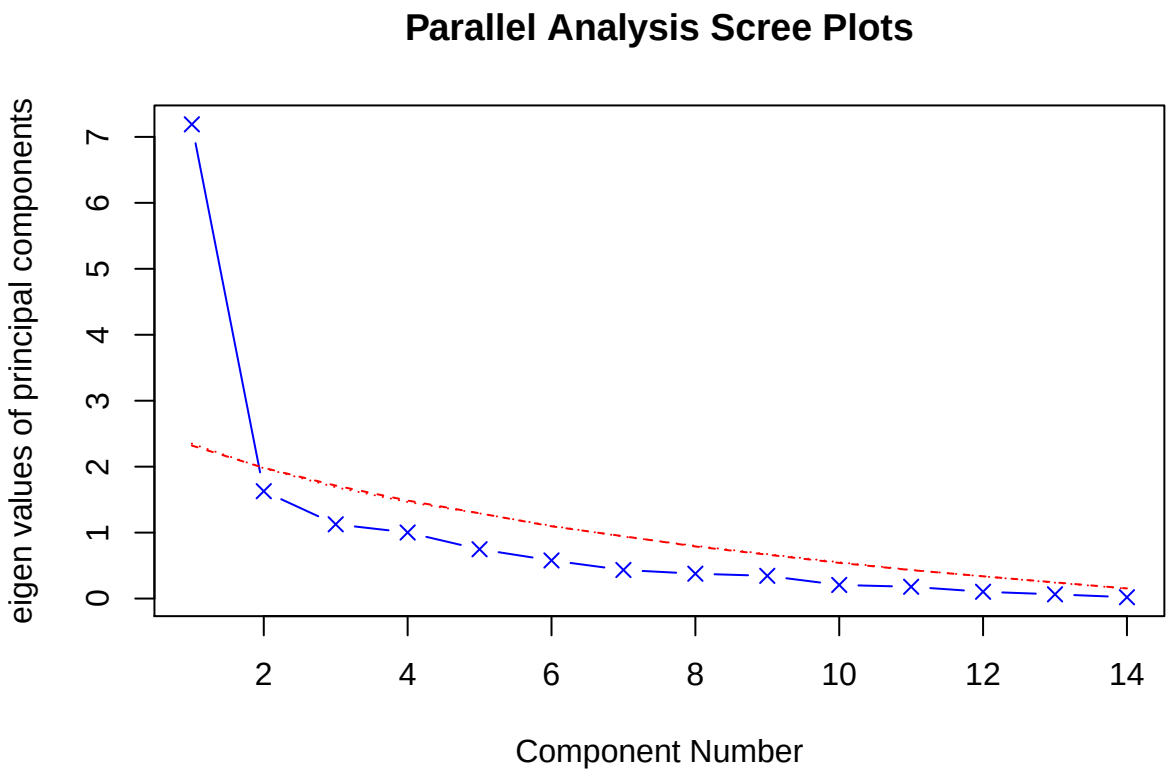
## Problem 1

Perform principal component analysis on NHL.xlsx, which contains statistics of 30 teams in the National Hockey League. The description of the variables is provided in the 'Description' sheet of the file. Focus only on the variables 12 through 25, and create a new data frame. • Input the new data frame to fa.parallel() function to determine the number of components to extract • Input the new data frame to principal() function to extract the components. If raw data is input, the correlation matrix is automatically calculated by principal() function. • Rotate the components • Compute component scores • Graph an orthogonal solution using factor.plot() • Interpret the results

### First, import all the required libraries

```
library(dplyr)
library(readxl)
library(psych)
```

```
# Import the NHL excel file as a dataframe
NHL <- data.frame(read_xlsx("./data/NHL.xlsx", sheet = "Data"))

# Select the columns 13-26, the 1st column is the index column from excel
df <- NHL[, 13:26]

# Use Parallel Analysis Scree Plots to figure out the number of factors to extract
fa.parallel(df, fa = "pc", n.iter = 100, show.legend = FALSE)
```




```
## Parallel analysis suggests that the number of factors =  NA  and the number of components =  1
```

From the above plot it will be appropriate to use 1 factor, but we will use 2 factors since a single factor might not be enough to capture enough of the variance.

```
# Perform PCA with varimax orthogonal rotation
pc <- principal(df, nfactors = 2, rotate = "varimax", scores = TRUE)

pc
```

```
## Principal Components Analysis
## Call: principal(r = df, nfactors = 2, rotate = "varimax", scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##         RC1   RC2   h2   u2 com
## gg     0.77  0.29 0.68 0.32 1.3
## gag   -0.79 -0.21 0.67 0.33 1.1
## five   0.90  0.16 0.83 0.17 1.1
## PPP   -0.07  0.78 0.61 0.39 1.0
## PKP    0.73  0.01 0.53 0.47 1.0
## shots  0.48  0.51 0.49 0.51 2.0
## sag   -0.48 -0.59 0.57 0.43 1.9
## sc1    0.81  0.13 0.68 0.32 1.1
## tr1    0.75  0.09 0.57 0.43 1.0
## lead1  0.77  0.28 0.67 0.33 1.3
## lead2  0.75  0.13 0.57 0.43 1.1
## wop    0.73  0.02 0.54 0.46 1.0
## wosp   0.88  0.02 0.77 0.23 1.0
```

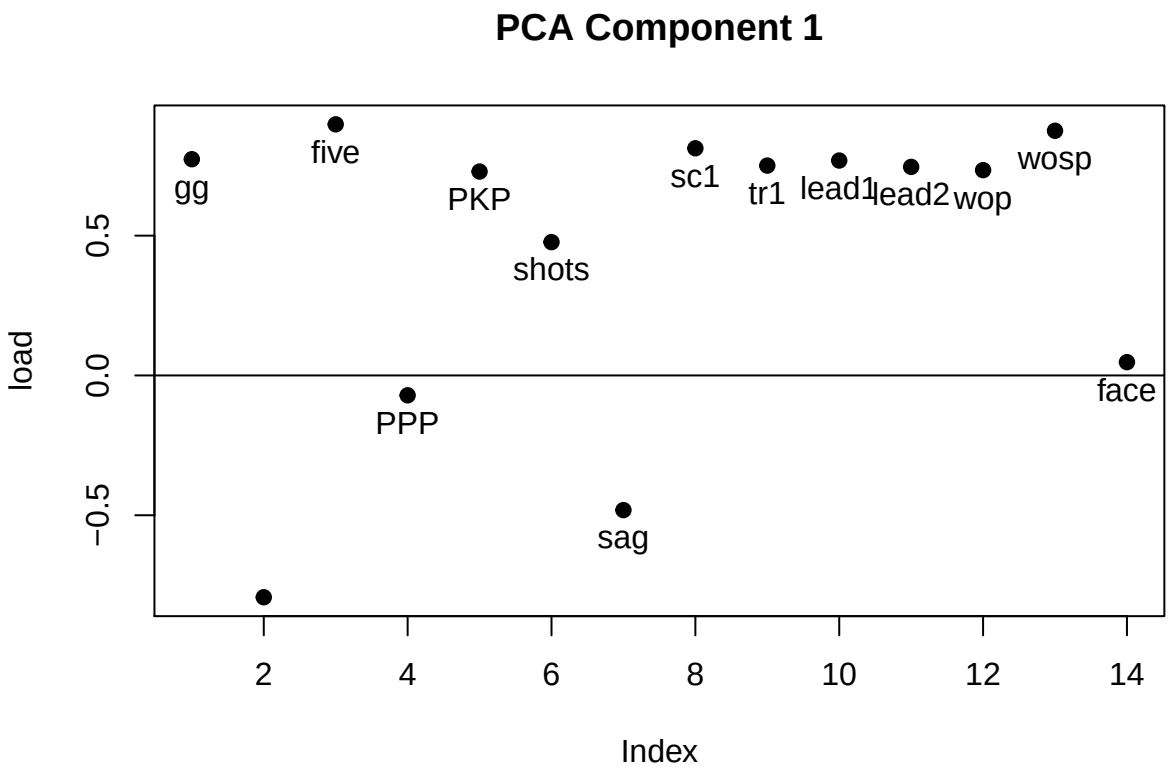
```
## face   0.05  0.79 0.63 0.37 1.0
##
##                        RC1  RC2
## SS loadings           6.71 2.11
## Proportion Var        0.48 0.15
## Cumulative Var        0.48 0.63
## Proportion Explained  0.76 0.24
## Cumulative Proportion 0.76 1.00
##
## Mean item complexity =  1.2
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.1
##  with the empirical chi square  50.08  with prob <  0.9
##
## Fit based upon off diagonal values = 0.96
```

Let's see the component scores

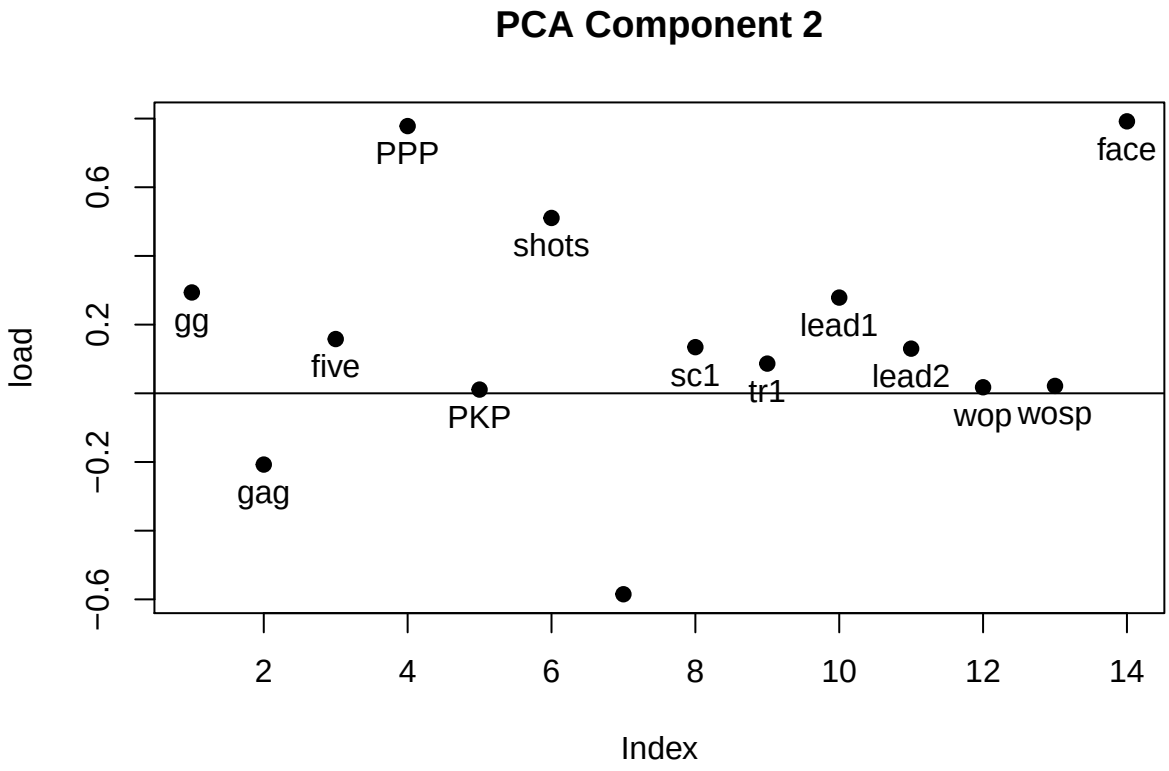
```
head(pc$scores)
```

```
##            RC1         RC2
## [1,] 1.7495821 -1.09131873
## [2,] 0.6973249  1.50788425
## [3,] 0.8153068 -0.24718608
## [4,] 0.9013922 -0.46220814
## [5,] 1.2681472 -0.05501335
## [6,] 1.1122392  0.40217809
```

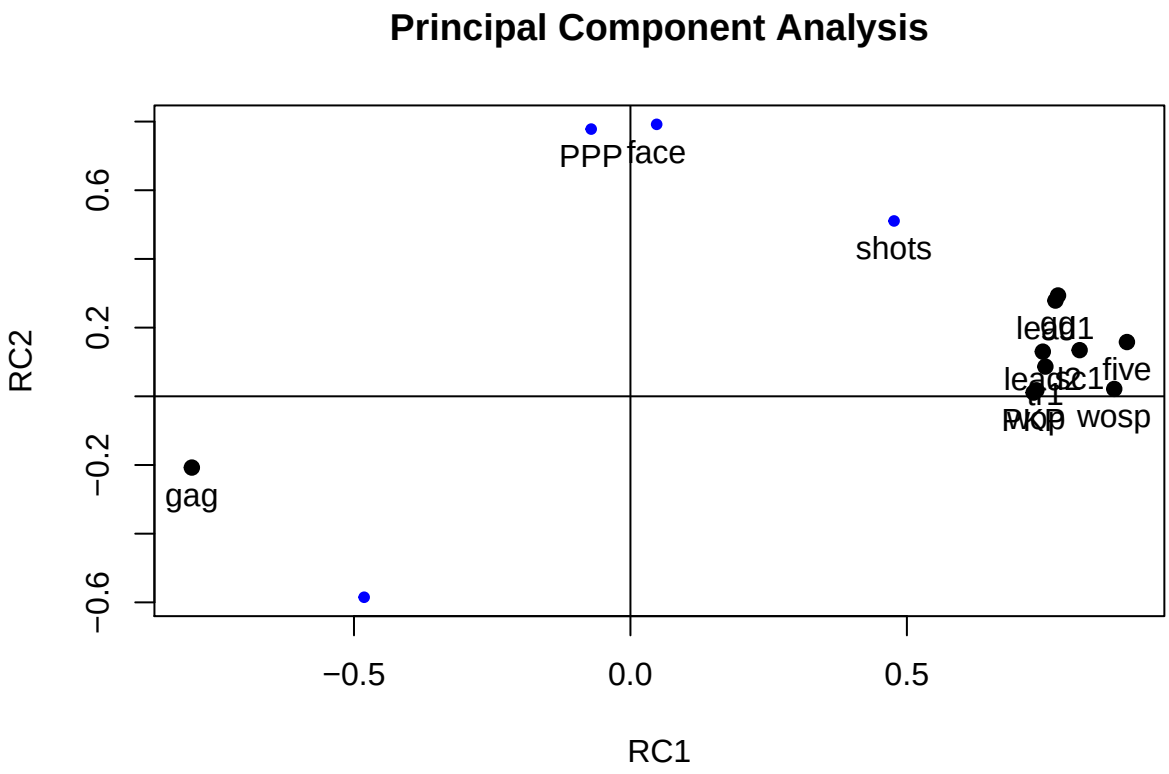
```
# Plot the components and analyze them
factor.plot(pc, choose = c(1), labels = colnames(df), title = "PCA Component 1")
```




```
factor.plot(pc, choose = c(2), labels = colnames(df), title = "PCA Component 2")
```




```
factor.plot(pc, labels = colnames(df))
```




```
rm(list = ls())
```

Observations made from the plots - - Face and PPP loads only on Component-2. - gg, gag, five, PKP, sc1, tr1, lead1, lead2, wop and wosp load only on Component-1. - shots and sag load on both Component-1 and Component-2.

---

## Problem 2

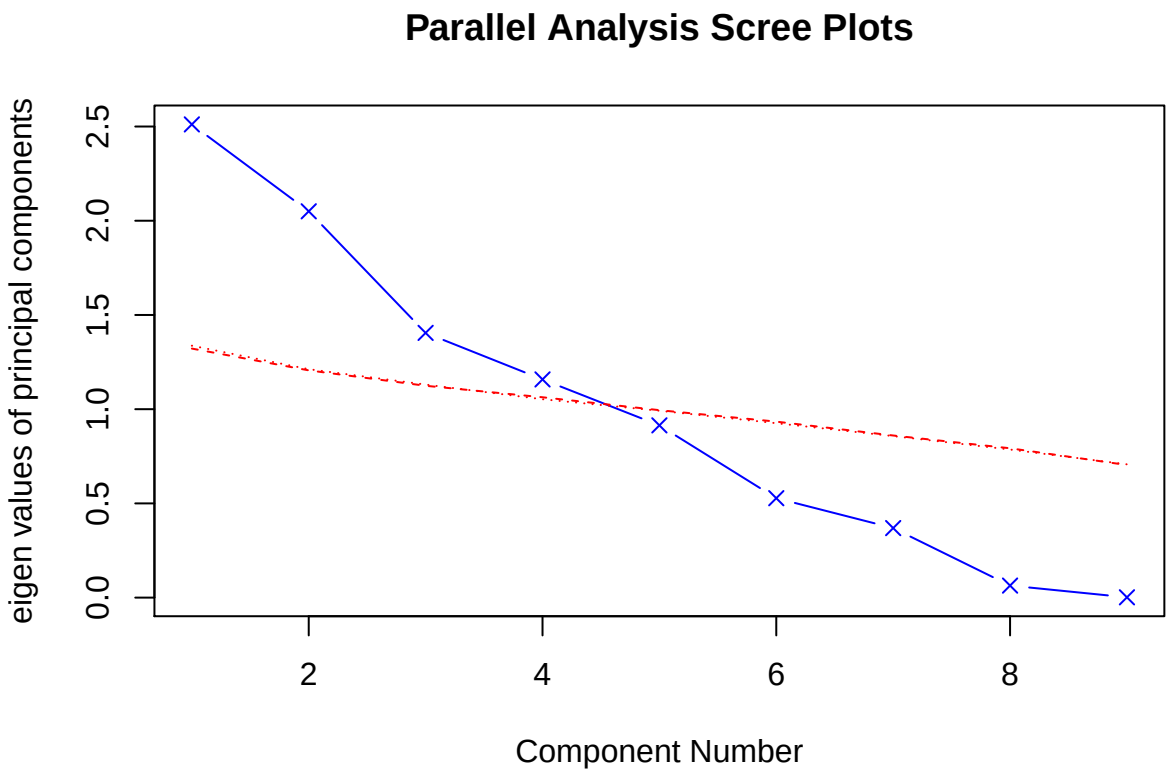
Perform principal component analysis on Glass Identification Data.xlsx • Input the raw data matrix to fa.parallel() function to determine the number of components to extract • Input the raw data matrix to principal() function to extract the components. If raw data is input, the correlation matrix is automatically calculated by principal() function. • Rotate the components • Compute component scores • Graph an orthogonal solution using factor.plot() • Interpret the results

```
glass_identification_data <- data.frame(read_xlsx("./data/Glass Identification Data.xlsx",
                                                  sheet = "Glass Data"))

df <- glass_identification_data[, 2:10]

fa.parallel(df, fa = "pc", n.iter = 100, show.legend = FALSE)
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect. Try a
## different factor score estimation method.
```

```
## Warning in fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate = rotate, : An
```

```
## ultra-Heywood case was detected. Examine the results carefully
```




```
## Parallel analysis suggests that the number of factors =  NA  and the number of components =  4
```

The Skree Plot and the fa.parallel() function suggests nfactors = 4.

```
# Perform PCA with
pc <- principal(df, nfactors = 4, rotate = "equamax", scores = TRUE)
```

```
## Loading required namespace: GPArotation
```

```
pc
```

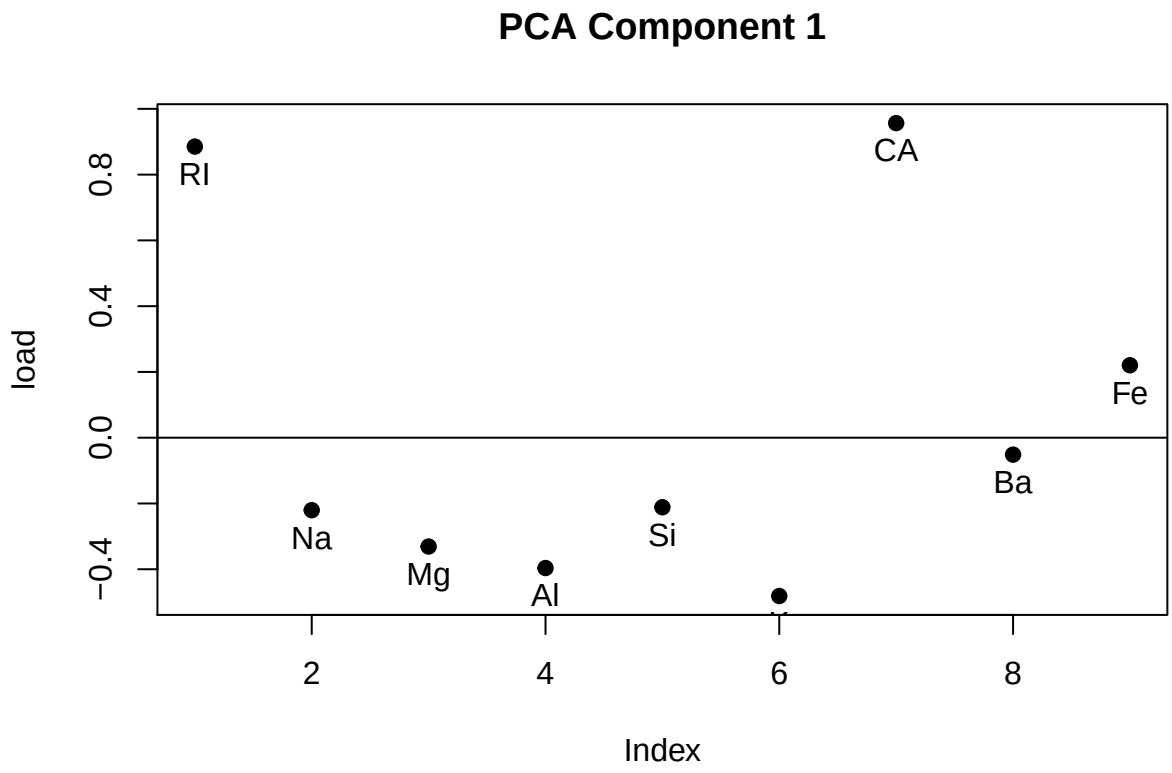
```
## Principal Components Analysis
## Call: principal(r = df, nfactors = 4, rotate = "equamax", scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##      RC1   RC2   RC3   RC4   h2    u2 com
## RI  0.89 -0.09 -0.02  0.40 0.95 0.051 1.4
## Na -0.22  0.25 -0.82  0.12 0.80 0.195 1.4
## Mg -0.33 -0.87  0.09  0.22 0.92 0.081 1.4
## Al -0.40  0.80  0.13  0.07 0.81 0.186 1.5
## Si -0.21  0.03  0.00 -0.96 0.97 0.031 1.1
## K  -0.48  0.20  0.63  0.35 0.79 0.212 2.7
## CA  0.96  0.11  0.12 -0.02 0.94 0.058 1.1
## Ba -0.05  0.73 -0.30  0.20 0.67 0.333 1.5
## Fe  0.22 -0.06  0.46  0.05 0.27 0.730 1.5
##
##                        RC1  RC2  RC3  RC4
## SS loadings           2.34 2.05 1.42 1.31
```

```
## Proportion Var        0.26 0.23 0.16 0.15
## Cumulative Var        0.26 0.49 0.65 0.79
## Proportion Explained  0.33 0.29 0.20 0.18
## Cumulative Proportion 0.33 0.62 0.82 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.08
##  with the empirical chi square  102.53  with prob <  7.4e-20
##
## Fit based upon off diagonal values = 0.92
```

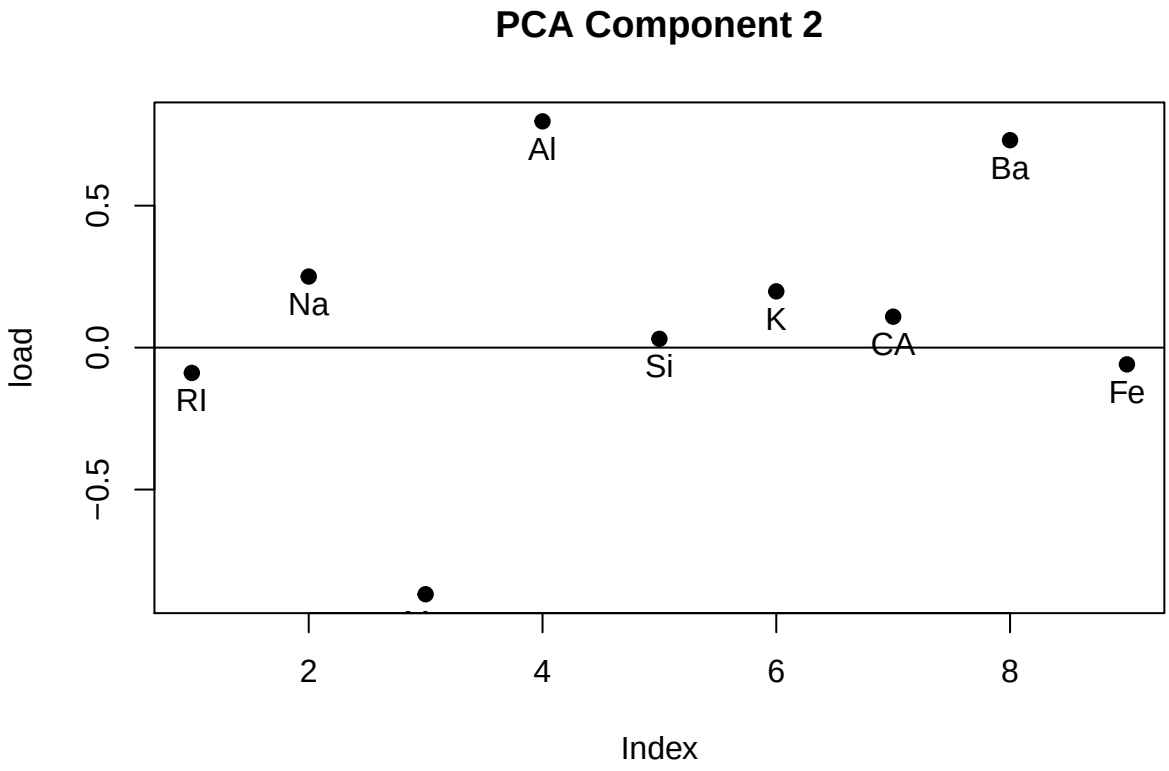
```
head(pc$scores)
```

```
##             RC1        RC2        RC3        RC4
## [1,]  0.1618402 -1.1279706 -0.8956385  1.1082699
## [2,] -0.6432782 -0.5585107 -0.6234778  0.1134193
## [3,] -0.7933298 -0.4140292 -0.3324192 -0.3369684
## [4,] -0.4559932 -0.6220157 -0.0816441  0.1174616
## [5,] -0.5111960 -0.6305315 -0.1247296 -0.3512540
## [6,] -0.5009036 -0.2487297  1.3010988 -0.3917468
```

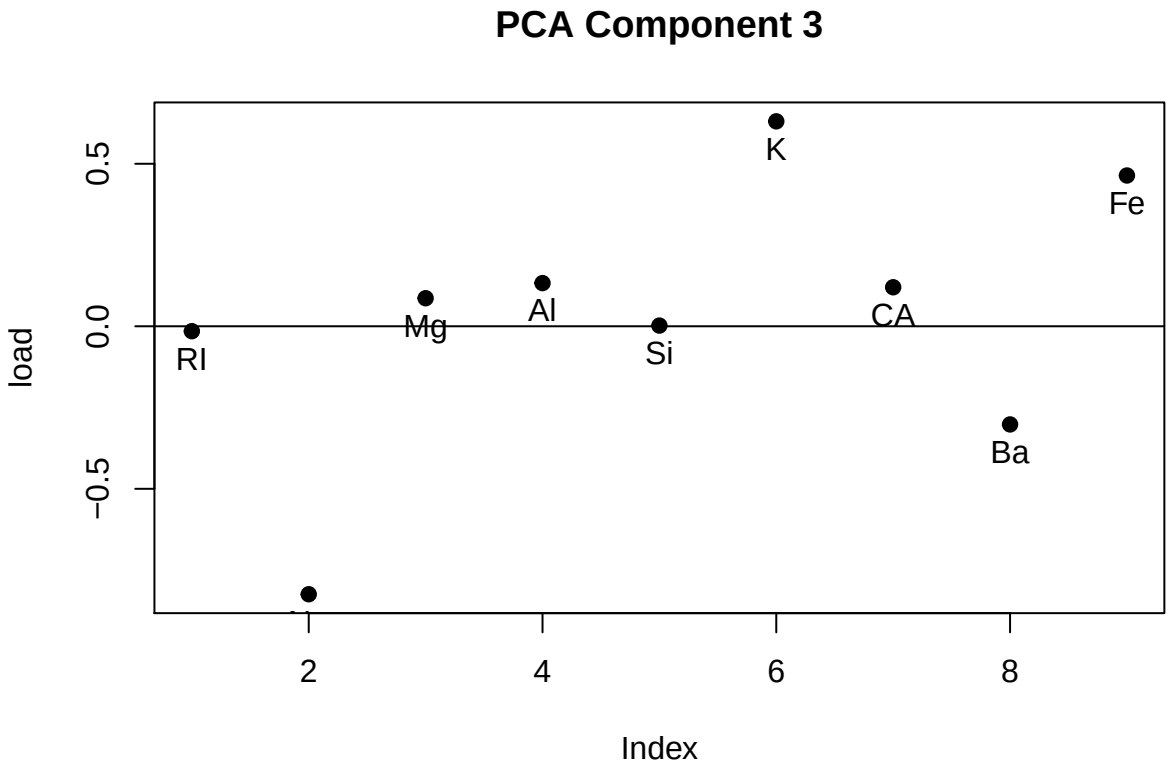
```
# Plot the components and analyze them
factor.plot(pc, choose = c(1), labels = colnames(df), title = "PCA Component 1")
```



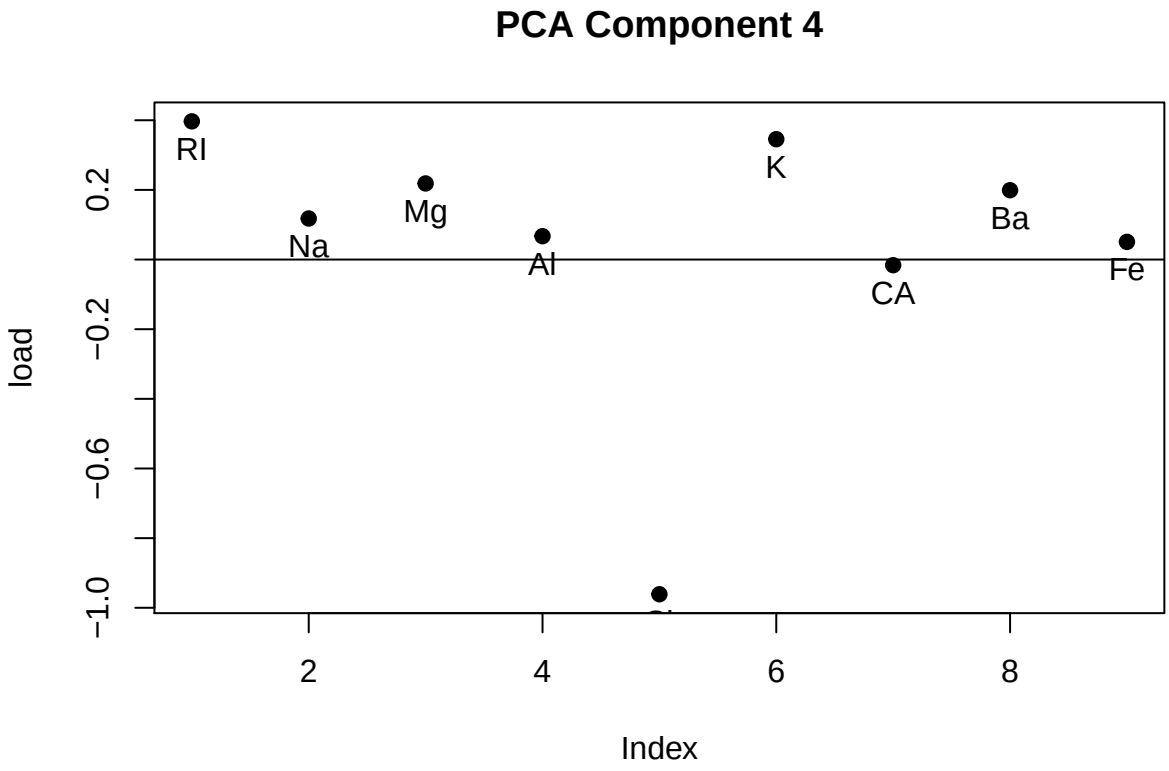

```
factor.plot(pc, choose = c(2), labels = colnames(df), title = "PCA Component 2")
```



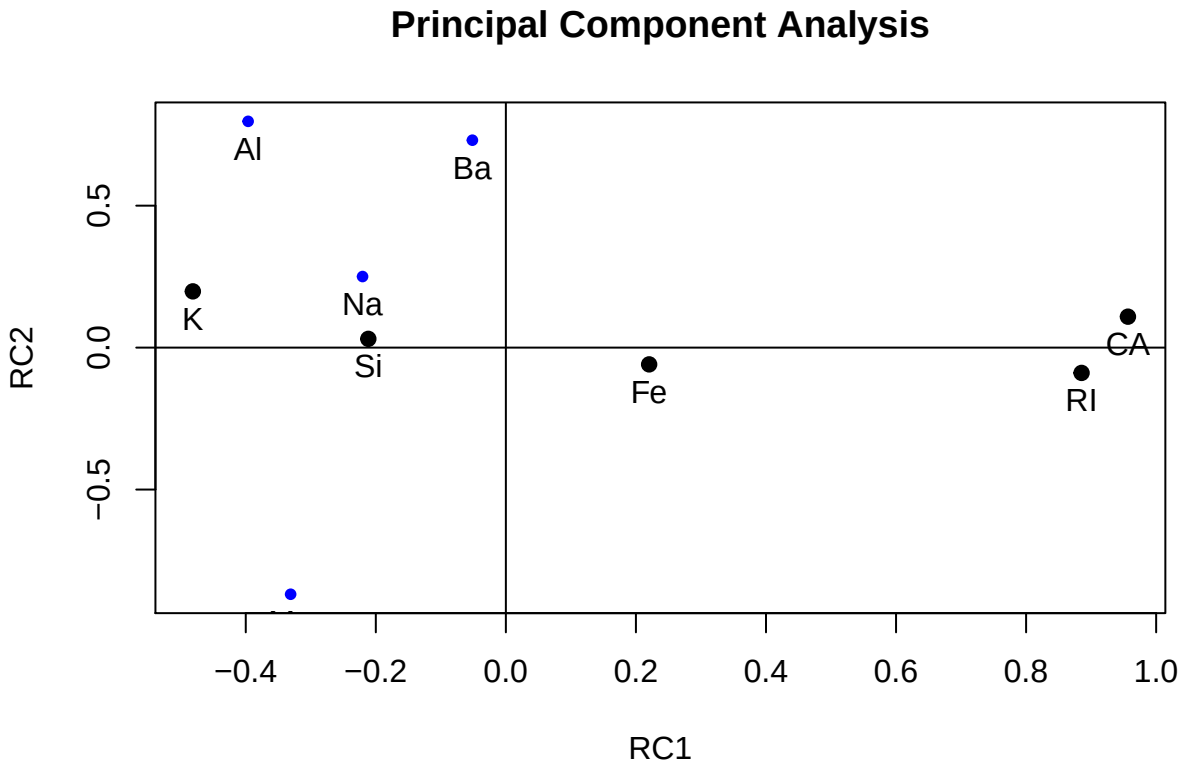

```
factor.plot(pc, choose = c(3), labels = colnames(df), title = "PCA Component 3")
```



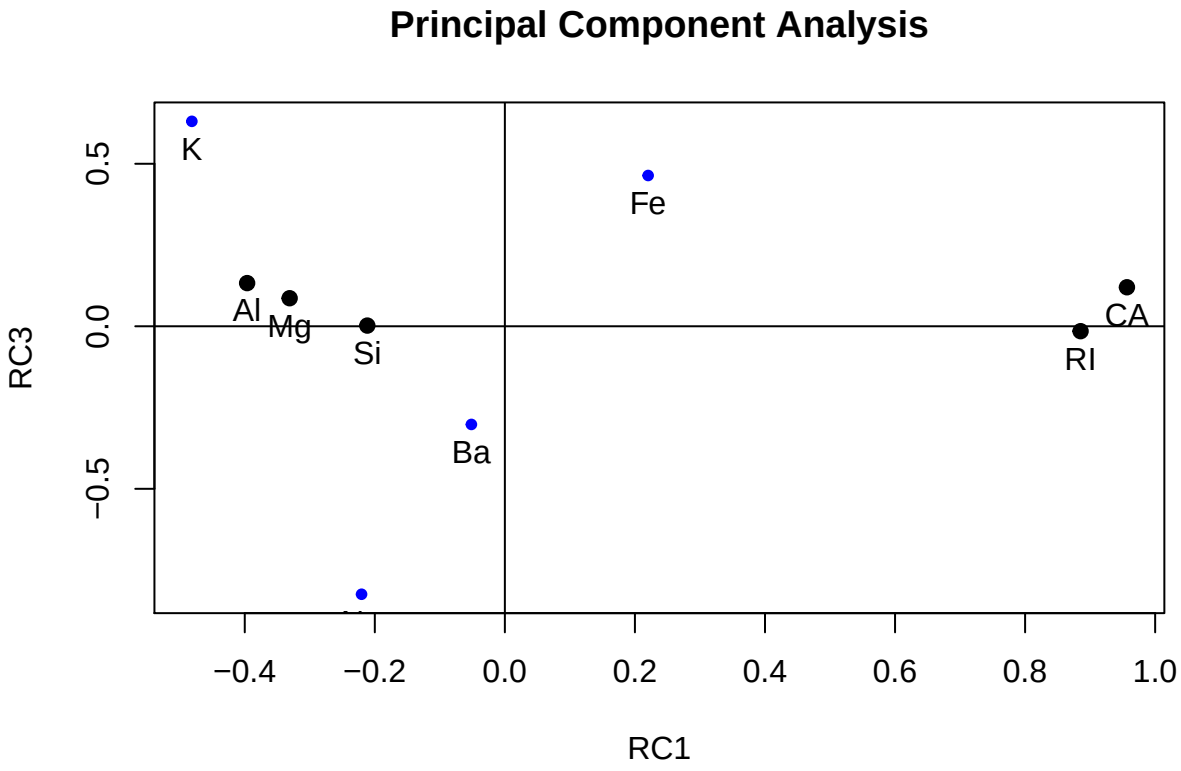

```
factor.plot(pc, choose = c(4), labels = colnames(df), title = "PCA Component 4")
```



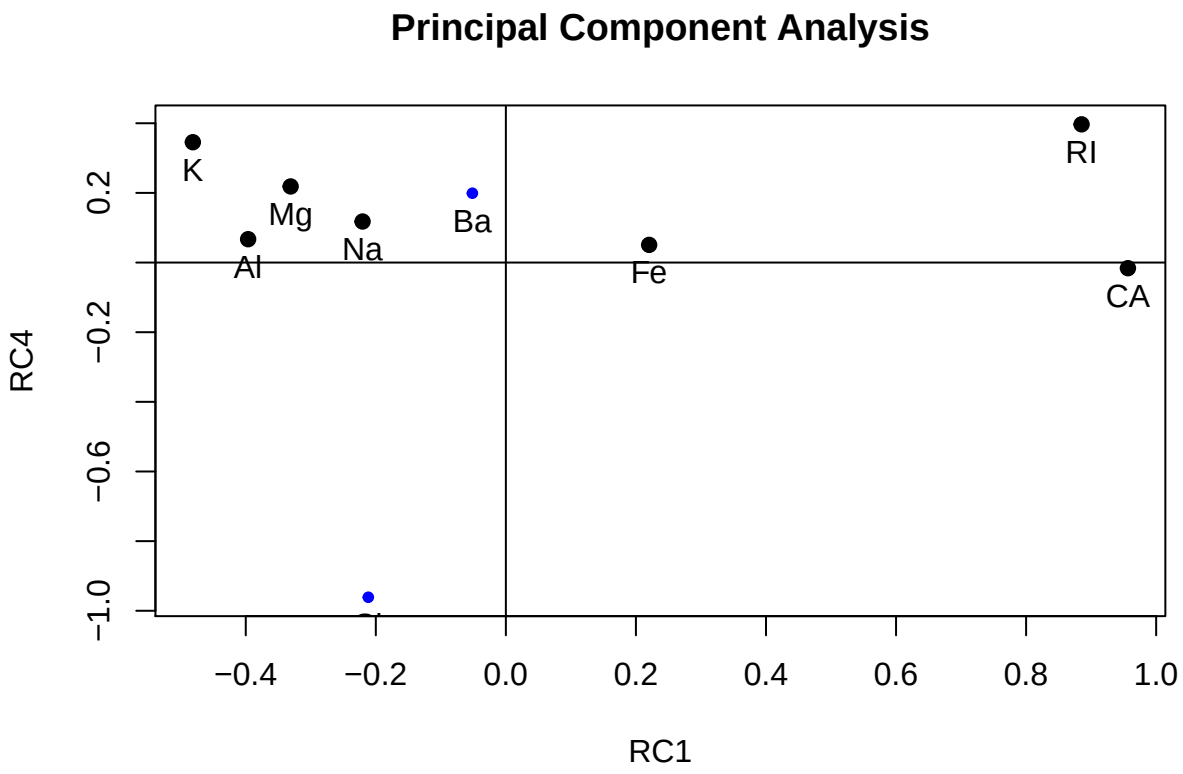

```
factor.plot(pc, choose = c(1, 2), labels = colnames(df))
```



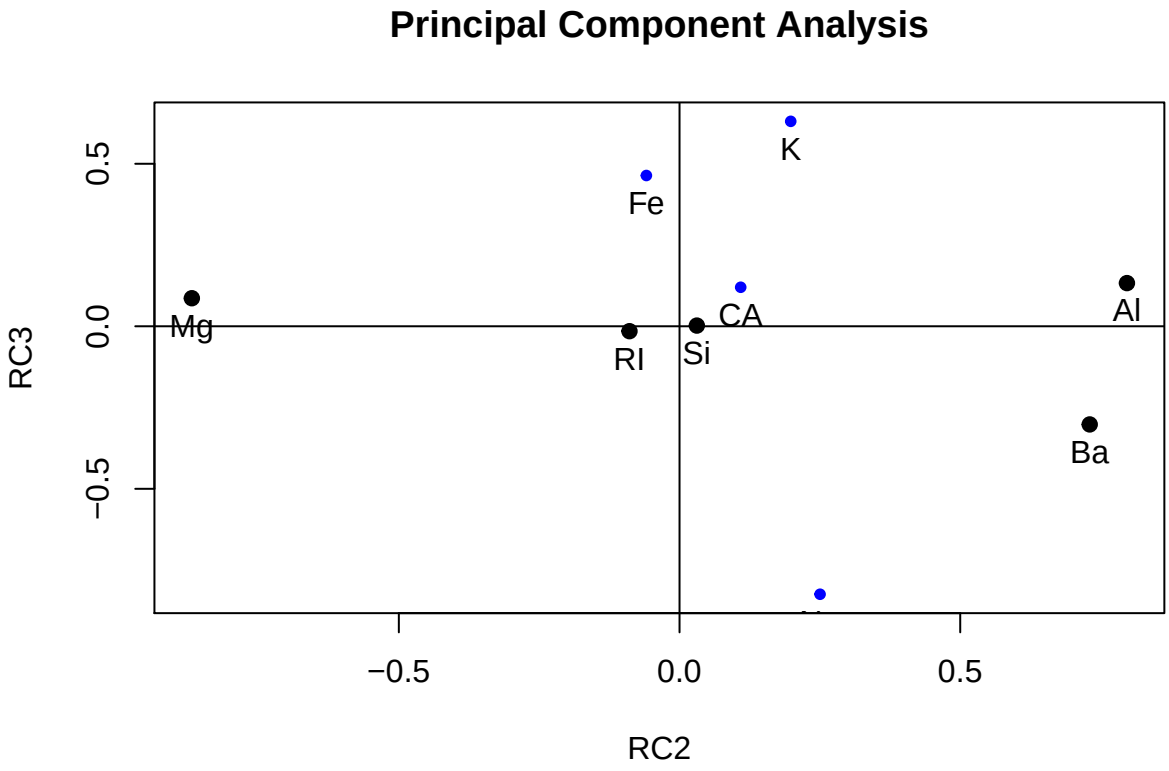

```
factor.plot(pc, choose = c(1, 3), labels = colnames(df))
```



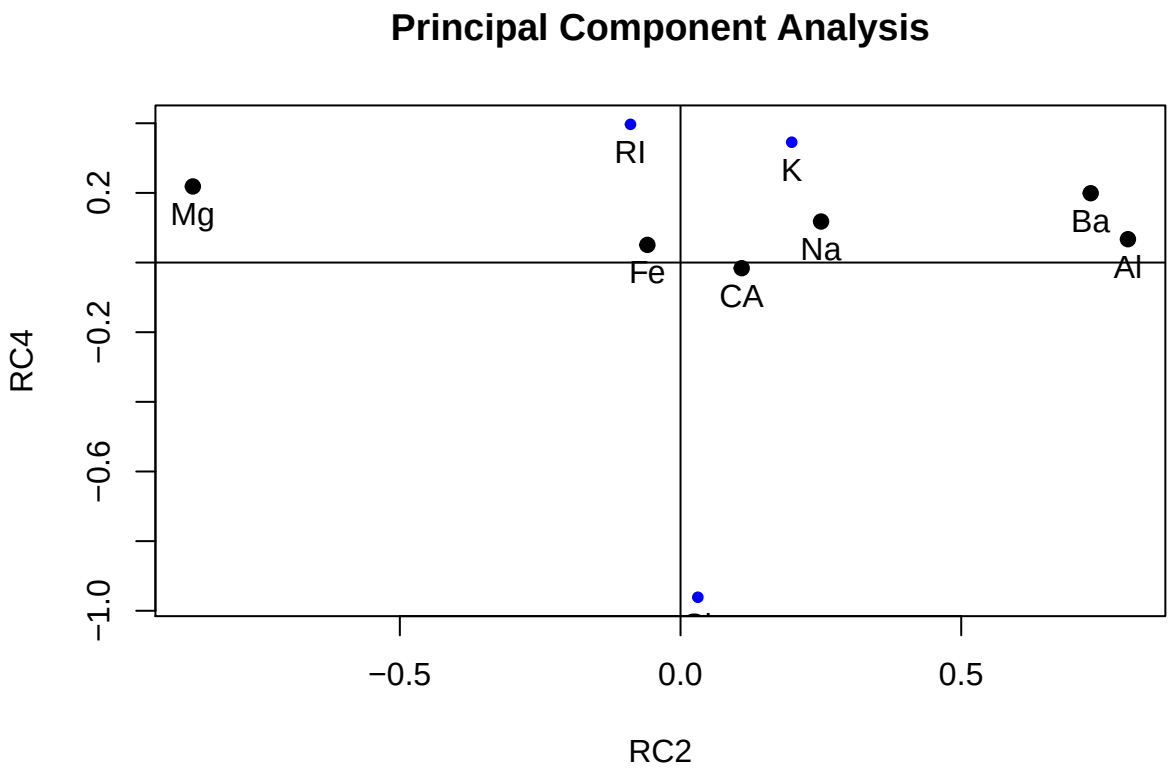

```
factor.plot(pc, choose = c(1, 4), labels = colnames(df))
```



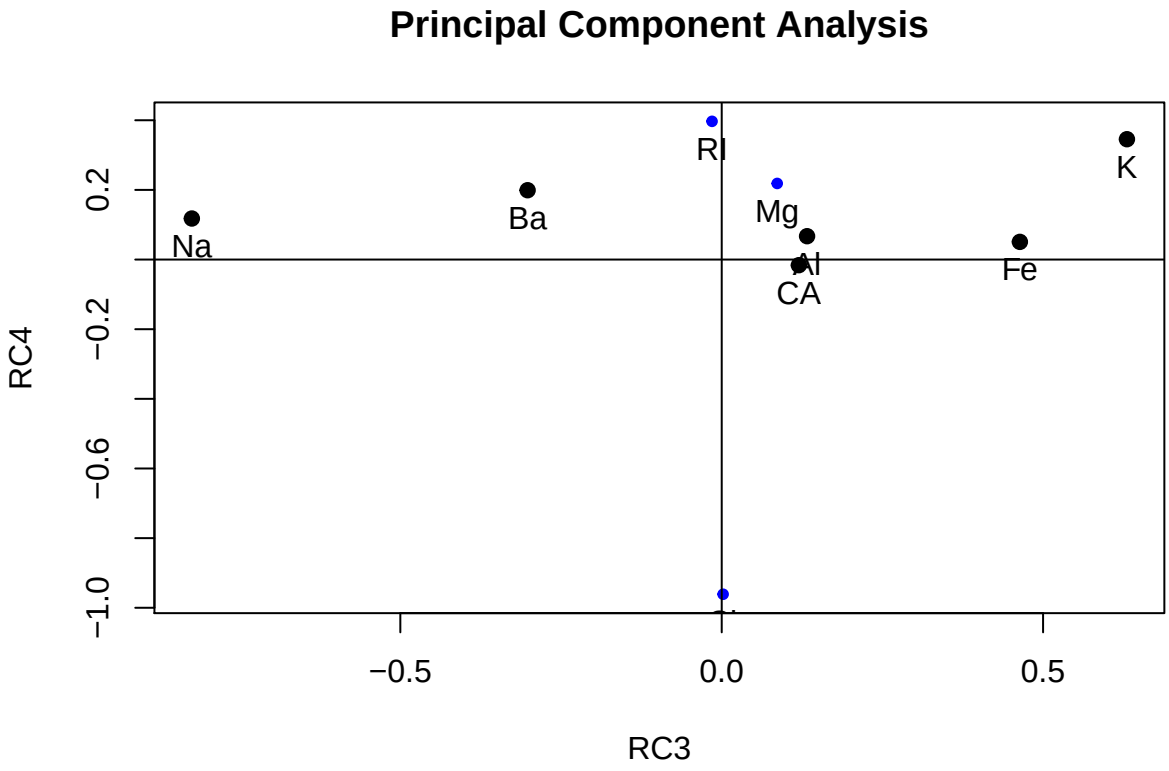

```
factor.plot(pc, choose = c(2, 3), labels = colnames(df))
```




```
factor.plot(pc, choose = c(2, 4), labels = colnames(df))
```




```
factor.plot(pc, choose = c(3, 4), labels = colnames(df))
```




```
rm(list = ls())
```

Observations made - - PC1: loads RI, CA, Mg, Al and K.

- PC2: loads Mg, Al and Ba.
- PC3: loads Na, K, Ba and Fe.
- PC4: loads RI and Si.

Hence - 1. PC1 signifies Calcium, Potassium, Magnesium and Aluminum heavy glass. It also signifies glass with high refractive index. 2. PC2 signifies glass with high Magnesium, Aluminum and Barium concentration. 3. PC3 signifies glass with high Sodium, Potassium, Barium and Iron concentration. 4. PC4 signifies glass with high Refractive Index and Silicon concentration.

---