Data Science Methods - Assignment 1

M. Alberti, 2020162 N.R. Ceschin

February 21, 2020

First we upload all relevant libraries:

```
library(readxl)
library(ggplot2)
library(ggfortify)
library(dplyr)
library(tidyr)
library(RCurl)
library(ggrepel)
```

Upload dataset:

```
setwd("C:/Users/Mr Nobody/Desktop/Uni/EME/Data science Methods/Assignments")
#setwd("~/Tilburg/Courses/Data Science Methods/Assignment1/DATA-SCIENCE-ASSIGNMENTS")
data<-read_excel("env_air_emis.xls")
```

After a quick glimse to the data we realized that data for the five pollutant are presented in separated tables, the separation contains some information in the first column and NA cells in the rest. To be sure not to drop NAs in the middle of the dataset, we first proceed to drop all raws containing at least 5 NA values and we assign to *df*:

```
dim(data)
df<-data[rowSums(is.na(data))<length(data)-5,]
#df<-data[complete.cases(data), ]
```

Given the data structure and the exercises a-c requests, we decided that the optimal approach would be looping over the chunks of data containing information for each pollutant, producing without repeting the code the outputs all in one step. First we create some variables that will be used in the loop:

```
#build 'index' for your loop
interval<-c(1,30,59,88,117)   #number of the first row of each inividiidual dataset
pollutants<-c("ammonia","nmvoc","smallpart","largepart","sulphur")
index<-data.frame(interval,pollutants)

PC1<-data.frame(matrix(ncol=5,nrow=28))  #empty data frames that will be filled with the scores
PC2<-data.frame(matrix(ncol=5,nrow=28))  # of the PC 1 and 2 for each pollutant and country

df['Short name']<- substr(df[[1]], start = 1, stop = 3) #Create country name abbreviation
```

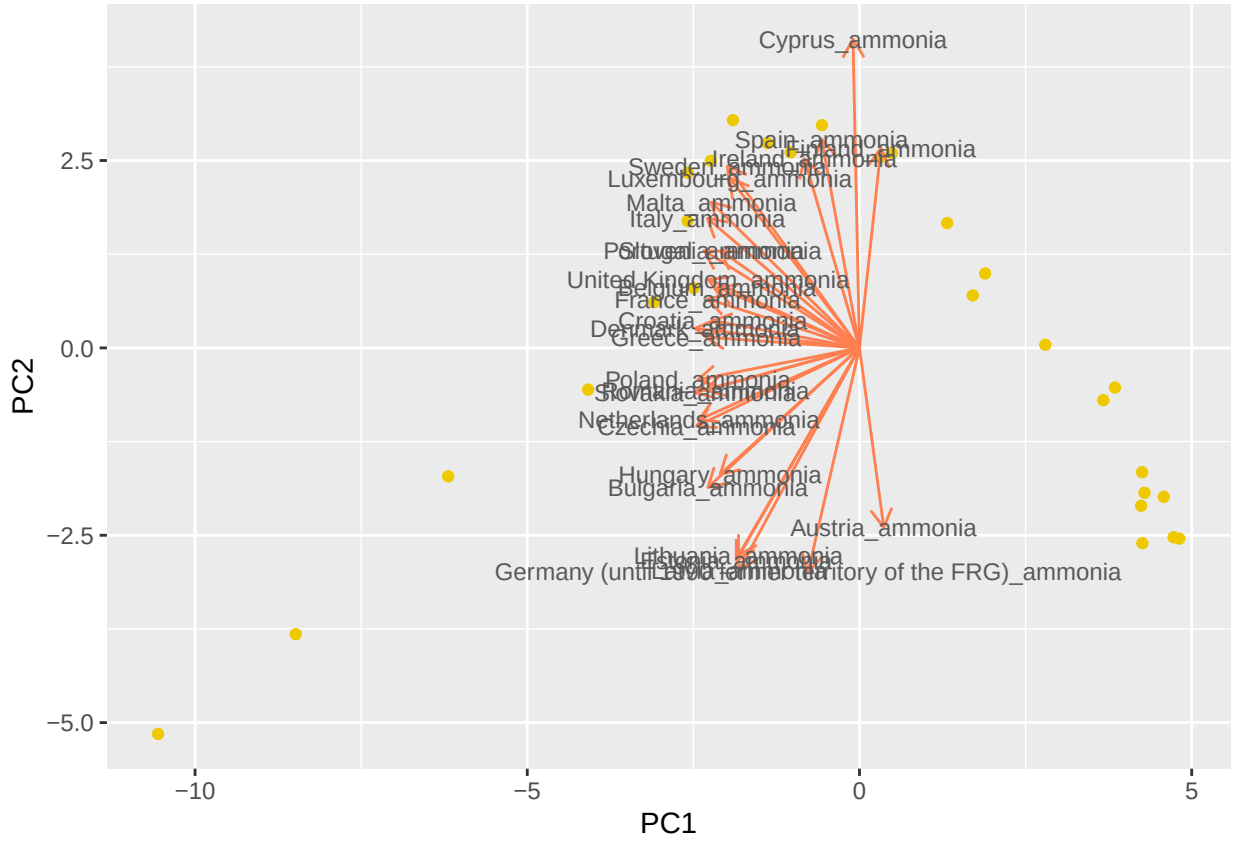
```
for (i in 1:5){

  #data chunk preparation
  begin<-index[i,1]
  end<-index[i,1]+28      #each chunk has 27 countries plus the first raw with years
  dfx<-df[begin:end,]            #slice portion of the dataframe, 'according to begin' and 'end'
```

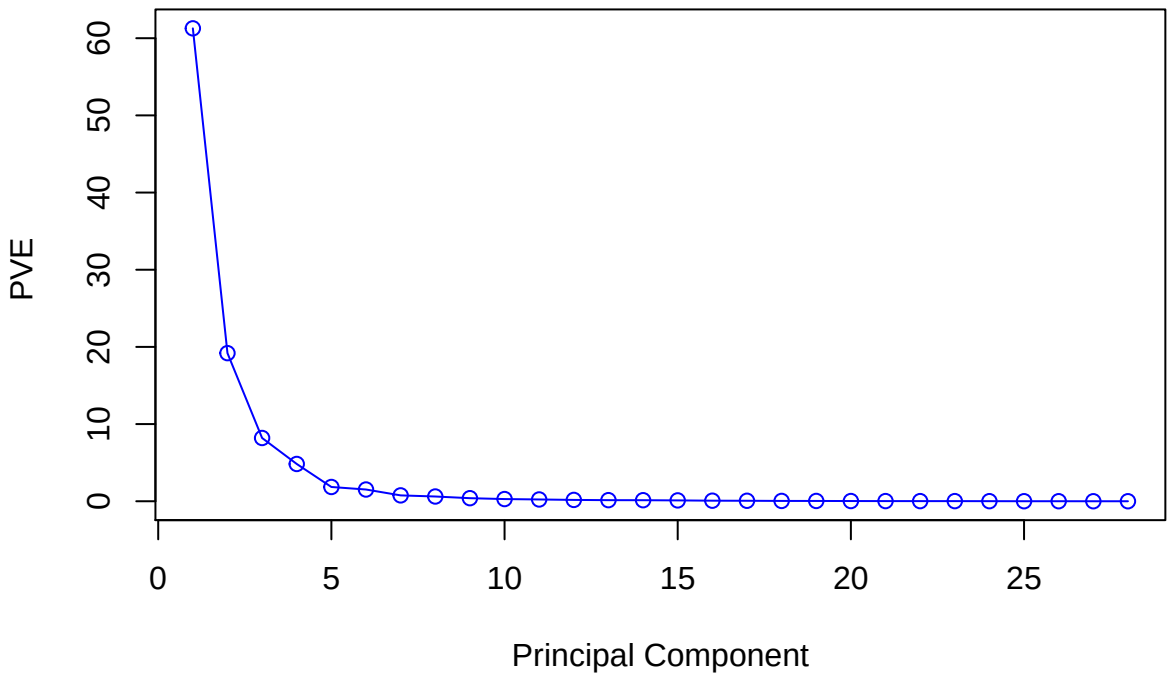
```
dfx[[1]]<-paste(dfx[[1]],index[i,2],sep="_")  #rename first column with the name of the pollutant
dfx<-as.data.frame(dfx)
colnames(dfx)<-dfx[1,]          #set first column as observations' names and first row as variables' n
rownames(dfx)<-dfx[,1]
dfx<-dfx[c(2:29),c(2:29)]       #drop first column and obtain the final datset
dfx<-as.data.frame(t(dfx))      #convert factor columns into numeric to apply prcomp
indx <- sapply(dfx, is.factor)
dfx[indx] <- lapply(dfx[indx], function(x) as.numeric(as.character(x)))

#Principal Component Analysis
pr.out<-prcomp(dfx, scale=TRUE)
print(pr.out$rotation[,1:2])           # print first two PC loadings and plot first two PC
graph<-autoplot(pr.out,variance_percentage=FALSE,loadings=TRUE,
         loadings.label=TRUE,loadings.colour="coral",loadings.label.size=3,
         loadings.label.colour="grey35", scale=0,
         colour="gold2")    # to get labels nicely plotted use ggplot+geom_text_repel()
print(graph)
pve =100* pr.out$sdev ^2/ sum(pr.out$sdev ^2)  #screeplot
scree<-plot(pve , type ="o", ylab="PVE ", xlab=" Principal Component ",
     col =" blue")
print(scree)

#compute vector of BIC for first 27 principal components
BIC<-c(1:27)   #initialize a numeric vector to be filled with BIC(k) values. set max k=p-1
for (j in 1:27) {
  f<-pr.out$x[,1:j]%*%t(pr.out$rotation[,1:j]) #compute aF in X=aF+e
  res_mat<-scale(dfx)-f                        #compute matrix of residuals
  res_mat_sq<-res_mat*res_mat                  #square residuals
  res<-(sum(rowSums(res_mat_sq))/28^2)         #residuals sum of squares
  k<-j
  BICk<-log(res)+k*(log(28^2)/(28^2))          #BIC for each k
  BIC[j]<-BICk                                 #fill BIC vector at each iteration
}
min<-min(BIC)
num_pc<-match(min,BIC)                         #find and print k, the index of the min of BIC
cat("According to the BIC criterion, the optimal number of principal components is ", num_pc)

###potential issue: smallest value for BIC is always the one with ###
###the max number of principal components...strange!I checked the calculations###
###and they seem fine. I think the issue is that the penalty part of BIC is really###
###trivial compared to the log(SSR) part####

#save first two PC in separate dataset for point d)
PC1[i]<-pr.out$x[,1]
colnames(PC1)[i]<-as.character(index[i,2])
PC2[i]<-pr.out$x[,2]
colnames(PC2)[i]<-as.character(index[i,2])

#save relevant objects with their respective name
assign(paste0("BIC_", index[i,2]), BIC)
assign(paste0("df_", index[i,2]), dfx)
assign(paste0("prcomp_",index[i,2]),pr.out)
assign(paste0("Screeplot_",index[i,2]),scree)
```

```
  assign(paste0("PC1-PC2_",index[i,2]),graph)
  #remove non relevant objects
  rm(dfx)
  rm(BIC)
  rm(pr.out)

}
```

```
##                                                                 PC1
## Belgium_ammonia                                          -0.204354811
## Bulgaria_ammonia                                         -0.217650067
## Czechia_ammonia                                          -0.233873843
## Denmark_ammonia                                          -0.236223091
## Germany (until 1990 former territory of the FRG)_ammonia -0.073673803
## Estonia_ammonia                                          -0.176900050
## Ireland_ammonia                                          -0.078943807
## Greece_ammonia                                           -0.219939409
## Spain_ammonia                                            -0.054272914
## France_ammonia                                           -0.217905915
## Croatia_ammonia                                          -0.210606402
## Italy_ammonia                                            -0.218250515
## Cyprus_ammonia                                           -0.009275632
## Latvia_ammonia                                           -0.172880552
## Lithuania_ammonia                                        -0.173839084
## Luxembourg_ammonia                                       -0.186278414
## Hungary_ammonia                                          -0.200010020
## Malta_ammonia                                            -0.212663056
## Netherlands_ammonia                                      -0.230512687
## Austria_ammonia                                           0.034912426
## Poland_ammonia                                           -0.231969539
## Portugal_ammonia                                         -0.223104123
## Romania_ammonia                                          -0.221232865
## Slovenia_ammonia                                         -0.199775456
## Slovakia_ammonia                                         -0.235965094
## Finland_ammonia                                           0.030855084
## Sweden_ammonia                                           -0.189550097
## United Kingdom_ammonia                                   -0.217133926
##                                                                 PC2
## Belgium_ammonia                                           0.07698565
## Bulgaria_ammonia                                         -0.17772705
## Czechia_ammonia                                          -0.09957224
## Denmark_ammonia                                           0.02481319
## Germany (until 1990 former territory of the FRG)_ammonia -0.28500382
## Estonia_ammonia                                          -0.27035189
## Ireland_ammonia                                           0.24095290
## Greece_ammonia                                            0.01359771
## Spain_ammonia                                             0.26604599
## France_ammonia                                            0.06239290
## Croatia_ammonia                                           0.03470130
## Italy_ammonia                                             0.16498716
## Cyprus_ammonia                                            0.39311813
## Latvia_ammonia                                           -0.28296826
## Lithuania_ammonia                                        -0.26431909
## Luxembourg_ammonia                                        0.21616528
```

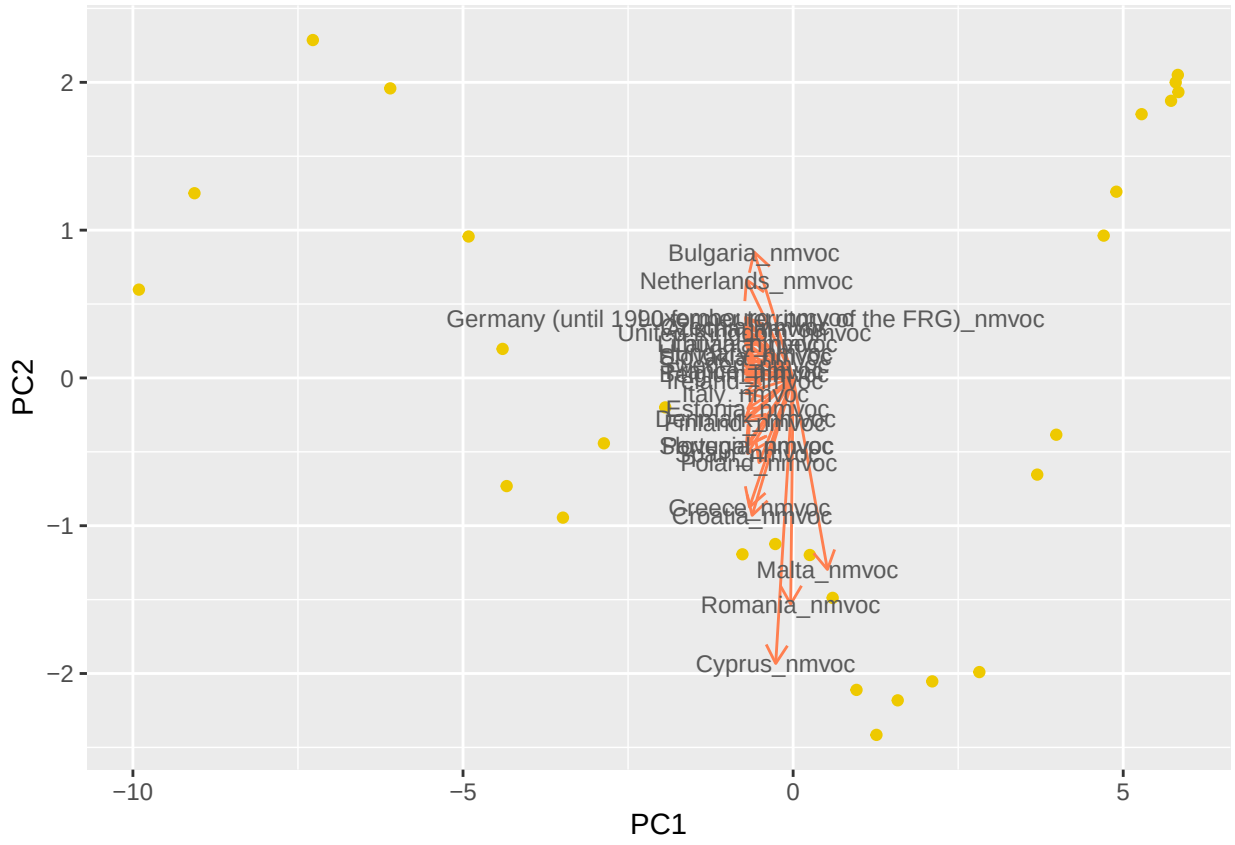
```
## Hungary_ammonia                -0.16080157
## Malta_ammonia                   0.18537648
## Netherlands_ammonia            -0.09143779
## Austria_ammonia                -0.22847877
## Poland_ammonia                 -0.04048120
## Portugal_ammonia                0.12387264
## Romania_ammonia                -0.05394295
## Slovenia_ammonia                0.12388136
## Slovakia_ammonia               -0.05691858
## Finland_ammonia                 0.25417417
## Sweden_ammonia                  0.23151697
## United Kingdom_ammonia          0.08731169
```

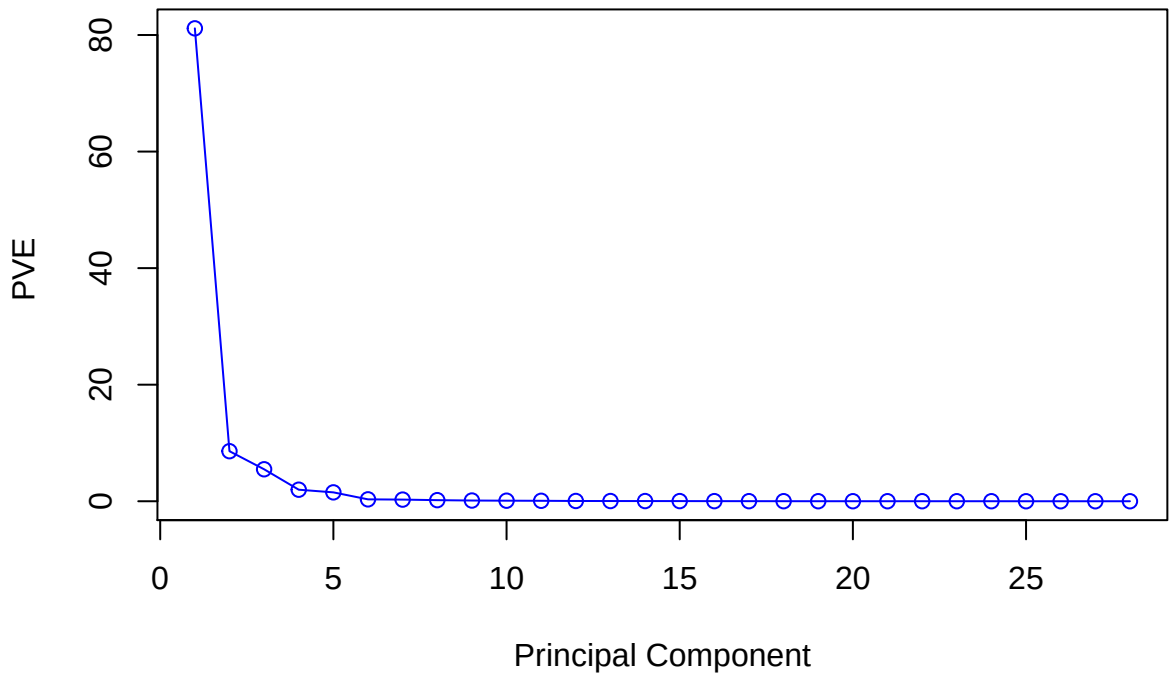





```
## NULL
## According to the BIC criterion, the optimal number of principal components is  27
## Belgium_nmvoc                                              -0.20868663  0.009564652
## Bulgaria_nmvoc                                             -0.16653105  0.239965911
## Czechia_nmvoc                                              -0.20428408  0.098589407
## Denmark_nmvoc                                              -0.20249450 -0.076285018
## Germany (until 1990 former territory of the FRG)_nmvoc -0.20267433  0.112955121
## Estonia_nmvoc                                              -0.19369069 -0.058552040
## Ireland_nmvoc                                              -0.20559374 -0.004281944
## Greece_nmvoc                                               -0.18605115 -0.246202597
## Spain_nmvoc                                                -0.19417498 -0.144155964
## France_nmvoc                                               -0.20760772  0.014466543
## Croatia_nmvoc                                              -0.17486347 -0.261968082
## Italy_nmvoc                                                -0.20323693 -0.028812924
## Cyprus_nmvoc                                               -0.07423076 -0.544267165
## Latvia_nmvoc                                               -0.20363462  0.065851086
## Lithuania_nmvoc                                            -0.19447347  0.065529887
## Luxembourg_nmvoc                                           -0.20265197  0.115754718
## Hungary_nmvoc                                              -0.20154415  0.049216850
## Malta_nmvoc                                                 0.14631305 -0.365109031
## Netherlands_nmvoc                                          -0.19877329  0.186788387
## Austria_nmvoc                                              -0.20659348  0.095247909
## Poland_nmvoc                                               -0.14556406 -0.161347534
## Portugal_nmvoc                                             -0.19234631 -0.128478752
## Romania_nmvoc                                              -0.01045727 -0.431015340
## Slovenia_nmvoc                                             -0.19935598 -0.128586836
```

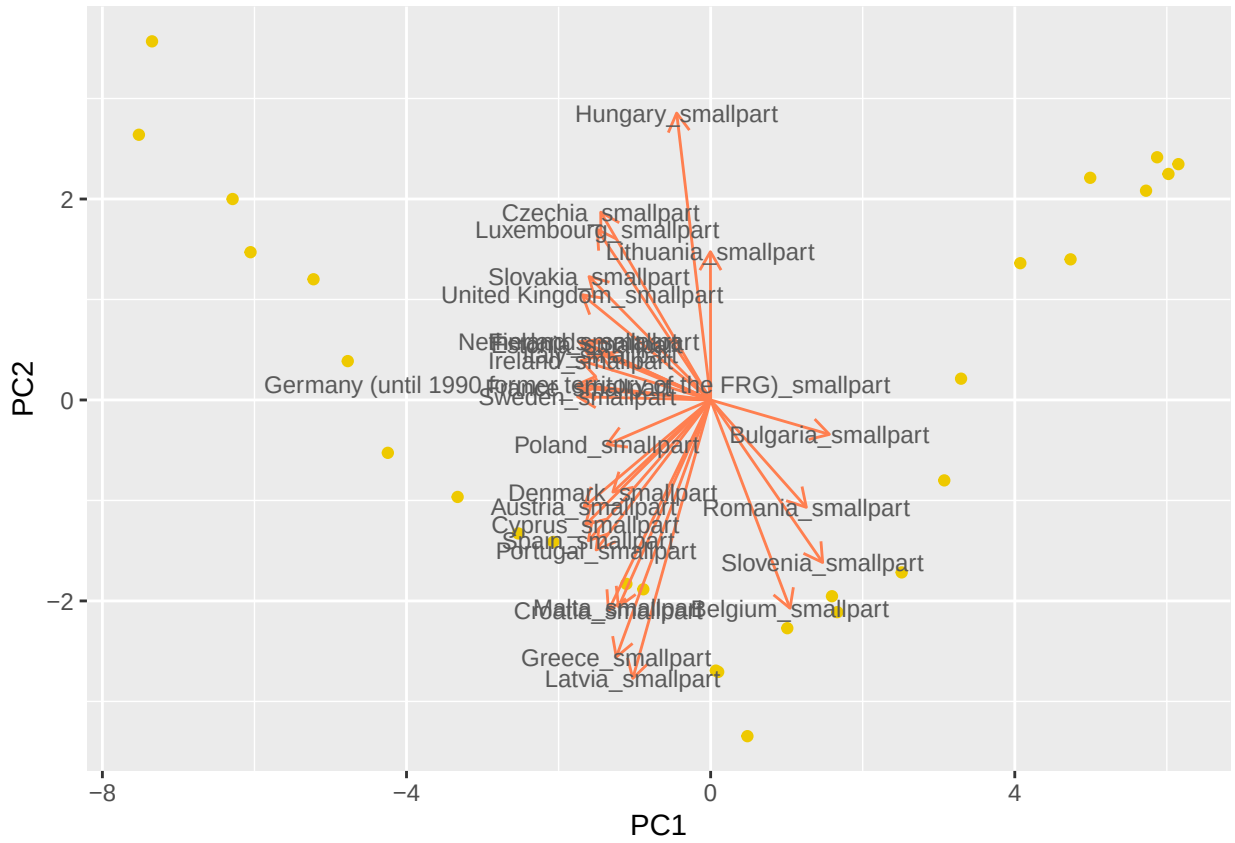
```
## Slovakia_nmvoc                    -0.20393298  0.040312704
## Finland_nmvoc                     -0.20497337 -0.083817829
## Sweden_nmvoc                      -0.20780079  0.025063338
## United Kingdom_nmvoc              -0.20714925  0.086278245
```

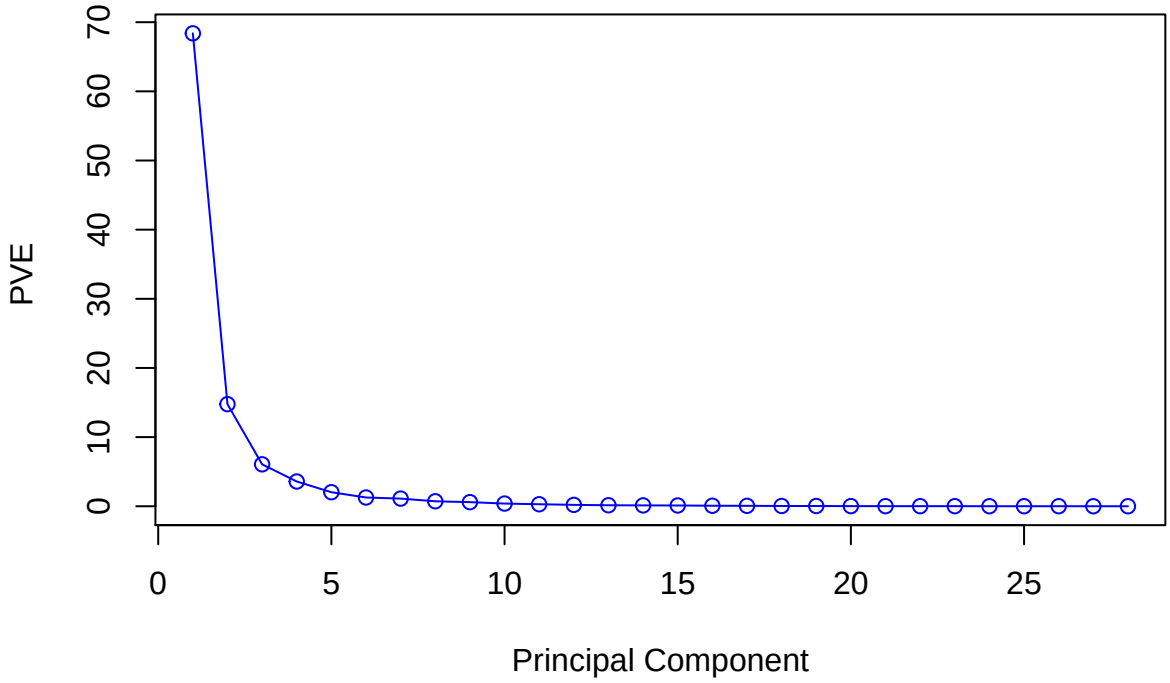





```
## NULL
## According to the BIC criterion, the optimal number of principal components is  27
## Belgium_smallpart                                              0.1352071876
## Bulgaria_smallpart                                             0.2022789094
## Czechia_smallpart                                             -0.1871950042
## Denmark_smallpart                                             -0.1666006870
## Germany (until 1990 former territory of the FRG)_smallpart -0.2271983117
## Estonia_smallpart                                             -0.2109031675
## Ireland_smallpart                                             -0.2213990409
## Greece_smallpart                                              -0.1607633268
## Spain_smallpart                                               -0.2084616844
## France_smallpart                                              -0.2257128025
## Croatia_smallpart                                             -0.1740512341
## Italy_smallpart                                               -0.1881282794
## Cyprus_smallpart                                              -0.2131205911
## Latvia_smallpart                                              -0.1325885817
## Lithuania_smallpart                                           -0.0003864404
## Luxembourg_smallpart                                          -0.1931251572
## Hungary_smallpart                                             -0.0578604185
## Malta_smallpart                                               -0.1570096919
## Netherlands_smallpart                                         -0.2245401637
## Austria_smallpart                                             -0.2158454484
## Poland_smallpart                                              -0.1767932219
## Portugal_smallpart                                            -0.1951488938
## Romania_smallpart                                              0.1629671421
## Slovenia_smallpart                                             0.1907706921
```

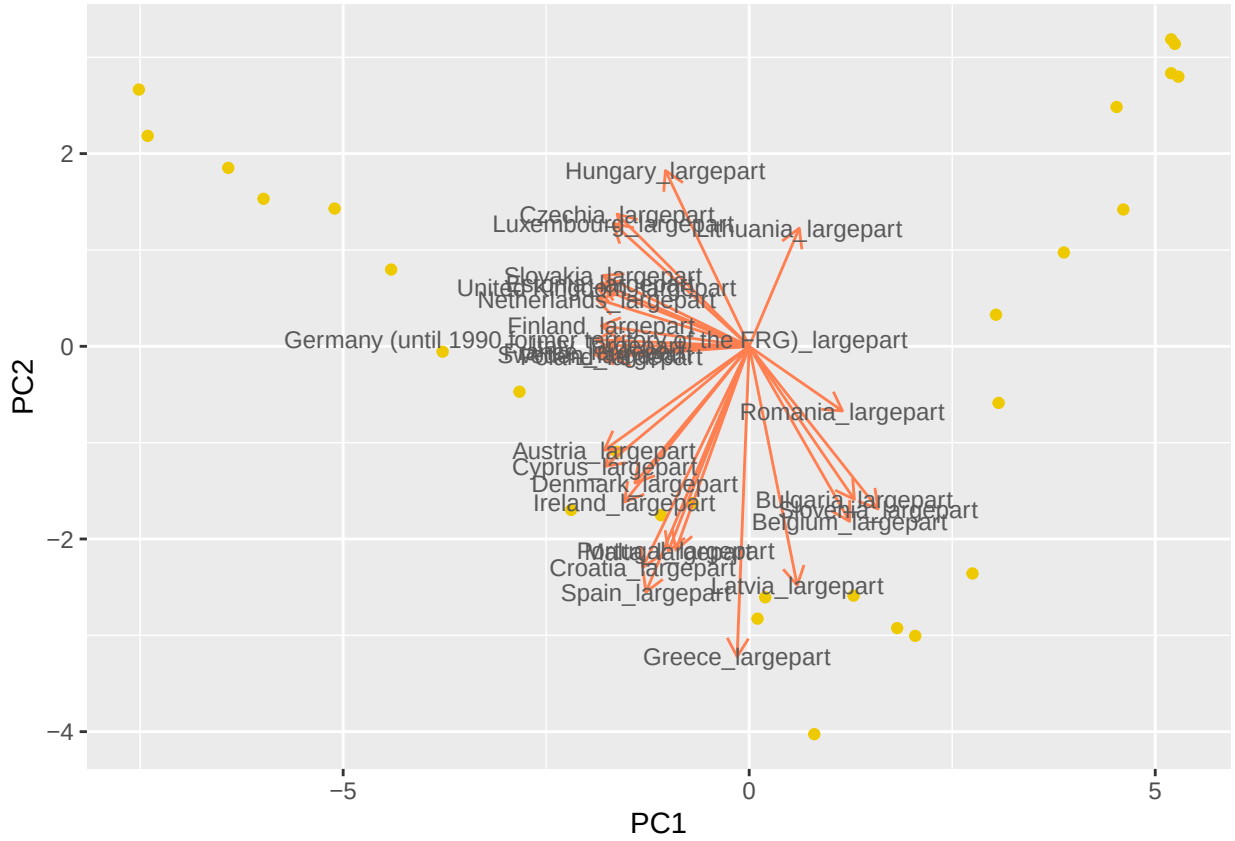
```
## Slovakia_smallpart                                          -0.2072447390
## Finland_smallpart                                           -0.2171427781
## Sweden_smallpart                                            -0.2267045231
## United Kingdom_smallpart                                    -0.2187744897
##                                                                        PC2
## Belgium_smallpart                                           -0.268443368
## Bulgaria_smallpart                                          -0.044369302
## Czechia_smallpart                                            0.242026480
## Denmark_smallpart                                           -0.118723939
## Germany (until 1990 former territory of the FRG)_smallpart  0.020019380
## Estonia_smallpart                                            0.072089099
## Ireland_smallpart                                            0.050870983
## Greece_smallpart                                            -0.331397401
## Spain_smallpart                                             -0.181343121
## France_smallpart                                             0.015996940
## Croatia_smallpart                                           -0.271383475
## Italy_smallpart                                              0.059095767
## Cyprus_smallpart                                            -0.160779709
## Latvia_smallpart                                            -0.358869394
## Lithuania_smallpart                                          0.191100120
## Luxembourg_smallpart                                         0.219789275
## Hungary_smallpart                                            0.369802996
## Malta_smallpart                                             -0.266773853
## Netherlands_smallpart                                        0.074925668
## Austria_smallpart                                           -0.137794278
## Poland_smallpart                                            -0.057511189
## Portugal_smallpart                                          -0.193269224
## Romania_smallpart                                           -0.138397012
## Slovenia_smallpart                                          -0.209494406
## Slovakia_smallpart                                           0.159016157
## Finland_smallpart                                            0.074933099
## Sweden_smallpart                                             0.004294029
## United Kingdom_smallpart                                     0.135631035
```

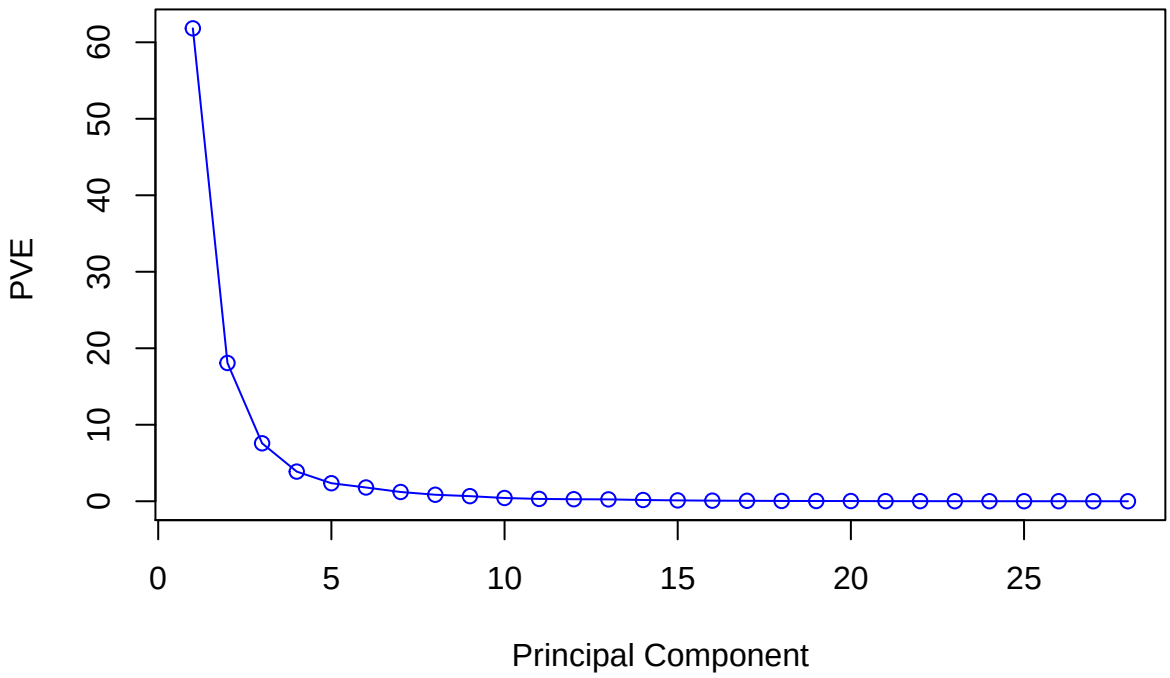

Hungary_smallpart
Czechia_smallpart
Luxembourg_smallpart
Lithuania_smallpart
Slovakia_smallpart
United Kingdom_smallpart
Germany (until 1990 former territory of the FRG)_smallpart
Poland_smallpart
Bulgaria_smallpart
Denmark_smallpart
Austria_smallpart
Romania_smallpart
Cyprus_smallpart
Portugal_smallpart
Slovenia_smallpart
Belgium_smallpart
Greece_smallpart
Latvia_smallpart
PC1
PC2
2
0
−2
−8
−4
0
4




```
## NULL
## According to the BIC criterion, the optimal number of principal components is  27
## Belgium_largepart                                              0.15477068
## Bulgaria_largepart                                             0.16213206
## Czechia_largepart                                             -0.20369439
## Denmark_largepart                                             -0.17618165
## Germany (until 1990 former territory of the FRG)_largepart -0.23607049
## Estonia_largepart                                             -0.22928018
## Ireland_largepart                                             -0.19231607
## Greece_largepart                                              -0.01862135
## Spain_largepart                                               -0.15932480
## France_largepart                                              -0.23755011
## Croatia_largepart                                             -0.16439488
## Italy_largepart                                               -0.21858399
## Cyprus_largepart                                              -0.22275505
## Latvia_largepart                                               0.07406570
## Lithuania_largepart                                            0.07719293
## Luxembourg_largepart                                          -0.20976966
## Hungary_largepart                                             -0.12920025
## Malta_largepart                                               -0.12479029
## Netherlands_largepart                                         -0.23507202
## Austria_largepart                                             -0.22364693
## Poland_largepart                                              -0.21232687
## Portugal_largepart                                            -0.11308466
## Romania_largepart                                              0.14384474
## Slovenia_largepart                                             0.19869857
```

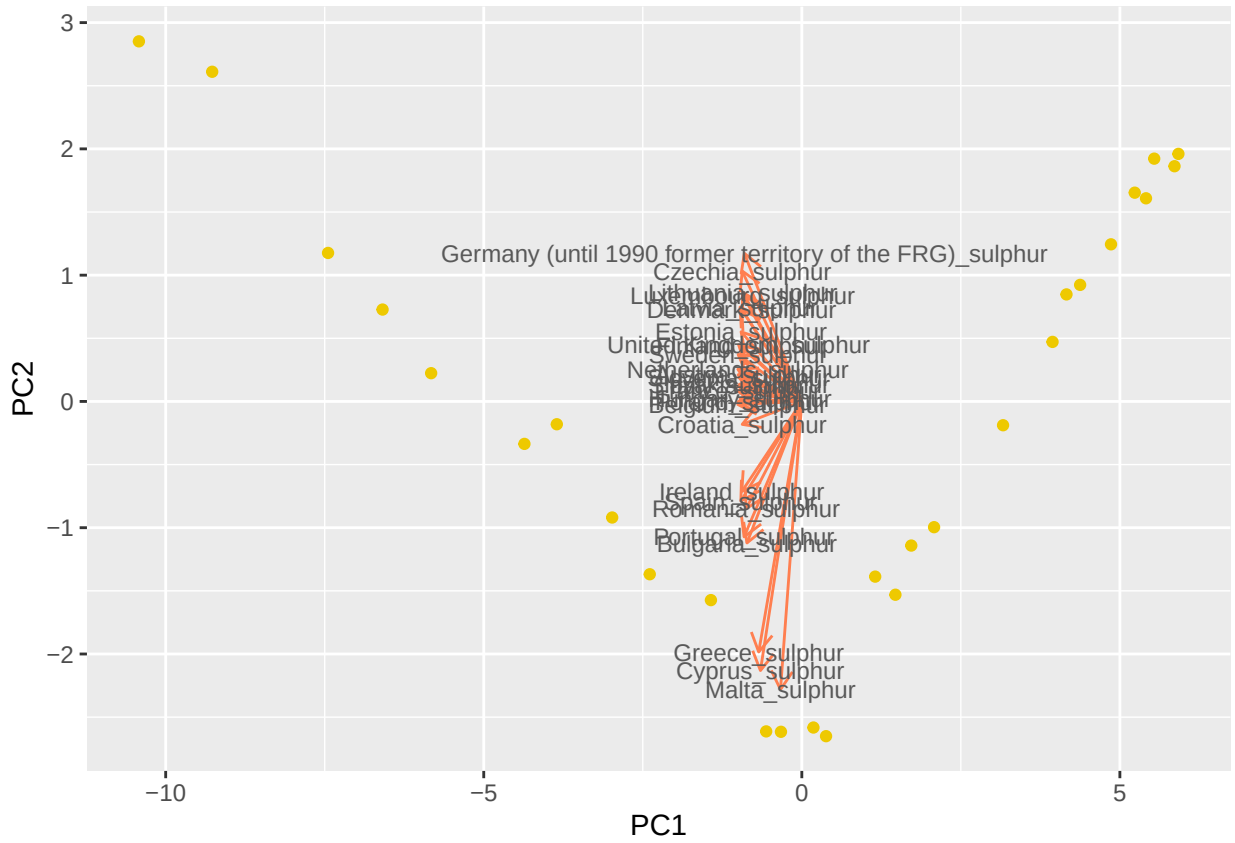
```
## Slovakia_largepart                                          -0.22501216
## Finland_largepart                                           -0.22816963
## Sweden_largepart                                            -0.23748525
## United Kingdom_largepart                                    -0.23501868
##                                                                     PC2
## Belgium_largepart                                           -0.2275145202
## Bulgaria_largepart                                          -0.1989601328
## Czechia_largepart                                            0.1718536070
## Denmark_largepart                                           -0.1776925695
## Germany (until 1990 former territory of the FRG)_largepart  0.0086072939
## Estonia_largepart                                            0.0816498373
## Ireland_largepart                                           -0.2023164257
## Greece_largepart                                            -0.4033132044
## Spain_largepart                                             -0.3195571312
## France_largepart                                            -0.0070766960
## Croatia_largepart                                           -0.2876333544
## Italy_largepart                                              0.0009109591
## Cyprus_largepart                                            -0.1560451588
## Latvia_largepart                                            -0.3104973798
## Lithuania_largepart                                          0.1535279442
## Luxembourg_largepart                                         0.1597616390
## Hungary_largepart                                            0.2284050943
## Malta_largepart                                             -0.2662982445
## Netherlands_largepart                                        0.0616098157
## Austria_largepart                                           -0.1349881988
## Poland_largepart                                            -0.0134585984
## Portugal_largepart                                          -0.2648549618
## Romania_largepart                                           -0.0840427285
## Slovenia_largepart                                          -0.2113784344
## Slovakia_largepart                                           0.0921581549
## Finland_largepart                                            0.0268618869
## Sweden_largepart                                            -0.0108128048
## United Kingdom_largepart                                     0.0771321031
```

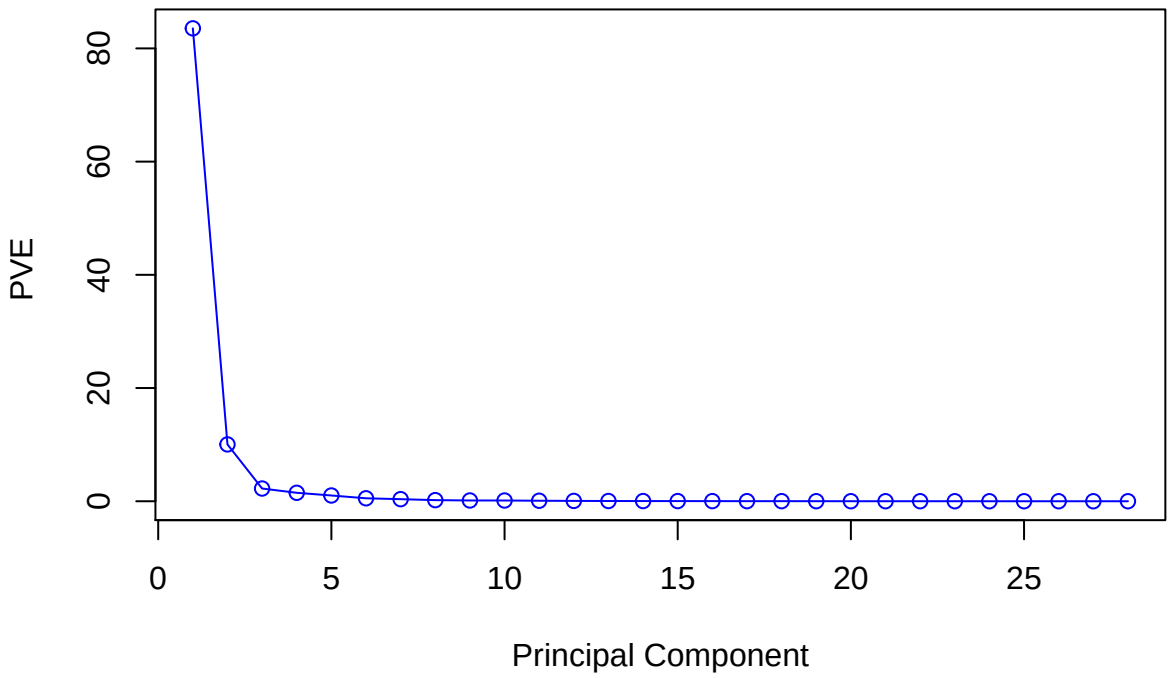

Hungary_largepart
Czechia_largepart
Luxembourg_largepart
Lithuania_largepart
Slovakia_largepart
Netherlands_largepart
Finland_largepart
Germany (until 1990 former territory of the FRG)_largepart
Romania_largepart
Austria_largepart
Cyprus_largepart
Denmark_largepart
Ireland_largepart
Bulgaria_largepart
Slovenia_largepart
Belgium_largepart
Croatia_largepart
Spain_largepart
Latvia_largepart
Greece_largepart
PC1
PC2
2
0
−2
−4
−5
0
5




```
## NULL
## According to the BIC criterion, the optimal number of principal components is  27
## Belgium_sulphur                                          -0.20548849
## Bulgaria_sulphur                                         -0.17619740
## Czechia_sulphur                                          -0.19129706
## Denmark_sulphur                                          -0.19273871
## Germany (until 1990 former territory of the FRG)_sulphur -0.18443143
## Estonia_sulphur                                          -0.19491370
## Ireland_sulphur                                          -0.19250944
## Greece_sulphur                                           -0.13837500
## Spain_sulphur                                            -0.19559416
## France_sulphur                                           -0.20547702
## Croatia_sulphur                                          -0.19128927
## Italy_sulphur                                            -0.20485962
## Cyprus_sulphur                                           -0.13281147
## Latvia_sulphur                                           -0.19813959
## Lithuania_sulphur                                        -0.18903137
## Luxembourg_sulphur                                       -0.18985987
## Hungary_sulphur                                          -0.19794826
## Malta_sulphur                                            -0.06845403
## Netherlands_sulphur                                      -0.20530940
## Austria_sulphur                                          -0.20434806
## Poland_sulphur                                           -0.20496878
## Portugal_sulphur                                         -0.18532539
## Romania_sulphur                                          -0.17879710
## Slovenia_sulphur                                         -0.20162493
```

```
## Slovakia_sulphur                                          -0.20018385
## Finland_sulphur                                           -0.19312189
## Sweden_sulphur                                            -0.20441943
## United Kingdom_sulphur                                    -0.20283591
##                                                                    PC2
## Belgium_sulphur                                           -0.0049280295
## Bulgaria_sulphur                                          -0.228139110
## Czechia_sulphur                                            0.2108385297
## Denmark_sulphur                                            0.1483398058
## Germany (until 1990 former territory of the FRG)_sulphur  0.2394686704
## Estonia_sulphur                                            0.1126171495
## Ireland_sulphur                                           -0.1456112594
## Greece_sulphur                                            -0.4050950618
## Spain_sulphur                                             -0.1610401147
## France_sulphur                                             0.0171261179
## Croatia_sulphur                                           -0.0370739269
## Italy_sulphur                                              0.0254222122
## Cyprus_sulphur                                            -0.4351474861
## Latvia_sulphur                                             0.1526505567
## Lithuania_sulphur                                          0.1763573985
## Luxembourg_sulphur                                         0.1721398760
## Hungary_sulphur                                            0.0044383948
## Malta_sulphur                                             -0.4658046487
## Netherlands_sulphur                                        0.0521329679
## Austria_sulphur                                            0.0431956530
## Poland_sulphur                                            -0.0005621382
## Portugal_sulphur                                          -0.2189733797
## Romania_sulphur                                           -0.1733716692
## Slovenia_sulphur                                           0.0387593527
## Slovakia_sulphur                                           0.0277410090
## Finland_sulphur                                            0.0904022402
## Sweden_sulphur                                             0.0759933914
## United Kingdom_sulphur                                     0.0915862551
```


Germany (until 1990 former territory of the FRG)_sulphur
Czechia_sulphur
Estonia_sulphur
Croatia_sulphur
Ireland_sulphur
Greece_sulphur
Cyprus_sulphur
Malta_sulphur
PC1
PC2
3
2
1
0
−1
−2
−10
−5
0
5




```
## NULL
## According to the BIC criterion, the optimal number of principal components is  27
```