# PRINCIPAL COMPONENT ANALYSIS

```
> dpca <- read.table("C:/R/rmmva/foodpricedata.txt", header=T)
> attach(dpca)
> # dpca                      #without # to print and check the data
> # plot(dpca)                #without # to have a look at the pairwise scatterplots
> # zdpca <- scale(dpca[,2:6], center = TRUE, scale = TRUE) #if you want to standardize the data
>
> summary(dpca)
```

```
         city        bread           burger           milk          oranges         tomatoes
 ATLANTA   : 1   Min.   :20.30   Min.   : 77.70   Min.   :51.50   Min.   : 74.6   Min.   :35.40
 BALTIMORE : 1   1st Qu.:23.70   1st Qu.: 86.90   1st Qu.:57.65   1st Qu.: 95.4   1st Qu.:42.80
 BOSTON    : 1   Median :25.30   Median : 91.00   Median :62.50   Median :105.9   Median :46.80
 BUFFALO   : 1   Mean   :25.29   Mean   : 91.86   Mean   :62.30   Mean   :103.0   Mean   :48.77
 CHICAGO   : 1   3rd Qu.:26.50   3rd Qu.: 94.15   3rd Qu.:66.00   3rd Qu.:111.3   3rd Qu.:52.85
 CINCINNATI: 1   Max.   :30.80   Max.   :110.70   Max.   :80.20   Max.   :133.2   Max.   :62.60
 (Other)   :17
```

```
> s=cor(dpca[,2:6])          #compute the correlation matrix
> eigen(s)                   #compute eigenvalues and eigenvectors
```

```
$values
[1] 2.4224680 1.1046749 0.7384805 0.4936113 0.2407653

$vectors
           [,1]        [,2]        [,3]        [,4]         [,5]
[1,] -0.4961487  0.30861972  0.38639398 -0.50930459  0.499898868
[2,] -0.5757023  0.04380176  0.26247227  0.02813712 -0.772635014
[3,] -0.3395696  0.43080905 -0.83463952 -0.04910000 -0.007882237
[4,] -0.2249898 -0.79677694 -0.29160659 -0.47901574  0.005966796
[5,] -0.5064340 -0.28702846  0.01226602  0.71270629  0.391201387
```

```
> library(stats)
> pcaout<- princomp(dpca[,-1], scores = TRUE, cor= TRUE)
> pcaout  #the eigenvalues are not given here but their square roots!
```

```
Call:
princomp(x = dpca[, -1], cor = TRUE, scores = TRUE)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
1.5564279 1.0510352 0.8593489 0.7025748 0.4906784

 5  variables and  23 observations.
```

```
> names(pcaout)              #find what kind of information carried by "princomp" function
```

```
[1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"   "call"
```

```
> pcaout$scores[1:10,1:2]
```

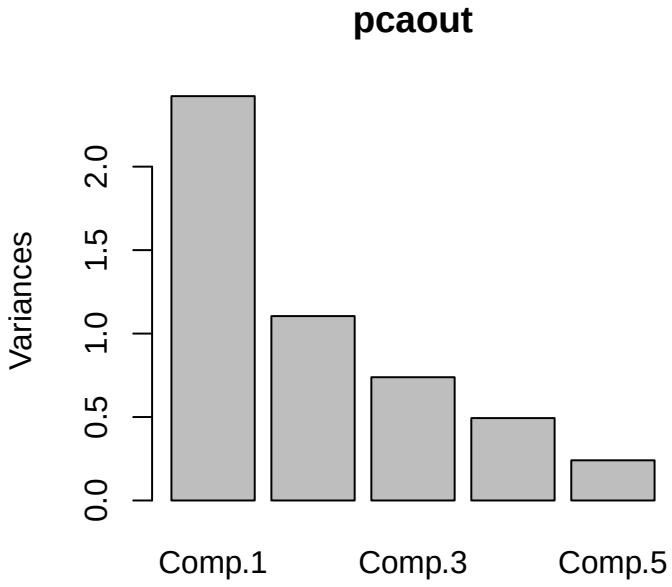
```
          Comp.1      Comp.2
 [1,]  0.2323147  2.23781857
 [2,] -0.2880227  1.92623541
```

```
 [3,] -2.2984921  0.07524344
 [4,]  0.3488520 -1.12992744
 [5,] -0.1163224 -0.08802682
 [6,] -0.6059976  0.46122164
 [7,]  1.2427141 -1.33550534
 [8,]  1.1215615 -0.85950113
 [9,]  0.2807925 -1.34737468
[10,] -4.1688508 -0.50508332
```

```
> #different scores in SAS and R:
> #in R the population variance is equal to the eigenvalue
> #in SAS the sample variance is equal to the eigenvalue, cfr:
> sassc <- pcaout$scores*sqrt(22/23)
> sassc[1:10,1:2]
```

```
          Comp.1      Comp.2
 [1,]  0.2272083  2.18862973
 [2,] -0.2816918  1.88389539
 [3,] -2.2479696  0.07358954
 [4,]  0.3411840 -1.10509084
 [5,] -0.1137656 -0.08609193
 [6,] -0.5926774  0.45108366
 [7,]  1.2153983 -1.30614999
 [8,]  1.0969088 -0.84060868
 [9,]  0.2746205 -1.31775844
[10,] -4.0772166 -0.49398124
```

```
> screeplot(pcaout)
```



```
> xlabsn <- as.character(dpca[,1])                    #to get the cities printed in the biplot
> biplot(pcaout, choices=1:2,xlabs=xlabsn ,cex=.5)    #ask for biplot with the first 2 components
```