# Reducing Dimensionality Home work

*Blessing*

*2/28/2020*

**1. One of the most famous data sets in statistics is Fisher's iris data. The data set (available in file iris.csv) contains measurements of 50 specimens from each of three different species of iris — Iris setosa, Iris versicolor, and Iris virginica — on the following dimensions (measurements are in millimeters):**

- X1 species (1 = Iris setosa, 2 = Iris versicolor, 3 = Iris virginica)
- X2 sepal length
- X3 sepal width
- X4 petal length
- X5 petal width

*(a) Analyze the iris data (variables X2–X5) using principal components analysis. How many components do you need to adequately describe the data? How would you interpret them? (b) Plot the average principal component scores for each of the three different types of iris for the first two principal components. Describe your findings.*

## Installing and loading the neccessary Packages

```
#install.packages('psych')
library(psych)
```

```
## Warning: package 'psych' was built under R version 3.6.2
```

```
library( tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.1

## -- Attaching packages --------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.1     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'readr' was built under R version 3.6.1

## Warning: package 'dplyr' was built under R version 3.6.1

## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

## Reading the data

For the sake of clarity, I replaced the denotation of the variables in the iris csv to their real meaining ( X1 to species, e.t.c)

```
iris <- read.csv('iris.csv')
```

## Exploring the data and a brief summary statistics on it.

```
head(iris)
```

```
##   species sepal_length sepal_width petal_length petal_width
## 1       1          5.1         3.5          1.4         0.2
## 2       1          4.9         3.0          1.4         0.2
## 3       1          4.7         3.2          1.3         0.2
## 4       1          4.6         3.1          1.5         0.2
## 5       1          5.0         3.6          1.4         0.2
## 6       1          5.4         3.9          1.7         0.4
```

```
describe(iris)
```

```
##              vars   n mean   sd median trimmed  mad min max range  skew
## species         1 150 2.00 0.82   2.00    2.00 1.48 1.0 3.0   2.0  0.00
## sepal_length    2 150 5.84 0.83   5.80    5.81 1.04 4.3 7.9   3.6  0.31
## sepal_width     3 150 3.06 0.44   3.00    3.04 0.44 2.0 4.4   2.4  0.31
## petal_length    4 150 3.76 1.77   4.35    3.76 1.85 1.0 6.9   5.9 -0.27
## petal_width     5 150 1.20 0.76   1.30    1.18 1.04 0.1 2.5   2.4 -0.10
##              kurtosis   se
## species         -1.52 0.07
## sepal_length    -0.61 0.07
## sepal_width      0.14 0.04
## petal_length    -1.42 0.14
## petal_width     -1.36 0.06
```

```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ species     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ sepal_length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ sepal_width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ petal_length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ petal_width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
```

## Analyzing the correlation matrix

Now we dive into the data set by first finding the correlation among our variables of interest.

```
vars <- scale(iris[,-1]) # the specie column is removed because it adds no value to the analysis in thi
cor <- cor(vars)
```

We can see the correlation matrix as a table with only the lower part shown (since it is symmetric), using the following code:

```
upper<-round(cor,3) # we round the results to the 3d digit after comma
upper[upper.tri(cor)]<-""
upper<-as.data.frame(upper)
upper
```

```
##              sepal_length sepal_width petal_length petal_width
## sepal_length            1
## sepal_width        -0.118           1
## petal_length        0.872      -0.428            1
## petal_width         0.818      -0.366        0.963           1
```

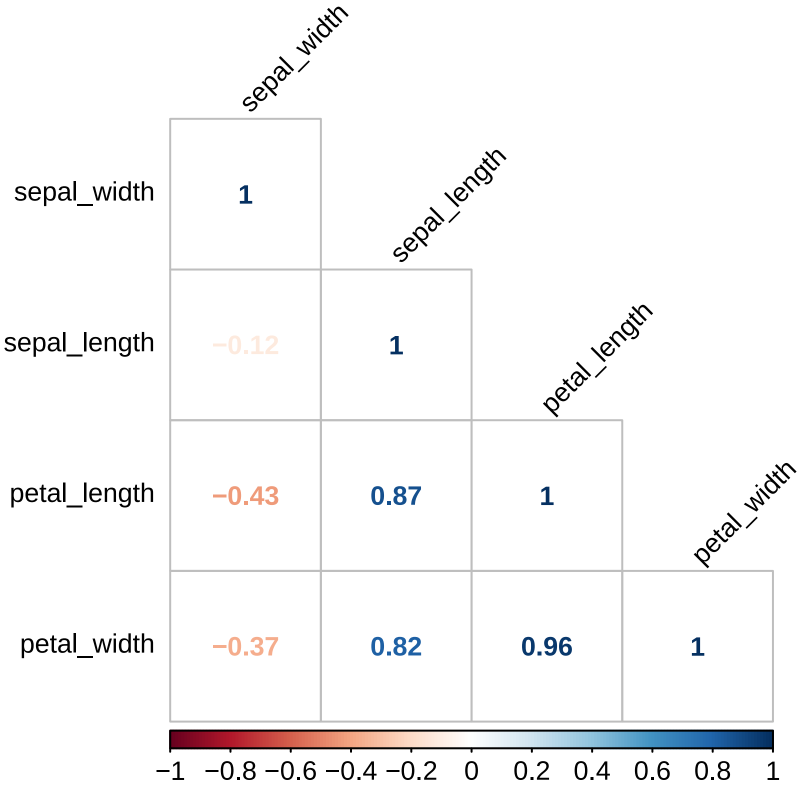
The correlation matrix is not so difficult to understand but additional insights can be obtained by visualizing its heatmap (a correlogram) generated with the following code:

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.6.2
```

```
## corrplot 0.84 loaded
```

```
corrplot(cor,
         method = "number",
         type = "lower",
         order = "hclust", # reorder by the size of the correlation coefficients
         tl.cex = 0.8, # font size of the variable labels
         tl.col = "black", # color of the variable labels
         tl.srt = 45, # rotation angle for the variable labels
         number.cex = 0.8 # font size of the coefficients
)
```



We note the relatively high correlations between sepal length and petal width, sepal length and petal length, and petal length and petal width. Some correlation coefficients are as high as 0.96, but some are in the range of -0.1 to -0.4.

## Finding the eigenvalues

The next step is to find the initial eigenvalues and their individual and cumulative percentages of variance explained. This will help us decide how many dimensions should be retained as relevant. We can compute the eigenvalues with the following code:

```
EV = eigen(cor)$values
EV
```

```
## [1] 2.91849782 0.91403047 0.14675688 0.02071484
```

Note that the last eigenvalue is effectively zero, due to the fact that the sum of the original variables (expressed as shares) is equal to 1.Thus, despite the fact that there are 4 metric of the flower measurements, there are only 3 independent dimensions (factors, or components) The individual percentage of variance explained by a dimension is found by dividing the eigenvalue of the dimension by the total number of variables (in our case, 4). The total number of variables corresponds to the total variance in the data because the variables are standardized so that each has a variance of 1.

```
# length() - to find the total number of elements in a vector
# sum() - to find a sum of the vector's elements
# EV/length(EV) is equivalent to EV/sum(EV)

EV/length(EV)
```

```
## [1] 0.729624454 0.228507618 0.036689219 0.005178709
```

The first eigenvalue (2.92) suggests that the first dimension (component) accounts for almost 72.96 percent of the variance in the original data, as shown in the percentages of variance. The second eigenvalue (0.94) suggests that the second dimension accounts for almost 22.9 percent of the variance in the data, and so on. The cumulative percentages of variance explained are:

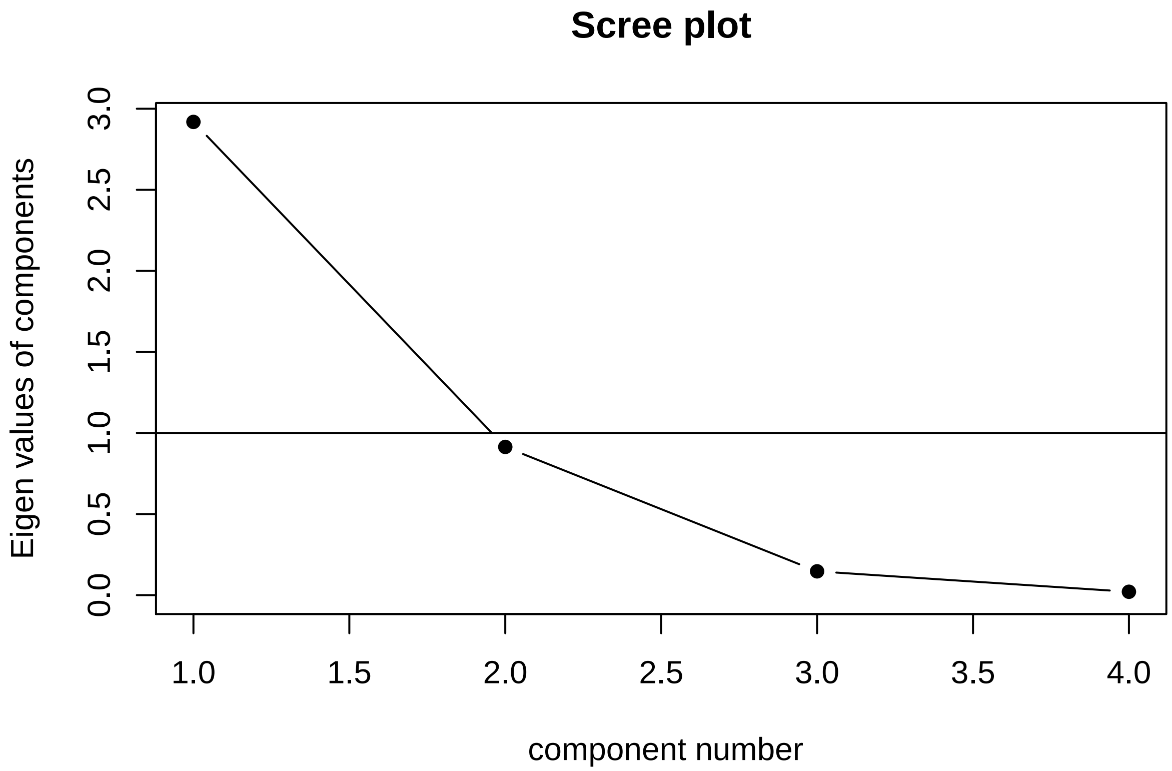
```
cumsum(EV/length(EV))
```

```
## [1] 0.7296245 0.9581321 0.9948213 1.0000000
```

The first two dimensions together account for 95.8(72.9+ 22.8) percent of the variance in the data, the first three dimensions together account for 99.5 percent,and so on.

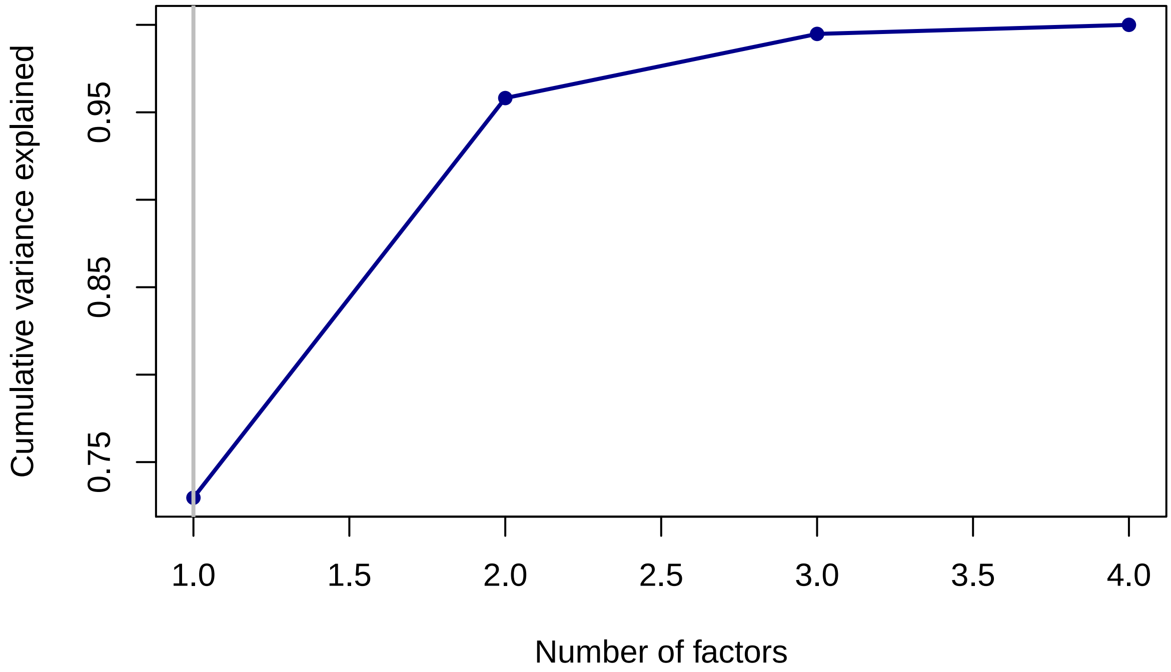
**How many components do you need to adequately describe the data? How would you interpret them?**

```
scree(cor, pc = TRUE, factors = FALSE)
```

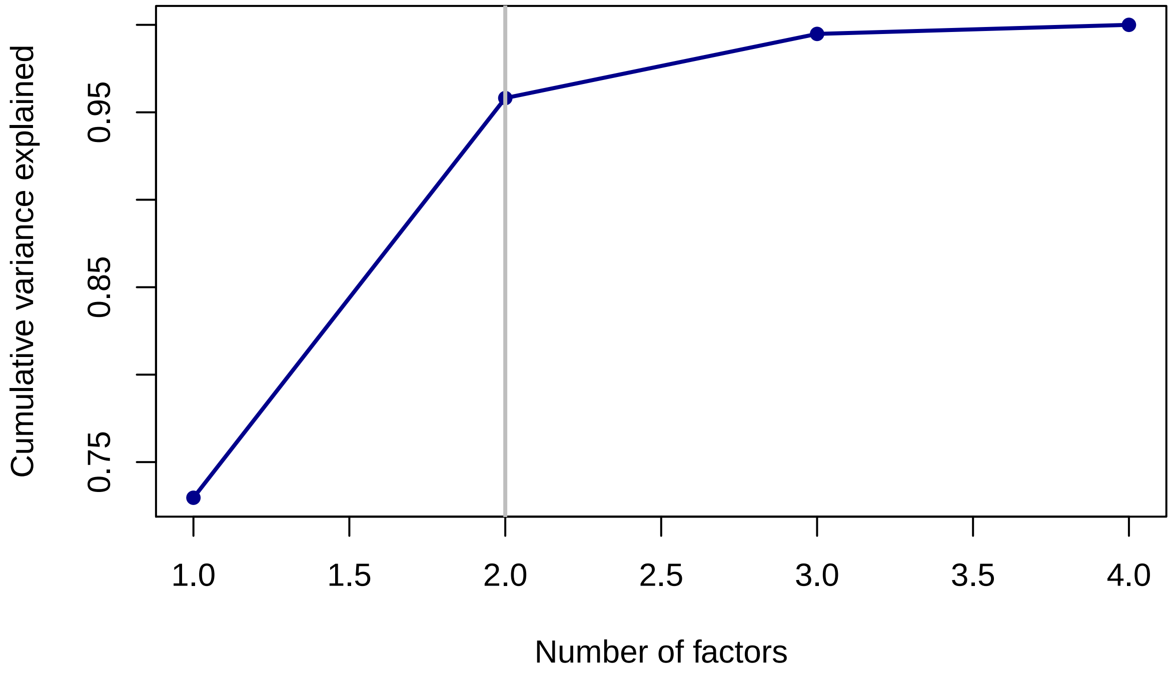


There are several criteria for deciding on the number of dimensions to retain. The first criterion retains only those eigenvalues that are greater than one (this is also called Kaiser's rule), and in this case it suggests retaining the first dimension. based on percentage of variance.In this approach, the number of factors to retain is determined by the cummulative percentage of variance explained by these factors, if it reaches a satisfactory level of 60%, which in this case is achieved with only the first factor that account for 73% of the variance. To answer the first part of the number one question on the homework, the first factor PC1 is enough to describe the data since it explains more than 73% of the variance but to see the bigger picture in my opinion it is advisable to include a second factor PC2.For PC1 the variable petal length has a loading of 0.99 and less than 0.35 loading score for PC2 and petal width has loading of 0.96 and less than 0.35 loading score fo PC2, a good intepretation/description for the component PC1 would be Petal features. For PC2 the variable sepal width loading of 0.89 and the variable sepal length has a loading of 0.346 which is greater than 0.35, a good intepretation/decription of PC2 would be sepal features or just Sepal width(since it loads pretty high for the width and relatively low for the length)

```r
# Shares for the cumulative variance explained
plot(cumsum(EV/length(EV)),
     type = "o", # type of plot: "o" for points and lines 'overplotted'
     col = "darkblue",
     pch = 16, # plot symbol: 16 = filled circle
     cex = 1, # size of plot symbols
     xlab = "Number of factors", # a title for the x axis
     ylab = "Cumulative variance explained", # a title for the y axis
     lwd = 2) # line width
abline(v = 1, lwd = 2, col = "grey") # draw a vertical line at v = 1
```



```
plot(cumsum(EV/length(EV)),
     type = "o", # type of plot: "o" for points and lines 'overplotted'
     col = "darkblue",
     pch = 16, # plot symbol: 16 = filled circle
     cex = 1, # size of plot symbols
     xlab = "Number of factors", # a title for the x axis
     ylab = "Cumulative variance explained", # a title for the y axis
     lwd = 2) # line width
abline(v = 2, lwd = 2, col = "grey") # draw a vertical line at v = 2
```



Executing The PCA so we can explain this Factor.

For the sake of convenience We shall extract the first 2 factors (or components/dimensions) from the correlation matrix using principal component analysis (PCA). The following code computes the loadings, communalities, specificities, and several other measures, retains three components and saves the component scores. Note that there is no rotation of the components involved.

```
PCA.iris <- principal(r = cor,
                      nfactors = 2,
                      rotate="none",
                      scores = TRUE)
```

In the output generated with the following code, it is useful to require sorting the component loadings by size, so that the structure of the solution is easier to see. To simplify the loadings matrix, we can also use the option cut = 0.35 to display only those component loadings that are larger than 0.35. While this unclutters the loadings matrix and helps interpretation, the 0.35 value is rather arbitrary and you can decide to change it to something else (how high is a “high” loading?).

```
print(PCA.iris,
      digits = 2, # to round numbers to the third digit
      cut = 0.35, # to show only values > 0.35
      sort = TRUE # to sort rows by loading size
)
```

```
## Principal Components Analysis
## Call: principal(r = cor, nfactors = 2, rotate = "none", scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##               item   PC1   PC2   h2     u2 com
## petal_length    3  0.99       0.98 0.0163 1.0
## petal_width     4  0.96       0.94 0.0647 1.0
## sepal_length    1  0.89  0.36 0.92 0.0774 1.3
## sepal_width     2 -0.46  0.88 0.99 0.0091 1.5
##
##                        PC1  PC2
## SS loadings           2.92 0.91
## Proportion Var        0.73 0.23
## Cumulative Var        0.73 0.96
## 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  0.03
##
## Fit based upon off diagonal values = 1
```

The proportion of variance in each of the original variables accounted for by the first two principal components are the communalities and can be displayed with the following code (note that this represents the information in the h2 column of the previous output):

```
PCA.iris$communality
```

```
## sepal_length  sepal_width petal_length  petal_width
##    0.9225986    0.9909193    0.9837300    0.9352804
```

For additional insights into how the different metric of the iris flower relate to each other, it is also helpful to plot the principal component loadings. The following code generates a scatter plot of the loadings for the first two principal components (also called a “factor loading plot”):

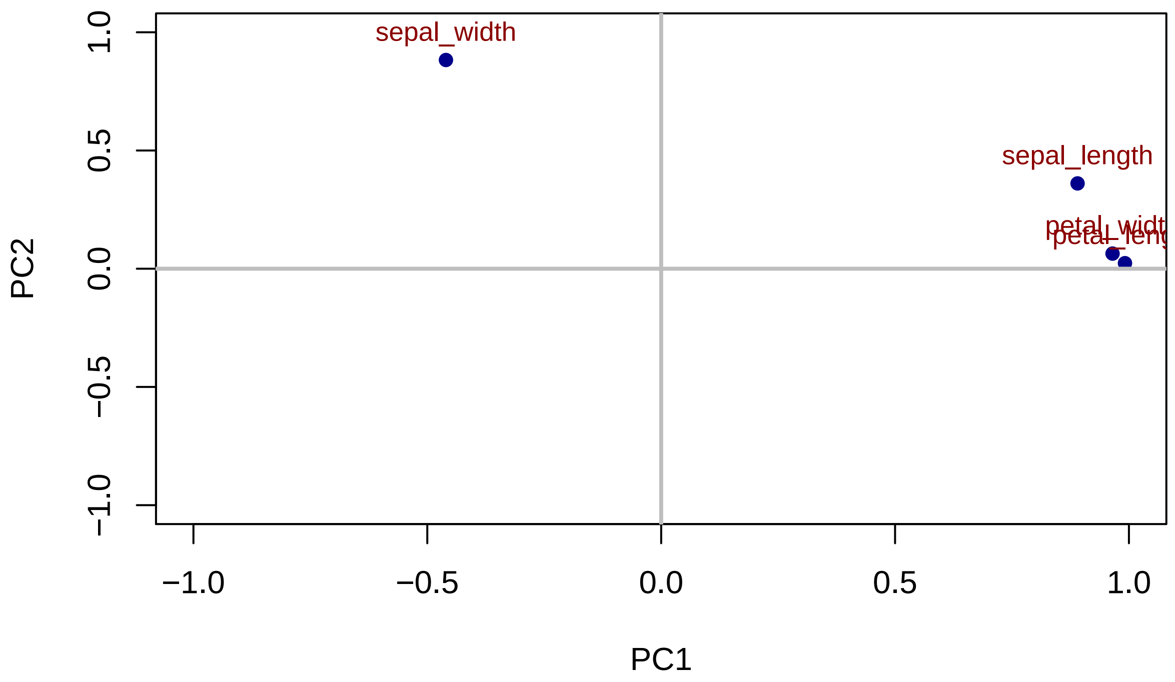
```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
L <- as.data.table(unclass(PCA.iris$loadings), keep.rownames = T)

plot(x = L$PC1, y = L$PC2,
     col ="darkblue",
     pch = 16,         # plot symbol: 16 = filled circle
     cex = 1,          # size of plot symbols
     xlab = "PC1",     # a title for the x axis
     ylab = "PC2",     # a title for the y axis
     xlim = c(-1,1),   # x axis values from -1 to 1
     ylim = c(-1,1))   # y axis values from -1 to 1

# add point labels
text(L$PC1, L$PC2,
     labels = L$rn,
     pos = 3,
     cex = 0.8,
     col = "darkred")

# add vertical and horizontal lines
abline(h = 0, lwd = 2, col = "grey") # draw a horizontal line at h = 0
abline(v = 0, lwd = 2, col = "grey") # draw a vertical line at v = 0
```



## the perceptual Map

The component score matrix for each observation (in our case, each of the 150 flowers) can be computed with the following code. As discussed earlier, the scores are the projections of the original data units into the new (2-dimensional) component space, instead of the original 2-dimensional variable space.

```
PCA.iris.scores = factor.scores(vars, unclass(PCA.iris$loadings))$scores
head(PCA.iris.scores,4)
```

```
##             PC1        PC2
## [1,] -1.321232  0.5004175
## [2,] -1.214037 -0.7027698
## [3,] -1.379296 -0.3564318
## [4,] -1.341465 -0.6227710
```

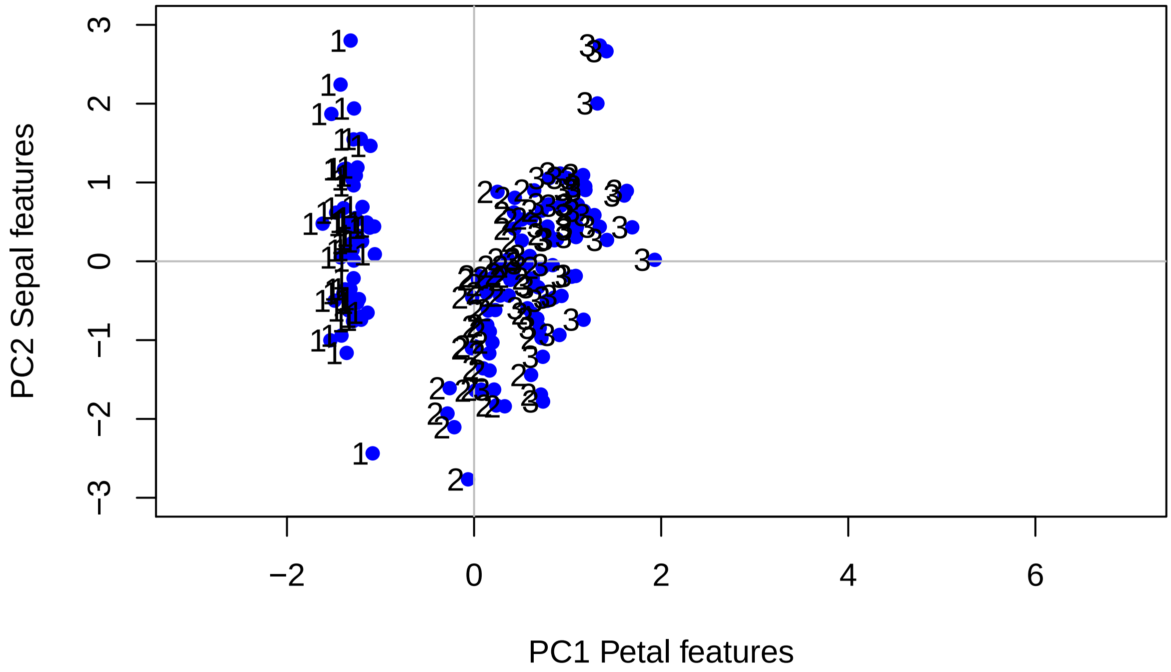
The component (factor) scores are standardized values with an average of 0 and a standard deviation of 1. We can add the component scores to the original data set in two new columns, so that for each observation (each flower) in the data we also have recorded its scores on the new two dimensional space:

```
iris.scores <- cbind(iris, PCA.iris.scores)
```

Plotting the principal component scores on a (perceptual) map can give us an idea of the location of the various iris flowers in the principal component space. The plot axes are the factor score values for a pair of the extracted components. Since we can look only at two components at once, in order to get a full understanding, when we extract more than two factors we need to consider plotting all pairs of factors (in our case this leads to two plots). Below we have a perceptual map for PC1 and PC2;

**(b) Plot the average principal component scores for each of the three different types of iris for the first two principal components. Describe your findings.**

```
plot(x = iris.scores$PC1,
     y = iris.scores$PC2,
     xlab = "PC1 Petal features", ylab = "PC2 Sepal features",
     xlim = c(-3, 7), ylim = c(-3, 3),
     pch = 16, cex = 1, col = "blue")

abline(h = 0, col = "grey")
abline(v = 0, col = "grey")

# add point labels
text(x = iris.scores$PC1,
     y = iris.scores$PC2,
     labels = iris.scores$species,
     cex = 1,
     adj = 1.2,
     col = "black")
```

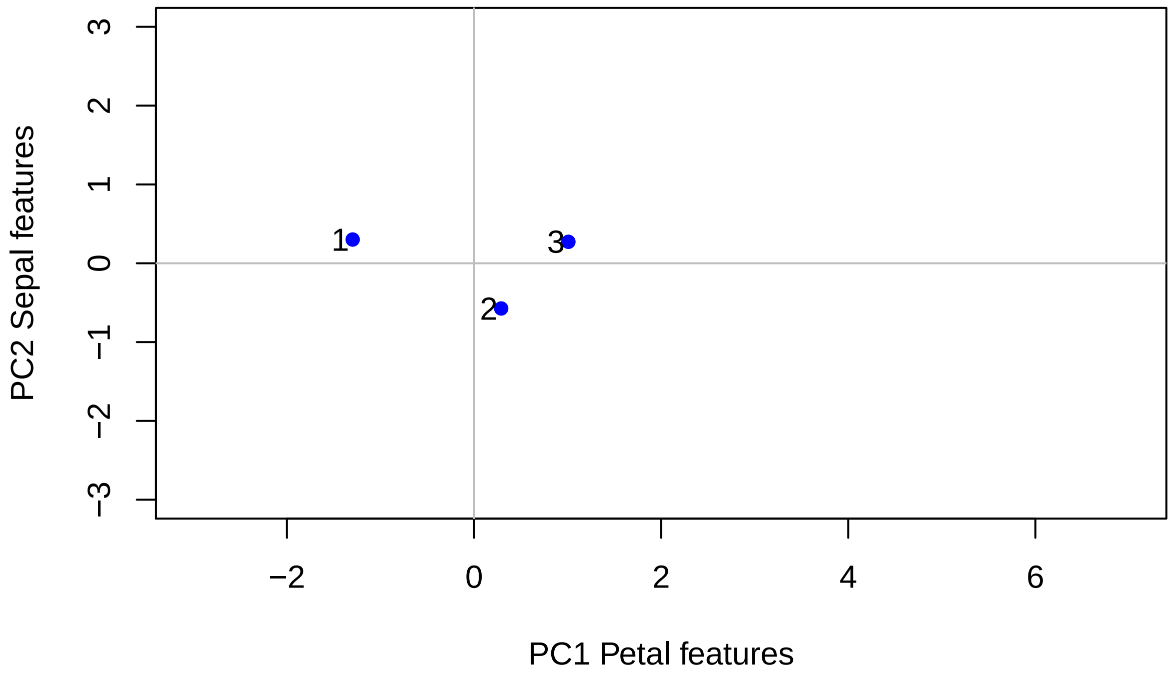


aggregating the plot data points.

```
# Average vales for factor scores for each iris specie
Average <- group_by(iris.scores, species) %>%
           summarise(PC1_mean= mean(PC1),PC2_mean= mean(PC2))
Average
```

```
## # A tibble: 3 x 3
##   species PC1_mean PC2_mean
##     <int>    <dbl>    <dbl>
## 1       1   -1.30     0.301
## 2       2    0.290   -0.574
## 3       3    1.01     0.272
```

```
Average <- as.data.frame(Average)
plot(x = Average$PC1,
     y = Average$PC2,
     xlab = "PC1 Petal features", ylab = "PC2 Sepal features",
     xlim = c(-3, 7), ylim = c(-3, 3),
     pch = 16, cex = 1, col = "blue")

abline(h = 0, col = "grey")
abline(v = 0, col = "grey")

# add point labels
text(x = Average$PC1,
     y = Average$PC2,
     labels = as.factor(Average$species),
     cex = 1,
     adj = 1.2,
     col = "black")
```



The first thing apparent from the plot is that specie1 with value close to -1.29 on the first dimension (principal component) is highly distinguished from the other two species( 2 and 3). This difference can be traced to low average value on the two petal dimensions of the iris flower: length and width in comparison to other species. We can also see that specie 2 and 3 are close together, this stems from the fact that they have almost the same average sepal features(mean sepal length (5.9 and 6.5) and width(2.7 and 2.9. which is the most important variable for PC2)) and their average petal features (length and width) are also almost the same and these values are significantly different from that of sepecie 1.

Here is a summary stats on the iris data to confirm these findings

```
summary <- group_by(iris, species) %>%
           summarise(sp_length_mean = mean(sepal_length),sp_width_mean = mean(sepal_width),p_length_mea


summary
```

```
## # A tibble: 3 x 13
##   species sp_length_mean sp_width_mean p_length_mean p_width_mean
##     <int>          <dbl>         <dbl>         <dbl>        <dbl>
## 1       1           5.01          3.43          1.46        0.246
## 2       2           5.94          2.77          4.26        1.33
## 3       3           6.59          2.97          5.55        2.03
## # ... with 8 more variables: sp_length_max <dbl>, sp_length_min <dbl>,
## #   sp_width_max <dbl>, sp_width_min <dbl>, p_length_max <dbl>,
## #   p_length_min <dbl>, p_width_max <dbl>, p_width_min <dbl>
```

**2. Golding and Seidman (1974) studied the vocational interests of 231 undergraduate males. Each respondent rated the strength of his interests in 22 vocational areas, listed below:**

*The correlation matrix is available in the file vocations.csv. Analyze the data using principal components. Does there appear to be more than one dimension describing vocational interests among undergraduate males? How would you describe the underlying dimension(s)? Which vocational interests seem to go together? Which seem most different?*

For clarity and convenient purpose I have modified the names of the variables to actually denote the interest they stand for.

# Reading in the correlation matrix

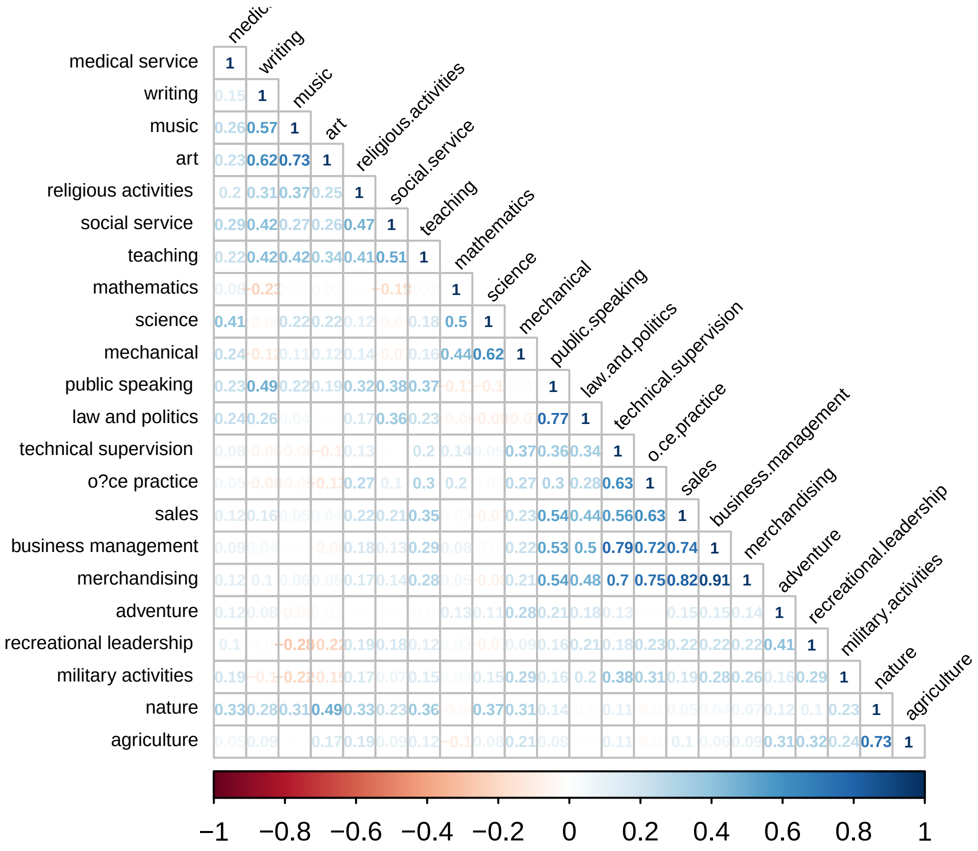
```
cor.voc <- read.csv('vocations.csv', row.names = 1)
```

I modified the data earlier adding the upper triangular part to it using excel but We can see the correlation matrix as a table with only the lower part shown (since it is symmetric), using the following code:

```
upper.voc<-round(cor.voc,3) # we round the results to the 3d digit after comma
upper.voc[upper.tri(cor.voc)]<-""
upper.voc<-as.data.frame(upper.voc)
upper.voc
```

```
##                          public.speaking law.and.politics
## public speaking                     1.00
## law and politics                    0.77                1
## business management                 0.53              0.5
## sales                               0.54             0.44
## merchandising                       0.54             0.48
## o?ce practice                       0.30             0.28
## military activities                 0.16              0.2
## technical supervision               0.36             0.34
## mathematics                        -0.11            -0.05
## science                            -0.10            -0.09
## mechanical                         -0.02            -0.07
## nature                              0.14            -0.02
## agriculture                         0.09            -0.01
## adventure                           0.21             0.18
## recreational leadership             0.16             0.21
## medical service                     0.23             0.24
## social service                      0.38             0.36
## religious activities                0.32             0.17
## teaching                            0.37             0.23
## music                               0.22             0.04
## art                                 0.19            -0.01
## writing                             0.49             0.26
##                          business.management sales merchandising
## public speaking
## law and politics
## business management                         1
## sales                                    0.74     1
## merchandising                           0.91  0.82             1
## o?ce practice                           0.72  0.63          0.75
## military activities                     0.28  0.19          0.26
## technical supervision                   0.79  0.56           0.7
## mathematics                             0.08  0.02          0.05
## science                                -0.03 -0.07         -0.08
## mechanical                              0.22  0.23          0.21
## nature                                  0.04  0.05          0.07
## agriculture                             0.06   0.1          0.09
## adventure                               0.15  0.15          0.14
## recreational leadership                 0.22  0.22          0.22
## medical service                         0.09  0.12          0.12
## social service                          0.13  0.21          0.14
## religious activities                    0.18  0.22          0.17
## teaching                                0.29  0.35          0.28
## music                                  -0.01  0.05          0.06
## art                                    -0.06  0.04          0.05
## writing                                 0.04  0.16           0.1
##                          o.ce.practice military.activities
## public speaking
## law and politics
## business management
## sales
## merchandising
## o?ce practice                        1
## military activities               0.31                   1
## technical supervision             0.63                0.38
## mathematics                        0.2                0.03
## science                           0.02                0.15
## mechanical                        0.27                0.29
## nature                           -0.03                0.23
## agriculture                      -0.03                0.24
## adventure                        -0.01                0.16
## recreational leadership           0.23                0.29
## medical service                   0.05                0.19
## social service                     0.1                0.07
## religious activities              0.27                0.17
## teaching                           0.3                0.15
## music                            -0.05               -0.22
## art                              -0.13               -0.15
## writing                          -0.08                -0.1
##                          technical.supervision mathematics science
## public speaking
## law and politics
## business management
## sales
## merchandising
## o?ce practice
## military activities
## technical supervision                        1
## mathematics                               0.14           1
## science                                   0.05         0.5       1
## mechanical                                0.37        0.44    0.62
## nature                                    0.11       -0.04    0.37
## agriculture                               0.11        -0.1    0.08
## adventure                                 0.13        0.13    0.11
## recreational leadership                   0.18        0.03   -0.07
## medical service                           0.08        0.08    0.41
## social service                               0       -0.19   -0.04
## religious activities                      0.13       -0.01    0.12
## teaching                                   0.2       -0.03    0.18
## music                                    -0.06        0.01    0.22
## art                                       -0.1        0.02    0.22
## writing                                  -0.06       -0.23   -0.04
##                          mechanical nature agriculture adventure
## public speaking
## law and politics
## business management
## sales
## merchandising
## o?ce practice
## military activities
## technical supervision
## mathematics
## science
## mechanical                        1
## nature                         0.31      1
## agriculture                    0.21   0.73           1
## adventure                      0.28   0.12        0.31         1
## recreational leadership        0.09    0.1        0.32      0.41
## medical service                0.24   0.33        0.05      0.12
## social service                -0.07   0.23        0.09     -0.01
## religious activities           0.14   0.33        0.19         0
## teaching                       0.16   0.36        0.12     -0.02
## music                          0.11   0.31           0     -0.05
## art                            0.12   0.49        0.17      0.02
## writing                       -0.12   0.28        0.09      0.08
##                          recreational.leadership medical.service
## public speaking
## law and politics
## business management
## sales
## merchandising
## o?ce practice
## military activities
## technical supervision
## mathematics
## science
## mechanical
## nature
## agriculture
## adventure
## recreational leadership                         1
## medical service                               0.1               1
## social service                               0.18            0.29
## religious activities                         0.19             0.2
## teaching                                     0.12            0.22
## music                                       -0.28            0.26
## art                                          -0.22            0.23
## writing                                      -0.02            0.15
##                          social.service religious.activities teaching
## public speaking
## law and politics
## business management
## sales
## merchandising
## o?ce practice
## military activities
## technical supervision
## mathematics
## science
## mechanical
## nature
## agriculture
## adventure
## recreational leadership
## medical service
## social service                        1
## religious activities               0.47                    1
## teaching                           0.51                 0.41        1
## music                              0.27                 0.37     0.42
## art                                0.26                 0.25     0.34
## writing                            0.42                 0.31     0.42
##                          music  art writing
## public speaking
## law and politics
## business management
## sales
## merchandising
## o?ce practice
## military activities
## technical supervision
## mathematics
## science
## mechanical
## nature
## agriculture
## adventure
## recreational leadership
## medical service
## social service
## religious activities
## teaching
## music                        1
## art                       0.73    1
## writing                   0.57 0.62       1
```

```
corrplot(data.matrix(cor.voc),
         method = "number",
         type = "lower",
         order = "hclust", # reorder by the size of the correlation coefficients
         tl.cex = 0.6, # font size of the variable labels
         tl.col = "black", # color of the variable labels
         tl.srt = 45, # rotation angle for the variable labels
         number.cex = 0.5 # font size of the coefficients
)
```

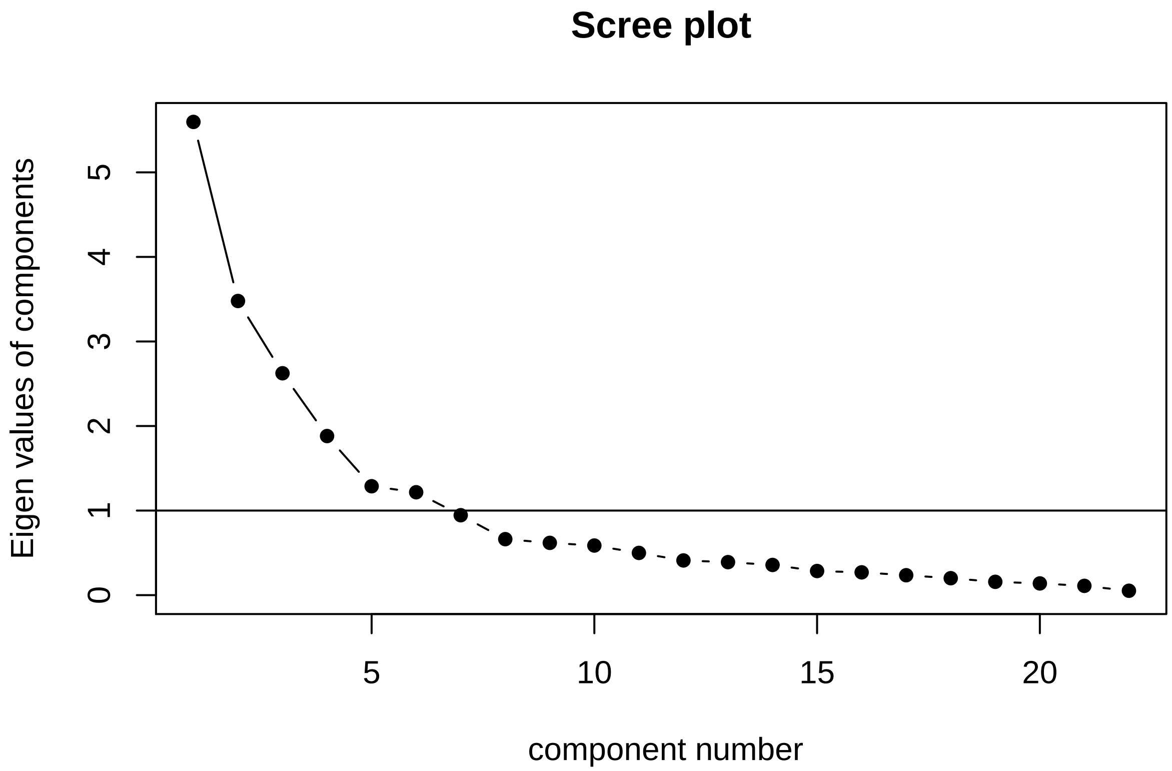


**Does there appear to be more than one dimension describing vocational interests among undergraduate males?** the answer is YES

## Eigenvalues and scree plot

Next, we examine a scree plot of the eigenvalues to determine the appropriate number of factors.

```
scree(cor.voc, pc = TRUE, factors = FALSE)
```



The scree plot suggests that we may be justified in extracting the first six factors(dimensions); note that there appears to be an "elbow" in the amount of variance accounted for after the fifth eigenvalue so we are also justified if we were to take just the first 5 factors. Using a proportionality criterion (i.e., each common factor should account for at least as much variation as one of the original variables in the analysis, which is directly analogous to the rationale underlying the rule for retaining factors with eigenvalues larger than one), the inclusion of the sixth factor is marginal at best.

The individual percentage of variance explained by a variable is found by dividing the eigenvalue of the variable by the total number of variables (in our case, 22).

```
EV.voc = eigen(data.matrix(cor.voc))$values
EV.voc/length(EV.voc)
```

```
##  [1] 0.254380785 0.158115061 0.119269913 0.085515833 0.058543166
##  [6] 0.055307859 0.042950685 0.030084982 0.028103042 0.026656421
## [11] 0.022689152 0.018645669 0.017734345 0.016223818 0.012959025
## [16] 0.012249085 0.010696009 0.009098696 0.007151948 0.006288344
## [21] 0.004992442 0.002343720
```

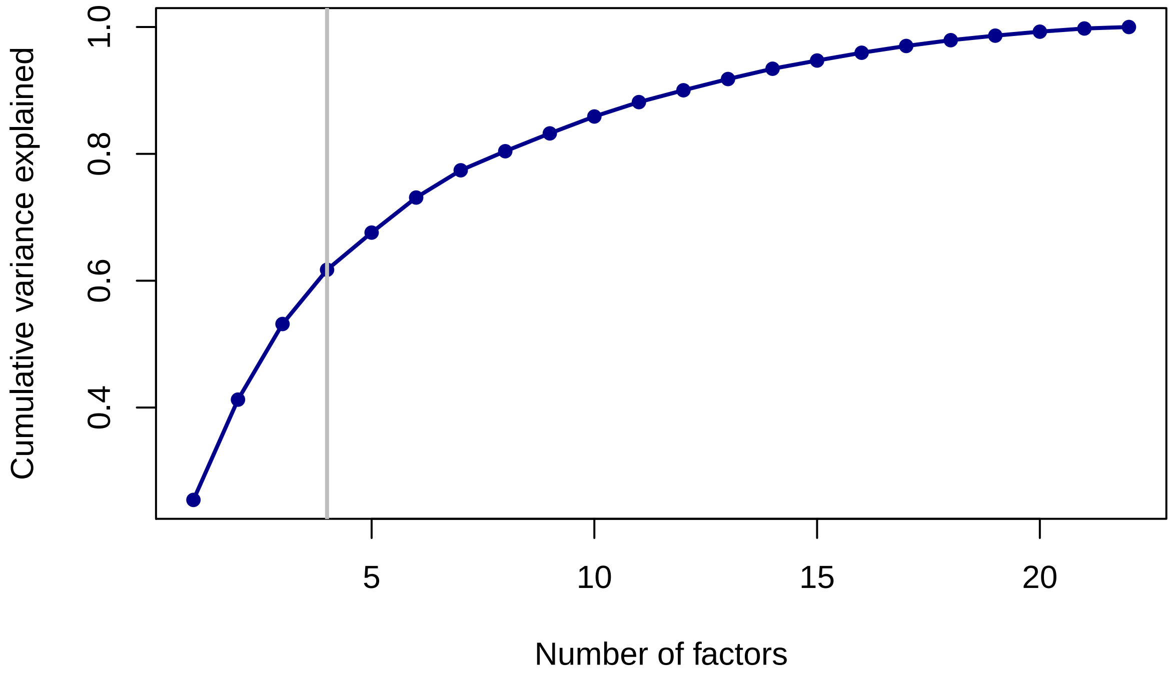
We can also compute and plot the cumulative percentages of variance explained. This is done with the following command:

```
cumsum(EV.voc/length(EV.voc))
```

```
##  [1] 0.2543808 0.4124958 0.5317658 0.6172816 0.6758248 0.7311326 0.7740833
##  [8] 0.8041683 0.8322713 0.8589277 0.8816169 0.9002626 0.9179969 0.9342207
## [15] 0.9471798 0.9594288 0.9701248 0.9792235 0.9863755 0.9926638 0.9976563
## [22] 1.0000000
```

Now we can as well retain only the first 4 factors, because the cumulative variance explained by these four factors (61.72%) is greater than the satisfactory level of 60%

```
# Shares for the cumulative variance explained
plot(cumsum(EV.voc/length(EV.voc)),
     type = "o", # type of plot: "o" for points and lines 'overplotted'
     col = "darkblue",
     pch = 16, # plot symbol: 16 = filled circle
     cex = 1, # size of plot symbols
     xlab = "Number of factors", # a title for the x axi
     ylab = "Cumulative variance explained", # a title for the y axis
     lwd = 2) # line width
abline(v = 4, lwd = 2, col = "grey") # draw a vertical line at v = 4
```



# Execute the PCA

```
PCA.Voc <- principal(r = cor.voc,
                 nfactors = 4,
                 rotate="none",
                 scores = T)
```

```
print(PCA.Voc,
      digits = 4, # to round numbers to the third digit
      cut = 0.35, # to show only values > 0.35
      sort = TRUE # to sort rows by loading size
)
```

```
## Principal Components Analysis
## Call: principal(r = cor.voc, nfactors = 4, rotate = "none", scores = T)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                          item    PC1     PC2     PC3     PC4     h2     u2
## merchandising               5 0.8249 -0.3567                 0.8484 0.1516
## business management         3 0.8104 -0.4179                 0.8632 0.1368
## sales                       4 0.7783                         0.7113 0.2887
## public speaking             1 0.7233         -0.3892         0.6865 0.3135
## technical supervision       8 0.6879 -0.4265                 0.6926 0.3074
## o?ce practice               6 0.6753 -0.4351                 0.7100 0.2900
## law and politics            2 0.6149         -0.3660         0.5303 0.4697
## teaching                   19 0.5779  0.4004                 0.5056 0.4944
## religious activities       18 0.4804  0.3731                 0.3753 0.6247
## social service             17 0.4346  0.4231                 0.4958 0.5042
## military activities         7 0.4031                         0.4200 0.5800
## medical service            16 0.3511                         0.2952 0.7048
## art                        21         0.7760                 0.7062 0.2938
## music                      20         0.7357         -0.3945 0.7681 0.2319
## writing                    22 0.3504  0.6499 -0.3821         0.6912 0.3088
## nature                     12 0.3615  0.5603  0.4089         0.7215 0.2785
## mechanical                 11 0.3578          0.7584         0.7363 0.2637
## science                    10                 0.7488         0.7414 0.2586
## mathematics                 9                 0.5846 -0.4111 0.5331 0.4669
## agriculture                13                         0.6482 0.6611 0.3389
## recreational leadership    15                         0.6145 0.5587 0.4413
## adventure                  14                         0.4232 0.3284 0.6716
##                            com
## merchandising           1.499
## business management     1.611
## sales                   1.358
## public speaking         1.588
## technical supervision   1.861
## o?ce practice           2.035
## law and politics        1.746
## teaching                1.861
## religious activities    1.938
## social service          3.145
## military activities     3.478
## medical service         2.819
## art                     1.355
## music                   1.835
## writing                 2.224
## nature                  3.345
## mechanical              1.557
## science                 1.677
## mathematics             1.951
## agriculture             2.206
## recreational leadership  1.964
## adventure               2.512
##
##                          PC1    PC2    PC3    PC4
## SS loadings           5.5964 3.4785 2.6239 1.8813
## Proportion Var        0.2544 0.1581 0.1193 0.0855
## Cumulative Var        0.2544 0.4125 0.5318 0.6173
## Proportion Explained  0.4121 0.2561 0.1932 0.1385
## Cumulative Proportion 0.4121 0.6682 0.8615 1.0000
```

```
##
## Mean item complexity =  2.1
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.0725
##
## Fit based upon off diagonal values = 0.9362
```

```r
sort(PCA.Voc$communality, decreasing = T )
```

```
##     business.management           merchandising                   music
##               0.8631899               0.8484447               0.7681034
##                 science              mechanical                  nature
##               0.7414440               0.7363333               0.7215051
##                   sales           o.ce.practice                     art
##               0.7113038               0.7099934               0.7061800
##   technical.supervision                 writing         public.speaking
##               0.6925865               0.6912150               0.6864729
##             agriculture recreational.leadership             mathematics
##               0.6611104               0.5586898               0.5331349
##        law.and.politics                teaching          social.service
##               0.5302822               0.5055699               0.4957999
##     military.activities    religious.activities               adventure
##               0.4199745               0.3752616               0.3284178
##         medical.service
##               0.2951819
```

The first four principal components capture more information in some areas of vocational interests (e.g., Business management and mechandising greater than 80%) than in others (e.g., medical services and adventure lower than 30 percent). The communalities (h2) are all roughly between 0.30 and 0.86, which means that between30% and 86% of the variation in each of the original input variables has been retained by the first four principal components.In 19 out of 22 variables the amount of variance explained is nearly 50 percent or more and less than 40 percent of the original variance is captured in 3 of the 22 variables,this in my opinionis sufficient.

### How would you describe the underlying dimension(s)?

```r
L.voc <- as.data.table(unclass(PCA.Voc$loadings), keep.rownames = T)
```

We can examine the extracted dimensions to get some idea of their interpretation. Looking at the loadings,we see that the first dimension is positively correlated with merchandising(0.82), business management (0.81) sales(0.78) and public speaking (0.72) this variables loads quite high on this component hence i intepret PC1 as bussiness and commerce interest. The second dimension is positively correlated with art(0.78), music (0.74) and writting(0.64) and negatively correlated to mechandising (-0.36), bussiness management(-0.42), technical supervision and law and office pratice(-0.43) hence i intepret PCA2 as Art and music interest. The third dimension is positively correlated with mechanical, science and mathematics and negatively correlated to public speaking (-0.39), writing(-0.38) and law and politics(-0.36) hence i intepret PCA3 has Engineering and science interest.

The fourth dimension is positively correlated with agriculture(0.65) and recreational leadership( 0.61) and negatively corellated to mathematics and music, hence i intepret PCA4 as Other Type of interests.

the loading structure is rather difficult to interpret. To simplify the interpretation of the components, some type of rotation might be helpful;

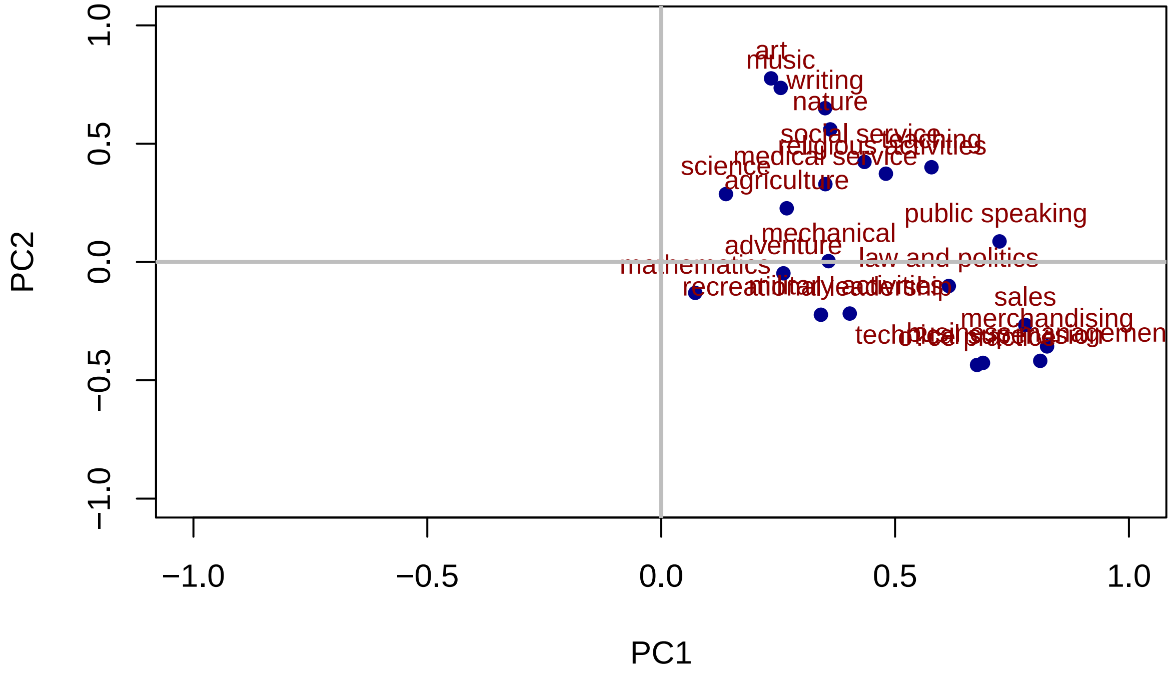
```r
RCA.Voc <- principal(r = cor.voc,
                 nfactors = 4,
                 rotate="varimax",
                 scores = T)

print(RCA.Voc,
      digits = 4, # to round numbers to the third digit
      cut = 0.35, # to show only values > 0.35
      sort = TRUE # to sort rows by loading size
)
```

```
## Principal Components Analysis
## Call: principal(r = cor.voc, nfactors = 4, rotate = "varimax", scores = T)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                          item     RC1     RC2     RC3     RC4     h2
## business management         3  0.9219                         0.8632
## merchandising               5  0.9113                         0.8484
## sales                       4  0.8228                         0.7113
## o?ce practice               6  0.8189                         0.7100
## technical supervision       8  0.7804                         0.6926
## public speaking             1  0.6058  0.4666                 0.6865
## law and politics            2  0.6001                         0.5303
## music                      20          0.8007                 0.7681
## art                        21          0.7989                 0.7062
## writing                    22          0.7848                 0.6912
## teaching                   19          0.6338                 0.5056
## social service             17          0.6033                 0.4958
## nature                     12          0.5666          0.5535 0.7215
## religious activities       18          0.5444                 0.3753
## medical service            16          0.4125                 0.2952
## science                    10                  0.8198         0.7414
## mechanical                 11                  0.7889         0.7363
## mathematics                 9                  0.7030         0.5331
## agriculture                13                          0.7755 0.6611
## recreational leadership    15                          0.6873 0.5587
## adventure                  14                          0.5598 0.3284
## military activities         7                          0.5407 0.4200
##                              u2   com
## business management      0.1368 1.031
## merchandising            0.1516 1.044
## sales                    0.2887 1.102
## o?ce practice            0.2900 1.117
## technical supervision    0.3074 1.281
## public speaking          0.3135 2.482
## law and politics         0.4697 1.973
## music                    0.2319 1.403
## art                      0.2938 1.219
## writing                  0.3088 1.243
## teaching                 0.4944 1.500
## social service           0.5042 1.750
## nature                   0.2785 2.561
## religious activities     0.6247 1.549
## medical service          0.7048 2.356
## science                  0.2586 1.210
## mechanical               0.2637 1.377
## mathematics              0.4669 1.161
## agriculture              0.3389 1.200
## recreational leadership  0.4413 1.378
## adventure                0.6716 1.097
## military activities      0.5800 1.852
##
##                           RC1    RC2    RC3    RC4
## SS loadings           4.8002 3.9097 2.5039 2.3664
## Proportion Var        0.2182 0.1777 0.1138 0.1076
## Cumulative Var        0.2182 0.3959 0.5097 0.6173
## Proportion Explained  0.3535 0.2879 0.1844 0.1743
## Cumulative Proportion 0.3535 0.6414 0.8257 1.0000
##
## Mean item complexity =  1.5
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.0725
##
## Fit based upon off diagonal values = 0.9362
```

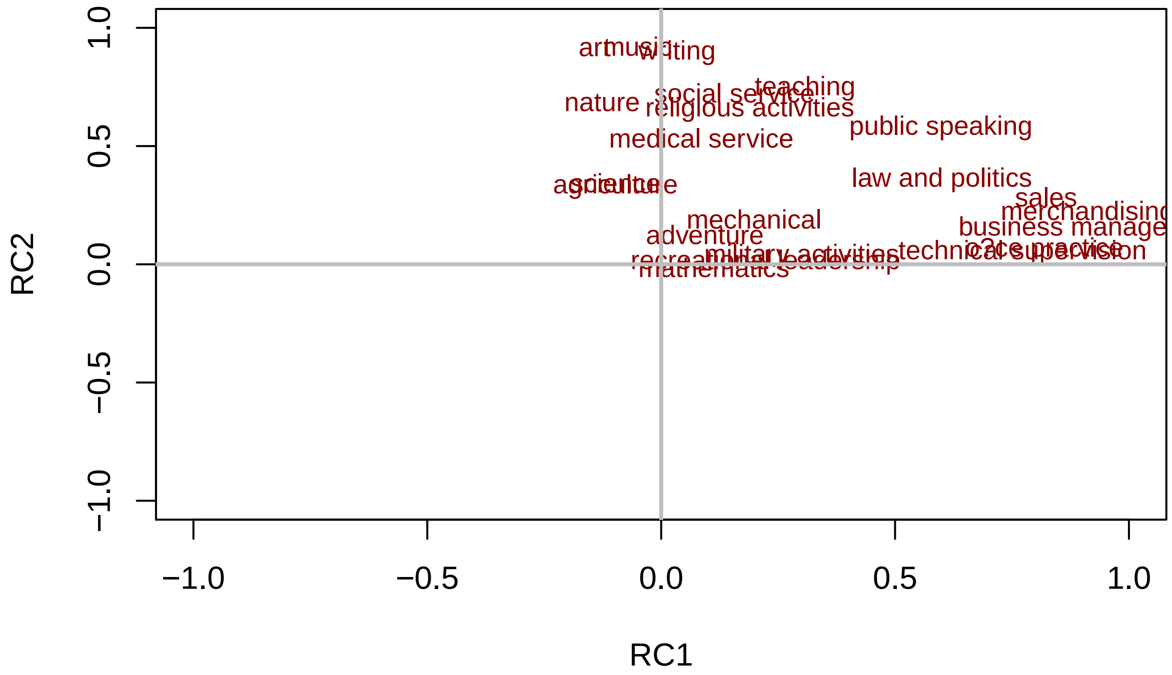
Now the loading structure is clearer and we can see our intepretation come to live in their glaring form, the intepretation for PC1-PC3 quite align but we might change the intepretation of PC4 from other type of interest to interest in Agriculture.

For additional insights into how the different measures of economic activity relate to each other, it is also helpful to plot the principal component loadings. The following code generates a scatter plot of the loadings for the first two principal components (also called a “factor loading plot”):

```r
plot(x = L.voc$PC1, y = L.voc$PC2,
     col ="darkblue",
     pch = 16,        # plot symbol: 16 = filled circle
     cex = 1,         # size of plot symbols
     xlab = "PC1",    # a title for the x axis
     ylab = "PC2",    # a title for the y axis
     xlim = c(-1,1),  # x axis values from -1 to 1
     ylim = c(-1,1))  # y axis values from -1 to 1

# add point labels
text(L.voc$PC1, L.voc$PC2,
     labels = L.voc$rn,
     pos = 3,
     cex = 0.8,
     col = "darkred")

# add vertical and horizontal lines
abline(h = 0, lwd = 2, col = "grey") # draw a horizontal line at h = 0
abline(v = 0, lwd = 2, col = "grey") # draw a vertical line at v = 0
```



```r
LR.voc <- as.data.table(unclass(RCA.Voc$loadings), keep.rownames = T)
plot(x = LR.voc$PC1, y = LR.voc$PC2,
     col ="darkred",
     pch = 16,        # plot symbol: 16 = filled circle
     cex = 1,         # size of plot symbols
     xlab = "RC1",    # a title for the x axis
     ylab = "RC2",
    xlim = c(-1,1),  # x axis values from -1 to 1
     ylim = c(-1,1)) # a title for the y axis
     # y axis values from -1 to 1

# add point labels
text(LR.voc$RC1, LR.voc$RC2,
     labels = LR.voc$rn,
     pos = 3,
     cex = 0.8,
     col = "darkred")

# add vertical and horizontal lines
abline(h = 0, lwd = 2, col = "grey") # draw a horizontal line at h = 0
abline(v = 0, lwd = 2, col = "grey") # draw a vertical line at v = 0
```



**Which vocational interests seem to go together?Which seem most different?**

from the graph of rotated RC1 against rotated RC2 components we can see 3 clear clusters all almost aligning with our intepretation: cluster1 (sales, mechandising,business management, office pratice and technical supervision) cluster2(teaching, medical services, art,music,wrting, nature, social services, religious activities) cluster3( agriculture,science, military activities, recreational leadership, mathematics, adventure and mechanical)

and two seperate variables standing alone, public speaking, and law and politics. This answers Question 2 of the homework.

**3. Sixty students rated 10 brands of soft drinks (Coke, Diet Pepsi, Dr. Pepper, Mt. Dew, Pepsi, Royal Crown, 7Up, Sprite, Diet 7Up, Tab) on four attributes (calories, sweetness, thirst–quenching, and popularity with others) at two different times during the semester (September and November). The variables in the data set are defined as follows:**

*The correlation matrix is available in the file soft drinks.csv. Analyze the data using factor analysis.*

in order for clarity and convenient purpose I have modified the names of the variables to actualy denote the interest they stand for.

## Reading in the correlation matrix

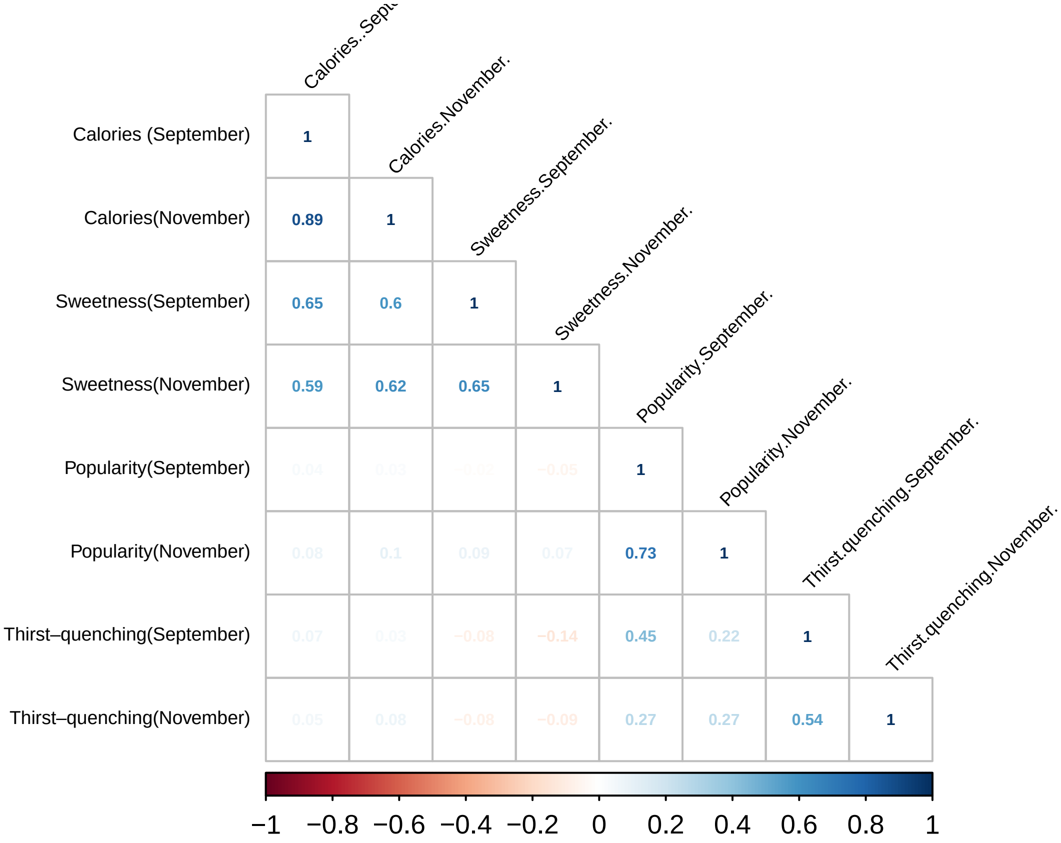
```
cor.soft <- read.csv('soft_drinks.csv', row.names = 1)
```

I modified the data earlier adding the upper triangular part to it using excel but We can see the correlation matrix as a table with only the lower part shown (since it is symmetric), using the following code:

```
upper.soft<-round(cor.soft,3) # we round the results to the 3d digit after comma
upper.soft[upper.tri(cor.soft)]<-""
upper.soft<-as.data.frame(upper.soft)
upper.soft
```

```
##                             Calories..September. Calories.November.
## Calories (September)                       1.000
## Calories(November)                         0.886                  1
## Sweetness(September)                       0.649              0.597
## Sweetness(November)                        0.588              0.621
## Thirst-quenching(September)                0.067              0.034
## Thirst-quenching(November)                 0.054              0.076
## Popularity(September)                      0.037              0.029
## Popularity(November)                       0.075              0.102
##                             Sweetness.September. Sweetness.November.
## Calories (September)
## Calories(November)
## Sweetness(September)                           1
## Sweetness(November)                        0.649                   1
## Thirst-quenching(September)                -0.08              -0.136
## Thirst-quenching(November)                -0.075              -0.092
## Popularity(September)                     -0.018              -0.054
## Popularity(November)                       0.089               0.069
##                             Thirst.quenching.September.
## Calories (September)
## Calories(November)
## Sweetness(September)
## Sweetness(November)
## Thirst-quenching(September)                           1
## Thirst-quenching(November)                        0.542
## Popularity(September)                             0.446
## Popularity(November)                              0.225
##                             Thirst.quenching.November.
## Calories (September)
## Calories(November)
## Sweetness(September)
## Sweetness(November)
## Thirst-quenching(September)
## Thirst-quenching(November)                          1
## Popularity(September)                           0.274
## Popularity(November)                            0.267
##                             Popularity.September. Popularity.November.
## Calories (September)
## Calories(November)
## Sweetness(September)
## Sweetness(November)
## Thirst-quenching(September)
## Thirst-quenching(November)
## Popularity(September)                           1
## Popularity(November)                         0.73                    1
```

```
corrplot(data.matrix(cor.soft),
         method = "number",
         type = "lower",
         order = "hclust", # reorder by the size of the correlation coefficients
         tl.cex = 0.6, # font size of the variable labels
         tl.col = "black", # color of the variable labels
         tl.srt = 45, # rotation angle for the variable labels
         number.cex = 0.5 # font size of the coefficients
)
```

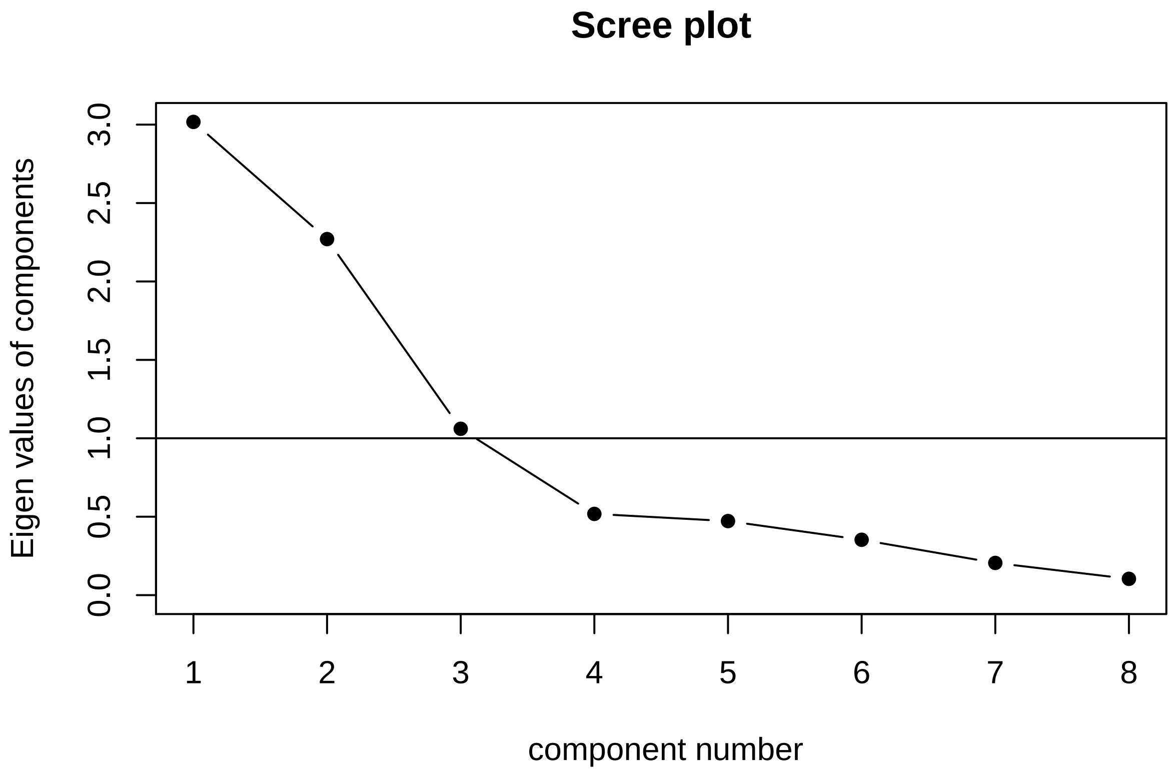


**How many factors are there?**

## Eigenvalues and scree plot

Next, we examine a scree plot of the eigenvalues to determine the appropriate number of factors.

```
scree(cor.soft, pc = TRUE, factors = FALSE)
```



The scree plot suggest to retain the first 3 factors.

```
EV.soft = eigen(data.matrix(cor.soft))$values
EV.soft/length(EV.soft)
```

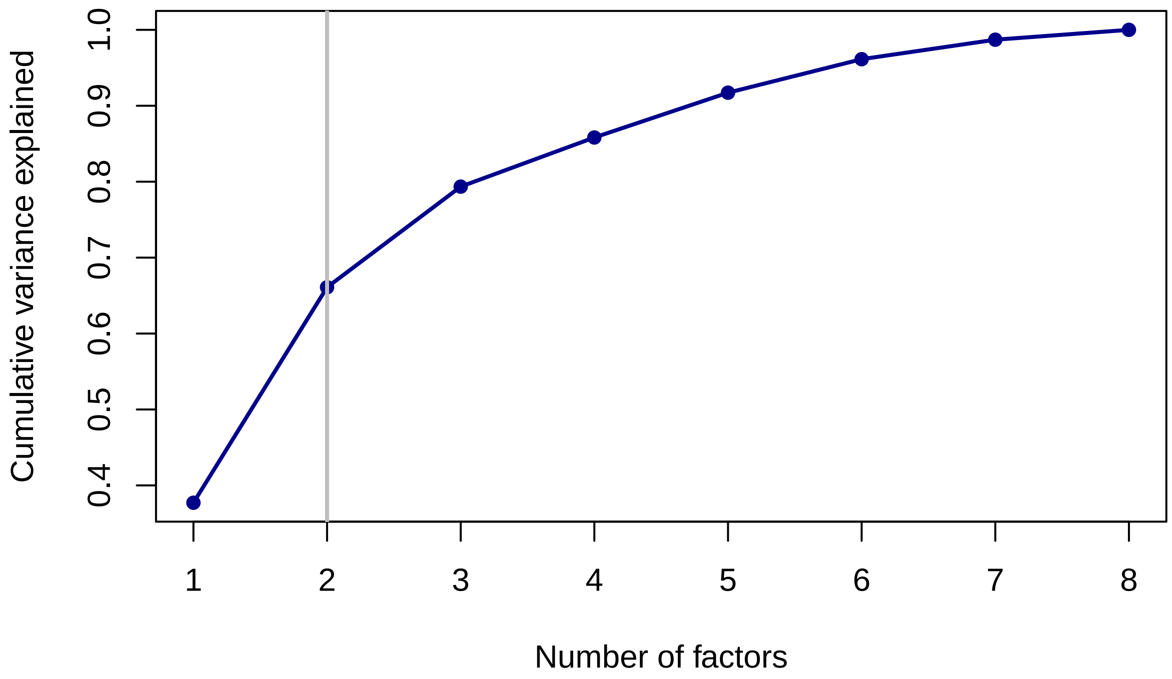
```
## [1] 0.37715038 0.28376785 0.13256927 0.06474078 0.05901568 0.04412334
## [7] 0.02565517 0.01297753
```

```
cumsum(EV.soft/length(EV.soft))
```

```
## [1] 0.3771504 0.6609182 0.7934875 0.8582283 0.9172440 0.9613673 0.9870225
## [8] 1.0000000
```

Now we can as well retain only the first 2 factors, because the cumulative variance explained by these two factors (66.09%) is greater than the satisfactory level of 60%. we want the model to be as simple as possible.

```
# Shares for the cumulative variance explained
plot(cumsum(EV.soft/length(EV.soft)),
     type = "o", # type of plot: "o" for points and lines 'overplotted'
     col = "darkblue",
     pch = 16, # plot symbol: 16 = filled circle
     cex = 1, # size of plot symbols
     xlab = "Number of factors", # a title for the x axis
     ylab = "Cumulative variance explained", # a title for the y axis
     lwd = 2) # line width
abline(v = 2, lwd = 2, col = "grey") # draw a vertical line at v = 2
```



## Executing the EFA

We shall now extract the first two factors (or components/dimensions) from the correlation matrix using exploratory factor analysis (EFA). We will first analyze the unrotated solution, then use factor rotations to see if interpretability improves.

Unrotated factor solution The following code computes the loadings, communalities, specificities, and several other measures, retains four factors and saves the factor scores.so far there is no rotation of the factors involved.

```
EFA1 <- fa(r = cor.soft,
           nfactors = 2,
           fm = "pa",
           rotate = "none")
```

```
print(EFA1,
      digits = 3, # to round numbers to the third digit
      cut = 0.35, # to show only values > 0.35
      sort = TRUE # to sort rows by loading size
)
```

```
## Factor Analysis using method =  pa
## Call: fa(r = cor.soft, nfactors = 2, rotate = "none", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                               item   PA1    PA2    h2    u2  com
## Calories..September.             1 0.902        0.814 0.186 1.00
## Calories.November.               2 0.893        0.797 0.203 1.00
## Sweetness.September.           3 0.743        0.562 0.438 1.03
## Sweetness.November.            4 0.725        0.546 0.454 1.08
## Popularity.September.          7        0.855 0.734 0.266 1.01
## Popularity.November.           8        0.664 0.462 0.538 1.09
## Thirst.quenching.September.    5        0.572 0.327 0.673 1.00
## Thirst.quenching.November.     6        0.486 0.237 0.763 1.00
##
##                          PA1   PA2
## SS loadings           2.714 1.766
## Proportion Var        0.339 0.221
## Cumulative Var        0.339 0.560
## Proportion Explained  0.606 0.394
## Cumulative Proportion 0.606 1.000
##
## Mean item complexity =  1
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  28  and the objective function was  4.315
## The degrees of freedom for the model are 13  and the objective function was  0.844
##
## The root mean square of the residuals (RMSR) is  0.084
## The df corrected root mean square of the residuals is  0.123
##
## Fit based upon off diagonal values = 0.951
## Measures of factor score adequacy
##                                                      PA1   PA2
## Correlation of (regression) scores with factors   0.954 0.905
## Multiple R square of scores with factors          0.911 0.819
## Minimum correlation of possible factor scores     0.821 0.638
```

### How would you interpret them?

The atributes calories( september and november) and sweetness( september and november) they both load high on the first factor(component) hence i shall proceed to intepretе/decribe PA1 as nutritional benefits. the attributes popularity( september and november) loads quite high for PA2( popularity september 0.86). hence a good decription for PA2 would be popularity amongst students.

### Which of the four attributes has the highest reliability? How do you tell?

```r
sort(EFA1$communality)
```

```
##  Thirst-quenching(November) Thirst-quenching(September)
##                   0.2370589                   0.3272534
##        Popularity(November)         Sweetness(November)
##                   0.4618882                   0.5460686
##        Sweetness(September)       Popularity(September)
##                   0.5618312                   0.7340943
##          Calories(November)        Calories (September)
##                   0.7974572                   0.8138723
```

calories september as the highest communality(0.81) which means the two factors were able to explain above 80% of the variance which the atribute calories september explains, on this rationale we can say calories(september) has the highest reliability.

```r
L.soft <- unclass(EFA1$loadings)
L.soft <- as.data.table(unclass(EFA1$loadings), keep.rownames = T)
```

```r
EFA1$loadings
```

```
##
## Loadings:
##                             PA1    PA2
## Calories..September.         0.902
## Calories.November.           0.902
## Sweetness.September.         0.743
## Sweetness.November.          0.725 -0.142
## Thirst.quenching.September.         0.572
## Thirst.quenching.November.          0.486
## Popularity.September.               0.855
## Popularity.November.         0.144  0.664
##
##                   PA1   PA2
## SS loadings    2.714 1.766
## Proportion Var 0.339 0.221
## Cumulative Var 0.339 0.560
```

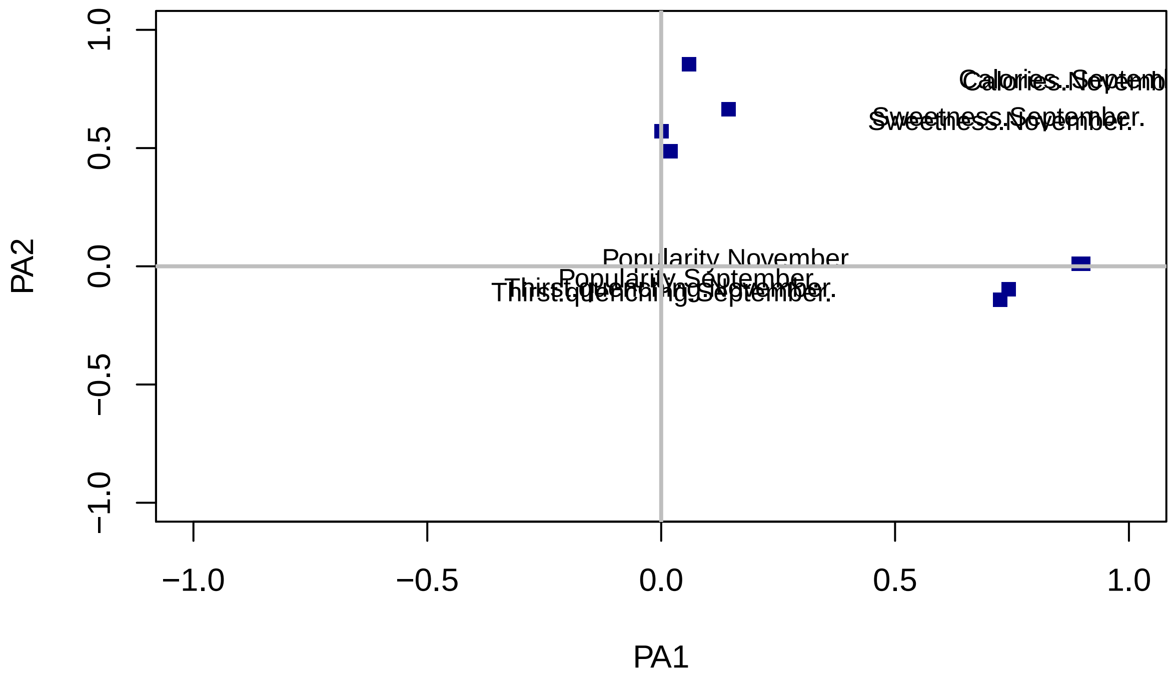
## Rotated factor solution

```r
EFA2 <- fa(r = cor.soft,
           nfactors = 2,
           fm = "pa",
           rotate = "varimax")

print(EFA2,
      digits = 2, # to round numbers to the third digit
      cut = 0.35, # to show only values > 0.35
      sort = TRUE # to sort rows by loading size
)
```

```
## Factor Analysis using method =  pa
## Call: fa(r = cor.soft, nfactors = 2, rotate = "varimax", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                             item   PA1   PA2   h2   u2 com
## Calories..September.           1  0.90       0.81 0.19   1
## Calories.November.             2  0.89       0.80 0.20   1
## Sweetness.September.           3  0.75       0.56 0.44   1
## Sweetness.November.            4  0.73       0.55 0.45   1
## Popularity.September.          7        0.86 0.73 0.27   1
## Popularity.November.           8        0.67 0.46 0.54   1
## Thirst.quenching.September.    5        0.57 0.33 0.67   1
## Thirst.quenching.November.     6        0.49 0.24 0.76   1
##
##                         PA1  PA2
## SS loadings           2.71 1.77
## Proportion Var        0.34 0.22
## Cumulative Var        0.34 0.56
## Proportion Explained  0.60 0.40
## Cumulative Proportion 0.60 1.00
##
## Mean item complexity =  1
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  28  and the objective function was  4.32
## The degrees of freedom for the model are 13  and the objective function was  0.84
##
## The root mean square of the residuals (RMSR) is  0.08
## The df corrected root mean square of the residuals is  0.12
##
## Fit based upon off diagonal values = 0.95
## Measures of factor score adequacy
##                                                    PA1  PA2
## Correlation of (regression) scores with factors   0.95 0.91
## Multiple R square of scores with factors          0.91 0.82
## Minimum correlation of possible factor scores     0.82 0.64
```

There appears to be little or no significant difference in the loadings after rotation is performed.

```
plot(x = L.soft$PA1, y = L.soft$PA2,
     col ="darkblue",
     pch = 15,        # plot symbol: 16 = filled circle
     cex = 1,         # size of plot symbols
     xlab = "PA1",    # a title for the x axis
     ylab = "PA2",    # a title for the y axis
     xlim = c(-1,1),  # x axis values from -1 to 1
     ylim = c(-1,1))  # y axis values from -1 to 1

# add point labels

text(L.soft$PA1 , L.soft$PA1,
     labels = L.soft$rn,
     pos = 1,
     cex = 0.8,
     col = "black")
# add vertical and horizontal lines
abline(h = 0, lwd = 2, col = "grey") # draw a horizontal line at h = 0
abline(v = 0, lwd = 2, col = "grey") # draw a vertical line at v = 0
```



this confirms our intepretation as we can see two distincts clusters in the above graph.