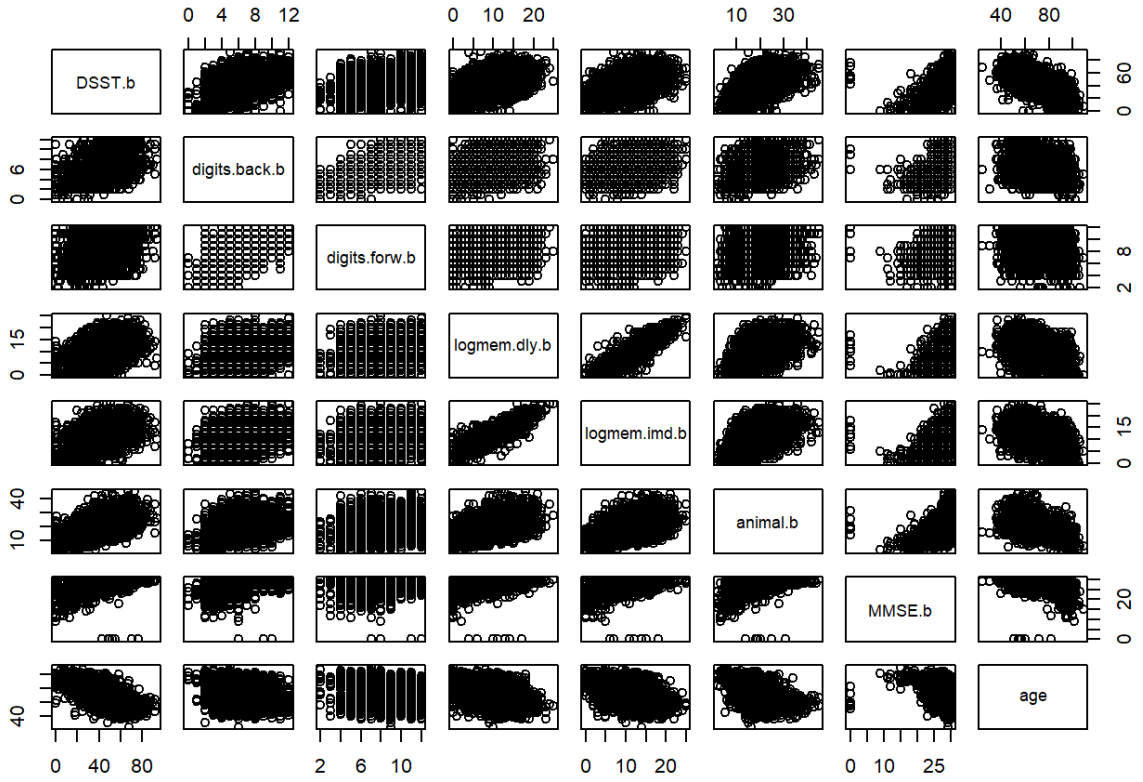
A data set of measurements of cognitive functions of 4010 individuals aged 30 and older was analyzed using principal component analysis. The cognitive functions were measured through these 7 tests:

- **DSST**= digit symbol substitution test, to measure attention (scale 0-60)
- **Digits back** and **digits forward** = two tests of memory (sale 0 -14)
- Logical memory immediate (**logmem.imd**) and delayd (**logmem.dly**) = two tests of logical memory for immediate recollection of facts or delayed recollection (scale 0-25)
- **Animal** fluency = a test of semantic fluency (scale 0-40)
- Mini-mental state exam (**MMSE**): a test geared to see whether a person has dementia (scale 0.25)

The following pairwise scatter plots were generated in Rstudio. What are the tests that appear to be most correlated? Can you conclude that as people age their cognitive functions tend to decrease? [10 pts. Be complete]



Logical memory immediate and logical memory delayed are the most correlated tests. From the scatterplots alone, I cannot conclude that age is correlated to any of the cognitive function tests because there do not appear to be any trend in the points.

The following output was produced in RStudio:

```
pca.1 <- prcomp(cog.data[,2:9] )
 pca.1
```

```
## Standard deviations:
## [1] 21.333876  8.485916  5.888948  4.689534  2.537513  2.220114  1.521568
## [8]  1.409889
##
## Rotation:
##                       PC1          PC2         PC3          PC4
## DSST.b         0.69177471 -0.697950422  0.16425920  0.005186485
## digits.back.b  0.04035604 -0.054112631 -0.04757837  0.030014717
## digits.forw.b  0.02908120 -0.048901107 -0.01328163  0.046871040
## logmem.dly.b   0.15136275  0.036827604 -0.53452460  0.456160206
## logmem.imd.b   0.13318107  0.026224863 -0.50822294  0.430884355
## animal.b       0.19806273  0.052599800 -0.59362693 -0.775504257
## MMSE.b         0.07113664 -0.002477848 -0.11401339  0.030084390
## age           -0.65881429 -0.709033722 -0.24734535 -0.028634567
##                       PC5          PC6          PC7          PC8
## DSST.b         0.083573293 -0.01617668  0.006597841  0.003906131
## digits.back.b -0.670946254 -0.14040189 -0.235652062 -0.683275325
## digits.forw.b -0.700567558 -0.15697753  0.162472371  0.672722492
## logmem.dly.b   0.104071173 -0.07687364 -0.651574938  0.201603808
## logmem.imd.b   0.029810858 -0.06586870  0.702317940 -0.197851739
## animal.b       0.001389212 -0.06007061  0.007530866  0.023184282
## MMSE.b        -0.199316122  0.97004597 -0.011018603  0.013849614
## age            0.024563861  0.02319017 -0.006985026  0.006729987
```

Why are there only 8 principal components?

There are only 8 principal components because there were only 8 variables in the dataset.

Write down the factor loadings of principal component 1. What are the variables that appear to have the largest weight in principal component 1?

$$PC1 = \begin{bmatrix} 0.6918 \\ 0.0404 \\ 0.0291 \\ 0.1514 \\ 0.1332 \\ 0.1981 \\ 0.0711 \\ -0.6588 \end{bmatrix}$$

DSST.b and age have the largest weight in the first principal component.

Compute the proportion of variance explained by principal components 1, 2 and 3.

$$variance = (21.333876)^2 + (8.485916)^2 + (5.888948)^2 + (4.689534)^2 + (2.537513)^2 + (2.220114)^2 + (1.521568)^2 + (1.409889)^2 = 599.4873$$

$$var(PC1, PC2, PC3) = (21.333876)^2 + (8.485916)^2 + (5.888948)^2 = 561.82474$$

$$\frac{561.82474}{599.4873} = 0.9371$$

The first three principal components explain 93.71% of the variability of the data.

How many principal components are necessary to explain at least 80% of the variance in the data?

$$var(PC1) = (21.333876)^2 = 455.13426$$

$$\frac{455.13426}{599.4873} = 0.7592$$

$$var(PC1, PC2) = (21.333876)^2 + (8.485916)^2 = 527.14503$$

$$\frac{527.14503}{599.4873} = 0.8793$$

Only the first two principal components are needed to explain 80% of the variability in the data.

Next the investigators analyzed the same data set as follows:

```
pca.1 <- prcomp(cog.data[,2:9], scale=T )
 pca.1
```

```
## Standard deviations:
## [1] 2.0134978 1.1404442 0.8889967 0.7675114 0.7070583 0.6330662 0.5202579
## [8] 0.3073308
##
## Rotation:
##                      PC1         PC2         PC3          PC4         PC5
## DSST.b         0.4015423 -0.02814745  0.38784162 -0.155517206  0.36222284
## digits.back.b  0.2691472 -0.60677316 -0.04902854 -0.043383130 -0.24108619
## digits.forw.b  0.2095507 -0.69600084 -0.07599602  0.007416434  0.09722192
## logmem.dly.b   0.4102855  0.20425375 -0.51211632 -0.094572393  0.06430731
## logmem.imd.b   0.4035603  0.18326924 -0.55713750 -0.076493380  0.04474600
## animal.b       0.3542503  0.18818345  0.30072939 -0.244685285 -0.79454143
## MMSE.b         0.3401616  0.09702789  0.13945449  0.922580714 -0.05625459
## age           -0.3873161 -0.16273366 -0.39855820  0.219216930 -0.40077633
##                       PC6         PC7           PC8
## DSST.b        -0.140095337  0.71582207  0.0114647428
## digits.back.b -0.684155134 -0.16975568  0.0096142143
## digits.forw.b  0.674702718 -0.01882191 -0.0284749906
## logmem.dly.b  -0.012587712  0.01128720 -0.7171595985
## logmem.imd.b   0.025734616  0.03733415  0.6951544512
## animal.b       0.236252718  0.04089325 -0.0009679021
## MMSE.b         0.007019259 -0.03220572 -0.0046786800
## age           -0.019898582  0.67394124 -0.0372138929
```

What is the difference between this analysis and the previous analysis?

This analysis scaled the variables to have unit standard deviations. In the previous unscaled analysis, DSST.b and age had the largest factor loadings because they had the biggest variance among the eight variables. In this new scaled analysis, the difference in factor loadings between the variables aren't as large because the standard deviations of each variable were scaled to 1.

Write down the factor loadings of principal component 1 and compare the results with those of the previous analysis.

$$PC1 = \begin{bmatrix} 0.6918 \\ 0.0404 \\ 0.0291 \\ 0.1514 \\ 0.1332 \\ 0.1981 \\ 0.0711 \\ -0.6588 \end{bmatrix} \qquad PC1' = \begin{bmatrix} 0.4015 \\ 0.2691 \\ 0.2096 \\ 0.4103 \\ 0.4036 \\ 0.3543 \\ 0.3402 \\ -0.3873 \end{bmatrix}$$

DSST and age have the largest weight in the PC1, but that's the case anymore in PC1'. The weights have much less variability in PC1' because the standard deviations have been scaled to 1. In PC1', digits backwards and digits forward have the smallest weights.

How many principal components do you need to explain at least 80% of the data variability?

$$variance = (2.0134978)^2 + (1.1404442)^2 + (0.8889967)^2 + (0.7675114)^2 + (0.7070583)^2 + (0.6330662)^2 + (0.5202579)^2 + (0.3073308)^2 = 8$$

$$var(PC1) = (2.0134978)^2 = 4.05417$$

$$\frac{4.05417}{8} = 0.5068$$

$$var(PC1, PC2) = (2.0134978)^2 + (1.1404442)^2 = 5.35478$$

$$\frac{5.35478}{8} = 0.6693$$

$$var(PC1, PC2, PC3) = (2.0134978)^2 + (1.1404442)^2 + (0.8889967)^2 = 6.1451$$

$$\frac{6.1451}{8} = 0.7681$$

$$var(PC1, PC2, PC3, PC4) = (2.0134978)^2 + (1.1404442)^2 + (0.8889967)^2 + (0.7675114)^2 = 6.73417$$

$$\frac{6.73417}{8} = 0.8417$$

The first four principal components are needed to explain 80% of the variability in the data.

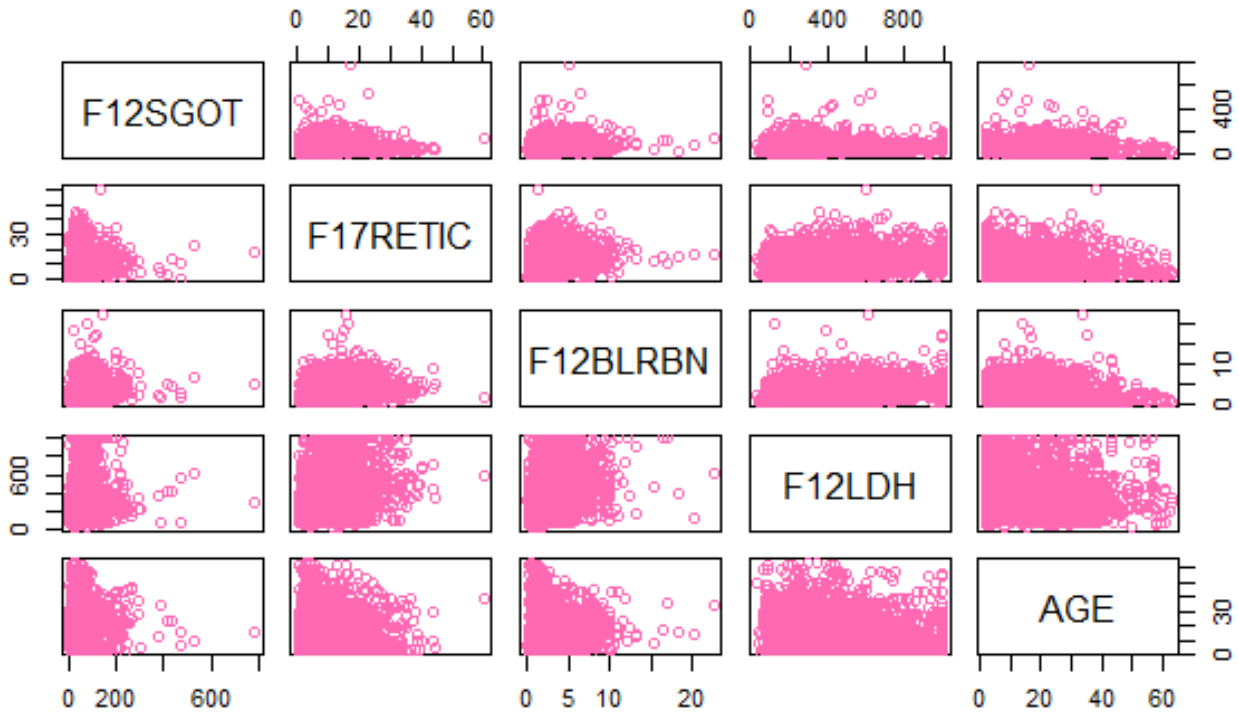
Hemolytic anemia (red blood cells that break) is a serious complication of sickle cell anemia. Typically doctors base their diagnosis on values of blood tests:

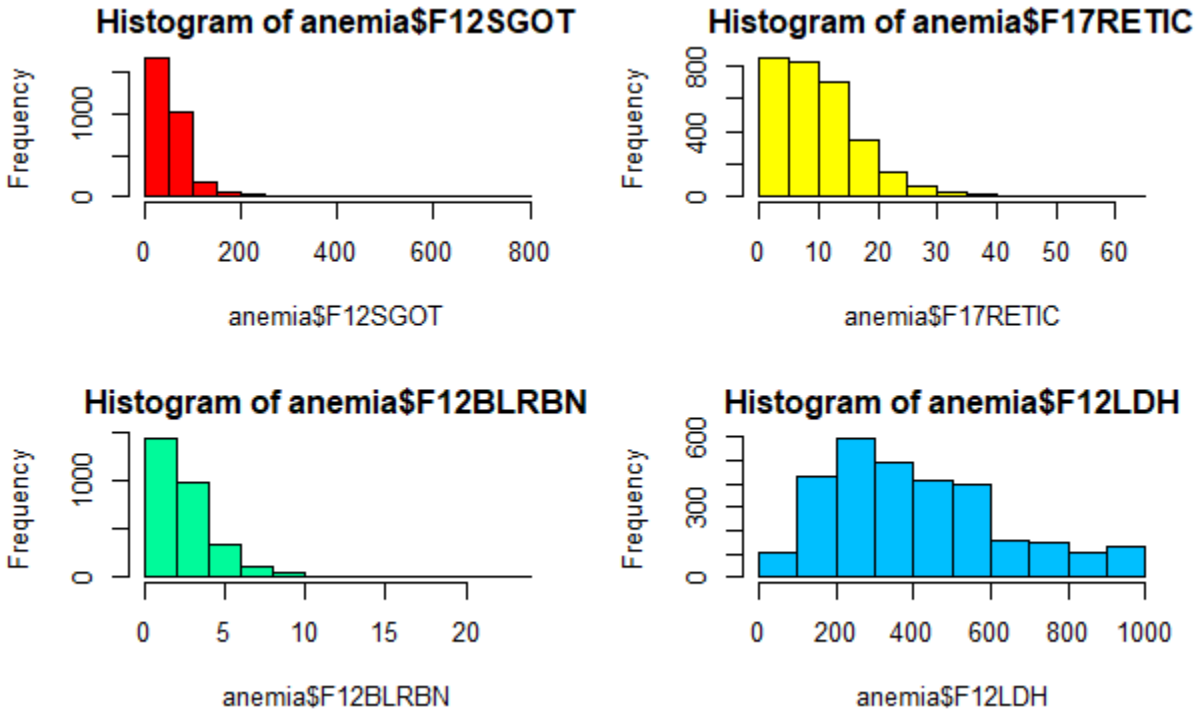
- **SGOT** – evaluates how much liver enzyme is in blood
- **LDH** (lactate dehydrogenase) evaluates enzyme required during the process of turning sugar into energy
- **Bilirubin (BLRBN)**– indicator of anemia
- **Reticulocyte counts (RETIC)**: a measure of how fast red blood cells are made from bone marrow.

Perform the following analyses using only data collected **at visit 1** in the **sca.csv**.

Produce pairwise scatter plots of the biomarkers named "**F12SGOT**","**F17RETIC**","**F12BLRBN**","**F12LDH**" (these are the untransformed data) and "**AGE**". Also produce histograms of these 4 biomarkers. Include in your write-up the plots and describe [10pts]

i. the distribution of the 4 biomarkers,
ii. the transformation that would be needed to make the biomarker data symmetrically distributed, and
iii. the mutual relations between biomarkers.





The distributions for F12SGOT, F17RETIC, F12BLRBN, and F12LDH are right-skewed, so would need to be transformed to have a normal distribution using logarithmic transformation. There do not appear to be any obvious trends between the biomarkers.

Run a PCA analysis of the biomarker data "F12SGOT","F17RETIC","F12BLRBN","F12LDH" **without rescaling the data.** Include in your write up the loading vectors of the 4 principal components, and the number of principal components that are necessary to explain 80% of the variability in the data.

$$PC1 = \begin{bmatrix} -0.9998 \\ -0.0095 \\ -0.0152 \\ -0.0028 \end{bmatrix} \quad PC2 = \begin{bmatrix} -0.0154 \\ 0.0188 \\ 0.9997 \\ 0.0102 \end{bmatrix} \quad PC3 = \begin{bmatrix} -0.0094 \\ 0.9959 \\ -0.0198 \\ 0.0882 \end{bmatrix} \quad PC4 = \begin{bmatrix} -0.0018 \\ -0.0883 \\ -0.0085 \\ 0.9961 \end{bmatrix}$$

Only the first principal component is needed to explain 80% of the variability in the data

Explain why the first principal component is highly correlated with the variable F12LDH.

The first principal component is highly correlated with F12LDH because that variable has a standard deviation of 230.1828, which is an order of magnitude away from the standard deviations of the other biomarkers, (F17RETIC: 7.1444, F12SGOT: 43.9104, F12BLRBN: 2.0624). The higher the variance, the more weight the variable has in that principal component.

Run a PCA analysis of the biomarker data "F12SGOT","F17RETIC","F12BLRBN","F12LDH" **now rescaling the data.** Include in your write up the loading vectors of the 4 principal components, and the number of principal components that are necessary to explain at least 80% of the variability in the data.

$$PC1' = \begin{bmatrix} -0.4957 \\ -0.5505 \\ -0.3350 \\ -0.5822 \end{bmatrix} \quad PC2' = \begin{bmatrix} -0.4339 \\ 0.2098 \\ 0.8738 \\ 0.0651 \end{bmatrix} \quad PC3' = \begin{bmatrix} -0.7389 \\ 0.5968 \\ -0.2387 \\ 0.2021 \end{bmatrix} \quad PC4' = \begin{bmatrix} -0.1415 \\ -0.5447 \\ -0.2595 \\ 0.7848 \end{bmatrix}$$

The first three principal component are needed to explain 80% of the variability in the data

Based on these analyses, do you think it is important to rescale the data before conducting a PCA?

Based on these analyses, I think it is very important to rescale the data before conducting a PCA. F12LDH has a disproportionately high standard variance compared to the other biomarkers, so it will have the largest loading in the first principal component. Scaling the standard deviations to 1 would remove that correlation and weigh all the biomarkers more equally.