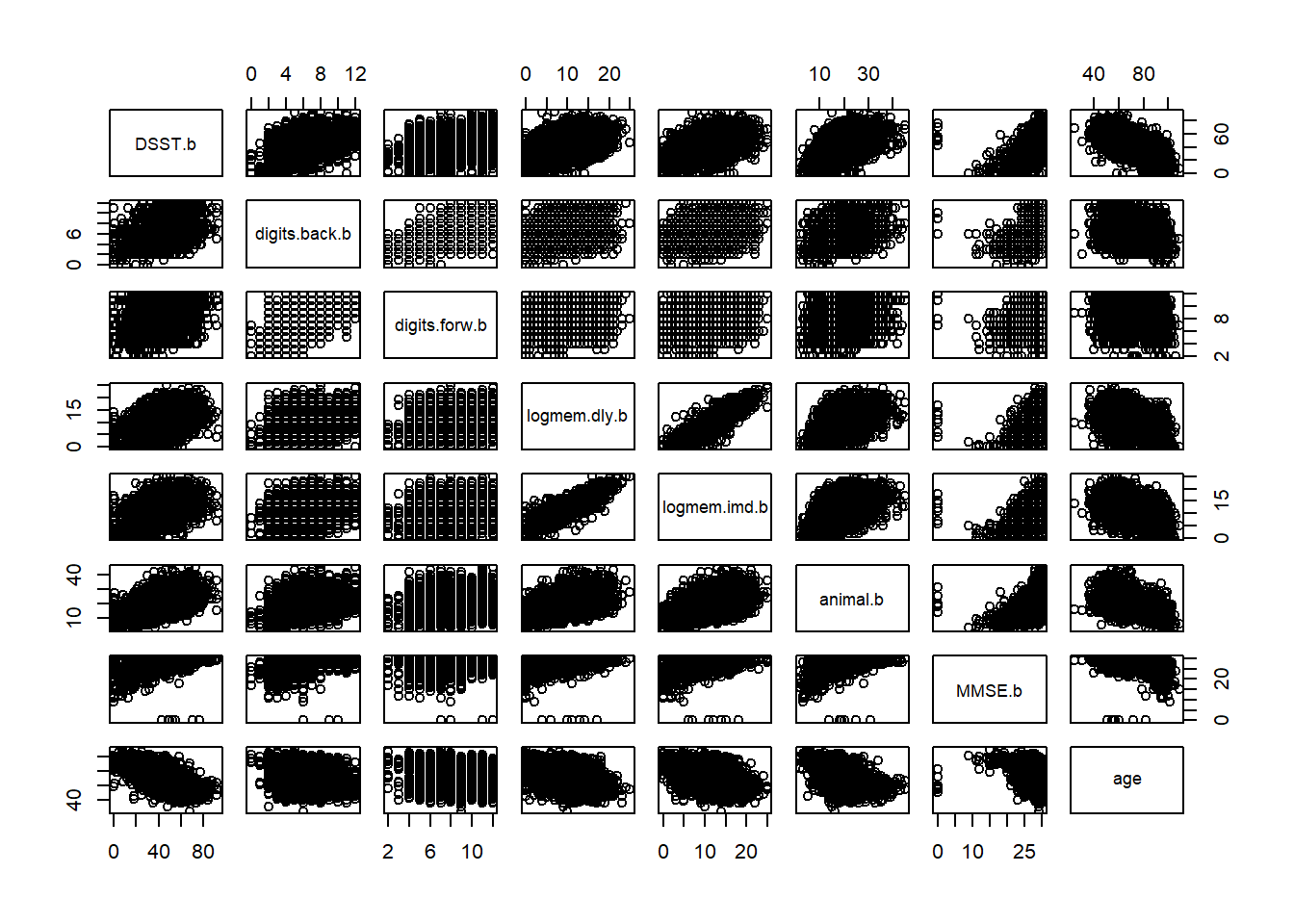
# **BS806 --- Homework 9**

**Reading**. Chapter 10 of (ISLR) Introduction to statistical learning. Pages 373—385, 401—404.

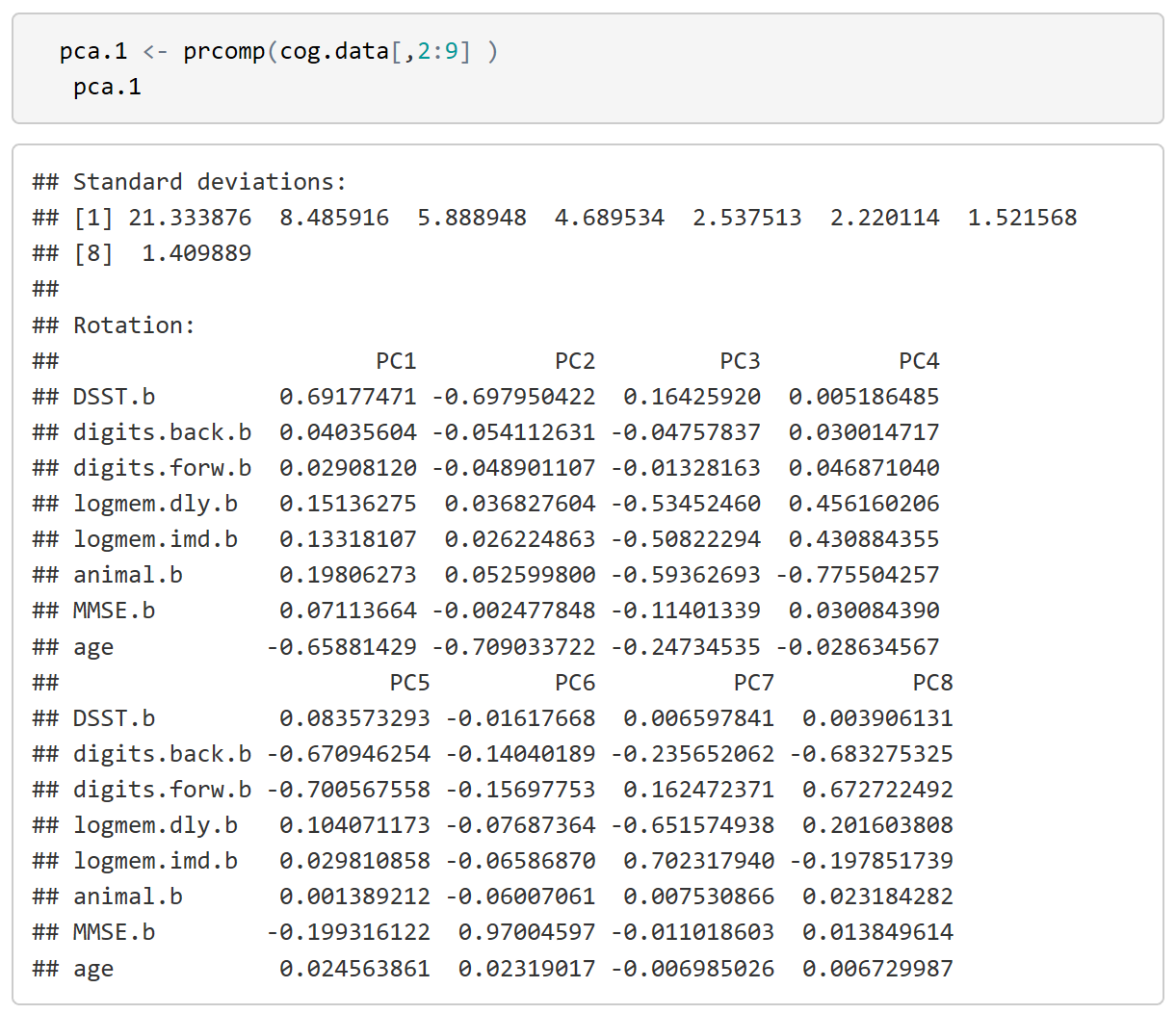
**Question 1**. (**60 points**) A data set of measurements of cognitive functions of 4010 individuals aged 30 and older was analyzed using principal component analysis. The cognitive functions were measured through these 7 tests:

* **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)

1. 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]

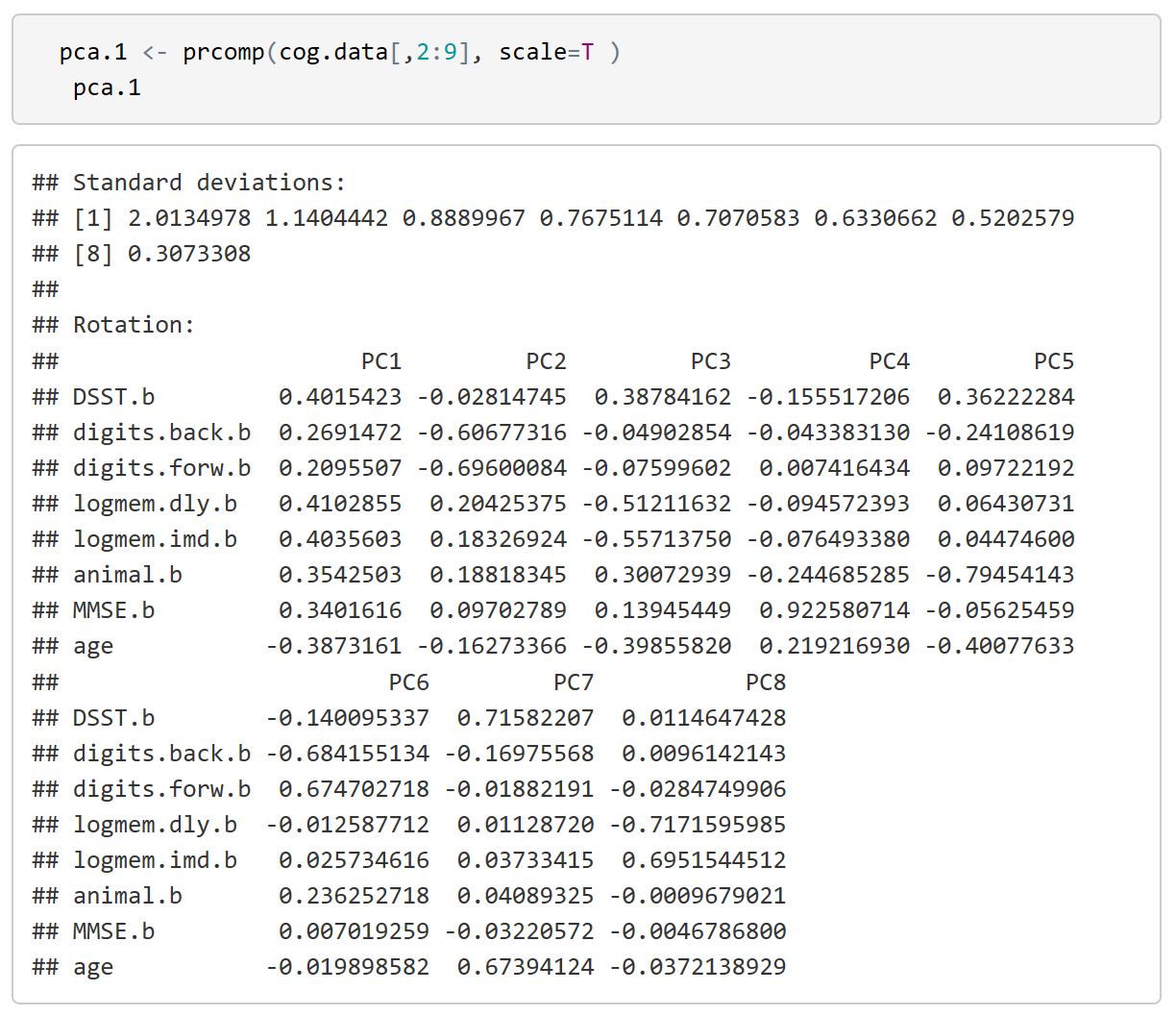


The following output was produced in RStudio:



1. Why are there only 8 principal components? [5pts]
2. Write down the factor loadings of principal component 1. What are the variables that appear to have the largest weight in principal component 1? [10pts]
3. Compute the proportion of variance explained by principal components 1, 2 and 3. [10pts]
4. How many principal components are necessary to explain at least 80% of the variance in the data? [5 pts]

Next the investigators analyzed the same data set as follows:



1. What is the difference between this analysis and the previous analysis? [5pts]
2. Write down the factor loadings of principal component 1 and compare the results with those of the previous analysis (part c). [10pts]
3. How many principal components do you need to explain at least 80% of the data variability? [5pts]

**Question 2. (40 points)** 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** which is available on the course website.

1. 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]
   1. the distribution of the 4 biomarkers,
   2. the transformation that would be needed to make the biomarker data symmetrically distributed, and
   3. the mutual relations between biomarkers.
2. 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. [10pts]
3. Explain why the first principal component is highly correlated with the variable F12LDH. [5pts]
4. 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. [10pts]
5. Based on these analyses, do you think it is important to rescale the data before conducting a PCA? [5pts]

[Be complete]