BS 806
Homework 9
Irene Hsueh

Question 1

Part A

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

Part B

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

Part C

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

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

Part D

$$\begin{aligned} variance = {} & (21.333876)^2 + (8.485916)^2 + (5.888948)^2 + (4.689534)^2 + (2.537513)^2 \\ & + (2.220114)^2 + (1.521568)^2 + (1.409889)^2 = 599.4873 \end{aligned}$$

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

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

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

Part E

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

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

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

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

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

Part F

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.

Part G

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

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.

Part H

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

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

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

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

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

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

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

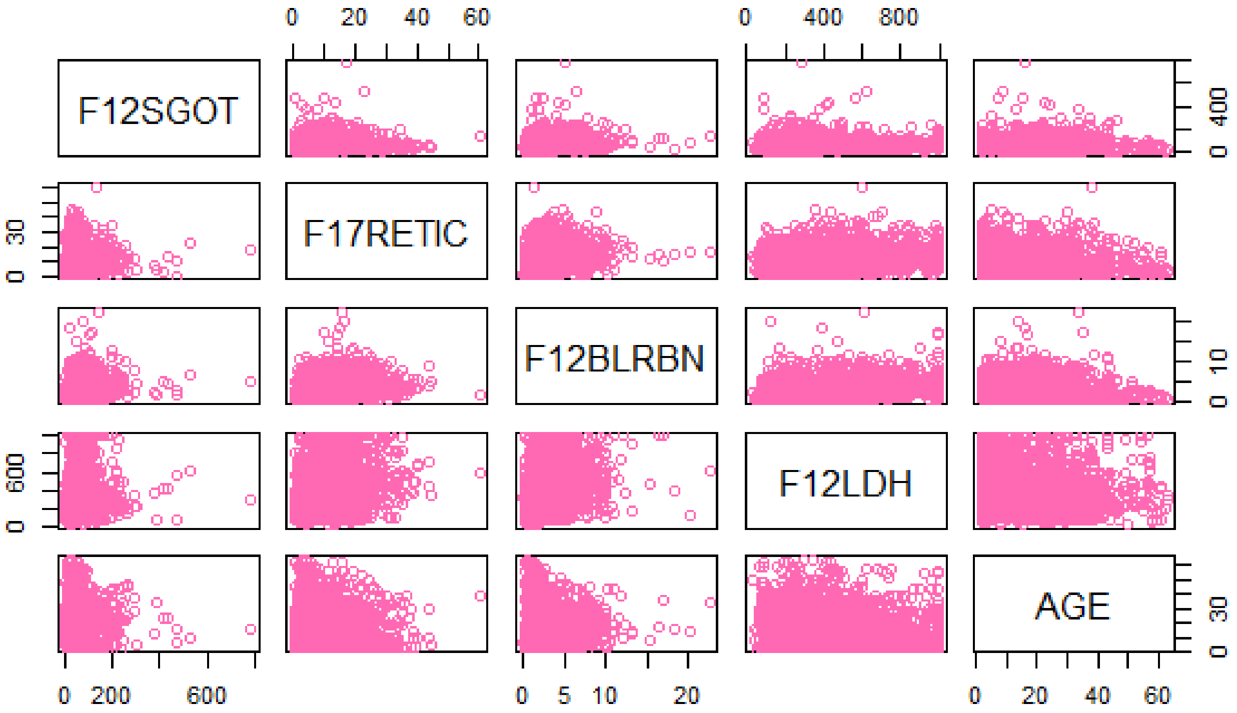
$$var(PC1, PC2, PC3, PC4) = (2.0134978)^2 + (1.1404442)^2 + (0.8889967)^2 + (0.7675114)^2 \\ = 6.73417$$

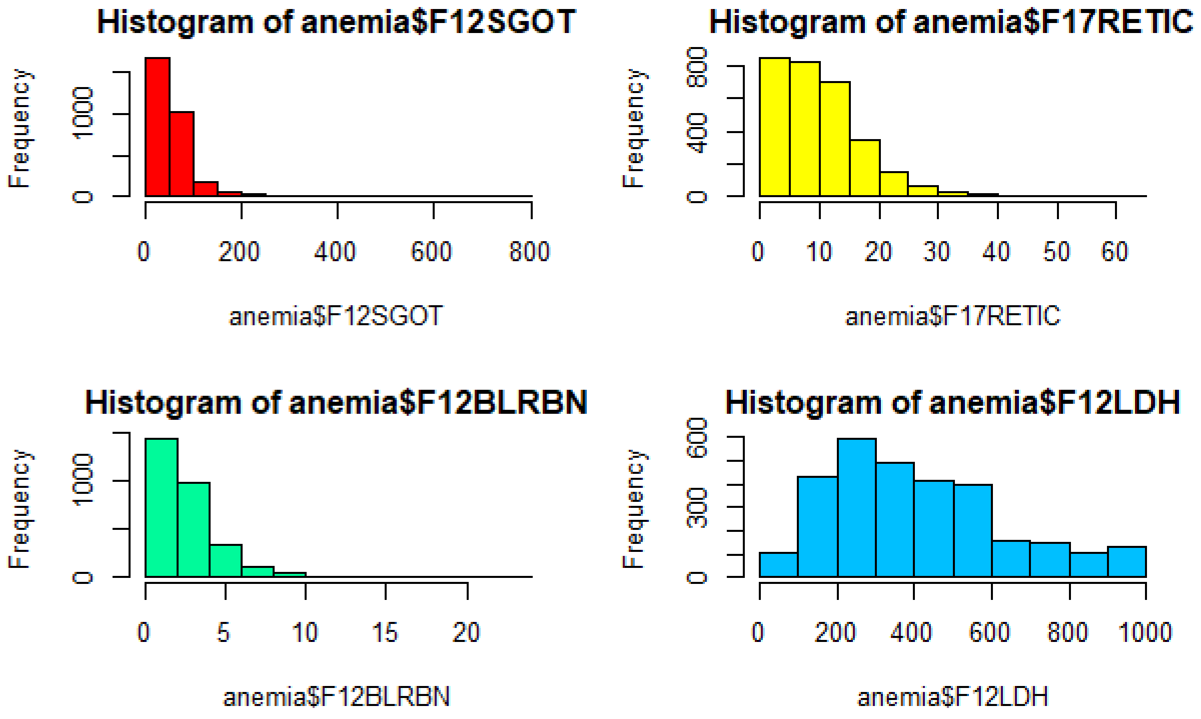
$$\frac{6.73417}{8} = 0.8417$$

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

Question 2

Part A





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.

Part B

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

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

Part C

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.

Part D

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

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

Part E

Based on these analyses, I think it’s 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.