Make sure you answer in a structured way, make sure your handwriting is readable.

1) Describe in general terms what the code below is doing? Golub is a gene-patient dataset. The dataset contains 38 patients and 3051 genes. Explain why you need to perform the different steps. Explain what you expect in the output.

a)

dim(golub)

#[1] 3051 38

row\_mean = apply(golub, 1, mean, na.rm=TRUE)

golub\_meancentered = golub - row\_mean

dim(golub\_meancentered)

[1] 3051 38

SD = apply(golub\_meancentered, 1, sd, na.rm=TRUE)

golub\_rescaled = golub\_meancentered/SD

dim(golub\_rescaled)

[1] 3051 38

d.euclidian <- dist(golub\_rescaled, method="euclidian")

m <- as.matrix(dist(golub))

dim(m)

#(3051,3051)

kmeangolub100= kmeans(d.euclidian,100)

2) PCA was performed on a dataset that contains the measurements of 154 metabolites profiled at several time points after treatment (more specifically at time 0 h, 4, 12, 24, 48 and 96 h after treatment). For each time point several replicates were measured) : The total datamatrix mDC (dim 154 X 52) contains 52 columns (corresponding to the timepoints and their replicates) and 154 metabolites (rows).

1. Explain the purpose of PCA and explain in this context.
2. Explain what you observe in the results below for a), b) and c) and relate what you describe to your previous answer

The lines below illustrate how the data look like: X12h.1 refers to e.g. timepoint 12 first replicate

mDC = metaboliteDataComplete

dim(mDC)

[1] 154 52

(a)

mDC[1,]

X0h X0h.1 X0h.2 X0h.3 X0h.4 X0h.5 X0h.6 X1h X1h.1 X1h.2 X1h.3 X1h.4 X1h.5 X1h.6 X1h.7 X4h X4h.1 X4h.2 X4h.3 X4h.4

0.0015 0.0252 0.0159 0.0116 -0.0431 -0.0053 -0.0058 0.0127 0.0309 0.0092 0.0430 -0.0457 -0.0108 -0.0126 -0.0831 0.0192 0.0156 -0.0015 -0.0500 -0.0481

X4h.5 X12h X12h.1 X12h.2 X12h.3 X12h.4 X12h.5 X12h.6 X12h.7 X24h X24h.1 X24h.2 X24h.3 X24h.4 X24h.5 X24h.6 X48h X48h.1 X48h.2 X48h.3

0.0127 0.1563 0.1174 0.1112 0.1411 0.0287 0.0597 0.0641 0.0839 0.1388 0.1160 0.1008 0.0276 0.0917 0.0629 0.0574 0.2348 0.2672 0.2599 0.2314

X48h.4 X48h.5 X48h.6 X48h.7 X96h X96h.1 X96h.2 X96h.3 X96h.4 X96h.5 X96h.6 X96h.7

0.1901 0.2015 0.2209 0.1881 0.3410 0.3206 0.2923 0.3276 0.2742 0.2560 0.2424 0.2655

>

PCA was applied to this dataset as follows

PCAres<-prcomp(t(mDC), scale = TRUE, center=TRUE)

(b)

head(PCAres$rotation)

…

PC48 PC49 PC50 PC51

Xylose methoxyamine (4TMS) -0.07859999 0.19098032 -0.017244269 -0.026338329

Tyramine (3TMS) -0.04106291 -0.01065654 0.026613024 0.075533621

trans-Sinapinic acid (2TMS) -0.12356559 -0.09788160 -0.066118945 0.107147173

Threonic acid-1,4-lactone (2TMS), trans- -0.04258868 0.12562553 -0.008027899 -0.104774412

Threonic acid (4TMS) 0.03251710 0.02498274 0.005758170 0.021452485

Succinic acid (2TMS) 0.06073680 -0.04754981 0.003804854 -0.006230646

PC52

Xylose methoxyamine (4TMS) -0.06446234

Tyramine (3TMS) 0.03340616

trans-Sinapinic acid (2TMS) -0.42494195

Threonic acid-1,4-lactone (2TMS), trans- 0.04360600

Threonic acid (4TMS) 0.07763435

Succinic acid (2TMS) -0.06637592

And the results were plotted as follows

(c)

X11

plot(predict(PCAres)[,1],predict(PCAres)[,2])

abline(v=0, col="gray")

abline(h=0, col="gray")

text(predict(PCAres)[,1],predict(PCAres)[,2], labels=sub("X(.+h)(\\..)?","\\1",rownames(t(mDC))),cex=1, adj = c(0,0))

