1. Given a leukemia dataset (golub), containing the expression of 3051 genes in 38 patients having either type 1 or type 2 leukemia. PCA is performed on this dataset (see code below).

gt = t(golub)

dim(gt)

#38 X 3051

Perform PCA

PCAres\_t<-prcomp(gt, center = TRUE, scale = TRUE)

dim(as.matrix(PCAres\_t$rotation[,1]))

# 3051 1

predict(PCAres\_t)[,1]

dim(as.matrix(predict(PCAres\_t)))

#38 38

dim(as.matrix(predict(PCAres\_t)[,1]))

#38 1

Plotted results (type 1= ‘0’, type 2 =’1’)



1. What are the variables and what are the observations in this example (how do you know)
2. What does ‘PCAres\_t$rotation[,1]’ represent
3. What does ‘predict(PCAres\_t)[,1]’ represent
4. Explain why you would perform PCA in this example

2) Yes or no. Motivate you answer (try to be concise and complete):

1. Applying K-means clustering with a ‘correlation-derived distance’ metric on non-rescaled data (with rescaling = mean centering + variance rescaling) gives exactly the same results as applying K-means clustering with a Euclidian distance on the rescaled data.
2. Increasing the parameter in K Means that determines the number of clusters (N) results in reducing the cluster size (i.e. the number of entities in a cluster).

kmeangolub10 = kmeans(d.correlation,N)