a. Perform ANOVA test on the discriminant analysis scores of nuclear localization signals of both nuclear and non-nuclear proteins by class variables (Target).

yeast <- read.table(url("http://archive.ics.uci.edu/ml/machine-learning-databases/yeast/yeast.data"), header = FALSE)

names(yeast)<- c("SequenceName", "mcg", "gvh", "alm", "mit", "erl", "pox", "vac", "nuc", "LocalizationSite")

pca <- princomp(yeast[, 2:9], cor=T) # principal components analysis using correlation matrix

pc.comp <- pca$scores

PrincipalComponent1 <- -1\*pc.comp[,1] # principal component 1 scores (negated for convenience)

PrincipalComponent2 <- -1\*pc.comp[,2] # principal component 2 scores (negated for convenience)

clustering.data <- cbind(PrincipalComponent1, PrincipalComponent2)

b. Which class is significantly different from others?

set.seed(100)

km <- kmeans(clustering.data, 8, iter.max = 30, nstart=30)

km

km$cluster

plot(PrincipalComponent1, PrincipalComponent2, col=km$cluster)

points(km$centers, pch=16)

aggregate(yeast[, 2:9],by=list(km$cluster),mean)

table(km$cluster, yeast$LocalizationSite)