Naman Code

# R code solution for Group 10

# Importing various libraries

# install.packages(c("tidyverse","factoextra", "cluster"))

library(tidyverse)  # data manipulation

library(cluster)    # clustering algorithms

library(factoextra) # clustering algorithms & visualization

# Q1

dungaree <- read.csv("dungaree.csv")

# dungaree

summary(dungaree)

# a) As there would be need of all the columns to cluster the data, we will not drop any column.

# b) Examine the input variables: are there any unsual data values? Are there missing values that should be replaced

# finding if there are any missing values

sum(is.na(dungaree))

# there are no missing values so we don't have to replace any values

# Normalizing the data as the values in a few columns are larger which can cause abnormal behaviour in the scattering

normalized\_dungaree <- sapply(dungaree, scale)

# normalized\_dungaree

set.seed(42)

# Implementing K means here with seed = 42 and k=20

km20 <- kmeans(normalized\_dungaree, 20)

# km20

# We will try to plot the kmeans cluster with k=20 with the following code

plot\_20<-fviz\_cluster(km20, geom = "point", data = normalized\_dungaree) + ggtitle("k = 20")

# Determining the optimal clusters with the help of Elbow method

# fviz\_nbclust is the function used to do the elbow method for K means

# here with the elbow method approach we can see that the optimal number of clusters required for the above normalized data would be 3.

# As the number of clusters given were 20 which is way ahead of the optimal cluster, this will cause over fitting problem in the dataset. Thus, we would try to avoid this and use k=5 in the problem further.

#  we will try to plot the graph with k=5 to see the difference

# Trying to plot the scatter plot for k=6

km6 <- kmeans(normalized\_dungaree,6)

plot\_6<-fviz\_cluster(km6, geom = "point", data = normalized\_dungaree) + ggtitle("k = 6")

# k=3 plot for kmeans

km3 <- kmeans(normalized\_dungaree, 3)

plot\_3<-fviz\_cluster(km3, geom = "point", data = normalized\_dungaree) + ggtitle("k = 3")

# printing all the plots together

# install.packages("gridExtra")

library(gridExtra)

grid.arrange(plot\_20, plot\_6, plot\_3, nrow = 2)

# printing the centroids of the clusters with k=6

km6$centers

# plotting the centroids of the cluster

# plot centroids

plot(c(0), xaxt = 'n', ylab = "", type = "l", ylim = c(min(km6$centers), max(km6$centers)), xlim = c(0, 6))

# label x-axes

axis(1, at = c(1:6), labels = names(dungaree))

for (i in c(1:6))

    lines(km6$centers[i,], lty = i, lwd = 3, col = ifelse(i %in% c(1, 3, 5),"black", "dark grey"))

# name clusters

text(x =0.5, y = km6$centers[, 1], labels = paste("Cluster", c(1:6)))

# km6 centroids

km6$centers

dist(km6$centers)

Khushi Code

Q2

setwd("E:/Khushi/UTD/BawithR/Datasets")

pharma.df <- read.csv("Pharmaceuticals.csv")

#--------------------2a------------------------

row.names(pharma.df) <- pharma.df[,1]

#Remove the symbol column

pharma.df <- pharma.df[,-1]

#Remove the Name column

pharma.df <- pharma.df[,-1]

#Remove the Median\_Recommendation column

pharma.df <- pharma.df[,-10]

#Remove the Location column

pharma.df <- pharma.df[,-10]

#Remove the Exchange column

pharma.df <- pharma.df[,-10]

#--------------------2b------------------------

#Normalize input variables

pharma.df.norm <- sapply(pharma.df, scale)

#Add row names: Symbols

row.names(pharma.df.norm) <- row.names(pharma.df)

#--------------------2c------------------------

#Compute normalized distance based on all 9 variables

d.norm <- dist(pharma.df.norm, method = "euclidean")

#In hclust(), set argument method = to "ward.D", "single", "complete", "average", "median", or

"centroid"

hc1 <- hclust(d.norm, method = "single")

plot(hc1, hang = -1, ann = TRUE)

#--------------------2d------------------------

memb <- cutree(hc1, k = 6)

memb

#Using hclust to show better grouping; dendexten

plot(hc1, cex = 1)

rect.hclust(hc1, k = 6, border = 2:5)

#--------------------2e------------------------

hc2 <- hclust(d.norm, method = "complete")

plot(hc2, hang = -1, ann = TRUE)

#--------------------2f------------------------

memb <- cutree(hc2, k = 6)

memb

#Using hclust to show better grouping; dendexten

plot(hc2, cex = 1)

rect.hclust(hc2, k = 6, border = 2:5)

#--------------------2g------------------------

dend1 <- as.dendrogram (hc1)

dend2 <- as.dendrogram (hc2)

library(dendextend)

tanglegram(dend1, dend2)