Cluster Analysis

utilities.df <- read.csv("C:/MA 299/R/Utilities.csv")

# set row names to the utilities column

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

# remove the utility column

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

# normalize input variables

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

# add row names: utilities

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

# compute normalized distance based on variables Sales and FuelCost

# compute Euclidean distance (to compute other distance measures, change the value in method = )

d.norm <- dist(utilities.df.norm[,c(6,8)], method = "euclidean")

d.norm

##===============================================================================

# compute normalized distance based on all 8 variables

d.norm <- dist(utilities.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 = FALSE)

##===============================================================================

memb <- cutree(hc1, k = 6)

memb

##===============================================================================

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

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

##===============================================================================