Title: Assignment #4: (*Clustering datasets – Categorical and Mixed(both categorical and numerical)*)

Purpose: *This assignment is to study clustering datasets in R. Clustering datasets of different categories in R and also clustering mixed datasets which consists of both categorical and numeric datasets.*

Dataset(s): *The dataset is obtained from the UCI Machine Learning Repository. The link for dataset can be found here:*

https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

https://archive.ics.uci.edu/ml/datasets/Wholesale+customers

Approach:

For Dataset -1 – Congressional Voting Records - Categorical Dataset

- *First the dataset is loaded, as a matrix dataframe.*
- *It is checked for NA's if any. This dataset had no NA's, but it had a "?" on the values. The "?" cannot be assumed to be NA because the "?" actually means that they are not answering "yes" or "no" and it slightly in-between as per the dataset description. So I tried changing the "?" to "m" but it didn't help much.*
- *The first column is removed, which indicates "democrat" or "republican".*
- *After removing the first column, every other column is converted into a factor. (Since all are "y" or "n")*
- *Data frame of the factored dataset is produced.*
- *Binary and Euclidian distance matrix is computed and saved.*
- *Hierarchical clustering using both Euclidian and Binary distance matrix is computed and saved.*
- *The clustering is computed for 2 clusters and a summary is plotted.*
- *Same is done using ade4 clustering mechanism for hierarchical clustering.*
- *It is observed that ade4 clustering mechanism is more better in producing better results compared to dist matrix.*
- *In addition to it, I observed that the Euclidian distance matrix produces slightly better results compared to binary. I assume it might be due the presence of "?" as a added variable.*

For Dataset 2 – Wholesale Customers – Mixed dataset of categorical and numerical values:

- *First the dataset is loaded. Checked for NA's if any (this dataset had none).*
- *It is said that the column 1 and column 2 are Categorical and the remaining columns are numerical.*

- *I used the daisy package to categorize and cluster the dataset. The metric is used as “bower” and then the type is listed as type = list(ordratio = c(1,2)).*
- *The hierarchical clustering is computed on it and then the results are observed.*
- *It is observed that though the cluster is computed and categorized, it clusters somewhat shaky because the “3” region is not clustered computed properly.*

Summary:

It is observed that if it is a completely categorized data, converting to factors and obtaining the results are easier. But if it is mixed, we need to use the bower metric, and since the channel is binary factor and the Region is categorized into 3 factors and it is slightly difficult for the clustering to categorize it accurately.

Appendix :

This includes the R code that I have done to obtain the analysis for the text mining:

```
library(readr)
library("klaR")
library("ade4")


house_votes_84 <- read_csv("~/Downloads/Education/Spring 17/R for Data
Scientists/Assignments/house-votes-84.txt")
View(house_votes_84)
dim(house_votes_84)
myData <- house_votes_84
myData = as.data.frame(unclass(house_votes_84))
head(myData)
myData$republican<-NULL
myDataClean <- na.omit((myData))
dim(myDataClean)
length(myDataClean)

for ( i in 1 : length(myDataClean)){
myDataClean[,i] = as.factor(myDataClean[,i])
}

head(myDataClean)

dataBinMM = data.frame(model.matrix(~.,data = myDataClean)[,-1])
head(dataBinMM)

dataDistBinaryMM <- dist(dataBinMM,method = "binary")
dataDistEucMM <- dist(dataBinMM,method = "euclidian")
```

```
fitMMBi <- hclust(d=dataDistBinaryMM, method ="ward.D2")
fitMMEu <- hclust(d=dataDistEucMM, method ="ward.D2")

plot(fitMMBi)

xMMbi = cutree(fitMMBi,2)
y = myDataClean[xMMbi ==1,]
summary(y)
y=myDataClean[xMMbi == 2,]
summary(y)

xMMEu=cutree(fitMMEu,2)
y=myDataClean[xMMEu == 1,]
summary(y)
y=myDataClean[xMMEu == 2,]
summary(y)

table(xMMbi,xMMEu)

newBinaryData = acm.disjonctif(myDataClean)
head(newBinaryData)

distBinaryade4 = dist(newBinaryData, method = "binary")
distEucade4 = dist(newBinaryData, method = "euclidian")
fitade4Bi = hclust(d=distBinaryade4,method="ward.D2")
fitade4Eu = hclust(d=distEucade4,method="ward.D2")

xade4bi=cutree(fitade4Bi,2)
y=myDataClean[xade4bi == 1,]
summary(y)
y=myDataClean[xade4bi == 2,]
summary(y)

xade4eu=cutree(fitade4Eu,2)
y=myDataClean[xade4eu == 1,]
summary(y)
y=myDataClean[xade4eu == 2,]
summary(y)

table(xade4bi,xade4eu)

km = kmodes(myDataClean, 2)

table(xade4bi,km$cluster)
table(xade4eu,km$cluster)

table(xMMbi,km$cluster)

table(xMMEu,km$cluster)
```

```
library(readr)
library(dplyr)
library(cluster)
library(Rtsne)
library(Rtsne)
library(ggplot2)

Wholesale_customers_data <- read_csv("~/Downloads/Education/Spring 17/R for Data
Scientists/Assignments/Wholesale customers data.csv")
View(Wholesale_customers_data)
myData <- Wholesale_customers_data

dim(myData)
myDataClean <- na.omit((myData))
dim(myDataClean)
glimpse(myDataClean)

bower_dist <- daisy(myDataClean, metric = "gower", type =list(ordratio=c(1,2)))
summary(bower_dist)

gower_mat <- as.matrix(bower_dist)

sil_width <- c(NA)
for(i in 2:10){
pam_fit <- pam(bower_dist,
diss = TRUE,
k = i)
sil_width[i] <- pam_fit$silinfo$avg.width
}

plot(1:10, sil_width,
xlab = "Number of clusters",
ylab = "Silhouette Width")
lines(1:10, sil_width)

pam_fit <- pam(bower_dist, diss = TRUE, k = 3)
```

```
tsne_obj <- Rtsne(bower_dist, is_distance = TRUE)

tsne_data <- tsne_obj$Y %>%
data.frame() %>%
setNames(c("X", "Y")) %>%
mutate(cluster = factor(pam_fit$clustering),
name = myDataClean$Region)

ggplot(aes(x = X, y = Y), data = tsne_data) +
geom_point(aes(color = cluster))

# tsne_data %>%
# filter(X > 15 & X < 25,
# Y > -15 & Y < -10) %>%
# left_join(myDataClean, by = "Region") %>%
# collect %>%
# .[["Region"]]
```

Results obtained from the first Dataset and second dataset:

```
    xMMEu
xMMbi   1   2
   1 155  42              Euclidian vs Binary Data Matrix
   2   0 237



     xade4eu
xade4bi   1   2             Euclidian vs Binary ade4
     1 186  29
     2   0 219

xade4bi   1   2
     1  21 194              Binaryade4 vs Kmclustering
     2 215   4


xade4eu   1   2
     1   7 179              Euclidianade4 vs kmClustering
     2 229  19

xMMbi   1   2
   1  15 182                Binary Dist matrix vs Km clustering
   2 221  16

xMMEu   1   2
   1   0 155
   2 236  43
```

Euclidian dist matrix vs Km clustering

Second dataset results:

```
X
     1   2   3
  1  18   0  59
  2  19   0  28
  3 105 211   0
```