---
title: "Assignment5_Hierarchical Clustering"
author: "Ram"
date: "11/29/2021"
output: word_document

Setting up working directory
```{r}
setwd("C:/Users/ramne/Desktop/ML Assignment/Hierarchical Clustering")
set.seed(123)
```

Loading required libraries.

```{r}
library(cluster)
library(caret)
library(dendextend)
library(knitr)
library(factoextra)
```

 Data Importing
 cereals dataset

```{r}
library(readr)
cereals<-read.csv("Cereals.csv")
DataFrame <- data.frame(cereals[,4:16])
```

 Data Pre-Processing

To remove any missing value that might be present in the data.

```{r}
OmitMissing <- na.omit(DataFrame)
```

 Data Normalization & Data Scaling:

 Normalizing the Data using Scale function.
```{r}
Normalise <- scale(OmitMissing)
```

Using the euclidean distance to measure the distance:

Computing the dissimilarity matrix values by using Dist and the method is Euclidean.

```{r}
d <- dist(Normalise, method = "euclidean")
```

Perform Hierarchical Clustering using complete linkage.

```{r}
HC <- hclust(d, method = "complete")
plot(HC)
```
Plotting the dendogram.

```{r}
round(HC$height, 3)
```
 Determining Optimal Clusters:

Highliting the clusters directly in dendogram
```{r}
plot(HC)
rect.hclust(HC,
  k = 4, # k is used to specify the number of clusters
  border = "Blue"
)
```

 We can also use agnes() function to perform clustering.

 Performing clustering using agnes() with single, complete, average and ward.

```{r}
HCsingle <- agnes(Normalise, method = "single")
HCcomplete <- agnes(Normalise, method = "complete")
HCaverage <- agnes(Normalise, method = "average")
HCward <- agnes(Normalise, method = "ward")
```

Now we will compare the agglomerative coefficients for Single,complete, average and ward.

```{r}
print(HCsingle$ac)
print(HCcomplete$ac)
print(HCaverage$ac)
print(HCward$ac)
```

 The results say that the wards method is the best with the value of 0.904.
 Plotting the agnes using ward method and Cutting the Dendrogram. We will take k = 4 by observing the distance.
```{r}
pltree(HCward, cex = 0.6, hang = -1, main = "Dendrogram of agnes-Ward")
```

 Hierarchical clustering using ward method.

```{r}
HC1 <- hclust(d, method = "ward.D2" )
subgrp <- cutree(HC1, k = 4)
table(subgrp)
dataframe <- as.data.frame(cbind(Normalise,subgrp))
```

To visualiZe the results in scatter plot.

```{r}
fviz_cluster(list(data = Normalise, cluster = subgrp))
```

To check the structure of the clusters and on their stability.

We will partition the data and apply one part to the other part
```{r}
Datapart1 <- OmitMissing[1:50,]
Datapart2 <- OmitMissing[51:74,]
```

Performing Hierarichal Clustering using agnes() with single, complete, average and ward with partitioned data , plotting dendrogram and then cutting the dendrogram by taking k = 4.

```{r}
Award <- agnes(scale(Datapart1), method = "ward")
Aaverage <- agnes(scale(Datapart1), method = "average")
Acomplete <- agnes(scale(Datapart1), method = "complete")
Asingle <- agnes(scale(Datapart1), method = "single")
cbind(ward=Award$ac, average=Aaverage$ac, complete=Acomplete$ac,
      single=Asingle$ac)
```

Plot dendogram for the partitioned data.
```{r}
pltree(Award, cex = 0.6, hang = -1, main = "Dendogram of Agnes-Ward")
rect.hclust(Award, k = 4, border = 2:5)
```

Using Cutree to divide into groups cluster = 4.

```{r}
c <- cutree(Award, k = 4)
print(c)
```

Calculating centers to assess the consistency of data.

```{r}
Assess <- as.data.frame(cbind(Datapart1,c))
Assess[Assess$c==1,]
c1 <- colMeans(Assess[Assess$c==1,])
Assess[Assess$c==2,]
c2 <- colMeans(Assess[Assess$c==2,])
Assess[Assess$c==3,]
c3 <- colMeans(Assess[Assess$c==3,])
Assess[Assess$c==4,]
c4 <- colMeans(Assess[Assess$c==4,])
```

Binding the 4 centers.

```{r}
```

```
centers <- rbind(c1,c2,c3,c4)
centers
```

Calculating Distance and comparing the record in B with the closest centroid in A

```{r}
d1 <- as.data.frame(rbind(centers[,-14],Datapart2))
d2 <- get_dist(d1)
matrix <- as.matrix(d2)
df1 <-
data.frame(data=seq(1,nrow(Datapart2),1),clusters=rep(0,nrow(Datapart2)))
for(i in 1:nrow(Datapart2)) {

  df1[i,2] <- which.min(matrix[i+4, 1:4])
}
df1
cbind(dataframe$subgrp[51:74], df1$clusters)
table(dataframe$subgrp[51:74] == df1$clusters)
```

From above Results , 12 are True and 12 are False, so we can say the model may be stable.

```{r}
```

Selecting the cluster that is best cereal for breakfast, which will have high protein, fiber and low in sugar, sodium.

Choosing the Cluster of Healthy Cereals.

```{r}
newdata <- cereals
newdata_omit <- na.omit(newdata)
Clust <- cbind(newdata_omit, subgrp)
Clust[Clust$subgrp==1,]
Clust[Clust$subgrp==2,]
Clust[Clust$subgrp==3,]
Clust[Clust$subgrp==4,]
```

Calculating mean ratings to determine the best cluster.

```{r}
mean(Clust[Clust$subgrp==1,"rating"])
mean(Clust[Clust$subgrp==2,"rating"])
mean(Clust[Clust$subgrp==3,"rating"])
mean(Clust[Clust$subgrp==4,"rating"])
```

As we can see that the mean ratings for the subgrp==1 is the highest(73.84), it's the best option to choose cluster 1 and the cereals in the cluster 1 for healthy diet.