ML Assignment-5

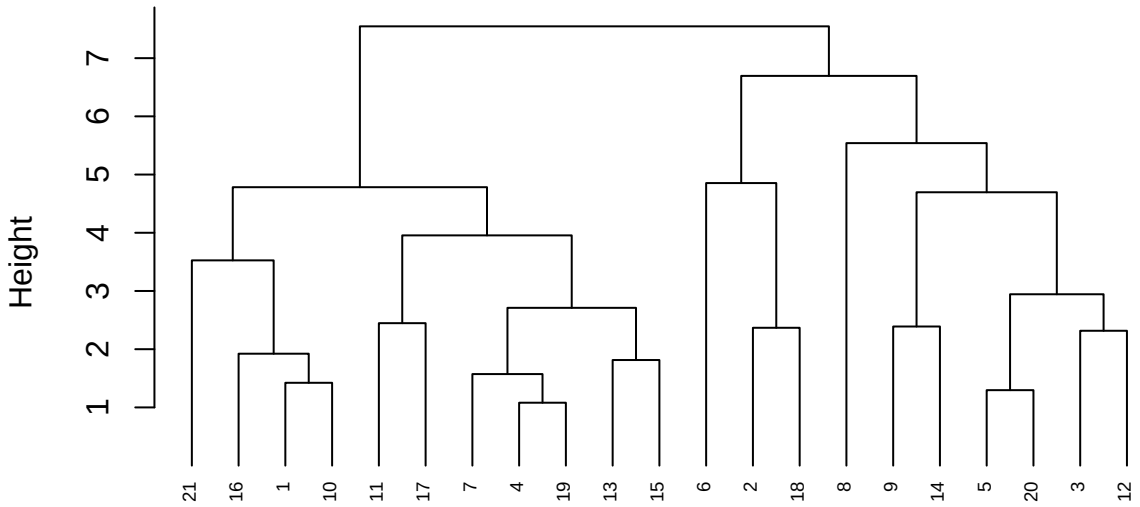
Riba Khan

17/04/2022

```
# Importing the dataset
library(readr)
Cereals <- read_csv("Cereals.csv")
```

```
## Rows: 77 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): name, mfr, type
## dbl (13): calories, protein, fat, sodium, fiber, carbo, sugars, potass, vitamins, shelf, weight,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(Cereals)

#Importing required libraries
library(cluster)
library(caret)
library(dendextend)
library(knitr)
library(factoextra)

# tASK 1
CerealsData <- data.frame(Cereals[,4:16])

#Preprocessing the data
#Remmoving missing values
CerealsData <- na.omit(CerealsData)

#Data Normalization
CerealsDataScale <- scale(CerealsData)

#Applying hierarchical clustering to the data using Euclidean distance
Distance <- dist(CerealsDataScale, method = "euclidean")
#Using complete
HClustering_complete <- hclust(distance, method = "complete")

#Plotting the dendogram
plot(HClustering_complete, cex = 0.6, hang = -1)
```




```
#Using agnes function to compare clustering with
 #single linkage, complete linkage, average linkage and Ward.

#Single
HClusteringSingle <- agnes(CerealsDataScale, method = "single")
#Complete
HClusteringComplete <- agnes(CerealsDataScale,method = "complete")
#Average
HClusteringAverage <- agnes(CerealsDataScale, method = "average")
#Ward
HClusteringWard <- agnes(CerealsDataScale, method = "ward")

#Comparing the agglomerative coefficients for all the above agnes
print(HClusteringSingle$ac)
```

```
## [1] 0.6067859
```

```
print(HClusteringComplete$ac)
```

```
## [1] 0.8353712
```

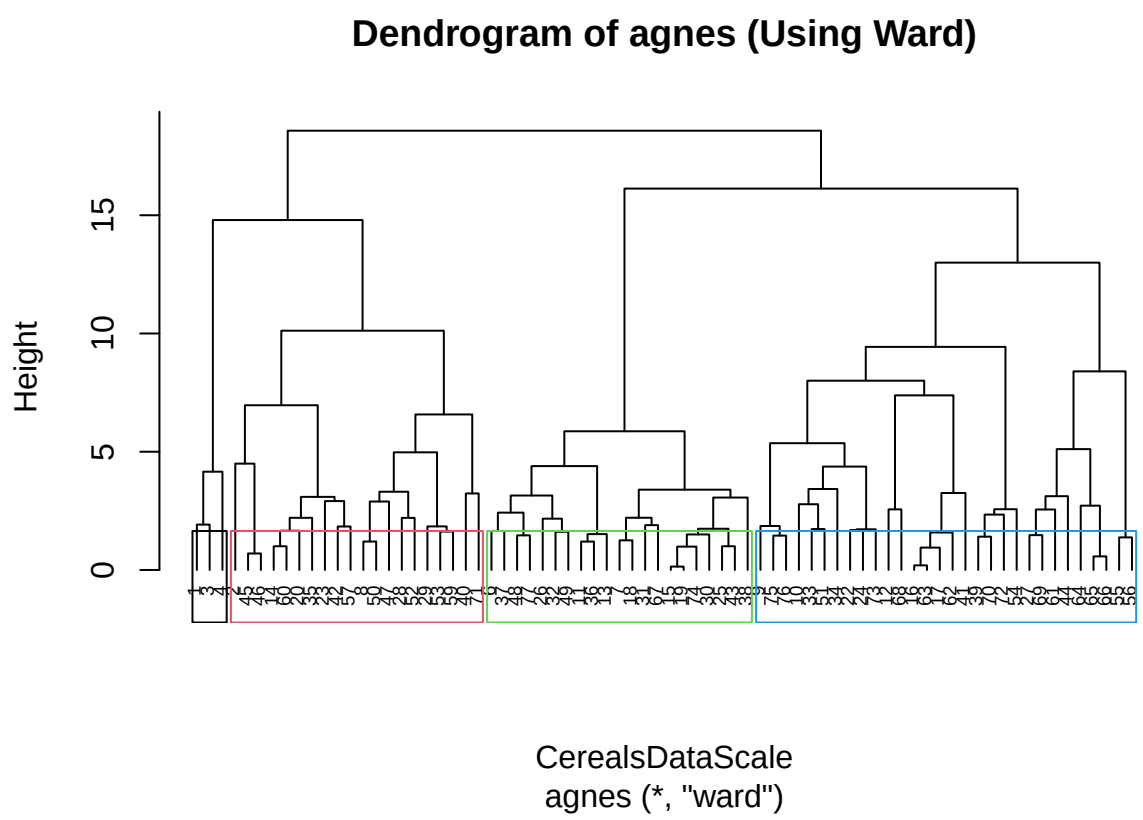
```
print(HClusteringAverage$ac)
```

```
## [1] 0.7766075
```

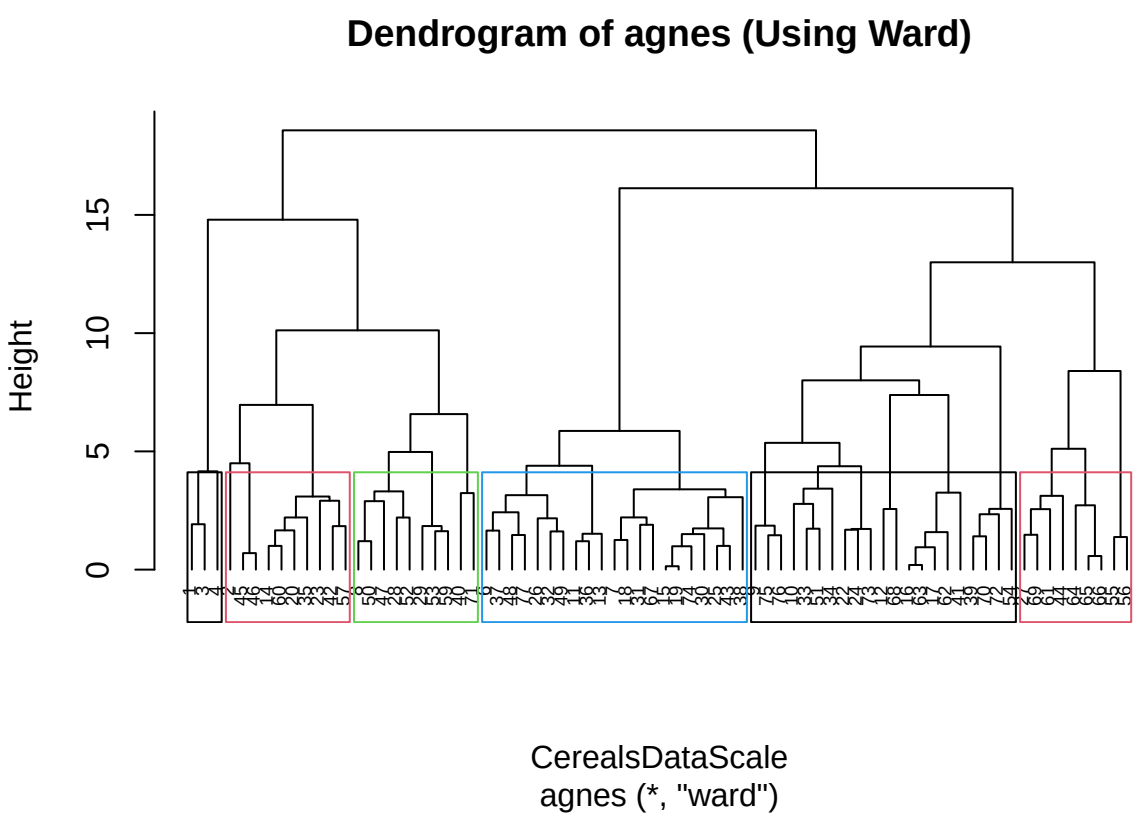
```
print(HClusteringWard$ac)
```

```
## [1] 0.9046042
```

```
#Results show ward agnes is the best method with value of 0.904


# TASK 2
# To choose the number of Clusters

# I will choose k = 4 and k = 6 and then compare the results
#k = 4
pltree(HClusteringWard, cex = 0.6, hang = -1, main = "Dendrogram of agnes (Using Ward)")
rect.hclust(HClusteringWard, k = 4, border = 1:4)
```



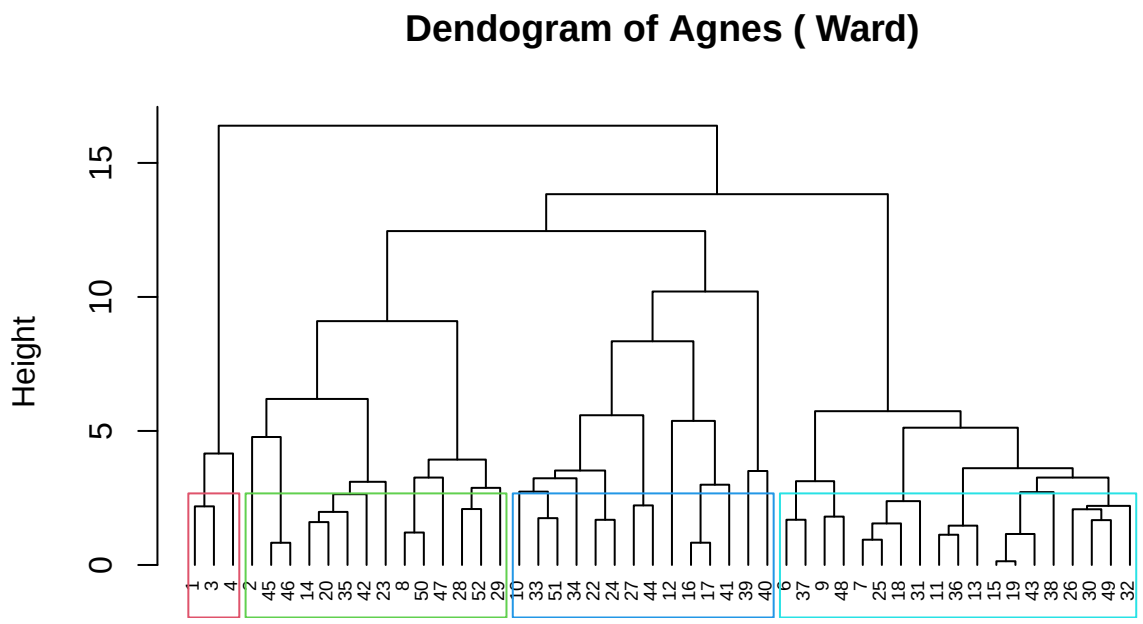

```
Cluster1 <- cutree(HClusteringWard, k=4)
DataFrame1 <- as.data.frame(cbind(CerealsDataScale,Cluster1))

# k = 6
pltree(HClusteringWard, cex = 0.6, hang = -1, main = "Dendrogram of agnes (Using Ward)")
rect.hclust(HClusteringWard, k = 6, border = 1:4)
```




```
Cluster2 <- cutree(HClusteringWard, k=6)
DataFrame2 <- as.data.frame(cbind(CerealsDataScale,Cluster2))

#According to my understanding I would choose k = 4 as the cluster
 #height apppears to be close

# To check the structure of the clusters and on their stability
#Creating Partitions
#set seed
set.seed(123)
Part1 <- CerealsData[1:50,]
Part2 <- CerealsData[51:74,]


#Performing Hierarchial Clustering,consedering k = 4.
Ag_single <- agnes(scale(Part1), method = "single")

Ag_complete <- agnes(scale(Part1), method = "complete")

Ag_average <- agnes(scale(Part1), method = "average")

Ag_ward <- agnes(scale(Part1), method = "ward")

cbind(single=Ag_single$ac , complete=Ag_complete$ac ,
average= Ag_average$ac , ward= Ag_ward$ac)
```

```
##         single  complete   average      ward
## [1,] 0.6393338 0.8138238 0.7408904 0.8764323
```

```
# Creating dendogram of the partitioned data
pltree(Ag_ward, cex = 0.6, hang = -1,
       main = "Dendogram of Agnes ( Ward)")
rect.hclust(Ag_ward, k = 4, border = 2:5)
```




```
c <- cutree(Ag_ward, k = 4)


#Calculating centers to assess the consistency of data.

answer <- as.data.frame(cbind(Part1, c))
answer[answer$c==1,]
```

```
##   calories protein fat sodium fiber carbo sugars potass vitamins shelf weight cups   rating c
## 1       70       4   1    130    10     5      6    280       25     3      1 0.33 68.40297 1
## 3       70       4   1    260     9     7      5    320       25     3      1 0.33 59.42551 1
## 4       50       4   0    140    14     8      0    330       25     3      1 0.50 93.70491 1
```

```
centroid_1 <- colMeans(answer[answer$c==1,])
answer[answer$c==2,]
```

```
##    calories protein fat sodium fiber carbo sugars potass vitamins shelf weight cups   rating c
```

```
## 2       120       3   5     15   2.0   8.0       8    135         0      3   1.00 1.00 33.98368 2
## 8       130       3   2    210   2.0  18.0       8    100        25      3   1.33 0.75 37.03856 2
## 14      110       3   2    140   2.0  13.0       7    105        25      3   1.00 0.50 40.40021 2
## 20      110       3   3    140   4.0  10.0       7    160        25      3   1.00 0.50 40.44877 2
## 23      100       2   1    140   2.0  11.0      10    120        25      3   1.00 0.75 36.17620 2
## 28      120       3   2    160   5.0  12.0      10    200        25      3   1.25 0.67 40.91705 2
## 29      120       3   0    240   5.0  14.0      12    190        25      3   1.33 0.67 41.01549 2
## 35      120       3   3     75   3.0  13.0       4    100        25      3   1.00 0.33 45.81172 2
## 42      100       4   2    150   2.0  12.0       6     95        25      2   1.00 0.67 45.32807 2
## 45      150       4   3     95   3.0  16.0      11    170        25      3   1.00 1.00 37.13686 2
## 46      150       4   3    150   3.0  16.0      11    170        25      3   1.00 1.00 34.13976 2
## 47      160       3   2    150   3.0  17.0      13    160        25      3   1.50 0.67 30.31335 2
## 50      140       3   2    220   3.0  21.0       7    130        25      3   1.33 0.67 40.69232 2
## 52      130       3   2    170   1.5  13.5      10    120        25      3   1.25 0.50 30.45084 2
```

```
centroid_2 <- colMeans(answer[answer$c==2,])
answer[answer$c==3,]
```

```
##    calories protein fat sodium fiber carbo sugars potass vitamins shelf weight cups   rating c
## 6       110       2   2    180   1.5  10.5      10     70        25     1      1 0.75 29.50954 3
## 7       110       2   0    125   1.0  11.0      14     30        25     2      1 1.00 33.17409 3
## 9        90       2   1    200   4.0  15.0       6    125        25     1      1 0.67 49.12025 3
## 11      120       1   2    220   0.0  12.0      12     35        25     2      1 0.75 18.04285 3
## 13      120       1   3    210   0.0  13.0       9     45        25     2      1 0.75 19.82357 3
## 15      110       1   1    180   0.0  12.0      13     55        25     2      1 1.00 22.73645 3
## 18      110       1   0     90   1.0  13.0      12     20        25     2      1 1.00 35.78279 3
## 19      110       1   1    180   0.0  12.0      13     65        25     2      1 1.00 22.39651 3
## 25      110       2   1    125   1.0  11.0      13     30        25     2      1 1.00 32.20758 3
## 26      110       1   0    200   1.0  14.0      11     25        25     1      1 0.75 31.43597 3
## 30      110       1   1    135   0.0  13.0      12     25        25     2      1 0.75 28.02576 3
## 31      100       2   0     45   0.0  11.0      15     40        25     1      1 0.88 35.25244 3
## 32      110       1   1    280   0.0  15.0       9     45        25     2      1 0.75 23.80404 3
## 36      120       1   2    220   1.0  12.0      11     45        25     2      1 1.00 21.87129 3
## 37      110       3   1    250   1.5  11.5      10     90        25     1      1 0.75 31.07222 3
## 38      110       1   0    180   0.0  14.0      11     35        25     1      1 1.33 28.74241 3
## 43      110       2   1    180   0.0  12.0      12     55        25     2      1 1.00 26.73451 3
## 48      100       2   1    220   2.0  15.0       6     90        25     1      1 1.00 40.10596 3
## 49      120       2   1    190   0.0  15.0       9     40        25     2      1 0.67 29.92429 3
```

```
centroid_3 <- colMeans(answer[answer$c==3,])
answer[answer$c==4,]
```

```
##    calories protein fat sodium fiber carbo sugars potass vitamins shelf weight cups   rating c
## 10       90       3   0    210     5    13       5    190        25     3    1.0 0.67 53.31381 4
## 12      110       6   2    290     2    17       1    105        25     1    1.0 1.25 50.76500 4
## 16      110       2   0    280     0    22       3     25        25     1    1.0 1.00 41.44502 4
## 17      100       2   0    290     1    21       2     35        25     1    1.0 1.00 45.86332 4
## 22      110       2   0    220     1    21       3     30        25     3    1.0 1.00 46.89564 4
## 24      100       2   0    190     1    18       5     80        25     3    1.0 0.75 44.33086 4
## 27      100       3   0      0     3    14       7    100        25     2    1.0 0.80 58.34514 4
## 33      100       3   1    140     3    15       5     85        25     3    1.0 0.88 52.07690 4
## 34      110       3   0    170     3    17       3     90        25     3    1.0 0.25 53.37101 4
## 39      110       2   1    170     1    17       6     60       100     3    1.0 1.00 36.52368 4
```

```
## 40      140       3   1    170     2    20       9     95       100     3    1.3 0.75 36.47151 4
## 41      110       2   1    260     0    21       3     40        25     2    1.0 1.50 39.24111 4
## 44      100       4   1      0     0    16       3     95        25     2    1.0 1.00 54.85092 4
## 51       90       3   0    170     3    18       2     90        25     3    1.0 1.00 59.64284 4
```

```
centroid_4 <- colMeans(answer[answer$c==4,])

#binding the four centers
centroids <- rbind(centroid_1, centroid_2, centroid_3, centroid_4)
centers <- as.data.frame(rbind(centroids[,-14], Part2))

#Calculating the Distance
D1 <- get_dist(centers)
Matrix1 <- as.matrix(D1)
df1 <- data.frame(data=seq(1,nrow(Part2),1), Clusters = rep(0,nrow(Part2)))
for(i in 1:nrow(Part2))
{df1[i,2] <- which.min(Matrix1[i+4, 1:4])}
df1
```

```
##    data Clusters
## 1     1        1
## 2     2        4
## 3     3        3
## 4     4        2
## 5     5        2
## 6     6        1
## 7     7        2
## 8     8        2
## 9     9        3
## 10   10        3
## 11   11        2
## 12   12        2
## 13   13        2
## 14   14        3
## 15   15        4
## 16   16        2
## 17   17        3
## 18   18        2
## 19   19        4
## 20   20        4
## 21   21        3
## 22   22        4
## 23   23        4
## 24   24        3
```

```
# Task 3
# Determing the healthiest cluster

# I am selecting that is best cereal for breakfast which will contain
 # low sugar and sodium and high protien and fiber
Healthy_Cluster <- Cereals
Healthy_Cluster_na <- na.omit(Healthy_Cluster)
Clust <- cbind(Healthy_Cluster_na, Cluster1)
Clust[Clust$Cluster1==1,]
```

```
##                        name mfr type calories protein fat sodium fiber carbo sugars potass vitamins
## 1                 100%_Bran   N    C       70       4   1    130    10     5      6    280       25
## 3                  All-Bran   K    C       70       4   1    260     9     7      5    320       25
## 4 All-Bran_with_Extra_Fiber   K    C       50       4   0    140    14     8      0    330       25
##   shelf weight cups   rating Cluster1
## 1     3      1 0.33 68.40297        1
## 3     3      1 0.33 59.42551        1
## 4     3      1 0.50 93.70491        1
```

```
Clust[Clust$Cluster1==2,]
```

```
##                                    name mfr type calories protein fat sodium fiber carbo sugars
## 2                     100%_Natural_Bran   Q    C      120       3   5     15   2.0   8.0       8
## 8                               Basic_4   G    C      130       3   2    210   2.0  18.0       8
## 14                             Clusters   G    C      110       3   2    140   2.0  13.0       7
## 20                   Cracklin'_Oat_Bran   K    C      110       3   3    140   4.0  10.0       7
## 23                 Crispy_Wheat_&_Raisins   G    C      100       2   1    140   2.0  11.0      10
## 28 Fruit_&_Fibre_Dates,_Walnuts,_and_Oats   P    C      120       3   2    160   5.0  12.0      10
## 29                          Fruitful_Bran   K    C      120       3   0    240   5.0  14.0      12
## 35                   Great_Grains_Pecan   P    C      120       3   3     75   3.0  13.0       4
## 40                Just_Right_Fruit_&_Nut   K    C      140       3   1    170   2.0  20.0       9
## 42                                  Life   Q    C      100       4   2    150   2.0  12.0       6
## 45       Muesli_Raisins,_Dates,_&_Almonds   R    C      150       4   3     95   3.0  16.0      11
## 46      Muesli_Raisins,_Peaches,_&_Pecans   R    C      150       4   3    150   3.0  16.0      11
## 47                   Mueslix_Crispy_Blend   K    C      160       3   2    150   3.0  17.0      13
## 50              Nutri-Grain_Almond-Raisin   K    C      140       3   2    220   3.0  21.0       7
## 52                   Oatmeal_Raisin_Crisp   G    C      130       3   2    170   1.5  13.5      10
## 53                 Post_Nat._Raisin_Bran   P    C      120       3   1    200   6.0  11.0      14
## 57                     Quaker_Oat_Squares   Q    C      100       4   1    135   2.0  14.0       6
## 59                           Raisin_Bran   K    C      120       3   1    210   5.0  14.0      12
## 60                      Raisin_Nut_Bran   G    C      100       3   2    140   2.5  10.5       8
## 71                    Total_Raisin_Bran   G    C      140       3   1    190   4.0  15.0      14
##    potass vitamins shelf weight cups   rating Cluster1
## 2     135        0     3   1.00 1.00 33.98368        2
## 8     100       25     3   1.33 0.75 37.03856        2
## 14    105       25     3   1.00 0.50 40.40021        2
## 20    160       25     3   1.00 0.50 40.44877        2
## 23    120       25     3   1.00 0.75 36.17620        2
## 28    200       25     3   1.25 0.67 40.91705        2
## 29    190       25     3   1.33 0.67 41.01549        2
## 35    100       25     3   1.00 0.33 45.81172        2
## 40     95      100     3   1.30 0.75 36.47151        2
## 42     95       25     2   1.00 0.67 45.32807        2
## 45    170       25     3   1.00 1.00 37.13686        2
## 46    170       25     3   1.00 1.00 34.13976        2
## 47    160       25     3   1.50 0.67 30.31335        2
## 50    130       25     3   1.33 0.67 40.69232        2
## 52    120       25     3   1.25 0.50 30.45084        2
## 53    260       25     3   1.33 0.67 37.84059        2
## 57    110       25     3   1.00 0.50 49.51187        2
## 59    240       25     2   1.33 0.75 39.25920        2
## 60    140       25     3   1.00 0.50 39.70340        2
## 71    230      100     3   1.50 1.00 28.59278        2
```

```
Clust[Clust$Cluster1==3,]
```

```
##                      name mfr type calories protein fat sodium fiber carbo sugars potass vitamins
## 6  Apple_Cinnamon_Cheerios   G    C      110       2   2    180   1.5  10.5     10     70       25
## 7               Apple_Jacks   K    C      110       2   0    125   1.0  11.0     14     30       25
## 11            Cap'n'Crunch   Q    C      120       1   2    220   0.0  12.0     12     35       25
## 13   Cinnamon_Toast_Crunch   G    C      120       1   3    210   0.0  13.0      9     45       25
## 15             Cocoa_Puffs   G    C      110       1   1    180   0.0  12.0     13     55       25
## 18               Corn_Pops   K    C      110       1   0     90   1.0  13.0     12     20       25
## 19           Count_Chocula   G    C      110       1   1    180   0.0  12.0     13     65       25
## 25             Froot_Loops   K    C      110       2   1    125   1.0  11.0     13     30       25
## 26          Frosted_Flakes   K    C      110       1   0    200   1.0  14.0     11     25       25
## 30          Fruity_Pebbles   P    C      110       1   1    135   0.0  13.0     12     25       25
## 31            Golden_Crisp   P    C      100       2   0     45   0.0  11.0     15     40       25
## 32          Golden_Grahams   G    C      110       1   1    280   0.0  15.0      9     45       25
## 36        Honey_Graham_Ohs   Q    C      120       1   2    220   1.0  12.0     11     45       25
## 37      Honey_Nut_Cheerios   G    C      110       3   1    250   1.5  11.5     10     90       25
## 38              Honey-comb   P    C      110       1   0    180   0.0  14.0     11     35       25
## 43            Lucky_Charms   G    C      110       2   1    180   0.0  12.0     12     55       25
## 48    Multi-Grain_Cheerios   G    C      100       2   1    220   2.0  15.0      6     90       25
## 49        Nut&Honey_Crunch   K    C      120       2   1    190   0.0  15.0      9     40       25
## 67                  Smacks   K    C      110       2   1     70   1.0   9.0     15     40       25
## 74                    Trix   G    C      110       1   1    140   0.0  13.0     12     25       25
## 77     Wheaties_Honey_Gold   G    C      110       2   1    200   1.0  16.0      8     60       25
##    shelf weight cups   rating Cluster1
## 6      1      1 0.75 29.50954        3
## 7      2      1 1.00 33.17409        3
## 11     2      1 0.75 18.04285        3
## 13     2      1 0.75 19.82357        3
## 15     2      1 1.00 22.73645        3
## 18     2      1 1.00 35.78279        3
## 19     2      1 1.00 22.39651        3
## 25     2      1 1.00 32.20758        3
## 26     1      1 0.75 31.43597        3
## 30     2      1 0.75 28.02576        3
## 31     1      1 0.88 35.25244        3
## 32     2      1 0.75 23.80404        3
## 36     2      1 1.00 21.87129        3
## 37     1      1 0.75 31.07222        3
## 38     1      1 1.33 28.74241        3
## 43     2      1 1.00 26.73451        3
## 48     1      1 1.00 40.10596        3
## 49     2      1 0.67 29.92429        3
## 67     2      1 0.75 31.23005        3
## 74     2      1 1.00 27.75330        3
## 77     1      1 0.75 36.18756        3
```

```
Clust[Clust$Cluster1==4,]
```

```
##                        name mfr type calories protein fat sodium fiber carbo sugars potass
## 9                 Bran_Chex   R    C       90       2   1    200     4    15      6    125
## 10              Bran_Flakes   P    C       90       3   0    210     5    13      5    190
## 12                 Cheerios   G    C      110       6   2    290     2    17      1    105
```

```
## 16                  Corn_Chex   R    C      110       2   0    280     0    22      3     25
## 17                Corn_Flakes   K    C      100       2   0    290     1    21      2     35
## 22                    Crispix   K    C      110       2   0    220     1    21      3     30
## 24                Double_Chex   R    C      100       2   0    190     1    18      5     80
## 27         Frosted_Mini-Wheats   K    C      100       3   0      0     3    14      7    100
## 33           Grape_Nuts_Flakes   P    C      100       3   1    140     3    15      5     85
## 34                  Grape-Nuts   P    C      110       3   0    170     3    17      3     90
## 39 Just_Right_Crunchy__Nuggets   K    C      110       2   1    170     1    17      6     60
## 41                         Kix   G    C      110       2   1    260     0    21      3     40
## 44                       Maypo   A    H      100       4   1      0     0    16      3     95
## 51           Nutri-grain_Wheat   K    C       90       3   0    170     3    18      2     90
## 54                 Product_19   K    C      100       3   0    320     1    20      3     45
## 55                Puffed_Rice   Q    C       50       1   0      0     0    13      0     15
## 56               Puffed_Wheat   Q    C       50       2   0      0     1    10      0     50
## 61              Raisin_Squares   K    C       90       2   0      0     2    15      6    110
## 62                   Rice_Chex   R    C      110       1   0    240     0    23      2     30
## 63               Rice_Krispies   K    C      110       2   0    290     0    22      3     35
## 64              Shredded_Wheat   N    C       80       2   0      0     3    16      0     95
## 65      Shredded_Wheat_'n'Bran   N    C       90       3   0      0     4    19      0    140
## 66   Shredded_Wheat_spoon_size   N    C       90       3   0      0     3    20      0    120
## 68                 Special_K   K    C      110       6   0    230     1    16      3     55
## 69     Strawberry_Fruit_Wheats   N    C       90       2   0     15     3    15      5     90
## 70           Total_Corn_Flakes   G    C      110       2   1    200     0    21      3     35
## 72           Total_Whole_Grain   G    C      100       3   1    200     3    16      3    110
## 73                     Triples   G    C      110       2   1    250     0    21      3     60
## 75                  Wheat_Chex   R    C      100       3   1    230     3    17      3    115
## 76                    Wheaties   G    C      100       3   1    200     3    17      3    110
##    vitamins shelf weight cups   rating Cluster1
## 9        25     1   1.00 0.67 49.12025        4
## 10       25     3   1.00 0.67 53.31381        4
## 12       25     1   1.00 1.25 50.76500        4
## 16       25     1   1.00 1.00 41.44502        4
## 17       25     1   1.00 1.00 45.86332        4
## 22       25     3   1.00 1.00 46.89564        4
## 24       25     3   1.00 0.75 44.33086        4
## 27       25     2   1.00 0.80 58.34514        4
## 33       25     3   1.00 0.88 52.07690        4
## 34       25     3   1.00 0.25 53.37101        4
## 39      100     3   1.00 1.00 36.52368        4
## 41       25     2   1.00 1.50 39.24111        4
## 44       25     2   1.00 1.00 54.85092        4
## 51       25     3   1.00 1.00 59.64284        4
## 54      100     3   1.00 1.00 41.50354        4
## 55        0     3   0.50 1.00 60.75611        4
## 56        0     3   0.50 1.00 63.00565        4
## 61       25     3   1.00 0.50 55.33314        4
## 62       25     1   1.00 1.13 41.99893        4
## 63       25     1   1.00 1.00 40.56016        4
## 64        0     1   0.83 1.00 68.23588        4
## 65        0     1   1.00 0.67 74.47295        4
## 66        0     1   1.00 0.67 72.80179        4
## 68       25     1   1.00 1.00 53.13132        4
## 69       25     2   1.00 1.00 59.36399        4
## 70      100     3   1.00 1.00 38.83975        4
```

```
## 72      100     3   1.00 1.00 46.65884        4
## 73       25     3   1.00 0.75 39.10617        4
## 75       25     1   1.00 0.67 49.78744        4
## 76       25     1   1.00 1.00 51.59219        4
```

```
# CalculatingMean ratings to determine the best cluster.
mean(Clust[Clust$Cluster1==1,"rating"])
```

```
## [1] 73.84446
```

```
mean(Clust[Clust$Cluster1==2,"rating"])
```

```
## [1] 38.26161
```

```
mean(Clust[Clust$Cluster1==3,"rating"])
```

```
## [1] 28.84825
```

```
mean(Clust[Clust$Cluster1==4,"rating"])
```

```
## [1] 51.43111
```

```
# Cluster 1 is the healthiest because of the mean rating being
 # the highest i.e 73.84 , hence we choose cluster 1
```