Hierarchical clustering

1. overview of dataset

The dataset describes relationship between customers with their buying behaviours. dataset selects 60 customers with their sex, age, visit times and average expense.

```
> ###data2
> rm(list = ls())
> data2 <- read.table("C:/Users/Hao/Desktop/2018fall/6231/presentation/data2.txt", header = T)
> dim(data2)
[1] 60  5
> attach(data2)
The following objects are masked from data2 (pos = 3):

    Age, Average.Expense, ID, Sex, Visit.Time

The following objects are masked from data2 (pos = 5):

    Age, Average.Expense, ID, Sex, Visit.Time

> head(data2,5)
  ID Visit.Time Average.Expense Sex Age
1  1          3             5.7   0  10
2  2          5            14.5   0  27
3  3         16            33.5   0  32
4  4          5            15.9   0  30
5  5         16            24.9   0  23
> str(data2)
'data.frame':   60 obs. of  5 variables:
 $ ID             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Visit.Time     : int  3 5 16 5 16 3 12 14 6 3 ...
 $ Average.Expense: num  5.7 14.5 33.5 15.9 24.9 12 28.5 18.8 23.8 5.3 ...
 $ Sex            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Age            : int  10 27 32 30 23 15 33 27 16 11 ...
```

From code above, it is obvious that we have 60 pieces of record with 5 dimensions. 'visit.time' describes the frequency of a customer come to the shop. 'average expense' describes the consumption capacity. 'Age' is a binary variable which 0 represents female and 1 represents male. The other 4 variables are all numerical.

2. preparation on data

Before clustering, we need to check the scale of variables. In this data, we have both categorical variable (age) and numerical variables. values in different variables vary differently. some variables have large range and some have narrow range. In order to eliminate the effect of dimension, we need to uniform dimension first. In R, function scale() will help us process the standardization.

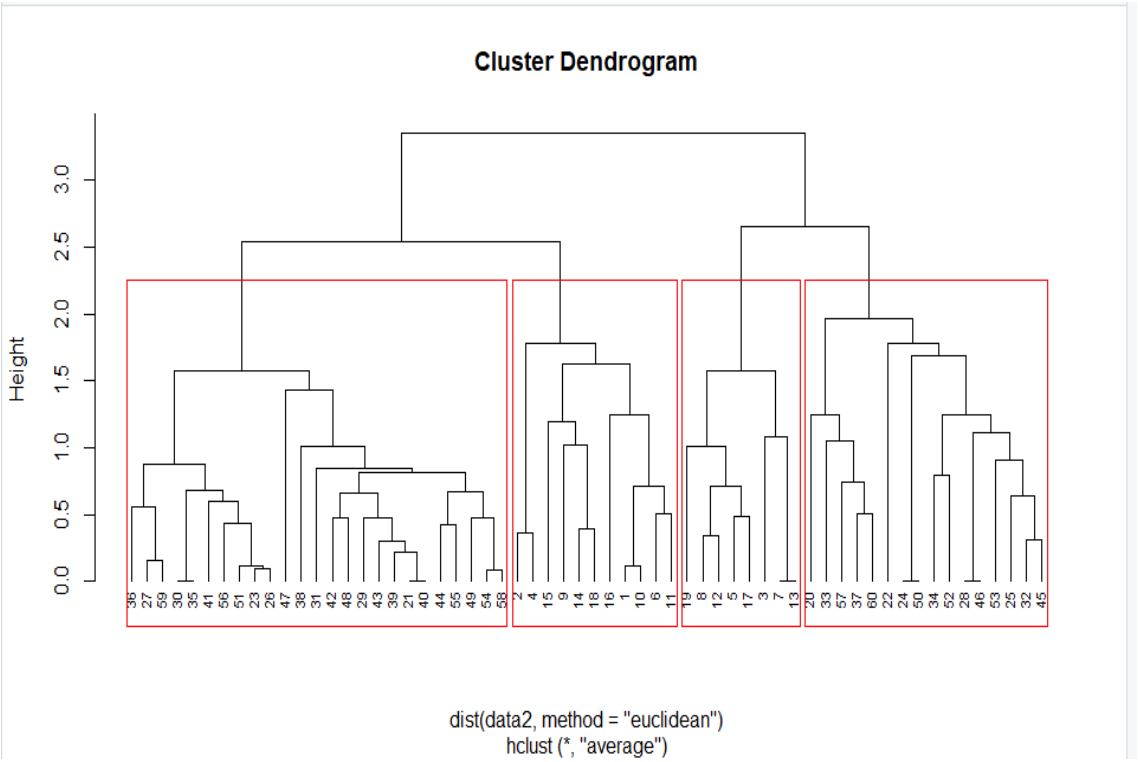
```
> data2 = scale(data2[,-1])
> data2
       Visit.Time Average.Expense        Sex         Age
 [1,] -1.20219054     -1.35237652 -1.4566845 -1.23134396
 [2,] -0.75693479     -0.30460718 -1.4566845  0.59951732
 [3,]  1.69197187      1.95762206 -1.4566845  1.13800594
 [4,] -0.75693479     -0.13791661 -1.4566845  0.92261049
 [5,]  1.69197187      0.93366567 -1.4566845  0.16872643
 [6,] -1.20219054     -0.60226893 -1.4566845 -0.69285535
 [7,]  0.80146036      1.36229858 -1.4566845  1.24570366
 [8,]  1.24671612      0.20737101 -1.4566845  0.59951732
 [9,] -0.53430691      0.80269450 -1.4566845 -0.58515763
[10,] -1.20219054     -1.40000240 -1.4566845 -1.12364624
[11,] -0.97956266     -1.00708890 -1.4566845 -0.90825080
[12,]  1.24671612      0.46931335 -1.4566845  0.38412188
```

1.3 Agglomerativing clustering and dendrogram

```
> hc2 = hclust(dist(data2,method = "euclidean"),method = "average")
> hc2

Call:
hclust(d = dist(data2, method = "euclidean"), method = "average")

Cluster method   : average
Distance         : euclidean
Number of objects: 60

> plot(hc2,hang = -0.01,cex =0.7)
> re2 <- rect.hclust(hc2, k = 4)
> re2
[[1]]
 [1] 21 23 26 27 29 30 31 35 36 38 39 40 41 42 43 44 47 48 49 51 54 55 56 58 59

[[2]]
 [1]  1  2  4  6  9 10 11 14 15 16 18

[[3]]
[1]  3  5  7  8 12 13 17 19

[[4]]
 [1] 20 22 24 25 28 32 33 34 37 45 46 50 52 53 57 60
```

hclust(x, method) function process hierarchical clustering. the 'method' command in dist() defines the dissimilarity measurement between observations is 'euclidean distance' and 'method' command in hclust() defines the distance between clusters is 'average linkage'.

rec.hclust(x, k) function divides the data into four cluster and circles out different clusters with rectangles.


Cluster Dendrogram
Height
0.0
0.5
1.0
1.5
2.0
2.5
3.0
dist(data2, method = "euclidean")
hclust (*, "average")