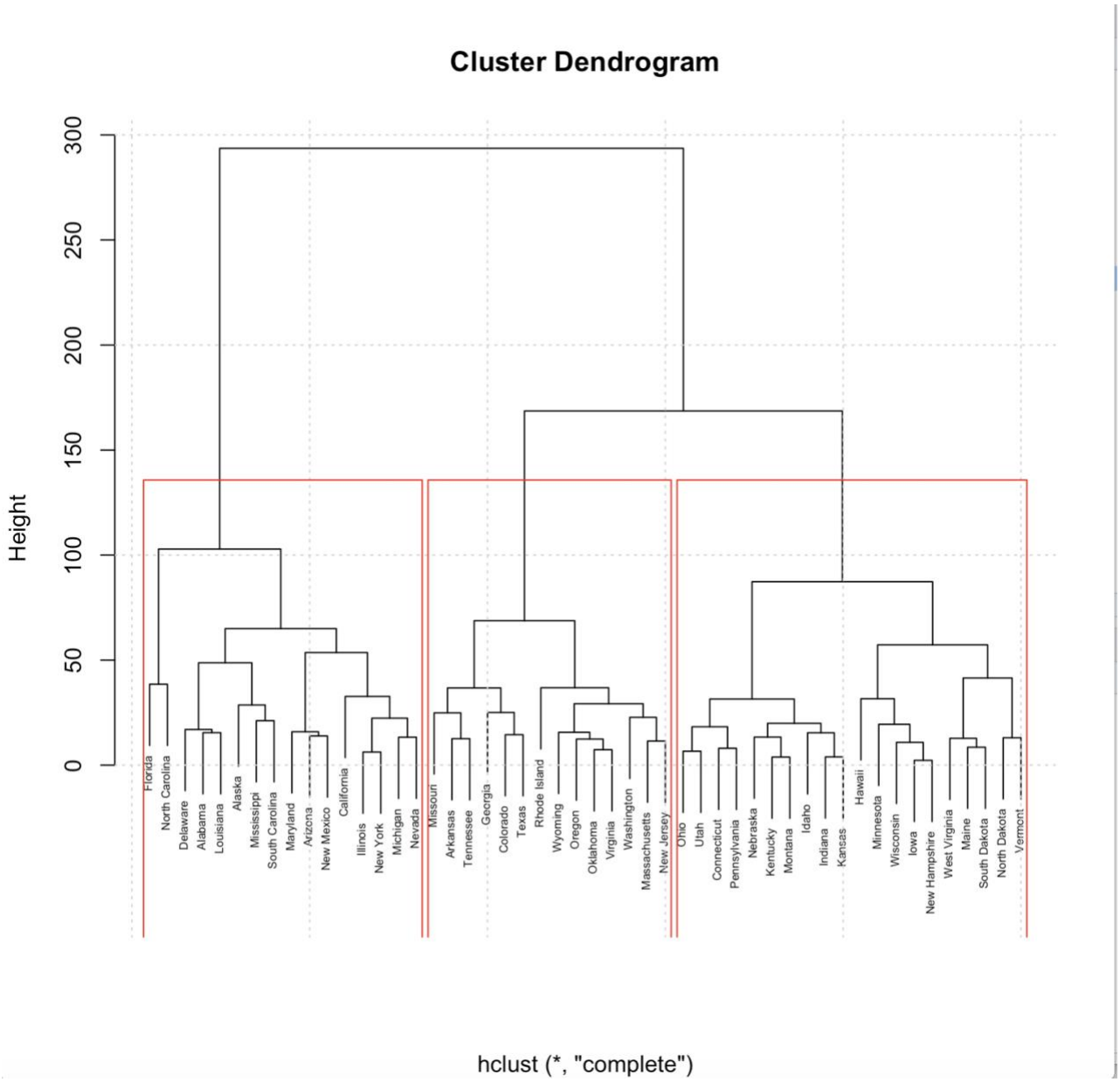
Shuting Chen

Final

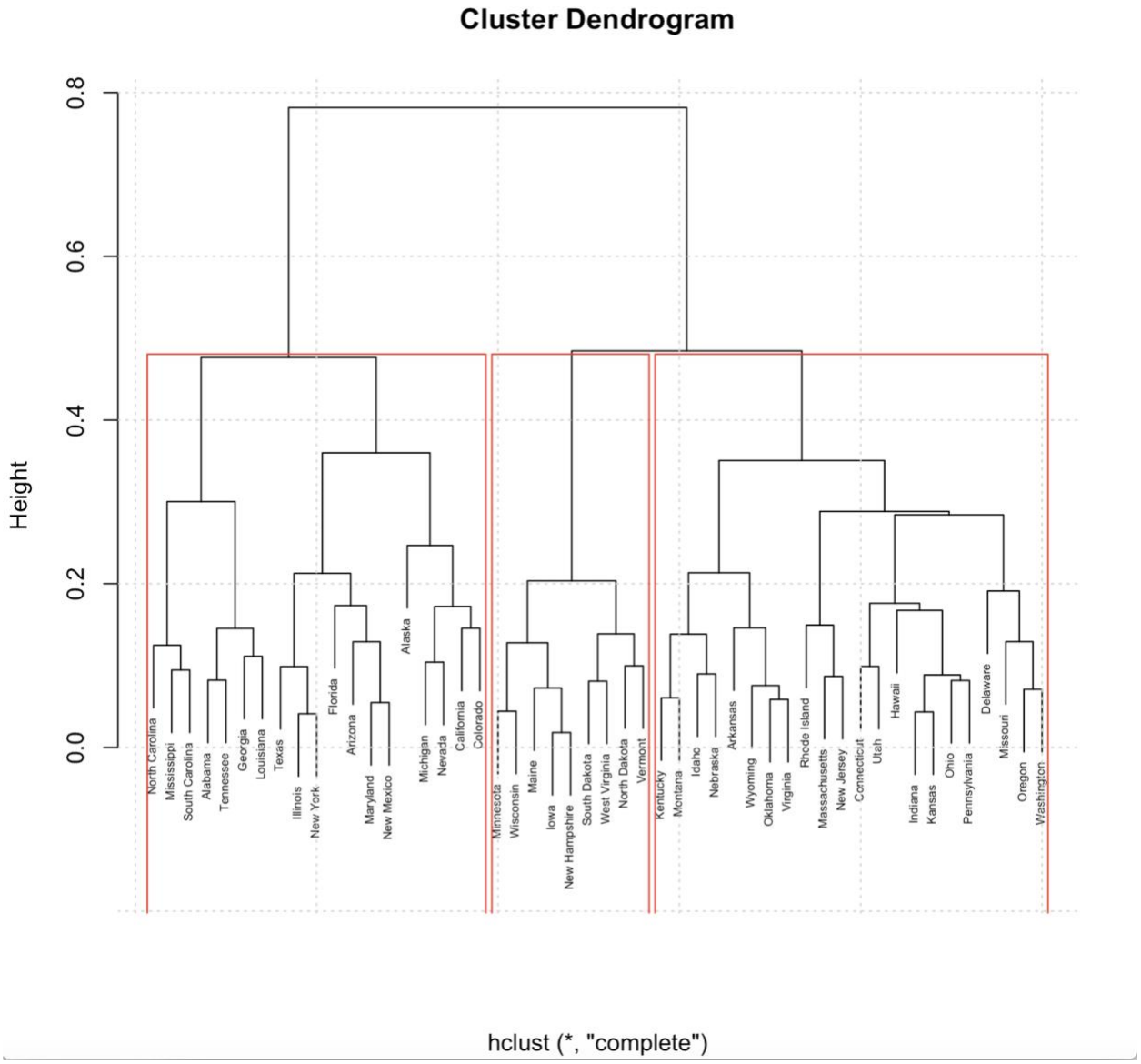
```
> #1 a)
> d0=data.frame(USArrests)
> d1=daisy(d0)
> seg.hc1=hclust(d1,method = "complete")
> plot(seg.hc1,cex=0.5,xlab = "")
> grid()
>
> # b)
> cut1=rect.hclust(seg.hc1,k=3,border = "red")
> seg.hc1.segment=cutree(seg.hc1,k=3)
```



#from cut1 we can see which states belong to which clusters

```
> cut1
[[1]]
     Alabama       Alaska      Arizona   California     Delaware      Florida     Illinois
           1            2            3            5            8            9           13
   Louisiana     Maryland     Michigan  Mississippi       Nevada   New Mexico     New York
          18           20           22           24           28           31           32
North Carolina South Carolina
          33           40

[[2]]
    Arkansas     Colorado      Georgia Massachusetts     Missouri    New Jersey      Oklahoma
```

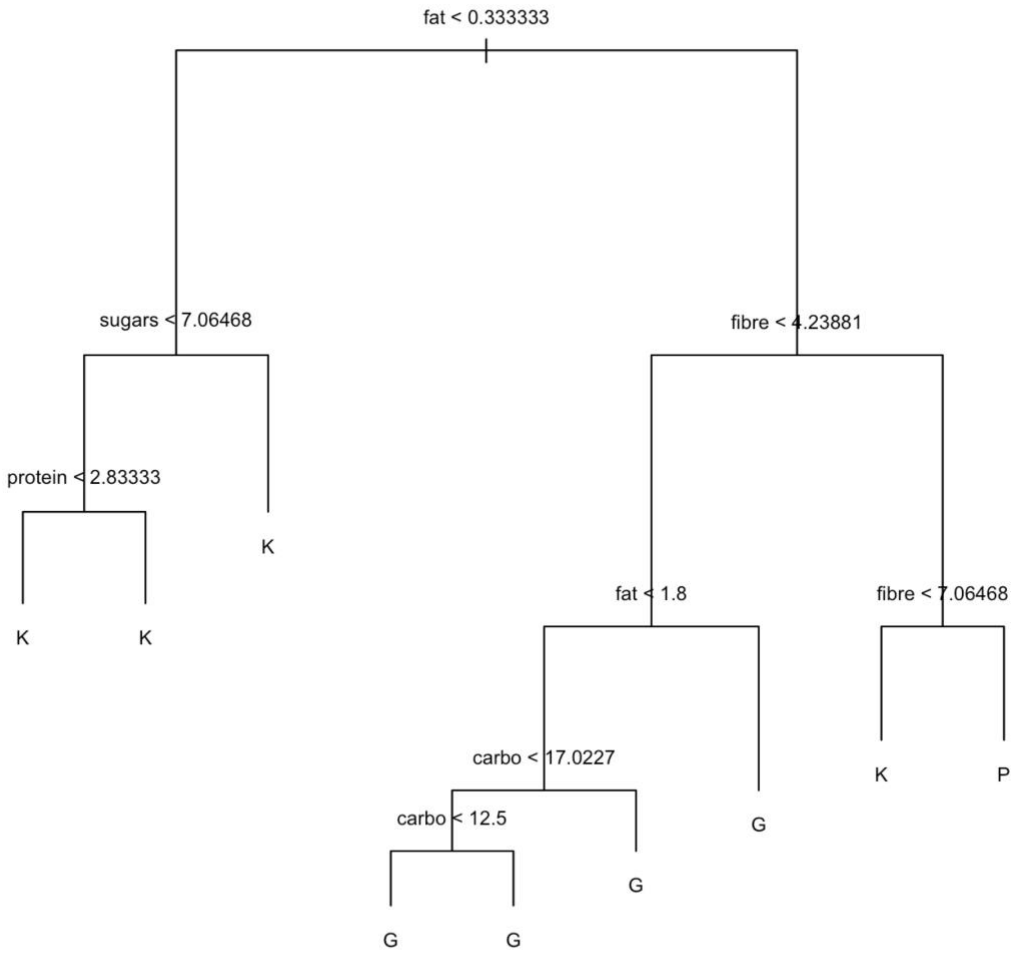
```
         4          6         10         21         25         30         36
    Oregon Rhode Island  Tennessee      Texas   Virginia Washington    Wyoming
        37         39         42         43         46         47         50

[[3]]
 Connecticut      Hawaii       Idaho     Indiana        Iowa      Kansas    Kentucky
         7         11         12         14         15         16         17
       Maine   Minnesota     Montana    Nebraska New Hampshire North Dakota        Ohio
        19         23         26         27         29         34         35
Pennsylvania South Dakota       Utah     Vermont West Virginia   Wisconsin
        38         41         44         45         48         49

>
>
> # c)
> d2=daisy(d0,metric = "gower")
> seg.hc2=hclust(d2,method = "complete")
> plot(seg.hc2,cex=0.5,xlab = "")
> grid()
> cut2=rect.hclust(seg.hc2,k=3,border = "red")
```




```
> seg.hc2.segment=cutree(seg.hc2,k=3)
> cut2
[[1]]
     Alabama       Alaska      Arizona   California     Colorado      Florida      Georgia
           1            2            3            5            6            9           10
    Illinois    Louisiana     Maryland     Michigan  Mississippi       Nevada   New Mexico
          13           18           20           22           24           28           31
    New York North Carolina South Carolina    Tennessee        Texas
          32           33           40           42           43

[[2]]
        Iowa        Maine    Minnesota New Hampshire  North Dakota  South Dakota      Vermont
          15           19           23           29           34           41           45
West Virginia    Wisconsin
```

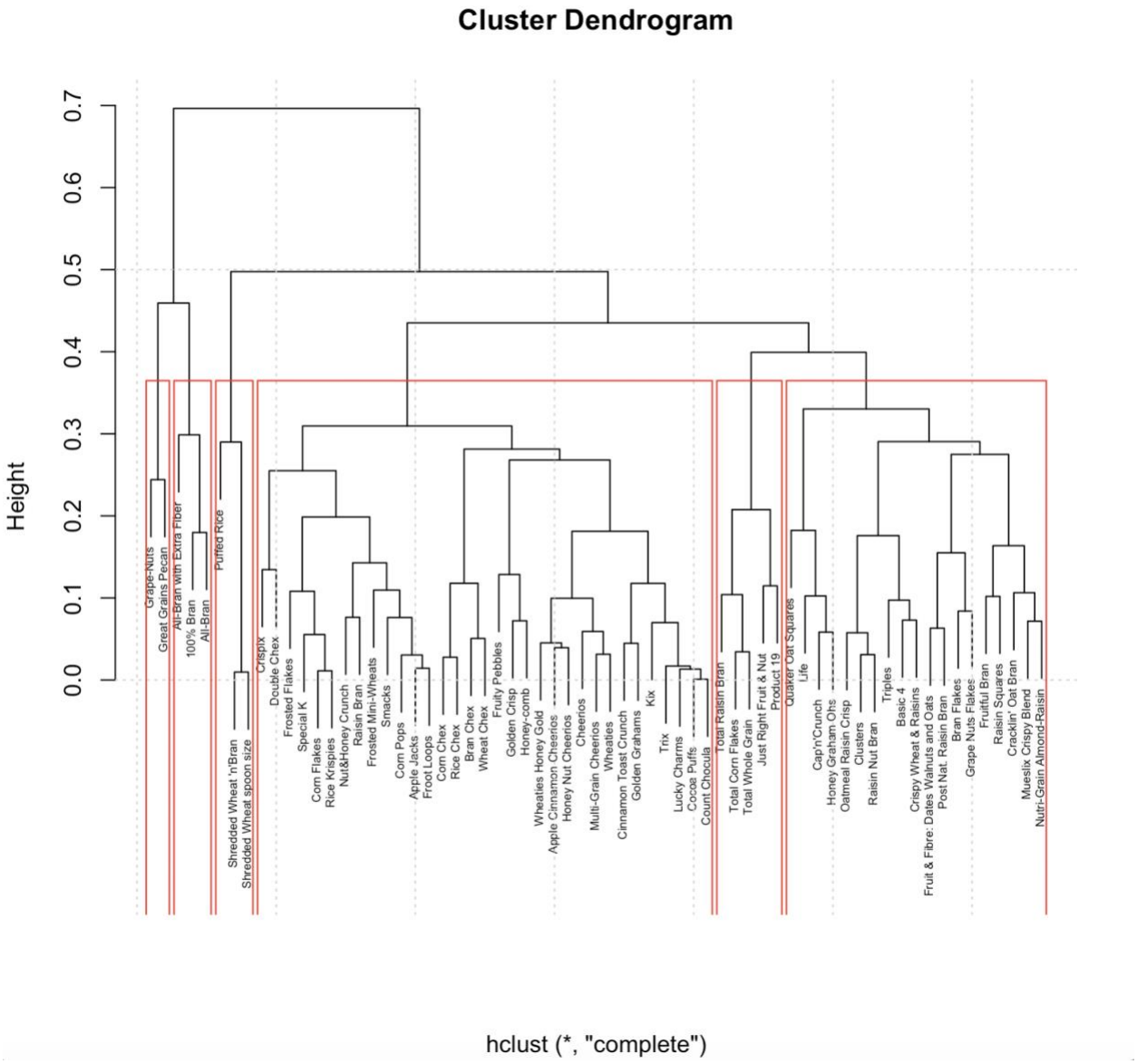
```
        48          49

[[3]]
    Arkansas   Connecticut      Delaware       Hawaii        Idaho      Indiana       Kansas
           4           7           8          11           12           14           16
    Kentucky Massachusetts      Missouri       Montana     Nebraska    New Jersey         Ohio
          17          21          25          26           27           30           35
    Oklahoma       Oregon  Pennsylvania  Rhode Island         Utah     Virginia    Washington
          36          37          38          39           44           46           47
     Wyoming
          50

>
> table(seg.hc1.segment, seg.hc2.segment)
             seg.hc2.segment
seg.hc1.segment  1  2  3  4  5  6
              1  3  0  0  0  0  0
              2  0 33  0  0  0  0
              3  0  0 19  0  0  0
              4  0  0  0  2  0  0
              5  0  0  0  0  5  0
              6  0  0  0  0  0  3
>
> #as table show above we don't get the same clusters so it effects the hierarchical clustering.
>
> #see some similarity of  Illinois 13 and New York 32
> #two merge very early in cluster 1 from cluster dendrogram using not scaling data
> d0[c(13,32),]
         Murder Assault UrbanPop Rape
Illinois   10.4     249       83 24.0
New York   11.1     254       86 26.1
> #see some similarity of Iowa 15 and New Hampshire 29
> #two merge very early in cluster 1 from cluster dendrogram using scaling data
> d0[c(15,29),]
             Murder Assault UrbanPop Rape
Iowa            2.2      56       57 11.3
New Hampshire   2.1      57       56  9.5
> #the difference of most similiar one in each group from cluster dendrogram
> #using not scaling data is bigger than scaling data
>
> #so the data should be scaled before the inter-observation dissimilarities are computed
> #since different variables are measure in different units
>
>
> #2 a)
> library(MASS)
> library(tree)
> d0=data.frame(UScereal)
```

```
> str(d0)
'data.frame':	65 obs. of  11 variables:
 $ mfr      : Factor w/ 6 levels "G","K","N","P",..: 3 2 2 1 2 1 6 4 5 1 ...
 $ calories : num  212 212 100 147 110 ...
 $ protein  : num  12.12 12.12 8 2.67 2 ...
 $ fat      : num  3.03 3.03 0 2.67 0 ...
 $ sodium   : num  394 788 280 240 125 ...
 $ fibre    : num  30.3 27.3 28 2 1 ...
 $ carbo    : num  15.2 21.2 16 14 11 ...
 $ sugars   : num  18.2 15.2 0 13.3 14 ...
 $ shelf    : int  3 3 3 1 2 3 1 3 2 1 ...
 $ potassium: num  848.5 969.7 660 93.3 30 ...
 $ vitamins : Factor w/ 3 levels "100%","enriched",..: 2 2 2 2 2 2 2 2 2 2 ...
> #method1
> #to test whether we can classify the manufacture by using classification tree method
> tree0=tree(mfr~.,d0)
> plot(tree0)
> text(tree0,cex=0.75)
> summary(tree0)
```



Classification tree:
tree(formula = mfr ~ ., data = d0)
Variables actually used in tree construction:
[1] "fat" "sugars" "protein" "fibre" "carbo"
Number of terminal nodes: 9
Residual mean deviance: 1.683 = 94.25 / 56
Misclassification error rate: 0.4 = 26 / 65
> #Variables actually used in tree construction: "fat" "sugars" "protein" "fibre" "carbo"
> #but as we can see from the cereal characteristics discriminate the K,G,P manufacturers are fat, fibre
> #other 3 manufacturers do not shown in this tree
> #so in general, we can conclude that they each have a balanced portfolio of cereals
>
> #method2

```
> #use cluster dendrogram method to cluster the cereals
> #and see whether different cluster are mainly produced by same manufacture
> d1=daisy(d0)
> seg.hc1=hclust(d1,method = "complete")
> str(seg.hc1)
List of 7
 $ merge      : int [1:64, 1:2] -13 -54 -15 -39 -5 -62 -14 -16 -12 -41 ...
 $ height     : num [1:64] 0.000952 0.009681 0.011263 0.013299 0.014351 ...
 $ order      : int [1:65] 31 32 3 1 2 47 54 55 19 21 ...
 $ labels     : chr [1:65] "100% Bran" "All-Bran" "All-Bran with Extra Fiber" "Apple Cinnamon Cheerios" ...
 $ method     : chr "complete"
 $ call       : language hclust(d = d1, method = "complete")
 $ dist.method: NULL
 - attr(*, "class")= chr "hclust"
> plot(seg.hc1,cex=0.5,xlab = "")
> grid()
> cut1=rect.hclust(seg.hc1,k=6,border = "red")
```




```
> seg.hc1.segment=cutree(seg.hc1,k=6)
> table(d0$mfr)

 G  K  N  P  Q  R
22 21  3  9  5  5
> (22+21+9)/65
[1] 0.8
> #as shown in table, the major manufactures are G, K and P, they produce 80% of cereals
> table(seg.hc1.segment,d0$mfr)

seg.hc1.segment  G  K  N  P  Q  R
              1  0  2  1  0  0  0
              2 13 12  0  3  0  5
              3  6  5  0  4  4  0
```

```
        4 0 0 0 2 0 0
        5 3 2 0 0 0 0
        6 0 0 2 0 1 0
> #analyze from that table, most of cereals manufactures G, K and P produce are in cluster 2 and 3
> #the small difference is that G and K are also produce few cluster 5 cereals(which is only produce by
them)
> #and P is the only manufacture produce 2 cereals in cluster 4
> #since major manufacturers G, K and P are all produce cereals in cluster 2 and 3,
> round(prop.table(table(seg.hc1.segment,d0$mfr),2),3)

seg.hc1.segment     G     K     N     P     Q     R
        1 0.000 0.095 0.333 0.000 0.000 0.000
        2 0.591 0.571 0.000 0.333 0.000 1.000
        3 0.273 0.238 0.000 0.444 0.800 0.000
        4 0.000 0.000 0.000 0.222 0.000 0.000
        5 0.136 0.095 0.000 0.000 0.000 0.000
        6 0.000 0.000 0.667 0.000 0.200 0.000
> #and the different cereals they produce have only small proportion so we still hard to discriminate
them by cereal characteristics
> #so we can conclude that they each have a balanced portfolio of cereals
>
>
> #2 b)
> table(seg.hc1.segment)
seg.hc1.segment
 1  2  3  4  5  6
 3 33 19  2  5  3
> summary(d0)
 mfr      calories        protein           fat            sodium          fibre
 G:22   Min.   : 50.0   Min.   : 0.7519   Min.   :0.000   Min.   :  0.0   Min.   : 0.000
 K:21   1st Qu.:110.0   1st Qu.: 2.0000   1st Qu.:0.000   1st Qu.:180.0   1st Qu.: 0.000
 N: 3   Median :134.3   Median : 3.0000   Median :1.000   Median :232.0   Median : 2.000
 P: 9   Mean   :149.4   Mean   : 3.6837   Mean   :1.423   Mean   :237.8   Mean   : 3.871
 Q: 5   3rd Qu.:179.1   3rd Qu.: 4.4776   3rd Qu.:2.000   3rd Qu.:290.0   3rd Qu.: 4.478
 R: 5   Max.   :440.0   Max.   :12.1212   Max.   :9.091   Max.   :787.9   Max.   :30.303
    carbo          sugars          shelf        potassium          vitamins
 Min.   :10.53   Min.   : 0.00   Min.   :1.000   Min.   : 15.00   100%   : 5
 1st Qu.:15.00   1st Qu.: 4.00   1st Qu.:1.000   1st Qu.: 45.00   enriched:57
 Median :18.67   Median :12.00   Median :2.000   Median : 96.59   none   : 3
 Mean   :19.97   Mean   :10.05   Mean   :2.169   Mean   :159.12
 3rd Qu.:22.39   3rd Qu.:14.00   3rd Qu.:3.000   3rd Qu.:220.00
 Max.   :68.00   Max.   :20.90   Max.   :3.000   Max.   :969.70
> #as we get from a) cut1
> #cluster 1, 2, 3, 5 only have 2, 3, 3, 5 cereal seperately,
> #after I look at their measurements indivisually
> #cluster1, cereal row number is 31, 32
> d0[31,]
        mfr calories protein fat sodium fibre carbo sugars shelf potassium vitamins
```

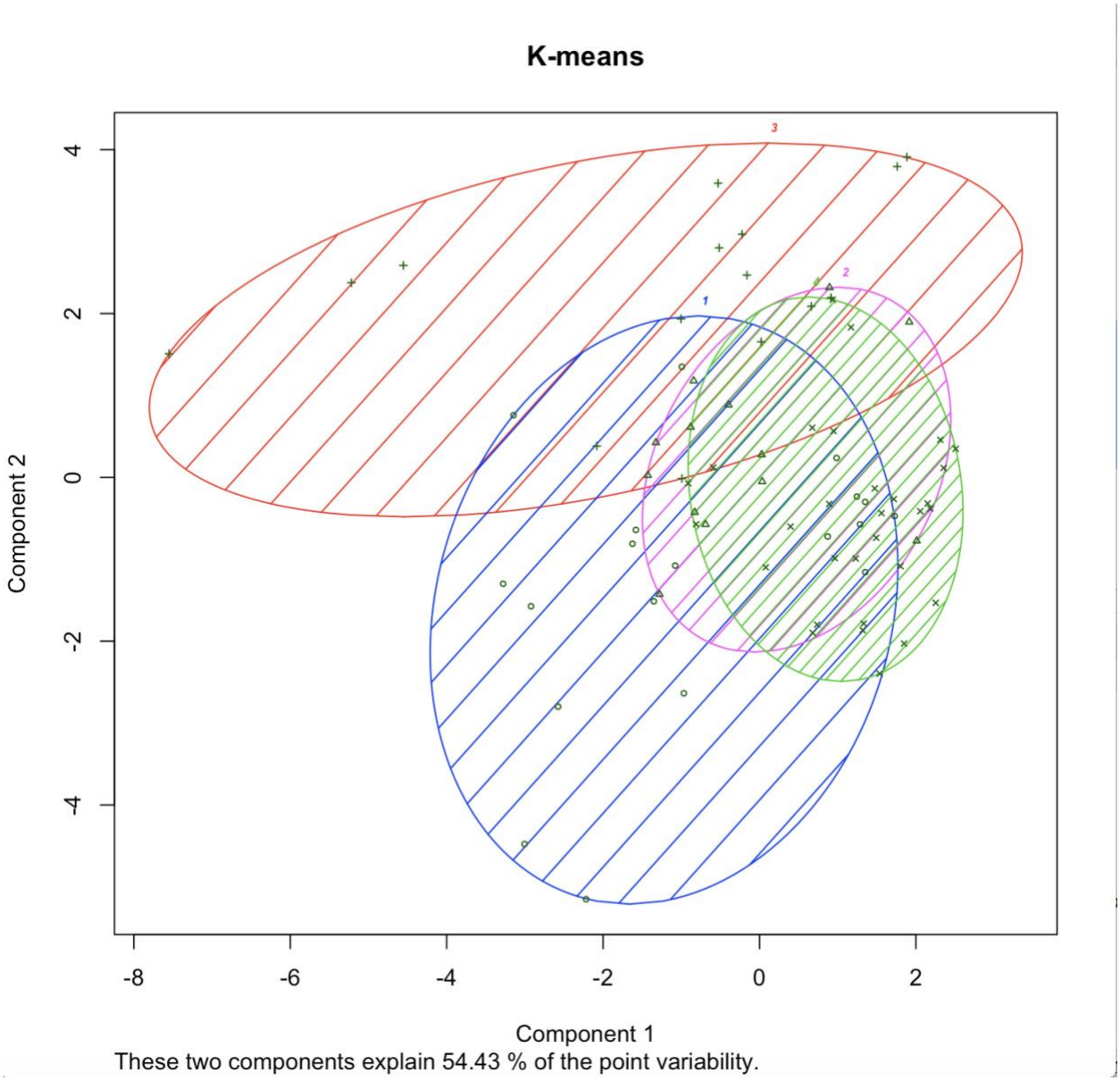
```
Grape-Nuts   P     440     12  0    680    12    68     12     3      360 enriched
> d0[32,]
                 mfr calories  protein     fat   sodium    fibre    carbo   sugars shelf potassium
Great Grains Pecan   P 363.6364 9.090909 9.090909 227.2727 9.090909 39.39394 12.12121     3
303.0303
                 vitamins
Great Grains Pecan enriched
> which.max(d0$calories)
[1] 31
> d1=d0
> d1[31,]$calories=0
> which.max(d1$calories)
[1] 32
> #for cluster 1 they are the two have highest two calories,
> #high protein, high fibre, both in shelf 3 and enriched vitamins
>
>
> #cluster2, cereal row number is 1, 2, 3
> d0[1,]
        mfr calories  protein     fat   sodium    fibre    carbo   sugars shelf potassium vitamins
100% Bran   N 212.1212 12.12121 3.030303 393.9394 30.30303 15.15152 18.18182     3  848.4849
enriched
> d0[2,]
       mfr calories  protein     fat   sodium    fibre    carbo   sugars shelf potassium vitamins
All-Bran   K 212.1212 12.12121 3.030303 787.8788 27.27273 21.21212 15.15151     3   969.697 enriched
> d0[3,]
                     mfr calories protein fat sodium fibre carbo sugars shelf potassium vitamins
All-Bran with Extra Fiber   K      100       8  0    280    28    16      0     3       660 enriched
> #for cluster 2, they all have high protein, high fibre and high potassium
> #they all almost achieve the max of all cereals
> #and they are all in shelf 3 and enriched vitamins
>
> #cluster3, cereal row number is 47, 54, 55
> d0[47,]
          mfr calories protein fat sodium fibre carbo sugars shelf potassium vitamins
Puffed Rice   Q       50       1  0      0     0    13      0     3        15     none
> d0[54,]
                  mfr calories  protein fat sodium    fibre    carbo sugars shelf potassium
Shredded Wheat 'n'Bran   N 134.3284 4.477612   0      0 5.970149 28.35821      0     1  208.9552
                  vitamins
Shredded Wheat 'n'Bran     none
> d0[55,]
                    mfr calories  protein fat sodium    fibre    carbo sugars shelf potassium
Shredded Wheat spoon size   N 134.3284 4.477612   0      0 4.477612 29.85075      0     1  179.1045
                    vitamins
Shredded Wheat spoon size     none
> #for cluster 3, they are all 0 fat, 0 suger, none vitamins
>
```

```
> #cluster5, cereal row number is 36, 46, 58, 59, 60
> a=d0[c(36,46,58,59,60),]
> a
                    mfr calories protein      fat   sodium    fibre    carbo sugars shelf potassium
Just Right Fruit & Nut   K 186.6667       4 1.333333 226.6667 2.666667 26.66667     12     3  126.6667
Product 19               K 100.0000       3 0.000000 320.0000 1.000000 20.00000      3     3   45.0000
Total Corn Flakes        G 110.0000       2 1.000000 200.0000 0.000000 21.00000      3     3   35.0000
Total Raisin Bran        G 140.0000       3 1.000000 190.0000 4.000000 15.00000     14     3  230.0000
Total Whole Grain        G 100.0000       3 1.000000 200.0000 3.000000 16.00000      3     3  110.0000
                    vitamins
Just Right Fruit & Nut     100%
Product 19                 100%
Total Corn Flakes          100%
Total Raisin Bran          100%
Total Whole Grain          100%
> #for cluster 5, these cereals are all low fat, low fibre, in shelf 3 and 100% vitamins
>
>
> #for cluster 4 and cluster 6, which have most of cereals
> table(seg.hc1.segment)
seg.hc1.segment
 1  2  3  4  5  6
 3 33 19  2  5  3
> #as is shown, the cluster number 2 from seg.hc1.segment is corresponding with cut1 cluster 4
> #cluster number 3 is corresponding with cut1 cluster 6
>
> cmeans=function(data,groups) aggregate(data,list(groups),function(x)mean(as.numeric(x)))
> cmeans(d0,seg.hc1.segment)
  Group.1      mfr calories   protein      fat   sodium     fibre    carbo    sugars    shelf potassium
1       1 2.333333 174.7475 10.747475 2.0202020 487.2727 28.525252 17.45455 11.111111 3.000000
826.06061
2       2 2.393939 122.2033  2.340626 0.8609377 213.5223  1.296902 16.82716  9.459648 1.575758
71.87564
3       3 2.736842 178.7172  4.416399 2.3457473 258.2894  4.264527 21.73504 13.088517 2.842105
204.20785
4       4 4.000000 401.8182 10.545455 4.5454545 453.6364 10.545454 53.69697 12.060606 3.000000
331.51515
5       5 1.400000 127.3333  3.000000 0.8666667 227.3333  2.133333 19.73333  7.000000 3.000000
109.33333
6       6 3.666667 106.2189  3.318408 0.0000000   0.0000  3.482587 23.73632  0.000000 1.666667
134.35323
  vitamins
1        2
2        2
3        2
4        2
5        1
6        3
```

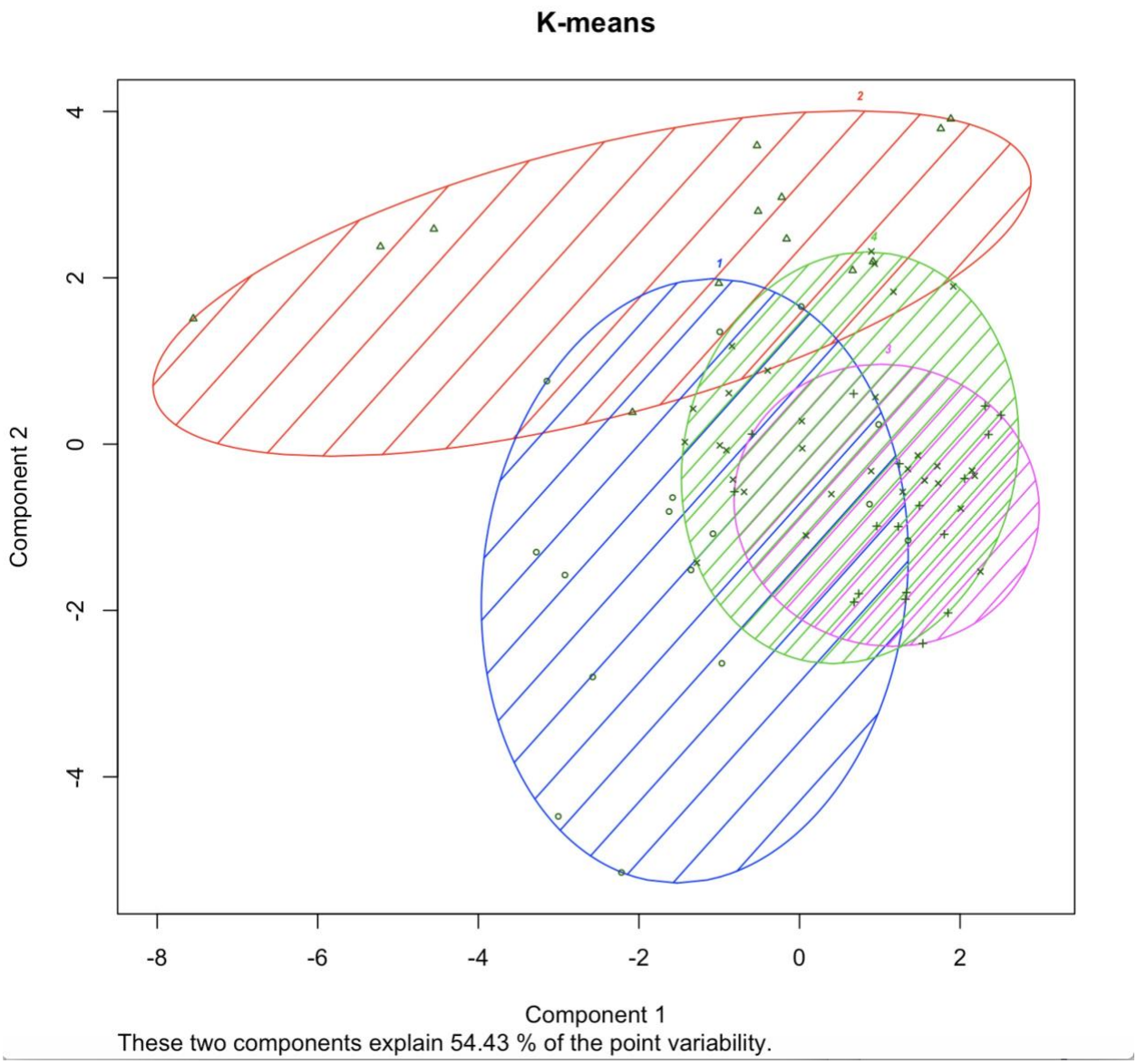
```
> #for cluster 4, they are low protein, low fat, low fibre and low potassium cereals
> #for cluster 6, they are the cereals which have medium value in most measurements
>
>
> #2 c)
> table(d0$shelf,seg.hc1.segment)
  seg.hc1.segment
   1  2  3  4  5  6
 1  0 16  0  0  0  2
 2  0 15  3  0  0  0
 3  3  2 16  2  5  1
> round(prop.table(table(d0$shelf,seg.hc1.segment),1),3)
  seg.hc1.segment
      1     2     3     4     5     6
 1 0.000 0.889 0.000 0.000 0.000 0.111
 2 0.000 0.833 0.167 0.000 0.000 0.000
 3 0.103 0.069 0.552 0.069 0.172 0.034
> #as table show shelf do not have much relation with previous 6 clusters
> table(d0$shelf,d0$mfr)

   G  K  N  P  Q  R
 1  6  4  2  2  0  4
 2  7  7  0  1  3  0
 3  9 10  1  6  2  1
> #as table show shelf do not have much relation with manufactures
>
> #so further analyze whether it depends on characteristics of cereals
> d1=subset(d0,shelf==1)
> summary(d1)
 mfr      calories        protein           fat            sodium          fibre
 G:6   Min.   : 82.71   Min.   :0.7519   Min.   :0.0000   Min.   :  0.0   Min.   :0.000
 K:4   1st Qu.:100.00   1st Qu.:2.0000   1st Qu.:0.0000   1st Qu.:203.1   1st Qu.:0.250
 N:2   Median :111.82   Median :2.6667   Median :0.0000   Median :236.0   Median :1.467
 P:2   Mean   :119.48   Mean   :2.9330   Mean   :0.6621   Mean   :216.1   Mean   :2.009
 Q:0   3rd Qu.:143.58   3rd Qu.:4.3582   3rd Qu.:1.3333   3rd Qu.:287.5   3rd Qu.:2.750
 R:4   Max.   :149.25   Max.   :6.0000   Max.   :2.6667   Max.   :343.3   Max.   :5.970
    carbo          sugars          shelf     potassium          vitamins
 Min.   :10.53   Min.   : 0.000   Min.   :1   Min.   : 25.00   100%   : 0
 1st Qu.:15.08   1st Qu.: 2.250   1st Qu.:1   1st Qu.: 35.00   enriched:16
 Median :19.51   Median : 3.739   Median :1   Median : 82.00   none   : 2
 Mean   :19.18   Mean   : 6.295   Mean   :1   Mean   : 89.18
 3rd Qu.:22.00   3rd Qu.:10.239   3rd Qu.:1   3rd Qu.:117.50
 Max.   :29.85   Max.   :17.045   Max.   :1   Max.   :208.96
> #calories        protein           fat            sodium          fibre
> #Median :111.82   Median :2.6667   Median :0.0000   Median :236.0   Median :1.467
> #carbo          sugars          shelf     potassium
> #Median :19.51   Median : 3.739   Median :1   Median : 82.00
> #low fat, low sugars
```

```
>
> d2=subset(d0,shelf==2)
> summary(d2)
 mfr     calories        protein          fat            sodium          fibre          carbo
 G:7   Min.   : 73.33   Min.   :1.000   Min.   :0.000   Min.   :  0.0   Min.   :0.000   Min.   :11.00
 K:7   1st Qu.:110.00   1st Qu.:1.083   1st Qu.:1.000   1st Qu.:128.8   1st Qu.:0.000   1st Qu.:12.00
 N:0   Median :122.50   Median :1.333   Median :1.167   Median :180.0   Median :0.000   Median :13.50
 P:1   Mean   :129.82   Mean   :2.058   Mean   :1.341   Mean   :190.0   Mean   :1.041   Mean   :14.95
 Q:3   3rd Qu.:148.61   3rd Qu.:2.500   3rd Qu.:1.453   3rd Qu.:266.0   3rd Qu.:1.000   3rd Qu.:17.46
 R:0   Max.   :179.10   Max.   :5.970   Max.   :4.000   Max.   :373.3   Max.   :6.667   Max.   :22.39
    sugars         shelf     potassium         vitamins
 Min.   : 2.00   Min.   :2   Min.   : 20.00   100%    : 0
 1st Qu.:12.00   1st Qu.:2   1st Qu.: 30.83   enriched:18
 Median :12.50   Median :2   Median : 54.17   none    : 0
 Mean   :12.51   Mean   :2   Mean   : 69.53
 3rd Qu.:13.86   3rd Qu.:2   3rd Qu.: 60.00
 Max.   :20.00   Max.   :2   Max.   :320.00
> #calories        protein          fat            sodium          fibre
> #Median :122.50   Median :1.333   Median :1.167   Median :180.0   Median :0.000
> #carbo           sugars         shelf     potassium
> #Median :13.50   Median :12.50   Median :2   Median : 54.17
> #low protein low fibre, low carbo and low potassium
> d3=subset(d0,shelf==3)
> summary(d3)
 mfr     calories        protein          fat            sodium          fibre
 G: 9   Min.   : 50.0   Min.   : 1.000   Min.   :0.000   Min.   :  0.0   Min.   : 0.000
 K:10   1st Qu.:133.3   1st Qu.: 3.000   1st Qu.:0.000   1st Qu.:220.0   1st Qu.: 2.667
 N: 1   Median :179.1   Median : 4.478   Median :1.333   Median :280.0   Median : 4.000
 P: 6   Mean   :180.1   Mean   : 5.159   Mean   :1.945   Mean   :281.0   Mean   : 6.783
 Q: 2   3rd Qu.:212.1   3rd Qu.: 6.000   3rd Qu.:2.985   3rd Qu.:320.0   3rd Qu.: 7.463
 R: 1   Max.   :440.0   Max.   :12.121   Max.   :9.091   Max.   :787.9   Max.   :30.303
    carbo          sugars         shelf     potassium         vitamins
 Min.   :13.00   Min.   : 0.000   Min.   :3   Min.   : 15.0   100%    : 5
 1st Qu.:17.05   1st Qu.: 5.682   1st Qu.:3   1st Qu.:110.0   enriched:23
 Median :21.00   Median :12.000   Median :3   Median :220.0   none    : 1
 Mean   :23.57   Mean   :10.857   Mean   :3   Mean   :258.1
 3rd Qu.:26.67   3rd Qu.:14.925   3rd Qu.:3   3rd Qu.:298.5
 Max.   :68.00   Max.   :20.896   Max.   :3   Max.   :969.7
> #calories        protein          fat            sodium          fibre
> #Median :179.1   Median : 4.478   Median :1.333   Median :280.0   Median : 4.000
> #carbo           sugars         shelf     potassium
> #Median :21.00  Median :12.000   Median :3   Median :220.0
> #high calories, high protein, high fibre
>
> #so in general, cereals display on shelf 1 are low fat and low sugars,
> #cereals display on shelf 2 are low protein low fibre, low carbo and low potassium
> #cereals display on shelf 3 are high calories, high protein, high fibre
>
```

```
>
> # 3)
> setwd("/Users/shutingchen/Desktop/ISE 529             Data Analytics/L13")
> set.seed(1)
> d0=read.csv("cereals.csv")
> d1=d0[,-which(names(d0)%in% "shelf")]
> d1=d1[,-c(1,2,3)]
> myKmeans<-function(dataSet,k){
+   p=ncol(dataSet)
+   n=nrow(dataSet)
+   #randomly assign the cluster to each row
+   t=sample(1:k,size=nrow(dataSet),replace = T)
+   #create matrices to store the cluster, mean and dist
+   pointProperty=matrix(data=t,nrow =n ,ncol = 1)
+   centerPointSet<-matrix(data = NA,nrow = k,ncol = p)
+   dist<-matrix(data=NA,nrow =n,ncol = k)
+   #create assign use for condition
+   assign=matrix(0,nrow=n,ncol = 1)
+   #stop when cluster assignments stop changing
+   if(identical(assign,pointProperty)==FALSE){
+     assign=pointProperty
+     for (i in 1:n) {
+       for (cent in 1:k) {
+         #calculate the p predictors' mean for each cluster
+         d=dataSet[which(pointProperty==cent),]
+         centerPointSet[cent,]=apply(d,2,mean)
+         #calculate the eudist for each row to k cluster
+         dist[i,cent]=daisy(rbind(dataSet[i,],centerPointSet[cent,]))
+         #assign the cluster which have min eudist to that row
+         pointProperty[i,]=as.numeric(which.min(dist[i,]))
+       }
+     }
+   }
+   return(pointProperty)
+ }
>
> m=myKmeans(d1,4)
> clusplot(d1,m,color=T,shade=T,labels=4,lines=0,main = "K-means",cex=0.5)
```




```
> table(m)
m
 1  2  3  4
19 13 15 30
>
>
> library(cluster)
> m1=kmeans(d1,centers=4)
> clusplot(d1,m1$cluster,color=T,shade=T,labels=4,lines=0,main = "K-means",cex=0.5)
```




```
> table(m1$cluster)

 1  2  3  4
16 13 18 30
```