A company is offering a subscription-based service (such as cable television or membership in a warehouse club) and have collected data from N = 300 respondents on age, gender, income, number of children, whether they own or rent their homes, and whether they currently subscribe to the offered service or not. We are interested in how measures such as household income and gender vary for the different segments. The objective is to find groups (clusters) of customers that differ in response to marketing efforts. By understanding the differences among groups the company can make a better strategy about product, promotion, positioning, etc.

It is interest to identify cluster of potential customers. To find the clusters go through the following steps

a) Download the data frame `segment.csv` available on blackboard.

b) Use function `hclust()` to group observations into clusters

c) Use function `cusplot` to plot observations in the first two PCs plane.

```
library(cluster)         # daisy()
x = matrix(1:12,nrow=3,ncol=4,byrow=T)
x
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
dist(x, diag = T)
   1  2  3
1  0
2  8  0
3 16  8  0

sqrt((9-1)^2 + (10-2)^2 + (11-3)^2 + (12-4)^2)   # 16

a=daisy(x)
b = as.matrix(a)
#    1 2  3
# 1  0 8 16
# 2  8 0  8
# 3 16 8  0

# daisy uses gower whenever non-numeric cols are

a=daisy(x,metric = "gower")
b = as.matrix(a)
#    1   2   3
#1 0.0 0.5 1.0
#2 0.5 0.0 0.5
#3 1.0 0.5 0.0

# columns are standardized by
# subtracting the minimum value, and
# dividing each entry by the range of the column
# rescaled column has range [0,1]

fun = function(x){max(x)-min(x)}
mini = apply(x,2,min)
rang = apply(x,2,fun)
x2 = scale(x,center=mini,scale=rang)
#      [,1] [,2] [,3] [,4]
# [1,]  0.0  0.0  0.0  0.0
# [2,]  0.5  0.5  0.5  0.5
# [3,]  1.0  1.0  1.0  1.0

# distance from obs 3 to obs 1 is 1.0
# distance from obs 3 to obs 2 is 0.5
```

```
library(cluster)         # daisy, clusplot()
d1=read.csv("segment.csv",header=T)
dim(d1)
# 300 customers with 6 attributes each

head(d1)
#  age gender income kids ownHome subscribe
#1  47   Male  49483    2   ownNo     subNo
#2  31   Male  35546    1  ownYes     subNo
#3  43   Male  44169    0  ownYes     subNo
#4  37 Female  81042    1   ownNo     subNo
#5  41 Female  79353    3  ownYes     subNo
#6  43   Male  58143    4  ownYes     subNo

summary(d1)
#      age           gender        income            kids        ownHome     subscribe
# Min.   :19.00   Female:157   Min.   : -5183   Min.   :0.00   ownNo :159   subNo :260
# 1st Qu.:33.00   Male  :143   1st Qu.: 39656   1st Qu.:0.00   ownYes:141   subYes: 40
# Median :39.50                Median : 52014   Median :1.00
# Mean   :41.17                Mean   : 50937   Mean   :1.27
# 3rd Qu.:48.00                3rd Qu.: 61404   3rd Qu.:2.00
# Max.   :80.00                Max.   :114278   Max.   :7.00

# dissimilarities
#=======================================================================

d2 = daisy(d1)         # 300(299)/2 = 44850   distances
d3 = as.matrix(d2)
dim(d3)                # 300 300
d3[1:5, 1:5]

#          1          2          3         4         5
#1 0.0000000 0.25363632 0.23262853 0.2618283 0.4152096
#2 0.2536363 0.00000000 0.06862683 0.4132008 0.3027257
#3 0.2326285 0.06862683 0.00000000 0.4249799 0.2926469
#4 0.2618283 0.41320077 0.42497987 0.0000000 0.2275711
#5 0.4152096 0.30272569 0.29264687 0.2275711 0.0000000

# largest and smallest dissimilarity
max(d3)
# 0.811403

diag(d3)=rep(1,300)
min(d3)
# 0.0002078782

# which are the two more/less dissimilar customers?
# dendrogram
#========================================================================

seg.hc = hclust(d2, method="complete")
class(seg.hc)
# "hclust"
str(seg.hc)
#List of 7
# $ merge      : int [1:299, 1:2] -85 -60 -126 -74 -14 -218 -175 -170 -145 -120 ...
# $ height     : num [1:299] 0.000208 0.002278 0.002395 0.00294 0.00296 ...
# $ order      : int [1:300] 128 137 102 101 107 173 219 298 256 287 ...
# $ labels     : NULL
# $ method     : chr "complete"
# $ call       : language hclust(d = d2, method = "complete")
# $ dist.method: NULL

# seg.hc is not a dendrogram

# Fig1 plot dendrogram from seg.hc
plot(seg.hc,cex=0.4,xlab="")
grid()

# cut at dissimilarity h=0.5
cut1=cut(as.dendrogram(seg.hc),h=0.5)
str(cut1)    # a list of two

# upper portion
cut1up = cut1$upper
str(cut1up)

#--[dendrogram w/ 2 branches and 5 members at h = 0.811]
#  |--[dendrogram w/ 2 branches and 2 members at h = 0.618]
#  |  |--leaf "Branch 1" (h= 0.359 midpoint = 9.75, x.member = 22 )
#  |  `--leaf "Branch 2" (h= 0.472 midpoint = 7.7, x.member = 18 )
#  `--[dendrogram w/ 2 branches and 3 members at h = 0.663]
#     |--leaf "Branch 3" (h= 0.466 midpoint = 54.8, x.member = 136 )
#     `--[dendrogram w/ 2 branches and 2 members at h = 0.506]
#        |--leaf "Branch 4" (h= 0.317 midpoint = 20.9, x.member = 58 )
#        `--leaf "Branch 5" (h= 0.344 midpoint = 14.7, x.member = 66 )

# Fig 2 dendrogram cut at h=0.50
plot(cut1up)
abline(h=0.5,lty=2,col="red")
grid()
```

```
# lower portion
cut1low = cut1$lower
str(cut1low)   # a list of five

# Fig 3 branch 1 dendrogram
plot(cut1low[[1]])
grid()

# some similarities shown
d1[c(101, 107), ]
#     age gender income kids ownHome subscribe
# 101  25   Male  18458    1   ownNo    subYes
# 107  23   Male  17510    1   ownNo    subYes

d1[c(278, 294), ]
#     age gender income kids ownHome subscribe
# 278  36 Female  46541    1   ownNo    subYes
# 294  36 Female  52353    1   ownNo    subYes

d1[c(173, 141), ]
#    age gender income kids ownHome subscribe
#173  65   Male  45517    0   ownNo    subYes
#141  25 Female  20126    2   ownNo    subYes



# dendogram distances
? cophenetic
# Details Section, 2nd paragraph
# ... dendrogram is an appropriate summary of the data
# if the correlation between original distances and the cophenetic distances is high

d4 = cophenetic(seg.hc)
class(d4)
# "dist"
length(d4)
# 44850
head(d4)
# 0.5061364 0.5061364 0.6629606 0.6629606 0.5061364 0.2270273

# daisy distances
d2 = daisy(d1)
length(d2)
# 44850
head(d2)
# 0.25363632 0.23262853 0.26182831 0.41520959 0.23729672 0.09053478
# compare dendogram and daisy distances
plot(d4,d2,pch=19,cex=0.25,xlab="dendogram distance",ylab="daisy distance")
grid()
# dendogram distances larger than daisy distances

cor(d4,d2)       #  0.7681604
# mild correlation

# cut to create k=4 groups
plot(seg.hc,cex=0.3,xlab="",main="")
grid()
cut4 = rect.hclust(seg.hc, k=4, border="red")

str(cut4)
#List of 4
# $ : int [1:22] 65 89 101 102 107 121 128 129 137 141 ...
# $ : int [1:18] 20 53 84 95 108 130 185 194 199 204 ...
# $ : int [1:136] 4 5 9 10 12 13 17 18 19 21 ...
# $ : int [1:124] 1 2 3 6 7 8 11 14 15 16 ...

# 22+18+136+124 = 300
# each component has the row numbers of each group
# group 1 has rows 1 2 3 6 7 8 11 14 15 16 ...
# group 2 has rows 4 5 9 10 12 13 17 18 19 21 ...

# vector of group assigments
seg.hc.segment = cutree(seg.hc, k=4)

head(seg.hc.segment)
# 1 1 1 2 2 1
table(seg.hc.segment)
# seg.hc.segment
#   1   2   3   4
# 124 136  18  22
# groups 1, 2 are most populated
# cluster means
aggregate(d1,by=list(seg.hc.segment),mean)
#  Group.1      age gender   income     kids ownHome subscribe
#1       1 40.74194     NA 49454.09 1.314516      NA        NA
#2       2 42.00735     NA 53759.61 1.235294      NA        NA
#3       3 44.27778     NA 52628.33 1.388889      NA        NA
#4       4 35.81818     NA 40456.09 1.136364      NA        NA
# There were 12 warnings (use warnings() to see them)

# avoid NAs
cmeans = function(data,groups) aggregate(data,list(groups),function(x) mean(as.numeric(x)))

cmeans(d1,seg.hc.segment)
#   Group.1      age   gender   income     kids  ownHome subscribe
# 1       1 40.74194 2.000000 49454.09 1.314516 1.467742         1
# 2       2 42.00735 1.000000 53759.61 1.235294 1.477941         1
# 3       3 44.27778 1.388889 52628.33 1.388889 2.000000         2
# 4       4 35.81818 1.545455 40456.09 1.136364 1.000000         2

levels(d1$gender)
# "Female" "Male"
levels(d1$ownHome)
# "ownNo"  "ownYes"
levels(d1$subscribe)
# "subNo"  "subYes"

# groups 1,2 different from 3,4 by subscription
# among non-subscribers, groups 1,2 diff by gender
# among subscribers, groups 3,4 diff by ownership

# clusterplot
library(cluster)
clusplot(d1,seg.hc.segment,color=T,shade=T,labels=4,lines=0,main="",cex=0.5,xlim=c(-4,4))

# 3,4 overlapping
# 1,2 more differentiated
```

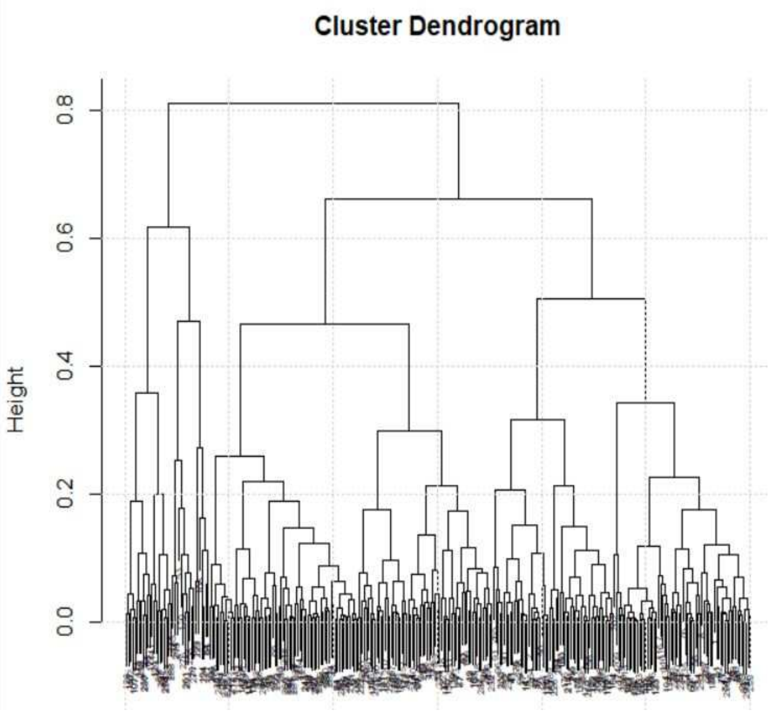

Figure 1: Complete dendrogram

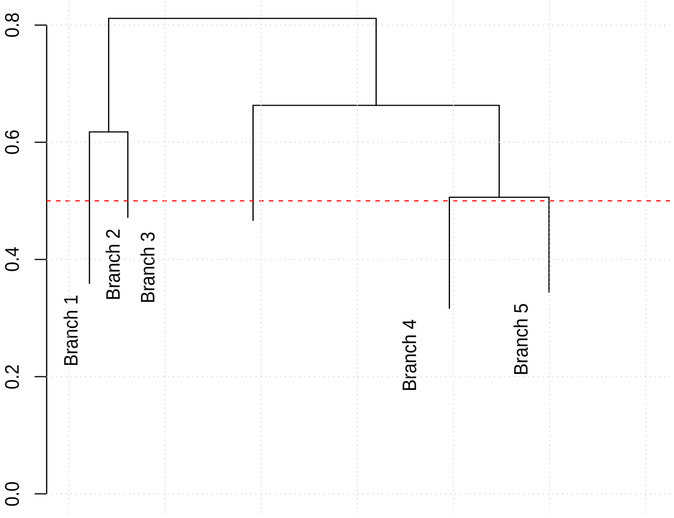

Figure 2: Cut at $h = 0.50$ resulting in 5 clusters (branches)

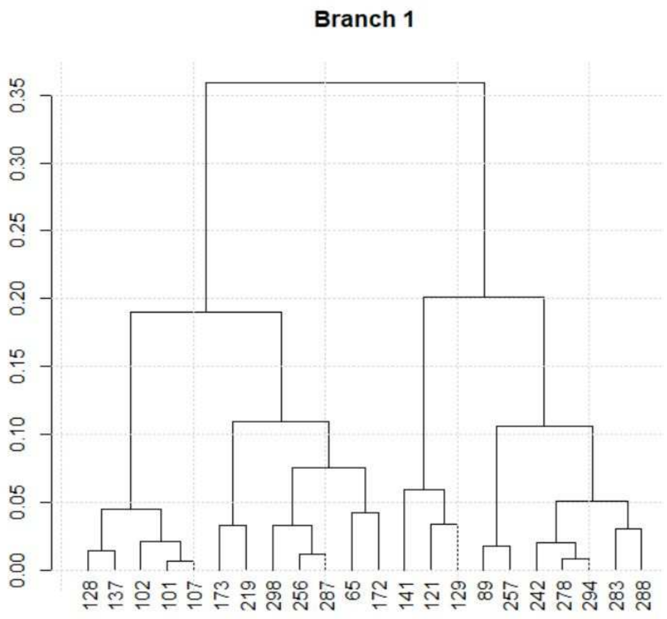

Figure 3: Branch 1 dendrogram

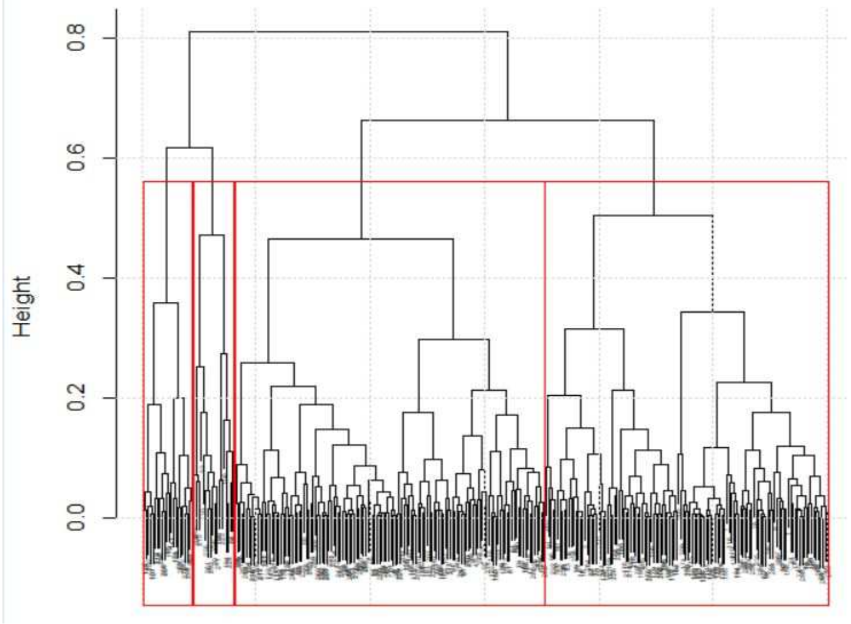

Figure 4: Cut to create $k = 4$ clusters

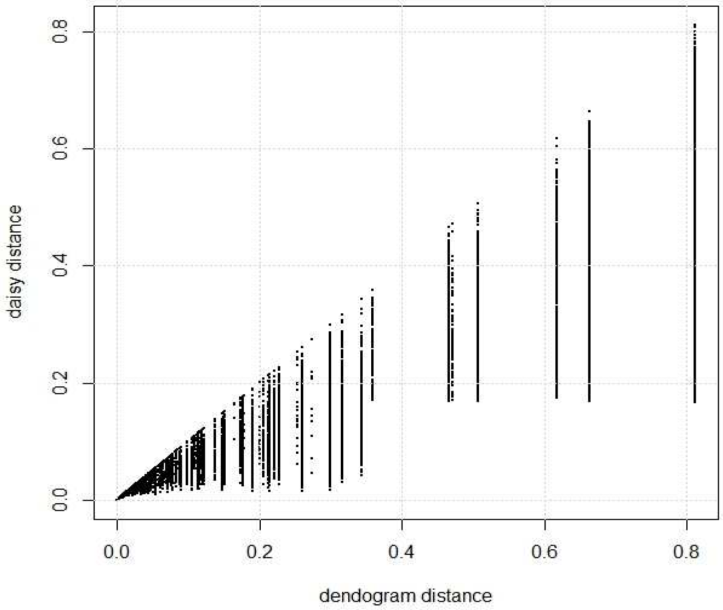

Figure 5: dendogram and daisy distances

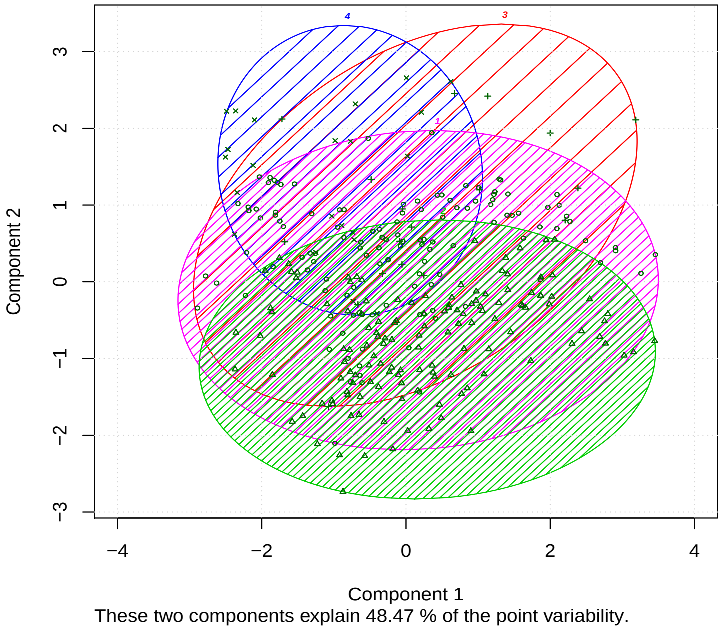

Figure 6: Clusters found by hclust in PC axes