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 company 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) Create a data frame by converting categorical variables to numerical

c) Use function `kmeans()` to group observations into 4 clusters

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

```
# kmeans.r

library(cluster)         # daisy(), clusplot()
setwd("C:/Users/USC Guest/Downloads2")
d1=read.csv("segment.csv",header=T)
dim(d1)
# 300   6
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

# file segment.csv rounded age and income

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

# means by categorical variable 'groups'
seg.summ <- function(data, groups) {
  aggregate(data, list(groups), function(x) mean(as.numeric(x)))
}

# k-means
#=========================================================
# k-means require numeric vars
# convert factors to (only) binary vars
d1.num <- d1
d1.num$gender    <- ifelse(d1$gender=="Male", 0, 1)
d1.num$ownHome   <- ifelse(d1$ownHome=="ownNo", 0, 1)
d1.num$subscribe <- ifelse(d1$subscribe=="subNo", 0, 1)

# all cols numeric
head(d1.num)
#   age gender income kids ownHome subscribe
# 1  47      0  49483    2       0         0
# 2  31      0  35546    1       1         0
# 3  43      0  44169    0       1         0
# 4  37      1  81042    1       0         0
# 5  41      1  79353    3       1         0
# 6  43      0  58143    4       1         0
```

```
# make window all wide
summary(d1.num)
#      age            gender           income            kids          ownHome       subscribe
# Min.   :19.00   Min.   :0.0000   Min.   : -5183   Min.   :0.00   Min.   :0.00   Min.   :0.0000
# 1st Qu.:33.00   1st Qu.:0.0000   1st Qu.: 39656   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.0000
# Median :39.50   Median :1.0000   Median : 52014   Median :1.00   Median :0.00   Median :0.0000
# Mean   :41.17   Mean   :0.5233   Mean   : 50937   Mean   :1.27   Mean   :0.47   Mean   :0.1333
# 3rd Qu.:48.00   3rd Qu.:1.0000   3rd Qu.: 61404   3rd Qu.:2.00   3rd Qu.:1.00   3rd Qu.:0.0000
# Max.   :80.00   Max.   :1.0000   Max.   :114278   Max.   :7.00   Max.   :1.00   Max.   :1.0000

# create 4 groups
set.seed(96743)
seg.k <- kmeans(d1.num, centers=4)
summary(seg.k)   # components of seg.k
#             Length Class  Mode
#cluster      300    -none- numeric
#centers       24    -none- numeric
#totss          1    -none- numeric
#withinss       4    -none- numeric
#tot.withinss   1    -none- numeric
#betweenss      1    -none- numeric
#size           4    -none- numeric
#iter           1    -none- numeric
#ifault         1    -none- numeric

#$ cluster has the assignments for each row

table(seg.k$cluster)
#  1   2   3   4
# 21  63  95 121

# cluster 4 highly populated


# cluster means
seg.summ(d1, seg.k$cluster)
#  Group.1      age   gender   income      kids  ownHome subscribe
#1       1 56.33333 1.428571 92287.10 0.4285714 1.857143  1.142857
#2       2 29.57143 1.571429 21631.76 1.0634921 1.301587  1.158730
#3       3 44.38947 1.452632 64703.78 1.2947368 1.421053  1.073684
#4       4 42.04132 1.454545 48208.83 1.5041322 1.528926  1.165289

# univariate segmentation

boxplot(d1.num$income ~ seg.k$cluster)
boxplot(d1.num$income ~ seg.k$cluster, ylab="Income", xlab="Cluster")
boxplot(d1.num$age ~ seg.k$cluster)

# groups are more differentiated by income
```

```
# bivariate segmentation

table(seg.k$cluster,d1.num$kids)
#     0  1  2  3  4  5  6  7
#  1 17  2  0  1  1  0  0  0
#  2 24 19 13  6  1  0  0  0
#  3 40 15 19 15  5  1  0  0
#  4 40 34 19 14  6  5  2  1

# groups 1,4 diff by n. kids

table(seg.k$cluster,d1$subscribe)
#    subNo subYes
#  1    18      3
#  2    53     10
#  3    88      7
#  4   101     20

# 1,3 few subscribers

table(seg.k$cluster,d1$gender)
#    Female Male
#  1     12    9
#  2     27   36
#  3     52   43
#  4     66   55

# all gender balanced

table(seg.k$cluster,d1$ownHome)
#    ownNo ownYes
#  1     3     18
#  2    44     19
#  3    55     40
#  4    57     64

# 1 more owners


# clusterplot
library(cluster)
clusplot(d1,seg.k$cluster,color=T,shade=T,labels=4,lines=0,main="K-means",cex=0.5)

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

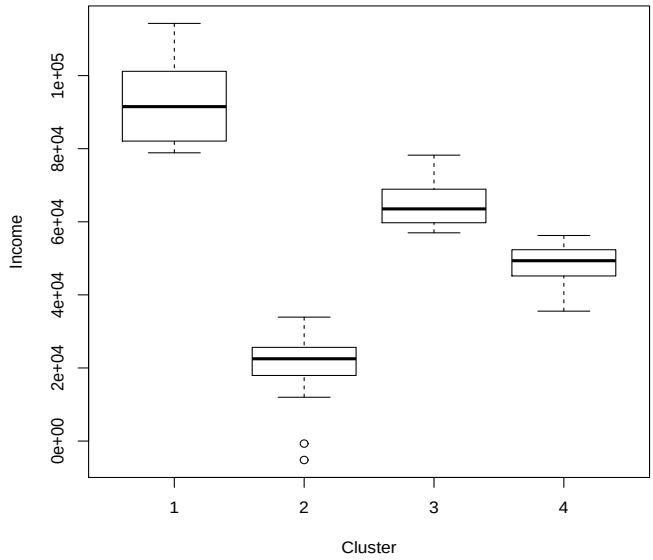

Figure 1: Boxplots per income group

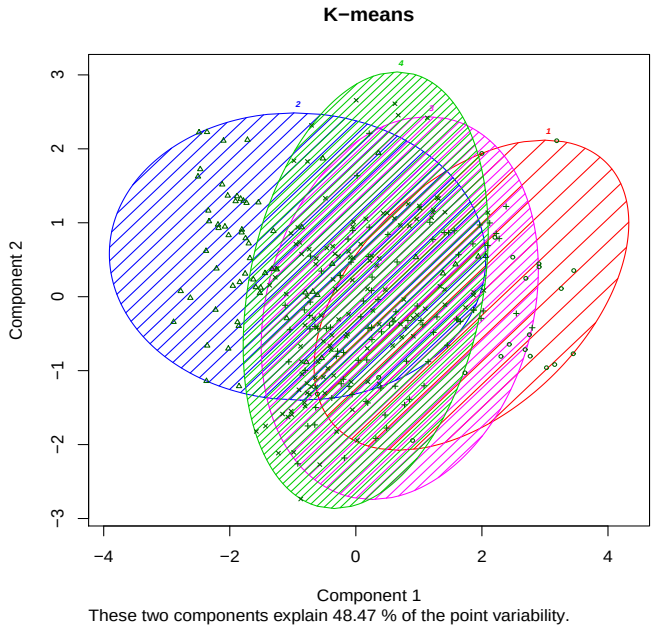

Figure 2: Clusters found by kmeans in PC axes