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) 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

setwd("C:/Users/USC Guest/Downloads2")
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
```

```
# k-means
#=========================================================
# k-means require numeric vars
# convert 2-level factors to binary vars
d2 = d1
d2$gender    = ifelse(d1$gender=="Male", 0, 1)
d2$ownHome   = ifelse(d1$ownHome=="ownNo", 0, 1)
d2$subscribe = ifelse(d1$subscribe=="subNo", 0, 1)

str(d2)
#'data.frame':   300 obs. of  6 variables:
# $ age      : int  47 31 43 37 41 43 38 28 44 35 ...
# $ gender   : num  0 0 0 1 1 0 0 0 1 1 ...
# $ income   : int  49483 35546 44169 81042 79353 58143 19282 47245 48333 52568 ...
# $ kids     : int  2 1 0 1 3 4 3 0 1 0 ...
# $ ownHome  : num  0 1 1 0 1 1 0 0 0 1 ...
# $ subscribe: num  0 0 0 0 0 0 0 0 0 0 ...

str(d1)
#'data.frame':   300 obs. of  6 variables:
# $ age      : int  47 31 43 37 41 43 38 28 44 35 ...
# $ gender   : Factor w/ 2 levels "Female","Male": 2 2 2 1 1 2 2 2 1 1 ...
# $ income   : int  49483 35546 44169 81042 79353 58143 19282 47245 48333 52568 ...
# $ kids     : int  2 1 0 1 3 4 3 0 1 0 ...
# $ ownHome  : Factor w/ 2 levels "ownNo","ownYes": 1 2 2 1 2 2 1 1 1 2 ...
# $ subscribe: Factor w/ 2 levels "subNo","subYes": 1 1 1 1 1 1 1 1 1 1 ...

# make window all wide
summary(d2)
#      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)
m1 = kmeans(d2, centers=4)
m1
# K-means clustering with 4 clusters of sizes 21, 63, 95, 121

# Cluster means:
#       age    gender   income      kids   ownHome  subscribe
#1 56.33333 0.5714286 92287.10 0.4285714 0.8571429 0.14285714
#2 29.57143 0.4285714 21631.76 1.0634921 0.3015873 0.15873016
#3 44.38947 0.5473684 64703.78 1.2947368 0.4210526 0.07368421
#4 42.04132 0.5454545 48208.83 1.5041322 0.5289256 0.16528926

#Clustering vector:
#  [1] 4 4 4 1 1 3 2 4 4 4 4 3 3 3 4 3 4 4 3 4 4 4 4 4 3 4 3 3 4 3 3 4 4 3 3 4 3 4 2 3 4 4 4 3 4 4 3
# [56] 4 4 3 2 3 4 3 4 4 3 3 4 3 4 3 4 3 4 4 4 1 4 3 4 4 1 4 3 4 3 4 4 4 3 3 3 3 3 4 4 3 3 4 2 2 2
#[111] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 3 2 3 2 4 3
#[166] 4 1 4 3 3 3 3 4 4 4 4 3 4 3 4 1 4 1 1 4 3 3 3 1 3 4 3 4 4 4 3 3 3 1 2 4 1 3 1 3 3 1 3 3 1 4 3
#[221] 2 3 1 4 1 1 3 1 3 4 4 4 4 4 3 3 4 4 3 3 4 3 4 3 3 4 4 3 3 4 4 3 3 3 4 4 3 4 3 3 4 2 4 4 4 4 3
#[276] 4 4 4 4 3 4 3 4 4 4 4 4 4 3 4 3 4 3 4 4 4 3 4 3 4

# Within cluster sum of squares by cluster:
# 2699877950  2820791226  3380699642  3433530327
# (between_SS / total_SS =  89.8 %)

# Available components:
# "cluster"  "centers"  "totss"  "withinss"   "tot.withinss" "betweenss" "size"  "iter"  "ifault"

# --- Comments  ----
# Clustering vector is m1$cluster
# Sum squares are m1$totss, m1$withinss, m1$betweenss, m1$tot.withinss

m1$withinss
# 2699877950 2820791226 3380699642 3433530327
sum(m1$withinss)
# 12334899144
m1$tot.withinss
# 12334899144

# total variance in the dataset that is explained by the clustering
m1$betweenss/m1$totss
# [1] 0.8982697

# variance not explained by the clustering
m1$tot.withinss/m1$totss
# [1] 0.1017303
```

```
# beware that increasing n. of clusters (centers) this ratio decreases

# From Cluster means Section
# 1 and 2 different by age, income
# gender about same across all clusters
# 1 and 4 different by kids

# components of m1
summary(m1)
#             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(m1$cluster)
#  1   2   3   4
# 21  63  95 121

# cluster 4 highly populated

# univariate segmentation

boxplot(d2$income ~ m1$cluster)
boxplot(d2$income ~ m1$cluster, ylab="Income", xlab="Cluster")
boxplot(d2$income/1000 ~ m1$cluster, ylab="Income (000s)", xlab="Cluster")

boxplot(d2$age ~ m1$cluster)

# groups are more differentiated by income


table(m1$cluster,d1$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(m1$cluster,d1$subscribe)
#    subNo subYes
#  1    18      3
#  2    53     10
#  3    88      7
#  4   101     20

# 1,3 few subscribers (absolute terms)
# group 3 few subscribers (relative terms)

table(m1$cluster,d1$gender)
#    Female Male
#  1     12    9
#  2     27   36
#  3     52   43
#  4     66   55

# all gender balanced

table(m1$cluster,d1$ownHome)
#    ownNo ownYes
#  1     3     18
#  2    44     19
#  3    55     40
#  4    57     64

# segment 1 with much more owners than no-owners


# clusterplot
library(cluster)
clusplot(d1,m1$cluster,color=T,shade=T,labels=4,lines=0,main="K-means",cex=0.5)

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

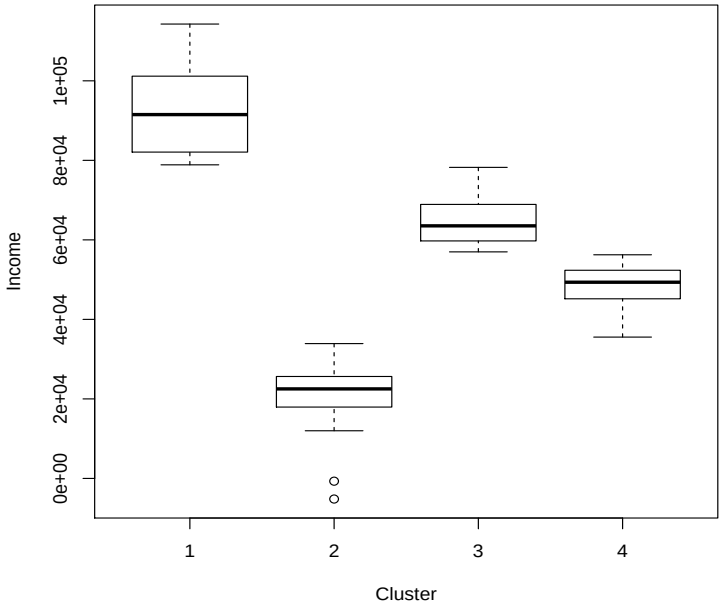

Figure 1: Income by group

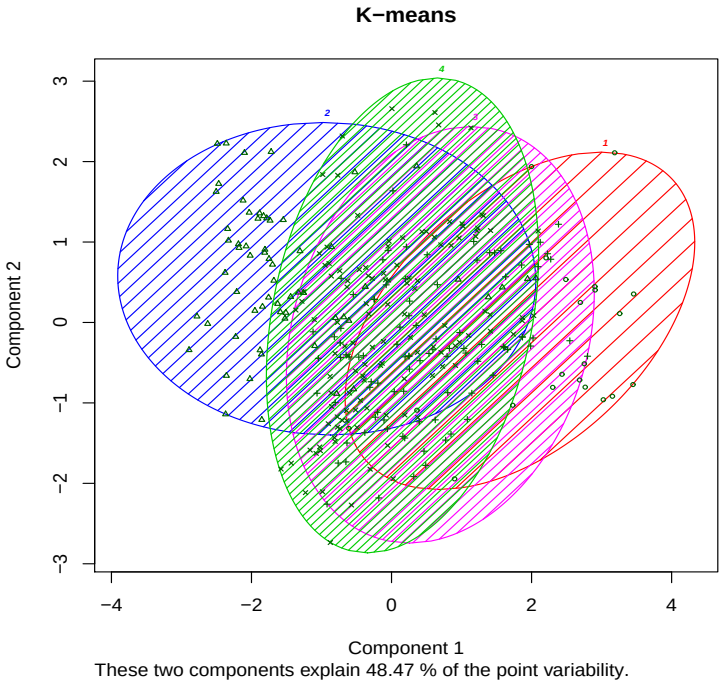

Figure 2: Clusters found by kmeans in PC axes