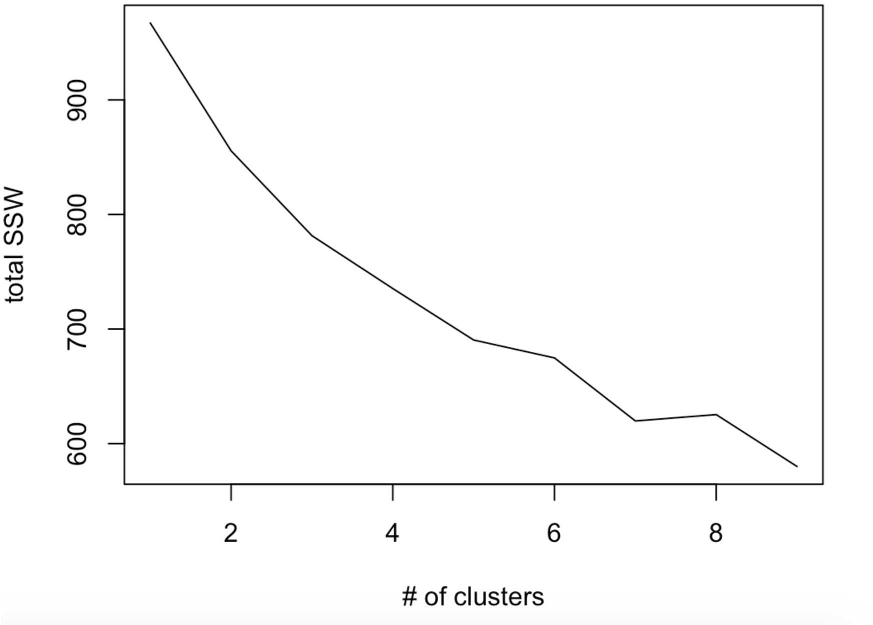
ASSIGNMENT 9: (partitioning by k-means) KEY

Using the data on members of the Classification Society posted on Moodle (files "csna.mlt" (data) and "csna.doc"(documentation)):

1. Using R or SPSS, use kmeans to partition the data set into k=1 to 9 clusters, using only the 14 binary variables dealing with research interests. You can use any method for choosing initial cluster seeds that you wish (or use the default in the package), but NOTE and state what option you are using. For each solution, save the (total) within-cluster sum of squares, and plot that against the number of clusters. Do you notice any anomalies? If yes, try to explain them.

R code:

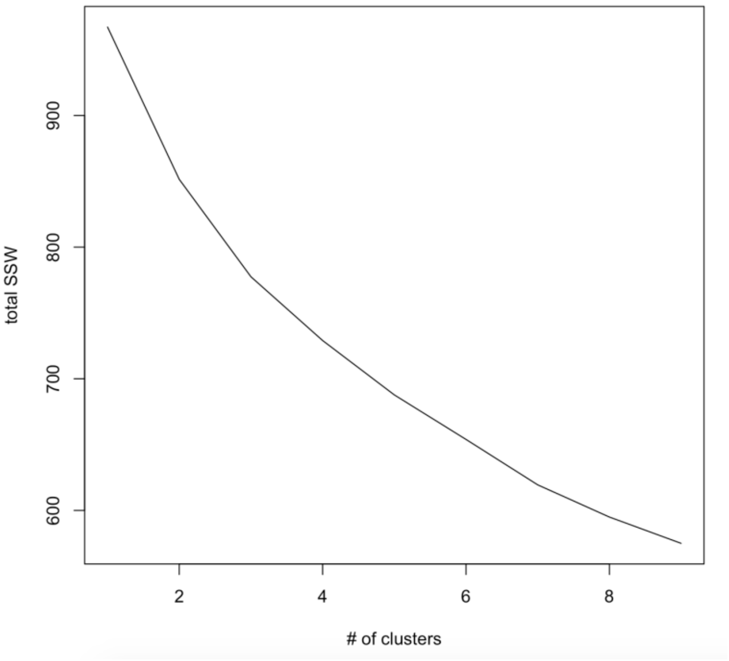
```
assign9<-read.csv("C:/Users/corter/Desktop/mdscstuff/CSNA_MLT.csv",header=TRUE,sep=",",
fill=TRUE)
head(assign9)
#put numeric interest variables into matrix x
x<-assign9[, 4:17]

# save WSS for each run, k=1 to 9
WSSvector<-c(0,0,0,0,0,0,0,0,0)
for (i in 1:9)
  { cl<-kmeans(x, i)
    WSSvector[i]<-cl$tot.withinss
  }
# plot WSS for each value of k=1 to 9
plot(WSSvector, type="l", xlab="# of clusters", ylab="total SSW")
```



PLOT: There does seem to be an anomaly: WSS should not increase from k=7 to k=8. The k=8 solution must be a local minimum. Try multiple random starts (maybe n=25).

# NOW SPECIFY MULTIPLE RANDOM STARTS FOR EACH VALUE OF k:

```
K=9
WSSvector<-c(rep(0,K))
for (k in 1:K) {
   cl<-kmeans(x, k, nstart=25)
   WSSvector[k]<-cl$tot.withinss
   }
WSSvector
plot(WSSvector, type="l", xlab="# of clusters", ylab="total SSW")
```



2. Using the plot you created in step 1, choose what seems to be the optimal number of clusters, k.
Re-run that solution, and save the cluster membership of each case in a vector or new variable.

→No elbow is obvious (maybe a faint one at k=7, so try k=7 clusters). Or, use some method to find optimal number of clusters (e.g. one of the stopping rules described in Milligan & Cooper, 1985).

```
cl<-kmeans(x, 7, nstart=25)
```

3. To help you interpret the k-cluster solution, run the following analyses:

A. INTERNAL VALIDITY: calculate means (and, optionally, the standard deviations) of the 14 interest variables, BY cluster. This basically shows the cluster centroids (with SDs) on the clustering variables.

```
> int<-t(cl$centers)
names<-c("appd","clus","comb","soft","disc","fact","grpt","grph","mdsc","numm","ntax","othr","patt","p
rob")
> rownames(int)<-names
> int
              1            2            3            4            5            6        7
appd0.13461538   0.95238095   0.63636364   0.22222222   0.30303030   0.84705882   0.58
clus  0.32692308   0.84126984   0.74545455   0.91666667   0.86363636   0.78823529   0.88
comb 0.01923077   0.00000000   0.07272727   0.75000000   0.01515152   0.04705882   0.02
soft  0.00000000   0.12698413   0.03636364   0.22222222   0.27272727   1.00000000   0.36
disc  0.00000000   0.17460317   0.63636364   0.05555556   0.15151515   0.18823529   1.00
fact  0.05769231   0.60317460   0.18181818   0.05555556   0.21212121   0.08235294   1.00
grpt  0.01923077   0.06349206   0.03636364   0.91666667   0.18181818   0.05882353   0.02
grph 0.03846154   0.44444444   0.21818182   0.08333333   0.22727273   0.35294118   0.08
mdsc 0.07692308   0.98412698   0.23636364   0.27777778   0.40909091   0.24705882   0.40
numm 0.07692308  0.12698413   0.21818182   0.05555556   0.19696970   0.16470588   0.04
ntax 0.05769231   0.06349206   0.00000000   0.25000000   1.00000000   0.21176471   0.22
othr  0.07692308   0.06349206   0.09090909   0.13888889   0.04545455   0.03529412   0.08
patt  0.01923077   0.04761905   0.90909091   0.27777778   0.43939394   0.45882353   0.02
prob 0.11538462   0.04761905   0.25454545   0.19444444   0.07575758   0.02352941   0.00
```

Some interpretations: Cluster 1 seems to be statisticians interested in cluster analysis. Cluster 2 may be psychologists, because highly interested in FA & PCA. Cluster 3 people are into pattern analysis. Cluster 4 adds graph theory. Cluster 5 may be biologists, because interested in numerical taxonomy. Cluster 6 = software. Cluster 7 = multivariate stats interests.

B. EXTERNAL VALIDITY: do a cross-tabulation of the obtained cluster memberships with the "academic specialty or discipline" variable ("spec"). Are your obtained clusters interpretable in terms of particular disciplines?

4. Using your results from part 3, interpret each cluster.

I read the table into Excel to label it. Almost every cluster contains statisticians, and many contain psychologists. But there are some patterns as to academic specialty, for example cluster 1 (and cluster 3) people are mostly statisticians, and cluster 5 mostly biologists, as hypothesized. Also Cluster 2 has many psychologists, as predicted. Computer scientists tend to be in Cluster 6 (a main interest there was “software”). Cluster 4 contains many of the mathematicians. The last cluster has many of the business school (marketing) people, perhaps mainly because they are interested in “business applications”.

[see table below]

		CLUSTER:						
specname	spec	**1**	**2**	**3**	**4**	**5**	**6**	**7**
01 agric	1	0	0	1	0	1	1	1
03 archeology	3	0	0	1	0	1	1	0
04 biology	4	3	3	3	3	19	9	6
05 busappl	5	5	3	1	1	2	3	8
06 chemistry	6	0	1	0	0	0	2	0
07 compsci	7	4	2	5	6	5	20	0
08 economics	8	0	0	0	0	0	2	2
09 education	9	1	6	1	0	0	2	1
10 engin	10	1	1	9	1	3	6	0
11 geography	11	0	0	0	0	2	0	0
12 geology	12	1	1	0	0	0	0	1
14 libsci	14	0	0	0	0	1	0	0
15 ling	15	0	1	0	0	0	0	0
16 math	16	5	0	3	9	2	1	0
17 medicine	17	1	0	2	0	2	2	0
18 other	18	2	5	3	2	5	6	3
20 polsci	20	2	0	0	1	0	0	0
21 psychiatry	21	0	0	0	0	1	1	1
22 psychology	22	7	21	4	11	11	13	13
23 sociology	23	0	3	0	1	0	2	1
24 soilsci	24	0	0	0	0	0	0	1
25 statistics	25	13	16	22	1	11	14	12
99 (missing)	99	7	0	0	0	0	0	0