FES758b / S&DS363 / S&DS 563
Multivariate Statistics
Homework #3: Cluster Analysis
Due: Friday, 2/22/2019 11:59pm on CANVAS

Answers should be complete and concise. You may use any statistics program for calculations that you wish (multiple programs is fine too).

IF YOU WORK IN A GROUP, YOU MAY TURN IN ONE ASSIGNMENT FOR YOUR GROUP. BE SURE ALL NAMES ARE ON THE ASSIGNMENT.

Examples of Cluster Analysis in SAS and R are up on the CANVAS home page

SAMPLE DATA SET

The example below is JUST FOR YOUR PRACTICE. NOTHING TO TURN IN HERE!

The file senate104.xls (or senate104.csv) contains the voting records for US senators during the 104th congressional session. This file actually only contains the first 198 votes. For each vote, senators may respond yes (1) or no(0). They may also abstain or not be present. For this dataset, I have replaced abstain/not present votes with 0.5 (hierarchical clustering in many software packages will not work with missing values). You task is to cluster senators.

Note : there are more than 100 senators because some senators resigned during the session and were replaced by others. Also, note that one Colorado senator switched from being a democrat to being a republican and is thus listed twice.

1). Get a measure of the standard deviation of each vote. What do you observe? (no need to turn in all the output). Votes with low standard deviation will not have much affect on the clustering process (i.e. no differentiation between senators based on these votes).

Note : in SPSS, use Analyze → Descriptive Statistics → Descriptives. In SAS, use something like

```
proc means data=in.senate104;
  var v1-v198;
run;
```

In R, assuming the votes are in an object called senate, use

```
#get the data from a CSV file
senate=read.csv("http://reuningscherer.net/stat660/data/sen
ate104.csv",header=T)

#get standard deviation for each vote
round(sqrt(apply(senate[,-1],2,var)),2)
```

Standard deviations were largely in the .3, .4 range; however, some are lower in the .1 and less range – that is, some votes had more agreement.

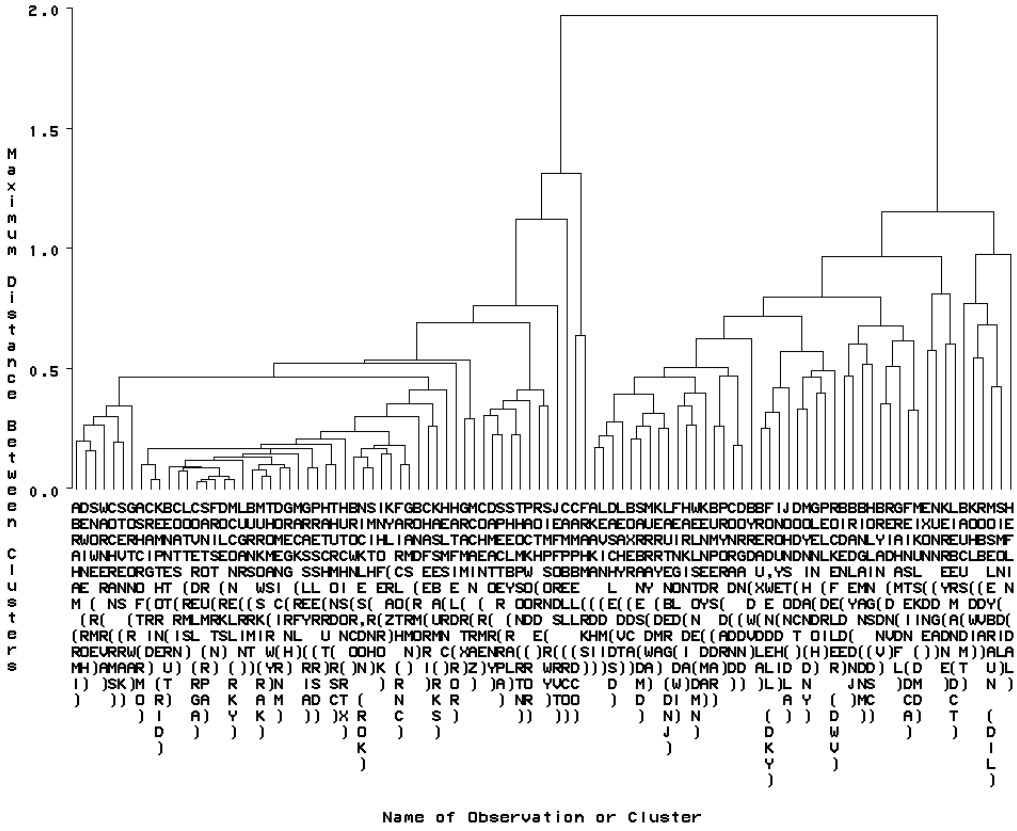
2). Use hierarchical clustering on the data. Try two metrics and two agglomeration procedures, one of which should be complete linkage (furthest neighbor). You might try some measure appropriate to binary data (which this is approximately).

Note – in SAS, I've given some code below. By default, Proc Cluster only calculates Euclidean distance. However, you can use a macro called 'distance' to calculate other distances. Copy the file `distnew.sas` *to you computer from the Software folder on the classes server in the materials folder.*

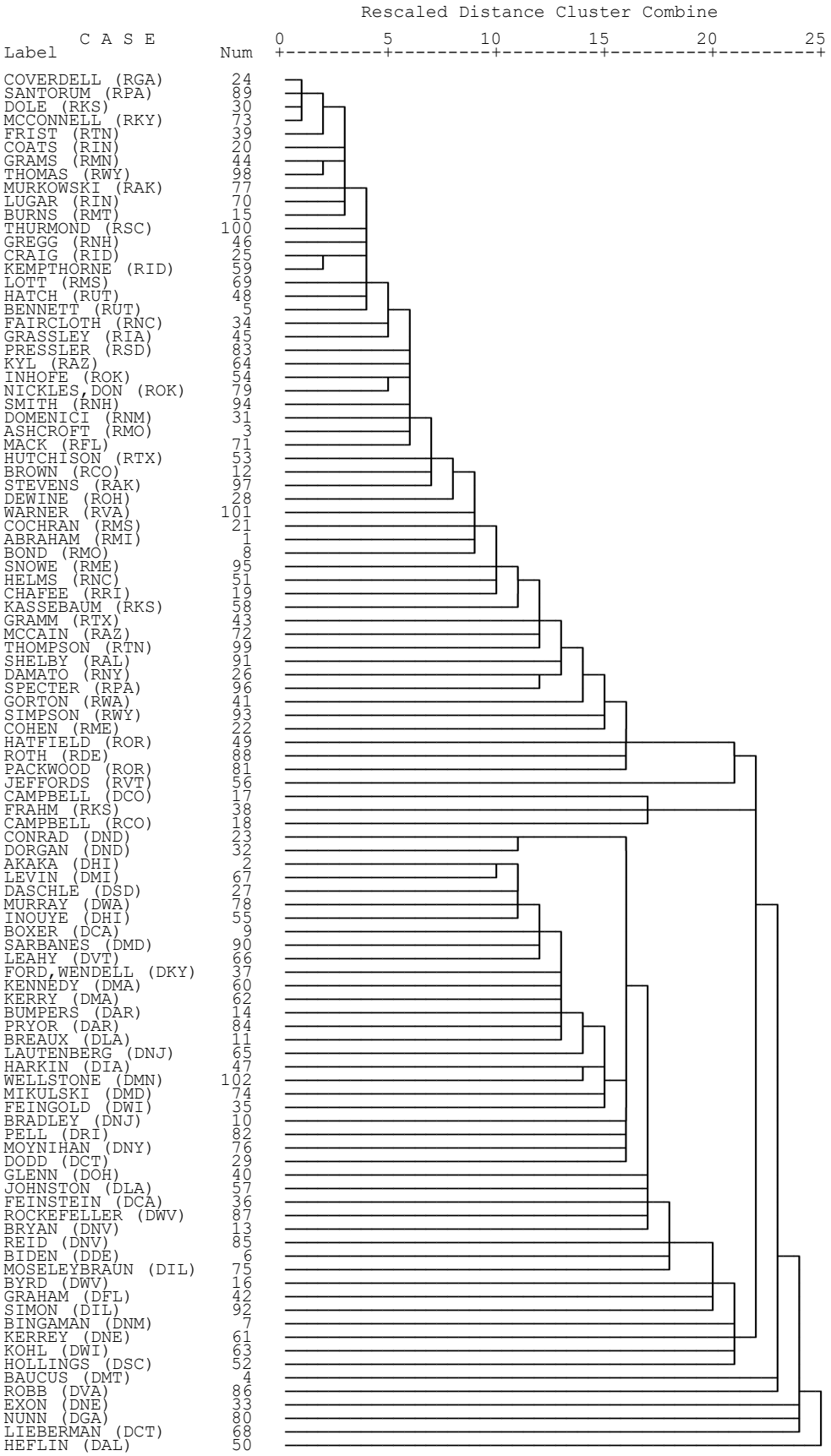
The usage is as follows :

```
*change method= for other metrics.  See SAS help file for options.
*Use SNORM option to standardize variables.  Need to use other options
*in the VAR statement for categorical data;



PROC DISTANCE DATA= MYLIB.senate OUT=OUTDIST  METHOD=CITYBLOCK
SNORM;
 VAR INTERVAL (V1-V198);
 COPY CEREAL;
RUN;

*RUN CLUSTERING PROCEDURE AND SAVE RESULTS;
PROC CLUSTER DATA=outdist METHOD=compact RMSSTD RSQ OUTTREE=TREE;
id senator;
RUN;


*MAKE A DENDOGRAM;
PROC TREE DATA=TREE;
RUN;
```

Obviously, there are many possibilities here. Here is compact clustering using a manhattan metric (from SAS – see program above).

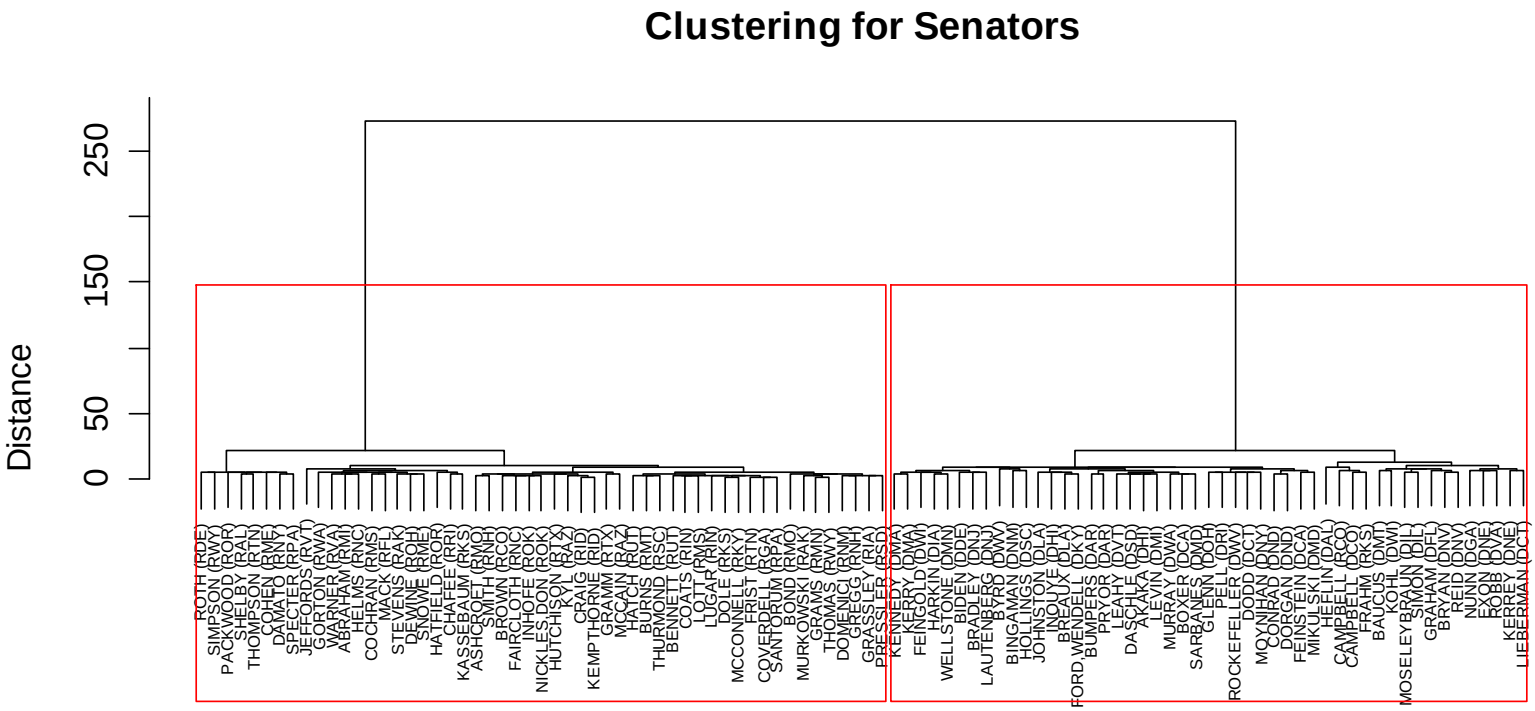


Here is single linkage using using Euclidean distance in SPSS : (Note that this is just a text picture, so you can change font size (below is 8) and change line spacing (below is set to 6 point) to make the picture fit better.


Rescaled Distance Cluster Combine
C A S E
Label Num
0 5 10 15 20 25
COVERDELL (RGA) 24
SANTORUM (RPA) 89
DOLE (RKS) 30
MCCONNELL (RKY) 73
FRIST (RTN) 39
COATS (RIN) 20
GRAMS (RMN) 44
THOMAS (RWY) 98
MURKOWSKI (RAK) 77
LUGAR (RIN) 70
BURNS (RMT) 15
THURMOND (RSC) 100
GREGG (RNH) 46
CRAIG (RID) 25
KEMPTHORNE (RID) 59
LOTT (RMS) 69
HATCH (RUT) 48
BENNETT (RUT) 5
FAIRCLOTH (RNC) 34
GRASSLEY (RIA) 45
PRESSLER (RSD) 83
KYL (RAZ) 64
INHOFE (ROK) 54
NICKLES,DON (ROK) 79
SMITH (RNH) 94
DOMENICI (RNM) 31
ASHCROFT (RMO) 3
MACK (RFL) 71
HUTCHISON (RTX) 53
BROWN (RCO) 12
STEVENS (RAK) 97
DEWINE (ROH) 28
WARNER (RVA) 101
COCHRAN (RMS) 21
ABRAHAM (RMI) 1
BOND (RMO) 8
SNOWE (RME) 95
HELMS (RNC) 51
CHAFEE (RRI) 19
KASSEBAUM (RKS) 58
GRAMM (RTX) 43
MCCAIN (RAZ) 72
THOMPSON (RTN) 99
SHELBY (RAL) 91
DAMATO (RNY) 26
SPECTER (RPA) 96
GORTON (RWA) 41
SIMPSON (RWY) 93
COHEN (RME) 22
HATFIELD (ROR) 49
ROTH (RDE) 88
PACKWOOD (ROR) 81
JEFFORDS (RVT) 56
CAMPBELL (DCO) 17
FRAHM (RKS) 38
CAMPBELL (RCO) 18
CONRAD (DND) 23
DORGAN (DND) 32
AKAKA (DHI) 2
LEVIN (DMI) 67
DASCHLE (DSD) 27
MURRAY (DWA) 78
INOUYE (DHI) 55
BOXER (DCA) 9
SARBANES (DMD) 90
LEAHY (DVT) 66
FORD,WENDELL (DKY) 37
KENNEDY (DMA) 60
KERRY (DMA) 62
BUMPERS (DAR) 14
PRYOR (DAR) 84
BREAUX (DLA) 11
LAUTENBERG (DNJ) 65
HARKIN (DIA) 47
WELLSTONE (DMN) 102
MIKULSKI (DMD) 74
FEINGOLD (DWI) 35
BRADLEY (DNJ) 10
PELL (DRI) 82
MOYNIHAN (DNY) 76
DODD (DCT) 29
GLENN (DOH) 40
JOHNSTON (DLA) 57
FEINSTEIN (DCA) 36
ROCKEFELLER (DWV) 87
BRYAN (DNV) 13
REID (DNV) 85
BIDEN (DDE) 6
MOSELEYBRAUN (DIL) 75
BYRD (DWV) 16
GRAHAM (DFL) 42
SIMON (DIL) 92
BINGAMAN (DNM) 7
KERREY (DNE) 61
KOHL (DWI) 63
HOLLINGS (DSC) 52
BAUCUS (DMT) 4
ROBB (DVA) 86
EXON (DNE) 33
NUNN (DGA) 80
LIEBERMAN (DCT) 68
HEFLIN (DAL) 50

In R, use something like the following :

```
sennorm<-scale(senate[,-1])
#get the distance matrix
dist1<-dist(sennorm [,-1],method="euclidean")
#now do clustering;
clust1<-hclust(dist1,method="ward")
#draw the dendrogram
plot(clust1,labels=senate[,1],cex=0.5,xlab="",ylab="Distance"
,main="Clustering for Senators")
rect.hclust(clust1,k=2)
```



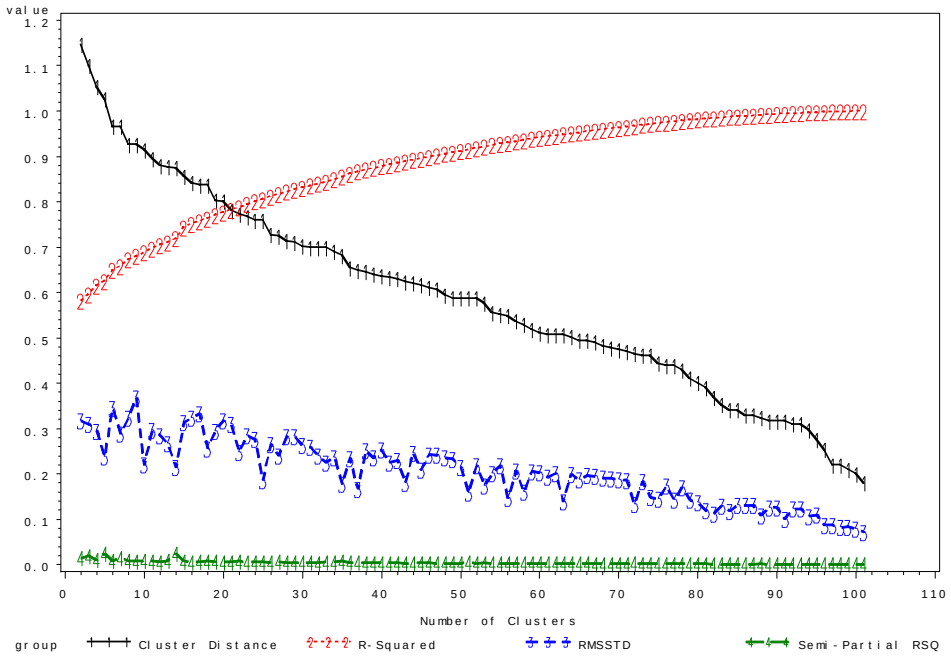
3). Below is a print of four measures evaluating hierarchical clustering using Euclidean distance and complete linkage (I'm not saying this is the best hierarchical clustering method to use, just the most common). Comment on these four statistics and decide on a suggested number of clusters.

Here is the SAS Code used to get the plot below. NOTE THAT YOU CANNOT GET THIS PLOT IF YOU START BY USING THE DISTANCE MACRO (I.E. YOU CAN ONLY GET THE PLOT BELOW USING EUCLIDEAN DISTANCE – SEEMS SILLY, BUT I HAVEN'T WORKED OUT A WAY AROUND THIS!!)

```
PROC CLUSTER DATA=IN.SENATE104 METHOD=compact RMSSTD RSQ OUTTREE=TREE;
id senator;
RUN;
```

```
*INCLUDE THE FILE WITH THE MACRO FOR EVALUATING CLUSTER NUMBER;
*NEED TO CHANGE FILE LOCATION TO MATCH THAT ON YOUR COMPUTER!!!!;
%INCLUDE 'C:\CLUSTER.SAS';
*RUN THE MACRO - ONLY ARGUMENT IS THE NAME OF THE OUTPUT DATASET FROM
THE CLUSTERING PROCEDURE;
%CLUSTPLOT(TREE);
RUN;
```

There is no obvious number of clusters suggested by these metrics. Cluster distance perhaps suggest a number of clusters in the 5 to 7 range, but this is not a well-defined break point.



4). Use k-means clustering on the data. Try somewhere between 2 and 10 groups. Make a scree plot of internal SS versus k to look for an elbow.

HOMEWORK ASSIGNMENT

PLEASE turn in the following answers for YOUR DATASET! If Cluster Analysis is not appropriate for your data, use one of the two loaner datasets described at the end of the assignment.

List your Name(s – if a group) and a one sentence reminder of which dataset your are using.

1. Think about what metrics are appropriate for your data based on data type. Write a few sentences about this. Also think about whether you should standardize or transform your data (comment as appropriate).

2. Try various forms of hierarchical cluster analysis. Try at least two different metrics and two agglomeration methods. Produce dendrograms and comment on what you observe.

3. If possible, run the SAS macro or the R function to think about how many groups you want to retain (i.e. get the plots of cluster distance, R-squared, etc). If you can't run this, discuss how many groups you think are present.

4. Run k-means clustering on your data. Compare results to what you got in 3.) Include a sum of squares vs. k (number of clusters) plot and comment on how many groups exist.

5. Comment on the number of groups that seem to be present based on what you find above.

LOANER DATASET 1
(if Cluster Analysis is not appropriate for your data)

The file `stream.xls` contains data on the prevalence of 11 species of microcrustacea in seven streams in Alaska. Five measurements were made at each site. The species are

sp1	*Nitocra hibernica*
sp2	*Atheyella illinoisensis*
sp3	*Atheyella idahoenis,*
sp4	*Bryocamptus hiemalis*
sp5	*Bryocamptus zschokkei,*
sp6	*Acanthocyclops vernalis*
sp7	*Alona guttata,*
sp8	*Graptoleberis*
sp9	*Chydorus*
sp10	*macrothricidae*
sp11	*Maraenobiotus insegnipes*

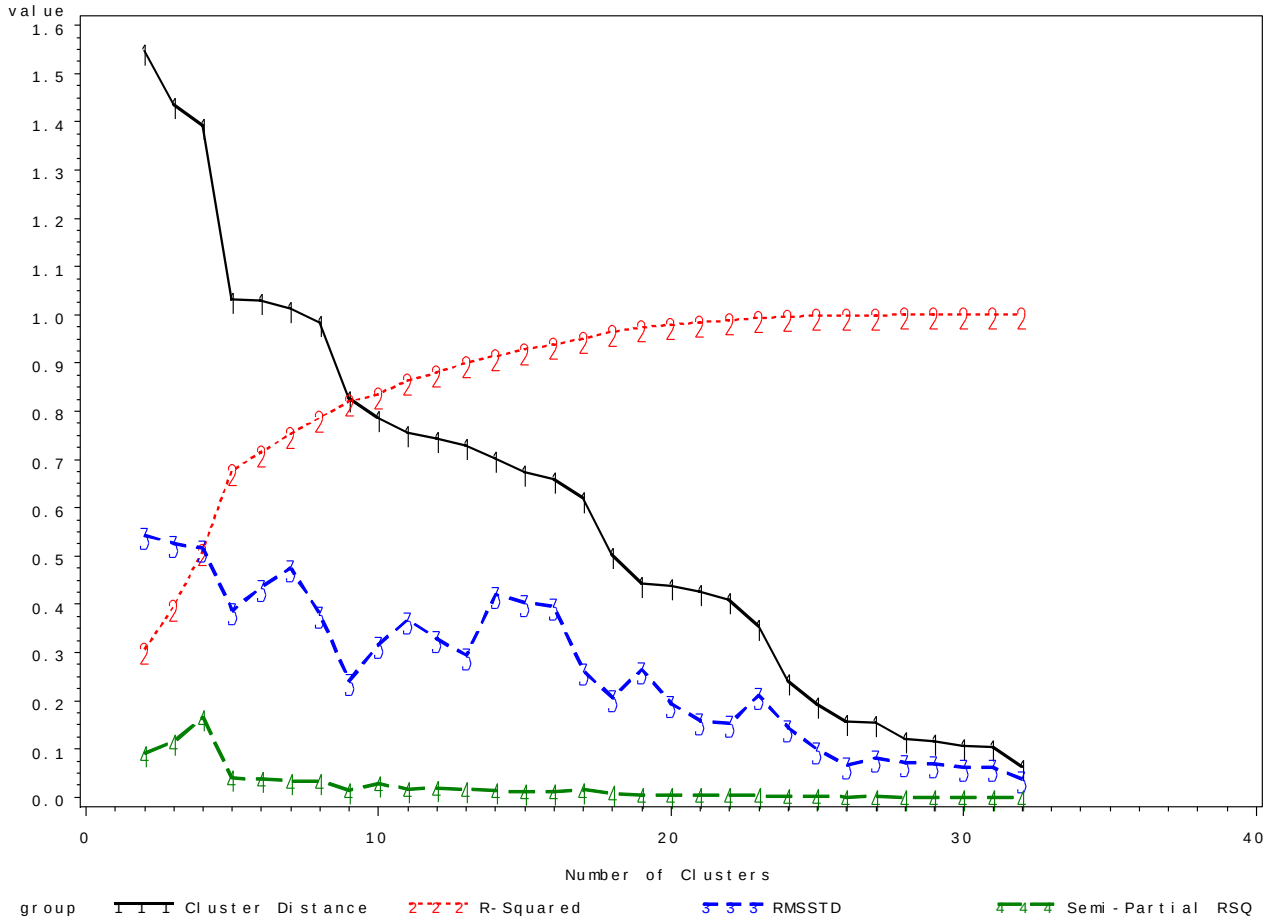
Stream age, the Pfankuch index (a measure of stability – higher values are lower stability), Temperature in degrees C, turbitidy, conductivity, and alkalinity were also measured for each stream – these values are also included.

Data is taken from Peter Shaw's Multivariate Statistics for the Environmental Sciences (2003), p. 24

Your goal is to use cluster analysis to identify similar streams.

You'll want to make a transformation of the species data before clustering. I recommend a log transformation, but you'll need to add one before taking logs (because . . . . *can't take log of zero!*)

Below is a SAS printout of statistics used to evaluate the number of clusters based on Euclidean distance and Ward's Method. How many clusters do you think exist?


value
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
Number of Clusters
group
Cluster Distance
R-Squared
RMSSTD
Semi-Partial RSQ

LOANER DATASET 2
(if Cluster Analysis is not appropriate for your data)

The file `University.csv` contains data on from 1995 on 25 Universities. The variables are

- SAT Score
- Percent of Class in Top 10% of high school class
- Acceptance Rate
- Student/Faculty Ratio
- Expenses (dollars)
- Graduation Rate (%)

Use cluster analysis to find groups of Universities (follow instructions for other datasets)