Multivariate Statistics (**MARS6300**) - Homework 2 **Name:** _____KEY_____

**Distributed:** Friday, February 6, 2018 **Due:** Thursday, February 22, 2018

**Instructions:** Copy and paste your answers below and turn in a word file and two excel files by the end of due day via email to khyrenba@gmail.com. Please use email title "MARS 6300 hw#1" and label all files with you're a suffix including your name (e.g., MARS6300_hw2_hyrenbach). Unlabeled emails / files will be penalized 10% of points.

You are free to use any reference materials of your choice. While you are encouraged to work together, make sure you turn your own assignment. This homework is worth 5 points. Make sure you leave the formulas showing all of your calculations in the excel file, and explain your reasoning, to get partial credit.

**The objectives of this homework are:**
A) To review and practice data transformations.
B) To calculate dissimilarities for a simple species / sample matrix.
C) To perform a clustering analysis.
D) To practice reporting the results of clustering analyses.

- Instruction file: "MARS6300_hw2.doc" (open with word file) – turn in
- "desserts_colors.xls" data file: (open with excel) – do not turn in
- "MossStems1M.WK1" data file: (open with PC-ORD) – do not turn in

## 1) Data Transformations:

A) Download and open the file "bMossStems1M.WK1"dataset (available on the SampleDatasets.zip folder). Summarize the data to look for "problematic" species distributions. Inspect "result.txt" file for the following information:

Copy and paste the kurtosis values for all 50 species below:

| | Skewness | Kurtosis |
|---|---|---|
| 1 Ancu | 5.146 | 27.716 |
| 2 Clad | 6.958 | 52.692 |
| 3 Cloc | 22.450 | 504.000 |
| 4 Clcr | 8.032 | 77.795 |
| 5 Deab | 18.061 | 343.594 |
| 6 Difu | 19.098 | 383.096 |
| 7 Disc | 9.227 | 94.995 |
| 8 Dita | 22.450 | 504.000 |
| 9 Euor | 4.886 | 27.398 |
| 10 Frbo | 17.016 | 329.079 |
| 11 Frni | 3.669 | 16.015 |

```
   12 Hofu        17.754   342.766
   13 Honu         8.714    83.939
   14 Hyci        22.450   504.000
   15 Hysu        12.700   168.171
   16 Ismy        -0.142    -1.349
   17 Leco        22.450   504.000
   18 Leme        15.869   268.165
   19 Lepo        18.043   341.240
   20 Loor        20.186   428.174
   21 Lopu        16.549   298.141
   22 Lobi        22.450   504.000
   23 Mete        22.450   504.000
   24 Meme         6.407    41.422
   25 Metz        11.358   147.858
   26 Mosi        20.810   447.755
   27 Nedo         1.137     0.400
   28 Nela        21.974   488.625
   29 Nere        22.450   504.000
   30 Oraf        22.450   504.000
   31 Orly         8.348    83.754
   32 Orob        22.450   504.000
   33 Orpu        19.092   390.884
   34 Pasu        11.609   144.303
   35 Peco        14.677   248.753
   36 Peme        22.450   504.000
   37 Plin        11.694   145.301
   38 Plve        22.450   504.000
   39 Plun        10.221   110.085
   40 Poco        12.358   163.791
   41 Pogl         9.206    96.639
   42 Pona         2.998     9.844
   43 Psan        12.700   168.171
   44 Rafa        11.884   142.604
   45 Rhgl        13.070   200.129
   46 Rhlo         6.435    46.719
   47 Spgl        16.549   288.930
   48 Stli        22.450   504.000
   49 Ulcr        20.263   433.650
   50 Usne        10.694   131.597
-----------------------------------
   Averages:      14.493   265.217
-----------------------------------
```

What is the average skewness (across all 50 species) ? 14.493

How many species have values > 1 or < -1? 49 species>1, 1 species<-1

We need to make some transformations to get the species abundance data to be more normally distributed. What transformation do you recommend? To answer this question, check the ranges of possible values (Hint: are there any negative values, are there values smaller than 1)?

There are many 0 values, but there cannot be any negative values. There are some values less than 1. Log transformation will not work on its own because there are many 0 values. Thus, a constant must be added.

What is the most numerous species? Ismy

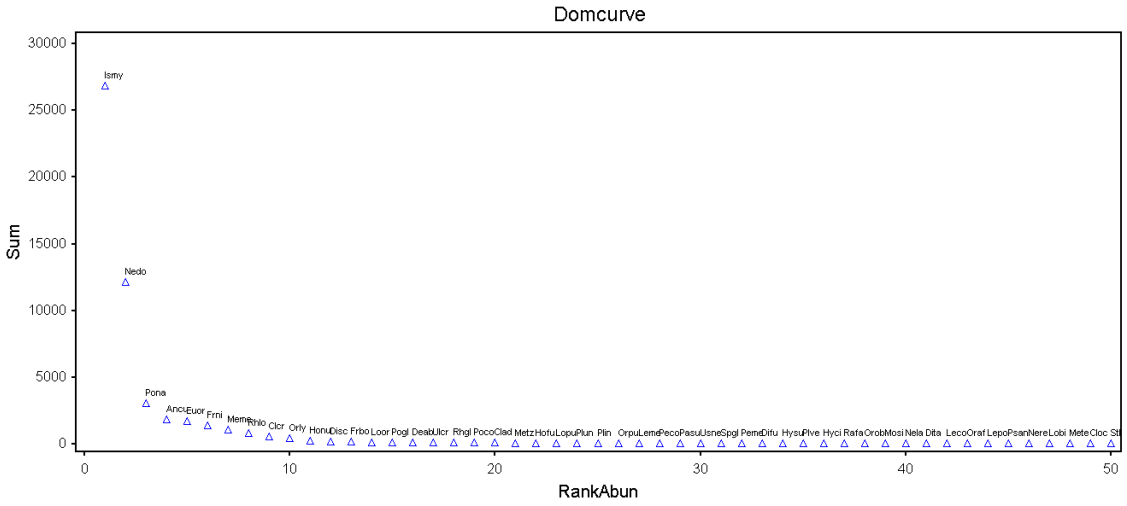
What is its maximum abundance? 100  (SUM = 26831.6)

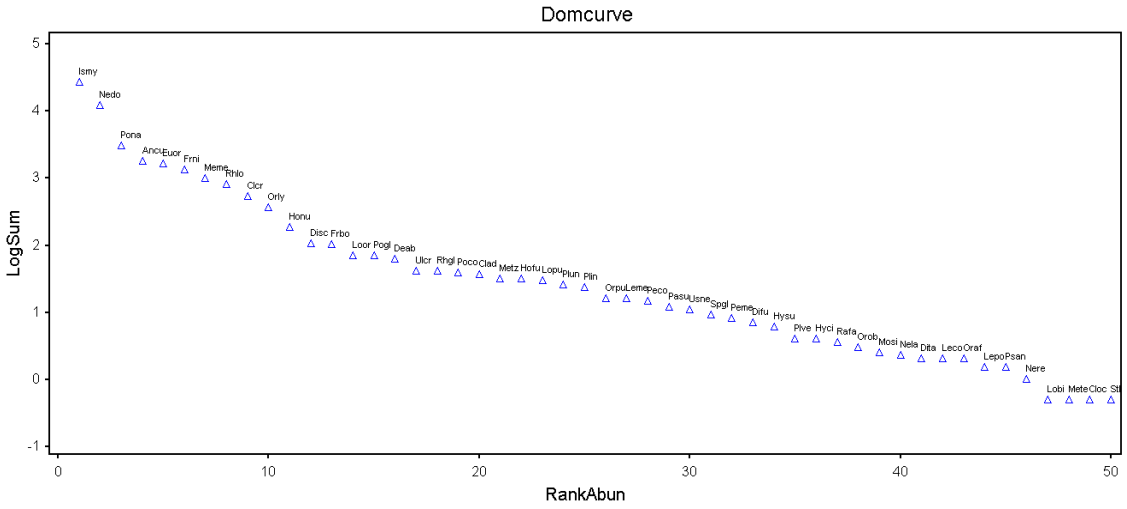
What is the least numerous species? Lobi, Mete, Cloc, and Stli (all four species are tied)

What is its maximum abundance? 0.5 (SUM = 0.5)

Note:  You can inspect the species abundances using the 'Graph > Dominance Curves" menu. Create two graphs of the sum and the log(sum) of species abundance versus species rank. Label the axes and the species.  Finally, using these figures and the output summary table as reference, discuss the relative abundance of the species in this sample.

(Hint:  append the new output to your "result.txt" file – it creates a log of your steps for you)



Let's try the log transform (of all the species data at once). Use the "Modify Data > Transformations > Logarithmic" menu (select: ln(x) with base 10). What happened? Why?

Error: Cannot transform 0's or negative numbers!

If the previous transformation did not work, you can fix this problem by adding a constant to all of the data. Think carefully about what value would you add. Base your answer on the results you obtained from the screening of the data.

Since the smallest values that are not zero are smaller than one the constant should be an order of magnitude smaller than the smallest value. I added .05 to all the values. Note, you could also add '1" to all the data, so log (0 + 1) = 0.

Once you have added a constant to all of the data (to all of the species), perform the log transform and re-summarize the data, to see what are the skewness values for the 50 species.

Copy and paste the kurtosis values for all 50 species (after the transformation) below:

| | Skewness | Kurtosis |
|---|---|---|
| 1 Ancu | 1.956 | 2.306 |
| 2 Clad | 4.797 | 22.041 |
| 3 Cloc | 22.450 | 503.993 |
| 4 Clcr | 3.299 | 9.617 |
| 5 Deab | 9.654 | 102.057 |
| 6 Difu | 16.103 | 261.159 |
| 7 Disc | 5.900 | 34.007 |
| 8 Dita | 22.450 | 503.996 |
| 9 Euor | 2.080 | 2.667 |
| 10 Frbo | 5.581 | 31.081 |
| 11 Frni | 0.962 | -0.865 |
| 12 Hofu | 11.832 | 142.670 |
| 13 Honu | 5.285 | 27.193 |
| 14 Hyci | 22.450 | 503.996 |
| 15 Hysu | 11.289 | 127.139 |
| 16 Ismy | -1.936 | 2.322 |
| 17 Leco | 22.450 | 503.996 |
| 18 Leme | 11.704 | 139.358 |
| 19 Lepo | 16.158 | 263.482 |
| 20 Loor | 13.239 | 177.197 |
| 21 Lopu | 11.907 | 146.234 |
| 22 Lobi | 22.450 | 503.993 |
| 23 Mete | 22.450 | 503.993 |
| 24 Meme | 5.132 | 24.954 |
| 25 Metz | 5.544 | 30.503 |
| 26 Mosi | 16.855 | 292.651 |
| 27 Nedo | -0.793 | -0.983 |
| 28 Nela | 18.216 | 348.188 |
| 29 Nere | 22.450 | 503.995 |
| 30 Oraf | 22.450 | 503.996 |

| | | | |
|---|---|---|---|
| 31 | Orly | 2.214 | 3.455 |
| 32 | Orob | 22.450 | 503.996 |
| 33 | Orpu | 11.820 | 143.598 |
| 34 | Pasu | 8.024 | 65.884 |
| 35 | Peco | 9.104 | 84.464 |
| 36 | Peme | 22.450 | 503.997 |
| 37 | Plin | 8.585 | 75.174 |
| 38 | Plve | 22.450 | 503.996 |
| 39 | Plun | 8.478 | 70.962 |
| 40 | Poco | 8.418 | 71.890 |
| 41 | Pogl | 5.559 | 29.824 |
| 42 | Pona | 0.487 | -1.453 |
| 43 | Psan | 11.469 | 132.487 |
| 44 | Rafa | 11.271 | 126.441 |
| 45 | Rhgl | 6.832 | 47.138 |
| 46 | Rhlo | 3.118 | 8.279 |
| 47 | Spgl | 13.382 | 181.429 |
| 48 | Stli | 22.450 | 503.993 |
| 49 | Ulcr | 9.289 | 90.199 |
| 50 | Usne | 7.337 | 53.923 |
| Averages: | | 11.391 | 188.332 |

What is the average skewness (across all 50 species)? 11.391

How many species have values > 1 or < -1? 46 species>1, 1 species<-1

Did the log transform decrease the skewness for all species? Explain why / why not?

The log transform did decrease the skewness. By adding a constant we dealt with all the 0 values, which were contributing greatly to the negative skewness. The log transform also dealt with the large tails (positive skewness). Altogether, this transformation allowed us to diminish the extreme positive skew (to the right); yet, the transformation did not solve the skewness of all the species involved, and many remain highly non-normal.

Finally, transform these species abundances into presence / absence data using another transformation. Which one would you use to achieve this? Explain what you are actually doing to transform these abundance values into a binary response: 0 (absent) and 1 (present).

To transform into a presence / absence matrix we have two options: (i) use the presence-absence transformation in the modify data menu, or (ii) use the power transformation (with power of 0). Both transformations will keep all 0 values unchanged (signaling the species was absent), while and all samples with a positive number will become a 1 (present).

**2) Distance measures:**

Import file "desserts colors.xls" (dataset 2.2 in the web-site) and use the data provided to explore distance measures:

- To begin, calculate the correlations between all of the samples (the people), in terms of their color and dessert choices. Use the "Summary > Write Distance Matrix" command to calculate the correlations. Make sure you request an output file – its easier to work with than the Result.txt screen.

  Open the file and report how many pairwise correlations are reported: __91_____

  Show why this number of pairwise matches makes sense: ___14 * 13 / 2_____

  Calculate and report the following statistics for the correlation:

| | |
|---|---|
| mean | 0.131 |
| STD | 0.183 |
| min | -0.282 |
| max | 0.600 |

Calculate and report the following statistics for the correlation-based distance:

| | |
|---|---|
| Mean | 0.434 |
| STD | 0.092 |
| Min | 0.200 |
| Max | 0.641 |

Use the formula provided in the lecture to convert the maximum and the minimum correlation coefficients into correlation-based distance measures (r distance):

$$r_{distance} = (1 - r)/2$$

| Correlation Coefficient | r distance |
|---|---|
| Maximum r = + 0.600 | 0.200 |

| | |
|---|---|
| Minimum r = - 0.282 | 0.641 |

What happened to the correlations showing the two most correlated and the two least correlated pairs of samples. Explain: The maximum correlation (most similar samples) yielded the minimum r distance value (least dissimilar samples). The minimum correlation (least similar samples) yielded the maximum r distance value (most dissimilar samples).

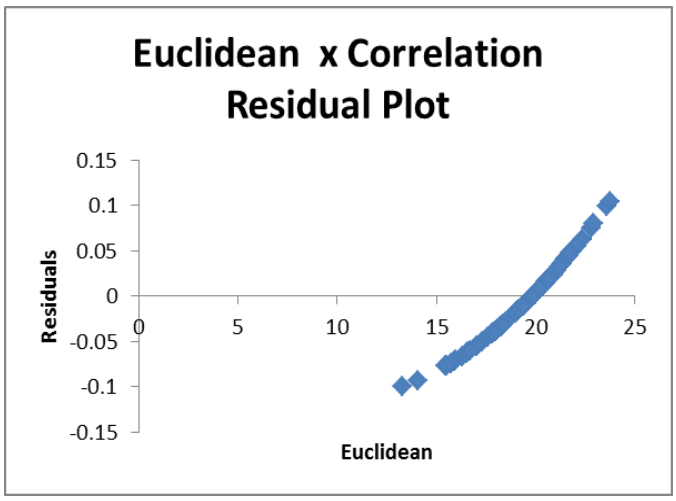
Next, calculate the following statistics for four distance measures:

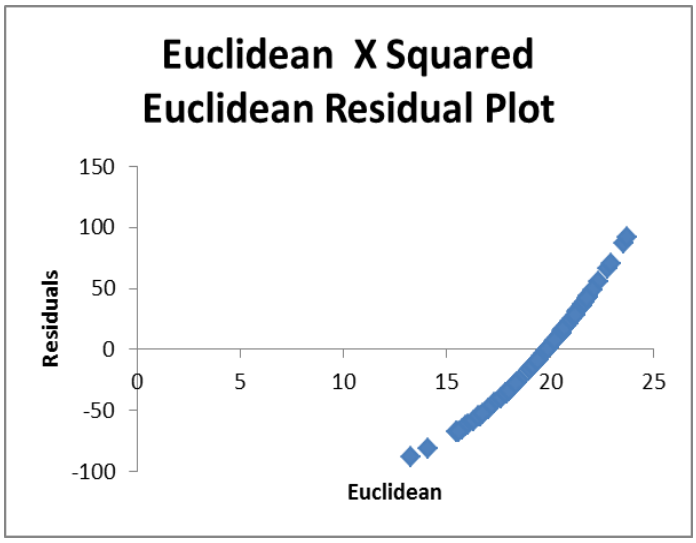
| | Correlation | Euclidean | SQ_Euclidean | Rel_Euclidean |
|---|---|---|---|---|
| mean | 0.434 | 19.435 | 382.176 | 1.310 |
| STD | 0.092 | 2.122 | 80.541 | 0.143 |
| min | 0.200 | 13.266 | 176.000 | 0.894 |
| max | 0.641 | 23.749 | 564.000 | 1.601 |

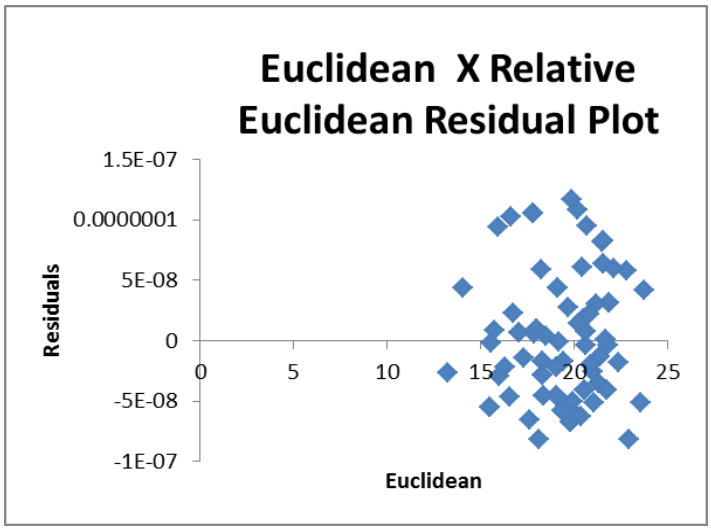
Using the same pair of samples, make three separate regressions of the Euclidean distances (x axis) versus the distances of: (i) correlation distance, (ii) Squared Euclidean distances, and (iii) Relative Euclidean distances. For each regression, fit the best-fit line (using linear regression) and discuss how the distances are distorted. Hint: Ask for the residual plot and paste it here. Can you see any patterns in the residuals?

Fill in this table:

| **Distance Measure Comparison** | **Adjusted R squared (from regression)** | **Pattern in residuals**<br><br>**(Do they appear random: yes / no?)** |
|---|---|---|
| Euclidean X Correlation | 0.994619 | No – curvilinear pattern |
| Euclidean X SQ Euclidean | 0.979116 | No – curvilinear pattern |
| Euclidean X Rel Euclidean | 0.988889 | YES |


Euclidean x Correlation Residual Plot
Residuals
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
0
5
10
15
20
25
Euclidean


Euclidean X Squared Euclidean Residual Plot
Residuals
150
100
50
0
-50
-100
0
5
10
15
20
25
Euclidean


Euclidean X Relative Euclidean Residual Plot
Residuals
1.5E-07
0.0000001
5E-08
0
-5E-08
-1E-07
0
5
10
15
20
25
Euclidean

Finally, calculate the Sorensen distance matrix for the "desserts colors" dataset.

What happened? Why?


It would not work because we have negative numbers in the dataset.

**3) Clustering analysis of a small dataset:**

A) Enter this main data matrix, showing the abundance of two species on four different sample plots, into PC ORD. You can use "File > New > Main Matrix" menu. Note: remember the formatting conventions for rows and columns, and add sample numbers and species numbers. Save data set for later using "Save As > Main" menu.

B) Create scatterplot, showing the four samples in "species space". Label each sample. Paste figure below. Using this figure, describe which two samples are "closer" to each other? Why?

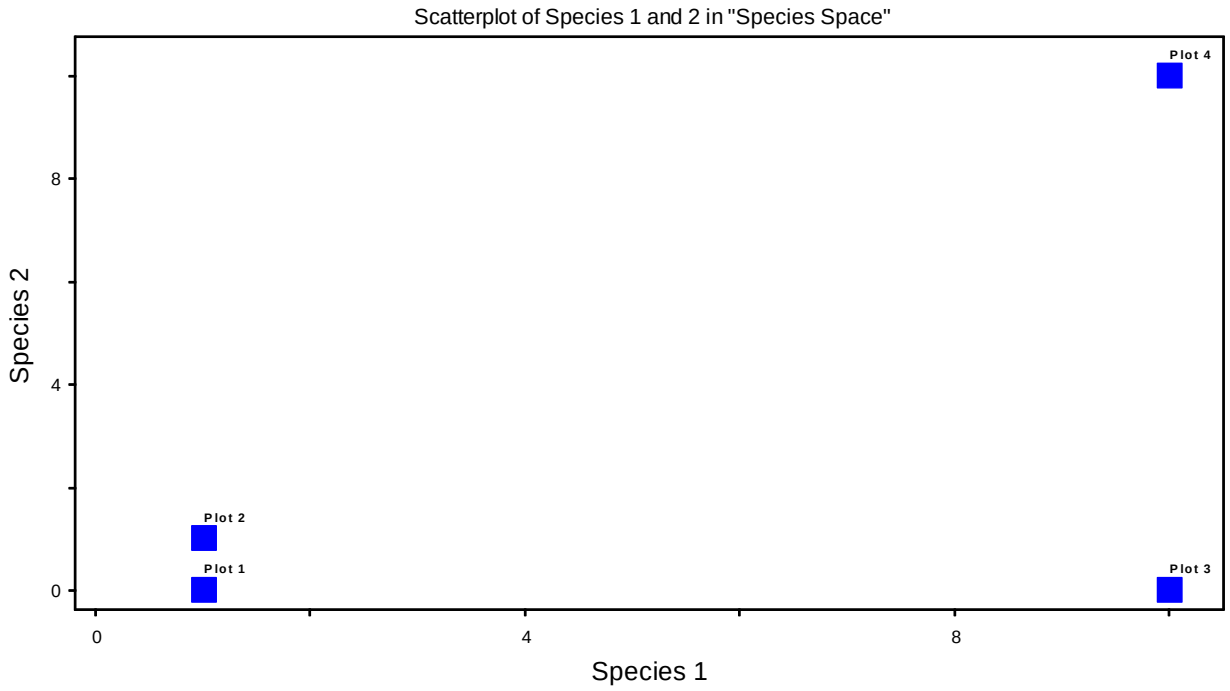



*Figure 2. Scatterplot of species 1 and 2 in "species space."*

The 2 samples that are "closest" together in "species space" are Plot 1 and Plot 2. This is because the distance between these two samples is much smaller than the distance between the combination of any 2 other samples.

Using this same logic, draw a diagram (bubbles and arrows) showing the way you think the clustering of these samples would look like. Do not worry about the size of the x and the y axis (the distances), I just want you to show me the order in which samples would be added to the dendrogram (Hint: look at slide 3 from lecture 8). Create figure with ppt, draw, or else. You can even draw figure manually and insert digital image below:

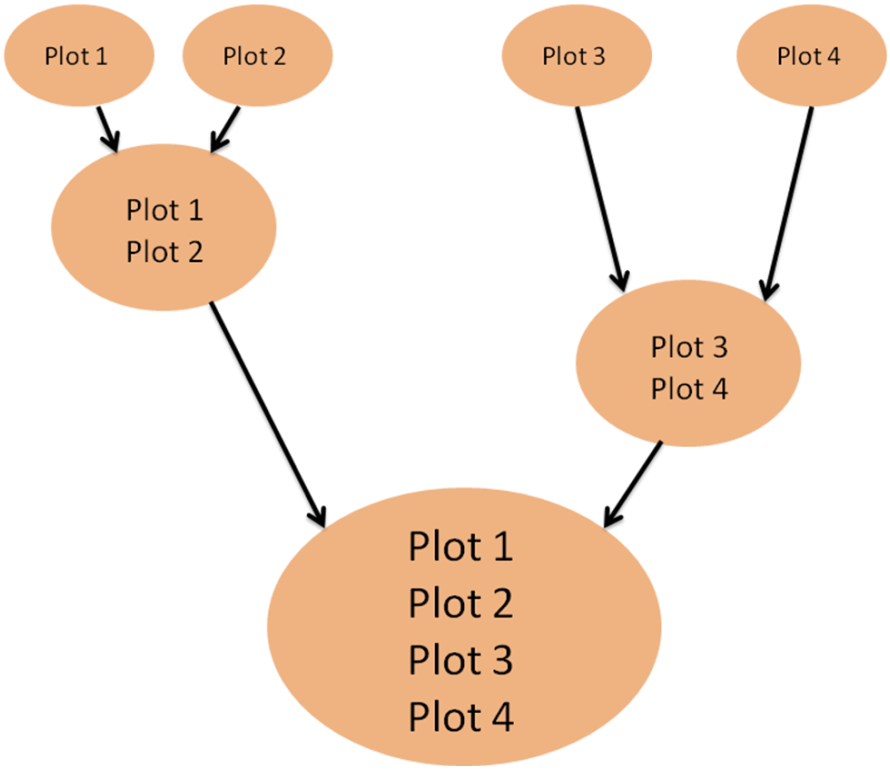



*Figure 3. Diagram showing the clustering of Plots 1, 2, 3, and 4. As plots 1 and 2 are closest together (Euclidian distance), these 2 would come together first. Next, plots 3 and 4 would join as they are closest together along the species 1 axis, but not as close together along the species 2 axis as plots 1 and 2 were. Finally, plots 1+2 will join with plots 3+4 to make a big cluster of all 4 plots.*

C) Calculate the Euclidean distances between plot 1 and plot 2, between plot 1 and plot 3, and between plot 1 and plot 4. Hint, use formula from lecture 6 (dimensions = k = 2). Report the Euclidean distances below:

The equation for the Euclidian distance is: $D= (x^k + y^k)^{(1/k)}$. Here we are saying that k = 2, meaning that we are taking the square root of $(x^k + y^k)$. "X" represents the horizontal distance between the 2 points (parallel to the x-axis), and "Y" represents the vertical distance between the 2 points (parallel to the y-axis).

plot 1 vs plot 2: First we must find the values of X and Y. As the value of plot 1 is (1,0) and the value of plot 2 is (1,1), then the x-value = $x_2-x_1$ = 1-1 = 0. The y-value = $y_2-y_1$ = 1-0 = 1. So, the equation for Euclidian distance = $D = (0^2 + 1^2)^{(1/2)}$ = **1 = Euclidian distance between Plot 1 and Plot 2**.

plot 1 vs plot 3: First we must find the values of X and Y. As the value of plot 1 is (1,0) and the value of plot 3 is (10,0), then the x-value = $x_2$-$x_1$ = 10-1 = 9. The y-value = $y_2$-$y_1$ = 0-0 = 0. So, the equation for Euclidian distance = D = $(9^2 + 0^2)^{(1/2)}$ = **9 = Euclidian distance between Plot 1 and Plot 3**.

plot 1 vs plot 4: First we must find the values of X and Y. As the value of plot 1 is (1,0) and the value of plot 3 is (10,10), then the x-value = $x_2$-$x_1$ = 10-1 = 9. The y-value = $y_2$-$y_1$ = 10-0 = 10. So, the equation for Euclidian distance = D = $(9^2 + 10^2)^{(1/2)}$ = **13.45 = Euclidian distance between Plot 1 and Plot 4**.

Next, use PC ORD to calculate the dissimilarity matrix using the Euclidean distance measure. Use "Summary > Write Distance matrix" menu. Request both a "result.txt" file and a "full matrix" as a spreadsheet (.wk1) so you can re-load it into PC ORD). Copy and paste table of pair-wise distances from result.txt in table provided below:

Distance measure = Euclidean (also called Pythagorean)

D I S T A N C E M A T R I X (Euclidean Distances)

| | **P1** | **P2** | **P3** | **P4** |
|---|---|---|---|---|
| **P1** | 0.000 | 1.000 | 9.000 | 13.45 |
| **P2** | 1.000 | 0.000 | 9.055 | 12.73 |
| **P3** | 9.000 | 9.055 | 0.000 | 10.00 |
| **P4** | 13.45 | 12.73 | 10.00 | 0.000 |

*Table 1. Distance matrix (Euclidian distances) between all combinations of points, calculated by PC ORD.*

Check your answers against the values calculated by PC ORD.
The distances I calculated were the same as calculated by PC ORD! GREAT

Remember: Clustering uses squared distances. Calculate the squared distances ($d^2$) below:

Distance measure = $d^2$ = (Euclidean) $^2$

To fill in the values on the following table, each value in Table 1 was squared. YES

D I S T A N C E M A T R I X (Euclidean Distance Squared)

| | **P1** | **P2** | **P3** | **P4** |
|---|---|---|---|---|
| **P1** | 0.000 | 1.000 | 81.00 | 180.90 |
| **P2** | 1.000 | 0.000 | 81.99 | 162.05 |
| **P3** | 81.00 | 81.99 | 0.000 | 100.00 |
| **P4** | 180.90 | 162.05 | 100.00 | 0.000 |

*Table 2. Distance matrix (Euclidian distance squared) between all combinations of points. PC ORD was used to calculate the Euclidian distances (Table 1), then each value was squared from Table 1 to create the values in Table 2.*

D) Next, you will combine the data of plot 1 and plot 2 because they are the most similar. Calculate the mean number of species 1 and species 2 in the union of plot 1 and plot 2.
Mean_species1 (mean of plot 1 and plot 2): Species 1, plot 1 = 1. Species 1, plot 2 = 1.
Mean_species1 = (1+1)/2 = 1. YES

Mean_species2 (mean of plot 1 and plot 2): Species 2, plot 1 = 0. Species 2, plot 1 = 1.
Mean_species2 = (1+0)/2 = 0.5. YES

Calculate E resulting from creation of this group (union of plot 1 and plot2) using this equation:

$$E_1 = \sum_{i=1}^{2}\sum_{j=1}^{2}(x_{ij1} - \bar{x}_{j1})^2$$

Where the number "1" means this is the first "step" in the clustering procedure.
And where i is the sample (plot 1 or 2) and j is the species (1 or 2).
Hint: Remember, you just calculated averages for both species abundance across both samples.
Write out the terms you are summing to get E1 (above) and report the value:
$E_1 = (x_{1,1} - \bar{x}_1)^2 + (x_{2,1} - \bar{x}_1)^2 + (x_{1,2} - \bar{x}_2)^2 + (x_{2,2} - \bar{x}_2)^2$
$E_1 = (1\text{-}1)^2 + (1\text{-}1)^2 + (0\text{-}0.5)^2 + (1\text{-}0.5)^2$
$E_1 = 0 + 0 + 0.25 + 0.25$
**$E_1 = 0.5$ GREAT**

E) Now that you created the first cluster (union of 1 and 2), we need to calculate the distances between this cluster and the other two samples (p3 and p4) using the Ward linkage method.
Use the formulas below:

$$d_{ir}^2 = \alpha_p d_{ip}^2 + \alpha_q d_{iq}^2 + \beta d_{pq}^2 + \gamma |d_{ip}^2 - d_{iq}^2|$$

where values of $\alpha_p$, $\alpha_q$, $\beta$, and $\gamma$ determine the type of sorting strategy (Use Ward coefficients, listed below).

| | Coefficient | | | |
|---|---|---|---|---|
| **Linkage method** | $\alpha_p$ | $\alpha_q$ | $\beta$ | $\gamma$ |
| **Ward's method** | $\frac{n_i + n_p}{n_i + n_r}$ | $\frac{n_i + n_q}{n_i + n_r}$ | $\frac{-n_i}{n_i + n_r}$ | 0 |

**NOTE:** n is the number of samples (n = 1 unless you are looking at a group)
Subscript "i" refers to the remaining plots not in the group (plot 3 or plot 4) (Hint: ni = 1)
Subscript "r" refers to the group you created (union of plot 1 and plot 2) (Hint: nr = 2)
Subscript "p" refers to plot 1 Subscript "q" refers to plot 2

So:

$$\alpha_1 = \frac{1+1}{1+2} = \frac{2}{3} \qquad \alpha_2 = \frac{1+1}{1+2} = \frac{2}{3}$$

$$\beta = -\frac{1}{3} \qquad \gamma = 0$$

For example, for objects 1, 3 and 4:

$d^2_{ip}$ = distance between plot 1 and plot 3: 81.0 from Table 2

$d^2_{iq}$ = distance between plot 1 and plot 4: 180.9 (181) from Table 2

$d^2_{pq}$ = distance between plot 3 and plot 4: 100.0 from Table 2

For example, for objects 3 and 1+2:

$d^2_{ip}$ = distance between plot 1 and plot 3: 81.00 from Table 2
$d^2_{iq}$ = distance between plot 2 and plot 3: 81.99 (82) from Table 2
$d^2_{pq}$ = distance between plot 1 and plot 2: 1.000 from Table 2

So, the squared distance between plot 3 and plot 1+2 (the union of plot 1 and 2) is:

Go to your table of squared distances and extract the following values. Paste values below:

$$d^2_{3,1+2} = \frac{2}{3}(81) + \frac{2}{3}(82) - \frac{1}{3}(1) = \frac{325}{3} = 108.3$$

Enter coefficients and distances into the same equation shown above to calculate the squared distance between plot 4 and plot 1+2. Show the sum for calculating the distance squared below:

$$d^2_{4,1+2} =$$

$= (2/3)(d^2_{ip}) + (2/3)(d^2_{iq}) - (1/3)(d^2_{pq}) + (0)(d^2_{ip} - d^2_{iq})$
$= (2/3)(180.9) + (2/3)(162.05) - (1/3)(1) + 0$
$= 120.6 + 108.03 - 0.333$
**= 228.297**

(Hint, I show you the other squared distances – in matrix below - for reference)

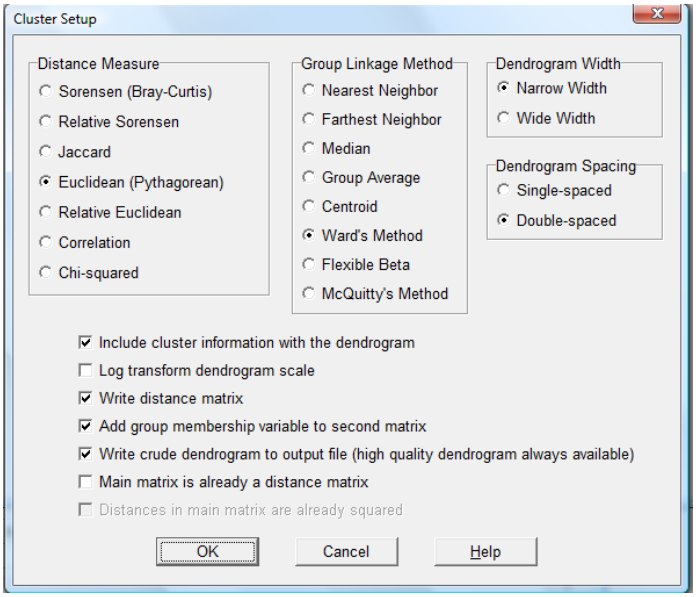
Revised distance matrix after the first fusion.

| | Plots 1+2 | Plot 3 | Plot 4 |
|---|---|---|---|
| Plots 1+2 | 0 | 108.3 | |
| Plot 3 | 108.3 | 0 | 100 |
| Plot 4 | | 100 | 0 |

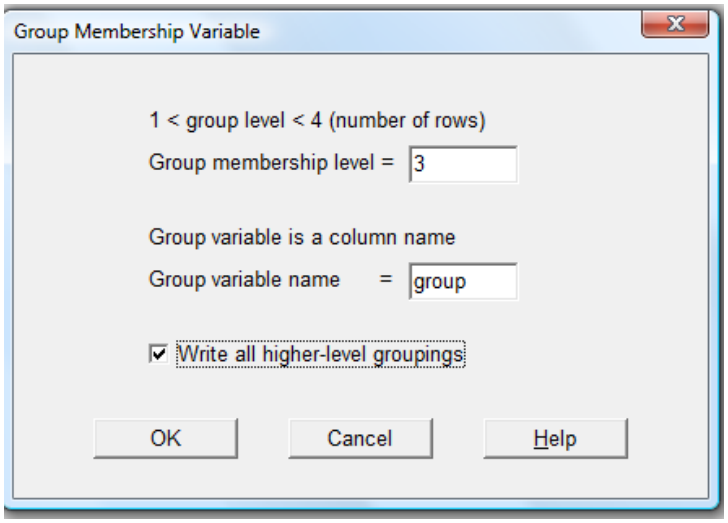
**4) Now, you will let PC ORD do all of this for you.**
Load the original dataset you created previously back into PC ORD, and run clustering analysis using "Groups > Cluster Analysis" menu.

| | Sp1 | Sp2 |
|---|---|---|
| Plot 1 | 1 | 0 |
| Plot 2 | 1 | 1 |
| Plot 3 | 10 | 0 |
| Plot 4 | 10 | 10 |

Use the following set up:




Once




you run clustering analysis, PC ORD will create a graph.

Inspect graph after looking at the "result.txt" file.
Copy and paste dendrogram – from graphs. Use "Graph> Dendrogram" menu to look at figure.
(Hint: play with graph settings, and try "save as" and "copy / paste" menus)

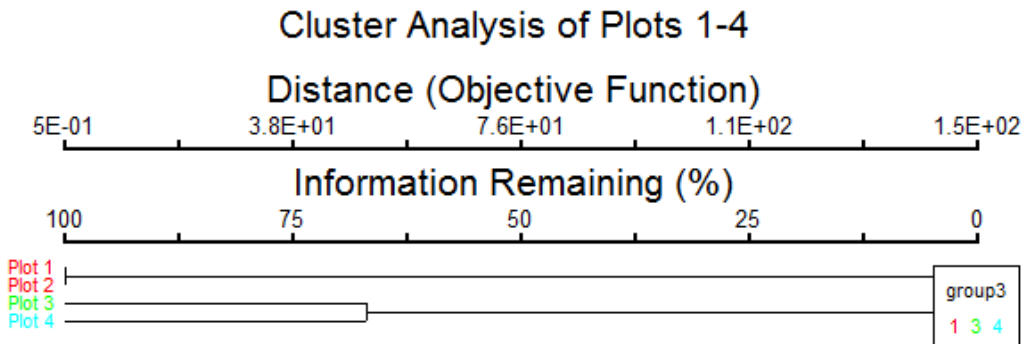



*Figure 4. Dendrogram of cluster analysis of Plots 1-4, created using PC ORD.*

**5) Review of Clustering Papers:**

Read the three papers on clustering and report the following information, pertaining to "hierarchical clustering analyses", for each paper. Please be brief but provide enough detail. A word or a sentence will suffice. NOTE: please do not report results from K-means clustering.

A) He et al. 1997:

- Describe data Matrix: how many samples / species, what are these data?

Samples: 46,961 long-lining sets. Species: 11 species or species groups (fish or fishes)

- List data transformations and explain rationale:

Arcsin-square-root transformed, in order to normalize the distribution.

- Distance Measure employed:

Euclidian Distance

- Linkage Method employed:

Ward's method

- If dendrogram plotted, was it re-scaled, explain what method used:

Dendrogram was plotted, and rescaled so the smaller distance was on the right (starting at 0.0) and the larger distance was on the left (0.2). Did not mention which method was used or why

- If groups were defined, was rule for "pruning" the tree provided:

Authors expected the 2 clusters of tuna and swordfish sets, but made more clusters to allow other categories to be seen; these extra clusters were counted until each one had less than 10% of the total number of sets

- If dendrogram was plotted, did it display the amount of information retained:

No, dendrogram does not display amount of information retained.

- Finally, were the resulting groups characterized? How?

Yes, into 5 resulting groups – Clusters 1-3 had large catches of bigeye tuna, clusters 4-5 caught more swordfish: Cluster 1 had most bigeye and yellowfin and also high swordfish; Cluster 2 had lots of bigeye, other fishes, and striped marlin; Cluster 3 had an even mixture of tuna and

swordfish, but lots of mahimahi; Cluster 4 had swordfish catches and blue shark catches; Cluster 5 had swordfish catches, and little blue shark catches. Final groups were characterized by catch composition, characteristic of fishing operation, spatial distribution, and CPUE.

B) Wolter 1987:
- Describe data Matrix: how many samples / species, what are these data?
Samples: 391 time series; Species: 8 different meteorological element-seasons (e.g. SST, SLP, wind speed, cloudiness in either summer or winter, and in the Atlantic, Indian, or Pacific oceans)

- List data transformations and explain rationale:
No transformation as data are already normally distributed

- Distance Measure employed:
Correlation

- Linkage Method employed:
Average linkage method

- If dendrogram plotted, was it re-scaled, explain what method used:
No dendrogram plotted.

- If groups were defined, was rule for "pruning" the tree provided:
They used Monte Carlo simulations to find whether clusters were significant using the (95% field significance, or 95% noise threshold). A cluster remained on a certain level of complexity if there were other clusters of LOWER levels that merge into it. The authors used four complexity levels to differentiate between different oceanwide clustering groups.

- If dendrogram was plotted, did it display the amount of information retained:
No dendrogram plotted

- Finally, were the resulting groups characterized? How?
Clusters were characterized as "strong" or "weak", if they exceed the 95% threshold or not. The clusters or groups were characterized spatially by how many 5 degree squares they contain (that helps to best describe the relationship between element-seasons and location or ocean basin). Groups were also characterized by characteristics of SLP, cloudiness, and SST based on the SOI were also characterized by similarities of SST, SLP, cloudiness, and wind, and location. For example, P2 is a strong SST and SLP cluster in the pacific.

C) Diehr et al. 1982:
- Describe data Matrix: how many samples / species, what are these data?
Samples: 726 headache patients; Species = 57 headache symptoms, reduced down to 19 used in the analysis.

- List data transformations and explain rationale:

None reported, but the symptom values were standardized so that each symptom had a standard deviation of 1.

- Distance Measure employed:
Sum of the squared distances = dissimilarity between symptoms in 2 patients; within group sum of squares (WGSS) = Euclidian Distance

- Linkage Method employed:
Never mentioned, but one can assume Ward's method as Euclidian Distance was used.

- If dendrogram plotted, was it re-scaled, explain what method used:
No dendrogram plotted, just a hypothetical diagnostic chart.

- If groups were defined, was rule for "pruning" the tree provided:
When split into 2 clusters, the tree was pruned in that the first 4 symptoms were more common in cluster 1 (tension headaches), and the rest of the symptoms were found more frequently in cluster 2 (migraine headaches).

- If dendrogram was plotted, did it display the amount of information retained:
No dendrogram plotted.

- Finally, were the resulting groups characterized? How?
When split into 2 clusters, the resulting groups were either migraine or tension headaches. However, it was also split into 8 clusters for more specific headache diagnoses. When split into 8 clusters, the tree was pruned again by how common certain symptoms were described in each patient.

**6) Perform Clustering Analysis:**

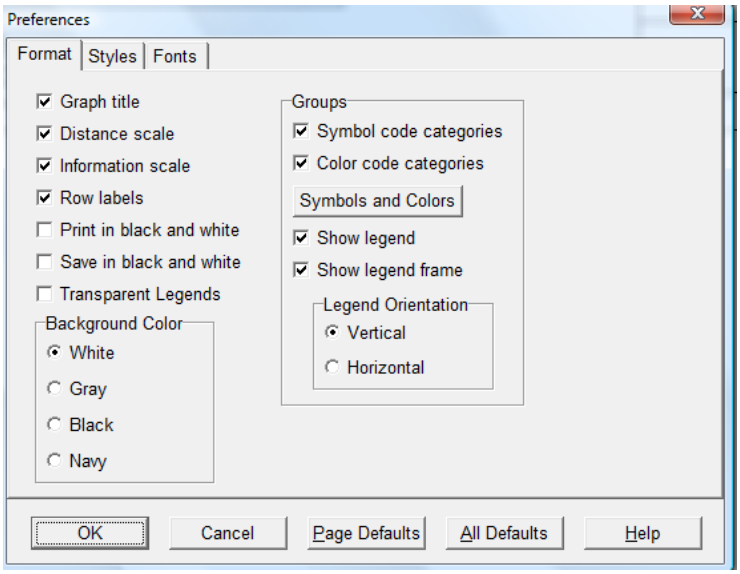
Analyze "dessert color" dataset. Download and import "desertscolors.xls" file using "File > Import Matrix > Main" menu. Note: data are on first sheet. (This is dataset 2.2 from web-site).

Calculate Euclidean distance and perform a clustering analysis with the Ward Method.

Request same set up as you used in question 4, but request 13 groups (instead of 3, as you did before). Make sure you select "write all higher level groupings".

Why is the maximum number of groups 13? Explain: This is because there are 14 samples, and the maximum number of groups must be the number of samples minus 1, otherwise there wouldn't be any groups, just samples. Here, n-1 = 13.

Look at the graph of the dendrogram and make sure you select “groups to be labeled” under the “Options> Preferences” menu.




Cycle through the different groups, using the “Groups > Select Grouping Variable” menu.

What is the “group” grouping variable showing you?
(Hint: How does the figure change as you increase the “grouping variable”).

This grouping variable defines the how the rest of the groups are defined. This organizes each group by similarity one step at a time as the number of the grouping variable increases. It seems that changing the “grouping” variable changes how each different sample is related to other samples. For instance, if “group2” is selected as the grouping variable, then all 14 samples are grouped into 2 separate groups, as indicated by a specific color and symbol. The dendrogram itself does not change (the lines connecting separate samples), but shows how all samples would be related if we only wanted to use 2 groups rather than 13. If “group13” is selected, then it assumes we want to keep all 14 samples as separate entities.

Open the “result.txt” file and look for the levels for the various cluster cycles.
For instance, the level is 88 (`8.8000E+01`)`for cluster 1 and 2484.143 for cycle 13`

```
------------ C L U S T E R   C Y C L E     1 -------------
 Combined group    12   into group     4   at level 8.8000E+01

------------ C L U S T E R   C Y C L E    13 -------------
 Combined group     7   into group     1   at level 2.4841E+03
```

Copy and paste the values into the table below:

| cycle number | Level | %_explained | %_info_remaining |
|---|---|---|---|
| 1 | 88 | 3.54 | 96.46 |
| 2 | 187 | 7.527756 | 92.47224 |
| 3 | 308 | 12.39866 | 87.60134 |
| 4 | 435 | 17.51109 | 82.48891 |

| | | | |
|---|---|---|---|
| 5 | 568.33 | 22.87834 | 77.12166 |
| 6 | 725.33 | 29.19843 | 70.80157 |
| 7 | 884.33 | 35.59904 | 64.40096 |
| 8 | 1078 | 43.3953 | 56.6047 |
| 9 | 1274.9 | 51.32158 | 48.67842 |
| 10 | 1484.4 | 59.75509 | 40.24491 |
| 11 | 1754.3 | 70.62001 | 29.37999 |
| 12 | 2107.5 | 84.81809 | 15.18191 |
| 13 | 2484.143 | 100 | 0 |

To determine %_explained value, the "level" value should be divided by the largest "level" value, where 100% of the information is explained, then multiplied by 100. The %_info_remaining is found by subtracting the %_explained value from 100.

At what cycle did we explain over 50% of the pattern? What was the exact $r^2$?
This occurred at cycle #9. As $r^2$ is defined, in this case, as the value ranging from 0 to 1 (or 0 to 100%) which shows the % of information explained, the exact $r^2$ value at cycle #9 = 0.513.

If cycle 1 creates two groups, how many groups did we have at the cycle when over 50 of the pattern was explained?
As this is cycle #9, this means there could be up to 10 groups at this point. With each new cycle, a new group is being added.

However, at the 40% level, there are 5 groups (see dendrogram below)

Select that number of groups using the "Groups > Select Grouping Variable" menu, and copy and paste the figure below.

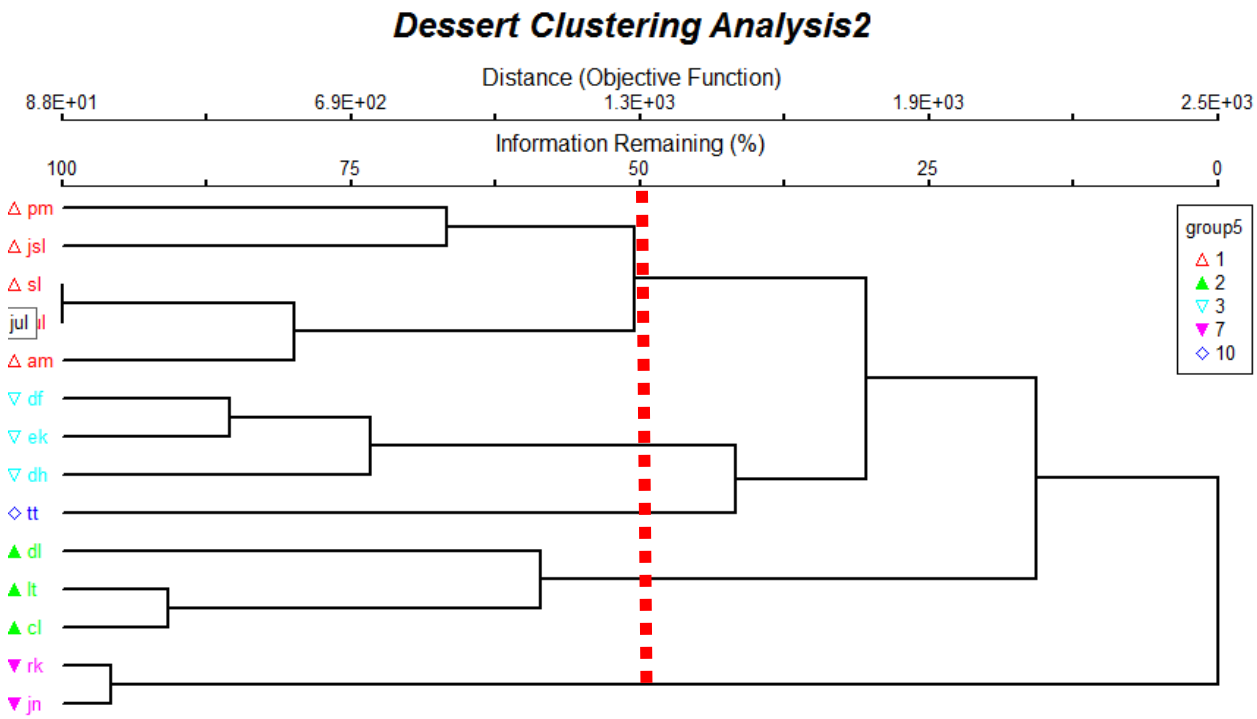



Figure 5. Cluster analysis dendrogram of dessert and color preferences of 14 students in class. The clustering analysis shows a grouping variable of 5, which was the point at which the "50% of information retained" threshold was crossed.