STSCI 5010, Homework 5

Pete Rigas (pbr43 cornell.edu)

November 27, 2020

---

1 Problem: Variable selection procedure

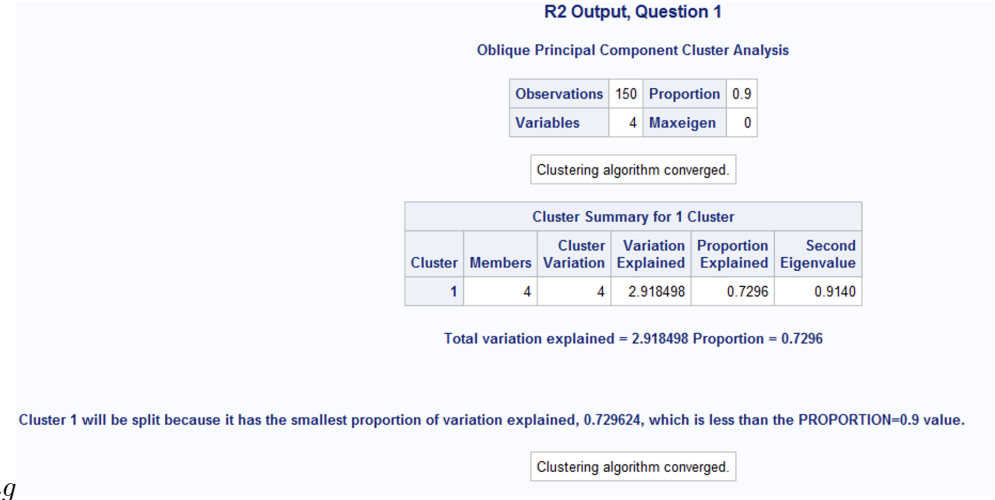
1.1 SAS Code

```
libname HW5 '\\rschfs1x\usercl\pbr43_STSCI5010\Desktop';

title "R2 Output, Question 1";
proc varclus data = HW5.flowers proportion = 0.9 outtree = tree;
var sl sw pl pw;
run;
```

1.2 SAS Code Output

- We have SAS output the proportions from the $1 - R^2$ ratio. Notably, we observe that the clusters that we would predict to see from the flowers data are 2, or 3 because from the cluster analysis plot later in the question, we see that the largest variance is observed through the sl and sw clusters. Other plots will show additional results from the clusters as we plot the PCA.



R2 Output, Question 1

Oblique Principal Component Cluster Analysis

Observations	150	Proportion	0.9
Variables	4	Maxeigen	0

Clustering algorithm converged.

Cluster Summary for 1 Cluster

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	4	4	2.918498	0.7296	0.9140

Total variation explained = 2.918498 Proportion = 0.7296

Cluster 1 will be split because it has the smallest proportion of variation explained, 0.729624, which is less than the PROPORTION=0.9 value.

Clustering algorithm converged.

(44).*png*

2 Clusters		R-squared with		
Cluster	Variable	Own Cluster	Next Closest	1-R**2 Ratio
Cluster 1	sl	0.8675	0.0138	0.1344
	pl	0.9690	0.1836	0.0379
	pw	0.9332	0.1340	0.0771
Cluster 2	sw	1.0000	0.1021	0.0000

Standardized Scoring Coefficients		
Cluster	1	2
sl	0.33627	0.00000
sw	0.00000	1.00000
pl	0.35541	0.00000
pw	0.34879	0.00000

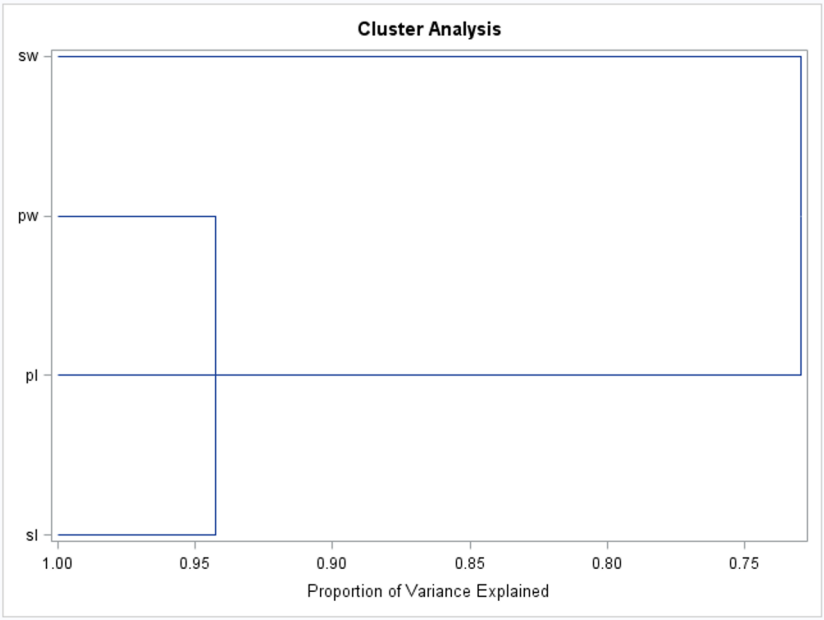
Cluster Structure		
Cluster	1	2
sl	0.93139	-0.11757
sw	-0.31951	1.00000
pl	0.98439	-0.42844
pw	0.96605	-0.36613

Inter-Cluster Correlations		
Cluster	1	2
1	1.00000	-0.31951
2	-0.31951	1.00000

(43).*png*

No cluster meets the criterion for splitting.

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	2.918498	0.7296	0.7296	0.914030	0.2117	
2	3.769741	0.9424	0.9232	0.199412	0.8675	0.1344



(42).*png*

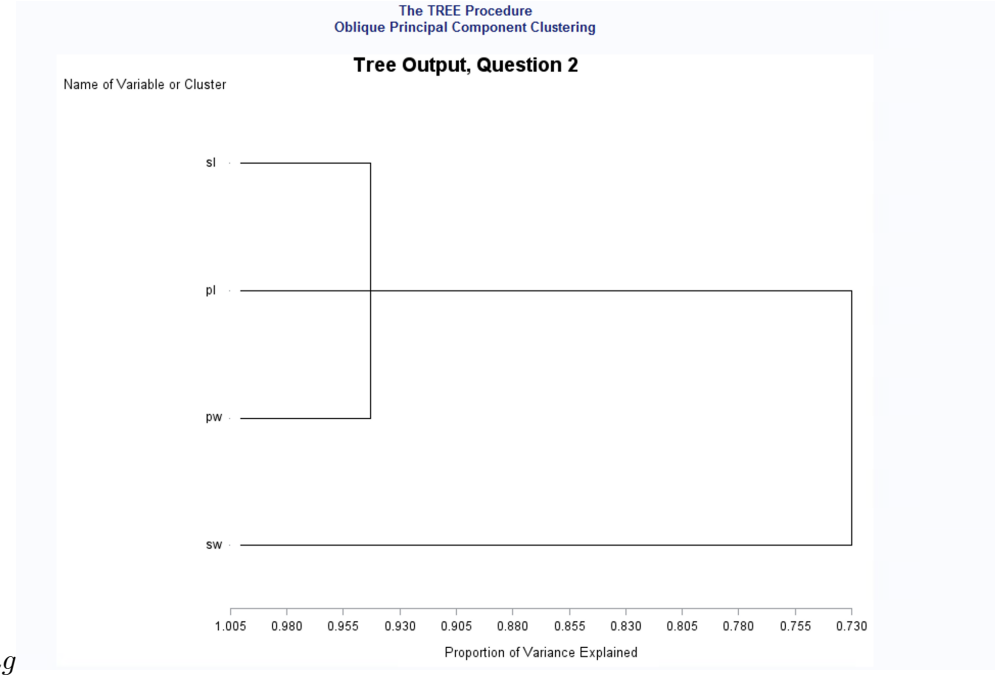
2 Problem: Producing a horizontal tree

2.1 SAS Code

```
    title "Tree Output, Question 2";
proc print data = tree;
run;
/* We will now use the tree SAS procedure */
proc tree data = tree horizontal;
height _propor_;
run;
```

2.2 SAS Code Output

- We have the code below output the tree.



(41).*png*

3 Problem: PCA Analysis

3.1 SAS Code

```
    title "Output, Question 3";
title2 "Principal Components";
proc princomp data = HW5.flowers out = flowerpca;
var sl sw pl pw;
run;
/* From the PCAs above, we will now produce
each of the color coded plots, with the code
below. */
title3 "Flowers Data Set PCA";
proc gplot data= flowerpca;
plot prin2*prin1=species;
symbol1 v = square color = red;
symbol2 v = star color = green;
symbol3 v = star color = blue;
run;
quit;
```

3.2 SAS Code Output

- The output below demonstrates the correlation matrix which can help our interpretation of how many clusters we expect to see from the data. From the eigenvalues and plots of the 3 clusters from the Flower Set PCA, we observe that the 3 clusters that we have labeled, are visible. From the cluster that is the most to the left, we observe that the cluster associated to these points is quite distinct, and that if we were to assign some point as the center of this cluster, that the points, with respect to a suitably defined metric, would all appear to be the closest to this point only. However, for the remaining 2 clusters that I have plotted, it appears that the points displayed with the green and blue stars most likely appear to be distinct. Again, because we cannot tell which number of clusters that the data is giving is necessarily correct, we would say from this clustering method that the most likely number of clusters is 2 or 3.

Output, Question 3
Principal Components

The PRINCOMP Procedure

Observations	150
Variables	4

Simple Statistics

	sl	sw	pl	pw
Mean	5.843333333	3.057333333	3.758000000	1.199333333
StD	0.828066128	0.435866285	1.765298233	0.762237669

Correlation Matrix

	sl	sw	pl	pw
sl	1.0000	-.1176	0.8718	0.8179
sw	-.1176	1.0000	-.4284	-.3661
pl	0.8718	-.4284	1.0000	0.9629
pw	0.8179	-.3661	0.9629	1.0000

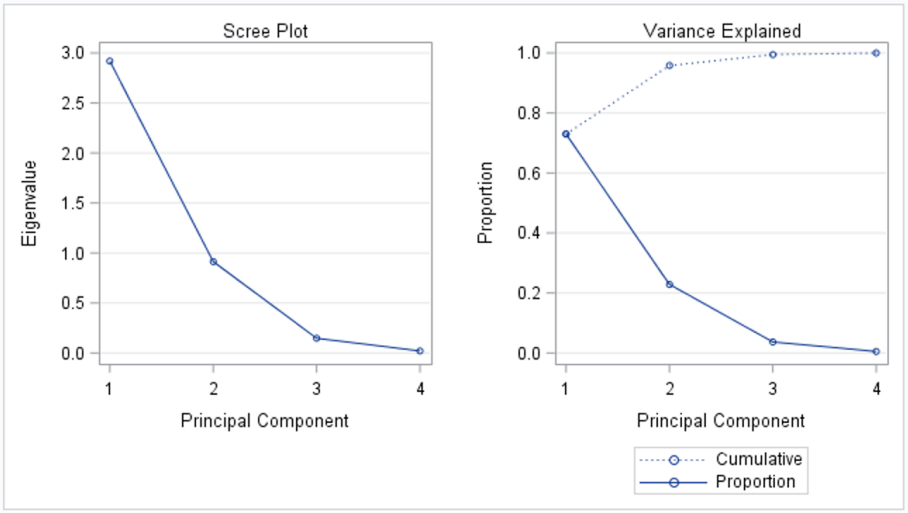
Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.91849782	2.00446735	0.7296	0.7296
2	0.91403047	0.76727360	0.2285	0.9581
3	0.14675688	0.12604204	0.0367	0.9948
4	0.02071484		0.0052	1.0000

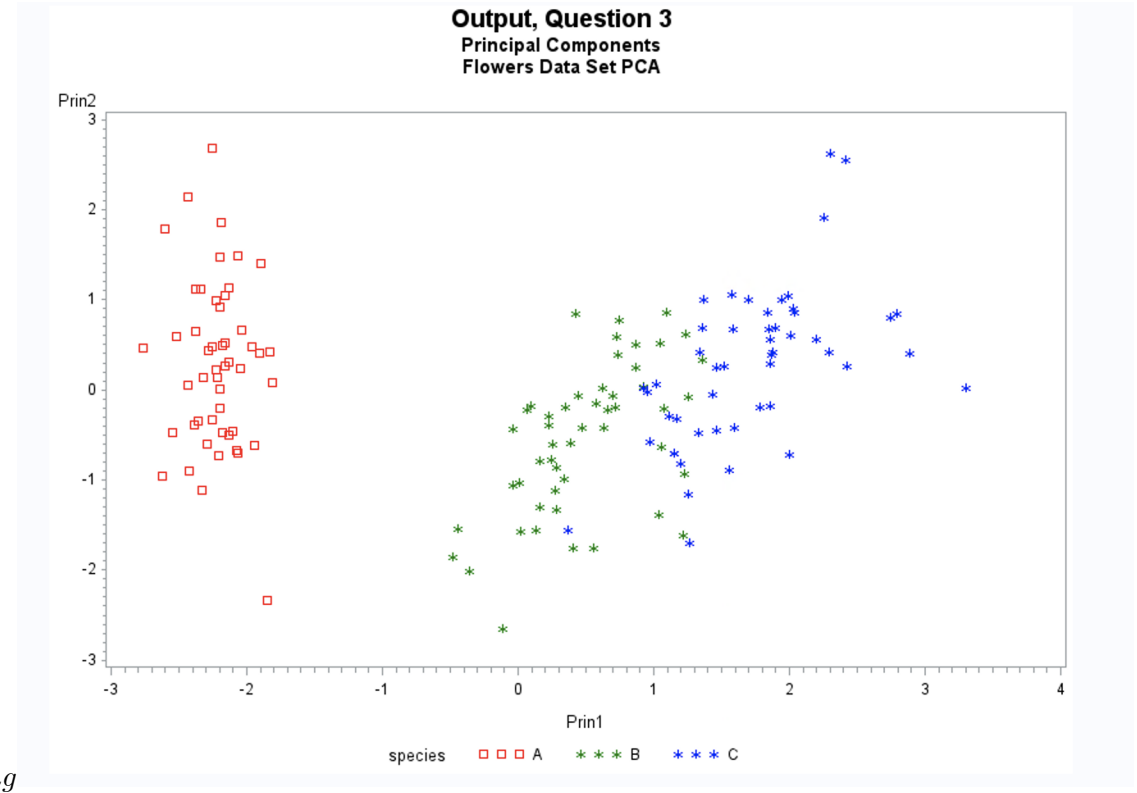
(40).*png*

	Eigenvalue	Difference	Proportion	Cumulative
1	2.91849782	2.00446735	0.7296	0.7296
2	0.91403047	0.76727360	0.2285	0.9581
3	0.14675688	0.12604204	0.0367	0.9948
4	0.02071484		0.0052	1.0000

Eigenvectors				
	Prin1	Prin2	Prin3	Prin4
sl	0.521066	0.377418	-.719566	-.261286
sw	-.269347	0.923296	0.244382	0.123510
pl	0.580413	0.024492	0.142126	0.801449
pw	0.564857	0.066942	0.634273	-.523597



(39).*png*



(38).*png*

4 Problem: Standardizing the data set

4.1 SAS Code

```
proc stdize data = HW5.flowers method = range out = t;
var sl sw pl pw;
run;
```

5 Problem: Average Linkage method

5.1 SAS Code

```
title "Clusters, Question 5 output";
title2 "Results from the Average Linkage Method";
proc cluster data = t method = average ccc pseudo outtree = tree_1;
var sl sw pl pw;
copy sl sw pl pw species;
run;
```

5.2 SAS Code Output

- For this output, we specified the pseudo method for our output, from which we obtained a cluster analysis plot. The number of clusters that we have given in the plot below coincides with the number of clusters that we give in the Dendogram plot for **6**.

(37).*png*

Clusters, Question 5 output
Results from the Average Linkage Method

The CLUSTER Procedure
Average Linkage Cluster Analysis

Eigenvalues of the Covariance Matrix

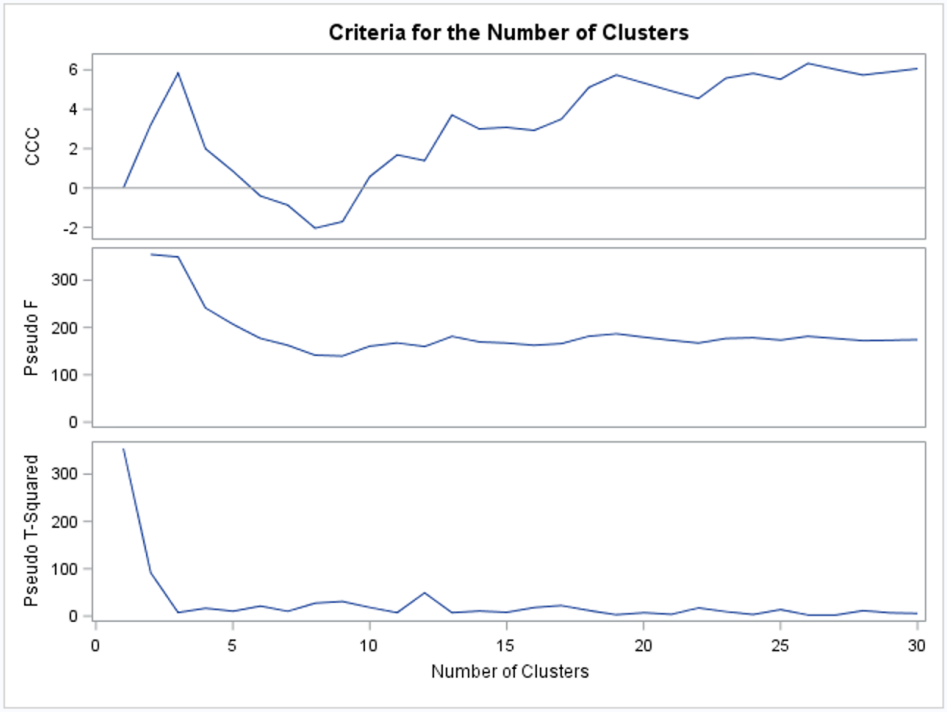
	Eigenvalue	Difference	Proportion	Cumulative
1	0.23245325	0.19998505	0.8414	0.8414
2	0.03246820	0.02287136	0.1175	0.9589
3	0.00959685	0.00783253	0.0347	0.9936
4	0.00176432		0.0064	1.0000

Root-Mean-Square Total-Sample Standard Deviation	0.262813

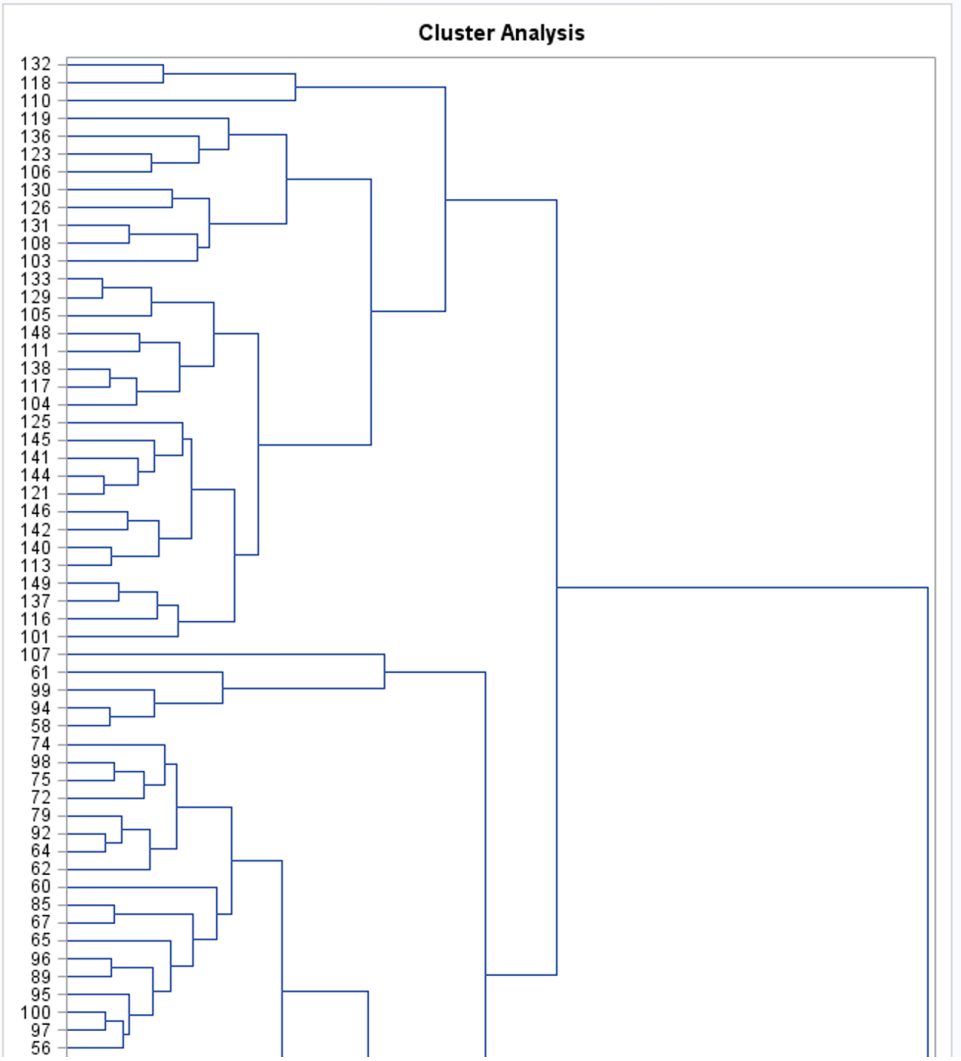
Root-Mean-Square Distance Between Observations	0.743347

Cluster History

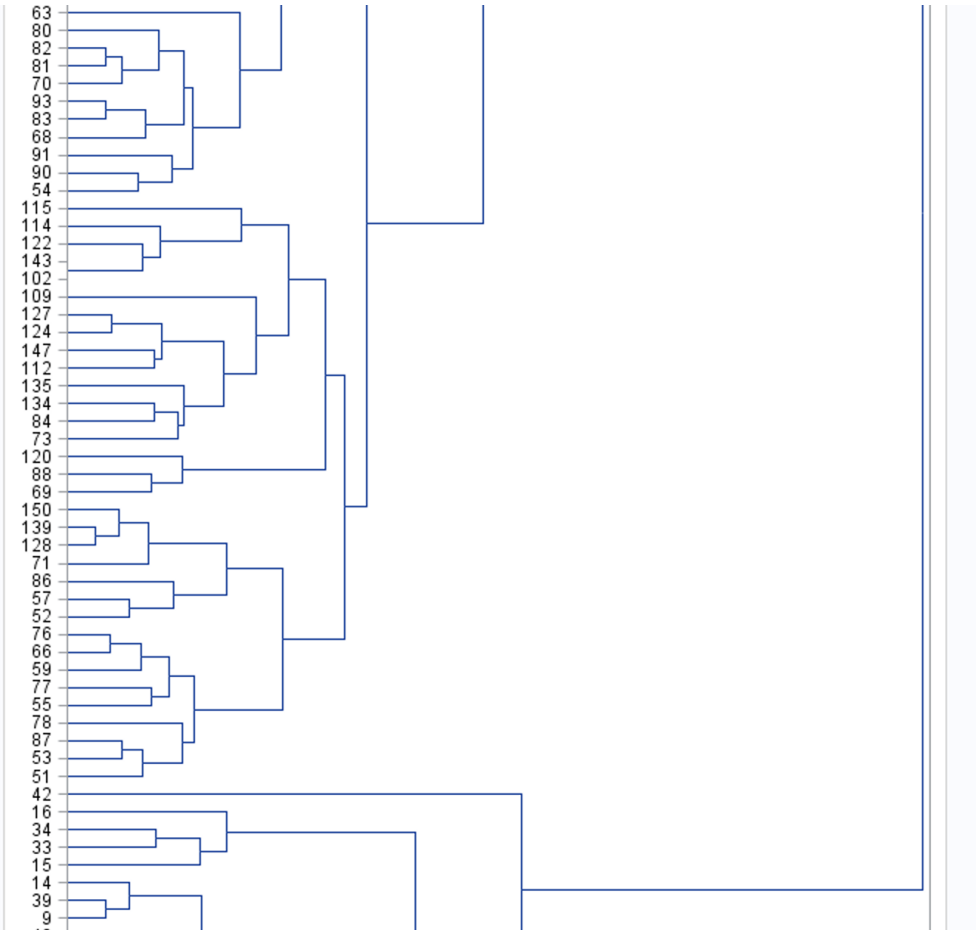
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Norm RMS Distance	Tie
149	OB102	OB143	2	0.0000	1.00	.	.	.	.	0	
148	OB8	OB40	2	0.0000	1.00	.	.	1452	.	0.0374	T
147	OB11	OB49	2	0.0000	1.00	.	.	1096	.	0.0374	
146	OB18	OB41	2	0.0000	1.00	.	.	873	.	0.0438	T
145	OB128	OB139	2	0.0000	1.00	.	.	781	.	0.0438	T
144	OB3	OB48	2	0.0000	1.00	.	.	732	.	0.0438	T
143	OB1	OB28	2	0.0000	1.00	.	.	702	.	0.0438	T
142	OB31	OB35	2	0.0000	1.00	.	.	683	.	0.0438	


Criteria for the Number of Clusters
CCC
6
4
2
0
-2
Pseudo F
300
200
100
0
Pseudo T-Squared
300
200
100
0
0
5
10
15
20
25
30
Number of Clusters

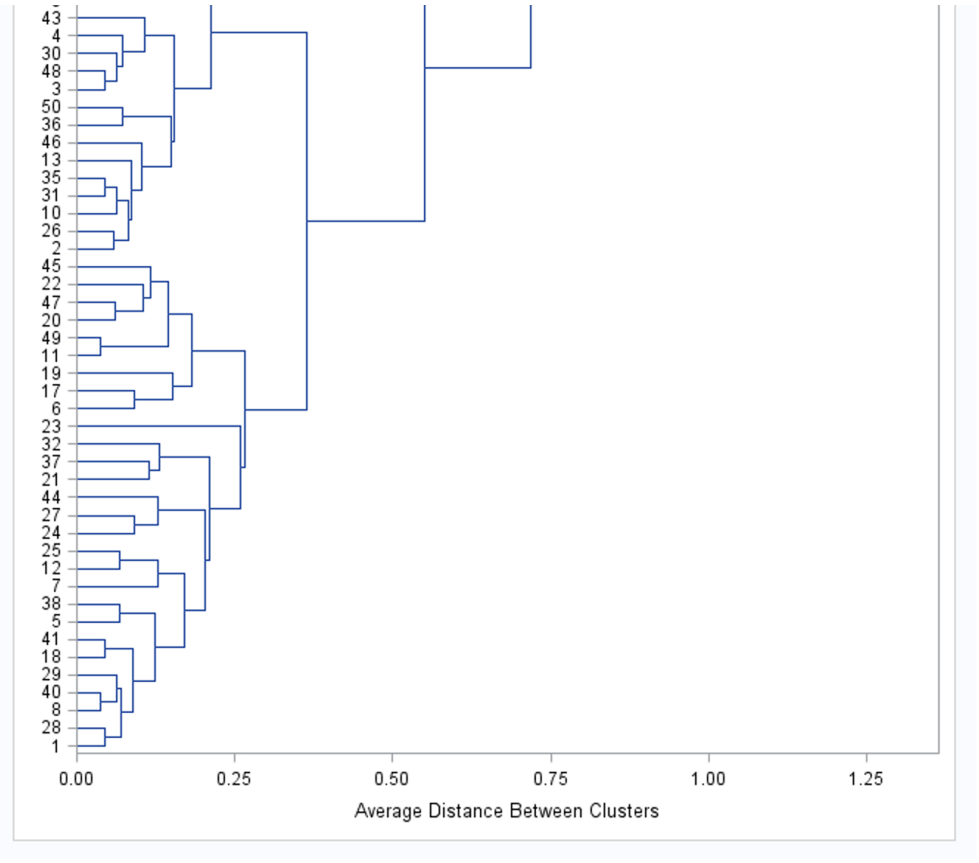
(36).*png*


Cluster Analysis
132
118
110
119
136
123
106
130
126
131
108
103
133
129
105
148
111
138
117
104
125
145
141
144
121
146
142
140
113
149
137
116
101
107
61
99
94
58
74
98
75
72
79
92
64
62
60
85
67
65
96
89
95
100
97
56

(33).*png*


63
80
82
81
70
93
83
68
91
90
54
115
114
122
143
102
109
127
124
147
112
135
134
84
73
120
88
69
150
139
128
71
86
57
52
76
66
59
77
55
78
87
53
51
42
16
34
33
15
14
39
9

(34).*png*


43
4
30
48
3
50
36
46
13
35
31
10
26
2
45
22
47
20
49
11
19
17
6
23
32
37
21
44
27
24
25
12
7
38
5
41
18
29
40
8
28
1
0.00
0.25
0.50
0.75
1.00
1.25
Average Distance Between Clusters

(35).*png*

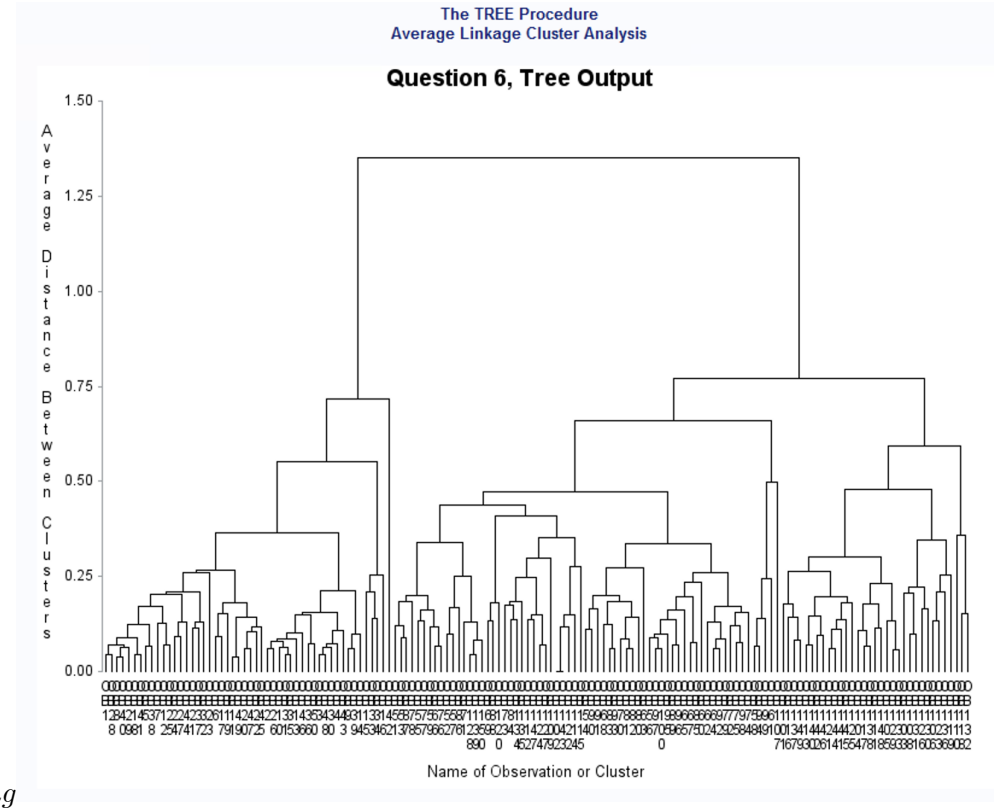
6 Problem: Plotting the tree

6.1 SAS Code

```
    title "Question 6, Tree Output";
proc tree data = tree_1;
run;
```

6.2 SAS Code Output: Dendogram

- The Dendogram below shows the number of clusters that have been pasted from the output. I tried specifying the number of clusters in the output but it still gave me several lines in the tree output.



(32).*png*