Hindi Vidya Prachar Samiti's
Ramniranjan Jhunjhunwala College of Arts, Science and Commerce
(Empowered Autonomous)

Programme: MSc. (Statistics) **Part-1** **Semester-2**

Practical based on Multivariate Analysis & its application
Cluster Analysis

1) The vocabulary "richness" of a test can be quantitatively described by counting the words used once, the words used twice, and so forth. Based on these counts, a linguist proposed the following distances between chapters of the Old Testament book Lamentation (data courtesy of Y.T. Radday and M.A. Pollatschek):

		Lamentation Chapter				
		1	2	3	4	5
Lamentation Chapter	1	0				
	2	0.76	0			
	3	2.97	0.8	0		
	4	4.88	4.17	0.21	0	
	5	3.86	1.92	1.51	0.51	0

Cluster the chapters of Lamentations using the three linkage hierarchical methods we have discussed. Draw the dendrograms and compare the results.

2) Sample correlations for five stocks rounded to two decimal places, are as follows

	Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
Allied Chemical	1				
Du Pont	0.58	1			
Union Carbide	0.51	0.6	1		
Exxon	0.39	0.39	0.44	1	
Texaco	0.46	0.32	0.43	0.52	1

Treating the sample correlations as similarity measures, cluster the stocks using the single linkage and complete linkage hierarchical procedures. Draw the dendrograms and compare the results.

3) Suppose we measure two variables X1 and X2 for four items A, B, C and D. The data are as follows:

Observations		
Item	x_1	x_2
A	5	4
B	1	-2
C	-1	1
B	3	1

Use the K-means clustering technique to divide the items into K = 2 clusters. Start with the initial groups (AC) and (BD).

4) Following table shows measurements on 8 variables for 43 breakfast cereals:

Brand	x1	x2	x3	x4	x5	x6	x7	x8
1	110	2	2	180	1.5	10.5	10	70
2	110	6	2	290	2	17	1	105
3	110	1	1	180	0	12	13	55
4	110	1	1	180	0	12	13	65
5	110	1	1	280	0	15	9	45
6	110	3	1	250	1.5	11.5	10	90
7	110	2	1	260	0	21	3	40
8	110	2	1	180	0	12	12	55
9	100	2	1	220	2	15	6	90
10	130	3	2	170	1.5	13.5	10	120
11	100	3	2	140	2.5	10.5	8	140
12	110	2	1	200	0	21	3	35
13	140	3	1	190	4	15	14	230
14	100	3	1	200	3	16	3	110
15	110	1	1	140	0	13	12	25
16	100	3	1	200	3	17	3	110
17	110	2	1	200	1	16	8	60
18	70	4	1	260	9	7	5	320
19	110	2	0	125	1	11	14	30
20	100	2	0	290	1	21	2	35
21	110	1	0	90	1	13	12	20
22	110	3	3	140	4	10	7	160
23	110	2	0	220	1	21	3	30
24	110	2	1	125	1	11	13	30
25	110	1	0	200	1	14	11	25
26	100	3	0	0	3	14	7	100
27	120	3	0	240	5	14	12	90
28	110	2	1	170	1	17	6	60
29	160	3	2	150	3	17	13	60

30	120	2	1	190	0	15	9	40
31	140	3	2	220	3	21	7	130
32	90	3	0	170	3	18	2	90
33	100	3	0	320	1	20	3	45
34	120	3	1	210	5	14	12	240
35	110	2	0	290	0	22	3	35
36	110	2	1	70	1	9	15	40
37	110	6	0	230	1	16	3	55
38	120	1	2	220	0	12	12	35
39	120	1	2	220	1	12	11	45
40	100	4	2	150	2	12	6	95
41	50	1	0	0	0	13	0	15
42	50	2	0	0	1	10	0	50
43	100	5	2	0	2.7	1	1	110

a) Using the data in the table, calculate the Euclidean distances between pairs of cereal brands.
b) Treating the distances calculated in (a) as measures of similarity, cluster the cereals using the single linkage and complete linkage hierarchical process. Construct dendograms and compare the results.

5) Use the data from the above table into a K-means clustering program. Cluster the cereals into K = 2, 3 and 4 groups. Compare the results with those in Q.4.

6) Case Study
