Types of linkages

ELBOW METHOD

→ 3 important concept - i) Total Error ii) Variance / Total Squared Error
iii) Within cluster sum of squares (WSS)

→ Example - we have 3 columns — Length, width and height & contain 20 variables

Obs	Length	length − mean(length)	[len − mean(len)]²
1	170.2	−2.04	4.16
2	188.8	16.56	274.23
3	171.6	2.36	5.56
4	186.7	14.46	209.09
5	166.3	−5.94	35.28
6	172.6	0.36	0.12
7	173.2	0.96	0.92
8	187.5	15.26	232.86
9	176.2	3.96	15.68
10	157.9	−14.34	205.63
11	157.1	−15.14	229.2
12	208.1	35.86	1285.93
13	150	22.24	494.61
14	177.8	5.56	30.91
15	167.5	−4.74	22.46
16	157.3	−14.94	24.40
17	187.8	15.56	223.20
18	141.1	−31.14	242.11
19	176.8	4.56	969.69
20			20.7
	mean = 172.24	Total error = 0	Total square error = 4526

Mean of 20 value of length = 172.24
So from mean if we add all the difference we will get 0.
So total error = 0.
(+ve & −ve value will cancel each other)

→ Indirectly we made that distribution as normal distribution.

→ Total square error & variance mean the same.

mean variance = $\frac{\text{total variance}}{\text{number of rows} - 1}$

$= \frac{4526}{19} = 238.26$

Same way calculate for width & height.

mean variance (width) = 6.89
mean variance (height) = 6.96

Total variance in data = 19 * [mean variance (length) + mean variance (width) + mean variance (height)] = 19 * [238.26 + 6.89 + 6.96]

Total variance in data = 4790

Suppose we took cluster = 2
In cluster 1 = 11 and cluster 2 = 9.
variance of cluster 1 = 1241.53
variance of cluster 2 = 781.31
Total within cluster sum of square = 2025.76
2025.76 is total variance in data when we divide the data into two clusters.

Note → Total variance of any cluster is called the within cluster sum of square. (WCSS)

Total within cluster sum of square = within cluster sum of square 1 + within cluster sum of square 2

Follow the same process, for 3 cluster = 849.35

So as we increase the cluster number, within sum of square (WSS) will decrease but after certain point WSS will become constant, that number of cluster is the optimal number of cluster. (Choose the drop point, significant change in WSS)

ELBOW DIG ← 3 is optimal number of cluster in the diagram