$$VIF = \frac{1}{1-R^2} \Rightarrow [1, \infty]$$

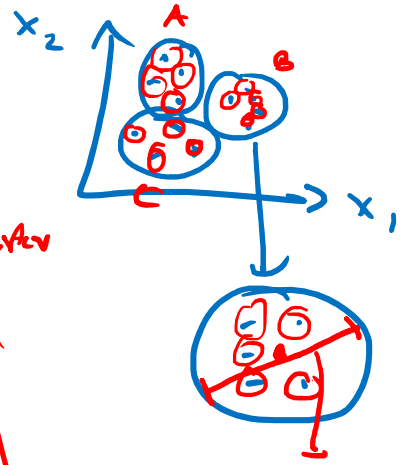
$[0, 1)$

$\geq 5, \infty \rightarrow$ high corr

< 5, No or wiek corr

K - means ?

→ Choosing the best value of K → {1, 20}



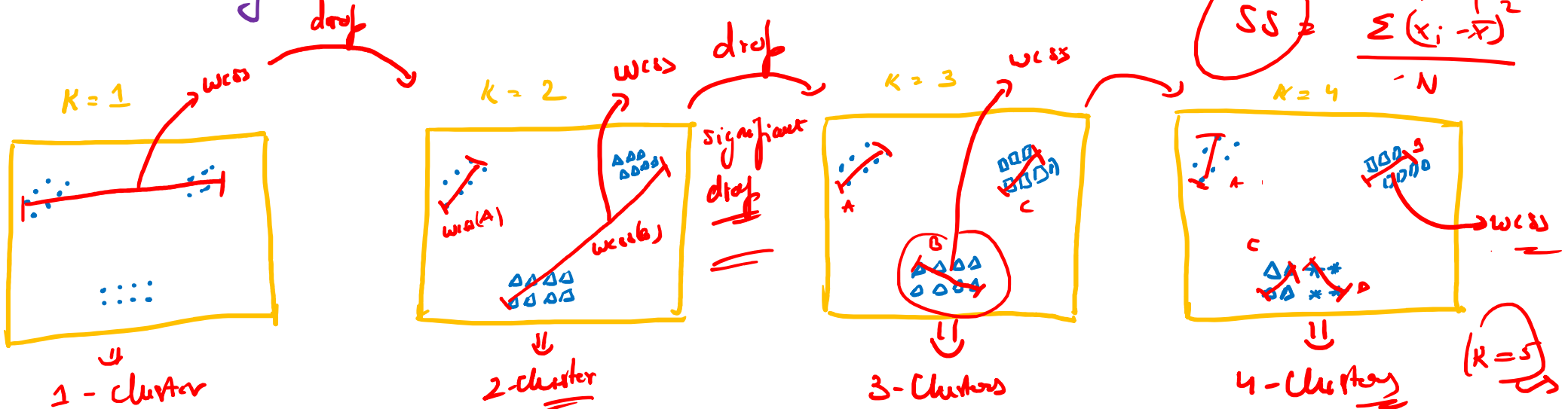
Spread → $\frac{\Sigma (x_i - \bar{x})^2}{N}$ (variance)

within the cluster

WCSS / WSS
(within - cluster Sum of sq)

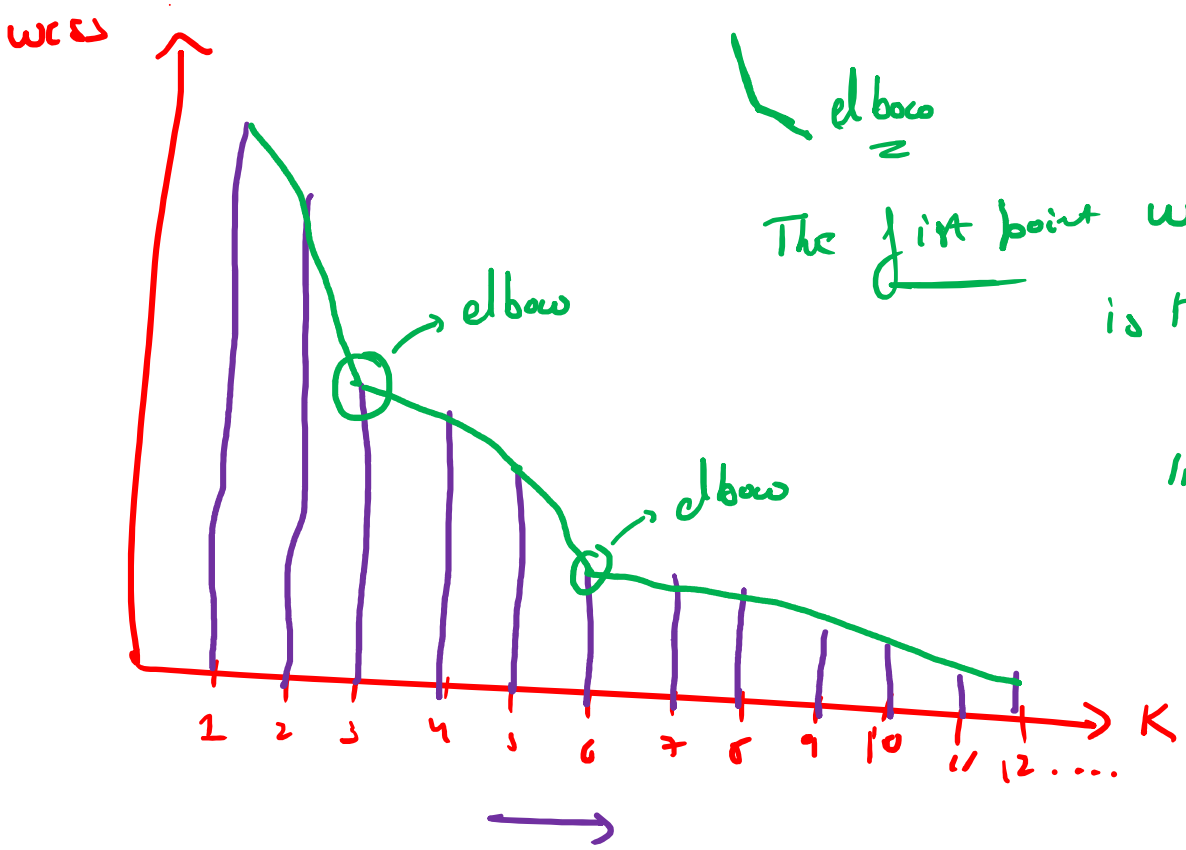
SS

Determined using WCSS method / elbow method



For multiple values we will choose the max (WCSS1, WCSS2 .....)

Plotting the value of wcss vs K

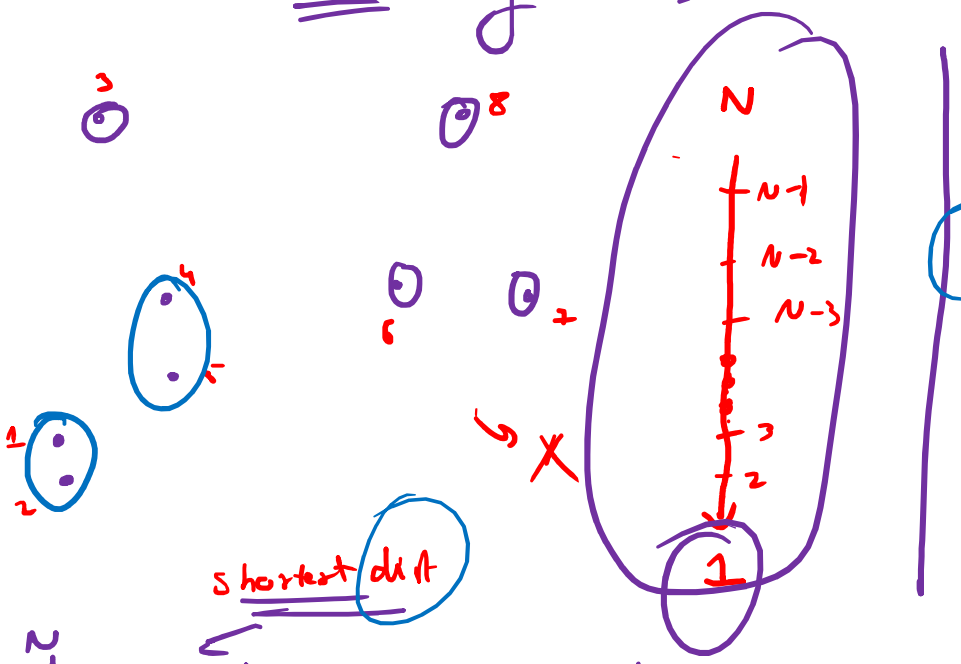


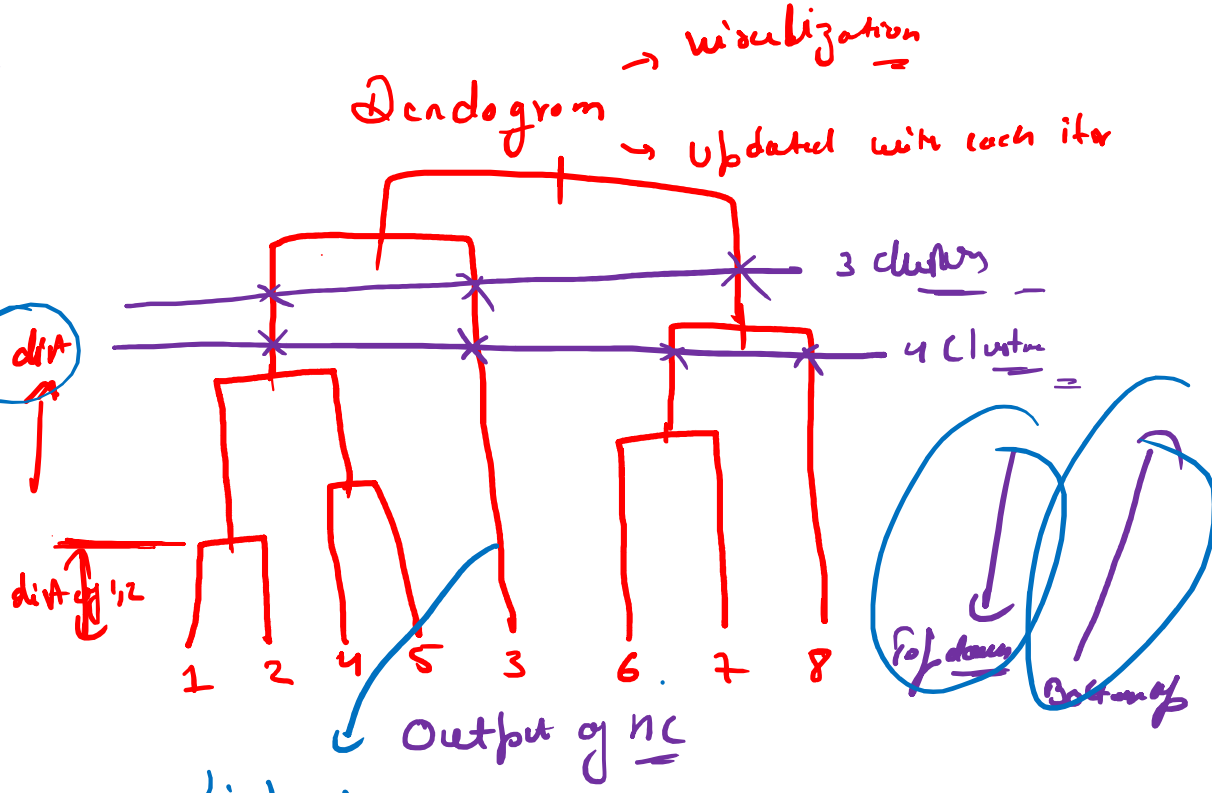
elbow

The first point where the elbow is formed is the best value of K

In case, the elbows are formed close to each other then business needs to decide on which one to take ↓ SME

→ Hierarchical Clustering (HC) → Flow





Linkages

linking the two groups together using certain calculations

→ Difference in HC, K-means

① HC cannot be used for a large dataset → too much time → $O(n^2)$ → Dendogram will be too complex → difficult to analyze

K-means can be used for a large dataset → $O(n)$ → faster than HC

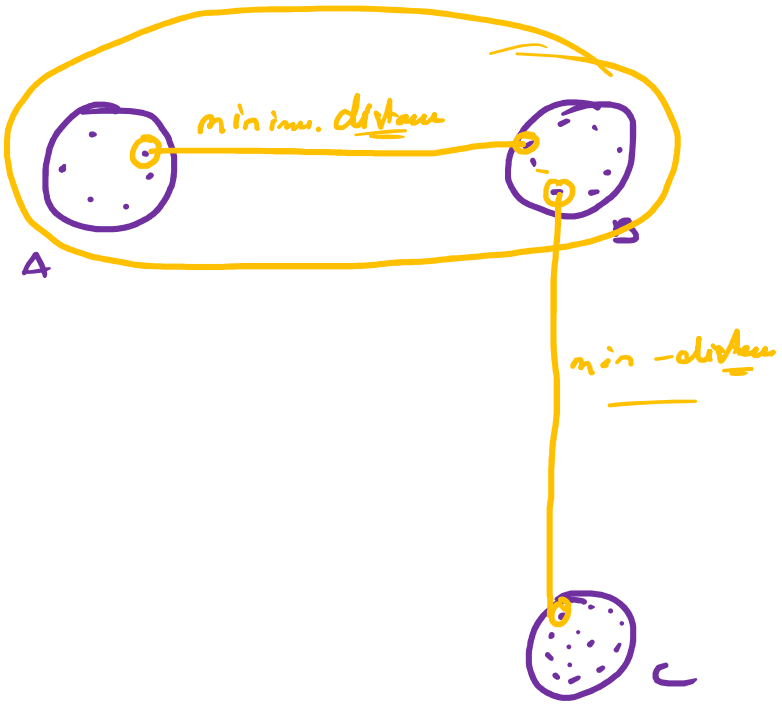
(In 100s)

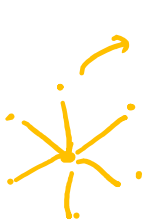
② HC you can reproduce the same clusters on same data everytime

In K-means you can/cannot reproduce the result, because of the random centroid allocation

→ Types of linkages

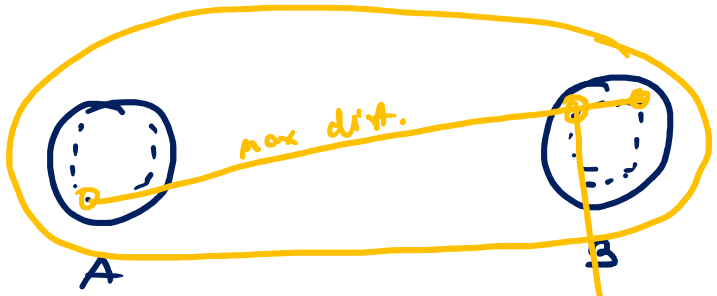
① Single linkage → closest minimum distance

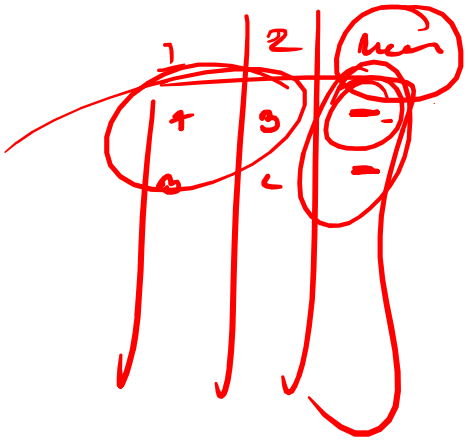




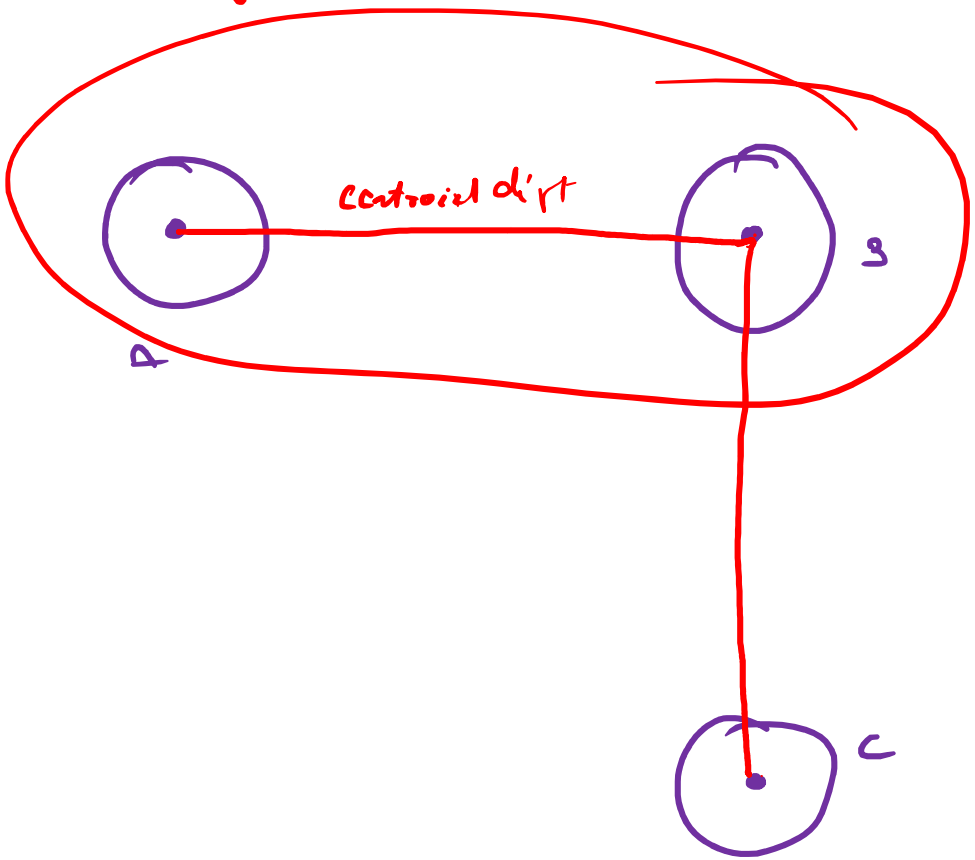
Any one can be choosen randomly in case of tie

(2) Complete linkage → closest — Maximum distance

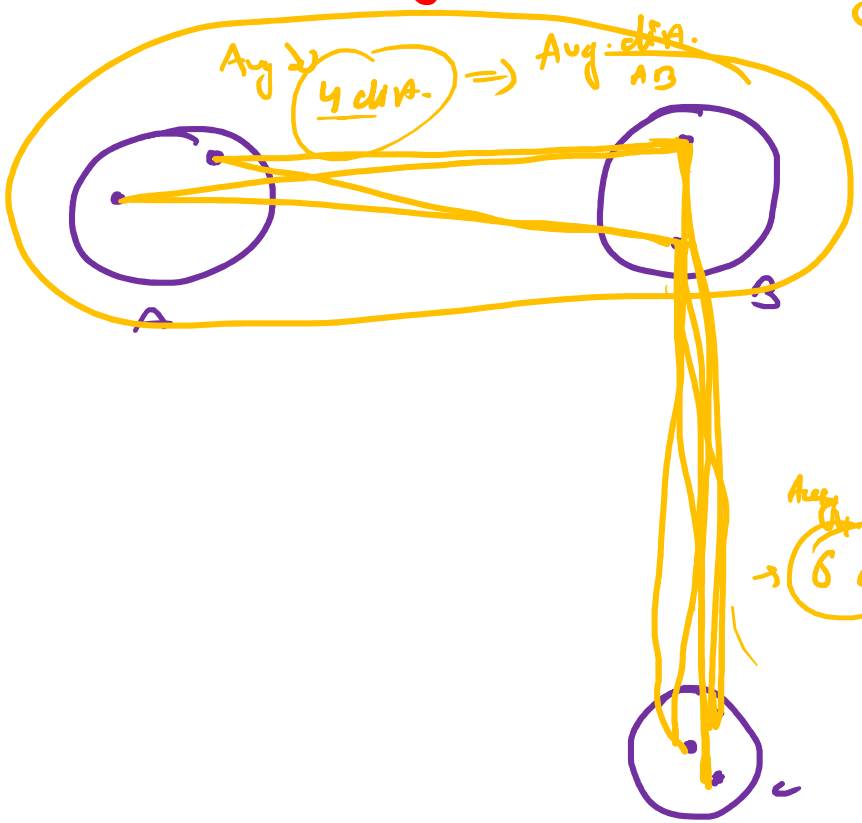


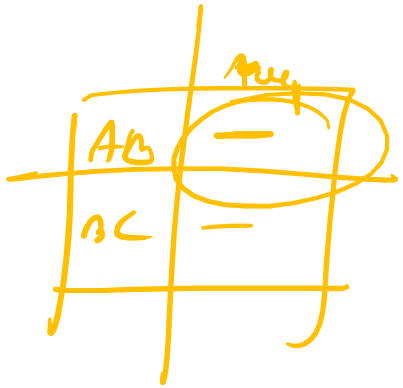


③ Centroid linkage → lowest Centroid distance

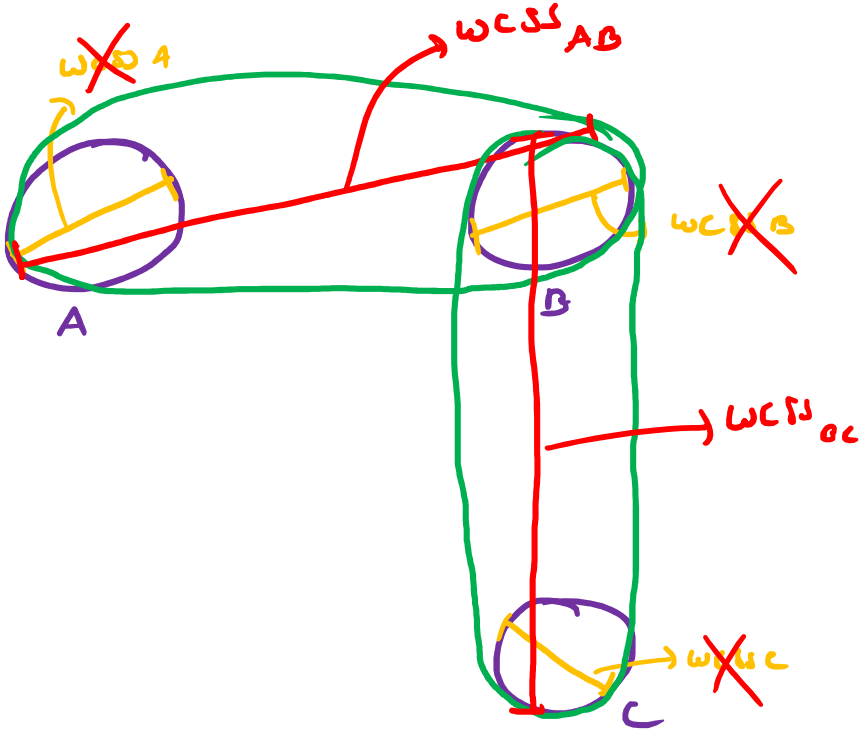


→ Average linkage → lowest Avg distance → Avg of Every possible distance from all the points of two clusters





(5) Ward linkage ⇒ lowest wcss/ess value → most used



lowest ($wcss_{AB}$, $wcss_{BC}$)

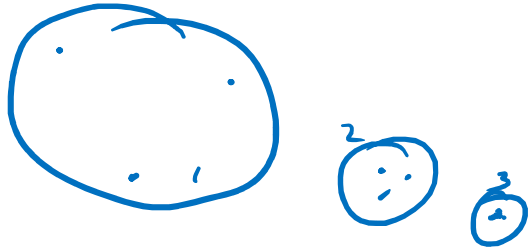
group having the lowest wcss will be grouped

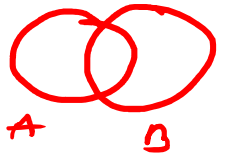
for Example:

for linkage code:

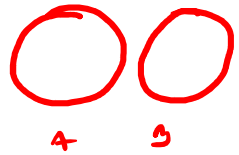
→ Quality of clusters → Silhouette score

$$S = \frac{b - a}{\max(a, b)}$$

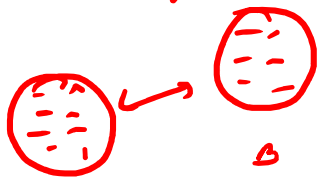




X
Incorrect identification



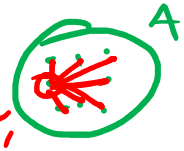
Bad quality



good quality

How much closely packed

a = Mean distance b/w a randomly selected point of the cluster & all other points in the same cluster

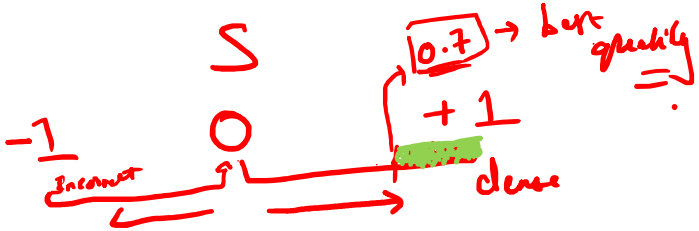


Avg is 'a'

b = Mean dist. b/w the same selected pt. as 'a' of the cluster & all the points in the nearest neighbour cluster

↳ Avg is 'b'

S



Best value of 'a' = 0

$$S = \frac{b - 0}{\max(b, 0)} = \frac{b}{b} = 1$$

↓ Best quality

b = a

↳ $S = \frac{b - a}{\max(b, a)} = \frac{0}{\max} = 0$

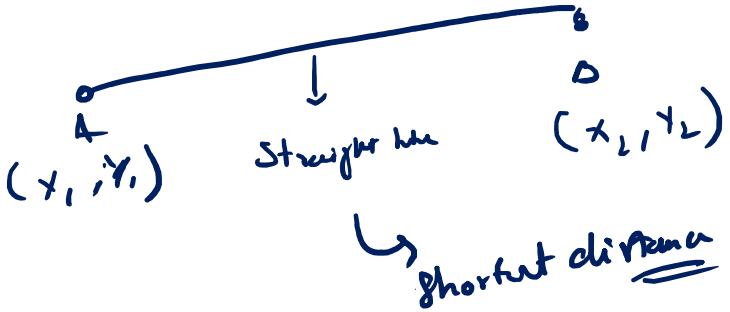
Bad quality

a > b

↳ over lapping clusters

b = a

↳ S = -1 → Incorrect clusters



↳ shortest distance

→ $AB = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

⇓

Euclidean dist^{n}.