

Q1a: Given the task of predicting blue preference, what's the Total Weighted Gini Impurity of this split?

Diagram of a split on 'purple' (Y) vs 'blue' (N):

```

    purple
   /   \
  Y     N
 /       \
blue     blue

```

Y	N	Gini index
39	31	$1 - \left(\left(\frac{39}{70}\right)^2 + \left(\frac{31}{70}\right)^2\right) = 0.49$

Y	N	Gini index
29	11	$1 - \left(\left(\frac{29}{40}\right)^2 + \left(\frac{11}{40}\right)^2\right) = 0.40$
20	11	$1 - \left(\left(\frac{20}{31}\right)^2 + \left(\frac{11}{31}\right)^2\right) = 0.45$

• total weighted Gini Impurity

$$= 0.40 \times \frac{39}{70} + 0.45 \times \frac{31}{70} = 0.42$$

Q1b: What would be the predicted labels of the left and right leaves?
blue

Left leaf: Blue

Right Leaf: Blue

Q1c: Given the task of predicting percentage of white t-shirts, what's the Total Weighted Variance of this split?

range	0-5%	6-30%	31-40%	41-45%	46-100%	
white/purple Y	7	16	10	5	1	=39
white/purple N	7	12	10	1	1	=31

• avg for first row =

$$\frac{(2.5 \times 7) + (18 \times 16) + (50.5 \times 10) + (83.5 \times 5) + (98 \times 1)}{39} = 33.94$$

• avg for second row =

$$\frac{(2.5 \times 7) + (18 \times 12) + (50.5 \times 10) + (83.5 \times 1) + (98 \times 1)}{31} = 29.61$$

• Variance for first row =

$$\frac{((2.5 - 33.94)^2 \times 7) + ((18 - 33.94)^2 \times 16) + \dots + ((98 - 33.94)^2 \times 1)}{39} = 765.17$$

• Variance for second row =

$$\frac{((2.5 - 29.61)^2 \times 7) + ((18 - 29.61)^2 \times 12) + \dots + ((98 - 29.61)^2 \times 1)}{31} = 601.13$$

• Total Weighted Variance = $765.17 \times \frac{39}{70} + 601.13 \times \frac{31}{70} = 693.12$

Q1d: What would be the predicted values of the left and right leaves?

Left leaf: 6-30%

Right leaf: 6-30%

Q2a.

I would not recommend 'k-Means Clustering'.

Since the distributions of the data intersect each other, k-Means, which evaluates the distance from the center point, is not suitable. k-Means is suitable for circular distributions. As in the example, when the shape of the cluster is not circular, accurate results cannot be derived.

Q2b.

I would not recommend 'k-Means Clustering'.

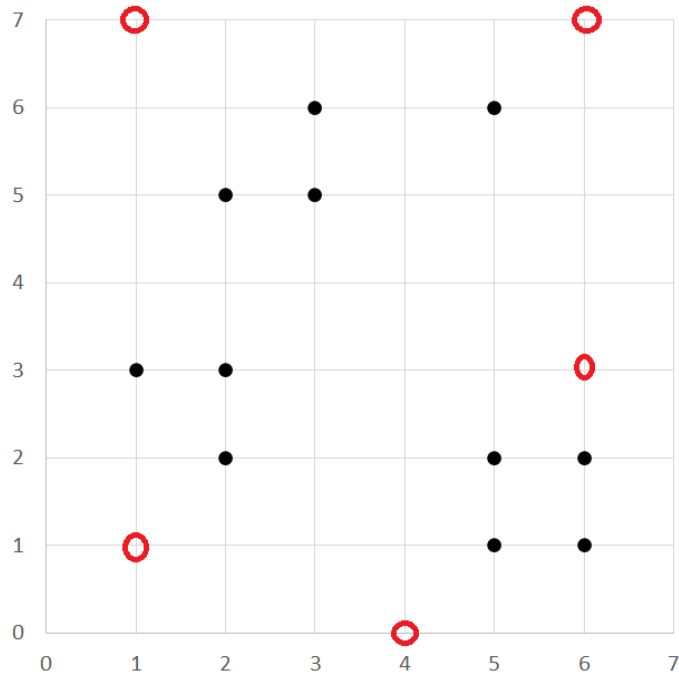
k-Means is not suitable if the distribution of the data is not linearly divisible.

Q2c.

I would not recommend 'k-Means Clustering'

Although the distribution of data can be classified as linear, k-Means is not suitable because the distribution is elongated rather than circular. k-Means is suitable for circular distributions.

QC. Draw a distortion plot of the slide page 15 in Clustering.pptx



Let's pick random k position.

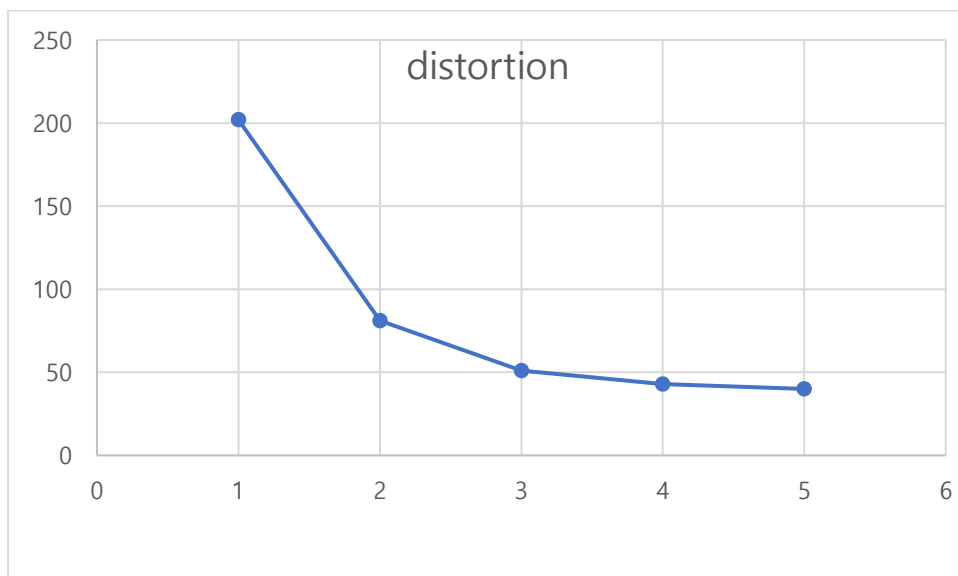
k = 1 (1,1): distortion = 202

k = 2 (1,1) (6,3): distortion = 81

k = 3 (1,1) (6,3) (1,7): distortion = 51

k = 4 (1,1) (6,3) (1,7) (6,7): distortion = 43

k = 5 (1,1) (6,3) (1,7) (6,7) (4,0): distortion = 40



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	x	y	x	y					x	y	x	y			
2	1	3	1	1	0	2	4		1	3	1	1	0	2	4
3	2	2	1	1	1	1	2		2	2	1	1	1	1	2
4	2	3	1	1	1	2	5		2	3	1	1	1	2	5
5	2	5	1	1	1	4	17		2	5	1	1	1	4	17
6	3	5	1	1	2	4	20		3	5	6	3	-3	2	13
7	3	6	1	1	2	5	29		3	6	6	3	-3	3	18
8	5	1	1	1	4	0	16		5	1	6	3	-1	-2	5
9	5	2	1	1	4	1	17		5	2	6	3	-1	-1	2
10	5	6	1	1	4	5	41		5	6	6	3	-1	3	10
11	6	1	1	1	5	0	25		6	1	6	3	0	-2	4
12	6	2	1	1	5	1	26		6	2	6	3	0	-1	1
13					k = 1	(1,1)	202						k = 2	(6,3)	81
14	x	y	x	y					x	y	x	y			
15	1	3	1	1	0	2	4		1	3	1	1	0	2	4
16	2	2	1	1	1	1	2		2	2	1	1	1	1	2
17	2	3	1	1	1	2	5		2	3	1	1	1	2	5
18	2	5	1	7	1	-2	5		2	5	1	7	1	-2	5
19	3	5	1	7	2	-2	8		3	5	1	7	2	-2	8
20	3	6	1	7	2	-1	5		3	6	1	7	2	-1	5
21	5	1	6	3	-1	-2	5		5	1	6	3	-1	-2	5
22	5	2	6	3	-1	-1	2		5	2	6	3	-1	-1	2
23	5	6	6	3	-1	3	10		5	6	6	7	-1	-1	2
24	6	1	6	3	0	-2	4		6	1	6	3	0	-2	4
25	6	2	6	3	0	-1	1		6	2	6	3	0	-1	1
26					k = 3	(1,7)	51						k = 4	(6,7)	43
27	x	y	x	y											
28	1	3	1	1	0	2	4								
29	2	2	1	1	1	1	2								
30	2	3	1	1	1	2	5								
31	2	5	1	7	1	-2	5								
32	3	5	1	7	2	-2	8								
33	3	6	1	7	2	-1	5								
34	5	1	4	0	1	1	2								
35	5	2	6	3	-1	-1	2								
36	5	6	6	7	-1	-1	2								
37	6	1	6	3	0	-2	4								
38	6	2	6	3	0	-1	1								
39					k = 5	(4,0)	40								