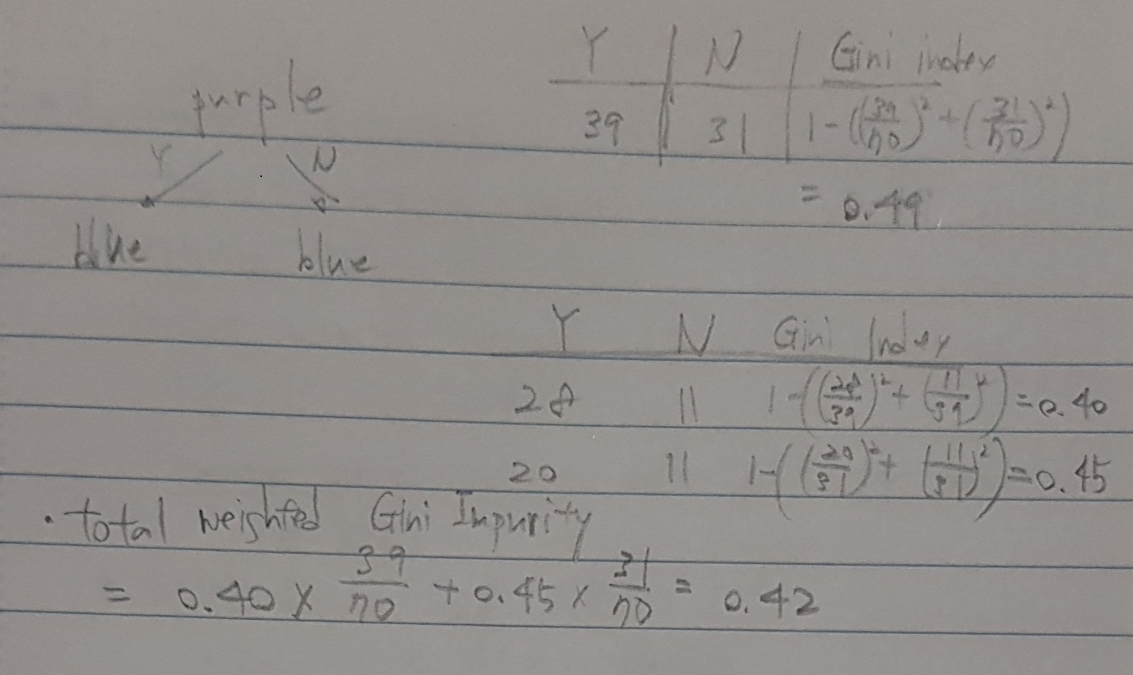
**Q1a: Given the task of predicting blue preference, what’s the Total Weighted Gini Impurity of this split?**

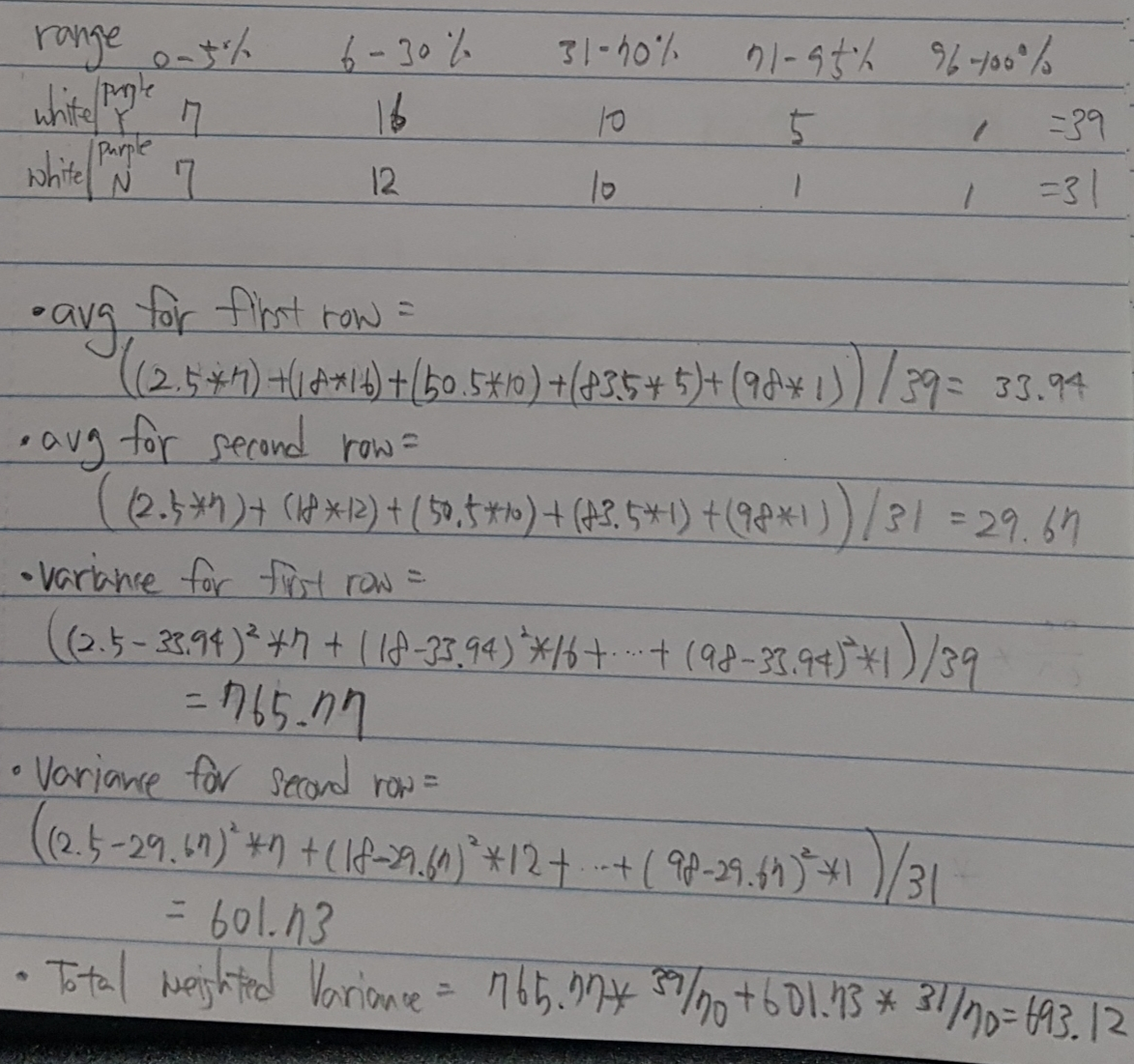


**Q1b: What would be the predicted labels of the left and right leaves?blue**

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?**



**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’.

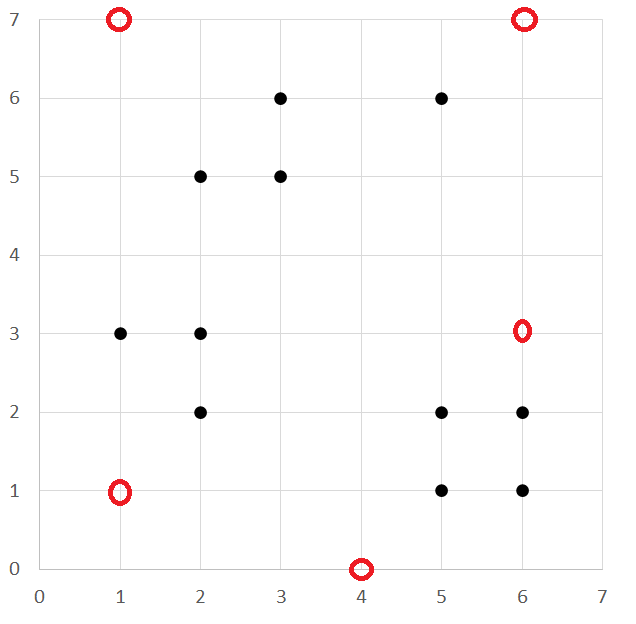
k-Means is not suitable if the distribution of the data is not linearly divisible.

**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

