Clustering Techniques [Segmentation]

Why and Where?

What we have done so far

- We had a definite response
- We had data points which possibly led to that response
- We extracted information through these modelling techniques, how the data points contributed to that response

No Response?

- Now what if we don't really have a response. We just have data. Is that any good? What kind of information would you be interested in extracting?

- Unsupervised Learning?

In absence of a target

- We can try to find out if there is some pattern in the data.
- What do we mean by pattern?
- If all the observations or data points are similar, all information that we can extract is by summarizing the data.

Contd..

- More information would be if some of the data points are different from others, or in other words there exist some groups within the population, different than each other.

- Examples?

Class Case : Fine Wine

Wine Tasters

- Wine tasters are used to rate wines on various parameters.
- Its not only becoming tough to find good wine tasters, its almost impossible get consistent feedbacks from multiple wine tasters
- For mass production companies, this process needs to be automated

Segmentation to Rescue

- Instead of relying on subjective opinions of wine tasters we can measure chemical properties of wines
- We can then group similar wines together based on their chemical properties
- Although final labeling on these groups will have to be a manual process

Quantifying Difference

What do you mean by different?

How do you “quantify” the difference? How do you make the group distinguishable based on data, more importantly “numbers’?

A small example

- Lets plot this small data set and try to figure out if there is way to quantify our intuition.

Age	Height (in cms)
35	160
42	172
35	172
42	160

So distance it is then. Scale matters?

- As it turns out scale matters, especially if variables in consideration are measured on different scales.

- The way out?

Standardization is the saviour Again!

- Or , is it?
- Although standardization is recommended, do we always standardize?

Combining Groups

Aggregation of groups : Methods

- Linkage Mehods
  - Single
  - Complete
  - Average
  - Centroid
  - Wards

Methods contd..

$$d_{single}(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$$

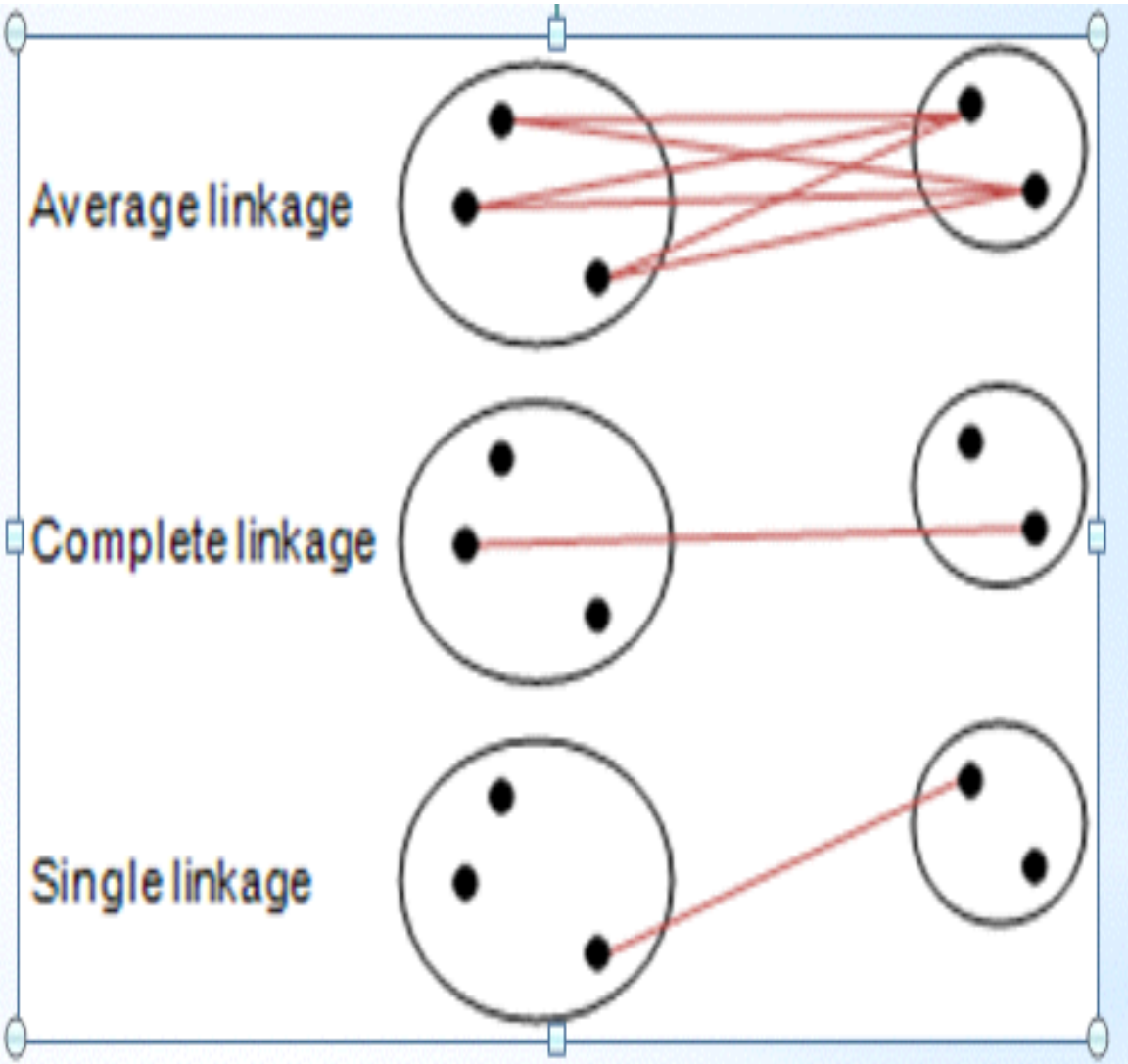
$$d_{complete}(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y)$$

$$d_{average}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

$$d_{centroid}(C_i, C_j) = d(m_i, m_j)$$

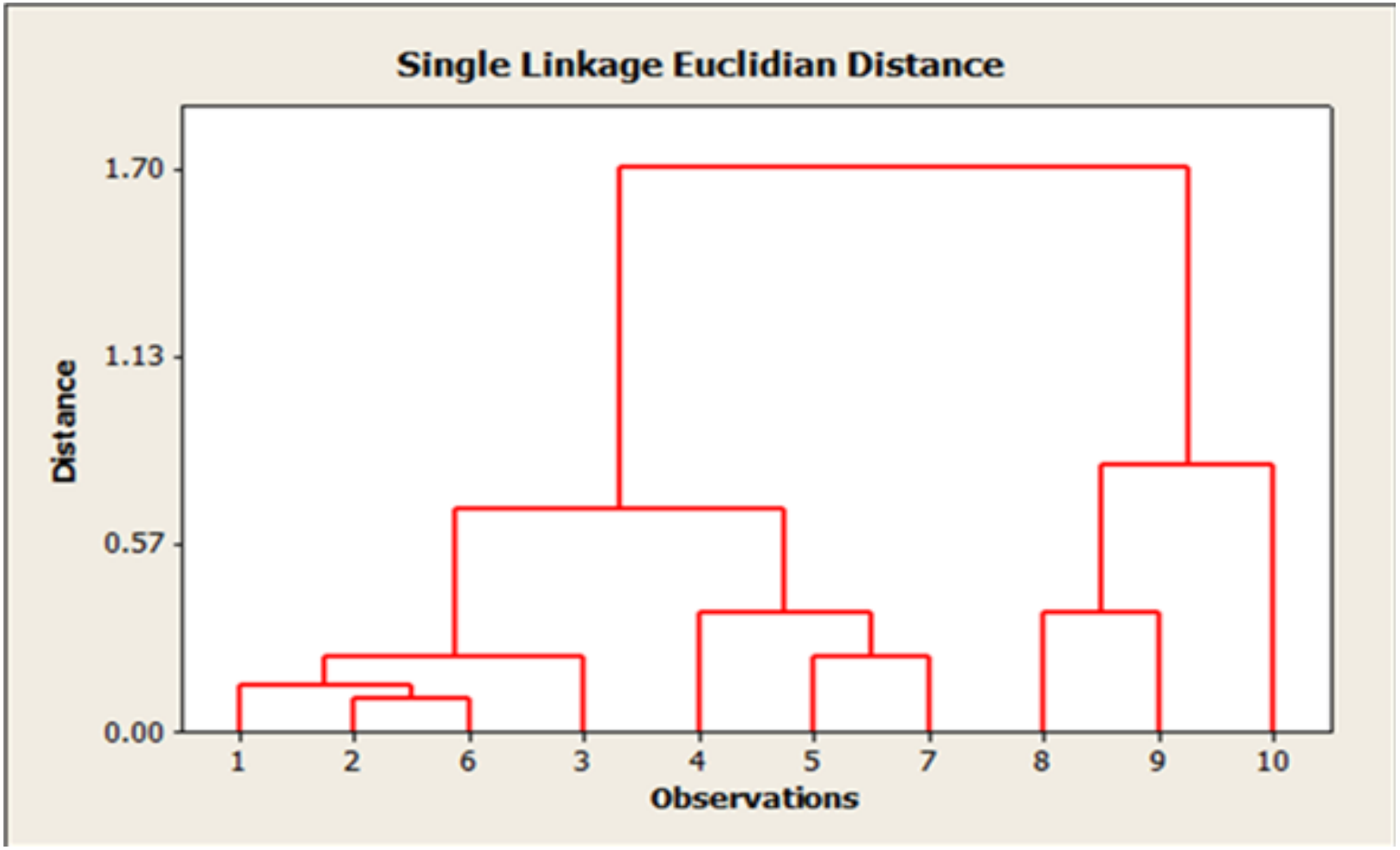
m_i - centroid of a cluster C_i

$|C_i|$ - number of samples in a cluster C_i



Hierarchical Clustering

Hierarchical Clustering



Problems with Hierarchical Clustering

- With increasing data; process becomes too slow and resource intensive
- Tree diagrams become too cluttered to make any sense out of them
- K-means clustering comes to the rescue

K-Means Clustering

K means Clustering

- You already tell the algorithm, how many clusters there are going to be in the data, that is the number K.

- You start with K random observations from the data, as K clusters.

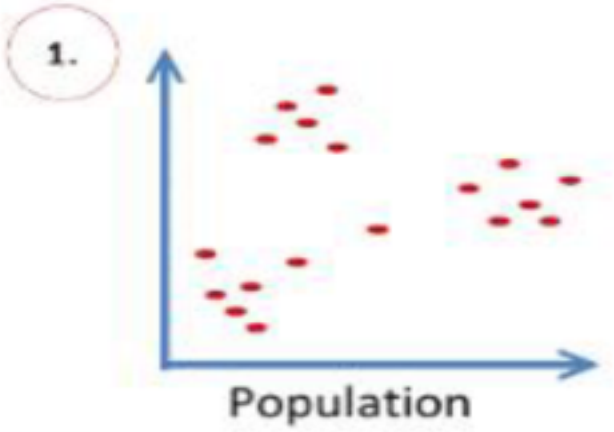
Contd..

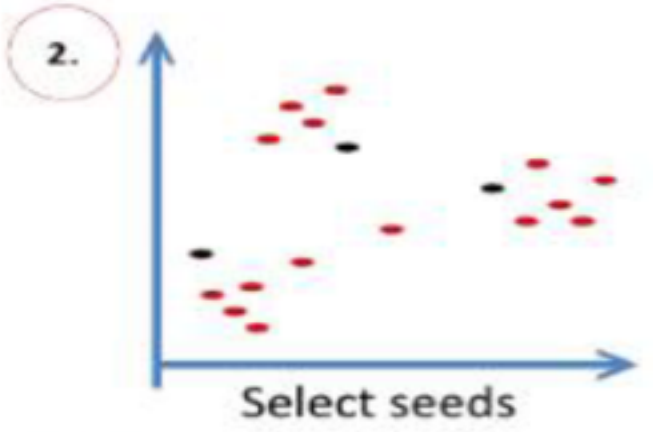
- Next observation is added to one of these clusters with the criterion which you have chosen.

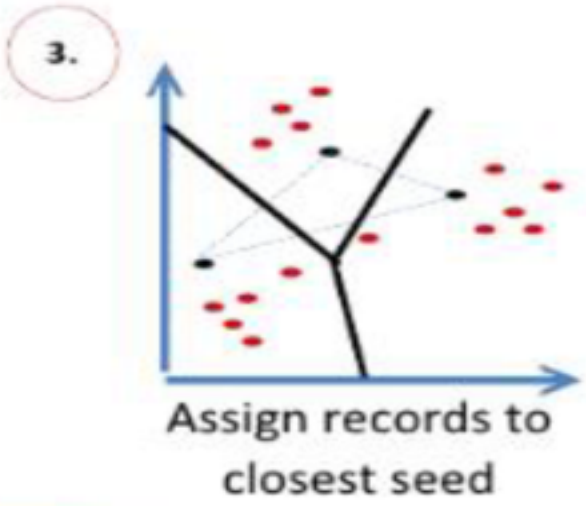
- Centroid for the said data is calculated, this would be the new point from which distance from the cluster will be calculated.

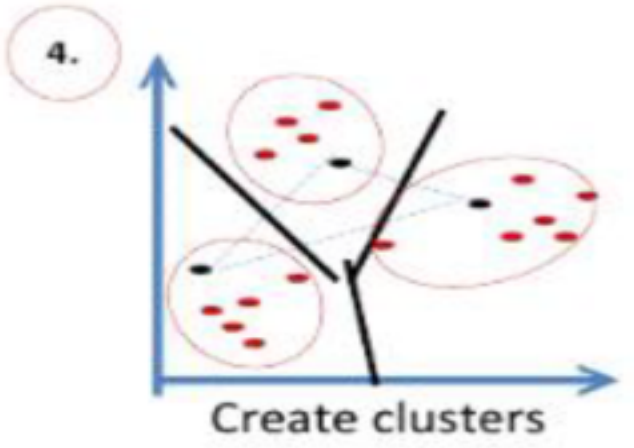
- This procedure is repeated until all the data points are assigned to clusters

K Means Clustering

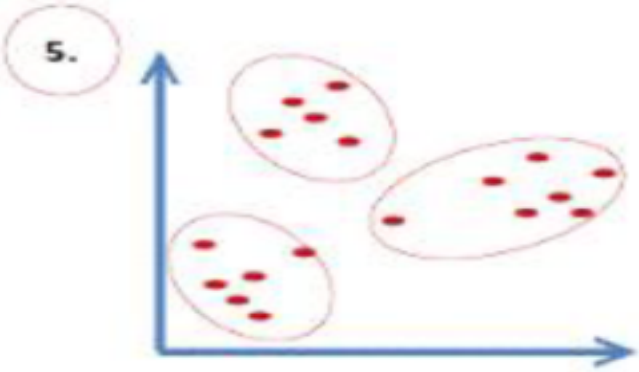




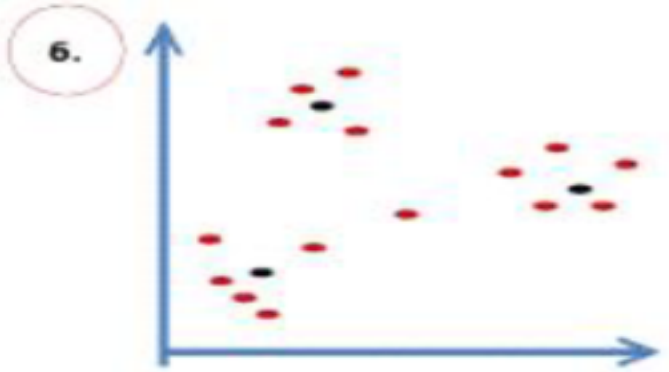




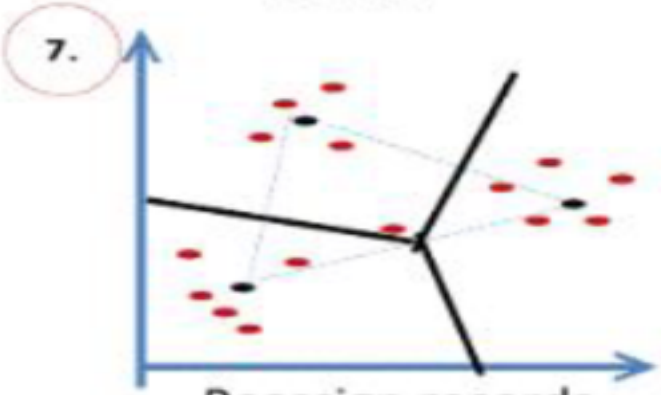
K means Clustering



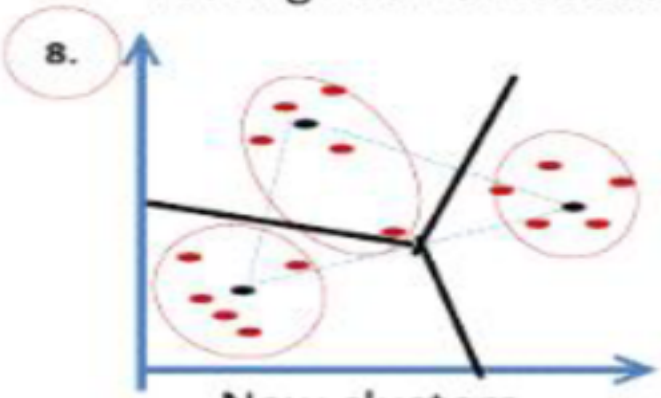
Initial clusters formed



Calculate centroids and reassign as new seeds



Reassign records based on new seeds



New clusters formed

Problems?

We'll take care of these problems in coming sections. Lets briefly touch upon these for now

- Sensitive to the initial seeds [the first K points]

K?

- What value of K is appropriate?
  - R^2 or WSS will keep on increasing /decreasing respectively with increase in value of K, until K equals the number of data points [We'll learn how wss plays a role in clustering in next section]
  - Where do we stop increasing K?

Variable Selection and results of clustering

- Which variable should be selected? All of them? What can be the problem with including all the variables?

- Fine my clustering is done; now what?

Sum of Squares & sons!

- For any given data Total sum of squares or SST is constant

- SST= SSW + SSB, that is if we break our data into multiple groups. If these groups are formed, SSB is much higher w.r.t. To SSW.

Contd..

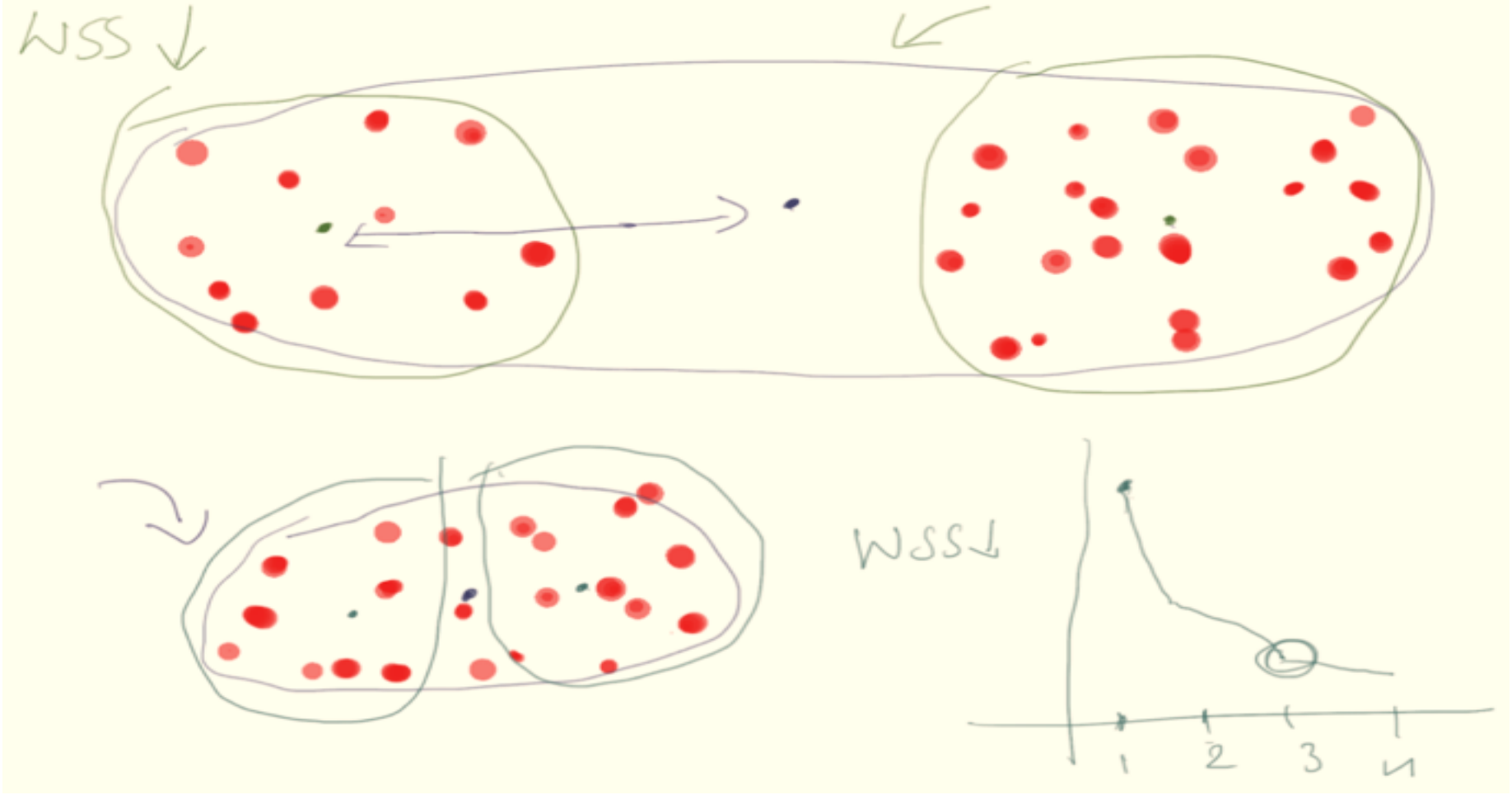
- As we increase number of groups SSW goes down.

- With increase in K if fall in SSW is not rapid/ steep, it implies the higher number of groups are not resulting in better formed groups.

Effect on R^2

R^2 = SSB/SST , as opposed to SSW, SSB goes up and equals SST if we have number of groups equal to data points. With increasing K, R^2 approaches to one.

Effect on WSS – Deciding on no. of clusters



Other Considerations

What to do with categorical variables?

- Make ordinal if possible
- Or Dummy variables
- Keep in mind that those ordinal variables should not be circular. Ok. Fine. Wait!....Circular?

Is multicollinearity an issue?

- There is no hard and fast statistical theory at play here
- No hypothesis testing
- No beta {the symbol!} , no DV, no hassle, why exactly multicollinearity can be an issue here?

Applications

Further applications of Cluster analysis

- Marketing & Media
- Banking & Insurance
- Medical and Pharmaceuticals
- Socio-economic