accenture

Strategy | Digital | Technology | Operations

Retail Segmentation Techniques

High performance. Delivered.

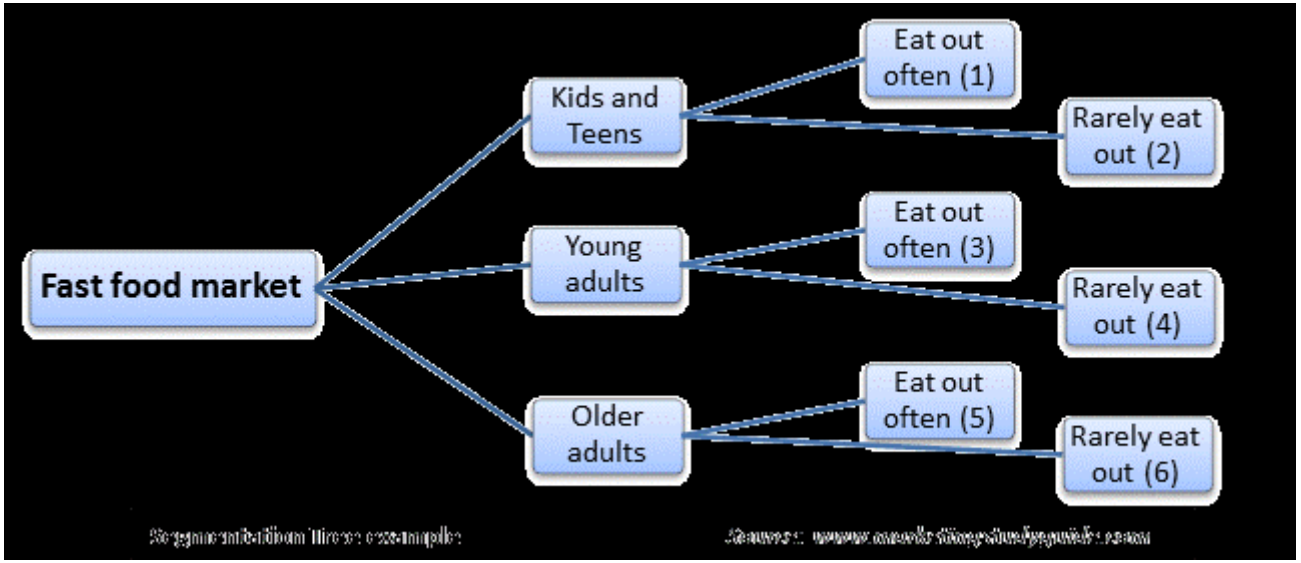
Segmentation:

Do you know before you purchase anything either Online or In Store, you have been targeted by Retailers and been put under one of the segments that they have already created using your past historical purchases or during your first purchase you move to one of the pre defined segments.

- ✓ Happy Customer ☺
- ✓ Easy to Strategize
- ✓ Maximize Profit

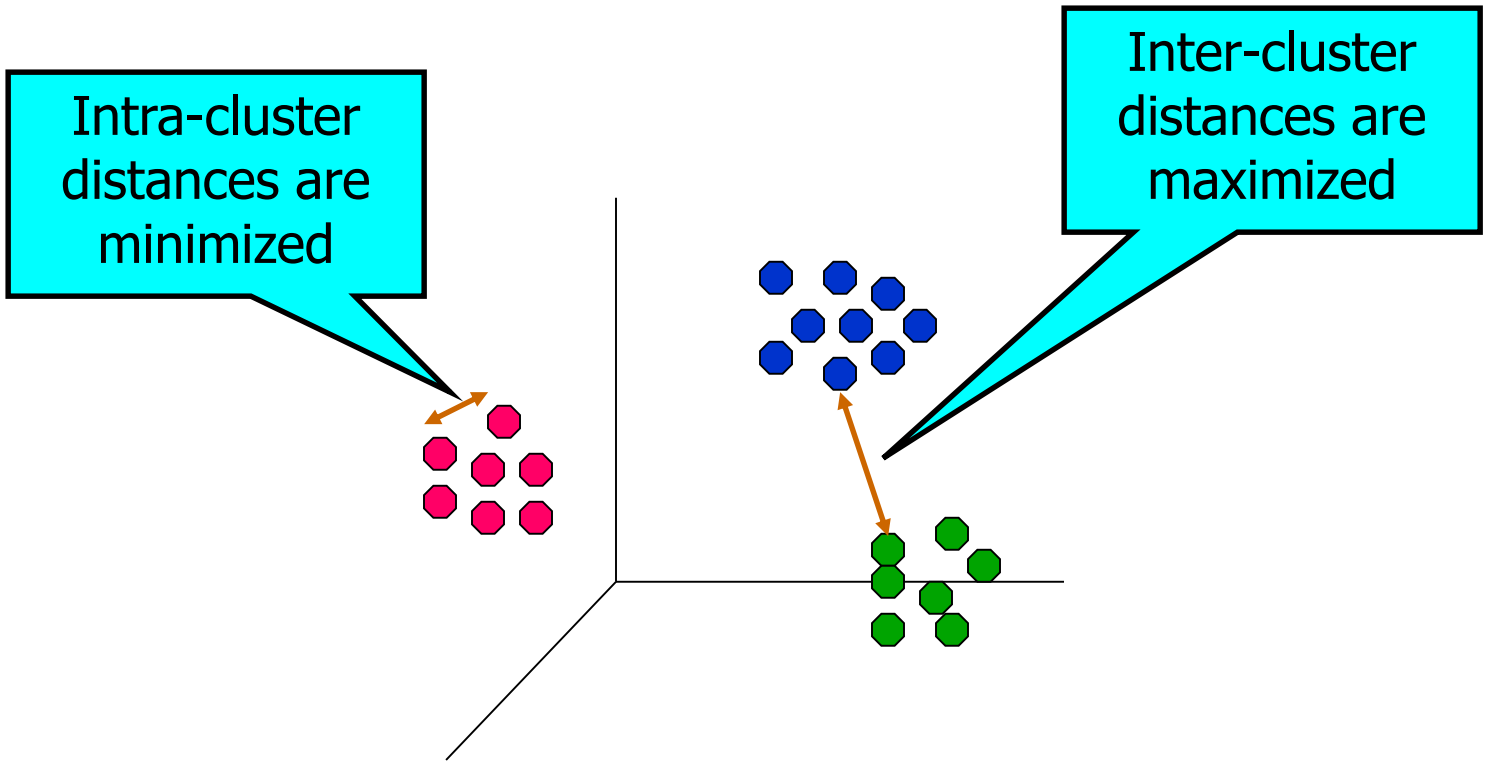
Segmentation Tree:

- *A segmentation tree is a variation of a decision tree, which visually shows the division of a problem into smaller possible problem segments*
- *Helps to divide problem based on knowledge and understanding*
- *Helpful when valid statistical or research data is not available*
- *Decision node - specifies some rules to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome*
- *Leaf node - indicates the value of the target attribute*



Cluster Analysis:

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



What is not Cluster Analysis?

Supervised classification

- Have class label information

Simple segmentation

- Dividing students into different registration groups alphabetically, by last name

Results of a query

- Groupings are a result of an external specification

Graph partitioning

- Some mutual relevance and synergy, but areas are not identical

Similarity and Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- *Minkowski distance*:

  where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer
- If $q = 1$, d is Manhattan distance

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

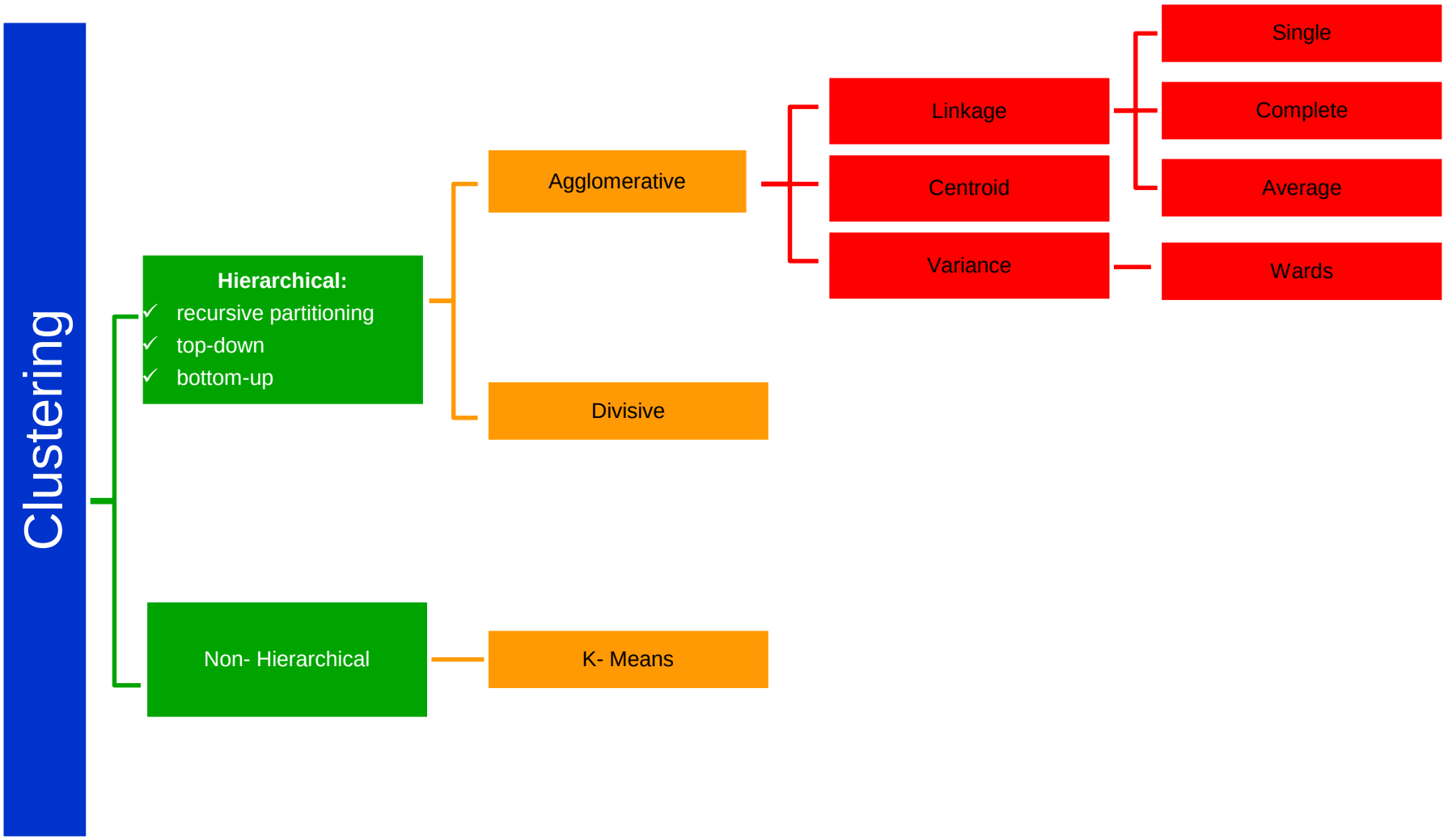
- *If* $q = 2$, d is Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

Requirements of Clustering:

- Scalability
- Ability to deal with different types of attributes
- Ability to handle dynamic data
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

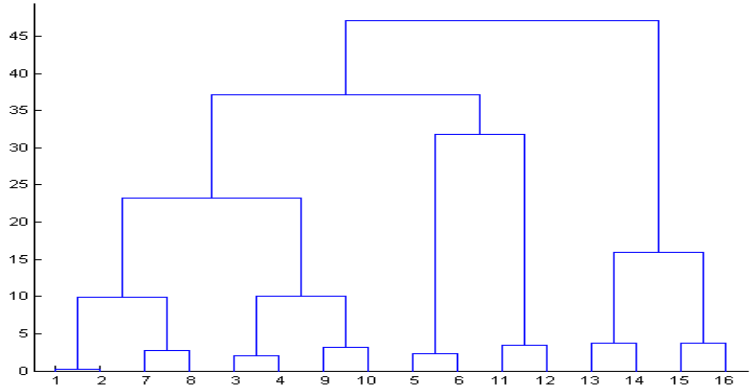
Clustering Methods:



Hierarchical Clustering:

- Agglomerative (Bottom-up)
  - Compute all pair-wise pattern-pattern similarity coefficients
  - Place each of *n* patterns into a class of its own
  - Merge the two most similar clusters into one
    - Replace the two clusters into the new cluster
    - Re-compute inter-cluster similarity scores w.r.t. the new cluster
  - Repeat the above step until there are *k* clusters left (*k* can be 1)

- Divisive (Top-down)
  - Start at the top with all patterns in one cluster
  - The cluster is split using a flat clustering algorithm
  - This procedure is applied recursively until each pattern is in its own singleton cluster



Agglomerative Clustering:

Single linkage: smallest distance between an element in one cluster and an element in the other, i.e., $dis(K_i, K_j) = min(t_{ip}, t_{jq})$

Complete linkage: largest distance between an element in one cluster and an element in the other, i.e., $dis(K_i, K_j) = max(t_{ip}, t_{jq})$

Average linkage: avg distance between an element in one cluster and an element in the other, i.e., $dis(K_i, K_j) = avg(t_{ip}, t_{jq})$

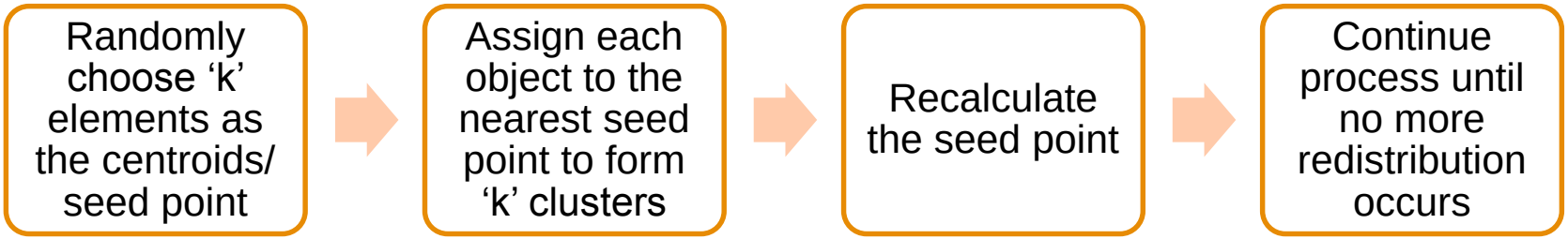
Centroid: distance between the centroids of two clusters, i.e., $dis(K_i, K_j) = dis(C_i, C_j)$

Wards:

- Uses variance, instead of distances
- Method starts out with *n* clusters of size 1 and continues until all the observations are included into one cluster
- At each step observations are combined to minimize the results of error from the squares or maximize the r^2
- It stops when all sample units are combined into a single large cluster of size *n*
- This method is most appropriate for quantitative variables, and not binary variables.

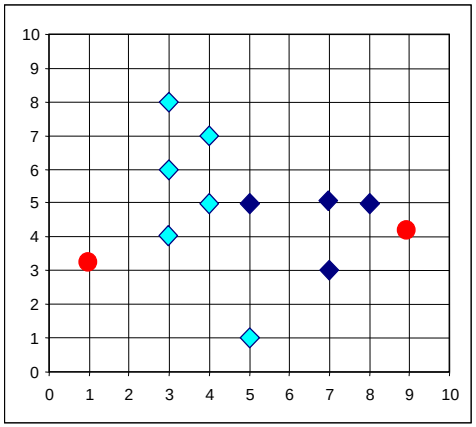
Non-Hierarchical Clustering:

- K- Means
  - Most commonly used **non hierarchical** clustering method.
  - Pre-decided number of clusters (k)
  - ‘Closeness’ is measured by Euclidean distance, cosine similarity, correlation, etc.
  - K-means will converge for common similarity measures mentioned above.
  - Most of the convergence happens in the first few iterations.



The K-Means Clustering Method:

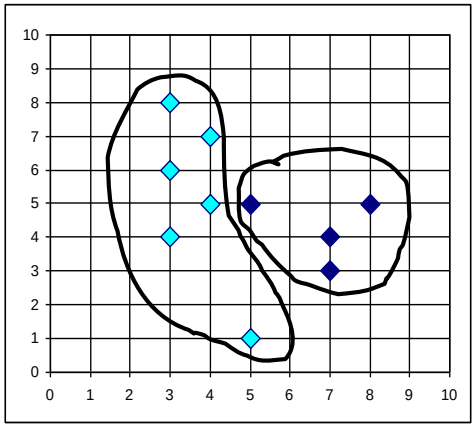
Example



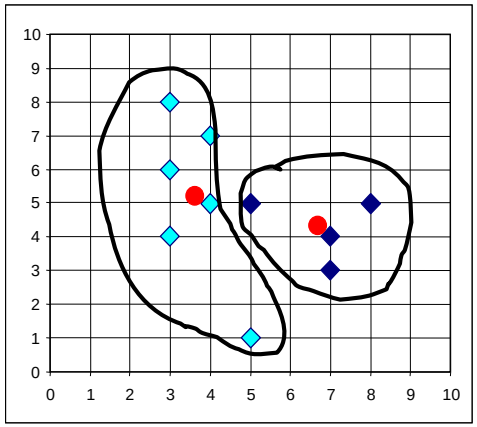
K=2

Arbitrarily choose K object as initial cluster center

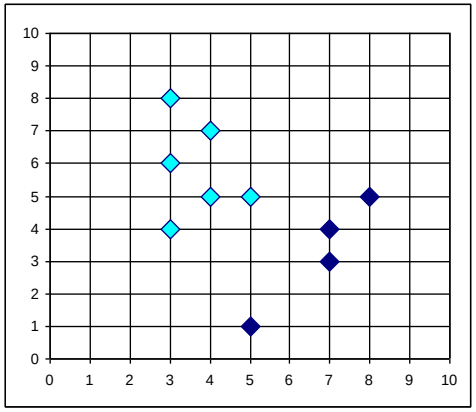
Assign each objects to most similar center



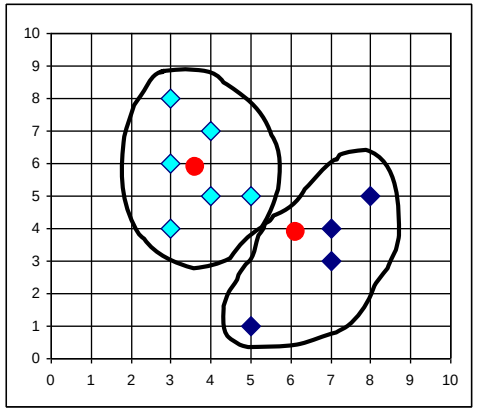
Update the cluster means



reassign



Update the cluster means



reassign

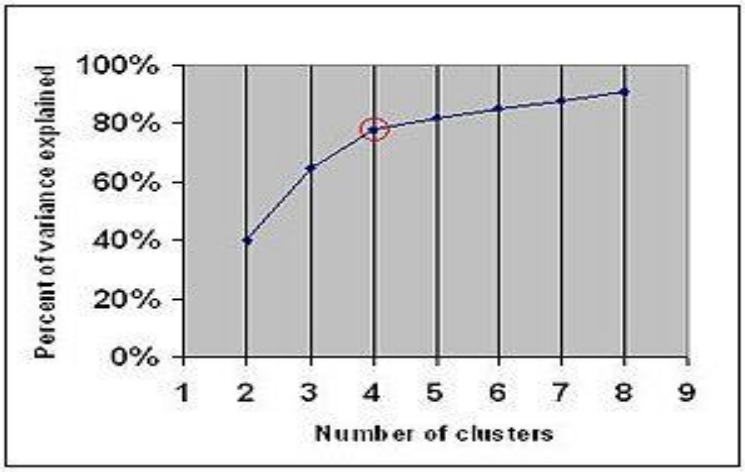
Optimal Cluster:

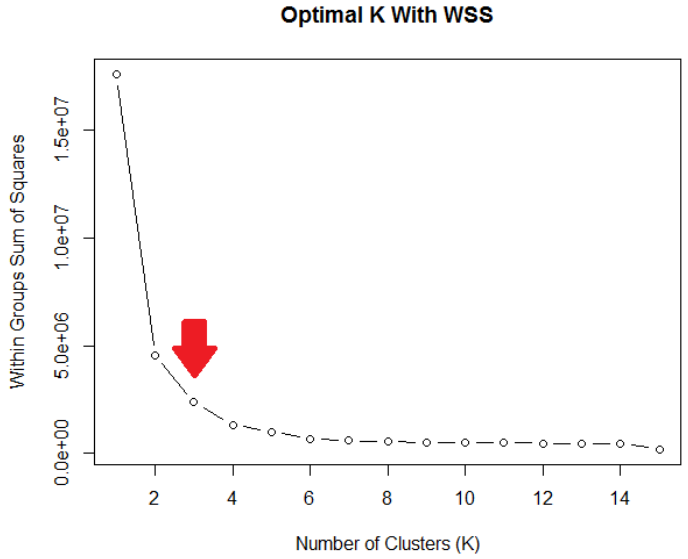
- Choice of *k* is ambiguous and crucial to analysis
  - Increasing k reduces the amount of error but makes computation tough
- **Rule of thumb**

  Set k=$\sqrt{n/2}$
- Plot Following Statistics against number of cluster
  - R-sq (The Overall R-Square)
  - ERSQ (The Expected R-Square)
  - SPRSQ (The Semi Partial R-Square)
  - CCC (The Cubic Clustering Criterion)

The RSQ and ERSQ increase as number of cluster increases, but SPRSQ will gradually decrease. Take that point (elbow Value) after which graphs of these against number of cluster become flat

Choose number of clusters as that value where CCC attains a peak. Large negative values can indicate outliers





Running Cluster Analysis:

- Data Understanding:
- Treatment to Categorical Variables: Create n dummy variables for a categorical variable with n categories
- Missing Value Treatment
- Multicollinearity
- Remove Outliers
- Standardize Variables
- Choose Clustering Method : For small datasets use hierarchical clustering process using ward‟s method or K-means method can be used when the size of the dataset is large.
- Create Training & Validation sample