Agenda

R Setup
Clustering Overview
Hierarchical Clustering
Non-Hierarchical Clustering

Python Setup for Clustering

- Python 3.2.3 or higher version should be installed
- Following Libraries are installed. Check by running the below command

  ## it is okay if you get Warning Message, but you should not get Error Message

  sklearn

  seaborn

matplotlib

Clustering Overview

Intro: Data – Information – Knowledge – Wisdom (DIKW)

Where is the Life… we have lost in Living?
Where is the wisdom… we have lost in Knowledge?
Where is the Knowledge…we have lost in Information?

- T. S. Eliot

- Where is the Information… we have lost in Data?

In order to go from

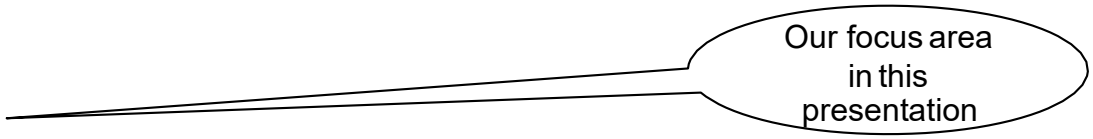
Data to Information to Knowledge and to Wisdom

We need to simplify Data

How do we simplify data

- Simplify data equals to eliminating data complexity
  - too many variables to a few variables
  - too many records to a few records

- Dimension Reduction
  - Reduce number of variable by using techniques like Principal Component Analysis, Collinearity Check, Business Rule Based, etc



- **Clustering / Segmentation**
  - Case reduction

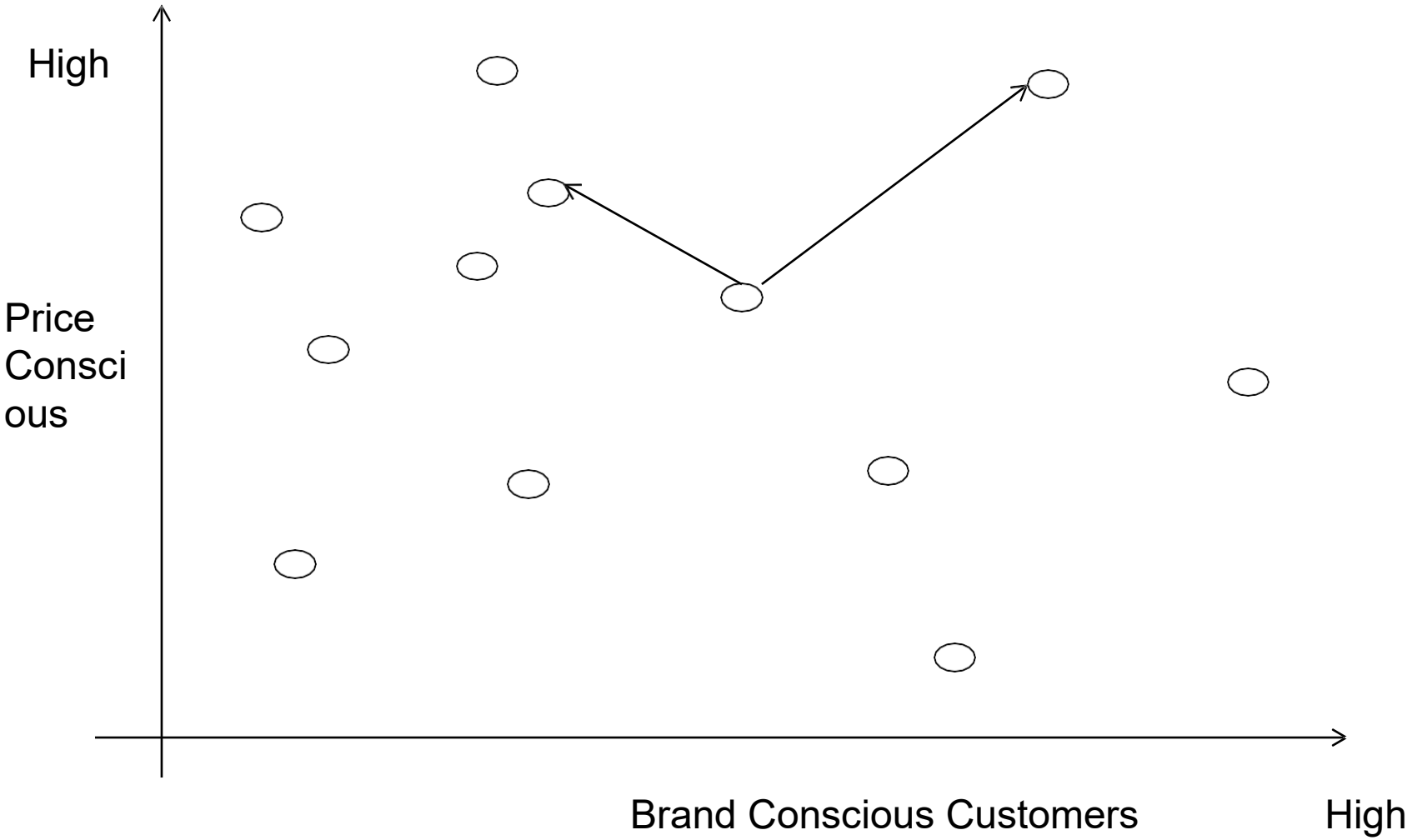
– Reduce the number of records by identifying similar groups and representing them as a cluster

Clustering

- Clustering is a technique for finding similar groups in data, called clusters.
- It groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters.

- A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters

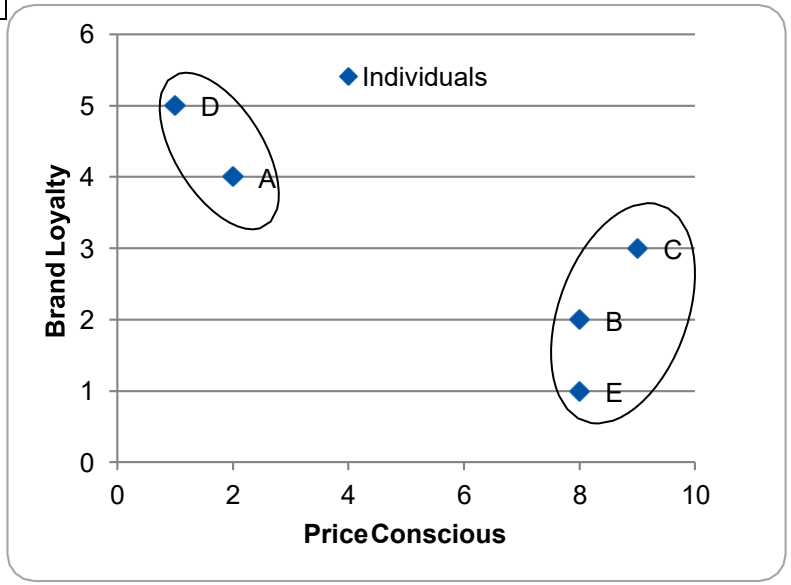
- How do we define "Similar" in clustering?


High
Price Consci ous
Brand Conscious Customers
High

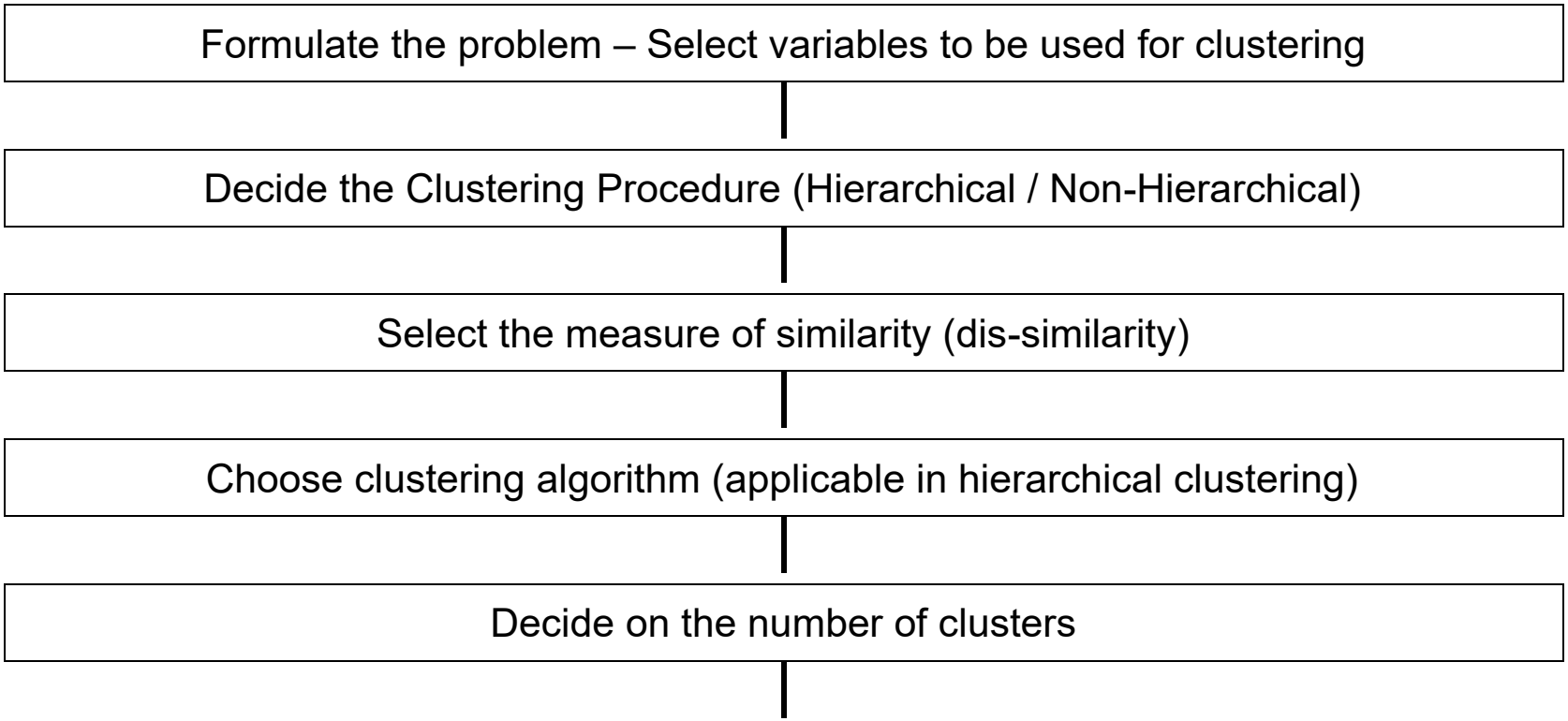
Simple Clustering e.g.

Individuals	Price Conscious	Brand Loyalty
A	2	4
B	8	2
C	9	3
D	1	5
E	8	1

- From Scatter Plot we can see that
  - A & D form one segment of customers who are very high on Brand Loyalty
  - B, C, & E is another segment which is very Price Conscious

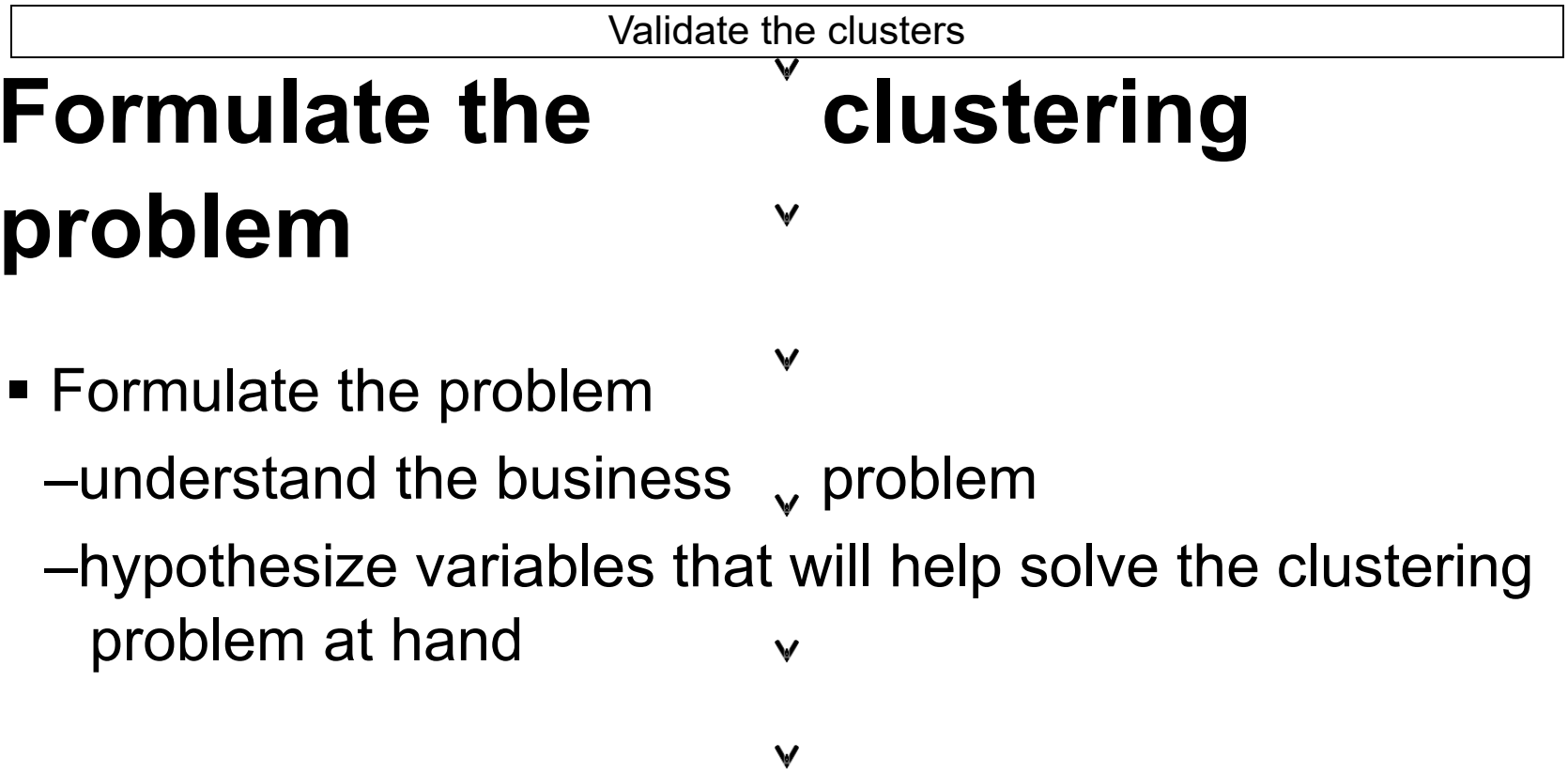


Note: The above e.g. is just an hypothetical data to introduce the subject of clustering. It is not related to the below url: http://globalbizresearch.org/chennai_conference/pdf/pdf/ID_C405_Formatted.pdf



Interpret the cluster output (Profile the clusters)

Steps involved in Clustering Analysis

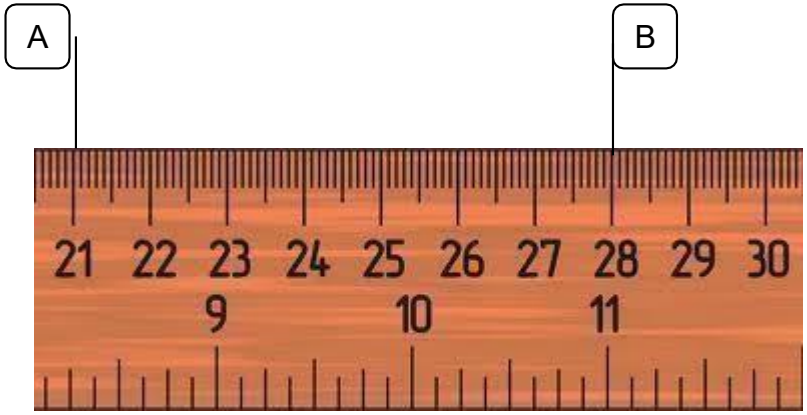


- Formulate the problem
  - –understand the business problem
  - –hypothesize variables that will help solve the clustering problem at hand

- Applications of Clustering Technique
  - –Store Clustering
  - –Customer Clustering
  - –Village Affluency Categorization

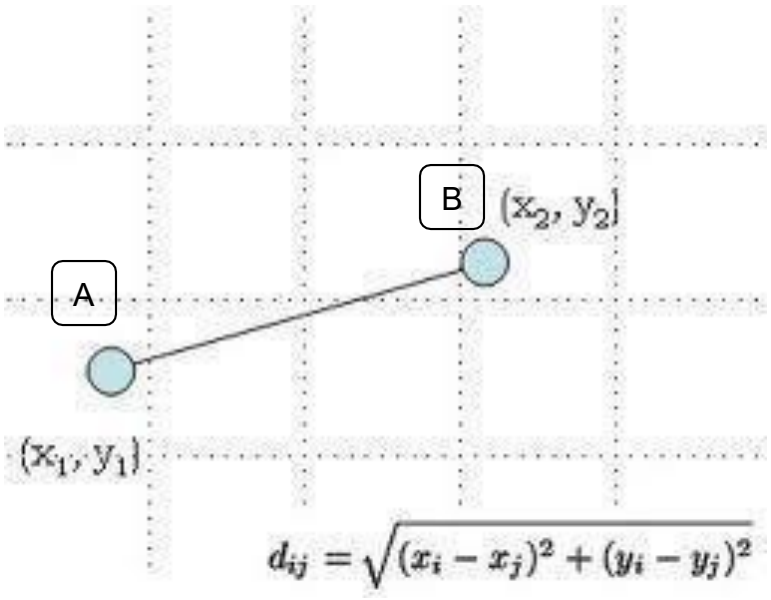
Note: It is preferable to do Factor Analysis / Principal Component Analysis before clustering

Distance Computation



What is the distance between Point A and B?

Ans: 7

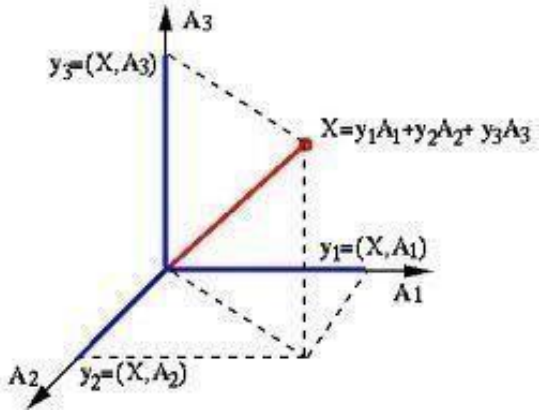


What is the distance between Point A and B?

Ans: $\sqrt{[(x2-x1)_2 + (y2-y1)_2]}$

(Remember the Pythagoras Theorem)

Distance Computation Contd…



- What is the distance between Point A and B in n-Dimension Space?

- If A (a_1, a_2, … a_n) and B (b_1, b_2, … b_n) are cartesian coordinates
- By using Euclidean Distance (which is an extension of Pythagoras Theorem), we get Distance AB as

- $D_{AB} = \sqrt{[(a_1-b_1)^2 + (a_2-b_2)^2 + \ldots + (a_n-b_n)^2]}$

Chebyshev Distance

- In mathematics, **Chebyshev distance** is a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension
- Assume two vectors: A (a_1, a_2, … a_n) & B (b_1, b_2, … b_n)
- Chebyshev Distance

$$= \text{Max} (| a_1 - b_1 | , | a_2 - b_2 | , ... | a_n - b_n |)$$

- Application: Survey / Research Data where the responses are Ordinal

https://en.wikipedia.org/wiki/Chebyshev_distance

Manhattan Distance

- Manhattan Distance also called City Block Distance
- Assume two vectors: A (x1, x2, …xn) & B (y1, y2,…yn)
- Manhattan Distance

$$= | a_1 - b_1 | + | a_2 - b_2 | + \ldots\ldots | a_n - b_n |$$

Block Manhattan Distance = 8 + 4 = 12

Block Chebyshev Distance = Max (8, 4) = 8

Block Eucledian Distance = sqrt (8^2 + 4^2) = 8.94

Block Block Block Block Block Block Block Block **B**

K Means Clustering

- K-Means is the most used, non-hierarchical clustering technique
- It is not based on Distance…
- It is based on within cluster Variation, in other words Squared Distance from the Centre of the Cluster
- The algorithm aims at segmenting data such that within cluster variation is reduced

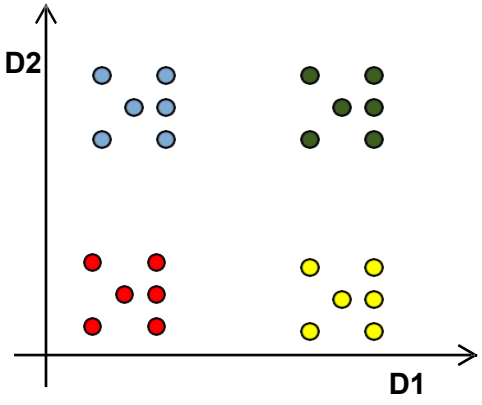
K Means Algorithm

- Input Required : No of Clusters to be formed. (Say K)
- Steps
  1. Assume K Centroids (for K Clusters)
  2. Compute Eucledian distance of each objects with these Centroids.
  3. Assign the objects to clusters with shortest distance
  4. Compute the new centroid (mean) of each cluster based on the objects assigned to each clusters. The K number of means obtained will become the new centroids for each cluster
  5. Repeat step 2 to 4 till there is convergence
     - i.e. there is no movement of objects from one cluster to another
     - Or threshold number of iterations have occurred

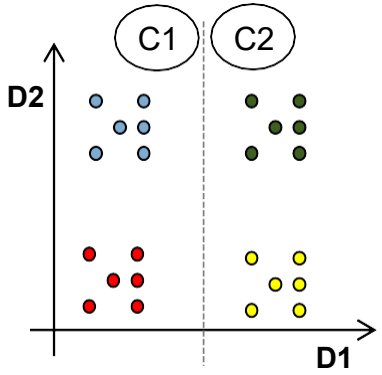
K-means advantages

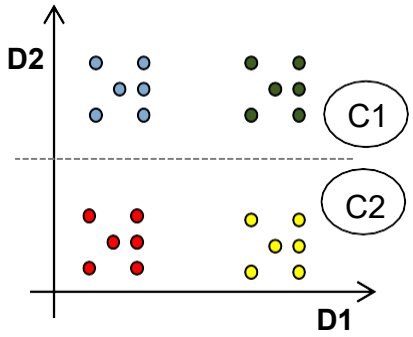
- K-means is superior technique compared to Hierarchical technique as it is less impacted by outliers
- Computationally it is more faster compared to Hierarchical
- Preferable to use on interval or ratio-scaled data as it uses Eucledian distance… desirable to avoid using on ordinal data
- **Challenge – Number of clusters are to be pre-defined and to be provided as input to the process**

Data to be clustered



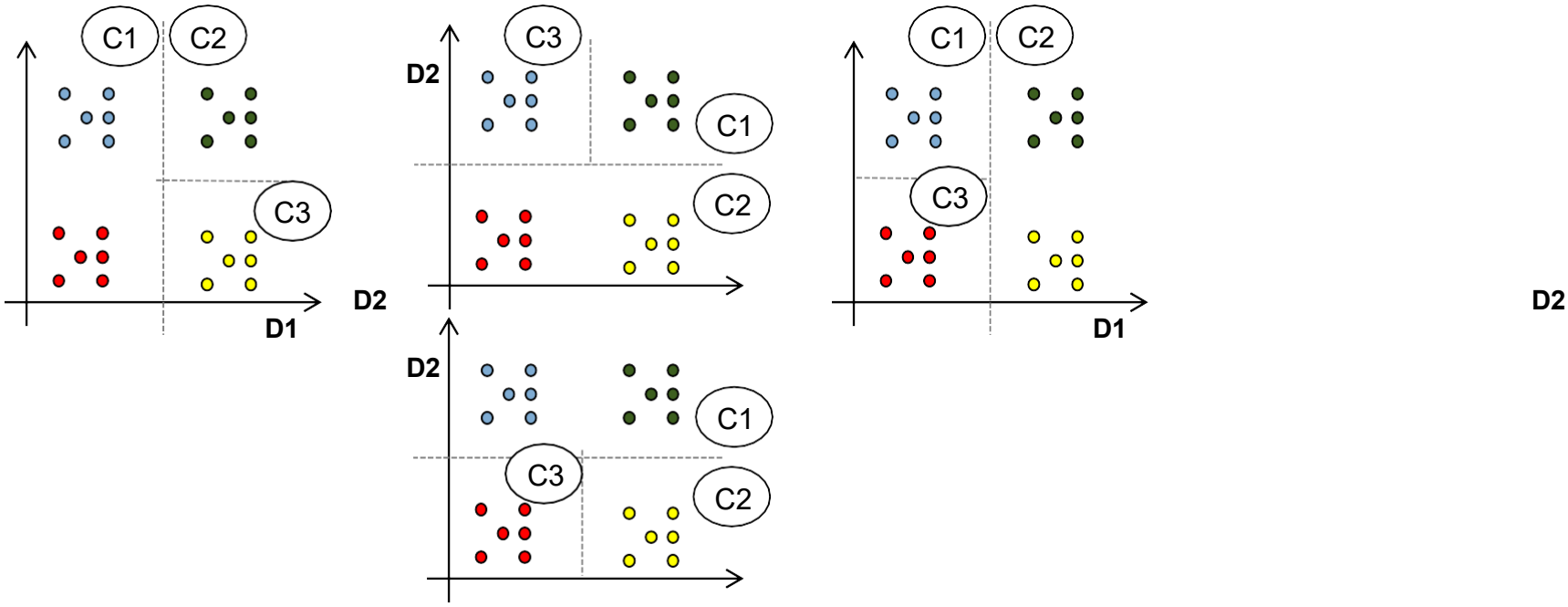
- Two Clusters – 2 possible solution





Why find optimal No. of Clusters?

- Three Clusters – Multiple possible solution

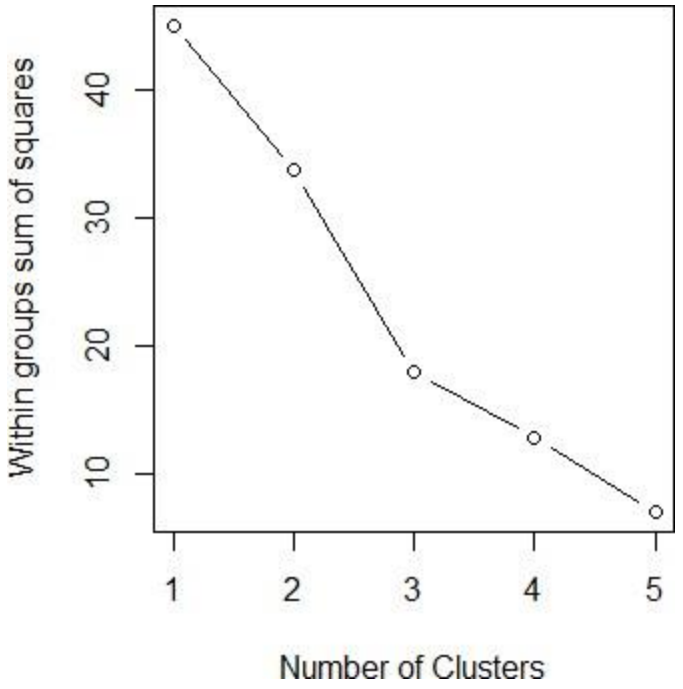


Optimal No. of Clusters

WSS Plot or Within Sum of Square Error Plot is used to identify the optimal number of clusters.

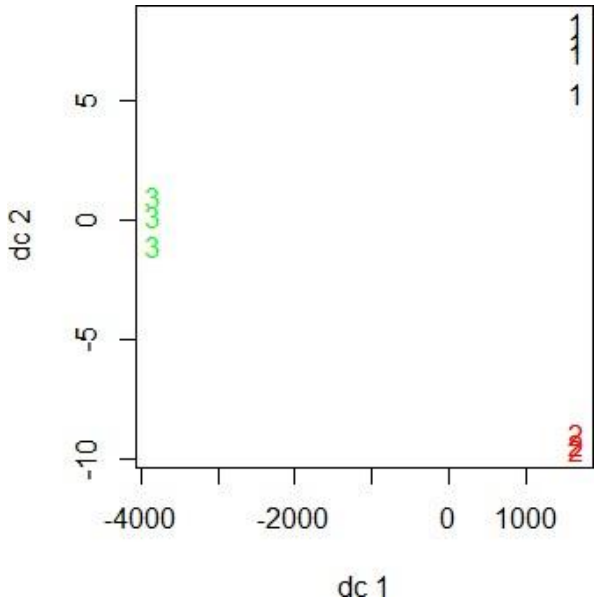
WSS plot is also called as Scree Plot or Elbow Curve within the Analytics Industry.

Elbow in the graph represents the optimal value of the clusters.


Within groups sum of squares
10
20
30
40
1
2
3
4
5
Number of Clusters

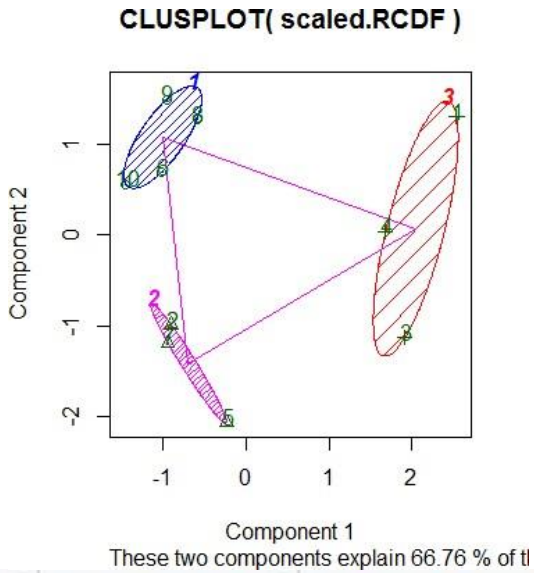
Plotting the clusters

plotting the clusters



More better plot

Profiling the clusters



Cluster	Freq	Avg_Mthly_Spend	No_Of_Visits	Apparel_Items	FnV_Items	Staples_Items
1	4	2375.000	3.000000	0	1.250000	4.500000
2	3	7833.333	4.666667	1	1.666667	2.666667
3	3	5166.667	4.666667	0	11.000000	4.666667

Next steps after clustering

- Clustering provides you with clusters in the given dataset
- Clustering does not provide you rules to classify future records
- To be able to classify future records you may do the following
  - –Build Discriminant Model on Clustered Data
  - –Build Classification Tree Model on Clustered Data

References

- Chapter 9 : Cluster Analysis (http://www.springer.com)
  - Google search : "www.springer.com cluster analysis chapter 9"

Thank you

Pratima