Clustering Overview

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

Our focus area in this presentation

- **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

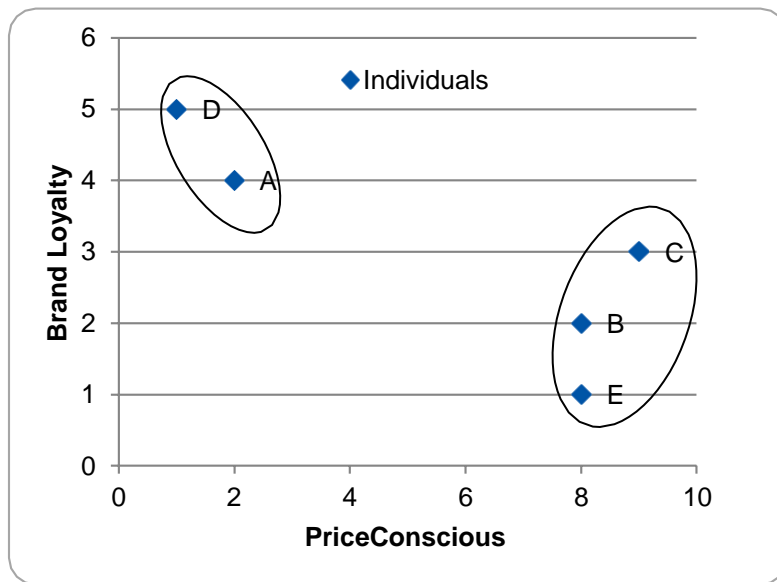- How do we define "Similar" in clustering?

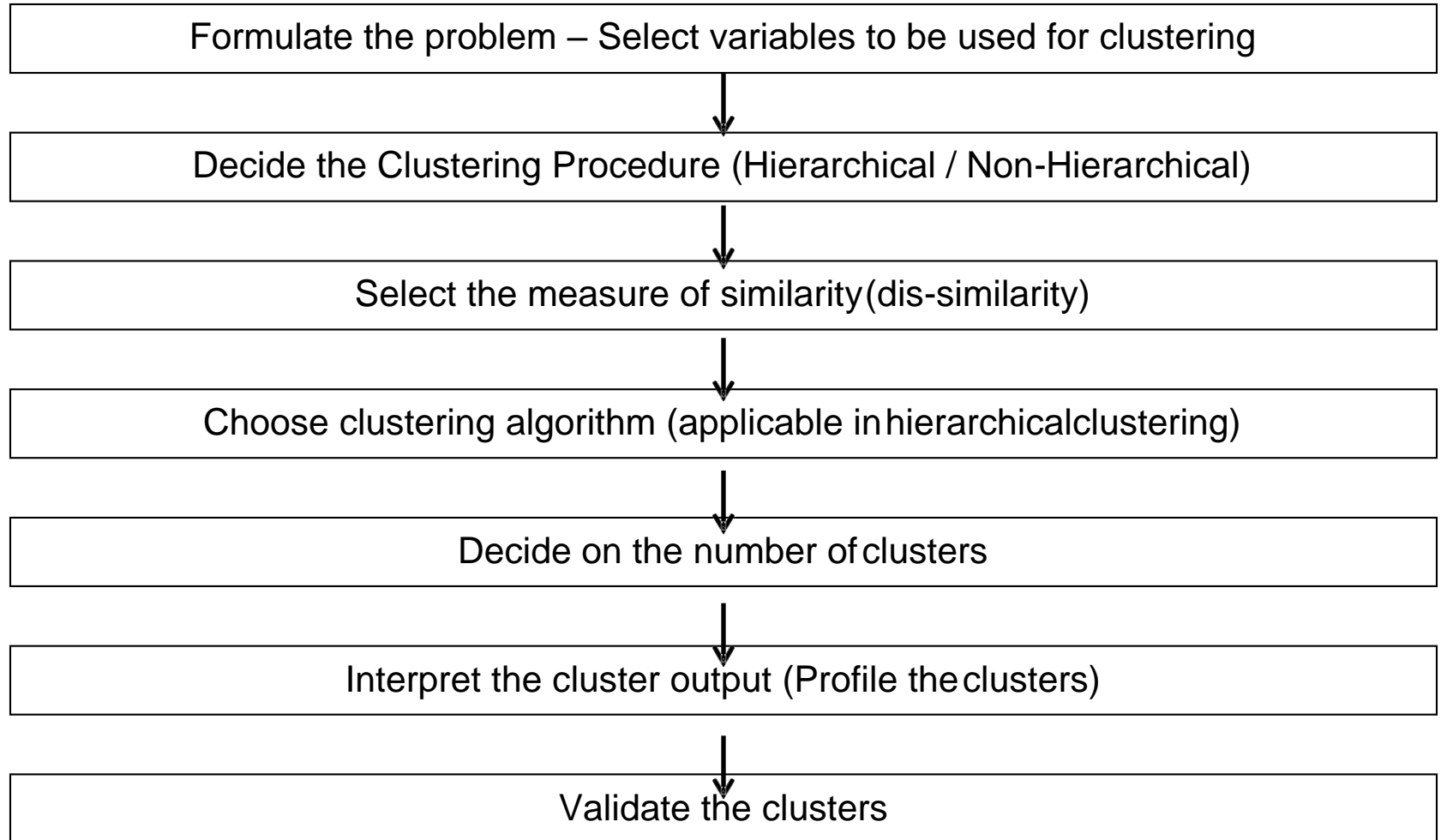**Note: Clustering is an unsupervised learning technique**

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

Steps involved in Clustering Analysis

Formulate the problem – Select variables to be used for clustering

↓

Decide the Clustering Procedure (Hierarchical / Non-Hierarchical)

↓

Select the measure of similarity (dis-similarity)

↓

Choose clustering algorithm (applicable in hierarchical clustering)

↓

Decide on the number of clusters

↓

Interpret the cluster output (Profile the clusters)

↓

Validate the clusters

Formulate the clustering problem

- 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

Decide the clustering procedure
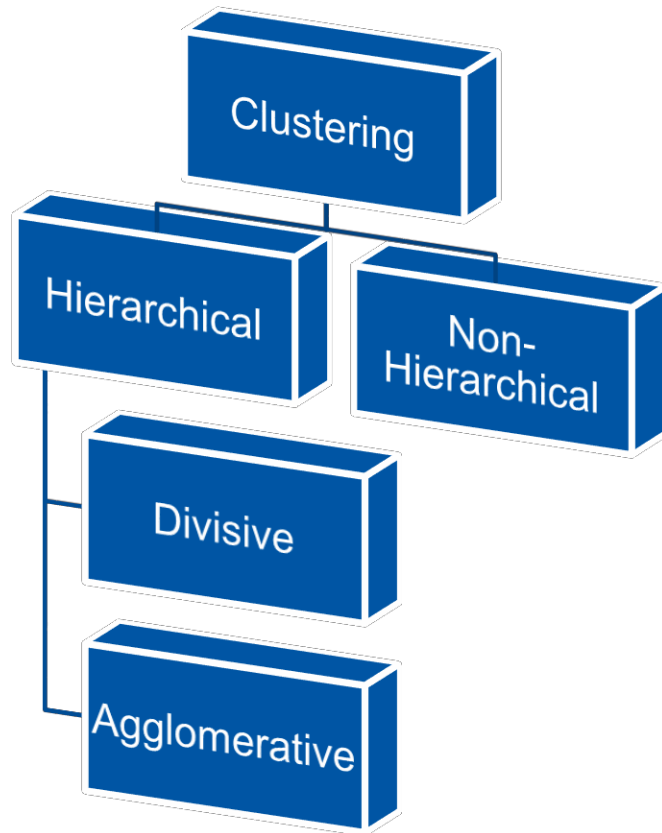
- Types of Clustering Procedures
  - Hierarchical Clustering
  - Non-Hierarchical Clustering

- Hierarchical clustering is characterized by a tree like structure and uses distance as a measure of (dis)similarity

- Non-hierarchical clustering techniques uses partitioning methods and within cluster variance as a measure to form homogeneous groups

Hierarchical Vs. Non-Hierarchical clustering

Hierarchical Clustering

- Relatively very slower

- Agglomerative clustering is mostused algorithm

- Uses distance as a measurefor (dis)similarity

- Helps suggest optimal numbersof clusters in data.

- Object assigned to a cluster remainsin that cluster

Non-Hierarchical Clustering

- Fast and preferable to usewith large datasets

- K-means is a very popular non-hierarchical clusteringtechnique

- Uses within cluster varianceas a measure of similarity

- Non-hierarchical clusteringrequire number of clusters as an input parameter for starting

- Objects can be reassigned to other clusters during the clusteringprocess

Hierarchical Clustering

## Hierarchical Clustering - (dis)similarity measure

- Hierarchical Clustering is based on dis(similarity) measure

- Most software package calculate a measure of dissimilarity by estimating the distance between pair of objects

- Objects with shorter distance are considered similar, whereas objects with larger distance are dissimilar

- Measures of similarity
  - Euclidean distance (most commonly used)
  - City Block or Manhattan distance
  - Chebyshev distance

Distance Computation

A   B
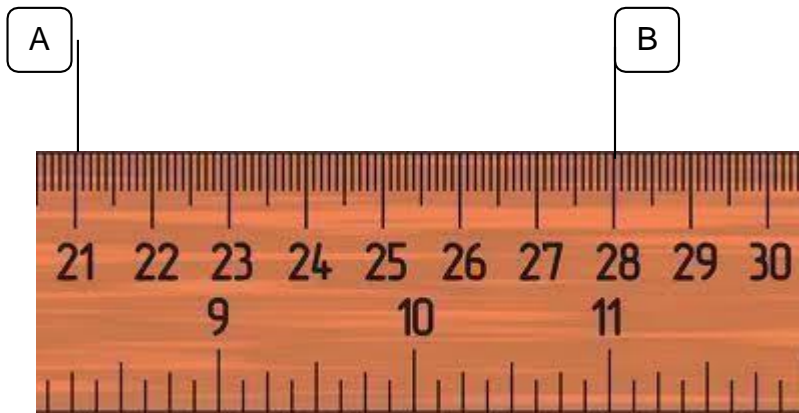
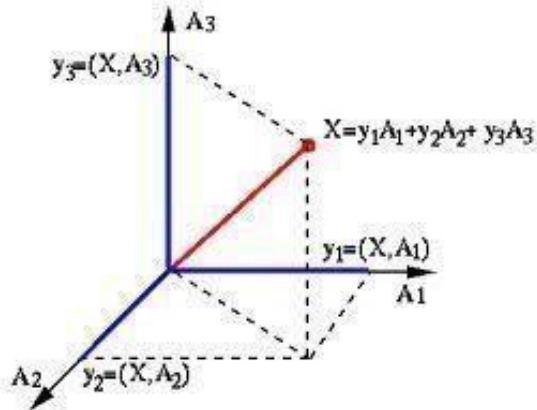What is the distance between Point A and B?

Ans: 7
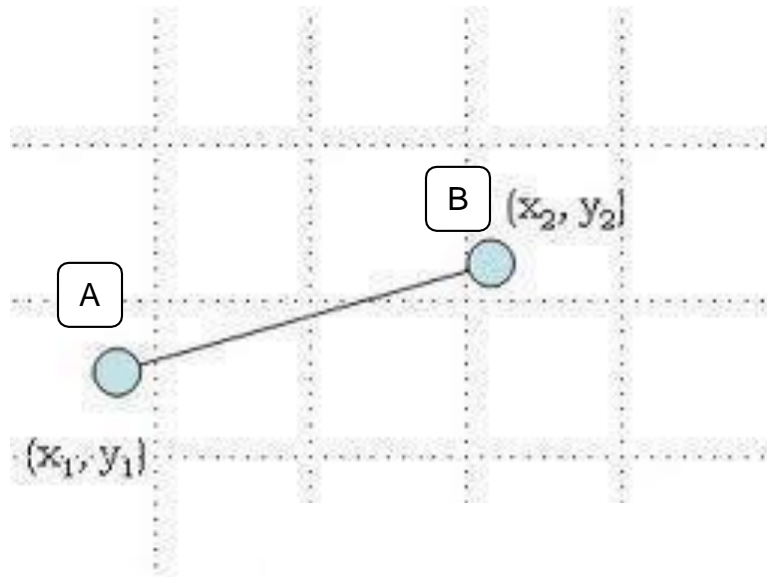
What is the distance between Point A and B?

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 $(a_1, a_2, \ldots a_n)$ and B $(b_1, b_2, \ldots b_n)$ 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 $(a_1, a_2, \ldots a_n)$ & B $(b_1, b_2, \ldots b_n)$

- Chebyshev Distance

$$= \text{Max} \left(\lvert a_1 - b_1 \rvert, \lvert a_2 - b_2 \rvert, \ldots \lvert a_n - b_n \rvert \right)$$

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

Manhattan Distance

- Manhattan Distance also called City Block Distance

- Assume two vectors: A (x1, x2, …xn) & B (y1, y2,…yn)

- Manhattan Distance

$$= |a_1 - b_1| \quad + \quad |a_2 - b_2| \quad +….. \quad |a_n - b_n|$$

A

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

Hierarchical Clustering | Agglomerative Clustering

- Starts with each record as a cluster of one record each

- Sequentially merges 2 closest records by distance as a measure of (dis)similarity to form a cluster. This reduces the number of records by 1

- Repeat the above step with new cluster and all remaining clusters till we have one big cluster

Step 0   Step 1   Step 2   Step 3   Step 4

a

a b

b

a b c d e

c

c d e

d

d e

e
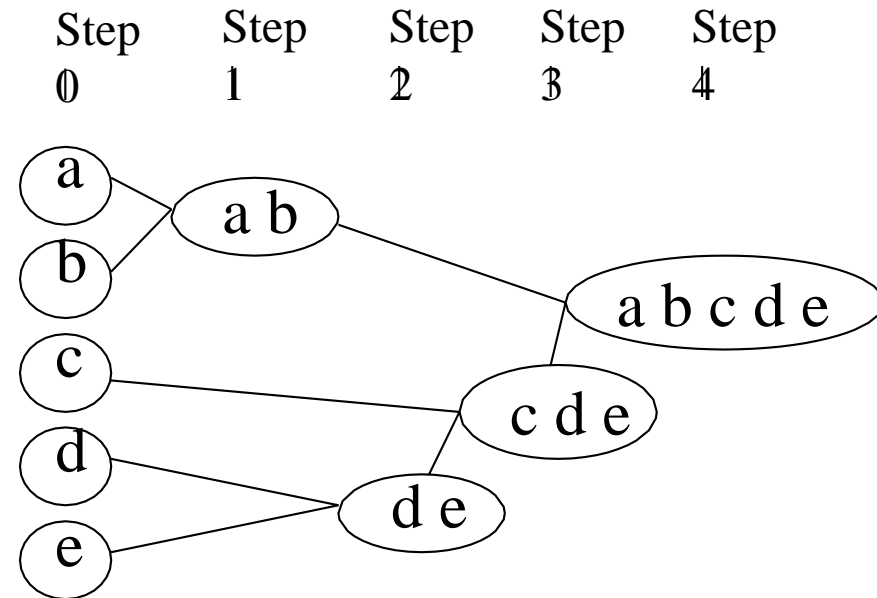
How do you measure the distance between cluster (a,b) and (c) or the cluster (a,b) and (d,e) ????

Agglomerative Clustering Procedures

- Single linkage – Minimum distance or Nearest neighbour rule

- Complete linkage – Maximum distance or Farthest distance

- Average linkage – Average of the distances between all pairs

- Centroid method – combine cluster with minimum distance between the centroids of the two clusters

- Ward's method – Combine clusters with which the increase in within cluster variance is to the smallest degree

http://sites.stat.psu.edu/~ajw13/stat505/fa06/19_cluster/09_cluster_wards.html
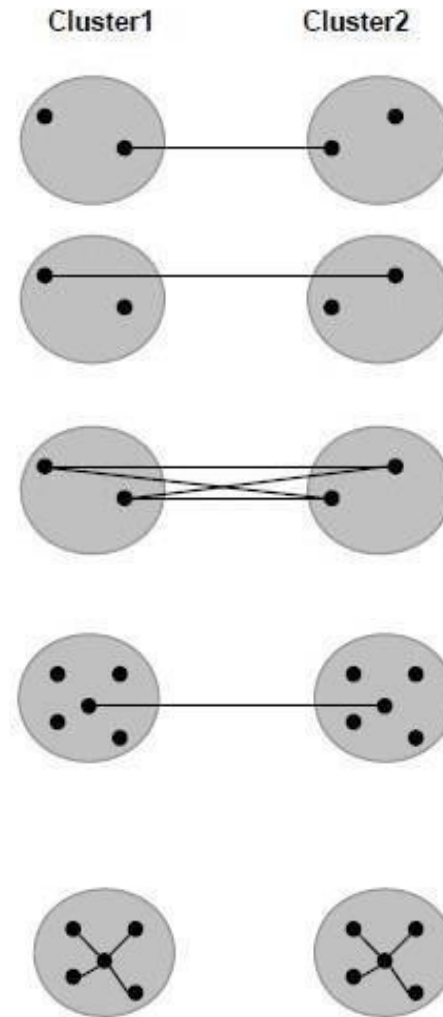
Clustering e.g. 1 : Clustering for Retail Customers

Let us find the clusters in given Retail Customer Spends data

We will use Hierarchical Clustering technique

Let us first set the working directory path and import the data

RCDF <- pd.read_csv("datafiles/Cust_Spend_Data.csv", header=TRUE)
View(RCDF)

Cust_ID	Name	Avg_Mthly_Spend	No_Of_Visits	Apparel_Items	FnV_Items	Staples_Items
1	A	10000	2	1	1	0
2	B	7000	3	0	10	9
3	C	7000	7	1	3	4
4	D	6500	5	1	1	4
5	E	6000	6	0	12	3
6	F	4000	3	0	1	8
7	G	2500	5	0	11	2
8	H	2500	3	0	1	1
9	I	2000	2	0	2	2
10	J	1000	4	0	1	7

HyperMarket CustomerSpend
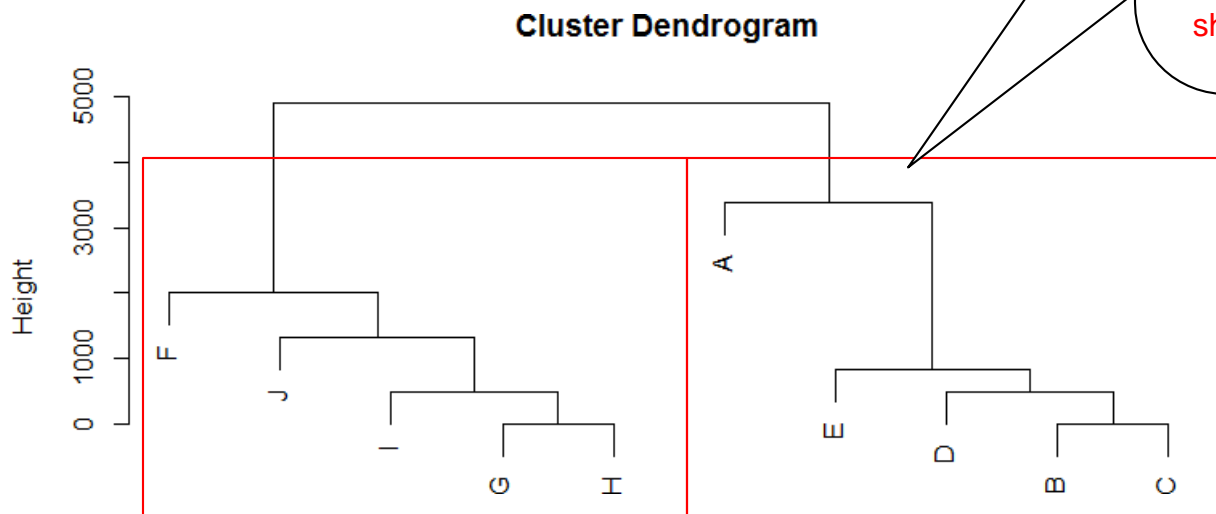
Data - Metatdata

AVG_Mthly_Spend: The average
monthly amount spent by customer

No_of_Visits: The number of times
a customer visited the HyperMarketi

Count of **Apparel, Fruits and
Vegetable, Staple Items** purchased
in a month

Building the hierarchical clusters (without variablescaling)

Dendogram looks like Decision Tree but its not s Tree or classification

Note: The two clustersformed are primarily on the basis of AVG_MTHLY_SPEND

Eucledian Distance computation in this case is influenced by AVG_MTHLY_SPEND variable as the range of this variable is too large compared to the other variables

To avoid this problem, we should scale the variables used for clustering

Cluster Dendrogram

Building the hierarchical clusters (with variablescaling)

scale function standardizes the values

Scaling of features can change the shape and sizes of the
clusters. It can even generate various numbers of clusters.

Cluster Dendrogram
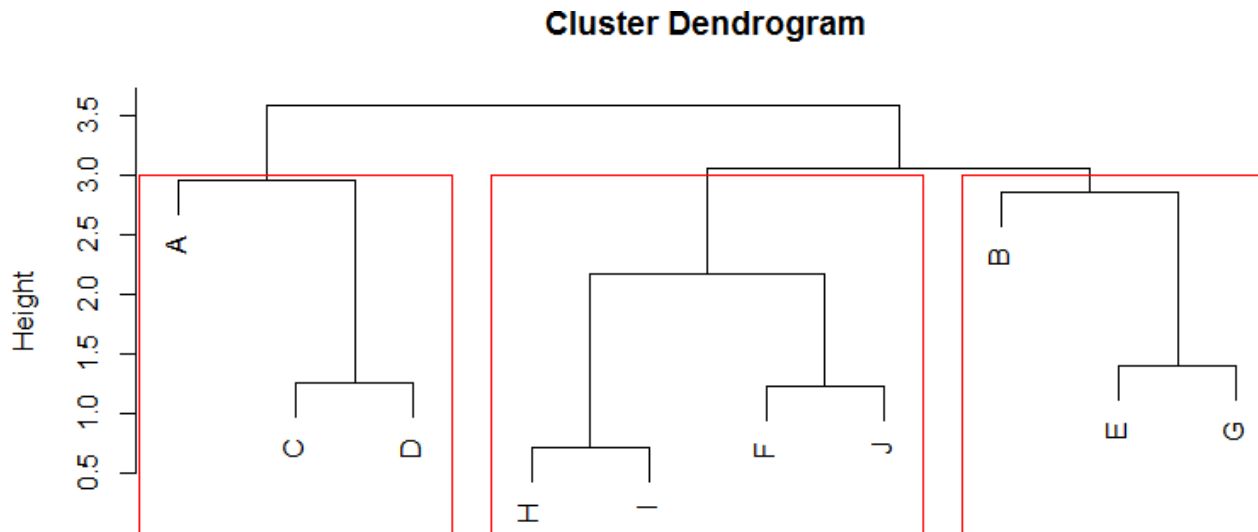
Understanding the Height Calculation in Clustering

Let us see the distance matrix

Dist.	A	B	C	D	E	F	G	H	I
B	4.25								
C	3.41	3.84							
D	2.51	3.47	1.26						
E	4.27	2.70	2.92	3.20					
F	3.98	2.21	3.58	2.85	3.43				
G	4.38	3.02	3.38	3.35	1.41	3.17			
H	3.40	3.60	3.66	2.93	3.24	2.35	2.46		
I	3.53	3.39	4.05	3.21	3.48	2.18	2.61	0.73	
J	4.55	2.97	3.59	3.04	3.41	1.24	2.80	2.12	2.06

Cluster Dendrogram

Dist.	A	B	C	D	E	F	G	H	I
B	4.25								
C	3.41	3.84							
D	2.51	3.47	1.26						
E	4.27	2.70	2.92	3.20					
F	3.98	2.21	3.58	2.85	3.43				
G	4.38	3.02	3.38	3.35	1.41	3.17			
H	3.40	3.60	3.66	2.93	3.24	2.35	2.46		
I	3.53	3.39	4.05	3.21	3.48	2.18	2.61	0.73	
J	4.55	2.97	3.59	3.04	3.41	1.24	2.80	2.12	2.06

C, D	1.26

H, I	0.73

F, J	1.24

E, G	1.41

A, (C,D)	
A, C	3.41
A, D	2.51
A, (C,D)	2.96

(H,I), (F,J)		
	F	J
H	2.35	2.12
I	2.18	2.06
((H,I), (F,J))	2.17	

B, (E,G)	
B, E	2.70
B, G	3.02
B, (E,G)	2.86

(H,I, F,J) , (B, E,G)			
	B	E	G
H	3.60	3.24	2.46
I	3.39	3.48	2.61
F	2.21	3.43	3.17
J	2.97	3.41	2.80
(H,I, F,J) , (B, E,G)	3.06		

(A, C, D) , (H, I, F, J, B, E, G)							
	H	I	F	J	B	E	G
A	3.40	3.53	3.98	4.55	4.25	4.27	4.38
C	3.66	4.05	3.58	3.59	3.84	2.92	3.38
D	2.93	3.21	2.85	3.04	3.47	3.20	3.35
(A, C, D) , (H, I, F, J, B, E, G)	3.59						

Profiling the clusters

profiling the clusters

Profiling of clusters means the aggregation of the patterns identified within the each cluster.
Also, The characteristics of the clusters

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

Non-Hierarchical Clustering (K Means)

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**
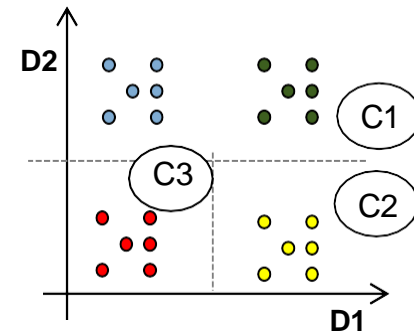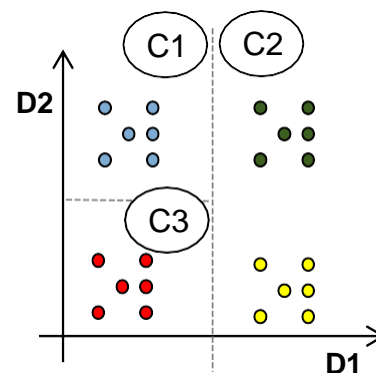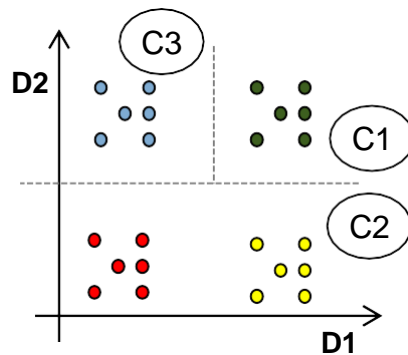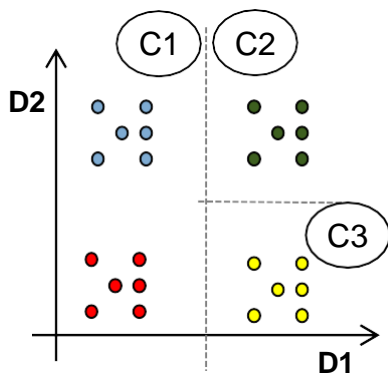
Why find optimal No. of Clusters?

Data to be clustered
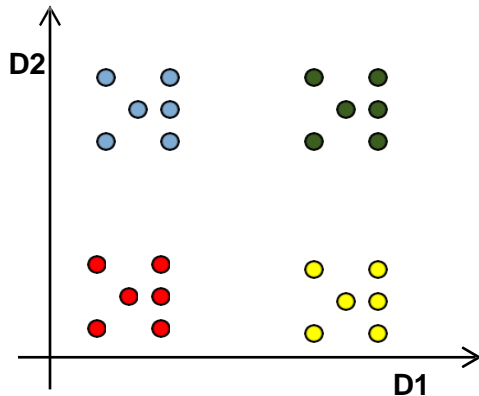
■ Two Clusters – 2 possible solution
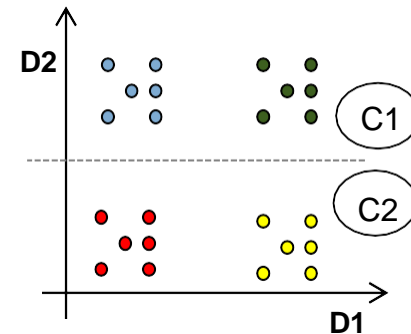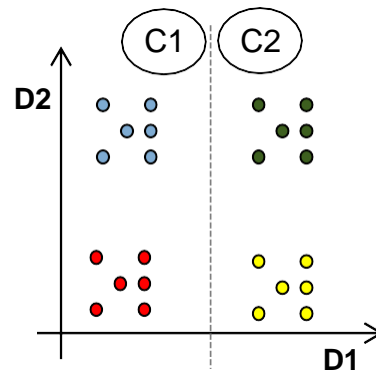
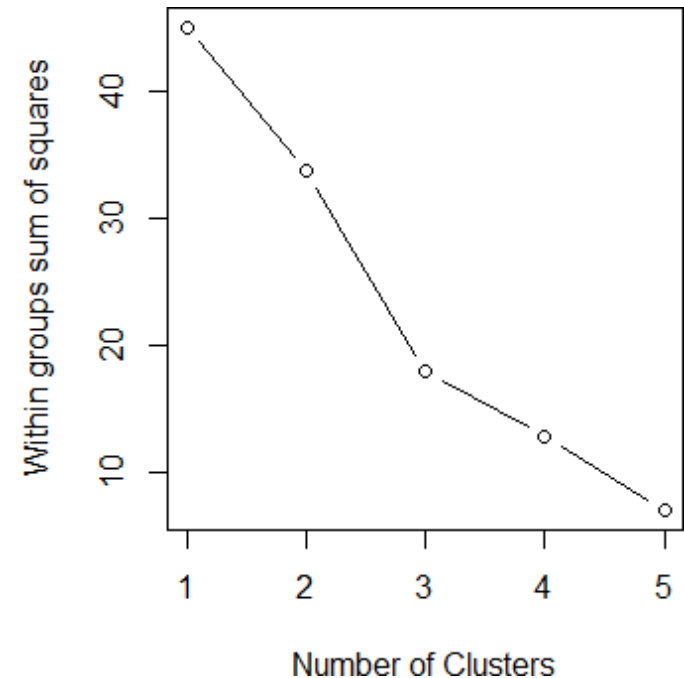**■ Three Clusters – Multiple possible solution**

Optimal No. of Clusters

WSS Plot or Within Sum of Square Error Plot is used to identify the optimal number of clusters.
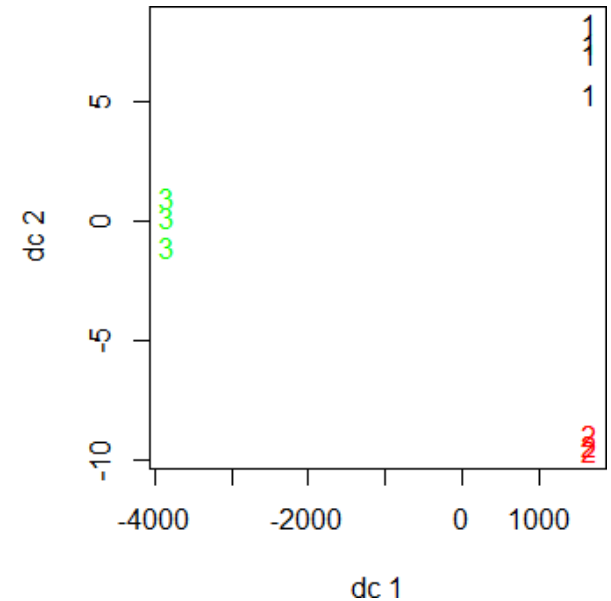
WSS plot is also called as Scree Plot or Elbow Curve within the Analytics Industry.

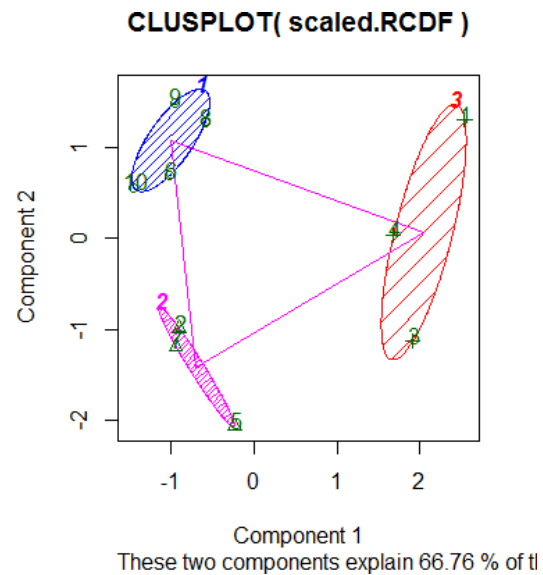Elbow in the graph represents the optimal value of the clusters.

Plotting the clusters

plotting the clusters

More better plot

CLUSPLOT(scaled.RCDF)

Component 1
These two components explain 66.76 % of t[...]

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