SUPERVISED & UNSUPERVISED LEARNING

- *Supervised learning algorithms are those used in classification and* prediction

We must have data available in which the value of the outcome of interest (e.g., purchase or no purchase) is known

e.g. segmentation, rule based grouping

- *Unsupervised learning algorithms are those used where there is no outcome variable* to predict or classify

E.g. Association rules, data reduction methods, and clustering

CLUSTERING TECHNIQUES?

- Hierarchical Clustering
  - Agglomerative
  - Divisive
- K-means Clustering

Thumb rule –

Use K-means clustering only when more than 100 data points are available, else go for hierarchical clustering

Pre-Requisites

- All variables should be numeric
- The variables need to be standardised to remove scale effect
- Multicollinearity is a problem in Cluster analysis. So correlated variables should be excluded from the model
- Outliers also need to be treated

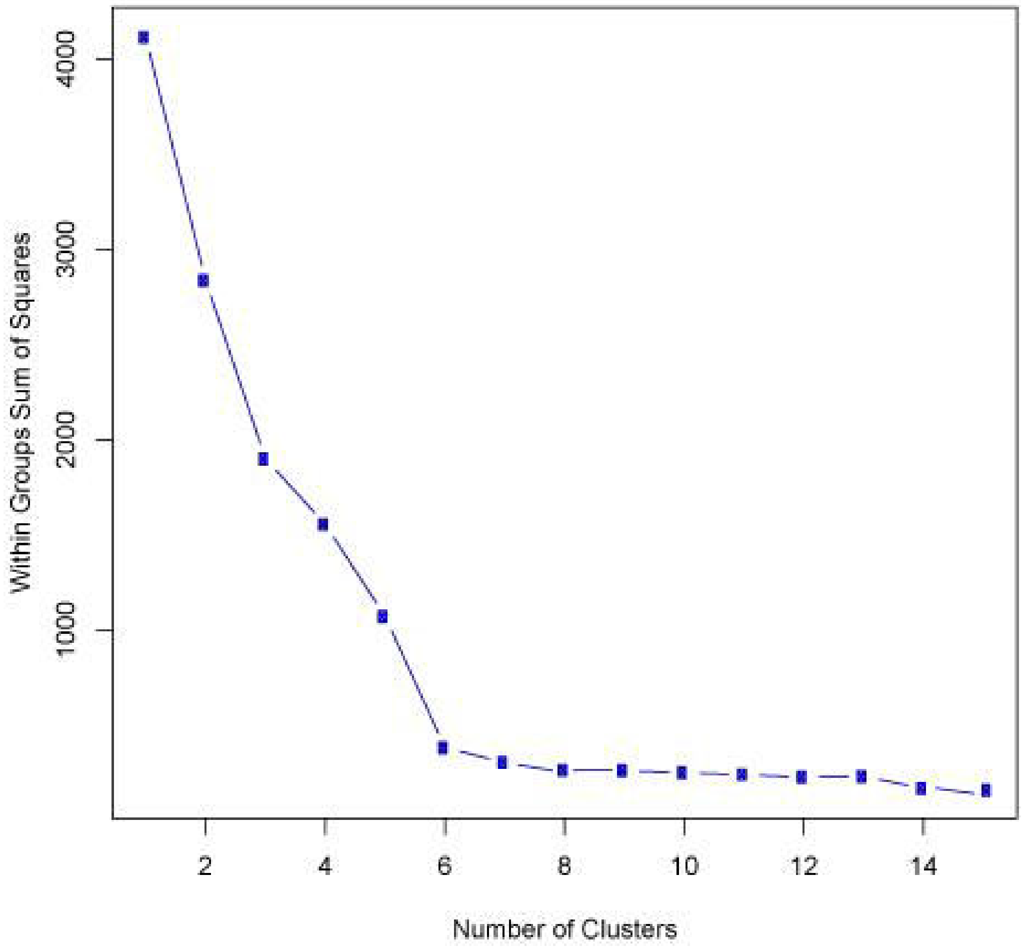
K-MEANS CLUSTERING

- K refers to number of clsuters that we want to build
- The *k-means* algorithm assumes partitioning criteria : minimize intra-cluster similarity and maximizeinter-cluster similarity
- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative reallocation technique to improve the partitioning by moving objects from one group to other.
- Initial centroids are often chosen randomly.
- Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- Closeness is measured by Eucleadian distance

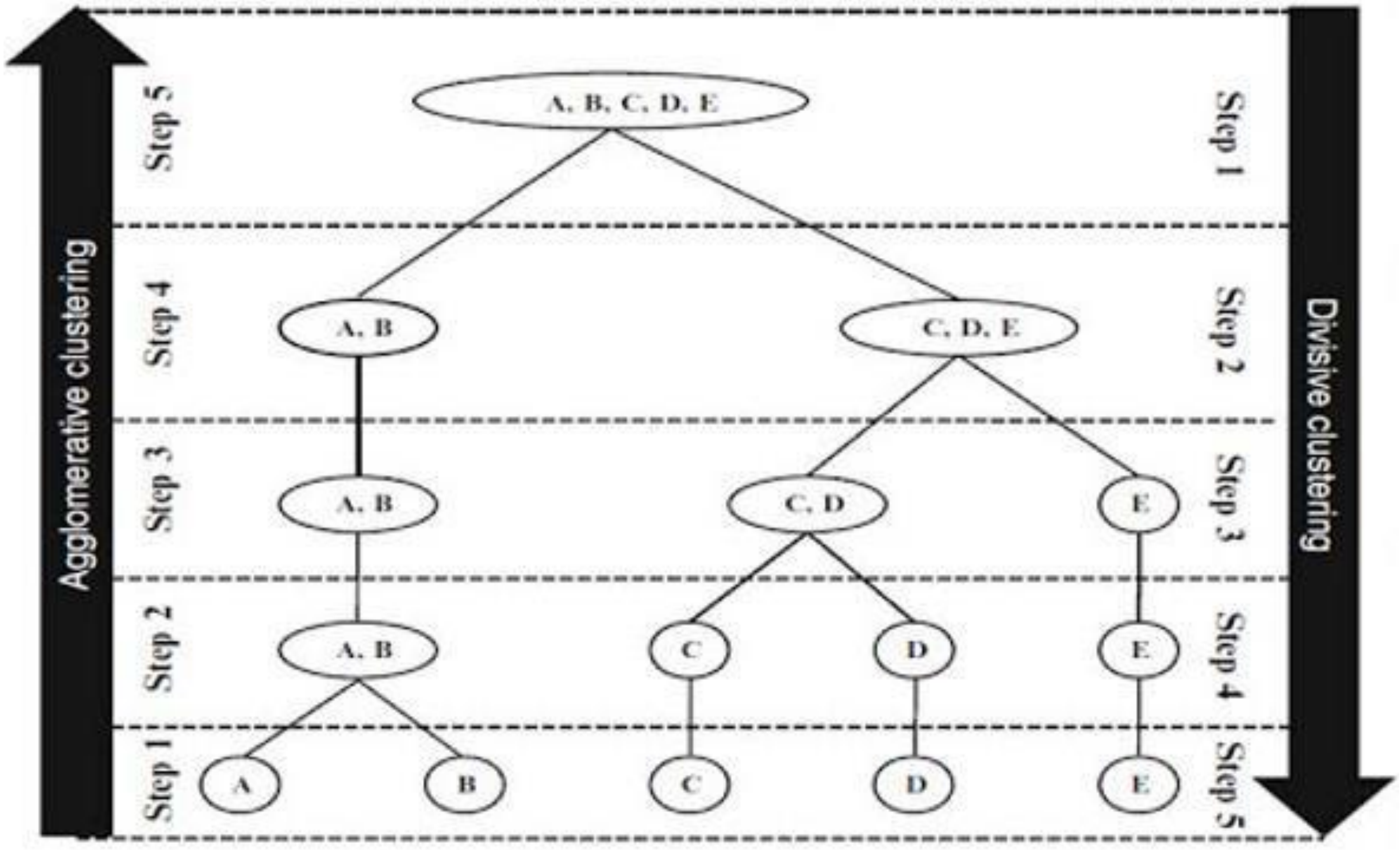
DETERMINING OPTIMAL CLUSTERS

- Since its mandatory to pass the number of clusters for kmeans it becomes difficult to decide for the optimal number of clusters
- For this purpose we make use of the Elbow chart which helps in determining the optimal number of clusters based on the within sum of squares.
- The Elbow chart s also known as a Scree Plot

- The x-axis is the number of clusters
- The y-axis is the within sum of squares(wss)
- The point at which the chart bends or the wss becomes small would be considered as the optimal number of clusters



HIERARCHICAL CLUSTERING



STEPS IN HIERARCHICAL CLUSTERING

- Normalize the data so that all the variables are on the same scale

(X-μ)/σ

- Calculate the distances using Euclidean distance,

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

- Pass the distance and agglomeration method to the algorithm(hclust in R)
- Plot Dendrogram to visualize the clustering

SIMPLE EXAMPLE TO UNDERSTAND HIERARCHICAL CLUSTERING

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

	BOS/NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY	0	223	1308	802	2815	2934	2786	1771
DC	223	0	1075	671	2684	2799	2631	1616
MIA	1308	1075	0	1329	3273	3053	2687	2037
CHI	802	671	1329	0	2013	2142	2054	996
SEA	2815	2684	3273	2013	0	808	1131	1307
SF	2934	2799	3053	2142	808	0	379	1235
LA	2786	2631	2687	2054	1131	379	0	1059
DEN	1771	1616	2037	996	1307	1235	1059	0

SIMPLE EXAMPLE TO UNDERSTAND HIERARCHICAL CLUSTERING

	BOS/NY/DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY/DC	0	1075	671	2684	2799	2631	1616
MIA	1075	0	1329	3273	3053	2687	2037
CHI	671	1329	0	2013	2142	2054	996
SEA	2684	3273	2013	0	808	1131	1307
SF	2799	3053	2142	808	0	379	1235
LA	2631	2687	2054	1131	379	0	1059
DEN	1616	2037	996	1307	1235	1059	0

	BOS/ NY/DC	MIA	CHI	SEA	SF/LA	DEN
BOS/NY/DC	0	1075	671	2684	2631	1616
MIA	1075	0	1329	3273	2687	2037
CHI	671	1329	0	2013	2054	996
SEA	2684	3273	2013	0	808	1307
SF/LA	2631	2687	2054	808	0	1059
DEN	1616	2037	996	1307	1059	0

SIMPLE EXAMPLE TO UNDERSTAND HIERARCHICAL CLUSTERING

	BOS/NY/DC/ CHI	MIA	SEA	SF/LA	DEN
BOS/NY/DC/CHI	0	1075	2013	2054	996
MIA	1075	0	3273	2687	2037
SEA	2013	3273	0	808	1307
SF/LA	2054	2687	808	0	1059
DEN	996	2037	1307	1059	0

	BOS/NY/DC/CHI	MIA	SF/LA/SEA	DEN
BOS/NY/DC/CHI	0	1075	2013	996
MIA	1075	0	2687	2037
SF/LA/SEA	2054	2687	0	1059
DEN	996	2037	1059	0

	BOS/NY/DC/CHI/DEN	MIA	SF/LA/SEA
BOS/NY/DC/CHI/DEN	0	1075	1059
MIA	1075	0	2687
SF/LA/SEA	1059	2687	0

	BOS/NY/DC/CHI/DEN/SF/LA/SEA	MIA
BOS/NY/DC/CHI/DEN/SF/LA/SEA	0	1075
MIA	1075	0

DENDOGRAM

- The x-axis are the observations
- The y-axis is a measure of closeness of either individual data points or clusters
- Longer the line indicates the clusters are clearly apart from each other
- This helps in determining the number of optimal clusters
- The red line indicates the number of clusters