

Operative Marketing Research

Dr. Rohit Trivedi • Universität Hamburg •
Arbeitsbereich Marketing & Innovation

Von-Melle-Park 5 • Raum 3071 • 20146 Hamburg

Tel: +49 40 42838-4643 • Fax: +49 40 42838-5250

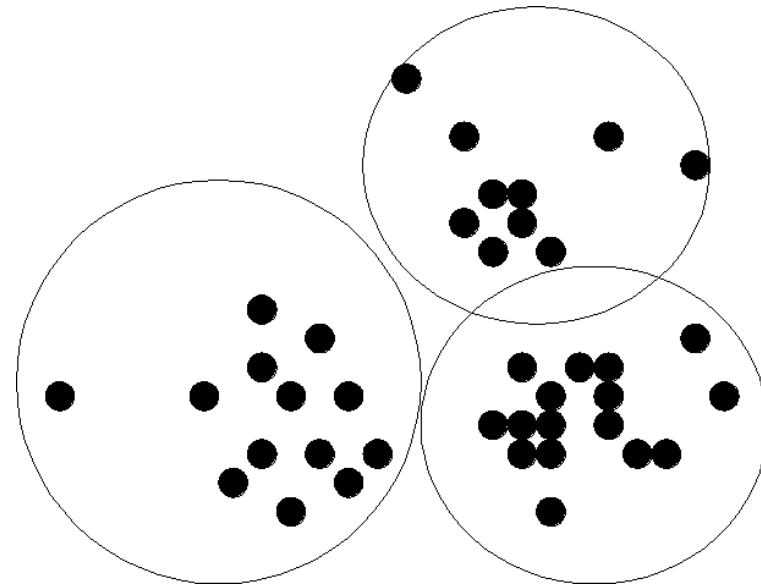
Email: ami@wiso.uni-hamburg.de

Objectives of Today's Lecture

- To understand the phases and benefits of market segmentation
- To introduce cluster analysis and its use in marketing research (e.g. for market segmentation, consumer or product profiling)
- To show the application of cluster analysis with SPSS

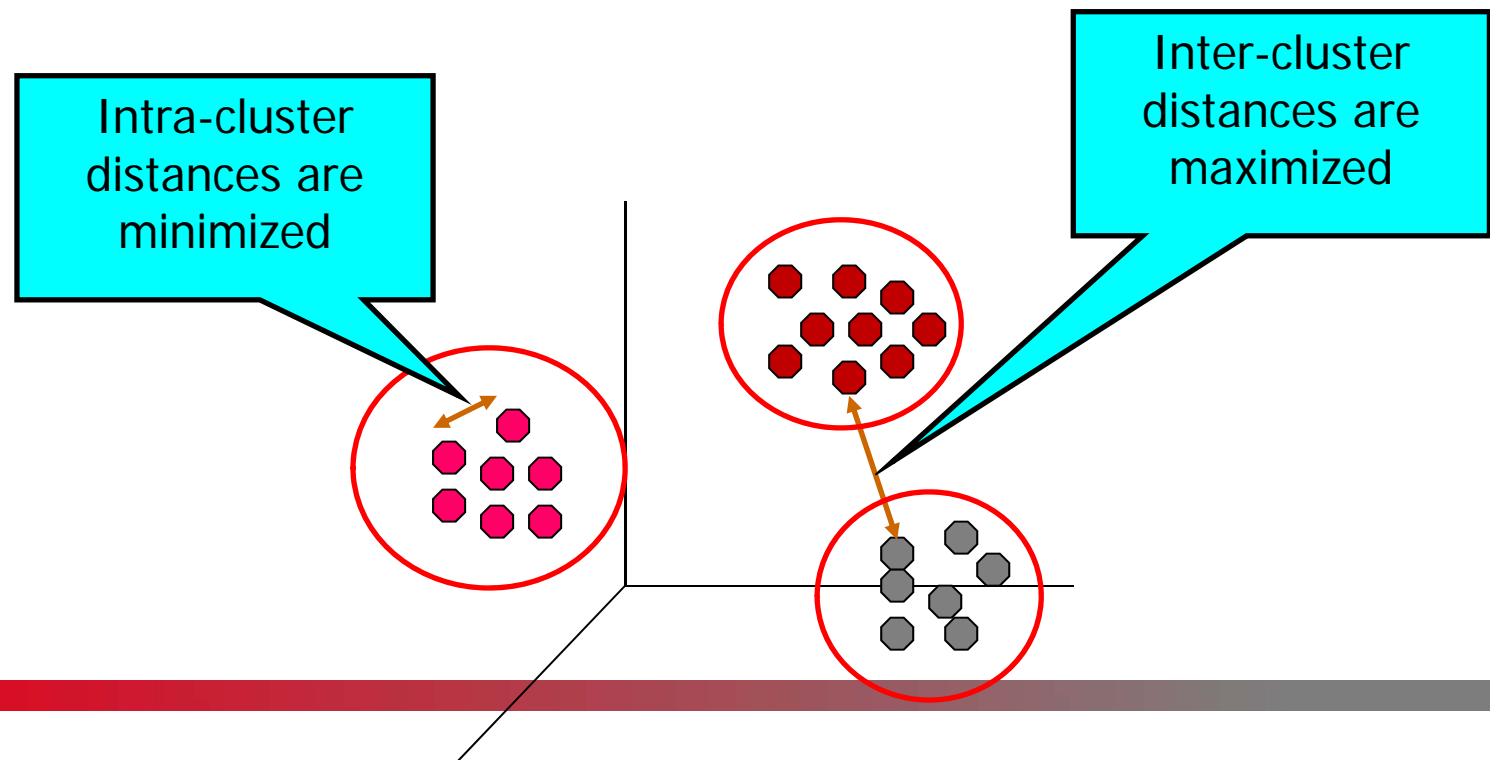
What is Cluster ?

- Cluster is a group of objects/respondents that are similar to each other and distant from other objects in a larger group based upon selected variable/s



Cluster analysis

- It is a class of techniques used to classify objects into groups that are
 - relatively homogeneous within themselves and
 - heterogeneous between each other



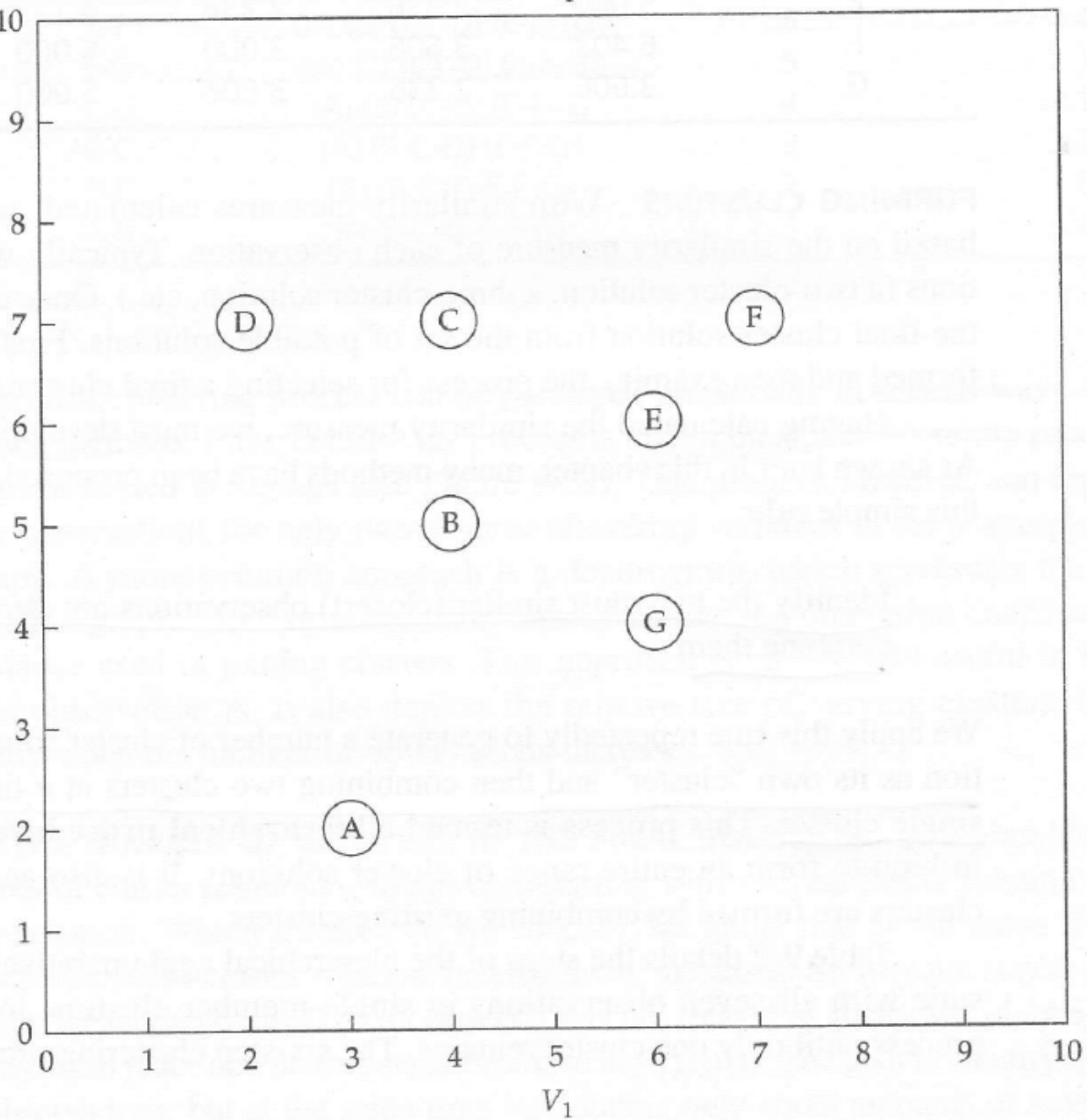
- To perform market segmentation
- To group companies with similar financial health indicators
- To divide employees into various sub-groups based on their productivity and past achievement records
- To perform inventory analysis
- To take visual merchandising decision

Clustering based upon Brand Liking and Purchase Intention

Data Values

Clustering Variable	Respondents						
	A	B	C	D	E	F	G
V_1	3	4	4	2	6	7	6
V_2	2	5	7	7	6	7	4

Scatterplot



- Problem formulation and Variable Selection
- Measuring similarity/distance
- Select a clustering algorithm
- Define the distance between two clusters
- Determine the number of clusters
- Validate the analysis

- Are the customers needs and behavior significantly differ from each other?
- Are there segments among these seven customers?
- If yes, what are the segments? How different are they?

Variable Selection

- To select variables based upon which we cluster objects.
- Inclusion of one or two irrelevant variables may distort an otherwise useful clustering solution.
- Variables should be selected based on past research, theory, or a consideration of the hypotheses being tested.

- Similarity is the degree of correspondence among objects across all of the characteristics used in the analysis
- Inter-subject/object similarity is an empirical measure of correspondence, or resemblance, between objects to be clustered.
- Two of the most widely used method to measure similarity are:
 - Correlational Measures.
 - Distance Measures.

Distance measures for individual observations

- With a single variable, *similarity* is straightforward
 - Income – two individuals are similar if they belong to the same Income group and the level of dissimilarity increases as the income gap increases
- Multiple variables require an aggregate distance measure
 - Many characteristics (e.g. income, age, consumption habits, family composition, owning a car, education level, job...), it becomes more difficult to define similarity with a single value
- The most known measure of distance is the *Euclidean distance*, which is the concept we use in everyday life for spatial coordinates.

- Measuring distance with Euclidean method:

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

Clustering based upon Brand Liking and Purchase Intention

Data Values

Clustering Variable	Respondents						
	A	B	C	D	E	F	G
V_1	3	4	4	2	6	7	6
V_2	2	5	7	7	6	7	4

Distance measurement with Euclidean method

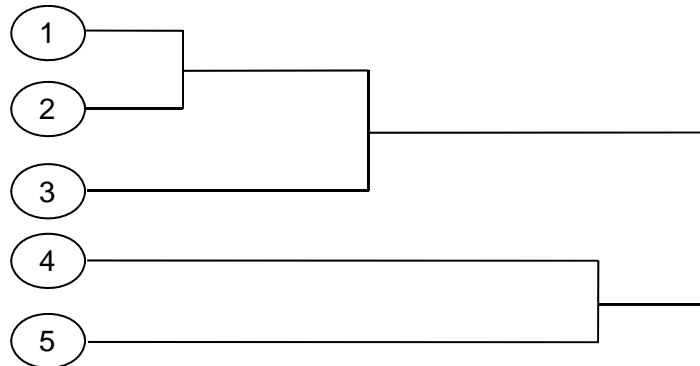
TABLE 9-1 Proximity Matrix of Euclidean Distances Between Observations

Observation	Observation						
	A	B	C	D	E	F	G
A	—						
B	3.162	—					
C	5.099	2.000	—				
D	5.099	2.828	2.000	—			
E	5.000	2.236	2.236	4.123	—		
F	6.403	3.606	3.000	5.000	1.414	—	
G	3.606	2.236	3.606	5.000	2.000	3.162	—

- **Hierarchical procedures**
 - Develop the exhaustive list of all possible number of clusters and decide to choose the appropriate number of clusters.
 - **Agglomerative** (start from n clusters to get to 1 cluster)
 - **Divisive** (start from 1 cluster to get to n clusters)
- **K-mean Clustering**
 - Decide the number of clusters and form the clusters based on similarity

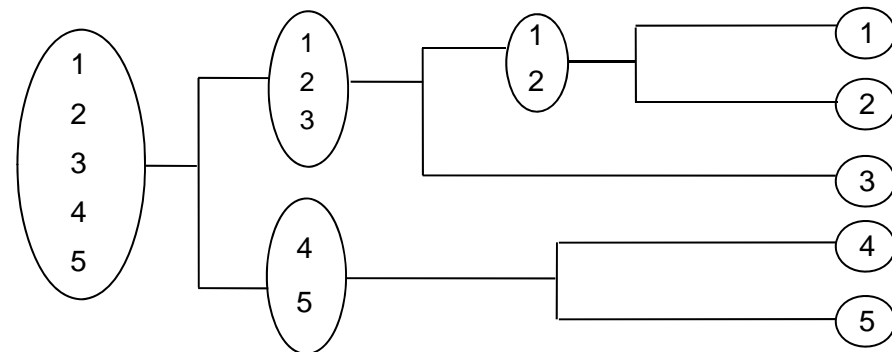
Agglomerative:

- Start from n clusters to get to 1 cluster
- There is a merging in each step until all observations end up in a single cluster in the final step
- Successive change to gross decomposition
- Stops, when defined criteria is reached
- **Short** computational times, **good practical** application



Divisive:

- Start from 1 cluster to get to n clusters
- All observations are initially assumed to belong to a single cluster
- Successive change to fine composition
- Stops, when defined criteria is reached



Hierarchical Clustering

1. Identify the most similar subject/objects and group them
2. Repeat step 1 and prepare the exhausted list of the clusters
3. Select the most distinct clusters

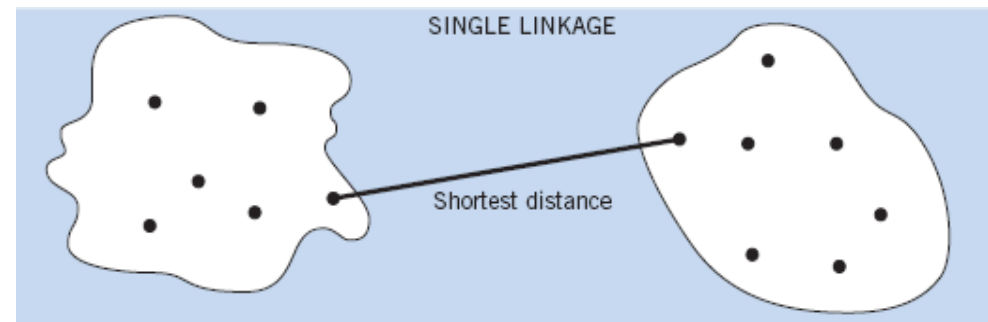
- There are various methods of measuring distance between objects like single linkage, Complete linkage, Average linkage etc.
- Measuring distance with Euclidean method:

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

- Single Linkage (Nearest neighbor)

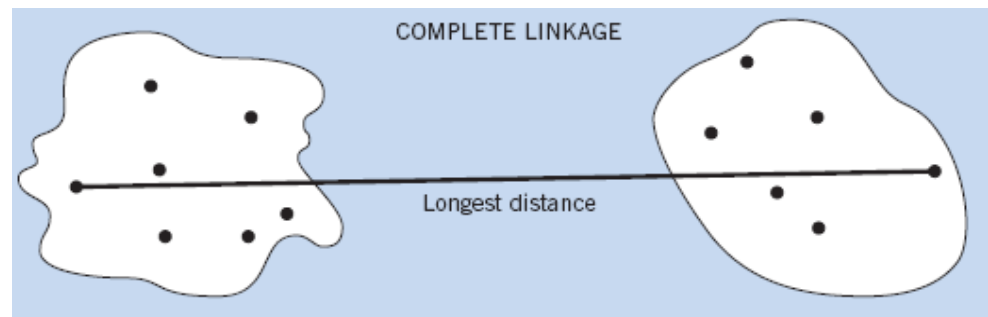
- Clustering criterion based on the shortest distance

$$D_{JM} = \min(D_{JK}, D_{JL})$$



- Complete Linkage (Furtherest neighbor)

- Clustering criterion based on the longest distance



$$D_{JM} = \max(D_{JK}, D_{JL})$$

- Average Linkage
(Between groups)
 - Clustering criterion based on the average distance

$$D_{JM} = \frac{N_K D_{JK} + N_L D_{JL}}{N_M}$$

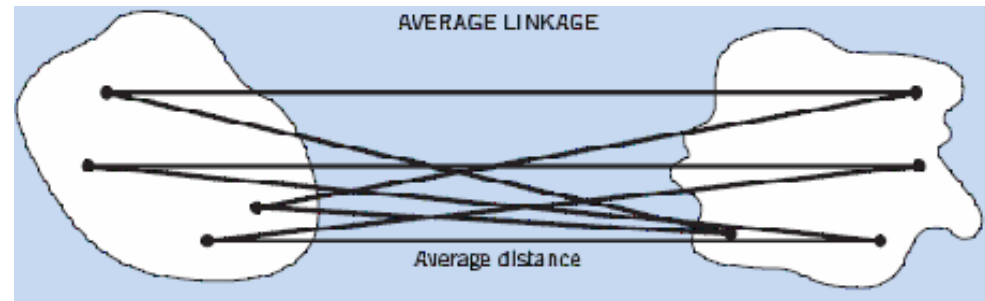


TABLE 9-1 Proximity Matrix of Euclidean Distances Between Observations

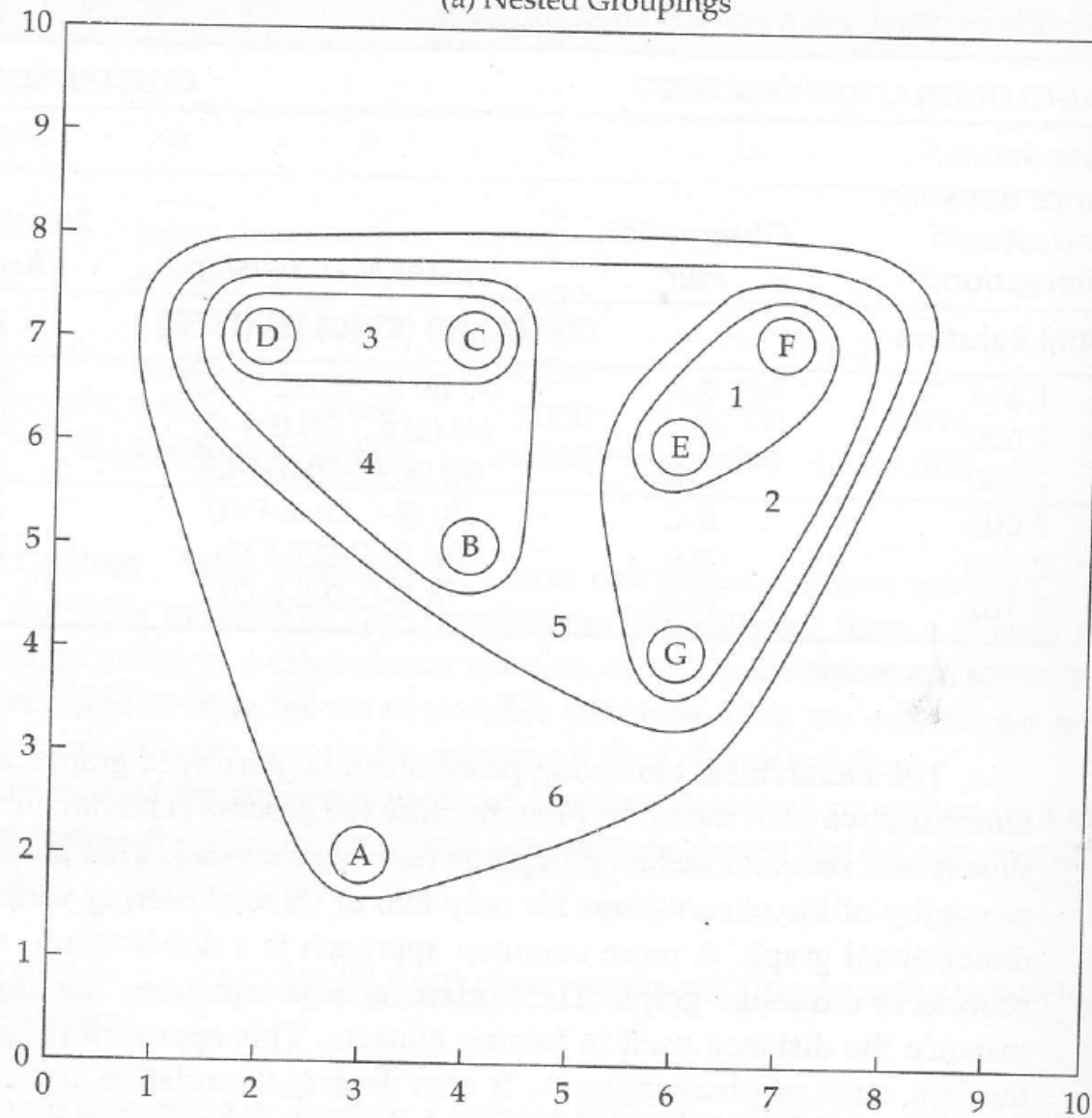
Observation	Observation						
	A	B	C	D	E	F	G
A	—						
B	3.162	—					
C	5.099	2.000	—				
D	5.099	2.828	2.000	—			
E	5.000	2.236	2.236	4.123	—		
F	6.403	3.606	3.000	5.000	1.414	—	
G	3.606	2.236	3.606	5.000	2.000	3.162	—

TABLE 9-2 Agglomerative Hierarchical Clustering Process

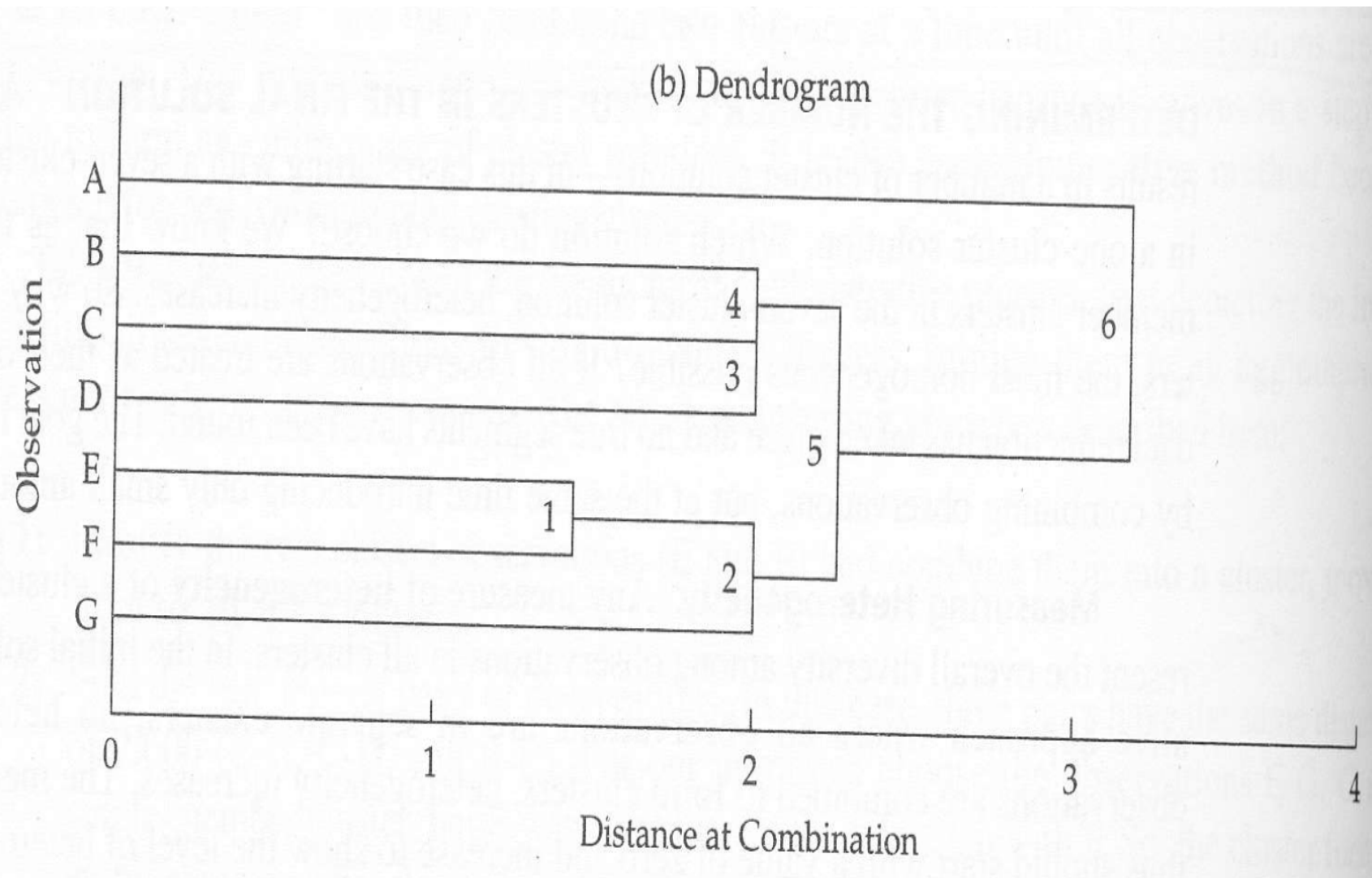
Step	AGGLOMERATION PROCESS		CLUSTER SOLUTION		
	Minimum Distance Between Unclustered Observations ^a	Observation Pair	Cluster Membership	Number of Clusters	Overall Similarity Measure (Average Within-Cluster Distance)
	Initial Solution		(A) (B) (C) (D) (E) (F) (G)	7	0
1	1.414	E-F	(A) (B) (C) (D) (E-F) (G)	6	1.414
2	2.000	E-G	(A) (B) (C) (D) (E-F-G)	5	2.192
3	2.000	C-D	(A) (B) (C-D) (E-F-G)	4	2.144
4	2.000	B-C	(A) (B-C-D) (E-F-G)	3	2.234
5	2.236	B-E	(A) (B-C-D-E-F-G)	2	2.896
6	3.162	A-B	(A-B-C-D-E-F-G)	1	3.420

Euclidean distance between observations.

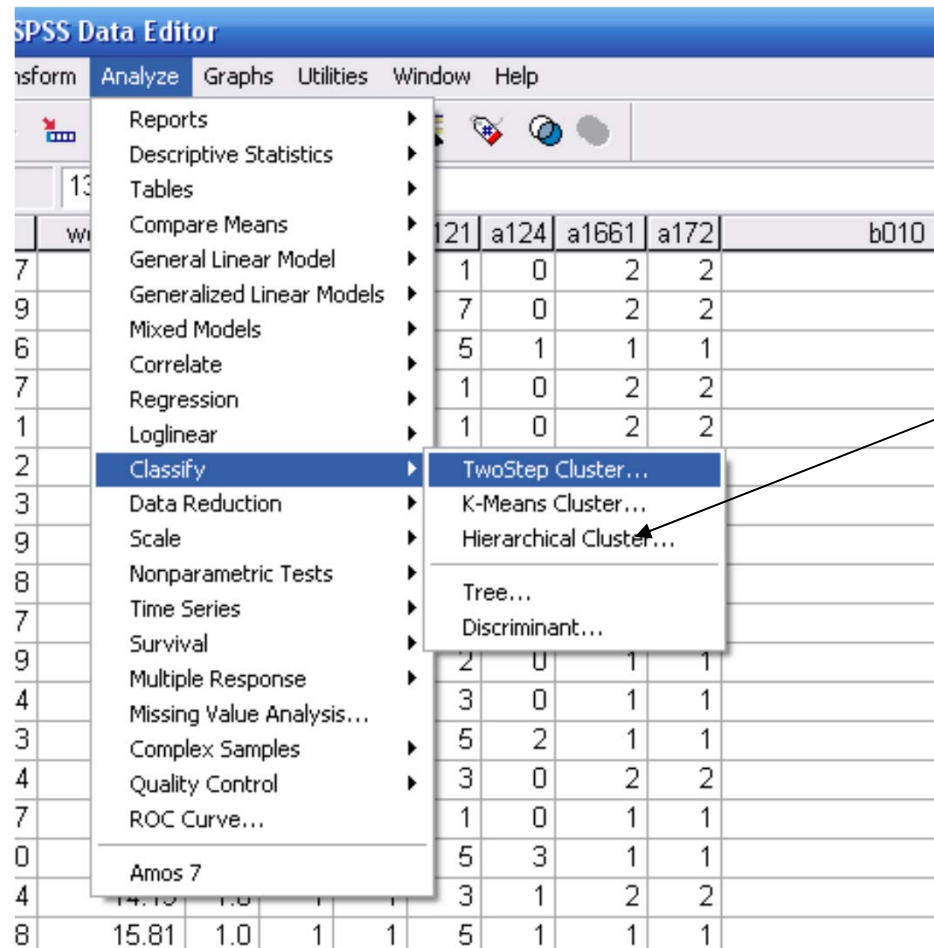
(a) Nested Groupings



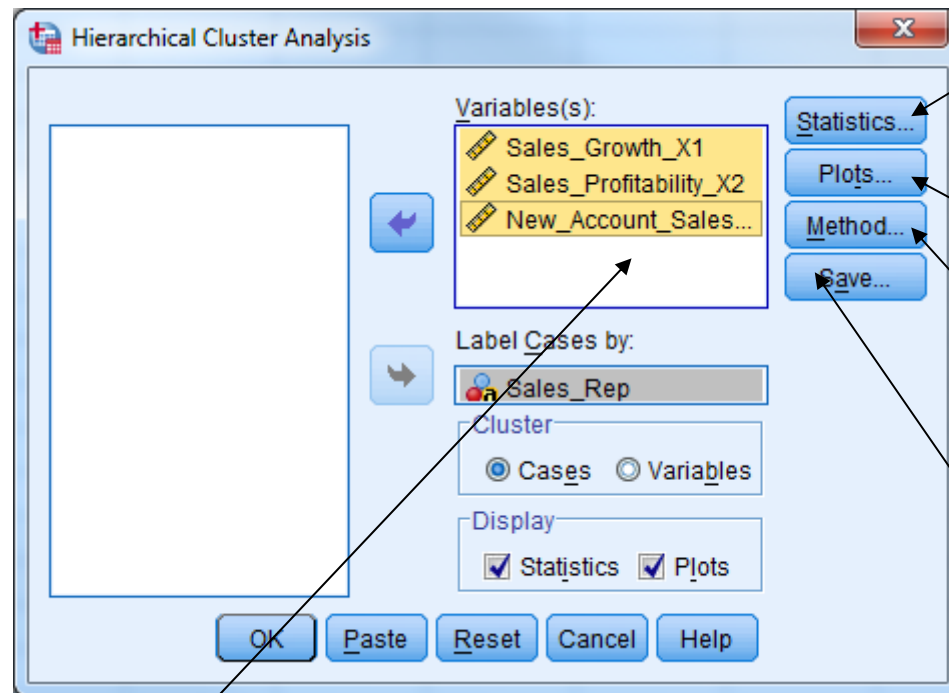
Dendrogram



- At stage 5 there is big jump in distance between the clusters to be clubbed
- We choose 3 clusters



Three types of cluster analysis are available in SPSS



Variables selected
for the analysis

Statistics required
in the analysis

Graphs (dendrogram)

Clustering method
and options

Create a new variable
with cluster
membership for each
case

The agglomeration schedule is a table which shows the steps of the clustering procedure, indicating which cases (clusters) are merged and the merging distance

The proximity matrix contains all distances between cases (it may be huge)

Statistics...

Hierarchical Cluster Analysis: Statistics

☒ Agglomeration schedule

☐ Proximity matrix

Cluster Membership

☒ None

☐ Single solution

Number of clusters:

☐ Range of solutions

Minimum number of clusters:

Maximum number of clusters:

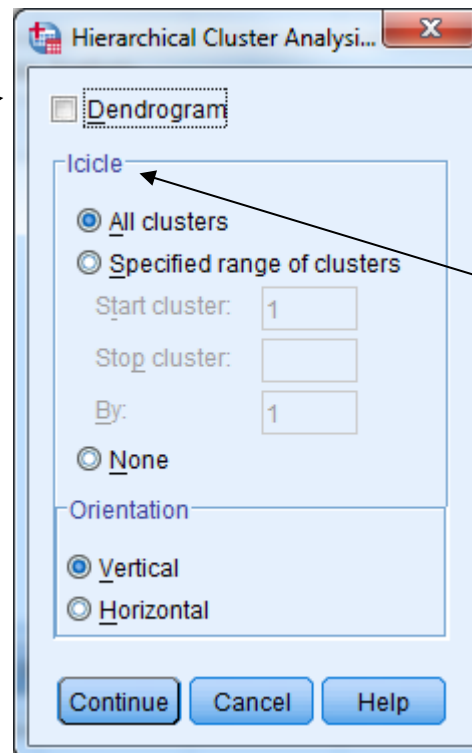
Continue Cancel Help

Shows the cluster membership of individual cases only for a sub-set of solutions

Plots...

Shows the clustering process, indicating which cases are aggregated and the merging distance

With many cases, the dendrogram is hardly readable



The icicle plot (which can be restricted to cover a small range of clusters), shows at what stage cases are clustered. The plot is cumbersome and slows down the analysis (advice: no icicle)

Method...

Choose a hierarchical algorithm

Hierarchical Cluster Analysis: Method

Cluster Method: Ward's method

Measure

☒ Interval: Euclidean distance

Power: 2 Root: 2

☐ Counts: Chi-square measure

☐ Binary: Squared Euclidean distance

Present: 1 Absent: 0

Transform Values

Standardize: None

None
Z scores
Range -1 to 1
Range 0 to 1
Maximum magnitude of 1

Transform Measures

☐ Absolute values
☐ Change sign
☐ Rescale to 0-1 range

Cluster Method: Between-groups linkage

Measure

☒ Interval: Euclidean distance

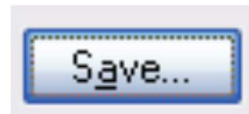
Measure

☒ Interval: Euclidean distance

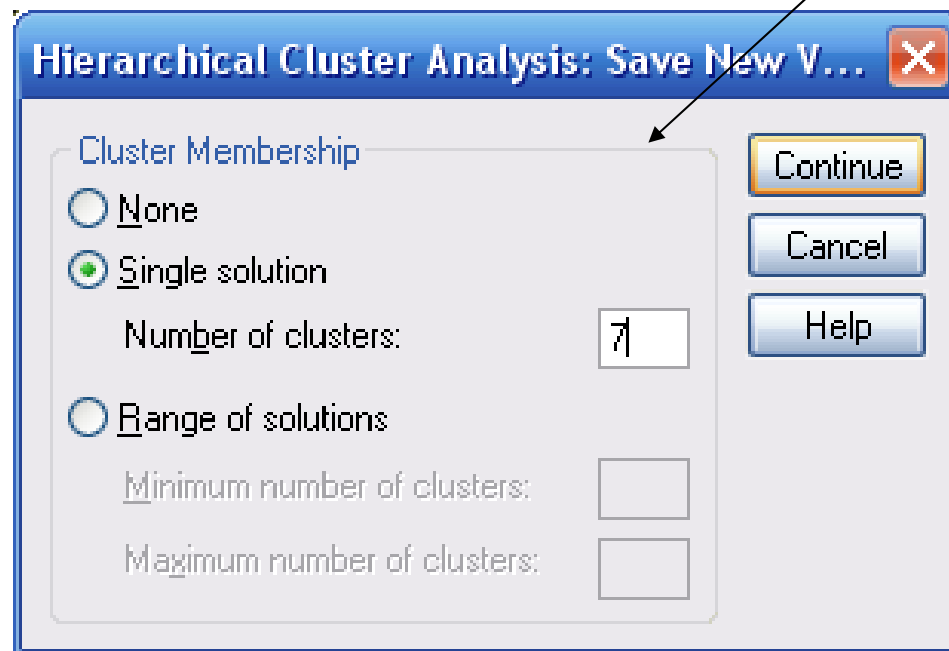
Squared Euclidean distance
Cosine
Pearson correlation
Chebychev
Block
Minkowski
Customized

Choose the type of data (interval, counts binary) and the appropriate measure

Specify whether the variables (values) should be standardized before analysis. Z-scores return variables with zero mean and unity variance. Other standardizations are possible. Distance measures can also be transformed



If the number of clusters has been decided (or at least a range of solutions), it is possible to save the cluster membership for each case into new variables



The example: agglomeration schedule

Last 10 stages of
the process
(10 to 1 clusters)

39	30	31	6,724	16	33	41
40	18	29	7,662	29	31	44
41	4	30	8,873	36	39	48
42	1	6	10,260	0	28	46
43	19	20	11,672	25	30	47
44	18	41	13,193	40	34	47
45	2	3	15,321	32	38	46
46	1	2	19,920	42	45	48
47	18	19	30,538	44	43	49
48	1	4	46,013	46	41	49
49	1	18	147,000	48	47	0

As the
algorithm
proceeds
towards the
end, the
distance
increases

Hierarchical Methods	Non-hierarchical methods
<ul style="list-style-type: none">• No decision about the number of clusters• Problems when data contain a high level of error• Can be very slow, preferable with small data-sets• Initial decisions are more influential (one-step only)• At each step they require computation of the full proximity matrix	<ul style="list-style-type: none">• Faster, more reliable, works with large data sets• Need to specify the number of clusters• Need to set the initial seeds• Only cluster distances to seeds need to be computed in each iteration

Non-hierarchical Clustering

K- Means Cluster

- This is non-hierarchical method of clustering
- Decide the number of clusters in advance
- Number of Clusters are formed based on similarity
- To check if clusters are distinct with reference to each variable, ANOVA is performed

Non-hierarchical clustering: K-means method

1. The number k of clusters is fixed
2. An initial set of k “seeds” (aggregation centres) is provided
 - First k elements
 - Other seeds (randomly selected or explicitly defined)
3. Given a certain fixed threshold, all units are assigned to the nearest cluster seed
4. New seeds are computed
5. Go back to step 3 until no reclassification is necessary

Units can be reassigned in successive steps (optimising partitioning)

- Two alternatives
 - Determined by the analysis
 - Fixed by the researchers
- In segmentation studies, the c represents the number of potential separate segments.
- Preferable approach: “let the data speak”
 - Hierarchical approach and optimal partition identified through statistical tests (stopping rule for the algorithm)
 - However, the detection of the optimal number of clusters is subject to a high degree of uncertainty
- If the research objectives allow a choice rather than estimating the number of clusters, non-hierarchical methods are the way to go.

K-means solution (4 clusters)

K-Means Cluster Analysis

Variables:

Label Cases by:

Number of Clusters: 2

Method

☒ Iterate and classify ☐ Classify only

Cluster Centers

☐ Read initial:

☐ Open dataset

☐ External data file

☐ Write final:

☐ New dataset

☐ Data file

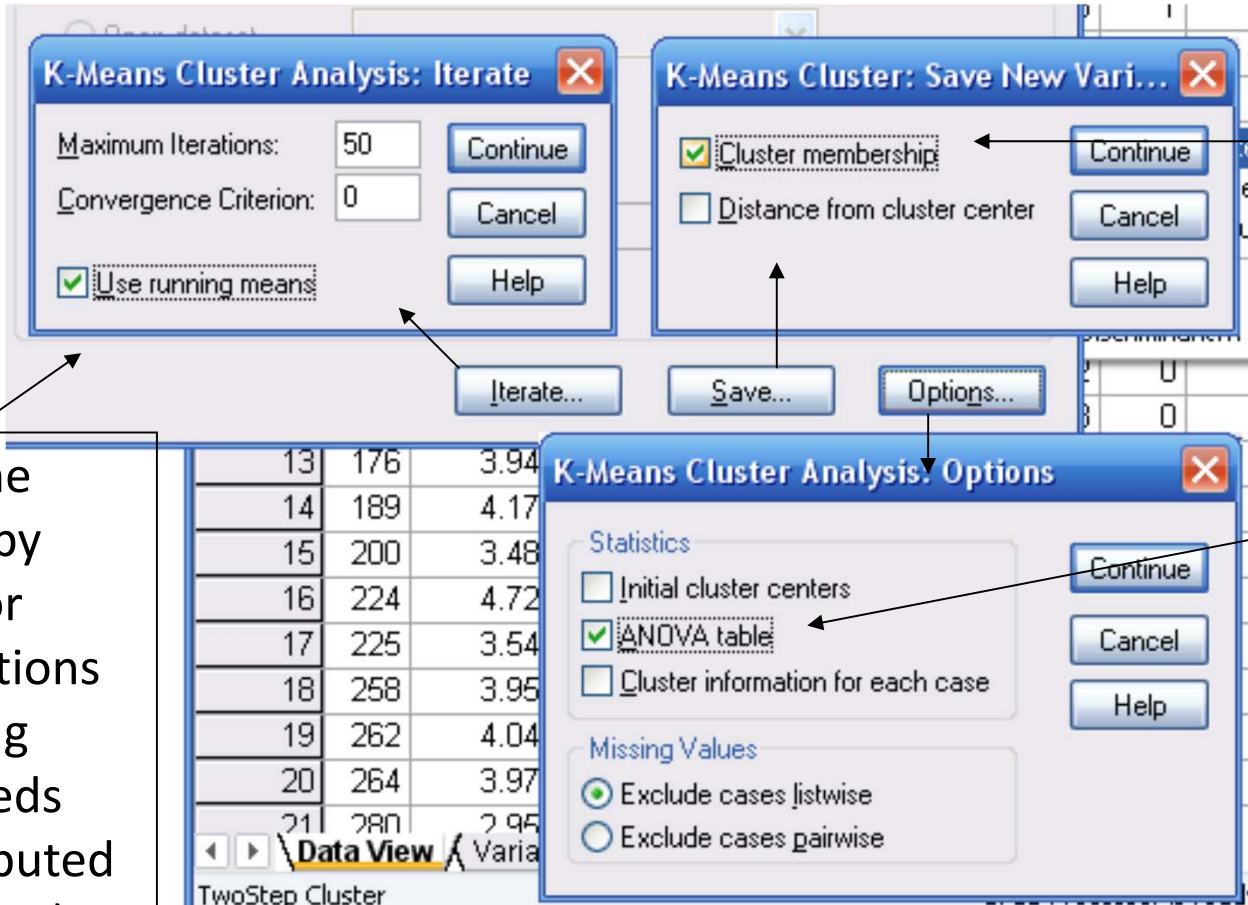
OK Paste Reset Cancel Help

Variables

Number of clusters (fixed)

Ask for one (classify only) or more iterations before stopping the algorithm

It is possible to read a file with initial seeds or write final seeds on a file



The screenshot shows three SPSS dialog boxes for K-Means Cluster Analysis. The 'Iterate' box has 'Maximum Iterations' set to 50, 'Convergence Criterion' set to 0, and 'Use running means' checked. The 'Save New Variables' box has 'Cluster membership' checked. The 'Options' box has 'ANOVA table' checked under 'Statistics' and 'Exclude cases listwise' selected under 'Missing Values'. Arrows point from text boxes to these specific options.

Improve the algorithm by allowing for more iterations and running means (seeds are recomputed at each stage)

Creates a new variable with cluster membership for each case

More options including an ANOVA table with statistics

Case	Cluster	Centroid
13	176	3.94
14	189	4.17
15	200	3.48
16	224	4.72
17	225	3.54
18	258	3.95
19	262	4.04
20	264	3.97
21	280	2.95

Results from k-means (initial seeds chosen by SPSS)

Final Cluster Centers

	Cluster		
	1	2	3
Sales_Growth_X1	105,8	98,7	88,3
Sales_Profitability_X2	118,0	104,7	92,8
New_Account_Sales_X3	107,4	102,3	96,8

Number of Cases in each Cluster

Cluster	1	17,000
	2	22,000
	3	11,000
Valid		50,000
Missing		,000

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Sales_Growth_X1	1025,748	2	12,479	47	82,200	,000
Sales_Profitability_X2	2198,896	2	13,471	47	163,232	,000
New_Account_Sales_X3	379,474	2	7,002	47	54,194	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

- Goodness-of-fit of a cluster analysis
 - ratio between the sum of squared errors and the total sum of squared errors (similar to R^2)
 - root mean standard deviation within clusters
- Validation: if the identified cluster structure (number of clusters and cluster characteristics) is real, it should not be c
- Validation approaches
 - use of different samples to check whether the final output is similar
 - Split the sample into two groups when no other samples are available
 - Check for the impact of initial seeds / order of cases (hierarchical approach) on the final partition
 - Check for the impact of the selected clustering method

- In practice use of both the methods is recommended when researcher has no idea about the existence of clusters in sample
- In the first phase, perform hierarchical clustering and decide possible number of clusters
- Perform K-means clustering for the number of clusters decided in first phase. (could be more than one option)
- K-means clustering is useful for simplifying interpretation and identification of significance of all the variables
- Remove the members who are behaving as outliers or are forming clusters of relatively very small size.
- Final decision on number of clusters using K-means clustering approach is done for which more number of the variables are significant in ANOVA table

- Having identified the clusters of individuals, it is essential to know the characteristics/profile of the clusters
- Clusters can be characterized by considering the demographic variables and or by psychographic variables
 - This can be done by developing cross tab for cluster membership and relevant demographic variable
 - By comparing responses on psychographic variables at the centers.

- Cluster analysis is descriptive, a-theoretical, and non-inferential.
- Will always create clusters, regardless of the actual existence of any structure in the data.
- The cluster solution can not be generalized because it is totally dependent upon the variables used as the basis for the similarity measure.

Regarding 4.2. Cluster Analysis

- Chapter 12 of main course book (required)