

# CLUSTER ANALYSIS

Krishan K Pandey (Ph.D.)



# What is Cluster analysis?

- Cluster analysis is a group of multivariate techniques whose primary purpose is to group objects (e.g., respondents, products, or other entities) based on the characteristics they possess.
- It is a means of grouping records based upon attributes that make them similar. If plotted geometrically, the objects within the clusters will be close together, while the distance between clusters will be farther apart.

- \* *Cluster Variate*

- represents a mathematical representation of the selected set of variables which compares the object's similarities.

# Cluster analysis vs Factor analysis

---

## Cluster Analysis

- grouping is based on the distance (proximity)

## Factor Analysis

- grouping is based on patterns of variation (correlation)

Factor analysis, we form group of variables based on the several people's responses to those variables. In contrast to Cluster analysis, we group people based on their responses to several variables.

# Cluster analysis vs Discriminant analysis

---

## Cluster Analysis

- you don't know who or what belongs to which group. Not even the number of groups.

## **Discriminant analysis**

- requires you to know group membership for the cases used to derive classification rule.

# Application:

- **Field of psychiatry** – where the characterization of patients on the basis of clusters of symptoms can be useful in the identification of an appropriate form of therapy.
- **Biology** – used to find groups of genes that have similar functions.
- **Information Retrieval** – The world Wide Web consists of billions of Web pages, and the results of a query to a search engine can return thousands of pages. Clustering can be used to group these search results into small number of clusters, each of which captures a particular aspect of the query. For instance, a query of “movie” might return Web pages grouped into categories such as reviews, trailers, stars and theaters. Each category (cluster) can be broken into subcategories (sub-clusters), producing a hierarchical structure that further assists a user’s exploration of the query results.
- **Climate** – Understanding the Earth’s climate requires finding patterns in the atmosphere and ocean. To that end, cluster analysis has been applied to find patterns in the atmospheric pressure of polar regions and areas of the ocean that have a significant impact on land climate.

# Common Roles Cluster Analysis can play:

## ‣ *Data Reduction*

---

–A researcher may be faced with a large number of observations that can be meaningless unless classified into manageable groups. CA can perform this data reduction procedure objectively by reducing the info. from an entire population of sample to info. about specific groups.

## ‣ *Hypothesis Generation*

– Cluster analysis is also useful when a researcher wishes to develop hypotheses concerning the nature of the data or to examine previously stated hypotheses.

# Most Common Criticisms of CA

- *Cluster analysis is descriptive, atheoretical, and noninferential.* Cluster analysis has no statistical basis upon which to draw inferences from a sample to a population, and many contend that it is only an exploratory technique. Nothing guarantees unique solutions, because the cluster membership for any number of solutions is dependent upon many elements of the procedure, and many different solutions can be obtained by varying one or more elements.
- *Cluster analysis will always create clusters, regardless of the actual existence of any structure in the data.* When using cluster analysis, the researcher is making an assumption of some structure among the objects. The researcher should always remember that just because clusters can be found does not validate their existence. Only with strong conceptual support and then validation are the clusters potentially meaningful and relevant.
- *The cluster solution is not generalizable because it is totally dependent upon the variables used as the basis for the similarity measure.* This criticism can be made against any statistical technique, but cluster analysis is generally considered more dependent on the measures used to characterize the objects than other multivariate techniques. With the cluster variate completely specified by the researcher. As a result, the researcher must be especially cognizant of the variables used in the analysis, ensuring that they have strong conceptual support.

# Objectives of cluster analysis

- ▶ Cluster analysis used for:
  - ***Taxonomy Description.*** Identifying groups within the data
  - ***Data Simplification.*** The ability to analyze groups of similar observations instead all individual observation.
  - ***Relationship Identification.*** The simplified structure from *CA* portrays relationships not revealed otherwise.
- ▶ Theoretical, conceptual and practical considerations must be observed when selecting clustering variables for *CA*:
  - Only variables that relate specifically to objectives of the *CA* are included.
  - Variables selected characterize the individuals (objects) being clustered.



# How does Cluster Analysis work?

The primary objective of cluster analysis is to define the structure of the data by placing the most similar observations into groups. To accomplish this task, we must address three basic questions:

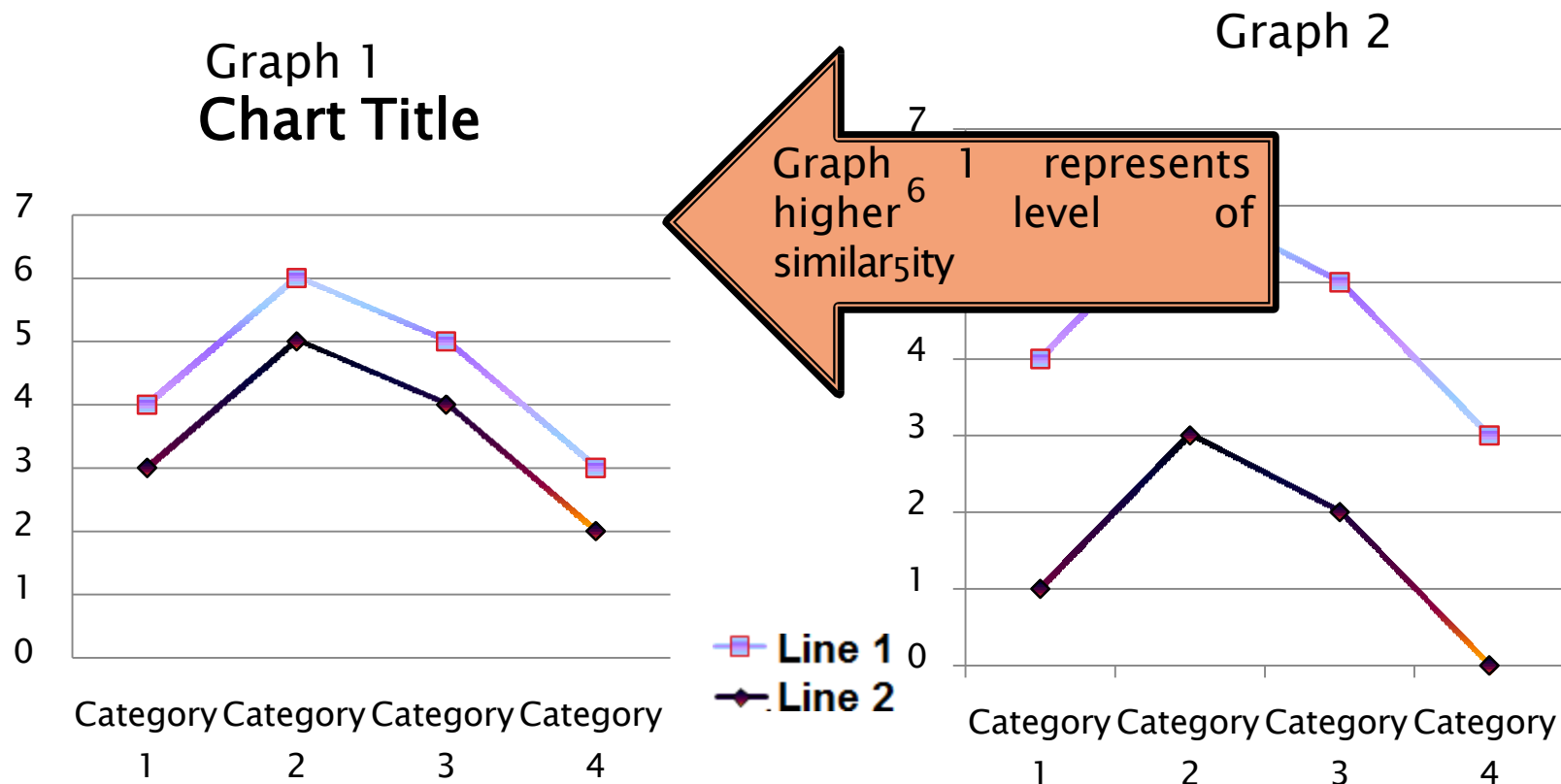
- How do we measure similarity?
- How do we form clusters?
- How many groups do we form?

# Measuring Similarity

- ▶ *Similarity* represents the degree of correspondence among objects across all of the characteristics used in the analysis. It is a set of rules that serve as criteria for grouping or separating items.
- **Correlational measures.**
  - Less frequently used, where large values of  $r$ 's do indicate similarity
- **Distance Measures.**

Most often used as a measure of similarity, with higher values representing greater dissimilarity (distance between cases), not similarity.

# Similarity Measure



- Both graphs have the same  $r = 1$ , which implies they have the same pattern. But the distances ( $d$ 's) are not equal.

# Distance Measures

- several distance measures are available, each with specific characteristics.
- ***Euclidean distance.*** The most commonly recognized to as straight-line distance.

$$d_{Euclidean}(B, C) = \sqrt{(x_B - x_C)^2 + (y_B - y_C)^2}$$

- ***Squared Euclidean distance.*** The sum of the squared differences without taking the square root.
- ***City-block (Manhattan) distance. Euclidean distance.*** Uses the sum of the variables" absolute differences

$$d_{City-block}(B, C) = |x_B - x_C| + |y_B - y_C|$$

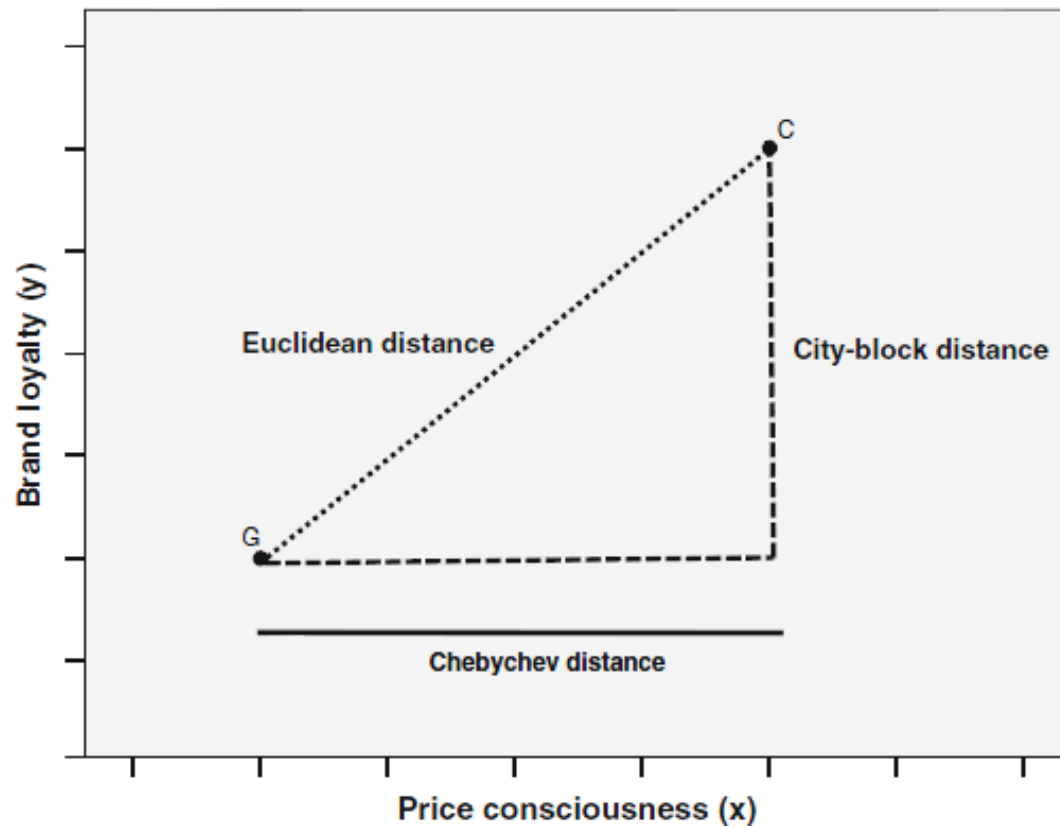
- ***Chebychev distance.***

Is the maximum of the absolute difference in the clustering variables' values. Frequently used when working with metric (or ordinal) data.

$$d_{Chebychev}(B, C) = \max(|x_B - x_C|, |y_B - y_C|)$$

- ***Mahalanobis distance ( $D^2$ ).*** Is a generalized distance measure that accounts for the correlations among variables in a way that weights each variables equally.
-

# Illustration:



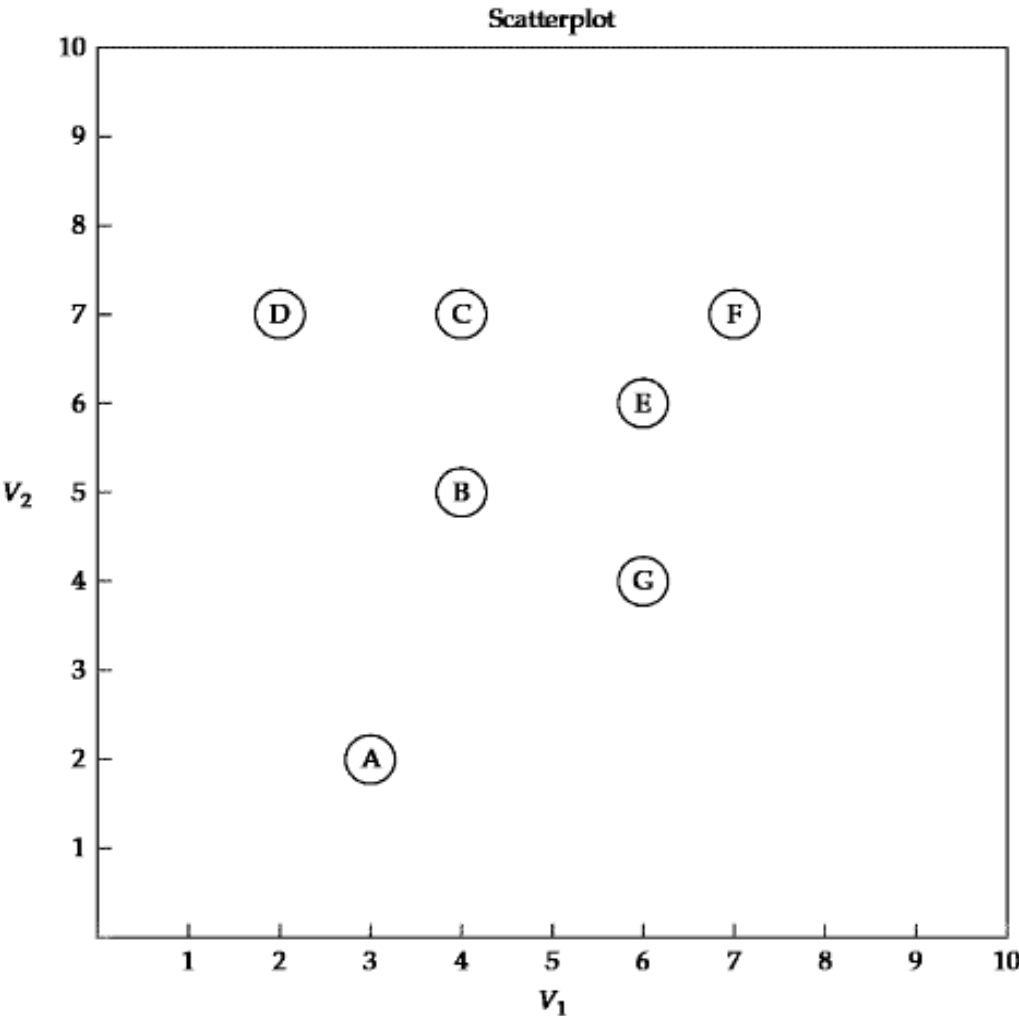
# Simple Example

- ▶ Suppose a marketing researcher wishes to determine market segments in a community based on patterns of loyalty to brands and stores a small sample of seven respondents is selected as a pilot test of how cluster analysis is applied. Two measures of loyalty-  $V_1$ (store loyalty) and  $V_2$ (brand loyalty)- were measured for each respondents on 0-10 scale.

**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

# Simple Example





# How do we measure similarity?

Proximity Matrix of Euclidean Distance Between Observations

Observation	Observations						
	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	3.16	---
						2	

$$d_{Euclidean}(A, B) = \sqrt{(V_{1(A)} - V_{1(B)})^2 + (V_{2(A)} - V_{2(B)})^2}$$

$$d_{Euclidean}(A, B) = \sqrt{(3 - 4)^2 + (2 - 5)^2} = 3.162$$

# How do we form clusters?

## ► SIMPLE RULE:

---

- Identify the two most similar(closest) observations not already in the same cluster and combine them.
- We apply this rule repeatedly to generate a number of cluster solutions, starting with each observation as its own “cluster” and then combining two clusters at a time until all observations are in a single cluster. This process is termed a **hierarchical procedure** because it moves in a stepwise fashion to form an entire range of cluster solutions. It is also an **agglomerative method** because clusters are formed by combining existing clusters

# How do we form clusters?

AGGLOMERATIVE PROCESS			CLUSTER SOLUTION		
Step	Minimum	Observation Pair	Cluster Membership	Number of Clusters	Overall Similarity
	Unclustered Observations <sup>a</sup> Distance				Within-Cluster Measure (Average Distance)
	Initial Solution		(A)(B)(C)(D)(E)(F)(G)	7	Distance
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

In steps 1,2,3 and 4, the OSM does not change substantially, which indicates that we are forming other clusters with essentially the same heterogeneity of the existing clusters.

When we get to step 5, we see a large increase. This indicates that joining clusters (B-C-D) and (E-F-G) resulted a single cluster that was markedly less homogenous.

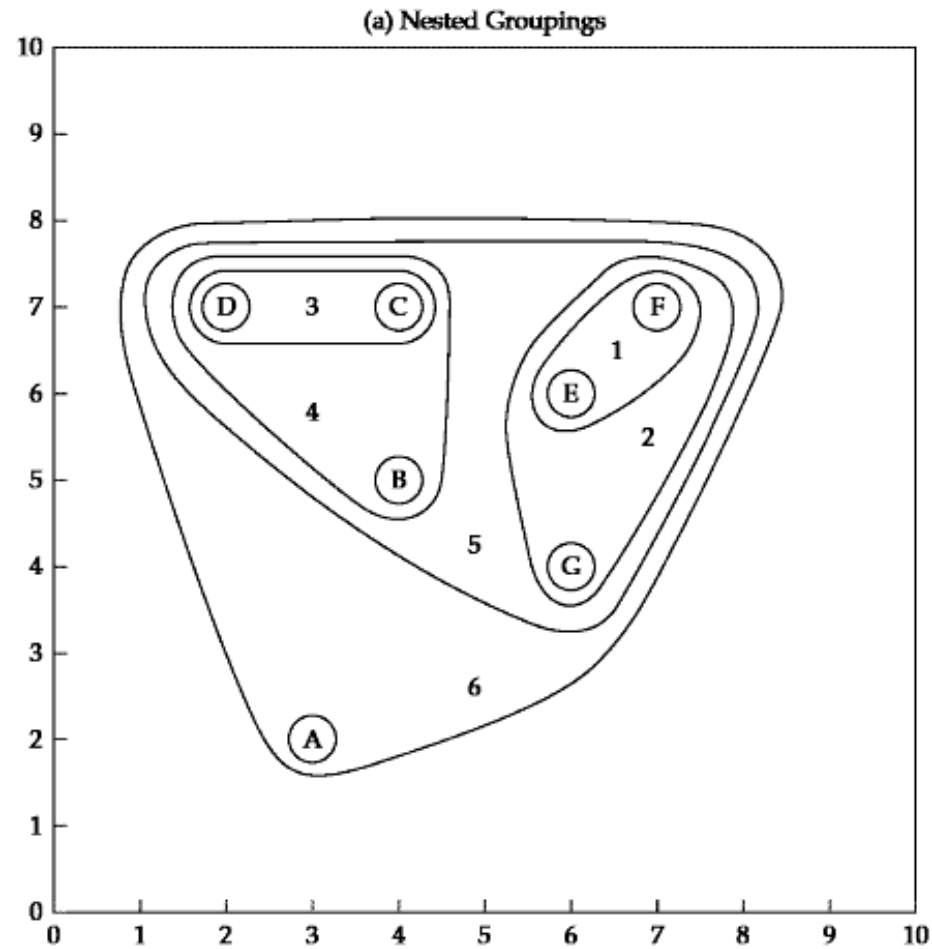
# How many groups do we form?

---

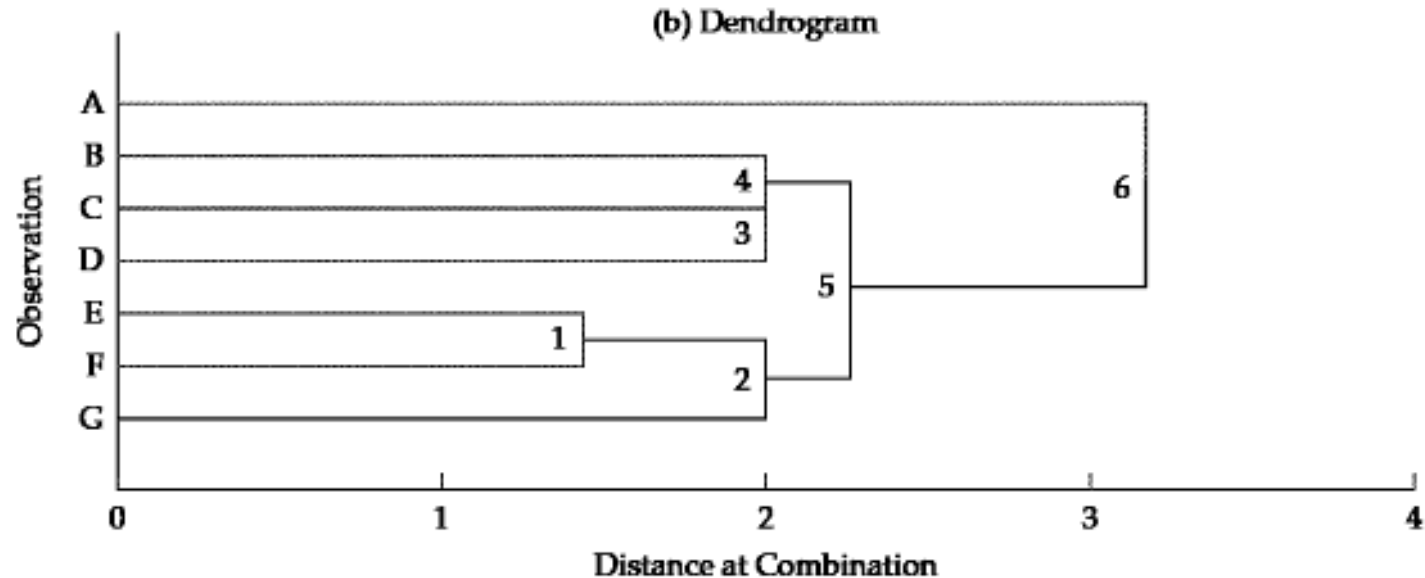
- ▶ Therefore, the three – cluster solution of Step 4 seems the most appropriate for a final cluster solution, with two equally sized clusters, (B–C–D) and (E–F–G), and a single outlying observation (A).

*This approach is particularly useful in identifying outliers, such as Observation A. It also depicts the relative size of varying clusters, although it becomes unwieldy when the number of observations increases.*

# Graphical Portrayals



# Graphical Portrayals



## ► Dendrogram

– Graphical representation (tree graph) of the results of a hierarchical procedure. Starting with each object as a separate cluster, the dendrogram shows graphically how the clusters are combined at each step of the procedure until all are contained in a single cluster

# Outliers: Removed or Retained?

- ▶ Outliers can severely distort the representativeness of the results if they appear as structure (clusters) inconsistent with the objectives.
- They should be removed if the outliers represents:
  - Abberant observations not representative of the population
  - Observations of small or insignificant segments within the population and of no interest to the research objectives
- They should be retained if a undersampling/poor representation of relevant groups in the population; the sample should be augmented to ensure representation of these group.

# Detecting Outliers

---

- ▶ Outliers can be identified based on the similarity measure by:
  - Finding observations with large distances from all other observations.
  - Graphic profile diagrams highlighting outlying cases.
  - Their appearance in cluster solutions as single – member or small clusters.



# Sample Size

- ▶ The researcher should ensure that the sample size is large enough to provide sufficient representation of all relevant groups of the population
- ▶ The researcher must therefore be confident that the obtained sample is representative of the population.

# Standardizing the Data

- ▶ Clustering variables that have scales using widely differing numbers of scale points or that exhibit large differences in standard deviations should be standardized.
- The most common standardization conversion is Z score (with mean equals to 0 and standard deviation of 1).

# Deriving Clusters

- There are number of different methods that can be used to carry out a cluster analysis; these methods can be classified as follows:

- ❖ **Hierarchical Cluster Analysis**

- ❖ **Nonhierarchical Cluster Analysis**

- ❖ **Combination of Both Methods**

# Hierarchical Cluster Analysis

# Hierarchical Cluster Analysis

- ▶ The stepwise procedure attempts to identify relatively homogeneous groups of cases based on selected characteristics using an algorithm either agglomerative or divisive, resulting to a construction of a hierarchy or treelike structure (dendogram) depicting the formation of clusters. This is one of the most straightforward method.
- ▶ HCA are preferred when:
  - The sample size is moderate (under 300 – 400, not exceeding 1000).

# Two Basic Types of HCA

- **Agglomerative Algorithm**
- **Divisive Algorithm**

*\*Algorithm*– defines how similarity is defined between multiple – member clusters in the clustering process.

# Agglomerative Algorithm

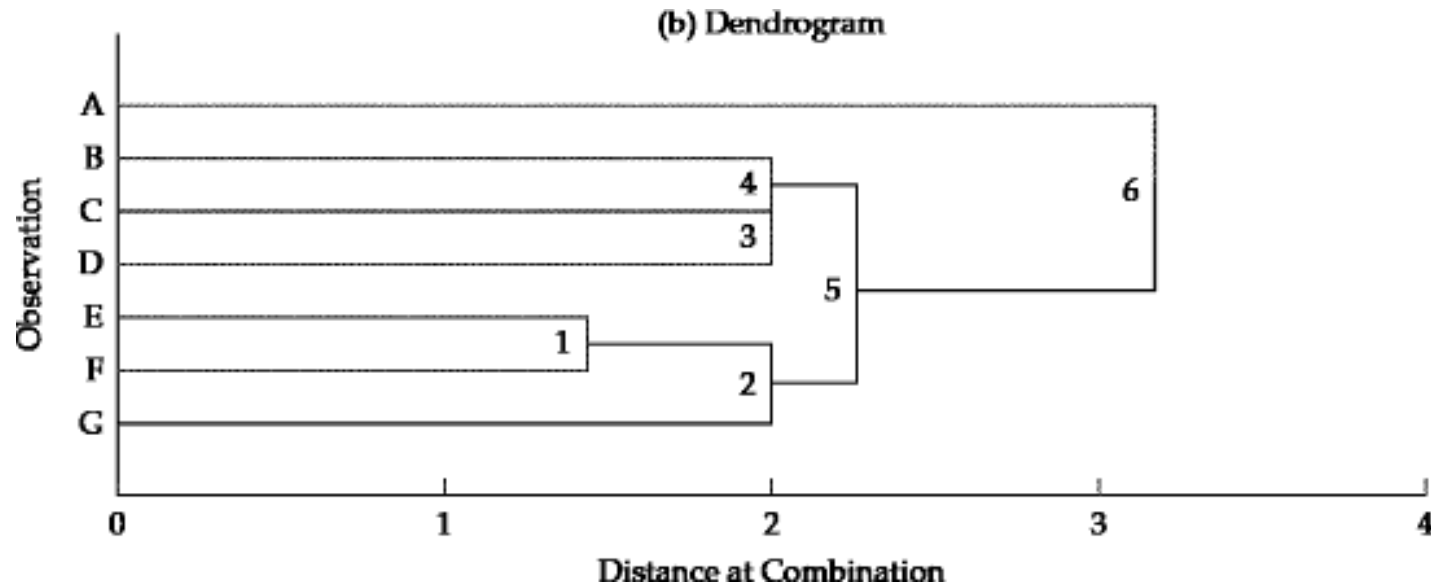
- ▶ Hierarchical procedure that begins with each *object* or observation in a separate cluster. In each subsequent step, the two clusters that are most similar are combined to build a new aggregate cluster. The process is repeated until all objects are finally combined into a single cluster. From  $n$  clusters to 1.
- ▶ Similarity decreases during successive steps. Clusters can't be split.

# Divisive Algorithm

- ▶ Begins with all *objects* in single cluster, which is then divided at each step into two additional clusters that contain the most dissimilar objects. The single cluster is divided into two clusters, then one of these clusters is split for a total of three clusters. This continues until all observations are in a single – member clusters. From 1 cluster to  $n$  sub clusters



# Dendrogram/ Tree Graph



Divisive Method

Aglomerative  
Method

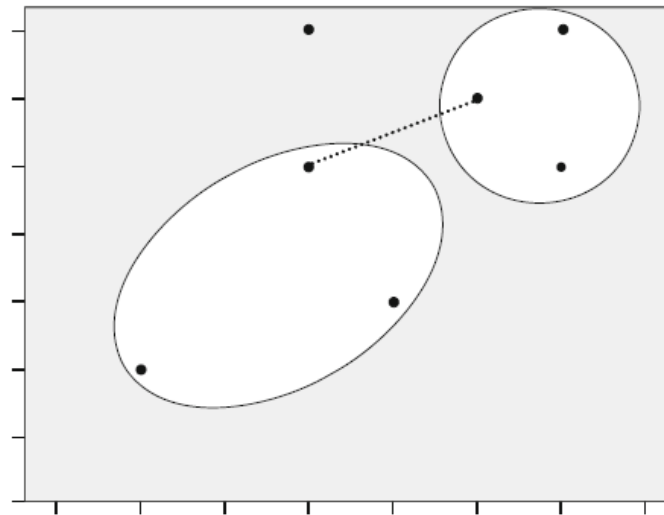
# Agglomerative Algorithms

- ▶ Among numerous approaches, the five most popular agglomerative algorithms are:
  - Single – Linkage
  - Complete – Linkage
  - Average – Linkage
  - Centroid Method
  - Ward's Method
  - Mahalanobis Distance

# Agglomerative Algorithms

## ▶ Single - Linkage

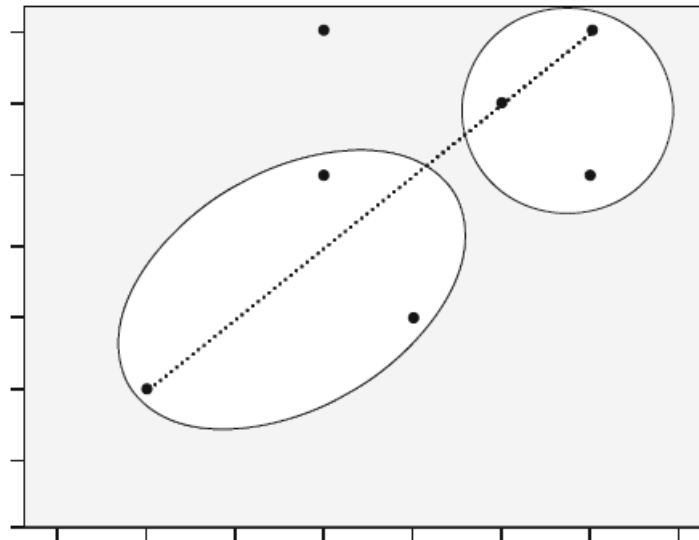
- Also called the *nearest - neighbor method*, defines similarity between clusters as the shortest distance from any object in one cluster to any object in the other.



# Agglomerative Algorithms

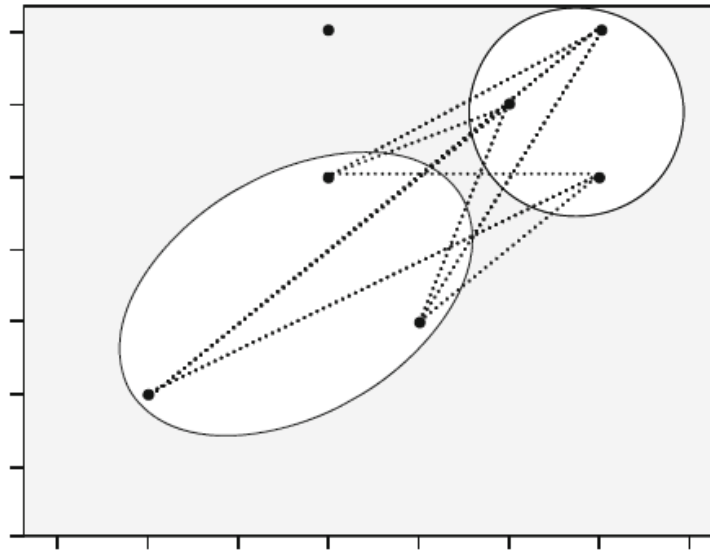
## ‣ *Complete Linkage*

- Also known as the *farthest – neighbor method*.
- The oppositional approach to single linkage assumes that the distance between two clusters is based on the maximum distance between any two members in the two clusters.



# Agglomerative Algorithms

- ▶ *Average Linkage*
  - ▶ The distance between two clusters is defined as the average distance between all pairs of the two clusters' members
- 

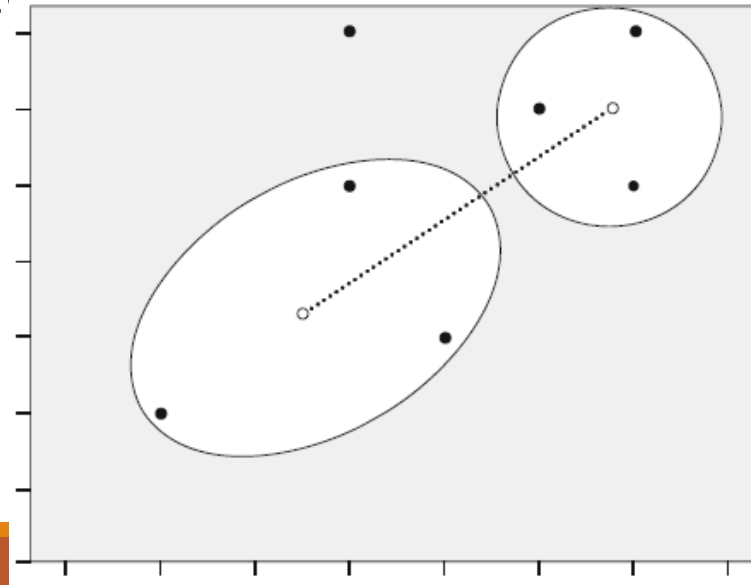


# Agglomerative Algorithms

## ► *Centroid Method*

---

- Cluster Centroids
  - are the mean values of the observation on the variables of the cluster.
- The distance between the two clusters equals the distance between



# Agglomerative Algorithms

## ▶ *Ward's Method*

---

- The similarity between two clusters is the sum of squares within the clusters summed over all variables.
- Ward's method tends to join clusters with a small number of observations, and it is strongly biased toward producing clusters with the same shape and with roughly the same number of observations.

# Hierarchical Cluster Analysis

- ▶ The Hierarchical Cluster Analysis provides an excellent framework with which to compare any set of cluster solutions.
- ▶ This method helps in judging how many clusters should be retained or considered.



# Non Hierarchical Cluster Analysis

# Non Hierarchical Cluster Analysis

- ▶ In contrast to Hierarchical Method, the NCA do not involve the treelike construction process. Instead, they assign objects into clusters once the number of clusters is specified.
  - Two steps in Non HCA
    - 1.) *Specify Cluster Seed* – identify starting points
    - 2.) *Assignment* – assign each observation to one of the cluster seeds.

# Non Hierarchical Clustering Algorithm

- Sequential Threshold Method
  - Parallel Threshold Method
  - Optimizing Procedures
- ▶ All of this belongs to a group of clustering algorithm known as *K – means*.
- K – means Method
    - This method aims to partition n observation into k clusters in which each observation belongs to the cluster with the nearest mean.
    - K – means is so commonly used that the term is used by some to refer to Nonhierarchical cluster analysis in general.

# Advantages of HCA

- ▶ *Simplicity.* With the development of *Dendogram*, the HCA so afford the researcher with a simple, yet comprehensive portrayal of clustering solutions.
- ▶ *Measures of Similarity.* HCA can be applied to almost any type of research question.
- ▶ *Speed.* HCA have the advantage of generating an entire set of clustering solutions in an expedient manner.

# Disadvantages of HCA

- ▶ To reduce the impact of outliers, the researcher may wish to cluster analyze the data several times, each time deleting problem observations or outliers.
- ▶ Hierarchical Cluster Analysis is not amenable to analyze large samples.

# Advantages of NonHCA

- ▶ The results are less susceptible to outliers in the data, the distance measure used, and the inclusion of irrelevant or inappropriate variables.
- ▶ Non Hierarchical Cluster Analysis can analyze extremely large data sets.

# Disadvantages of NonHCA

- ▶ Even a nonrandom starting solution does not guarantee an optimal clustering of observations. In fact, in many instances, the researcher will get a different final solution for each set of specified seed points. How is the researcher to select the optimum answer? Only by analysis and validation can the researcher select what is considered the best representation of structure, realizing that many alternatives may be acceptable.
- ▶ Nonhierarchical methods are also not so efficient when a large number of potential cluster solutions. Each cluster solution is a separate analysis, in contrast to the hierarchical techniques that generate all possible cluster solutions in a single analysis. Thus, nonhierarchical techniques are not as well suited to exploring a wide range of solutions based on varying elements such as similarity measures, observations included, and potential seed points.

# Combination of Both Method

A combination approach using a hierarchical approach followed by a nonhierarchical approach is often advisable.

- First, a hierarchical technique is used to select the number of clusters and profile clusters centers that serve as initial cluster seeds in the nonhierarchical procedure.
- A nonhierarchical method then clusters all observations using the seed points to provide more accurate cluster memberships.

In this way, the advantages of hierarchical methods are complemented by the ability of the nonhierarchical methods to refine the results by allowing the switching of cluster membership.



# Interpretation of Clusters

- ▶ The cluster centroid, a mean profile of the cluster on each clustering variable, is particularly useful in the interpretation stage:
  - Interpretation involves examining the distinguishing characteristics of each cluster's profile and identifying substantial differences between clusters.
  - Cluster solutions failing to show substantial variation indicate other cluster solutions should be examined.
  - The cluster centroid should also be assessed for correspondence with the researcher's prior expectations based on theory or practical experience.

# Validation of Clusters

- ▶ Validation is essential in cluster analysis because the clusters are descriptive of structure and require additional support for their relevance:
  - Cross – validation empirically validates a cluster solution by creating two subsamples (randomly splitting the sample) and then comparing the two cluster solutions for consistency with respect to the number of clusters and the clusters profiles.
  - Validation is also achieved by examining differences on variables not included in the cluster analysis but for which a theoretical and relevant reason enables the expectation of variation across the clusters.

---

# Cluster Analysis on SPSS

# Types of Clustering Procedure

- ▶ Hierarchical Cluster Analysis
- ▶ Non Hierarchical Cluster Analysis
- ▶ Two – Step Cluster Analysis

# Hierarchical Cluster Analysis

# Hierarchical Cluster Analysis

Example:

This file only includes 20 cases, each responding to items on demographics (gender, qualifications, days absence from work, whether they smoke or not), on their attitudes to smoking in public places (subtest totals for pro and anti), plus total scale score for self-concept. We are attempting to determine how many natural groups exist and who belongs to each group.

\*SPSS Chap 23 Data File A.sav [DataSet1] - PASW Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Reports  
Descriptive Statistics  
Tables  
Compare Means  
General Linear Model  
Generalized Linear Models  
Mixed Models  
Correlate  
Regression  
Loglinear  
Neural Networks  
**Classify**  
Dimension Reduction  
Scale  
Nonparametric Tests  
Forecasting  
Survival  
Multiple Response  
Missing Value Analysis...  
Multiple Imputation  
Complex Samples  
Quality Control  
ROC Curve...

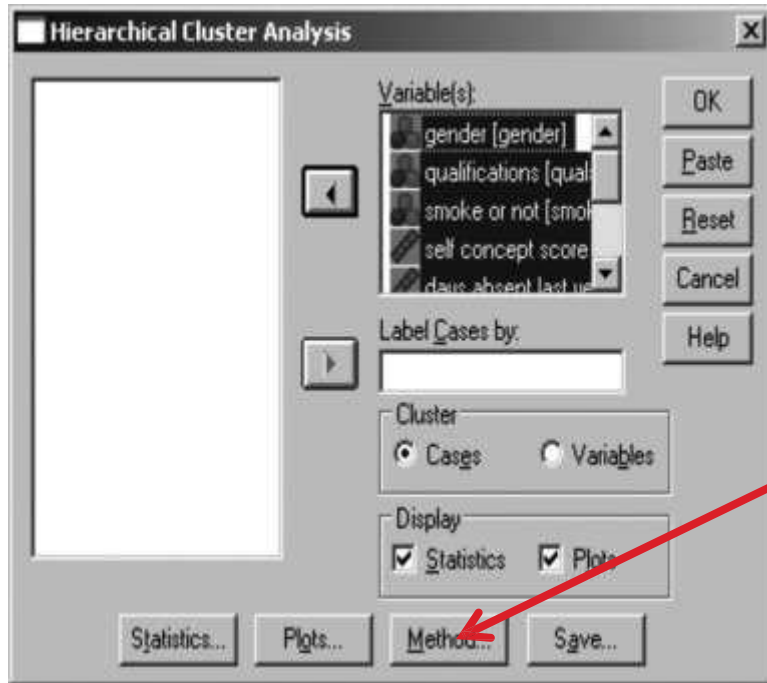
absentb subtestb var var var

1 3.00 30.00  
2 .0 29.00  
3 21.00 23.00  
4 14.00 14.00  
5 9.00 16.00  
6 5.00 20.00  
7 8.00 22.00  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27

TwoStep Cluster...  
K-Means Cluster...  
**Hierarchical Cluster...**  
Tree...  
Discriminant...  
Nearest Neighbor...

	gender	qual	absentb	subtestb	var	var	var
1	1						
2	1						
3	1						
4	1						
5	1						
6	1						
7	1						
8	2						
9	1						
10	2						
11	1						
12	1						
13	1						
14	1						
15	2						
16	1						
17	2						
18	1						
19	1	1.00	1.00	34.00	.0	20.00	
20	2	3.00	2.00	45.00	21.00	13.00	
21	.	.	.	.	.	.	
22	.	.	.	.	.	.	
23	.	.	.	.	.	.	
24	.	.	.	.	.	.	
25	.	.	.	.	.	.	
26	.	.	.	.	.	.	
27	.	.	.	.	.	.	

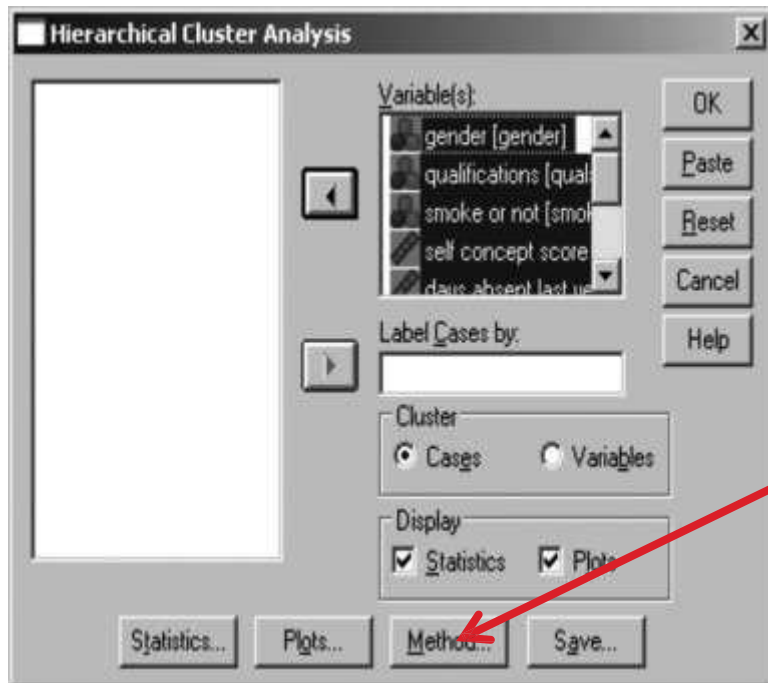
The initial step is determining how many groups exist. The SPSS hierarchical analysis actually calculates every possibility between everyone forming their own group (as many clusters as there are cases) and everyone belonging to the same group, giving a range in our data of from 1 to 20 clusters.



Click „Plots“

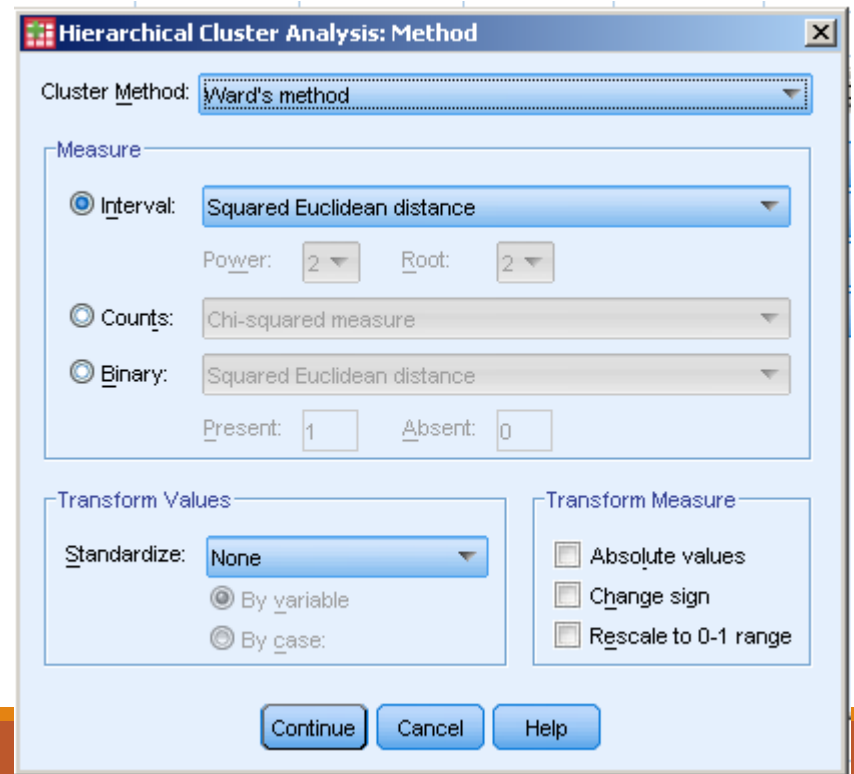




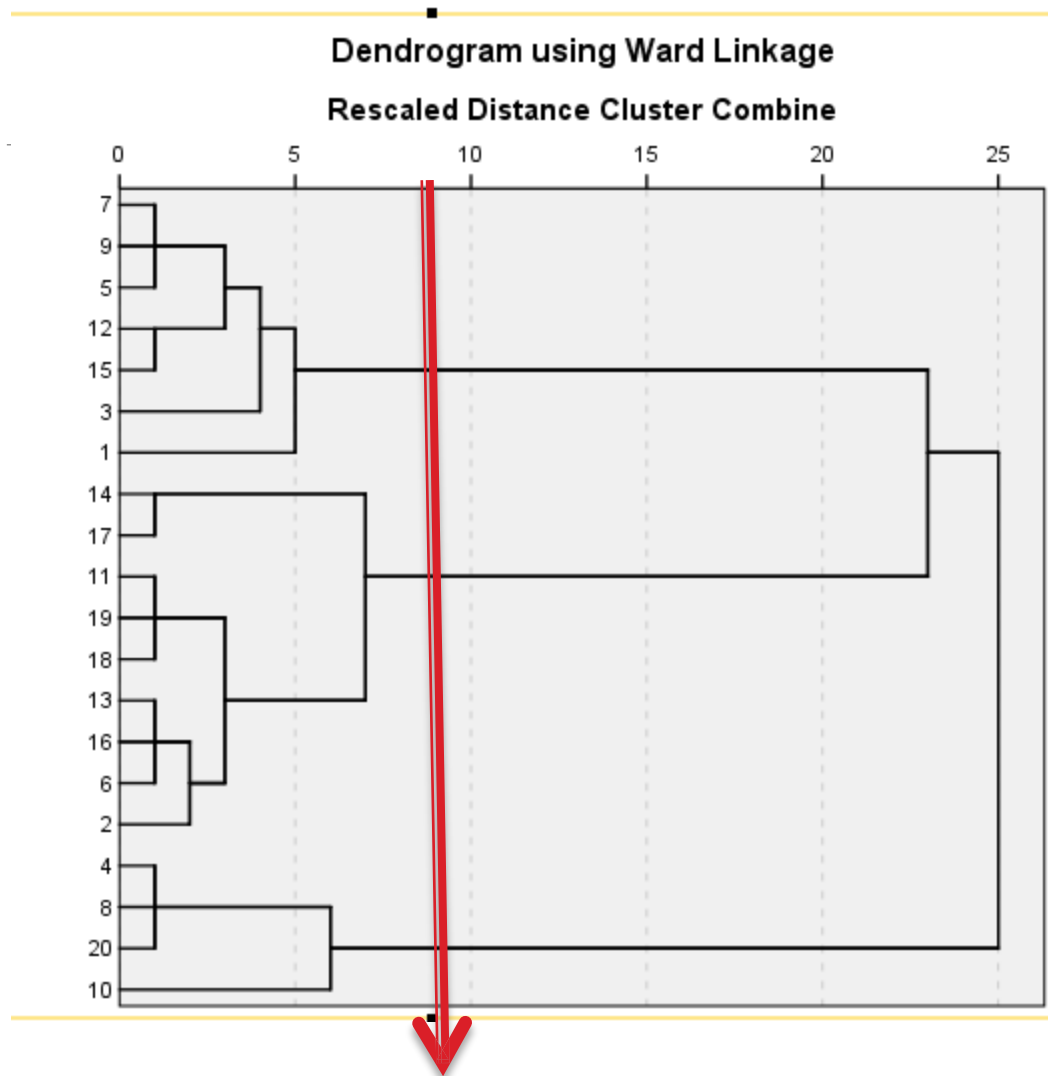


Click „Method“

Select *Continue* then *OK*.



# DENDROGRAM



Parsing the classification tree to determine the number of clusters is a subjective process. Generally, you begin by looking for "gaps" between joining along the horizontal axis. Starting from the right, there is a gap between 5 and 10, which splits the data into two clear clusters and a minor one.

# AGGLOMERATION SCHEDULE

Table 23.1 Agglomeration schedule

Stage	Cluster combined		Coefficients	Stage cluster first appears		Next stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	7	9	4.000	0	0	8
2	11	19	13.500	0	0	5
3	4	8	23.000	0	0	10
4	13	16	34.000	0	0	7
5	11	18	46.500	2	0	12
6	14	17	61.000	0	0	17
7	6	13	79.333	0	4	11
8	5	7	107.333	0	1	13
9	12	15	137.333	0	0	13
10	4	20	183.167	3	0	16
11	2	6	231.583	0	7	12
12	2	11	324.976	11	5	17
13	5	12	442.976	8	9	14
14	3	5	586.810	0	13	15
15	1	3	806.405	0	14	18
16	4	10	1055.071	10	0	19
17	2	14	1361.651	12	6	18
18	1	2	2362.438	15	17	19
19	1	4	3453.150	18	16	0

The agglomeration schedule is a numerical summary of the cluster solution.

At the first stage, cases 7 and 9 are combined because they have the smallest distance

The cluster created by their joining next appears in stage 8

In stage 8, the observations 5 and 7 were joined. The resulting cluster next appears in stage 13.

When there are many cases, this table becomes rather long, but it may be easier to scan the coefficients column for large gaps rather than scan the dendrogram.

# AGLOMERATION SCHEDULE

Table 23.2 Re-formed agglomeration table

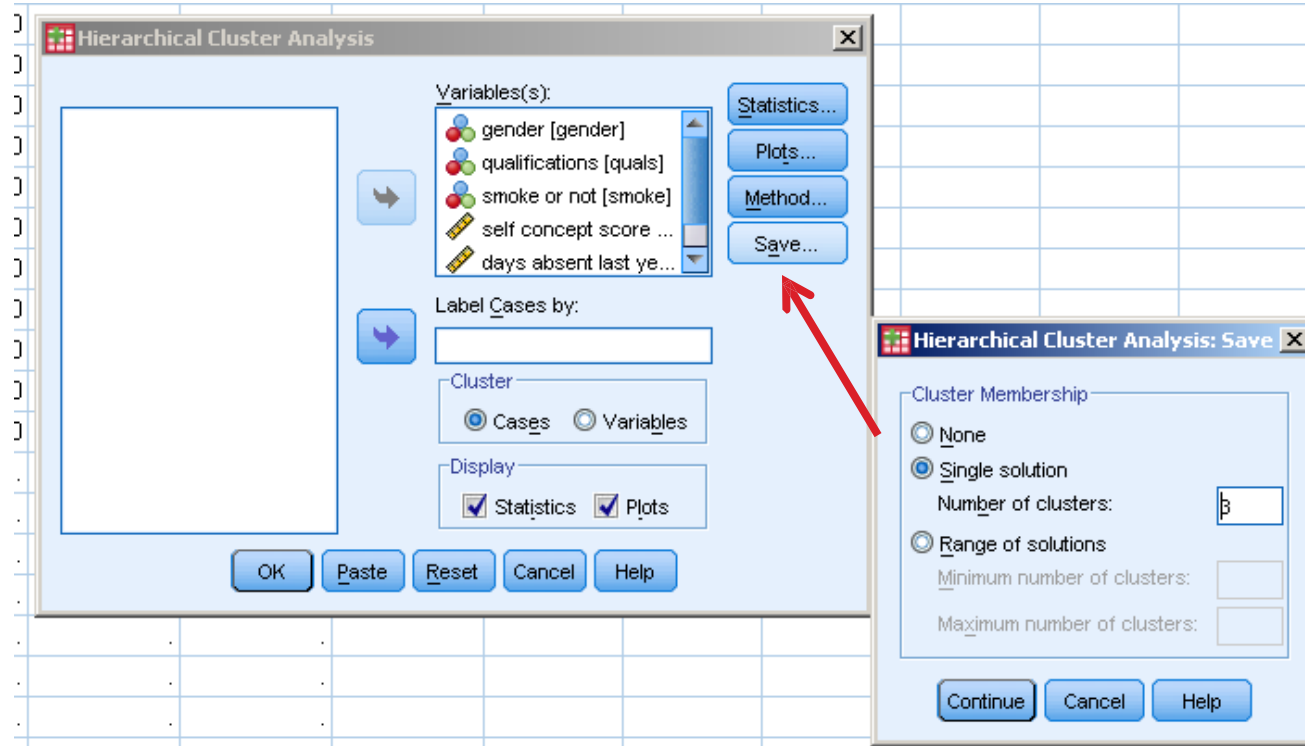
No. of clusters	Agglomeration last step	Coefficients this step	Change
2	3453.150	2362.438	1090.712
3	2362.438	1361.651	1000.787
4	1361.651	1055.071	306.634
5	1055.071	806.071	248.946
6	806.405	586.810	219.595

A clear demarcation point seems to be here.

A good cluster solution sees a sudden jump (gap) in the distance coefficient. The solution before the gap indicates the good solution.

Table 23.2 is a reformed table to see the changes in the coefficients as the number of clusters increase. The final column, headed 'Change', enables us to determine the optimum number of clusters. In this case it is 3 clusters as succeeding clustering adds very much less to distinguishing between cases.

- ▶ Repeat step 1 to 3 to place cases into one of three clusters.



The number you place in the box is the number of clusters that seem best to represent the clustering solution in a parsimonious way.

Finally click **OK**.

Ch 18 data file SPSS A - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

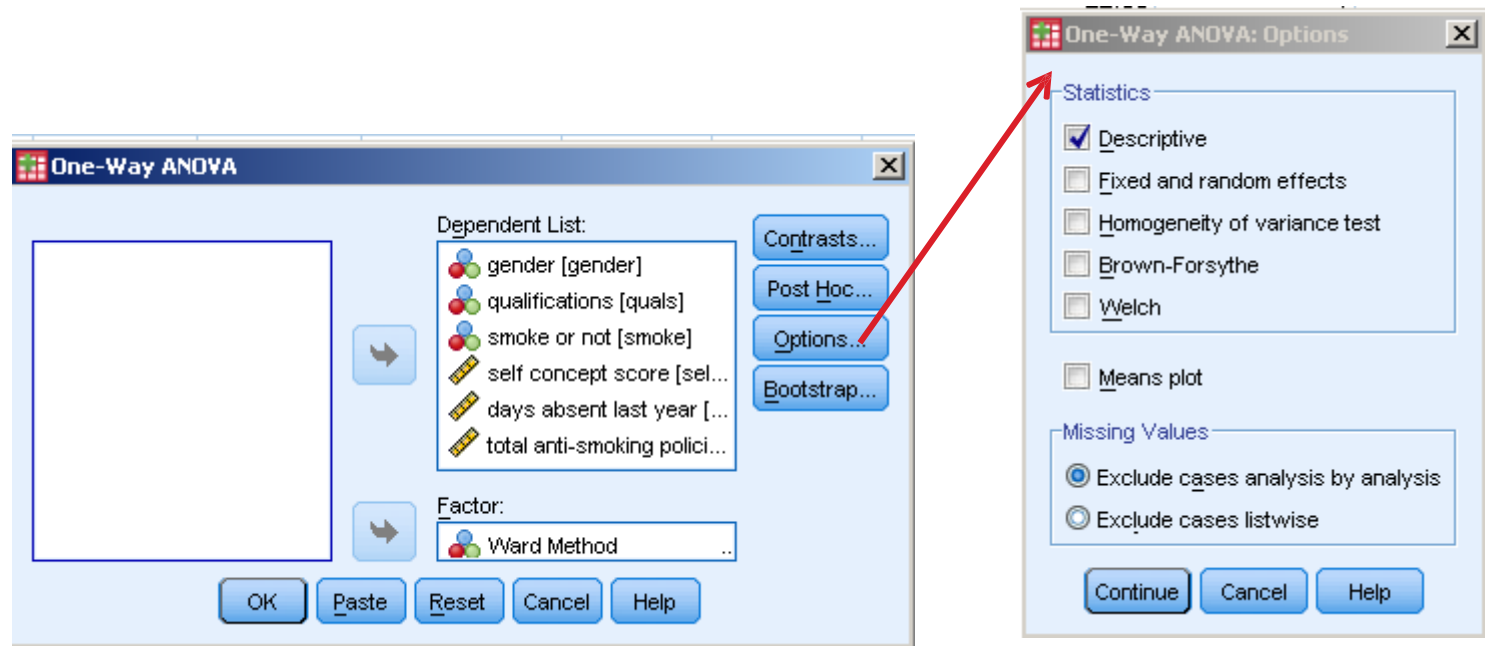
15:

	gender	quals	smoke	selfcon	absence	subtestb	clu3_1	var
1	1	1.00	1.00	22.00	3.00	30.00	1	
2	1	1.00	1.00	45.00	.00	29.00	2	
3	1	1.00	1.00	36.00	21.00	23.00	1	
4	1	2.00	2.00	45.00	14.00	14.00	3	
5	1	1.00	1.00	32.00	9.00	16.00	1	
6	1	1.00	2.00	41.00	5.00	20.00	2	
7	1	1.00	1.00	35.00	8.00	22.00	1	
8	2	1.00	2.00	42.00	12.00	12.00	3	
9	1	3.00	1.00	33.00	8.00	22.00	1	
10	2	1.00	2.00	54.00	30.00	18.00	3	
11	1	2.00	1.00	37.00	.00	23.00	2	
12	1	1.00	1.00	26.00	12.00	22.00	1	
13	1	1.00	1.00	39.00	4.00	24.00	2	
14	1	1.00	1.00	53.00	.00	20.00	2	
15	2	1.00	1.00	23.00	13.00	15.00	1	
16	1	1.00	2.00	43.00	3.00	26.00	2	
17	2	2.00	2.00	52.00	.00	15.00	2	
18	1	1.00	1.00	39.00	.00	19.00	2	
19	1	1.00	1.00	34.00	.00	20.00	2	
20	2	3.00	2.00	45.00	21.00	13.00	3	
21	.	.	.	.	.	.	.	

A new variable has been generated at the end of your SPSS data file called clu3\_1 (labelled Ward method in variable view). This provides the cluster membership for each case in your sample

Nine respondents have been classified in cluster 2, while there are seven in cluster 1 and four in cluster 3.

We now proceed by conducting a **one-way ANOVA** to determine on which classifying variables are significantly different between the groups



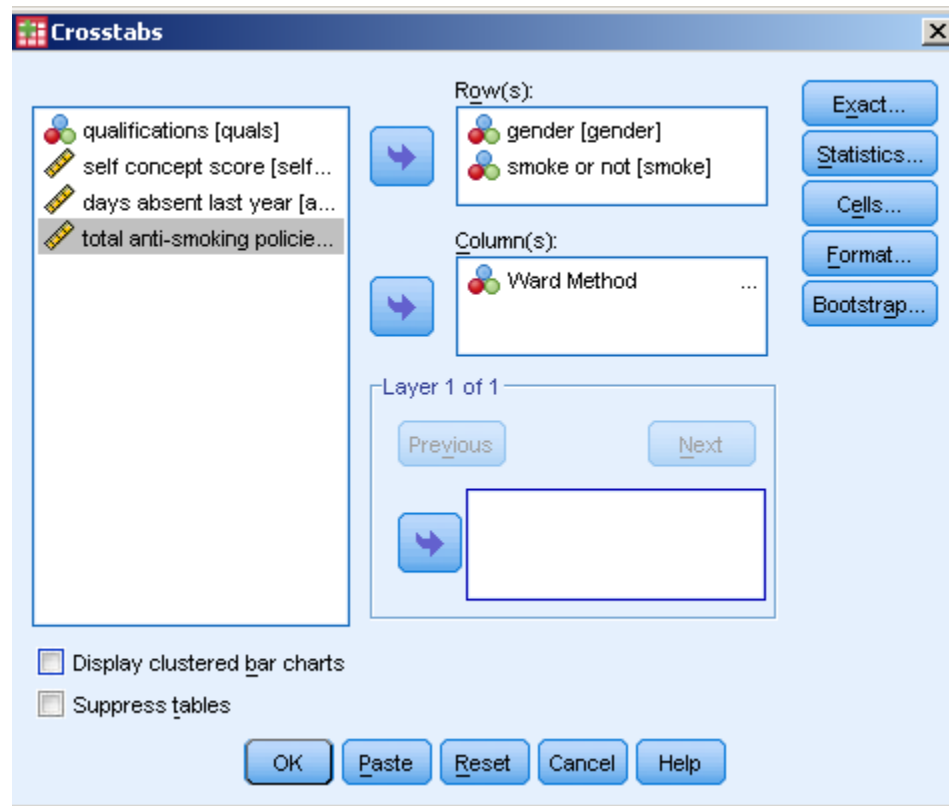
ANOVA						
		Sum of Squares	df	Mean Square	F	Sig.
gender	Between Groups	1.254	2	.627	4.270	.031
	Within Groups	2.496	17	.147		
	Total	3.750	19			
qualifications	Between Groups	.816	2	.408	.897	.426
	Within Groups	7.734	17	.455		
	Total	8.550	19			
smoke or not	Between Groups	2.550	2	1.275	10.838	.001
	Within Groups	2.000	17	.118		
	Total	4.550	19			
self concept score	Between Groups	960.263	2	480.132	13.188	.000
	Within Groups	618.937	17	36.408		
	Total	1579.200	19			
days absent last year	Between Groups	952.086	2	476.043	19.156	.000
	Within Groups	422.464	17	24.851		
	Total	1374.550	19			
total anti-smoking policies subtest B	Between Groups	174.530	2	87.265	4.816	.022
	Within Groups	308.020	17	18.119		
	Total	482.550	19			

*Qualifications* did not produce any significant associations.



# Interpreting Clusters

- ▶ Crosstab analysis of the nominal variables *gender*, *qualifications* and *whether smoke or not* produced some significant associations with clusters.



gender * Ward Method		Crosstabulation			
Count					
		Ward Method			Total
		1	2	3	
gender	male	6	8	1	15
	female	1	1	3	5
Total		7	9	4	20

smoke or not * Ward Method		Crosstabulation			
Count					
		Ward Method			Total
		1	2	3	
smoke or not	non-smoker	7	6	0	13
	smoker	0	3	4	7
Total		7	9	4	20

Cluster 1 : consists of male non smokers

Cluster 2 : consists of smoking and non smoking males

Cluster 3 : consists of smoking females

Descriptives									
		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
self concept score	1	7	29.5714	5.79819	2.19151	24.2090	34.9339	22.00	36.00
	2	9	42.5556	6.48288	2.16096	37.5724	47.5387	34.00	53.00
	3	4	46.5000	5.19615	2.59808	38.2318	54.7682	42.00	54.00
	Total	20	38.8000	9.11679	2.03858	34.5332	43.0668	22.00	54.00
days absent last year	1	7	10.5714	5.62308	2.12533	5.3709	15.7719	3.00	21.00
	2	9	1.3333	2.06155	.68718	-.2513	2.9180	.00	5.00
	3	4	19.2500	8.13941	4.06971	6.2984	32.2016	12.00	30.00
	Total	20	8.1500	8.50557	1.90190	4.1693	12.1307	.00	30.00
total anti-smoking policies subtest B	1	7	21.4286	4.96176	1.87537	16.8397	26.0174	15.00	30.00
	2	9	21.7778	4.17665	1.39222	18.5673	24.9882	15.00	29.00
	3	4	14.2500	2.62996	1.31498	10.0652	18.4348	12.00	18.00
	Total	20	20.1500	5.03958	1.12688	17.7914	22.5086	12.00	30.00

**Cluster 1** is characterized by low self-concept, average absence rate, average attitude score to anti-smoking, non-smoking males.

**Cluster 2** is characterized by moderate self-concept, low absence rate, average attitude score to anti-smoking, smoking and non-smoking males.

**Cluster 3** is characterized by high self-concept, high absence rate, low score to antismoking, smoking females.

# K – means Cluster Procedure

# K – means Cluster Procedure

## Example:

The telecommunication provider wants to segment its customer base by service usage patterns. If customer can be classified by usage, the company can offer more attractive package to its customers. The following variables are;

- Multiple lines
- Voice mail
- Paging service
- Internet
- Caller ID
- Call waiting
- Call forwarding
- 3-way calling
- Electronic billing

telco\_extra.sav [DataSet1] - PASW Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

1: region 2

region tenure

1 2  
2 3  
3 3  
4 2  
5 2  
6 2  
7 3  
8 2  
9 3  
10 1  
11 2  
12 3  
13 1  
14 2  
15 2  
16 1  
17 3  
18 3  
19 2  
20 1  
21 3  
22 1  
23 3

Reports  
Descriptive Statistics  
Tables  
Compare Means  
General Linear Model  
Generalized Linear Models  
Mixed Models  
Correlate  
Regression  
Loglinear  
Neural Networks  
Classify  
Dimension Reduction  
Scale  
Nonparametric Tests  
Forecasting  
Survival  
Multiple Response  
Missing Value Analysis...  
Multiple Imputation  
Complex Samples  
Quality Control  
ROC Curve...

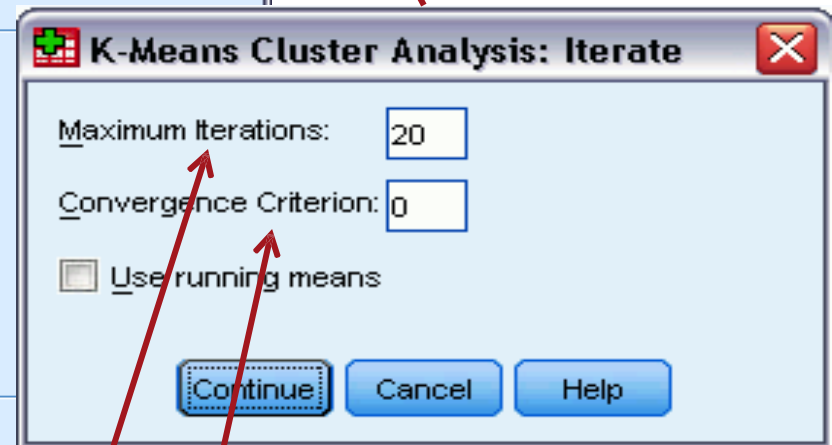
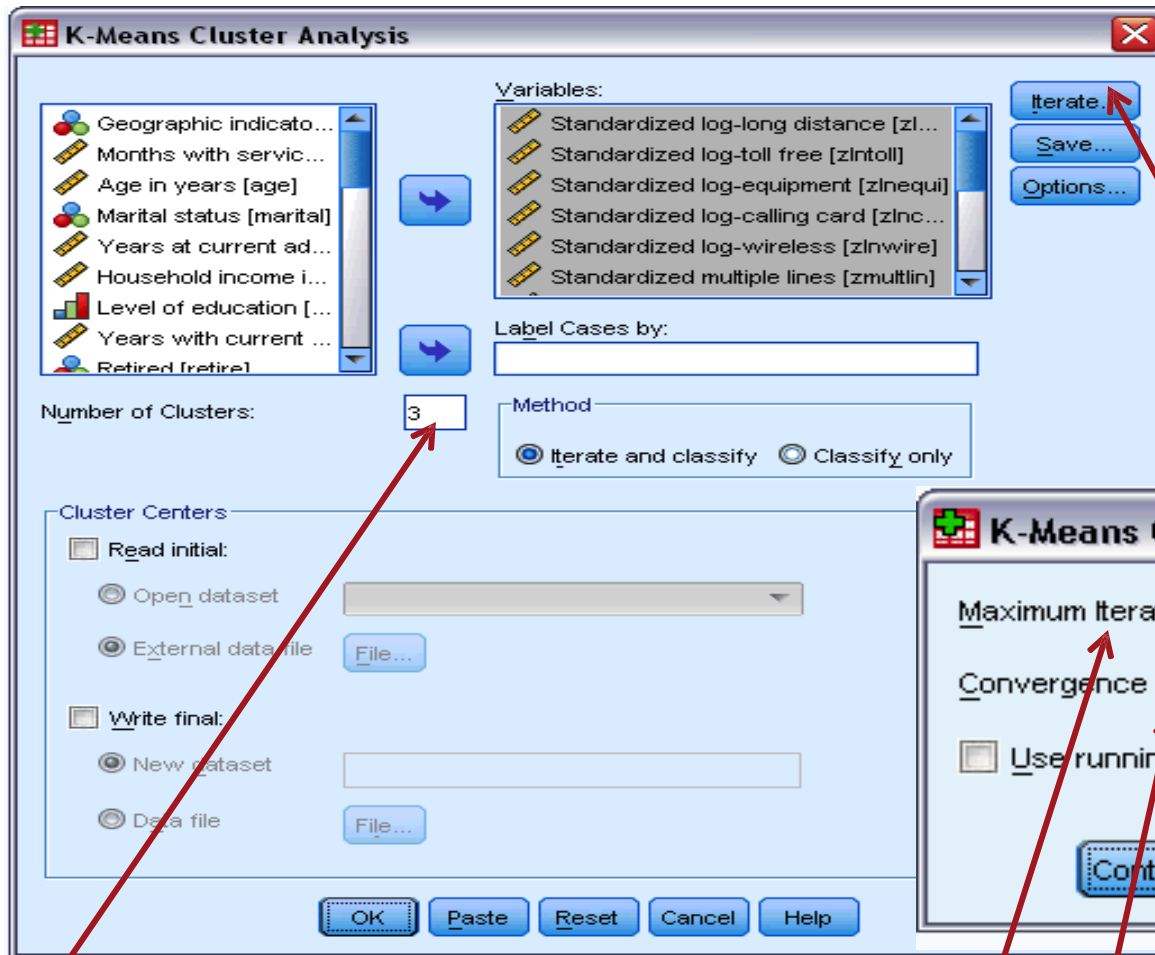
TwoStep Cluster...  
K-Means Cluster...  
Hierarchical Cluster...  
Tree...  
Discriminant...  
Nearest Neighbor...

dress income ed employ retire gender reside tollfree equip callcard wireless lo

9 64.00 4 5 .0 0 2 .0 .0 1.00 .0  
7 136.00 5 5 .0 0 6 1.00 .0 1.00 1.00  
24 116.00 1 29 .0 1 2 1.00 .0 1.00 .0  
12 33.00 2 0 .0 1 1 .0 .0 .0 .0  
9 30.00 1 2 .0 0 4 .0 .0 .0 .0  
17 78.00 2 16 .0 1 1 1.00 .0 1.00 .0  
2 19.00 2 4 .0 1 5 .0 .0 1.00 .0  
5 76.00 2 10 .0 0 3 1.00 1.00 1.00 1.00  
4 31 .0 0 5 1.00 .0 1.00 .0  
1 22 .0 0 3 .0 .0 1.00 .0  
4 5 .0 1 1 .0 1.00 .0 .0  
2 15 .0 1 1 1.00 .0 1.00 .0  
2 9 .0 1 3 .0 .0 1.00 .0  
4 23 .0 1 3 1.00 1.00 .0 1.00  
1 8 .0 1 2 .0 .0 .0 .0  
12 75.00 5 1 .0 0 4 .0 1.00 1.00 .0  
38 162.00 2 30 .0 0 1 1.00 1.00 1.00 .0  
3 49.00 2 6 .0 1 3 1.00 .0 .0 .0  
3 20.00 1 3 .0 0 1 .0 .0 .0 .0  
3 77.00 4 2 .0 0 4 .0 1.00 1.00 1.00  
7 16.00 3 1 .0 1 1 .0 1.00 .0 1.00  
17 120.00 1 24 .0 0 2 .0 .0 1.00 .0  
10 101.00 5 4 .0 1 2 .0 1.00 1.00 1.00

Visible: 46 of 46 Variables

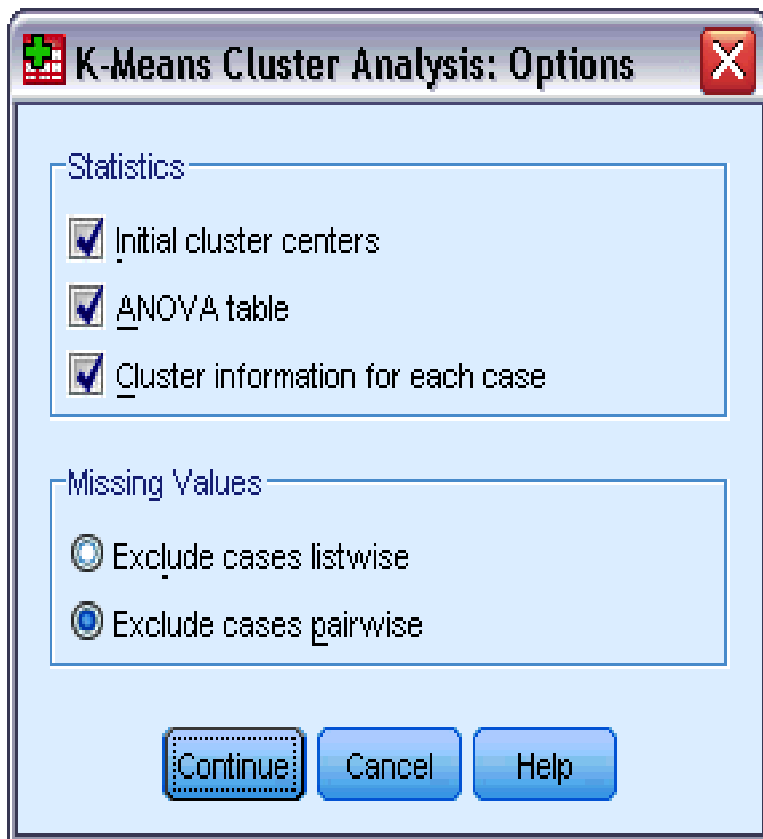
Data View Variable View



Specify number of clusters

Number *iteration or repetition* of combining different clusters.

Determines when iteration cease and represent a proportion of the min. distance bet. Initial cluster center.



**STATISTICS**  
it will show the  
information for each  
group.



# Initial Cluster center

	Cluster		
	1	2	3
Standardized log-long distance	2.48	-1.70	.12
Standardized log-toll free	2.34	-.20	-.39
Standardized log-equipment	1.34	-.65	.59
Standardized log-calling card	2.49	-.86	-1.28
Standardized log-wireless	1.14	-1.75	1.42
Standardized multiple lines	1.05	-.95	1.05
Standardized voice mail	1.51	1.51	1.51
Standardized paging	1.68	1.68	1.68
Standardized internet	1.31	-.76	1.31
Standardized caller id	1.04	1.04	-.96
Standardized call waiting	1.03	-.97	1.03
Standardized call forwarding	1.01	1.01	-.99
Standardized 3-way calling	1.00	1.00	-1.00
Standardized electronic billing	-.77	-.77	1.30

The initial cluster centers are the variable values of the k well-spaced observations.

# Iteration History

The iteration history shows the *progress* of the clustering process at each step.

Iteration	Change in Cluster Centers		
	1	2	3
1	3.298	3.590	3.491
2	1.016	.427	.931
3	.577	.320	.420
4	.240	.180	.195
5	.119	.125	.108
6	.093	.083	.027
7	.069	.094	.032
8	.059	.051	.018
9	.035	.085	.063
10	.025	.359	.333
11	.068	.439	.287
12	.079	.368	.177
13	.125	.139	.078
14	.077	.096	.020
15	.041	.047	.015
16	.014	.027	.000
17	.019	.038	.000
18	.000	.000	.000

In early iterations, the cluster centers shift quite a lot.

Iteration	Change in Cluster Centers		
	1	2	3
1	3.298	3.590	3.491
2	1.016	.427	.931
3	.577	.320	.420
4	.240	.180	.195
5	.119	.125	.108
6	.093	.083	.027
7	.069	.094	.032
8	.059	.051	.018
9	.035	.085	.063
10	.025	.359	.333
11	.068	.439	.287
12	.079	.368	.177
13	.125	.139	.078
14	.077	.096	.020
15	.041	.047	.015
16	.014	.027	.000
17	.019	.038	.000
18	.000	.000	.000

By the 14th iteration, they have settled down to the general area of their final location, and the last four iterations are minor adjustments.

# ANOVA

# ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Standardized log-long distance	13.063	2	.976	997	13.387	.000
Standardized log-toll free	43.418	2	.820	472	52.932	.000
Standardized log-equipment	99.056	2	.488	383	202.999	.000
Standardized log-calling card	6.301	2	.984	675	6.402	.002
Standardized log-wireless	52.879	2	.646	293	81.873	.000
Standardized multiple lines	38.032	2	.926	997	41.084	.000
Standardized voice mail	236.301	2	.528	997	447.554	.000
Standardized paging	298.992	2	.402	997	743.348	.000
Standardized internet	123.447	2	.754	997	163.642	.000
Standardized caller id	308.104	2	.384	997	802.474	.000
Standardized call waiting	294.674	2	.411	997	717.172	.000
Standardized call forwarding	288.343	2	.424	997	680.718	.000
Standardized 3-way calling	262.397	2	.476	997	551.678	.000
Standardized electronic billing	112.782	2	.776	997	145.381	.000

The ANOVA table indicates which variables contribute the most to your cluster solution.

Variables with large F values provide the greatest separation between clusters.

# FINAL CLUSTER CENTERS

	Cluster		
	1	2	3
Standardized log-long distance	.05	.22	-.16
Standardized log-toll free	.24	.12	-1.05
Standardized log-equipment	.81	-.19	-.69
Standardized log-calling card	.17	.02	-.17
Standardized log-wireless	.42	-.75	-1.00
Standardized multiple lines	.48	-.29	-.05
Standardized voice mail	1.26	-.24	-.44
Standardized paging	1.43	-.38	-.44
Standardized internet	.81	-.59	-.02
Standardized caller id	.82	.71	-.81
Standardized call waiting	.76	.72	-.80
Standardized call forwarding	.78	.69	-.79
Standardized 3-way calling	.74	.67	-.75
Standardized electronic billing	.70	-.63	.05

The final cluster centers are computed as the mean for each variable within each final cluster. The final cluster centers reflect the characteristics of the typical case for each cluster.

Customers in cluster 1 tend to be big spenders who purchase a lot of services.

## FINAL CLUSTER CENTERS

## FINAL CLUSTER CENTERS

	Cluster		
	1	2	3
Standardized log-long distance	.05	.22	-.16
Standardized log-toll free	.24	.12	-1.05
Standardized log-equipment	.81	-.19	-.69
Standardized log-calling card	.17	.02	-.17
Standardized log-wireless	.42	-.75	-1.00
Standardized multiple lines	.48	-.29	-.05
Standardized voice mail	1.26	-.24	-.44
Standardized paging	1.43	-.38	-.44
Standardized internet	.81	-.59	-.02
Standardized caller id	.82	.71	-.81
Standardized call waiting	.76	.72	-.80
Standardized call forwarding	.78	.69	-.79
Standardized 3-way calling	.74	.67	-.75
Standardized electronic billing	.70	-.63	.05

Customers in cluster 2 tend to be moderate spenders who purchase the "calling" services.

## FINAL CLUSTER CENTERS

	Cluster		
	1	2	3
Standardized log-long distance	.05	.22	-.16
Standardized log-toll free	.24	.12	-1.05
Standardized log-equipment	.81	-.19	-.69
Standardized log-calling card	.17	.02	-.17
Standardized log-wireless	.42	-.75	-1.00
Standardized multiple lines	.48	-.29	-.05
Standardized voice mail	1.26	-.24	-.44
Standardized paging	1.43	-.38	-.44
Standardized internet	.81	-.59	-.02
Standardized caller id	.82	.71	-.81
Standardized call waiting	.76	.72	-.80
Standardized call forwarding	.78	.69	-.79
Standardized 3-way calling	.74	.67	-.75
Standardized electronic billing	.70	-.63	.05

Customers in cluster 3 tend to spend very little and do not purchase many services.

## DISTANCES BETWEEN FINAL CLUSTER CENTER

Cluster	1	2	3
1		3.500	4.863
2	3.500		3.396
3	4.863	3.396	

This table shows the Euclidean distances between the final cluster centers. Greater distances between clusters correspond to greater dissimilarities.

Clusters 1 and 3 are most different.

Cluster 2 is approximately equally similar to clusters 1 and 3.

Cluster	1	2	3
1		3.500	4.863
2	3.500		3.396
3	4.863	3.396	

# Number of Cases in Each Cluster

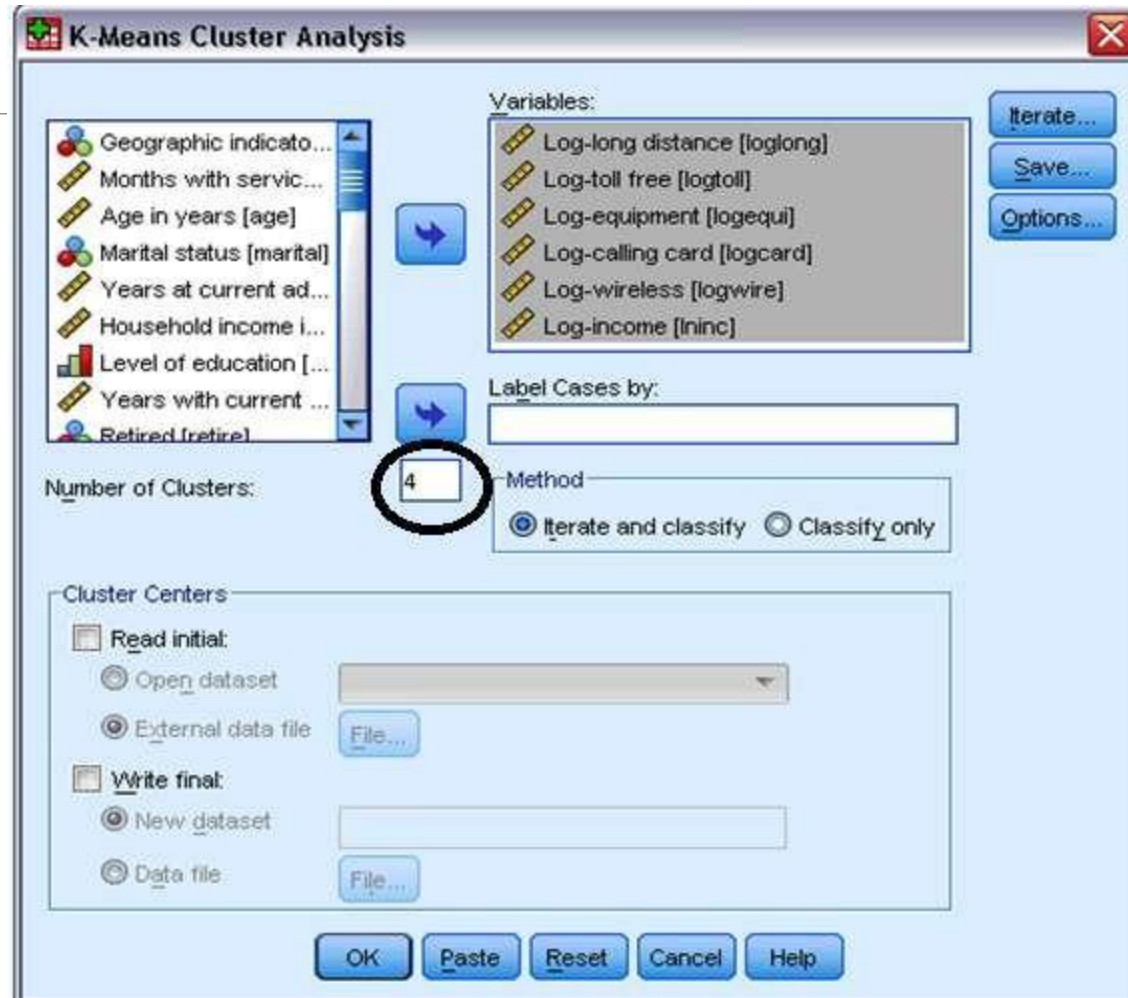
Cluster	1	226.000
	2	292.000
	3	482.000
Valid		1000.000
Missing		.000

A large number of cases were assigned to the third cluster, which unfortunately is the least profitable group.

Perhaps a fourth, more profitable, cluster could be extracted from this "basic service" group.

# MAIN DIALOG BOX

# MAIN DIALOG BOX



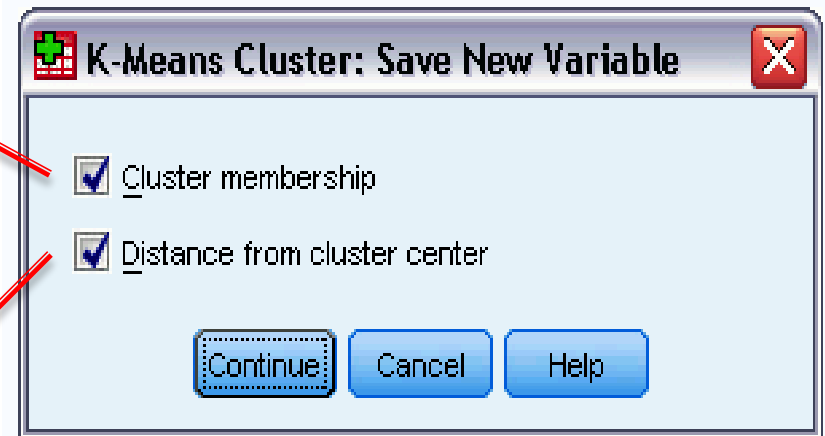


# SAVE DIALOG BOX

## SAVE DIALOG BOX

Creates a new variable indicating the final cluster membership of each case.

Creates a new variable indicating the Euclidean distances bet. Each cases and its classification center.



# FINAL CLUSTER CENTERS

This table shows that an important grouping is missed in the three-cluster solution

	Cluster			
	1	2	3	4
Standardized log-long distance	.23	-.48	.05	.24
Standardized log-toll free	-.75	-1.10	.26	.11
Standardized log-equipment	-.36	-1.28	.86	-.20
Standardized log-calling card	-.06	-.37	.18	.06
Standardized log-wireless	-.84	-1.54	.45	-.71
Standardized multiple lines	.75	-.74	.45	-.26
Standardized voice mail	-.21	-.59	1.30	-.23
Standardized paging	-.32	-.51	1.50	-.37
Standardized internet	.57	-.52	.77	-.56
Standardized caller id	-.78	-.73	.86	.72
Standardized call waiting	-.76	-.73	.80	.74
Standardized call forwarding	-.71	-.82	.83	.76
Standardized 3-way calling	-.69	-.79	.84	.72
Standardized electronic billing	.58	-.38	.67	-.63

Members of clusters 1 and 2 are drawn from cluster 3 in the three-cluster solution, and they are *unlikely* to be big spenders.

	Cluster			
	1	2	3	4
Standardized log-long distance	.23	-.48	.05	.24
Standardized log-toll free	-.75	-1.10	.26	.11
Standardized log-equipment	-.36	-1.28	.86	-.20
Standardized log-calling card	-.06	-.37	.18	.06
Standardized log-wireless	-.84	-1.54	.45	-.71
Standardized multiple lines	.75	-.74	.45	-.26
Standardized voice mail	-.21	-.59	1.30	-.23
Standardized paging	-.32	-.51	1.50	-.37
Standardized internet	.57	-.52	.77	-.56
Standardized caller id	-.78	-.73	.86	.72
Standardized call waiting	-.76	-.73	.80	.74
Standardized call forwarding	-.71	-.82	.83	.76
Standardized 3-way calling	-.69	-.79	.84	.72
Standardized electronic billing	.58	-.38	.67	-.63

Clusters 3 and 4 seem to correspond to clusters 1 and 2 from the three-cluster solution.

# DISTANCE BETWEEN FINAL CLUSTER CENTERS

Cluster 4 is still equally similar to the other clusters.

Cluster	1	2	3	4
1		2.568	4.454	3.631
2	2.568		5.746	3.675
3	4.454	5.746		3.515
4	3.631	3.675	3.515	

Cluster	1	2	3	4
1		2.568	4.454	3.631
2	2.568		5.746	3.675
3	4.454	5.746		3.515
4	3.631	3.675	3.515	

Clusters 1 and 2 are the most similar, which makes sense because they were combined into one cluster in the three-cluster solution.

# NUMBER OF CASES IN EACH CLUSTER

Cluster	1	236.000
	2	272.000
	3	212.000
	4	280.000
Valid		1000.000
Missing		.000

Nearly 25% of cases belong to the newly created group of "E-service" customers, which is very significant to your profits.

# Two – Step Cluster Analysis

# Two Step Cluster Analysis

- ▶ The Two Step Cluster Analysis procedure is an exploratory tool designed to reveal natural groupings (or clusters) within a data set that would otherwise not be apparent. The algorithm employed by this procedure has several desirable features that differentiate it from traditional clustering techniques:
  - The ability to create clusters based on both categorical and continuous variables.
  - Automatic selection of the number of clusters.
  - The ability to analyze large data files efficiently.

# Clustering Principles

In order to handle categorical and continuous variables, the TwoStep Cluster Analysis procedure uses a likelihood distance measure which assumes that variables in the cluster model are independent.

Further, each continuous variable is assumed to have a normal (Gaussian) distribution and each categorical variable is assumed to have a multinomial distribution. Empirical internal testing indicates that the procedure is fairly robust to violations of both the assumption of independence and the distributional assumptions, but you should try to be aware of how well these assumptions are met.

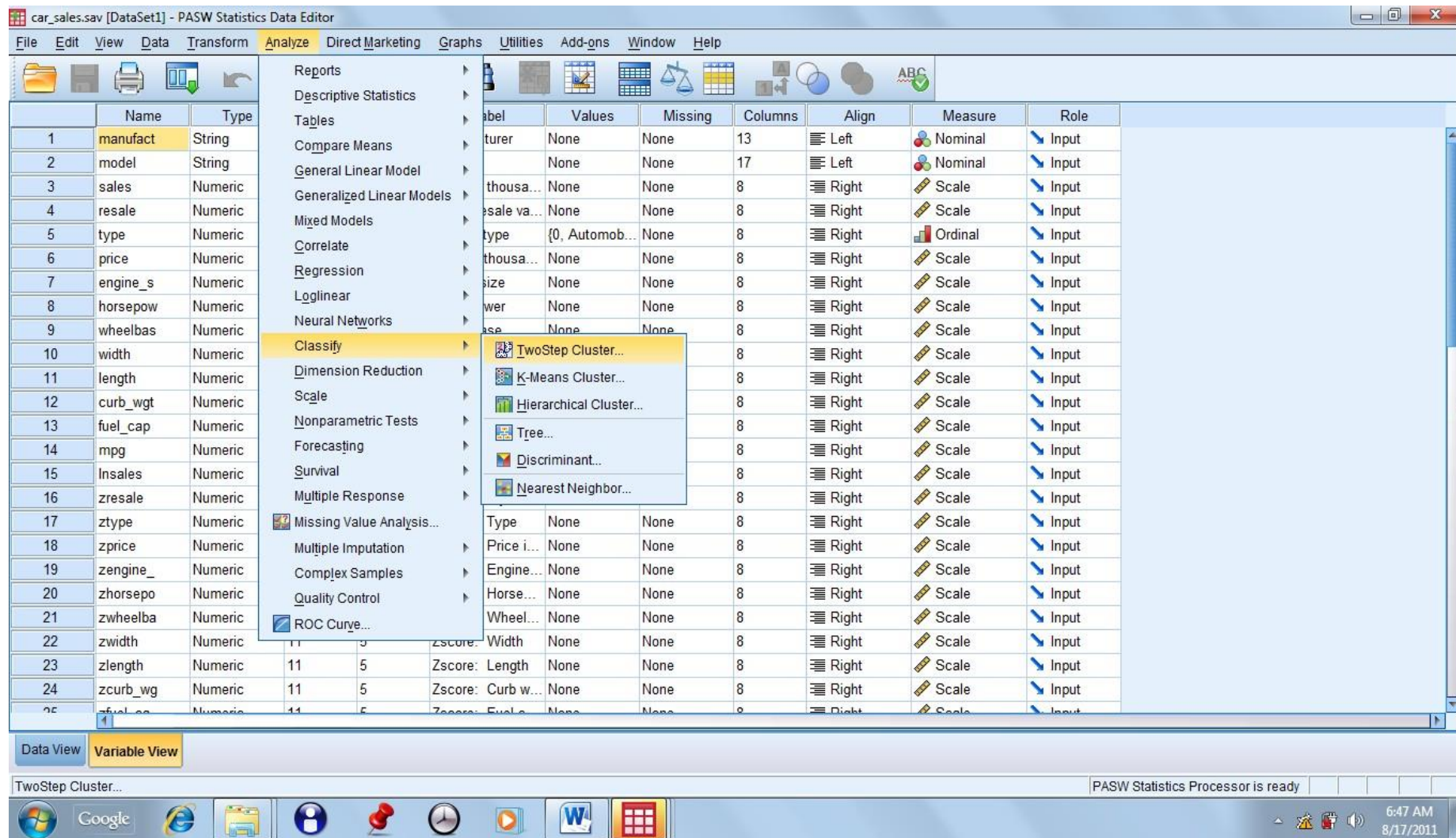
The two steps of the TwoStep Cluster Analysis procedure's algorithm can be summarized as follows:

- ▶ **Step 1.** The procedure begins with the construction of a Cluster Features (CF) Tree. The tree begins by placing the first case at the root of the tree in a leaf node that contains variable information about that case. Each successive case is then added to an existing node or forms a new node, based upon its similarity to existing nodes and using the distance measure as the similarity criterion. A node that contains multiple cases contains a summary of variable information about those cases. Thus, the CF tree provides a capsule summary of the data file.
- ▶ **Step 2.** The leaf nodes of the CF tree are then grouped using an agglomerative clustering algorithm. The agglomerative clustering can be used to produce a range of solutions. To determine which number of clusters is "best", each of these cluster solutions is compared using Schwarz's Bayesian Criterion (BIC) or the Akaike Information Criterion (AIC) as the clustering criterion.

## Using TwoStep Cluster Analysis to Classify Motor Vehicles

Car manufacturers need to be able to appraise the current market to determine the likely competition for their vehicles. If cars can be grouped according to available data, this task can be largely automatic using cluster analysis.





## Running the Analysis

**TwoStep Cluster Analysis**

**Variables:**

- Manufacturer [manufa...]
- Model [model]
- Sales in thousands [s...]
- 4-year resale value [r...]
- Log-transformed sale...
- Zscore: 4-year resal...
- Zscore: Type [ztype]
- Zscore: Price in thou...
- Zscore: Engine size I...

**Categorical Variables:**

- Vehicle type [type]

**Continuous Variables:**

- Price in thousands [pri...]
- Engine size [engine\_s]
- Horsepower [horsepo...]
- Miles per gallon [milespergall...]

**Distance Measure**

☒ Log-likelihood

☐ Euclidean

**Count of Continuous Variables**

To be Standardized: 9

Assumed Standardized: 0

**Number of Clusters**

☒ Determine automatically

Maximum: 15

☐ Specify fixed

Number: 5

**Clustering Criterion**

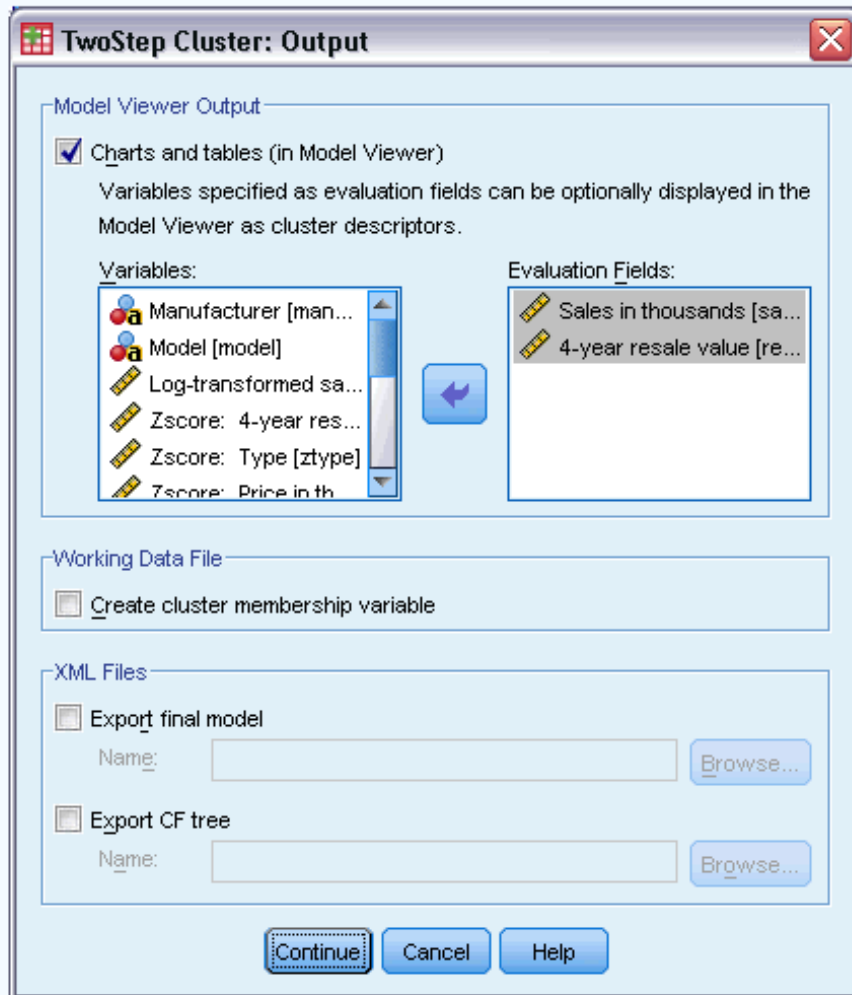
☒ Schwarz's Bayesian Criterion (BIC)

☐ Akaike's Information Criterion (AIC)

OK Paste Reset Cancel Help

- If the variable list does not display variable labels in file order, right-click anywhere in the variable list and from the context menu choose **Display Variable Labels** and **Sort by File Order**.
- Select *Vehicle type* as a categorical variable.
- Select *Price in thousands* through *Fuel efficiency* as continuous variables.
- Click **Output**.

## Running the Analysis



The dialog box is titled "TwoStep Cluster: Output" and contains three main sections: "Model Viewer Output", "Working Data File", and "XML Files".

**Model Viewer Output**

- ☒ **Charts and tables (in Model Viewer)**  
Variables specified as evaluation fields can be optionally displayed in the Model Viewer as cluster descriptors.
- Variables:**
  - Manufacturer [man...]
  - Model [model]
  - Log-transformed sa...
  - Zscore: 4-year res...
  - Zscore: Type [ztype]
  - Zscore: Price in th...
- Evaluation Fields:**
  - Sales in thousands [sa...]
  - 4-year resale value [re...]

**Working Data File**

- ☐ **Create cluster membership variable**

**XML Files**

- ☐ **Export final model**  
Name:
- ☐ **Export CF tree**  
Name:

Buttons:

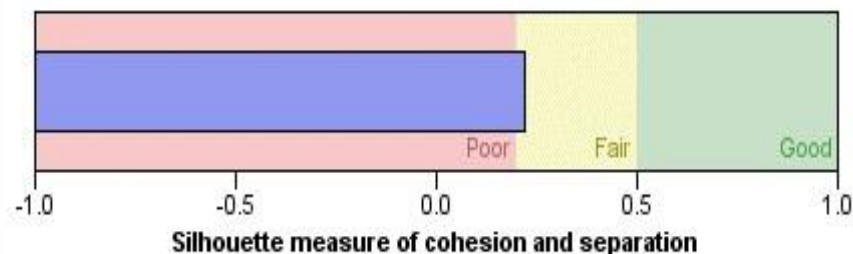
- Select *Sales in thousands [sales]* and *4-year resale value [resale]* as evaluation fields. These fields will not be used to create the cluster model, but can give you further insight to the clusters created by the procedure.
- Click **Continue** and then click **OK**.

## Model Summary and Cluster Quality

### Model Summary

Algorithm	TwoStep
Inputs	7
Clusters	3

### Cluster Quality

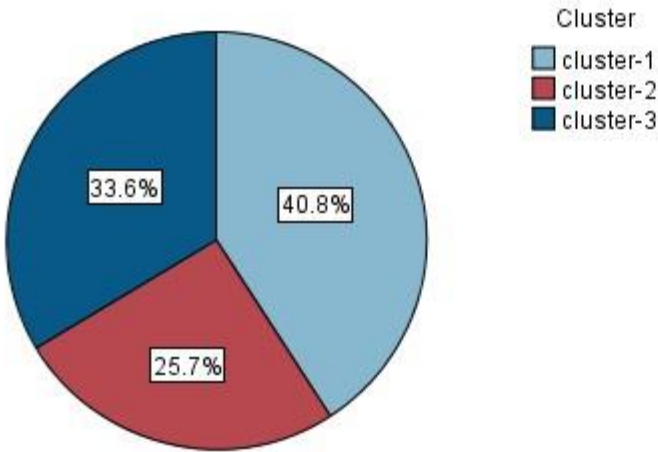


The Viewer contains a Model Viewer object. By activating (double-clicking) this object, you gain an interactive view of the model. The default main view is the Model Summary view.

- The model summary table indicates that three clusters were found based on the ten input features (fields) you selected.
- The cluster quality chart indicates that the overall model quality is "Fair".

Cluster Distribution

Cluster Sizes



Size of Smallest Cluster	39 (25.7%)
Size of Largest Cluster	62 (40.8%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	1.59

The Cluster Sizes view shows the frequency of each cluster. Hovering over a slice in the pie chart reveals the number of records assigned to the cluster. 40.8% (62) of the records were assigned to the first cluster, 25.7% (39) to the second, and 33.6% (51) to the third.



## Cluster Profiles

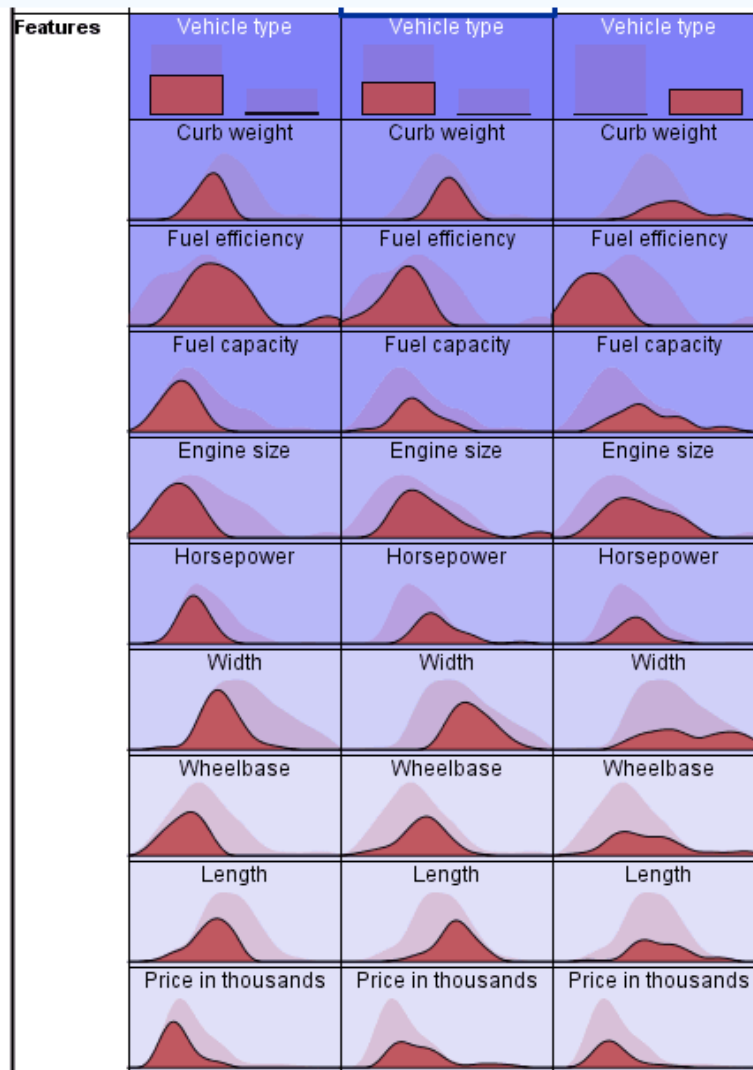
Features	Vehicle type Automobile (98.4%)	Vehicle type Automobile (100.0%)	Vehicle type Truck (100.0%)
	Curb weight 2.84	Curb weight 3.58	Curb weight 3.97
	Fuel efficiency 27.24	Fuel efficiency 23.02	Fuel efficiency 19.51
	Fuel capacity 15.00	Fuel capacity 18.40	Fuel capacity 22.10
	Engine size 2.20	Engine size 3.70	Engine size 3.60
	Horsepower 143.24	Horsepower 232.96	Horsepower 187.92
	Width 68.50	Width 72.90	Width 72.70
	Wheelbase 102.60	Wheelbase 109.00	Wheelbase 113.00
	Length 178.20	Length 194.70	Length 191.10
	Price in thousands 19.62	Price in thousands 37.30	Price in thousands 26.56

The cluster means suggest that the clusters are well separated.

- Motor vehicles in cluster 1 are cheap, small, and fuel efficient automobiles, except for a single truck (the 1.6% of the cluster not comprised of automobiles).
- Motor vehicles in cluster 2 (column 3) are moderately priced, heavy, and have a large gas tank, presumably to compensate for their poor fuel efficiency. Cluster 2 is also entirely comprised of trucks.
- Motor vehicles in cluster 3 (column 2) are expensive, large, and are moderately fuel efficient automobiles.



## Cluster Profiles

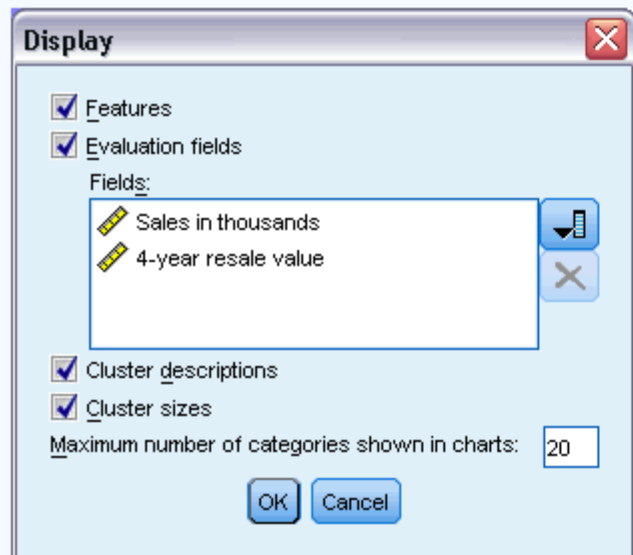


- The cluster means (for continuous fields) and modes (for categorical fields) are useful, but only give information about the cluster centers. In order to get a visualization of the distribution of values for each field by cluster, click on the **Cells show absolute distributions** button in the toolbar.

Now you can see, for example, that there is some overlap between clusters 1 and 3 on curb weight, engine size, and fuel capacity. There is considerably more overlap between clusters 2 and 3 on these fields, with the difference that the vehicles with the very highest curb weight and fuel capacity are in cluster 2 (column 3) and the vehicles with the very highest engine size appear to be in cluster 3 (column 2).



## Cluster Profiles



- To see this information for the evaluation fields, click on the **Display** button in the toolbar.
- Select **Evaluation fields**.
- Click **OK**.

The evaluation fields should now appear in the cluster table.

## Cluster Profiles



The distribution of sales is similar across clusters, except that clusters 1 and 2 have longer tails than cluster 3 (column 2). There is a fair amount of overlap in the distributions of 4-year resale value, but clusters 2 and 3 are centered on a higher value than cluster 1, and cluster 3 has a longer tail than either cluster 1 or 2.

## Cluster Profiles

## Cluster Comparison

■ 1 ■ 3 ■ 2

Vehicle type



Automobile



Truck

Curb weight



Fuel efficiency



Fuel capacity



- For another way to compare clusters, select (control-click) on the cluster numbers (column headings) in the clusters table.

- In the auxiliary view, select **Cluster Comparison** from the dropdown.

For each categorical field, this shows a dot plot for the modal category of each cluster, with dot size corresponding to the percentage of records. For continuous fields, this shows a boxplot for the distribution of values within each cluster overlaid on a boxplot for the distribution of values overall. These plots generally confirm what you've seen in the Clusters view. The Cluster Comparison view can be especially helpful when there are many clusters, and you want to compare only a few of them.

## Summary

Using the TwoStep Cluster Analysis procedure, you have separated the motor vehicles into three fairly broad categories. In order to obtain finer separations within these groups, you should collect information on other attributes of the vehicles. For example, you could note the crash test performance or the options available.

