Principal Components Analysis & Cluster Analysis

Quantitative Data Analysis

José Novais Santos

Summary

- Factor analysis
  - Principal components analysis
- Cluster analysis
  - Hierarchical
  - K-Means
- Cluster analysis & Principal component analysis

Multivariate

- Open file “Oh8 – survey analysis v5.sav”

Factor analysis

- Is about data reduction
  - Whiteout losing significant information
- Identifies unobserved variables (factors) that explain patterns of correlations within a set of observed variables
  - Identify a small number of factors that explain most of the variance embedded in a larger number of variables
- Herein we focusing exploratory factor analysis, particularly, principal components analysis

Principal components analysis

- The basic idea is to summarize highly correlated items within latent variables
  - Latent variables (also called factor or component) are not directly observable but each is inferred and based on several items
    - It is standard practice for factors to be illustrated by circles, whereas rectangles are used for items
- Factors are, by definition, uncorrelated
  - Each factor covers distinct and unrelated aspects
  - If we use them in regression analysis (at the expense of accuracy), collinearity is not an issue

Principal components analysis

- Assumptions
  - Large number of response categories (five or more) for the items (e.g. Likert or higher)
    - Using ordinal data requires the items' scale steps to be equidistant
  - Sample size rule of thumb: at least ten times the number of items used for analysis (after excluding cases with missing values, which is recommended)
  - Items should be sufficiently correlated

Principal components analysis

- Checking items correlation
  - Look at the correlation matrix
  - Examine the “anti-image”
    - The anti-image describes the portion of an item’s variance that is independent of another item in the analysis
    - Kaiser–Meyer–Olkin (KMO) statistic
      - Also called the measure of sampling adequacy (MSA), indicates whether the correlations between variables can be explained by the other variables in the dataset
    - Bartlett’s test of sphericity
      - Test the null hypothesis that the correlation matrix is a diagonal matrix (i.e., all non-diagonal elements are zero) in the population
      - This test statistic values go hand in hand with KMO values

Principal components analysis

- Checking items correlation

Mooi (2011)
Table 8.2
Threshold values for
KMO and MSA

KMO/MSA value	Adequacy of the correlations
Below 0.50	Unacceptable
0.50–0.59	Miserable
0.60–0.69	Mediocre
0.70–0.79	Middling
0.80–0.89	Meritorious
0.90 and higher	Marvelous

– How do we proceed if this is not the case?
  - We can try to affect the results negatively (examining the correlation matrix) or to remove items with MSA values below 0,5 from the analysis (looking at the anti-image correlation matrix)

    (see e.g. Mooi, 2011, p.208)

Principal components analysis

- Determining the number of factors using Kaiser criterion
  - Extracting all factors with an Eigenvalue greater than one (1)
    - We should include (or exclude) additional factors with a smaller (or higher) Eigenvalue than 1 if it is beneficial for interpreting the solution in a meaningful way
  - Eigenvalue
    - Describes how much variance is accounted for by a certain factor

Principal components analysis

- Communality
  - Indicates how much of each variable's variance is captured (or reproduced) by the factors extracted
  - Should lie above 0.30 (30%)
  - Every additional factor extracted will increase this variance and if we extract as many factors as there are items, the communality of each variable would be 1 (100%)
    - But we aim to reduce the number of variables

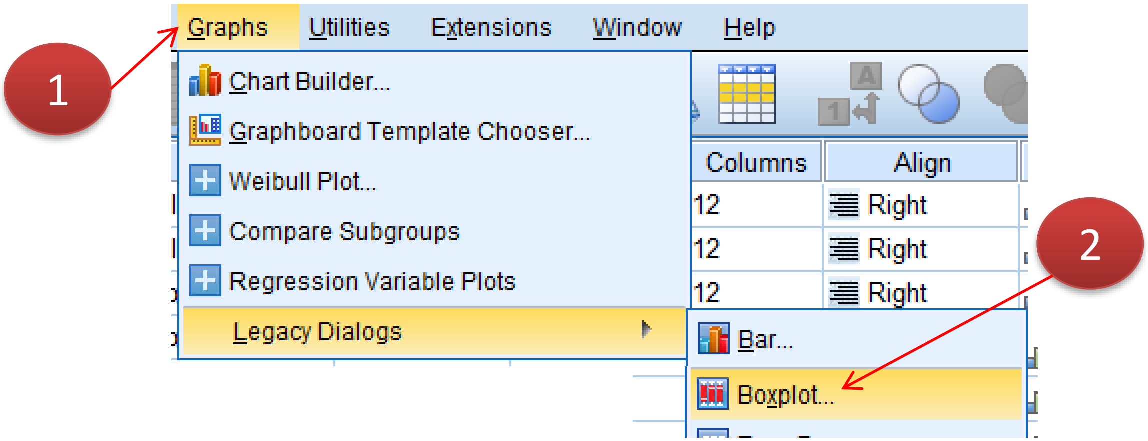
Principal components analysis

- Interpreting the factor solution
  - We have to determine which variables strongly relate to each of the factors extracted
  - Examining the factor loadings
    - Which represent the correlations between the factors and the variables and, thus, can take values from -1 to +1)
  - Making use of factor rotation (it facilitate interpretation)
    - Varimax rotation enhances the interpretability of the results

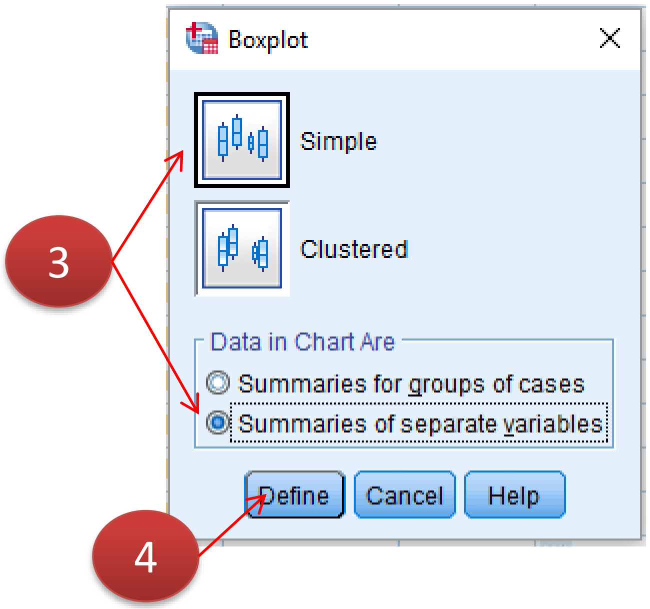
Principal components analysis

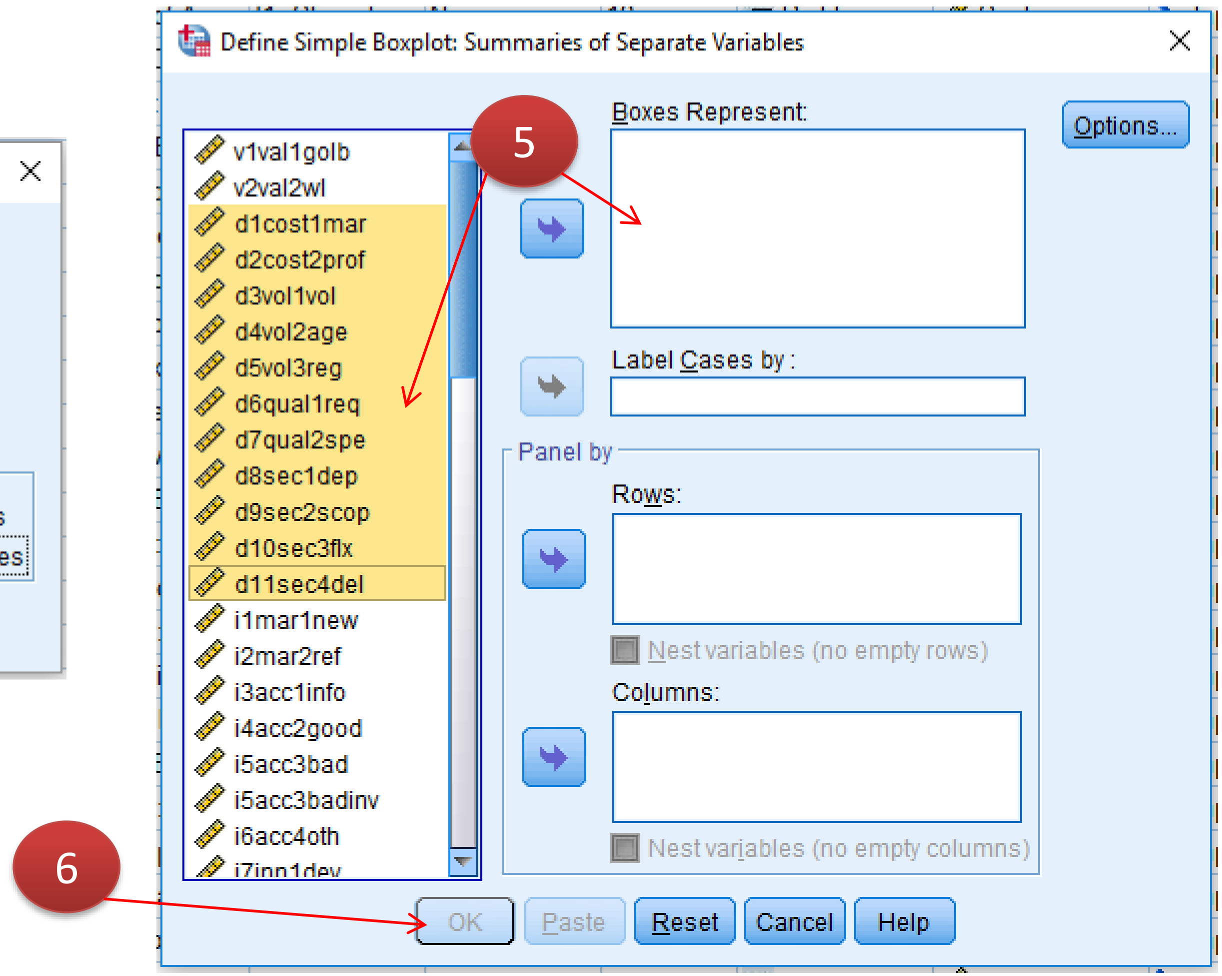
- We want to identify the latent principal components in the 11 direct functions

  (before this analysis we will detect the presence of outliers and responses variability creating boxplots)



Principal components analysis





Principal components analysis

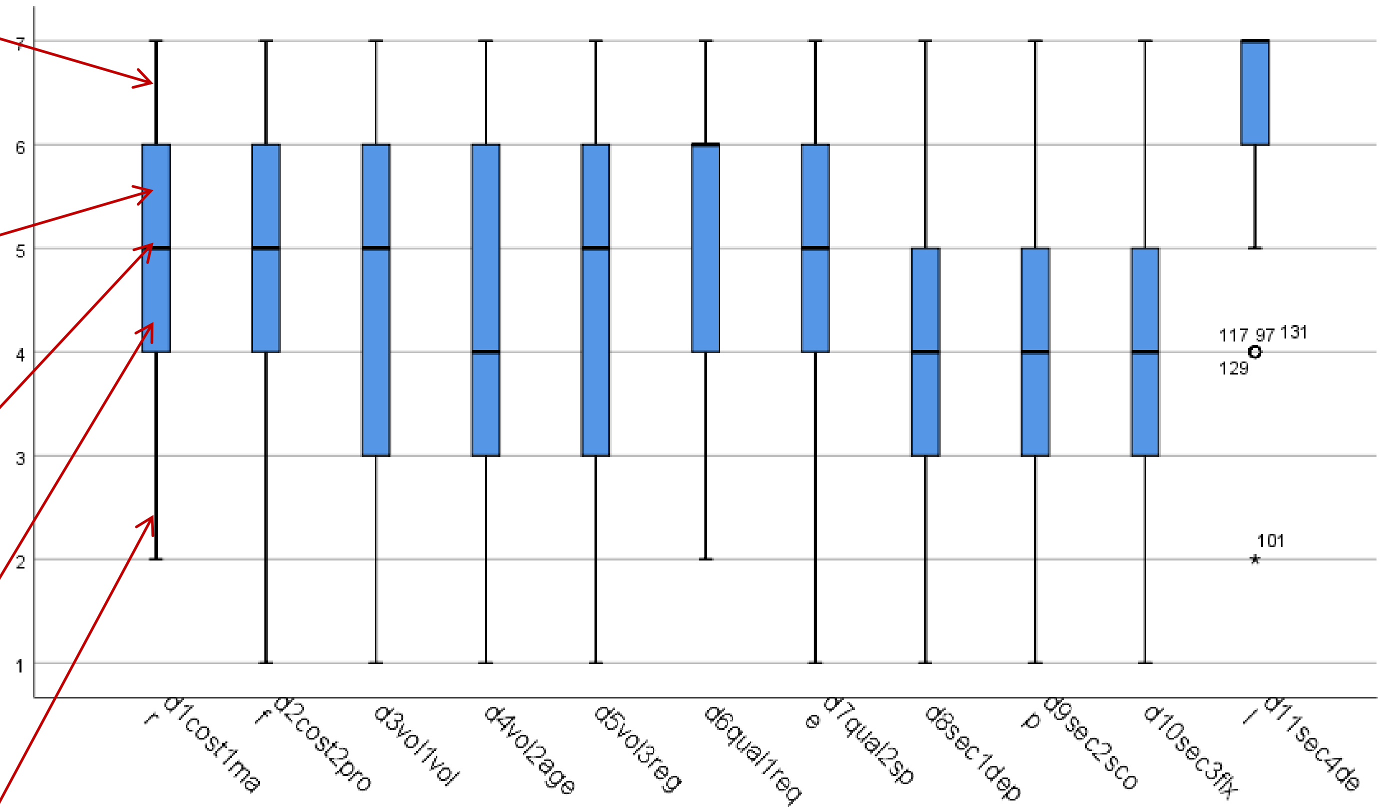
Upper whisker
25%
Quartile group 4

Upper quartile
25%
Quartile group 3

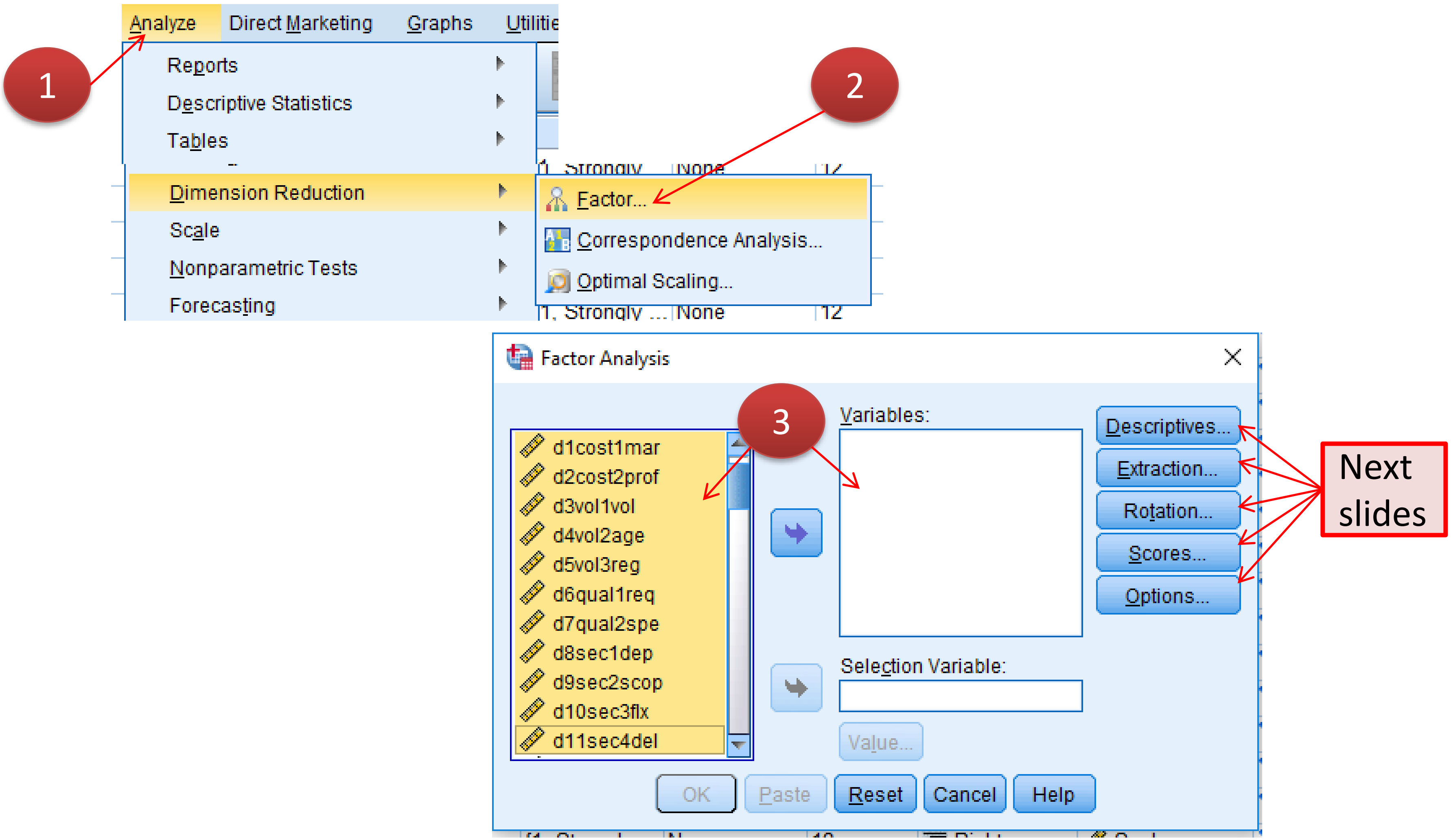
Median
(medium quartile)

Lower quartile
25%
Quartile group 2

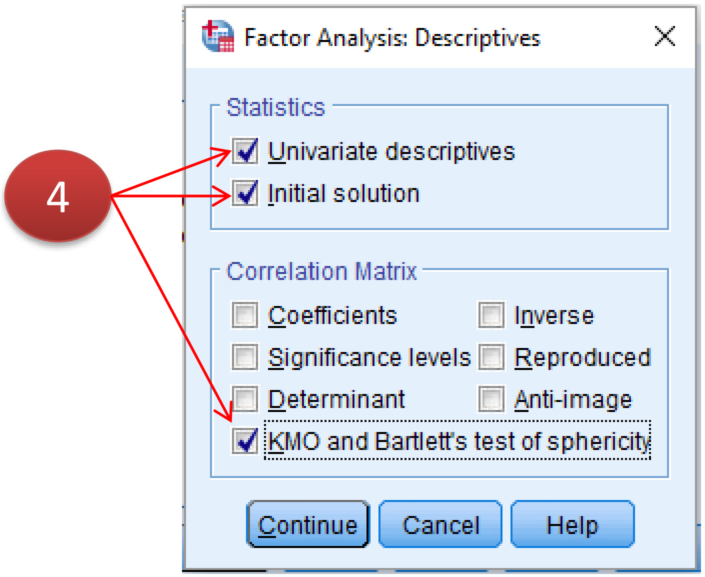
Lower whisker
25%
Quartile group 1

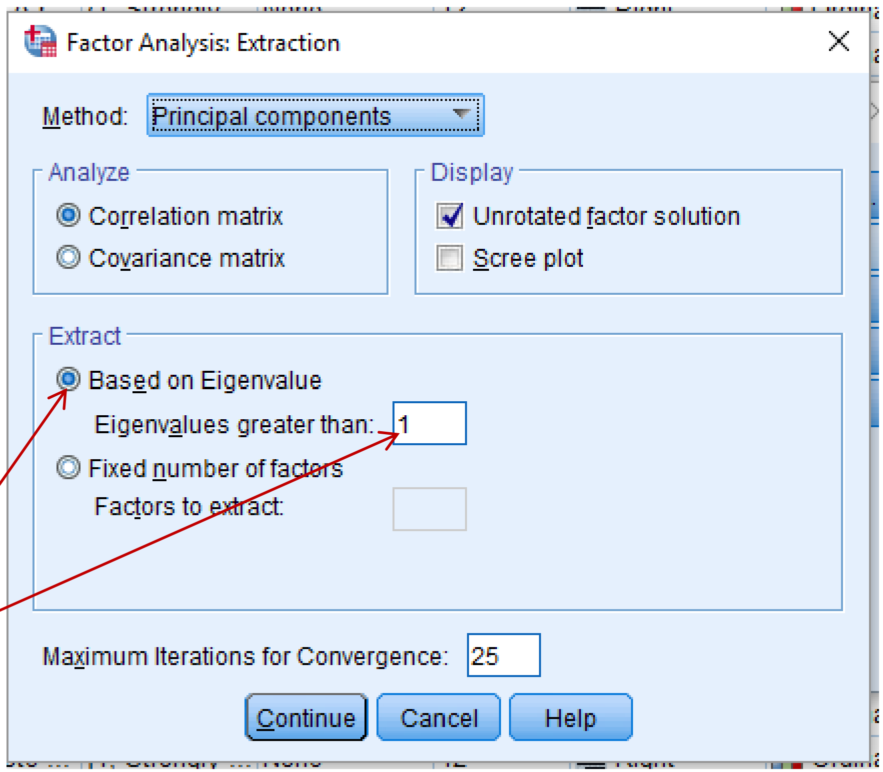


Principal components analysis



Principal components analysis

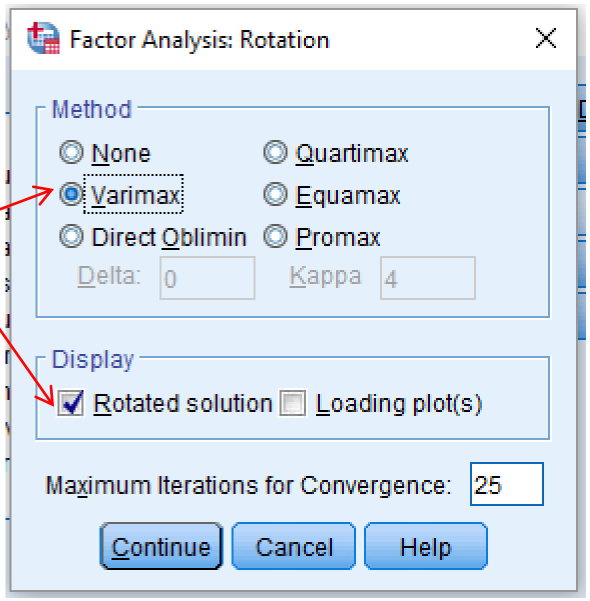




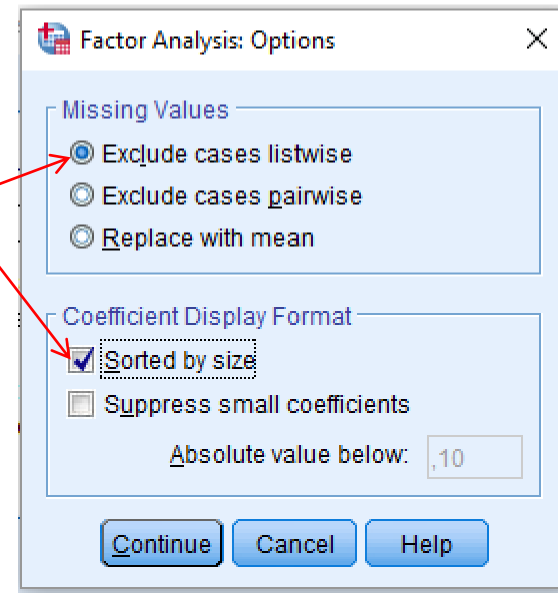
Check the Kaiser criterion (Eigenvalues > 1)

Principal components analysis

5

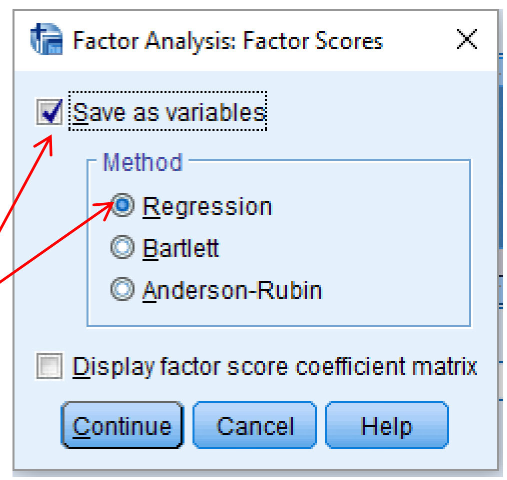


7



We can chose to save factor scores only after checking the PCA adequacy

6



Principal components analysis

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,824
Bartlett's Test of Sphericity	Approx. Chi-Square	1400,968
	df	55
	Sig.	,000

KMO
The KMO statistic value of 0,824 is meritorious, indicating a good adequacy of the PCA

Bartlett's test
The Bartlett's test of sphericity confirms the PCA overall good adequacy, χ^2 (55) = 1400,968; p<0,001

Communalities
The extracted factors should account for at least 30% of a variable's variance (if it was not the case, we could consider removing those variables)

Communalities

	Initial	Extraction
d1cost1mar:	1,000	,789
d2cost2prof	1,000	,841
d3vol1vol	1,000	,817
d4vol2age	1,000	,636
d5vol3reg	1,000	,707
d6qual1req	1,000	,788
d7qual2spe	1,000	,802
d8sec1dep	1,000	,611
d9sec2scop	1,000	,845
d10sec3flx	1,000	,830
d11sec4del	1,000	,593

Extraction Method: Principal Component Analysis.

Principal components analysis

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5,346	48,600	48,600	5,346	48,600	48,600	3,644	33,129	33,129
2	1,606	14,598	63,198	1,606	14,598	63,198	2,377	21,610	54,739
3	1,306	11,877	75,075	1,306	11,877	75,075	2,237	20,336	75,075
4	,681	6,191	81,266						
5	,536	4,869	86,135						
6	,395	3,588	89,723						
7	,319	2,897	92,620						
8	,269	2,445	95,066						
9	,248	2,253	97,318						
10	,188	1,707	99,025						
11	,107	,975	100,000						

Extraction Method: Principal Component Analysis.

Total variance explained (& Eigenvalues)
As requested, SPSS extracted factors with eigenvalue above 1... Accordingly 3 factors were extracted. This 3 extracted factors explain about 75% of the total variance.
If we wanted to extract a 4th factor (with an eigenvalue of 0,681) we would increase the total variance explained in 6% (if it were the case, in the menu “Extraction” we could select “number of factors: 4”

Principal components analysis

Presenting results… e.g. table

Rotated Component Matrix[a]

	Component 1	Component 2	Component 3
d3vol1vol	,879	,136	,161
d2cost2prof	,878	,135	,226
d1cost1mar	,861	,185	,117
d5vol3reg	,763	,207	,284
d4vol2age	,707	,370	-,002
d10sec3flx	,189	,885	,104
d9sec2scop	,226	,872	,179
d8sec1dep	,287	,650	,326
d7qual2spe	,142	,274	,841
d6qual1req	,286	-,045	,839
d11sec4del	,071	,287	,711

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Direct functions

Principal Component Analysis with Varimax rotation

(KMO=0,824; Bartlett's test: χ^2 (55) = 1400,968; p<0,001)

	Profit	Safeguard	Quality
d3vol1vol	**0,879**	0,136	0,161
d2cost2prof	**0,878**	0,135	0,226
d1cost1mar	**0,861**	0,185	0,117
d5vol3reg	**0,763**	0,207	0,284
d4vol2age	**0,707**	0,370	-0,002
d10sec3flx	0,189	**0,885**	0,104
d9sec2scop	0,226	**0,872**	0,179
d8sec1dep	0,287	**0,650**	0,326
d7qual2spe	0,142	0,274	**0,841**
d6qual1req	0,286	-0,045	**0,839**
d11sec4del	0,071	0,287	**0,711**
Explained variance (%)	33%	22%	20%

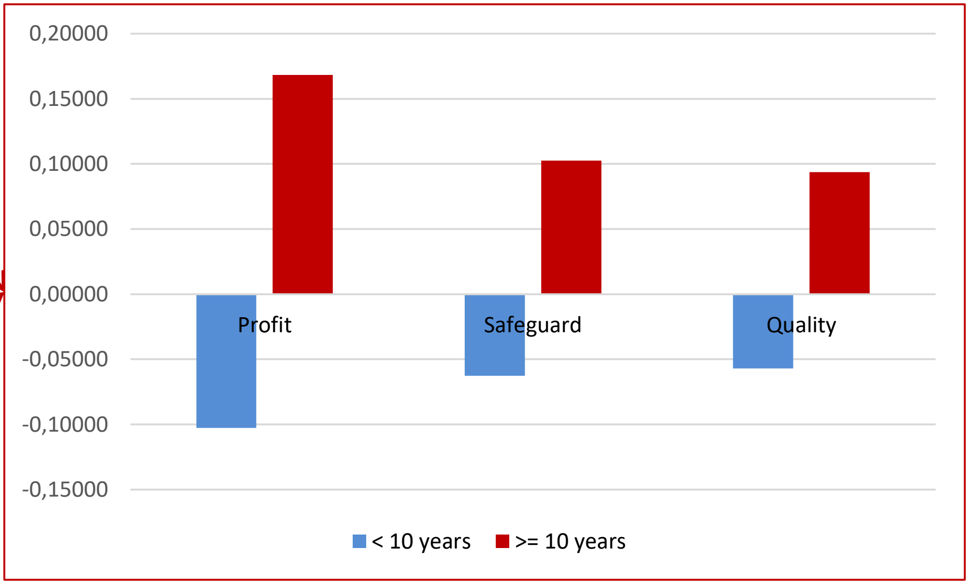
We can name the components, e.g. Profit (1), Safeguard (2) & Quality (3)

Principal components analysis

Custom Tables

Analyse > Tables > Custom tables

		REGR factor score 1 for analysis 1 Mean	REGR factor score 2 for analysis 1 Mean	REGR factor score 3 for analysis 1 Mean
relationship age with more (or less) than 10 years	< 10 years	-,10264	-,06254	-,05706
	>= 10 years	,16832	,10257	,09358



We can make several charts with Excel (such as this one) considering other variables

Cluster analysis

- Method for identifying homogenous groups of observations (called clusters)
  - Observations (or cases, objects) in a specific cluster share many characteristics, but are very dissimilar to objects not belonging to that cluster
  - Grouping similar customers and products is a fundamental marketing activity
    - The segmentation of customers is a standard application of cluster analysis

Cluster analysis

- Procedure in brief
  1. Decide on the characteristics that we will use to segment, for instance, costumers
     - Chose clustering variables to include in the analysis
  2. Select the clustering procedure
     - Hierarchical methods, partitioning methods (k-means), and two-step clustering
  3. Decide on the number of clusters
  4. Defining and labelling clusters

Cluster analysis

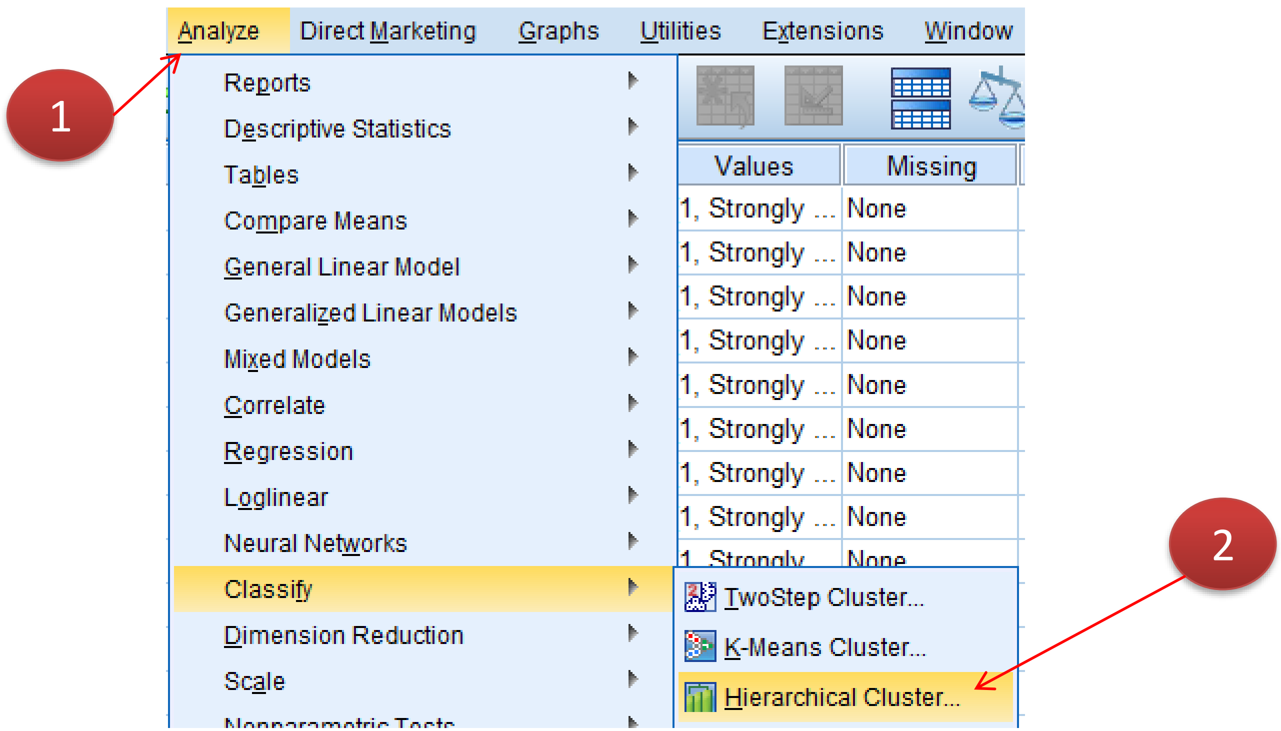
- Hierarchical methods
  - Most hierarchical techniques fall into a category called agglomerative clustering
    - Starts with each object representing an individual cluster then these are sequentially merged according to their similarity
    - Ward's method
      - Based on the sum-of-squares approach – combines objects whose merger increases the overall within-cluster variance to the smallest possible degree
      - Tends to produces equally sized clusters
      - Affected by outliers
      - Most commonly used method in marketing

Cluster analysis

- Decide on the number of clusters
  - One potential way is to use a chart
    - We plot the number of clusters on the x-axis (starting with the one-cluster solution at the very left) against the distance at which objects or clusters are combined on the y-axis (plot 30 differences between coefficients with the highest value)
    - We then search for the distinctive break (elbow)
      - SPSS does not produce this plot automatically – we have to use the distances provided by SPSS to draw a line chart by using a common spreadsheet program (e.g. MS Excel)
  - We can make use of the dendrogram
    - SPSS provides a dendrogram rescaling the distances to a range of 0–25

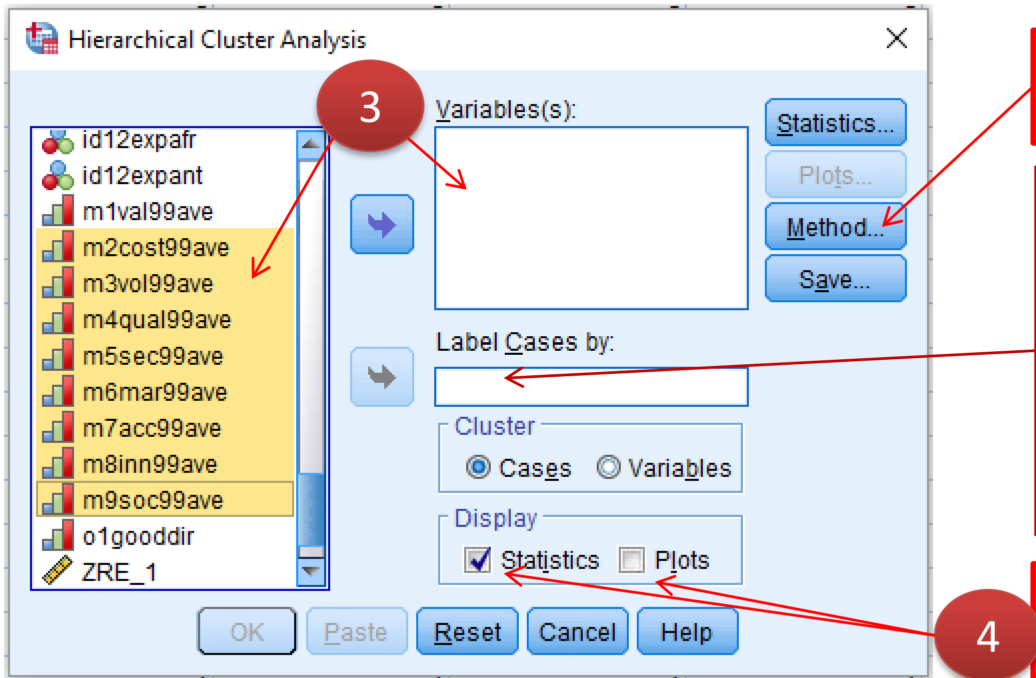
Cluster analysis

- Hierarchical method



Cluster analysis

- Hierarchical method



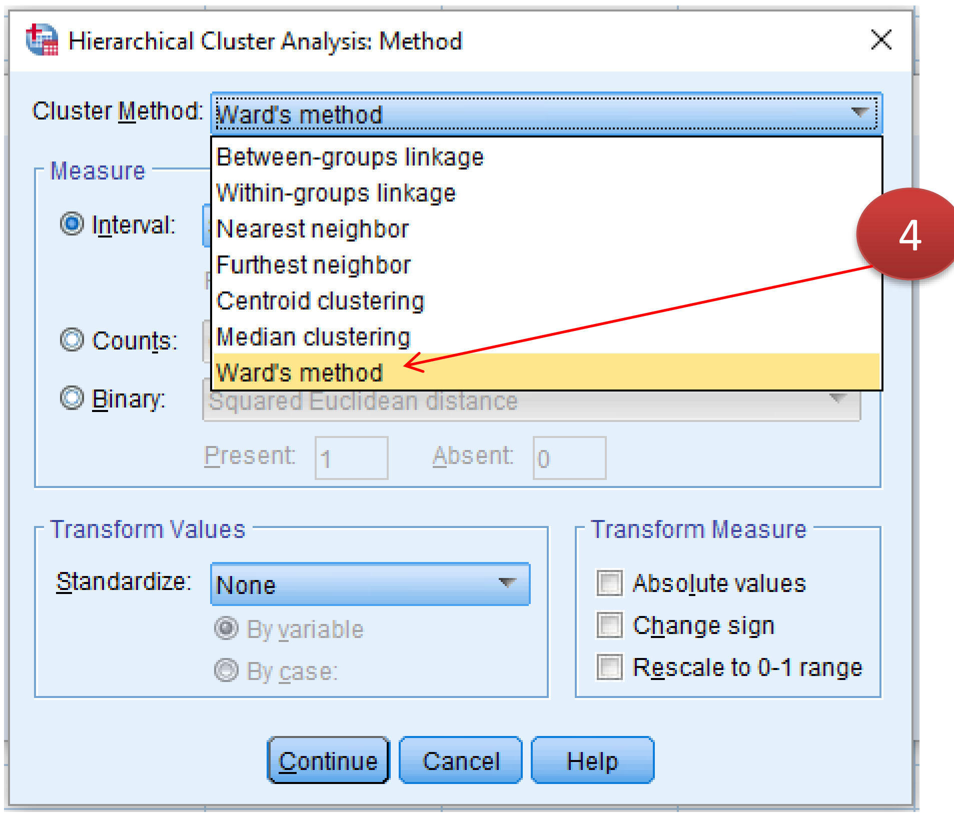
Next slides

For clustering aggregated data… we need to select the variable to identify cases (e.g. countries, brands, firms,…)
We can use the dendrogram to determine the number of clusters

We want the clustering coefficients, the dendrogram is expected to be excessively large for interpretation

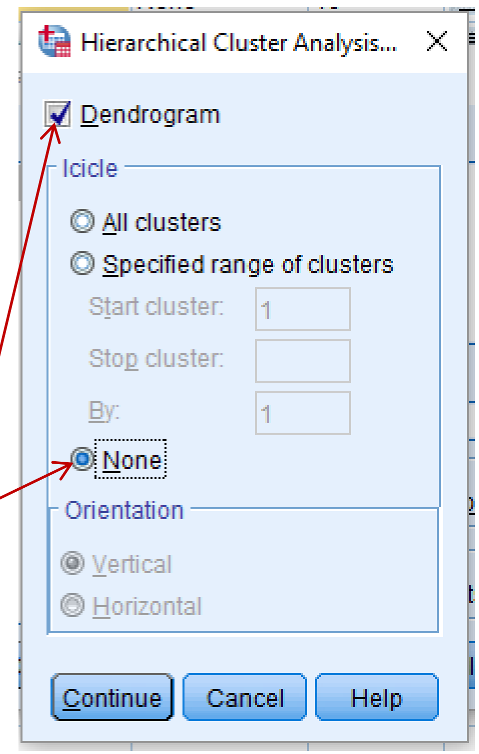
Cluster analysis

- Hierarchical method



We can select the ward's method... expecting clusters with similar sizes

For clustering aggregated data... to select the dendrogram



Cluster analysis

- Decide on the number of clusters

Agglomeration Schedule

Stage	Cluster Combined: Cluster 1	Cluster Combined: Cluster 2	Coefficients	Stage Cluster First Appears: Cluster 1	Stage Cluster First Appears: Cluster 2	Next Stage
1	169	198	,000	0	0	141
2	168	197	,000	0	0	143
3	166	195	,000	0	0	130
...	...	...	...	...	...	...
193	1	2	1194,340	187	184	194
194	1	5	1344,962	193	192	197
195	4	26	1510,102	191	159	196
196	4	14	1745,380	195	190	197
197	1	4	2945,913	194	196	194

SPSS output

Sort coefficients from largest to smallest

Assess 30 distances (differences between coefficients)
H4=C4-C5
H5=C6-C7
...

K21

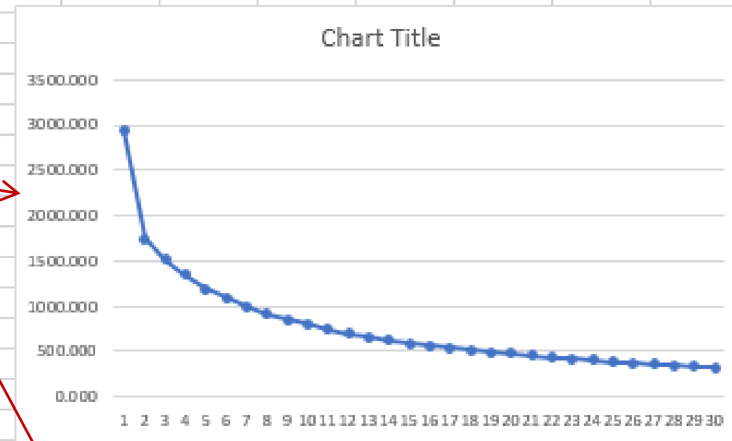
	A	B	C	D	E	F	G	H
1	Agglomeration Schedule							
2		Cluster Combined			Appears			
3	Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage	Differences betweer
4	197	1	4	2945.913	194	196	0	1200.53
5	196	4	14	1745.380	195	190	197	235.28
6	195	4	26	1510.102	191	159	196	165.14
7	194	1	5	1344.962	193	192	197	150.62
8	193	1	2	1194.340	187	184	194	99.13
9	192	5	23	1095.212	189	176	194	93.27
10	191	4	12	1001.937	188	186	195	84.02

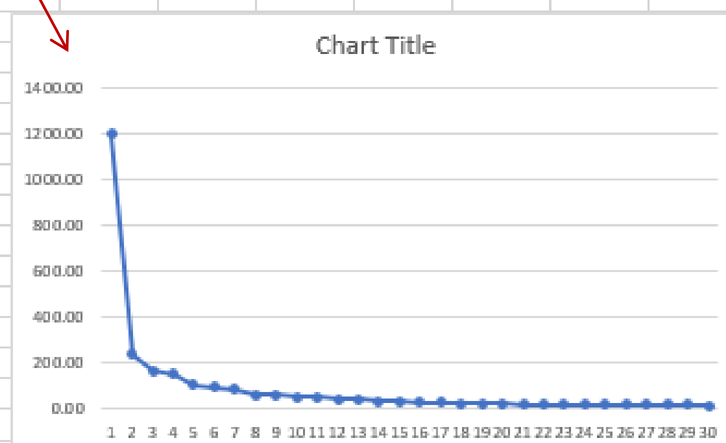
Excel sheet

Cluster analysis

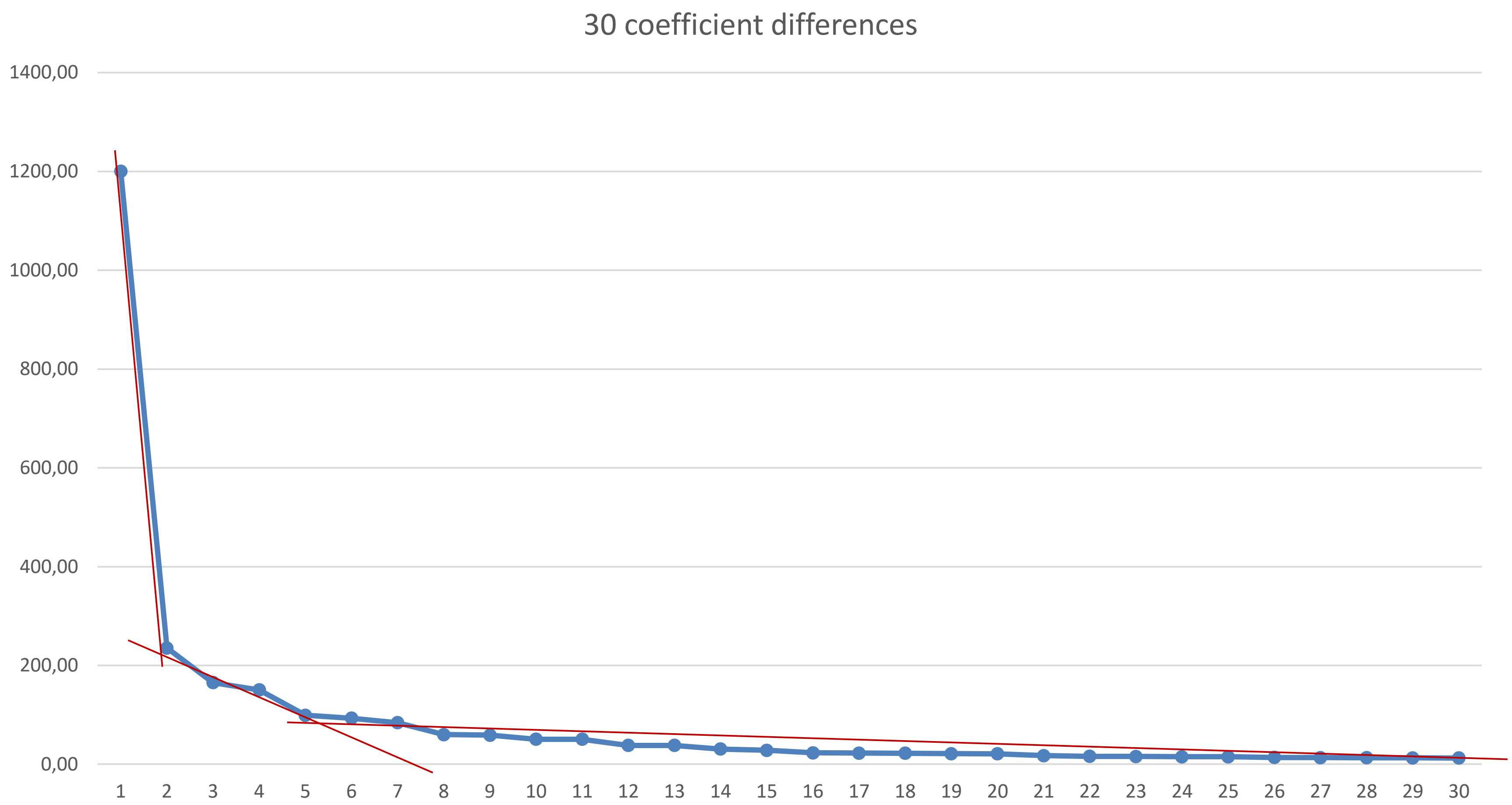
Agglomeration Schedule

Stage	Cluster Combined: Cluster 1	Cluster Combined: Cluster 2	Coefficients	Stage Cluster First Appears: Cluster 1	Stage Cluster First Appears: Cluster 2	Next Stage	Differences between coefficients
197	1	4	2945.913	194	196	0	1200.53
196	4	14	1745.380	195	190	197	235.28
195	4	26	1510.102	191	159	196	165.14
194	1	5	1344.962	193	192	197	150.62
193	1	2	1194.340	187	184	194	99.13
192	5	23	1095.212	189	176	194	93.27
191	4	12	1001.937	188	186	195	84.62
190	14	16	917.312	174	185	196	59.95
189	5	39	857.965	181	151	192	58.87
188	4	13	799.099	183	178	191	50.65
187	1	118	748.446	179	160	193	50.58
186	12	27	697.863	182	146	191	38.11
185	16	21	659.750	173	165	190	38.06
184	2	8	621.690	168	163	193	30.74
183	4	31	590.951	180	171	188	28.11
182	12	30	562.840	162	161	186	23.22
181	5	10	539.615	169	177	189	22.53
180	4	17	517.081	175	166	183	22.36
179	1	37	494.725	172	170	187	21.46
178	13	101	473.269	149	0	188	21.09
177	10	51	452.174	167	143	181	17.40
176	23	116	434.773	150	69	192	15.96
175	4	145	418.818	153	132	180	15.82
174	14	19	403.002	138	140	190	15.30
173	16	117	387.705	164	106	185	15.05
172	1	6	372.659	145	158	179	13.82
171	31	32	358.838	135	154	183	13.57
170	37	115	345.271	45	27	179	13.25
169	5	20	332.021	152	148	181	12.96
168	2	40	319.059	155	111	184	12.64
167	10	93	306.423	112	137	177	





Cluster analysis

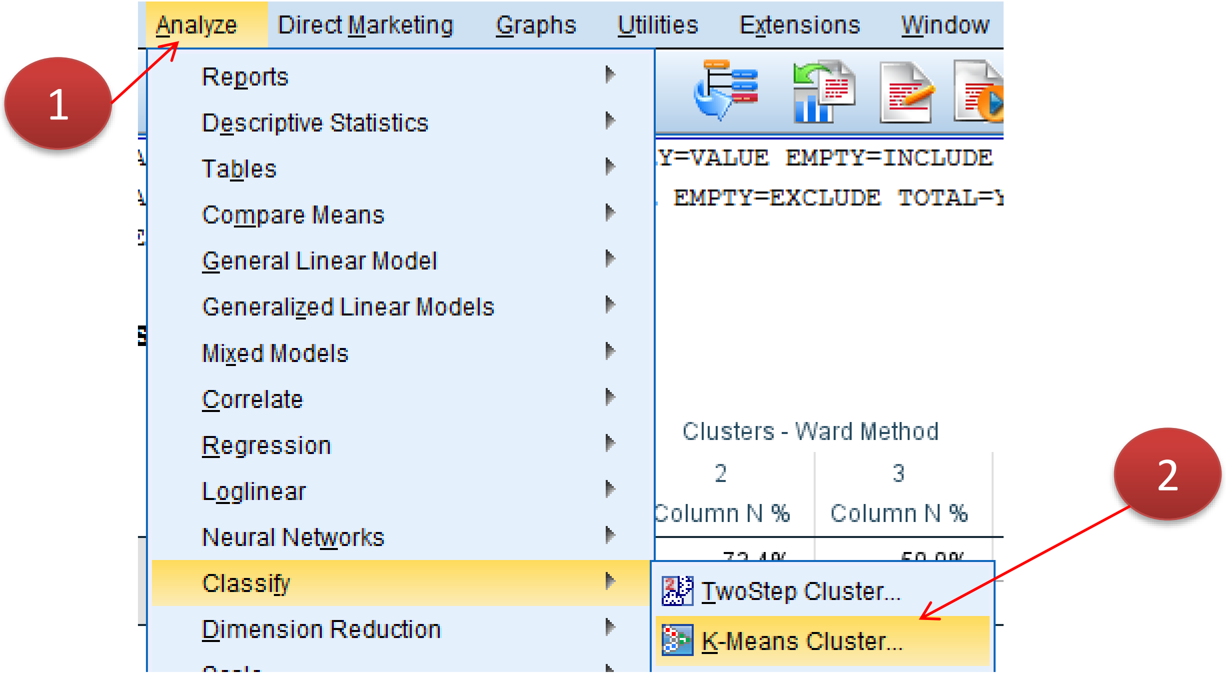


Cluster analysis

- K-Means procedure
  - This algorithm is not based on distance measures
  - Uses the within-cluster variation as a measure to form homogenous clusters
    - Aims at segmenting the data in such away that the within-cluster variation is minimized
  - K-means does not build a hierarchy
    - With the hierarchical methods, an object remains in a cluster once it is assigned to it, but with k-means, cluster affiliations can change in the course of the clustering process
  - K-means is (generally) superior to hierarchical methods
    - Is less affected by outliers and the presence of irrelevant clustering variables
    - Can be applied to very large datasets
    - Recommended for sample sizes above 500

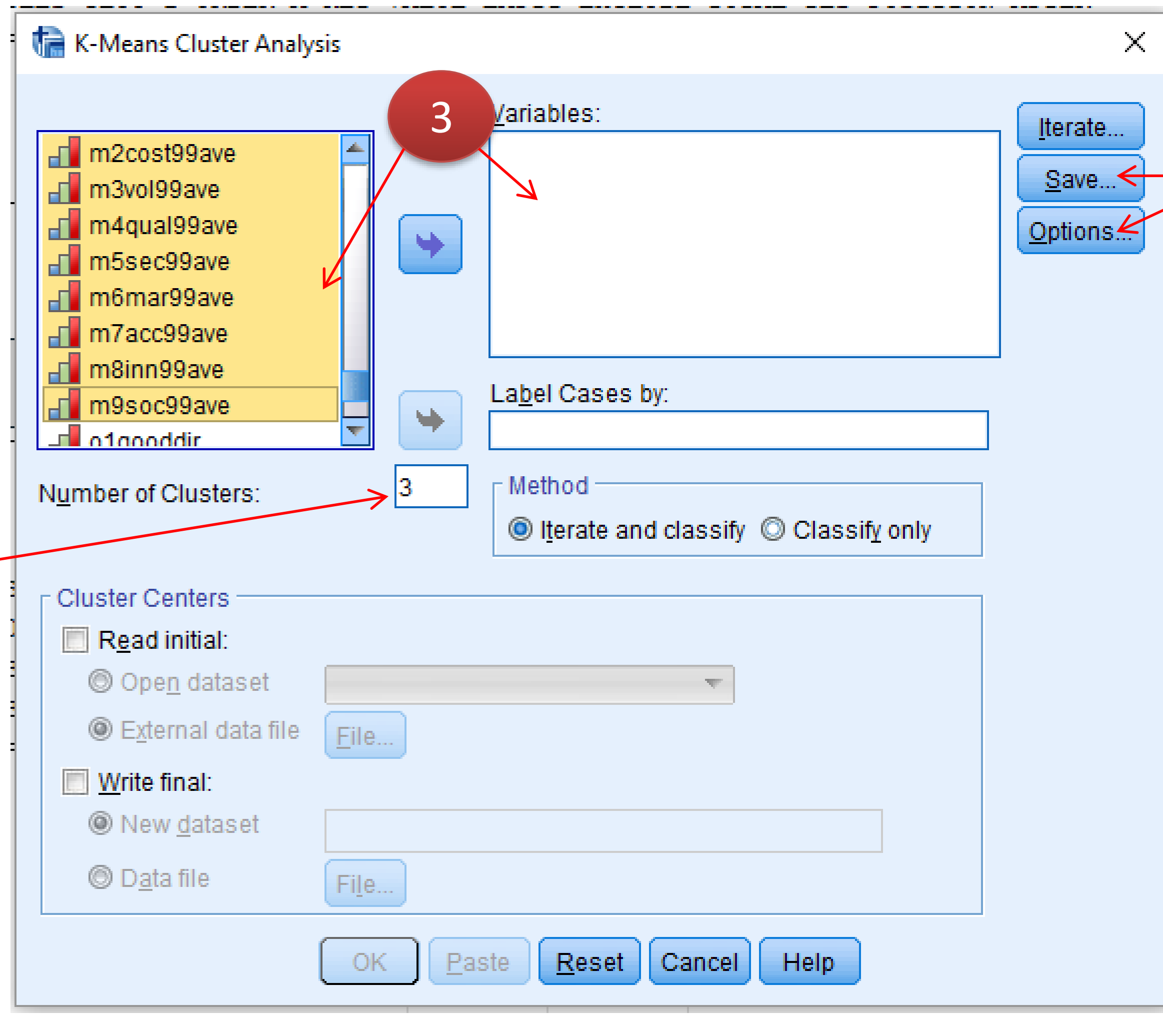
Cluster analysis

- K-Means



Cluster analysis

- K-Means



Next slides

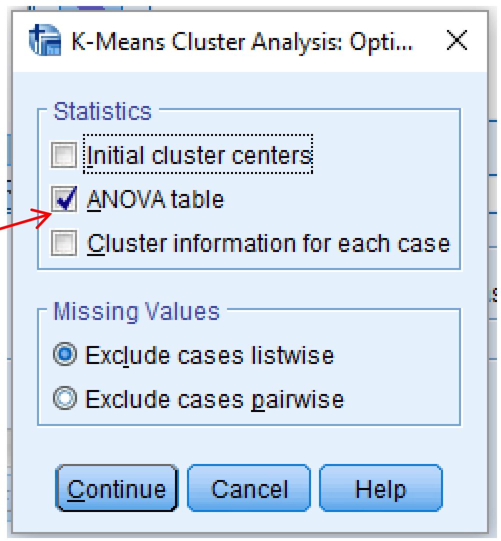
5

We need to select the number of clusters (3)… the number We got on the hierarchical method

Cluster analysis

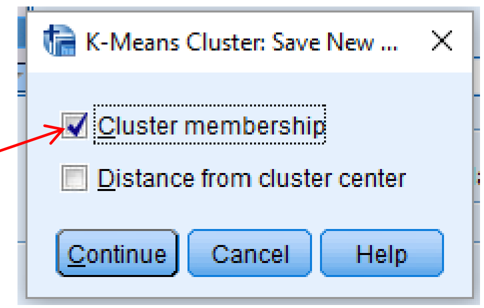
- K-Means

6



All clusters should have significant differences

7



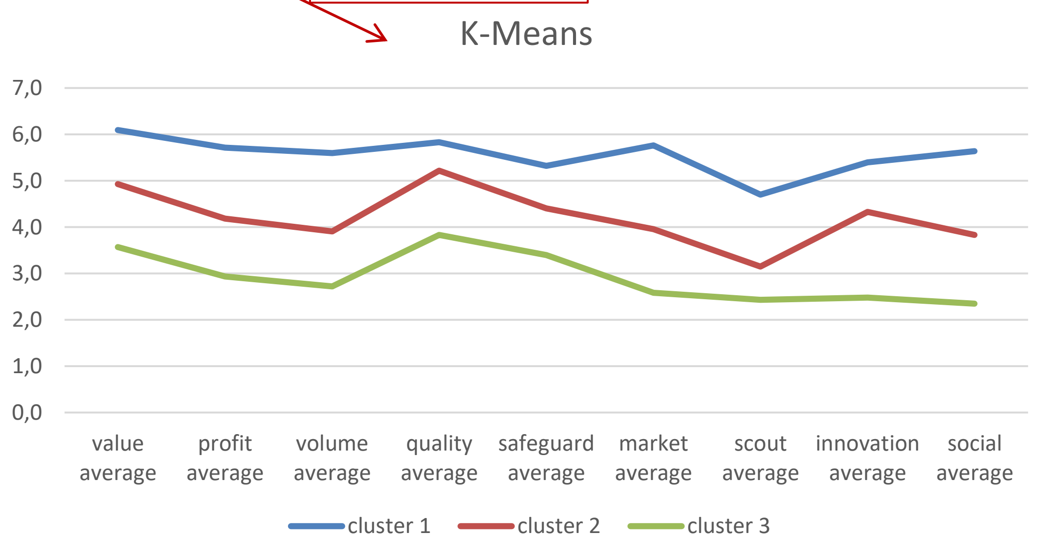
Cluster analysis

Custom Tables

	Cluster Number of Case			
	1	2	3	Total
	Mean	Mean	Mean	Mean
value average	6,1	4,9	3,6	5,2
profit average	5,7	4,2	2,9	4,7
volume average	5,6	3,9	2,7	4,5
quality average	5,8	5,2	3,8	5,3
safeguard average	5,3	4,4	3,4	4,7
market average	5,8	4,0	2,6	4,6
scout average	4,7	3,2	2,4	3,7
innovation average	5,4	4,3	2,5	4,5
social average	5,6	3,8	2,4	4,4

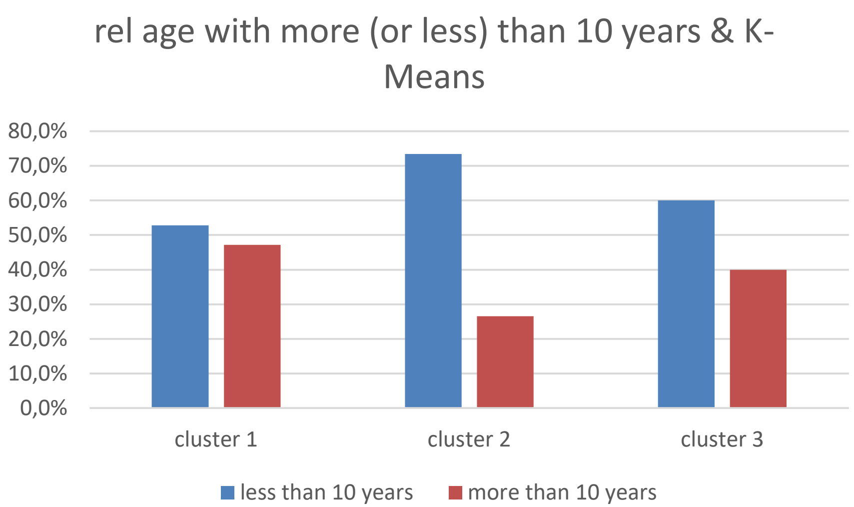
Custom tables results

Excel chart

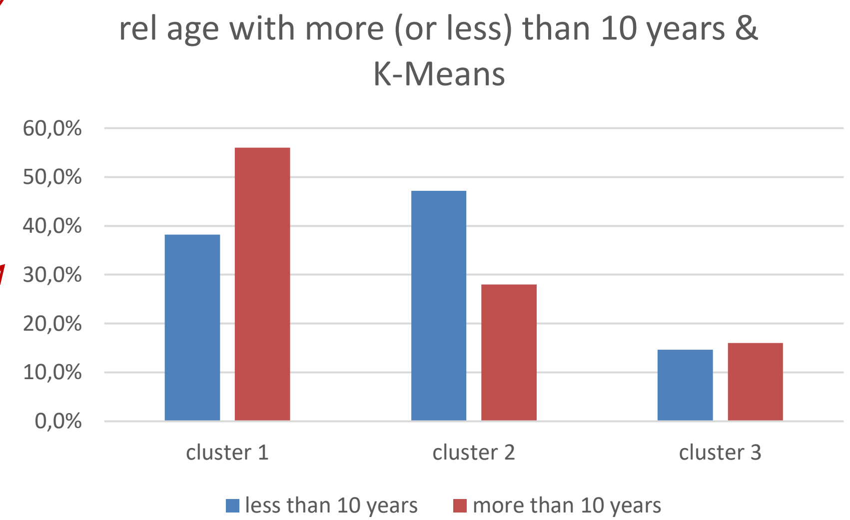


Cluster analysis

- K-Means



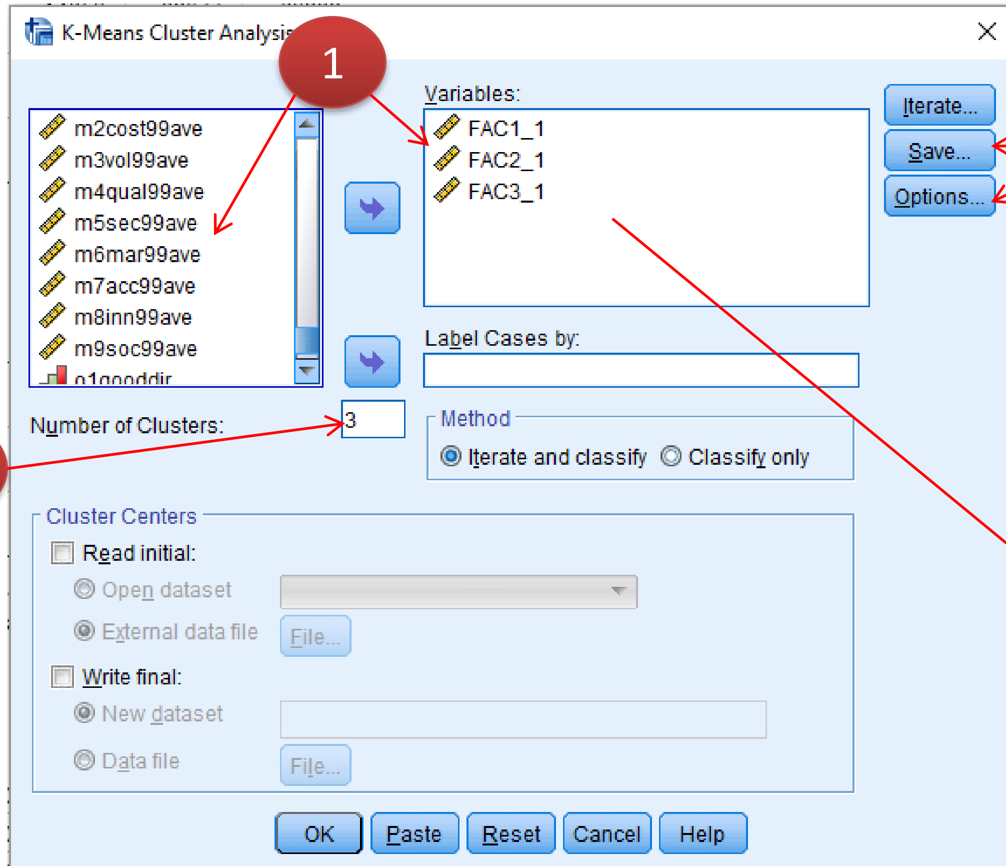
Column N%
(clusters (social) characterization)



Row N%
(cases distribution in clusters)

Cluster analysis & PCA

- K-Means & Principal component analysis



Next slides

We need to select the number of clusters (3)

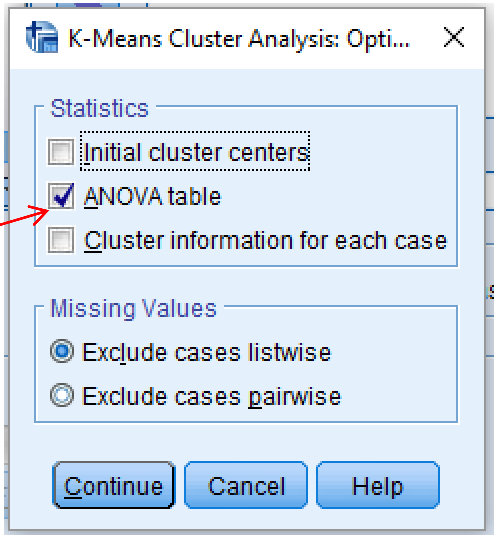
PCA components / factors: Profit (1), Safeguard (2) & Quality (3)

Cluster analysis & PCA

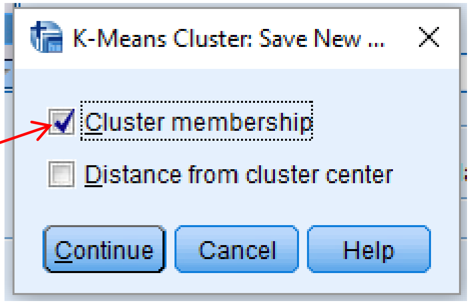
- K-Means & Principal component analysis

3

All clusters should have significant differences



4



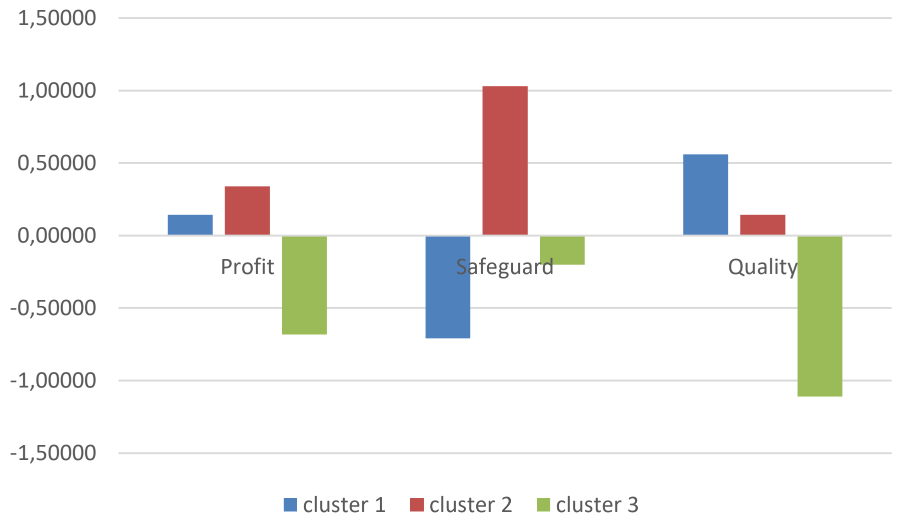
Cluster analysis & PCA

- K-Means & Principal component analysis

Cluster Number of Case

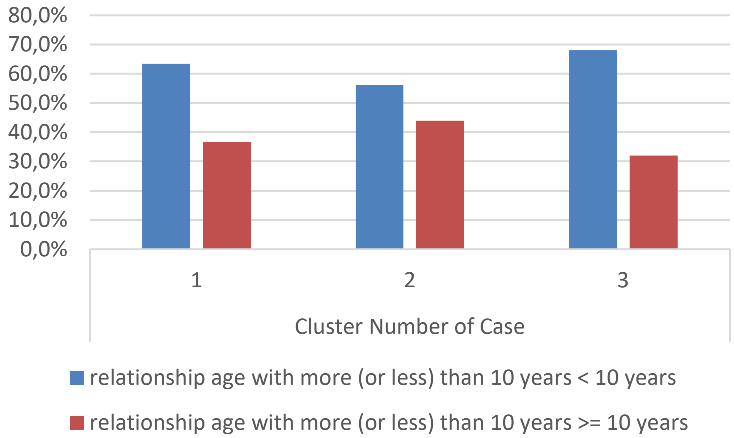
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	82	41.4	41.4	41.4
	2	66	33.3	33.3	74.7
	3	50	25.3	25.3	100.0
	Total	198	100.0	100.0	

		REGR factor score 1 for analysis 1	REGR factor score 2 for analysis 1	REGR factor score 3 for analysis 1
		Mean	Mean	Mean
Cluster Number of Case	1	.14258	-.70780	.56099
	2	.33977	1.03116	.14360
	3	-.68233	-.20035	-1.10958



Cluster analysis & PCA

- K-Means & Principal component analysis



		Cluster Number of Case		
		1	2	3
		Column N %	Column N %	Column N %
relationship age with more (or less) than 10 years	< 10 years	63.4%	56.1%	68.0%
	>= 10 years	36.6%	43.9%	32.0%

		Cluster Number of Case		
		1	2	3
		Row N %	Row N %	Row N %
relationship age with more (or less) than 10 years	< 10 years	42.3%	30.1%	27.6%
	>= 10 years	40.0%	38.7%	21.3%

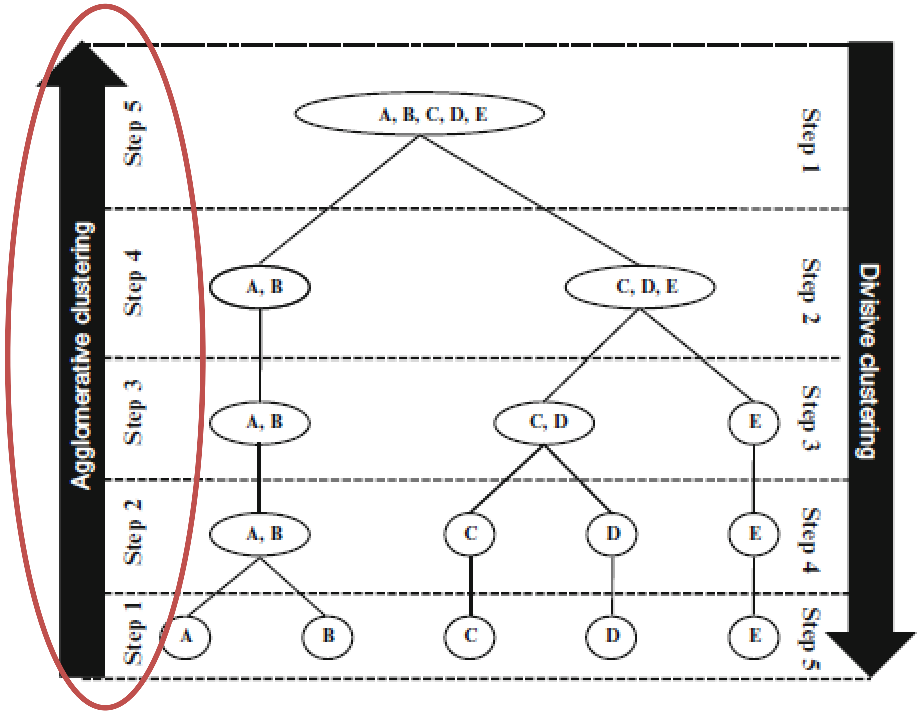
Annex

- Extension…
  - Cluster analysis

Cluster analysis

- Hierarchical methods
  - Most hierarchical techniques fall into a category called agglomerative clustering
    - Starts with each object representing an individual cluster then these are sequentially merged according to their similarity
  - A cluster hierarchy can also be generated top-down (divisive clustering)
    - Rarely used in market research
  - Are based on measures of similarity or dissimilarity
    - Euclidean distance, city-block distance, Chebychev distance, among others (Mooi, 2011, p.245)

Cluster analysis



Mooi (2011, p.244)

Fig. 9.3 Agglomerative and divisive clustering

Cluster analysis

- Hierarchical methods
  - Transform variables
    - If the variables under consideration are measured on different scales or levels we can standardize the data prior to the analysis (Mooi, 2011, p.247)
  - Agglomerative method / Clustering Algorithm
    - We need to select the Linkage algorithms, these can yield totally different results when used on the same dataset, as each has its specific properties

      (Mooi, 2011, p.250)

Cluster analysis

- Hierarchical methods
  - Clustering Algorithm
    - Between-groups linkage (Average linkage)
      - The distance between two clusters is defined as the average distance between all pairs of the two clusters' members
      - Produces clusters with low within-cluster variance and similar sizes
      - Affected by outliers (though not as much as complete linkage)
      - Good results with scattered data
    - Within-groups Linkage
      - Similar to between-groups linkage
      - Aims to produce clusters with low within-cluster variance

Cluster analysis

- Hierarchical methods
  - Clustering Algorithm
    - Nearest neighbour (Single linkage)
      - The distance between two clusters corresponds to the shortest distance between any two members in the two clusters
      - Tends to form one large cluster with the other clusters containing only one or few objects each
      - Tends to create chains of clusters (it helps in identifying outliers)
      - Not “greatly” affected by outliers – as these will be merged with the remaining objects in the last steps of the analysis
      - Works best with long chains of clusters, unsuitable when the clusters are not clearly separated
      - Considered the most versatile algorithm

Cluster analysis

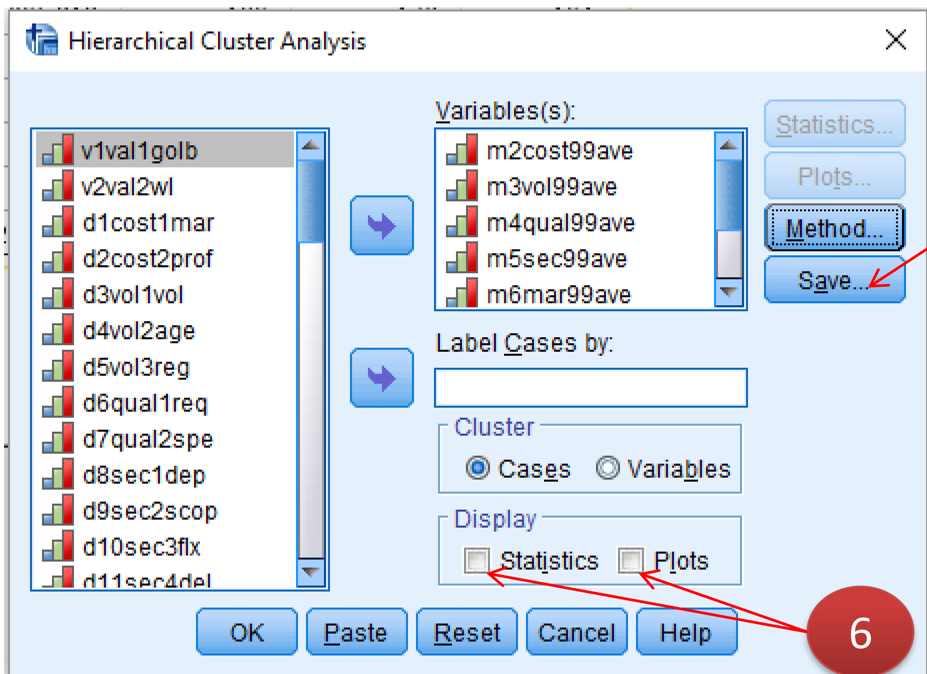
- Hierarchical methods
  - Clustering Algorithm
    - Furthest neighbour (complete linkage)
      - The oppositional approach to single linkage assumes that the distance between two clusters is based on the longest distance between any two members in the two clusters
      - Strongly affected by outliers
      - Works best with dense blobs of clusters
      - Produces rather compact and tightly clusters (of similar cases)
    - Centroid clustering
      - The geometric center (centroid) of each cluster is computed first
      - The distance between the two clusters equals the distance between the two centroids
      - Produces clusters with low within-cluster variance and similar sizes
      - Affected by outliers and by the (different) size of the clusters

Cluster analysis

- Hierarchical methods
  - Clustering Algorithm
    - Median Clustering
      - Very similar to centroid clustering, uses the median distance instead of the mean to determine the cluster centroid
      - This method takes into consideration the size of a cluster
    - Ward's method
      - Based on the sum-of-squares approach – combines objects whose merger increases the overall within-cluster variance to the smallest possible degree
      - Tends to produces equally sized clusters
      - Affected by outliers
      - Most commonly used method in marketing

Cluster analysis

- To continue using the hierarchical method after knowing the number of clusters



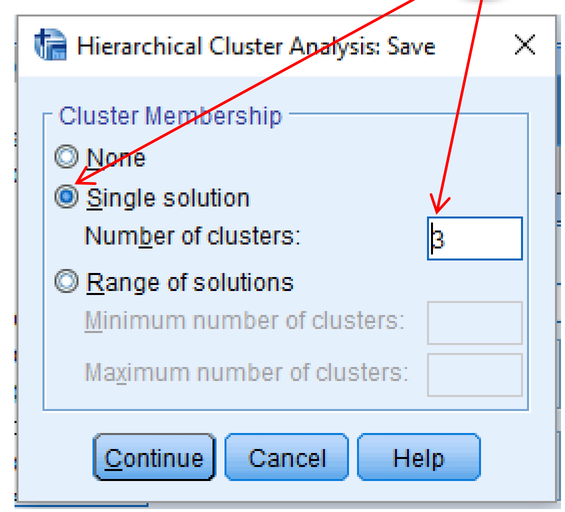
4

6

Now we don’t need the output

We want to save the 3 clusters variables

5



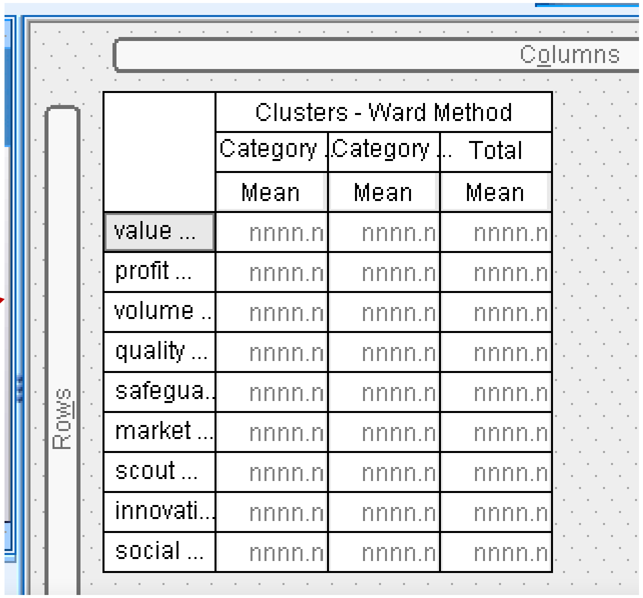
Cluster analysis

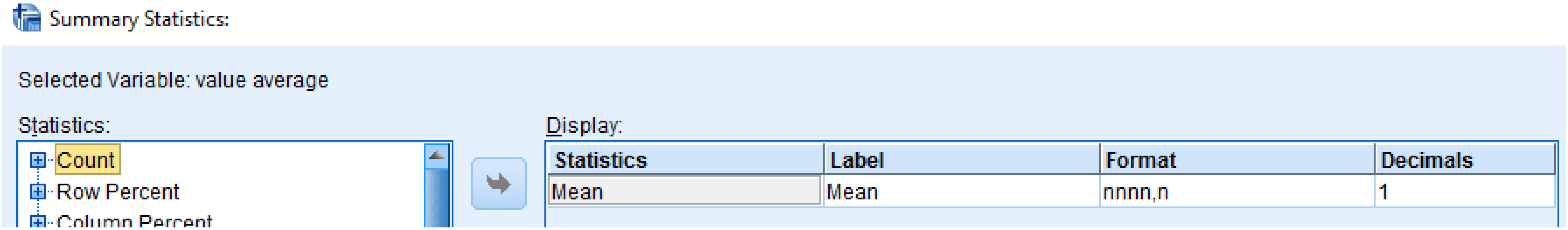
Ward Method

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	95	48,0	48,0	48,0
	2	79	39,9	39,9	87,9
	3	24	12,1	12,1	100,0
	Total	198	100,0	100,0	

"New variable" frequencies

Analyse > Tables > Custom Tables





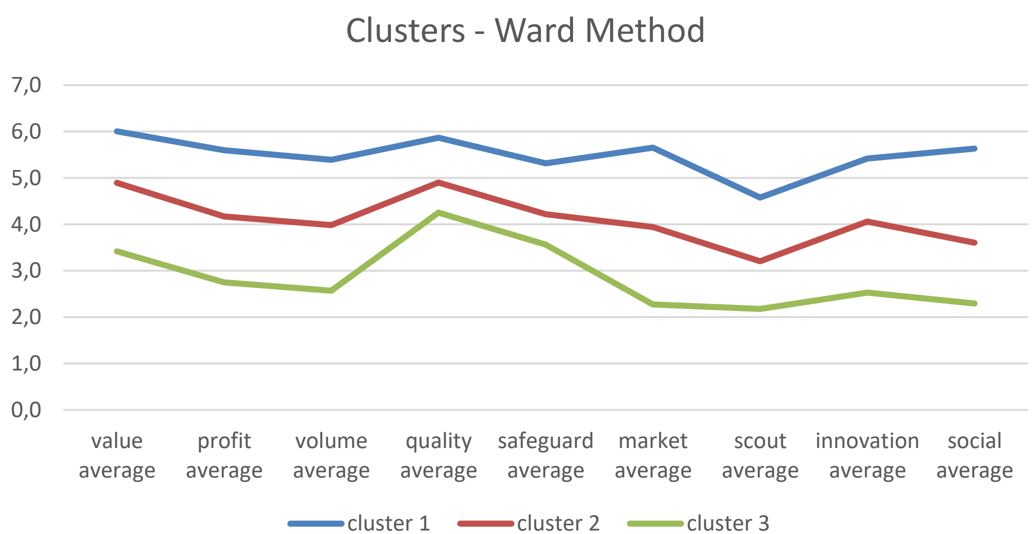
Cluster analysis

Custom Tables

	Clusters - Ward Method 1 Mean	2 Mean	3 Mean	Total Mean
value average	6,0	4,9	3,4	5,2
profit average	5,6	4,2	2,8	4,7
volume average	5,4	4,0	2,6	4,5
quality average	5,9	4,9	4,3	5,3
safeguard average	5,3	4,2	3,6	4,7
market average	5,7	3,9	2,3	4,6
scout average	4,6	3,2	2,2	3,7
innovation average	5,4	4,1	2,5	4,5
social average	5,6	3,6	2,3	4,4

Custom tables results

Excel chart



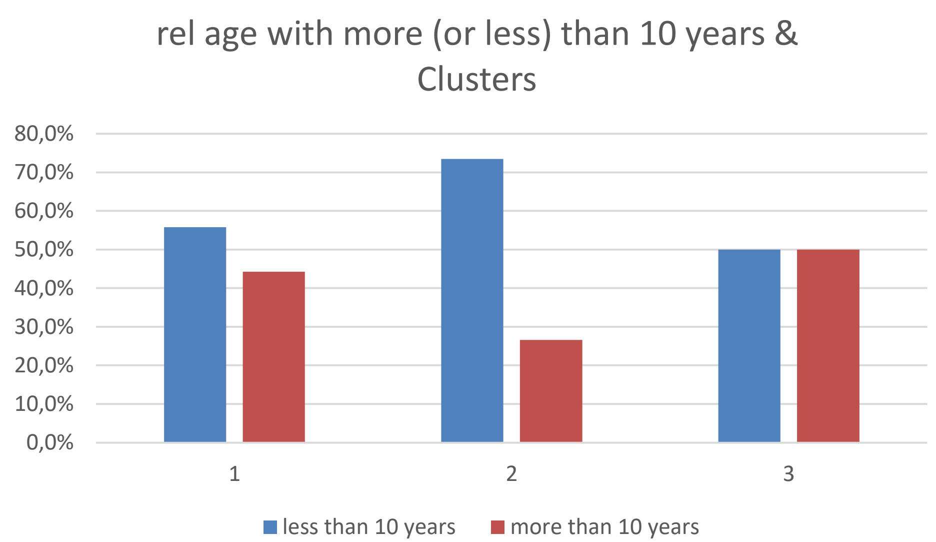
Cluster analysis

Custom Tables

		Clusters - Ward Method 1 Column N %	2 Column N %	3 Column N %	Total Column N %
rel age with more (or less) than 10 years	less than 10 years	55,8%	73,4%	50,0%	62,1%
	more than 10 years	44,2%	26,6%	50,0%	37,9%

Custom tables results

Excel chart



Clusters (social) characterization Column N%

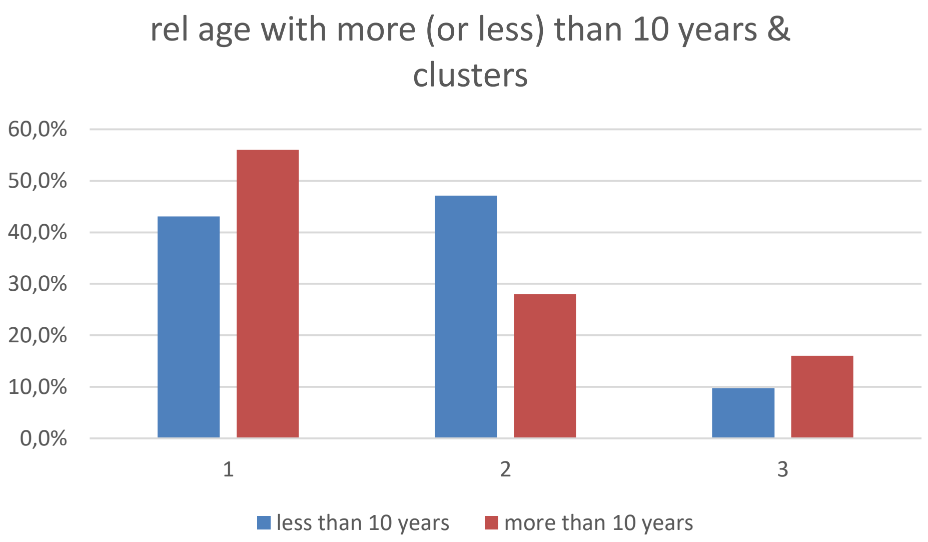
Cluster analysis

Custom Tables

		Clusters - Ward Method 1 Row N %	2 Row N %	3 Row N %	Total Row N %
rel age with more (or less) than 10 years	less than 10 years	43,1%	47,2%	9,8%	100,0%
	more than 10 years	56,0%	28,0%	16,0%	100,0%

Custom tables results

Excel chart



Cases distribution in clusters
Row N%