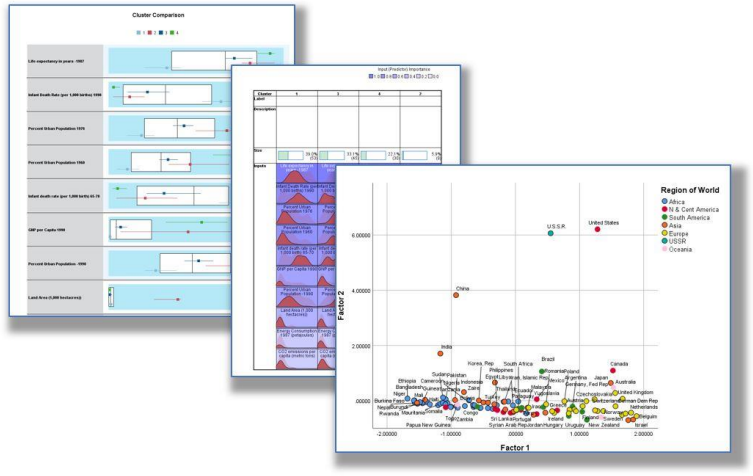
SMARTVISION Europe

Training Webinar



Factor and Cluster Analysis with IBM SPSS Statistics

Jarlath Quinn

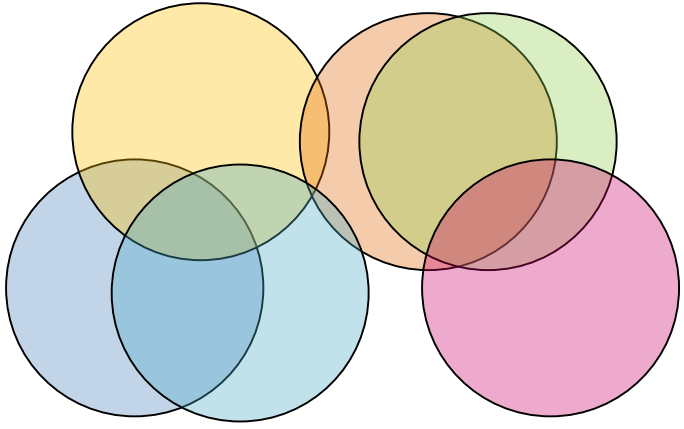
Materials

- All demonstrations are documented in these slides with accompanying SPSS data files
- If you have any follow-up questions, please email us at: **info@sv-europe.com**

Agenda

- Introducing Factor Analysis
- Exploring Correlations
- Performing Principal Component Analysis
- Rotated Solutions
- Performing Factor Analysis
- Analysing Component Scores
- Introducing Cluster Analysis
- Comparing Cluster Methods
- Performing Cluster Analysis
- Interpreting Output
- Creating Cluster Groupings

SMARTVISION Europe



Introducing Factor Analysis

Introducing Factor Analysis

- Factor Analysis is a statistical approach known as ‘data reduction’.
- At the heart of this technique, is the idea that if two or more variables correlate strongly with one another, then they may be measuring the same underlying ‘factor’
- It works by identifying these correlations between groups of variables with the aim of distilling the underlying ‘factors’ that account for their variation
- The procedure can then turn these selected factors into variables that show where each individual data record lies on the factor scale

Introducing Factor Analysis

- Examples of Factor Analysis include:
- Analysis that attempts to uncover the underlying ‘themes’ in a dataset: for example, what topics tend to co-occur when someone is asked about customer service?
- Researchers designing a personality test using a questionnaire based on rating scales. The procedure then produces factor scores that show how extroverted or introverted each respondent is.
- Data Scientists use a form of Factor Analysis (PCA) to reduce the number of variables used in building a predictive model, by combining groups of variables together to create linear combinations of them.

- Variables in a dataset are like cocktails
- They are distinct yet they *share common ingredients*
- Therefore if two cocktails are similar, it's because they share the same components or characteristics

SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

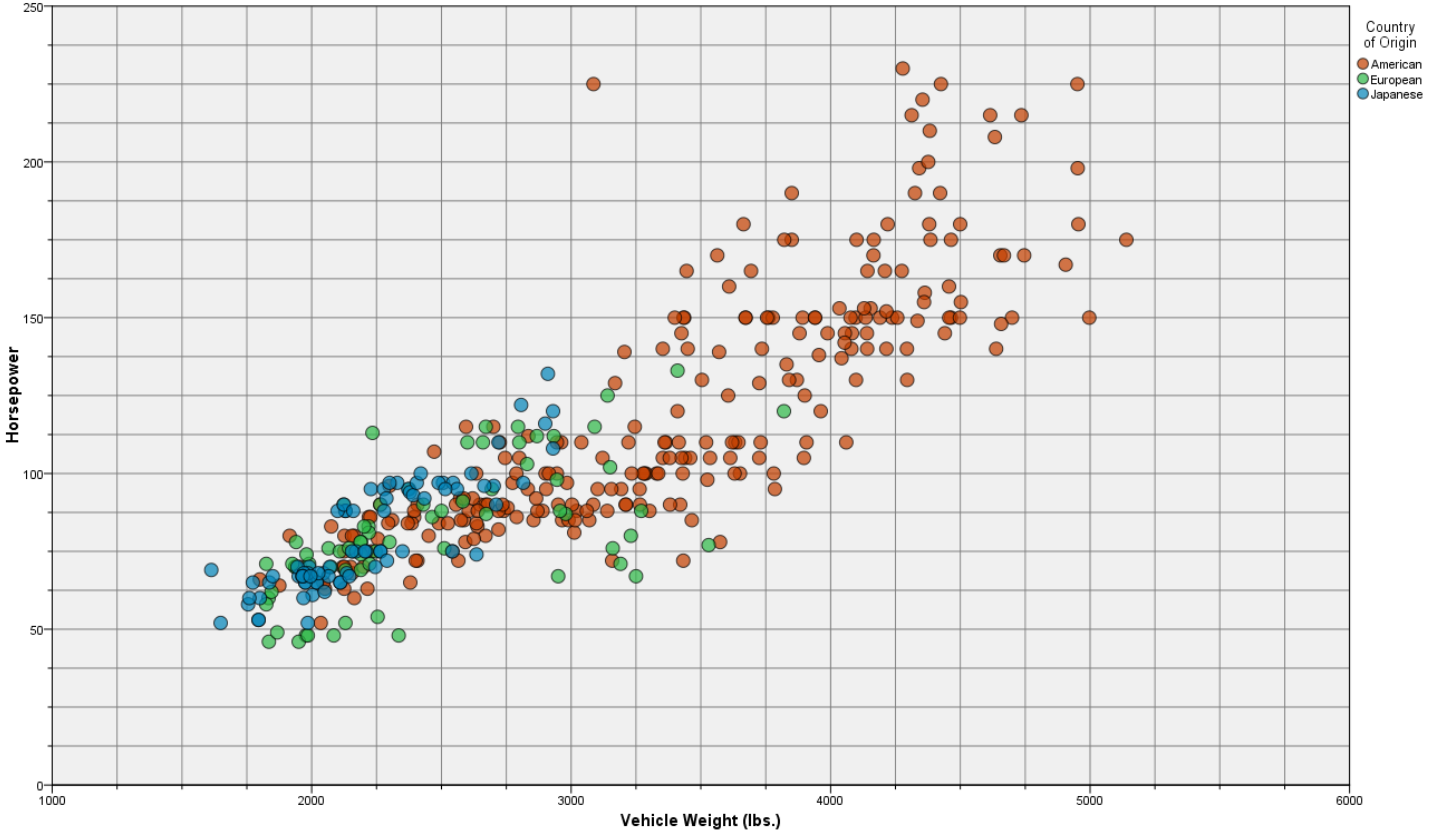
- Factor Analysis allows us to separate out the components of the variables that correlate with each other and combine them together to create new variable ‘factors’

SMARTVISION Europe

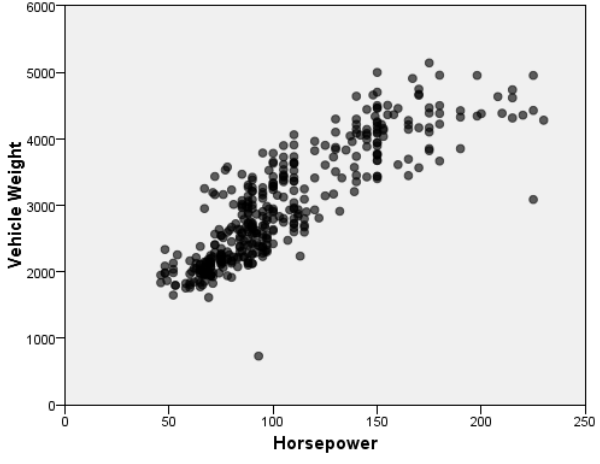
A SELECT INTERNATIONAL COMPANY

Scatterplot of different cars showing horsepower by weight

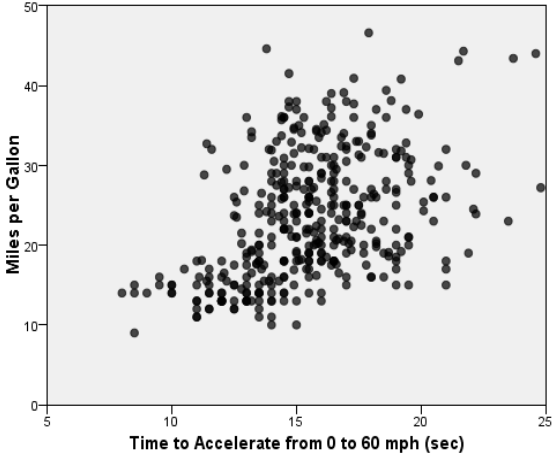
- Is there an *underlying factor* that explains the relationship in between the cars in this scatterplot?
- How can we measure the strength of this relationship?



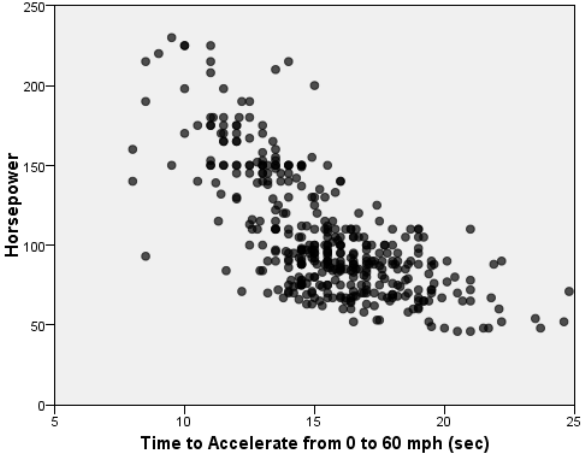
Correlations measure the strength of linear relationships



0.859



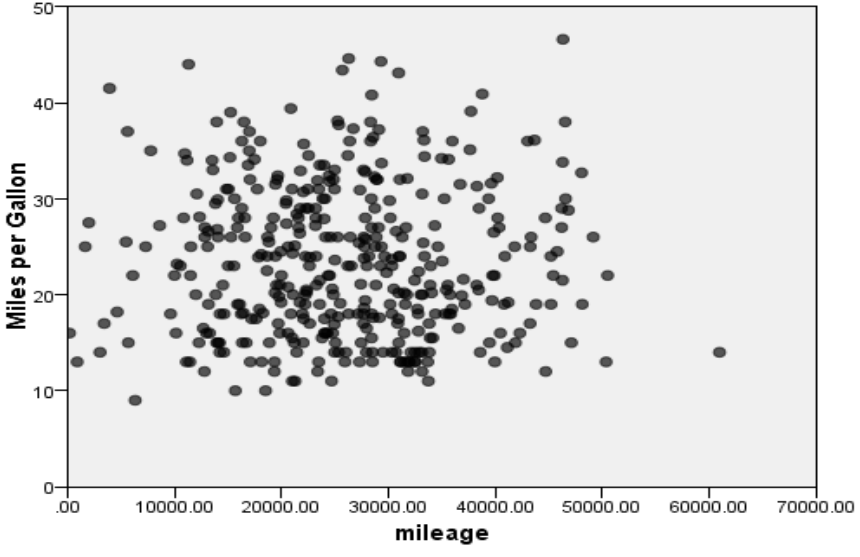
0.434

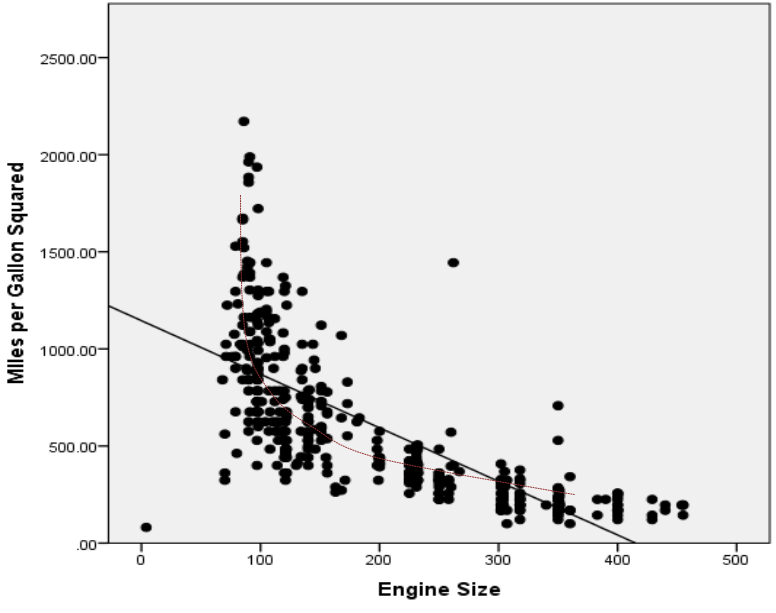


-.701

Pearson Correlation Values

Non-linear relationships are not accounted for





0.005

-.671

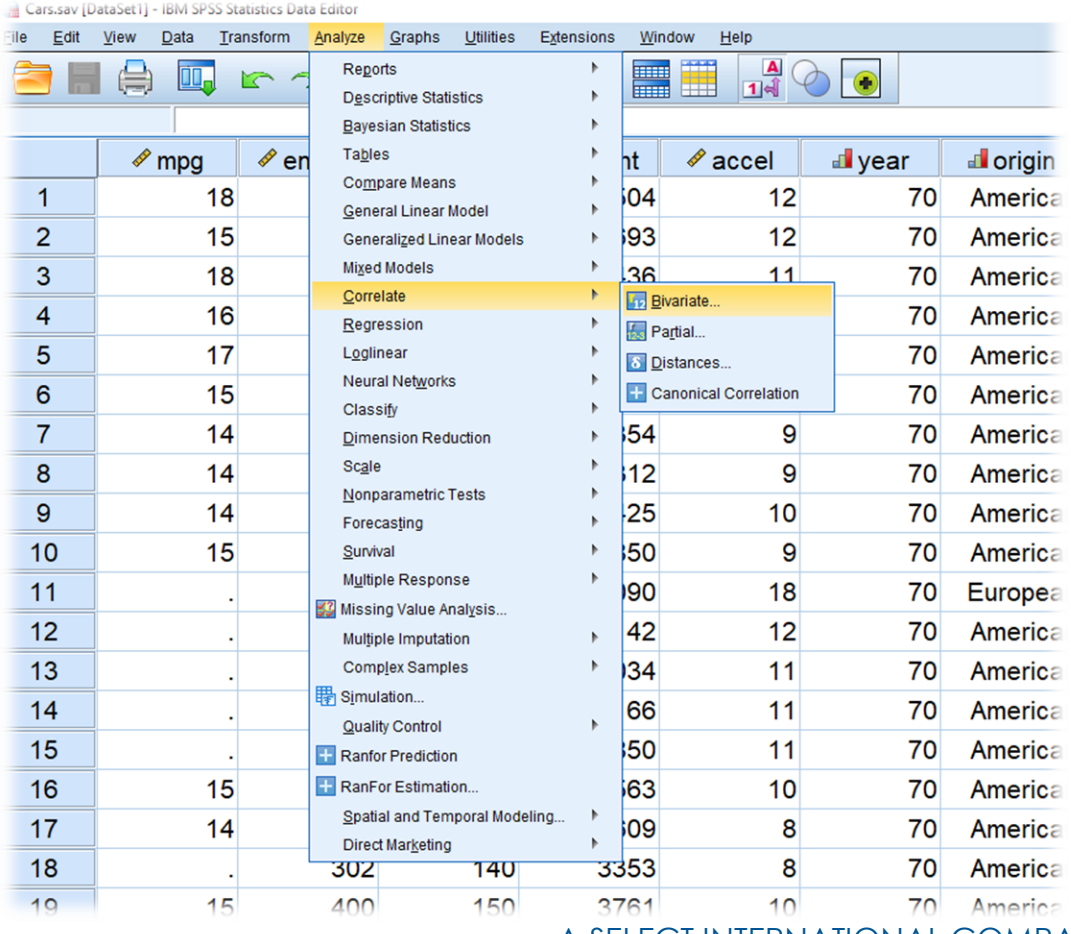
Pearson Correlation Values

SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

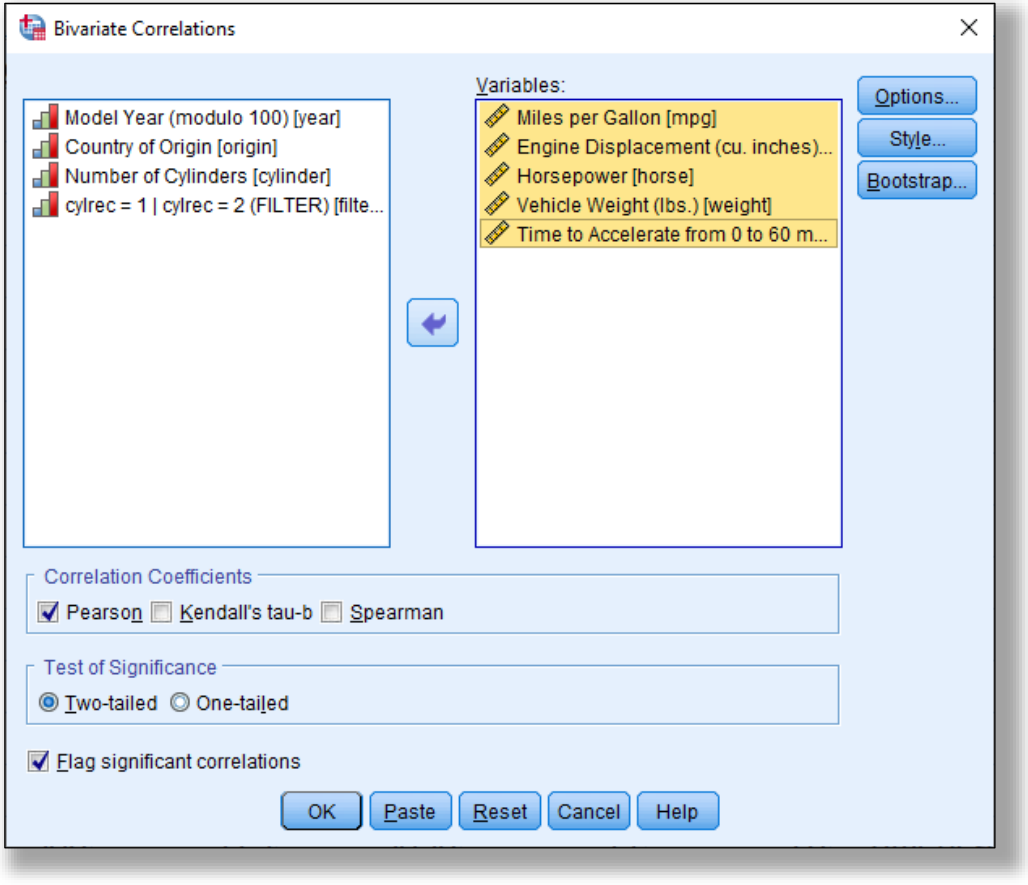
Requesting Correlations

- Using the **Cars.sav** dataset we can request bivariate correlations.
- From the main menu, click:
- **Analyze**
  - **Correlate**
    - **Bivariate**



Requesting Correlations

- Selecting only the continuous (scale) variables, we can request Pearson's bivariate correlations
- This will result in a correlation matrix showing the correlations between each combination of the variables
- Click:
  - **OK**



Requesting Correlations

Correlations

		Miles per Gallon	Engine Displacement (cu. inches)	Horsepower	Vehicle Weight (lbs.)	Time to Accelerate from 0 to 60 mph (sec)
Miles per Gallon	Pearson Correlation	1	-.789**	-.771**	-.807**	.434**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	398	398	392	398	398
Engine Displacement (cu. inches)	Pearson Correlation	-.789**	1	.897**	.933**	-.545**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	398	406	400	406	406
Horsepower	Pearson Correlation	-.771**	.897**	1	.859**	-.701**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	392	400	400	400	400
Vehicle Weight (lbs.)	Pearson Correlation	-.807**	.933**	.859**	1	-.415**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	398	406	400	406	406
Time to Accelerate from 0 to 60 mph (sec)	Pearson Correlation	.434**	-.545**	-.701**	-.415**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	398	406	400	406	406

**. Correlation is significant at the 0.01 level (2-tailed).

Requesting Correlations

Correlations

		Miles per Gallon	Engine Displacement (cu. inches)	Horsepower	Vehicle Weight (lbs.)	Time to Accelerate from 0 to 60 mph (sec)
Miles per Gallon	Pearson Correlation	1	-.789**	-.771**	-.807**	.434**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	398	398	392	398	398
Engine Displacement (cu. inches)	Pearson Correlation	-.789**	1	.897**	.933**	-.545**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	398	406	400	406	406
Horsepower	Pearson Correlation	-.771**	.897**	1	.859**	-.701**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	392	400	400	400	400
Vehicle Weight (lbs.)	Pearson Correlation	-.807**	.933**	.859**	1	-.415**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	398	406	400	406	406
Time to Accelerate from 0 to 60 mph (sec)	Pearson Correlation	.434**	-.545**	-.701**	-.415**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	398	406	400	406	406

**. Correlation is significant at the 0.01 level (2-tailed).

Every value is shown twice as the half of the matrix is a mirror of the other part

Correlation values are highlighted in yellow

Requesting Correlations

Correlations

		Miles per Gallon	Engine Displacement (cu. inches)	Horsepower	Vehicle Weight (lbs.)	Time to Accelerate from 0 to 60 mph (sec)
Miles per Gallon	Pearson Correlation	1	-.789**	-.771**	-.807**	.434**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	398	398	392	398	398
Engine Displacement (cu. inches)	Pearson Correlation	-.789**	1	.897**	.933**	-.545**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	398	406	400	406	406
Horsepower	Pearson Correlation	-.771**	.897**	1	.859**	-.701**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	392	400	400	400	400
Vehicle Weight (lbs.)	Pearson Correlation	-.807**	.933**	.859**	1	-.415**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	398	406	400	406	406
Time to Accelerate from 0 to 60 mph (sec)	Pearson Correlation	.434**	-.545**	-.701**	-.415**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	398	406	400	406	406

**. Correlation is significant at the 0.01 level (2-tailed).

There a number of very strong correlations in this table…

…and even a correlation as low as - 0.415 is fairly notable.

Requesting Correlations

Correlations

		Miles per Gallon	Engine Displacement (cu. inches)	Horsepower	Vehicle Weight (lbs.)	Time to Accelerate from 0 to 60 mph (sec)
Miles per Gallon	Pearson Correlation	1	-.789**	-.771**	-.807**	.434**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	398	398	392	398	398
Engine Displacement (cu. inches)	Pearson Correlation	-.789**	1	.897**	.933**	-.545**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	398	406	400	406	406
Horsepower	Pearson Correlation	-.771**	.897**	1	.859**	-.701**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	392	400	400	400	400
Vehicle Weight (lbs.)	Pearson Correlation	-.807**	.933**	.859**	1	-.415**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	398	406	400	406	406
Time to Accelerate from 0 to 60 mph (sec)	Pearson Correlation	.434**	-.545**	-.701**	-.415**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	398	406	400	406	406

**. Correlation is significant at the 0.01 level (2-tailed).

Horsepower, Vehicle Weight and Engine Size all correlate very strongly with one another….

…would they form their own factor?

Introducing Factor Analysis

- So Factor Analysis produces a solution showing which variables strongly correlate with each factor. The job of the analyst is try to interpret the strongest factors in a meaningful way.
- But Factor Analysis also shows, in descending order, how much variation each factor can explain.
- Imagine, you have a dataset with 10 fields. You want to reduce the data to uncover the primary factors that explain the variation in the file.
- The analysis shows the following result.
  - 100% of the variation can be explained by creating 10 factors
  - But, 70% of the variation can be explained using only the top 4 factors
  - And 45 % of the variation can be explained by the top 2 factors
- Which factor solution is the most appropriate?

Introducing Factor Analysis

- Most analysts think a good solution with Factor Analysis is one that meets two main criteria:
  - Firstly, a relatively, small number of factors can explain a relatively large proportion of variation*
  - Secondly, the factors are easy to interpret and make sense, because we can see *why* the underlying variables correlate with each other

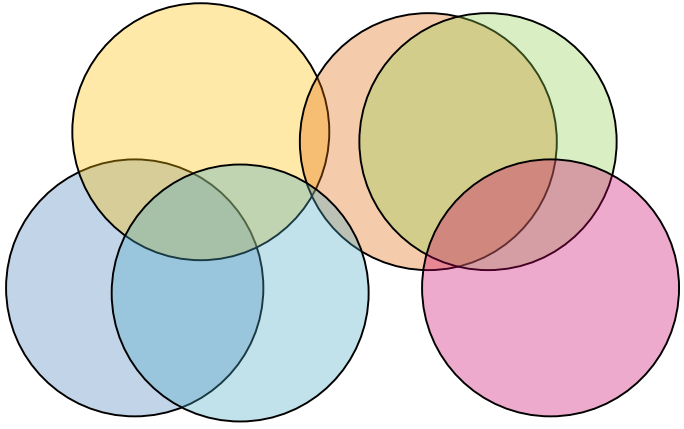
* so 4 factors explaining 70% variation is pretty good

Introducing Principal Components Analysis

- Before we look at *a true* Factor Analysis example, we will introduce a commonly-used, but related method: **Principal Components Analysis** (or PCA)
- The aim of PCA is to reduce the number of variables by creating linear combinations of them to form their *principal components*
- The resulting principal components when expressed as variables *have no correlation* with one another (they are *orthogonal components*)
- In a lot of statistical analyses, such as regression and clustering, using highly correlated variables together can cause problems, as you are entering a component that measures the same thing multiple times
- Therefore, PCA is often used because of its practical benefits: i.e. when analysts need to work with fewer, uncorrelated fields without sacrificing the variation in a dataset

* so 4 factors explaining 70% variation is pretty good

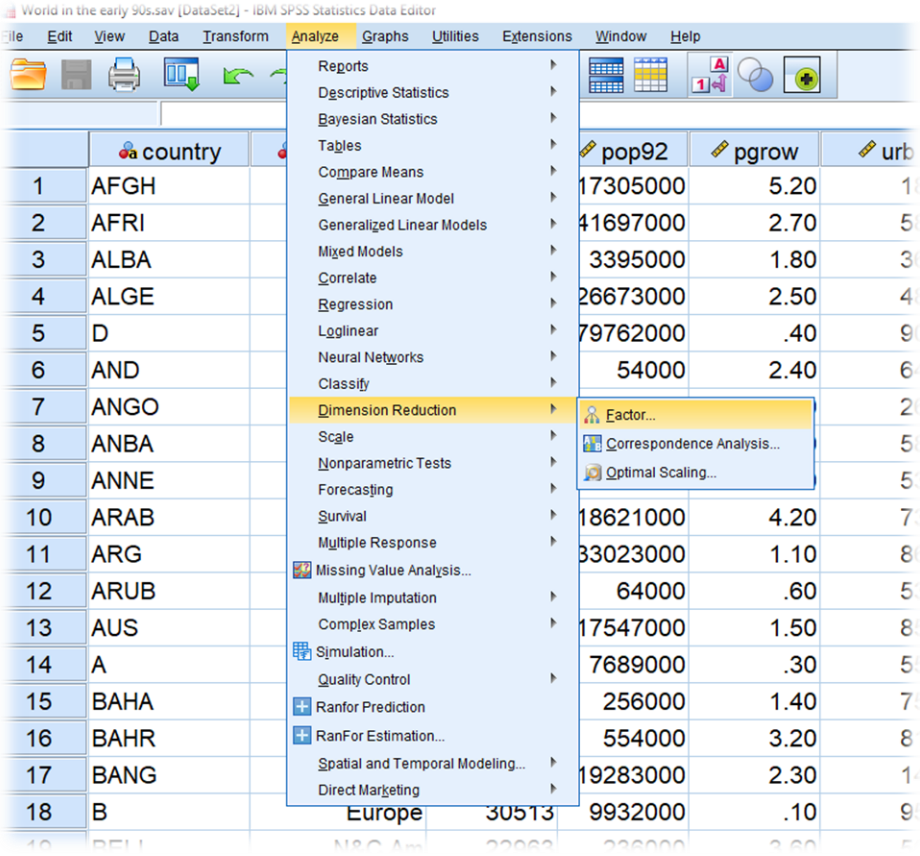
SMARTVISION
Europe



Performing Principal Components Analysis

Performing Principal Components Analysis

- We are using the **World in the early 90s.sav** dataset
- This dataset is comprised of several continuous fields showing different social, health and economic measurements for 183 countries in the early 1990s.
- To request the **Factor** procedure, from the main menu, click:
- **Analyze**
  - **Dimension Reduction**
    - **Factor**

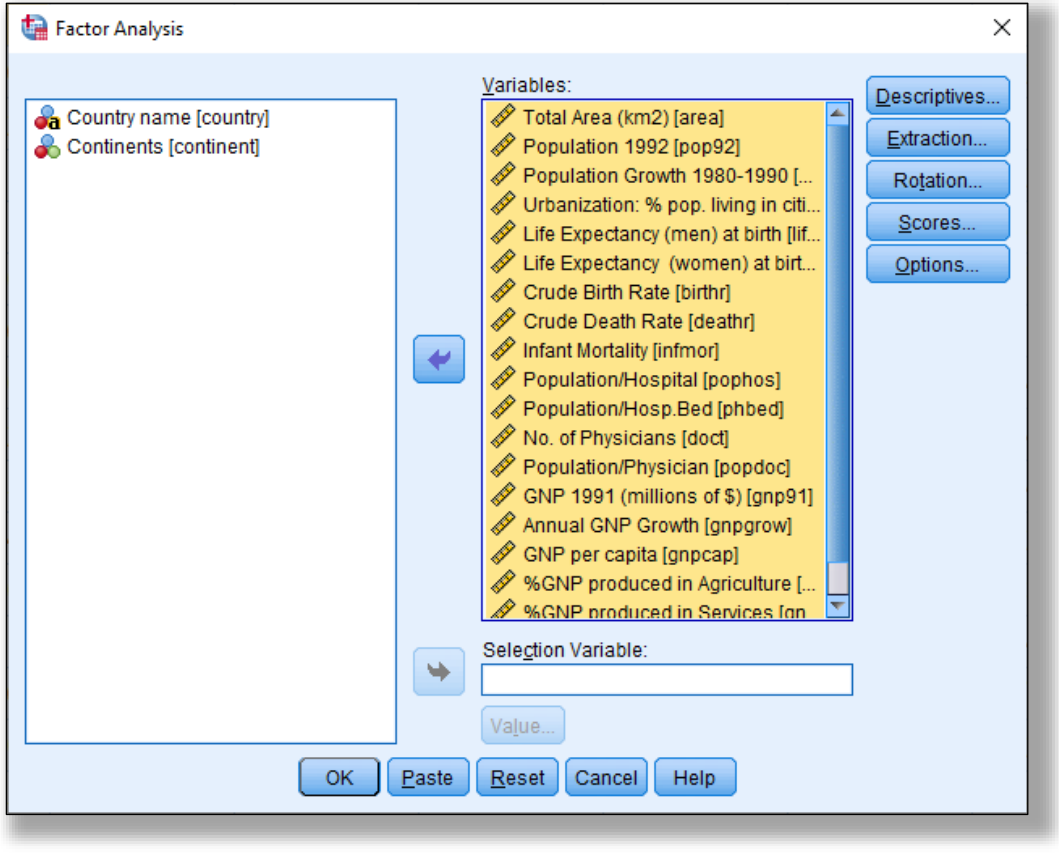


SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Performing Principal Components Analysis

- Despite the dialog's title, the default procedure is a Principal Components Analysis
- Selecting only the continuous (scale) variables we can run the procedure and view the default output
- Click:
  - **OK**



Performing Principal Components Analysis

- You can see that the first table produced is the one labelled 'Communalities'
- These values refer to the amount variation that the solution explains
- With PCA, the 'Initial' column is always equal to 1 (meaning 100%) . This is because PCA adds all of the variance to the procedure. Even if some of it is random.
- If we were to combine *all* of the resulting principal components together they would account for 100% of the variance in the submitted data

Communalities

	Initial	Extraction
Total Area (km2)	1.000	.695
Population 1992	1.000	.842
Population Growth 1980-1990	1.000	.729
Urbanization: % pop. living in cities	1.000	.710
Life Expectancy (men) at birth	1.000	.913
Life Expectancy (women) at birth	1.000	.936
Crude Birth Rate	1.000	.874
Crude Death Rate	1.000	.862
Infant Mortality	1.000	.908
Population/Hospital	1.000	.626
Population/Hosp.Bed	1.000	.593
No. of Physicians	1.000	.892
Population/Physician	1.000	.639
GNP 1991 (millions of $)	1.000	.625
Annual GNP Growth	1.000	.853
GNP per capita	1.000	.673
%GNP produced in Agriculture	1.000	.791
%GNP produced in Services	1.000	.595
Literacy Rate (%)	1.000	.772

Extraction Method: Principal Component Analysis.

- But we want *fewer* components than there are variables
- So the 'Extraction' column shows how much variance can be accounted for once the *default number* of principal components has been extracted
- So for 'Total Area (km2), for example, this is 0.695 (69.5%)

Performing Principal Components Analysis

- The 'Total Variance Explained' table shows what percentage of variance each component explains
- Note the first component explains 44%, the second 14.4% etc
- We can also see, that after extraction, the total amount of variance explained is 76.5% based on the top 5 components
- If we wanted to explain 100% of the variance, then we would need 19 components, as there are 19 variables
- PCA by default, extracts components if they have *eigenvalues greater than 1*

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.391	44.161	44.161	8.391	44.161	44.161
2	2.705	14.234	58.395	2.705	14.234	58.395
3	1.303	6.859	65.255	1.303	6.859	65.255
4	1.108	5.833	71.088	1.108	5.833	71.088
5	1.022	5.380	76.468	1.022	5.380	76.468
6	.868	4.569	81.037			
7	.687	3.618	84.655			
8	.584	3.074	87.729			
9	.472	2.486	90.216			
10	.455	2.393	92.608			
11	.417	2.194	94.802			
12	.281	1.481	96.284			
13	.222	1.166	97.450			
14	.171	.901	98.351			
15	.110	.580	98.931			
16	.088	.465	99.396			
17	.065	.341	99.737			
18	.042	.221	99.958			
19	.008	.042	100.000			

Extraction Method: Principal Component Analysis.

Performing Principal Components Analysis

- Eigenvalues are used by PCA and Factor analysis to extract components/factors that explain more than their 'fair share' of the variance
- If you have 10 variables, then PCA will produce 10 components
- If everything was equal, we might expect the components to explain 10% of the variance each
- But if the first component explained 20% of the variance, then it would have an eigenvalue of 2 – as its explaining twice as much 10%
- If the second component explained 15%, then it would have an eigenvalue of 1.5
- If the third component explained only 5%, then it would have an eigenvalue of 0.5
- This is a fairly crude way to figure out how many components or factors should be extracted and analysts are free to overrule the default setting if they wish

Performing Principal Components Analysis

- Finally, we have the Component Matrix table which is very like the correlation matrix we saw earlier, except that here we see the correlation (or 'loading') of each variable against each of the *extracted components*
- In this form however, it's not easy to make sense of as there are many weak correlations as well as strong ones
- Looking at Component 1 we can see a number of loadings greater than 0.7 in absolute magnitude that seem to be related to mortality and the measures of 'economic development'
- Let's re-run the process with a minor modification to make this particular output table easier to view

Component Matrix[a]

	Component				
	1	2	3	4	5
Total Area (km2)	-.049	.788	-.084	.236	.098
Population 1992	.000	.866	-.206	-.200	.101
Population Growth 1980-1990	.605	-.100	-.431	.190	.362
Urbanization: % pop. living in cities	-.755	-.056	.096	.355	.023
Life Expectancy (men) at birth	-.949	-.024	-.078	-.034	.073
Life Expectancy (women) at birth	-.965	-.031	-.049	-.021	.034
Crude Birth Rate	.893	-.077	-.226	.037	.135
Crude Death Rate	.770	.095	.409	.100	-.287
Infant Mortality	.949	.043	.047	.055	-.039
Population/Hospital	.393	.042	.467	.250	.435
Population/Hosp.Bed	.672	.072	.145	.025	.338
No. of Physicians	-.142	.922	-.128	-.064	.023
Population/Physician	.739	.000	.258	.152	-.057
GNP 1991 (millions of $)	-.254	.580	.239	.399	-.089
Annual GNP Growth	-.118	.165	.430	-.715	.340
GNP per capita	-.634	.090	.497	.066	-.109
%GNP produced in Agriculture	.803	.111	.068	-.227	-.279
%GNP produced in Services	-.529	-.245	.117	.129	.474
Literacy Rate (%)	-.868	-.033	.007	-.027	-.129

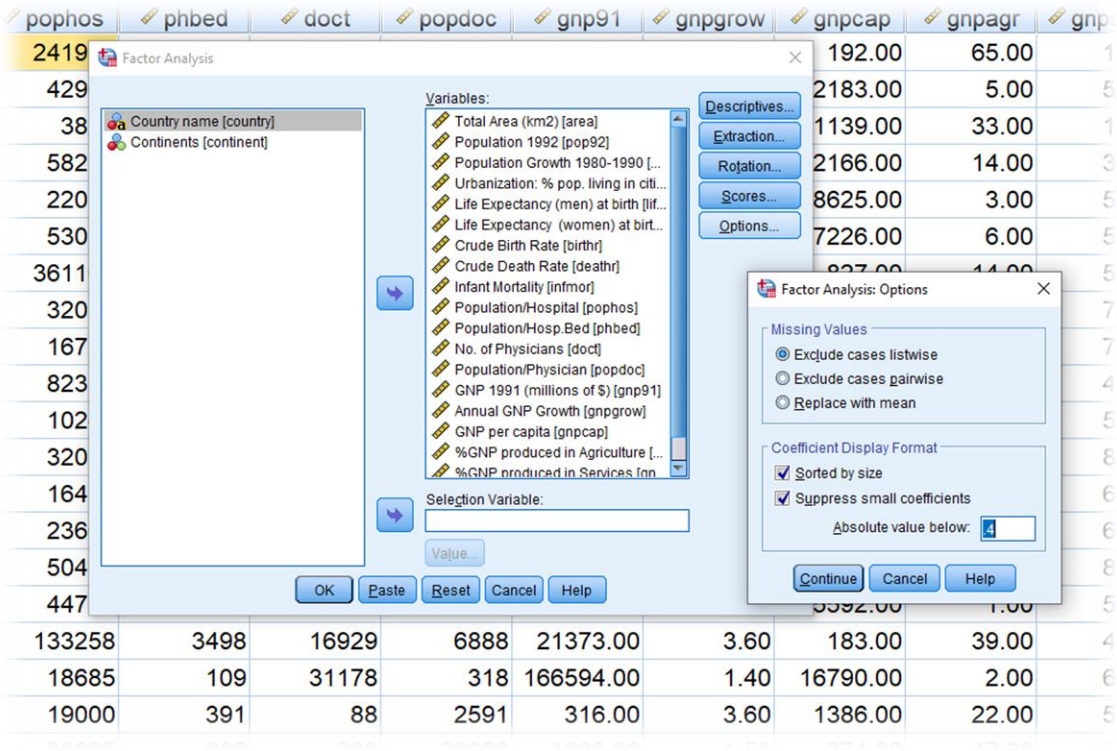
Extraction Method: Principal Component Analysis.

a. 5 components extracted.

SMARTVISION Europe

Performing Principal Components Analysis

- Returning to the Factor Analysis dialog, we can request that the coefficients are displayed in a clearer fashion.
- Within the dialog, click:
  - **Options**
- Check the boxes marked:
  - **Sorted by size**
  - **Suppress small coefficients**
- In the **Absolute value below** box change the value to:
  - **0.4**
- Click:
  - **Continue**
    - **OK**



SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Performing Principal Components Analysis

- By changing the display we can now see the strongest loadings by variable against the 5 extracted components
- Component 1 – accounting for 44% of the variation seems to be related to how developed the nations are
- Component 2 - accounting for 14% of the variation seems to relate to size and population of the country
- Component 3 - accounting for 6.8% of the variation is harder to interpret as has variables that also load on Component 1 but has reasonable loading population per hospital and GNP growth
- Component 4 accounting for 5.8% of the variation has only one strong negative loading suggesting it measures lack of GNP growth
- Component 5 accounting for 5.3% of the variation has only two variables with reasonably strong loadings. It's unclear what this measures but it might be related to developed nations with strong service-based economies

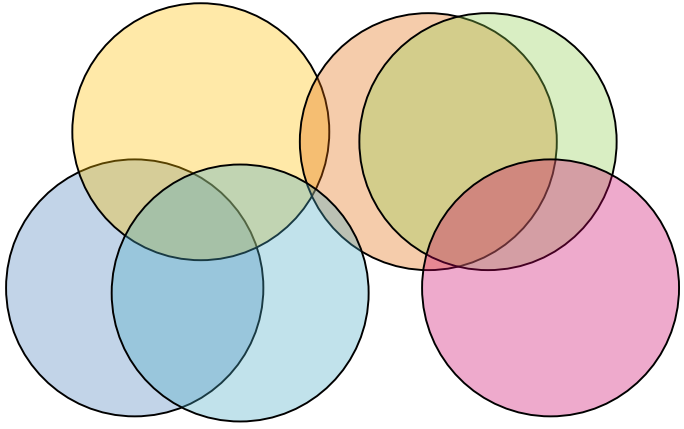
Component Matrix[a]

	Component				
	1	2	3	4	5
Life Expectancy (women) at birth	-.965				
Life Expectancy (men) at birth	-.949				
Infant Mortality	.949				
Crude Birth Rate	.893				
Literacy Rate (%)	-.868				
%GNP produced in Agriculture	.803				
Crude Death Rate	.770		.409		
Urbanization: % pop. living in cities	-.755				
Population/Physician	.739				
Population/Hosp.Bed	.672				
GNP per capita	-.634		.497		
Population Growth 1980-1990	.605		-.431		
%GNP produced in Services	-.529				.474
No. of Physicians		.922			
Population 1992		.866			
Total Area (km2)		.788			
GNP 1991 (millions of $)		.580			
Population/Hospital			.467		.435
Annual GNP Growth			.430	-.715	

Extraction Method: Principal Component Analysis.

a. 5 components extracted.

SMARTVISION
Europe



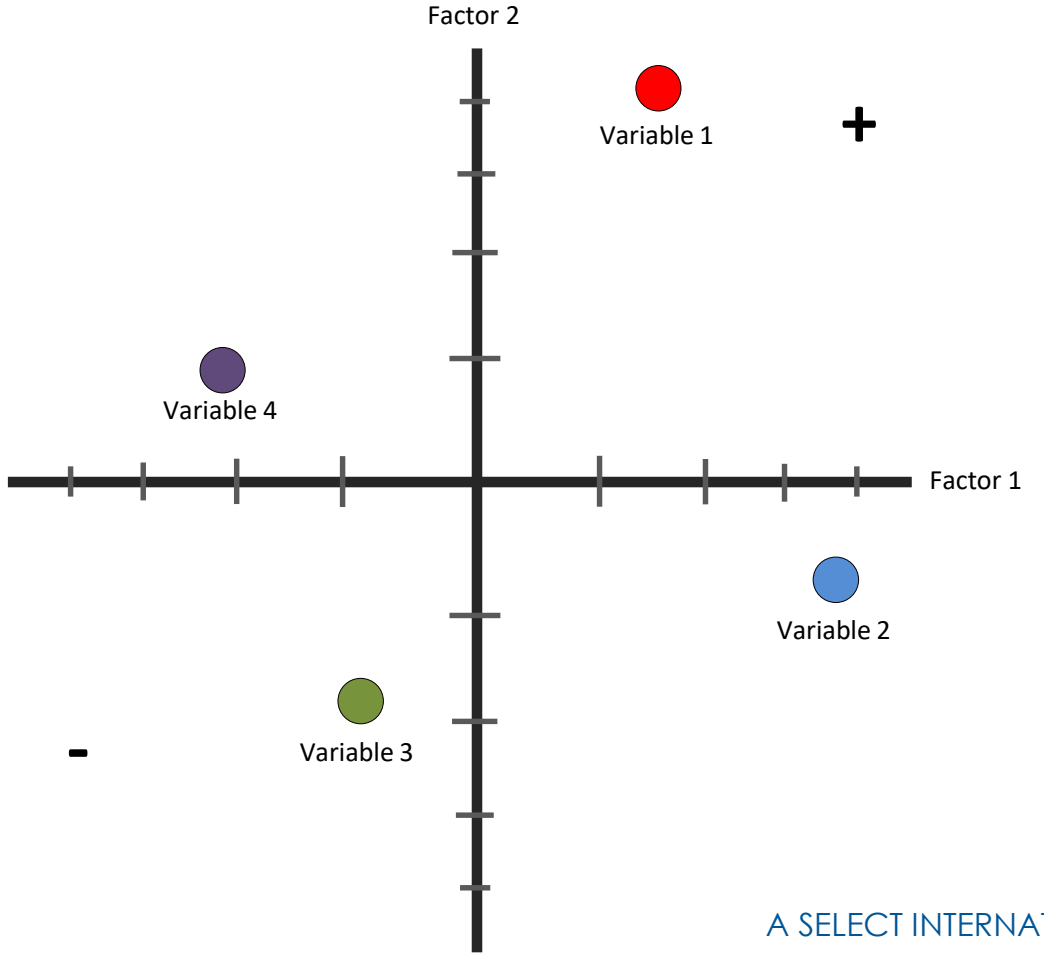
Rotated Solutions

Rotated Solutions

- Both PCA and Factor Analysis have an additional way to help us understand the relationships between the variables and the components/factors: this approach is known as 'Rotation'
- The goal of rotation is to create a simpler structure between the variables and the components/factors to aid interpretation
- This is done by rotating the axes of the components/factors mathematically so that we *maximise the loadings* of the variables on one components/factor or another
- Doing this in such a way that the rotated components/factors remain *uncorrelated* with one another, is known as an **Orthogonal** rotation
- However, a rotation that allows the components/factors to *correlate* with one another is known as an **Oblique** al rotation

Rotated Solutions

- Note that all the variables load at least a little bit on both Factor axes
- For example, Variable 1 loads high on Factor 1 but also a little on Factor 2
- Variable 3 loads negatively on Factor 2 but also somewhat negatively on Factor 1

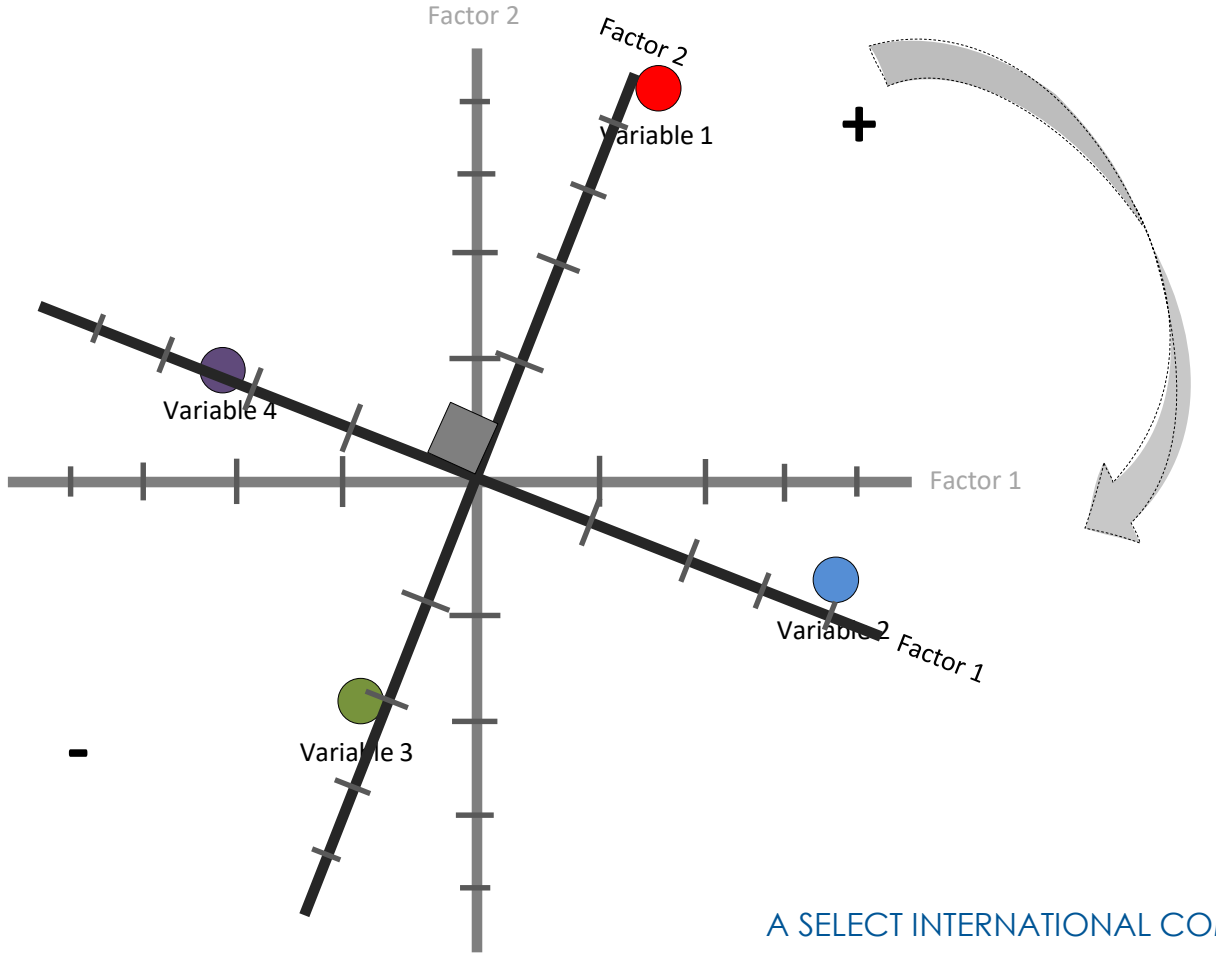


SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Rotated Solutions

- If we perform an Orthogonal rotation…
- We can see that the variables now load higher on one factor axis and less on the other
- The factors are simply scales that can be manipulated to allow this to occur
- It means that the extracted factor solution *may* be easier to interpret

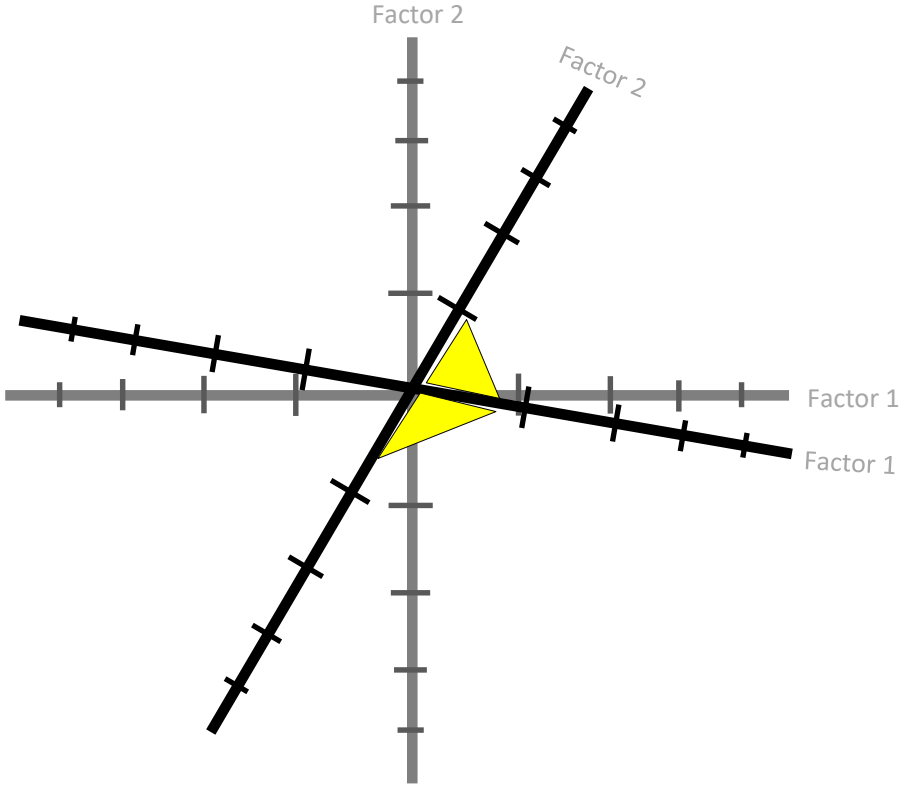


SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Rotated Solutions

- Oblique rotations can also be performed
- These are when the rotated factor axes are not kept at right angles to one another
- The result is that the rotated factors can be correlated with each other

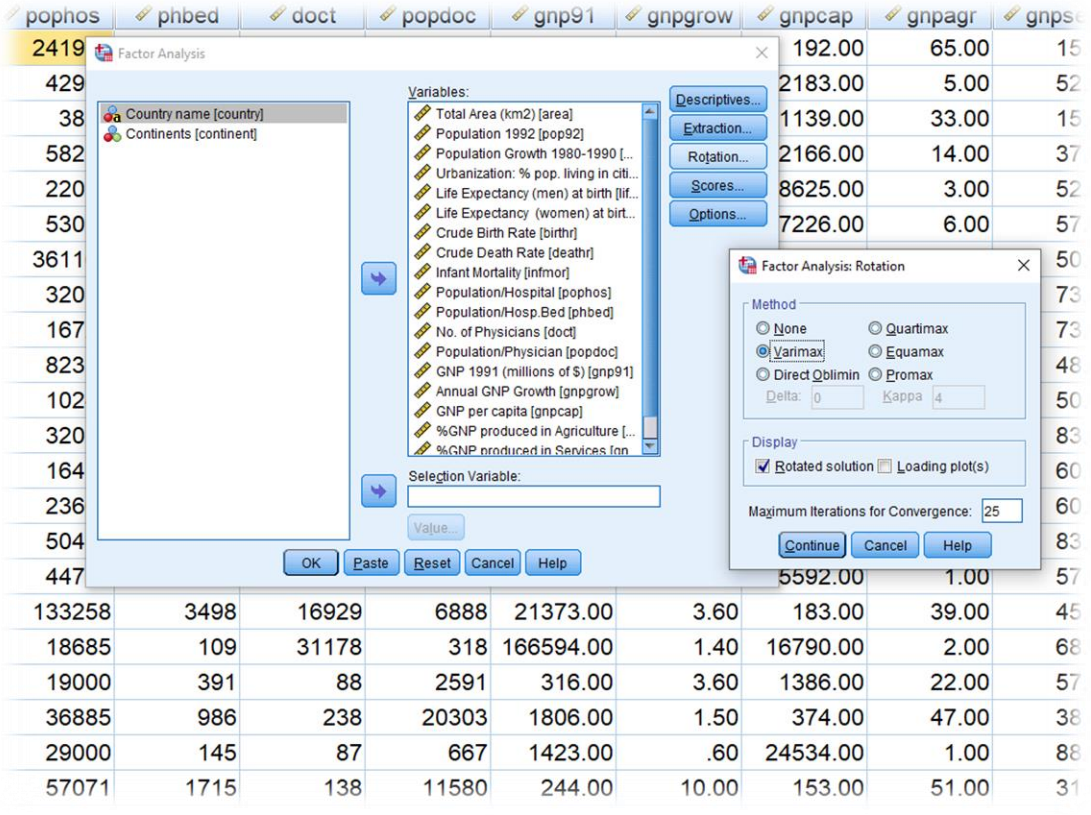


Rotated Solutions

- Within PCA and Factor Analysis there are 3 main kinds of **Orthogonal** rotation
  - **Varimax** rotation is perhaps the most well known orthogonal method. It tries to spread the loadings across the different factors in order to make the factors easier to interpret.
  - **Quartermax** rotation instead tries to minimize the number of factors needed to explain each variable. Often this approach leads to a solution where many variables dominate the loadings on one factor.
  - **Equamax** rotation is a compromise between the previous two orthogonal methods.
- The methods of **Oblique** rotation available are:
  - **Direct Oblimin** allows the user to control the degree to which the solution is non-orthogonal with the delta value. Using delta values between 0 and .8 will result in factors that are increasingly oblique. But the result may be factors that are so highly correlated (positively or negatively) as to be indistinguishable. On the other hand, large negative values of delta will lead to factors which are nearly orthogonal.
  - **Promax** is another oblique rotation method. This rotation can be calculated more quickly than a Direct Oblimin rotation, so it is useful for large datasets.

Performing Principal Components Analysis

- Returning to the Factor Analysis dialog, we can request that the coefficients are displayed in a clearer fashion.
- Within the dialog, click:
  - **Rotation**
    - **Varimax**
- Click:
  - **Continue**
    - **OK**



SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Performing Principal Components Analysis

- Scrolling to the end of the output, we can now see an additional table showing the loadings of each variable against the extracted components. It is labelled **Rotated Component Matrix**
- Component 1 is now particularly strongly related to mortality measures and the degree of education and urbanization
- Component 2 seems to measure the degree to which the population size is changing and GNP per capita
- Component 3 may be related to size of the country in terms of its land mass, population and GNP
- Component 4 focusses on the availability of medical care such as doctors and hospitals per head of the population
- Component 5 is most strongly related to annual GNP growth

Rotated Component Matrix[a]

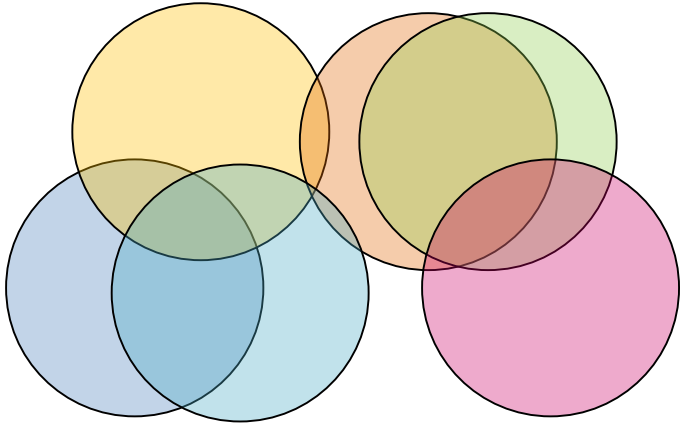
	Component				
	1	2	3	4	5
Crude Death Rate	.873				
%GNP produced in Agriculture	.855				
Life Expectancy (men) at birth	-.839				
Life Expectancy (women) at birth	-.829	.422			
Infant Mortality	.816	-.409			
%GNP produced in Services	-.684				
Population/Physician	.683				
Urbanization: % pop. living in cities	-.663	.445			
Literacy Rate (%)	-.658	.494			
Population Growth 1980-1990		-.782			
GNP per capita		.732			
Crude Birth Rate	.607	-.674			
No. of Physicians			.928		
Population 1992			.870		
Total Area (km2)			.816		
GNP 1991 (millions of $)		.428	.574		
Population/Hospital				.757	
Population/Hosp.Bed	.440			.478	
Annual GNP Growth					.897

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

a. Rotation converged in 7 iterations.

SMARTVISION Europe

SMARTVISION Europe



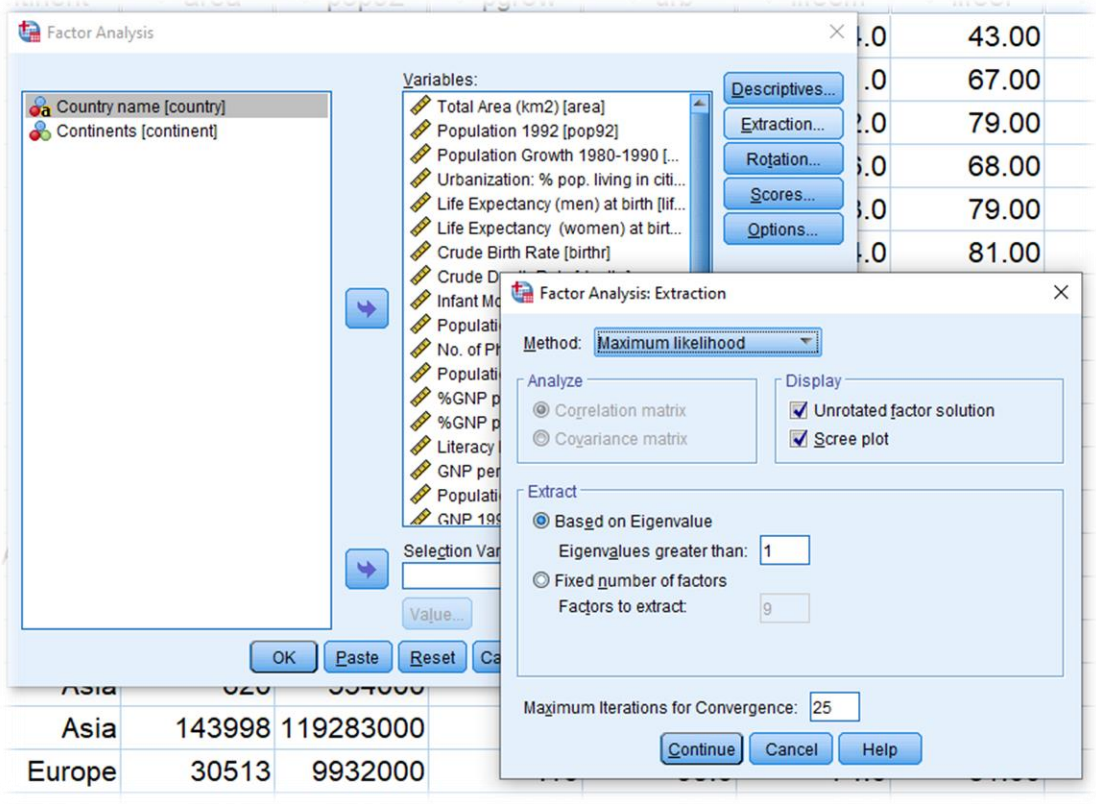
Performing Factor Analysis

Performing Factor Analysis

- To many people, the differences between **Principal Components Analysis (PCA)** and true **Factor Analysis** may appear subtle.
- **PCA** is regarded as a more practical form of the family of techniques that Factor Analysis belongs to.
- **PCA** has a strong focus on reducing the data down to a smaller number of dimensions by creating *linear combinations* of the variables.
- **PCA** is particularly used to create new uncorrelated variables so that they can be used in analysis procedures that are sensitive to highly correlated data.
- **Factor Analysis** is more often used to model *latent variables.* These are factors that can't be measured directly, such as extroversion or introversion. Instead a number of variables are used to measure different, correlated aspects these latent factors.
- **Factor Analysis** is a more formal approach usually used within a wider theoretical context of the subject matter under investigation (such as personality testing)
- Unlike **PCA**, in **Factor Analysis** the initial communality values of only the variables *are not equal to 1*. Because Factor Analysis considers only the commonly shared variance, the initial values represent the total variance that be accounted for using only the other variables in the procedure.

Performing Factor Analysis

- To demonstrate a Factor Analysis procedure, return to the dialog and click:
  - **Extraction**
- There are a number of different extraction procedures. Choose:
  - **Maximum Likelihood**
- This is a reasonable, all-round version of Factor Analysis, also click the box marked:
  - **Scree Plot**
- Now click:
  - **Continue**
    - **OK**



SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Performing Factor Analysis

- You can see that the first 'Communalities' table is very different
- Now the 'Initial' values are similar to the 'Extracted' values in the PCA procedure earlier
- The initial values represent the common variability that can be accounted for by all of the variables *before* they are entered into the analysis
- The 'Extraction' column shows how much variance can be accounted after the factors with eigen values greater than 1 are extracted.

Communalities[a]

	Initial	Extraction
Total Area (km2)	.462	.411
Population 1992	.817	.789
Population Growth 1980-1990	.743	.999
Urbanization: % pop. living in cities	.665	.641
Life Expectancy (men) at birth	.983	.985
Life Expectancy (women) at birth	.987	.995
Crude Birth Rate	.887	.866
Crude Death Rate	.870	.905
Infant Mortality	.945	.941
Population/Hosp.Bed	.533	.379
No. of Physicians	.850	1.000
Population/Physician	.584	.530
%GNP produced in Agriculture	.748	1.000
%GNP produced in Services	.466	.420
Literacy Rate (%)	.769	.733
GNP per capita	.635	.622
Population/Hospital	.327	.145
GNP 1991 (millions of $)	.435	.351
Annual GNP Growth	.249	.051

Extraction Method: Maximum Likelihood.

a. One or more communalitiy estimates greater than 1 were encountered during iterations. The resulting solution should be interpreted with caution.

- Note that some of the variables have extraction values equal to 1. Meaning that all of their accountable variance can be predicted from the extracted factors
- Meanwhile the variable Annual GNP Growth has an extracted value of .051 meaning only 5% of it's accounted for variance can be predicted using the extracted factors

Performing Factor Analysis

- Unlike the PCA output, after extraction, the total amount of variance explained is 67% (rather than) 76.4% based on the top 5 components.
- So this solution doesn't appear to account for as much variation as the PCA approach

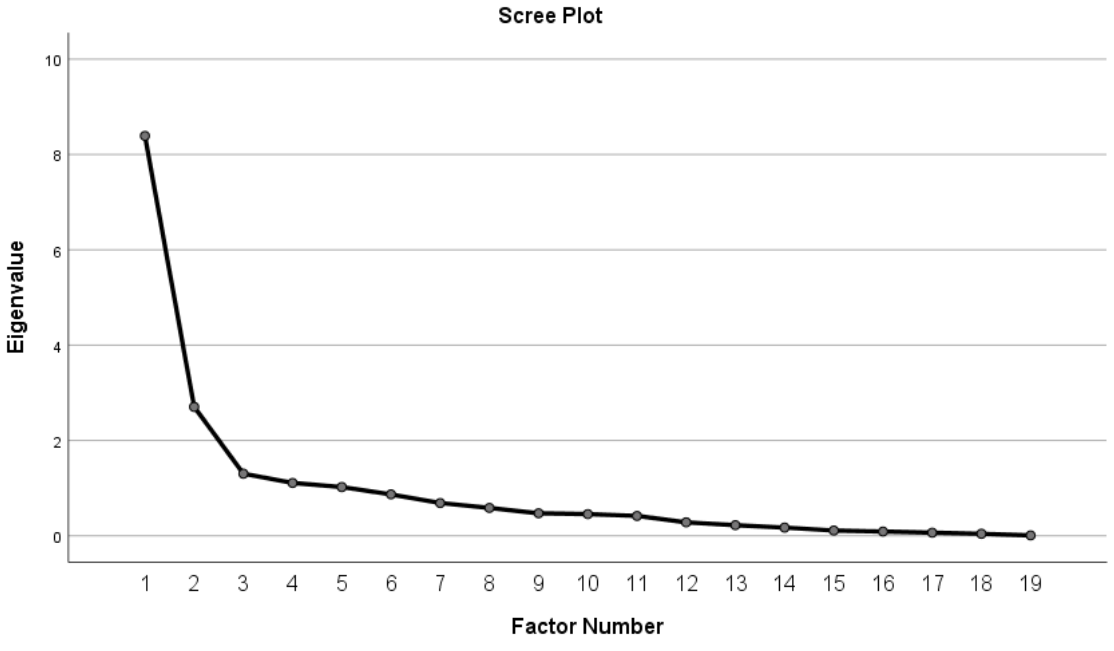
Total Variance Explained

Factor	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	8.391	44.161	44.161	4.702	24.748	24.748	5.438	28.621	28.621
2	2.705	14.234	58.395	3.472	18.271	43.019	2.380	12.525	41.146
3	1.303	6.859	65.255	1.919	10.102	53.121	1.743	9.172	50.318
4	1.108	5.833	71.088	2.001	10.533	63.654	1.697	8.933	59.251
5	1.022	5.380	76.468	.668	3.513	67.167	1.504	7.916	67.167
6	.868	4.569	81.037						
7	.687	3.618	84.655						
8	.584	3.074	87.729						
9	.472	2.486	90.216						
10	.455	2.393	92.608						
11	.417	2.194	94.802						
12	.281	1.481	96.284						
13	.222	1.166	97.450						
14	.171	.901	98.351						
15	.110	.580	98.931						
16	.088	.465	99.396						
17	.065	.341	99.737						
18	.042	.221	99.958						
19	.008	.042	100.000						

Extraction Method: Maximum Likelihood.

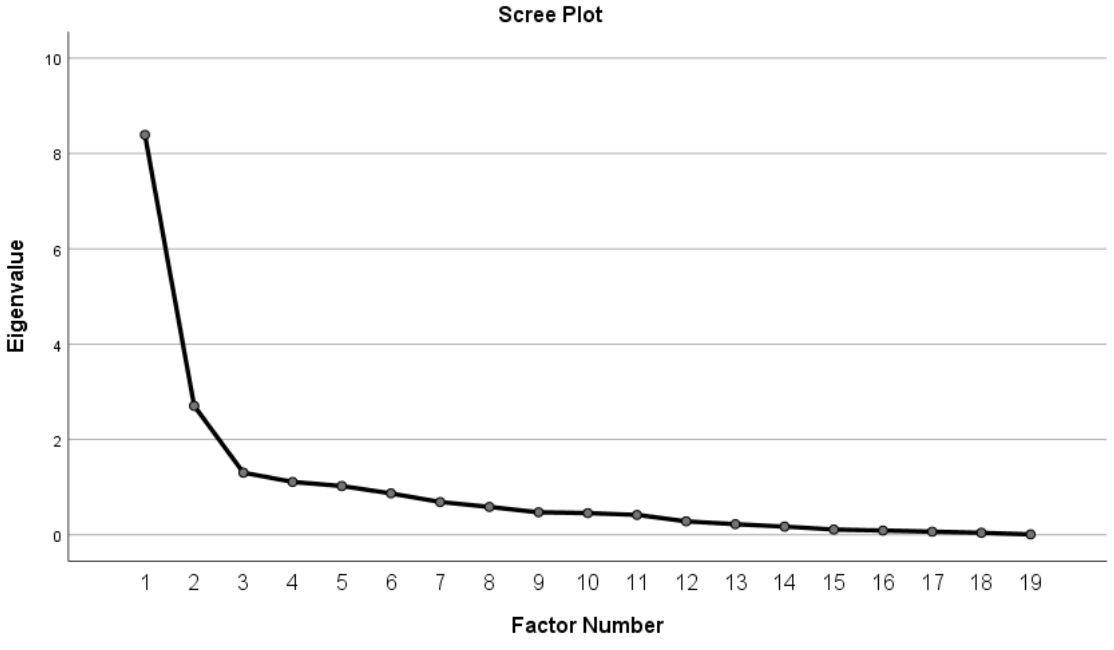
Performing Factor Analysis

- We also requested a Scree Plot. This charts the relationship between the eigenvalues and the number of factors/components
- The idea is that rather than simply extracting factors based on the rule that their eigenvalue should be greater than 1, a visual representation might help.



Performing Factor Analysis

- The convention is that the number of factors we should request corresponds to where ‘elbow’ of the scree plot seems level off
- In this case, the plot indicates that this might be at the point 3 factors are extracted.
- The obvious criticism of scree plots is that they are rather subjective



Performing Factor Analysis

- Scrolling to the end of the output can see a **Goodness of Fit Test** and a **Rotated Component Matrix** for our Factor Analysis
- The **Chi-Square Goodness of Fit Test** compares the how well the five extracted factors explain the variation in the data.
- In this test, values below 0.05 in the Significance column indicate the model does not fit the data well
- Here, the test indicates that the 5 factor solution *does not* fit the data well *
- We might resolve this issue by increasing the number of factors that are extracted. But that might defeat the point of the exercise if we are particularly interested in data reduction
- Perhaps Maximum Likelihood Factor Analysis is not an appropriate method for this dataset unless we are trying to detect latent variables as part of a wider theoretical model of global economics and poverty

Goodness-of-fit Test

Chi-Square	df	Sig.
215.452	86	.000

Rotated Factor Matrix[a]

	Factor				
	1	2	3	4	5
Crude Death Rate	.923				
Life Expectancy (men) at birth	-.887				
Life Expectancy (women) at birth	-.866				
Infant Mortality	.855				
Literacy Rate (%)	-.704				
Population/Physician	.610				
Crude Birth Rate	.605			-.440	.482
Population/Hosp.Bed	.458				
Population/Hospital					
No. of Physicians		.988			
Population 1992		.862			
Total Area (km2)		.637			
GNP 1991 (millions of $)		.461			
%GNP produced in Agriculture	.523		.825		
%GNP produced in Services			-.528		
GNP per capita				.697	
Urbanization: % pop. living in cities	-.444		-.441	.495	
Population Growth 1980-1990					.920
Annual GNP Growth					

Extraction Method: Maximum Likelihood.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

SMARTVISION Europe

* Increasing the number of extracted factors to 8 causes this test to not be shown

Performing Factor Analysis

- The Rotated Factor Matrix shows the following results
- Factor 1 – accounting for 28.6% of the variation, is again comprised of mortality measures and variables associated with the degree to which a country is a developed nation
- Factor 2 - accounting for 12.5% of the variation again seems to relate to size and population of the country
- Factor 3 - accounting for 9.1 % of the variation appears to relate to the degree to which the economy is agriculturally or service-based
- Factor 4 accounting for 8.9% of the is related to the wealth of the country and the degree of urbanization
- Factor 5 accounting for 7.9% of the variation is related to population growth
- Note that the variable **Annual GNP Growth**, for which the extracted solution could account for 5% its of variance, has no loadings greater than 0.4, so its values are not shown

Goodness-of-fit Test

Chi-Square	df	Sig.
215.452	86	.000

Rotated Factor Matrix[a]

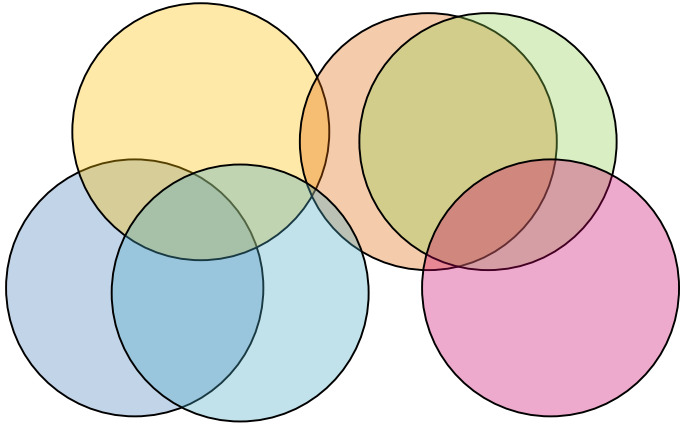
	Factor				
	1	2	3	4	5
Crude Death Rate	.923				
Life Expectancy (men) at birth	-.887				
Life Expectancy (women) at birth	-.866				
Infant Mortality	.855				
Literacy Rate (%)	-.704				
Population/Physician	.610				
Crude Birth Rate	.605			-.440	.482
Population/Hosp.Bed	.458				
Population/Hospital					
No. of Physicians		.988			
Population 1992		.862			
Total Area (km2)		.637			
GNP 1991 (millions of $)		.461			
%GNP produced in Agriculture	.523		.825		
%GNP produced in Services			-.528		
GNP per capita				.697	
Urbanization: % pop. living in cities	-.444		-.441	.495	
Population Growth 1980-1990					.920
Annual GNP Growth					

Extraction Method: Maximum Likelihood.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

SMARTVISION Europe

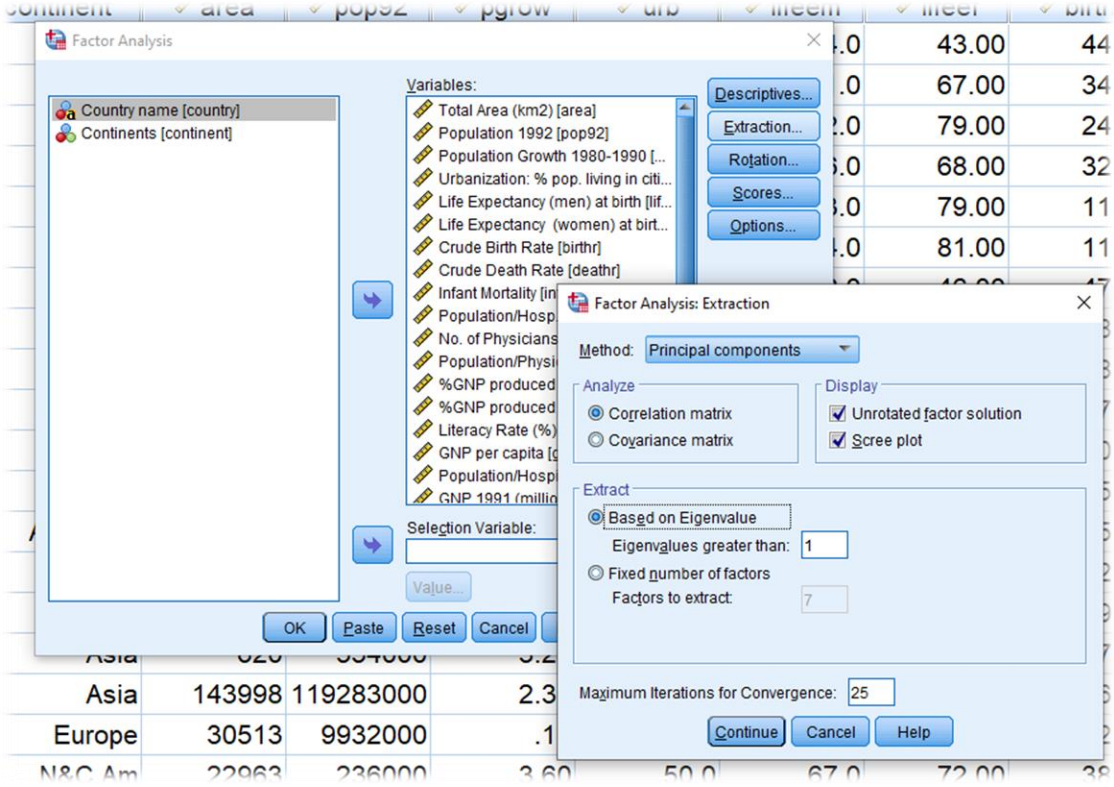
SMARTVISION Europe



Analysing Component Scores

Analysing Component Scores

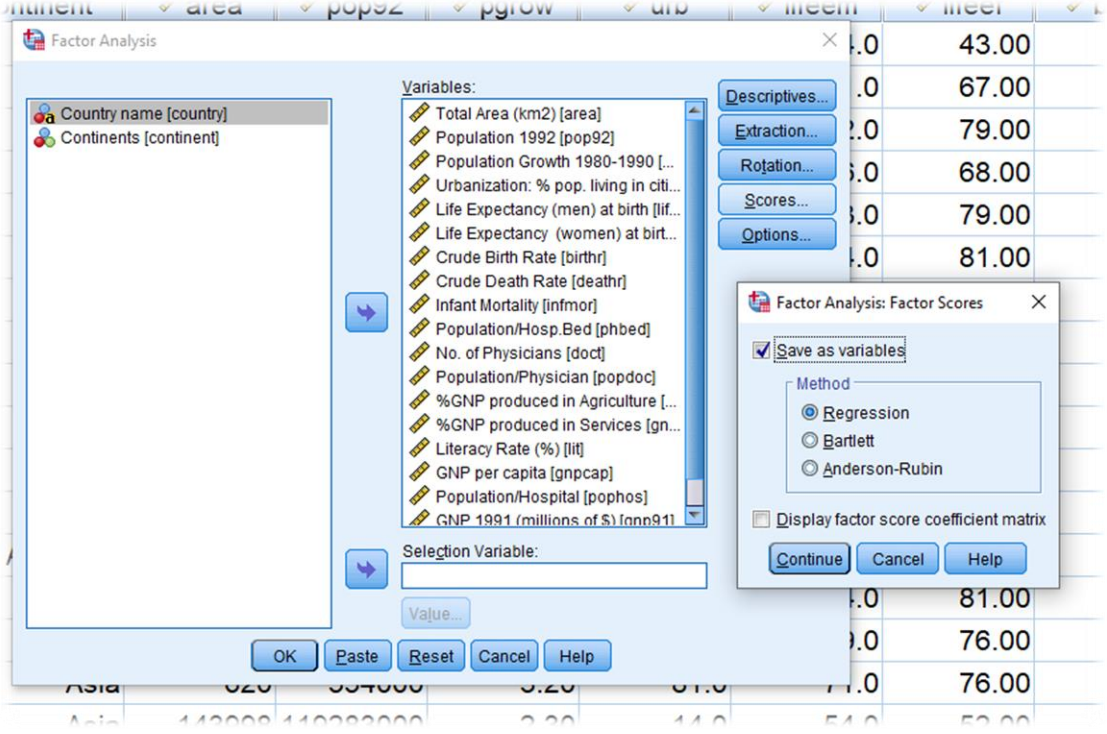
- To demonstrate creating Facto/Component scores, lets return to the procedure dialog and change extraction method back to PCA. To do so, click:
  - **Extraction**
  - **Principal components**
- Now click:
  - **Continue**



Analysing Component Scores

- To request that the procedure generates component scores in the form of new variables, now click:
  - **Scores**
- Check the box marked

  Save as variables
- Now click:
  - **Continue**
    - **OK**



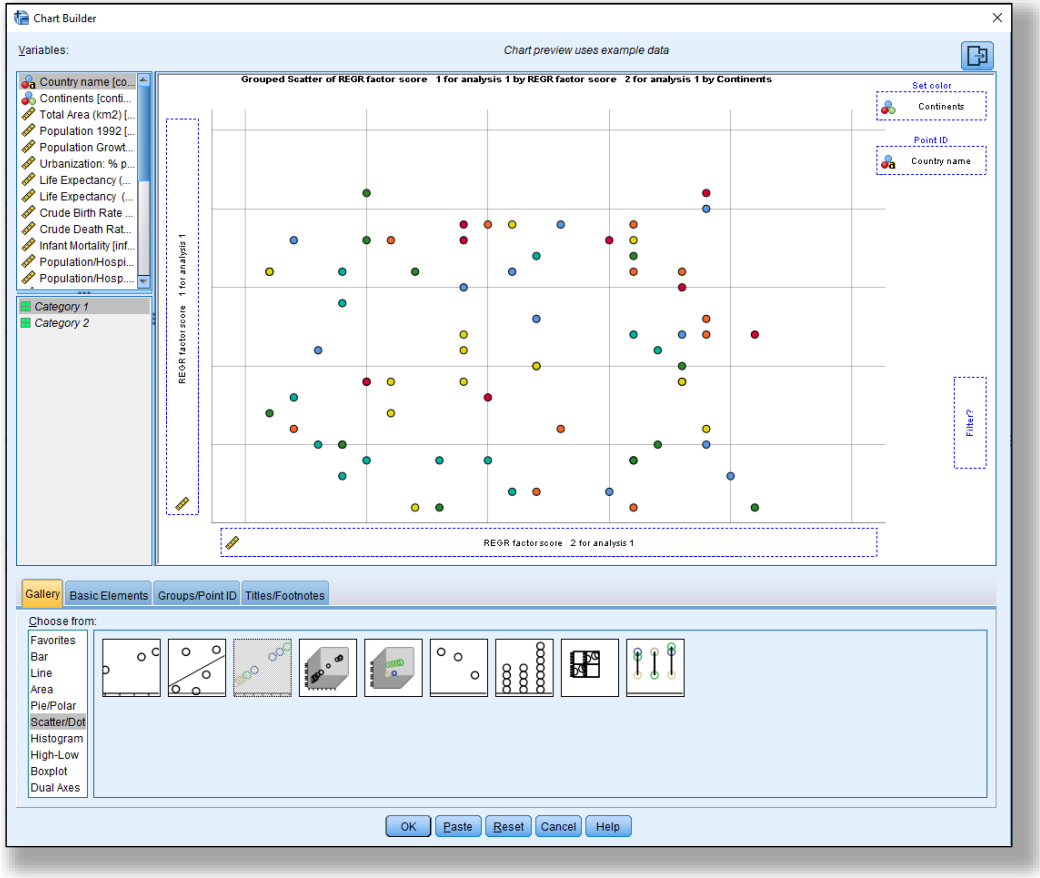
Analysing Component Scores

- The PCA analysis is re-run as before but now 5 new variables have been added to the dataset
- These variables show how strongly each case (or in this case 'country') loads on a standardised scale
- This is how the procedure can be used to create dimensions for a personality test, showing, for example, how highly someone scored on an introversion-extroversion scale
- In the same sense we can show how highly countries in the early 1990's scored in Component 1: a scale related to high life-expectancy and literacy rates and economic development

	lit	FAC1_1	FAC2_1	FAC3_1	FAC4_1	FAC5_1
0	29.00	2.02989	-.85621	.12990	.89655	-.482
0	76.00	-.50861	-.69883	.21113	.01169	-.547
0	72.00	.37863	-.04462	-.15253	-2.03314	-.389
0	50.00	-.20542	-.66107	.43724	-.23806	-.188
0	99.00	-.13798	2.35712	1.05142	.27136	-1.034
0	100.00	-1.18909	-.07282	-.29228	-.09634	.157
0	42.00	1.11317	-.05696	-.08430	2.39161	1.058
0	89.00	-.80371	.32930	-.45501	-.03826	.784
0	94.00	-.99954	.31498	-.44487	-.28447	.377
0	62.00	-.90719	-1.31271	.44776	.49847	-1.640
0	95.00	-.35309	.38807	.57405	-.52714	-1.088
0	95.00	-1.09242	.78774	-.55799	.66122	3.251
0	100.00	-1.07672	.70594	1.46415	.63283	-.887
0	99.00	-.28920	1.64375	-.34093	-.22679	-.078
0	90.00	-1.25111	.25803	-.44223	.67725	.452
0	77.00	-1.39679	-.88475	-.20506	.09306	-.920
0	35.00	.84989	-.93118	.26492	1.22353	1.186
0	99.00	-.57371	1.75185	-.32580	.17961	-.491
0	91.00	-.79436	-1.15393	-.24501	-.37569	.139
0	23.00	1.56223	-.45992	-.24837	.00134	-.461
0	98.00	-1.29899	1.07846	-.50245	.97233	-.597
0	18.00	2.01482	.11549	-.30280	.07645	1.848
0	78.00	.04803	-.55765	-.02082	-.44161	-1.060
0	23.00	-.04887	-.83552	-.12768	.23738	1.549
0	81.00	-.62315	-.44163	2.65939	.31716	-.783

Analysing Component Scores

- Under the **Graphs** menu, we can use the **Chart Builder** to request a **Grouped Scatterplot** of the first two score variables.
- Assign:
  - **Component 1 (Fac1_1)** to the vertical axis
  - **Component 2 (Fac2_1)** to the horizontal axis
  - The variable **Continents** to the **Set color** box
- To generate the chart click:
  - **OK**

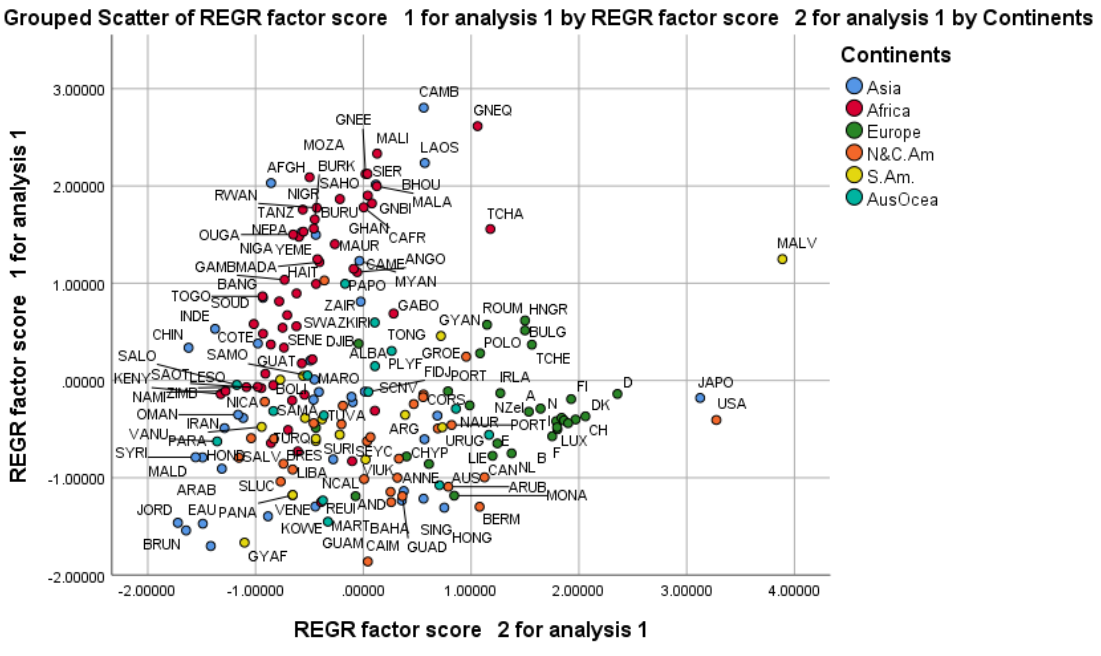


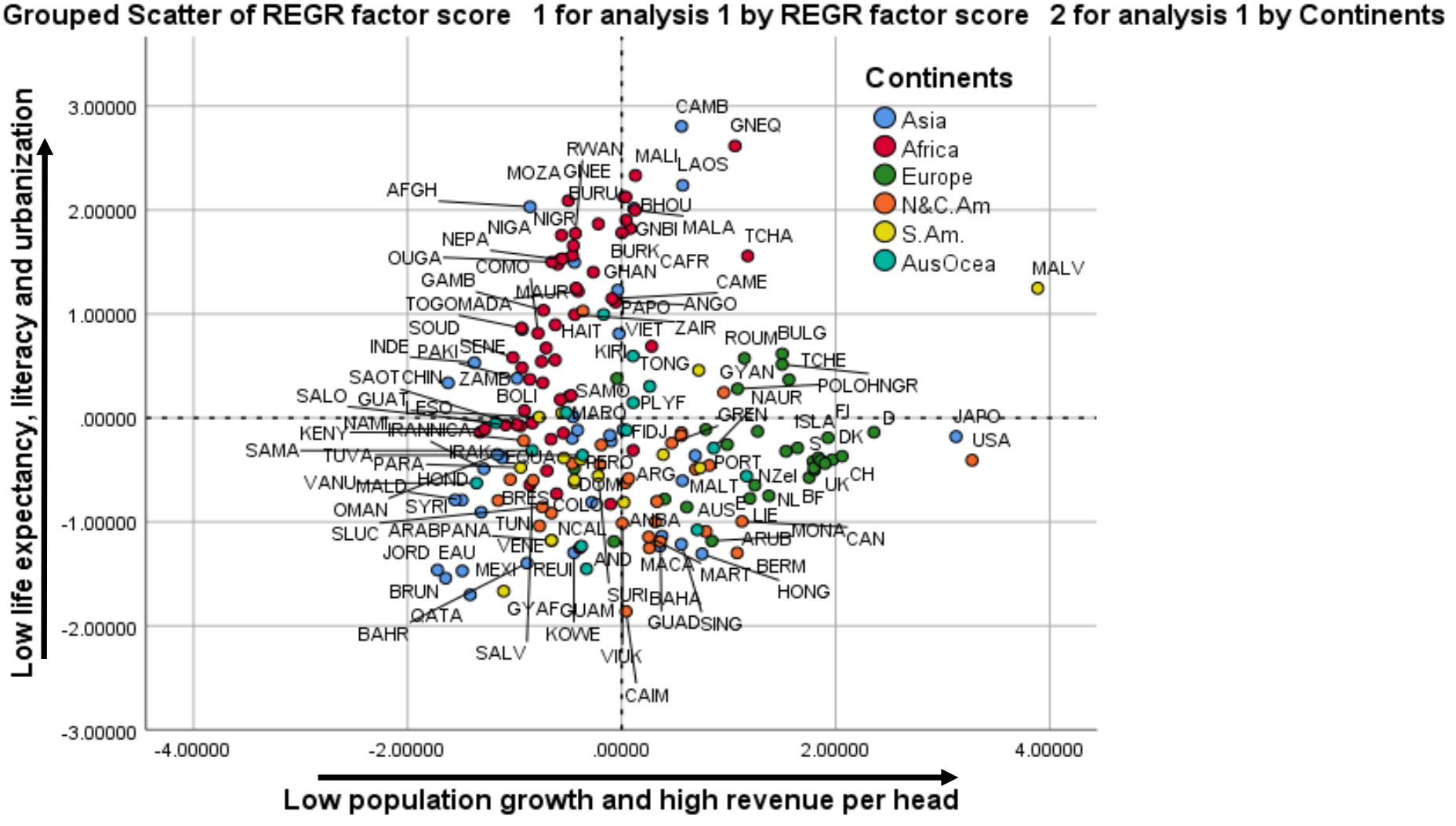
SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

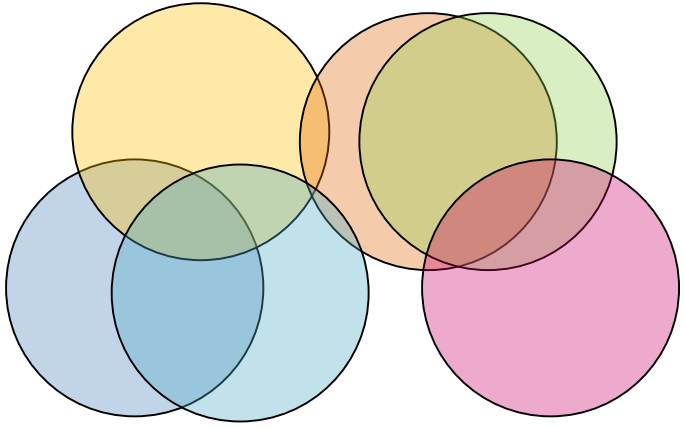
Analysing Component Scores

- The resulting chart shows the positions of the various countries with respect to their values on the first two component score variables.
- Remember that those that score high on the first component are countries with low values for life expectancy, literacy rates and urbanisation and high values for agrarian economies
- Whereas those that score high on the second component are those countries with low population growth and high GNP per capita




Grouped Scatter of REGR factor score 1 for analysis 1 by REGR factor score 2 for analysis 1 by Continents
Continents
Asia
Africa
Europe
N&C.Am
S.Am.
AusOcea
Low life expectancy, literacy and urbanization
Low population growth and high revenue per head
3.00000
2.00000
1.00000
.00000
-1.00000
-2.00000
-3.00000
-4.00000
-2.00000
.00000
2.00000
4.00000
CAMB
GNEQ
RWAN
MALI
LAOS
MOZA
GNEE
AFGH
BURU
BHOU
NIGR
NIGA
GNBI
MALA
TCHA
NEPA
BURK
OUGA
COMO
GHAN
CAFR
MALV
GAMB
MAUR
CAME
TOGO
MADA
PAPO
ANGO
SOUD
HAIT
VIET
ZAIR
ROUM
BULG
INDE
PAKI
SENE
KIRI
TONG
TCHE
SAOT
CHIN
ZAMB
GYAN
POLO
HNGR
SALO
GUAT
BOLI
SAMO
PLYF
NAUR
LESO
MARO
GREN
FJ
ISLA
D
JAPO
NAMI
IRAN
NICA
FIDJ
S
DK
USA
KENY
SAMA
TUVA
IRAQ
ECUA
PORT
PARA
RERO
NZel
CH
VANU
MALD
HOND
ARG
MALT
UK
DOMI
BRES
NL
BF
OMAN
SYRI
COLO
AUS
LIE
SLUC
ARAB
PANA
TUNI
NCAL
ANBA
MONA
CAN
ARUB
JORD
EAU
VENE
AND
MACA
MART
BERM
MEXI
REUI
HONG
BRUN
SURI
BAHA
QATA
GYAF
GUAM
GUAD
SING
BAHR
KOWE
SALV
VIUK
CAIM

SMARTVISION
Europe



Introducing Cluster Analysis

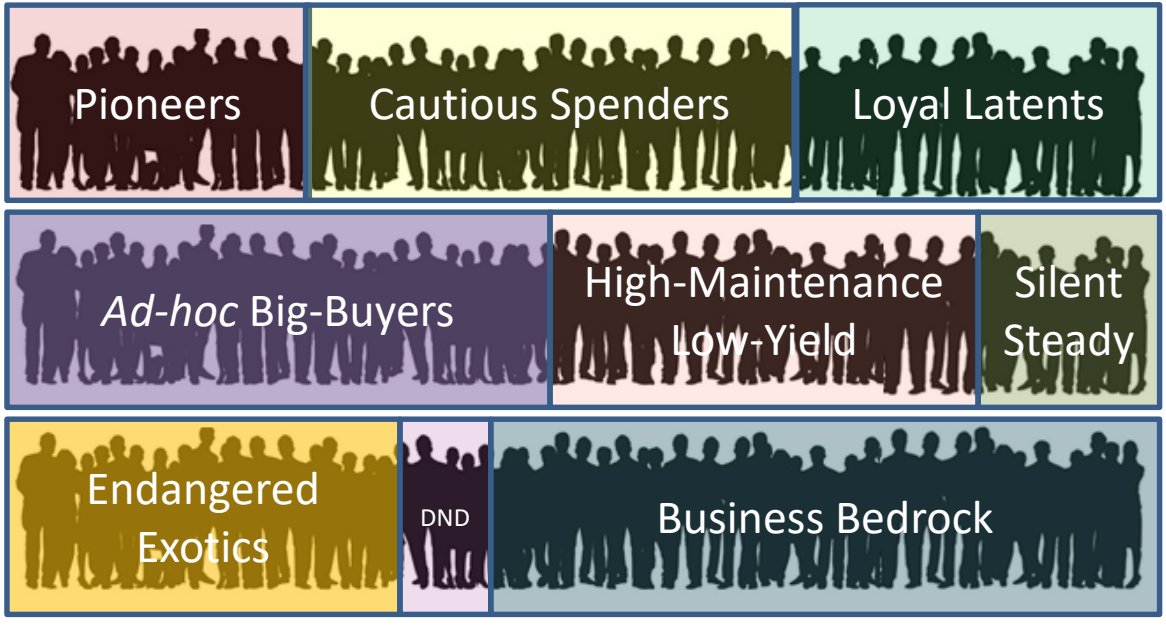
Introducing Cluster Analysis

- So far we have seen how techniques like Factor Analysis can be use to detect latent factors and create linear combinations of variables by identifying how groups of fields that are related to each other
- In contrast, Cluster Analysis refers to a family of techniques that focus on how records (or cases) can be grouped together to create new category groups in a data file
- Cluster Analysis attempts to establish a distance measurement between the records in a dataset and to use this to find records which are most similar to one another

Introducing Cluster Analysis

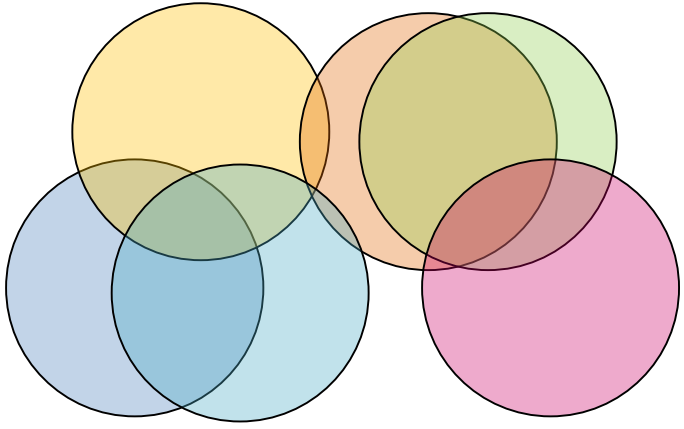
- Cluster Analysis is used in multiple applications such as:
  - Using demographics and shopping behaviour within a customer database to find similar segments in order to drive deeper customer insight and provide more compelling products or services.
  - Using data from smart meters to establish the distinct ways in which different kinds of households consume electricity
  - Creating groups of similar retail stores together in terms of their sales patterns to more effectively resource them and to uncover different modes of shopper behaviour
  - Finding clusters of patients with similar symptoms and characteristics in order to detect otherwise hidden illnesses and disorders
  - Uncovering separate groups of user interactions with a website or an app in order to improve the layout design and information delivery

Finding groups to drive deeper insight



Techniques like cluster analysis can uncover subtle differences and previously hidden groups

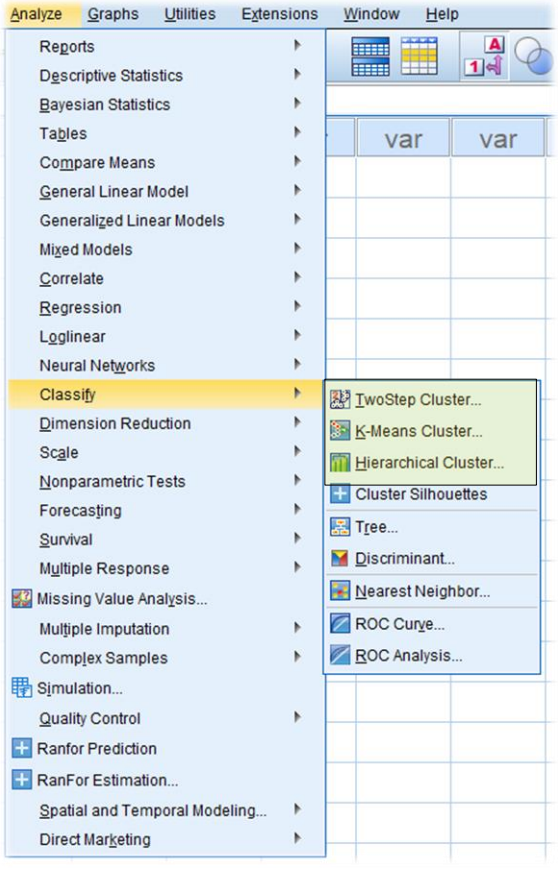
SMARTVISION
Europe



Comparing Cluster Methods

Introducing Cluster Analysis

- IBM SPSS Statistics has three Cluster Analysis algorithms:
  - K-Means Cluster
  - Hierarchical Cluster Analysis
  - Two- Step Cluster



Introducing Cluster Analysis

- **K-Means Cluster**
- This is a classic form of cluster analysis designed to work with continuous variables. The K-Means algorithm (the SPSS Quick Cluster command) identifies *k* number of centroid positions in the data and then allocates each case point to the cluster with the nearest centroid.
  1. It begins this process by calculating an average value for each case in the dataset
  2. Then it tries to find *k* cases have very different average values.
  3. It uses these *k* cases as its start points (i.e. the initial centroids)
  4. It calculates which of these initial *k* centroids are the nearest neighbours of the remaining cases and assigns each case to the nearest centroid to form a *k* clusters
  5. Having formed *k* clusters it *recalculates* the centroid of each cluster as then *reassigns* all the cases to the nearest new centroids
  6. It repeats this process. As it does so, the centroids change positions a little less each time until the centroid positions stop moving and the final clusters are established

Introducing Cluster Analysis

- **Limitations of K-Means Cluster**

- It requires the user to choose how many clusters they want *before* they run the analysis
- You can't use variables like region or product category. It really only works with continuous data. Dichotomous (binary) variables can be used, but it's better if all the data are made up the same variable types.
- In SPSS, the variables need to be *standardised* in some way. This means that variables like income and age need to be converted to something like percentiles or z-scores so that they are measured on the same scale. Otherwise, income will have a much greater influence as it's normally measured in thousands whereas age is normally measured in single or double digits.
- There is a random element to the procedure as the initial cluster centre is randomly chosen. So sorting the data and re-running the procedure may produce slightly different results.
- Because the clusters are formed by centroid calculations, outlier cases may have an increased influence on the cluster formation

Introducing Cluster Analysis

- **Hierarchical Cluster Analysis**
- This is another classic cluster analysis algorithm again designed to work with continuous variables. Hierarchical clustering (sometimes referred to as 'Agglomerative' clustering) works by calculating the 'distance' between every possible combination of cases in the dataset.
- The normal approach to calculating the distance between two cases is to use Squared Euclidean Distance

Student ID	Sociology	Economics	Statistics	Spanish	Geography
1	65%	78%	56%	82%	72%
Squared Difference	**100** +	**16** +	**100** +	**49** +	**4**

= 269 Squared Euclidean Distance

Student ID	Sociology	Economics	Statistics	Spanish	Geography
2	55%	82%	66%	75%	74%

Note that all the variables are in the same units for simplicity sake. As with K-Means Clustering, Hierarchical Cluster Analysis requires that the data values are in the same units.

Introducing Cluster Analysis

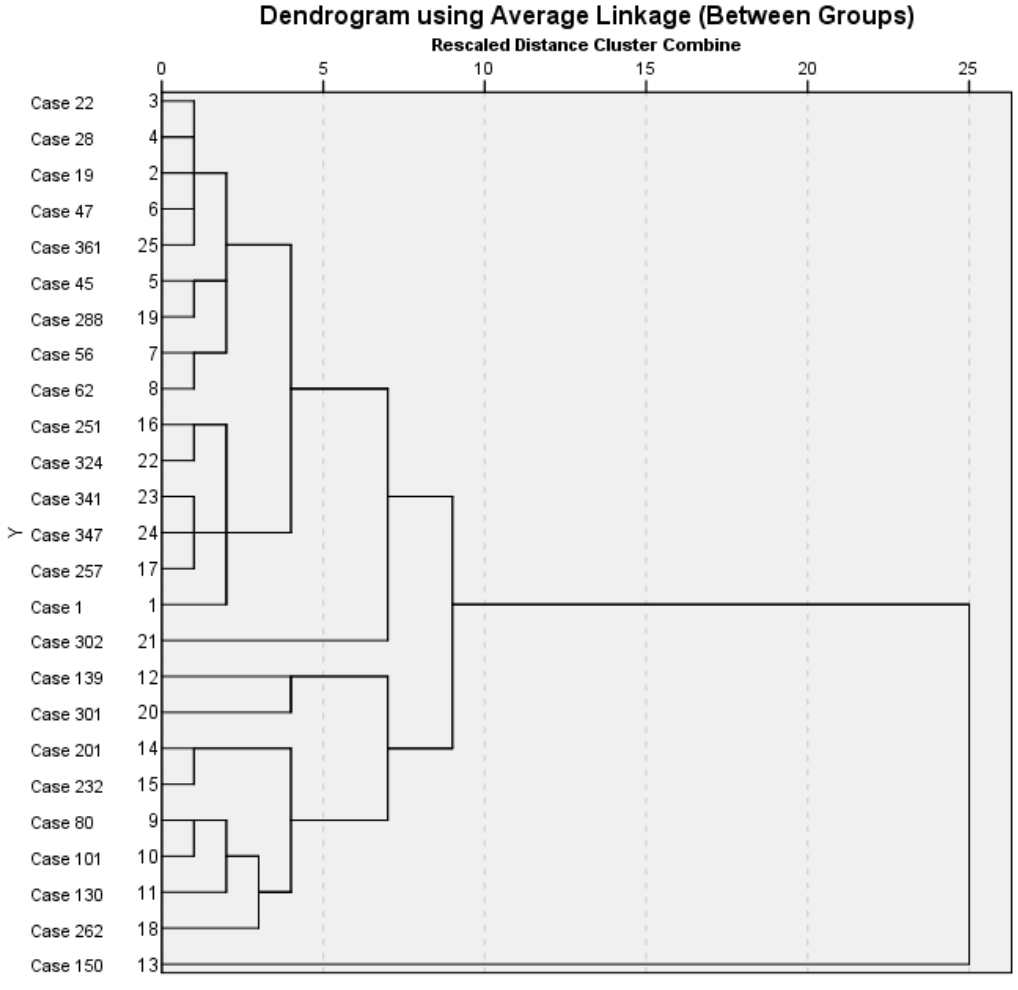
- Hierarchical Cluster Analysis also requires the input data to be standardised, however the procedure includes a setting that will do this for you.
- Because Hierarchical Cluster Analysis needs to calculate the distance between every combination of cases, it means that even with modest data sets of 1,000 records, the number of unique distance calculations in a matrix is equal to 499,500.
- So this technique becomes impractical when dealing with large data files.

Matrix showing distance calculations between 5 cases

	Case 1	Case 2	Case 3	Case 4	Case 5
Case 1	0	5.0	7.8	13.9	11.6
Case 2	5.0	0	7.0	12.4	5.8
Case 3	7.8	7.0	0	9.2	4.9
Case 4	13.9	12.4	9.2	0	22.3
Case 5	11.6	5.8	**4.9**	22.3	0

Introducing Cluster Analysis

- Nevertheless, because this technique can calculate the distances at the finest degree of granularity in the data (i.e. between *individual cases*), it means that it can start to form hierarchies of clusters as the individual cases are grouped together.
- So users can request a *range* of cluster solutions (e.g. 4 to 9 clusters)
- This hierarchical structure is shown in a dendrogram chart.



SMARTVISION Europe

Introducing Cluster Analysis

- **Limitations of Hierarchical Cluster Analysis**

- It requires the user to at least choose the range clusters they want *before* the analysis is run
- It only works with continuous data, count data or binary variables. In whichever case, all the data should be comprised of the same variable type.
- It needs to calculate a matrix that stores the distances between individual cases. This is written out as a temporary file in the background. The size of this file increases exponentially as the number of rows in the original data file increases.

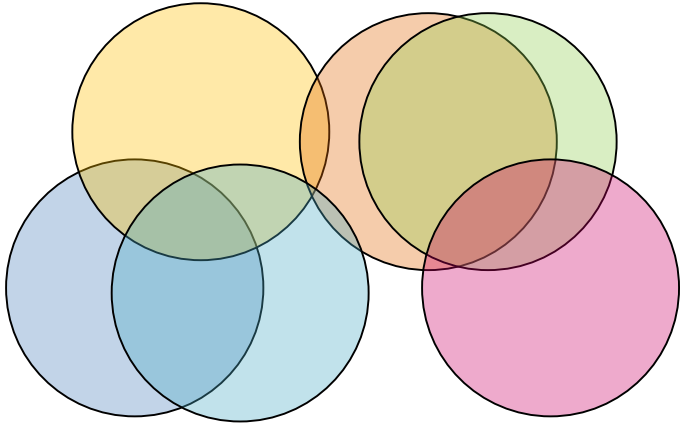
Introducing Cluster Analysis

- **TwoStep Cluster**
- The 2 Step Cluster method avoids the issue of poor scalability that Hierarchical Cluster suffers from. Meaning that it can be used against large datasets.
  1. It gets around this by initially scanning the data to create a ‘tree’ of many ‘leaf’ clusters containing only a few cases in each. This is the first step.
  2. The second step takes the form of a hierarchical clustering process based on the micro-clusters generated in step one.
- Further advantages of the 2 Step Cluster method are that:
  - It allows us to use both continuous and categorical variables
  - It attempts to recommend the ‘optimal’ number of clusters based on a criterion known as the silhouette value. The silhouette value is a measure of how similar a case is to its own cluster (cohesion) compared to other clusters (separation)
  - It automatically standardises the variables so that they are all in the same units
  - It has its own Cluster Viewer application to help us understand the differences between the clusters in the resulting cluster solution

Introducing Cluster Analysis

- **Limitations of TwoStep Cluster**
- Despite its ability to use both categorical and continuous variables, in practice, the results of the TwoStep Cluster method are more heavily affected by the inclusion of categorical variables. Meaning that different combinations of the categorical variables can dominate the results.
- Although TwoStep Clustering will automatically select the 'optimum' number of clusters, often this results in a solution containing only two clusters. Analysts should therefore be prepared to over-rule this automatic setting and experiment with specifying their own required number of clusters.
- The algorithm contains a random element meaning that sorting the data in a different order may produce slightly different cluster solutions.

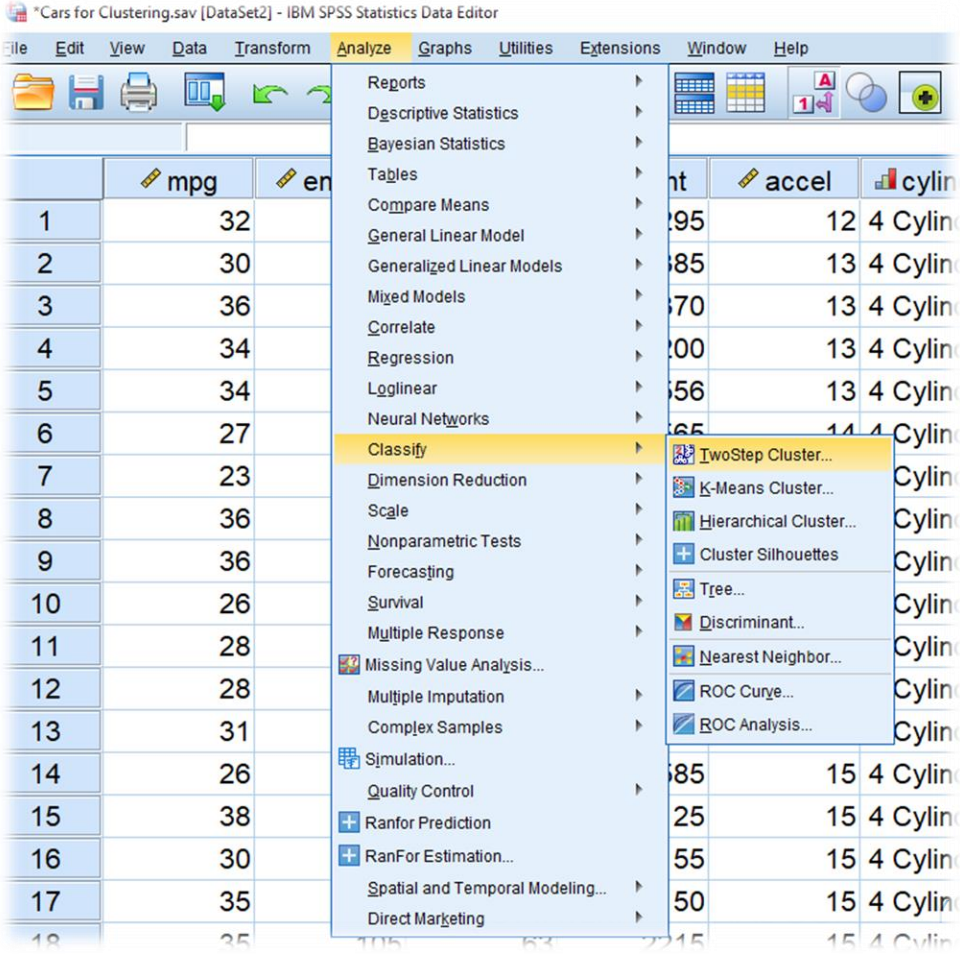
SMARTVISION Europe



Performing Cluster Analysis

Performing Cluster Analysis

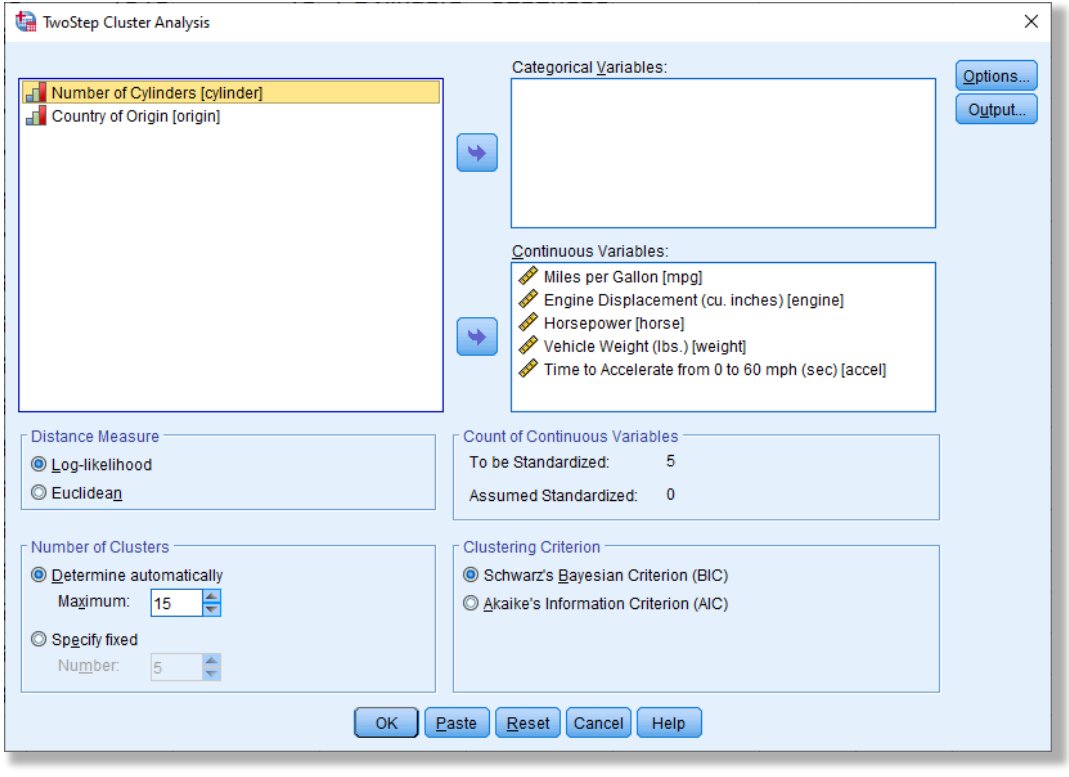
- We are using the **Cars for Clustering.sav** dataset
- This dataset is comprised of variables showing different physical and technical performance measures from 379 automobiles
- To request the **TwoStep Cluster** procedure, from the main menu, click:
- **Analyze**
  - **Classify**
    - **TwoStep Cluster**

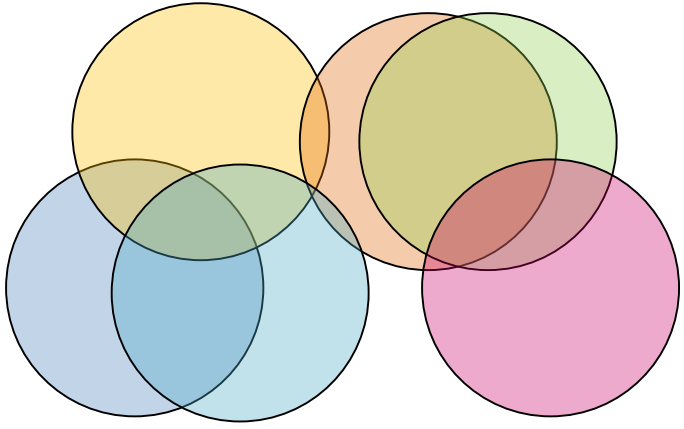


SMARTVISION Europe

Performing Cluster Analysis

- To introduce the procedure, in this example we will only add the continuous variables in the dataset to the box marked:
  - **Continuous Variables**
- To generate the cluster solution, click:
  - **OK**





Interpreting Output

Interpreting Output

- The initial output, displays only the Model Summary showing the three cluster groups were found
- Also the Silhouette measure is above 0.5 meaning that in terms of creating clusters that are cohesive and distinct the solution can be nominally regarded as ‘good’.
- In practice, it’s rare to get a Silhouette measure of this magnitude and analysts should continue to explore the output even if the results are shown to only be ‘fair’.
- To activate the Cluster Viewer:
  - ***Double click on the Model Summary output***

Model Summary

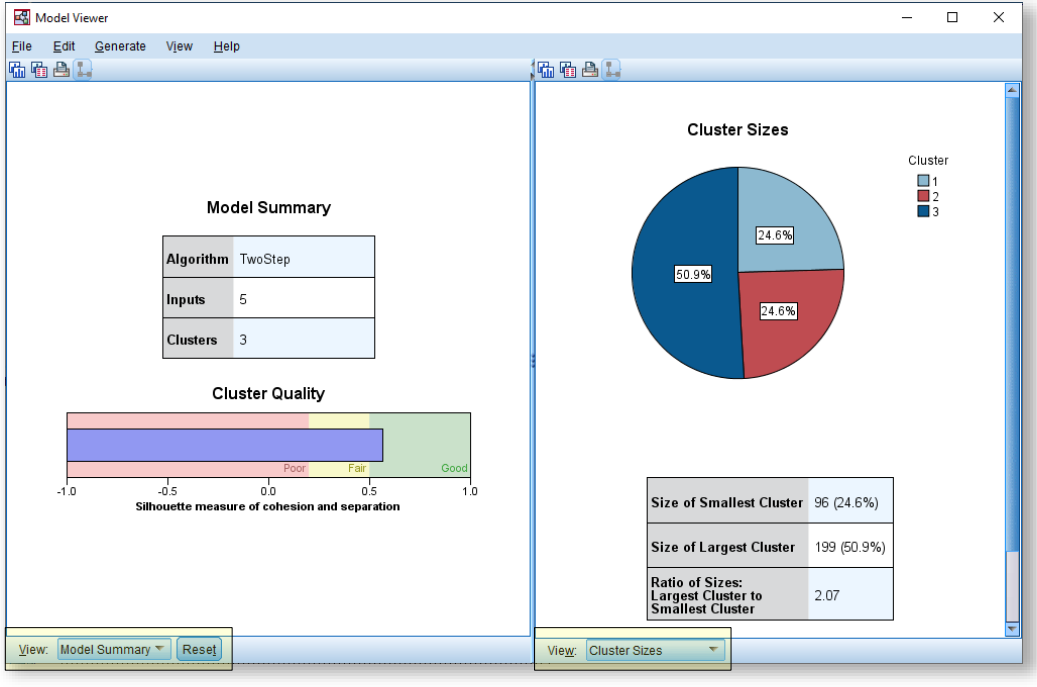
Algorithm	TwoStep
Inputs	5
Clusters	3

Cluster Quality

The silhouette value ranges from −1 to +1, where a high value indicates that the case is well matched to its own cluster and poorly matched to neighbouring clusters. If most cases have a high value, then the clustering configuration is considered appropriate.

Interpreting Output

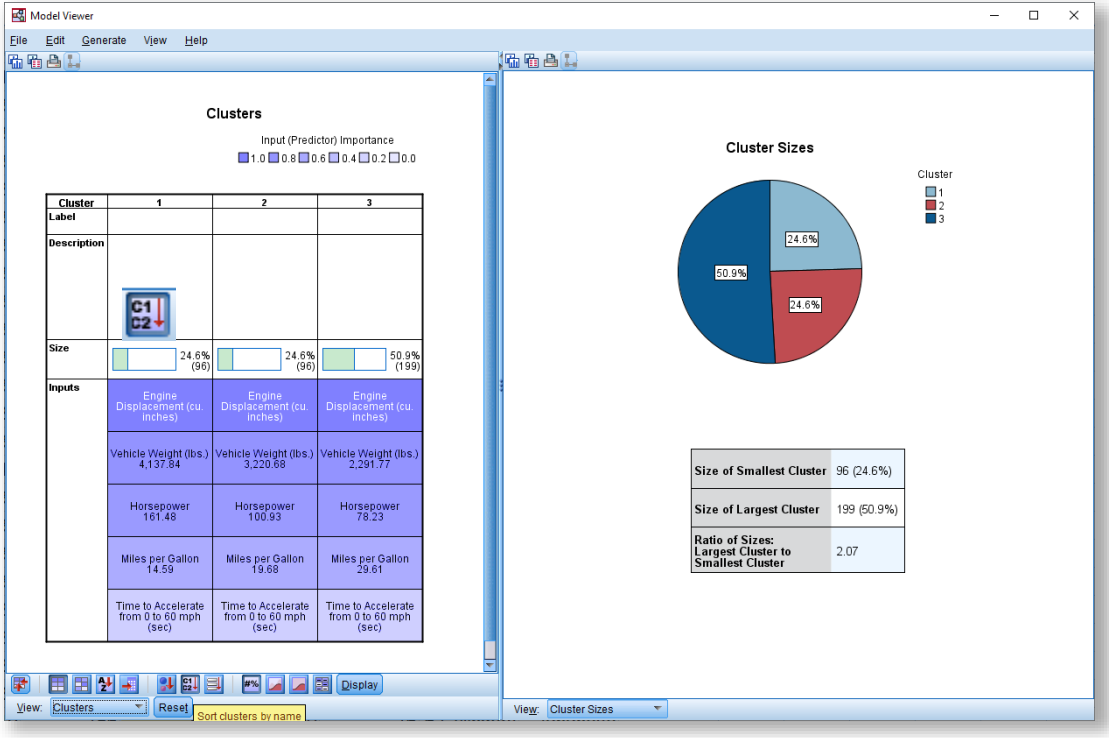
- The Model Viewer opens into its own window. Initially, it only shows us some additional information regarding the relative sizes of the cluster groups.
- Note however that both panes have their own drop down menus allowing us to access additional output.
- To see more details about the Cluster solution, click the drop-down menu labelled:
  - **Model Summary**
- …and select
  - **Clusters**



Interpreting Output

- We now see more information about the clusters. The cluster table shows each cluster's average values for the input variables .
- There are several buttons along the bottom of this pane. To sort the clusters in order of their cluster number click:

- You may notice that the input variables are also sorted, from top to bottom, in terms of their importance to cluster solution



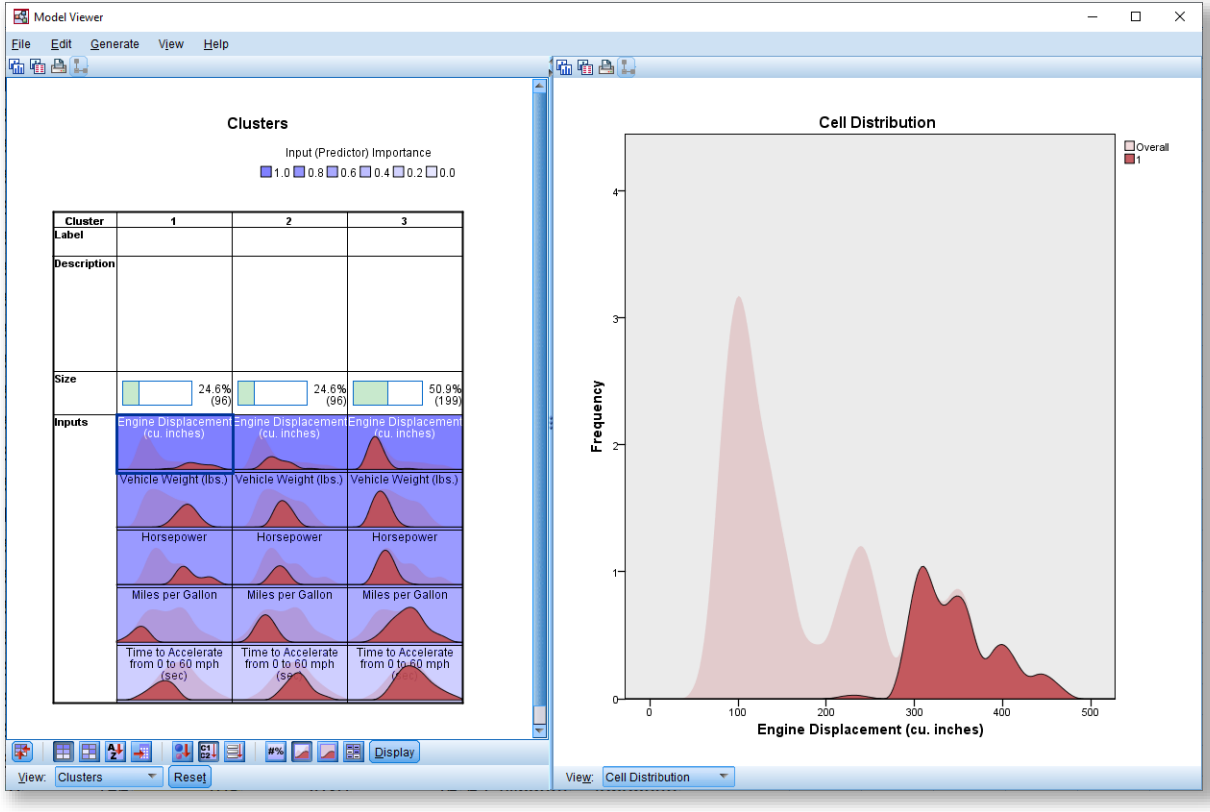
SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Interpreting Output

- There are many different ways in which we can compare the clusters in the Model Viewer. To see a visual representation of their distributions, click the Absolute Distributions button:

- Now click the first cell in the cluster table.
- The Cell Distribution chart is now shown. Indicating that Cluster 1 is comprised of cars with a larger than average engine compared to the overall distribution.

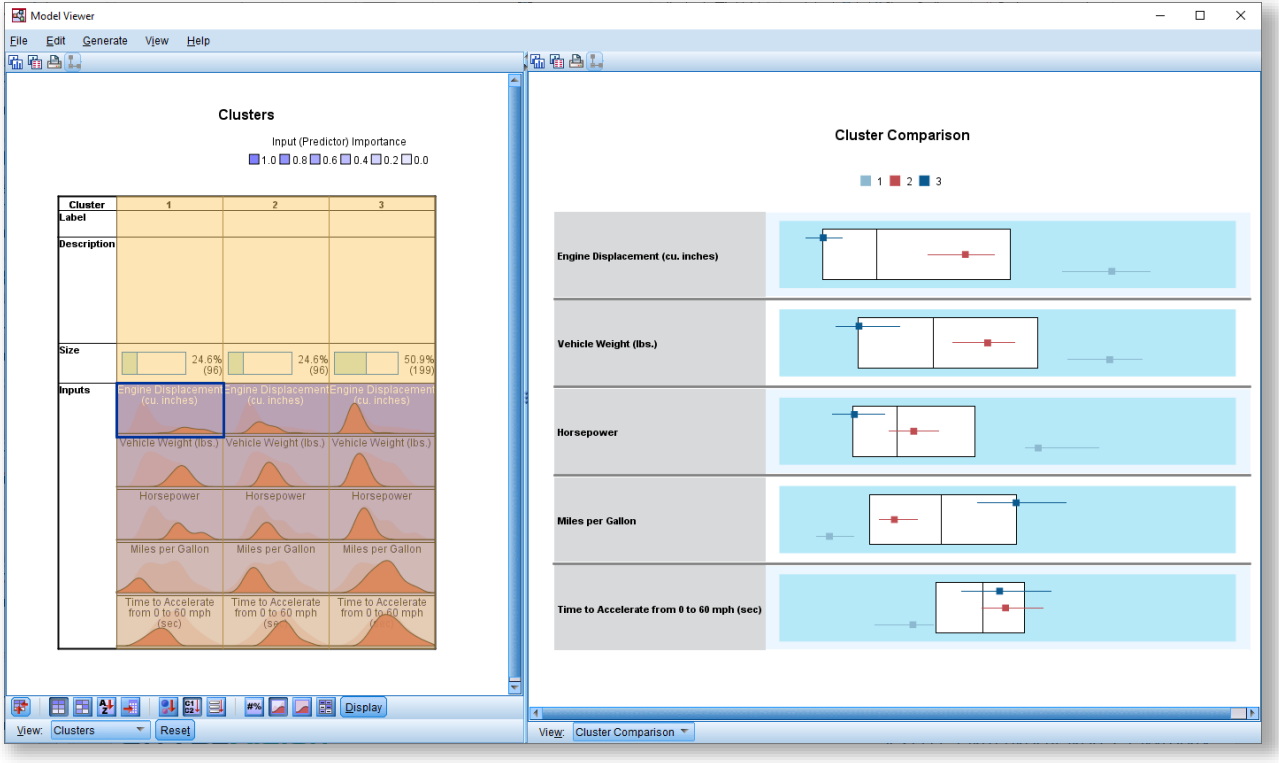


SMARTVISION Europe

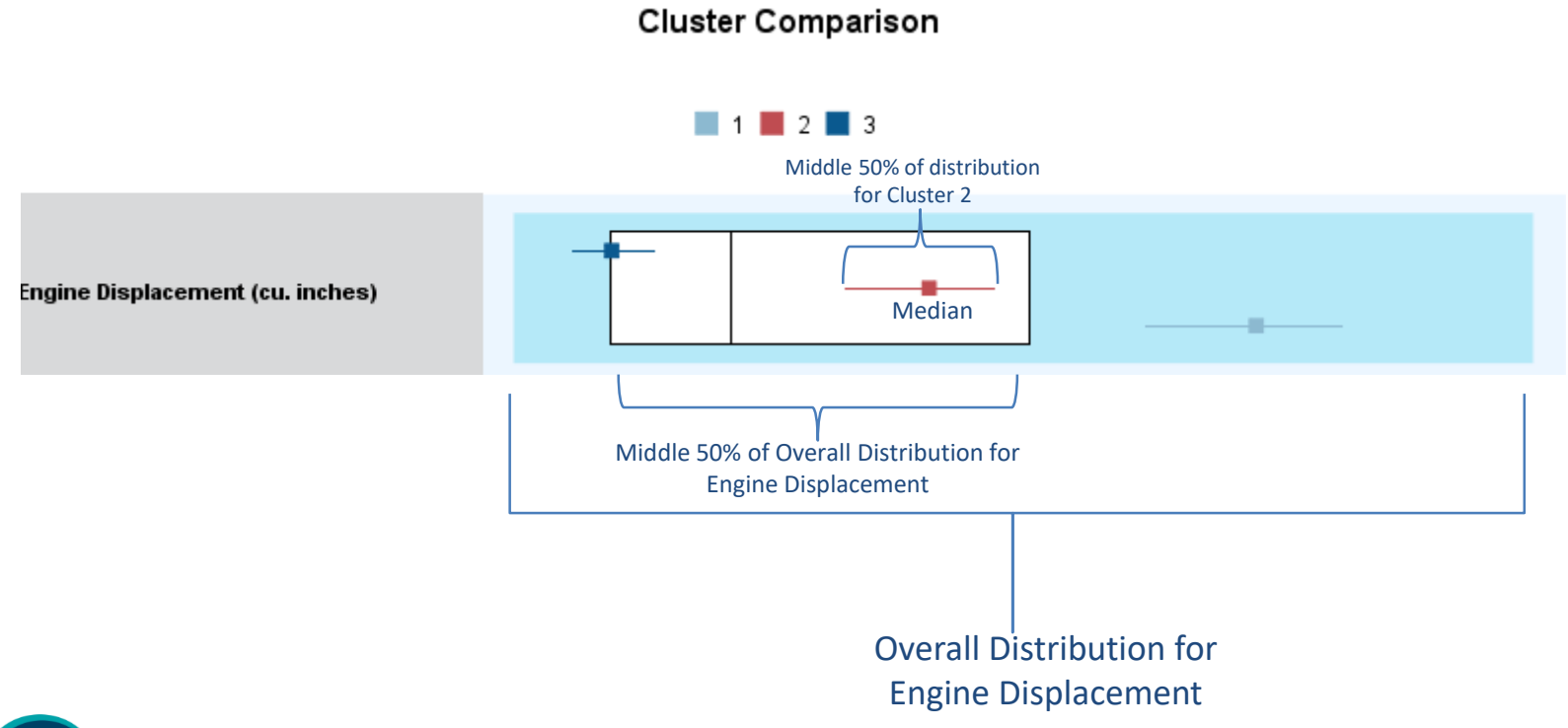
A SELECT INTERNATIONAL COMPANY

Interpreting Output

- One way for us to compare the Cluster groups is to *ctrl-click* the header of each column in the cluster table.
- We can now see the interquartile range of each cluster group depicted as series of charts with the median value shown in the middle

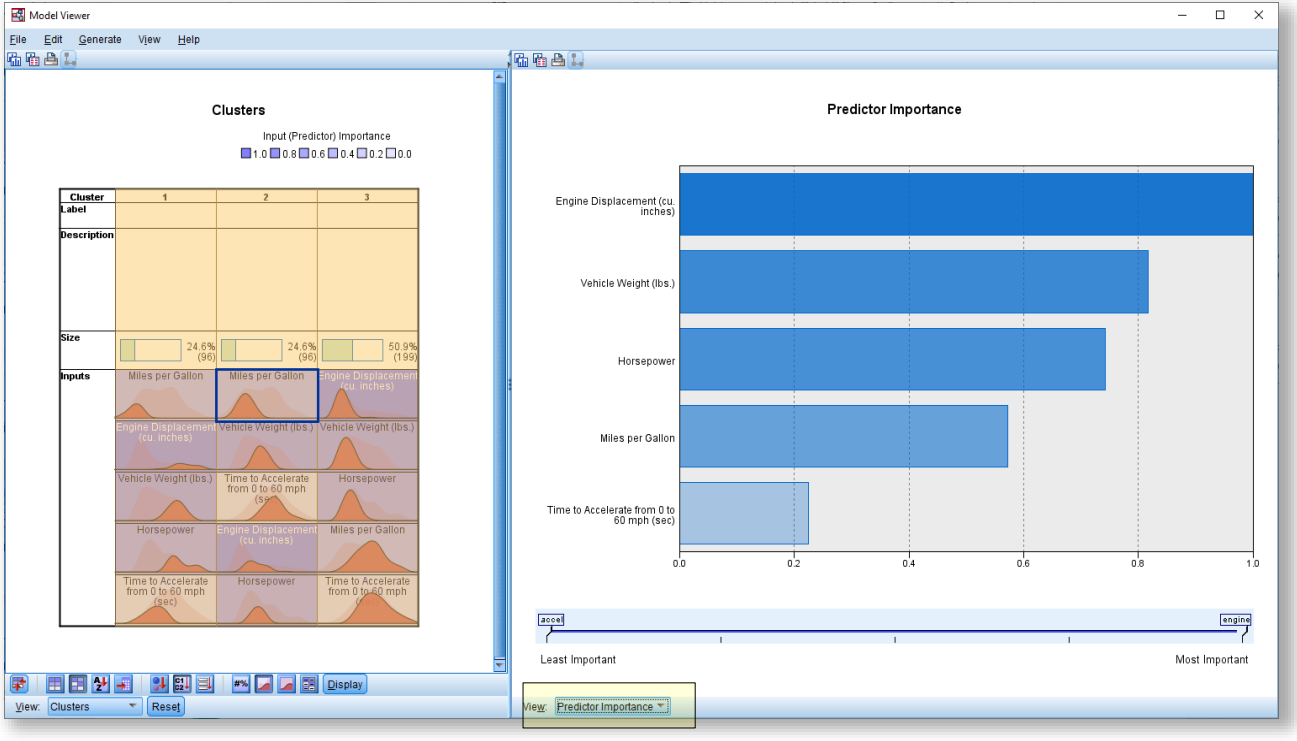


Interpreting Output

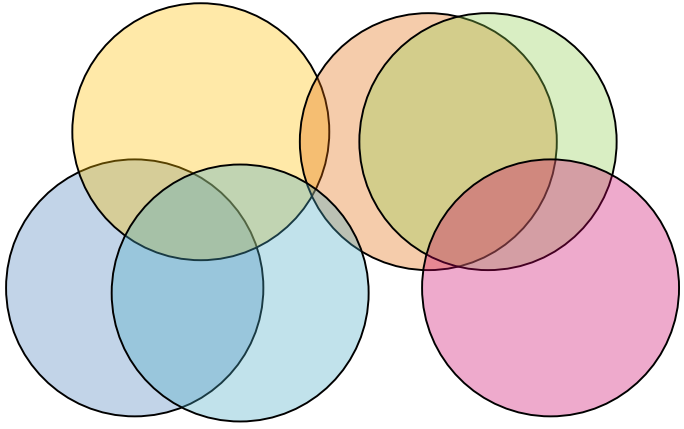


Interpreting Output

- We can also request that the Model Viewer displays a ‘predictor’ importance chart showing the relative importance of each variable in the analysis.
- Finally, we could have requested that the procedure saved the clusters as a cluster membership variable in the dataset so that we could perform our own further analyses, but we will introduce this in our next example.



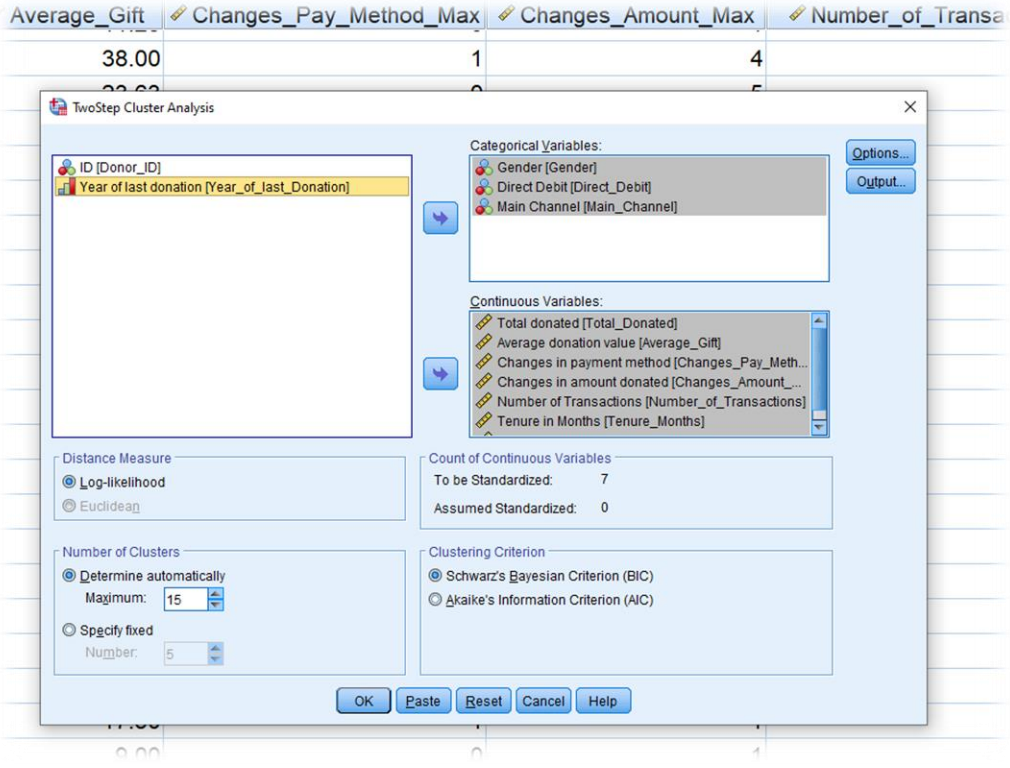
SMARTVISION Europe



Creating Cluster Groupings

Creating Cluster Groupings

- In this last example, we are using the **Charity_Donors_Clustering.sav** dataset
- This file is comprised of variables showing donation data from 4,227 charity supporters over a three year period
- Again employing the TwoStep procedure, we are using all of the continuous variables and assigning them to the Continuous Variables box
- As well as three of the categorical variables: **Gender, Direct Debit**, and **Main Channel**
- Click:
  - **OK**



Creating Cluster Groupings

- The Model Summary shows that the procedure has automatically produced a cluster solution comprised of three groups.
- Let's assume that the analyst suspects that there are more 'natural' supporter segments in the dataset.
- With that in mind we will return to the Cluster dialog and overrule the automatic setting, instead requesting five cluster groups

Model Summary

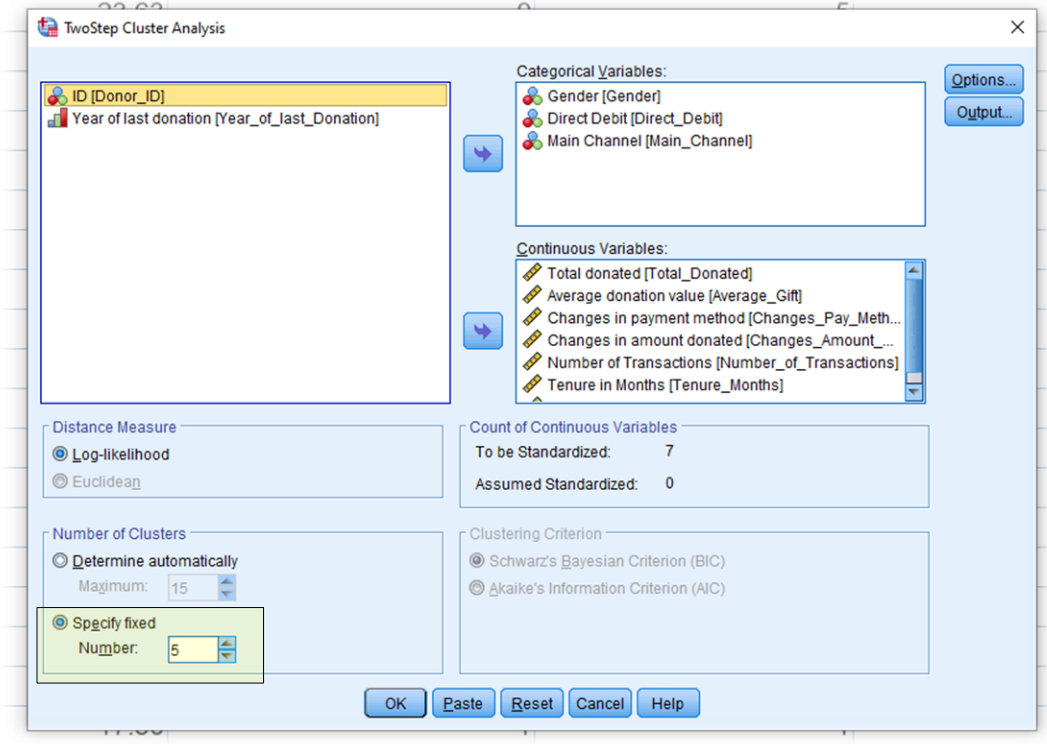
Algorithm	TwoStep
Inputs	10
Clusters	3

Cluster Quality

Creating Cluster Groupings

- In the area of the dialog marked **Number of Clusters**, click the radio button marked **Specify fixed** and enter the value: **5**

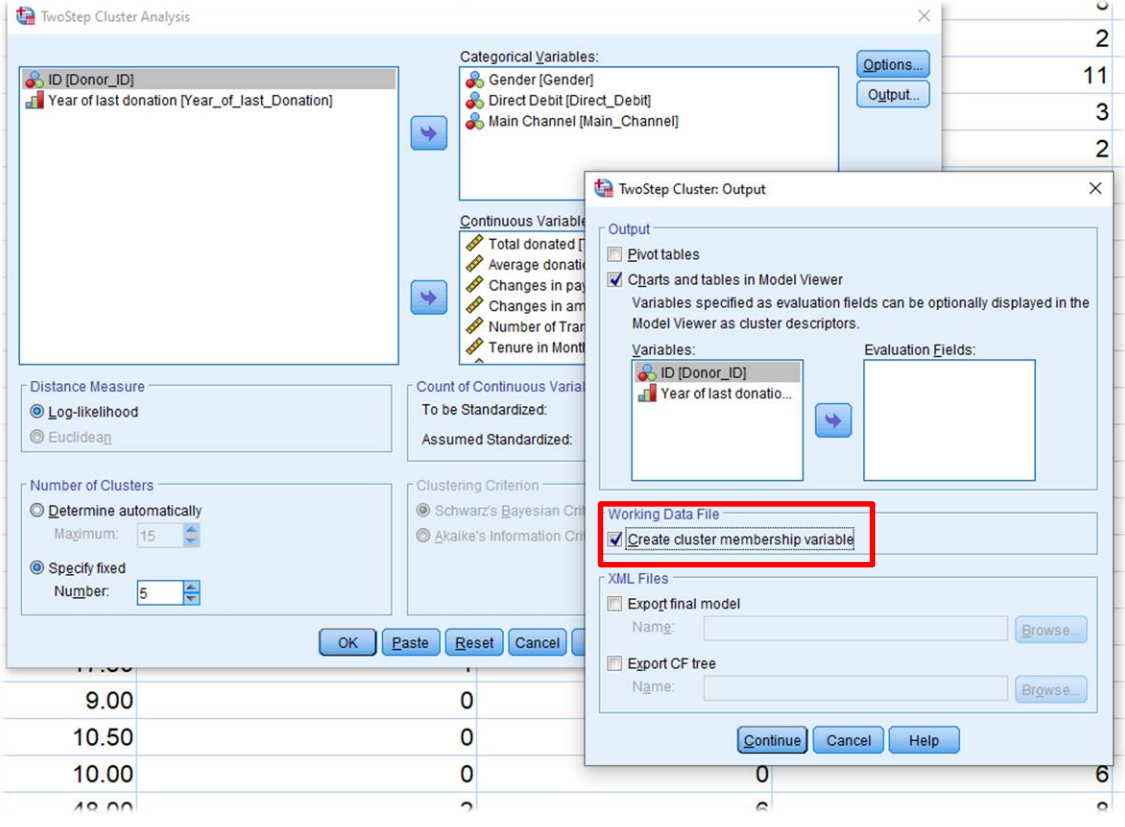
- Before we run the procedure again however, we can request that the analysis also saves the cluster membership as a variable. To do so, click the button marked:
  - **Output**



Creating Cluster Groupings

- In the resulting sub-dialog, check the box marked:
  - **Create cluster membership variable**

- Now click:
  - **Continue**
    - **OK**



Creating Cluster Groupings

- The Model Summary shows a five cluster solution and interestingly, the Silhouette measure has increased slightly.
- Once again, to access the Model Viewer application, double-click on the Model Summary output.

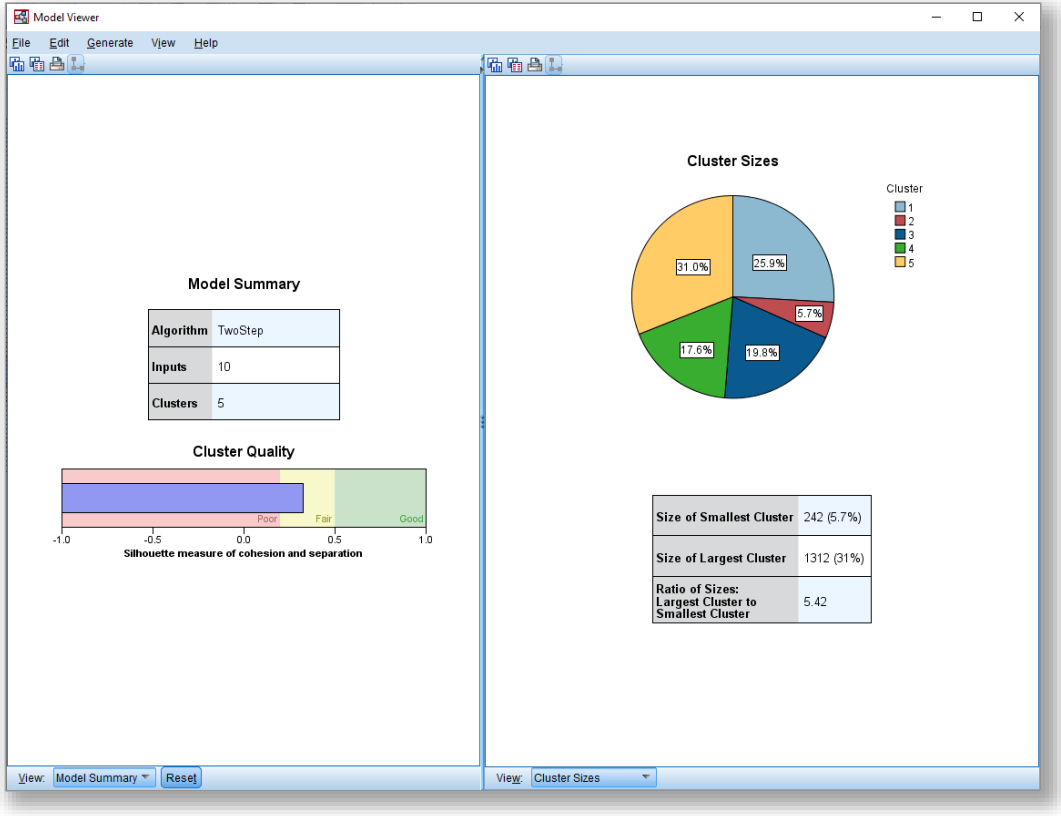
Model Summary

Algorithm	TwoStep
Inputs	10
Clusters	5

Cluster Quality

Creating Cluster Groupings

- The Model Viewer opens into its own window. It shows the proportions of cases that have fallen into the different cluster groups. The smallest group accounts for 5.7% of the data and the largest 31%.
- To see more details about the Cluster solution, click the drop-down menu labelled:
  - **Model Summary**
- …and select
  - **Clusters**



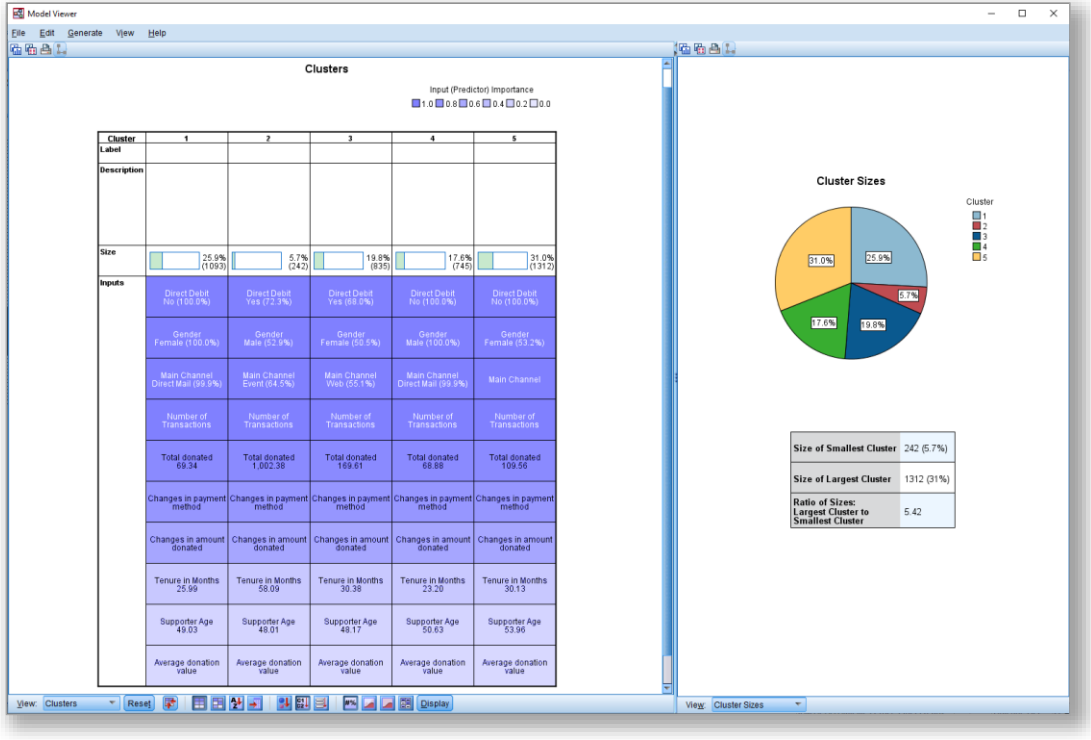
SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Creating Cluster Groupings

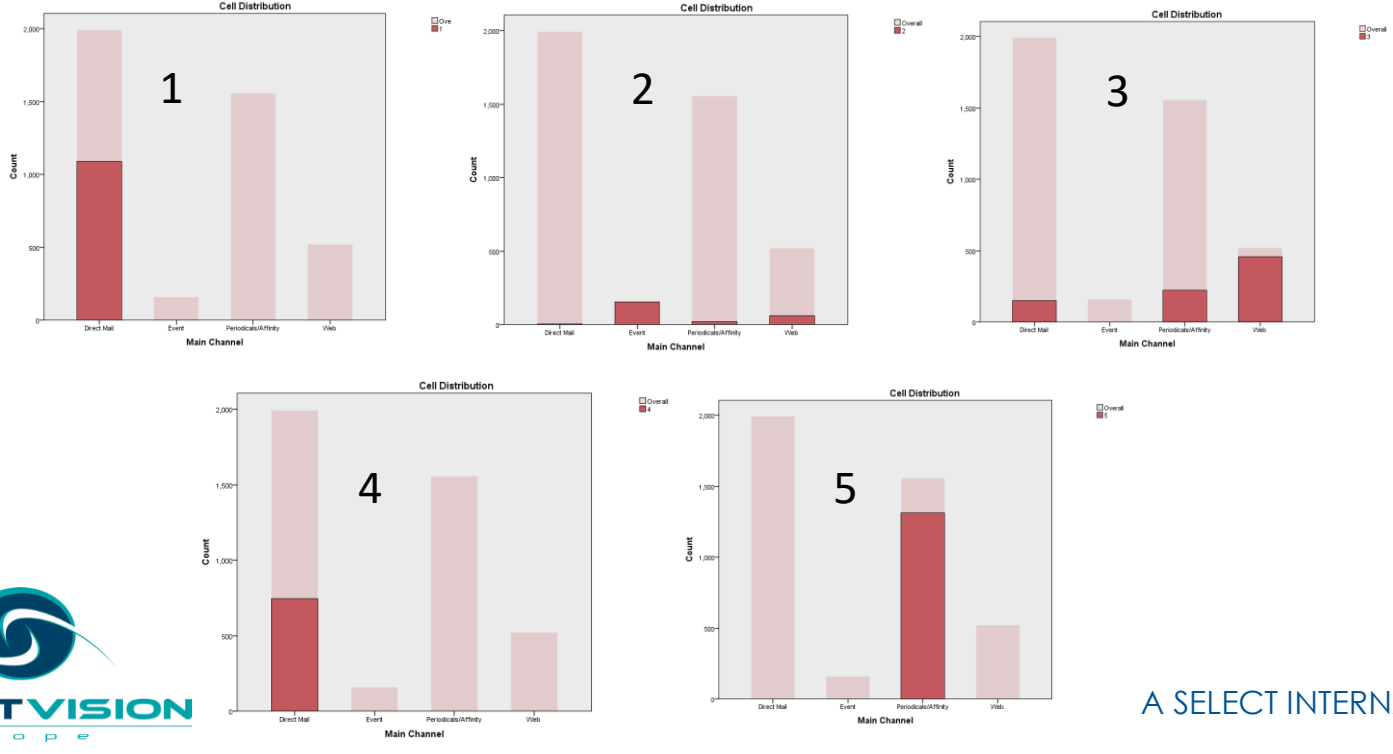
- Again, we can sort the clusters in order of their cluster number by clicking:

- The categorical variables seem to be dominating the cluster solution. We can see that:
  - Cluster groups 2 and 3 are the only ones using direct debit
  - Cluster group 1 is 100% female and cluster group 4 is 100% male



Creating Cluster Groupings

- Cluster groups 1 and 4 are almost exclusively Direct Mail channel users
- Cluster group 2 dominates the Events channel
- Cluster group 3 dominates the web channel
- Cluster group 5 almost exclusively use Periodicals/Affinity channels

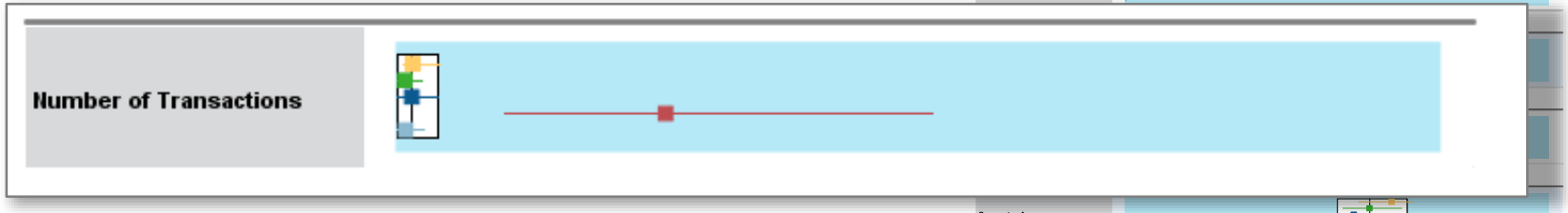


SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Creating Cluster Groupings

- Further analysis shows that cluster group 2 is associated with significantly larger donations, multiple transactions, multiple donation methods and long tenure
- You may recall that this cluster group is also strongly associated with charity events
- Perhaps this group contain supporters who are heavily invested with the charity and directly take part in, or organise sponsored events to raise money



SMARTVISION Europe

Creating Cluster Groupings

- The five cluster solution has also been saved as a new variable in the data file.
- This variable simply marks which cluster group each record belongs to
- It is common practice for analysts to rename these cluster groupings so that they have more meaningful descriptions

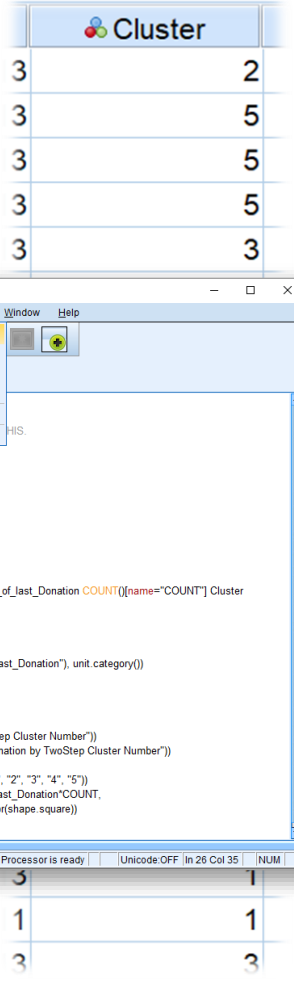
TwoStep Cluster Number

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1093	25.9	25.9	25.9
	2	242	5.7	5.7	31.6
	3	835	19.8	19.8	51.3
	4	745	17.6	17.6	69.0
	5	1312	31.0	31.0	100.0
	Total	4227	100.0	100.0	

	TSC_7459
3	2
3	5
3	5
3	5
3	3
2	1
3	3
1	5
3	3
3	5
2	5
2	5
2	1
3	5
2	1
3	2
3	5
3	3
3	1
1	1
3	3
2	3
2	1
2	1

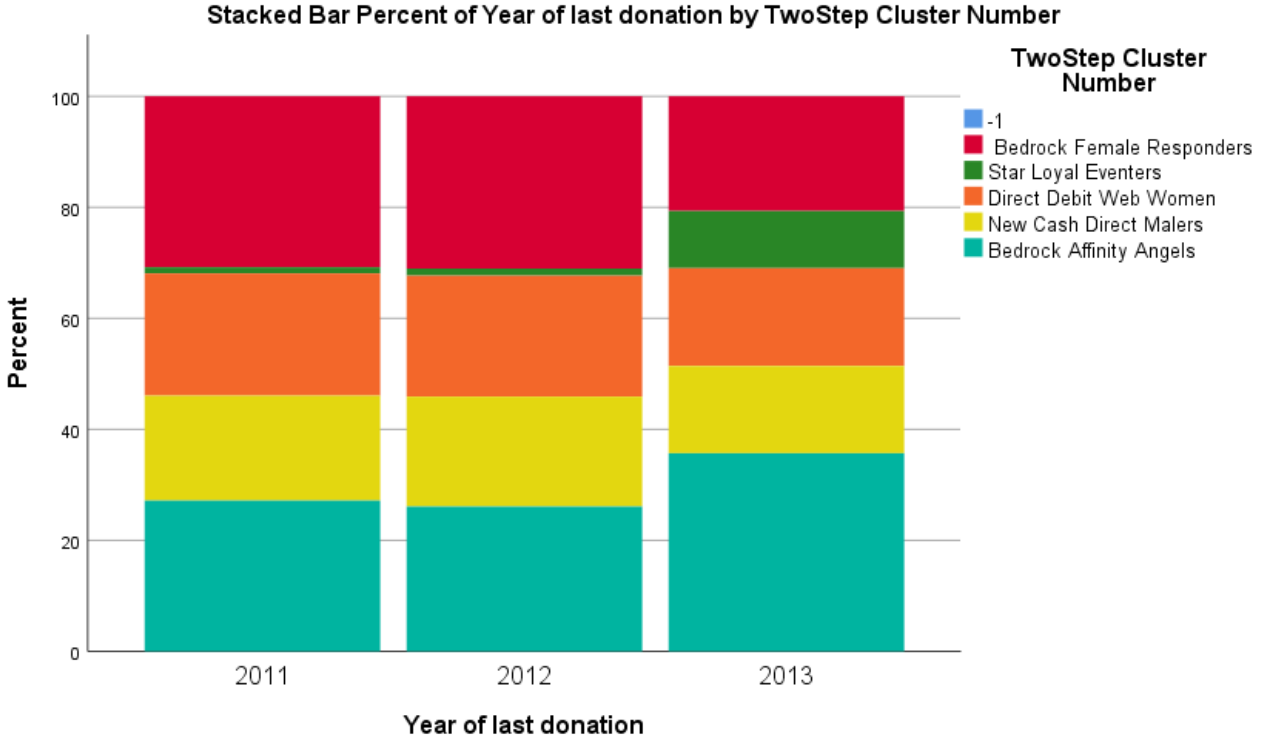
Creating Cluster Groupings

- We can run some prepared SPSS syntax to add some value labels generate a chart.
- Before we do so, we should rename the cluster variable first (as the name changes with each analysis).
    - Double- click on the variable name in the column header and change the name simple to **'cluster'.**
    - Open the SPSS syntax file **'Cluster Bar Chart Charity Group by Years.sps'** and execute it
    - To do so, from the main menu in the syntax window, click:
        - **Run**
            - **All**



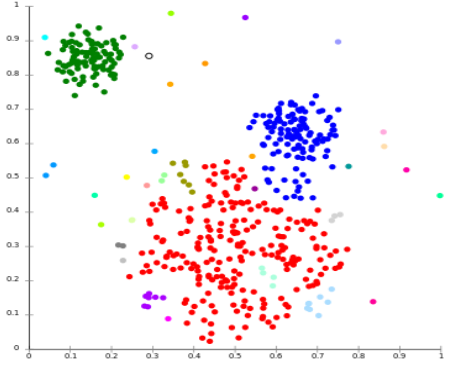
Creating Cluster Groupings

- The resulting stacked bar chart shows that the group ‘Bedrock Affinity Angels’ (formerly cluster group 5) increased their presence in 2013.
- Also, in the same year, we can see the emergence of the ‘Star Loyal Eventers’ as a significant segment



Benefits and Challenges of Cluster Analysis

- **Benefits**
  - Can find very subtly different groups that might otherwise have been missed
  - Lets the dataset tell its own story
  - Can utilise multiple dimensions/variables
  - Can be combined with Factor Analysis
- **Challenges**
  - Need to determine how many segments there should be
  - Needs care when interpreting
  - Can't automatically select the most useful variables
  - Highly correlated variables may be measuring the same thing
  - Not always obvious how to exploit the results



Working with Smart Vision Europe Ltd.

- **Sourcing Software**
  - You can buy your analytical software from us often with discounts
  - Assist with selection, pilot, implementation & support of analytical tools
  - http://www.sv-europe.com/buy-spss-online/
- **Training and Consulting Services**
  - Guided consulting & training to develop in house skills
  - Delivery of classroom training courses / side by side training support
  - Identification & recruitment of analytical skills into your organisation
- **Advice and Support**
  - offer 'no strings attached' technical and business advice relating to analytical activities
  - Technical support services

SMARTVISION Europe

Contact us:

+44 (0)207 786 3568
info@sv-europe.com
Twitter: @sveurope
Follow us on Linked In
Sign up for our Newsletter

Thank you