FACTOR ANALYSIS USING IBM SPSS – OVERVIEW WITH MARKET SEGMENTATION EXAMPLE.

USE OF FACTOR ANALYSIS IN MARKET SEGMENTATION STUDIES 1
WHAT TO LOOK FOR WHEN RUNNING FACTOR ANALYSIS 2
PRINCIPLES 2
Factor Analysis and Principal Component Analysis 3
NUMBER OF FACTORS 4
ROTATIONS 4
FACTOR SCORES 5
SAMPLE SIZE 5
METHODS 6
Overall Recommendations 6
AN EXAMPLE: IMPORTANCE OF BENEFITS 7
Looking at Correlations 9
Running Principal Components Analysis 10
Extraction and Rotation 13
Reviewing the Principal Components Analysis 16
Clustering Based on Components 17
Creating Factor Score Variables 17
K-Means Clustering of Component Scores 19

Introduction

In this session we will discuss factor analysis primarily as a data reduction technique in support of segmentation analyses (cluster) and response modeling (for example, logistic regression and discriminant analysis). We will first review the role of factor in segmentation studies (when to use it), and then discuss what to look for when running factor. Some background principles of factor analysis will be covered along with comments about popular factor methods, and overall recommendations are made. Next we perform a principal components analysis on customer ratings of benefits/features and create factor score variables that can be used as the basis of additional analysis (for example, cluster analysis). Finally, we will explore the value of clustering of the factor score variables as a further step in understanding these techniques.

Use of Factor Analysis in Market Segmentation Studies

In the area of market segmentation, factor analysis typically serves in the ancillary role of reducing the many variables available for the purpose of segmentation to a core set of composite variables (factors) that are used by cluster, discriminant or logistic regression.

In some surveys done for segmentation purposes, dozens of customer attitude measures or product attribute ratings may be collected. Although cluster analysis can be run using a large number of cluster variables, two complications can develop. First, if several variables measure the same or very similar characteristics, and are included in a cluster analysis, then what they measure is weighted more heavily in the analysis. For example, suppose a set of rating questions about technical support are used in a cluster analysis with other unrelated questions. Since distance calculations are based on the differences between observations on each variable, then other things being equal, the set of related items would carry more weight in the analysis.

To exaggerate to make a point, if two variables were identical copies of each other and both were used in a cluster analysis, the effect would be to double the influence of what they measure. In practice you rarely ask the same number of rating questions about each attribute (or psychographic) area. So factor is used to explicitly combine the variables into independent composite variables, to guide the analyst in constructing subscales, or to aid in selection of representative sets of variables (some analysts select three variables strongly related to each factor to be used in cluster analysis). Cluster is then performed on these variables.

Similarly, in response-based segmentation methods (see Chapter 5) such as discriminant and logistic regression, highly correlated predictor variables can yield an unstable solution (the problem of multicollinearity). Prior data reduction using factor or principal components analysis is one approach to reducing this risk.

A second reason factor might be run prior to clustering or response-based segmentation is for conceptual clarify and simplification. If a cluster analysis is based on forty variables, it is difficult to look at so large a table of means or a line chart and make much sense of them. As mentioned in the last chapter, you can perform ANOVA or discriminant analysis on the clusters, or ask for importance plots in TwoStep in order to identify the influential variables and then summarize those. If factor analysis is run first, then the clustering or logistic regression is done based on the themes or concepts measured by the factors. Or as mentioned above, clustering can be done on equal-sized sets of variables, where each set is based on a factor. If the factors have a ready interpretation, it can be much easier to understand a solution based on five or six factors, compared to one based on forty variables. As you might expect, factor analysis is more often performed on “soft” measures—attitudes, beliefs, and attribute ratings— and less often on behavioral measures like usage and purchasing patterns.

Keep in mind that factor analysis is considered an exploratory data technique (although there are confirmatory factor methods; for example, *Amos* can be used to test specific factor models). So as with cluster, do not expect a definitive, unassailable answer. When deciding on the number and interpretation of factors, knowledge of the customer base from which the data are taken, common sense, and a dose of hard thinking are very valuable.

What to Look for When Running Factor Analysis

There are two main questions that arise when running factor analysis: how many (if any) factors are there, and what do they represent? These questions are related because in the practice of market research you rarely retain factors that you cannot identify and name. Although the naming of factors has rarely stumped a creative market researcher for long, which has led to some very odd-sounding “factors,” it is accurate enough to say that interpretability is one of the criteria when deciding to keep or drop a factor. When choosing the number of factors there are some technical aids (eigenvalues, percentage of variance accounted for) we will discuss, but they are guides and not absolute criteria.

To interpret the factors a set of coefficients, called factor loadings or lambda coefficients, relating the factors to the variables, are very important. They provide information as to which factors are highly related to which variables and thus give insight into what the factors represent.

Principles

Factor analysis operates on the correlation matrix relating the variables to be factored. The basic argument is that the variables are correlated because they share one or more common components, and if they didn’t correlate there would be no need to perform factor analysis. Mathematically, a one-factor model for three variables can be represented as follows (Vs are variables, Fs are factors, Es represent variation that is unique to each variable…uncorrelated with the E component of the others):

$$\mathbf{V_1 = L_1 * F_1 + E_1}$$

$$V_2 = L_2 * F_1 + E_2$$
$$V_3 = L_3 * F_1 + E_3$$

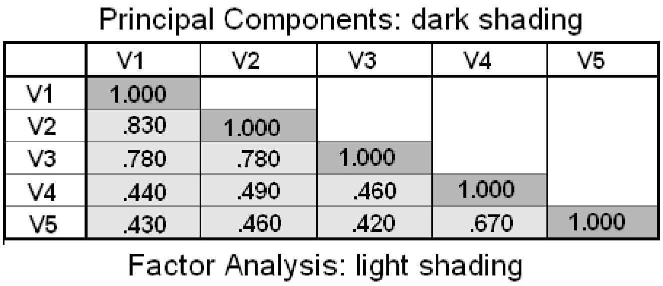
Each variable is composed of the common factor (F_1) multiplied by a loading coefficient (L_1, L_2, L_3 - the lambdas) plus a unique or random component. If the factor were measurable directly (which it isn't) this would be a simple regression equation. Since these equations can't be solved as given (the Ls, Fs and Es are unknown), factor analysis takes an indirect approach. If the equations above hold, then consider why variables V_1 and V_2 correlate. Each contains a random or unique component that cannot contribute to their correlation (Es are assumed to have 0 correlation). However, they share the factor F_1, and so if they correlate the correlation should be related to L_1 and L_2 (the factor loadings). If this logic is applied to all the pairwise correlations, the loading coefficients can be estimated from the correlation data. One factor may account for the correlations between the variables, and if not, the equations can be easily generalized to accommodate additional factors. There are a number of approaches to fitting factors to a correlation matrix (least squares, generalized least squares, maximum likelihood), which has given rise to a number of factor methods.

What is a factor? In market research factors are usually taken to be underlying traits, attitudes or beliefs that are reflected in specific rating questions. You need not believe that factors actually exist in order to perform a factor analysis, but in practice the factors are usually interpreted, given names and generally spoken of as real things.

Factor Analysis and Principal Component Analysis

Within the general area of data reduction there are two highly related techniques: factor analysis and principal components analysis. They can both be applied to correlation matrices with data reduction as a goal. They differ in a technical way having to do with how they attempt to fit the correlation matrix. We will pursue the distinction since it is relevant to which method you choose. The diagram below is a correlation matrix composed of five variables.

Figure 1 Correlation Matrix



Principal Components: dark shading

	V1	V2	V3	V4	V5
V1	1.000				
V2	.830	1.000			
V3	.780	.780	1.000		
V4	.440	.490	.460	1.000	
V5	.430	.460	.420	.670	1.000

Factor Analysis: light shading

Principal components analysis attempts to account for the maximum amount of variation in the set of variables. Since the diagonal of a correlation matrix (the ones) represents standardized variances, each principal component can be though of as accounting for as much as possible of the variation remaining in the diagonal. Factor analysis, on the other hand, attempts to account for correlations between the variables, and its focus is on the off-diagonal elements (the correlations). So while both methods attempt to fit a correlation matrix with fewer components or factors than variables, they differ in what they focus on when fitting. Of course, if a principal component accounts for most of the variance in variables V_1 and V_2 it must also account for much of the correlation between them. And if a factor accounts for the correlation between V_1 and V_2 it must account for at least some of their variance. Thus, there is overlap in the methods and in practice they usually yield similar results.

Often factor is used when there is interest in studying relations among the variables, while principal components is used when there is a more emphasis on data reduction and less on interpretation. In market research, principal components is very popular because it can run even when the data are multicollinear (one variable can be perfectly predicted from the others), while most factor methods cannot. Since in marketing studies many related attribute questions are asked of relatively few subjects the data are likely to be multicollinear or near multicollinear, principal components is used routinely. Both methods are available in the Factor procedure in SPSS; in fact, by default, Factor will run the principal components method.

Number of Factors

When factor or principal components analysis is run, there are several technical measures that can guide you in choosing a tentative number of factors or components. The first indicator would be the ***eigenvalues.*** *Eigenvalues are fairly technical measures, but in principal components analysis, and some factor methods (under orthogonal rotations), their values represent the amount of variance in the variables that is accounted for by the component (or factor).*

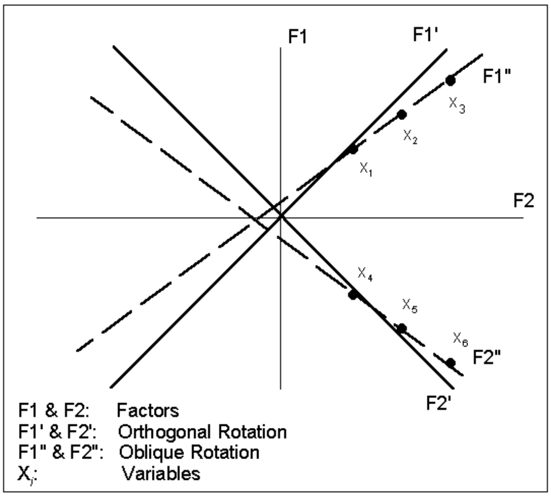
If we turn back to the correlation matrix in Figure 1, there are five variables and therefore 5 units of standardized variance to be explained. Each eigenvalue measures the amount of this variance accounted for by a factor. This leads to a rule of thumb and a useful measure to evaluate a given number of factors. The rule of thumb is that there are as many factors as there are eigenvalues greater than 1. Why? If the eigenvalue represents the amount of standardized variance in the variables accounted for by the factor, then if it is above 1, it must represent variance contained in more than one variable. This is because the maximum amount of standardized variance contained in a single variable is 1. Thus if in our five-variable analysis the first eigenvalue is 3, it must account for variation in several variables. An eigenvalue can be less than 1 and still account for variation shared among several variables (for example 1/3 of the variation of each of three variables for an eigenvalue of .9), so the eigenvalue of 1 rule is only applied as a rule of thumb and not the final word.

Another aspect of eigenvalues (for principal components and some factor methods) is that their sum is the same as the number of variables, which is equal to the total standardized variance in the variables. Thus you can convert the eigenvalue into a measure of percentage of variance accounted for, which is helpful when evaluating a solution. Finally it is important to mention that in market research and segmentation work, the factors must make sense. ***For this reason, factors with eigenvalues over 1 that cannot be interpreted may be dropped and those with eigenvalues less than 1 may be retained.***

Rotations

When factor analysis succeeds, you obtain a relatively small number of interpretable factors that account for much of the variation in the original set of variables. Suppose you have six variables and factor returns a two-factor solution. Formally, the factor solution represents a two-dimensional space. Such a space can be represented on this page with a pair of axis as shown below.

Figure 2 Two Factors in a Two Dimensional Space



While each pair of axes defines the same two-dimensional space, the coordinates of a point would vary depending on which pair of axes was applied. This creates a problem for factor since the values for the loadings or lambda coefficients vary with the orientation of axes and there is no unique orientation defined by the factor analysis itself. Principal components does not suffer from this problem since its method produces a unique orientation. This difficulty for factor is a fundamental mathematical problem. The solutions to it are designed to simplify the task of interpretation for the analyst. Most involve, in some fashion, finding a rotation of the axes that maximizes the variance of the loading coefficients, so some are large and some small. This makes it easier for the analyst to interpret the factors. The fact that factor loadings are not uniquely determined by the method is a valid criticism leveled against it by some statisticians. We will discuss the various rotational schemes in the Methods section below.

Factor Scores

If you are satisfied with a factor analysis or principal components solution, you can request that a new set of variables be created that represent the scores of each observation on the factors. These are calculated by summing the product of each original variable (in z-score form) and a weight coefficient (derived from the lambda coefficients). These factor variables can then be used as the input variables for segmentation analysis. They are usually normalized to have a mean of zero and standard deviation of one. An alternative some analysts prefer is to use the lambda coefficients to judge which variables are highly related to a factor, and then compute a new variable which is the sum or mean of that set of variables. This method, while not optimal in a technical sense, keeps (if means are used) the new scores on the same scale as the original variables (this of course assumes the variables themselves share the same scale), which can make the interpretation and more importantly, the presentation, more straightforward. Essentially, subscale scores are created based on the factor results, and these scores are used for segmentation purposes.

Sample Size

Since factor is a multivariate statistical method, the rule of thumb for sample size (commonly violated) is that there should be from 10 to 25 times as many observations as there are variables used in the factor analysis. This is because factor is based on correlations and for p variables there are p*(p–1)/2 possible correlations. Think of this as a desirable goal and not a formal requirement (technically if there are p variables there must be p+1 observations for factor to run, but don't expect reasonable results in that situation). **If your sample size is very small relative to the number of variables, you should turn to principal components.**

Methods

There are several popular methods within the domain of factor and principal components analysis. The common factor methods differ in how they go about fitting the correlation matrix. A traditional method that has been around for many years, and for some it means factor analysis, is **principal axis** factoring (often abbreviated as PAF). A more modern method that carries some technical advantages is **maximum likelihood** factor analysis. If the data are ill behaved (say near multicollinear), maximum likelihood, the more refined method, is more prone to give strange solutions. In most cases results using the two methods will be very close, so either is fine under general circumstances. If you suspect there are problems with your data, then principal axis may be a safer bet. The other factor methods are considerably less popular.

One factor method, called **Q factor** analysis, involves transposing the data matrix and then performing a factor analysis on the respondents instead of the variables. Essentially, correlations are calculated for each pair of subjects based on their responses to the variables. This technique is related to cluster analysis, but is used infrequently today. Besides the factor methods, principal components can be run and, as mentioned earlier, must be run when the variables are multicollinear.

Similarly, there are several choices in rotations. The most popular is the **varimax** rotation, which attempts to simplify the interpretation of the factors by maximizing the variances of the variable loadings on each factor. In other words, it attempts to finds a rotation in which some variables have high and some low loadings on each factor which makes it easier to understand and name each factor. The **quartimax** rotation attempts to simplify the interpretation of each variable in terms of the factors by finding a rotation yielding high and low loadings across factors for each variable. The **equimax** rotation is a compromise between the other two rotation methods. These three rotations are **orthogonal**, which means the axes are perpendicular to each other and the factors will be uncorrelated. This is considered a desirable feature in market research since statements can be made about independent factors or aspects of the data.

There are non-orthogonal rotations available (axes are not perpendicular), the most popular ones are **oblimin** and **promax** (runs faster than oblimin). Such rotations are rarely used in market research, since the point of data reduction is to obtain relatively independent composite measures and it is easier to speak of independent effects when the factors are uncorrelated.

Principal components does not require a rotation, since there is a unique solution associated with it. However, in the practice of market research a varimax rotation is often done to facilitate the interpretation of the components.

Overall Recommendations

For market segmentation and market research in general, principal components is usually performed because of the expected high correlations among the many attribute measures that are commonly analyzed. Varimax rotation is usually done (although it is not necessary for principal components) to simplify the interpretation. If there are not many highly correlated variables (or other sources for ill-behaved data, for example, much missing data), then either principal axis or maximum likelihood factor can be performed. Maximum likelihood has technical advantages, but can produce a strange solution if the data are not well conditioned.

An Example: Importance of Benefits

We will apply factor analysis to a set of 23 survey questions in which customers rated the importance of different benefits of shopping at a superstore. Data are contained in an SPSS data file called *Benefit_factor.sav*. The data file contains 361 respondents who rated benefits on a 1 (Not Important at All) to 7 (Very Important) scale. Only the endpoints (1,7) were labeled.

The list contains questions with potentially some overlap in the areas of convenience, service, economics, quality and availability of goods, and so forth. We intend to use these questions to segment the customers, and cluster analysis can be applied directly to these items. To the extent that each question represents a distinct characteristic, this is just what you want to do. However, if you suspect or discover high correlation or redundancy among the items, then for reasons of overweighing issues created by correlated questions and potentially greater clarity in clustering by a few themes, you will likely want to apply factor analysis.

Table 1 Variables and Labels for Survey Questions

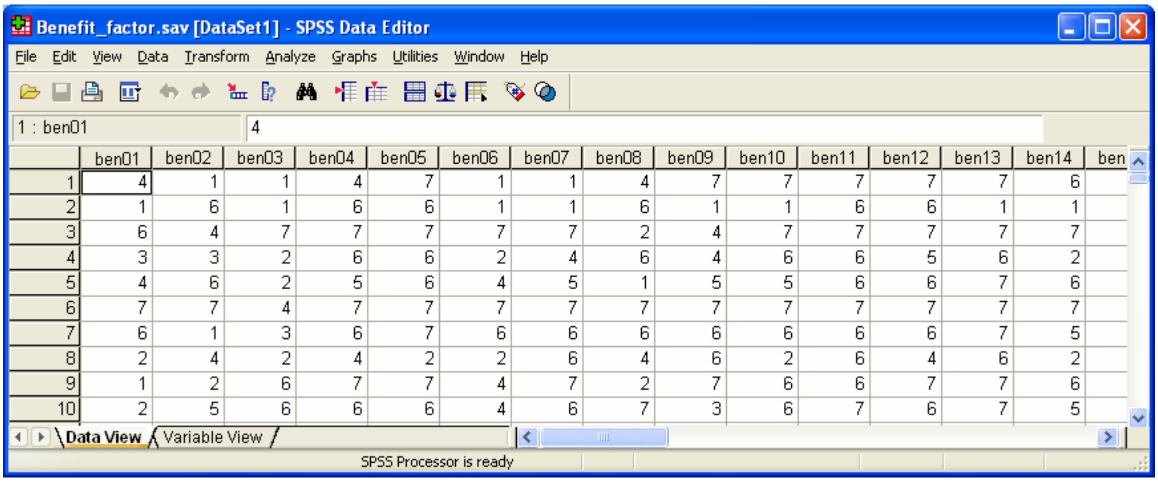
ben01	Customer services readily available
ben02	Knowledgeable personnel
ben03	Good selection of DVDS
ben04	Clean, inviting appearance
ben05	Well laid out -easy to locate things
ben06	Employees circulating to help
ben07	Good selection of household goods
ben08	Shopping for whole family
ben09	Wide selection of manufacturers
ben10	Advertised items in stock
ben11	Open convenient hours
ben12	Carrying sizes you need
ben13	Regular lower prices
ben14	Helpful store employees
ben15	Easy to complete transaction
ben16	Good selection of electronics
ben17	Stocking high quality goods
ben18	Good selection of holiday goods
ben19	Easy parking near store
ben20	Short wait at checkout
ben21	Store located nearby
ben22	Friendly environment
ben23	Roomy aisles

Assuming useful factors emerge, you may then cluster the factor scores, cluster subscale scores based on the factors, or cluster balanced sets of variables derived from the factors. Some situations are clear-cut: If only four questions are asked then factor analysis is unlikely to be necessary, while if there are 60 product attribute ratings then factor is appropriate. In fact, nothing but limited time prevents you from running cluster both ways (on original variables and on factor scores) when you are unsure what to do. In our example, we will factor first, and then cluster the factor scores.

You might ask yourself, or your client: should the segments be based on specific questions (stick to the original variables) or on more general characteristics and themes (factor first)?

Click File..Open..Data
Move to the **c:\Train\Mkseg** directory (if necessary)
Double-click on Benefit_factor.sav

Figure 3 Benefit Study File in Data Editor


Benefit_factor.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : ben01
4
Data View
Variable View
SPSS Processor is ready

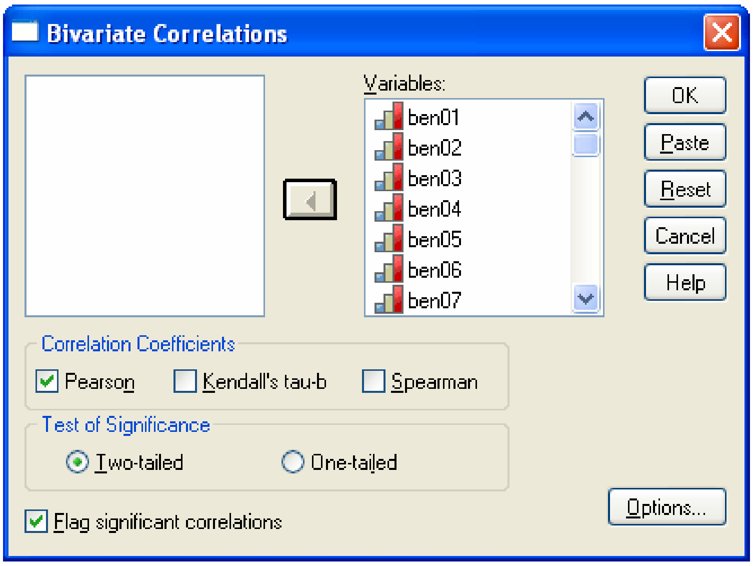
	ben01	ben02	ben03	ben04	ben05	ben06	ben07	ben08	ben09	ben10	ben11	ben12	ben13	ben14	ben
1	4	1	1	4	7	1	1	4	7	7	7	7	7	6	
2	1	6	1	6	6	1	1	6	1	1	6	6	1	1	
3	6	4	7	7	7	7	7	2	4	7	7	7	7	7	
4	3	3	2	6	6	2	4	6	4	6	6	5	6	2	
5	4	6	2	5	6	4	5	1	5	5	6	6	7	6	
6	7	7	4	7	7	7	7	7	7	7	7	7	7	7	
7	6	1	3	6	7	6	6	6	6	6	6	6	7	5	
8	2	4	2	4	2	2	6	4	6	2	6	4	6	2	
9	1	2	6	7	7	4	7	2	7	6	6	7	7	6	
10	2	5	6	6	6	4	6	7	3	6	7	6	7	5	

Looking at Correlations

In many marketing studies where many attribute ratings are collected, the analyst knows that the ratings will be correlated. Although we certainly expect this here, we will first run correlations to gauge how strong the relationships are among the benefit questions. It may give us a hint whether factor analysis will work.

Click Analyze..Correlate..Bivariate
Move **ben01** to **ben23** into the **Variables** list box

Figure 4 Completed Correlation Dialog



Variables:
ben01
ben02
ben03
ben04
ben05
ben06
ben07

Correlation Coefficients
☑ Pearson ☐ Kendall's tau-b ☐ Spearman

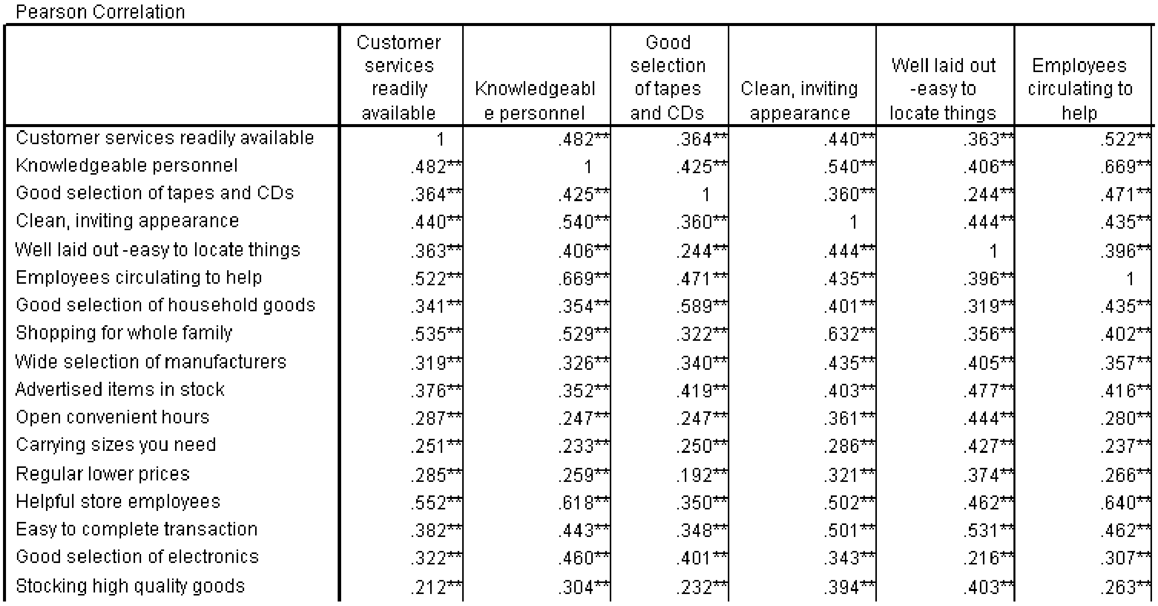
Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations

By default, Pearson product moment correlation coefficients will be calculated. For rank ordered data, SPSS can produce Spearman rank correlations and Kendall's coefficients of concordance.

Click **OK**

Figure 5 Correlations of Benefit Questions (Beginning)



Pearson Correlation

	Customer services readily available	Knowledgeable personnel	Good selection of tapes and CDs	Clean, inviting appearance	Well laid out -easy to locate things	Employees circulating to help
Customer services readily available	1	.482**	.364**	.440**	.363**	.522**
Knowledgeable personnel	.482**	1	.425**	.540**	.406**	.669**
Good selection of tapes and CDs	.364**	.425**	1	.360**	.244**	.471**
Clean, inviting appearance	.440**	.540**	.360**	1	.444**	.435**
Well laid out -easy to locate things	.363**	.406**	.244**	.444**	1	.396**
Employees circulating to help	.522**	.669**	.471**	.435**	.396**	1
Good selection of household goods	.341**	.354**	.589**	.401**	.319**	.435**
Shopping for whole family	.535**	.529**	.322**	.632**	.356**	.402**
Wide selection of manufacturers	.319**	.326**	.340**	.435**	.405**	.357**
Advertised items in stock	.376**	.352**	.419**	.403**	.477**	.416**
Open convenient hours	.287**	.247**	.247**	.361**	.444**	.280**
Carrying sizes you need	.251**	.233**	.250**	.286**	.427**	.237**
Regular lower prices	.285**	.259**	.192**	.321**	.374**	.266**
Helpful store employees	.552**	.618**	.350**	.502**	.462**	.640**
Easy to complete transaction	.382**	.443**	.348**	.501**	.531**	.462**
Good selection of electronics	.322**	.460**	.401**	.343**	.216**	.307**
Stocking high quality goods	.212**	.304**	.232**	.394**	.403**	.263**

Pivoting was used on this table to show the coefficients only.

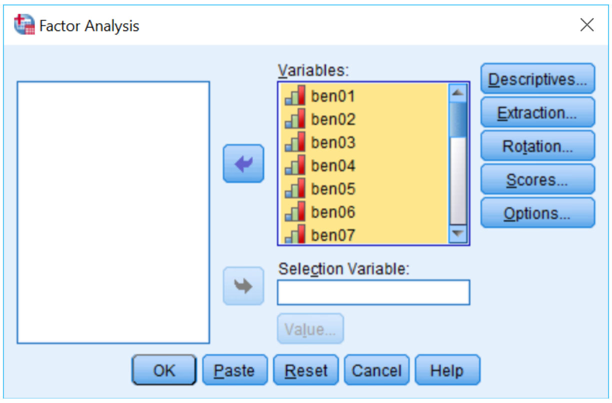
We see that all the visible correlations are positive, significant (at the .01 level indicated by the asterisks), and many are in the range of .3 to .6. Thus there are definitely relations among questions. Therefore, Factor Analysis as a data reduction method can be applied.

Running Principal Components Analysis

We will run principal components analysis instead of one of the factor methods because principal components is the data reduction method used most often in market research studies. We will approach this first step as an exploratory analysis by requesting a diagnostic plot (scree plot).

Click Analyze..Dimension Reduction ..Factor
Move **ben01** to **ben23** into the **Variables** list box

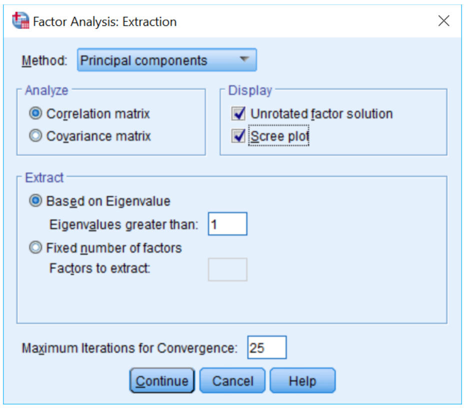
Figure 6 Factor Analysis Dialog



We could run the analysis at this point, but we will request some options that should help us in deciding on the solution. The *Descriptives* button can be used to provide such descriptive statistics as correlations and means. Since we have already viewed some of the correlations we will skip this. The *Extraction* button leads us to the dialog in which we choose a factor method.

Click the **Extraction** button
Click the **Scree Plot** check box in the **Display** area

Figure 7 Factor Analysis: Extraction Dialog Box



Factor Analysis: Extraction

Method: Principal components

Analyze
- ● Correlation matrix
- ○ Covariance matrix

Display
- ☑ Unrotated factor solution
- ☑ Scree plot

Extract
- ● Based on Eigenvalue
  - Eigenvalues greater than: 1
- ○ Fixed number of factors
  - Factors to extract:

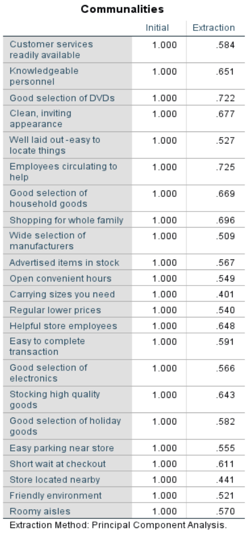
Maximum Iterations for Convergence: 25

The default method used by the Factor procedure is the principal components method. If we wished to run a different factor method, such as principal axis or maximum likelihood, we could choose it from the *Method* drop-down list. Note that by default, correlations will be analyzed.

The *Extract* area indicates that SPSS will select as many factors as there are eigenvalues over 1 (we discussed this rule of thumb earlier in the chapter). Notice you can change this rule or specify a number of factors. You might do this if you prefer more or fewer factors than the eigenvalue rule provides. We ask for a scree plot, which some analysts use to guide their choice of the number of factors. By default, the unrotated solution will appear.

Click **Continue**, click **OK**

Figure 8 Communalities



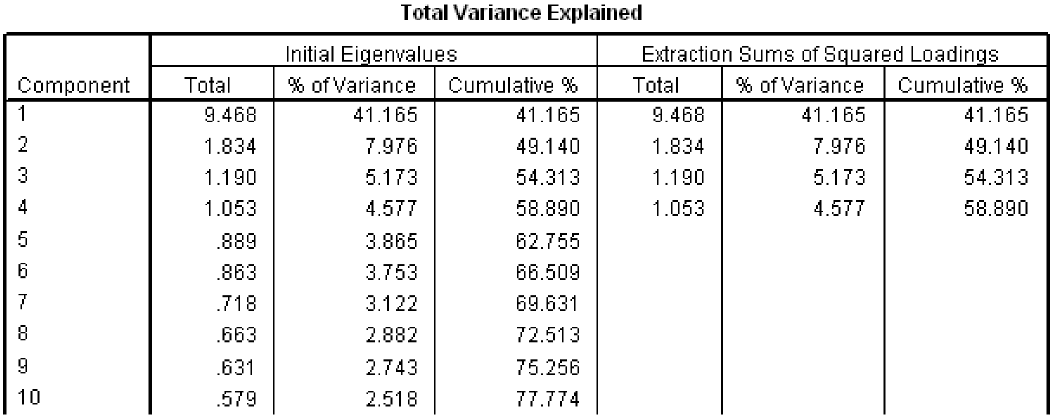
Communalities

	Initial	Extraction
Customer services readily available	1.000	.584
Knowledgeable personnel	1.000	.651
Good selection of DVDs	1.000	.722
Clean, inviting appearance	1.000	.677
Well laid out -easy to locate things	1.000	.527
Employees circulating to help	1.000	.725
Good selection of household goods	1.000	.669
Shopping for whole family	1.000	.696
Wide selection of manufacturers	1.000	.509
Advertised items in stock	1.000	.567
Open convenient hours	1.000	.549
Carrying sizes you need	1.000	.401
Regular lower prices	1.000	.540
Helpful store employees	1.000	.648
Easy to complete transaction	1.000	.591
Good selection of electronics	1.000	.566
Stocking high quality goods	1.000	.643
Good selection of holiday goods	1.000	.582
Easy parking near store	1.000	.555
Short wait at checkout	1.000	.611
Store located nearby	1.000	.441
Friendly environment	1.000	.521
Roomy aisles	1.000	.570

Extraction Method: Principal Component Analysis.

The communalities represent the proportion of variance in a variable explained by the factors (here principal components). Since initially there are as many components as there are variables, the communalities in the *Initial* column are trivially 1. They are of interest when a solution is reached (*Extraction* column). Here the communalities are below 1 and measure the percentage of variance in each variable that is accounted for by the selected number of components (four as we will see shortly). Any variables having very small communalities (say .3 or below) have little in common with the other variables, and are neither explained by the components (or factors), nor contribute to their definition.

Figure 9 Total Variance Explained (Beginning)

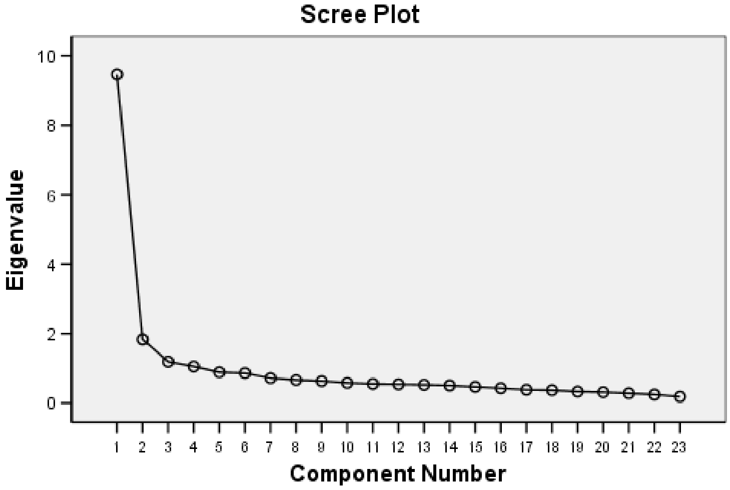


Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	9.468	41.165	41.165	9.468	41.165	41.165
2	1.834	7.976	49.140	1.834	7.976	49.140
3	1.190	5.173	54.313	1.190	5.173	54.313
4	1.053	4.577	58.890	1.053	4.577	58.890
5	.889	3.865	62.755			
6	.863	3.753	66.509			
7	.718	3.122	69.631			
8	.663	2.882	72.513			
9	.631	2.743	75.256			
10	.579	2.518	77.774			

The *Initial Eigenvalues* area contains all (23) eigenvalues, along with the percentage of variance (of the variables) explained by each, and a cumulative percentage of variance. We see in the section *Extracted Sums of Squared Loadings* that there are four eigenvalues over 1. The first component is quite large. Four components were selected and they collectively account for about 59 percent of the variance of the 23 variables. Also, although SPSS selects four principal components, this is not clear-cut, since one eigenvalue is barely above 1 (1.053). As explained above, since all our eigenvalues sum to 23, any factor or component with an eigenvalue of less than 1 explains less variance than a single variable, which is the reason for the eigenvalue>1 selection rule. As you would expect, if there were three or four components with large eigenvalues, we would pay little attention to the eigenvalues hovering near 1. Unfortunately, this is a relatively rare occurrence in survey and market research.

Figure 10 Scree Plot



The scree plot is used to aid the decision about the number of factors (or components) to select. It consists of a plot containing eigenvalues on the vertical axis and the factor (component) numbers on the horizontal axis. What we look for is an elbow, or a flattening bend **(scree is the loose rock debris at the base of a cliff),** which would suggest a transition from large eigenvalues on the left to the very small ones appearing on the right side of the plot. In our graph there is such a bend between the second and third factor, and an even greater flattening after the third, which might indicate a two-factor or a three-factor solution.

The number of components to choose is still not obvious. One reason is that the first factor is so much larger than all the others that it creates a very steep drop-off (essentially there is one large component, then everything else), which distorts the elbow. Also, the first two components by themselves account for nearly half the variance. However, if we are performing principal components analysis with the goal of employing the results in clustering for segmentation purposes, cluster analysis based on just two components may not be provide opportunity for segments to express themselves. So we reject accepting only two.

Looking back to the Total Variance Explained table in Figure 9, there is an argument for choosing three components. By the third component, total variance explained has surpassed 50%. And the fourth component adds only 5%.

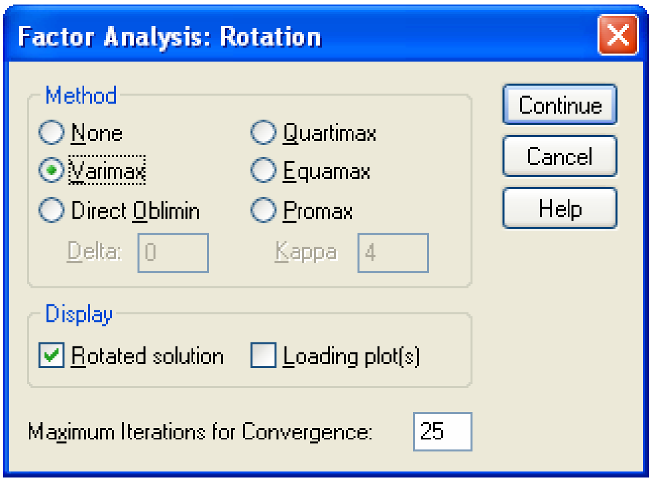
However, another strategy is to continue with the analysis of all four components by examining the results after rotation to see if all four are interpretable.

Extraction and Rotation

To follow this line of analysis, we'll need to rerun Factor Analysis and ask for a rotated solution.

Click the Dialog Recall tool, click Factor Analysis (not shown)
Click the **Rotation** button
Click the **Varimax** option button

Figure 11 Factor Analysis: Rotation Dialog



Factor Analysis: Rotation

Method
○ None ○ Quartimax
● Varimax ○ Equamax
○ Direct Oblimin ○ Promax
Delta: 0 Kappa 4

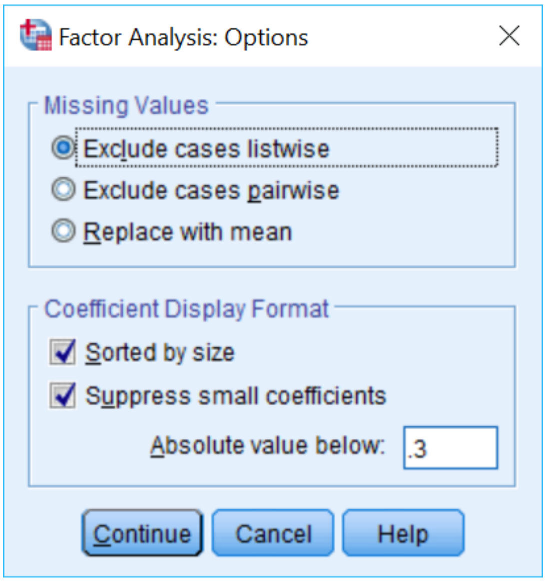
Display
☑ Rotated solution □ Loading plot(s)

Maximum Iterations for Convergence: 25

The most popular rotation (Varimax) has been checked to simplify the task of interpreting the factors. Recall that principal components does have a unique rotation associated with it, so we are rotating for the benefit of easier interpretation of the components. By default, the solution (loadings) will appear in a table. The loadings can be plotted as well, but with so many variables the plot might be too cluttered to be of much use. As the final step, we will pick some formatting options that organize and simplify the factor loading results.

Click Continue
Click the **Options** button
Click the **Sorted by size** check box
Click the Suppress absolute values less than check box
Type .3 in the Suppress absolute values less than: text box

Figure 12 Factor Analysis: Options Dialog



Factor Analysis: Options

Missing Values

● Exclude cases listwise

○ Exclude cases pairwise

○ Replace with mean

Coefficient Display Format

☑ Sorted by size

☑ Suppress small coefficients

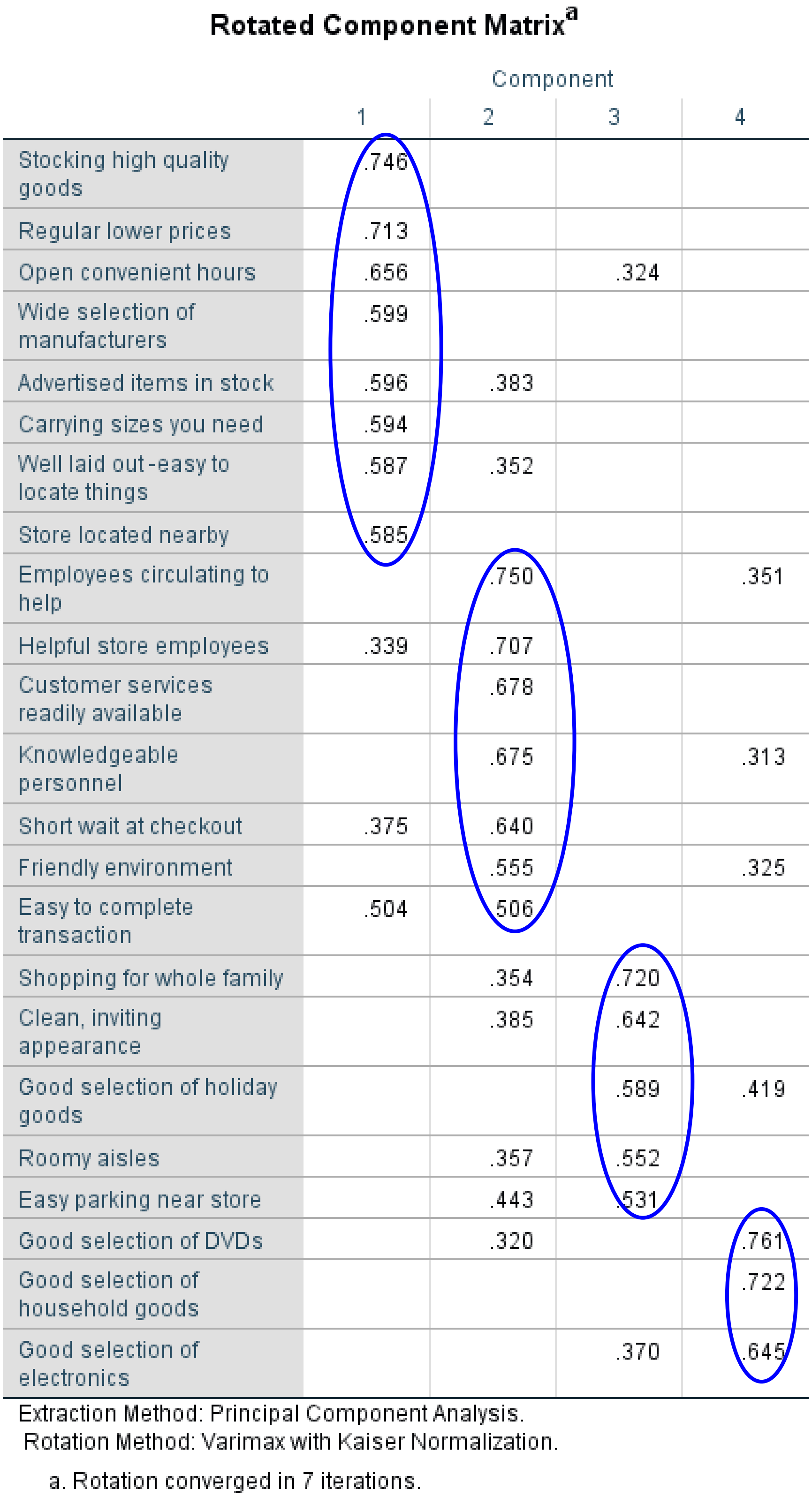
Absolute value below: .3

Factor has several missing value options, but since it will only calculate factor score variables (which is a goal of this analysis) for cases with complete data (339 out of the 361 respondents) we leave the listwise default in place. Note that the SPSS Missing Values option provides summaries of missing value patterns and contains sophisticated methods for estimating missing values.

The *Sorted by size* option lists the variables in descending order by their loading coefficients on the factor for which they load highest. This makes it easier to see which variables relate to which factors. Furthermore, by suppressing loading coefficients less that .3 in absolute value we will only see the larger loadings (small values are replaced with blanks) and not be distracted by small values. These options make the interpretive task much easier when many variables are involved.

Factor presents the *Rotated Component (or Factor) Matrix*, which contains the rotated loadings. Although an experienced factor analyst can work directly from unrotated loading, the point of rotations is to simplify the solution and we will move directly to the rotated loadings.

Figure 13 Rotated Component Loadings



Rotated Component Matrix[a]

	Component			
	1	2	3	4
Stocking high quality goods	.746			
Regular lower prices	.713			
Open convenient hours	.656		.324	
Wide selection of manufacturers	.599			
Advertised items in stock	.596	.383		
Carrying sizes you need	.594			
Well laid out -easy to locate things	.587	.352		
Store located nearby	.585			
Employees circulating to help		.750		.351
Helpful store employees	.339	.707		
Customer services readily available		.678		
Knowledgeable personnel		.675		.313
Short wait at checkout	.375	.640		
Friendly environment		.555		.325
Easy to complete transaction	.504	.506		
Shopping for whole family		.354	.720	
Clean, inviting appearance		.385	.642	
Good selection of holiday goods			.589	.419
Roomy aisles		.357	.552	
Easy parking near store		.443	.531	
Good selection of DVDs		.320		.761
Good selection of household goods				.722
Good selection of electronics			.370	.645

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

a. Rotation converged in 7 iterations.

The variables form the rows and the components (or factors if a factor method were run) form the columns. The values in the table are loadings. Notice the blank areas throughout the table; these occur where loadings are less than .3 in absolute value, and are suppressed due to our option choice. In this way we can focus on the larger (absolute value closer to 1) loadings.

In the *Component 1* column, the variables from *Stocking high quality goods* to *Store located nearby* appear as a group and are sorted in descending order by their loadings. They constitute the variables whose greatest loading is on *Component 1* and thus are the variables most associated with the first principal component. Few of these variables have loadings above .3 on the other components shown here, and few other variables have loadings above .3 on Component 1 (*Easy to complete transaction* is an exception to this).

Similarly *Component 2* is most strongly related to the set of variables *Employees circulating to help* to *Friendly environment*. Again there is some ambiguity, particularly with *Easy parking near store*, which actually loads higher on *Component 3*. And, of course, *Easy to complete transaction* has an equal loading on the first two components.

Looking at the other two components, we see that *Component 3* primarily groups variables *Shopping for whole family* through *Easy parking near store*, whereas *Component 4* groups three *Good selection* questions.

This presentation method of grouping the variables based on which component they relate to most strongly, sorting them in descending order by loadings within the group, and not displaying small loadings, make it considerably easier to read the output. Now can we interpret the components?

The first component has high loadings on questions relating to the *convenience* of shopping at the superstore (low prices, hours, wide selection, sizes, location). The second component focuses attention on *service* (circulating, helpful, available, knowledgeable). The third component is a bit harder to interpret, but seems to relate to shopping *comfort* (family, appearance, room, parking). Finally, the component we nearly rejected is directly related to the availability of specific *goods*. The substantive coherence of the last component is a good argument for keeping it regardless of its eigenvalue and variance explained.

We can now also understand why *Easy to complete transaction* loads on both the *convenience* and *service* components, since this question is relevant to both concepts.

The examination of these components brings into relief the point that factors or components are constructs based on the variables analyzed, **so asking several questions relating to one issue/concept will generate a component (this is used in constructing personality assessments for example)**. Even if we decide to run cluster analysis on the original questions, it is useful to see which are related and obtain a sense of which themes will be more influential in that analysis. If it seems useful to run a cluster analysis based on the four components, we must rerun the Factor procedure, this time creating factor variables (you normally don’t save factor scores until you are satisfied with a factor solution).

Reviewing the Principal Components Analysis

From the perspective of a good factor analysis example, this data set falls a bit short in that in the initial summary, one component dominates, followed by a series of small components (although this may be typical for market research data). We could have alleviated this pattern by selecting only two components for extraction (see the Extraction dialog). This has the advantage of parsimony (simplification), but it has the disadvantage of often creating potentially ambiguous components or factors, and it may not explain enough variance in customer attitudes. The classic examples presented in texts have a more even distribution of variance accounted for among the components or factors.

However, our situation certainly does occur in practice, so it is useful to see what issues emerge from such a data file.

For more advanced users: If this data were to be used for a response-based segmentation study using Discriminant or Logistic Regression, then the high correlations among some of the 23 variables could result in difficulties in the solution (unstable estimated coefficients). One solution to this problem would be to first run factor or principal components and use the factor score variables in the later modeling. Alternatively, the factor or principal component solution could guide creation of subscale variables or the selection of proxy predictor variables (a proxy would be one variable that represents other variables with which it is highly correlated). Such variables would replace the original set in the analysis.

Clustering Based on Components

The 23 variables we began with are not too many to use in a cluster analysis, although some variable selection would have to be done when presenting mean profiles and other cluster results. The question is whether to cluster on specific items or on more general themes. Clustering of factor scores loses some richness of the original data (components accounted for only 60% of the variance of the variables), but reduces the scope of the interpretation and avoids the problem of weighting some issues more heavily than others due to the number of questions per issue.

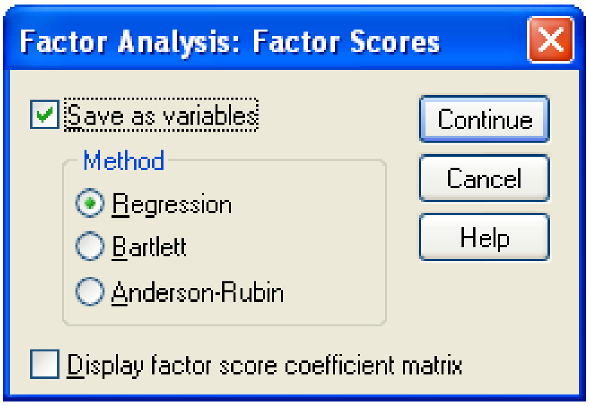
If you are in doubt which strategy to pursue, you can perform cluster with and without factor analysis and see which method provides more insight. Another possibility is to select a fixed number of variables (the top three or four for example) loading on each factor, and use them in the cluster analysis. This preserves the uniqueness of individual questions, yet retains and partially balances the influence of the factors.

Creating Factor Score Variables

To explore the usefulness of our Factor Analysis results, we need to return to Factor and request the factor (component) score variables.

Click the **Dialog Recall** tool, then click **Factor Analysis**
Click the **Scores** button
Click the **Save as variables** check box

Figure 14 Factor Analysis: Factor Scores Dialog



Factor Analysis: Factor Scores

☑ Save as variables

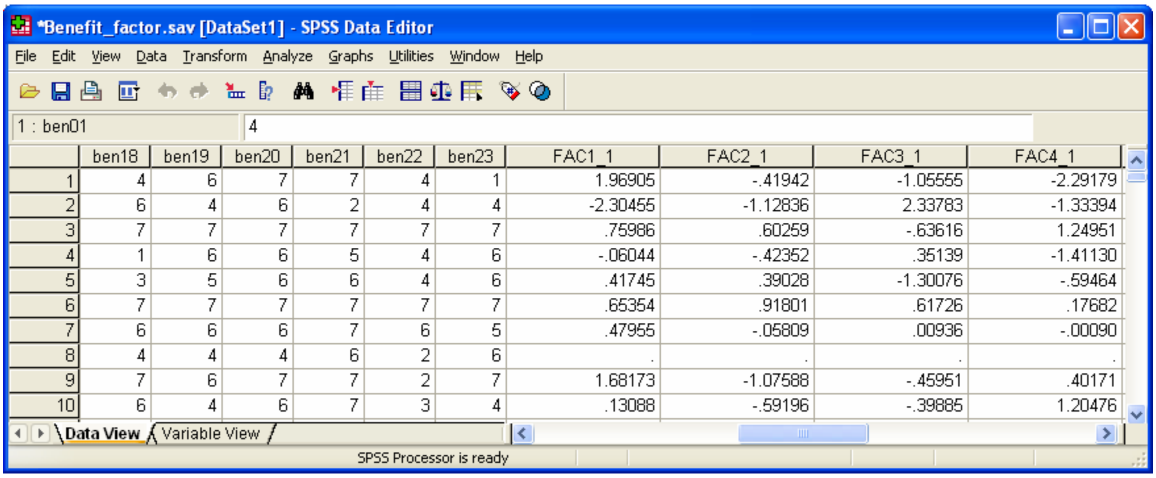
Method
● Regression
○ Bartlett
○ Anderson-Rubin

□ Display factor score coefficient matrix

With *Save as variables* checked, SPSS will create four new variables that will contain each customer's scores on the components. These are derived from the loadings viewed above. This process involves solving for the factors as a function of the variables (the loadings are from equations expressing the variables as a function of the factors). There are several methods for doing this and they differ in technical ways. The default method (Regression) is quite adequate.

Click **Continue**, then click **OK**
Switch to the **Data Editor** window
Scroll right to variable **FAC1_1**

Figure 15 Component Variables in Data Editor


*Benefit_factor.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is ready

1 : ben01 4

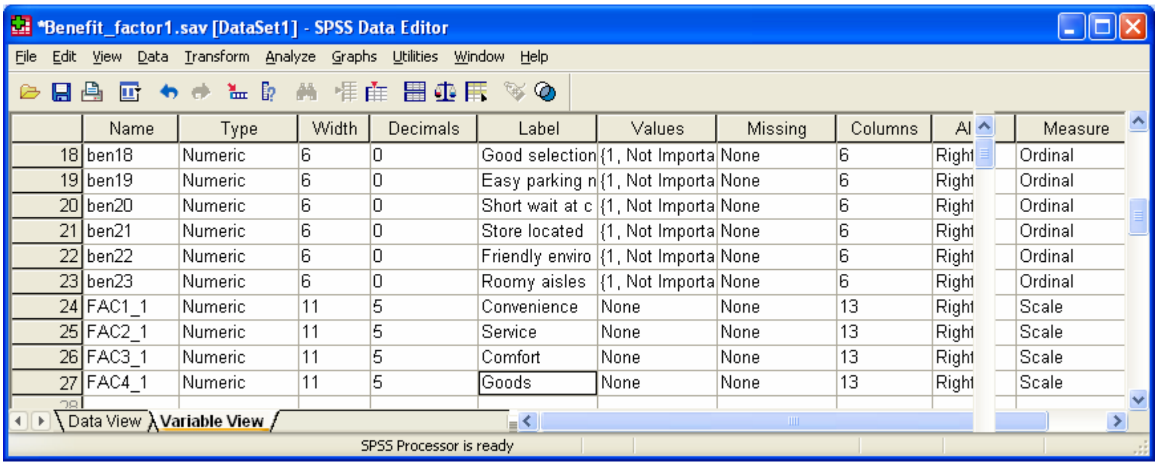
	ben18	ben19	ben20	ben21	ben22	ben23	FAC1_1	FAC2_1	FAC3_1	FAC4_1
1	4	6	7	7	4	1	1.96905	-.41942	-1.05555	-2.29179
2	6	4	6	2	4	4	-2.30455	-1.12836	2.33783	-1.33394
3	7	7	7	7	7	7	.75986	.60259	-.63616	1.24951
4	1	6	6	5	4	6	-.06044	-.42352	.35139	-1.41130
5	3	5	6	6	4	6	.41745	.39028	-1.30076	-.59464
6	7	7	7	7	7	7	.65354	.91801	.61726	.17682
7	6	6	6	7	6	5	.47955	-.05809	.00936	-.00090
8	4	4	4	6	2	6	.	.	.	.
9	7	6	7	7	2	7	1.68173	-1.07588	-.45951	.40171
10	6	4	6	7	3	4	.13088	-.59196	-.39885	1.20476

We created four component score variables. They are in z-score form having means of 0 and standard deviations of 1. For example, the first customer's score for the first component (*FAC1_1*) is very high (1.96905), meaning that he or she rates convenience higher than other customers. On the other hand, the fourth customer's first component score is near zero (–.06044) indicating that he or she falls near the mean in rating the convenience benefits of the store. Therefore, we wouldn't expect these two respondents to end up in the same cluster if our explorations worked.

For easier interpretation of clustering results, labels should be supplied for the four new score variables: Convenience, Service, Comfort and Goods in our case.

Click the **Variable View** tab in the Data editor
Enter labels **Convenience, Service, Comfort, Goods** in that order for the score variables

Figure 16 Variable Labels

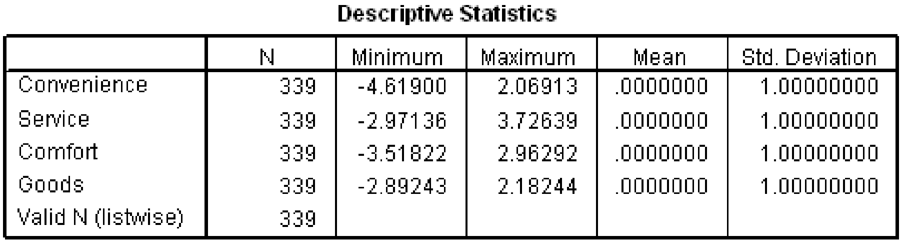

*Benefit_factor1.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is ready

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Al	Measure
18	ben18	Numeric	6	0	Good selection	{1, Not Importa	None	6	Righ	Ordinal
19	ben19	Numeric	6	0	Easy parking n	{1, Not Importa	None	6	Righ	Ordinal
20	ben20	Numeric	6	0	Short wait at c	{1, Not Importa	None	6	Righ	Ordinal
21	ben21	Numeric	6	0	Store located	{1, Not Importa	None	6	Righ	Ordinal
22	ben22	Numeric	6	0	Friendly enviro	{1, Not Importa	None	6	Righ	Ordinal
23	ben23	Numeric	6	0	Roomy aisles	{1, Not Importa	None	6	Righ	Ordinal
24	FAC1_1	Numeric	11	5	Convenience	None	None	13	Righ	Scale
25	FAC2_1	Numeric	11	5	Service	None	None	13	Righ	Scale
26	FAC3_1	Numeric	11	5	Comfort	None	None	13	Righ	Scale
27	FAC4_1	Numeric	11	5	Goods	None	None	13	Righ	Scale

To check on the four component score variables, we can look at their distributions, verifying that respondents who answered all 23 questions (339) can be clustered on the four new variables, and that the variables have a z-score distribution with mean 0 and standard deviation 1.

Click Analyze..Descriptive Statistics..Descriptives (not shown)
Move **FAC1_1** through **FAC4_1** into the **Variable(s)** list box
Click **OK**

Figure 19 Descriptives Table



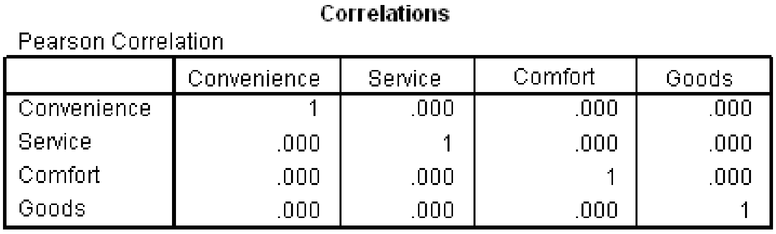
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Convenience	339	-4.61900	2.06913	.0000000	1.00000000
Service	339	-2.97136	3.72639	.0000000	1.00000000
Comfort	339	-3.51822	2.96292	.0000000	1.00000000
Goods	339	-2.89243	2.18244	.0000000	1.00000000
Valid N (listwise)	339				

While the minimums and maximums vary among the score variables, each has a standardized distribution.

Furthermore, by rotating these components before generating the scores, they are uncorrelated. The table below (produced by Correlations - not shown) demonstrates this consequence.

Figure 20 Correlations Among Component Score Variables



Correlations

Pearson Correlation

	Convenience	Service	Comfort	Goods
Convenience	1	.000	.000	.000
Service	.000	1	.000	.000
Comfort	.000	.000	1	.000
Goods	.000	.000	.000	1

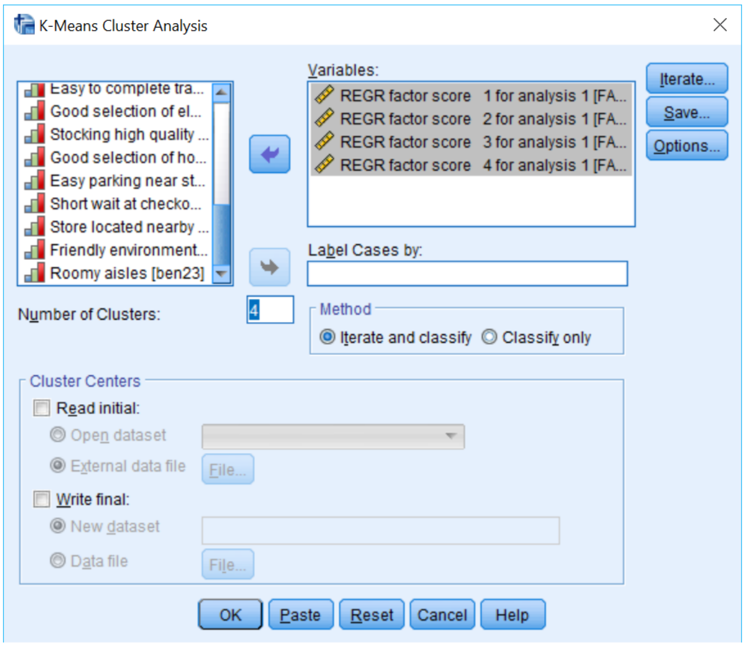
Thus we are prepared to use the four component score variables for clustering.

K-Means Clustering of Component Scores

We choose the K-Means clustering procedure, and we choose to specify four clusters. The choice of four clusters is somewhat arbitrary, but in this case we can say that it is an attempt to see if there are four clusters based on each of the four components we derived. That is, we are looking for clusters of respondents who have the same opinions about each of the overall benefits of shopping at the superstore.

Click Analyze..Classify..K-Means Cluster
Place the four component variables in the Variables box
Change the **Number of Clusters** value to **4**

Figure 21 K-Means Dialog



K-Means Cluster Analysis

Easy to complete tra...
Good selection of el...
Stocking high quality ...
Good selection of ho...
Easy parking near st...
Short wait at checko...
Store located nearby ...
Friendly environment...
Roomy aisles [ben23]

Variables:
REGR factor score 1 for analysis 1 [FA...
REGR factor score 2 for analysis 1 [FA...
REGR factor score 3 for analysis 1 [FA...
REGR factor score 4 for analysis 1 [FA...

Label Cases by:

Number of Clusters: 4

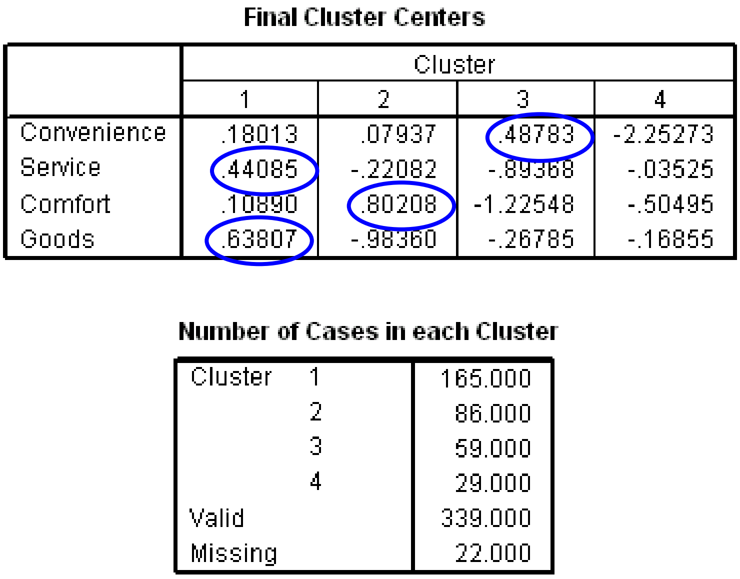
Method
● Iterate and classify ○ Classify only

Cluster Centers
□ Read initial:
● Open dataset
● External data file File...
□ Write final:
● New dataset
○ Data file File...

K-Means iterates to a solution, and you must be certain that the procedure has converged before using a clustering solution. The default number of iterations (*iteration = repetition of a mathematical or computational procedure applied to the result of a previous application, typically as a means of obtaining successively closer approximations to the solution of a problem*). is 10, which is normally adequate, but not for these data. To save time:

Click Iterate
Change the value of **Maximum Iterations** to **50** (not shown
Click Continue, then click **OK**
Scroll to the Final Custer Centers table

Figure 21 K-Means Cluster Results



Final Cluster Centers

	Cluster			
	1	2	3	4
Convenience	.18013	.07937	.48783	-2.25273
Service	.44085	-.22082	-.89368	-.03525
Comfort	.10890	.80208	-1.22548	-.50495
Goods	.63807	-.98360	-.26785	-.16855

Number of Cases in each Cluster

Cluster	1	165.000
	2	86.000
	3	59.000
	4	29.000
Valid		339.000
Missing		22.000

We can see in the Final Cluster Centers table that K-Means Cluster 1 (165 respondents) seems to be characterized by high ratings for *Service* benefits of shopping at the superstore and for the selection of *Goods*. But notice that it has no factor score means which are negative, unlike the other clusters. This implies that they rate all aspects of shopping as of moderate importance, or more. In comparison, Cluster 4 has negative means on all four factors, especially on *Convenience*. This group, fortunately the smallest segment with 29 respondents, rates all aspects of moderate importance, or less.

Cluster 2 (86 respondents) can be characterized as those customers who rate *Comfort* issues high, but not *Goods*, and Cluster 3 (59 respondents) by customers who rate *Convenience* high, but not *Service* nor *Comfort*.

Looking across the rows of the Goods component, we can see that there are differences among the four clusters that are large enough to justify our earlier decision to keep it as a component.

Given our ability to see distinct differences among the four clusters based on their attitudes on the four overall benefit issues, we should deem successful both our principal components factoring and our K-Means clustering of the factor scores.