

Springer Texts in Business and Economics

Marko Sarstedt
Erik Mooi

A Concise Guide to Market Research

The Process, Data, and Methods
Using IBM SPSS Statistics

Third Edition

MULTIMEDIA



Springer



Cluster Analysis

- 9.1 Introduction – 302**
- 9.2 Understanding Cluster Analysis – 302**
- 9.3 Conducting a Cluster Analysis – 305**
 - 9.3.1 Select the Clustering Variables – 305
 - 9.3.2 Select the Clustering Procedure – 309
 - 9.3.3 Select a Measure of Similarity or Dissimilarity – 322
 - 9.3.4 Decide on the Number of Clusters – 328
 - 9.3.5 Validate and Interpret the Clustering Solution – 331
- 9.4 Example – 336**
 - 9.4.1 Hierarchical Cluster Analysis – 336
 - 9.4.2 Two-Step Clustering – 349
- 9.5 Oh, James! (Case Study) – 352**
- 9.6 Review Questions – 353**
- References – 353**

Electronic supplementary material

The online version of this chapter (https://doi.org/10.1007/978-3-662-56707-4_9) contains additional material that is available to authorized users. You can also download the “Springer Nature More Media App” from the iOS or Android App Store to stream the videos and scan the image containing the “Play button”.

Learning Objectives

After reading this chapter you should understand:

- The basic concepts of cluster analysis.
- How basic cluster algorithms work.
- How to compute simple clustering results manually.
- The different types of clustering procedures.
- The SPSS clustering outputs.

Keywords

Agglomerative clustering • Average linkage • Centroid linkage • Chaining effect • Chebychev distance • City-block distance • Clusters • Clustering variables • Complete linkage • Dendrogram • Distance matrix • Divisive clustering • Euclidean distance • Factor-cluster segmentation • Hierarchical clustering methods • *k*-means • *k*-means++ • *k*-medians • *k*-medoids • Label switching • Linkage algorithm • Local optimum • Manhattan metric • Market segmentation • Matching coefficients • Non-hierarchical clustering methods • Partitioning methods • Profiling • Russel and Rao coefficient • Silhouette measure of cohesion and separation • Simple matching coefficient • Single linkage • Straight line distance • Two-step clustering • Variance ratio criterion • Ward's linkage

9

9.1 Introduction

Market segmentation is one of the most fundamental marketing activities. To successfully match products and services to customer needs, companies have to divide markets into groups (segments) of consumers, customers, and clients with similar needs and wants. Firms can then target each of these segments by positioning themselves in a unique segment (e.g., Ferrari in the high-end sports car market). Market segmentation “is essential for marketing success: the most successful firms segment their markets carefully” (Lilien and Rangaswamy 2004, p. 61) and “tools such as segmentation [...] have the largest impact on marketing decisions” (Roberts et al. 2014, p. 127). While market researchers often form market segments based on practical grounds, industry practice and wisdom, cluster analysis uses data to form segments, making segmentation less dependent on subjectivity.

9.2 Understanding Cluster Analysis

Cluster analysis is a method for segmentation and identifies homogenous groups of objects (or cases, observations) called **clusters**. These objects can be individual customers, groups of customers, companies, or entire countries. Objects in a certain cluster should be as similar as possible to each other, but as distinct as possible from objects in other clusters.

Let's try to gain a basic understanding of cluster analysis by looking at a simple example. Imagine that you are interested in segmenting customers A to G in order to better target them through, for example, pricing strategies.

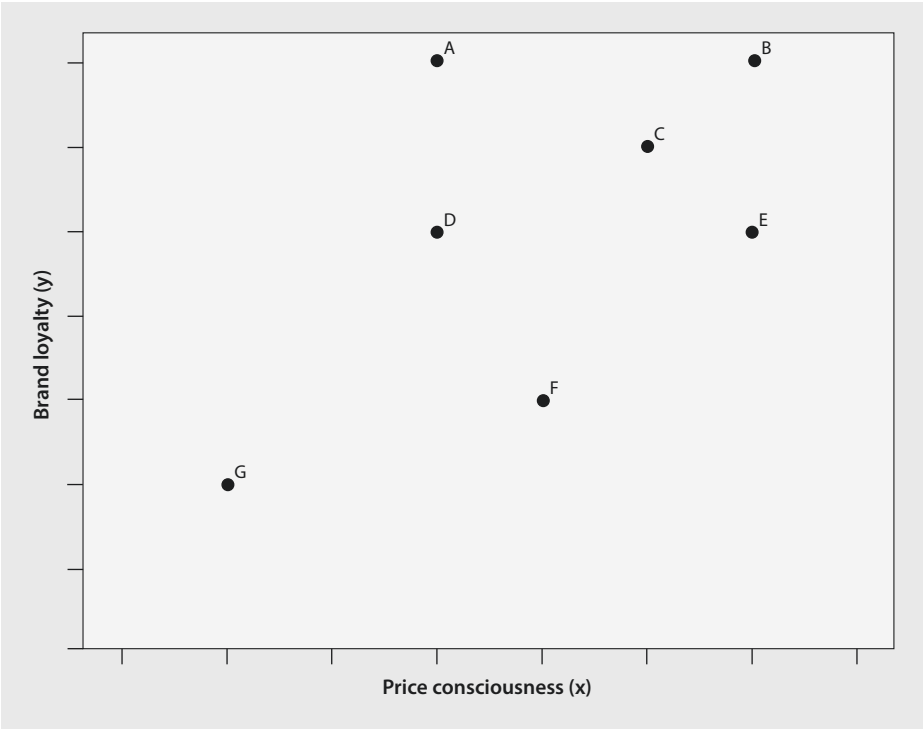
The first step is to decide on the characteristics that you will use to segment your customers A to G. In other words, you have to decide which **clustering variables** will

be included in the analysis. For example, you may want to segment a market based on customers’ price consciousness (x) and brand loyalty (y). These two variables can be measured on a scale from 0 to 100 with higher values denoting a higher degree of price consciousness and brand loyalty. ■ Table 9.1 and the scatter plot in ■ Fig. 9.1 show the values of seven customers (referred to as objects).

The aim of cluster analysis is to identify groups of objects (here, customers) that are very similar regarding their price consciousness and brand loyalty, and assign them to clusters. After having decided on the clustering variables (here, price consciousness and brand loyalty), we need to decide on the clustering procedure to form our groups of objects. This step is crucial for the analysis, as different procedures require different decisions prior to

■ Table 9.1 Data

Customer	A	B	C	D	E	F	G
x	33	82	66	30	79	50	10
y	95	94	80	67	60	33	17



■ Fig. 9.1 Scatter plot

analysis. There are many different approaches and little guidance on which one to use. We will discuss the most popular approaches in market research, including:

- hierarchical methods,
- partitioning methods (especially *k*-means), and
- two-step clustering.

While the basic aim of these procedures is the same, namely grouping similar objects into clusters, they take different routes, which we will discuss in this chapter. An important consideration before starting the grouping is to determine how similarity should be measured. Most methods calculate measures of (dis)similarity by estimating the distance between pairs of objects. Objects with smaller distances between one another are considered more similar, whereas objects with larger distances are considered more dissimilar. The decision on how many clusters should be derived from the data is a fundamental issue in the application of cluster analysis. This question is explored in the next step of the analysis. In most instances, we do not know the exact number of clusters and then we face a trade-off. On the one hand, we want as few clusters as possible to make the clusters easy to understand and actionable. On the other hand, having many clusters allows us to identify subtle differences between objects.

9

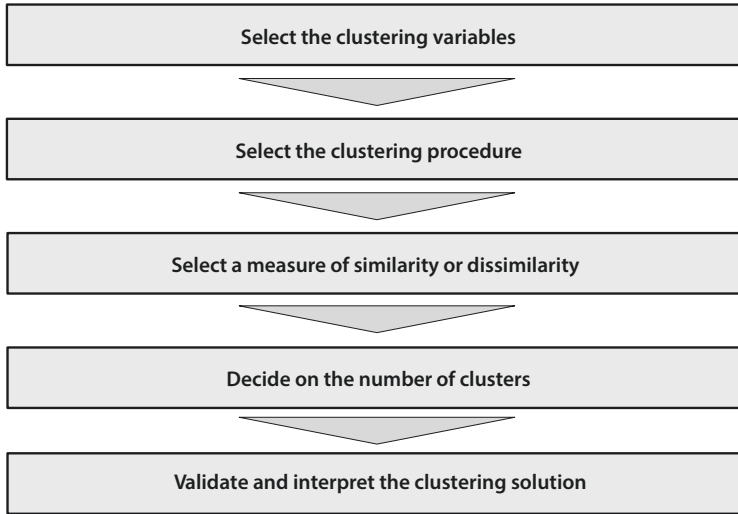
Megabus is a hugely successful bus line in the US. They completely rethought the nature of their customers and concentrated on three specific segments of the market: College kids, women travelling in groups, and active seniors. To meet these customer segments' needs, Megabus reimaged the entire driving experience by developing double-decker buses with glass roofs and big windows, and equipped with fast WiFi. Megabus's success of segmenting and targeting efforts has led to practitioners talk about the "Megabus Effect"—how one company has shaped an entire industry.



© Stagecoach Group plc.

<https://www.youtube.com/watch?v=mnrblywmSEo>

In the final step, we need to interpret the clustering solution by defining and labeling the obtained clusters. We can do this by comparing the mean values of the clustering variables across the different clusters, or by identifying explanatory variables to profile the



■ Fig. 9.2 Steps involved in a cluster analysis

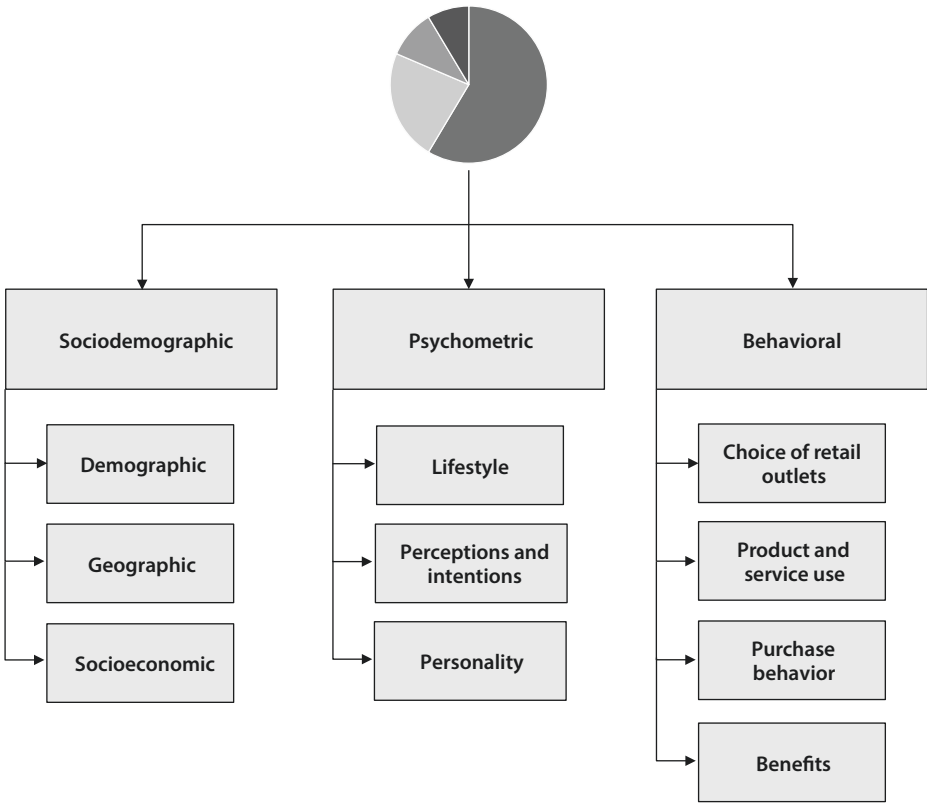
clusters. Ultimately, managers should be able to identify customers in each cluster on the basis of easily measurable variables. This final step also requires us to assess the clustering solution's stability and validity. ■ Figure 9.2 illustrates the steps associated with a cluster analysis; we will discuss these steps in more detail in the following sections.

9.3 Conducting a Cluster Analysis

9.3.1 Select the Clustering Variables

At the beginning of the clustering process, we have to select appropriate variables for clustering. Even though this choice is critical, it is rarely treated as such. Instead, a mixture of intuition and data availability guide most analyses in marketing practice. However, faulty assumptions may lead to improper market segmentation and, consequently, to deficient marketing strategies. Thus, great care should be taken when selecting the clustering variables! There are several types of clustering variables, as shown in ■ Fig. 9.3. Sociodemographic variables define clusters based on people's demographic (e.g., age, ethnicity, and gender), geographic (e.g., residence in terms of country, state, and city), and socioeconomic (e.g., education, income, and social class) characteristics. Psychometric variables capture unobservable character traits such as people's personalities or lifestyles. Finally, behavioral clustering variables typically consider different facets of consumer behavior, such as the way people purchase, use, and dispose of products. Other behavioral clustering variables capture specific benefits which different groups of consumers look for in a product.

The types of variables used for cluster analysis provide different solutions and, thereby, influence targeting strategies. Over the last decades, attention has shifted from more traditional sociodemographic clustering variables towards behavioral and psychometric variables. The latter generally provide better guidance for decisions on marketing instruments'



■ Fig. 9.3 Types of clustering variables

effective specification. Generally, clusters based on psychometric variables are more homogenous and these consumers respond more consistently to marketing actions (e.g., Wedel and Kamakura 2000). However, consumers in these clusters are frequently hard to identify as such variables are not easily measured. Conversely, clusters determined by sociodemographic variables are easy to identify but are also more heterogeneous, which complicates targeting efforts. Consequently, researchers frequently combine different variables such as lifestyle characteristics and demographic variables, benefiting from each one's strengths.

In some cases, the choice of clustering variables is apparent because of the task at hand. For example, a managerial problem regarding corporate communications will have a fairly well defined set of clustering variables, including contenders such as awareness, attitudes, perceptions, and media habits. However, this is not always the case and researchers have to choose from a set of candidate variables. But how do we make this decision? To facilitate the choice of clustering variables, we should consider the following guiding questions:

- Do the variables differentiate sufficiently between the clusters?
- Is the relation between the sample size and the number of clustering variables reasonable?
- Are the clustering variables highly correlated?
- Are the data underlying the clustering variables of high quality?

■ Do the variables differentiate sufficiently between the clusters?

It is important to select those clustering variables that provide a clear-cut differentiation between the objects.¹ More precisely, *criterion validity* is of special interest; that is, the extent to which the “independent” clustering variables are associated with one or more criterion variables not included in the analysis. Such criterion variables generally relate to an aspect of behavior, such as purchase intention or willingness-to-pay. Given this relationship, there should be significant differences between the criterion variable(s) across the clusters (e.g., consumers in one cluster exhibit a significantly higher willingness-to-pay than those in other clusters). These associations may or may not be causal, but it is essential that the clustering variables distinguish significantly between the variable(s) of interest.

■ Is the relation between the sample size and the number of clustering variables reasonable?

When choosing clustering variables, the sample size is important. From a statistical perspective, every additional variable requires an over-proportional increase in observations to ensure valid results. Unfortunately, there is no generally accepted guideline regarding minimum sample sizes or the relationship between the objects and the number of clustering variables used. Recent rules-of-thumb are as follows:

- In the simplest case where clusters are of equal size, Qiu and Joe (2009) recommend a sample size at least ten times the number of clustering variables multiplied by the number of clusters.
- Dolnicar et al. (2014) recommend using a sample size of 70 times the number of clustering variables.
- Dolnicar et al. (2016) find that increasing the sample size from 10 to 30 times the number of clustering variables substantially improves the clustering solution. This improvement levels off subsequently, but is still noticeable up to a sample size of approximately 100 times the number of clustering variables.

These rules-of-thumb are approximate as the required sample size depends on many factors, such as survey data characteristics (e.g., nonresponse, sampling error, response styles), relative cluster sizes, and the degree to which the clusters overlap (Dolnicar et al. 2016). Qiu and Joe (2009) suggest a minimum sample size of 10 times the number of clustering variables. Keep in mind that no matter how many variables are used and no matter how small the sample size, cluster analysis will almost always provide a result. At the same time, increasing the sample size has decreasing marginal returns on the quality of results. In addition, we need to be able to find clusters that are managerially relevant as the cluster sizes need to be substantial to ensure that the targeted marketing programs are profitable.

■ Are the clustering variables highly correlated?

If there is strong correlation between the variables, they are not sufficiently unique to identify distinct market segments. If highly correlated variables—0.90 and over—are used for cluster

1 Tonks (2009) provides a discussion of segment design and the choice of clustering variables in consumer markets.

Box 9.1 Issues with factor-cluster segmentation

Dolnicar and Grün (Dolnicar and Grün 2009) identify several problems of the factor-cluster segmentation approach (see ► Chap. 8 for a discussion of principal component and factor analysis and related terminology):

1. The data are pre-processed and the clusters are identified on the basis of transformed values, not on the original information, which leads to different results.
2. In factor analysis, the factor solution does not explain all the variance; information is thus discarded before the clusters have been identified or constructed.
3. Eliminating variables with low loadings on all the extracted factors means that, potentially, the most important pieces of information for the identification of niche clusters are discarded, making it impossible to ever identify such groups.
4. The interpretations of clusters based on the original variables become questionable, given that these clusters were constructed by using factor scores.

Several studies have shown that the factor-cluster segmentation reduces the success of finding useable clusters significantly.² Consequently, you should reduce the number of items in the questionnaire's pre-testing phase, retaining a reasonable number of relevant, non-overlapping questions that you believe differentiate the clusters well. However, if you have doubts about the data structure, factor-clustering segmentation may still be a better option than discarding items.

9

analysis, the specific aspects that these variables cover will be overrepresented in the clustering solution. For example, if we were to add another variable called *brand preference* to our analysis, it would almost cover the same aspect as *brand loyalty*. The concept of being attached to a brand would therefore be overrepresented in the analysis, because the clustering procedure does not conceptually differentiate between the clustering variables. Researchers frequently handle such correlation problems by applying cluster analysis to the observations' factor scores, derived from a principal component or factor analysis. However, this **factor-cluster segmentation** approach is subject to several limitations, which we discuss in Box 9.1.

■ Are the data underlying the clustering variables of high quality?

Ultimately, the choice of clustering variables always depends on contextual influences, such as the data availability or the resources to acquire additional data. Market researchers often overlook that the choice of clustering variables is closely connected to data quality. Only those variables that ensure that high quality data can be used should be included in the analysis (Dolnicar and Lazarevski 2009). Following our discussions in ► Chaps. 3–5, data are of high quality if the questions ...

- ... have a strong theoretical basis,
- ... are not contaminated by respondent fatigue or response styles, and
- ... reflect the current market situation (i.e., they are recent).

The requirements of other functions in the organization often play a major role in the choice of clustering variables. Consequently, we have to be aware that the choice of clustering variables should lead to segments acceptable to the different functions in the organization.

2 See Arabie and Hubert (1994), Sheppard (1996), and Dolnicar and Grün (2009).

9.3.2 Select the Clustering Procedure

By choosing a specific clustering procedure, we determine how clusters should be formed. This forming of clusters always involves optimizing some kind of criterion, such as minimizing the within-cluster variance (i.e., the clustering variables' overall variance of the objects in a specific cluster), or maximizing the distance between the clusters. The procedure could also address the question of how to determine the (dis)similarity between objects in a newly formed cluster and the remaining objects in the dataset.

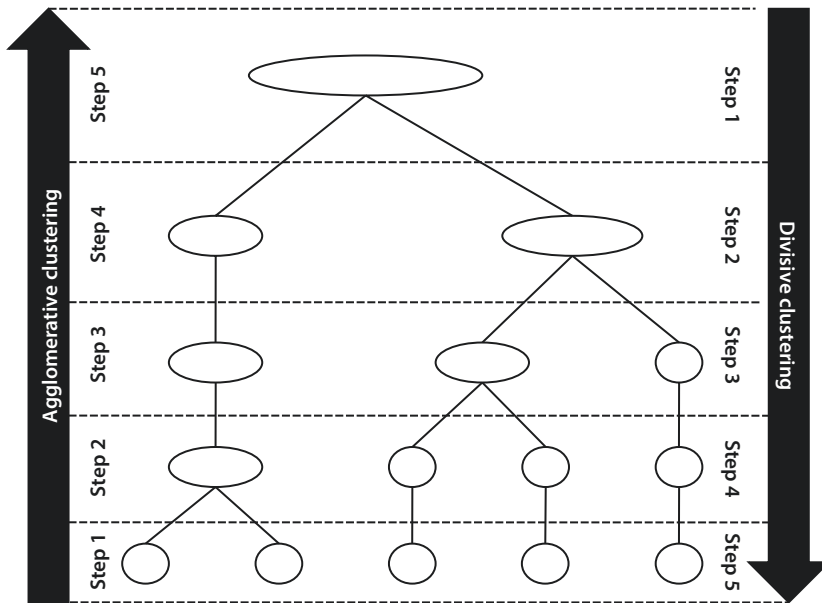
There are many different clustering procedures and also many ways of classifying these (e.g., overlapping versus non-overlapping, unimodal versus multimodal, exhaustive versus non-exhaustive). Wedel and Kamakura (2000), Dolnicar (2003), and Kaufman and Rousseeuw (2005) offer reviews of clustering techniques. A practical distinction is the differentiation between hierarchical and partitioning methods (especially *k*-means), which we will discuss in the next sections.

9.3.2.1 Hierarchical Clustering Methods

■ Understanding Hierarchical Clustering Methods

Hierarchical clustering methods are characterized by the tree-like structure established in the course of the analysis. Most hierarchical methods fall into a category called **agglomerative clustering**. In this category, clusters are consecutively formed from objects. Agglomerative clustering starts with each object representing an individual cluster. The objects are then sequentially merged to form clusters of multiple objects, starting with the two most similar objects. Similarity is typically defined in terms of the distance between objects. That is, objects with smaller distances between one another are considered more similar, whereas objects with larger distances are considered more dissimilar. After the merger of the first two most similar (i.e., closest) objects, the agglomerative clustering procedure continues by merging another pair of objects or adding another object to an already existing cluster. This procedure continues until all the objects have been merged into one big cluster. As such, agglomerative clustering establishes a hierarchy of objects from the bottom (where each object represents a distinct cluster) to the top (where all objects form one big cluster). The left-hand side of ■ Fig. 9.4 shows how agglomerative clustering merges objects (represented by circles) step-by-step with other objects or clusters (represented by ovals).

Hierarchical clustering can also be interpreted as a top-down process, where all objects are initially merged into a single cluster, which the algorithm then gradually splits up into smaller clusters. This approach to hierarchical clustering is called **divisive clustering**. The right-hand side of ■ Fig. 9.4 illustrates the divisive clustering concept. As we can see, in both agglomerative and divisive clustering, a cluster on a higher level of the hierarchy always encompasses all clusters from a lower level. This means that if an object is assigned to a certain cluster, there is no possibility of reassigning this object to another cluster (hence, the name hierarchical clustering). This is an important distinction between hierarchical and partitioning methods, such as *k*-means, which we will explore later in this chapter.



■ Fig. 9.4 Agglomerative and divisive clustering

Divisive procedures are rarely used in market research and not implemented in statistical software programs such as SPSS as they are computationally very intensive for all but small datasets.³ We therefore focus on (agglomerative) hierarchical clustering.

■ Linkage algorithms

When using agglomerative hierarchical clustering, you need to specify a **linkage algorithm**. Linkage algorithms define the distance from a newly formed cluster to a certain object, or to other clusters in the solution. The most popular linkage algorithms include the following:

- **Single linkage** (*nearest neighbor* in SPSS): The distance between two clusters corresponds to the shortest distance between any two members in the two clusters.
- **Complete linkage** (*furthest neighbor* in SPSS): The oppositional approach to single linkage assumes that the distance between two clusters is based on the longest distance between any two members in the two clusters.
- **Average linkage** (*between-groups linkage* in SPSS): The distance between two clusters is defined as the average distance between all pairs of the two clusters' members.
- **Centroid linkage**: In this approach, the geometric center (centroid) of each cluster is computed first. This is done by computing the clustering variables' average values of all the objects in a certain cluster. The distance between the two clusters equals the distance between the two centroids.

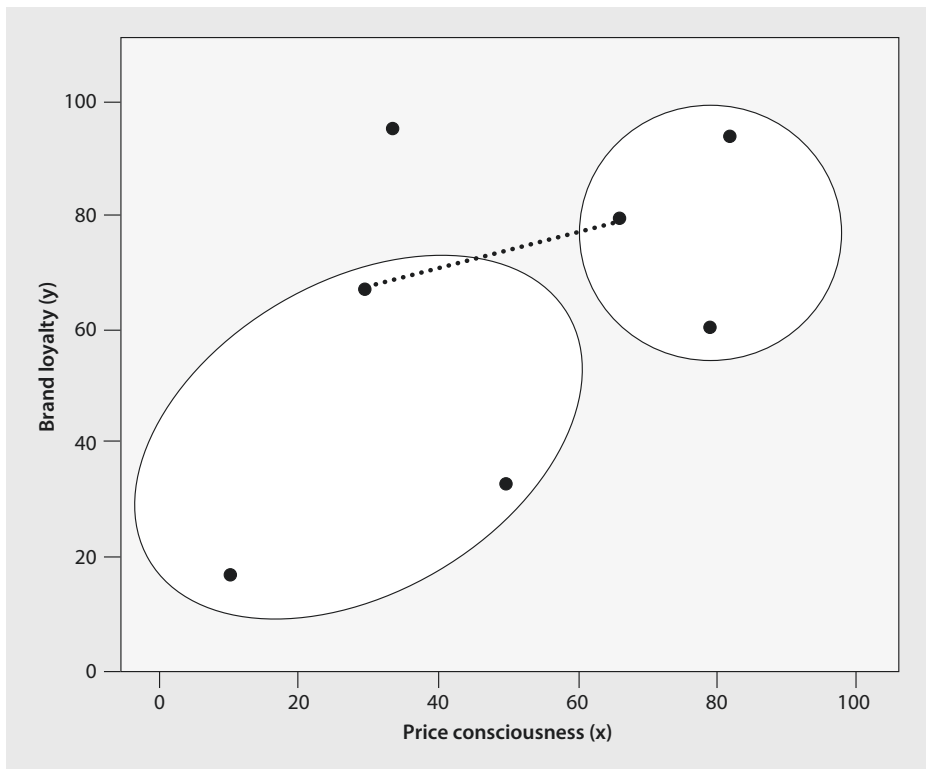
³ Whereas agglomerative methods have the large task of checking $N(N-1)/2$ possible first combinations of observations (note that N represents the number of observations in the dataset), divisive methods have the almost impossible task of checking $2^{(N-1)}-1$ combinations.

- **Ward's linkage:** This approach differs from the previous ones in that it does not combine the two closest or most similar objects successively. Instead, Ward's linkage combines those objects whose merger increases the overall within-cluster variance (i.e., the homogeneity of clusters) to the smallest possible degree. The approach is generally used in combination with (squared) Euclidean distances, but can be used in combination with any other (dis)similarity measure.

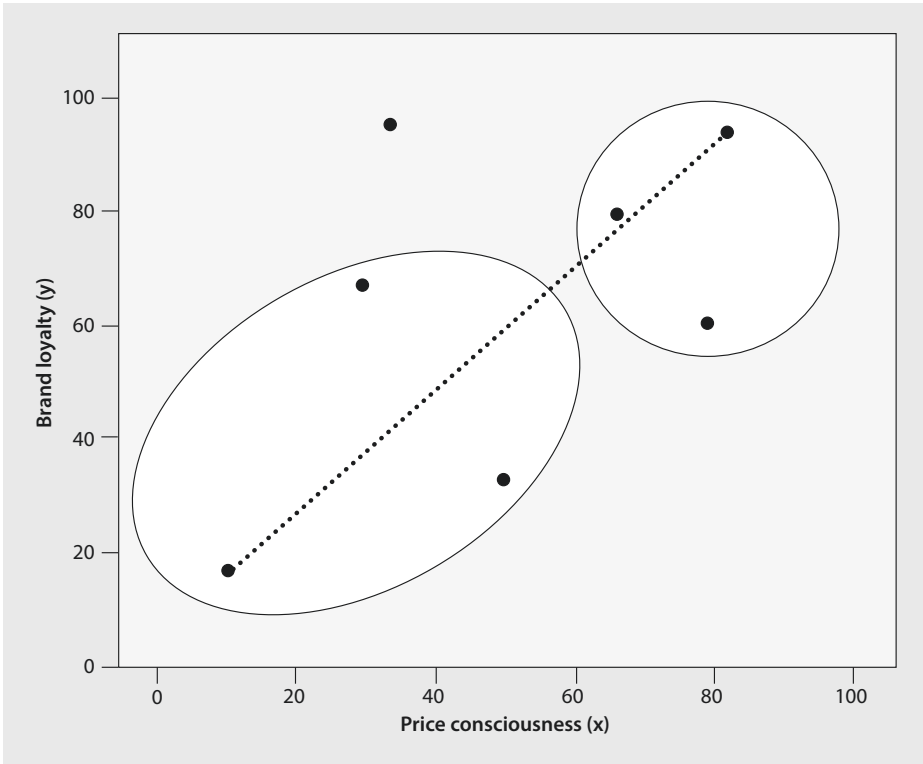
■ Figs. 9.5, 9.6, 9.7, 9.8 and 9.9 illustrate these linkage algorithms for two clusters, which are represented by white circles surrounding a set of objects.

Each of these linkage algorithms can yield different results when used on the same dataset, as each has specific properties:

- The single linkage algorithm is based on minimum distances; it tends to form one large cluster with the other clusters containing only one or a few objects each. We can make use of this **chaining effect** to detect *outliers*, as these will be merged with the remaining objects—usually at very large distances—in the last steps of the analysis. Single linkage is considered the most versatile algorithm.
- The complete linkage method is strongly affected by outliers, as it is based on maximum distances. Clusters produced by this method are likely to be compact and tightly clustered.



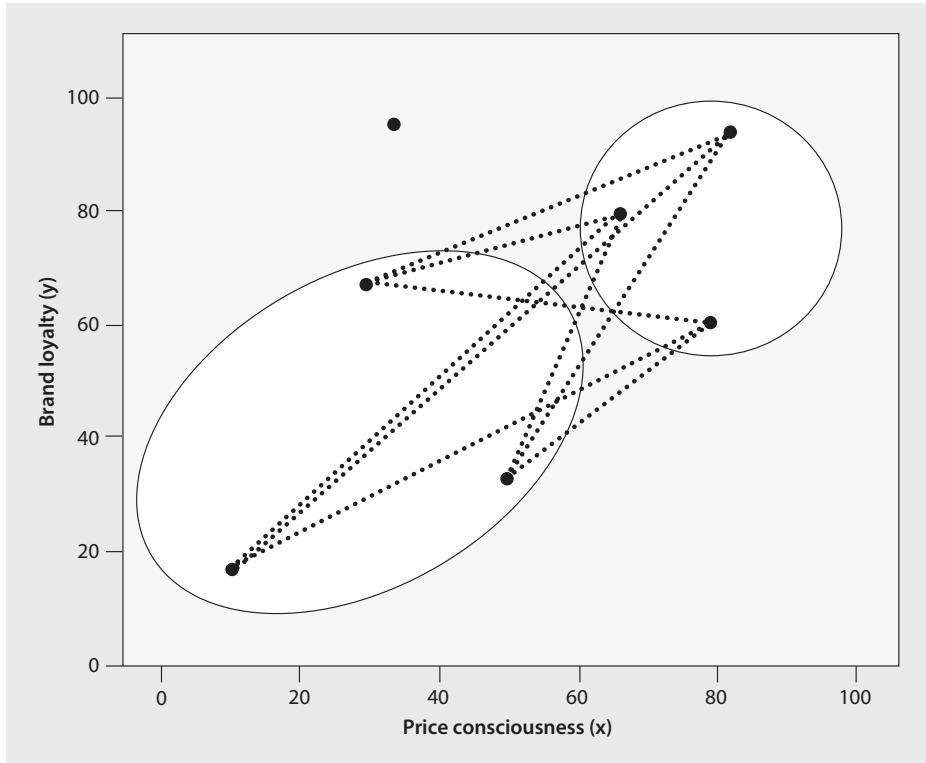
■ Fig. 9.5 Single linkage



■ Fig. 9.6 Complete linkage

- The average linkage and centroid linkage algorithms tend to produce clusters with low within-cluster variance and with similar sizes. The average linkage is affected by outliers, but less than the complete linkage method.
- Ward's linkage yields clusters of similar size with a similar degree of tightness. Prior research has shown that the approach generally performs very well. However, outliers and highly correlated variables have a strong impact on the results.

To better understand how the linkage algorithms work, let's manually examine some calculation steps using single linkage as an example. Let's start by looking at the distance matrix in ■ Table 9.2, which shows the distances between objects A-G from our initial example. In this distance matrix, the non-diagonal elements express the distances between pairs of objects based on the Euclidean distance—we will discuss this distance measure in the following section. The diagonal elements of the matrix represent the distance from each object to itself, which is, of course, 0. In our example, the distance matrix is an 8×8 table with the lines and rows representing the objects under consideration (see ■ Table 9.1). As the distance between objects B and C (in this case, 21.260 units; printed in bold in ■ Table 9.2) is the same as between C and B, the distance matrix is symmetrical.



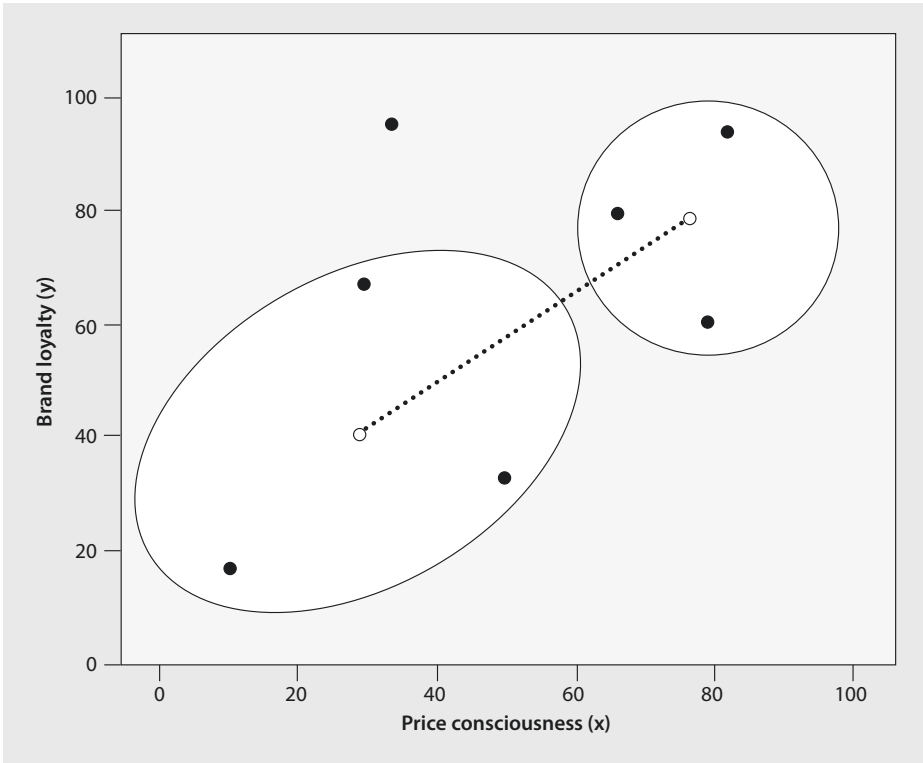
■ Fig. 9.7 Average linkage

Furthermore, since the distance between an object and itself is 0, you only need to look at either the lower or upper non-diagonal elements.

In the very first step, the two objects exhibiting the smallest distance in the matrix are merged. Since the smallest distance occurs between B and C ($d(B,C) = 21.260$), we merge these two objects in the first step of the analysis.

➤ **Agglomerative clustering procedures always merge those objects with the smallest distance, regardless of the linkage algorithm used (e.g., single or complete linkage).**

In the next step, we form a new distance matrix by considering the single linkage decision rule as discussed above. Using this linkage algorithm, we need to compute the distance from the newly formed cluster $[B,C]$ (clusters are indicated by squared brackets) to all the other objects. For example, with regard to the distance from the cluster $[B,C]$ to object A, we need to check whether A is closer to object B or to object C. That is, we look for the minimum value in $d(A,B)$ and $d(A,C)$ from ■ Table 9.2. As $d(A,C) = 36.249$ is smaller than $d(A,B) = 49.010$, the distance from A to the newly formed cluster is equal to $d(A,C)$; that is, 36.249. We also compute the distances from cluster $[B,C]$ to all the other



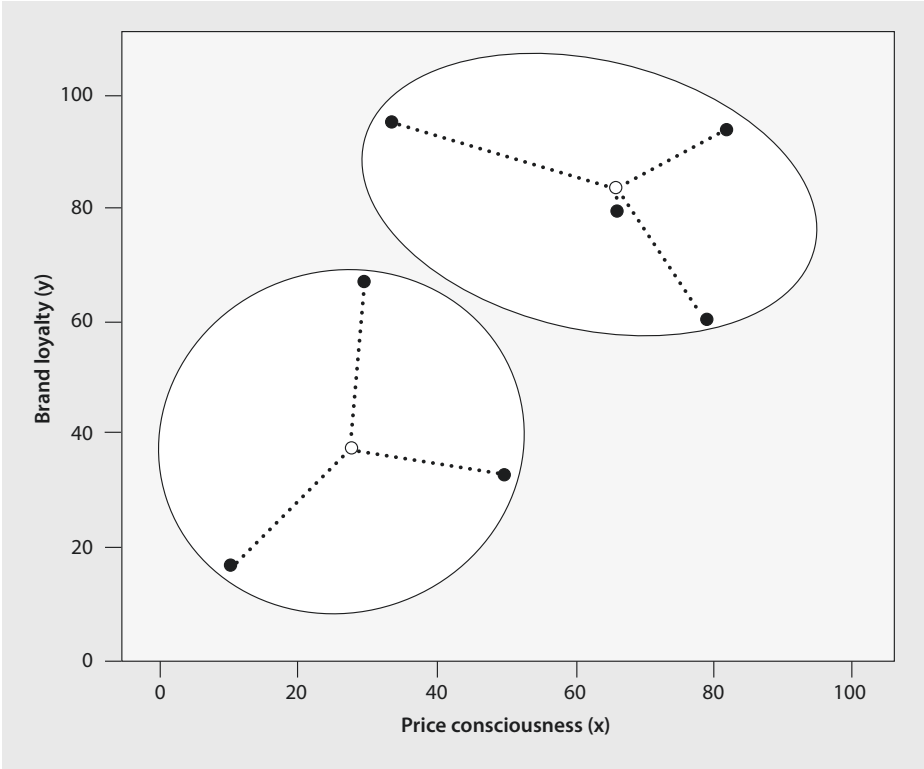
■ Fig. 9.8 Centroid linkage

objects (i.e., D, E, F, G). For example, the distance between $[B, C]$ and D is the minimum of $d(B, D) = 58.592$ and $d(C, D) = 38.275$ (■ Table 9.2). Finally, there are several distances, such as $d(D, E)$ and $d(E, F)$, which are not affected by the merger of B and C. These distances are simply copied into the new distance matrix. This yields the new distance matrix shown in ■ Table 9.3.

Continuing the clustering procedure, we simply repeat the last step by merging the objects in the new distance matrix that exhibit the smallest distance and calculate the distance from this new cluster to all the other objects. In our case, the smallest distance (23.854, printed in bold in ■ Table 9.3) occurs between the newly formed cluster $[B, C]$ and object E. The result of this step is described in ■ Table 9.4.

Try to calculate the remaining steps yourself and compare your solution with the distance matrices in the following ■ Tables 9.5, 9.6 and 9.7.

By following the single linkage procedure, the last steps involve the merger of cluster $[A, B, C, D, E, F]$ and object G at a distance of 43.081. Do you get the same results? As you can see, conducting a basic cluster analysis manually is not that hard at all—not if there are only a few objects.



■ Fig. 9.9 Ward's linkage

■ Table 9.2 Euclidean distance matrix							
Objects	A	B	C	D	E	F	G
A	0						
B	49.010	0					
C	36.249	21.260	0				
D	28.160	58.592	38.275	0			
E	57.801	34.132	23.854	40.497	0		
F	64.288	68.884	49.649	39.446	39.623	0	
G	81.320	105.418	84.291	53.852	81.302	43.081	0

Note: Smallest distance is printed in bold.

Table 9.3 Distance matrix after first clustering step (single linkage)

Objects	A	B, C	D	E	F	G
A	0					
B, C	36.249	0				
D	28.160	38.275	0			
E	57.801	23.854	40.497	0		
F	64.288	49.649	39.446	39.623	0	
G	81.320	84.291	53.852	81.302	43.081	0

Note: Smallest distance is printed in bold.

Table 9.4 Distance matrix after second clustering step (single linkage)

Objects	A	B, C, E	D	F	G
A	0				
B, C, E	36.249	0			
D	28.160	38.275	0		
F	64.288	39.623	39.446	0	
G	81.320	81.302	53.852	43.081	0

Note: Smallest distance is printed in bold.

Table 9.5 Distance matrix after third clustering step (single linkage)

Objects	A, D	B, C, E	F	G
A, D	0			
B, C, E	36.249	0		
F	39.446	39.623	0	
G	53.852	81.302	43.081	0

Note: Smallest distance is printed in bold.

Table 9.6 Distance matrix after fourth clustering step (single linkage)

Objects	A, B, C, D, E	F	G
A, B, C, D, E	0		
F	39.446	0	
G	53.852	43.081	0

Note: Smallest distance is printed in bold.

■ **Table 9.7** Distance matrix after fifth clustering step (single linkage)

Objects	A, B, C, D, E, F	G
A, B, C, D, E, F	0	
G	43.081	0

9.3.2.2 Partitioning Methods: *k*-means

Partitioning clustering methods are another important group of procedures. As with hierarchical clustering, there is a wide array of different algorithms; of these, *k*-means is the most popular for market research.

■ **Understanding *k*-means Clustering**

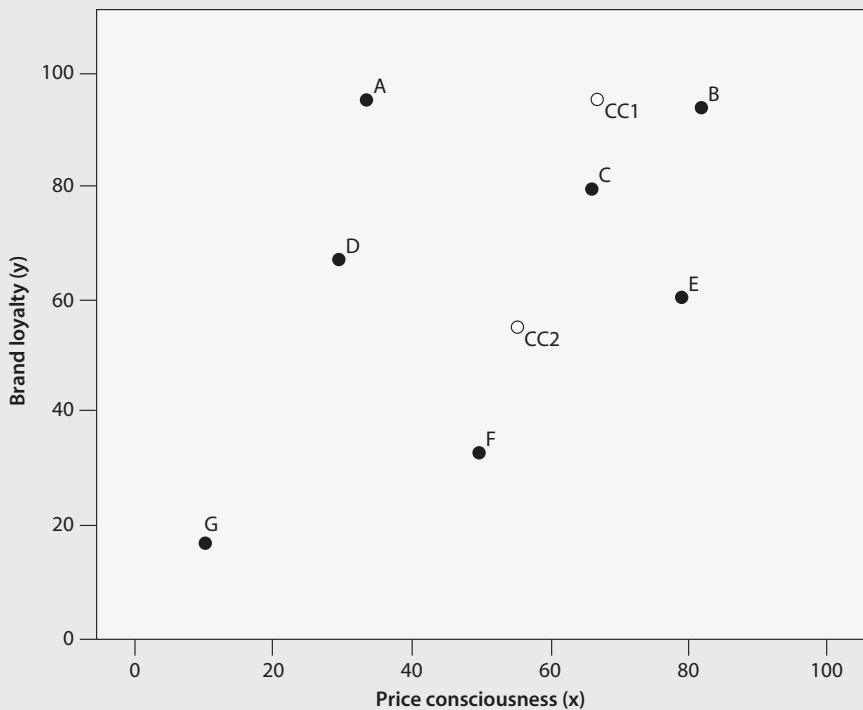
The ***k*-means** method follows an entirely different concept than the hierarchical methods discussed above. The initialization of the analysis is one crucial difference. Unlike with hierarchical clustering, we need to specify the number of clusters to extract from the data prior to the analysis. Using this information as input, *k*-means starts by randomly assigning all objects to the clusters. In the next step, *k*-means successively reassigns the objects to other clusters with the aim of minimizing the within-cluster variation. This within-cluster variation is equal to the squared distance of each observation to the center of the associated cluster (i.e., the centroid). If the reallocation of an object to another cluster decreases the within-cluster variation, this object is reassigned to that cluster.

Since cluster affiliations can change in the course of the clustering process (i.e., an object can move to another cluster in the course of the analysis), *k*-means does not build a hierarchy as hierarchical clustering does (■ Fig. 9.4). Therefore, *k*-means belongs to the group of **non-hierarchical clustering methods**.

For a better understanding of the approach, let's take a look at how it works in practice. ■ Figs. 9.10, 9.11, 9.12 and 9.13 illustrate the four steps of the *k*-means clustering process—research has produced several variants of the original algorithm, which we briefly discuss in Box 9.2.

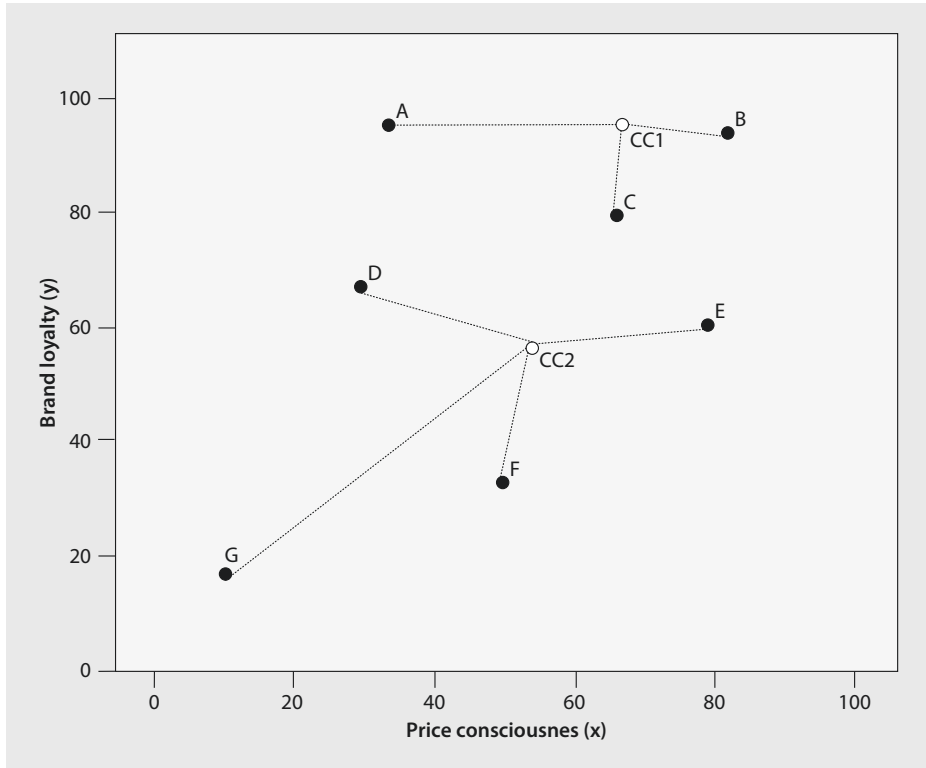
Box 9.2 Variants of the original *k*-means method

***k*-medians** is a popular variant of *k*-means, which essentially follows the same logic and procedure. However, instead of using the cluster mean as a reference point for the calculation of the within cluster variance, *k*-medians minimizes the absolute deviations from the cluster medians, which equals the city-block distance. Thus, *k*-medians does *not* optimize the squared deviations from the mean as in *k*-means, but absolute distances. Thereby *k*-median avoids the possible effect of extreme values on the cluster solution. Other variants use other cluster centers (e.g., ***k*-medoids**; Kaufman and Rousseeuw 2005; Park and Jun 2009), or optimize the initialization process (e.g., ***k*-means++**; Arthur and Vassilvitskii 2007). However, neither of these variants is menu-accessible in SPSS.



■ Fig. 9.10 *k*-means procedure (step 1: placing random cluster centers)

- **Step 1:** The researcher needs to specify the number of clusters that *k*-means should retain from the data. Using this number as the input, the algorithm selects a center for each cluster. In our example, two cluster centers are randomly initiated, which CC1 (first cluster) and CC2 (second cluster) represent in ■ Fig. 9.10.
- **Step 2:** Euclidean distances are computed from the cluster centers to every object. Each object is then assigned to the cluster center with the shortest distance to it. In our example (■ Fig. 9.11), objects A, B, and C are assigned to the first cluster, whereas objects D, E, F, and G are assigned to the second. We now have our initial partitioning of the objects into two clusters.
- **Step 3:** Based on the initial partition in step 2, each cluster's geometric center (i.e., its centroid) is computed. This is done by computing the mean values of the objects contained in the cluster (e.g., A, B, C in the first cluster) in terms of each of the variables (price consciousness and brand loyalty). As we can see in ■ Fig. 9.12, both clusters' centers now shift to new positions (CC1' in the first and CC2' in the second cluster; the inverted comma indicates that the cluster center has changed).



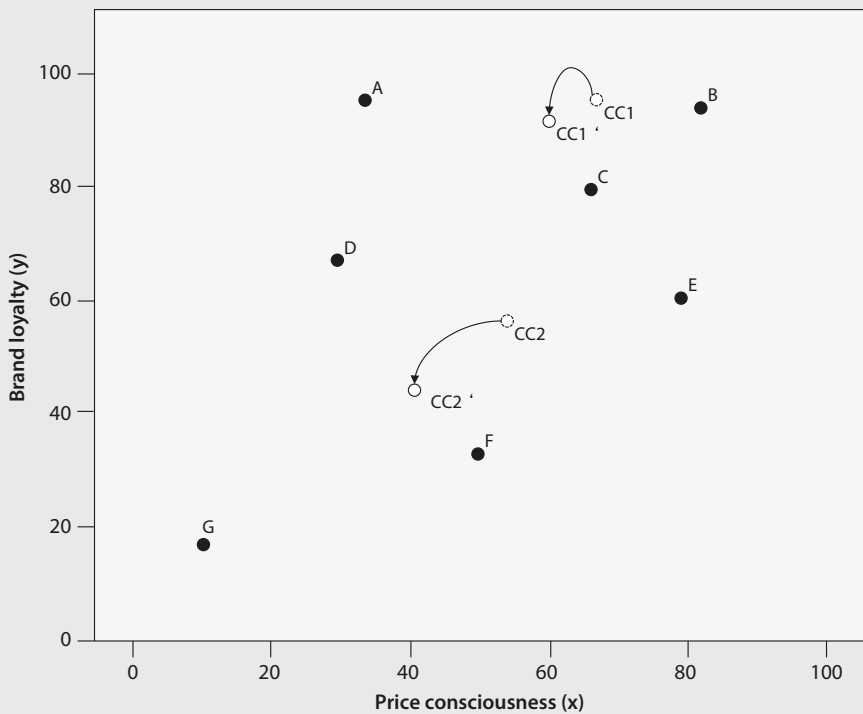
■ **Fig. 9.11** *k*-means procedure (step 2: assigning objects to the closest cluster center)

- **Step 4:** The distances are computed from each object to the newly located cluster centers and the objects are again assigned to a certain cluster on the basis of their minimum distance to other cluster centers (CC1' and CC2'). Since the cluster centers' position changed with respect to the initial situation, this could lead to a different cluster solution. This is also true of our example, because object E is now—unlike in the initial partition—closer to the first cluster center (CC1') than to the second (CC2'). Consequently, this object is now assigned to the first cluster (■ [Fig. 9.13](#)).

The *k*-means procedure is now repeated until a predetermined number of iterations are reached, or convergence is achieved (i.e., there is no change in the cluster affiliations).

Tip

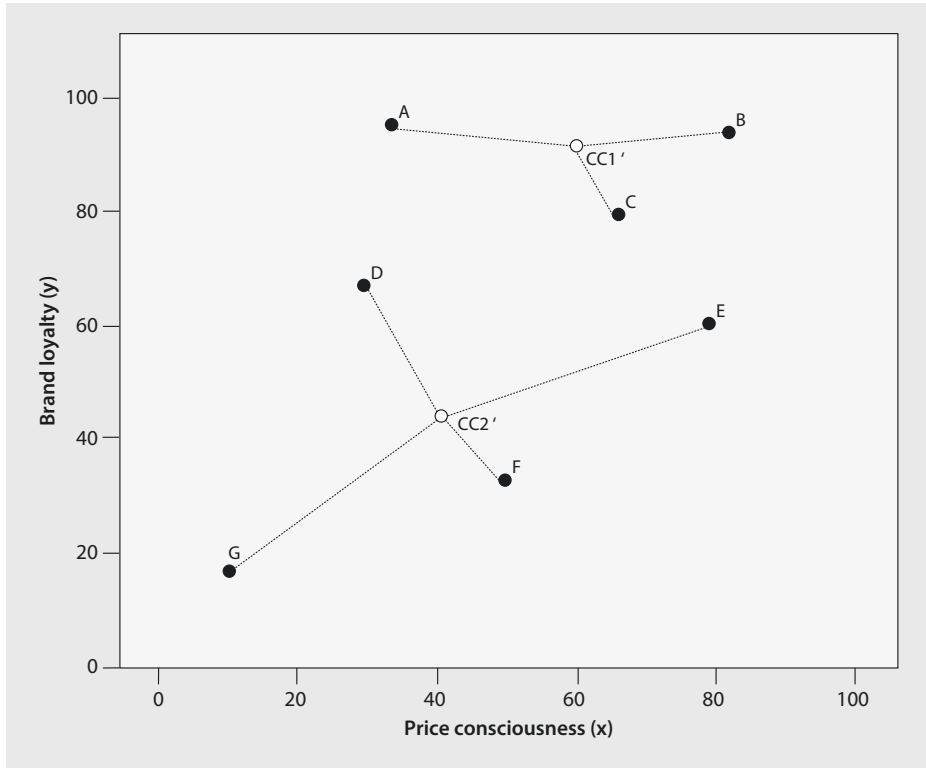
Naftali Harris's website offers a nice visualization of *k*-means clustering:
<https://www.naftaliharris.com/blog/visualizing-k-means-clustering/>



■ Fig. 9.12 *k*-means procedure (step 3: re-computing cluster centers)

Three aspects are worth noting in terms of using *k*-means:

- *k*-means is implicitly based on pairwise Euclidean distances, because the sum of the squared distances from the centroid is equal to the sum of the pairwise squared Euclidean distances divided by the number of objects. Hence, SPSS does not allow for selecting a distance measure—as in hierarchical clustering—but uses Euclidean distances. Therefore, the method should only be used with metric and, in case of equidistant scales, ordinal variables.
- Results produced by *k*-means depend on the starting partition. That is, *k*-means produce different results, depending on the starting partition chosen by the researcher or initiated by the software. In SPSS, the initialization depends on the ordering of the objects. As a result, *k*-means may converge in a **local optimum**, which means that the solution is only optimal compared to similar solutions, but not globally. Therefore, you should run *k*-means multiple with objects sorted in different random orders to verify the stability of a given solution.
- *k*-means is less computationally demanding than hierarchical clustering techniques. The method is therefore generally preferred for sample sizes above 500, and particularly for *big data* applications.
- Running *k*-means requires specifying the number of clusters to retain prior to running the analysis. We discuss this issue in the next section.



■ Fig. 9.13 *k*-means procedure (step 4: reassigning objects to the closest cluster center)

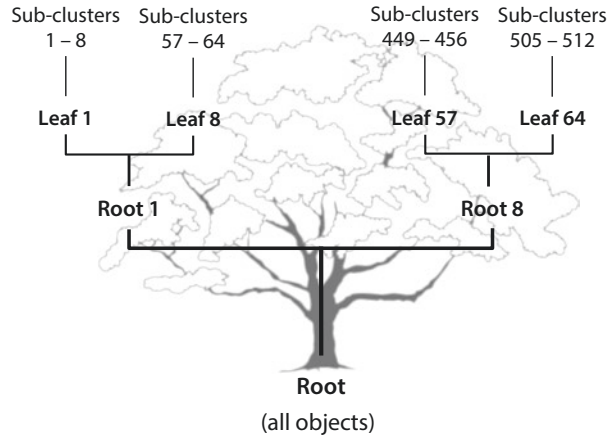
9.3.2.3 Two-Step Cluster Analysis

Chiu et al.'s (2001) **two-step cluster analysis** is an alternative to *k*-means for very large datasets. As its name implies, the method follows a two-stage approach.

In the first stage, the method merges all objects into sub-clusters. To do so, the method successively screens all objects to decide whether an object is merged with an existing cluster or establishes a new sub-cluster. Thereby two-step clustering establishes a **cluster feature tree** with roots and leaves (■ Fig. 9.14). Each of potentially eight roots consists of a maximum number of eight leaves. Each leaf has a maximum number of eight sub-clusters. Hence, two-step clustering allows for a maximum number of $8 \cdot 8 \cdot 8 = 512$ sub-clusters. Sub-clusters in one leaf are similar to each other, as defined by the distance measure, whereas sub-clusters in different leaves are distinct. By establishing a cluster feature tree, two-step cluster analysis reduces computing time, which is an issue for very large datasets. In the second stage, two-step cluster analysis uses a modified hierarchical agglomerative clustering procedure to merge the sub-clusters.

One crucial advantage of the two-step cluster analysis is that it can handle categorical and continuous variables simultaneously. Hierarchical clustering and *k*-means are clearly limited in this regard as these methods require continuous variables (*k*-means) or variables measured on either a categorical, ordinal, or continuous scale (hierarchical clustering). Furthermore, two-step clustering allows for automatically selecting the number of

■ Fig. 9.14 Cluster feature tree



clusters based on statistical criteria. The procedure also indicates each variable's importance for the construction of a specific cluster. Finally, two-step cluster analysis also offers an overall goodness-of-fit measure called **silhouette measure of cohesion and separation**. It is essentially based on the average distances between the objects and can vary between -1 and $+1$. A value of less than 0.20 indicates a poor solution quality, a value between 0.20 and 0.50 a fair solution, whereas values higher than 0.50 indicate a good solution. These desirable features make the somewhat less popular two-step clustering a good alternative to the traditional methods.

9.3.3 Select a Measure of Similarity or Dissimilarity

In the previous section, we discussed different linkage algorithms used in agglomerative hierarchical clustering, the k -means procedure as well as two-step clustering. All these clustering procedures rely on measures that express the (dis)similarity between pairs of objects. In the following section, we introduce different measures for metric, ordinal, nominal, and binary variables.

9.3.3.1 Metric and Ordinal Variables

■ Distance Measures

A straightforward way to assess two objects' proximity is by drawing a straight line between them. For example, on examining the scatter plot in ■ Fig. 9.1, we can easily see that the length of the line connecting observations B and C is much shorter than the line connecting B and G. This type of distance is called **Euclidean distance** or **straight line distance**; it is the most commonly used type for analyzing metric variables and, if the scales are equidistant (► Chap. 3), ordinal variables. Researchers also often use the **squared Euclidean distance**.

In order to use a clustering procedure, we need to express these distances mathematically. Using the data from Table 9.1, we can compute the Euclidean distance between customer B and customer C (generally referred to as $d(B,C)$) by using variables x and y with the following formula:

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

As can be seen, the Euclidean distance is the square root of the sum of the squared differences in the variables' values. Using the data from ■ Table 9.1, we obtain the following:

$$d_{Euclidean}(B,C) = \sqrt{(82 - 66)^2 + (94 - 80)^2} = \sqrt{452} \approx 21.260$$

This distance corresponds to the length of the line that connects objects B and C. In this case, we only used two variables, but we can easily add more under the root sign in the formula. However, each additional variable will add a dimension (e.g., with six clustering variables, we have to deal with six dimensions), making it difficult to represent the solution graphically. Similarly, we can compute the distance between customer B and G, which yields the following:

$$d_{Euclidean}(B,G) = \sqrt{(82 - 10)^2 + (94 - 17)^2} = \sqrt{11,113} \approx 105.418$$

We should also compute the distance between all other pairs of objects and summarize them in a distance matrix. ■ Table 9.2 shows the Euclidean distance matrix for objects A-G.

There are also alternative distance measures: The **city-block distance** uses the sum of the variables' absolute differences. This distance measure is referred to as the **Manhattan metric** as it is akin to the walking distance between two points in a city like New York's Manhattan district, where the distance equals the number of blocks in the directions North-South and East-West. Using the city-block distance to compute the distance between customers B and C (or C and B) yields the following:

$$d_{City-block}(B,C) = |x_B - x_C| + |y_B - y_C| = |82 - 66| + |94 - 80| = 30$$

The resulting distance matrix is shown in ■ Table 9.8.

■ Table 9.8 City-block distance matrix

Objects	A	B	C	D	E	F	G
A	0						
B	50	0					
C	48	30	0				
D	31	79	49	0			
E	81	37	33	56	0		
F	79	93	63	54	56	0	
G	101	149	119	70	112	56	0

Lastly, when working with metric (or ordinal) data, researchers frequently use the **Chebychev distance**, which is the maximum of the absolute difference in the clustering variables' values. For customers B and C, this is calculated as:

$$d_{Chebychev}(B,C) = \max(|x_B - x_C|, |y_B - y_C|) = \max(|82 - 66|, |94 - 80|) = 16$$

■ **Figure 9.15** illustrates the interrelation between these three distance measures regarding two objects (here: B and G) from our example.

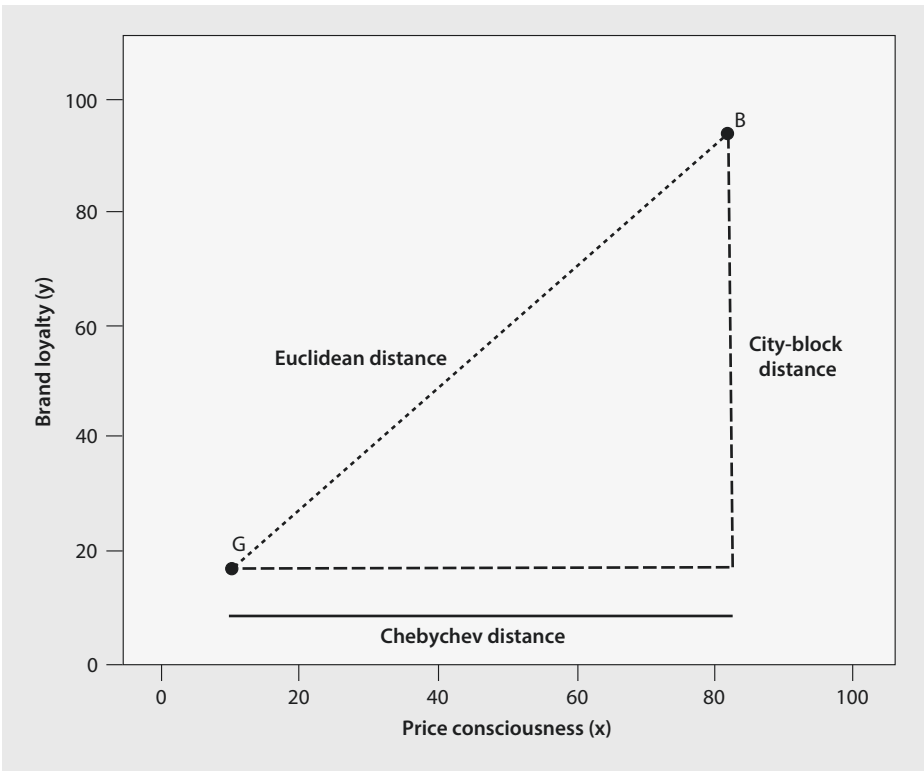
Tip

Different distance measures typically lead to different cluster solutions. Thus, it is advisable to use several measures, check for the stability of results, and compare them with theoretical or known patterns.

■ **Association Measures**

The (dis)similarity between objects can also be expressed using *association measures* (e.g., correlations). For example, suppose a respondent rated price consciousness 2 and brand loyalty 3, a second respondent indicated 5 and 6, whereas a third rated these variables 3 and 3. Euclidean and city-block distances would indicate that the first respondent is more

9



■ **Fig. 9.15** Distance measures

similar to the third than to the second. Nevertheless, one could convincingly argue that the first respondent's ratings are more similar to the second's, as both rate brand loyalty higher than price consciousness. This can be accounted for by computing the correlation between two vectors of values as a measure of similarity (i.e., high correlation coefficients indicate a high degree of similarity). Consequently, similarity is no longer defined as the difference between the answer categories, but as the similarity of the answering profiles.

Tip

Whether you use one of the distance measures or correlations depends on whether you think the relative magnitude of the variables within an object (which favors correlation) matters more than the relative magnitude of each variable across the objects (which favors distance). Some researchers recommended using correlations when applying clustering procedures that are particularly susceptible to outliers, such as complete linkage, average linkage, or centroid linkage. Furthermore, correlations implicitly standardize the data, as differences in the scale categories do not have a strong bearing on the interpretation of the response patterns. Nevertheless, distance measures are most commonly used for their intuitive interpretation. Distance measures best represent the concept of proximity, which is fundamental to cluster analysis. Correlations, although having widespread application in other techniques, represent patterns rather than proximity.

■ Standardizing the Data

In many analysis tasks, the variables under consideration are measured in different units with hugely different variance. This would be the case if we extended our set of clustering variables by adding another metric variable representing the customers' gross annual income. Since the absolute variation of the income variable would be much higher than the variation of the remaining two variables (remember, x and y are measured on a scale from 0 to 100), this would significantly change our analysis results. We can resolve this problem by standardizing the data prior to the analysis (► Chap. 5).

Different standardization methods are available, such as z -standardization, which rescales each variable to a mean of 0 and a standard deviation of 1 (see ► Chap. 5). In cluster analysis, however, *range standardization* (e.g., to a range of 0 to 1) typically works better (Milligan and Cooper 1988).

9.3.3.2 Binary and Nominal Variables

Whereas the distance measures presented thus far can be used for variables measured on a metric and, in general, on an ordinal scale, applying them to binary and nominal variables is problematic. When nominal variables are involved, you should instead select a similarity measure expressing the degree to which the variables' values share the same category. These **matching coefficients** can take different forms, but rely on the same allocation scheme as shown in ■ Table 9.9. In this crosstab, cell a is the number of characteristics present in both objects A and B, whereas cell d describes the number of characteristics absent in both objects. Cells b and c describe the number of characteristics present in one, but not the other, object.

■ **Table 9.9** Allocation scheme for matching coefficients

		Second object	
		Presence of a characteristic (1)	Absence of a characteristic (0)
First object	Presence of a characteristic (1)	a	b
	Absence of a characteristic (0)	c	d

The allocation scheme in ■ **Table 9.9** applies to binary variables (i.e., nominal variables with two categories). For nominal variables with more than two categories, you need to convert the categorical variable into a set of binary variables in order to use matching coefficients. For example, a variable with three categories needs to be transformed into three binary variables, one for each category (see the following example).

Based on the allocation scheme in ■ **Table 9.9**, we can compute different matching coefficients, such as the **simple matching (SM) coefficient**:

$$SM = \frac{a + d}{a + b + c + d}$$

This coefficient takes both the joint presence and the joint absence of a characteristic (as indicated by cells *a* and *d* in ■ **Table 9.9**) into account. This feature makes the simple matching coefficient particularly useful for symmetric variables where the joint presence and absence of a characteristic carry an equal degree of information. For example, the binary variable *gender* has the possible states “male” and “female.” Both are equally valuable and carry the same weight when the simple matching coefficient is computed. However, when the outcomes of a binary variable are not equally important (i.e., the variable is asymmetric), the simple matching coefficient proves problematic. An example of an asymmetric variable is the presence, or absence, of a relatively rare attribute, such as customer complaints. While you say that two customers who complained have something in common, you cannot say that customers who did not complain have something in common. The most important outcome is usually coded as 1 (present) and the other is coded as 0 (absent). The agreement of two 1s (i.e., a positive match) is more significant than the agreement of two 0s (i.e., a negative match). Similarly, the simple matching coefficient proves problematic when used on nominal variables with many categories. In this case, objects may appear very similar, because they have many negative matches rather than positive matches.

Given this issue, researchers have proposed several other matching coefficients, such as the **Jaccard coefficient (JC)** and the **Russell and Rao coefficient**, which (partially) omit the *d* cell from the calculation. Like the simple matching coefficient, these coefficients range from 0 to 1 with higher values indicating a greater degree of similarity.⁴ They are defined as follows:

4 There are many other matching coefficients, with exotic names such as *Yule's Q*, *Kulczynski*, or *Ochiai*, which are also menu-accessible in SPSS. As most applications of cluster analysis rely on metric or ordinal data, we will not discuss these. See Wedel and Kamakura (2000) for more information on alternative matching coefficients.

$$JC = \frac{a}{a + b + c}$$

$$RR = \frac{a}{a + b + c + d}$$

To provide an example that compares the three coefficients, consider the following three variables:

- *gender*: male, female
- *customer*: yes, no
- *country of residence*: GER, UK, USA

We first transform the measurement data into binary data by recoding the original three variables into seven binary variables (i.e., two for *gender* and *customer*; three for *country of residence*). ■ **Table 9.10** shows a binary data matrix for three objects A, B, and C. Object A is a male customer from Germany; object B is a male non-customer from the United States; object C is a female non-customer, also from the United States.

Using the allocation scheme from ■ **Table 9.9** to compare objects A and B yields the following results for the cells: $a = 1$, $b = 2$, $c = 2$, and $d = 2$. This means that the two objects have only one shared characteristic ($a = 1$), but two characteristics, which are absent from both objects ($d = 2$). Using this information, we can now compute the three coefficients described earlier:

$$SM(A, B) = \frac{1 + 2}{1 + 2 + 2 + 2} = 0.571,$$

$$JC(A, B) = \frac{1}{1 + 2 + 2} = 0.2$$

, and

$$RR(A, B) = \frac{1}{1 + 2 + 2 + 2} = 0.143$$

As we can see, the simple matching coefficient suggests that objects A and B are reasonably similar. Conversely, the Jaccard coefficient, and particularly the Russel Rao coefficient, suggests that they are not.

■ **Table 9.10** Recoded measurement data

Object	Gender (binary)		Customer (binary)		Country of residence (binary)		
	Male	Female	Yes	No	GER	UK	USA
A	1	0	1	0	1	0	0
B	1	0	0	1	0	0	1
C	0	1	0	1	0	0	1

Try computing the distances between the other object pairs. Your computation should yield the following: $SM(A,C) = 0.143$, $SM(B,C) = 0.714$, $JC(A,C) = 0$, $JC(B,C) = 0.5$, $RR(A,C) = 0$, and $RR(B,C) = 0.286$.

9.3.3.3 Mixed Variables

Most datasets contain variables that are measured on multiple scales. For example, a market research questionnaire may require the respondent's gender, income category, and age. We therefore have to consider variables measured on a nominal, ordinal, and metric scale. How can we simultaneously incorporate these variables into an analysis?

Often research use the distance measures discussed in the context of metric (and ordinal) data. Even though this approach may slightly change the results compared to using matching coefficients, it should not be rejected. Cluster analysis is mostly an exploratory technique whose results only provide guidance for making decisions but are no substitute for decision-making.

An alternative is to dichotomize all the variables and apply the matching coefficients discussed above. For metric variables, this involves specifying categories (e.g., low, medium, and high age) and converting these into sets of binary variables. In most cases, the specification of categories is somewhat arbitrary. Furthermore, this procedure leads to a severe loss in precision, as we disregard more detailed information on each object. For example, we lose precise information on each respondent's age when scaling this variable down into age categories. Given such issues, you should avoid combining metric and nominal variables in a single cluster analysis.

Another way to handle variables measured on different scale levels is to use the two-step cluster analysis (see ► Sect. 9.3.2.3). This method uses a distance measure that draws on probability distributions. Specifically, this distance defines the distance between two objects in terms of the decrease of the likelihood value when merging them.

9.3.4 Decide on the Number of Clusters

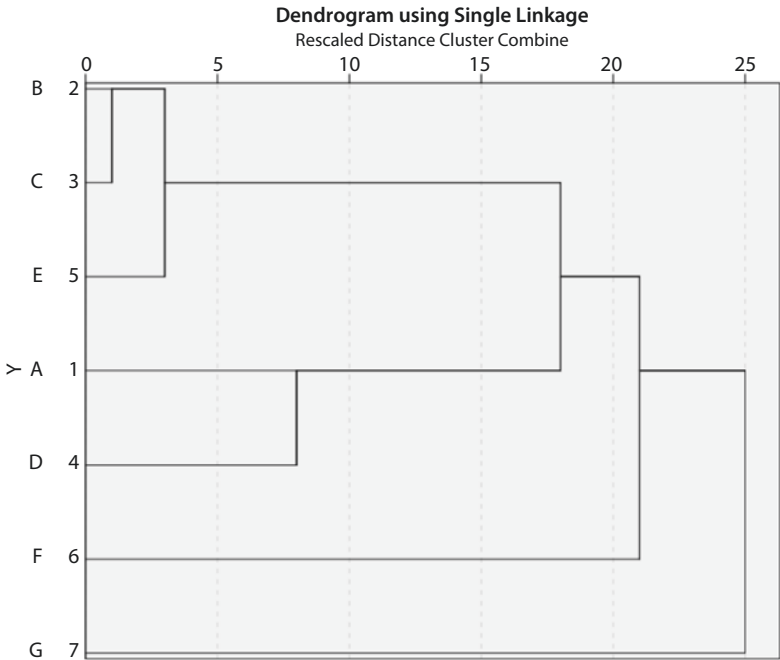
An important question we haven't yet addressed is how to decide on the number of clusters. A misspecified number of clusters results in under- or oversegmentation, which easily leads to inaccurate management decisions on, for example, customer targeting, product positioning, or determining the optimal marketing mix (Becker et al. 2015).

We can select the number of clusters pragmatically, choosing a grouping that “works” for our analysis, but sometimes we want to select the “best” solution that the data suggest. However, different clustering methods require different approaches to decide on the number of clusters. Hence, we discuss hierarchical and portioning methods separately.

9.3.4.1 Hierarchical Methods

To guide the decision of how many clusters to extract from the data, we can draw on the distances at which the objects were combined. More precisely, we can seek a solution in which an additional combination of clusters or objects would occur at a greatly increased distance. This raises the issue of what a great distance is.

We can seek an answer by plotting the distance level at which the mergers of objects and clusters occur by using a **dendrogram**. ■ Figure 9.16 shows the dendrogram for our example as produced by SPSS. We read the dendrogram from the left to the right. The



■ Fig. 9.16 Dendrogram

horizontal lines indicate the distances at which the objects were merged. Note that in SPSS, these distances do not correspond to the actual merging distances as computed in Tables 9.2, 9.3, 9.4, 9.5 and 9.7. Instead, SPSS rescales the distances to a range of 0–25 (i.e., the last merging step to a one-cluster solution takes place at a rescaled distance of 25). The rescaling on the x-axis facilitates the decision on how many clusters to extract from the data. Specifically, to decide on the number of clusters, we cut the dendrogram vertically in the area where no merger has occurred for a long distance. In our example, this is done when moving from a four-cluster solution, which occurs at a rescaled distance of 8, to a three-cluster solution, which occurs at a distance of 18. This result suggests a four-cluster solution [A,D], [B,C,E], [F], and [G], but this conclusion is not clear-cut. In fact, the dendrogram often does not provide a clear indication, because it is generally difficult to identify where the cut should be made. This is particularly true of large sample sizes when the dendrogram becomes unwieldy.

As an alternative to the dendrogram, we can also contrast the distances against the number of clusters to produce a **scree plot**, similar to the one used to decide on the number of factors in factor analysis (► Chap. 8). Specifically, we can plot the number of clusters on the x-axis (starting with the one-cluster solution at the very left) against the distance at which objects or clusters are merged on the y-axis. Using this plot, we then search for the distinctive break (*elbow*), which indicates the number of clusters to retain. Note that—unlike in factor analysis—we do not pick the solution with one cluster less than indicated by the elbow. Furthermore, the distances typically sharply increase when switching from a two-cluster solution to a one-cluster solution. However, this break should not be viewed as a reliable indicator for the decision regarding the number of segments.

Research has produced several other criteria for determining the number of clusters in a dataset. One of the most prominent criteria is Calinski and Harabasz's (1974) **variance ratio criterion (VRC)**. For a solution with n objects and k clusters, the VRC is defined as:

$$VRC_k = (SS_B / (K - 1)) / (SS_W / (n - K)),$$

where SS_B is the sum of the squares between the clusters and SS_W is the sum of the squares within the clusters. The criterion should seem familiar, as it is equivalent to the F -value of a one-way ANOVA (► Chap. 6). To determine the appropriate number of clusters, you should choose the number that maximizes the VRC. However, as the VRC usually decreases with a greater number of clusters, you should compute the difference in the VRC values ω_k of each cluster solution, using the following formula:

$$\omega_k = (VRC_{k+1} - VRC_k) - (VRC_k - VRC_{k-1}).$$

The number of clusters k that minimizes the value in ω_k indicates the best cluster solution. Prior research has shown that the VRC reliably identifies the correct number of clusters across a broad range of constellations (Miligan and Cooper 1985). However, owing to the term VRC_{k-1} , which is not defined for a one-cluster solution, the minimum number of clusters that can be selected is three, which is a disadvantage when using the ω_k statistic.

To compute the VRC, we need to run a series of ANOVAs using the clustering variables as dependent variables and the cluster affiliation as the factor variable. The VRC for a certain number of clusters k results from summing all the F -value across the different ANOVAs. Note that the computation of the VRC values is more straightforward when running k -means clustering as SPSS allows running ANOVAs on the clustering variables as part of this clustering procedure.

➤ **Overall, the above criteria can often only provide rough guidance regarding the number of clusters that should be selected—you should also take practical considerations into account. Occasionally, you might have a priori knowledge, or a theory on which you can base your choice. However, first and foremost, you should ensure that your results are interpretable and meaningful. Not only must the number of clusters be small enough to ensure manageability, but each segment should also be large enough to warrant strategic attention.**

9.3.4.2 Partitioning Methods

When running partitioning methods, such as k -means, you have to pre-specify the number of clusters to retain from the data. There are varying ways of guiding this decision:

- Compute the VRC (see discussion in the context of hierarchical clustering) for an alternating number of clusters and select the solution that maximizes the VRC or minimizes ω_k . For example, compute the VRC for a three- to five-cluster solution and select the number of clusters that minimizes ω_k .
- Run a hierarchical procedure to determine the number of clusters by using the dendrogram and run k -means afterwards.⁵ This approach also enables you to find

⁵ See Punji and Stewart (1983) for additional information on this sequential approach.

starting values for the initial cluster centers to handle a second problem, which relates to the procedure's sensitivity to the initial classification (we will follow this approach in the example application).

- Rely on prior information, such as earlier research findings.

9.3.4.3 Two-step Clustering

One crucial advantage of two-step clustering is that the method allows for automatically selecting the number of clusters based on statistical criteria. In doing so, two-step clustering follows a two-stage approach (Bacher et al. 2004).

In the first stage, the method determines a maximum number of clusters based on *Akaike's Information Criterion* (AIC; Akaike 1973) or the *Bayes Information Criterion* (BIC; Schwarz 1978), depending on the researcher's specification. These criteria add different terms to the log likelihood value resulting from the analysis, which penalize the complexity of the solution as expressed by the number of clusters—solutions with a more clusters entail a stronger penalty term. In SPSS, the maximum number of clusters is determined by the ratio between AIC (or BIC) for a solution with k clusters and a one-cluster solution. The solution for which this ratio is smaller than a certain threshold assumed by the program is the maximum number of clusters.

In the second stage, two-step clustering computes the ratio of distances between different cluster solutions using the AIC (or BIC) values as input. The resulting ratio determines the final number of clusters to extract.

9.3.5 Validate and Interpret the Clustering Solution

Before interpreting the cluster solution, we need to assess the stability of the results. Stability means that the cluster membership of individuals does not change, or only changes a little when different clustering methods are used to cluster the objects. Thus, when different methods produce similar results, we claim stability.

The aim of any cluster analysis is to differentiate well between the objects. The identified clusters should therefore differ substantially from each other and the members of different clusters should respond differently to different marketing-mix elements and programs.

Lastly, we need to profile the cluster solution by using observable variables. **Profiling** ensures that we can easily assign new objects to clusters based on observable traits. For example, we could identify clusters based on loyalty to a product, but in order to use these different clusters, their membership should be identifiable according to tangible variables, such as income, location, or family size, in order to be actionable.

The key to successful segmentation is to critically revisit the results of different cluster analysis set-ups (e.g., by using different algorithms on the same data) in terms of managerial relevance. The following criteria help identify a clustering solution (Kotler and Keller 2015; Tonks 2009).

- **Substantial:** The clusters are large and sufficiently profitable to serve.
- **Reliable:** Only clusters that are stable over time can provide the necessary basis for a successful marketing strategy. If clusters change their composition quickly, or their members' behavior, targeting strategies are not likely to succeed. Therefore, a certain degree of stability

is necessary to ensure that marketing strategies can be implemented and produce adequate results. Reliability can be evaluated by critically revisiting and replicating the clustering results at a later date.

- *Accessible*: The clusters can be effectively reached and served.
- *Actionable*: Effective programs can be formulated to attract and serve the clusters.
- *Parsimonious*: To be managerially meaningful, only a small set of substantial clusters should be identified.
- *Familiar*: To ensure management acceptance, the cluster composition should be easy to relate to.
- *Relevant*: Clusters should be relevant in respect of the company's competencies and objectives.

9.3.5.1 Stability

Stability is evaluated by using different clustering procedures on the same data and considering the differences that occur. For example, you may first run a hierarchical clustering procedure, followed by k -means clustering to check whether the cluster affiliations of the objects change. Alternatively, running a hierarchical clustering procedure, you can use different distance measures and evaluate their effect on the stability of the results. However, note that it is common for results to change even when your solution is adequate. As a rule of thumb, if more than 20 % of the cluster affiliations change from one technique to the other, you should reconsider the analysis and use, for example, a different set of clustering variables, or reconsider the number of clusters. Note, however, that this percentage is likely to increase with the number of clusters used.

When the data matrix exhibits identical values (referred to as *ties*), the ordering of the objects in the dataset can influence the results of the hierarchical clustering procedure. For example, when computing the distance matrix based on the city-block distance for the data from ■ Table 9.1, object pairs (D,E), (E,F), and (F,G) have the same distance of 56 units. Ties can prove problematic when they occur for the minimum distance in a distance matrix, as the decision about which objects to merge then becomes ambiguous (i.e., should we merge objects D and E, E and F, or F and G if 56 was the smallest distance in the matrix?). To handle this problem, Van Der Kloot et al. (2005) recommend re-running the analysis with a different input order of the data. The downside of this approach is that the labels of a cluster may change from one analysis to the next. This issue is referred to as **label switching**. For example, in the first analysis, cluster 1 may correspond to cluster 2 in the second analysis. Ties are, however, more the exception than the rule in practical applications—especially when using (squared) Euclidean distances—and generally don't have a pronounced impact on the results. However, if changing the order of the objects also drastically changes the cluster compositions (e.g., in terms of cluster sizes), you should reconsider the set-up of the analysis and, for example, re-run it with different clustering variables.

9.3.5.2 Differentiation of the Data

To examine whether the final partition differentiates the data well, we need to examine the cluster centroids. This step is highly important, as the analysis sheds light on whether the clusters are truly distinct. Only if objects across two (or more) clusters exhibit significantly different means in the clustering variables (or any other relevant variable) can

they be distinguished from each other. This can be easily ascertained by comparing the means of the clustering variables across the clusters with independent *t*-tests or ANOVA (see ► Chap. 6).

Furthermore, we need to assess the solution's *criterion validity* (see ► Chap. 4). We do this by focusing on the criterion variables that have a theoretical relationship with the clustering variables, but were not included in the analysis. In market research, criterion variables are usually managerial outcomes, such as the sales per person, or willingness-to-pay. If these criterion variables differ significantly, we can conclude that the clusters are distinct groups with criterion validity.

9.3.5.3 Profiling

As indicated at the beginning of the chapter, cluster analysis usually builds on unobservable clustering variables. This creates an important problem when working with the final solution: How can we decide to which cluster a new object should be assigned if its unobservable characteristics, such as personality traits, personal values, or lifestyles, are unknown? We could survey these attributes and make a decision based on the clustering variables. However, this is costly and researchers therefore usually try to identify observable variables (e.g., demographics) that best mirror the partition of the objects. More precisely, these observable variables should partition the data into similar groups as the clustering variables do. Using these observable variables, it is then easy to assign a new object (whose cluster membership is unknown) to a certain cluster. For example, assume that we used a set of questions to assess the respondents' values and learned that a certain cluster contains respondents who appreciate self-fulfillment, enjoyment of life, and a sense of accomplishment, whereas this is not the case in another cluster. If we were able to identify explanatory variables, such as gender or age, which distinguish these clusters adequately, then we could assign a new person to a specific cluster on the basis of these observable variables whose value traits may still be unknown.

9.3.5.4 Interpret the Clustering Solution

The interpretation of the solution requires characterizing each cluster by using the criterion or other variables (in most cases, demographics). This characterization should focus on criterion variables that convey why the cluster solution is relevant. For example, you could highlight that customers in one cluster have a lower willingness to pay and are satisfied with lower service levels, whereas customers in another cluster are willing to pay more for a superior service. By using this information, we can also try to find a meaningful name or label for each cluster; that is, one that adequately reflects the objects in the cluster. This is usually a challenging task, especially when unobservable variables are involved.

While companies develop their own market segments, they frequently use standardized segments, based on established buying trends, habits, and customers' needs to position their products in different markets. The *PRIZM* lifestyle by Nielsen is one of the most popular segmentation databases. It combines demographic, consumer behavior, and geographic data to help marketers identify, understand, and reach their customers and prospective customers. *PRIZM* defines every US household in terms of more than 60 distinct segments to help marketers discern these consumers' likes, dislikes, lifestyles, and purchase behaviors.

An example is the segment labeled “Connected Bohemians,” which Nielsen characterizes as a “collection of mobile urbanites, Connected Bohemians represent the nation’s most liberal lifestyles. Its residents are a progressive mix of tech savvy, young singles, couples, and families ranging from students to professionals. In their funky row houses and apartments, Bohemian Mixers are the early adopters who are quick to check out the latest movie, nightclub, laptop, and microbrew.” Members of this segment are between 25 and 44 years old, have a midscale income, own a hybrid vehicle, eat at Starbucks, and go skiing/snowboarding. (www.MyBestSegments.com).

■ **Table 9.11** summarizes the steps involved in a hierarchical, *k*-means, and two-step clustering using SPSS.

■ **Table 9.11** Steps involved in carrying out a cluster analysis in SPSS

Theory	Action
<i>Research problem</i>	
Identification of homogenous groups of objects in a population	
Select clustering variables to form segments	Select relevant variables that potentially exhibit high degrees of criterion validity with regard to a specific managerial objective.
<i>Requirements</i>	
Sufficient sample size	Make sure that the relationship between the objects and the clustering variables is reasonable. Ten times the number of clustering variables is the bare minimum, but 30 to 70 times is recommended. Ensure that the sample size is large enough to guarantee substantial segments.
Low levels of collinearity among the variables	► Analyze ► Correlate ► Bivariate In case of highly correlated variables (correlation coefficients > 0.90), delete one variable of the offending pair.
<i>Specification</i>	
Choose the clustering procedure	If there is a limited number of objects in your dataset or you do not know the number of clusters: ► Analyze ► Classify ► Hierarchical Cluster
	If there are many observations (> 500) in your dataset and you have a priori knowledge regarding the number of clusters: ► Analyze ► Classify ► K-Means Cluster
	If there are many observations in your dataset and the clustering variables are measured on different scale levels: ► Analyze ► Classify ► Two-Step Cluster
Choose clustering algorithm (only hierarchical clustering)	► Analyze ► Classify ► Hierarchical Cluster ► Method ► Cluster Method Use Ward’s method if equally sized clusters are expected and no outliers are present. Preferably use single linkage, also to detect outliers.

■ **Table 9.11** (Continued)

Theory	Action
Select a measure of (dis)similarity	<p><i>Hierarchical methods:</i></p> <p>► Analyze ► Classify ► Hierarchical Cluster ► Method ► Measure</p> <p>Depending on the scale level, select the measure; convert variables with multiple categories into a set of binary variables and use matching coefficients; standardize variables if necessary (on a range of 0 to 1).</p>
	<p><i>k-means clustering:</i></p> <p>Uses Euclidean distances per default.</p>
	<p><i>Two-step clustering:</i></p> <p>► Analyze ► Classify ► Two-Step Cluster ► Distance Measure</p> <p>Use Euclidean distances when all variables are continuous; for mixed variables, you have to use the log-likelihood.</p>
Deciding on the number of clusters	<p><i>Hierarchical clustering:</i></p> <p>Examine the dendrogram:</p> <p>► Analyze ► Classify ► Hierarchical Cluster ► Plots ► Dendrogram</p> <p>Draw a scree plot: Double-click on the Agglomeration Schedule in the output window, highlight all coefficients in the column and right-click the mouse button. In the menu that opens up, select Create Graph ► Line</p> <p>Compute the VRC using an ANOVA:</p> <p>► Analyze ► Compare Means ► One-Way ANOVA</p> <p>Move the cluster membership variable in the Factor box and the clustering variables in the Dependent List box;</p> <p>Compute VRC for each segment solution and compare values.</p> <p>Include practical considerations in your decision.</p>
	<p><i>k-means:</i></p> <p>Run a hierarchical cluster analysis and decide on the number of segments based on a dendrogram or scree plot; use this information to run <i>k-means</i> with <i>k</i> clusters.</p> <p>Compute the VRC using an ANOVA:</p> <p>► Analyze ► Classify ► K-Means Cluster ► Options ► ANOVA table;</p> <p>Compute VRC for each segment solution and compare values.</p> <p>Include practical considerations in your decision.</p>
	<p><i>Two-step clustering:</i></p> <p>Specify the maximum number of clusters:</p> <p>► Analyze ► Classify ► Two-Step Cluster ► Number of Clusters</p> <p>Run separate analyses using the AIC and BIC as clustering criteria:</p> <p>► Analyze ► Classify ► Two-Step Cluster ► Clustering Criterion</p> <p>Examine the model summary output.</p> <p>Include practical considerations in your decision.</p>

Table 9.11 (Continued)

Theory	Action
<i>Validating and interpreting the cluster solution</i>	
Stability	Re-run the analysis using different clustering procedures, algorithms or distance measures.
	Change the order of objects in the dataset.
Differentiation of the data	Compare the cluster centroids across the different clusters for significant differences.
	If possible, assess the solution's criterion validity.
Profiling	Identify observable variables (e.g., demographics) that best mirror the partition of the objects based on the clustering variables.
Interpreting of the cluster solution	Identify names or labels for each cluster and characterize each cluster using observable variables.

9

9.4 Example

Let's go back to the Oddjob Airways case study and run a cluster analysis on the data. Our aim is to identify a manageable number of segments that differentiates the customer base well. To do so, we first select a set of clustering variables, taking the sample size and potential collinearity issues into account. Next, we apply hierarchical clustering based on the squared Euclidean distances, using the Ward's linkage algorithm. This analysis will help us determine a suitable number of segments and a starting partition, which we will then use as the input for *k*-means clustering.

9.4.1 Hierarchical Cluster Analysis

9.4.1.1 Select the Clustering Variables

The Oddjob Airways dataset (↓ Web Appendix → Downloads) offers several variables for segmenting its customer base. Our analysis draws on the following set of variables, which we consider promising for identifying distinct segments based on customers' expectations regarding the airline's service quality (variable names in parentheses):

- With Oddjob Airways you will arrive on time (*e1*),
- Oddjob Airways provides you with a very pleasant travel experience (*e5*),
- Oddjob Airways gives you a sense of safety (*e9*),
- Oddjob Airways makes traveling uncomplicated (*e21*), and
- Oddjob Airways provides you with interesting on-board entertainment, service, and information sources (*e22*).

With five clustering variables, our analysis meets even the most conservative rule-of-thumb regarding minimum sample size requirements. Specifically, according to Dolnicar et al.

(2016), the cluster analysis should draw on 100 times the number of clustering variables to optimize cluster recovery. As our sample size of 1065 is clearly higher than $5 \cdot 100 = 500$, we can proceed with the analysis. Note, however, that the actual sample size used in the analysis may be substantially lower when using casewise deletion. This also applies to our analysis, which draws on 969 objects (i.e., after casewise deletion) as we can see in ■ Table 9.16.

To begin with, we examine the variable correlations by clicking on ► Analyze ► Correlate ► Bivariate. Next, enter the variables *e1*, *e5*, *e9*, *e21*, and *e22* into the **Variables** box (■ Fig. 9.17). Click on **OK** and SPSS will display the results (■ Table 9.12).

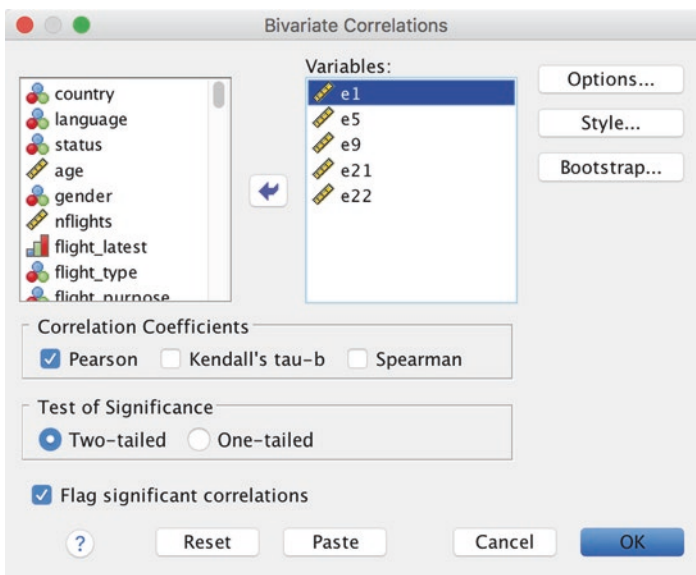
The results show that collinearity is not at a critical level. The variables *e1* and *e21* show the highest correlation of 0.613, which is clearly lower than the 0.90 threshold. We can therefore proceed with the analysis, using all five clustering variables.

9.4.1.2 Select the Clustering Procedure and a Measure of Similarity or Dissimilarity

To initiate hierarchical clustering, go to ► Analyze ► Classify ► Hierarchical Cluster, which opens a dialog box similar to ■ Fig. 9.18.

Move the variables *e1*, *e5*, *e9*, *e21*, and *e22* into the **Variable(s)** box. The **Statistics** option gives us the opportunity to request the distance matrix (labeled proximity matrix in this case) and the agglomeration schedule, which provides information on the objects being combined at each stage of the clustering process. Furthermore, we can specify the number or range of clusters to retain from the data. As we do not yet know how many clusters to retain, just check the box **Agglomeration schedule** and continue.

Under **Plots**, check the box **Dendrogram** to graphically display the distances at which objects and clusters are joined. SPSS also offers the option to display an **Iceberg** diagram (**All clusters**), which is yet another graph for displaying clustering solutions. Its name stems



■ Fig. 9.17 Bivariate correlations dialog box

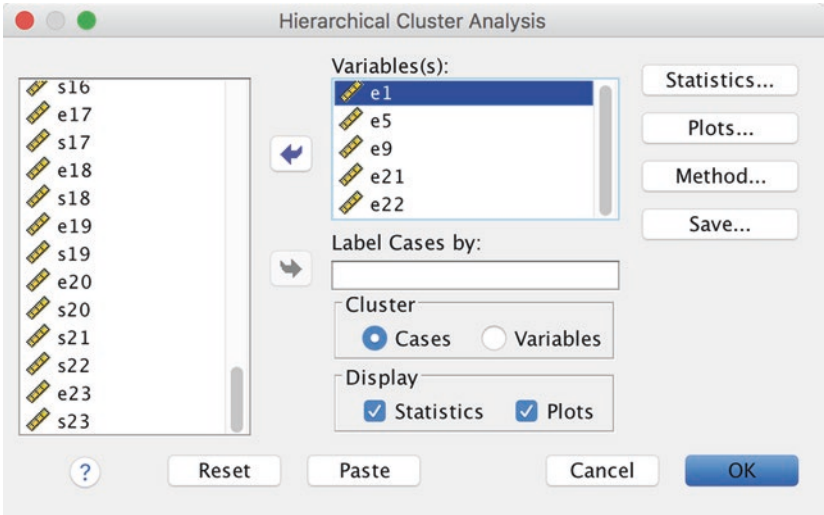
Table 9.12 Bivariate correlations output

		Correlations				
		e1	e5	e9	e21	e22
e1	Pearson Correlation	1	.515**	.533**	.613**	.370**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	1038	1026	1023	1018	997
e5	Pearson Correlation	.515**	1	.525**	.574**	.530**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	1026	1041	1023	1017	998
e9	Pearson Correlation	.533**	.525**	1	.522**	.417**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	1023	1023	1036	1016	996
e21	Pearson Correlation	.613**	.574**	.522**	1	.425**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	1018	1017	1016	1028	989
e22	Pearson Correlation	.370**	.530**	.417**	.425**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	997	998	996	989	1012

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

from the analogy to rows of icicles hanging from the eaves of a house. The diagram is read from the bottom to the top; the columns correspond to the objects being clustered, and the rows represent the number of clusters. Given the great number of objects, we do not request the icicle diagram in our example.

The option **Method** allows us to specify the cluster method, the distance measure, and the type of standardization of values. Because of its versatility and general performance, we choose the **Ward's method** and **Squared Euclidean distance** as distance measure. Even though all the variables used in our analysis are measured on a scale from 0 to 100, we standardize the data to account for differences in the variables' variances. To do so, go to the **Transform Values** drop-down menu and select **Range 0 to 1**.



■ Fig. 9.18 Hierarchical cluster analysis dialog box

Finally, the **Save** option enables us to save cluster memberships for a single solution or a range of solutions. Saved variables can then be used in subsequent analyses to explore differences between groups. As a start, we will skip this option, so continue and click on **OK** in the main menu.

9.4.1.3 Decide on the Number of Clusters

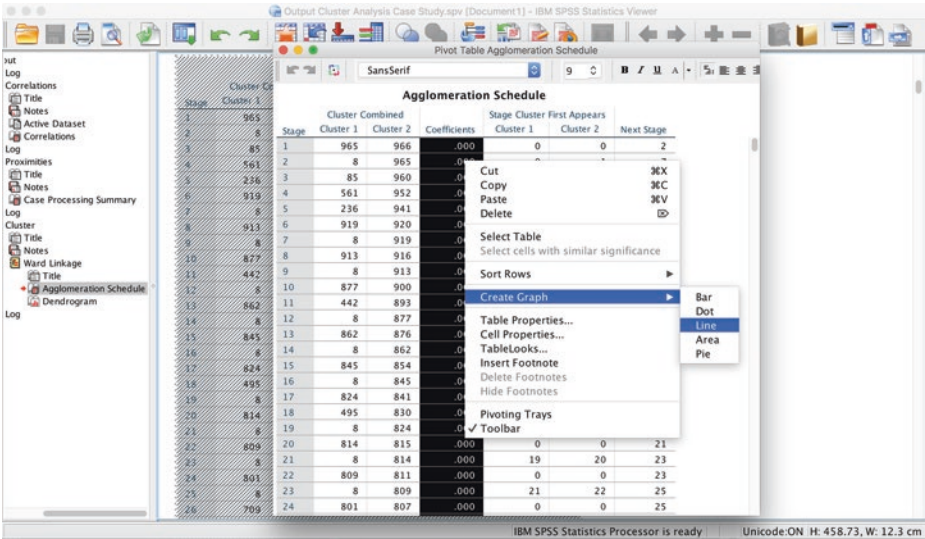
First, we take a closer look at the agglomeration schedule (■ Table 9.13), which displays the objects or clusters combined at each stage (columns **Cluster 1** and **Cluster 2**) and the distances at which this merger takes place (column **Coefficients**). Given the great number of objects, we limit the display of the agglomeration schedule to the merger stages 200 to 210. The table shows that in stage 200, objects 133 and 684 are merged at a distance of 0.046. From here onward, the resulting cluster is labeled as indicated by the first object involved in this merger, which is object 133. The last column on the very right tells you in which stage of the algorithm this cluster will appear next. In this case, this happens in stage 350, where this object is merged with object 409 at a distance of 0.359 (not shown).

Next, we use the agglomeration schedule to determine the number of segments to retain from the data. To do so, we generate a scree plot by plotting the distances (**Coefficients** column in ■ Table 9.13) against the number of clusters. The distinct break (elbow) indicates the solution regarding where an additional combination of two objects or clusters would occur at a greatly increased distance. Thus, the number of clusters prior to this merger is the most probable solution. SPSS does not automatically provide this plot. To generate a scree plot we have to double-click the **Agglomeration Schedule** in the output window. Next, highlight all coefficients in the column and right-click the mouse button. In the menu that opens up, select **Create Graph ▶ Line** (■ Fig. 9.19). SPSS will add a line chart to the output, which represents the scree plot.

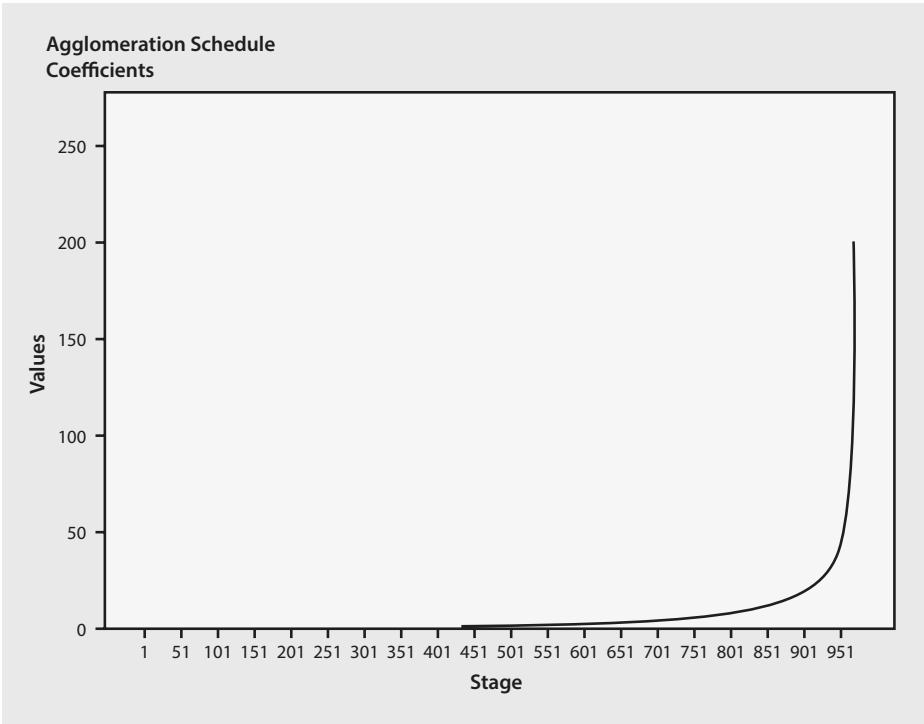
The scree plot in ■ Fig. 9.20 shows that there is no clear elbow indicating a suitable number of clusters to retain. This result is quite common for datasets with several hundred objects.

■ Table 9.13 Agglomeration schedule (partial screenshot)

Agglomeration Schedule						
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
...
200	133	684	.046	0	119	350
201	330	478	.047	91	0	429
202	723	881	.048	0	0	391
203	536	835	.049	0	0	319
204	250	712	.050	0	0	257
205	624	631	.051	0	0	363
206	370	505	.052	0	0	427
207	67	112	.053	0	0	370
208	444	853	.054	0	83	362
209	48	767	.055	0	0	325
210	563	572	.057	0	112	385
...



■ Fig. 9.19 Generating a scree plot



■ Fig. 9.20 Scree plot

Next, we should take a look at the dendrogram. We don't display the dendrogram here because of the size of the dataset. Reading the dendrogram from left to right, we find that the vast majority of objects are merged at very small distances. The dendrogram also shows that the step from a three-cluster solution to a two-cluster solution occurs at a greatly increased distance. Hence, we assume a three-cluster solution and continue with the analysis.

9.4.1.4 Validate and Interpret the Clustering Solution

To get a first impression of the size and nature of the three clusters, let's re-run the hierarchical cluster analysis, but this time, we pre-specify the number of segments. To do so, go back to ► Analyze ► Classify ► Hierarchical Cluster and select the **Save** option. In the dialog box that opens, select **Single solution** and enter **3** next to **Number of clusters**. Click on **Continue** followed by **OK**. When running the analysis, SPSS generates the same output but also adds one additional variable to your dataset (*CLU3_1*), which reflect each object's cluster membership. SPSS automatically places *CLU* in front, followed by a 3 to identify the total number of clusters. The variable's values (1, 2, and 3) identify each object's cluster membership.

To learn about the size of the clusters, go to ► Analyze ► Descriptive Statistics ► Frequencies and enter *CLU3_1* into the **Variable(s)** box. When clicking on **OK**, SPSS will open an output similar to ■ Table 9.14.

■ Table 9.14 Frequencies

		CLU3_1			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	516	48.5	53.3	53.3
	2	238	22.3	24.6	77.8
	3	215	20.2	22.2	100.0
	Total	969	91.0	100.0	
Missing	System	96	9.0		
Total		1065	100.0		

The output in ■ Table 9.14 shows that the cluster analysis assigned 969 objects to the three segments; 96 objects were not assigned to any segment due to missing values. The first cluster is the largest among the three clusters with 516 objects, which translates into a relative cluster size of 53.3 %. Clusters 2 and 3 are smaller and similar in size with 238 and 215 objects, respectively.

Next, we would like to compute the centroids of our clustering variables. To do so, split up the dataset using the **Split File** command (► Data ► Split File) (see ► Chap. 5). Choose *CLU3_1* as the grouping variable and select the option **Compare groups**. Next, go to ► Analyze ► Descriptive Statistics ► Descriptives (see ► Chap. 5) and request the mean, minimum, and maximum, as well as the standard deviations for the clustering variables *e1*, *e5*, *e9*, *e21*, and *e22*. ■ Table 9.15 shows the resulting output. The first column in the table indicates the cluster number with the first element (labeled with a dot) representing the group of missing values. However, we focus our analysis of the results on the first three groups and particularly the clustering variables' mean values.

Comparing the variable means across the three clusters, we find that respondents in the first cluster have extremely high expectations regarding all five performance features, as evidenced in average values of around 90 and higher. Respondents in the second cluster strongly emphasize punctuality (*e1*), while comfort (*e5*) and, particularly, entertainment aspects (*e22*) are less important. Finally, respondents in the third cluster do not express high expectations in general, except in terms of security (*e9*). Based on these results, we could label the first cluster “the demanding traveler,” the second cluster “on-time is enough,” and the third cluster “no thrills.” We could further check whether these differences in means are significant by using a one-way ANOVA as described in ► Chap. 6.

In a further step, we can try to profile the clusters using sociodemographic variables. Specifically, we use crosstabs (see ► Chap. 5) to contrast our clustering with the variable *flight_purpose*, which indicates whether the respondents primarily fly for business purposes (*flight_purpose* = 1) or private purposes (*flight_purpose* = 2). Before doing so, we need to turn off the **Split File** command by going to ► Data ► Split File and clicking on **Analyze all cases, do not create groups**, followed by **OK**. Next, click on ► Analyze ► Descriptive Statistics ► Crosstabs. In the dialog box that opens, enter *CLU3_1* into the

■ **Table 9.15** Descriptive statistics

		Descriptive Statistics				
CLU3_1		N	Minimum	Maximum	Mean	Std. Deviation
.	e1	69	2	100	79.09	23.522
	e5	72	1	100	70.65	26.768
	e9	67	43	100	81.99	18.709
	e21	59	1	100	71.47	27.555
	e22	43	2	100	61.35	23.931
	Valid N (listwise)	0				
1	e1	516	69	100	95.13	7.202
	e5	516	25	100	86.98	14.519
	e9	516	28	100	94.38	10.035
	e21	516	50	100	89.89	11.507
	e22	516	50	100	87.61	12.195
	Valid N (listwise)	516				
2	e1	238	53	100	92.58	9.165
	e5	238	5	100	76.65	20.048
	e9	238	19	100	89.77	15.189
	e21	238	1	100	83.37	17.343
	e22	238	1	75	47.16	15.865
	Valid N (listwise)	238				
3	e1	215	1	100	59.42	21.327
	e5	215	1	100	58.28	19.658
	e9	215	1	100	71.63	20.414
	e21	215	1	100	56.73	19.303
	e22	215	2	100	58.03	20.175
	Valid N (listwise)	215				

Row(s) box and *flight_purpose* into the Column(s) box. Also click on **Statistics** and select **Chi-square** and **Contingency coefficient** and click on **Continue** followed by **OK**. The results in ■ **Table 9.16** show that the first cluster primarily consists of leisure travelers, whereas the majority of respondents in the second and third cluster are business travelers. With a p -value of **0.003**, the χ^2 -test statistic indicates a significant relationship between these two variables. However, the strength of the variables' association is rather small, as indicated by the **Contingency Coefficient** of **0.108**.

Table 9.16 Crosstab

CLU3_1 * flight_purpose Crosstabulation

Count		flight_purpose		Total
		1	2	
CLU3_1	1	232	284	516
	2	137	101	238
	3	114	101	215
Total		483	486	969

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	11.463 ^a	2	.003
Likelihood Ratio	11.493	2	.003
Linear-by-Linear Association	6.432	1	.011
N of Valid Cases	969		

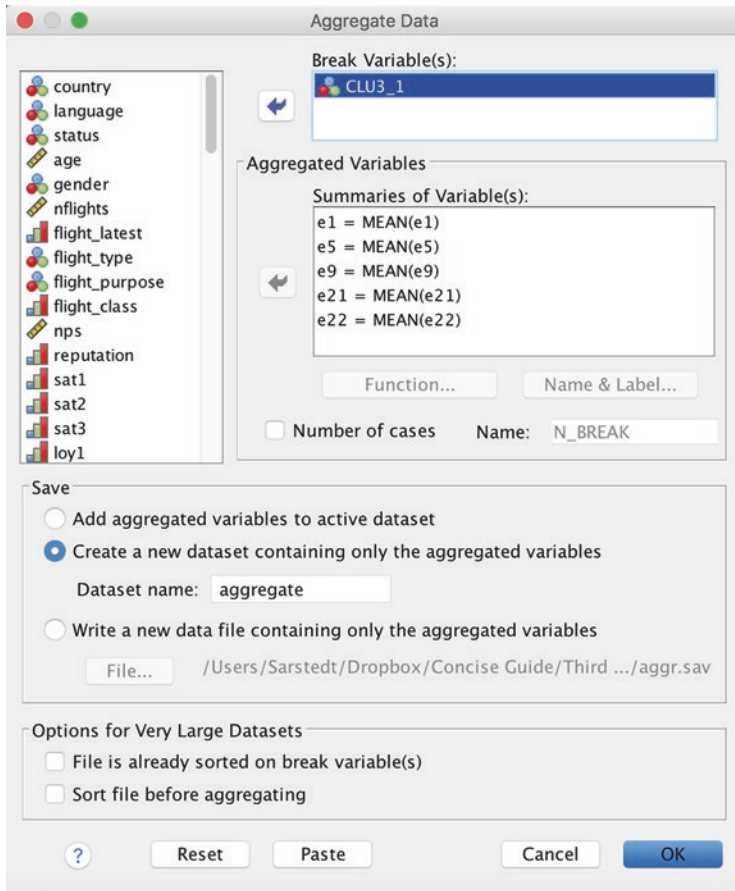
^a 0 cells (0.0%) have expected count less than 5. The minimum expected count is 107.17.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Contingency Coefficient	.108	.003
N of Valid Cases		969	

The Oddjob Airways dataset offers various other variables such as *age*, *gender*, or *status*, which could be used to further profile the cluster solution. However, instead of testing these variables' efficacy step-by-step, we proceed and assess the solution's stability by running an alternative clustering procedure on the data. Specifically, we apply *k*-means clustering, using the cluster centers produced by the Ward's linkage analysis as input for the starting partition, instead of letting *k*-means choose the centers.

To do so, we need to do some data management in SPSS, as the cluster centers have to be supplied in a specific format. Specifically, we need to aggregate the data first (briefly introduced in ► Chap. 5). By going to ► Data ► Aggregate, SPSS opens a dialog box similar to Fig. 9.21. Proceed by entering *CLU3_1* into the **Break Variable(s)** box as well as *e1*, *e5*, *e9*, *e21*, and *e22* into the **Aggregated Variables** box. When using the default settings, SPSS computes the variables' mean values along the lines of the break variable, which correspond to the cluster centers that we need for the *k*-means clustering. SPSS indicates this circumstance by the postfix *_mean*, added to each aggregate variable's name. For *k*-means to process the cluster centers, we need to delete the postfix *_mean* using the **Name**



■ Fig. 9.21 Aggregate data dialog box

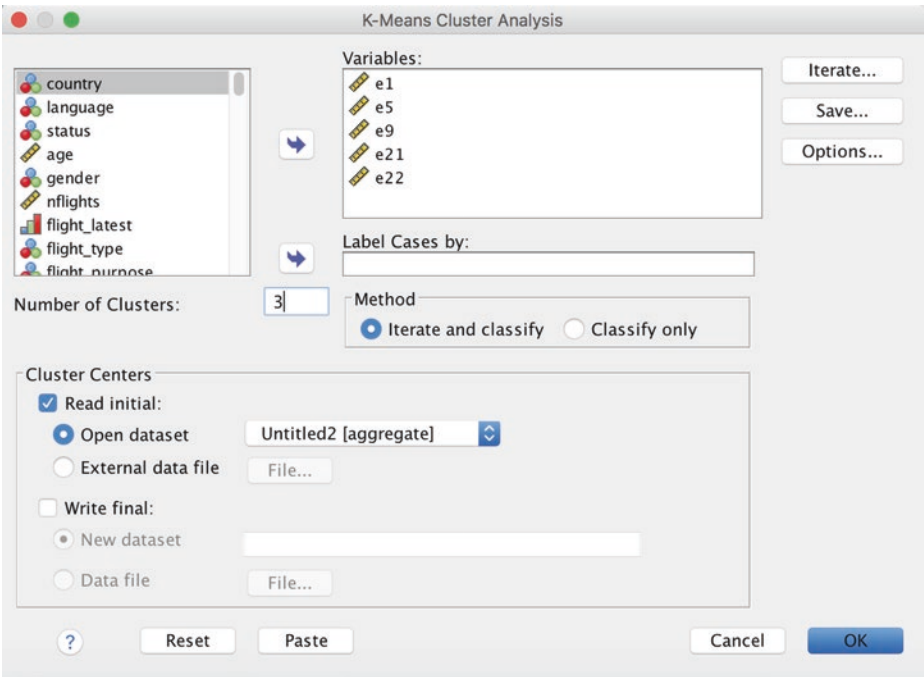
& Label. Finally, we do not want to add the aggregated variables to the active dataset, but rather need to create a new dataset comprising only the aggregated variables. Hence, select **Create a new dataset containing only the aggregated variables** and specify a dataset label such as *aggregate* (■ Fig. 9.21). When clicking on **OK**, SPSS creates and opens a new dataset labeled *aggregate*.

The new dataset is almost in the right format—but we still need to change the break variable's name from *CLU3_1* to *cluster_*. SPSS will issue a warning but this can be safely ignored. Furthermore, we need to delete the first object, which includes the cluster centers of the missing values. The final dataset should have the form shown in ■ Fig. 9.22.

Everything is now set for the *k*-means cluster analysis. To run the analysis, select the original dataset *Oddjob.sav* and go to ► **Analyze** ► **Classify** ► **K-Means Cluster**. In the dialog box that opens (■ Fig. 9.23), first move the five clustering variables into the **Variables** box. To use the cluster centers from our previous analysis, check the box **Read initial** and click on **Open dataset**. You can now choose the dataset labeled *aggregate*. In the **Number of Clusters** box, specify 3, which corresponds to the result of the hierarchical cluster analysis.

	cluster_	e1	e5	e9	e21	e22
1	1	95,13	86,98	94,38	89,89	87,61
2	2	92,58	76,65	89,77	83,37	47,16
3	3	59,42	58,28	71,63	56,73	58,03

■ Fig. 9.22 Aggregated data file



■ Fig. 9.23 *k*-means cluster analysis dialog box

Next, click on **Save** and check the box **Cluster Membership** in order to create a new variable indicating each object’s cluster membership as produced by *k*-means clustering. Under **Options**, you can request several statistics and specify how missing values should be treated. Ensure to request the **Initial cluster centers** as well as the **ANOVA table**. Now start the analysis.

The *k*-means procedure generates ■ Tables 9.17 and 9.18, which show the initial and final cluster centers. As we can see, there is a high degree of agreement between the initial cluster centers produced by the Ward’s linkage and the final cluster centers produced by *k*-means clustering. While some cluster centers changed (also indicated in the **Iteration History** output, not shown here), the clusters’ nature, as expressed by the cluster labels “the demanding traveler,” “on-time is enough,” and “no thrills,” remains intact.

To further check for the solution’s stability, we next explore the overlap in the two cluster solutions, by contrasting the objects’ cluster affiliations using crosstabs. To do so, go to ► **Analyze** ► **Descriptive Statistics** ► **Crosstabs** and select *CLU3_1* under **Row(s)**

■ **Table 9.17** Initial cluster centers

Initial Cluster Centers			
		Cluster	
	1	2	3
e1	95	93	59
e5	87	77	58
e9	94	90	72
e21	90	83	57
e22	88	47	58

Input from FILE Subcommand

■ **Table 9.18** Final cluster centers

Final Cluster Centers				
		Cluster		
		1	2	3
e1		95	90	59
e5		92	74	55
e9		96	90	67
e21		92	81	54
e22		91	55	54

■ **Table 9.19** Comparison of clustering results

CLU3_1 * QCL_1 Crosstabulation					
		QCL_1			Total
		1	2	3	
CLU3_1	1	410	100	6	516
	2	14	213	11	238
	3	10	36	169	215
Total		434	349	186	969

and *QCL_1* under **Column(s)**. The latter variable represents the objects' cluster affiliations as produced by the *k*-means clustering. After clicking on **OK**, SPSS will produce an output similar to ■ [Table 9.19](#).

■ Table 9.20 ANOVA output

	ANOVA					
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
e1	91964.042	2	170.733	966	538.643	.000
e5	94966.114	2	230.012	966	412.875	.000
e9	58156.159	2	164.349	966	353.857	.000
e21	96081.135	2	202.743	966	473.905	.000
e22	158600.747	2	227.709	966	696.508	.000

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

The results show that there is a strong degree of overlap between the two cluster analyses. For example, 410 objects that fall into the first cluster in the Ward's linkage analysis also fall into this cluster in the k -means clustering. At the same time, however, 100 objects now appear in the second k -means cluster. This divergence is considerably lower in the second and third cluster. Overall, the two analyses have an overlap of $(410 + 213 + 169)/969 = 81.73\%$, which is very satisfactory as less than 20 % of all objects appear in a different cluster when using k -means.

In contrast to hierarchical clustering, the k -means outputs provide us with an ANOVA of the cluster centers (■ Table 9.20). Since all the values in the final column **Sig.** are below 0.05, we can conclude that all the clustering variables' means differ significantly across at least two of the three segments.

Since we used the prior analysis results from hierarchical clustering as an input for the k -means procedure, the problem of selecting the correct number of segments is not problematic in this example. Complementing our prior analyses, we now compute the VRC for different numbers of clusters based on the k -means results. Specifically, we want use the VRC values to compute the ω_k statistics for a three-, four-, and five-cluster solution. Since determining a suitable number clusters using the ω_k statistic involves comparing the VRC values of solutions with one segment less than k and with one cluster more than k , we need to run k -means for a two- to six-cluster solution. To do so, go back to ► Analyze ► Classify ► K-Means Cluster. As we seek to run k -means with different numbers of clusters, we cannot use the initial cluster centers from the Ward's linkage clustering. Hence, uncheck the box next to **Read initial**. Next, set the **Number of Clusters** to 2, run the analysis, and save the F -values for variables $e1$, $e5$, $e9$, $e21$, and $e22$ from the ANOVA table, which correspond to the VRC values. Repeat these steps for a three-, four-, five- and six-cluster solution, each time saving the F -values. ■ Table 9.21 summarizes the F -values from the ANOVA tables.

To compute the ω_k statistic, we enter the F -values—which again, correspond to the VRC values—from ■ Table 9.21 into the following formula:

$$\omega_k = (VRC_{k+1} - VRC_k) - (VRC_k - VRC_{k-1}).$$

■ **Table 9.21** F-values for different numbers of clusters

F-values	Number of clusters k				
	2	3	4	5	6
$e1$	448.182	555.639	425.992	391.870	474.269
$e5$	830.757	458.807	306.988	272.402	290.935
$e9$	456.818	373.290	264.142	237.495	186.223
$e21$	734.041	490.399	453.413	479.337	312.860
$e22$	578.393	581.981	707.882	548.408	446.994
Total	3,048.191	2,460.116	2,158.417	1,929.512	1,711.281

For example, for a three-cluster solution, we compute

$$\omega_3 = (2,158.417 - 2,460.116) - (2,460.116 - 3,048.191) = 286.376$$

Similarly, we can compute ω_k for four and five clusters resulting in $\omega_4 = 72.794$ and $\omega_5 = 10.674$, respectively. Comparing the values, we find that the minimum ω_k results for a five-cluster solution. However, looking into the cluster sizes of a five-cluster solution, shows that one cluster contains only 15 objects, which calls the relevance of this cluster into question. Similarly, when using a four-cluster solution, one cluster contains only 60 objects. Hence, it appears more reasonable to retain the three-cluster solution.

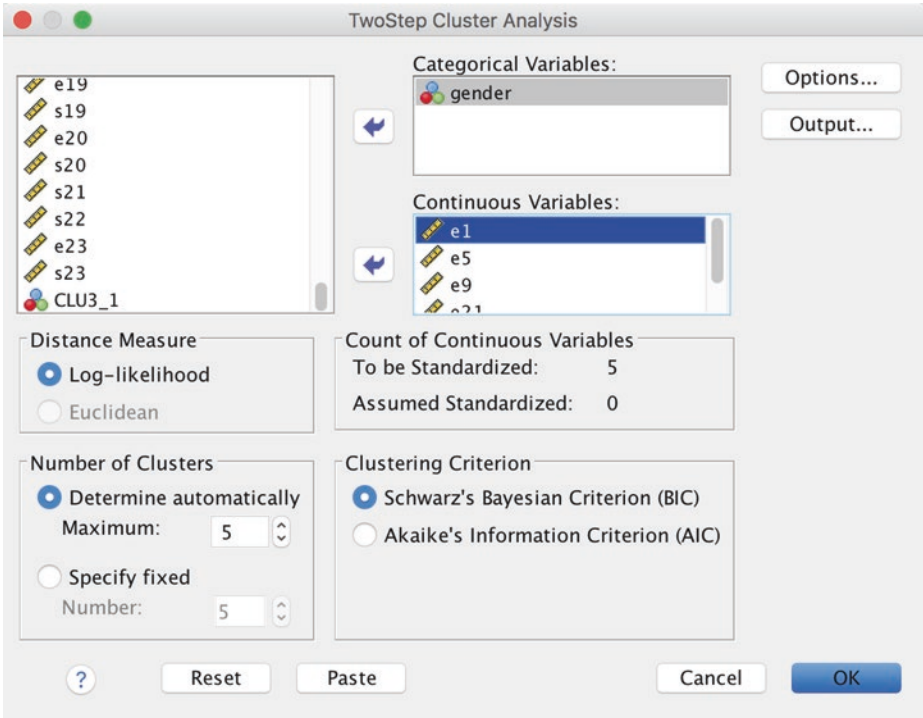
This analysis concludes our cluster analysis. However, we could further explore the solution's stability by running other linkage algorithms, such as centroid or complete linkage, on the data. Relatedly, we could use different (dis)similarity measures and assess their impact on the results. So go ahead and explore these options yourself!

9.4.2 Two-Step Clustering

In the last step of the analysis, we run two-step clustering on the data. As two-step clustering allows handling segmentation variables measured on different scale levels, we extend the prior set and now also consider *gender* as an additional (categorical) segmentation variable. To initiate the analysis, go to ► Analyze ► Classify ► Two-Step Cluster. A new dialog box opens, similar to that shown in ■ Fig. 9.24. First, move *gender* into the **Categorical Variables** box and $e1$, $e5$, $e9$, $e21$, and $e22$ into the **Continuous Variables** box.

Under **Distance Measure** we can choose between two options. While **Log-likelihood** can be used for categorical and continuous variables, the **Euclidean** distance requires variables measured on a continuous scale. Since our analysis contains both categorical and continuous variables, we have to use the **Log-likelihood** distance measure.

Under **Number of Clusters**, we can specify a fixed number or a maximum number of clusters to retain from the data. One of two-step clustering's major advantages is that it allows the automatic selection of the number of clusters on the grounds of information



■ Fig. 9.24 Two-step cluster analysis dialog box

criteria. In line with our previous analyses, we specify a maximum number of 5 clusters. Under **Clustering Criterion**, select **Schwarz's Bayesian Criterion (BIC)** but to test the stability of the solution, we will re-run the analysis using **Akaike's Information Criterion (AIC)**.

Under **Options**, we can select options related to outlier treatment, memory allocation, and variable standardization. Variables that are already standardized have to be assigned as such, but since this is not the case in our analysis, we can simply proceed.

Finally, under **Output**, we can specify additional variables for describing the resulting clusters. Select **Create cluster membership variable** and click on **Continue** followed by **OK**.

SPSS produces a very simple output, as shown in ■ Fig. 9.25. The upper part of the output describes the algorithm applied, the number of variables used (labeled input features) and the final number of clusters retained from the data. In our case, the number of clusters is chosen according to BIC, which indicates a three-segment solution (the same holds when using AIC instead of BIC).

The lower part of the output (■ Fig. 9.25) indicates the quality of the cluster solution. The silhouette measure of cohesion and separation reaches a value of less than 0.50, indicating a fair cluster quality. We proceed with the analysis by double-clicking on the output. This will open up the model viewer (■ Fig. 9.26), an evaluation tool that graphically presents the structure of the revealed clusters.

The model viewer provides us with two windows: The main view, which initially shows a model summary (left-hand side), and an auxiliary view, which initially features

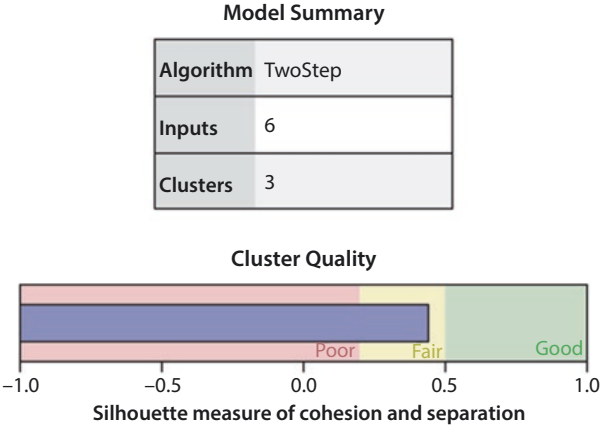


Fig. 9.25 Two-step cluster analysis output

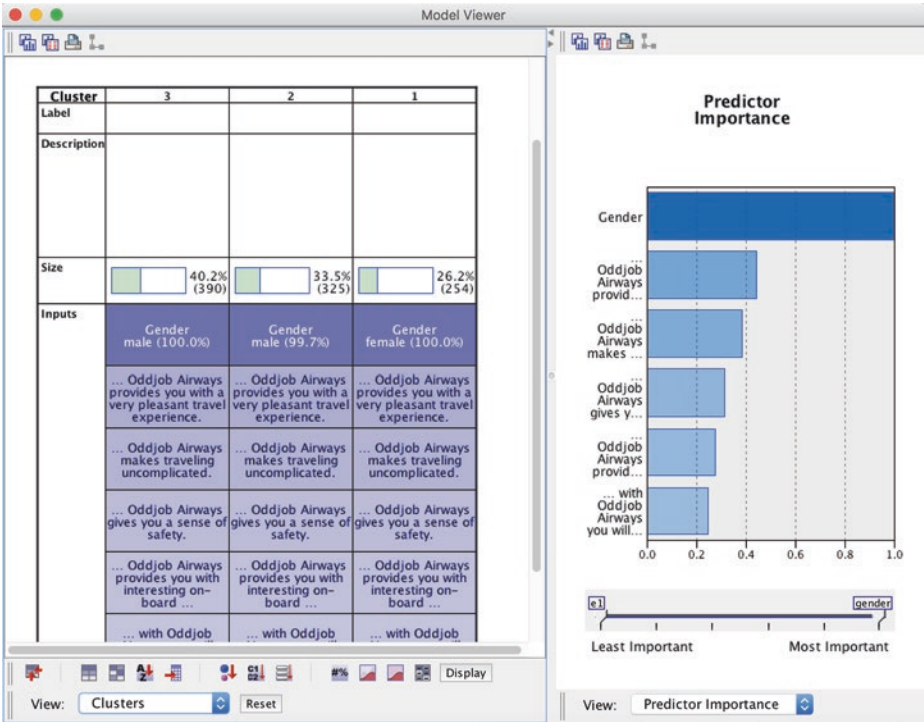



Fig. 9.26 Additional options in the model viewer

the cluster sizes (right-hand side). At the bottom of each window (next to **View**), you can request different information on each of the clusters. To further analyze the clusters, select **Clusters** in the main view and **Predictor Importance** in the auxiliary view (Fig. 9.26).

On the left of  Fig 9.26, we can now see a description of the three clusters, including their (relative) sizes. We find that the first cluster contains 40.2 % of the objects, the second cluster 33.5 % of the objects, and the third cluster contains 26.2 % of the objects. Further below, the output shows the distribution of the *gender* variable in each cluster. Moving the mouse over the boxes showing the clustering variable labels, SPSS shows their mean values as well as their relative importance in terms of predicting each object's membership per cluster. Darker shades (i.e., higher values in feature importance) denote the variable's greater importance for the clustering solution. Comparing the results, we learn that *gender* is by far the most important variable for each of the clusters, followed by *e5* ("Oddjob Airways provides you with a very pleasant travel experience"), *e21* ("Oddjob Airways makes traveling uncomplicated"), *e9* ("Oddjob Airways gives you a sense of safety"), *e22* ("Oddjob Airways provides you with interesting on-board entertainment, service, and information sources"), and *e1* ("with Oddjob Airways you will arrive on time").⁶ Clicking on one of the boxes will show a graph with the frequency distribution of each cluster.

The auxiliary view on the right-hand side shows an overview of the variables' overall importance for predicting the clustering solution (i.e., across all clusters). The model viewer provides us with additional options for visualizing the results or comparing clustering solutions. It is worthwhile to simply play around with the different self-explanatory options. So go ahead and explore the model viewer's features yourself!

9

9.5 Oh, James! (Case Study)

Case Study

The James Bond movie series is one of the success stories of filmmaking. The movies are the longest continually running and the third-highest-grossing film series to date, which started in 1962 with Dr. No, starring Sean Connery as James Bond. As of 2018, there have been 24 movies with six actors having played James Bond. Interested in the factors that contributed to this running success, you decide to investigate the different James Bond movies' characteristics. Specifically, you want to find out whether the movies can be grouped into clusters, which differ in their box-office revenues. To do so, you draw on Internet Movie Database (www.imdb.com) and collect data on all 24 movies based on the following variables (variable names in parentheses):

- Title. (*title*)
- Actor playing James Bond. (*actor*)
- Year of publication. (*year*)
- Budget in USD, adjusted for inflation. (*budget*)
- Box-office revenues in the USA, adjusted for inflation. (*gross_usa*)
- Box-office revenues worldwide, adjusted for inflation. (*gross_worldwide*)
- Runtime in minutes. (*runtime*)
- Native country of the villain actor. (*villain_country*)
- Native country of the bondgirl. (*bondgirl_country*)

⁶ The strong emphasis of gender in determining the solution supports prior research, which found that two-step clustering puts greater emphasis on categorical variables in the results computation (Bacher et al. 2004).

— Haircolor of the bondgirl. (*bondgirl_hair*)

Use the dataset *James Bond.sav* (↓ Web Appendix → Downloads) to run a cluster analysis—despite potential objections regarding the sample size. Answer the following questions:

1. Which clustering variables would you choose in light of the study objective, their levels of measurement, and correlations?
2. Given the levels of measurement, which clustering method would you prefer? Carry out a cluster analysis using this procedure.
3. Interpret and profile the obtained clusters by examining cluster centroids. Compare the differences across clusters on the box-office revenue variables.
4. Use a different clustering method to test the stability of your results.

9.6 Review Questions

1. In your own words, explain the objective and basic concept of cluster analysis.
2. What are the differences between hierarchical and partitioning methods? When do we use hierarchical or partitioning methods?
3. Repeat the manual calculations of the hierarchical clustering procedure from the beginning of the chapter, but use complete linkage as the clustering method. Compare the results with those of the single linkage method.
4. Explain the different options to decide on the number of clusters to extract from the data. Should you rather on statistical measures or rather on practical reasoning?
5. Run the two-step clustering analysis on the Oddjob Airways data again (*Oddjob.sav*, ↓ Web Appendix → Downloads) but with a prespecified number of four and five clusters. Compare your results with the original three-cluster solution.
6. Which clustering variables could be used to segment:
 - The market for smartphones?
 - The market for chocolate?
 - The market for car insurances?

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Selected papers of Hirotugu Akaike* (pp. 199–213). New York: Springer.
- Arabie, P., & Hubert, L. (1994). Cluster analysis in marketing research. In R. P. Bagozzi (Ed.), *Advanced methods in marketing research* (pp. 160–189). Cambridge: Basil Blackwell & Mott, Ltd.
- Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. *Proceedings of the 18th annual ACM-SIAM symposium on discrete algorithms*. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, pp. 1027–1035.
- Bacher, J., Wenzig, K., & Vogler, M. (2004). SPSS TwoStep Cluster – A first evaluation. Arbeits- und Diskussionspapiere/Universität Erlangen-Nürnberg, Sozialwissenschaftliches Institut, Lehrstuhl für Soziologie, 2004-2. <http://www.ssoar.info/ssoar/handle/document/32715>.
- Becker, J.-M., Ringle, C. M., Sarstedt, M., & Völckner, F. (2015). How collinearity affects mixture regression results. *Marketing Letters*, 26(4), 643–659.
- Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics—Theory and Methods*, 3(1), 1–27.

- Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. *Proceedings of the 7th ACM SIGKDD international conference in knowledge discovery and data mining*. Association for Computing Machinery, San Francisco, CA, USA, pp. 263–268.
- Dolnicar, S. (2003). Using cluster analysis for market segmentation—typical misconceptions, established methodological weaknesses and some recommendations for improvement. *Australasian Journal of Market Research*, 11(2), 5–12.
- Dolnicar, S., & Grün, B. (2009). Challenging “factor-cluster segmentation”. *Journal of Travel Research*, 47(1), 63–71.
- Dolnicar, S., & Lazarevski, K. (2009). Methodological reasons for the theory/practice divide in market segmentation. *Journal of Marketing Management*, 25(3–4), 357–373.
- Dolnicar, S., Grün, B., Leisch, F., & Schmidt, F. (2014). Required sample sizes for data-driven market segmentation analyses in tourism. *Journal of Travel Research*, 53(3), 296–306.
- Dolnicar, S., Grün, B., & Leisch, F. (2016). Increasing sample size compensates for data problems in segmentation studies. *Journal of Business Research*, 69(2), 992–999.
- Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data. An introduction to cluster analysis*. Hoboken, NY: Wiley.
- Kotler, P., & Keller, K. L. (2015). *Marketing management* (15th ed.). Upper Saddle River, NJ: Prentice Hall.
- Lilien, G. L., & Rangaswamy, A. (2004). *Marketing engineering. Computer-assisted marketing analysis and planning* (2nd ed.). Bloomington: Trafford Publishing.
- Milligan, G. W., & Cooper, M. (1988). A study of variable standardization. *Journal of Classification*, 5(2), 181–204.
- Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36(2), 3336–3341.
- Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, 20(2), 134–148.
- Qiu, W., & Joe, H. (2009). clusterGeneration: Random cluster generation (with specified degree of separation). R package version 1.2.7. <https://cran.r-project.org/web/packages/clusterGeneration/clusterGeneration.pdf>. Accessed 04 May 2018.
- Roberts, J. H., Kayande, U. K., & Stemersch, S. (2014). From academic research to marketing practice: Exploring the marketing science value chain. *International Journal of Research in Marketing*, 31(2), 127–140.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Sheppard, A. (1996). The sequence of factor analysis and cluster analysis: Differences in segmentation and dimensionality through the use of raw and factor scores. *Tourism Analysis*, 1, 49–57.
- Tonks, D. G. (2009). Validity and the design of market segments. *Journal of Marketing Management*, 25(3/4), 341–356.
- Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Boston, NJ: Kluwer Academic.
- Van Der Kloot, W. A., Spaans, A. M. J., & Heinser, W. J. (2005). Instability of hierarchical cluster analysis due to input order of the data: The PermuCLUSTER solution. *Psychological Methods*, 10(4), 468–476.

Further Reading

- Bottomley, P., & Nairn, A. (2004). Blinded by science: The managerial consequences of inadequately validated cluster analysis solutions. *International Journal of Market Research*, 46(2), 171–187.
- Dolnicar, S., Grün, B., & Leisch, F. (2016). Increasing sample size compensates for data problems in segmentation studies. *Journal of Business Research*, 69(2), 992–999.
- Dolnicar, S., & Leisch, F. (2017). Using segment level stability to select target segments in data-driven market segmentation studies. *Marketing Letters*, 28(3), 423–436.
- Ernst, D., & Dolnicar, S. (2017). How to avoid random market segmentation solutions. *Journal of Travel Research*, 57(1), 69–82.
- Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, 20(2), 134–148.
- Romesburg, C. (2004). *Cluster analysis for researchers*. Morrisville: Lulu Press.
- Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Boston: Kluwer Academic.