EDITED BY
RAHUL PRATAP SINGH KAURAV,
DOGAN GURSOY AND NIMIT CHOWDHARY

AN SPSS GUIDE FOR TOURISM, HOSPITALITY AND EVENTS RESEARCHERS



eRESOURCES

ROUTLEDGE

Chapter 14

Cluster analysis

Ana Brochado

Market segmentation of hostel guests using cluster analysis

Hostelling International is a hostel federation formed in 1932, with currently over 4,000 members around the world. This organisation defines a hostel as good-quality but budget-friendly accommodations that provide a restful night's sleep and a welcoming atmosphere, all at a reasonable price. Hostels also provide opportunities for guests to mingle with many travellers with similar needs in communal facilities such as bars or lounges and even kitchens. Hostels thus are a hybrid product, namely, accommodation services with an informal, welcoming atmosphere (Brochado & Rita, 2018). The hostels' financial business model relies on the multiple-bed factor, which generates a higher profit per square foot (Brochado, Gameiro & Rita, 2015).

Hostel businesses have recorded rapid growth in recent decades and hostel guests are viewed as an important tourist segment in terms of target marketing (Musa & Thirumoorthi, 2011). Although most hostel guests are young and budget-conscious, hostel clients also look for opportunities to meet new, open-minded friends and to experience host cities. The literature has reported new trends in this niche market. Some hostels have upgraded their offer by providing safe storage facilities and private rooms with a bathroom and upscale design, among other tangible improvements. Hostel managers are also facing new challenges due to increasing levels of competition and a stronger focus on business success as of crucial importance. Hostel guests who are pleased with the service quality provided will recommend hostels and these customers are more likely to return (Brochado & Rita, 2018).

For hostels, service quality is a construct with many dimensions. According to Brochado and Gameiro (2013), it includes the following characteristics: atmosphere, cleanliness, the existence of bar service, facilities, Internet facilities, location, opportunities to meet other travellers, price,

staff quality and security. Previous studies (e.g. Brochado & Rita, 2018) have found that hostel managers should consider specific demographic segments' perspectives and tailor their hostel's offer to meet the needs of different age, gender and nationality groups to strengthen perceived overall service quality.

Objectives and research propositions

This research's main objective was to use cluster analysis to conduct market segmentation (Dolnicar, 2002) of hostel guests. The present case study included two research propositions:

Proposition 1: Hostel guests can be segmented according to their perceptions of the importance of different service quality attributes.

Proposition 2: Hostel guest segments based on the importance of service quality attributes vary according to demographic characteristics: age, gender and nationality.

Methodology

Data collection

This study's data were collected by administrating a survey to hostel guests. The questionnaire had three sections. The first part comprised introductory questions about the respondents' stay in Lisbon. The second section asked guests to rate how important 10 attributes are to overall hostel experiences, using a scale ranging from 1 ('not important') to 7 ('extremely important'). The survey's last section elicited demographic information.

Market segmentation basis

The variables used were both general and observed variables (i.e. demographics) and inferred variables specific to the hostel product (i.e. service quality attributes) (see Table 14.1).

Market segmentation method

Cluster analysis was conducted to address the first proposition, that is, to determine whether different market segments exist based on the importance given different service quality items. The analysis comprised two steps. In the first, hierarchical clustering and a visual inspection of a dendrogram of the variables identified the number of segments to keep for further analysis. The second step involved using non-hierarchical clustering via the k-means method to isolate a final set of guest

Table 14.1 Variables

Variable	*Coding (category)*
Atmosphere	1 ('not important') to 7 ('extremely important')
Cleanliness	
Existence of bar service	
Facilities	
Internet facilities	
Location	
Opportunities to meet other travellers	
Price	
Staff quality	
Security	
Gender	1 (male), 2 (female)
Age group	1 (15–25), 2 (26–29), 3 (30 ≤)
Nationality	1 (United States/Canada), 2 (Latin America), 3 (Europe), 4 (Australia/New Zealand), 5 (Asia)

clusters. The analysis performed was thus post hoc as the tourist segments' number and size were determined after the data were processed. The second proposition was then addressed by examining the obtained segments to determine whether they exhibit different demographic profiles. This part of the research was performed using crosstabulation and Chi-square tests.

Cluster analysis decision process

The cluster analysis process included six steps. These included (i) preparing the data (i.e. detecting outliers and standardising the data), (ii) making assumptions (i.e. multicollinearity and other assumptions), (iii) selecting clustering algorithms, (iv) determining the number of clusters, (v) interpreting clusters and (vi) profiling clusters.

Data treatment

The above process comprised a two-step approach that combined Ward's hierarchical method with the k-means non-hierarchical method. This part of the analysis was based on the Euclidean squared distance metric. In the first step, a visual inspection of the variables' dendrogram and the agglomeration coefficient's evolution determined what would be an adequate number of clusters. In the second step, k-means clustering was performed to identify the best allocation of individual elements within each cluster. Kruskal-Wallis tests were then conducted to evaluate differences between groups in terms of the 10 attributes used to generate the clusters. To profile the clusters, this study relied on crosstabulation and the Chi-square test of independence.

Cluster analysis using SPSS software

Cluster analysis is a set of techniques used to partition a set of cases under study (e.g. guests, hotels and restaurants) into two or more homogeneous groups called 'clusters'. This analysis facilitates taxonomy description, data simplification and relationship identification. The present chapter only covers clustering procedures that assign each object to only one cluster (i.e. non-overlapping clustering procedures). The methods discussed are available on SPSS software's menu as follows: **Analyze > Classify > K-Means** or **Hierarchical Clustering**.

Data preparation

The most important part of cluster analysis is to select the variables used to generate clusters. These variables can be chosen based on previous studies, theory or considerations connected to the hypotheses or propositions to be tested. Alternatively, variables can be selected to reflect that the research is exploratory. In the present case study, the analysis used 10 variables that are attributes guests use to assess their hostel experiences. All the variables were defined based on the literature and/or previous studies. The 10 variables were thus selected because they are the most important to hostel guests.

The guests were asked to evaluate how significant the following attributes were to them on a 7-point scale (1 = 'not important'; 7 = 'extremely important'): atmosphere, cleanliness, existence of bar service, facilities, Internet facilities, location, opportunities to meet other travellers, price, staff quality and security. The data were gathered from a sample of 285 Lisbon hostel guests and input into an Excel file. This sample's data are used below to illustrate each step of the clustering procedure.

Cluster analysis's first step is to prepare the data by detecting outliers and standardising the data. Cluster analysis is sensitive to outliers, so an initial screening for outliers is of utmost importance. Outliers exhibit a different profile from all other data, registering extreme values for one or more variables. Outliers can be associated with measurement errors or cases that are not representative of the study population or underrepresented in the sample.

One graphical approach used to identify outliers is to analyse a box plot of all variables used to generate the taxonomy in question. Box plots show how the data for each variable are distributed. The dark line in the middle of each box is each variable's median. In the case under analysis, half of the hostel guests gave a value higher than the median and half gave a lower value to each variable. The top of the box represents the seventy-fifth percentile, and the bottom indicates the 25th percentile. The T-bars that extend from the boxes are the 'inner fences' or 'whiskers'. Points are outliers when their values do not fall inside the inner fences. The moderate outliers are called 'bubbles' and extreme outliers are termed 'asterisks' or 'stars'. These results help researchers decide whether to eliminate extreme outliers from their analysis.

To obtain a box plot for all the variables, these steps are followed:

1 From the menus choose: **Graphs > Box Plot**.
2 Select **Simple > Data in Chart > Summaries of Separate Variables**.

3 Click **Define**.
4 Select the variables to be represented by the boxes: atmosphere, cleanliness, existence of bar service, facilities, Internet facilities, location, opportunities to meet other travellers, price, staff quality and security (ten attributes).
5 Click **OK**.

The present analysis of the resulting box plot revealed that the median value of the variables' importance is high (i.e. on a 7-point scale) and that only moderate outliers are present, as denoted by the bubbles in the box plot (see Figure 14.1). Therefore, no data needed to be removed from subsequent analyses at this stage.

The boxplot for each variable of the descriptive statistics (i.e. average, median, minimum, maximum and standard deviation [SD]) are obtained as follows:

1 From the menu, choose **Analyze > Descriptive Statistics > Explore**.
2 Select the 10 attributes variables to be included in the dependent list.
3 Click **Statistics** and select **Descriptive and Outliers**.
4 Click **OK**.

Any differences in the metrics used to measure the variables used to generate clusters introduce bias into the analysis as larger variables will unduly influence the results. The present variables were measured using different scales, so their effects needed to be equalised by standardising the measurement instruments before beginning analyses. One of the most popular approaches to standardising variables is to calculate z-scores. The raw data are converted into standardised values with a

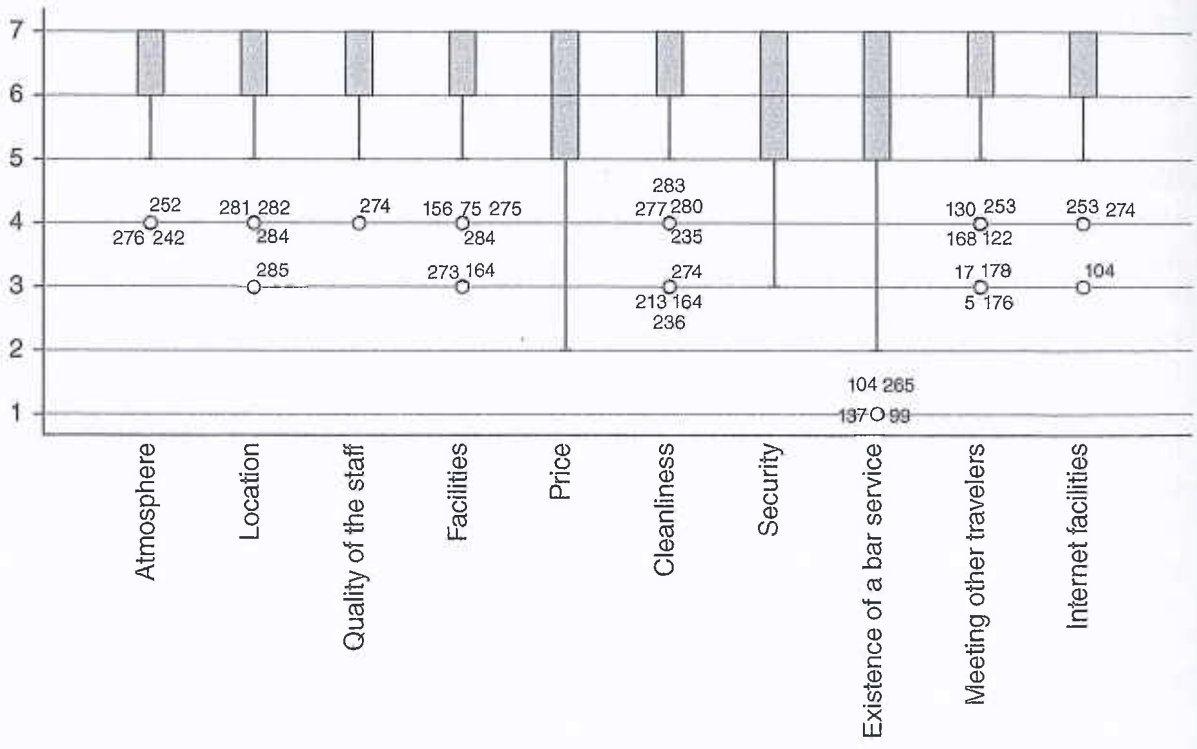


Figure 14.1 Boxplot for ten attributes under analysis

mean of 0 and SD of 1. These values are then input into the k-means cluster analysis, which is done by following the procedure described below.

To calculate new standardised values for the variables:

1 From the menu, select **Analyze > Descriptive Statistics > Descriptives.**
2 Select the variables under analysis: 10 attributes.
3 Select **Save Standardized Values as Variables.**
4 Click **OK.**

The 10 new standardised variables are created and available as the dataset appearing on the right of the screen.

Hierarchical methods further offer the possibility of standardising data for either cases or values before computing the proximity between objects, although this information is not available for binary data. To transform the variables to be introduced for hierarchical clustering, the following steps are followed:

1 From the menu, select **Analyze > Classify > Hierarchical Clustering.**
2 Select the variables under analysis: 10 attributes.
3 Click **Method** and select **Transform Variables > Standardize** (the available standardisation methods: z-scores, range –1 to 1, range 0 to 1, maximum magnitude of 1, mean of 1 and SD of 1).
4 Click **OK.**

In this case, as the ten variables under analysis were all measured on a scale from 1 ('not important') to 7 ('extremely important'), no standardisation was necessary before analyses were conducted.

Cluster analysis assumptions

The second step involves verifying assumptions, namely, the relevant variables, multicollinearity, data type and an adequate measure of similarity. Cluster analysis is an exploratory technique that facilitates data simplification and the identification of the study population's structural characteristics. Although assumptions made regarding other statistical techniques are not important in this context (i.e. normality and homoscedasticity), three critical issues should be considered in cluster analysis.

The first is the requirement that all influential variables are included in analyses. The omission of important variables can produce misleading results. The second critical issue is multicollinearity, defined here as how much each variable is explained by other variables included in the research model. Cluster analysis translates multicollinearity into greater weight and importance within clusters. Multicollinearity is evaluated using correlation analysis, based on the following steps:

1 From the menu, select **Analyze > Correlate > Bivariate.**
2 Select the variables: 10 attributes.
3 Select **Pearson** or **Spearman.**
4 Click **OK.**

In the present case, some correlations between pairs of variables are significant, but they are not strong (i.e. less than 0.50). Thus, multicollinearity was not considered a problem. This analysis was conducted with all 10 attributes.

If the multicollinearity is considered substantial, the Mahalanobis distance can be used to make adjustments. This distance is a standardised form of the Euclidean distance, which is utilised to make adjustments to eliminate intercorrelations between variables. When variables show strong intercorrelation, principal component analysis can also be performed and the factor scores obtained can be used to generate clusters.

The last issue is that an appropriate distance measure needs to be selected. As cluster analysis focuses on grouping similar items together, each object has to be evaluated for differences or similarities to other items. Most frequently, researchers quantify similarity as the distance detected between pairs of objects. Pairs with short distances between them are more alike. Currently, the most commonly used measure is Euclidean distance, namely, the square root of the sum of the squared differences between variable's values.

Researchers use SPSS software to find the necessary distance or similarity values when clustering items, selecting measures according to the type of data involved, which can be interval, count or binary. The alternatives for the first kind of data are cosine, block interval, Pearson's correlation, Chebychev's inequality, Minkowski space, customised interval and Euclidean distance and squared distance. The most popular measures at the moment are the last two.

The alternatives currently used to process count data are Chi-square and phi-square measures. Many alternatives have also been developed to deal with binary data including shape, dice, variance, dispersion, lambda, Hamann, simple matching, Anderberg's D, Jaccard index, size and pattern difference, phi 4-point correlation and Euclidean distance and squared distance. Other measures mentioned in the literature are Ochiai, Russel and Rao, Lance and Williams, Kulczynski 1 and 2, Rogers and Tanimoto, Yule's Y and Q and Sokal and Sneath 1, 2, 3, 4 and 5. Notably, hierarchical methods rely on binary, count and interval data, while non-hierarchical methods require, at the minimum, data measured on interval scales.

Cluster algorithms

The third step involves generating clusters and encompasses decisions such as selecting clustering algorithms. These algorithms can be defined as the process followed to place similar cases in clusters. Clustering algorithms are used to maximise differences between clusters and minimise variation within them. The algorithms can be classified as either hierarchical or non-hierarchical methods. Each method employs a different algorithm to create clusters.

Hierarchical clustering methods comprise successive agglomerations or divisions of cases into groups. Given n cases, the algorithm applied will produce $n - 1$ clustering solutions. These methods facilitate the computation of a range of possible cluster solutions, as well as saving each cluster's membership. Hierarchical clustering procedures' results are displayed in a graphical representation called a dendrogram. Dendrograms reveal how objects and clusters are combined at each step of an algorithm's application, ranging from solutions with n separated objects to

solutions with a single cluster. Hierarchical methods are most useful when no more than a few hundred objects are under analysis.

The clustering algorithms available in SPSS that partition databases differ in the rules applied to obtain clusters, namely, how the distance between two clusters is calculated:

- Complete linkage – the maximum distance and/or furthest neighbour clustering.
- Single linkage – the minimum distance and/or nearest neighbour clustering.
- Average linkage – the average of distances between all pairs of cases, in which each member of the pair is from a different cluster.
- Centroid method – the distance between centroids calculated as each variable's average value, for all items in each cluster.
- Ward's method – minimisation of Euclidean squared distances to clusters' means.

The methods of complete, single and average linkage are known as 'linkage methods' as they cluster objects based on the distance between them. The centroid and Ward's methods are called 'variance methods' because they seek to generated clusters that minimise within-cluster variance.

Hierarchical cluster analysis is recommended for smaller databases (i.e. several hundreds of objects to be clustered). This approach has some advantages compared to non-hierarchical methods, such as the ability to cluster both cases and variables and to allow analyses of different types of data (i.e. scale or binary) if all the variables use the same type of measurement instrument.

Non-hierarchical methods require researchers to specify the number of clusters in advance. The algorithms applied produce a single cluster solution based on the number of clusters specified. These procedures start with the calculation of the initial cluster seeds – also known as cluster centroids – which are each variable's mean value (i.e. centroid) in the first solution generated. Next, the algorithms assign cases to clusters based on their distance from the centroids (i.e. step 1) and update the centroids based on the mean values of each case (i.e. step 2). These two steps are repeated until further reassignment of cases does not improve the solution.

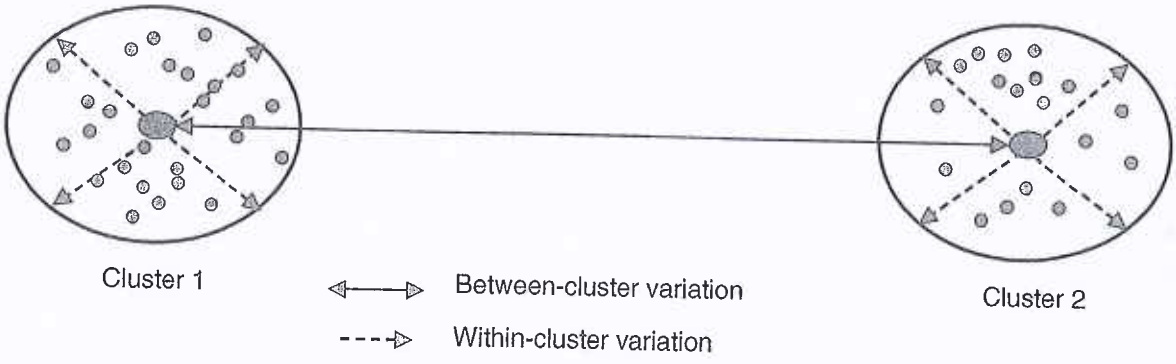
The non-hierarchical method most widely applied is k-means. This method is limited to quantitative analysis at the interval or ratio level, but k-means can be used to analyse large databases. This approach to cluster analysis assigns cases to a specific range of clusters without previously defined characteristics. The clusters are instead based on a set of specific variables.

Hierarchical and non-hierarchical methods can also be combined. Hierarchical cluster analysis is first conducted to identify the number of clusters. Next, k-means is applied to find a final solution using the number of clusters obtained from the first step. Figure 14.2 illustrates all the possible methods.

Number of clusters

The number of clusters cannot be defined a priori as it needs to be defined based on the data gathered. The number of clusters can be selected (i.e. the stopping rule) by combining various approaches as follows:

- A priori criteria.
- Theoretical or conceptual foundations.



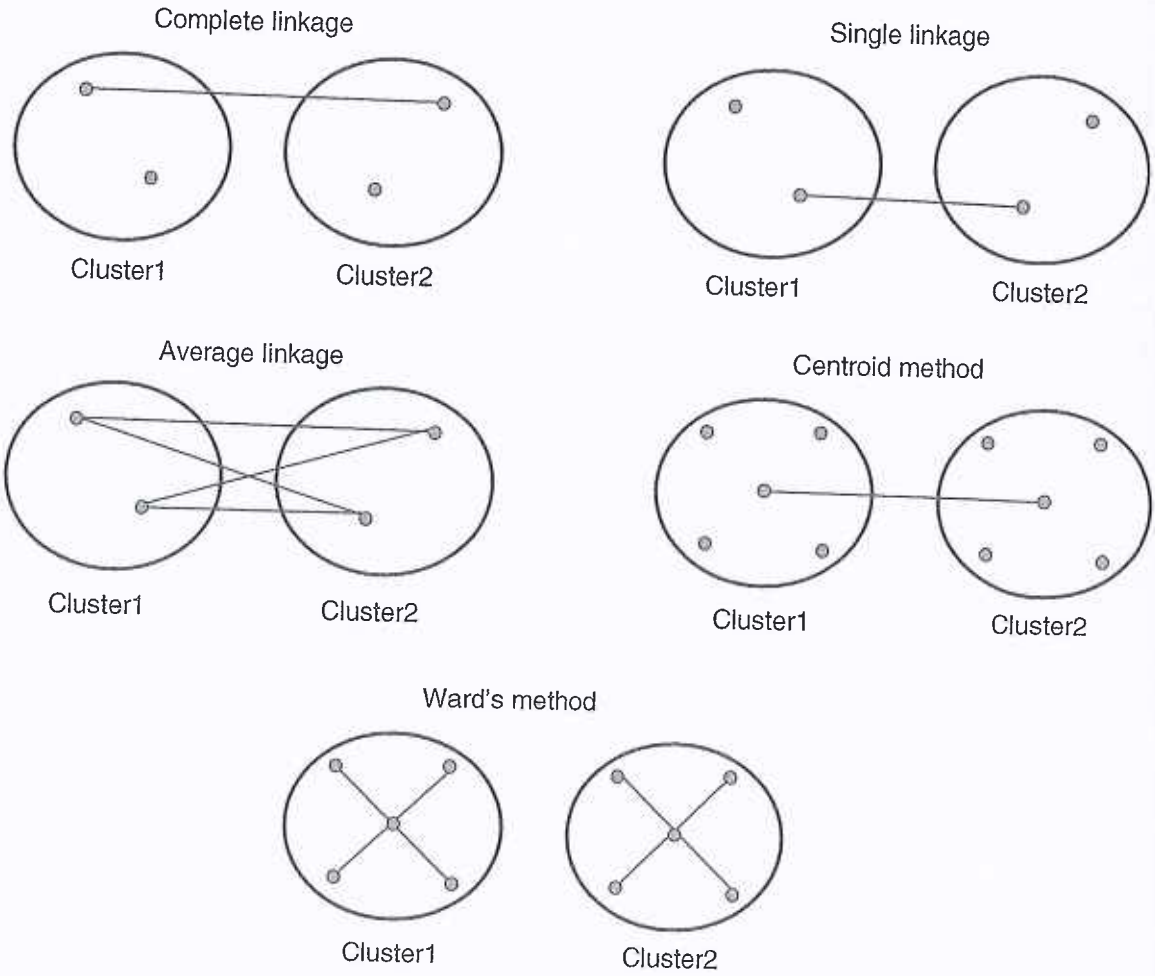


Figure 14.2 Clustering methods

- Managerial or practical judgment.
- Empirical judgment.

Groups' relative size can also help researchers select the number of clusters. Clusters containing a relatively small group of objects can be eliminated because of their insignificance. When non-hierarchical clustering is applied, researchers can use a

graphical representation of the ratio of total within-group to between-group variation to determine the number of clusters. The most suitable numbers are denoted by elbows in graphs.

The present study applied hierarchical methods to obtain the number of clusters, as well as empirical judgment based on a visual inspection of the dendrogram and variations in distances between clusters at each successive step of the algorithm's application. To obtain a hierarchical cluster solution, the steps below are followed (see Figure 14.3):

1 From the menu, choose **Analyze > Classify > Hierarchical Clustering.**
2 Select the variables to be used in the cluster analysis.
3 Click **Methods > Cluster Methods** and select **Ward's Method** (available alternatives are between- and within-groups linkage, nearest and furthest neighbour, centroid and median clustering and Ward's method).
4 From the menu, select **Measure > Interval > Euclidean Squared Distance.**

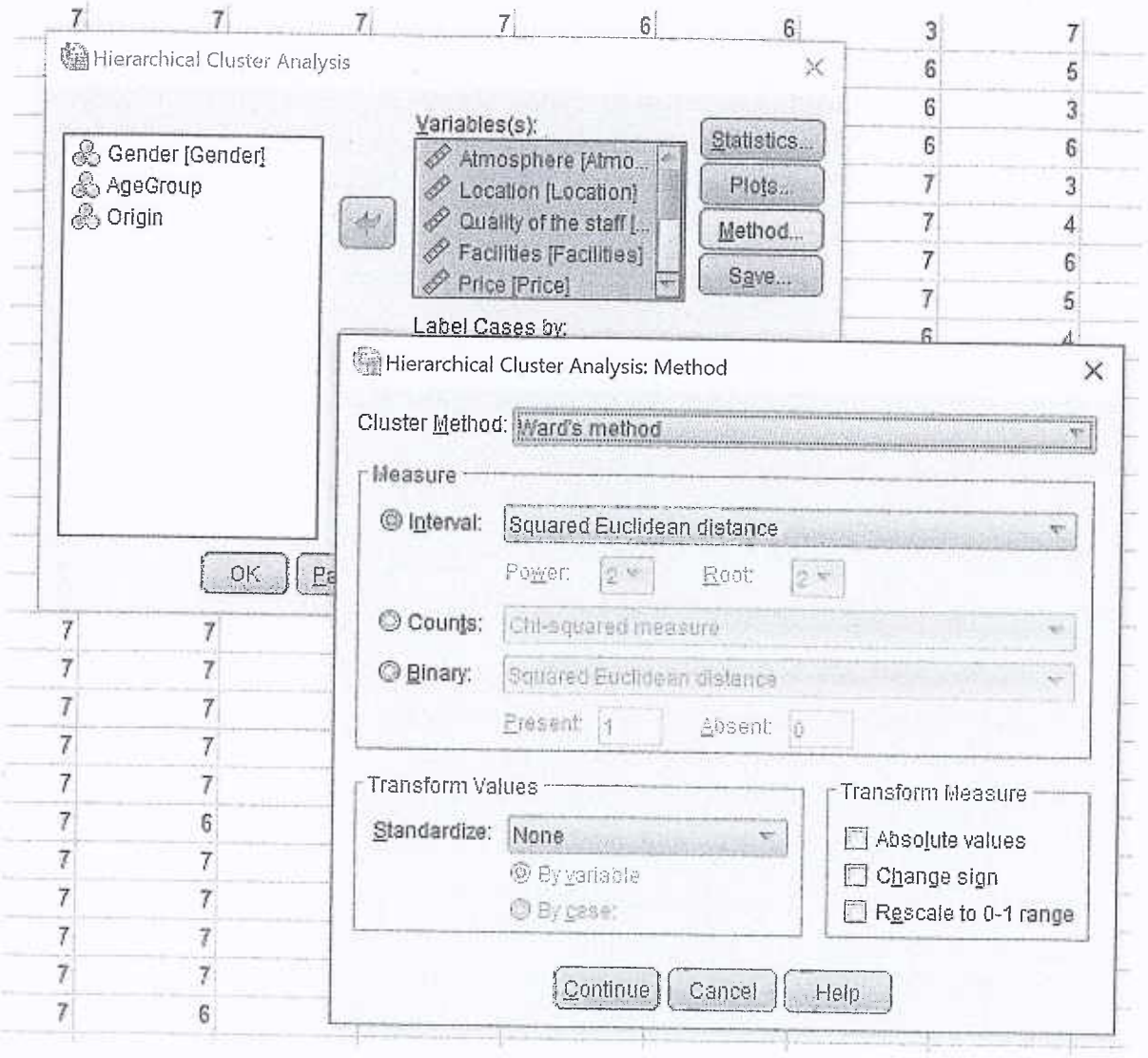


Figure 14.3 Screenshot of hierarchical cluster solution

5 Click **Statistics** and select **Agglomeration Schedule**.
6 Click **Plots** and select **Dendrogram**.
7 Click **OK**.

The agglomeration solution and dendrogram provide similar information. The former is a numerical summary of the cluster solution, while the second is a graphical representation. The present dendrogram offered a visualisation of the cluster solution, in which the hostel guests or cases are listed along the horizontal axis and the distance between the clusters are placed along the vertical axis. Researchers can use dendrograms to evaluate these clusters' cohesiveness in order to determine how many clusters should be kept. Both the visual inspection of the current dendrogram and the growth rate coefficient suggested that a four-cluster solution should be retained for further analysis.

To minimise the subjectivity of this process, the rate of growth in the agglomeration distance between two steps of the algorithm applied can also be calculated. A sudden jump in this coefficient indicates that two different cases or clusters have merged, so the solution before the jump should be worthwhile considering. The present research's agglomeration schedule indicated that, in the first stage, cases 249 and 250 were combined because they exhibited the smallest distance. The new cluster created appeared again in the second stage. In this stage, the cluster created in the first stage was joined with case 246. The resulting cluster subsequently appeared in the third stage (see Figure 14.4).

The dendrogram also facilitated the identification of outliers, such as any 'branch' that did not appear until quite late in the analysis. To obtain a k-means cluster analysis, these steps are followed (see Figure 14.5):

1 From the menu, choose **Analyze > Classify > K-Means Cluster**.
2 Select the variables to be used in the cluster analysis.

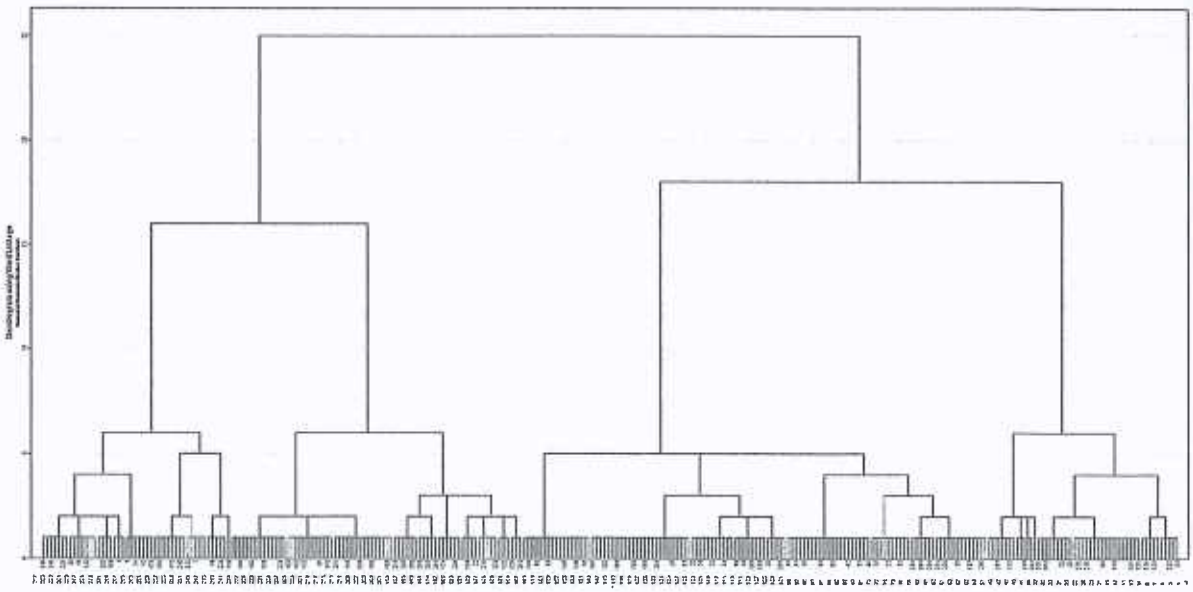


Figure 14.4 Number of clusters – dendrogram and agglomeration schedule for Ward's method solution

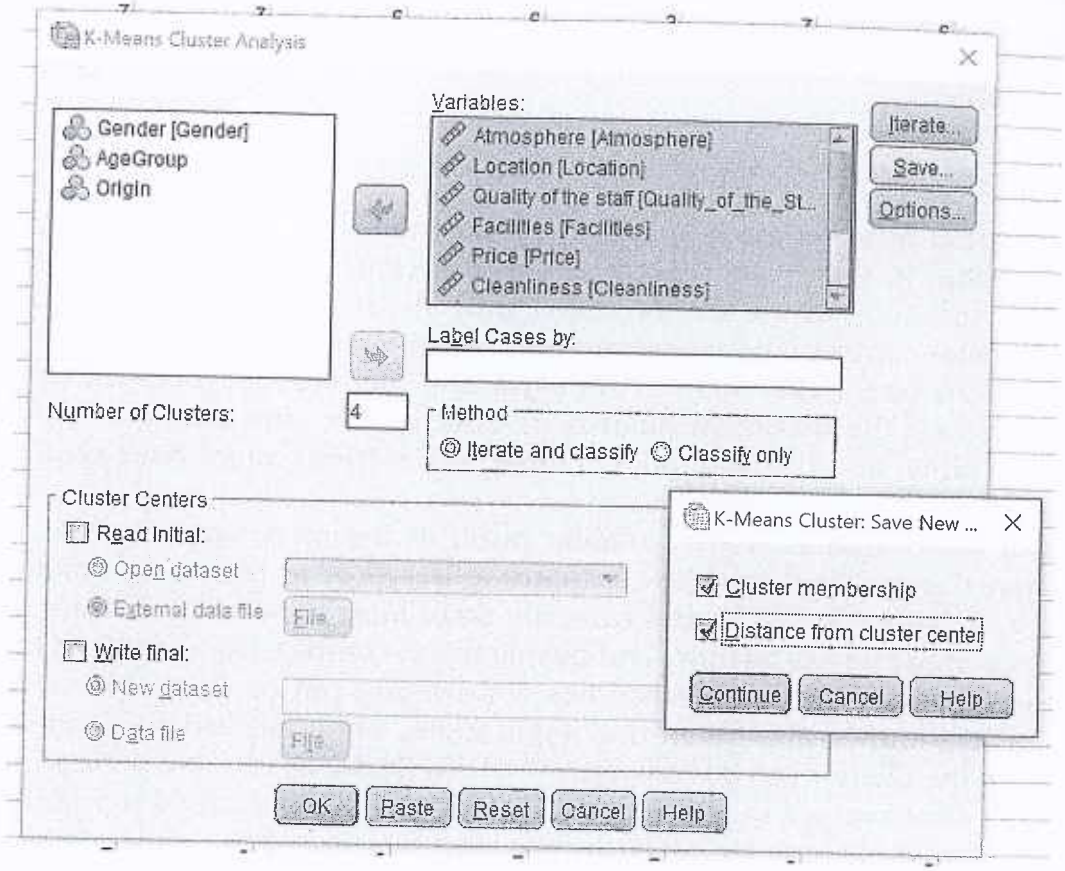


Figure 14.5 Screenshot of K-Means Cluster analysis

3 To specify the number of clusters, type 4 as the number of clusters (between two and the maximum number of cases in the database).
4 Click **Iterate** and set up the maximum number of iterations to 20.
5 Select either **Iterate** or **Classify**.
6 Click **Options** in the dialog box and select **ANOVA Table** and **Cluster Information for Each Case**.
7 Click **Save** and select **Cluster Membership** and **Distance from the Cluster Centre** (the first function generates a new variable representing each case's final cluster membership, with this variable's value potentially ranging from one to the final number of clusters; each case's distance from its cluster's core produces a new variable denoting the Euclidean distance between that item and its cluster's centre).
8 Another option is to select an identification variable to label cases.

Cluster analysis allows researchers to classify cases into homogeneous groups, making it a useful data reduction technique. However, researchers' selection of the final cluster solution is based on both objective and subjective decisions. The cluster solution's empirical robustness can be improved further by comparing the solution's stability. The comparison needs to consider (i) different distance measures, (ii) various clustering methods (i.e. hierarchical, non-hierarchical or other methods

using hierarchical procedures), (iii) the data split into two samples and (iv) different initial cluster centres when non-hierarchical methods are used.

Interpretation of clusters

The iteration history describes the progress of the clustering analysis algorithm at each step. In the present study, the cluster centres changed substantially in the first iterations. In the last iterations, only minor adjustments were observed in the cluster centres. The process ended before reaching the maximum number of iterations (see Table 14.2). If the algorithm had stopped because the algorithm reached the maximum number of observations, the solution might have been unstable, and the maximum number of iterations might have required an adjustment.

Table 14.2 indicates which variables result in the most significant separation between clusters. The variables contributing the most to the cluster solution are those with large F values. In this case, the three most important variables are the existence of bar service, security and cleanliness. In contrast, the variable of location appears to be unimportant. When hostel attributes can be deleted from further analyses without having theoretical implications, irrelevant variables not needed to obtain the clusters can be removed. The final cluster centres are defined as each variable's final average for each group. Each case within a cluster is represented by the group centroid, namely, each cluster's most typical case.

The interpretation stage involves analysing each cluster centre and similarities and differences between clusters. When the cluster analysis is performed based on raw data, the description can be done directly. However, when the data have been standardised, the analysis includes going back to the raw scores for the original variables and calculating the means for each cluster. For example, the present results revealed that the most important attribute for hostel guests in Cluster 1 is the existence of bar service; in Cluster 2, opportunities to meet other travellers; in Cluster 3, staff quality; and in Cluster 4, facilities. The least significant attribute for each group also varies so that, for Cluster 1, this is price; for Cluster 2, security; for Cluster 3, price; and, for Cluster 4, the existence of bar service.

Table 14.2 lists the distances between the final cluster centres and provides information about the Euclidean distance between these centres. The most different clusters exhibit greater distances. In this case, Clusters 2 and 3 are the most dissimilar. In addition, the results include the number of cases in each cluster. The table above also provides the number of hostel guests in each group: 106 guests in Cluster 1, 60 in Cluster 2, 39 in Cluster 3, and 80 in Cluster 4.

The variables retained in the final dataset, cluster membership and distance from cluster centre can be used to generate a useful box plot by following these steps:

1 From the menu, select **Graphs** > **Boxplot**.
2 Select **Simple** > **Data in Chart** > **Summaries of Separate Variables**.
3 Click **Define**.
4 Select the variables of distance of case from its cluster's centre and category axis of clusters' number of cases.

Table 14.2 K-means for cluster solution

Iteration history

	Change in cluster centres			
Iteration	1	2	3	4
1	3.223	3.359	3.576	3.565
2	.297	1.458	.447	.612
3	.316	1.019	.430	.480
4	.091	.440	.134	.154
5	.045	.433	.220	.139
6	.064	.546	.477	.323
7	.143	.333	.357	.366
8	.085	.111	.439	.353
9	.065	.100	.238	.170
10	.022	.055	.093	.083
11	.000	.000	.080	.050
12	.000	.000	.095	.056
13	.000	.000	.088	.049
14	.000	.000	.131	.069
15	.000	.000	.120	.057
16	.000	.000	.000	.000

Final cluster centres for four-centre solution

	Cluster			
	1.00	2.00	3.00	4.00
Atmosphere	6.26	6.57	5.87	6.25
Location	6.46	6.28	6.28	6.45
Staff quality	6.62	6.48	6.85	6.48
Facilities	6.52	5.73	6.72	6.59
Price	6.10	6.17	4.44	6.53
Cleanliness	6.62	4.88	6.44	6.45
Security	6.30	4.57	6.28	6.34
Existence of bar service	6.77	5.77	4.90	3.98
Opportunities to meet other travellers	6.63	6.70	4.95	6.21
Internet facilities	6.11	6.55	6.10	6.21

Distance between final cluster centres for four-cluster solution

Cluster	*1*	*2*	*3*	*4*
1		2.830	3.074	2.873
2	2.830		3.736	3.184
3	3.074	3.736		3.184
4	2.873	3.184	2.675	2.675

Note: Convergence achieved due to no or small changes in cluster centres; maximum absolute coordinate change for any centre is .000; current iteration is 16; minimum distance between initial centres is 6.557.

(*continued*)

Table 14.2 K-means for cluster solution (*continued*)

	Analysis of variance (ANOVA)					
	Cluster		Error			
	Mean square	df	Mean square	df	F	Sig.
Atmosphere	3.833	3	.757	281	5.065	.002
Location	.656	3	.649	281	1.011	.388
Staff quality	1.465	3	.316	281	4.631	.004
Facilities	11.640	3	.568	281	20.509	.000
Price	40.085	3	.874	281	45.838	.000
Cleanliness	43.345	3	.436	281	99.444	.000
Security	48.081	3	.516	281	93.270	.000
Existence of bar service	125.435	3	.672	281	186.652	.000
Opportunities to meet other travellers	31.038	3	.586	281	53.007	.000
Internet facilities	2.749	3	.571	281	4.814	.003

Number of guests in each cluster		
Cluster	1	106.000
	2	60.000
	3	39.000
	4	80.000
Valid		285.000
Missing		0.000

Note: F tests were only used to serve descriptive purposes as clusters were selected to amplify differences between cases in several clusters; significance levels observed in this study were not adjusted to reflect this, so they should not be treated as useful as a test of the hypothesis that clusters' means are equivalent values.

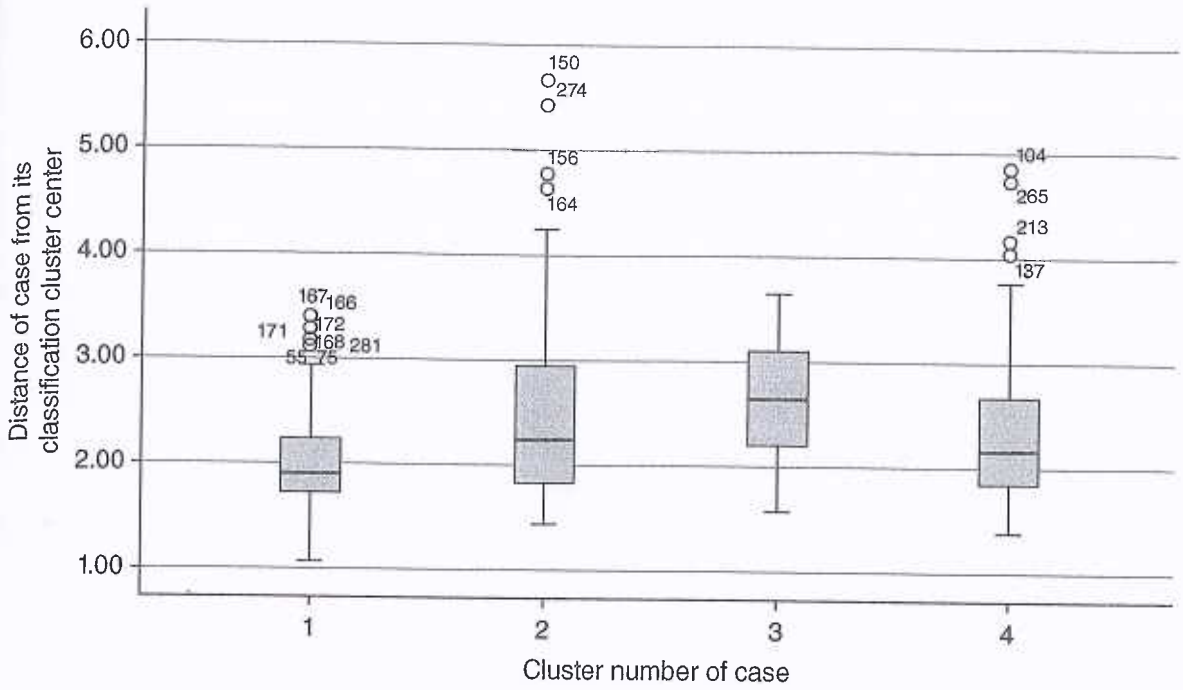


Figure 14.6 Plot of distances from cluster centre by cluster membership for four-cluster solution

The resulting box plot facilitates testing for outliers. In this case, some variability was observed in the clusters, but no extreme outliers were detected (see Figure 14.6).

Cluster analysis is by nature exploratory. The present study's solution was satisfactory, but, if this had not happened, the same solution could have been rerun without the outliers, variables that failed to contribute to the solution and a different number of clusters.

Cluster profiling with additional variables

A good cluster analysis solution is both efficient (i.e. defines as few clusters as possible) and effective (i.e. identifies groups that are significant market segments). The final cluster solution must be relevant in practice. Variables not used to generate the clusters' profile can also be used in subsequent analyses, such as demographics and psychographics.

Crosstabulation tables can highlight the relationship between two variables measured as nominal or ordinal. The variables' independence can also be tested. In the present study, the new variable – cluster membership – was a nominal variable with four categories. Gender and origin are nominal variables, and age group is ordinal. Three demographic and geographic variables that described the hostel guests in the sample were used to profile the clusters: gender, age

group and country of origin. The frequencies table for these three variables are obtained as follows:

1 Select **Analyze > Descriptive Statistics > Frequencies.**
2 Select the variables of gender, age group and origin.

The table generated lists the number of respondents and the relative frequency in each cluster of the three variables under analysis. To create a crosstabulation table for the clusters and demographic variables, the steps below are followed:

1 From the menu, choose **Analyze > Table > Custom Table.**
2 Select the variables to appear in rows to be gender, age group and origin and the variable to appear in columns to be clusters.
3 Another option is to select from the menu **Test Statistics: Tests of Independence (Chi-Square)** and the coefficient of association **Cramer's V.**

Interpretation of results

Sample profile

The current research's dataset included 149 males (52.3%) and 136 males (47.7%). Most respondents were 15 to 25 years old (173 guests, 60.7% of the sample), followed by guests 26 to 29 years old (71, 24.6%) and 30 years old or older (41, 14.4%). Due to the large number of nationalities in the sample, they were further grouped into major regions. More than half of the guests are from Europe (157, 55.1%), followed by the USA/Canada (61, 21.4%), Australia (33, 11.6%), Latin America (22, 7.7%) and Asia (12, 4.2%).

Attributes for assessment of overall hostel experiences

Although the hotel guests considered all the attributes important, they could be ranked from the most to least significant. The attributes that were given the highest importance are, in descending order, staff quality (mean = 6.58; SD = 0.573), facilities (mean = 6.40; SD = 0.827), location (mean = 6.40; SD = 0.805) and opportunities to meet other travellers (mean = 6.30; SD = 0.952). The attributes that received the lowest scores are, in ascending order, the existence of bar service (mean = 5.52; SD = 1.411), security (mean = 5.94; SD = 1.009), price (mean = 6.01; SD = 1.135), cleanliness (mean = 6.18; SD = 0.943) and Internet facilities (mean = 6.23; SD = 0.711). The three attributes with the lowest mean scores are also those that are associated with the highest variability.

Market segmentation of hostel guests

The visual inspection of the dendrogram and the agglomeration coefficient's evolution in the different steps of applying the algorithm used in Ward's method suggested a solution with four clusters. These clusters are heterogeneous in terms of

size. The largest is Cluster 1, with 106 hostel guests (37%); followed by Cluster 4, with 80 (28%); Cluster 2, with 60 (21%); and Cluster 3, with 39 guests (14%). Each market segment can be described by the cluster's centroid, which is the mean score given to indicate the importance of each attribute used to assess hostel experiences. Guests from Cluster 1 value most the existence of bar service (mean = 6.77), followed by opportunities to meet other travellers (mean = 6.63) and staff quality and cleanliness (both with mean = 6.62). This group's least valued attribute is price. These guests enjoy opportunities to party and socialise in hostels. The greater significance given to bar service differentiates this group's members from those of Clusters 2, 3 and 4.

Guests from Cluster 2 differed from other groups because these clients place the highest importance on opportunities to meet other travellers (mean = 6.70), atmosphere (mean = 6.57) and Internet facilities (mean = 6.70). The attribute of least importance to this group, as well as to the entire sample, is security (mean = 4.57). Guests from this group thus value hostels' social components.

Guests from Cluster 3 registered the highest scores in the sample for the attributes staff quality (mean = 6.85) and facilities (mean = 6.72). The third most important attribute for this group is cleanliness (mean = 6.44). Price is of the least significance to these guests (mean = 4.44), as well as for the entire sample. This group, therefore, values hostels' tangible components.

Guests from Cluster 4 give greater importance to facilities (mean = 6.59), price (mean = 6.53) and staff quality (mean = 6.48). This group also places the highest importance on price of all the clusters. The attribute of least importance to these guests is the existence of bar service (mean = 3.98). Thus, guests in this group are price sensitive.

The results obtained confirm the existence of four groups labelled as follows. Cluster 1 is party-oriented guests. Cluster 2 is social-oriented guests. Cluster 3 is tangibles-oriented guests. Cluster 4 is price-sensitive guests. Hotel guests are clearly not a homogeneous group, so Proposition 1 was verified. Hostel guests can be segmented according to their perceptions of different service quality attributes' importance.

Profiling of segments

Cluster 1 or the party-oriented group comprises mainly 15–25- and 26–29-year-old guests from Asia, Latin America, Australia and Europe. Cluster 2 was labelled the social-oriented group, which includes mainly males 26–29 years old from the USA/Canada, Europe and Australia. This cluster does not include any guests from Latin America and Asia. Cluster 3 is the tangibles-oriented group. These guests are mainly 15 –25 and 30 years old or older, and they are from Europe. The cluster comprises both males and females. Finally, Cluster 4 was termed the price-oriented cluster, which includes mainly individuals who are 30 years old or more and from the USA/Canada (Table 14.3).

The above geographic and demographic profile of the four segments confirms Proposition 2. Segments of hostel guests based on variations in the importance of service quality attributes also vary according to demographic characteristics, in this case, age, gender and nationality.

Table 14.3 Cluster solution – centroid and demographic and geographic profile

		Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*
		Average			
Atmosphere		6.26	6.57	5.87	6.25
Location		6.46	6.28	6.28	6.45
Staff quality		6.62	6.48	6.85	6.48
Facilities		6.52	5.73	6.72	6.59
Price		6.10	6.17	4.44	6.53
Cleanliness		6.62	4.88	6.44	6.45
Security		6.30	4.57	6.28	6.34
Existence of bar service		6.77	5.77	4.90	3.98
Opportunities to meet other travellers		6.63	6.70	4.95	6.21
Internet facilities		6.11	6.55	6.10	6.21
		Frequency			
		n = 106	n = 60	n = 39	n = 80
Gender	Male	51	39	20	39
	Female	55	21	19	41
Age group	15–25	74	39	13	47
	26–29	29	20	6	16
	30 ≤	3	1	20	17
Origin	US/Canada	14	18	5	24
	Europe	48	33	31	45
	Latin America	14	0	1	7
	Australia	20	9	1	3
	Asia	10	0	1	1

Note: n = number.

References

Brochado, A. & Gameiro, C. (2013). Toward a better understanding of backpackers' motivations. *Tékhne – Review of Applied Management Studies*, *11*, 92–99.

Brochado, A., Gameiro, C. & Rita, P. (2015). Exploring backpackers' perceptions of the hostel's service quality. *International Journal of Contemporary Hospitality Management*, *27*(8), 1839–1855.

Brochado, A. & Rita, P. (2018). Exploring heterogeneity among backpackers in hostels. *Current Issues in Tourism*, *21*(13), 1502–1520.

Dolnicar, S. (2002). A review of data-driven market segmentation in tourism. *Journal of Travel and Tourism Marketing*, *12*(1), 1–22.

Musa, G. & Thirumoorthi, T. (2011). Red Palm: Exploring the service quality and servicescape of the best backpacker hostel in Asia. *Current Issues in Tourism*, *14*(2), 103–120.

'The content of this book is superior to all others offering guidance on SPSS and will be of particular use to those studying Tourism, Hospitality and Events. It is a groundbreaking book which includes the rare addition of a separate section on writing interpretations. I recommend the purchase of this book for Tourism and Hospitality researchers.'
Professor S.P. Bansal, Vice Chancellor, Himachal Pradesh Technical University, India

'This is an excellent guide full of crucial and innovative information. The guidance on writing up the interpretation of SPSS outputs and the short case studies that open each chapter make the book quite interesting for those of us who want to enjoy research in Tourism, Hospitality and Events.'
Professor Justin Paul, University of PR, San Juan, USA

'This is an excellent guide for students and researchers who want to gain knowledge and skills about statistical analysis with SPSS. The relevant empirical examples help the reader understand not only the theoretical concepts in statistical analysis but help to see how they are applied in actual research projects. The screenshots guide the reader through the implementation of the respective analyses in easy-to-follow steps.'
Professor Stanislav Ivanov, Varna University of Management, Bulgaria

This is the first book to provide the student of tourism, hospitality and events with all that they need to undertake statistical analysis using SPSS for research in their industry. Employing examples directly from the tourism, hospitality and events sector, it provides a comprehensive explanation on how appropriate statistical tools and methods can be identified for this research context and provides a step-by-step demonstration on how to carry out the chosen statistical operations.

Each chapter opens with a sector-specific case study reflecting current research trends and issues from a range of different countries that are affecting the industry today. This is followed by an examination of the SPSS procedures relating to the case study and various solutions are offered. The implementation of clear, step-by-step demonstrations on how to carry out statistical operations using a combination of screenshots, diagrams and tables aids the reader's understanding. Chapters close with thorough guidance on how to appropriately write up interpretations of the research in a report. Research implications and recommendations for tourism and hospitality businesses are also provided, to enable them to successfully create and manage research strategies in action.

Adopting an interdisciplinary perspective and written by a range of industry experts from across the globe, this book will be essential for all students and researchers in the field of tourism, hospitality and events as well as all those in related fields with an interest in statistical data analysis.

Rahul Pratap Singh Kaurav is an Assistant Professor in Marketing and Tourism Management at the Prestige Institute of Management, Gwalior, India.

Dogan Gursoy is the Taco Bell Distinguished Professor in Hospitality Business Management in the School of Hospitality Business Management, Washington State University, USA.

Nimit Chowdhary is a Professor at the Department of Tourism and Hospitality Management, Jamia Millia Islamia, New Delhi, India.

TOURISM, HOSPITALITY AND EVENTS MANAGEMENT

Cover image: © Getty Images

Routledge
Taylor & Francis Group
www.routledge.com

eRESOURCES
www.routledge.com/9780367236588

an **informa** business
ISBN 978-0-367-23658-8

Routledge titles are available as eBook editions in a range of digital formats