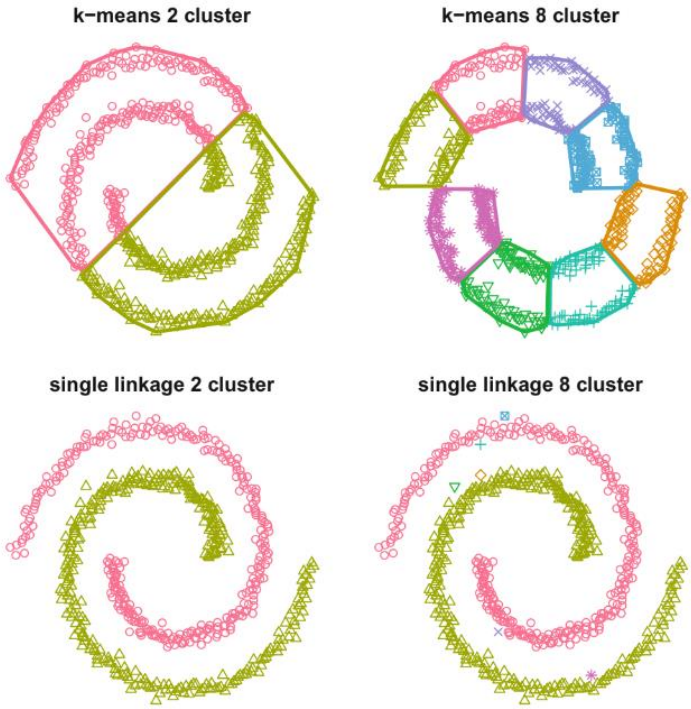
STEP 5

7.1 Grouping Consumers

Market segmentation analysis is a process of dividing consumers into distinct groups based on their characteristics and preferences. However, consumer data is often unstructured, and consumer preferences tend to be spread across a wide range rather than forming clear groups. Therefore, extracting market segments from such data requires the use of exploratory methods.

Cluster analysis is commonly used in market segmentation, where market segments correspond to clusters. The choice of a clustering method depends on the data and the researcher's specific requirements. Different clustering algorithms impose different structures on the extracted segments.

To obtain meaningful market segmentation solutions, it is important to explore different clustering methods and understand how they shape the segments. The choice of the algorithm should align with the specific requirements and characteristics of the data. By considering various methods and their implications, researchers can derive more accurate and insightful market segmentation analyses.



The example illustrated in Figure may give the impression that single linkage clustering is more powerful and should be preferred for market segmentation analysis. However, this is not the case. The data set used in this example was specifically designed to showcase the strengths of the single linkage algorithm in identifying the grouping corresponding to the spirals. In summary, the interaction between the data and the algorithm is critical in market segmentation analysis. Different algorithms have different tendencies and should be chosen based on the specific characteristics and requirements of the data set. There is no one-size-fits-all approach, and researchers need to carefully consider the data and algorithm interplay to obtain meaningful segmentation solutions.

Data set and segment characteristics informing extraction algorithm selection

Data set characteristics:

– Size (number of consumers, number of segmentation variables)
– Scale level of segmentation variables (nominal, ordinal, metric, mixed)
– Special structure, additional information

Segment characteristics:

– Similarities of consumers in the same segment
– Differences between consumers from different segments
– Number and size of segments

7.2 Distance-Based Methods

In the context of market segmentation for tourists with similar activity patterns on vacation, a fictitious data set is provided in Table 7.2. This data set includes information about seven individuals and the percentage of time they spend on different activities such as BEACH, ACTION, and CULTURE during their vacations.

To identify groups of tourists with similar activity patterns, a distance-based approach can be employed. The goal is to measure the similarity or dissimilarity between individuals based on their activity preferences. For instance, Anna and Bill share the same profile as they both prefer to relax on the beach. Thus, they should be grouped together in the same segment. On the other hand, Michael stands out from the others as he is not interested in the beach, which sets him apart.

To quantify the similarity or dissimilarity, a distance measure is required. This measure will mathematically capture the differences between individuals based on their activity preferences. By calculating the distances between individuals, it becomes possible to identify groups of similar tourists who have comparable vacation activity patterns.

In summary, in order to group tourists with similar activity patterns, a distance-based method is utilized. By employing an appropriate distance measure, such as Euclidean distance or cosine similarity, the similarities and dissimilarities between individuals can be quantified, enabling the identification of distinct segments of tourists with similar vacation activity preferences.

7.2.1 Distance Measures

The most common distance measures used in market segmentation analysis are:

- Euclidean distance
- Manhattan or absolute distance
- Asymmetric binary distance

Euclidean distance:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}$$

Manhattan or absolute distance:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{p} |x_j - y_j|$$

Asymmetric binary distance: applies only to binary vectors, that is, all xj and yj are either 0 or 1.

$$d(\mathbf{x}, \mathbf{y}) = \begin{cases} 0, & \mathbf{x} = \mathbf{y} = \mathbf{0} \\ (\#\{j | x_j = 1 \text{ and } y_j = 1\})/(\#\{j | x_j = 1 \text{ or } y_j = 1\}) \end{cases}$$

7.2.2 Hierarchical Methods

Hierarchical clustering methods are considered intuitive for grouping data because they mimic how a human would approach the task of dividing a set of observations (such as consumers) into distinct groups or segments. When applying hierarchical clustering to market segmentation, the goal is to find an appropriate number of segments (k) that balances between having one large segment (k = 1) and having each consumer in their own individual segment (k = n). Overall, hierarchical clustering is an intuitive method for market segmentation as it emulates the human approach to grouping data. It allows for the exploration of various segmentations, considering different numbers of segments, and provides valuable insights into consumer behaviour patterns.

Divisive hierarchical clustering methods start with the complete data set X and splits it into two market segments in a first step. Then, each of the segments is again split into two segments. This process continues until each consumer has their own market segment.

Agglomerative hierarchical clustering approaches the task from the other end. The starting point is each consumer representing their own market segment (n singleton clusters). Step-by-step, the two market segments closest to one another are merged until the complete data set forms one large market segment.

Both approaches result in a sequence of nested partitions. A partition is a grouping of observations such that each observation is exactly contained in one group. The sequence of partitions ranges from partitions containing only one group (segment) to n groups (segments). They are nested because the partition with k + 1 groups (segments) is obtained from the partition with k groups by splitting one of the groups.

In hierarchical clustering, both divisive and agglomerative methods rely on a distance measure between groups of observations (segments) to determine the similarity or dissimilarity between them. The linkage method is then used to generalize how distances between pairs of observations translate into distances between groups of observations.

Single linkage: The distance between two clusters is defined as the minimum distance between any two points, one from each cluster. It focuses on the closest pair of points between the two clusters.

$$l(\mathcal{X}, \mathcal{Y}) = \min_{\mathbf{x}\in\mathcal{X}, \mathbf{y}\in\mathcal{Y}} d(\mathbf{x}, \mathbf{y})$$

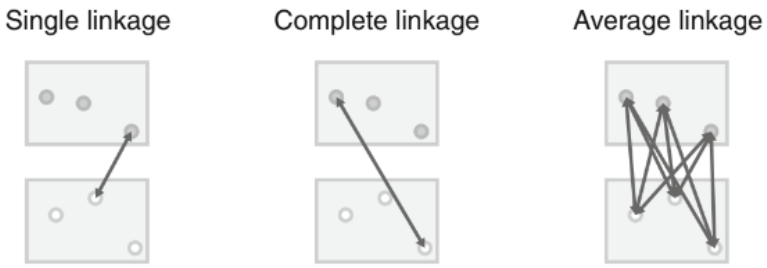
Complete linkage: The distance between two clusters is defined as the maximum distance between any two points, one from each cluster. It considers the farthest pair of points between the two clusters.

$$l(\mathcal{X}, \mathcal{Y}) = \max_{\mathbf{x}\in\mathcal{X}, \mathbf{y}\in\mathcal{Y}} d(\mathbf{x}, \mathbf{y})$$

Average linkage: The distance between two clusters is defined as the average distance between all pairs of points, one from each cluster. It takes into account the average similarity across all pairs of points between the clusters.

$$l(\mathcal{X}, \mathcal{Y}) = \frac{1}{|\mathcal{X}||\mathcal{Y}|} \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{y}\in\mathcal{Y}} d(\mathbf{x}, \mathbf{y}),$$

A comparison of different linkage methods between two sets of points



7.2.3 Partitioning Methods

Hierarchical clustering methods are well-suited for analyzing small datasets with up to a few hundred observations (consumers). However, for larger datasets, dendrograms become difficult to interpret, and the pairwise distance matrix may not fit into computer memory. In such cases, clustering methods that create a single partition, rather than a nested sequence of partitions, are more appropriate. Instead of computing all pairwise distances between

observations upfront, these partitioning methods calculate distances between each consumer and the center of the segments. This significantly reduces the number of distances to compute, making the analysis computationally feasible for larger datasets. Additionally, if the goal is to extract only a few segments, it is more efficient to optimize specifically for that objective rather than constructing the complete dendrogram and then cutting it into segments using heuristics.

7.2.3.1 k-Means and k-Centroid Clustering

k-Means and k-Centroid clustering methods are widely used in market segmentation to group consumers based on their similarities in preferences or behavior. These methods aim to partition the data into k distinct segments, with each segment represented by a centroid or a mean vector.

In k-Means clustering, the algorithm starts by randomly assigning k initial centroids and iteratively updates them to minimize the within-cluster sum of squares. Each consumer is then assigned to the nearest centroid, forming clusters based on proximity. The process continues until convergence, where the centroids stabilize and the clusters become consistent. k-Means clustering is computationally efficient and suitable for large datasets. However, it is sensitive to the initial centroid placement and can converge to suboptimal solutions.

k-Centroid clustering is a variation of k-Means clustering that allows for the use of different distance measures and centroid updating rules. It offers more flexibility in defining the similarity between consumers and adapting the clustering process to specific requirements. The choice of distance measure, such as Euclidean or Manhattan distance, affects the clustering results and should be aligned with the characteristics of the data.

Both k-Means and k-Centroid clustering methods require specifying the number of segments (k) in advance. Determining the optimal value of k is often a challenge and can be addressed through techniques like the elbow method or silhouette analysis. It is important to consider the interpretability and meaningfulness of the resulting segments in the context of the marketing problem at hand.

Overall, k-Means and k-Centroid clustering methods provide efficient and effective approaches for market segmentation, allowing businesses to identify distinct consumer groups based on their similarities and tailor their marketing strategies accordingly.

7.2.3.2 “Improved” k-Means

The k-means clustering algorithm can be improved by using "smart" starting values instead of randomly selecting initial centroids. Random selection may lead to suboptimal solutions as some randomly chosen consumers may be close to each other and not representative of the entire data space. This increases the risk of the algorithm getting stuck in local optima, which are good solutions but not the best possible ones. One way to address this is by evenly spreading the starting points across the data space to better represent the entire dataset.

Research by Steinley and Brusco (2007) compared 12 different strategies for initializing the k-means algorithm. Through extensive simulations using artificial datasets with known structures, they found that the best approach is to randomly draw multiple starting points and select the set that best represents the data. Good representatives are those that are close to their segment members, resulting in a small total distance of all segment members to their representatives. In contrast, bad representatives are far away from their segment members, leading to a high total distance. By selecting the best set of starting points, the k-means algorithm can be more likely to converge to a better solution.

7.2.3.3 Hard Competitive Learning

Hard competitive learning, also known as learning vector quantization, is a different approach to segment extraction compared to the standard k-means algorithm. While both methods aim to minimize the sum of distances between consumers and their closest segment representatives, they differ in the process by which this is achieved. In hard competitive learning, a random consumer is selected, and their closest segment representative is moved a small step towards that consumer's position.

This procedural difference can lead to different segmentation solutions, even when using the same initial starting points. It is also possible that hard competitive learning may find the globally optimal segmentation solution, while k-means may get trapped in a local optimum, or vice versa. Neither method is superior to the other; they simply offer distinct approaches to segmentation analysis.

Hard competitive learning has been applied in market segmentation analysis, such as in segment-specific market basket analysis. The "cclust" function from the "flexclust" package in R can be used to perform hard competitive learning clustering.

7.2.3.4 Neural Gas and Topology Representing Networks

A variation of hard competitive learning is the neural gas algorithm, which adjusts not only the location of the closest segment representative but also the second closest representative towards the randomly selected consumer. Neural gas clustering has been applied in market segmentation analysis and can be performed using the "cclust" function with the "neuralgas" method in the "flexclust" package in R.

Topology representing networks (TRN) is a further extension of neural gas clustering. TRN counts the frequency of pairs of segment representatives being closest and second closest to consumers and uses this information to construct a virtual map where similar representatives are placed next to each other. While the original TRN algorithm is not implemented in R, using neural gas in combination with segment neighbourhood graphs achieves similar results. The segment neighbourhood graph provides information about the relationships between segment representatives.

Neural gas and topology representing networks are not superior to the k-means algorithm or hard competitive learning; they simply offer different approaches to market segmentation analysis. Having a larger toolbox of algorithms available for exploration is valuable in data-driven segmentation analysis, which is inherently exploratory in nature.

7.2.3.5 Self-Organising Maps

Another variation of hard competitive learning is the self-organising map (SOM) or self-organising feature map. In SOM, segment representatives (centroids) are positioned on a regular grid, such as a rectangular or hexagonal grid. The algorithm works similarly to hard competitive learning, where a random consumer is selected, and the closest representative moves towards it. Additionally, the neighbouring representatives on the grid also adjust their positions. This process is repeated multiple times until a final solution is reached. The advantage of SOM is that the numbering of market segments aligns with the grid, providing a structured representation. However, the trade-off is that the sum of distances between segment members and representatives may be larger compared to other clustering algorithms. Studies have compared SOM with other algorithms like k-means and topology representing networks for market segmentation applications.

7.2.3.6 Neural Networks

Auto-encoding neural networks provide a different approach to cluster analysis compared to traditional clustering methods. These networks, particularly the single hidden layer perceptron, are widely used in this context. The network consists of an input layer, a hidden layer, and an output layer. The hidden layer computes weighted combinations of the input variables using non-linear functions. The network is trained to minimize the squared Euclidean distance between the inputs and outputs. The parameters connecting the hidden layer to the output layer can be interpreted as segment representatives, while the parameters connecting the input layer to the hidden layer represent consumer membership in segments. Auto-encoding neural networks generate fuzzy segmentations, where consumers can belong to multiple segments with membership values between 0 and 1. Various implementations of auto-encoding neural networks are available in R, such as the autoencoder package. Other clustering algorithms also produce fuzzy market segmentation solutions.

7.2.4 Hybrid Approaches

Hybrid segmentation approaches aim to combine the strengths of hierarchical and partitioning clustering algorithms while mitigating their weaknesses. Hierarchical clustering offers the advantage of not requiring the number of segments to be specified in advance and provides visualizations through dendrograms. However, it can be memory-intensive and challenging to interpret with large sample sizes. On the other hand, partitioning clustering algorithms are efficient for segmenting large datasets but necessitate specifying the number of segments in advance and lack the ability to track changes in segment membership across

different solutions. In hybrid approaches, a partitioning algorithm is initially applied to extract a larger number of segments. Then, only the segment representatives (centroids) and sizes are retained for input into hierarchical clustering, where the dendrogram can inform the decision on the appropriate number of segments to extract. This hybrid approach leverages the scalability of partitioning algorithms and the interpretability of hierarchical algorithms.

7.2.4.1 Two-Step Clustering

The two-step clustering procedure, implemented in IBM SPSS, combines a partitioning algorithm with a hierarchical algorithm. It has been widely used in various application areas, such as segmenting mobile phone users based on internet access types, segmenting potential nature-based tourists, identifying potential electric vehicle adopters, and segmenting travel-related risks.

The basic idea of the two-step clustering approach is to first apply a partitioning algorithm, such as k-means, to extract a larger number of clusters than the desired market segments. This helps reduce the size of the dataset by retaining only one representative member from each cluster. The extracted clusters can be visualized using a neighborhood graph, which shows cluster means as nodes and their similarities as edges.

In the second step, a hierarchical cluster analysis is performed using the representatives of the extracted clusters and their segment sizes. The resulting dendrogram from the hierarchical analysis reveals the existence of market segments based on the clustering solution. However, the original data's cluster memberships cannot be determined solely from the hierarchical analysis since the data was discarded. To link the original data with the segmentation solution, the twoStep() function from the MSA package can be used, taking the hierarchical clustering solution, the cluster memberships from the partitioning algorithm, and the desired number of segments as inputs.

Overall, the two-step clustering procedure combines the strengths of partitioning and hierarchical algorithms, allowing for scalable analysis of large datasets while providing insights into market segmentation through the hierarchical structure.

7.2.4.2 Bagged Clustering

Bagged clustering is a segmentation technique that combines hierarchical clustering and partitioning clustering with bootstrapping. It addresses the challenges of identifying market segments by repeating the clustering process using randomly drawn (bootstrapped) samples from the data set. The procedure starts with partitioning clustering on the bootstrapped data sets to obtain cluster centroids. These centroids are then used as input for hierarchical clustering, which provides insights into the optimal number of market segments to extract.

Bagged clustering offers several advantages. It can identify niche markets by capturing them as distinct branches in the dendrogram. It reduces the dependency on specific individuals in

the data by using bootstrapping. It is suitable when niche markets are suspected, when there is a concern about local solutions, or when the data set is large. The five steps of bagged clustering involve creating bootstrap samples, performing partitioning clustering on each sample, creating a derived data set with cluster centroids, conducting hierarchical clustering on the derived data set, and determining the final segmentation solution by assigning observations to the closest market segment representative.

Bagged clustering has been successfully applied to tourism data and has been used to segment winter vacation activities based on responses from tourists. This technique provides a robust and efficient approach to market segmentation, particularly when dealing with large data sets or when niche markets need to be identified**.**

7.3 Model-Based Methods

Model-based methods have emerged as an alternative to distance-based methods for market segmentation analysis. These methods, such as finite mixture models, are based on the assumption that market segments have specific characteristics and sizes. Unlike distance-based methods that rely on similarities or distances between consumers, model-based methods estimate the parameters of a finite mixture model that best reflects the data. Estimation is typically done using maximum likelihood estimation or Bayesian methods.

The number of segments in a finite mixture model needs to be specified in advance, which can be challenging. Information criteria such as AIC, BIC, and ICL are commonly used to guide the selection of the number of segments. These criteria balance model complexity with goodness-of-fit to the data. By comparing models with different numbers of segments, analysts can determine the most appropriate number.

Finite mixture models offer the advantage of capturing complex segment characteristics and can be extended in various ways. They provide an alternative approach to market segmentation and complement distance-based methods by offering a different extraction technique. While they may initially appear complex, they offer flexibility and can accommodate various modeling scenarios.

Overall, model-based methods, particularly finite mixture models, have gained significant interest in market segmentation analysis and are considered influential in the field. They provide a valuable tool for exploratory segment extraction, allowing analysts to uncover distinct market segments based on their specific characteristics and sizes.

7.3.1 Finite Mixtures of Distributions

In the simplest case of model-based clustering, independent variables are not included, and the focus is solely on fitting a distribution to the segmentation variable y. This approach is

comparable to distance-based methods, as both use the same segmentation variables to identify market segments. These segmentation variables represent consumer characteristics, such as vacation activities, without the inclusion of additional information like travel expenditures.

The formulation of the finite mixture model in this case reduces to the form:

$$\sum_{h=1}^{k} \pi_h f(y|\theta_h), \quad \pi_h \geq 0, \quad \sum_{h=1}^{k} \pi_h = 1.$$

The primary difference from Equation 7.1 is the absence of the independent variables x. The statistical distribution function f () used in the model depends on the measurement level or scale of the segmentation variables y.

In summary, the simplest model-based clustering approach focuses solely on fitting distributions to segmentation variables, similar to distance-based methods. It allows for the identification of market segments based on the characteristics captured by these variables, without considering additional information about consumers.

7.3.1.1 Normal Distributions

The most popular finite mixture model for metric data is a mixture of multivariate normal distributions. This model is suitable for variables that exhibit covariance between them, such as physical measurements on humans or prices in markets with multiple players. The multivariate normal distribution can capture the correlation between variables, allowing for an accurate representation of the data.

In this model, each segment is characterized by a mean vector and a covariance matrix. The mean vector contains the mean values for each segmentation variable, while the covariance matrix represents the variances and covariances between pairs of variables. The covariance matrix is symmetric and contains unique values for the variances and covariances.

To fit a mixture of normal distributions, the R package mclust is commonly used. It employs the EM algorithm and determines the number of segments based on the Bayesian Information Criterion (BIC). The BIC takes into account both the number of segments and the shape of the covariance matrices, allowing for model selection.

The mclust package provides various covariance models, including spherical and ellipsoidal shapes. Spherical covariance matrices assume equal variances for all variables, while ellipsoidal covariance matrices can have different shapes, areas, and orientations. By selecting an appropriate covariance structure, the number of parameters that need to be estimated can be reduced.

The BIC values obtained from fitting different models are used to determine the optimal number of segments. Lower BIC values indicate a better fit, and the BIC often recommends a parsimonious model with fewer segments if it provides a good representation of the data.

In summary, fitting a mixture of normal distributions is a common approach for market segmentation when the variables are metric. The model captures the covariance between variables and allows for the identification of distinct segments. The choice of covariance structure and the number of segments are determined based on the BIC, which guides the selection of an appropriate model that balances complexity and fit to the data.

7.3.1.2 Binary Distributions

In this, a mixture model of binary distributions, also known as latent class analysis or latent class models, is used to analyze binary data related to vacation activities of Austrian tourists. The model assumes that different segments of respondents have different probabilities of engaging in specific activities. For instance, some respondents may be interested in alpine skiing but not in sight-seeing, while others may be interested in sight-seeing but not in alpine skiing. The goal is to identify these segments and understand the association between the activities.

By fitting a mixture model using the flexmix R package, the observed frequencies of the activities can be compared to the expected frequencies from the model. The EM algorithm is used to estimate the parameters of the model. Multiple random restarts are performed to avoid local optima, and the model with the highest likelihood or information criteria (such as AIC or BIC) is selected.

The fitted model reveals the presence of two segments. Segment 1 consists of respondents with a high likelihood of engaging in sight-seeing and a low probability of participating in alpine skiing. On the other hand, Segment 2 represents respondents who prefer alpine skiing and are not interested in sight-seeing. The expected frequencies of the activities based on the fitted model closely match the observed frequencies, indicating that the association between the activities across all consumers is explained by the segmentation.

In summary, the mixture model of binary distributions provides insights into the different segments of tourists based on their activity preferences and explains the association between activities by capturing the distinct patterns within each segment.

7.3.2 Finite Mixtures of Regressions

Finite mixtures of regression models offer a distinct approach to market segmentation analysis. They assume that the relationship between a dependent variable (such as willingness to pay) and independent variables (such as the number of rides in a theme park) differs across different market segments. By fitting a finite mixture of regression models, it is possible to identify and characterize these segments. In the example given, the data set represents two

market segments with different preferences regarding the number of rides and their willingness to pay. The mixture model estimates regression coefficients for each segment, allowing for segment-specific insights. However, it is important to note that fitting mixtures with the EM algorithm may result in label switching, where the order of segments in the output may differ from the true underlying segments. Overall, finite mixtures of regression models provide a valuable tool for market segmentation analysis, enabling a deeper understanding of consumer behaviour and preferences.

7.3.3 Extensions and Variations

They can accommodate various types of data, including metric, binary, nominal, and ordinal variables, by using appropriate statistical models for each data type. Mixture models can handle complex data characteristics and capture the presence of distinct market segments while allowing for variation within segments. They can also account for response styles in ordinal variables and incorporate preference differences in conjunction with conjoint analysis. Additionally, mixture models can be extended to incorporate mixed-effects models or heterogeneity models, which acknowledge both distinct segments and variation within segments. Mixture models can also be applied to time series data to cluster consumers over time or track changes in brand choice using dynamic latent change models or Markov chains. Furthermore, mixture models enable the simultaneous inclusion of segmentation and descriptor variables, allowing for the modelling of segment sizes based on differences in the composition of descriptor variables. In summary, finite mixture models provide a versatile framework for market segmentation analysis, accommodating diverse data types and capturing the complexity of consumer behaviour.

7.4 Algorithms with Integrated Variable Selection

While most segmentation algorithms assume that all segmentation variables contribute to the segmentation solution, it is common for datasets to contain redundant or noisy variables. Pre-processing methods can help identify and remove such variables. For metric variables, the filtering approach by Steinley and Brusco (2008a) assesses the clusterability of each variable and only includes those above a certain threshold as segmentation variables. However, for binary variables, it is more challenging to identify redundant or noisy variables during pre-processing. In such cases, biclustering and the variable selection procedure for clustering binary data (VSBD) by Brusco (2004) are algorithms that extract segments while simultaneously selecting suitable segmentation variables. Another approach, factor-cluster analysis, compresses segmentation variables into factors before segment extraction. These methods enhance the segmentation process by identifying relevant and informative variables, particularly for binary data.

7.4.1 Biclustering Algorithms

Biclustering is a versatile method that simultaneously clusters consumers and variables, particularly in the binary data case. It has gained popularity with the rise of genetic and proteomic data analysis. Biclustering algorithms aim to identify groups of consumers who share a common value of 1 for a subset of variables, forming biclusters. The process involves rearranging the data matrix to create a large rectangle of 1s, assigning consumers within this rectangle to a bicluster, and repeating the procedure until no more biclusters of sufficient size can be found. The repeated Bimax algorithm is often used for this purpose, providing efficient and optimal solutions. Biclustering allows for the identification of different patterns within biclusters, such as constant column patterns representing consumers with identical socio-demographics. It is particularly useful for datasets with numerous segmentation variables and offers advantages such as preserving the original data, capturing niche markets, and accommodating large numbers of variables without data transformation. However, it does not group all consumers, leaving some ungrouped.

7.4.2 Variable Selection Procedure for Clustering Binary Data (VSBD)

Brusco (2004) proposed the Variable Selection Procedure for Clustering Binary Data (VSBD), which is based on the k-means algorithm and aims to identify the most relevant variables for clustering. The method assumes that not all variables are important for obtaining a good clustering solution and focuses on removing masking variables. The procedure starts by selecting a small subset of observations, determined by a parameter φ. Then, an exhaustive search is performed to find the subset of variables that minimizes the within-cluster sum-of-squares criterion. Additional variables are added one by one, selecting the variable that leads to the smallest increase in the within-cluster sum-of-squares. The procedure stops when the increase exceeds a threshold. The number of clusters, k, needs to be specified in advance, and the Ratkowsky and Lance index is recommended to determine the optimal value. The algorithm involves random initializations and can be computationally demanding, but using the Hartigan-Wong algorithm can reduce the number of initializations required.

7.4.3 Variable Reduction: Factor-Cluster Analysis

Factor-cluster analysis is a two-step procedure used for data-driven market segmentation. In the first step, segmentation variables are factor analyzed, and the resulting factor scores are used in the second step to extract market segments. While this approach can be justified in cases where the original variables are designed to load onto factors, such as in validated psychological test batteries, it is often used to deal with a high number of segmentation variables relative to the sample size. However, this approach lacks conceptual legitimacy and

comes with substantial costs. Factor analysis leads to a loss of information, as a significant portion of the variability in the data is discarded. Additionally, factor-cluster results are more difficult to interpret and do not accurately represent market segments derived from the raw segmentation variables. Empirical evidence suggests that factor-cluster analysis does not outperform cluster analysis using raw data in terms of identifying the correct market segment structure. Therefore, it is generally discouraged to use factor-cluster analysis for market segmentation purposes.

7.5 Data Structure Analysis

Validation in market segmentation is inherently exploratory, as it is not possible to determine an optimal solution with a clear criterion. Instead, validation is often focused on assessing the reliability and stability of segmentation solutions through repeated calculations or modifications to the data or algorithm. This stability-based data structure analysis provides insights into the properties of the data and helps guide methodological decisions. It can indicate whether natural, distinct, and well-separated market segments exist in the data or not. If structure is present, data structure analysis can also assist in selecting an appropriate number of segments to extract. Ultimately, this approach aids in identifying the most useful segments for an organization when exploring alternative solutions.

7.5.1 Cluster Indices

In market segmentation analysis, cluster indices are commonly used to provide guidance and insights for critical decisions, such as determining the number of market segments to extract. There are two types of cluster indices: internal cluster indices and external cluster indices.

Internal cluster indices are calculated based on a single market segmentation solution and utilize information within that solution. They offer insights into the characteristics of segments within the solution. For example, the sum of distances between pairs of segment members can indicate the similarity of members within the same segment, with lower values suggesting more similarity and more attractive segments.

External cluster indices, on the other hand, require another segmentation as additional input and measure the similarity between two segmentation solutions. However, the correct assignment of members to segments is typically unknown in real consumer data. External cluster indices are often used in artificially generated data where the correct assignments are known. Repeated calculations of market segmentation can be compared to assess stability and consistency of segment extraction. Common measures of similarity for external cluster indices include the Jaccard index, the Rand index, and the adjusted Rand index.

By utilizing cluster indices, analysts can gain insights into the characteristics of segments and compare different segmentation solutions to aid in decision-making during market segmentation analysis.

7.5.1.1 Internal Cluster Indices

Internal cluster indices are used to assess the compactness and separation of market segments within a single segmentation solution. These indices rely on distance measures between observations or groups of observations. One common internal cluster index is the sum of within-cluster distances, which measures how similar members within the same segment are. The scree plot, derived from this index, can help select the number of market segments by identifying an "elbow" where significant decreases in within-cluster distances occur. Another index, the Ball-Hall index, corrects for the monotonic decrease of the sum of within-cluster distances by dividing it by the number of segments.

Additionally, internal cluster indices can capture the aspect of dissimilarity between segments. For example, the Ratkowsky and Lance index calculates the weighted distances between centroids of segments. Various combinations of compactness and separation measures are used to derive internal cluster indices, providing insights into segment characteristics. The Calinski-Harabasz index is often recommended and compares the ratio of between-cluster distances to within-cluster distances.

There are several internal cluster indices available in software packages, allowing data analysts to explore and evaluate different aspects of market segmentation solutions. While internal cluster indices may not provide definitive guidance in consumer data with natural market segments, they are still valuable for understanding the properties of segmentations. In such cases, external cluster indices and stability analysis can offer additional insights.

7.5.1.2 External Cluster Indices

External cluster indices evaluate market segmentation solutions using additional external information beyond the segmentation itself. These indices are valuable when the true segment structure is known, although it is typically only available for artificially generated data. In the absence of true segment information, repeated calculations or variations of the original data can be used as external information. A common challenge when comparing segmentation solutions is label switching, where the arbitrary labels assigned to segments can vary. To address this, external indices focus on the consistency of segment assignments between pairs of consumers rather than specific segment labels.

The Jaccard index and the Rand index are two widely used external cluster indices. The Jaccard index measures the similarity between two solutions based on pairs of consumers assigned to the same segment, while the Rand index considers all four possibilities of pairwise segment assignments. Both indices range from 0 to 1, with 0 indicating complete dissimilarity and 1 indicating identical solutions. However, these indices can be challenging to interpret due to their dependence on segment sizes.

To address the issue of chance agreement given segment sizes, the adjusted Rand index was proposed by Hubert and Arabie. This index incorporates a correction for chance agreement and provides a more reliable measure of similarity. The adjusted Rand index is commonly used in the resampling-based data structure analysis approach.

In R, various packages offer functions to calculate external cluster indices, including the Jaccard index, Rand index, and adjusted Rand index. These indices play a crucial role in assessing the quality and stability of market segmentation solutions.

7.5.2 Gorge Plots

A simple method to assess the separation of segments is by examining the distances between each consumer and the segment representatives. The similarity of a consumer to a segment representative can be measured using the exponential function of the distance, where a hyperparameter γ controls the translation of distance differences into similarity differences. These similarity values range between 0 and 1, with a sum of 1 for each consumer across all segment representatives. The similarity values can be visualized using gorge plots, silhouette plots, or shadow plots.

Gorge plots, which are histograms of similarity values for each segment, provide a visual representation of the distribution of similarities. High similarity values indicate consumers who are close to the segment representative, while low values indicate consumers who are far away. In the case of distance-based methods, high similarities indicate proximity to the centroid, while in model-based methods, high similarities indicate a high probability of segment membership. Gorge plots ideally show a "gorge" shape with peaks on both sides, indicating well-separated segments.

Gorge plots are particularly useful for assessing the quality of market segmentation solutions. By examining the plots, it is possible to determine if segments are well-separated or if consumers exhibit similarities to multiple segments. However, generating and inspecting gorge plots for various numbers of segments can be time-consuming and may not account for sample randomness. To overcome these limitations, stability analysis can be conducted at the global or segment level, providing a more robust evaluation of the segmentation solution.

7.5.3 Global Stability Analysis

Resampling methods offer a valuable alternative approach to analyzing data structures in market segmentation. These methods involve generating new data sets using resampling techniques and extracting multiple segmentation solutions from them. By comparing the stability of these solutions across repeated calculations, insights into the global stability of the market segmentation can be gained. This approach is particularly useful when dealing with consumer data that lacks distinct, well-separated market segments or is unstructured. Resampling methods help identify whether natural segments, reproducible segments, or

artificially constructed segments exist in the data. The results of global stability analysis aid in determining the most suitable number of segments to extract from the data and guide decision-making in strategic planning and marketing. By considering the stability and reproducibility of segmentation solutions, organizations can gain a deeper understanding of the underlying data structure and make informed decisions for effective market segmentation strategies.

7.5.4 Segment Level Stability Analysis

When conducting global stability analysis for market segmentation, it is important to consider not only the overall stability of the segmentation solutions but also the stability of individual market segments within those solutions. Choosing a segmentation solution based solely on global stability may overlook the presence of highly stable and valuable individual segments. Therefore, it is advisable to assess segment-level stability in addition to global stability to avoid prematurely discarding solutions that contain interesting and relevant segments. Since most organizations typically aim to target a single segment, evaluating both global and segment-level stability helps ensure that the chosen segmentation solution includes a stable and meaningful target segment.

7.5.4.1 Segment Level Stability Within Solutions (SLSW)

Dolnicar and Leisch (2017) propose a method for assessing market segmentation solutions that focuses on segment-level stability rather than evaluating the overall stability of the entire solution. This approach ensures that a segmentation solution containing at least one suitable market segment is not discarded prematurely. For organizations that only need to target one segment, identifying a highly stable segment is crucial for their survival and competitive advantage. Segment-level stability within solutions (SLSW) measures the consistency of identifying market segments with the same characteristics across multiple calculations of segmentation solutions using bootstrap samples. By examining the stability of individual segments, interesting and stable niche markets can be identified within a segmentation solution, even if other segments are unstable. This assessment is done using the Jaccard index and generating boxplots to visualize the segment-level stability. It is important to consider segment-level stability within solutions in addition to global stability to avoid disregarding solutions that contain valuable individual segments. Analyzing the stability of segments within a solution provides insights into the suitability of different market segments for targeting, which is especially relevant in multi-dimensional datasets where a quick visual inspection is not feasible.

7.5.4.2 Segment Level Stability Across Solutions (SLSA)

The criterion of segment level stability across solutions (SLSA) proposed by Dolnicar and Leisch (2017) focuses on the re-occurrence of market segments across different market segmentation solutions with varying numbers of segments. High values of SLSA indicate the presence of naturally occurring segments in the data, which are more desirable for organizations as they reflect actual market characteristics without the need for artificial segment creation. To assess SLSA, a series of partitions with different numbers of segments is analyzed, and the stability of each segment is measured using entropy. The resulting SLSA plot visualizes the movement of segment members across different segmentation solutions, highlighting stable segments that persist across solutions. This information helps identify natural segments and avoid segments that are artificially created during the extraction process. The SLSA values can be used as numeric indicators of segment stability. In the case of the artificial mobile phone data set, the SLSA plot confirms the stability of the high-end mobile phone segment while showing the subdivision of the other segments as the number of segments increases. Overall, SLSA provides valuable insights into the stability and natural occurrence of segments within market segmentation solutions.

7.6 Step 5 Checklist

Task	Who is responsible?	Completed?
Pre-select the extraction methods that can be used given the properties of your data.		☐
Use those suitable extraction methods to group consumers.		☐
Conduct global stability analyses and segment level stability analyses in search of promising segmentation solutions and promising segments.		☐
Select from all available solutions a set of market segments which seem to be promising in terms of segment-level stability.		☐
Assess those remaining segments using the knock-out criteria you have defined in Step 2.		☐
Pass on the remaining set of market segments to Step 6 for detailed profiling.		☐