Steps of Market Segmentation Analysis (1, 2, 3, 4, 5 & 6)

By Mohammed Ahmed Hussain

(Batch-SB-23-9-5 MLI)

1. Deciding (not) to Segment

Implications of Committing to Market Segmentation

Market segmentation is a powerful marketing strategy, but it's not always the best choice. Before diving into it, organizations must understand its indications. Market segmentation requires long-term commitment and substantial investments. It involves costs like research, surveys, and product/advertising design. Cahill suggests only segmenting if it leads to more profitable marketing. This strategy often demands changes like new product development, pricing adjustments, and organizational restructuring to cater to different market segments. Croft recommends organizing around market segments rather than products. Due to these significant indications, the decision to explore market segmentation should be made at the highest executive level and consistently communicated throughout the organization.

Implementation Barriers

Several books on market segmentation discuss successful implementation in organizations and identify various barriers to its effective execution.

1. **Senior Management Barriers**: Lack of leadership, proactive involvement, and commitment from senior leadership can undermine market segmentation success.
2. **Organizational Culture Barriers**: Issues such as a lack of market or consumer orientation, resistance to change, poor communication, short-term thinking, and office politics can hinder market segmentation success.
3. **Training and Expertise**: The lack of understanding between senior management and the segmentation team regarding the fundamentals and consequences of market segmentation can lead to failure.
4. **Objective Restrictions**: Limitations like financial resources and the inability to make necessary structural changes can be obstacles.
5. **Process-related Barriers**: These include unclear segmentation objectives, lack of planning, unstructured processes, undefined responsibilities, and time pressure.

6. **Operational Acceptance**: Management might resist techniques they don't understand, so presenting market segmentation analysis in an easy-to-understand way is important.

These barriers should be identified early in a market segmentation study and addressed proactively. If they cannot be resolved, organizations may need to reconsider pursuing market segmentation as a strategy. But if they are, dedication, patience, and a resolute sense of purpose are essential for successful implementation.

Step 1 Checklist

This checklist includes not only tasks, but also a series of questions which need to be answered in the affirmative, serve as knock-out criteria. For example, if an organization is not market-oriented, even the finest of market segmentation analyses cannot be successfully implemented. The checklist includes –

- Ask if the organization's culture is market oriented. If yes, proceed. If not, seriously consider not to proceed.
- Ask if the organization is genuinely willing to change. If yes, proceed. If not, seriously consider not to proceed.
- Ask if the organization takes a long-term perspective. If yes, proceed. If not, seriously consider not to proceed.
- Ask if the organization is open to new ideas. If yes, proceed. If not, seriously consider not to proceed.
- Ask if communication across organizational units is good. If yes, proceed. If not, seriously consider not to proceed.
- Ask if the organization is in the position to make significant (structural) changes. If yes, proceed. If not, seriously consider not to proceed.
- Ask if the organization has sufficient financial resources to support a market segmentation strategy. If yes, proceed. If not, seriously consider not to proceed.
- Secure visible commitment to market segmentation from senior management.
- Secure active involvement of senior management in the market segmentation analysis.
- Secure required financial commitment from senior management.
- Ensure that the market segmentation concept is fully understood. If it is not: conduct training until the market segmentation concept is fully understood.
- Ensure that the implications of pursuing a market segmentation strategy are fully understood. If they are not: conduct training until the implications of pursuing a market segmentation strategy are fully understood.
- Put together a team of 2-3 people (segmentation team) to conduct the market segmentation analysis.
- Ensure that a marketing expert is on the team.

- Ensure that a data expert is on the team.
- Ensure that a data analysis expert is on the team.
- Set up an advisory committee representing all affected organizational units.
- Ensure that the objectives of the market segmentation analysis are clear.
- Develop a structured process to follow during market segmentation analysis.
- Assign responsibilities to segmentation team members using the structured process.
- Ensure that there is enough time to conduct the market segmentation analysis without time pressure.

2. Specifying the Ideal Target Segment

Segmentation Evaluation Criteria

The third step of market segmentation analysis depends heavily on user input and involvement throughout the process. This involvement shouldn't be limited to just the initial phase or the final development of a marketing mix but should extend across various stages of the analysis.

In Step 2 of the process, the organization plays a crucial role by establishing two sets of segment evaluation criteria. The first set, known as "knock-out criteria", outlines the non-negotiable features that segments must possess to be considered for targeting. The second set, "attractiveness criteria" is used to assess the relative appeal of the remaining segments that meet the knock-out criteria.

Several criteria for segment evaluation are suggested in the literature, including factors like size, growth, profitability, accessibility, and compatibility with the company's strengths. These criteria are used to determine which segments are most attractive.

The knock-out criteria are essential and not open to negotiation, whereas the attractiveness criteria offer a broader list from which the segmentation team can choose. The team also needs to weigh the relative importance of each attractiveness criterion.

This process ultimately guides the selection of target segments in Step 8 of the segmentation analysis.

Knock Out Criteria

Knock-out criteria are essential conditions used to determine whether market segments resulting from segmentation analysis are eligible for assessment using segment attractiveness criteria. These criteria, proposed by Kotler in 1994 and expanded upon by other authors, include the following:

1. Homogeneity: Segment members must be similar to each other.
2. Distinctiveness: Segment members should be significantly different from members of other segments.
3. Size: The segment must have a sufficient number of consumers to justify customizing the marketing mix for them.
4. Alignment with Organizational Strengths: The organization should have the capability to meet the needs of segment members.
5. Identifiability: Segment members should be identifiable in the marketplace.

6. Reachability: There should be a means of contacting segment members to make the customized marketing mix accessible to them.

These knock-out criteria are crucial and should be well understood by senior management, the segmentation team, and the advisory committee. While most of them do not require further specification, the specific minimum size for a viable target segment should be defined.

Attractiveness Criteria

In addition to the knock-out criteria, there are various segment attractiveness criteria. These criteria provide a wide range of factors for the segmentation team to consider when determining which ones are most relevant to their specific situation.

Unlike knock-out criteria, attractiveness criteria are not binary; segments are not simply classified as either complying or not complying with them. Instead, each market segment is assessed and rated in terms of how attractive it is with respect to each specific criterion. The collective attractiveness across all these criteria ultimately influences whether a market segment is chosen as a target segment in the later stages of market segmentation analysis.

Implementing a Structured Process

In the field of market segmentation, following a structured process is beneficial when evaluating market segments. A popular structured approach involves using a segment evaluation plot, where segment attractiveness is plotted along one axis, and organizational competitiveness is plotted along the other. The values for segment attractiveness and organizational competitiveness are determined by the segmentation team. It's important to note that there's no one-size-fits-all set of criteria, so these factors need to be agreed upon by the team. To do this, a wide range of potential criteria should be explored. It is advisable to limit the number of factors to no more than six for simplicity.

This process is best undertaken by a team primarily responsible for market segmentation analysis, they can propose initial criteria and discuss them with an advisory committee consisting of representatives from various organizational units. Including representatives from different units is essential because they bring unique perspectives. Although the segment evaluation plot can't be completed in Step 2 of the market segmentation analysis because no segments are available yet, selecting attractiveness criteria early in the process is highly beneficial. It ensures that relevant information is collected during data collection and streamlines the selection of a target segment in Step 8.

By the end of this step, the market segmentation team should have a list of around six segment attractiveness criteria, each with a weight indicating its importance to the organization compared to other criteria. The typical method for weighing involves team members distributing 100 points across the criteria, with allocations being negotiated until agreement is reached. Seeking approval from the advisory committee, which represents various organizational units and perspectives, is also recommended.

Step 2 Checklist

- Convene a segmentation team meeting.
- Discuss and agree on the knock-out criteria of homogeneity, distinctness, size, match, identifiability, and reachability. These knock-out criteria will lead to the automatic elimination of market segments which do not comply (in Step 8 at the latest).
- Present the knock-out criteria to the advisory committee for discussion and (if required) adjustment.
- Individually study available criteria for the assessment of market segment attractiveness.
- Discuss the criteria with the other segmentation team members and agree on a subset of no more than six criteria.
- Individually distribute 100 points across the segment attractiveness criteria you have agreed upon with the segmentation team. Distribute them in a way that reflects the relative importance of each attractiveness criterion.
- Discuss weightings with other segmentation team members and agree on a weighting.
- Present the selected segment attractiveness criteria and the proposed weights assigned to each of them to the advisory committee for discussion and (if required) adjustment.

3. Collecting Data

Segmentation Variables

In market segmentation, observed data forms the basis for both commonsense and data-driven approaches. This data is used to identify or create market segments and later to describe these segments in detail. Commonsense segmentation typically involves using one single characteristic, such as gender, as the variable. Each consumer is represented as a row in a data table. Descriptor variables, like age and benefits sought, are used to describe the segments.

In contrast, data-driven segmentation utilizes multiple segmentation variables. These variables are employed to identify naturally existing or artificially created market segments that are valuable to the organization. Descriptor variables are still used to describe these segments in detail. The quality of observed data is essential in both cases. For commonsense segments, data quality is critical for correctly assigning individuals to the correct segment and for accurately describing the segments. In data-driven segmentation, data quality determines the quality of the extracted segments and their descriptions, which are vital for product customization, pricing strategies, distribution channels, and advertising.

Observed data for segmentation studies can come from various sources, including surveys, observations like scanner data, or experimental studies. While surveys are common, they may not always reflect actual consumer behavior, especially when behavior is socially desirable.

Segmentation Criteria

Before actually extracting market, segments or collecting data for segmentation, an organization must make a critical decision: choosing the segmentation criterion. The term "segmentation criterion" is broader than "segmentation variable". It can also refer to a specific construct, such as benefits sought. This decision can't be easily delegated because it requires prior knowledge about the market. The most common segmentation criteria include geographic, sociodemographic, psychographic, and behavioral factors.

Differences between consumers that are most relevant for market segmentation often revolve around factors like profitability, bargaining power, product preferences, barriers to choose, and consumer interaction effects. With vast criteria available, it can be challenging to choose the best one. Generally, the recommendation is to use the simplest approach that works for the product or service in question. In essence, the goal is to use what works effectively for the least cost.

Geographic Segmentation

Geographic information is considered one of the earliest and simplest segmentation criteria used for market segmentation as it relies on the consumer's place of residence as the primary criterion to create market segments. The advantage of geographic segmentation is that it allows for easy assignment of consumers to specific geographic units, making it simple to target communication messages and select relevant communication channels, such as local newspapers and radio stations, to reach these geographic segments. However, the key disadvantage of geographic segmentation is that living in the same area or country doesn't necessarily imply that people share other relevant characteristics, such as their product preferences or the benefits they seek when making a purchase. People in the same geographic region can have diverse tastes and preferences, making it challenging to identify common characteristics solely based on location.

Despite its limitations, geographic segmentation has experienced a resurgence in international market segmentation studies that aim to create market segments across geographic boundaries.

Socio-Demographic Segmentation

Socio-demographic segmentation criteria, which typically include factors such as age, gender, income, and education, can be valuable in certain industries. For example, they are useful in luxury goods (associated with high income), cosmetics (associated with gender, even when targeting both men and women), baby products (linked to gender), retirement villages (linked to age), and tourism resort products (related to the presence of small children).

Similar to geographic segmentation, socio-demographic criteria have the advantage of easily determining segment membership for each consumer. In some cases, socio-demographics can provide explanations for specific product preferences (e.g., having children influences the choice of a family vacation), but in many instances, socio-demographic factors are not the primary drivers of product preferences.

Psychographic Segmentation

Psychographic segmentation categorizes people based on psychological criteria like beliefs, interests, preferences, aspirations, or the benefits they seek when making a purchase. These criteria are more intricate than geographic or socio-demographic ones because it's challenging to capture the psychological dimension with a single characteristic.

It often employs multiple segmentation variables to represent different psychological aspects, such as various travel motives or perceived risks during vacations. This approach is advantageous as it offers deeper insights into the underlying reasons for consumer behavior. For instance, travelers motivated by cultural exploration are more likely to choose destinations rich in cultural experiences. However, psychographic segmentation can be more complex in determining consumer segment memberships. Its effectiveness relies on the quality and reliability of the measures used to capture the psychological dimensions of interest.

Behavioral Segmentation

Behavioral Segmentation involves grouping individuals based on their actual behaviors or reported behaviors, such as prior product experience, purchase frequency, spending amounts, and information-seeking habits. This approach is advantageous because it uses the behavior of interest as the primary basis for segmenting individuals. Behavioral segmentation, when based on actual behavior, is particularly powerful, as it directly reflects consumer actions. However, obtaining behavioral data may be challenging, especially when trying to include potential customers who have not previously interacted with the product or brand, in contrast to existing customers.

Data from Survey Studies

Most market segmentation analyses are based on survey data. Survey data is cheap and easy to collect, making it a feasible approach for any organization. But survey data – as opposed to data obtained from observing actual behavior – can be contaminated by a wide range of biases. Such biases can, in turn, negatively affect the quality of solutions derived from market segmentation analysis. A few key aspects that need to be considered when using survey data are discussed below.

Choice of Variables

Careful selection of variables in market segmentation is crucial for the quality of the segmentation solution. In data-driven segmentation, it's essential to include all variables relevant to the segmentation criterion while avoiding unnecessary ones. Unnecessary variables can lead to longer, tedious surveys, causing respondent fatigue and lower data quality. They can also complicate the segmentation process and divert algorithms from identifying the correct solution, known as noisy or masking variables.

To avoid noisy variables, it's crucial to collect necessary and unique data while refraining from including redundant questions. Redundant questions, stemming from traditional psychometric principles, can hinder segmentation algorithms' ability to identify correct solutions. A good questionnaire development process involves both qualitative exploratory research and quantitative survey research to ensure all crucial variables are included.

Response Options

The answer options provided to survey respondents can impact the suitability of data for segmentation analysis. Binary data, where respondents choose one of two options, can be represented as 0s and 1s and are ideal for segmentation analysis. Nominal variables, where respondents select one option from a list of unordered categories (e.g., occupation), can be transformed into binary data. Metric data, generated by numerical responses like age or nights stayed at a hotel, are well-suited for segmentation analysis as they allow various statistical procedures, including distance measurement. However, ordinal data is common in surveys, where respondents use a limited number of ordered answer options (e.g., agreement scales). The issue with ordinal data is that the distance between adjacent options is undefined, making standard distance measures unsuitable.

Preferably, surveys should use binary or metric response options when meaningful to the question. Visual analogue scales, such as slider scales, are a useful option for capturing fine nuances while still generating metric data. Binary response options have been shown to perform well in many contexts, particularly when formulated in a level-free way.

Response Styles

Survey data can be affected by response biases and styles, which can lead to inaccurate segmentation results. Response biases are systematic tendencies in how respondents answer questions unrelated to the item's content. If a bias is consistent over time and across various survey questions, it becomes a response style.

Common response styles include using extreme answer options, choosing the midpoint, or agreeing with all statements. These response styles can impact segmentation results because they are often indistinguishable from genuine beliefs. To ensure accurate segmentation, it is essential to minimize the risk of capturing response styles during data collection. In cases where segments with potential response-style-influenced patterns emerge, additional analyses may be needed to confirm or exclude this possibility.

Sample Size

The success of market segmentation analysis depends on various factors, including sample size and data quality. Sample size is critical in segmentation analysis, with recommendations varying based on the algorithm and scale used. Some studies suggest that a sample size of at least five times the number of segmentation variables is necessary. Additionally, sample size requirements can be affected by market characteristics, such as the number and size of market segments and the extent of their overlap.

Data quality also plays a crucial role in segmentation analysis. High-quality data should include all necessary items, exclude unnecessary items, avoid correlated items, and have responses of good quality. Binary or metric response options are preferred to make the analysis easier. Response styles and biases should be minimized to prevent the misinterpretation of segments.

Data from Internal Sources

Organizations have access to significant amounts of internal data that can be utilized for market segmentation analysis. This data, such as scanner data in grocery stores, booking data from airline loyalty programs, and online purchase records, is advantageous because it reflects actual consumer behavior, in contrast to self-reported data that may be influenced by memory issues and response biases. Another benefit is that this data is often automatically generated, requiring minimal additional effort to collect. However, a potential drawback is that internal data may be biased toward existing customers, lacking information about potential future customers whose consumption patterns may differ from current customers.

Data from Experimental Studies

Experimental data, whether derived from field or laboratory experiments, can be a valuable source for market segmentation analysis. Such data can be generated by studying how people respond to various stimuli, such as advertisements, and using their responses as segmentation criteria. These experiments yield information about the impact of specific attributes on consumer choices, which can also be employed as segmentation criteria.

Step 3 Checklist

- Convene a market segmentation team meeting.

- Discuss which consumer characteristics could serve as promising segmentation variables.
- These variables will be used to extract groups of consumers from the data.
- Discuss which other consumer characteristics are required to develop a good understanding of market segments. These variables will later be used to describe the segments in detail.
- Determine how you can collect data to capture both the segmentation variables and the descriptor variables most validly.
- Design data collection carefully to keep data contamination through biases and other sources of systematic error to a minimum.
- Collect data.

4. Exploring Data

A first glimpse at the Data

The role of data exploration in identifying measurement levels, examining the distributions of variables, and assessing dependencies between variables. The introduction of a travel motives dataset for demonstration, which includes various attributes related to travel behavior. An example summary of the dataset, showcasing data on gender, age, income, and a modified income variable (Income2). An explanation that the dataset contains missing values (NAs) in the income variables.

In essence, data exploration is a crucial step in understanding the dataset and guiding the selection of suitable segmentation methods. The text uses the travel motives dataset to illustrate how to examine and understand the data's characteristics.

Data Cleaning

Data cleaning involves verifying the correctness of recorded values and ensuring consistent labels for categorical variables. It includes checking for implausible or erroneous values, which can indicate errors during data collection or data entry. The importance of checking the permissible values for categorical variables, such as gender, to ensure they match the expected categories. The example of an Australian travel motives dataset is used, and the text mentions that data cleaning is not required for variables like "Gender" and "Age." The text explains that issues related to the order of categories in a categorical variable can arise, especially when using functions like read.csv() or read.table() in R. An example is provided for re-ordering the categories of the "Income2" variable in R, ensuring that the levels are correctly re-ordered. The text emphasizes the benefits of documenting and recording every step of data cleaning and analysis to ensure reproducibility and consistency, making it easier for other analysts to replicate the process. The "save()" function is mentioned as a way to save the cleaned data frame, allowing for easy retrieval in future R work sessions using the "load()" function.

Descriptive Analysis

Descriptive numeric and graphic representations of data are crucial for gaining insights and avoiding misinterpretation of results. In R, the summary() command provides a numeric summary of data, including the range, quartiles, mean for numeric variables, and frequency counts for categorical variables. It also reports the number of missing values for each variable.

Graphical Tools for Numeric Data:

- Histograms: Visualize the distribution of numeric variables, showing how often observations occur within specific value ranges. Histograms help identify whether a variable's distribution is unimodal, symmetric, or skewed.
- Boxplots: Boxplots provide a summary of unimodal distributions, highlighting the minimum, first quartile, median, third quartile, and maximum values. They can reveal skewness in the data.
- Outlier Handling: Most statistical software, including R, restricts the length of whiskers in boxplots to avoid extreme outliers. Outliers are depicted as individual circles to preserve their information.
- Dot Chart Visualization: Dot charts are used to illustrate the percentage of agreement with various travel motives in the Australian travel motives dataset. The chart quickly reveals the range of agreement levels for each motive, indicating the heterogeneity of responses.

The graphical inspection of the data suggests that the Australian travel motives variables are suitable as segmentation variables due to differences in the importance attributed to different motives among respondents.

Pre Processing

Categorical Variables

Merging levels is useful when the original categories of a categorical variable are too numerous or unbalanced. For example, the income variable had several fine-grained categories, but some categories had very few respondents (e.g., only 10 to 15 people in the top income categories). By merging these categories into larger, more balanced groups, a new variable (e.g., "Income2") is created with improved data distribution.

Some data analysis methods assume numeric data measured on comparable scales. Ordinal data can be converted to numeric if it's reasonable to assume that the distances between adjacent scale points are approximately equal. For example, income categories can be reasonably transformed into numeric values because they represent ranges of equal length. Multi-category scales, such as Likert scales (e.g., "STRONGLY DISAGREE" to "STRONGLY AGREE"), can also be converted to numeric if the assumption of equal distances between answer options can be justified. However, this assumption may not always hold due to response styles. Binary answer options, on the other hand, are less prone to capturing response styles and do not require pre-processing. Binary variables can be converted to numeric with 0 and 1 values.

In the provided example, the travel motives (variables) are converted into a numeric matrix, with 0 and 1 representing "NO" and "YES" responses, respectively.

Numeric Values

The range of values of a segmentation variable can affect its relative influence on distance-based methods used for segment extraction. For example, if one segmentation variable is binary (0 or 1) to indicate whether a tourist likes to dine out and another variable indicates expenditure in dollars per person per day (ranging from 0 to $1000), a difference of one dollar in expenditure is treated as equally important as the difference in dining preferences. This can lead to unbalanced influence in segment extraction.

To balance the influence of segmentation variables, it's common to standardize them, which transforms variables to a common scale. The default standardization method subtracts the empirical mean ($\bar{x}$) from each data point and divides it by the empirical standard deviation (s).

Principal Component Analysis (PCA)

PCA transforms a dataset into a new set of variables (principal components) that are uncorrelated and ordered by importance. The first principal component contains the most variability, the second principal component the second most variability, and so on. PCA retains the data space but views it from a different angle.

PCA works with the covariance or correlation matrix of numeric variables. If variables have similar data ranges and are measured on the same scale, it doesn't matter whether the covariance or correlation matrix is used. However, when data ranges differ, the correlation matrix is preferred. Principal components can be used to project high-dimensional data into lower dimensions for visualization and analysis. The importance of each principal component is assessed using standard deviations. The first few principal components capture the most variation in the data.

Principal components can be visualized in scatter plots. In this case, the first and second principal components are often used. The rotation matrix specifies how the original variables contribute to each principal component.

The text demonstrates PCA using the Australian travel motives dataset, focusing on the second and third principal components for a perceptual map. Interpretations are made based on the loading patterns of original variables on principal components.

PCA is sometimes considered to reduce the number of segmentation variables before extracting market segments from consumer data. However, this approach can be problematic as it replaces original variables with a subset of principal components, potentially leading to a different segmentation space. While using PCA for segmentation variables is not recommended, it can be useful for exploring data and identifying highly correlated variables, which can then be reduced to achieve dimensionality reduction while still working with the original variables collected.

Step 4 Checklist

- Explore the data to determine if there are any inconsistencies and if there are any systematic contaminations.
- If necessary, clean the data.
- If necessary, pre-process the data.
- Check if the number of segmentation variables is too high given the available sample size. You should have information from a minimum of 100 consumers for each segmentation variable.
- If you have too many segmentation variables, use one of the available approaches to select a subset.
- Check if the segmentation variables are correlated. If they are, choose a subset of uncorrelated segmentation variables.
- Pass on the cleaned and pre-processed data to Step 5 where segments will be extracted from it.

5. Extracting Segments

Grouping Consumers

Market segmentation analysis is exploratory and often deals with unstructured consumer data. Consumer preferences are usually distributed across a plot rather than forming clear groups. The results of segmentation analysis are strongly influenced by the underlying data and the chosen extraction algorithm. Different segmentation methods make different assumptions about the structure of segments, and these assumptions shape the segmentation solution. Many segmentation methods are derived from cluster analysis, where market segments correspond to clusters. The choice of a suitable clustering method should match the data's characteristics and the researcher's requirements.

An illustrative example is provided with two different algorithms (k-means and single linkage hierarchical clustering) applied to a dataset with two spiraling segments. K-means fails to identify the spirals due to its focus on compact, equally sized clusters, while single linkage clustering correctly identifies them. The data's structure and the algorithm's tendency are critical. Well-structured data with distinct segments may make the choice of algorithm less important.

In contrast, for less structured data, the algorithm's tendency significantly influences the outcome. No single method is best for all situations; each method has advantages and disadvantages. Two main groups of methods are discussed: distance-based methods and model-based methods. Distance-based methods focus on finding groups of similar observations, while model-based methods formulate stochastic models for segments. Some algorithms aim to achieve multiple goals in one step, such as variable selection during segmentation. To select the appropriate extraction algorithm, data characteristics (size, scale level of variables, special structures) and segment characteristics (commonalities and differences) must be considered. For binary segmentation variables, the treatment of 0s and 1s (asymmetric or symmetric) can impact the choice of method and the resulting segments.

Distance based Methods

The data set contains information on seven tourists and the percentage of time they spend on three vacation activities: BEACH, ACTION, and CULTURE. Each row represents a tourist, and each column represents an activity. The goal is to group tourists with similar vacation activity patterns. For example, some tourists prefer the beach, while others enjoy both action and culture. To measure the similarity or dissimilarity between tourists, distance measures are used. These measures calculate the distance between two tourists based on their vacation activity profiles.

A distance measure should be symmetric, meaning that the distance from tourist A to tourist B is the same as from tourist B to tourist A. The distance between a tourist and themselves should be zero. The combined distance from tourist A to tourist B, via an intermediate tourist C, should be at least as long as going directly from A to B.

- Euclidean Distance: Measures the straight-line distance between two points in multi-dimensional space. It considers all dimensions of the vectors.
- Manhattan or Absolute Distance: Calculates the distance assuming travel on a grid (e.g., city streets). It also considers all dimensions.
- Asymmetric Binary Distance: Applies only to binary vectors (0s and 1s). It measures the proportion of common 1s over dimensions where at least one vector contains a 1.

Euclidean distance is commonly used in market segmentation analysis because it considers all dimensions and is applicable to both metric and binary variables.

Hierarchical Methods

Hierarchical clustering is an intuitive method for grouping data, mimicking how humans might approach dividing a set of observations (in this case, consumers) into segments. It provides a hierarchy of partitions from one large segment containing all consumers to n segments, each with one consumer.

Divisive Hierarchical Clustering: Starts with all observations in one segment and recursively divides them into smaller segments until each consumer has their own segment. Agglomerative Hierarchical Clustering: Starts with each consumer in their own segment and progressively merges the two closest segments until all observations are in one large segment. Distance measures are used to determine the dissimilarity between groups of observations (segments). They are specified by choosing a distance measure for individual observations and a linkage method to calculate distances between groups.

Linkage Methods:

- Single Linkage: The distance between two groups is determined by the closest observations from each group.
- Complete Linkage: The distance between two groups is based on the farthest observations from each group.

- Average Linkage: The mean distance between observations in two groups.

Ward clustering is based on squared Euclidean distances and minimizes the weighted squared Euclidean distance between cluster centers. It's a popular alternative for hierarchical clustering, particularly when dealing with continuous variables. The results of hierarchical clustering are often presented as a dendrogram, which is a tree diagram showing the sequence of partitions and distances between clusters. It can be used to visualize the hierarchy and select the number of segments.

The characteristics of segments are assessed by analyzing column-wise means within each cluster. A bar chart is used to visualize the differences between segment means and the total population. Segments are interpreted by comparing their characteristics to the total population.

Partitioning Methods

Hierarchical clustering methods are well-suited for analyzing small data sets containing up to a few hundred observations. They provide a detailed hierarchy of partitions. For larger data sets, dendrograms, which visualize hierarchical clustering results, become difficult to read and interpret. In cases where the data set contains more than 1000 observations, other clustering methods that aim to create a single partition become more suitable. These methods involve fewer distance calculations, making them computationally more efficient. Partitioning clustering algorithms, which aim to extract a specific number of segments, only require the computation of distances between consumers and the centers of the segments. This reduces the number of distance calculations compared to hierarchical clustering, especially for large data sets.

k-Means and k-Centroid Clustering

Partitioning clustering methods, such as k-means, aim to divide a set of observations (e.g., consumers) into subsets (market segments) based on their similarity. The representative of each market segment is known as the centroid. The k-means algorithm is a popular partitioning method that uses squared Euclidean distance to measure similarity. It divides consumers into k segments, where k is a user-specified value. The initial centroids are randomly selected, and consumers are assigned to the nearest centroid.

Five Steps of the K-Means Algorithm:

- Specify the desired number of segments, k.
- Randomly select k initial centroids.
- Assign each observation to the nearest centroid, forming an initial suboptimal segmentation.
- Recompute centroids while keeping cluster membership fixed.
- Repeat steps 3 and 4 until the segmentation converges.

Determining the optimal number of segments (k) can be challenging. One approach is to compare the sum of within-cluster distances across different k values. The "elbow" point in the scree plot often indicates the optimal k.

The choice of the optimal number of segments should be made based on the specific characteristics of the data, its purpose, and user needs. Stability analysis and other techniques can provide additional insights into the segmentation decision.

“Improved” k-Means

The k-means algorithm can be refined to produce better results. One common improvement is to replace the initial random selection of k data points with more strategic starting values. Randomly selecting k data points as initial centroids can lead to suboptimal outcomes. This is because these randomly chosen data points may end up being close to each other in the data space, which doesn't accurately represent the entire dataset. The risk of using randomly drawn starting points is that the k-means algorithm might converge to a local optimum rather than the global optimum.

A better approach is to initialize the algorithm with starting points that are evenly distributed across the entire data space. Such starting points provide a more representative view of the dataset. Steinley and Brusco conducted an extensive simulation study using artificial datasets with known structures. They compared 12 different strategies for initializing the k-means algorithm. The study's results suggested that the most effective approach was to randomly generate numerous starting points and then select the best set of starting points. The best starting points are those that closely represent the data; the total distance between segment members and their representatives should be minimal.

Hard Competitive Learning

Hard competitive learning, also known as learning vector quantization, differs from the k-means algorithm in how it extracts segments from data. Similar to k-means, hard competitive learning aims to minimize the sum of distances from each data point to its closest representative, which is typically referred to as a centroid or segment representative. The key difference lies in the process of achieving this minimization. K-means uses all consumers in the data set in each iteration to determine new segment representatives, whereas hard competitive learning randomly selects one consumer and moves that consumer's closest segment representative a small step toward the selected consumer. Due to these procedural differences, even if both algorithms are initialized with the same starting points, they may produce different segmentation solutions. It is also possible that hard competitive learning might find the globally optimal solution while k-means gets stuck in a local optimum or vice versa. Neither method is inherently superior to the other; they are simply different approaches to segmentation. The choice between them depends on the specific problem and data at hand. Hard competitive learning can be implemented in R using the cclust function with the method = "hardcl" option from the flexclust package.

Neural Gas and Topology Representing Networks

Neural gas, proposed by Martinetz et al. in 1993, is a variation of hard competitive learning. In this algorithm, not only the primary segment representative (centroid) is adjusted toward the randomly selected consumer, but also the location of the second closest segment representative is adjusted. However, the adjustment for the second closest representative is smaller.

Neural gas clustering can be performed in R using the cclust function with the method = "neuralgas" option from the flexclust package. An extension of neural gas clustering is the topology representing networks (TRN) introduced by Martinetz and Schulten in 1994. TRN uses a similar algorithm to neural gas but additionally tracks how often each pair of segment representatives is closest and second closest to a randomly drawn consumer. This information is used to construct a virtual map, placing "similar" representatives together. The segment neighborhood graph is generated based on this information.

The segment neighborhood graph is part of the default segment visualization functions of the flexclust package. It helps visualize the relationships between segment representatives. Neither neural gas nor topology representing networks are claimed to be superior to the k-means algorithm or hard competitive learning. They are different approaches to clustering, and their usage results in different market segmentation solutions. The choice of method depends on the specific needs and objectives of the analysis.

Self-Organizing Maps

Self-organizing maps, also known as self-organizing feature maps or Kohonen maps, are a variation of hard competitive learning. These maps position segment representatives (centroids) on a regular grid, such as a rectangular or hexagonal grid. The SOM algorithm is similar to hard competitive learning, where a random consumer is selected, and the closest representative to that consumer moves towards it. In addition, the grid neighbors of the closest representative also adjust their positions toward the selected consumer. This process is repeated multiple times with consumers selected randomly to adjust centroid locations. The adjustments become smaller over iterations until a final solution is reached.

SOMs provide an advantage in that the numbering of market segments is not random but aligns with the grid. However, the disadvantage is that the sum of distances between segment members and representatives can be larger than in other clustering algorithms due to the grid-imposed restrictions.

Neural Networks

Auto-encoding Neural Networks operate differently from traditional clustering methods. A popular method in this family of algorithms uses a single hidden layer perceptron. The neural network consists of three layers: the input layer, hidden layer, and output layer. The input layer receives data, and the output layer produces the network's response. In clustering, the output layer provides the same data as the input. The hidden layer has nodes named h1, h2, h3, with each node representing a weighted linear combination of input variables. The weights (αij) connecting input nodes to hidden nodes are depicted by arrows. The function fj ensures that the hidden nodes (hj) are between 0 and 1, with their sum equal to 1. The network predicts outputs (xˆi) as weighted combinations of hidden nodes using coefficients (βj i). Training the network involves adjusting the parameters (αij and βj i) to minimize the squared Euclidean distance between inputs and outputs for the training data (consumers). The network is named "auto-encoder" because it's trained to predict inputs as accurately as possible. When the number of hidden nodes is fewer than the available input nodes, the network learns to represent the data using segment representatives. Parameters connecting the hidden layer to the output layer are interpreted as segment representatives (centroids). Parameters linking the input layer to the hidden layer can be interpreted as membership values for consumers in different segments. Consumers with a high hj value near 1 belong to a specific segment. Unlike traditional clustering methods (e.g., k-means), which produce crisp segmentations where each consumer belongs to one segment, neural network clustering results in fuzzy segmentations. Membership values range from 0 (not a member of a segment) to 1 (exclusive membership in a segment), indicating partial membership in multiple segments. Various implementations of auto-encoding neural networks are available in R, with an example being the autoencode() function in the "autoencoder" package.

Hybrid approaches

Hierarchical clustering algorithms offer the advantage of not requiring the number of market segments to be specified in advance. They provide a visual representation of similarities between market segments through dendrograms. Standard implementations of hierarchical clustering can demand substantial memory capacity, limiting their application to large data sets.

Partitioning clustering algorithms have minimal memory requirements, making them suitable for segmenting large data sets. They require the number of market segments to be specified in advance. They do not allow for the tracking of changes in segment membership across solutions with different segment numbers because these solutions are not nested. The hybrid approach combines the strengths of both hierarchical and partitioning algorithms.

It begins by running a partitioning algorithm, which can handle data sets of any size. However, the initial partitioning algorithm generates a larger number of segments than required. The original data is then discarded, and only the centers of the resulting segments (centroids) and segment sizes are retained. These centroids and sizes are used as input for hierarchical cluster analysis, which can now handle the reduced data size effectively. The dendrogram produced by hierarchical clustering can assist in determining how many segments to extract, using a smaller, more manageable data set.

Two step Clustering

IBM SPSS implements a two-step clustering procedure, consisting of two stages: partitioning and hierarchical clustering. This procedure has been used in various application areas, such as segmenting mobile phone users, nature-based tourists, potential electric vehicle adopters, and travel-related risks.

Partitioning Clustering Step:

- In the first step, a partitioning clustering method (e.g., k-means) is applied with a larger number of clusters (k) than the actual number of market segments.
- The primary goal of this step is to reduce the data set's size by retaining only one representative member of each extracted cluster.
- This approach is sometimes referred to as vector quantization.

The results of the partitioning step are visualized using a neighborhood graph that shows cluster means as nodes. The neighborhood graph displays edges that represent the similarity between clusters and includes a scatter plot of the data with observations colored by cluster memberships and cluster hulls.

The representatives of the extracted clusters (centroids) and segment sizes serve as the input for the second step, which involves hierarchical clustering. A dendrogram is produced in this step, and the number of vertical lines in the dendrogram indicates the number of market segments in the data. The hierarchical clustering analysis cannot determine which consumers belong to which market segment since the original data was discarded.

To link the original data with the segmentation solution, the twoStep() function from the MSA package is used. It requires inputs like the hierarchical clustering solution, cluster memberships obtained in the partitioning step, and the desired number of segments (k). The number of segment members extracted from the two-step procedure should match the number of segment members generated for the artificial data set. The correctness of the segments can be confirmed by inspecting the plot of the data colored by segment memberships.

Bagged Clustering

Bagged clustering combines hierarchical and partitioning clustering methods with bootstrapping. Bootstrapping involves randomly drawing samples from the data with replacement, reducing dependence on specific data samples. The process starts with partitioning clustering of bootstrapped data sets, which do not have sample size restrictions. The original data and bootstrapped data sets are discarded, and only cluster centroids (representatives) are retained for the second step: hierarchical clustering.

Hierarchical clustering in the second step can suggest the ideal number of market segments by analyzing the dendrogram. Bagged clustering is suitable when standard algorithms may lead to suboptimal results or when the dataset is too large for traditional hierarchical clustering.

Bagged clustering involves five steps: bootstrapping, repeated partitioning clustering, creation of a derived data set with cluster centroids, hierarchical clustering, and final segmentation based on the dendrogram. Bagged clustering has been successfully applied in various domains, including tourism data segmentation. It can identify niche market segments and offer better solutions when standard algorithms may fail. It results in distinct market segments with varying sizes, providing insights into consumer preferences for various vacation activities and can uncover niche segments that other clustering algorithms might miss.

Model Based Methods

Model-based methods, pioneered by Wedel and Kamakura, are becoming increasingly popular in market segmentation analysis. These methods offer an alternative approach to segmenting data, allowing data analysts to explore multiple techniques for identifying market segments. They do not rely on distances or similarities to determine segment membership. They assume that the market segmentation has two general properties: specific segment sizes and unique characteristics for each segment. The exact values of these properties are unknown initially.

Model-based methods are based on finite mixture models, which consider the number of market segments to be finite. These models combine segment-specific characteristics (θ) with segment sizes (π) in a probabilistic framework. Parameter estimation is typically performed through maximum likelihood estimation or Bayesian methods. The goal is to find parameter values that make the observed data most likely to occur. After parameter estimation, probabilities of consumers belonging to specific segments are calculated. Consumers are assigned to segments based on these probabilities, with each consumer assigned to the segment with the highest probability.

Choosing the appropriate number of segments (k) is challenging because it is often unknown. Information criteria, such as AIC, BIC, and ICL, are used to guide the selection of the number of segments. These criteria balance goodness of fit and model complexity, helping to avoid overly complex models.

Benefits of Finite Mixture Models:

Finite mixture models can capture complex segment characteristics and can be extended in various ways. They offer a flexible approach to market segmentation that can accommodate different structures and characteristics for each segment.

Finite Mixture of Distributions

In this scenario, model-based clustering is applied without the use of additional independent variables (x). The objective is to fit a statistical distribution to a single variable of interest, y. This approach can be compared to distance-based methods used in market segmentation, where the same segmentation variables, such as consumer activities during a vacation, are employed without incorporating additional information about consumers. The choice of the statistical distribution function f() depends on the measurement level or scale of the segmentation variable y. Different types of variables may require different distribution functions based on their characteristics.

Normal Distributions

Model-based clustering is applied using multivariate normal distributions when the segmentation variables are continuous or metric in nature. Multivariate normal distributions are suitable because they can capture the covariance between multiple variables, making them a valuable choice for data where variables are correlated. In biology, physical measurements on humans, such as height, arm length, leg length, or foot length, can be well approximated by multivariate normal distributions. In business, prices in markets with multiple players can be modeled using (log-)normal distributions.

The multivariate normal distribution is mathematically represented with parameters for the mean and covariance matrix. In a segmentation context, if there are 'p' segmentation variables, each segment has a segment-specific mean vector (μh) of length 'p' and a covariance matrix (Σh) that describes the covariance structure between the variables. The number of parameters to estimate for each segment is 'p + p(p + 1)/2.'

The package 'mclust' is utilized for this purpose, which fits models for different numbers of segments using the EM algorithm. The best model is selected based on the Bayesian Information Criterion (BIC) and may include different covariance matrix structures for the segments. Various covariance matrix models are introduced, including spherical, diagonal, and ellipsoidal, each with different restrictions on volume, shape, and orientation. The choice of the covariance matrix model significantly affects the number of parameters to be estimated.

The BIC is used for model selection, recommending a model with the most favorable BIC value, which balances the goodness-of-fit and model complexity. It recommends a specific number of segments and covariance matrix structures for the best model. The passage includes visualizations to illustrate the fitted models, such as classification plots and uncertainty plots, which help assess the quality of segment assignments. In practical scenarios with empirical data, it is not always straightforward to assess the quality of model recommendations made by information criteria like the BIC.

Binary Distributions

Metric data refers to data with continuous values. Multivariate normal distributions are commonly used for segmenting such data. This distribution is suitable for data with multiple variables and accounts for covariance between them. It is illustrated using an artificial mobile phone data set, and the 'mclust' package in R is recommended for fitting models for different numbers of segments. The number of parameters to estimate for a mixture of normal distributions with p segmentation variables is $p + p(p + 1)/2$, making it essential to have large sample sizes for reliable estimates.

To simplify the estimation process, restrictions on the covariance matrices can be imposed, such as using spherical covariances. Binary data refers to data with two possible values (0 and 1). A mixture of binary distributions, also known as latent class models, is commonly used to segment binary data. It assumes that respondents in different segments have different probabilities of exhibiting certain behaviors or binary outcomes. For example, tourists' preferences for vacation activities could be represented as binary variables (e.g., 1 for choosing the activity and 0 for not).

The use of the 'flexmix' package in R is recommended to fit mixture models, and the EM algorithm is used with multiple random starts for robustness. Information criteria like AIC, BIC, and ICL are used to select the best-fitting model.

Example with Winter Vacation Activities: The passage walks through an example where a mixture of binary distributions is fitted to a dataset of winter activities of tourists. The analysis involves varying the number of segments and selecting the optimal number based on information criteria. A five-segment solution is chosen for further examination. A segment profile plot is generated to visualize the segment characteristics and behaviors of tourists.

Finite Mixture of Regressions

This technique is used when dealing with binary data, where data points are represented as 0s and 1s. It's applied to scenarios like vacation activities, where 0 represents non-participation, and 1 represents participation in a particular activity. It is used to identify segments of respondents with varying probabilities of engaging in specific activities, allowing for an association between binary variables within segments, which might not be present in the overall data.

In this method, the dependent target variable (e.g., willingness to pay for a theme park) is explained by independent variables (e.g., the number of rides available). It is assumed that different market segments have different regression relationships between the dependent and independent variables. The example illustrates how this approach can uncover distinct segments within a dataset with varying linear and non-linear associations.

Extensions and Variations

Finite mixture models are more complex than distance-based clustering methods. This complexity allows them to be highly flexible and adaptable. They can accommodate a wide range of data characteristics and types. They can be applied to various data types, making them suitable for different scenarios: For metric data, mixtures of normal distributions can be used. For binary data, mixtures of binary distributions are applicable. For nominal variables, mixtures of multinomial distributions or multinomial logit models can be employed. For ordinal variables, several models can be used to create mixtures, addressing the complexities of ordinal data.

In the context of ordinal variables, finite mixture models can help address response style effects. These effects can be disentangled from content-specific responses when extracting market segments. This allows for more accurate segmentation while accounting for individual response styles. Finite mixture models, when combined with conjoint analysis, can account for differences in preferences among consumers. This is particularly useful in market research. The debate in the segmentation literature is about whether consumer differences should be modeled using a continuous distribution or distinct market segments. An extension to mixture models, known as a mixture of mixed-effects models or heterogeneity model, can reconcile these positions. It acknowledges the existence of distinct segments while allowing for variation within each segment.

When dealing with data containing repeated observations over time, mixture models can be used to cluster time series and extract groups of similar consumers. Markov chains and dynamic latent change models can be employed to track changes in brand choices and buying decisions over time. Finite mixture models can incorporate both segmentation variables and descriptor variables. Segmentation variables are used for grouping and are included in segment-specific models. Descriptor variables help model differences in segment sizes, assuming that segments differ in their composition concerning these descriptor variables. These descriptor variables are known as concomitant variables and can be included in the analysis.

Algorithms with Integrated Variable Selection

In some cases, segmentation variables in the data may not all contribute meaningfully to the segmentation solution. These variables might contain redundant or noisy information, which can negatively impact the accuracy of the segmentation process. Preprocessing methods can help identify and select the most relevant segmentation variables. One such approach is the filtering approach proposed by Steinley and Brusco (2008a). It evaluates the clusterability of individual variables and includes only those that meet a certain threshold. This approach is effective for metric variables.

When dealing with binary segmentation variables, the problem of variable selection becomes more challenging. Binary variables might not provide meaningful information for clustering, making it difficult to pre-screen variables individually. To address the challenges of binary segmentation variables, some algorithms are designed to extract segments while simultaneously selecting suitable segmentation variables. Two such algorithms are mentioned:

Biclustering is an approach that identifies subsets of variables that are relevant for specific clusters or segments. It helps uncover which variables contribute to the segmentation. An alternative approach is discussed, where segmentation variables are compressed into factors before the segment extraction process. This two-step approach involves dimensionality reduction before applying clustering algorithms, which can help improve the quality of segmentation.

Biclustering Algorithms

Biclustering is a method that clusters both consumers and binary segmentation variables. It aims to extract market segments in which consumers share specific binary variable patterns, forming biclusters. The concept of biclustering is not new and has been used in various fields. However, its popularity increased with the emergence of large-scale genetic and proteomic data, where traditional clustering algorithms were not suitable due to the presence of many non-functional genes. Several biclustering algorithms exist, and they differ in how they define biclusters. In the simplest case, a bicluster is defined as a set of consumers (rows) with binary value 1 for a subset of variables (columns). The goal is to identify large groups of consumers who share as many activities as possible. The biclustering algorithm typically involves the following steps:

Step 1: Rearrange rows and columns in the data matrix to create a large rectangle with identical 1s.

Step 2: Assign the consumers within this rectangle to one bicluster.

Step 3: Remove assigned consumers from the data matrix and repeat the procedure until no more biclusters of sufficient size can be identified.

Control Parameters: Biclustering algorithms have control parameters, such as the minimum number of observations and variables required to form a bicluster. Biclustering is useful in market segmentation when dealing with a high number of segmentation variables. It doesn't require data transformation, preserving the original information. It can capture niche markets by allowing control over the level of matching required for segment formation.

Example - Australian Vacation Activities: An example of using the repeated Bimax algorithm for biclustering is presented. The data includes binary information about vacation activities for Australian tourists. Biclustering identifies segments of tourists based on their activity patterns.

The visualization results of biclustering, including a bicluster membership plot that displays market segments and the similarity of consumers within each segment regarding specific vacation activities. It also provides information about how distinct members of a segment are from the average consumer with respect to specific activities.

Variable Selection Procedure for Clustering Binary Data (VSBD)

The VSBD method is designed for clustering binary datasets. It assumes that not all variables are relevant for obtaining a good clustering solution and aims to identify and remove irrelevant variables, specifically masking variables. This variable selection process helps in achieving an accurate segment structure and makes interpretation easier. The VSBD procedure involves several steps to identify the best subset of variables for segment extraction. It is based on the k-means algorithm and uses the within-cluster sum-of-squares as the performance criterion, which is minimized by the k-means algorithm.

Steps in the VSBD Algorithm:

Step 1: Select a subset of observations, typically a fraction (φ) of the original data size. The choice of φ depends on the dataset size.

Step 2: Conduct an exhaustive search for the set of V variables (a small number) that results in the smallest within-cluster sum-of-squares. The value of V should be chosen carefully based on the number of clusters and variables.

Step 3: Identify the variable that, when added to the segmentation variables, leads to the smallest increase in within-cluster sum-of-squares.

Step 4: Add the variable if the increase in within-cluster sum-of-squares is smaller than a specified threshold (δ).

The algorithm requires parameters like φ, V, and δ to be specified. The number of clusters (k) also needs to be predetermined, and the Ratkowsky and Lance index is recommended to select an appropriate k.

Example - Australian Travel Motives: The passage provides an example of applying the VSBD algorithm to the Australian Travel Motives dataset. The example uses a k value of 6 and specific settings for φ, V, and δ.

The VSBD algorithm selects a subset of relevant variables from the original dataset. In this example, only 6 out of 20 variables are chosen. Using the selected variables, the algorithm creates a segmentation solution. The segments represent groups of individuals who share specific travel motives. The segments are interpreted based on the selected variables. The selection of variables using the VSBD algorithm results in an interpretable segmentation solution. For each segment, the chosen variables effectively differentiate between the segments, making the interpretation straightforward.

Variable Reduction: Factor-Cluster Analysis

Factor-cluster analysis is a two-step approach in market segmentation. In the first step, the segmentation variables are subjected to factor analysis. After factor analysis, the raw data, which are the original segmentation variables, are discarded. Factor-cluster analysis can be conceptually legitimate in cases where the empirical data result from validated psychological test batteries specifically designed with variables that load onto factors. For example, IQ tests contain items assessing general knowledge, and replacing original variables with factor scores for general knowledge can be justified. However, the factor scores should either be determined simultaneously with group extraction or be provided separately, not determined from the data itself. It is often used when the number of original segmentation variables is too high in relation to the sample size. A rule of thumb is that the sample size should be at least 100 times the number of segmentation variables, which can be challenging to achieve in practice.

One significant drawback of factor analysis is the substantial loss of information. When segmentation variables are factor analyzed, a portion of the data's variability is sacrificed, which can impact the quality of the segmentation. Factor analysis transforms the data, meaning segments are extracted from a modified version of the consumer data, not the original data. This transformation can change the nature of the data before segment extraction, affecting the quality of the results. Factor-cluster results are more challenging to interpret because they are based on factor space, which lacks concrete meaning. Profiling segments using factor-based data is not as straightforward as using the original variables, making practical recommendations for marketing mix adjustments more difficult. Empirical evidence suggests that factor-cluster analysis does not outperform cluster analysis using raw data. Studies have shown that, even when data is generated based on a factor-analytic model, factor-cluster analysis does not consistently identify the correct market segment structure in the data.

Data Structure Analysis

Extracting market segments is inherently exploratory, regardless of the segmentation algorithm used. Traditional validation methods with clear optimality criteria are not feasible because organizations cannot simultaneously implement multiple segmentation strategies to determine the most profitable or successful one. In the context of market segmentation, the term "validation" typically refers to assessing the reliability or stability of segmentation solutions. This involves repeatedly calculating segments with slight data modifications or algorithm variations. It is fundamentally different from validation using external criteria. Validation is often achieved through stability-based data structure analysis. This approach assesses the stability of segmentation solutions when data or algorithms are modified. It provides insights into the properties of the data and helps determine whether natural, distinct, and well-separated market segments exist. Data structure analysis offers valuable insights that guide methodological decisions in market segmentation. It helps determine if meaningful segments exist in the data and assists in selecting the appropriate number of segments to extract.

Cluster Indices

Market segmentation analysis is an exploratory process that requires data analysts to make critical decisions, such as selecting the number of market segments to extract. To aid in these decisions, cluster indices are commonly used. Cluster indices can be categorized into two main groups: internal cluster indices and external cluster indices. These indices are calculated based on a single market segmentation solution. They use information from that segmentation to offer insights. An example of an internal cluster index is the sum of distances between pairs of segment members. A lower value indicates that members within the same segment are more similar, making such segments attractive.

These indices require two segmentation solutions for calculation, making them external to a single solution. They measure the similarity between two different segmentation solutions. To use external cluster indices, the correct market segmentation needs to be known, which is often only possible with artificially generated data. In consumer data analysis, there is no definitive "correct" assignment of members to segments. In such cases, the analysis can be repeated, and the second solution can serve as the additional input for calculating external cluster indices. The commonly used measures of similarity for comparing two market segmentation solutions include the Jaccard index, the Rand index, and the adjusted Rand index. These measures help assess the extent to which two different segmentation solutions produce similar segment assignments.

Internal Cluster Indices

Internal cluster indices are used to evaluate the quality of a single segmentation solution. They assess the compactness and separation of segments within this solution. One common internal cluster index measures the compactness of segments by calculating the sum of distances between each segment member and their segment representative. This sum of within-cluster distances (Wk) is used to evaluate the segmentation solution. The k-means algorithm often leads to a monotonically decreasing Wk as the number of segments (k) increases. To help select the number of market segments in k-means clustering, a scree plot is commonly used. This plot shows the Wk values for different numbers of segments, and an "elbow" in the plot, indicating a point where Wk changes significantly, can guide the choice of the optimal number of segments. A variation of the internal cluster index for compactness is the Ball-Hall index (Wk/k), which aims to correct for the monotonous decrease of the internal cluster index with increasing numbers of segments.

In addition to compactness, it's important to evaluate the dissimilarity and separation between segments. An internal cluster index based on the weighted distances between centroids (Bk) captures the separation aspect. It assesses the difference between segments and their similarity to consumers. Internal cluster indices can combine both compactness (Wk) and separation (Bk) aspects. By relating these two values in different ways, data analysts can choose a suitable number of segments. The Ratkowsky and Lance index is recommended for variable selection procedures, and it is based on squared Euclidean distance. It calculates the sum of squares between segments for each variable and averages these ratios, then divides by the square root of the number of segments. The number of segments with the maximum index value is selected.

While calculating internal cluster indices is valuable and cost-effective, they may not provide clear guidance when working with consumer data, which often lacks naturally occurring market segments. In such cases, external cluster indices and stability analysis are recommended to complement the analysis.

External Cluster Indices

External cluster indices evaluate a market segmentation solution using additional external information beyond the data used for the analysis. The most valuable additional information is the true segment structure, but this is typically known only for artificially generated data. When comparing two segmentation solutions, a problem called "label switching" arises. This occurs because segment labels are arbitrary and can differ between solutions. To address this, comparisons focus on whether pairs of consumers are assigned to the same segments repeatedly, rather than segment labels.

The Jaccard index, proposed by Jaccard in 1912, assesses the similarity between two segmentation solutions. It considers pairs of consumers that are assigned to the same segments in both solutions. The index ranges between 0 (completely different solutions) and 1 (identical solutions). The Rand index, proposed by Rand in 1971, is similar to the

Jaccard index but includes all four values: a, b, c, and d. Like the Jaccard index, it ranges between 0 and 1. Both the Jaccard and Rand indices have the challenge of interpreting their absolute values, as the minimum values depend on the size of the market segments in the solutions. Applying the general correction to the Rand index results in the adjusted Rand index. It addresses the problem of chance agreement and provides a corrected measure of similarity between segmentation solutions.

Gorge Plots

Similarity values (sih) are used to measure the similarity of each consumer to the representative (centroid or cluster center) of a particular market segment. These values range between 0 and 1 and sum to 1 for each consumer over all segment representatives. The hyperparameter γ controls how differences in distance between consumers and segment representatives translate into differences in similarity. It allows for fine-tuning the similarity measure. Gorge plots are used to visualize similarity values. They contain histograms of similarity values (sih) for each segment. The x-axis represents similarity values, and the y-axis shows the frequency of occurrence. High similarity values indicate that a consumer is close to the segment representative, while low values suggest the consumer is far from the representative. They help assess the separation of segments. In ideal cases, where well-separated segments exist in the data, the plot resembles a gorge with distinct peaks to the left and right. The plot shape reflects the degree of separation among consumers in different segments.

Gorge plots are for three artificial data sets representing different types of market segmentation: natural, reproducible, and constructive. In the natural clustering case, a clear gorge plot indicates well-separated segments, while the other cases show less distinct gorge plots with consumers scattered across the similarity range. In real market segmentation analysis, gorge plots should be generated and inspected for different numbers of segments. This process can be time-consuming and may not account for randomness in the sample. To overcome the limitations of generating gorge plots for various segment numbers, stability analysis is introduced. This analysis can be conducted at both the global and segment levels and helps assess the stability and quality of segmentation solutions.

Global Stability Analysis

Resampling methods are used to assess the stability of a market segmentation solution by generating new data sets and extracting multiple segmentation solutions. The stability of these solutions is compared, and the one that can best be replicated is chosen. Several resampling methods have been proposed by different researchers, such as Breckenridge, Dudoit and Fridlyand, Grün and Leisch, Lange et al., Tibshirani and Walther, Gana Dresen et al., and Maitra et al. Real consumer data rarely contain distinct, well-separated market segments like artificial data. In the worst case, consumer data can be entirely unstructured, making segmentation challenging.

- Conceptually, consumer data can be categorized into three groups:
- Natural, well-separated segments exist and are easy to identify.
- Data is unstructured, making it impossible to reproduce segmentation.
- Data lacks distinct natural clusters but is not entirely unstructured. In this case, reproducible segmentation is possible.

Global stability analysis is used to determine which category a given data set falls into. It acknowledges the role of sample randomness and algorithm choice in segmentation. Multiple computations are necessary to account for randomness. Bootstrapping is recommended to assess global stability. It involves drawing multiple bootstrap samples from the original data and calculating replicate segmentation solutions for different numbers of segments. The adjusted Rand index or other external cluster indices are used to evaluate the similarity between the replication solutions. Boxplots are created to assess the global reproducibility of segmentation solutions. The passage provides guidelines for interpreting global stability boxplots based on their characteristics, which indicate whether natural segments, reproducible segments, or constructive segments are present. While global stability assesses the stability of the entire segmentation solution, the stability of individual segments within a solution is crucial for practical use. The next section discusses segment level stability.

Segment Level Stability Analysis

Selecting the best overall segmentation solution based on global stability analysis does not necessarily guarantee that it contains the single best market segment. Relying solely on global stability could result in choosing a segmentation solution that is globally stable but lacks individual segments with high stability. To avoid discarding solutions containing valuable individual segments prematurely, it's advisable to assess both the global stability of alternative segmentation solutions and the stability of individual market segments within those solutions. This approach ensures that valuable target segments are not overlooked since most organizations typically need to identify a single target segment for their marketing efforts.

Segment Level Stability Within Solutions (SLSw)

Dolnicar and Leisch (2017) propose a method for assessing segmentation solutions based on segment level stability within solutions (SLSW). This approach focuses on the stability of individual segments within a segmentation solution, rather than the overall solution. The aim is to prevent valuable individual segments from being discarded when choosing a segmentation solution, as many organizations typically target only one segment to secure their survival and competitive advantage. It is similar to global stability, but it calculates stability at the segment level. This means it can detect a highly stable segment within a segmentation solution even when other segments may be unstable.

The method involves the following steps:

- Compute a segmentation solution, extracting k segments using a chosen algorithm.
- Draw bootstrap samples from the original data.
- Cluster all bootstrap samples into k segments and assign observations to these segments.
- Calculate the maximum agreement between the original segments and bootstrap segments using the Jaccard index.
- Create boxplots of the agreement values to assess the segment level stability within solutions (SLSW).

This approach is illustrated using both an artificial mobile phone dataset and a real-world dataset containing Australian travel motives. For the artificial dataset, three distinct and well-separated segments exist, leading to high SLSW values when extracting three segments. However, for datasets without natural segments, this method helps identify stable segments even within unstable solutions.

Segment level stability within solutions (SLSW) analysis is valuable because it ensures that individual segments, which may be highly valuable to an organization, are not disregarded based solely on the stability of the overall segmentation solution. This approach is especially critical for multi-dimensional data where understanding the data structure is complex.

Segment Level Sability across Solutions (SLSa)

The second criterion proposed by Dolnicar and Leisch (2017) for assessing segmentation solutions is referred to as Segment Level Stability Across Solutions (SLSA). This criterion aims to determine the re-occurrence of market segments across different segmentation solutions that contain varying numbers of segments. High SLSA values indicate that a market segment occurs naturally in the data, rather than being artificially constructed. Natural segments are more attractive to organizations because they represent actual consumer groupings, eliminating the need for managerial judgment in creating segments. The SLSA analysis involves a series of m partitions (segmentation solutions), denoted as P1, P2, ..., Pm, with different numbers of segments (ranging from kmin to kmax), where m = kmax - kmin + 1. The user, in collaboration with the data analyst, specifies the minimum (kmin) and maximum (kmax) number of segments of interest.

SLSA can be calculated in combination with any algorithm used to extract segments. However, it is particularly useful for hierarchical clustering because it reflects the creation of a sequence of nested partitions. When using methods like k-means, k-medians, neural gas, or finite mixture models, where segment labels depend on random initialization, it's crucial to identify similar segments across solutions with different numbers of segments and assign consistent labels. This process is facilitated by relabeling the segments based on similarity. SLSA plot is created to visually represent the stability and changes in segments as the number of segments in the solution varies. In this plot, each column represents a segmentation solution with a specific number of segments, and lines between segments indicate the movement of segment members. Thick lines indicate stable, naturally occurring segments.

Entropy, a measure of uncertainty in a distribution, is used as a numeric indicator of SLSA. High entropy values represent less stability, while low entropy indicates high stability. The SLSA values are used to color the nodes and edges in the SLSA plot, providing a visual representation of segment stability. The usefulness of SLSA is demonstrated with examples using an artificial mobile phone dataset and a real-world dataset of Australian travel motives. In the artificial mobile phone dataset, SLSA shows that the high-end mobile phone segment remains stable across different segmentation solutions, while other segments are split into subsegments. In the Australian travel motives dataset, segments with high average SLSA values are identified as attractive target segments. For instance, segment 6 in the six-segment solution is stable across different solutions and represents a potentially natural market segment.

Step 5 Checklist

- 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.

6. Profiling Segments

Identifying Key Characteristics of Market Segments

The primary purpose is to understand the characteristics of the market segments resulting from the data-driven segmentation approach. Profiling is necessary when conducting data-driven market segmentation. In contrast, for commonsense segmentation, where predefined characteristics are used, profiling is not required. When using data-driven segmentation, the defining characteristics of the segments are initially unknown until the data analysis is completed. Profiling aims to identify and characterize these defining characteristics of the market segments, primarily with respect to the segmentation variables. It involves characterizing the market segments individually and comparing them to one another. For example, if winter tourists in Austria are surveyed about their vacation activities, many may indicate that they are going alpine skiing. While alpine skiing can characterize a segment, it may not be sufficient to differentiate that segment from others.

In the profiling stage, several alternative market segmentation solutions are examined, which is especially important when there are no natural segments in the data. Good profiling is essential for the accurate interpretation of the segmentation results, and this correct interpretation is crucial for making effective strategic marketing decisions. Data driven market segmentation solutions can be challenging to interpret, and many managers face difficulties in understanding them. Studies have shown that marketing managers often find segmentation results presented to them in a way that can be contradictory, lack a clear executive summary, or appears as a "black box." Graphical statistical approaches to segment profiling are discussed as a means to make profiling less tedious and less prone to misinterpretation.

Traditional Approaches to Profiling Market Segments

Data-driven segmentation solutions are typically presented to users, such as managers or clients, in one of two ways. The first is high-level summaries that oversimplify segment characteristics to the point of being misleadingly trivial. The second involves large tables that provide exact percentages for each segmentation variable within each segment. Using tables to interpret segment profiles can be challenging. These tables present exact percentages of segment members for each segmentation variable and require comparisons to identify defining characteristics. This process involves comparing each segment's values to the total sample and also to values of other segments. For a table with 20 rows and six segments, this can lead to a large number of comparisons (420 in total). When multiple alternative segmentation solutions are presented (e.g., five solutions with six segments each), users would have to make 2100 pairs of number comparisons to understand the defining characteristics of the segments. This can be a highly tedious and challenging task.

In some cases, information about the statistical significance of differences between segments for each segmentation variable is provided. However, this approach is not statistically correct, as segment membership is directly derived from the segmentation variables, and segments are created to be maximally different, making standard statistical tests inappropriate for assessing differences.

Segment Profiling with Visualizations

Traditional methods of presenting market segmentation solutions rely on highly simplified or very complex tabular representations. These representations often lack the utilization of graphics, despite the significance of data visualization in statistical data analysis. Graphics are particularly valuable in exploratory statistical analysis, such as cluster analysis, as they provide insights into the complex relationships between variables. Visualizations offer a means to monitor developments over time, especially in the era of big data. Prominent experts in the field, such as McDonald and Dunbar and Lilien and Rangaswamy, recommend the use of visualization techniques to enhance the interpretability of market segmentation analysis results. Graphical representations can convey the same information in a more insightful and intuitive manner compared to tabular forms. They simplify the understanding of complex data relationships.

Visualizations are particularly useful in the data-driven market segmentation process for inspecting individual segments within each segmentation solution. They aid in interpreting segment profiles and assessing the suitability of a given segmentation solution. As the process often results in numerous alternative solutions, visualizations help analysts and users make critical decisions when selecting the most appropriate solution.

Identifying Defining Characteristics of Market Segments

A segment profile plot visually displays how each market segment differs from the overall sample across all segmentation variables. It is a graphical representation of tabular data, such as that found in Table 8.1. The order of segmentation variables in figures and tables does not necessarily have to match their order in the original dataset. Variables can be rearranged for better visualization. One approach is to cluster variables based on similarity in response patterns. Variables can be reordered using hierarchical clustering techniques. By clustering the columns of the data matrix and visualizing the results, patterns of variable similarity become apparent. Marker variables are those that significantly deviate from the overall mean, indicating their importance in characterizing a segment. A threshold (e.g., a 0.25 absolute difference or 50% relative difference) can be used to identify marker variables. Segment profile plots typically contain panels, with each panel representing one segment. They display the cluster centers (centroids) for each segment, which are analogous to the values in the table. The plot also includes dots representing total mean values across all observations.

Segment profile plots, especially when marker variables are highlighted in color, are easier and faster to interpret than traditional tabular presentations of segmentation results. They provide a visual summary of segment characteristics, making it simpler for analysts and managers to understand the data. An eye-tracking study is mentioned, comparing the interpretation of segmentation results presented in tables, improved tables, and segment profile plots. The study found that segment profile plots required less effort and were easier to interpret. Well-designed visualizations, such as segment profile plots, can significantly aid managers in making long-term strategic decisions based on segmentation results. They offer a higher return on investment as they facilitate quicker and more efficient decision-making.

Assessing Segment Separation

Segment separation plots are used to visualize the degree of overlap or separation between market segments in multi-dimensional data space. These plots can be relatively simple when there are few segmentation variables but become more complex as the number of variables increases. Nevertheless, they provide data analysts and users with a quick overview of the segmentation solution.

In cases where there are many segmentation variables, projection techniques are used to reduce the data to a lower number of dimensions for plotting. Principal components analysis (PCA) is one such method. PCA allows the visualization of complex high-dimensional data by projecting it onto a smaller number of dimensions. The segment separation plots are used to assess the degree of separation or overlap between market segments. By examining the plots, analysts can make inferences about the distinctiveness of segments based on the chosen projection. The readability of segment separation plots can be improved by using different colors to distinguish segments and modifying plot elements. For example, it is possible to omit data points and focus solely on cluster hulls, making the plots cleaner and easier to interpret.

Segment separation plots help analysts identify whether segments are well-separated or overlap in the chosen projection. This information is valuable for understanding the distinctiveness of market segments. It's important to note that different projections can yield different results, and one projection does not represent the full range of possibilities. The choice of projection can influence the interpretation of segment separations.

Step 6 Checklist

- Use the selected segments from Step 5.
- Visualize segment profiles to learn about what makes each segment distinct.
- Use knock-out criteria to check if any of the segments currently under consideration should already be eliminated because they do not comply with the knock-out criteria.
- Pass on the remaining segments to Step 7 for describing.

McDonalds Case Study	GitHub: link